3

STATISTIQUES DESCRIPTIVES : MÉTHODES NUMÉRIQUES

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

STATISTIQUES APPLIQUÉES

Small Fry Design *
Santa Ana, Californie

Fondé en 1997, Small Fry Design est une société de jouets et accessoires qui crée et importe des produits pour enfants. La gamme de produits de la société comprend des ours en peluche, des mobiles, des jouets musicaux, des hochets et des doudous ; les jouets sont de très bonne qualité et une attention particulière est accordée à la couleur, à la texture et au son des objets. Les produits sont créés aux États-Unis et fabriqués en Chine.

Small Fry Design engage des représentants indépendants pour vendre ses produits à des détaillants de fournitures infantiles, à des magasins d'habillement et d'accessoires pour enfants, à des boutiques de cadeaux, aux grands magasins haut de gamme et aux principales sociétés de vente par correspondance. Actuellement, les produits Small Fry Design sont distribués dans plus de 1 000 points de vente à travers les États-Unis.

La gestion des liquidités est l'une des activités les plus importantes dans l'exploitation quotidienne de cette entreprise. La différence entre un succès et un échec commercial peut reposer sur la présence d'un flux de liquidités suffisant pour rembourser les dettes présentes et futures. Un facteur important dans la gestion des liquidités est l'analyse et le contrôle des créances. En estimant l'échéance moyenne et la valeur des factures impayées, les gestionnaires peuvent prévoir les disponibilités en liquidité. La société a fixé les objectifs suivants : l'échéance moyenne des impayés ne doit pas dépasser 45 jours et la valeur des impayés de plus de 60 jours ne doit pas dépasser 5 % de la valeur de toutes les créances.

Une étude récente des créances a fourni les statistiques suivantes concernant le délai de recouvrement des factures :

Moyenne	40 jours
Médiane	35 jours
Mode	31 jours

Selon ces statistiques, le délai moyen de recouvrement d'une facture est de 40 jours. La médiane indique que la moitié des factures restent impayées pendant au moins 35 jours. Le mode, c'est-à-dire le délai de recouvrement des factures le plus fréquent, est de 31 jours. Le résumé statistique révèle également que seulement 3 % de la valeur des comptes clients restent impayés pendant plus de 60 jours. Sur la base de cette information statistique, la direction se déclarait satisfaite du contrôle des créances et du flux de liquidité.

Dans ce chapitre, vous apprendrez à calculer et interpréter quelques mesures statistiques utilisées par Small Fry Design. En plus de la moyenne, de la médiane et du mode, vous vous familiariserez avec d'autres statistiques descriptives telles que l'étendue, la variance, l'écart type, les percentiles et la corrélation. Ces mesures numériques sont essentielles pour la compréhension et l'interprétation des données.

* Les auteurs remercient John A. McCarthy, président de Small Fry Design, de leur avoir fourni ce Statistiques Appliquées.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Dans le chapitre 2, nous avons discuté des méthodes graphiques et sous forme de tableaux utilisées pour résumer des données. Dans ce chapitre, nous présentons plusieurs méthodes numériques de statistiques descriptives qui permettent également de résumer les données.

Nous commencerons par présenter des méthodes numériques pour résumer des ensembles de données d'une seule variable. Lorsqu'un ensemble de données contient plus d'une variable, des mesures numériques similaires peuvent être calculées séparément pour chaque variable. Cependant dans le cas de deux variables, nous développerons également des mesures de la relation entre les variables.

Nous introduirons des mesures de tendance centrale, de dispersion, nous examinerons la forme des distributions et la relation entre les variables. Si les mesures sont calculées à partir de données issues d'un échantillon, on parle de **statistiques d'échantillon**. Si les mesures sont calculées à partir de données issues d'une population, on parle de **paramètres de la population**. En inférence statistique, une statistique d'échantillon est qualifiée d'**estimateur ponctuel** du paramètre de la population correspondant. Dans le chapitre 7, nous discuterons de façon plus détaillée du processus d'estimation ponctuelle.

Dans les trois annexes de ce chapitre, nous montrerons comment utiliser Minitab, Excel et StatTools pour calculer de nombreuses statistiques descriptives numériques décrites dans ce chapitre.

3.1 MESURES DE TENDANCE CENTRALE

3.1.1 Moyenne

La **moyenne**, ou valeur moyenne, est peut-être la mesure de tendance centrale la plus importante pour une variable. Si les données sont issues d'un échantillon, la moyenne est notée $\overline{x}$; si les données sont issues d'une population, la moyenne est notée μ.

La moyenne est parfois qualifiée de moyenne arithmétique.

En langage statistique, il est fréquent de noter la valeur de la première observation de la variable x x_1, la valeur de la deuxième observation x_2 et ainsi de suite. De façon générale, la valeur de la i^e observation est notée x_i. Pour un échantillon de n observations, la formule de la moyenne de l'échantillon est la suivante.

► **Moyenne d'échantillon**

$$\overline{x} = \frac{\sum x_i}{n} \tag{3.1}$$

La moyenne d'échantillon $\overline{x}$ est une statistique d'échantillon.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Dans la formule précédente, le numérateur correspond à la somme des valeurs des *n* observations. C'est-à-dire,

$$\sum x_i = x_1 + x_2 + \ldots + x_n$$

La lettre grecque $\sum$ est le signe somme.

Pour illustrer le calcul d'une moyenne d'échantillon, considérons les données suivantes relatives au nombre d'élèves d'un échantillon de cinq classes.

46 54 42 46 32

Nous utilisons les notations x_1, x_2, x_3, x_4, x_5 pour représenter le nombre d'élèves dans chacune des cinq classes.

$$x_1 = 46 \quad x_2 = 54 \quad x_3 = 42 \quad x_4 = 46 \quad x_5 = 32$$

Par conséquent, pour calculer la moyenne de l'échantillon, on peut écrire

$$\overline{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 54 + 42 + 46 + 32}{5} = 44$$

La taille moyenne des classes de l'échantillon est de 44 élèves.

Pour avoir une représentation graphique de la moyenne et montrer comment elle peut être influencée par des valeurs extrêmes, considérez le diagramme de points obtenu à partir des données sur la taille des classes, représenté à la figure 3.1. En considérant l'axe horizontal utilisé pour créer le diagramme de points comme une longue planche étroite sur laquelle chaque point a le même poids, la moyenne correspond au point d'appui qui permet de maintenir la planche en équilibre. Il s'agit du même principe que celui grâce auquel fonctionne une balançoire dans un jardin public, la seule différence étant que le point d'appui de la balançoire est situé au milieu de façon à ce que lorsque l'un se trouve en haut, l'autre se trouve en bas. Sur le diagramme de points, nous avons situé le point pivot en fonction de la localisation des points. Maintenant, imaginez ce qui se passerait si nous augmentions la valeur la plus élevée de 54 à 114. Nous devrions alors déplacer le point d'appui vers la droite pour rééquilibrer le diagramme de points. Pour déterminer jusqu'où déplacer le point d'appui, nous calculons simplement la moyenne d'échantillon avec les données révisées sur les tailles de classes.

$$\overline{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 114 + 42 + 46 + 32}{5} = \frac{280}{5} = 56$$

Ainsi, la moyenne pour les données révisées relatives à la taille des classes est de 56, soit 12 étudiants supplémentaires. En d'autres termes, nous devons déplacer le point d'équilibre de 12 unités vers la droite pour rétablir l'équilibre sous le nouveau diagramme de points.

L'exemple suivant est une autre illustration du calcul d'une moyenne d'échantillon. Supposez que le conseiller d'orientation d'un collège ait envoyé un questionnaire à un échantillon de diplômés d'une école de commerce afin de connaître leur salaire au début de leur carrière. Le tableau 3.1 regroupe les données collectées (fichier en ligne Salaire

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Tableau 3.1 *Salaire mensuel de départ d'un échantillon de 12 diplômés d'une école de commerce*

Salaire de départ 2012

Diplômé	Salaire mensuel de départ ($)
1	3850
2	3950
3	4050
4	3880
5	3755
6	3710
7	3890
8	4130
9	3940
10	4325
11	3920
12	3880

de départ 2012). La moyenne du salaire mensuel initial d'un échantillon de 12 diplômés d'une école de commerce est égale à

$$\overline{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \ldots + x_{12}}{12} = \frac{3\,850 + 3\,950 + \ldots + 3\,880}{12} = \frac{47\,280}{12} = 3\,940$$

La formule (3.1) illustre la manière dont la moyenne est calculée pour un échantillon de n observations. La formule pour calculer la moyenne d'une population est identique, mais les notations utilisées sont différentes, pour indiquer que nous travaillons avec la population entière. Le nombre d'observations dans une population est N et le symbole pour la moyenne d'une population est μ.

► **Moyenne de la population**

$$\mu = \frac{\sum x_i}{N} \tag{3.2}$$

La moyenne d'échantillon $\overline{x}$ est un estimateur ponctuel de la moyenne de la population μ.

3.1.2 Moyenne pondérée

Dans les formules de calcul de la moyenne d'un échantillon ou d'une population, chaque observation x_i a la même importance ou la même pondération. Par exemple, la formule de la moyenne d'un échantillon peut se réécrire de la façon suivante :

$$\overline{x} = \frac{\sum x_i}{n} = \frac{1}{n}\left(\sum x_i\right) = \frac{1}{n}(x_1 + x_2 + x_3 + \ldots + x_n) = \frac{1}{n}(x_1) + \frac{1}{n}(x_2) + \ldots + \frac{1}{n}(x_n)$$

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Cela montre que chaque observation de l'échantillon est pondérée par $1/n$. Bien que cette pratique soit la plus courante, dans certaines situations, la moyenne est calculée en donnant à chaque observation une pondération qui reflète son importance. Une moyenne calculée de cette manière est appelée **moyenne pondérée**. La moyenne pondérée est calculée de la façon suivante :

► **Moyenne pondérée**

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \tag{3.3}$$

où
w_i correspond à la pondération de l'observation i

Lorsque les données sont issues d'un échantillon, la formule (3.3) fournit la moyenne pondérée de l'échantillon. Lorsque les données sont issues d'une population, $\bar{x}$ est remplacé par μ et la formule (3.3) fournit la moyenne pondérée de la population.

Pour illustrer le calcul d'une moyenne pondérée, considérons l'échantillon suivant relatif à cinq achats de matière première au cours des trois derniers mois.

Achat	Coût par livre ($)	Nombre de livres
1	3,00	1 200
2	3,40	500
3	2,80	2 750
4	2,90	1 000
5	3,25	800

Notez que le coût par livre varie entre 2,80 et 3,40 dollars, et que la quantité achetée varie entre 500 et 2 750 livres. Supposons qu'un responsable veuille obtenir des informations sur le coût moyen par livre de matière première. Puisque les quantités commandées varient, nous devons utiliser la formule d'une moyenne pondérée. Les cinq valeurs des observations sur le coût par livre sont $x_1 = 3{,}00$, $x_2 = 3{,}40$, $x_3 = 2{,}80$, $x_4 = 2{,}90$ et $x_5 = 3{,}25$. Le coût moyen pondéré, par livre, est obtenu en pondérant chaque coût par la quantité correspondante. Dans cet exemple, les pondérations sont $w_1 = 1\,200$, $w_2 = 500$, $w_3 = 2\,750$, $w_4 = 1\,000$ et $w_5 = 800$. En utilisant la formule (3.3), la moyenne pondérée est égale à :

$$\bar{x} = \frac{1\,200(3{,}00) + 500(3{,}40) + 2\,750(2{,}80) + 1\,000(2{,}90) + 800(3{,}25)}{1\,200 + 500 + 2\,750 + 1\,000 + 800} = \frac{18\,500}{6\,250} = 2{,}96$$

Ainsi le calcul de la moyenne pondérée révèle que le coût moyen par livre de matière première est égal à 2,96 dollars. Notez que l'utilisation de la formule (3.1) au lieu de la formule de la moyenne pondérée aurait fourni des résultats erronés. Dans ce cas, la moyenne des cinq observations sur le coût par livre est égale à $(3{,}00 + 3{,}40 + 2{,}80 + 2{,}90 + 3{,}25)/5 = 15{,}35/5 = 3{,}07$ dollars, ce qui surestime le coût moyen par livre réel.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Le choix des pondérations dans le calcul d'une moyenne pondérée particulière dépend de l'étude. Un exemple bien connu des étudiants américains est le calcul de la moyenne des notes. Dans ce calcul, les valeurs généralement utilisées sont 4 pour un A, 3 pour un B, 2 pour un C, 1 pour un D et 0 pour un F. Les pondérations correspondent au nombre d'heures de travaux dirigés suivis. L'exercice 16, à la fin de cette section, fournit un exemple du calcul de cette moyenne pondérée. Dans d'autres calculs de moyenne pondérée, les quantités, exprimées en livres ou en dollars, sont fréquemment utilisées comme pondération. Dans tous les cas, lorsque les observations n'ont pas toutes la même importance, l'analyste doit choisir la pondération qui reflète le mieux l'importance de chaque observation dans la détermination de la moyenne.

3.1.3 Médiane

La **médiane** est une autre mesure de tendance centrale pour une variable. Lorsque les données sont classées en ordre croissant (de la plus petite à la plus grande valeur), la médiane correspond à la valeur centrale. Lorsque le nombre d'observations est impair, la médiane correspond à la valeur centrale. Un nombre pair d'observations n'a pas une unique valeur centrale. Dans ce cas, la convention consiste à définir la médiane comme la moyenne des valeurs des deux observations centrales. Par commodité la définition de la médiane est reformulée ci-dessous.

► **Médiane**
Classer les observations en ordre croissant (de la plus petite à la plus grande valeur).
(a) Pour un nombre d'observations impair, la médiane est la valeur centrale.
(b) Pour un nombre d'observations pair, la médiane est la moyenne des deux valeurs centrales.

Appliquons cette définition au calcul de la taille médiane des classes de l'échantillon considérées ci-dessus. Si l'on ordonne de façon croissante les cinq observations, on obtient la liste suivante.

32 42 46 46 54

Puisque le nombre d'observations ($n = 5$) est impair, la médiane correspond à la valeur centrale. Ainsi la taille médiane des classes est de 46 élèves. Bien que l'ensemble de données comporte deux observations qui ont pour valeur 46, chaque observation est traitée séparément lorsqu'on ordonne les données de façon croissante.

Calculons également le salaire initial médian des 12 jeunes diplômés d'une école de commerce. Tout d'abord, nous ordonnons de façon croissante les 12 observations du tableau 3.1.

3 710 3 755 3 850 3880 3 880 $\underbrace{3\,890 \quad 3\,920}_{\text{Deux valeurs centrales}}$ 3 940 3 950 4 050 4 130 4 325

Puisque le nombre d'observations ($n = 12$) est pair, les deux valeurs centrales sont : 3890 et 3920. La médiane correspond à la moyenne de ces deux valeurs.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

$$\text{Médiane} = \frac{3\,890 + 3\,920}{2} = 3\,905$$

La procédure que nous utilisons pour calculer la médiane, dépend du caractère pair ou impair du nombre d'observations. Décrivons maintenant une approche plus conceptuelle et visuelle en utilisant les données sur les salaires mensuels de départ de 12 diplômés. Comme précédemment, nous commençons par ordonner les données par ordre croissant.

3 710 3 755 3 850 3 880 3 880 3 890 3 920 3 940 3 950 4 050 4 130 4 325

Une fois les données ordonnées par ordre croissant, nous barrons successivement les valeurs les plus élevées et les plus faibles situées à chaque extrémité, jusqu'à ce qu'aucune paire supplémentaire de données ne puisse être barrée sans éliminer toutes les données. Par exemple, après avoir barré l'observation la plus faible (3 710) et l'observation la plus élevée (4 325), nous obtenons un nouvel ensemble de données avec 10 observations.

~~3 710~~ 3 755 3 850 3 880 3 880 3 890 3 920 3 940 3 950 4 050 4 130 ~~4 325~~

Nous barrons la plus faible valeur de ce nouvel ensemble (3 755) ainsi que la plus élevée (4 130) et obtenons un nouvel ensemble de données contenant huit observations.

~~3 710~~ ~~3 755~~ 3 850 3 880 3 880 3 890 3 920 3 940 3 950 4 050 ~~4 130~~ ~~4 325~~

En poursuivant ce processus, nous obtenons les résultats suivants.

~~3 710~~ ~~3 755~~ ~~3 850~~ 3 880 3 880 3 890 3 920 3 940 3 950 ~~4 050~~ ~~4 130~~ ~~4 325~~
~~3 710~~ ~~3 755~~ ~~3 850~~ ~~3 880~~ 3 880 3 890 3 920 3 940 ~~3 950~~ ~~4 050~~ ~~4 130~~ ~~4 325~~
~~3 710~~ ~~3 755~~ ~~3 850~~ ~~3 880~~ ~~3 880~~ 3 890 3 920 ~~3 940~~ ~~3 950~~ ~~4 050~~ ~~4 130~~ ~~4 325~~

Ici, il n'est plus possible de barrer des valeurs sans éliminer toutes les données. Aussi, la médiane correspond à la moyenne des deux valeurs restantes. Lorsqu'il y a un nombre pair d'observations, le processus d'élimination progressif des valeurs extrêmes conduira toujours à laisser deux valeurs, et la moyenne de ces valeurs sera égale à la médiane. Lorsque le nombre d'observations est impair, le processus d'élimination progressif conduira toujours à conserver une seule valeur et cette valeur correspondra précisément à la médiane. Ainsi, cette méthode fonctionne que le nombre d'observations soit pair ou impair.

La médiane est la mesure de tendance centrale la plus souvent utilisée lorsque l'on traite de données sur le revenu annuel et la valeur foncière, car quelques valeurs très élevées du revenu ou de la valeur foncière peuvent accroître la moyenne. Dans de telles situations, la médiane est une meilleure mesure de tendance centrale.

Bien que la moyenne soit la mesure de tendance centrale la plus souvent utilisée, dans certaines situations l'utilisation de la médiane est préférable. La moyenne est en effet

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

influencée par les valeurs extrêmement petites et extrêmement grandes. Par exemple, supposez que l'un des diplômés (cf. tableau 3.1) ait un salaire initial de 10 000 dollars par mois (la famille de cette personne possède peut-être la société). Si l'on remplace le salaire mensuel initial le plus élevé du tableau 3.1, égal à 4 325 dollars, par 10 000 dollars et que l'on recalcule la moyenne, cette dernière passera de 3 940 à 4 413 dollars. Par contre, la médiane égale à 3 905 dollars est inchangée puisque les valeurs centrales, 3 890 et 3 920 ne sont pas modifiées. Étant donnée cette valeur extrêmement élevée du salaire initial de l'un des jeunes diplômés, la médiane fournit une meilleure mesure de tendance centrale que la moyenne. De façon générale, lorsqu'un ensemble de données contient des valeurs extrêmes, la médiane est souvent une mesure préférable de la tendance centrale.

3.1.4 Moyenne géométrique

La moyenne géométrique est une mesure de tendance centrale qui est calculée en trouvant la racine n^{ième} du produit de n valeurs.

► **Moyenne géométrique**

$$\bar{x}_g = \sqrt[n]{(x_1)\,(x_2)\ldots(x_n)} = [(x_1)(x_2)\ldots(x_n)]^{1/n} \qquad (3.4)$$

La moyenne géométrique est souvent utilisée pour analyser les taux de croissance relatifs à des données financières. Dans ce type de situation, la moyenne arithmétique ou la valeur moyenne fournissent des résultats trompeurs.

Pour illustrer l'utilisation de la moyenne géométrique, considérons le tableau 3.2 qui fournit les rendements annuels en pourcentage, ou taux de croissance, d'un fond mutuel au cours des 10 dernières années. Supposons que nous voulions calculer combien 100 dollars investis dans ce fond au début de l'année 1 valent à la fin de l'année 10. Commençons par calculer le solde du fond à la fin de l'année 1. Puisque le rendement annuel en pourcentage durant l'année 1 était de -22,1 %, le solde à la fin de l'année 1 était de

$$100\ \$ - 0{,}221(100\ \$) = (0{,}779)100\ \$ = 77{,}90\ \$$$

Notez que 0,779 correspond au facteur de croissance de l'année 1 inscrit dans le tableau 3.2. Ce résultat révèle que nous pouvons calculer le solde à la fin de l'année 1 en multipliant la valeur investie dans le fond au début de l'année 1 par le facteur de croissance de l'année 1.

Le facteur de croissance pour chaque année est 1 plus 0,01 fois le rendement en pourcentage. Un facteur de croissance inférieur à 1 indique une croissance négative, alors qu'un facteur de croissance supérieur à 1 indique une croissance positive. Le facteur de croissance ne peut pas être inférieur à zéro.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Tableau 3.2 *Rendements annuels en pourcentage et facteurs de croissance du fond mutuel*

Année	Rendement (%)	Facteur de croissance
1	–22,1	0,779
2	28,7	1,287
3	10,9	1,109
4	4,9	1,049
5	15,8	1,158
6	5,5	1,055
7	–37,0	0,630
8	26,5	1,265
9	15,1	1,151
10	2,1	1,021

Le solde du fond à la fin de l'année 1, 77,90 dollars, correspond au montant présent dans le fond au début de l'année 2. Aussi, avec un rendement annuel en pourcentage de 28,7 % au cours de l'année 2, le solde à la fin de l'année 2 était de

$$77,90\ \$ + 0,287(77,90\ \$) = (1 + 0,287)77,90\ \$ = (1,287)77,90\ \$ = 100,2573\ \$$$

Notez que 1,287 correspond au facteur de croissance de l'année 2 figurant dans le tableau 3.2. Et, en substituant 77,90 $ par (0,779)100 $, nous voyons que le solde du fond à la fin de l'année 2 est

$$(0,779)(1,287)100\ \$ = 100,2573\ \$$$

En d'autres termes, le solde à la fin de l'année 2 correspond à l'investissement initial effectué au début de l'année 1 multiplié par le produit des deux premiers facteurs de croissance. Ce résultat peut être généralisé pour montrer que le solde à la fin de l'année 10 correspond à l'investissement initial multiplié par le produit des 10 facteurs de croissance.

$$100\ \$[(0,779)(1,287)(1,109)(1,049)(1,158)(1,055)(0,630)(1,265)(1,151)(1,021)] =$$
$$100\ \$(1,334493) = 133,4493\ \$$$

Ainsi, investir 100 dollars dans le fond au début de l'année 1 aurait rapporté 133,44 dollars à la fin de l'année 10. Notez que le produit des 10 facteurs de croissance est égal à 1,334493. Par conséquence, nous pouvons calculer le solde à la fin de l'année 10 pour n'importe quel montant investi au début de l'année 1 en multipliant la valeur de cet investissement initial par 1,334493. Par exemple, un investissement initial de 2 500 dollars au début de l'année 1 aurait rapporté $(1,334493) \times 2\ 500\ \$$ soit approximativement 3 336 dollars à la fin de l'année 10.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

La racine n^{ième} peut être calculée en utilisant de puissantes calculatrices ou la fonction PUISSANCE d'Excel. Par exemple, en utilisant Excel, la racine 10^{e} de 1,334493 = PUISSANCE (1,334493, 1/10) ou 1,029275.

Mais quel était le rendement annuel en pourcentage moyen ou le taux de croissance moyen de cet investissement sur les 10 années ? Voyons comment utiliser la moyenne géométrique des 10 facteurs de croissance pour répondre à cette question. Puisque le produit des 10 facteurs de croissance est égal à 1,334493, la moyenne géométrique correspond à la racine 10^{e} de 1,334493, soit

$$\overline{x_g} = \sqrt[n]{1{,}334\ 493} = 1{,}029275$$

La moyenne géométrique nous dit que les rendements annuels ont augmenté au taux annuel moyen de $(1{,}029275 - 1)100\ \%$, soit 2,9275 %. En d'autres termes, avec un taux de croissance annuel moyen de 2,9275 %, un investissement de 100 dollars au début de l'année 1 aurait rapporté $100(1{,}029275)^{10}\ \$ = 133{,}4493\ \$$ au bout de 10 ans.

Il est important de comprendre que la moyenne arithmétique des rendements annuels en pourcentage ne fournit pas le taux de croissance annuel moyen de cet investissement. La somme des 10 rendements annuels en pourcentage figurant dans le tableau 3.2 est égale à 50,4. Par conséquent, la moyenne arithmétique des 10 rendements annuels en pourcentage est égale à $50{,}4 / 10 = 5{,}04\ \%$. Un courtier pourrait essayer de vous convaincre d'investir dans ce fond en affirmant que le rendement annuel moyen en pourcentage est de 5,04 %. Une telle affirmation est non seulement trompeuse mais fausse. Un rendement annuel moyen en pourcentage de 5,04 % correspond à un facteur de croissance moyen de 1,0504. Si le facteur de croissance moyen avait réellement été de 1,0504, 100 dollars investis dans le fond au début de l'année 1 aurait rapporté $100\ \$(1{,}0504)^{10} = 163{,}51\ \$$ au bout des 10 années. Mais, en utilisant les rendements annuels en pourcentage figurant dans le tableau 3.2, nous avons montré qu'un investissement initial de 100 dollars rapportait 133,45 dollars au bout de 10 ans. L'affirmation du courtier d'un rendement annuel moyen en pourcentage de 5,04 % surestime grossièrement la croissance réelle de ce fond mutuel. Le problème est que la moyenne d'échantillon n'est pertinente que pour un processus additif. Pour un processus multiplicatif, comme pour des cas impliquant des taux de croissance, la moyenne géométrique est la mesure appropriée.

Alors que les applications de la moyenne géométrique aux problèmes relatifs à la finance, aux investissements ou aux opérations bancaires sont particulièrement courantes, la moyenne géométrique devrait être appliquée à chaque fois que vous souhaitez déterminer le taux d'évolution moyen sur plusieurs périodes successives. Des changements dans la population d'espèces, dans les rendements agricoles, les niveaux de pollution et les taux de naissance et de décès sont d'autres cas d'application courants de la moyenne géométrique. Notez également que la moyenne géométrique peut être appliquée quelle que soit le nombre de périodes considérées et quelle que soit leur durée. En plus des évolutions annuelles, la moyenne géométrique est souvent appliquée pour trouver le taux moyen d'évolution trimestriel, mensuel, hebdomadaire et même quotidien.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

3.1.5 Mode

Une autre mesure de tendance centrale est le mode. Le mode est défini de la façon suivante.

► **Mode**
Le mode correspond à la valeur de l'observation qui a la plus grande fréquence.

Considérons l'exemple de l'échantillon des cinq tailles de classe. La seule valeur qui apparaît plus d'une fois est 46. Puisque cette valeur, qui a une fréquence de 2, a la plus grande fréquence, il s'agit du mode. Considérons à présent l'échantillon des salaires initiaux des diplômés d'une école de commerce. Le seul salaire mensuel initial qui apparaît plus d'une fois est 3 880 dollars. Puisque cette valeur a la plus grande fréquence, il s'agit du mode.

Il est possible que plusieurs valeurs apparaissent avec la même fréquence et que cette fréquence soit la plus importante. Dans ce cas, plus d'un mode existe. Si les données ont exactement deux modes, on dit que les données sont *bimodales*. Si les données ont plus de deux modes, on dit qu'elles sont *multimodales*. Dans les cas multimodaux, le mode n'est presque jamais utilisé car énumérer trois modes ou plus n'est pas particulièrement utile pour décrire les données.

3.1.6 Percentiles

Un **percentile** fournit des informations sur la manière dont les observations sont réparties dans l'intervalle entre la plus petite et la plus grande valeur. Pour des données dont la valeur n'est pas répétée plusieurs fois, le p^e percentile divise l'ensemble de données en deux parties. Environ p pour cent des observations ont une valeur inférieure au p^e percentile ; environ $(100 - p)$ pour cent des observations ont une valeur supérieure au p^e percentile. Le p^e percentile est défini formellement de la façon suivante :

► **Percentile**
Le p^e percentile est la valeur telle qu'au moins p pour cent des observations sont inférieures ou égales à cette valeur, et au plus $(100 - p)$ pour cent des observations sont supérieures ou égales à cette valeur.

Les résultats des tests d'admission des grandes écoles et universités sont fréquemment rapportés en termes de percentiles. Par exemple, supposez qu'un candidat obtienne une note égale à 54 à l'oral du test d'admission. Les résultats de cet étudiant ne sont pas directement comparables à ceux obtenus par d'autres étudiants ayant effectué le même test. Cependant, si la note de 54 correspond au 70^e percentile, nous savons qu'approximativement 70 % des étudiants ont une note inférieure à celle de cet individu et qu'approximativement 30 % des étudiants ont une note supérieure.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

La procédure suivante peut être utilisée pour calculer le p^e percentile.

► **Calculer le p^e percentile**

Étape 1. Classer les données en ordre croissant (de la plus petite à la plus grande valeur).

Étape 2. Calculer un index i

$$i = \left(\frac{p}{100}\right) n$$

où p est le percentile considéré et n le nombre d'observations.

Étape 3. (a) Si i n'est pas un nombre entier, l'arrondir. La position du p^e percentile correspond à l'entier supérieur à i.
(b) Si i est un nombre entier, la position du p^e percentile correspond à la moyenne des valeurs des observations i et $i+1$.

Suivre ces étapes facilite le calcul des percentiles.

Pour illustrer cette procédure, déterminons le 85e percentile pour les données sur les salaires initiaux du tableau 3.1.

Étape 1. Classer les données en ordre croissant.

3 710 3 755 3 850 3 880 3 880 3 890 3 920 3 940 3 950 4 050 4 130 4 325

Étape 2.

$$i = \left(\frac{p}{100}\right) n = \left(\frac{85}{100}\right) 12 = 10,2$$

Étape 3. Puisque i n'est pas un nombre entier, on l'arrondit. La position du 85e percentile correspond au nombre entier supérieur à 10,2, soit la 11e position.

En reprenant les données, on s'aperçoit que le 85e percentile est égal à 4 130.

Considérons à présent le calcul du 50e percentile pour les données sur les salaires initiaux. En appliquant l'étape 2, on obtient

$$i = \left(\frac{50}{100}\right) 12 = 6$$

Puisque i est un nombre entier, d'après l'étape 3(b), le 50e percentile correspond à la moyenne des 6e et 7e observations ; ainsi le 50e percentile est égal à (3 890 + 3 920) / 2 = 3 905. Remarquez que *le 50e percentile est également la médiane*.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

3.1.7 Quartiles

Les quartiles sont des percentiles particuliers ; aussi, les étapes de calcul des percentiles peuvent être directement appliquées au calcul des quartiles.

Il est souvent utile de diviser les données en quatre parts, chacune contenant approximativement un quart, soit 25 % des observations. La figure 3.1 représente une distribution de données divisée en quatre parts. Les points de division sont appelés **quartiles** et sont définis de la façon suivante

Q_1 = premier quartile, ou 25^e percentile
Q_2 = deuxième quartile, ou 50^e percentile (aussi la médiane)
Q_3 = troisième quartile, ou 75^e percentile.

Pour calculer les quartiles des données sur les salaires initiaux, nous classons les données par ordre croissant.

3 710 3 755 3 850 3 880 3 880 3 890 3 920 3 940 3 950 4 050 4 130 4 325

Q_2, le deuxième quartile (la médiane), a déjà été calculé : il est égal à 3 905. Le calcul des quartiles Q_1 et Q_3 nécessite l'utilisation de la règle de calcul des 25^e et 75^e percentiles. Ces calculs sont présentés ci-dessous.

Pour Q_1,

$$i = \left(\frac{p}{100}\right)n = \left(\frac{25}{100}\right)12 = 3$$

Puisque i est un nombre entier, l'étape 3(b) indique que le premier quartile, ou 25^e percentile, est la moyenne de la 3^e et de la 4^e observation ; ainsi, $Q_1 = (3\,850 + 3\,880)/2 = 3\,865$.

Pour Q_3,

$$i = \left(\frac{p}{100}\right)n = \left(\frac{75}{100}\right)12 = 9$$

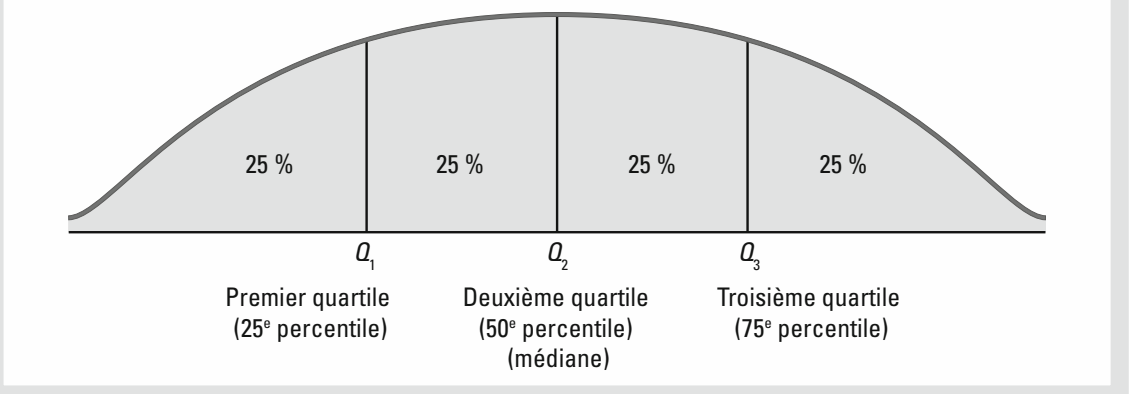


Figure 3.1 *Position des quartiles*

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

De nouveau, puisque i est un nombre entier, l'étape 3(b) indique que le troisième quartile, ou 75e percentile, est la moyenne de la 9e et de la 10e observation ; ainsi, $Q_3 = (3\,950 + 4\,050) / 2 = 4\,000$.

Les quartiles ont permis de diviser les données sur les salaires initiaux en quatre parties, chacune comportant 25 % des observations.

3 310 3 355 3 450 | 3 480 3 480 3 490 | 3 520 3 540 3 550 | 3 650 3 730 3 925

$Q_1 = 3\,465$ $Q_2 = 3\,505$ (Médiane) $Q_3 = 3\,600$

Nous avons défini les quartiles comme étant les 25e, 50e et 75e percentiles. Ainsi nous avons calculé les quartiles de la même façon que les percentiles. On peut utiliser d'autres conventions pour calculer les quartiles, leurs valeurs pouvant varier légèrement en fonction de la convention utilisée. Cependant quelle que soit la procédure de calcul des quartiles utilisée, l'objectif est de diviser l'ensemble des données en quatre parts égales.

REMARQUES

Il est préférable d'utiliser la médiane plutôt que la moyenne comme mesure de tendance centrale lorsque l'ensemble de données contient des valeurs extrêmes. Une autre mesure parfois utilisée, lorsque des valeurs extrêmes sont présentes, est la **moyenne tronquée**. Elle est obtenue en supprimant un certain pourcentage des observations les plus petites et des observations les plus grandes d'un ensemble de données puis en calculant la moyenne des valeurs restantes. Par exemple, la moyenne tronquée à 5 % est obtenue en supprimant 5 % des plus petites valeurs et 5 % des valeurs les plus grandes puis en calculant la moyenne des valeurs restantes. En utilisant l'échantillon contenant les 12 observations sur les salaires initiaux, $0{,}05 \times 12 = 0{,}6$. Si l'on arrondit cette valeur à 1, la moyenne tronquée à 5 % est obtenue en supprimant la plus petite et la plus grande valeur. Ainsi, la moyenne tronquée à 5 %, en utilisant les 10 observations restantes, est égale à 3 924,5.

D'autres percentiles couramment utilisés sont les quintiles (les 20e, 40e, 60e et 80e percentiles) et les déciles (les 10e, 20e, 30e, 40e, 50e, 60e, 70e, 80e et 90e percentiles).

EXERCICES

Méthode

1. Considérer un échantillon avec les observations suivantes : 10, 20, 12, 17 et 16. Calculer la moyenne et la médiane.

2. Considérer un échantillon avec les observations suivantes : 10, 20, 21, 17, 16 et 12. Calculer la moyenne et la médiane.

3. Considérer les données suivantes et les pondérations associées.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

x_i	Pondération (w_i)
3,2	6
2,0	3
2,5	2
5,0	8

a) Calculer la moyenne pondérée.

b) Calculer la moyenne d'échantillon des quatre observations sans tenir compte des pondérations. Notez la différence entre les deux résultats.

4. Considérer les données suivantes.

Période	Taux de rendement (%)
1	−6,0
2	−8,0
3	−4,0
4	2,0
5	5,4

Quel est le taux de croissance moyen au cours des cinq périodes ?

5. Considérer un échantillon avec les observations suivantes : 27, 25, 20, 15, 30, 34, 28 et 25. Calculer le 20e, 25e, 65e et 75e percentile.

6. Considérer un échantillon avec les observations suivantes : 53, 55, 70, 58, 64, 57, 53, 69, 57, 68 et 53. Calculer la moyenne, la médiane et le mode.

Applications

Temps trajet domicile-travail

7. Les Américains mettent en moyenne 27,7 minutes pour aller travailler (*Sterling's Best Places*, 13 avril 2012). Les temps moyens en minutes pour aller travailler pour 48 villes sont les suivants (fichier en ligne Temps trajet domicile-travail).

Albuquerque	23,3	Jacksonville	26,2	Phoenix	28,3
Atlanta	28,3	Kansas City	23,4	Pittsburgh	25,0
Austin	24,6	Las Vegas	28,4	Portland	26,4
Baltimore	32,1	Little Rock	20,1	Providence	23,6
Boston	31,7	Los Angeles	32,2	Richmond	23,4
Charlotte	25,8	Louisville	21,4	Sacramento	25,8
Chicago	38,1	Memphis	23,8	Salt Lake City	20,2
Cincinnati	24,9	Miami	30,7	San Antonio	26,1
Cleveland	26,8	Milwaukee	24,8	San Diego	24,8
Columbus	23,4	Minneapolis	23,6	San Francisco	32,6
Dallas	28,5	Nashville	25,3	San Jose	28,5
Denver	28,1	New Orleans	31,7	Seattle	27,3

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Detroit	29,3	New York	43,8	St. Louis	26,8
El Paso	24,4	Oklahoma City	22,0	Tucson	24,0
Fresno	23,0	Orlando	27,1	Tulsa	20,1
Indianapolis	24,8	Philadelphia	34,2	Washington, D.C.	32,8

a) Quel est le temps moyen pour aller travailler dans ces 48 villes ?

b) Calculer le temps médian.

c) Calculer le mode.

d) Calculer le troisième quartile.

8. Durant la saison 2007-2008 de basket de la NCAA, les équipes masculines de basket ont battu le record de tirs à 3 points, atteignant en moyenne 19,07 tirs par match (Associated Press Sports, 24 janvier 2009). Dans le but de décourager les tirs à 3 points et encourager davantage de jeu offensif, le comité des règles de la NCAA a reculé la ligne des tirs à 3 points de 19 pieds et 9 pouces à 20 pieds et 9 pouces au début de la saison 2008-2009. Des données sur les tirs à 3 points réalisés lors d'un échantillon de 19 matchs de la NCAA durant la saison 2008-2009 sont réunies dans le tableau suivant (fichier en ligne 3 points).

Tirs à trois points tentés	Tirs réussis	Tirs à trois points tentés	Tirs réussis
23	4	17	7
20	6	19	10
17	5	22	7
18	8	25	11
13	4	15	6
16	4	10	5
8	5	11	3
19	8	25	8
28	5	23	7
21	7		

a) Quel est le nombre moyen de tirs à 3 points tentés par match ?

b) Quel est le nombre moyen de tirs à 3 points réussis par match ?

c) En partant de la ligne des trois points la plus proche du panier, les joueurs réussissaient 35,2 % de leurs tirs. Quel pourcentage de tirs les joueurs réussissent-ils à partir de la nouvelle ligne des trois points ?

d) Quel fut l'impact du changement de règle de la NCAA qui repoussa la ligne des trois points à 20 pieds et 9 pouces durant la saison 2008-2009 ? Êtes-vous d'accord avec l'affirmation figurant dans l'article de l'Associated Press Sports selon laquelle « Le recul de la ligne de tir à trois points n'a pas fondamentalement changé la façon de jouer » ? Expliquez.

9. La dotation budgétaire est un élément critique des budgets annuels des grandes écoles et des universités. Selon une étude menée par l'Association nationale des gestionnaires d'universités et de grandes écoles auprès de 435 grandes écoles et universités, le budget

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

global de ces institutions s'élevait à 413 milliards de dollars. Les 10 universités les plus riches sont regroupées dans le tableau suivant (*The Wall Street Journal*, 27 janvier 2009). Les montants sont exprimés en milliards de dollars.

Université	Budget (milliards de dollars)	Université	Budget (milliards de dollars)
Columbia	7,2	Princeton	16,4
Harvard	36,6	Stanford	17,2
M.I.T.	10,1	Texas	16,1
Michigan	7,6	Texas A&M	6,7
Northwestern	7,2	Yale	22,9

a) Quel est le budget moyen de ces dix universités ?

b) Quel est le budget médian ?

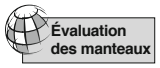
c) Quel est le mode ?

d) Calculer les premier et troisième quartiles.

e) Quel est le budget total de ces dix universités ? Ces universités représentent 2,3 % des 435 grandes écoles et universités interrogées. En pourcentage que représente le budget de ces dix universités sur les 413 milliards de dollars mentionnés dans l'étude ?

f) Le *Wall Street Journal* déclarait qu'au cours des cinq derniers mois, le ralentissement de l'économie avait entraîné une réduction des budgets de 23 %. Quelle est l'estimation de la réduction budgétaire (en milliards de dollars) que pourraient subir ces 10 universités ? Étant donnée la situation, quelles mesures les gestionnaires des universités pourraient-ils prendre ?

10. Pendant neuf mois, OutdoorGearLab a testé des manteaux conçus pour l'ascension des glaciers, l'alpinisme et la randonnée. Une note allant de 0 (la plus faible) à 100 (la plus élevée) a été attribuée à chaque manteau testé en fonction de son côté respirant, de sa durée de vie, de sa polyvalence, des possibilités de se mouvoir avec et de son poids. Les données suivantes correspondent aux évaluations des 20 meilleurs manteaux (OutdoorGearLab, 27 février 2013).



42	66	67	71	78	62	61	76	71	67
61	64	61	54	83	63	68	69	81	53

a) Calculer la moyenne, la médiane et le mode.

b) Calculer les premier et troisième quartiles.

c) Calculer et interpréter le 90^{e} percentile.

11. Selon l'Association nationale pour l'éducation (NEA), les enseignants passent généralement plus de 40 heures par semaine à des tâches éducatives (site Internet de NEA, avril 2012). Les données suivantes indiquent le nombre d'heures hebdomadaires d'enseignement d'un échantillon de 13 professeurs de sciences et de 11 professeurs d'anglais au lycée.

Professeurs de sciences :	53	56	57	57	88	58	49	61	54	54	52	53	54
Professeurs d'anglais :	52	47	50	46	47	48	49	46	55	44	47		

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

a) Quel est le nombre médian d'heures hebdomadaires de cours pour l'échantillon des 13 professeurs de sciences ?

b) Quel est le nombre médian d'heures hebdomadaires de cours pour l'échantillon des 11 professeurs d'anglais ?

c) Quel groupe a le nombre d'heures de cours par semaine médian le plus élevé ? Quel est l'écart entre le nombre d'heures de cours par semaine médian ?

12. *The Big Bang Theory*, une série mettant en scène Johnny Galecki, Jim Parsons et Kaley Cuoco, est un des programmes télévisés les plus regardés. Les deux premiers épisodes de la saison 2011-2012 ont été diffusés pour la première fois le 22 septembre 2011 ; le premier épisode a attiré 14,1 millions de téléspectateurs et le second épisode 14,7 millions. Le tableau suivant (fichier en ligne BigBangTheory) indique le nombre de téléspectateurs (en millions) qui ont regardé les 21 premiers épisodes de la saison 2011-2012 (site Internet de *The Big Bang Theory*, 17 avril 2012).

Date de diffusion	Nombre de téléspectateurs (millions)	Date de diffusion	Nombre de téléspectateurs (millions)
22 septembre 2011	14,1	12 janvier 2012	16,1
22 septembre 2011	14,7	19 janvier 2012	15,8
29 septembre 2011	14,6	26 janvier 2012	16,1
6 octobre 2011	13,6	2 février 2012	16,5
13 octobre 2011	13,6	9 février 2012	16,2
20 octobre 2011	14,9	16 février 2012	15,7
27 octobre 2011	14,5	23 février 2012	16,2
3 novembre 2011	16,0	8 mars 2012	15,0
10 novembre 2011	15,9	29 mars 2012	14,0
17 novembre 2011	15,1	5 avril 2012	13,3
8 décembre 2011	14,0		

a) Calculer le nombre minimum et maximum de téléspectateurs.

b) Calculer la moyenne, la médiane et le mode.

c) Calculer les premier et troisième quartiles.

d) L'audience a-t-elle augmenté ou diminué au cours de la saison 2011-2012 ? Discuter.

13. Pour tester la consommation d'essence, 13 automobiles ont parcouru 300 miles dans des conditions de conduite similaires à celles obtenues en ville et sur autoroute. Les données sur la consommation, en miles par gallon, sont présentées ci-dessous.

Ville :	16,2	16,7	15,9	14,4	13,2	15,3	16,8	16,0	16,1	15,3	15,2	15,3	16,2
Autoroute :	19,4	20,6	18,3	18,6	19,2	17,4	17,2	18,6	19,0	21,1	19,4	18,5	18,7

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Utiliser la moyenne, la médiane et le mode pour étudier les différences de performance entre la conduite en ville et sur autoroute.

14. Les données contenues dans le fichier en ligne nommé Taux de chômage indiquent les taux de chômage enregistrés en mars 2011 et en mars 2012 dans chaque État et dans le District de Columbia (site Internet du Bureau des statistiques de l'emploi, 10 avril 2012). Pour comparer les taux de chômage de mars 2011 avec ceux de mars 2012, calculer le premier quartile, la médiane et le troisième quartile pour les données de mars 2011 et de mars 2012. Que suggèrent ces statistiques à propos de l'évolution des taux de chômage au sein des États ?

15. Martinez Auto Supplies possède des magasins dans huit villes de Californie. Le prix qu'ils pratiquent pour un produit particulier dans chaque ville varie à cause des conditions concurrentielles différentes. Par exemple, le prix pratiqué pour un bidon d'huile de moteur d'une marque connue dans chaque ville est fourni ci-dessous. Les données indiquent également le nombre de bidons vendus au cours du dernier trimestre par Martinez Auto dans chaque ville.

Ville	Prix ($)	Ventes (nombre de bidons)
Bakersfield	34,99	501
Los Angeles	38,99	1 425
Modesto	36,00	294
Oakland	33,59	882
Sacramento	40,99	715
San Diego	38,59	1 088
San Francisco	39,59	1 644
San Jose	37,99	819

Calculer le prix moyen de vente d'un bidon d'huile au cours du dernier trimestre.

16. Le calcul de la moyenne des notes des étudiants correspond au calcul d'une moyenne pondérée. Dans la plupart des universités américaines, les notes ont les valeurs suivantes : A (4), B (3), C (2), D (1) et F (0). Sur un total de 60 heures de travaux dirigés, un étudiant d'une université a sanctionné 9 heures de TD par un A, 15 heures par un B, 33 heures par un C et 3 heures par un D.

a) Calculer la moyenne de cet étudiant.

b) Les étudiants d'une université publique doivent obtenir une moyenne de 2,5 pour leurs 60 premières heures de travaux dirigés pour pouvoir passer en deuxième année. Est-ce que cet étudiant sera admis ?

17. Morningstar enregistre le rendement total d'un grand nombre de fonds mutuels. Le tableau suivant indique le rendement total et le nombre de fonds pour quatre catégories de fonds mutuels (*Morningstar Funds 500*, 2008).

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Type de fonds	Nombre de fonds	Rendement total (%)
Fonds domestique	9 191	4,65
Fonds international	2 621	18,15
Action spécialisée	1 419	11,36
Fonds hybride	2 900	6,75

a) En utilisant le nombre de fonds comme pondération, calculer le rendement total moyen pondéré pour les fonds mutuels suivis par Morningstar.

b) Y a-t-il une difficulté à utiliser le nombre de fonds comme pondération pour calculer le rendement total moyen pondéré à la question (a) ? Discuter. Quel autre facteur pourrait être utilisé comme pondération ?

c) Supposez que vous ayez investi 10 000 dollars dans les fonds mutuels au début de 2007 et diversifié votre investissement en plaçant 2 000 dollars dans des fonds domestiques, 4 000 dollars dans des fonds internationaux, 3 000 dollars dans des actions spécialisées et 1 000 dollars dans des fonds hybrides. Quel est le rendement attendu de votre portefeuille ?

18. À partir d'une enquête sur 425 programmes de master dans des écoles de commerce, *U.S. News & World Report* a classé l'école de commerce Kelley de l'université de l'Indiana à la 20e place des meilleurs programmes du pays (*America's Best Graduate Schools*, 2009). Le classement était basé en partie sur des enquêtes réalisées auprès des doyens des écoles et des chasseurs de tête. Chaque personne interrogée devait attribuer une note à la qualité académique générale du programme de master sur une échelle allant de 1 « mauvaise » à 5 « remarquable ». Utiliser l'échantillon suivant de réponses pour calculer la note moyenne pondérée attribuée par les doyens et les chasseurs de tête. Discuter.

Note attribuée	Nombre de doyens des écoles	Nombre de chasseurs de tête
5	44	31
4	66	34
3	60	43
2	10	12
1	0	0

19. Le revenu annuel de Corning Supplies a augmenté de 5,5 % en 2007, 1,1 % en 2008, – 3,5 % en 2009, – 1,1 % en 2010 et 1,8 % en 2011. Quel est le taux annuel de croissance moyen sur cette période ?

20. Supposez qu'au début de l'année 2004 vous investissiez 10 000 dollars dans le fond mutuel Stivers et 5 000 dollars dans le fond mutuel Trippi. La valeur de chaque investissement à la fin de chaque année suivante est fournie dans le tableau ci-dessous. Quel est le fond le plus performant ?

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Année	Stivers	Trippi
2004	11 000	5 600
2005	12 000	6 300
2006	13 000	6 900
2007	14 000	7 600
2008	15 000	8 500
2009	16 000	9 200
2010	17 000	9 900
2011	18 000	10 600

21. Si la valeur d'un actif passe de 5 000 dollars à 3 500 dollars en neuf ans, quel est le taux de croissance annuel moyen de la valeur de cet actif au cours de ces neuf années ?

22. La valeur actuelle d'une société s'élève à 25 millions de dollars. Si la valeur de la société six ans auparavant était de 10 millions de dollars, quel est le taux de croissance annuel moyen de la valeur de cette société au cours des six dernières années ?

3.2 MESURES DE VARIABILITÉ

En plus des mesures de tendance centrale, il est souvent utile de considérer des mesures de variabilité ou de dispersion des données. Par exemple, supposons que vous êtes le directeur du service des achats d'une grande entreprise et que régulièrement vous passez commande à deux fournisseurs différents. Après plusieurs mois, vous vous apercevez que le nombre moyen de jours nécessaires aux deux fournisseurs pour honorer les commandes est de dix jours. Les histogrammes indiquant le nombre de jours nécessaires aux deux fournisseurs pour honorer une commande sont représentés à la figure 3.2. Bien que le nombre moyen de jours soit égal à 10 pour les deux fournisseurs, peut-on accorder le même degré de

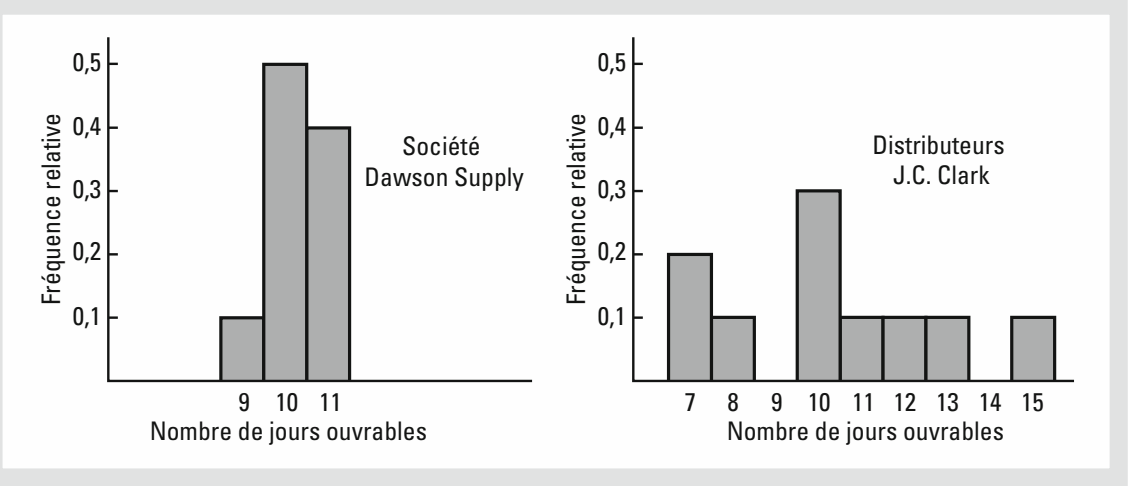


Figure 3.2 *Données historiques indiquant le nombre de jours nécessaires pour honorer les commandes*

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

confiance aux deux fournisseurs en termes de délais de livraison ? Notez la dispersion, ou variabilité, dans les délais de livraison, indiquée par les histogrammes. Quel fournisseur préfèreriez-vous ?

La variabilité des délais de livraison crée une incertitude dans le planning de production. Les méthodes présentées dans cette section aident à mesurer et à comprendre la variabilité.

Pour la plupart des entreprises, recevoir les matériaux et les marchandises dans les délais est important. Le délai de sept ou huit jours demandé par la société J. C. Clark peut être considéré comme acceptable ; par contre, un délai de treize ou quinze jours peut être désastreux en termes de gestion de la production. Cet exemple illustre une situation dans laquelle la variabilité des délais de livraison peut être un élément déterminant dans le choix d'un fournisseur. Pour la plupart des directeurs des achats, la plus faible dispersion des délais imposés par la société Dawson peut être un avantage pour ce fournisseur.

Nous discutons maintenant des mesures de dispersion les plus souvent utilisées.

3.2.1 Étendue

L'**étendue** est la mesure de dispersion la plus simple.

► **Étendue**

Étendue = Valeur la plus grande – Valeur la plus petite

Reprenons les données sur les salaires initiaux des diplômés d'une école de commerce du tableau 3.1. Le salaire initial le plus élevé est de 4 325 et le plus petit est de 3 710. L'étendue est égale à 4 325 − 3 710 = 615.

Bien que l'étendue soit la mesure de dispersion la plus simple à calculer, elle est rarement utilisée seule parce qu'elle est basée uniquement sur deux observations et donc est très influencée par les valeurs extrêmes. Supposons que l'un des diplômés ait un salaire initial de 10 000 dollars par mois. Dans ce cas, l'étendue serait égale à 10 000 − 3 710 = 6 290 au lieu de 615. Cette valeur importante de l'étendue ne décrit pas correctement la dispersion des données, qui contiennent 11 observations sur 12 comprises entre 3 710 et 4 130.

3.2.2 Étendue interquartile

L'étendue interquartile (EIQ) est une mesure de dispersion qui n'est pas dépendante des valeurs extrêmes. Cette mesure de dispersion est égale à l'écart entre le troisième quartile Q_3 et le premier quartile Q_1. En d'autres termes, l'intervalle interquartile mesure l'étendue de la moitié centrale des observations.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

► **Étendue interquartile**

$$EIQ = Q_3 - Q_1 \tag{3.5}$$

Pour les données sur les salaires mensuels initiaux, les 1er et 3e quartiles sont respectivement égaux à 4 000 et 3 865. Ainsi, l'étendue interquartile est égale à 4 000 − 3 865 = 135.

3.2.3 Variance

La **variance** est une mesure de dispersion qui utilise toutes les observations. La variance est basée sur la différence entre la valeur de chaque observation (x_i) et la moyenne ($\bar{x}$ pour un échantillon, μ pour la population). La différence entre chaque observation x_i et la moyenne est appelée *écart par rapport à la moyenne*. Pour un échantillon, un écart par rapport à la moyenne s'écrit $(x_i - \bar{x})$; pour une population, il s'écrit $(x_i - \mu)$. Pour calculer la variance, les écarts par rapport à la moyenne sont élevés au carré.

Si les données sont issues d'une population, la moyenne des écarts au carré est appelée *variance de la population*. La variance de la population est notée par le symbole grec σ^2. Dans le cadre d'une population comprenant N observations, de moyenne μ, la variance est définie par l'expression suivante :

► **Variance de la population**

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \tag{3.6}$$

Dans la plupart des études statistiques, les données à analyser sont issues d'un échantillon. Le calcul de la variance d'un échantillon nous permet généralement ensuite d'estimer la variance de la population σ^2. Bien qu'une explication détaillée ne soit pas l'objet de ce paragraphe, on peut souligner que si la somme des écarts par rapport à la moyenne au carré est divisée par $n - 1$ et non par n, la variance de l'échantillon fournira un estimateur sans

Tableau 3.3 *Calcul des écarts et des écarts au carré par rapport à la moyenne pour les données relatives à la taille des classes*

Nombre d'étudiants dans la classe (x_i)	Taille moyenne des classes ($\bar{x}$)	Écart par rapport à la moyenne ($x_i - \bar{x}$)	Écart au carré par rapport à la moyenne $(x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	12	144
		Somme = 0	Somme = 256

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

biais de la variance de la population. Pour cette raison, la *variance de l'échantillon*, notée s^2, est définie de la façon suivante :

► **Variance de l'échantillon**

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} \tag{3.7}$$

La variance d'échantillon s^2 est l'estimateur de la variance de la population σ^2.

Pour illustrer le calcul de la variance d'un échantillon, nous utiliserons les données sur la taille des classes fournies à la section 3.1. Un résumé des données, incluant le calcul des écarts par rapport à la moyenne et des écarts au carré, est présenté dans le tableau 3.3. La somme des écarts par rapport à la moyenne au carré $\sum (x_i - \bar{x})^2$ est égale à 256. Avec $n - 1 = 4$, la variance de l'échantillon est égale à

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{256}{4} = 64$$

Tableau 3.4 *Calcul de la variance d'échantillon pour les données sur les salaires initiaux des jeunes diplômés*

Salaire mensuel (x_i)	Moyenne d'échantillon ($\bar{x}$)	Écart par rapport à la moyenne ($x_i - \bar{x}$)	Écart au carré par rapport à la moyenne ($x_i - \bar{x}$)^2
3 450	3 540	−90	8 100
3 550	3 540	10	100
3 650	3 540	110	12 100
3 480	3 540	−60	3 600
3 355	3 540	−185	34 225
3 310	3 540	−230	52 900
3 490	3 540	−50	2 500
3 730	3 540	190	36 100
3 540	3 540	0	0
3 925	3 540	385	148 225
3 520	3 540	−20	400
3 480	3 540	−60	3 600
		Somme = 0	Somme = 301 850

En utilisant l'équation (3.5),

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{301\,850}{11} = 27\,440{,}91$$

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Avant de poursuivre, notez que les unités associées à la variance de l'échantillon sont souvent à l'origine de confusions. Puisque les valeurs additionnées dans le calcul de la variance, $(x_i - \bar{x})^2$, sont élevées au carré, les unités associées à la variance de l'échantillon sont également élevées au carré. Par exemple, la variance d'échantillon pour les données sur la taille des classes est égale à 64 (élèves)^2. Le fait que les unités associées à la variance soient élevées au carré, rend difficile l'interprétation intuitive de la valeur numérique de la variance. Nous vous recommandons de considérer la variance comme une mesure utile pour comparer le degré de dispersion de plusieurs variables. La variable qui a la plus grande variance, a la plus grande dispersion. Il n'est pas nécessaire de chercher d'autres interprétations à la valeur de la variance.

La variance est utile pour comparer la dispersion de plusieurs variables.

Considérons à présent l'exemple des salaires initiaux des 12 diplômés d'une école de commerce, énumérés dans le tableau 3.1, pour illustrer le calcul de la variance d'échantillon. Dans la section 3.1, nous avons montré que la moyenne d'échantillon des salaires initiaux était égale à 3 940. Le calcul de la variance d'échantillon (s^2 = 27 440,91) est décrit dans le tableau 3.4.

Dans les tableaux 3.3 et 3.4, nous avons indiqué à la fois la somme des écarts par rapport à la moyenne et la somme des écarts par rapport à la moyenne au carré. Pour tout ensemble de données, la somme des écarts par rapport à la moyenne est toujours égale à zéro. Ainsi, comme indiqué dans les tableaux 3.3 et 3.4, $\sum(x_i - \bar{x}) = 0$. On obtient toujours ce résultat car les écarts positifs et les écarts négatifs s'annulent, égalisant la somme des écarts par rapport à la moyenne à zéro.

3.2.4 Écart type

L'écart type correspond à la racine carrée de la variance. En utilisant les notations adoptées pour définir la variance d'échantillon et la variance de la population, on utilise s pour noter l'écart type de l'échantillon et σ pour noter l'écart type de la population. L'écart type est déduit de la variance de la façon suivante.

► **Écart type**

$$\text{Écart type de l'échantillon} = s = \sqrt{s^2} \tag{3.8}$$

$$\text{Écart type de la population} = \sigma = \sqrt{\sigma^2} \tag{3.9}$$

L'écart type de l'échantillon s est l'estimateur de l'écart type de la population σ.

Rappelons que la variance d'échantillon pour l'échantillon des cinq classes est égale à 64. Ainsi, l'écart type de l'échantillon est égal à $s = \sqrt{64} = 8$. Pour les données sur les salaires initiaux, l'écart type de l'échantillon est égal à $s = \sqrt{27\,440,91} = 165,65$.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

L'écart type est plus facile à interpréter que la variance puisqu'il est mesuré dans les mêmes unités que les données.

Quel est l'intérêt de convertir la variance en écart type ? Rappelons que les unités associées à la variance sont élevées au carré. Par exemple, la variance d'échantillon pour les données sur les salaires initiaux des 12 diplômés d'une école de commerce est égale à 27 440,91 (dollars)^2. Puisque l'écart type est la racine carrée de la variance, les unités de la variance, dollars au carré, sont converties en dollars dans l'écart type. Ainsi, l'écart type pour les données sur les salaires initiaux est de 165,65 dollars. En d'autres termes, l'écart type est mesuré dans les mêmes unités que les données originales. Pour cette raison, l'écart type est plus facilement comparable à la moyenne et à d'autres statistiques mesurées dans les mêmes unités que les données originales.

3.2.5 Coefficient de variation

Dans certaines situations, il est intéressant d'obtenir un indicateur du rapport entre l'écart type et la moyenne. Cette mesure est appelée *coefficient de variation* et est généralement exprimée en pourcentage.

Le coefficient de variation est une mesure de dispersion relative ; il mesure l'écart type relatif à la moyenne.

► **Coefficient de variation**

$$\frac{\text{Écart type}}{\text{Moyenne}} \times 100 \qquad (3.10)$$

Pour les données sur la taille des classes, nous avons trouvé une moyenne de 44 et un écart type de 8. Le coefficient de variation est donc égal à (8/44) □ 100 % = 18,2 %. Ce qui signifie que l'écart type d'échantillon représente 18,2 % de la valeur de la moyenne. Pour les données sur les salaires initiaux, la moyenne d'échantillon est égale à 3 540, l'écart type à 165,65 ; donc le coefficient de variation est égal à [(165,65 / 3 940) □ 100] % = 4,2 %, ce qui signifie que l'écart type représente seulement 4,2 % de la moyenne de l'échantillon. En général, le coefficient de variation est une statistique utile pour comparer la dispersion de variables qui ont des écarts type et des moyennes différentes.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

REMARQUES

1. Les logiciels statistiques et les tableurs peuvent être utilisés pour calculer les statistiques descriptives présentées dans ce chapitre. Après avoir enregistré les données dans une feuille de calcul, quelques commandes simples génèrent le résultat souhaité. Nous verrons comment utiliser Minitab, Excel et StatTools pour développer ces statistiques descriptives dans les trois annexes de ce chapitre.

2. L'écart type constitue une mesure très utilisée du risque associé aux investissements boursiers et aux fonds communs de placement (site Internet de *Morningstar*, 21 juillet 2012). Il fournit une mesure des fluctuations mensuelles des rendements par rapport au rendement moyen de long terme.

3. Arrondir la valeur de la moyenne d'échantillon $\bar{x}$ et les valeurs des écarts au carré $(x_i - \bar{x})^2$ peut générer des erreurs lorsqu'une calculatrice est utilisée pour calculer la variance et l'écart type. Pour réduire les erreurs d'arrondis, nous recommandons d'utiliser au moins six chiffres après la virgule dans les calculs intermédiaires. La variance (ou l'écart type) peut ensuite être arrondie à deux chiffres après la virgule.

4. Une formule alternative pour calculer la variance d'échantillon est

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n - 1}$$

où $\sum x_i^2 = x_1^2 + x_2^2 + \ldots + x_n^2$.

EXERCICES

Méthode

23. Considérer un échantillon avec les observations suivantes : 10, 20, 12, 17 et 16. Calculer l'étendue et l'étendue interquartile.

24. Considérer un échantillon avec les observations suivantes : 10, 20, 12, 17 et 16. Calculer la variance et l'écart type.

25. Considérer un échantillon avec les observations suivantes : 27, 25, 20, 15, 30, 34, 28 et 25. Calculer l'étendue, l'étendue interquartile, la variance et l'écart type.

Applications

26. Le score d'un joueur de boules lors de six parties était respectivement de 182, 168, 184, 190, 170 et 174 points. En considérant ces données comme celles d'un échantillon, calculer les statistiques descriptives suivantes :

a) L'étendue.
b) La variance.
c) L'écart type.
d) Le coefficient de variation.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

27. Les résultats d'une recherche pour trouver les vols aller-retour les moins chers vers Altanta et Salt Lake City à partir de 14 villes américaines sont indiqués dans le tableau ci-dessous. La date de départ était le 20 juin 2012 et la date de retour le 27 juin 2012 (fichier en ligne Vols).

	Coût d'un aller-retour ($)	
Ville de départ	**Atlanta**	**Salt Lake City**
Cincinnati	340,10	570,10
New York	321,60	354,60
Chicago	291,60	465,60
Denver	339,60	219,60
Los Angeles	359,60	311,60
Seattle	384,60	297,60
Detroit	309,60	471,60
Philadelphie	415,60	618,40
Washington	293,60	513,60
Miami	249,60	523,20
San Francisco	539,60	381,60
Las Vegas	455,60	159,60
Phoenix	359,60	267,60
Dallas	333,90	458,60

a) Calculer le prix moyen d'un vol aller-retour pour Atlanta et le prix moyen d'un vol aller-retour pour Salt Lake City. Est-il moins coûteux d'aller à Atlanta qu'à Salt Lake City par avion ? Si oui, qu'est-ce qui peut expliquer cette différence ?

b) Calculer l'étendue, la variance et l'écart type des deux échantillons. Que vous apprennent ces données concernant le prix des vols à destination de ces deux villes ?

28. L'Open d'Australie est le premier des quatre tournois du Grand Chlem de tennis professionnel qui ont lieu tous les ans. Victoria Azarenka a battu Maria Sharapova et a remporté l'Open d'Australie féminin en 2012 (*Washington Post*, 27 janvier 2012). Durant le tournoi, le service de Victoria Azarenka a atteint 178 kilomètres heure. Ci-dessous sont indiquées les vitesses des services des 20 plus rapides joueuses enregistrées au cours de l'Open d'Australie 2012 (fichier en ligne Open d'Australie).

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Open d'Australie

Joueuse	Vitesse du service (km/h)	Joueuse	Vitesse du service (km/h)
S. Williams	191	G. Arn	179
S. Lisichi	190	V. Azarenka	178
M. Keys	187	Ivanovic	178
L. Hradecka	187	P. Kvitova	178
J. Gajdosova	187	M. Krajicek	178
J. Hampton	181	V. Dushevina	178
B. Mattek-Sands	181	S. Stosur	178
F. Schiavone	179	S. Cirstea	177
P. Parmentier	179	M. Barthel	177
N. Petrova	179	P. Ormaechea	177

a) Calculer la moyenne, la variance et l'écart type des vitesses de service.

b) Un échantillon similaire des vitesses de service de 20 joueuses lors du tournoi de Wimbledon en 2011 révèle une vitesse de service moyenne de 182,5 km/h. La variance et l'écart type étaient respectivement de 33,3 et 5,77. Discuter des différences entre les vitesses de service des joueuses lors de l'Open d'Australie et du tournoi de Wimbledon.

29. Le *Los Angeles Times* rapporte régulièrement l'indice de la qualité de l'air pour plusieurs régions de la Californie du Sud. Un échantillon des indices de la qualité de l'air à Pomona fournit les données suivantes : 28, 42, 58, 48, 45, 55, 60, 49 et 50.

a) Calculer l'étendue et l'étendue interquartile.

b) Calculer la variance et l'écart type d'échantillon.

c) Un échantillon des indices de la qualité de l'air à Anaheim fournit une moyenne de 48,5, une variance de 136 et un écart type de 11,66. Quelles comparaisons pouvez-vous faire entre la qualité de l'air à Pomona et à Anaheim en vous basant sur ces statistiques descriptives ?

30. Les données ci-dessous ont servi à construire les histogrammes représentant le nombre de jours nécessaires aux sociétés Dawson Supply et J. C. Clark pour honorer les commandes (cf. figure 3.2).

Délai de livraison pour la société Dawson Supply : 11 10 9 10 11 11 10 11 10 10
Délai de livraison pour la société Clark Distributors : 8 10 13 7 10 11 10 7 15 12

Utiliser l'étendue et l'écart type pour soutenir l'observation précédente selon laquelle les délais de livraison de la société Dawson Supply sont plus acceptables.

31. Les résultats de la dernière enquête Workonomix de Accounting Principal indiquent que le travailleur américain moyen dépense 1 092 dollars en café par an (*The Consumerist*, 20 janvier 2012). Pour déterminer s'il existe des écarts dans les dépenses en café selon l'âge, des échantillons de 10 consommateurs ont été sélectionnés parmi trois classes d'âge (18-34 ans, 35-44 ans et 45 ans et plus). Le montant en dollar dépensé par chaque consommateur de l'échantillon l'an dernier est fourni ci-dessous (fichier en ligne Café).

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

18-34 ans	35-44 ans	45 ans et plus
1 355	969	1 135
115	434	956
1 456	1 792	400
2 045	1 500	1 374
1 621	1 277	1 244
994	1 056	825
1 937	1 922	763
1 200	1 350	1 192
1 567	1 586	1 305
1 390	1 415	1 510

Café

a) Calculer la moyenne, la variance et l'écart type pour chacun des trois échantillons.

b) Quelles observations peuvent être faites sur la base de ces données ?

32. *Advertising Age* liste chaque année les 100 sociétés qui dépensent le plus en publicité. La société de biens de consommation Procter & Gamble arrive souvent en tête du classement, dépensant des milliards de dollars chaque année (site Internet de *Advertising Age*, 12 mars 2013). Considérez les données qui se trouvent dans le fichier en ligne Advertising. Il contient les dépenses publicitaires annuelles d'un échantillon de 20 sociétés du secteur automobile et de 20 sociétés du secteur de la grande distribution.

Advertising

a) Quelle est la dépense moyenne en publicité pour chaque secteur ?

b) Quel est l'écart type pour chaque secteur ?

c) Quelle est l'étendue des dépenses publicitaires dans chaque secteur ?

d) Quelle est l'étendue interquartile dans chaque secteur ?

e) En vous basant sur cet échantillon et vos réponses aux questions (a) à (d), commenter les différences qui apparaissent dans les dépenses publicitaires des sociétés appartenant à ces deux secteurs.

33. Les scores obtenus par un golfeur amateur lors du championnat de golf Bonita Fairways, à Bonita Springs en Floride, en 2011 et 2012 sont les suivants :

Saison 2011 :	74	78	79	77	75	73	75	77
Saison 2012 :	71	70	75	77	85	80	71	79

a) Calculer la moyenne et l'écart type pour les performances du golfeur au cours des deux années.

b) Quelle est la principale différence entre les performances de 2011 et celles de 2012 ? Quelle amélioration, s'il y en a une, peut-on voir dans les scores de 2012 ?

34. Les temps ci-dessous correspondent aux temps mis par les coureurs d'une équipe universitaire pour parcourir un mile et un quart de mile (les temps sont en minutes).

Temps pour parcourir un quart de mille :	0,92	0,98	1,04	0,90	0,99
Temps pour parcourir un mille :	4,52	4,35	4,60	4,70	4,50

Après avoir observé cet échantillon, l'un des entraîneurs a souligné que les temps de parcours d'un quart de mile étaient plus réguliers. Utiliser l'écart type et le coefficient

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

de variation pour résumer la dispersion des données. Le coefficient de variation confirme-t-il les dires de l'entraîneur ?

3.3 INDICATEURS DE LA FORME D'UNE DISTRIBUTION, MESURES DE TENDANCE RELATIVE ET DÉTECTION DES VALEURS ABERRANTES

Nous avons décrit plusieurs mesures de tendance centrale et de dispersion pour les données. En outre, il est souvent important d'avoir une idée de la forme de la distribution des données. Dans le chapitre 2, nous avons évoqué le fait qu'un histogramme constitue une représentation graphique de la distribution. L'**asymétrie** est une mesure numérique importante permettant de déterminer la forme d'une distribution.

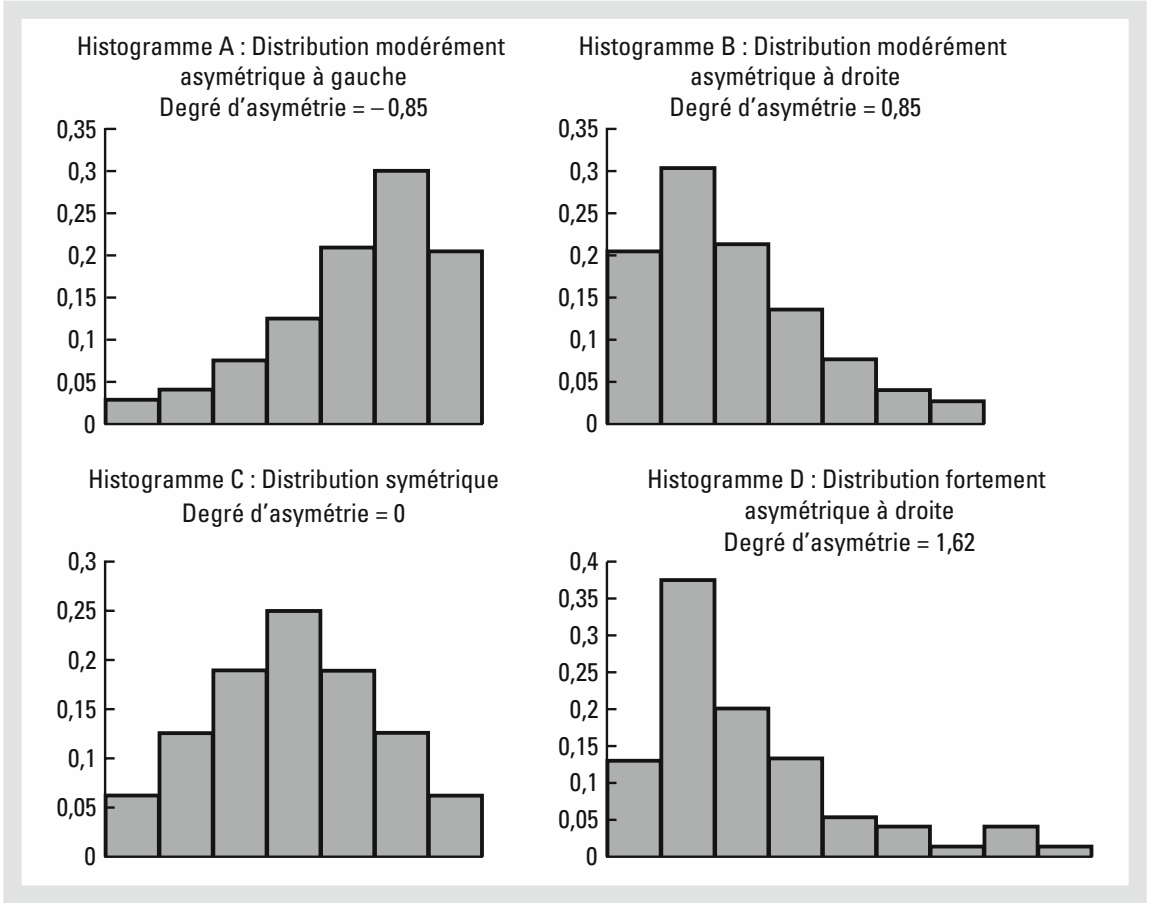


Figure 3.3 *Histogrammes illustrant le degré d'asymétrie de quatre distributions*

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

3.3.1 Forme d'une distribution

La figure 3.3 représente quatre histogrammes construits à partir de distributions de fréquence relative. Les exemples A et B illustrent des distributions modérément asymétriques. L'histogramme A est biaisé à gauche, son degré d'asymétrie est égal à -0,85. L'histogramme B est biaisé à droite, son degré d'asymétrie est égal à +0,85. L'histogramme C est symétrique, son degré d'asymétrie est nul. L'histogramme D est fortement biaisé à droite, son degré d'asymétrie est égal à + 1,62. La formule utilisée pour calculer le degré d'asymétrie est quelque peu complexe[1]. Cependant, le degré d'asymétrie peut être facilement calculé grâce aux logiciels statistiques. Lorsque les données sont biaisées à gauche, le degré d'asymétrie est négatif ; lorsqu'elles sont biaisées à droite, il est positif. Si les données sont symétriques, le degré d'asymétrie est nul.

La moyenne et la médiane d'une distribution symétrique sont égales. Lorsque les données sont positivement asymétriques (c'est-à-dire biaisées à droite), la moyenne est généralement supérieure à la médiane ; lorsque les données sont négativement asymétriques (c'est-à-dire biaisées à gauche), la moyenne est généralement inférieure à la médiane. Les données utilisées pour construire l'histogramme D correspondent aux dépenses de la clientèle d'un magasin d'habillement pour femme. Le montant moyen des achats s'élève à 77,60 dollars et le montant médian à 59,70 dollars. Les quelques achats d'un montant élevé tendent à accroître la moyenne, alors que la médiane n'est pas affectée par ces montants importants d'achat. La médiane constitue la mesure de tendance centrale la plus appropriée lorsque les données sont fortement asymétriques.

3.3.2 Variable centrée réduite

Outre les mesures de tendance centrale, de dispersion et d'asymétrie des données, la tendance relative mérite également notre attention. Les mesures de tendance relative nous permettent de déterminer l'écart d'une valeur particulière par rapport à la moyenne.

En utilisant la moyenne et l'écart type, on peut déterminer la position relative d'une observation. Supposons que nous ayons un échantillon de n observations, notées $x_1, x_2, \ldots, x_n$, dont la moyenne $\overline{x}$ et l'écart type s ont été calculés. En les associant à chaque observation x_i, on obtient une autre valeur appelée **variable centrée réduite**. L'équation (3.11) explique comment la variable centrée réduite est calculée pour chaque observation.

[1] La formule de calcul du degré d'asymétrie pour des données issues d'un échantillon est la suivante :

$$\frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \overline{x}}{s} \right)^3 .$$

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

► Variable centrée réduite *z*

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.11}$$

où z_i est la variable centrée réduite pour l'observation *i*
$\bar{x}$ est la moyenne d'échantillon
s est l'écart type d'échantillon

La variable centrée réduite *z* est souvent appelée *valeur standardisée*. La variable centrée réduite z_i peut être interprétée comme le nombre d'écarts type qui séparent x_i de la moyenne $\bar{x}$. Par exemple, $z_1 = 1,2$ signifie que x_1 se situe à 1,2 écart type au-dessus de la moyenne d'échantillon. De même, $z_2 = -0,5$ signifie que x_2 se situe à 1/2 écart type en-dessous de la moyenne d'échantillon. Les valeurs de la variable centrée réduite sont positives lorsque les observations sont supérieures à la moyenne et négatives lorsque les observations sont inférieures à la moyenne. Lorsque la valeur de la variable centrée réduite est nulle, l'observation est égale à la moyenne.

La variable centrée réduite peut être interprétée comme une mesure de tendance centrale relative des observations. Ainsi, des observations de deux ensembles de données différents, qui ont la même variable centrée réduite, peuvent être considérées comme ayant la même situation relative, c'est-à-dire comme étant placées à un même nombre d'écarts type par rapport à la moyenne.

Le processus de transformation de la valeur d'une variable en valeur centrée réduite est souvent appelé « transformation z ».

Les valeurs des variables centrées réduites pour les données sur la taille des classes (cf. section 3.1) sont énumérées dans le tableau 3.5. La moyenne d'échantillon, $\bar{x} = 44$, et l'écart type d'échantillon, $s = 8$, ont été calculés précédemment. La valeur de la variable centrée réduite de la 5e observation, égale à $-1,5$, indique que cette observation est la plus

Tableau 3.5 *Valeur de la variable centrée réduite pour les données sur la taille des classes*

Nombre d'étudiants dans la classe (x_i)	Écart par rapport à la moyenne ($x_i - \bar{x}$)	Valeur de la variable centrée réduite $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = 0,25
54	10	10/8 = 1,25
42	−2	−2/8 = −0,25
46	2	2/8 = 0,25
32	−12	−12/8 = −1,50

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

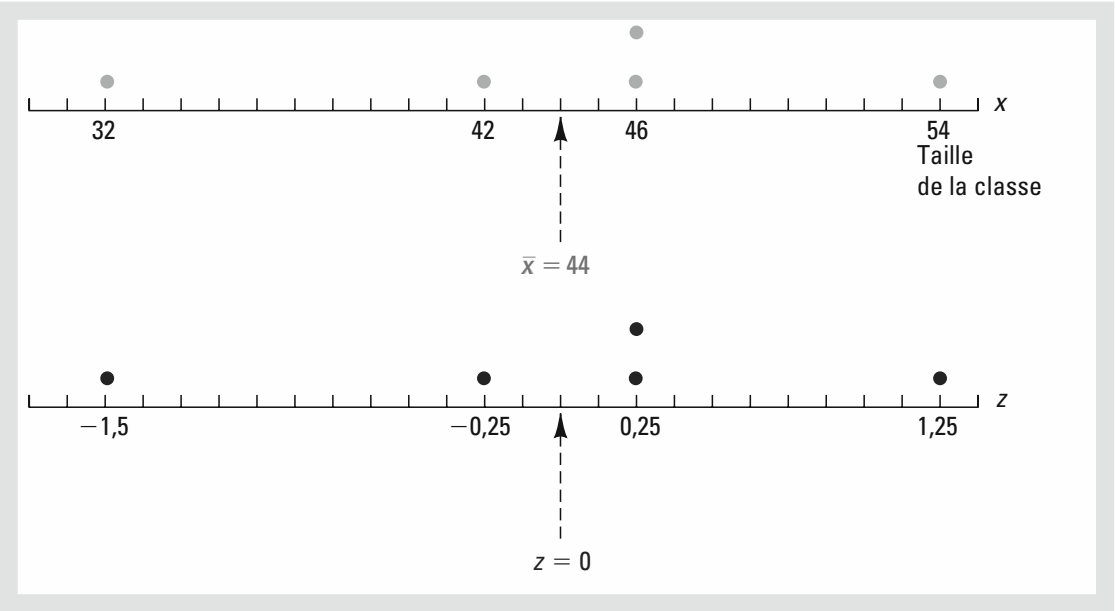


Figure 3.4 ***Diagramme de points des données sur la taille des classes et variables centrées réduites associées***

éloignée de la moyenne ; elle se situe à 1,5 écart type en-dessous de la moyenne. La figure 3.4 fournit un diagramme de points des données sur la taille des classes. Sur le second graphique sont indiquées les valeurs de la variable centrée réduite z associée aux données.

3.3.3 Le théorème de Chebyshev

Le **théorème de Chebyshev** nous permet de déterminer le pourcentage d'observations qui devraient se situer à un certain nombre d'écarts type de part et d'autre de la moyenne.

► **Théorème de Chebyshev**
Au moins $(1 - 1/z^2)$ des observations doivent se situer au plus à $|z|$ écarts type de part et d'autre de la moyenne (c'est-à-dire dans l'intervalle $[\bar{x} - zs ; \bar{x} + zs]$), avec z supérieur à 1.

Quelques conséquences de ce théorème, avec $z = 2$, 3 ou 4 écarts type, sont décrites ci-dessous.

- Au moins 0,75 ou 75 % des observations se situent, au plus, à 2 écarts type de part et d'autre de la moyenne (dans l'intervalle $[\bar{x} - 2s ; \bar{x} + 2s]$).
- Au moins 0,89 ou 89 % des observations se situent, au plus, à 3 écarts type de part et d'autre de la moyenne (dans l'intervalle $[\bar{x} - 3s ; \bar{x} + 3s]$).
- Au moins 0,94 ou 94 % des observations se situent, au plus, à 4 écarts type de part et d'autre de la moyenne (dans l'intervalle $[\bar{x} - 4s ; \bar{x} + 4s]$).

Pour illustrer le théorème de Chebyshev, supposons que la moyenne des notes de 100 étudiants d'une école de commerce, obtenues à l'examen de statistiques, soit égale à 70 et que l'écart type soit égal à 5. Combien d'étudiants ont obtenu une note

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

comprise entre 60 et 80 ? Combien d'étudiants ont obtenu une note comprise entre 58 et 82 ?

Pour les notes comprises entre 60 et 80, on peut remarquer que 60 correspond à la moyenne moins 2 fois l'écart type et 80 correspond à la moyenne plus 2 fois l'écart type. D'après le théorème de Chebyshev, au moins 75 % des observations doivent avoir une valeur distante d'au plus ± 2 écarts type de la moyenne. Aussi, au moins 75 % des étudiants doivent avoir obtenu une note comprise entre 60 et 80.

Pour les notes comprises entre 58 et 82, puisque $(58 - 70)/5 = -2,4$, 58 se situe à 2,4 écarts type en-dessous de la moyenne et puisque $(82 - 70)/5 = +2,4$, 82 se situe à 2,4 écarts type au-dessus de la moyenne. En appliquant le théorème de Chebyshev avec $z = 2,4$, on obtient

$$\left(1 - \frac{1}{z^2}\right) = \left[1 - \frac{1}{(2,4)^2}\right] = 0,826$$

Au moins 82,6 % des étudiants doivent avoir une note comprise entre 58 et 82.

Le théorème de Chebyshev exige que z soit supérieur à 1, mais z n'est pas forcément un nombre entier.

3.3.4 La règle empirique

L'un des avantages du théorème de Chebyshev est qu'il s'applique à tout ensemble de données, quelle que soit la forme de la distribution des données. En conséquence, il peut être utilisé pour toutes les distributions représentées à la figure 3.3. Dans la pratique, cependant, de nombreux ensembles de données ont une distribution en forme de cloche, ou de butte, semblable à celle représentée à la figure 3.5. Lorsque l'on pense que les données suivent une telle distribution, la **règle empirique** peut être utilisée pour déterminer le pourcentage d'observations qui se situent à une certaine distance, mesurée en écarts type, autour de la moyenne.

La règle empirique est fondée sur la distribution de probabilité normale, introduite au chapitre 6. La distribution normale est fréquemment utilisée à travers tout l'ouvrage.

► **Règle empirique**

Pour des données ayant une distribution en forme de cloche :

- Environ 68 % des observations se situent dans l'intervalle $[\overline{x} - s\ ;\ \overline{x} + s]$.
- Environ 95 % des observations se situent dans l'intervalle $[\overline{x} - 2s\ ;\ \overline{x} + 2s]$.
- Presque toutes les observations se situent dans l'intervalle $[\overline{x} - 3s\ ;\ \overline{x} + 3s]$.

Par exemple, les flacons de détergent liquide sont remplis automatiquement sur une chaîne de production. Les poids de remplissage ont fréquemment une distribution en forme de

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

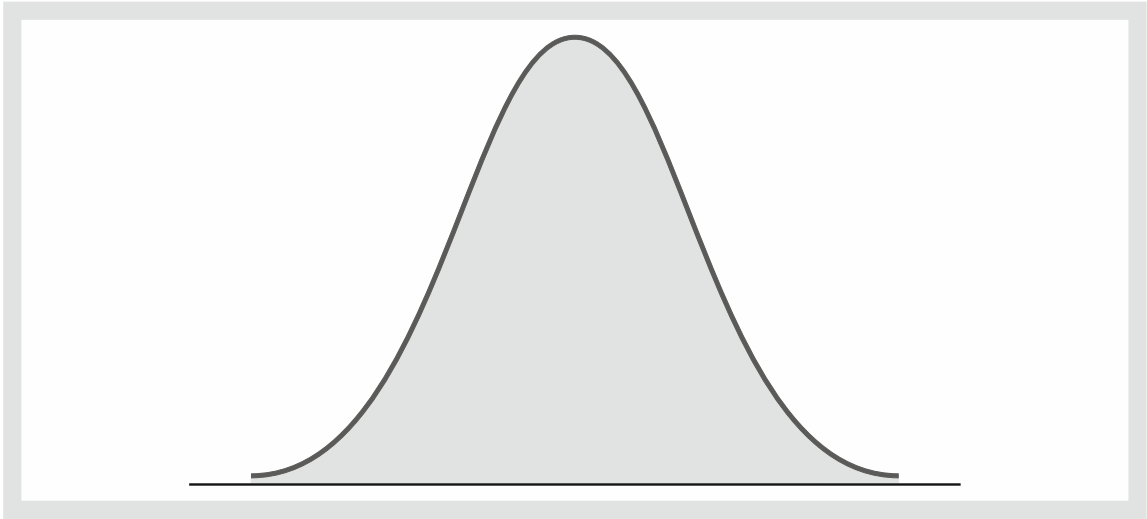


Figure 3.5 ***Une distribution symétrique en forme de cloche ou de butte***

cloche. Si le poids moyen de remplissage est de 16 onces et l'écart type de 0,25 once, on peut utiliser la règle empirique pour obtenir les conclusions suivantes.

- Approximativement 68 % des flacons remplis doivent peser entre 15,75 et 16,25 onces (la moyenne plus ou moins un écart type).
- Approximativement 95 % des flacons remplis doivent peser entre 15,50 et 16,50 onces (la moyenne plus ou moins 2 écarts type).
- Presque tous les flacons doivent peser entre 15,25 et 16,75 onces (la moyenne plus ou moins 3 écarts type).

3.3.5 Détection des valeurs aberrantes

Parfois un ensemble de données contient une ou plusieurs observations anormalement grandes ou petites. Ces valeurs extrêmes sont dites **aberrantes**. Les statisticiens expérimentés identifient les valeurs aberrantes et les reconsidèrent chacune attentivement. Une valeur aberrante peut provenir d'une erreur d'enregistrement. Si tel est le cas, elle doit être corrigée avant toute analyse supplémentaire. Une valeur aberrante peut également provenir d'une observation qui a été incluse par erreur dans l'ensemble de données ; si tel est le cas, elle doit être supprimée. Pour finir, une valeur aberrante peut être une valeur inhabituelle, correctement enregistrée et qui appartient à l'ensemble de données. Dans une telle situation, elle doit être conservée.

Les variables centrées réduites peuvent être utilisées pour identifier les valeurs aberrantes. Rappelons que la règle empirique nous permet de conclure que, pour des données distribuées en forme de cloche, presque toutes les observations sont comprises entre la moyenne et plus ou moins 3 écarts type. Ainsi, en utilisant les variables centrées réduites pour identifier les valeurs aberrantes, nous recommandons de considérer toute observation dont la variable centrée réduite z est inférieure à -3 ou supérieure à +3, comme

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

aberrante. De telles observations doivent être réexaminées avec attention pour déterminer si elles appartiennent bien à l'ensemble des données.

C'est une bonne idée de vérifier la présence de valeurs aberrantes avant de prendre des décisions en se basant sur l'analyse des données. Des erreurs sont souvent commises en collectant les données et en les enregistrant. Les valeurs aberrantes ne doivent pas nécessairement être supprimées, mais leur exactitude doit être vérifiée avant toute analyse supplémentaire des données.

Reprenons les variables centrées réduites pour les données sur la taille des classes du tableau 3.5. La valeur de $-1{,}5$, associée à la cinquième taille de classe, indique que cette observation est la plus éloignée de la taille moyenne. Cependant, cette valeur est comprise entre -3 et $+3$, limites au-delà desquelles l'observation est considérée comme aberrante. Aussi, les variables centrées réduites n'indiquent pas la présence de valeurs aberrantes dans l'ensemble de données sur la taille des classes.

Une autre approche d'identification des valeurs aberrantes est basée sur les valeurs des premier et troisième quartiles (Q_1 et Q_3) et de l'étendue interquartile (EIQ). Cette méthode consiste dans un premier temps à calculer les limites inférieure et supérieure suivante :

$$\text{Limite inférieure} = Q_1 - 1{,}5\ EIQ$$

$$\text{Limite supérieure} = Q_3 + 1{,}5\ EIQ$$

Une observation est considérée comme une valeur aberrante si sa valeur est inférieure à la limite inférieure ou supérieure à la limite supérieure. Pour les données sur les salaires mensuels initiaux figurant dans le tableau 3.1, $Q_1 = 3\,465$, $Q_3 = 3\,600$, $EIQ = 135$ et les limites inférieures et supérieures sont respectivement égales à :

$$\text{Limite inférieure} = Q_1 - 1{,}5\ EIQ = 3\,465 - 1{,}5(135) = 3\,262{,}5$$

$$\text{Limite supérieure} = Q_3 + 1{,}5\ EIQ = 3\,600 + 1{,}5(135) = 3\,802{,}5$$

En regardant les données du tableau 3.1, nous constatons qu'il n'y a aucune observation dont le salaire initial est inférieur à la limite inférieure égale à 3 262,5. Mais il y a un salaire initial, 3 925, qui est supérieur à la limite supérieure égale à 3 802,5. Aussi, 3 925 est considéré comme une valeur aberrante en utilisant cette approche alternative de détection des valeurs aberrantes.

L'approche qui utilise les premier et troisième quartiles et l'étendue interquartile pour identifier les valeurs aberrantes ne fournit pas nécessairement les mêmes résultats que l'approche basée sur les variables centrées réduites inférieures à -3 ou supérieures à $+3$. Chaque méthode séparément ou les deux simultanément peuvent être utilisées.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

REMARQUES

1. Le théorème de Chebyshev est applicable à tout ensemble de données et peut être utilisé pour déterminer le nombre minimum de données qui seront à une certaine distance, établie en écarts type, de part et d'autre de la moyenne. Si l'on pense que la distribution des données est en forme de cloche, on peut en dire plus. Par exemple, la règle empirique nous permet de dire qu'approximativement 95 % des observations seront dans l'intervalle $[\overline{x} - 2s\ ;\ \overline{x} + 2s]$; le théorème de Chebyshev nous permet seulement de conclure qu'au moins 75 % des observations seront dans cet intervalle.

2. Avant d'analyser un ensemble de données, les statisticiens effectuent habituellement diverses vérifications afin de garantir la validité des données. Dans une étude importante, il n'est pas rare de faire des erreurs en collectant les données ou en les enregistrant dans l'ordinateur. L'identification des valeurs aberrantes est l'un des outils utilisés pour vérifier la validité des données.

EXERCICES

Méthode

35. Considérer un échantillon avec les observations suivantes : 10, 20, 12, 17 et 16. Calculer les valeurs de la variable centrée réduite z pour chacune des cinq observations.

36. Considérer un échantillon de moyenne 500 et d'écart type 100. Quelle est la valeur de la variable centrée réduite z pour les observations suivantes : 520, 650, 500, 450 et 280 ?

37. Considérer un échantillon de moyenne 30 et d'écart type 5. Utiliser le théorème de Chebyshev pour déterminer le pourcentage d'observations comprises entre :

a) 20 et 40.
b) 15 et 45.
c) 22 et 38.
d) 18 et 42.
e) 12 et 48.

38. Des données, distribuées en forme de cloche, ont une moyenne de 30 et un écart type de 5. Utiliser la règle empirique pour déterminer le pourcentage d'observations comprises entre :

a) 20 et 40.
b) 15 et 45.
c) 25 et 35.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Applications

39. Les résultats d'une enquête nationale indiquent qu'en moyenne, les adultes dorment 6,9 heures par nuit. Supposons que l'écart type soit de 1,2 heure.

- **a)** Utiliser le théorème de Chebyshev pour calculer le pourcentage d'individus qui dorment entre 4,5 et 9,3 heures par nuit ?
- **b)** Utiliser le théorème de Chebyshev pour calculer le pourcentage d'individus qui dorment entre 3,9 et 9,9 heures par nuit ?
- **c)** Supposons que le nombre d'heures de sommeil suit une distribution normale (en forme de cloche). Utiliser la règle empirique pour calculer le pourcentage d'individus qui dorment entre 4,5 et 9,3 heures par nuit. Comparer ces résultats à la valeur obtenue en utilisant le théorème de Chebyshev à la question (a).

40. Le département d'information sur l'énergie indiquait que le prix moyen d'un gallon de gasoil était de 3,43 dollars (*Energy Information Administration*, juillet 2012). Supposons que l'écart type était de 0,10 dollar et que le prix du gasoil a une distribution normale (en forme de cloche).

- **a)** Quel est le pourcentage de gasoil vendu à un prix compris entre 3,33 et 3,53 dollars par gallon ?
- **b)** Quel est le pourcentage de gasoil vendu à un prix compris entre 3,33 et 3,63 dollars par gallon ?
- **c)** Quel est le pourcentage de gasoil vendu à un prix supérieur à 3,63 dollars par gallon ?

41. La moyenne nationale de l'épreuve de mathématiques d'un test d'aptitude au lycée est de 515 (*The World Almanac*, 2009). Le comité du lycée réévalue périodiquement le test de manière à ce que l'écart type soit à peu près égal à 100. Répondre aux questions suivantes en supposant la distribution des notes au test d'aptitude normale et en utilisant la règle empirique.

- **a)** Quel est le pourcentage d'élèves qui ont une note en maths supérieure à 615 ?
- **b)** Quel est le pourcentage d'élèves qui ont une note en maths supérieure à 715 ?
- **c)** Quel est le pourcentage d'élèves qui ont une note en maths comprise entre 415 et 515 ?
- **d)** Quel est le pourcentage d'élèves qui ont une note en maths comprise entre 315 et 615 ?

42. Beaucoup de familles en Californie utilisent leur abri de jardin comme bureau, studio artistique, aire de jeu ou espace de rangement supplémentaire. Supposez que le prix moyen d'un abri de jardin en bois soit de 3 100 dollars et que l'écart type soit de 1 200 dollars.

- **a)** Quelle est la valeur de la variable centrée réduite pour un abri de jardin coûtant 2 300 dollars ?
- **b)** Quelle est la valeur de la variable centrée réduite pour un abri de jardin coûtant 4 900 dollars ?
- **c)** Interpréter les valeurs des questions (a) et (b). Y a-t-il des valeurs aberrantes ?

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

d) Si le coût d'un bureau-abri de jardin construit à Albany, en Californie, s'élève à 13 000 dollars, cette valeur peut-elle être considérée comme aberrante ? Expliquer.

43. La société Florida Power & Light (FP&L) a acquis la réputation de réactiver rapidement ses installations électriques après des tempêtes. Toutefois, durant la saison des ouragans en 2004 et 2005, il est apparu que le processus historique de réparation d'urgence des systèmes électriques de la société n'était plus aussi performant (*The Wall Street Journal*, 16 janvier 2006). Les données indiquant le nombre de jours nécessaires pour rétablir le courant après sept ouragans en 2004 et 2005 sont présentées ci-dessous.

Ouragan	Nombre de jours nécessaires pour rétablir le courant
Charley	13
Frances	12
Jeanne	8
Dennis	3
Katrina	8
Rita	2
Wilma	18

À partir de cet échantillon de 7 observations, calculer les statistiques descriptives suivantes :

a) La moyenne, la médiane et le mode

b) L'étendue et l'écart type

c) L'ouragan Wilma devrait-il être considéré comme une valeur aberrante en termes de jours requis pour rétablir le courant ?

d) Les sept ouragans ont généré 10 millions d'interruptions de service électrique. Est-ce que les statistiques suggèrent que FP&L devrait revoir son processus de réparation d'urgence des systèmes électriques ? Discuter.

44. Un échantillon des résultats de 10 matchs de basket fournit les données suivantes (fichier en ligne NCAA).

Équipe gagnante	Points	Équipe perdante	Points	Écart de points
Arizona	90	Oregon	66	24
Duke	85	Georgetown	66	19
État de Floride	75	Wake Forrest	70	5
Kansas	78	Colorado	57	21
Kentucky	71	Notre Dame	63	8
Louisville	65	Tennessee	62	3
Oklahoma State	72	Texas	66	6
Purdue	76	Michigan State	70	6
Stanford	77	Southern Cal	67	10
Wisconsin	76	Illinois	56	20

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

a) Calculer la moyenne et l'écart type des points obtenus par l'équipe gagnante.

b) Supposons que la distribution des points obtenus par l'équipe gagnante pour tous les matchs soit en forme de cloche. En utilisant la moyenne et l'écart type calculés à la question (a), estimer le pourcentage de matchs au cours desquels l'équipe gagnante marque 84 points ou plus. Estimer le pourcentage de matchs au cours desquels l'équipe gagnante marque plus de 90 points.

c) Calculer la moyenne et l'écart type des données relatives à l'écart de points. Les données contiennent-elles des valeurs aberrantes ? Expliquer.

45. Selon le rapport de l'équipe Marketing de Associated Press, l'équipe des Cowboys de Dallas était l'équipe pour laquelle le ticket d'entrée à un match de la ligue nationale de football était le plus élevé (*USA Today*, 20 octobre 2009). Ci-dessous sont repris les prix moyens d'un billet pour un échantillon de 14 équipes de la ligue nationale de football (fichier en ligne Billets Ligue nationale de foot).

Équipe	Prix du billet (dollars)	Équipe	Prix du billet (dollars)
Atlanta Falcons	72	Green Bay Packers	63
Buffalo Bills	51	Indianapolis Colts	83
Carolina Panthers	63	New Orleans Saints	62
Chicago Bears	88	New York Jets	87
Cleveland Browns	55	Pittsburgh Steelers	67
Dallas Cowboys	160	Seattle Seahawks	61
Denver Broncos	77	Tennessee Titans	61

a) Quel est le prix moyen du billet ?

b) L'année précédente, le prix moyen du billet était de 72,20 dollars. Quelle a été l'augmentation moyenne du prix d'un billet en pourcentage sur un an ?

c) Calculer le prix médian du billet.

d) Calculer le premier et le troisième quartile.

e) Calculer l'écart type.

f) Quelle est la valeur de la variable centrée réduite associée au prix du billet des Dallas Cowboys ? Ce prix devrait-il être considéré comme une valeur aberrante ? Expliquer.

3.4 RÉSUMÉ EN CINQ CHIFFRES ET BOÎTES-À-PATTES

Les résumés statistiques et les graphiques faciles à représenter basés sur ces résumés statistiques peuvent être utilisés rapidement pour résumer de grande quantité de données. Dans cette section, nous montrons comment développer des résumés en cinq chiffres et des « boîtes-à-pattes » (*box plots*, en anglais) pour identifier plusieurs caractéristiques d'un vaste ensemble de données.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

3.4.1 Résumé en cinq chiffres

Dans un **résumé en cinq chiffres**, les cinq valeurs suivantes sont utilisées pour résumer les données.

1. Valeur la plus petite
2. Premier quartile (Q_1)
3. Médiane (Q_2)
4. Troisième quartile (Q_3)
5. Valeur la plus élevée

La façon la plus simple de construire un résumé en cinq chiffres est tout d'abord d'ordonner les observations de façon croissante. Ensuite, il est facile d'identifier la plus petite valeur, les trois quartiles et la plus grande valeur. Les salaires mensuels initiaux, présentés dans le tableau 3.1, pour un échantillon de 12 diplômés d'une école de commerce, sont réécrits ici en ordre croissant.

3 710	3 755	3 850	3 880	3 880	3 890	3 920	3 940	3 950	4 050	4 130	4 325
		$Q_1 = 3\,465$			$Q_2 = 3\,905$ (Médiane)			$Q_3 = 4\,000$			

La médiane égale à 3905 et les quartiles, Q_2 = 3 865 et Q_3 = 4 000, ont déjà été calculés (cf. section 3.1). La valeur la plus petite des données est 3 710, la plus grande 4 325. Ainsi le résumé en cinq chiffres pour les données sur les salaires comporte les chiffres suivants : 3 710, 3 865, 3 905, 4 000, 4 325. Approximativement un quart (25 %) des observations sont comprises entre deux nombres adjacents du résumé en cinq chiffres.

3.4.2 Boîte-à-pattes

La **boîte-à-pattes** est une illustration des données, basée sur le résumé en cinq chiffres. La médiane et les quartiles Q_1 et Q_3 sont les éléments clés de la construction d'une boîte-à-pattes. L'étendue interquartile, $EIQ = Q_3 - Q_1$, est également utilisée. La figure 3.6 correspond à la boîte-à-pattes obtenue pour les données sur les salaires mensuels initiaux. Les étapes de la construction d'une boîte-à-pattes sont détaillées ci-dessous.

1. On dessine une boîte ; les 1er et 3e quartiles constituent les deux extrémités de la boîte. Pour les données sur les salaires, Q_1 = 3 865 et Q_3 = 4 000. La boîte contient 50 % des observations centrales.
2. Une ligne verticale est tracée dans la boîte au niveau de la médiane (3 905 pour les données sur le salaire).
3. On fixe les limites en utilisant l'étendue interquartile, $EIQ = Q_3 - Q_1$. Les limites de la boîte-à-pattes sont situées aux points ($Q_1 - 1{,}5\,EIQ$) et ($Q_3 + 1{,}5\,EIQ$). Pour les données sur les salaires, $EIQ = Q_3 - Q_1 = 135$. Ainsi, les limites sont $3\,865 - 1{,}5(135) = 3\,662{,}5$ et $4\,000 + 1{,}5(135) = 4\,202{,}5$. Les valeurs situées hors de ces limites sont considérées comme des *valeurs aberrantes*.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

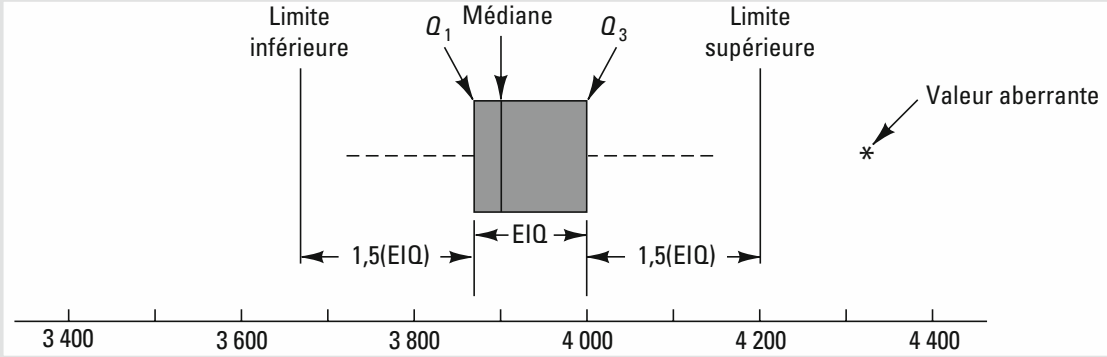


Figure 3.6 ***Boîte-à-pattes obtenue à partir des données relatives aux salaires mensuels initiaux des jeunes diplômés, avec matérialisation des limites inférieure et supérieure par des lignes***

4. Les lignes en pointillés sur la figure 3.6 constituent les pattes. Les pattes sont tracées depuis la fin de la boîte jusqu'à la plus petite valeur des observations comprises entre les limites calculées à l'étape 3, d'un côté, et jusqu'à la plus grande valeur des observations comprises entre les limites calculées à l'étape 3, de l'autre côté. Ainsi les pattes vont jusqu'à 3 710 et 4 130 de part et d'autre de la boîte.
5. Enfin, les valeurs aberrantes sont représentées par le symbole *. Dans la figure 3.6, on constate la présence d'une valeur aberrante, l'observation 4 325.

La boîte-à-pattes est un moyen de visualiser plusieurs caractéristiques d'un ensemble de données.

Sur la figure 3.6, nous avons représenté les limites par des lignes, de manière à expliciter les calculs et à bien visualiser leur position pour les données sur les salaires. Bien que ces limites soient toujours calculées, elles ne sont généralement pas représentées sur le graphique de la boîte-à-pattes. La figure 3.7 illustre l'apparence habituelle d'une boîte-à-pattes, pour les données sur les salaires.

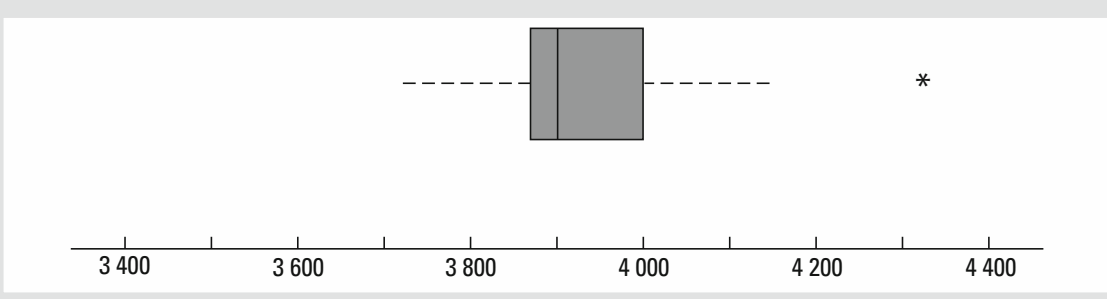


Figure 3.7 ***Boîte-à-pattes obtenue à partir des données sur les salaires initiaux***

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Salaires par discipline

Pour comparer les salaires mensuels initiaux des jeunes diplômés par discipline, un échantillon de 111 jeunes diplômés a été sélectionné (fichier en ligne Salaires par discipline). La discipline et le salaire mensuel initial ont été enregistrés pour chaque diplômé. La figure 3.8 représente les boîtes-à-pattes obtenues avec Minitab pour les diplômés en comptabilité, finance, systèmes d'information, management et marketing. Notez que la discipline est indiquée sur l'axe horizontal et que chaque boîte-à-pattes est représentée verticalement au-dessus de la discipline considérée. Représenter ainsi les boîtes-à-pattes est un excellent moyen graphique pour comparer plusieurs groupes.

Quelles observations pouvez-vous faire à propos des salaires mensuels initiaux par discipline à partir des boîtes-à-pattes représentées sur la figure 3.8 ? Nous pouvons en particulier relever les observations suivantes :

- Les salaires les plus élevés sont observés au sein des diplômés en comptabilité ; les salaires les plus faibles au sein des diplômés en management et marketing.
- Les salaires médians les plus élevés sont observés au sein des diplômés en comptabilité et en systèmes d'information ; ils sont par ailleurs similaires. Vient ensuite le salaire médian des diplômés en finance, puis en marketing et en management.
- Des valeurs aberrantes (salaires très élevés) apparaissent pour les diplômés en comptabilité, finance et marketing.
- Les salaires des diplômés en finance sont les moins variables, alors que les salaires des comptables présentent une forte dispersion.

Peut-être voyez-vous d'autres commentaires à faire à partir de ces boîtes-à-pattes.

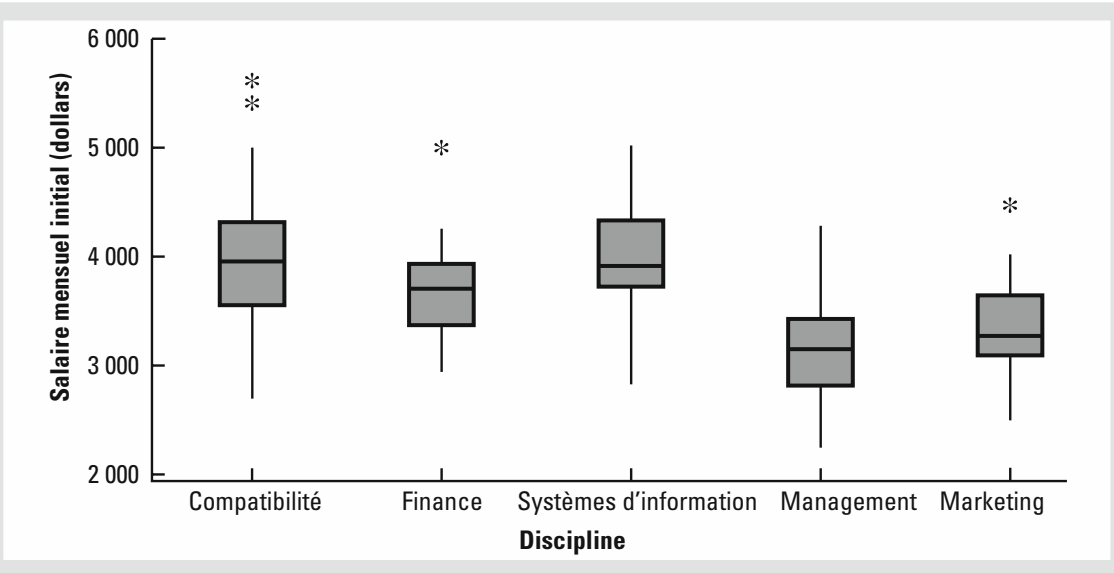


Figure 3.8 ***Boîtes-à-pattes obtenues à partir des données sur les salaires initiaux par discipline avec Minitab***

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

REMARQUES

Nous explicitons la procédure de construction d'une boîte-à-pattes grâce à Minitab dans l'annexe 3.1. La boîte-à-pattes obtenue est semblable à celle représentée à la figure 3.7 mais est dessinée verticalement.

EXERCICES

Méthode

46. Considérer un échantillon avec les observations suivantes : 27, 25, 20, 15, 30, 34, 28 et 25. Fournir le résumé en cinq chiffres de ces données.

47. Construire la boîte-à-pattes pour les données de l'exercice 46.

48. Fournir le résumé en cinq chiffres et construire la boîte-à-pattes pour les données suivantes : 5, 15, 18, 10, 8, 12, 16, 10, 6.

49. Les premier et troisième quartiles d'un ensemble de données sont respectivement égaux à 42 et 50. Calculer les limites inférieure et supérieure. Peut-on considérer la valeur 65 comme une valeur aberrante ?

Applications

50. La ville de Naples en Floride organise chaque année en janvier un semi-marathon (21,1 km). L'évènement attire des coureurs venant des quatre coins des États-Unis et du monde entier. En janvier 2009, 22 hommes et 31 femmes âgés de 19 à 24 ans ont participé à la course. Les temps de course en minutes de ces coureurs sont fournis ci-dessous (*Naples Daily News*, 19 janvier 2009). Les temps sont fournis par ordre d'arrivée (fichier en ligne Coureurs).

Arrivée	Homme	Femme	Arrivée	Homme	Femme	Arrivée	Homme	Femme
1	65,30	109,03	11	109,05	123,88	21	143,83	136,75
2	66,27	111,22	12	110,23	125,78	22	148,70	138,20
3	66,52	111,65	13	112,90	129,52	23		139,00
4	66,85	111,93	14	113,52	129,87	24		147,18
5	70,87	114,38	15	120,95	130,72	25		147,35
6	87,18	118,33	16	127,98	131,67	26		147,50
7	96,45	121,25	17	128,40	132,03	27		147,75
8	98,52	122,08	18	130,90	133,20	28		153,88
9	100,52	122,48	19	131,80	133,50	29		154,83
10	108,18	122,62	20	138,63	136,57	30		189,27
						31		189,28

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

a) George Towett de Marietta, en Géorgie, est arrivé le premier chez les hommes et Lauren Wald de Gainesville en Floride a terminé à la première place chez les femmes. Comparer les temps des vainqueurs masculin et féminin. Si les 53 coureurs hommes et femmes avaient concouru dans le même groupe, à quelle place Lauren aurait-elle été classée ?

b) Quel est le temps médian des coureurs de sexe masculin et des coureurs de sexe féminin ? Comparer les coureurs des deux sexes sur la base de leurs temps médians.

c) Fournir un résumé en cinq chiffres pour les hommes et un pour les femmes.

d) Y a-t-il des valeurs aberrantes ?

e) Construire la boîte-à-pattes pour chaque groupe. Qui des hommes ou des femmes ont la plus grande dispersion dans les temps de course ? Expliquer

51. Les ventes annuelles, en millions de dollars, de 21 entreprises pharmaceutiques sont fournies ci-dessous.

8 408	1 374	1 872	8 879	2 459	11 413
608	14 138	6 452	1 850	2 818	1 356
10 498	7 478	4 019	4 341	739	2 127
3 653	5 794	8 305			

a) Fournir le résumé en cinq chiffres.

b) Calculer les limites inférieure et supérieure.

c) Les données contiennent-elles des valeurs aberrantes ?

d) Les ventes de Johnson & Johnson sont les plus importantes de la liste ; elles s'élèvent à 14 138 millions de dollars. Supposez qu'il y ait eu une erreur lors de l'enregistrement des données et que le chiffre 41 138 ait été enregistré. Est-ce que la méthode de détection des valeurs aberrantes utilisée à la question (c) permet d'identifier cette erreur et de corriger les données ?

e) Dessiner une boîte-à-patte.

52. Le magazine *Consumer Reports* fournissait les taux de satisfaction des consommateurs vis-à-vis des services de téléphonie mobile proposés par AT&T, Sprint, T-Mobile et Verizon dans les principales zones urbaines américaines. La note attribuée à chaque service reflète la satisfaction générale des clients au regard de plusieurs facteurs tels que le tarif, les problèmes de connexion, les appels manqués, les interférences et le service client. Une échelle de notation de 0 à 100 a été utilisée, 0 indiquant une insatisfaction totale et 100 une satisfaction totale. Les notes attribuées aux quatre opérateurs de téléphonie mobile dans 20 zones urbaines (fichier en ligne Service mobile) sont fournies ci-dessous (*Consumer Reports*, janvier 2009).

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Zone urbaine	AT&T	Sprint	T-Mobile	Verizon
Atlanta	70	66	71	79
Boston	69	64	74	76
Chicago	71	65	70	77
Dallas	75	65	74	78
Denver	71	67	73	77
Detroit	73	65	77	79
Jacksonville	73	64	75	81
Las Vegas	72	68	74	81
Los Angeles	66	65	68	78
Miami	68	69	73	80
Minneapolis	68	66	75	77
Philadelphie	72	66	71	78
Phoenix	68	66	76	81
San Antonio	75	65	75	80
San Diego	69	68	72	79
San Francisco	66	69	73	75
Seattle	68	67	74	77
Saint Louis	74	66	74	79
Tampa	73	63	73	79
Washington	72	68	71	76

Service mobile

a) Considérez tout d'abord T-Mobile. Quelle est sa note médiane ?

b) Développer un résumé en cinq chiffres pour le service proposé par T-Mobile.

c) Y a-t-il des valeurs aberrantes dans les notes attribuées à T-Mobile ? Expliquer.

d) Répéter les questions (b) et (c) pour les trois autres opérateurs.

e) Représenter la boîte-à-pattes pour les quatre services de téléphonie mobile sur un graphique. Discuter de ce qu'une comparaison des boîtes-à-pattes nous apprend des quatre services. Quel service le magazine *Consumer Reports* recommandait-il comme étant le meilleur au regard de la satisfaction globale des clients ?

53. Les Phillies de Philadelphie ont battu les Bay Rays de Tampa 4 à 3 et ont gagné la coupe de la ligue principale de baseball lors de la coupe du monde en 2008. Plus tôt dans la saison, lors des jeux décisifs de la coupe de la ligue de baseball, les Phillies de Philadelphie avaient battu les Dodgers de Los Angeles et gagné le championnat national, alors que les Bay Rays de Tampa battaient les Red Sox de Boston et gagnaient le championnat américain. Le fichier Salaires MLB contient les salaires des 28 joueurs de chacune de ces quatre équipes (Base de données des salaires de *USA Today*, octobre 2008). Les données, exprimées en milliers de dollars, ont été ordonnées du plus élevé au plus faible salaire pour chaque équipe.

Salaires MLB

a) Analyser les salaires des champions mondiaux de Philadelphie. Quel est le revenu total pour l'équipe ? Quel est le salaire médian ? Fournir le résumé en cinq chiffres.

b) Y a-t-il des valeurs aberrantes dans les données sur les salaires des Phillies de Philadelphie ? Si oui, combien et quels sont les montants de ces salaires aberrants ?

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

c) Quel est le salaire moyen pour chacune des trois autres équipes ? Fournir le résumé en cinq chiffres pour chaque équipe et identifier les valeurs aberrantes.

d) Construire la boîte-à-pattes des salaires pour les quatre équipes. Quelle en est votre interprétation ? Est-ce que c'est l'équipe, parmi les quatre étudiées, qui a les salaires les plus élevés qui a gagné le championnat national et la coupe du monde ?

54. Le bureau des statistiques sur le transport surveille toutes les entrées et sorties du territoire américain aux différents postes frontières situés le long des frontières entre les États-Unis et le Canada et entre les États-Unis et le Mexique. Le fichier en ligne Frontières contient les données sur le nombre de véhicules personnels qui passent les frontières (arrondis au millier le plus proche) aux 50 postes frontières les plus empruntés durant le mois d'août (site Internet du département américain des transport, 28 février 2013).

a) Quels sont les nombres moyen et médian de véhicules se présentant à ces postes frontières ?

b) Quel est le premier quartile ? Le troisième quartile ?

c) Fournir le résumé en cinq chiffres

d) Y a-t-il des valeurs aberrantes ? Construire une boîte-à-pattes.

3.5 MESURES DE LA RELATION ENTRE DEUX VARIABLES

Jusqu'à présent, nous avons étudié les méthodes numériques utilisées pour résumer les données d'une variable à un moment donné. Souvent un responsable s'intéresse à la relation entre deux variables. Dans cette section, nous présenterons la covariance et la corrélation, deux mesures descriptives de la relation entre deux variables.

Tableau 3.6 ***Données d'échantillon pour le magasin de 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

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Reconsidérons tout d'abord l'exemple du magasin d'équipement hi-fi de San Francisco, présenté dans la section 2.4. Le responsable du magasin s'intéresse à la relation qui pourrait exister entre le nombre de spots publicitaires diffusés au cours d'un week-end et les ventes effectuées la semaine suivante. Le tableau 3.6 regroupe un échantillon de données sur les ventes, exprimées en centaines de dollars. Il fournit 10 observations ($n = 10$), une par semaine. Le nuage de points représenté à la figure 3.9 dévoile une relation positive, un plus important volume de vente (y) étant associé à un plus grand nombre de spots publicitaires (x). Le nuage de points suggère donc qu'une ligne droite caractérise la relation. Nous introduisons dans cette section la covariance en tant que mesure descriptive de la relation linéaire entre deux variables.

3.5.1 Covariance

Pour un échantillon de taille n composé des observations (x_1, y_1), (x_2, y_2), etc., la covariance de l'échantillon est définie par :

► **Covariance de l'échantillon**

$$s_{xy} = \frac{\sum (x_i - \bar{x})\ (y_i - \bar{y})}{n - 1} \tag{3.12}$$

Dans cette formule, à chaque observation x_i est associée une observation y_i. Les produits obtenus en multipliant l'écart de chaque observation x_i par rapport à sa moyenne d'échantillon $\bar{x}$, par l'écart entre l'observation y_i qui lui est associée, et sa moyenne d'échantillon $\bar{y}$, sont sommés. Cette somme est ensuite divisée par $n - 1$.

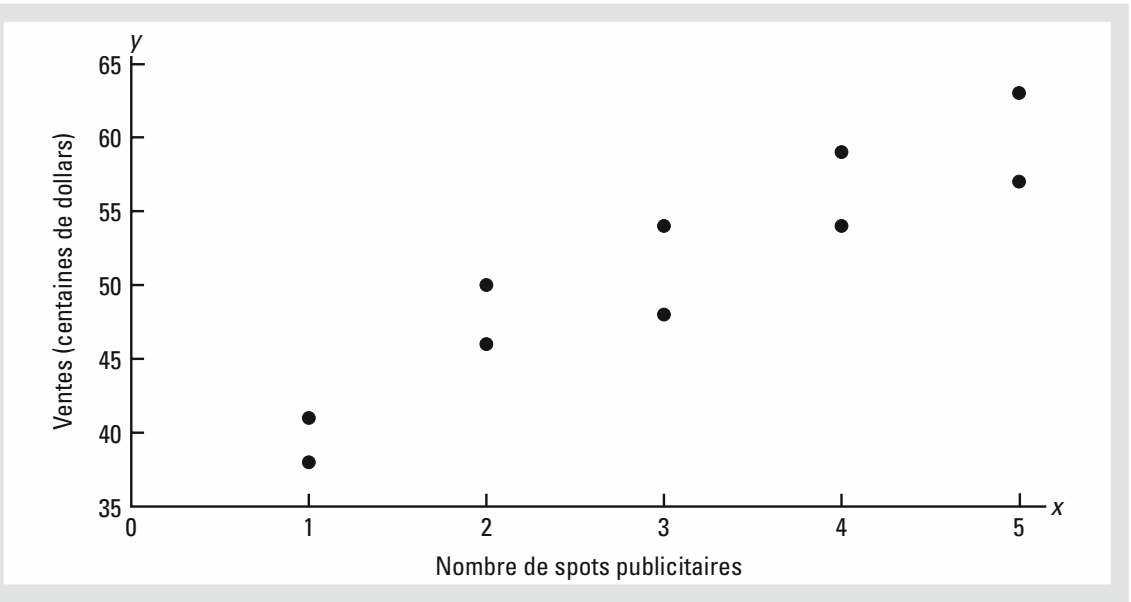


Figure 3.9 *Nuage de points pour le magasin de hi-fi*

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Tableau 3.7 ***Calcul de la covariance d'échantillon***

(x_i)	(y_i)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
2	50	−1	−1	1
5	57	2	6	12
1	41	−2	−10	20
3	54	0	3	0
4	54	1	3	3
1	38	−2	−13	26
5	63	2	12	24
3	48	0	−3	0
4	59	1	8	8
2	46	−1	−5	5
Total = 30	Total = 510	Total = 0	Total = 0	Total = 99

Pour mesurer la robustesse de la relation linéaire entre le nombre de spots publicitaires x et le volume des ventes y dans le problème du magasin d'équipement hi-fi, on utilise la formule (3.12) pour calculer la covariance de l'échantillon. Les calculs de $\sum(x_i - \bar{x})(y_i - \bar{y})$ sont détaillés dans le tableau 3.7. Notez que $\bar{x} = 30/10 = 3$ et $\bar{y} = 510/10 = 51$. En utilisant la formule (3.12), on obtient une covariance de l'échantillon égale à

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{9} = 11$$

La formule de calcul de la covariance pour une population de taille N est similaire à la formule (3.12) mais nous utilisons des notations différentes pour indiquer que nous travaillons avec la population entière.

► **Covariance de la population**

$$\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N} \qquad (3.13)$$

Dans la formule (3.13), nous utilisons la notation μ_x pour décrire la moyenne de la population de la variable x et μ_y pour décrire la moyenne de la population de la variable y. La covariance de la population σ_{xy} est définie pour une population de taille N.

3.5.2 Interprétation de la covariance

Pour interpréter plus facilement la covariance d'échantillon, considérons la figure 3.10. La figure est semblable au nuage de points présenté à la figure 3.9, avec une ligne verticale en pointillés tracée au point $\bar{x} = 3$ et une ligne horizontale en pointillés tracée au point

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

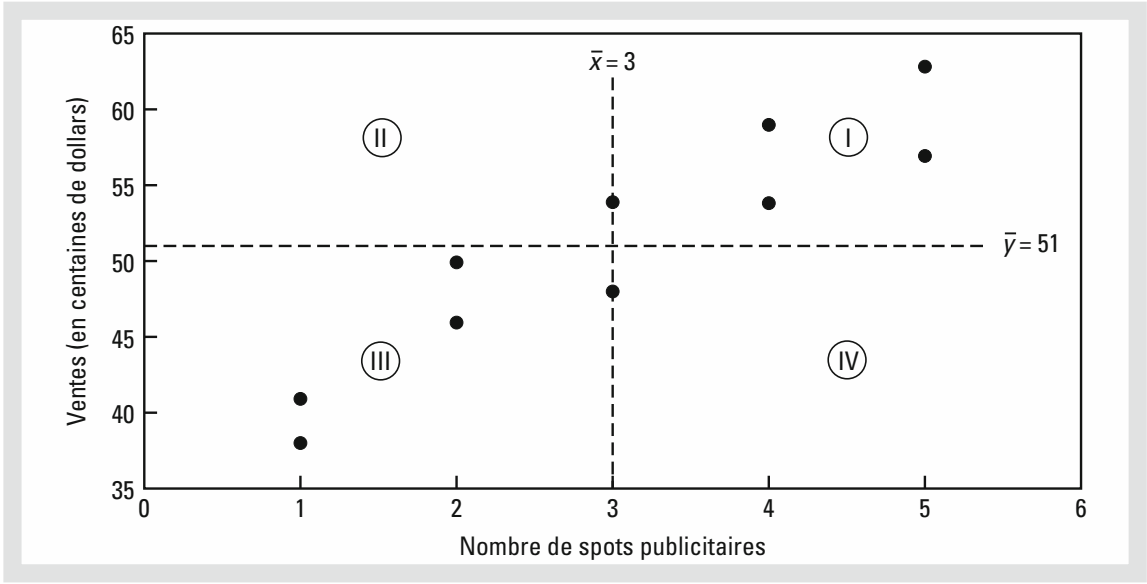


Figure 3.10 ***Partition du nuage de points pour le magasin de hi-fi***

$\bar{y} = 51$. Le graphique est maintenant découpé en quatre cadrans. Les points situés dans le cadran I sont caractérisés par une valeur x_i supérieure à $\bar{x}$ et une valeur y_i supérieure à $\bar{y}$; les points situés dans le cadran II sont caractérisés par une valeur x_i inférieure à $\bar{x}$ et une valeur y_i supérieure à $\bar{y}$; etc. Ainsi, la valeur de $(x_i - \bar{x})(y_i - \bar{y})$ est positive pour les points situés dans les cadrans I et III et négative pour les points situés dans les cadrans II et IV.

Si la valeur de s_{xy} est positive, les points qui ont la plus grande influence sur s_{xy} se trouvent dans les cadrans I et III. Ainsi, une valeur positive de s_{xy} révèle une relation linéaire positive entre x et y ; c'est-à-dire, lorsque la valeur de x augmente, la valeur de y augmente. Si la valeur de s_{xy} est négative, ce sont les points situés dans les cadrans II et IV qui ont la plus grande influence sur s_{xy}. Ainsi, une valeur négative de s_{xy} révèle une relation linéaire négative entre x et y ; c'est-à-dire, lorsque la valeur de x augmente, la valeur de y diminue. Si les points sont répartis de façon uniforme entre les quatre cadrans, la valeur de s_{xy} sera proche de zéro, indiquant l'absence d'une relation linéaire entre x et y. La figure 3.11 illustre les différentes valeurs que peut prendre s_{xy} pour trois types de nuage de points.

La covariance est une mesure de la relation linéaire entre deux variables.

En se référant de nouveau à la figure 3.10, nous remarquons que le nuage de points obtenu avec les données sur le magasin d'équipement hi-fi a la même forme que celui représenté en haut de la figure 3.11. Comme l'on s'y attendait, la valeur de la covariance indique une relation linéaire positive, avec $s_{xy} = 11$.

D'après la discussion précédente, une valeur positive élevée de la covariance semble indiquer une forte relation positive et une valeur négative élevée de la

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

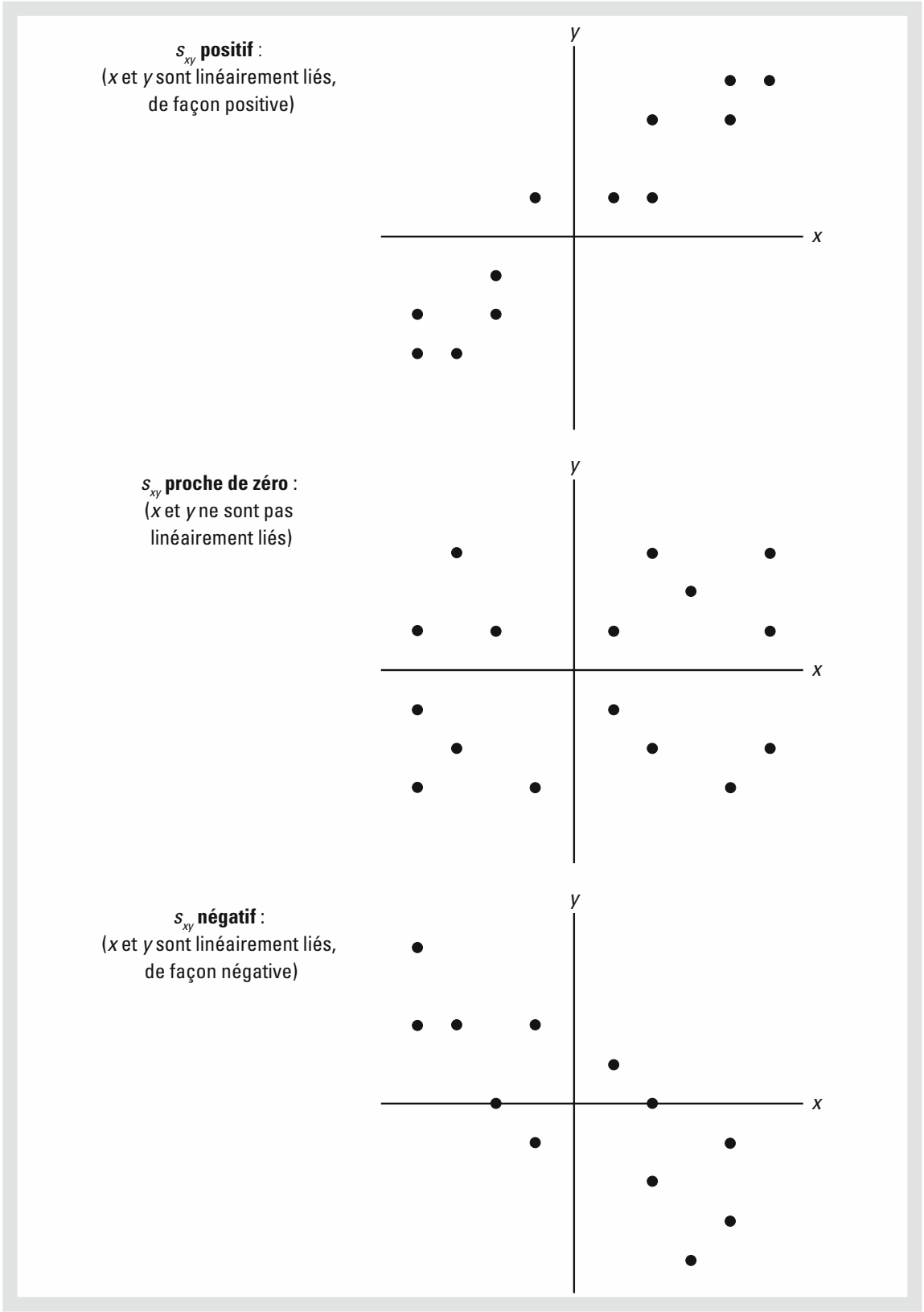


Figure 3.11 ***Interprétation de la covariance d'échantillon***

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

covariance semble indiquer une forte relation négative. Cependant, l'utilisation de la covariance comme mesure de la robustesse de la relation linéaire présente un inconvénient : la valeur de la covariance dépend de l'unité de mesure des variables x et y. Par exemple, supposons que nous nous intéressions à la relation entre la taille, x, et le poids, y, d'individus. La robustesse de la relation devrait être la même que la taille soit mesurée en mètres ou en centimètres. Cependant, lorsque la taille est mesurée en centimètres, les valeurs numériques $(x_i - \bar{x})$ sont supérieures à celles obtenues en mesurant la taille en mètres. Ainsi, lorsque la taille est mesurée en centimètres, on obtient une valeur supérieure au numérateur $\sum(x_i - \bar{x})(y_i - \bar{y})$ dans la formule (3.12) - et donc une covariance supérieure - alors qu'en fait, il n'y a pas de différence dans la relation. Le **coefficient de corrélation** est une mesure de la relation entre deux variables qui n'est pas exposée à ce type de problème.

3.5.3 Coefficient de corrélation

Pour un échantillon de données, le coefficient de corrélation de Pearson est défini par :

► **Coefficient de corrélation de Pearson : Données d'échantillon**

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3.14}$$

où

r_{xy} correspond au coefficient de corrélation de l'échantillon
s_{xy} correspond à la covariance de l'échantillon
s_x correspond à l'écart type d'échantillon de x
s_y correspond à l'écart type d'échantillon de y

D'après la formule (3.14), le coefficient de corrélation de Pearson pour un échantillon de données (appelé plus simplement coefficient de corrélation de l'échantillon) est calculé en divisant la covariance de l'échantillon par le produit des écarts type d'échantillon de x et de y.

Calculons le coefficient de corrélation d'échantillon pour l'exemple du magasin d'équipement hi-fi. En utilisant les données du tableau 3.6, nous pouvons calculer les écarts type des deux variables.

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{20}{9}} = 1{,}49$$

$$s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{566}{9}} = 7{,}93$$

Puisque $s_{xy} = 11$, le coefficient de corrélation est égal à

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{(1{,}49)(7{,}93)} = 0{,}93$$

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

La formule de calcul du coefficient de corrélation pour une population, noté ρ_{xy}, est donnée ci-dessous.

► **Coefficient de corrélation de Pearson : données issues d'une population**

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.15}$$

où

ρ_{xy} correspond au coefficient de corrélation de la population
σ_{xy} correspond à la covariance de la population
σ_x correspond à l'écart type de x, au niveau de la population
σ_y correspond à l'écart type de y, au niveau de la population

Le coefficient de corrélation de l'échantillon r_{xy} est l'estimateur du coefficient de corrélation de la population ρ_{xy}.

Le coefficient de corrélation de l'échantillon r_{xy} fournit une estimation du coefficient de corrélation de la population ρ_{xy}.

3.5.4 Interprétation du coefficient de corrélation

Considérons, tout d'abord un exemple simple pour illustrer une relation parfaitement linéaire et positive. Le nuage de points de la figure 3.12 décrit la relation entre x et y, basée sur les données suivantes.

x_i	y_i
5	10
10	30
15	50

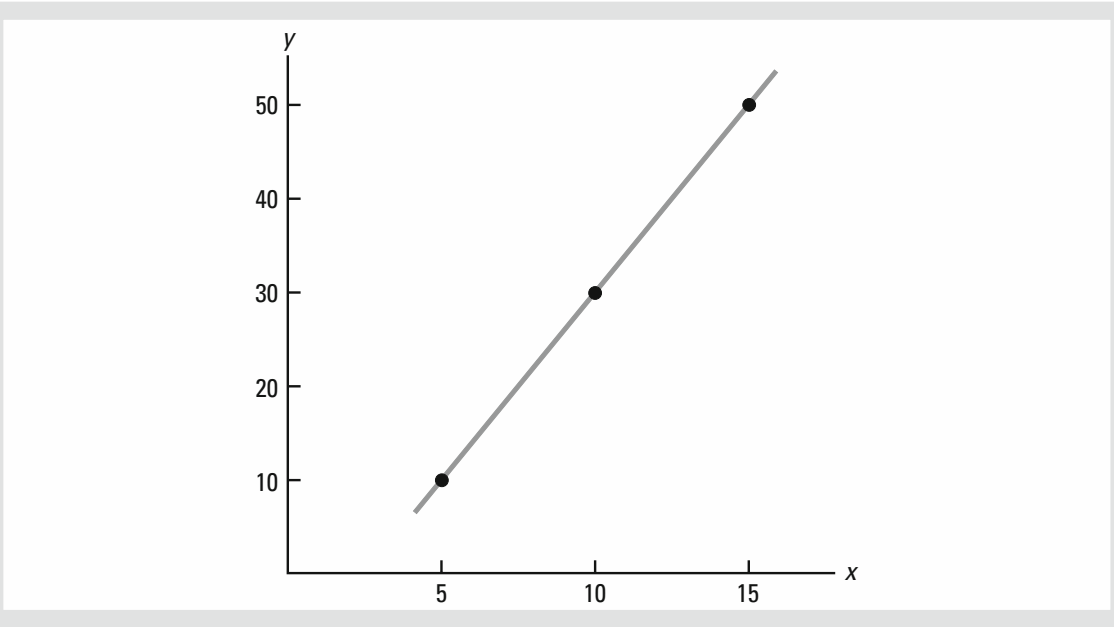


Figure 3.12 ***Nuage de points décrivant une relation positive parfaitement linéaire***

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

La ligne droite tracée entre les trois points illustre une relation parfaitement linéaire et positive entre x et y. Pour appliquer l'équation (3.14) et calculer le coefficient de corrélation de l'échantillon, il est nécessaire de calculer tout d'abord s_{xy}, s_x et s_y. Certains calculs sont présentés dans le tableau 3.8. En les utilisant, on obtient

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{200}{2} = 100$$

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{50}{2}} = 5$$

$$s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{800}{2}} = 20$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{100}{5(20)} = 1$$

Le coefficient de corrélation de l'échantillon est égal à 1.

> Le coefficient de corrélation varie entre – 1 et + 1. Des valeurs proches de – 1 ou de + 1 révèlent une forte relation linéaire. Plus le coefficient est proche de zéro, plus la relation est faible.

En général, si tous les points d'un ensemble de données sont alignés sur une droite de pente positive, la valeur du coefficient de corrélation de l'échantillon est +1 ; en d'autres termes, un coefficient de corrélation de +1 correspond à une relation parfaitement linéaire et positive entre x et y. À l'inverse, si les points d'un ensemble de données sont alignés sur une droite de pente négative, la valeur du coefficient de corrélation est – 1 ; en d'autres termes, un coefficient de corrélation de – 1 correspond à une relation parfaitement linéaire et négative entre x et y.

Supposons maintenant qu'un ensemble de données particulier révèle une relation linéaire positive entre x et y mais que cette relation n'est pas parfaitement linéaire.

Tableau 3.8 *Calculs utilisés pour déterminer le coefficient de corrélation de l'échantillon*

x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
5	10	– 5	25	– 20	400	100
10	30	0	0	0	0	0
15	50	5	25	20	400	100
Total = 30	Total = 90	Total = 0	Total = 50	Total = 0	Total = 800	Total = 200
$\bar{x} = 10$ $\bar{y} = 30$						

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

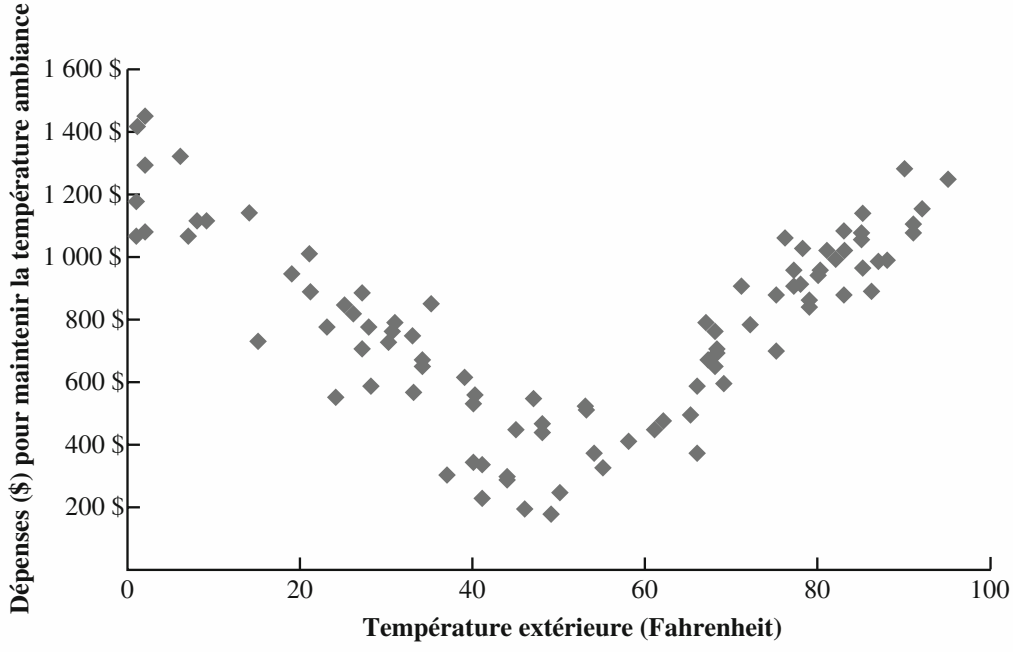
La valeur de r_{xy} sera inférieure à 1, indiquant que les points du nuage de points ne sont pas tous alignés sur une même droite. Plus les points dévient d'une relation positive parfaitement linéaire, plus la valeur de r_{xy} sera petite. Une valeur de r_{xy} égale à zéro indique l'absence de relation linéaire entre x et y, et des valeurs de r_{xy} proches de zéro révèlent une faible relation linéaire.

Pour les données sur le magasin d'équipement hi-fi, rappelons que $r_{xy} = 0{,}93$. Ainsi, on peut conclure qu'il existe une forte relation linéaire positive entre le nombre de spots publicitaires diffusés et les ventes. Plus précisément, une augmentation du nombre de spots publicitaires se traduira par une augmentation des ventes.

Pour conclure, soulignons que la corrélation fournit une mesure de la relation linéaire mais pas nécessairement une relation de causalité. Une corrélation importante entre deux variables ne signifie pas que des changements intervenant sur l'une des variables se traduiront par des changements sur l'autre variable. Par exemple, on pourrait trouver que la qualité et le prix d'un repas dans un restaurant sont positivement corrélés. Cependant, une augmentation du prix du repas n'impliquera pas forcément une augmentation de sa qualité.

REMARQUES

1. Dans la mesure où le coefficient de corrélation ne mesure que la robustesse d'une relation linéaire entre deux variables quantitatives, il est possible que le coefficient de corrélation soit proche de zéro, suggérant l'absence de relation linéaire, lorsque la relation entre les deux variables est non linéaire. Par exemple, le nuage de points



Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

ci-dessus indique la relation entre le montant dépensé par un petit magasin pour maintenir la température ambiante (chauffage et climatisation) et la température quotidienne extérieure maximale sur une période 100 jours.

2. Le coefficient de corrélation de l'échantillon est égal à $r_{xy} = -0,07$ et indique qu'il n'existe pas de relation linéaire entre ces deux variables. Toutefois, la forme du nuage de points indique l'existence d'une relation non linéaire. Nous pouvons en effet voir que lorsque les températures extérieures maximales augmentent, le montant dépensé pour maintenir une température ambiante sous contrôle commence par décroître dans la mesure où moins de chauffage est nécessaire puis augmente au fur et à mesure que les besoins de climatisation augmentent.

EXERCICES

Méthode

55. Cinq observations pour deux variables sont présentées ci-dessous.

x_i	4	6	11	3	16
y_i	50	50	40	60	30

a) Dessiner un nuage de points avec x sur l'axe des abscisses.

b) Quelle relation entre les deux variables le nuage de points de la question (a) indique-t-il ?

c) Calculer et interpréter la covariance de l'échantillon.

d) Calculer et interpréter le coefficient de corrélation de l'échantillon.

56. Cinq observations pour deux variables sont présentées ci-dessous.

x_i	6	11	15	21	27
y_i	6	9	6	17	12

a) Dessiner un nuage de points avec x sur l'axe des abscisses.

b) Quelle relation entre les deux variables le nuage de points de la question (a) indique-t-il ?

c) Calculer et interpréter la covariance de l'échantillon.

d) Calculer et interpréter le coefficient de corrélation de l'échantillon.

Applications

57. Dix matchs de football universitaire ont été joués en janvier 2010. L'université de l'Alabama a battu l'université du Texas 37 à 21 et est devenue le champion national universitaire. Les résultats (fichier en ligne BowlGames) des 10 matchs sont fournis dans le tableau suivant (*USA Today*, 8 janvier 2010). L'écart de points prévisionnel entre l'équipe gagnante et l'équipe perdante était estimé grâce aux paris effectués à Las Vegas environ une semaine avant que les matchs aient lieu. Par exemple, les paris désignaient

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Auburn gagnant sur Northwestern lors du championnat Outback Bowl par 5 points. L'écart de points réels en faveur de Auburn fut de 3. Un écart de points estimé négatif signifie que l'équipe qui a réellement gagné le match était l'outsider et aurait dû perdre selon les pronostics. Par exemple, dans le championnat Rose Bowl, les paris donnaient l'État de l'Ohio perdant avec un déficit de 2 points et finalement, l'État de l'Ohio a gagné par 9 points.

Championnat	Score	Écart de points attendu	Écart de points effectif
Outback	Auburn 38 Northwestern 35	5	3
Gator	État de Floride 33 Virginie Occidentale 21	1	12
Capital One	État de Pennsylvanie 19 LSU 17	3	2
Rose	État de l'Ohio 26 Oregon 17	−2	9
Sugar	Floride 51 Cincinnati 24	14	27
Cotton	État du Mississippi 21 état de l'Oklahoma 7	3	14
Alamo	Texas Tech 41 état du Michigan 31	9	10
Fiesta	État de Boise 17 TCU 10	−4	7
Orange	Iowa 24 Georgia Tech 14	−3	10
Championnat national	Alabama 37 Texas 21	4	16

a) Dessiner un nuage de points pour les données, avec l'écart de point attendu en abscisse.

b) Quelle est la relation entre l'écart de points attendu et l'écart de points effectif ?

c) Calculer et interpréter la covariance de l'échantillon.

d) Calculer le coefficient de corrélation de l'échantillon. Qu'indique cette valeur quant à la relation entre l'écart de points attendu par les parieurs de Las Vegas et l'écart de points effectif lors des matchs de football universitaire ?

58. Une étude du ministère des transports sur la vitesse et le kilométrage des véhicules de taille moyenne a fourni les données suivantes :

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

Calculer et interpréter le coefficient de corrélation de l'échantillon.

59. Au début de l'année 2009, la crise économique a entraîné la destruction d'emplois et l'augmentation des saisies immobilières. Le taux de chômage national s'élevait à 6,5 % et le pourcentage de saisies immobilières à 6,12 % (*The Wall Street Journal*, 27 janvier 2009). Pour prévoir quel serait l'état du marché immobilier au cours de l'année à venir, les économistes ont étudié la relation entre le taux de chômage et le pourcentage de saisies immobilières. Les économistes pensaient que si le taux de chômage continuait à augmenter, il y aurait également une augmentation des saisies immobilières. Les données suivantes fournissent le taux de chômage et les pourcentages de saisies immobilières sur 27 marchés immobiliers (fichier en ligne Logement).

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Zone urbaine	Taux de chômage (%)	Saisies immobilières (%)	Zone urbaine	Taux de chômage (%)	Saisies immobilières (%)
Atlanta	7,1	7,02	New York	6,2	5,78
Boston	5,2	5,31	Comté d'Orange	6,3	6,08
Charlotte	7,8	5,38	Orlando	7,0	10,05
Chicago	7,8	5,40	Philadelphie	6,2	4,75
Dallas	5,8	5,00	Phoenix	5,5	7,22
Denver	5,8	4,07	Portland	6,5	3,79
Detroit	9,3	6,53	Raleigh	6,0	3,62
Houston	5,7	5,57	Sacramento	8,3	9,24
Jacksonville	7,3	6,99	Saint Louis	7,5	4,40
Las Vegas	7,6	11,12	San Diego	7,1	6,91
Los Angeles	8,2	7,56	San Francisco	6,8	5,57
Miami	7,1	12,11	Seattle	5,5	3,87
Minneapolis	6,3	4,39	Tampa	7,5	8,42
Nashville	6,6	4,78			

Logement

a) Calculer le coefficient de corrélation de l'échantillon. Y a-t-il une corrélation positive entre le taux de chômage et le pourcentage de saisies immobilières ? Quelle est votre interprétation ?

b) Dessiner un nuage de points de la relation entre le taux de chômage et le pourcentage de saisies immobilières.

Russell

60. Le Russell 1000 est un indice financier composé des valeurs des plus grandes sociétés américaines. Le Dow Jones industriel moyen est basé sur 30 grandes sociétés. Le fichier en ligne Russell fournit les rendements annuels en pourcentage pour chacun de ces indices entre 1988 et 2012 (site Internet 1stock1).

a) Construire un nuage de points pour ces rendements.

b) Calculer la moyenne et l'écart type d'échantillon pour chaque indice.

c) Calculer le coefficient de corrélation de l'échantillon.

d) Discuter des similitudes et des différences entre ces deux indices.

61. Les températures journalières minimales et maximales de 14 villes à travers le monde sont regroupées dans le tableau suivant (La chaîne météo, 22 avril 2009 ; fichier en ligne Températures mondiales).

Ville	Maximales	Minimales	Ville	Maximales	Minimales
Athènes	68	50	Londres	67	45
Pékin	70	49	Moscou	44	29
Berlin	65	44	Paris	69	44
Le Caire	96	64	Rio de Janeiro	76	69
Dublin	57	46	Rome	69	51
Genève	70	45	Tokyo	70	58
Hong Kong	80	73	Toronto	44	39

Températures mondiales

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

a) Quelle est la température maximale moyenne ?

b) Quelle est la température minimale moyenne ?

c) Quel est le coefficient de corrélation entre les minimales et les maximales ? Discuter.

3.6 TABLEAU DE BORD : AJOUTER DES MESURES NUMÉRIQUES POUR AMÉLIORER SON EFFICACITÉ

Dans la section 2.5, nous avons présenté une introduction à la visualisation des données, un terme utilisé 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 communiquer des informations clés relatives à des données de façon aussi efficace et claire que possible. L'un des outils de visualisation des données les plus fréquemment utilisés est le tableau de bord, un ensemble de représentations visuelles qui organisent et présentent les informations utiles pour surveiller la performance d'une société ou d'une organisation d'une manière simple à lire, comprendre et interpréter. Dans cette section, nous étendons la discussion relative aux tableaux de bord de données pour montrer comment l'ajout de mesures numériques peut améliorer l'efficacité générale de la présentation.

L'ajout de mesures numériques, telles que la moyenne et l'écart type d'indicateurs de performance clés à un tableau de bord, est crucial dans la mesure où ces mesures numériques constituent souvent des benchmarks ou des objectifs par rapport auxquels les indicateurs clés de performance sont évalués. De plus, les représentations graphiques qui comprennent des mesures numériques sont également fréquemment incluses dans les tableaux de bord. Nous devons garder à l'esprit que le but d'un tableau de bord de données est de fournir des informations sur les indicateurs clés de performance d'une manière facile à lire, à comprendre et à interpréter. Ajouter des mesures numériques et des graphiques basés sur ces mesures numériques peut nous aider à atteindre cet objectif.

Pour illustrer l'utilisation de mesures numériques dans un tableau de bord de données, reprenons l'exemple de la société Grogan Oil développé dans la section 2.5 pour introduire le concept de tableau de bord des données. La société Grogan Oil possède des bureaux situés dans trois villes du Texas : Austin (son siège social), Houston et Dallas. Le centre d'appel informatique de Grogan, situé dans les bureaux d'Austin, traite des appels relatifs à des problèmes informatiques (lgociels, Internet et e-mail) rencontrés par les employés des trois bureaux. La figure 3.13 représente le tableau de bord développé par la société Grogan pour contrôler la performance du centre d'appel. Les éléments clés de ce tableau de bord de données sont les suivants :

– Le graphique en barres empilées dans le coin supérieur gauche du tableau de bord indique le volume d'appels pour chaque type de problème (logiciel, Internet ou e-mail) survenu au cours du temps.
– Le diagramme circulaire situé 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 problème ou le temps d'inactivité.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

– Pour chaque appel non résolu, qui a été reçu il y a plus de 15 minutes, le diagramme en barres figurant sur le côté gauche de la partie centrale du tableau de bord indique la durée qu'il a fallu pour résoudre ces cas.
– Le diagramme en barres situé côté droit de la partie centrale du tableau de bord indique le volume d'appels par bureau (Houston, Dallas et Austin) pour chaque type de problème.
– L'histogramme représenté en bas du tableau de bord indique la distribution du temps nécessaire pour résoudre un cas parmi l'ensemble des cas résolus par l'équipe en poste.

Dans le but d'en apprendre davantage sur la performance du centre d'appel, le responsable informatique de Grogan a décidé d'étendre le tableau de bord actuel en y ajoutant des boîtes-à-pattes relatives au temps nécessaire pour répondre aux appels reçus pour chaque type de problème (e-mail, Internet et logiciels). De plus, un graphique indiquant le temps nécessaire pour résoudre les cas individuels a été ajouté dans la partie inférieure gauche du tableau de bord. Enfin, le responsable informatique

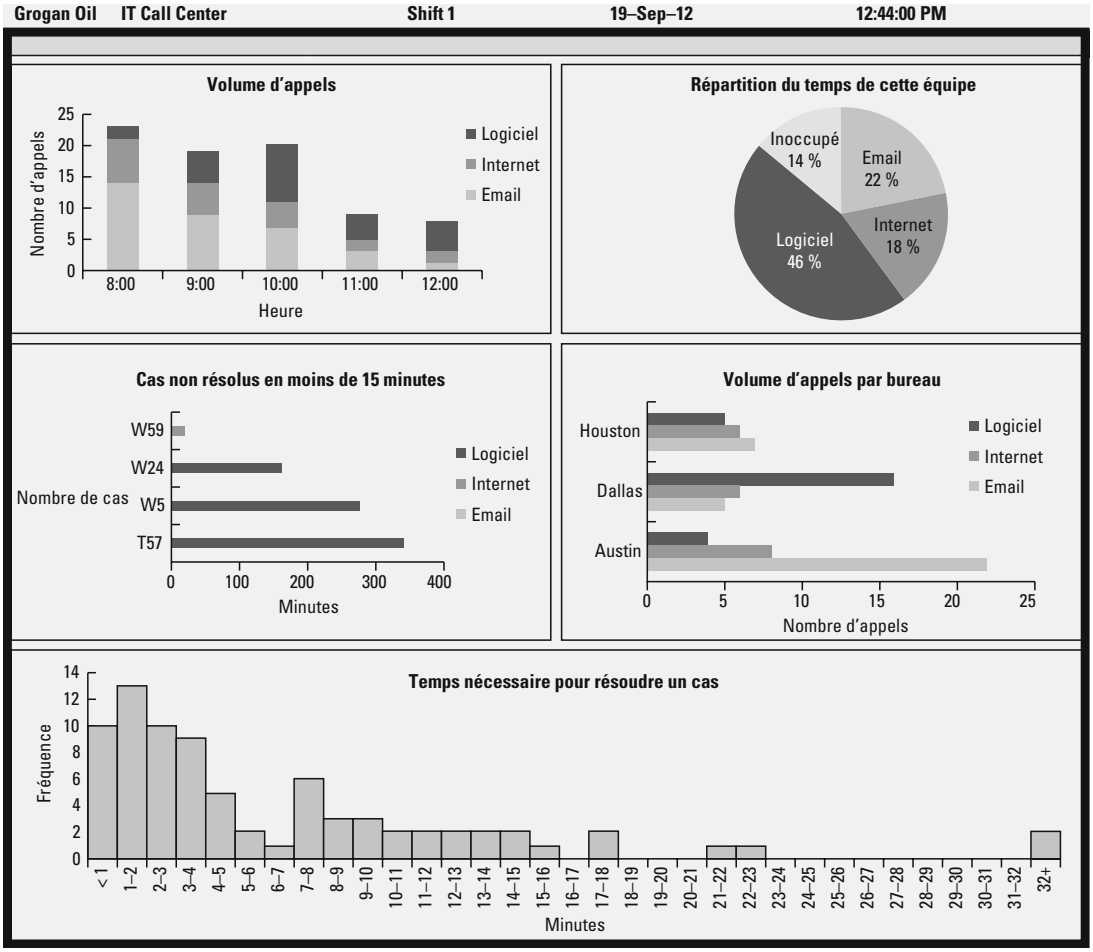
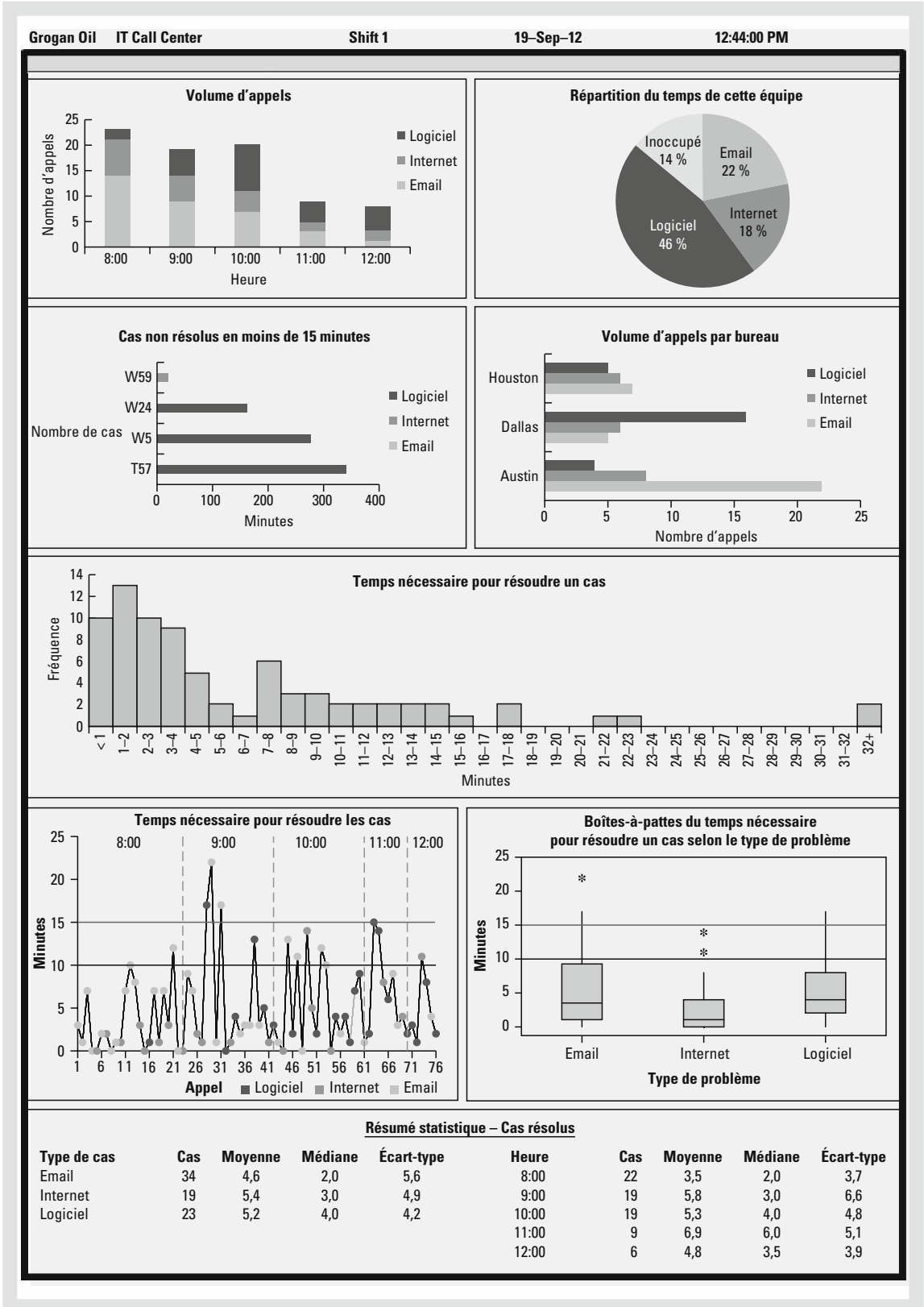


Figure 3.13 *Tableau de bord initial du centre d'appel informatique de la société Grogan Oil*

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.



Résumé statistique – Cas résolus

Type de cas	Cas	Moyenne	Médiane	Écart-type
Email	34	4,6	2,0	5,6
Internet	19	5,4	3,0	4,9
Logiciel	23	5,2	4,0	4,2

Heure	Cas	Moyenne	Médiane	Écart-type
8:00	22	3,5	2,0	3,7
9:00	19	5,8	3,0	6,6
10:00	19	5,3	4,0	4,8
11:00	9	6,9	6,0	5,1
12:00	6	4,8	3,5	3,9

Figure 3.14 ***Tableau de bord acutalisé du centre d'appel informatique de la société Grogan Oil***

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

a ajouté un résumé des statistiques pour chaque type de problème et pour chacune des premières heures de l'équipe. Le tableau de bord actualisé est présenté à la figure 3.14.

Le centre d'appel informatique s'est fixé comme objectif de performance de résoudre en moyenne un cas en 10 minutes. De plus, le centre a décidé qu'il n'était pas acceptable que la résolution d'un problème prenne plus de 15 minutes. Pour refléter ces objectifs, des lignes horizontales matérialisant respectivement l'objectif moyen de 10 minutes et le niveau maximal acceptable de 15 minutes ont été ajoutées sur le graphique indiquant la durée de résolution des cas et sur le graphique représentant la boîte-à-pattes du temps nécessaire pour répondre aux appels reçus pour chaque type de problème.

Le résumé statistique présent dans le tableau de bord de la figure 3.14 indique que la durée moyenne pour résoudre un cas concernant les e-mails est de 4,6 minutes, pour résoudre un cas concernant Internet de 5,4 minutes et un cas concernant un logiciel de 5,2 minutes. Ainsi, la durée moyenne pour résoudre chaque type de problème est inférieure à l'objectif fixé (10 minutes).

En examinant les boîtes-à-pattes, nous voyons que la boîte associée aux problèmes relatifs aux e-mails est « plus grande » que les boîtes associées aux deux autres types de problèmes. Le résumé statistique nous indique également que l'écart type de la durée nécessaire pour résoudre des problèmes liés aux e-mails est plus grand que les écarts types de la durée de résolution des deux autres types de problèmes. Cela nous conduit à examiner plus attentivement les cas relatifs à des problèmes de messagerie électronique dans les deux nouveaux graphiques. La boîte-à-pattes des cas relatifs à la messagerie électronique a une patte qui s'étend au-delà de 15 minutes et une valeur aberrante bien supérieure à 15 minutes. Le graphique représentant la durée de résolution des cas individuels (dans le cadran gauche le plus bas du tableau de bord) indique que cela est dû à deux appels pour des problèmes d'e-mail survenus entre 9 h et 10 h qui ont pris plus de 15 minutes pour être solutionnés. Cette analyse peut amener le responsable du centre d'appel informatique à chercher à comprendre pourquoi la durée pour résoudre des problèmes relatifs aux e-mails est plus variable que celle relative à des cas impliquant Internet ou des logiciels. En se fondant sur cette analyse, le responsable informatique peut également décider d'examiner les circonstances qui ont conduit à ces durées inhabituellement longues pour résoudre les deux cas relatifs à des problèmes de messagerie électronique qui ont pris plus de 15 minutes pour être résolus.

Le graphique indiquant la durée de résolution des cas individuels montre également que la plupart des appels reçus au cours de la première heure de prise de poste de l'équipe ont été solutionnés assez rapidement ; le graphique indique également que le temps nécessaire pour résoudre les problèmes a augmenté progressivement au cours de la matinée. Cela peut être lié à une tendance à l'apparition de problèmes complexes après la prise de poste de l'équipe ou au retard pris dans le traitement des appels qui s'accumulent. Bien que le résumé statistique suggère que les cas soumis entre 9 h et 10 h soient les plus longs à être résolus, le graphique relatif à la durée de résolution des cas individuels indique que deux cas chronophages relatifs à des problèmes d'e-mails et un

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

cas chronophage relatif à des problèmes de logiciel ont été enregistrés durant cette heure, et cela peut expliquer pourquoi le temps moyen de résolution des cas entre 9 et 10 h est plus important que durant les autres heures durant lesquelles l'équipe était en poste. Globalement, les cas reportés ont généralement été traités en 15 minutes au plus durant les heures de travail de cette équipe.

Les tableaux de bord de données comme celui de la société Grogan Oil sont souvent interactifs. Par exemple, lorsqu'un responsable utilise une souris ou touche un écran d'ordinateur pour positionner le curseur sur la représentation graphique ou pointer quelque chose sur le graphique, des informations supplémentaires telles que la durée pour résoudre le problème, l'heure à laquelle l'appel a été reçu, et l'individu ou le lieu d'où est émis l'appel peuvent apparaître. Cliquer sur l'individu peut également conduire l'utilisateur à un nouveau niveau d'analyse des cas individuels.

L'exploration plus approfondie fait référence à une fonctionnalité des tableaux de bord de donnés qui permet à l'usager d'accéder à des informations et des analyses à un niveau de plus en plus détaillé.

RÉSUMÉ

Dans ce chapitre, nous avons introduit plusieurs statistiques descriptives, utilisées pour résumer la tendance centrale, la dispersion et la forme de la distribution d'un ensemble de données. Contrairement aux procédures graphiques et sous forme de tableaux introduites dans le chapitre 2, les mesures introduites dans ce chapitre résument les données par des valeurs numériques. Lorsque les valeurs numériques obtenues sont issues d'un échantillon, on parle de statistiques d'échantillon. Lorsque les valeurs numériques sont issues d'une population, on parle de paramètres de la population. On a reproduit certaines notations utilisées pour les statistiques d'échantillon et les paramètres de la population ci-dessous :

	Statistiques d'échantillon	Paramètres de la population
Moyenne	$\bar{x}$	μ
Variance	s^2	σ^2
Écart type	s	σ
Covariance	s_{xy}	σ_{xy}
Corrélation	r_{xy}	ρ_{xy}

En inférence statistique, la statistique d'échantillon est appelée estimateur ponctuel du paramètre correspondant de la population.

Nous avons défini les mesures de tendance centrale suivantes : la moyenne, la médiane, le mode, la moyenne pondérée, la moyenne géométrique, les percentiles et les

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

quartiles. Puis, nous avons présenté l'étendue, l'étendue interquartile, la variance, l'écart type et le coefficient de variation comme mesures de dispersion. Notre mesure principale de la forme d'une distribution est fournie par le degré d'asymétrie des données. Des valeurs négatives indiquent une distribution biaisée à gauche. Des valeurs positives indiquent une distribution biaisée à droite. Nous avons ensuite décrit la façon d'utiliser la moyenne et l'écart type, en appliquant le théorème de Chebyshev et la règle empirique, pour obtenir plus d'informations sur la distribution des données et pour identifier les valeurs aberrantes.

Dans la section 3.4, nous avons montré comment construire un résumé en cinq chiffres et une boîte-à-pattes pour obtenir simultanément des informations sur la tendance centrale, la dispersion et la forme de la distribution. Dans la section 3.5, nous avons présenté la covariance et le coefficient de corrélation, deux mesures de la relation entre deux variables. Dans la dernière section, nous avons montré comment l'ajout de mesures numériques peut améliorer l'efficacité des tableaux de bord de données.

Les statistiques descriptives, présentées ici, peuvent être calculées en utilisant les logiciels statistiques et les feuilles de calcul. Dans les annexes de ce chapitre, nous montrerons comment développer les statistiques descriptives introduites dans ce chapitre en utilisant Minitab, Excel et StatTools.

GLOSSAIRE

STATISTIQUE D'ÉCHANTILLON. Valeur numérique utilisée comme mesure d'un échantillon (par exemple, la moyenne d'échantillon, $\bar{x}$, la variance d'échantillon, s^2, et l'écart type d'échantillon, s).

PARAMÈTRE DE LA POPULATION. Valeur numérique utilisée comme mesure de la population (par exemple, la moyenne de la population, μ, la variance de la population, σ^2 et l'écart type de la population, σ).

ESTIMATEUR PONCTUEL. Statistique d'échantillon, telle que $\bar{x}$, s^2 et s, utilisée pour estimer le paramètre correspondant de la population.

MOYENNE. Mesure de tendance centrale. Elle est obtenue en sommant la valeur des observations et en divisant par le nombre d'observations.

MOYENNE PONDÉRÉE. Moyenne obtenue en assignant à chaque observation une pondération qui reflète son importance.

MÉDIANE. Mesure de tendance centrale. Il s'agit de la valeur centrale de l'ensemble de données classé en ordre croissant.

MOYENNE GÉOMÉTRIQUE. Mesure de tendance centrale calculée en trouvant la racine $n^{\text{ième}}$ du produit de n valeurs.

MODE. Mesure de tendance centrale, définie comme la valeur de l'observation la plus fréquente.

PERCENTILE. Valeur telle qu'au moins p pour cent des observations ont une valeur inférieure ou égale à cette valeur et au moins $(100 - p)$ pour cent des observations ont une valeur supérieure ou égale à cette valeur. La médiane correspond au 50^{e} percentile.

QUARTILE. Les 25^{e}, 50^{e} et 75^{e} percentiles sont appelés respectivement premier quartile, deuxième quartile (médiane) et troisième quartile. Les quartiles divisent l'ensemble des données en quatre parties, chacune

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

contenant environ 25 % des données.

Étendue. Mesure de dispersion, égale à la différence entre la plus grande et la petite valeur.

Étendue interquartile (EIQ). Mesure de dispersion, égale à la différence entre le troisième et le premier quartile.

Variance. Mesure de dispersion, basée sur les écarts au carré des observations par rapport à la moyenne.

Écart type. Mesure de dispersion, égale à la racine carrée de la variance.

Coefficient de variation. Mesure de dispersion relative, égale au rapport de l'écart type à la moyenne, multiplié par 100.

Degré d'asymétrie. Mesure de la forme d'une distribution de données. Des données biaisées à gauche sont caractérisées par un degré d'asymétrie négatif. Une distribution symétrique a un degré d'asymétrie nul. Des données comportant un biais à droite sont caractérisées par un degré d'asymétrie positif.

Variable centrée réduite z. Valeur obtenue en divisant l'écart par rapport à la moyenne par l'écart type s. La variable centrée réduite mesure la distance, en nombre d'écarts type, entre l'observation x_i et la moyenne.

Théorème de Chebyshev. Théorème utilisé pour déduire le pourcentage d'observations qui se situent dans un intervalle de x écarts type de part et d'autre de la moyenne.

Règle empirique. Règle qui donne le pourcentage d'observations situées dans les intervalles de un, deux et trois écarts type autour de la moyenne, pour une distribution en forme de cloche (distribution dite « normale »).

Valeur aberrante. Observation anormalement grande ou petite.

Résumé en cinq chiffres. Technique d'analyse exploratoire des données qui utilise cinq chiffres pour résumer les données : la plus petite valeur, le premier quartile, la médiane, le troisième quartile et la plus grande valeur.

Boîte-à-pattes. Résumé graphique des données, à partir du résumé en cinq chiffres.

Covariance. Mesure de la relation linéaire entre deux variables. Des valeurs positives indiquent une relation positive ; des valeurs négatives indiquent une relation négative.

Coefficient de corrélation. Mesure de la relation linéaire entre deux variables, dont les valeurs sont comprises entre -1 et $+1$. Des valeurs proches de $+1$ indiquent une forte relation linéaire positive, des valeurs proches de -1 indiquent une forte relation linéaire négative, et des valeurs proches de zéro indiquent l'absence de relation linéaire.

FORMULES CLÉ

Moyenne d'échantillon

$$\bar{x} = \frac{\sum x_i}{n} \tag{3.1}$$

Moyenne de la population

$$\mu = \frac{\sum x_i}{N} \tag{3.2}$$

Moyenne pondérée

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \tag{3.3}$$

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Moyenne géométrique

$$\overline{x_g} = \sqrt[n]{(x_1)(x_2)\ldots(x_n)} = [(x_1)(x_2)\ldots(x_n)]^{1/n} \tag{3.4}$$

Étendue interquartile

$$EIQ = Q_3 - Q_1 \tag{3.5}$$

Variance de la population

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} \tag{3.6}$$

Variance de l'échantillon

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{3.7}$$

Écart type

$$\text{Écart type de l'échantillon} = s = \sqrt{s^2} \tag{3.8}$$

$$\text{Écart type de la population } \sigma = \sqrt{\sigma^2} \tag{3.9}$$

Coefficient de variation

$$\frac{\text{Écart type}}{\text{Moyenne}} \times 100 \tag{3.10}$$

Variable centrée réduite *z*

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.11}$$

Covariance de l'échantillon

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{3.12}$$

Covariance de la population

$$\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N} \tag{3.13}$$

Coefficient de corrélation de Pearson : données issues d'un échantillon

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3.14}$$

Coefficient de corrélation de Pearson : données issues d'une population

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.15}$$

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

EXERCICES SUPPLÉMENTAIRES

62. Le nombre moyen de fois où les Américains dînent à l'extérieur au cours d'une semaine est passé de 4,0 en 2008 à 3,8 en 2012 (Zagat.com, 1er avril 2012). Les données suivantes correspondent au nombre de fois où un échantillon de 20 familles a dîné à l'extérieur la semaine dernière.

6	1	5	3	7	3	0	3	1	3
4	1	2	4	1	0	5	6	3	1

- **a)** Calculer la moyenne et la médiane.
- **b)** Calculer les premier et troisième quartiles.
- **c)** Calculer l'étendue et l'étendue interquartile.
- **d)** Calculer la variance et l'écart type.
- **e)** Le degré d'asymétrie de ces données est de 0,34. Commenter la forme de cette distribution. Est-ce la forme à laquelle vous vous attendiez ? Pourquoi ?
- **f)** Les données contiennent-elles des valeurs aberrantes ?

63. Le magazine *USA Today* rapporte que les écoles et les universités NCAA offrent aujourd'hui de meilleurs salaires à un entraîneur de football nouvellement recruté, comparativement à ce que ces établissements offraient en termes de rémunération à leurs anciens entraîneurs (*USA Today*, 12 février 2013). Les salaires annuels de base des anciens et des nouveaux entraîneurs de 23 écoles sont fournis dans le fichier en ligne Entraîneurs.

- **a)** Déterminer le salaire annuel médian pour un ancien entraîneur et pour un nouvel entraîneur de football.
- **b)** Calculer l'étendue des salaires à la fois pour les anciens et les nouveaux entraîneurs.
- **c)** Calculer l'écart type des salaires à la fois pour les anciens et les nouveaux entraîneurs.
- **d)** En vous basant sur vos réponses aux questions (a) à (c), commenter toutes les différences qui apparaîtraient entre le salaire annuel de base qu'une école offre à un nouvel entraîneur de football comparativement à ce qu'elle offrait à un ancien entraîneur.

64. Le temps d'attente moyen d'un patient dans un cabinet médical d'El Paso est de l'ordre de 29 minutes, bien au-dessus de la moyenne nationale qui s'établit à 21 minutes. En fait, El Paso détient le record du temps d'attente chez un médecin des États-Unis (*El Paso Times*, 8 janvier 2012). Pour résoudre le problème des temps d'attente, certains cabinets médicaux utilisent des systèmes d'évaluation des temps d'attente pour informer les patients des temps d'attente attendus. Les patients peuvent adapter le moment de leur arrivée en se basant sur cette information et passer moins de temps dans les salles d'attente. Les données suivantes fournissent les temps d'attente (en minutes) d'un échantillon de patients dans des cabinets qui n'ont pas de systèmes d'évaluation des temps d'attente et les temps d'attente d'un échantillon de patients dans des cabinets qui possèdent un tel système (fichier en ligne Temps d'attente).

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Temps d'attente

Sans système d'évaluation des temps d'attente	Avec système d'évaluation des temps d'attente
24	31
67	11
17	14
20	18
31	12
44	37
12	9
23	13
16	12
37	15

a) Quels sont les temps d'attente moyen et médian des patients dans les cabinets possédant le système d'évaluation des temps d'attente ? Quels sont les temps d'attente moyen et médian des patients dans les cabinets ne possédant pas ce système ?

b) Quels sont la variance et l'écart type des temps d'attente des patients dans les cabinets possédant le système d'évaluation des temps d'attente ? Quels sont la variance et l'écart type des temps d'attente des patients dans les cabinets ne possédant pas le système d'évaluation des temps d'attente ?

c) Le temps d'attente des patients dans les cabinets possédant le système d'évaluation des temps d'attente est-il plus faible que celui des patients dans les cabinets ne possédant pas ce système ? Expliquer.

d) En ne tenant compte que des cabinets sans système d'évaluation des temps d'attente, quelle est la valeur de la variable centrée réduite pour le 10^{e} patient de l'échantillon ?

e) En ne tenant compte que des cabinets avec système d'évaluation des temps d'attente, quelle est la valeur de la variable centrée réduite pour le 6^{e} patient de l'échantillon ? Comparez-la à la valeur de la variable centrée réduite calculée à la question (d).

f) En vous basant sur les valeurs des variables centrées réduites, les données relatives aux cabinets sans système d'évaluation des temps d'attente contiennent-elles des valeurs aberrantes ? En vous basant sur les valeurs des variables centrées réduites, les données relatives aux cabinets avec système d'évaluation des temps d'attente contiennent-elles des valeurs aberrantes ?

65. Les sociétés américaines perdent chaque année 63,2 milliards de dollars à cause des travailleurs souffrant d'insomnies. Les travailleurs perdent en moyenne l'équivalent de 7,8 jours de productivité en moyenne par an, à cause du manque de sommeil (*Wall Street Journal*, 23 janvier 2013). Les données suivantes indiquent le nombre d'heures de sommeil effectives au cours d'une nuit récente d'un échantillon de 20 travailleurs (fichier en ligne Sommeil).

Sommeil

6	5	10	5	6	9	9	5	9	5
8	7	8	6	9	8	9	6	10	8

a) Quel est le nombre moyen d'heures de sommeil pour cet échantillon ?

b) Quelle est la variance ? L'écart-type ?

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

66. Une étude sur les utilisateurs de smartphones révèle que 68 % des utilisations de smartphone surviennent à la maison et qu'un utilisateur passe en moyenne 410 minutes par mois à utiliser un smartphone pour interagir avec d'autres personnes (*Harvard Business Review*, janvier-février 2013). Considérez les données suivantes qui indiquent le nombre de minutes par mois passées à interagir avec d'autres via un smartphone pour un échantillon de 50 utilisateurs (fichier en ligne Smartphone).

353	458	404	394	416
437	430	369	448	430
431	469	446	387	445
354	468	422	402	360
444	424	441	357	435
461	407	470	413	351
464	374	417	460	352
445	387	468	368	430
384	367	436	390	464
405	372	401	388	367

Smartphone

a) Quel est le nombre moyen de minutes passées à interagir avec d'autres pour cet échantillon ? Comparez-le à la moyenne rapportée dans l'étude ?

b) Quel est l'écart type pour cet échantillon ?

c) Y a-t-il des valeurs aberrantes dans cet échantillon ?

67. Chaque jour, pour aller travailler, un employé a le choix entre prendre les transports en commun ou son véhicule personnel. Un échantillon des temps de trajet avec chacun des deux modes de transport est présenté ci-dessous. Les temps sont exprimés en minutes.

Transport en commun :	28	29	32	37	33	25	29	32	41	34
Véhicule personnel :	29	31	33	32	34	30	31	32	35	33

a) Calculer le temps moyen du trajet effectué avec chacun des deux modes de transport.

b) Calculer l'écart type pour les deux méthodes.

c) Sur la base de vos résultats aux questions (a) et (b), quelle méthode de transport préconiseriez-vous ? Expliquer.

d) Construire une boîte-à-pattes pour chaque mode de transport. Est-ce que la comparaison des boîtes-à-pattes confirme votre réponse à la question (c) ?

68. Les consommateurs empruntent de l'argent pour diverses raisons, comme par exemple l'achat d'une maison, d'une voiture et d'appareils électroménagers ou pour aider à payer les études de leurs enfants. Environ 75 % des ménages américains sont endettés (*Wall Street Journal*, 25 février 2013). Considérez que le montant d'endettement d'un échantillon de 25 ménages est reporté ci-dessous (fichier en ligne Dette).

122 231	69 402	52 055	131 176	59 423
125 409	142 762	72 576	58 458	18 927
59 025	131 934	148 782	57 380	124 831
116 128	107 320	79 649	110 354	53 880
60 370	68 140	94 513	97 544	72 140

Dette

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

a) Quel est le montant d'endettement médian d'un ménage ?

b) Fournir un résumé à cinq chiffres de ces données d'échantillon.

c) Quel est le montant d'endettement moyen des ménages de cet échantillon ?

d) L'échantillon contient-il des valeurs aberrantes ?

e) Préférez-vous utiliser la moyenne ou la médiane pour décrire le niveau d'endettement des ménages ? Pourquoi ?

69. L'enquête sur les communautés américaines du bureau américain du recensement a fourni le pourcentage d'enfants de moins de 18 ans qui ont vécu sous le seuil de pauvreté au cours des 12 mois précédents (site Internet du bureau américain du recensement, août 2008). La région – Nord-Est (NE), Sud-Est (SE), Centre-Ouest (CO), Sud-Ouest (SO) et Ouest (O) – ainsi que le pourcentage d'enfants de moins de 18 ans qui ont vécu sous le seuil de pauvreté sont donnés pour chaque État (fichier en ligne Seuil de pauvreté).

État	Région	% pauvreté	État	Région	% pauvreté
Alabama	SE	23,0	Montana	O	17,3
Alaska	O	15,1	Nebraska	CO	14,4
Arizona	SO	19,5	Nevada	O	13,9
Arkansas	SE	24,3	New Hampshire	NE	9,6
Californie	O	18,1	New Jersey	NE	11,8
Colorado	O	15,7	Nouveau Mexique	SO	25,6
Connecticut	NE	11,0	New York	NE	20,0
Delaware	NE	15,8	Caroline du Nord	SE	20,2
Floride	SE	17,5	Dakota du Nord	CO	13,0
Géorgie	SE	20,2	Ohio	CO	18,7
Hawaii	O	11,4	Oklahoma	SO	24,3
Idaho	O	15,1	Oregon	O	16,8
Illinois	CO	17,1	Pennsylvanie	NE	16,9
Indiana	CO	17,9	Rhode Island	NE	15,1
Iowa	CO	13,7	Caroline du Sud	SE	22,1
Kansas	CO	15,6	Dakota du Sud	CO	16,8
Kentucky	SE	22,8	Tennessee	SE	22,7
Louisiane	SE	27,8	Texas	SO	23,9
Maine	NE	17,6	Utah	O	11,9
Maryland	NE	9,7	Vermont	NE	13,2
Massachusetts	NE	12,4	Virginie	SE	12,2
Michigan	CO	18,3	Washington	O	15,4
Minnesota	CO	12,2	Virginie Occidentale	SE	25,2
Mississippi	SE	29,5	Wisconsin	CO	14,9
Missouri	CO	18,6	Wyoming	O	12,0

a) Quel est le pourcentage médian d'enfants vivant en-dessous du seuil de pauvreté pour les 50 États ?

b) Quels sont les premier et troisième quartiles ? Quelle est votre interprétation des quartiles ?

c) Dessiner une boîte-à-pattes pour les données. Que vous apprend la boîte-à-pattes

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

quant au niveau de pauvreté des enfants aux États-Unis. Y a-t-il des États qui peuvent être considérés comme des valeurs aberrantes ? Discuter.

d) Identifier les États appartenant au quartile inférieur. Quelle est votre interprétation de ce groupe et quelle(s) région(s) est (sont) la (les) plus représentée(s) dans le quartile inférieur ?

70. Le magazine *Travel + Leisure* présentait sa liste annuelle des 500 meilleurs hôtels à travers le monde (*Travel + Leisure*, janvier 2009). Le magazine attribue une note à chaque hôtel ainsi qu'un bref descriptif qui inclut la taille de l'hôtel, les commodités et le tarif par nuit pour une chambre double. Un échantillon de 12 des meilleurs hôtels aux États-Unis est fourni ci-dessous (fichier en ligne Travel).

Hôtel	Lieu	Nombre de chambres	Tarif par nuit
Boulders Resort & Spa	Phoenix, AZ	220	499
Disney's Wilderness Lodge	Orlando, FL	727	340
Four Seasons Hotel Beverly Hills	Los Angeles, CA	285	585
Four Seasons Hotel	Boston, MA	273	495
Hay Adams	Washington, DC	145	495
Inn on Biltmore Estate	Asheville, NC	213	279
Loews Ventana Canyon Resort	Phoenix, AZ	398	279
Mauna Lani Bay Hotem	Hawaii	343	455
Montage Laguna Beach	Laguna Beach, CA	250	595
Sofitel Water Tower	Chicago, IL	414	367
St. Regis Monarch Beach	Dana Point, CA	400	675
The Broadmoor	Colorado Springs, CO	700	420

a) Quel est le nombre moyen de chambres ?

b) Quel est le tarif moyen par nuit pour une chambre double ?

c) Représenter un nuage de points avec le nombre de chambres sur l'axe horizontal et le tarif par nuit sur l'axe vertical. Une relation entre le nombre de chambres et le tarif par nuit apparaît-elle ? Discuter

d) Quel est le coefficient de corrélation de l'échantillon ? Que vous apprend t-il sur la relation entre le nombre de chambres et le tarif par nuit pour une chambre double ? Cela vous semble-t-il raisonnable ? Discuter.

71. Morningstar suit les performances d'un nombre important de sociétés et publie une évaluation de chacune d'entre elles. Parmi un ensemble de données financières, Morningstar fournit une estimation du juste prix qui devrait être payé pour une action de la société. Les données pour 30 sociétés sont disponibles dans le fichier en ligne intitulé Juste prix. Les données incluent l'estimation du juste prix par action, le prix de l'action le plus récent et le rendement des actions de la société (*Morningtar Stocks 500*, 2008).

a) Dessiner un nuage de points pour les données relatives au juste prix et au prix observé des actions, avec le prix observé des actions sur l'axe horizontal. Quel est le coefficient de corrélation de l'échantillon et que vous apprend-t-il sur la relation entre les variables ?

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

b) Dessiner un nuage de points pour les données relatives au juste prix et au rendement des actions, avec le rendement des actions sur l'axe horizontal. Quel est le coefficient de corrélation de l'échantillon et que vous apprend-t-il sur la relation entre les variables ?

72. Est-ce que les résultats d'une équipe de la ligue principale de baseball durant l'entraînement de printemps fournissent une indication sur les performances de jeu de l'équipe durant la saison de championnat ? Au cours des six dernières années, le coefficient de corrélation entre les pourcentages de matchs gagnés par une équipe durant l'entraînement de printemps et durant la saison de championnat était de 0,18 (*The Wall Street Journal*, 30 mars 2009). Le tableau ci-dessous regroupe les pourcentages de matchs gagnés par les 14 équipes de la ligue américaine durant la saison 2008 (fichier en ligne Entraînement de printemps).

Équipe	Entraînement de printemps	Saison de championnat	Équipe	Entraînement de printemps	Saison de championnat
Baltimore Oriole	0,407	0,422	Minnesota Twins	0,500	0,540
Boston Red Sox	0,429	0,586	New York Yankees	0,577	0,549
Chicago White Sox	0,417	0,546	Oakland A's	0,692	0,466
Cleveland Indians	0,569	0,500	Seattle Mariners	0,500	0,377
Detroit Tigers	0,569	0,457	Tampa Bay Rays	0,731	0,599
Kansas City Royals	0,533	0,463	Texas Rangers	0,643	0,488
Los Angeles Angels	0,724	0,617	Toronto Blue Jays	0,448	0,531

a) Quel est le coefficient de corrélation entre les résultats obtenus lors de l'entraînement de printemps et ceux obtenus lors du championnat ?

b) Quelle est votre conclusion : les performances d'une équipe lors de l'entraînement de printemps fournissent-elles une indication quant aux performances de l'équipe durant le championnat ? Quelles pourraient être les raisons d'une telle corrélation ? Discuter.

73. L'échéance (en nombre de jours) d'un échantillon de cinq placements sur le marché monétaire est indiquée ci-dessous. Les montants investis (en millions de dollars) dans ces placements sont également indiqués. Utiliser la moyenne pondérée pour déterminer l'échéance moyenne des cinq placements.

Échéance (en jours)	Valeur (millions de dollars)
20	20
12	30
7	10
5	15
6	10

74. Un système de radar de la police d'État contrôle la vitesse des automobiles roulant sur une route où la vitesse est limitée à 55 kilomètres par heure. La distribution de fréquence des vitesses est présentée ci-dessous.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Vitesse (km par heure)	Fréquence
45-49	10
50-54	40
55-59	150
60-64	175
65-69	75
70-74	15
75-79	10
Total	475

a) Quelle est la vitesse moyenne des automobiles roulant sur cette route ?

b) Calculer la variance et l'écart type.

75. La campagnie ferroviaire Panama a été créée en 1850 afin de construire le chemin de fer permettant de relier rapidement les océans Atlantique et Pacifique. Le tableau suivant (*The Big Ditch*, Mauer et Yu, 2011) fournit les rendements annuels de l'action de la Panama entre 1853 et 1880 (fichier en linge PanamaRailroad).

Année	Rendement de l'action de la Panama (%)
1853	– 1
1854	– 9
1855	19
1856	2
1857	3
1858	36
1859	21
1860	16
1861	– 5
1862	43
1863	44
1864	48
1865	7
1866	11
1867	23
1868	20
1869	– 11
1870	– 51
1871	– 42
1872	39
1873	42
1874	12
1875	26
1876	9
1877	– 6
1878	25
1879	31
1880	30

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

a) Créer un graphique des rendements annuels de l'action. Le rendement annuel moyen à la Bourse de New York était de 8,4 % entre 1853 et 1880. Pouvez-vous dire à partir du graphique si l'action de la Panama surperformait à la Bourse de New York ?

b) Calculer le rendement annuel moyen de l'action de la compagnie Panama entre 1853 et 1880. L'action était-elle plus rentable que la moyenne des actions à la Bourse de New York à la même époque ?

PROBLÈME 1 *Les magasins Pelican*

Les magasins Pelican, filiale de National Clothing, sont une chaîne de magasins de vêtements pour femme implantée aux États-Unis. Le magasin a récemment lancé 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, alors que la campagne de promotion était lancée. Le tableau 3.9 reprend une partie de cet ensemble de données. La méthode de paiement par carte de fidélité fait référence aux dépenses payées en utilisant la carte National Clothing. Les clients qui ont fait un achat en utilisant un bon de réduction sont identifiés comme des clients occasionnels et les clients qui ont effectué un achat mais sans utiliser un bon de réduction, sont identifiés comme 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 des personnes présentant des bons de réduction n'auraient pas été faits en l'absence de ces bons. Bien sûr, les responsables des magasins Pelican espèrent également que les clients occasionnels continueront à faire leurs achats dans leur magasin.

La plupart des variables contenues dans le tableau 3.12 sont compréhensibles. Deux nécessitent toutefois quelques éclaircissements.

Articles	Nombre d'articles achetés
Ventes globales	Montant total (en dollars) réglé par carte de crédit

La direction des magasins Pelican souhaite utiliser les données de cet échantillon pour mieux connaître ses clients et évaluer l'impact des promotions sous forme de bons de réduction.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Tableau 3.9 ***Échantillon de 100 achats réglés par carte de crédit dans les magasins Pelican***

Client	Type de client	Articles	Ventes globales	Méthode 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
6	Régulier	1	44,50	MasterCard	Femme	Marié	44
7	Occasionnel	2	78,00	Carte de fidélité	Femme	Marié	30
8	Régulier	1	22,50	Visa	Femme	Marié	40
9	Occasionnel	2	56,52	Carte de fidélité	Femme	Marié	46
10	Régulier	1	44,50	Carte de fidélité	Femme	Marié	36
…	…	…	…	…	…	…	…
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

Rapport

Utiliser les méthodes de statistiques descriptives présentées dans ce chapitre pour résumer les données et commenter vos résultats. Votre rapport doit contenir les résumés et discussions suivants.

1. Des statistiques descriptives sur les ventes globales en fonction des différentes catégories de clients.
2. Des statistiques descriptives concernant la relation entre l'âge des clients et les ventes.
3. Commenter les résultats qui vous paraissent présenter un intérêt pour la direction des magasins.

PROBLÈME 2 *L'industrie cinématographique*

L'industrie cinématographique est un secteur concurrentiel. Plus de 50 studios produisent plusieurs centaines de 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, les recettes globales (en millions de dollars), le nombre de cinémas

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Tableau 3.10 ***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 1re 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

Films 2011

projetant le film et le nombre de semaines au cours desquelles le film est classé dans le top 60 des entrées sont les variables généralement utilisées pour évaluer le succès d'un film. Les données collectées pour un échantillon de 100 films produits en 2011 (site Internet de Box Office Mojo, 17 mars 2012) sont regroupées dans le fichier en ligne intitulé Films2011. Le tableau 3.10 reprend les données pour les 10 premiers films de ce fichier. Notez que certains films, comme *War Horse*, sont sortis fin 2011 et sont toujours à l'affiche début 2012.

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.

1. Des statistiques descriptives pour chacune des quatre variables, accompagnées d'une discussion sur ce qu'elles nous apprennent à propos de l'industrie cinématographique.
2. Quels films, s'il y en a, devraient être considérés comme des valeurs aberrantes au regard de leur surperformance ? Expliquer.
3. Des statistiques descriptives décrivant la relation entre les ventes globales et chacune des autres variables. Discuter.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

PROBLÈME 3 *Les écoles de commerce d'Asie-Pacifique*

La poursuite d'études supérieures de commerce est devenue un phénomène international. Une étude montre que de plus en plus d'Asiatiques souhaitent devenir titulaire d'une maîtrise de gestion. En conséquence, le nombre de candidats aux cours MBA dans les écoles du Pacifique asiatique continue d'augmenter.

À travers la région, des milliers d'Asiatiques ont montré un intérêt croissant à interrompre provisoirement leur carrière pour obtenir en deux ans une formation commerciale théorique. Les cours suivis dans ces écoles sont réputés difficiles et incluent l'enseignement de l'économie, de la finance, du marketing, des sciences comportementales, des relations professionnelles, de la prise de décision, de la stratégie, du droit commercial, etc. L'ensemble de données du tableau 3.11 illustre certaines caractéristiques des principales écoles de commerce de la région du Pacifique asiatique (fichier en ligne Asie).

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Tableau 3.11 *Données sur 25 écoles de commerce asiatiques*

École de commerce	Inscription à plein temps	Nombre d'étudiants par enseignant	Frais de scolarité pour étudiants locaux ($)	Frais de scolarité pour étudiants étrangers ($)	Âge	% d'étrangers	Test d'admission	Test d'anglais	Expérience professionnelle	Salaire de départ ($)
École de commerce de Melbourne	200	5	24 420	29 600	28	47	Oui	Non	Oui	71 400
Université de New South Wales (Sydney)	228	4	19 993	32 582	29	28	Oui	Non	Oui	65 200
Institut indien de management (Ahmedabad)	392	5	4 300	4 300	22	0	Non	Non	Non	7 100
Université chinoise de Hong Kong	90	5	11 140	11 140	29	10	Oui	Non	Non	31 000
Université internationale du Japon (Niigata)	126	4	33 060	33 060	28	60	Oui	Oui	Non	87 000
Institut asiatique du management (Manille)	389	5	7 562	9 000	25	50	Oui	Non	Oui	22 800
Institut indien du management (Bangalore)	380	5	3 935	16 000	23	1	Oui	Non	Non	7 500
Université nationale de Singapour	147	6	6 146	7 170	29	51	Oui	Oui	Oui	43 300
Institut indien du management (Calcutta)	463	8	2 880	16 000	23	0	Non	Non	Non	7 400
Université nationale australienne (Canberra)	42	2	20 300	20 300	30	80	Oui	Oui	Oui	46 600
Université technologique de Nanyang (Singapour)	50	5	8 500	8 500	32	20	Oui	Non	Oui	49 300
Université de Queensland (Brisbane)	138	17	16 000	22 800	32	26	Non	Non	Oui	49 600
Université des sciences et des technologies de Hong Kong	60	2	11 513	11 513	26	37	Oui	Non	Oui	34 000
École de gestion Macquarie (Sydney)	12	8	17 172	19 778	34	27	Non	Non	Oui	60 100

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

École de commerce	Inscription à plein temps	Nombre d'étudiants par enseignant	Frais de scolarité pour étudiants locaux ($)	Frais de scolarité pour étudiants étrangers ($)	Âge	% d'étrangers	Test d'admission	Test d'anglais	Expérience professionnelle	Salaire de départ ($)
Université Chulalongkorn (Bangkok)	200	7	17 355	17 355	25	6	Oui	Non	Oui	17 600
École de commerce Monash Mt. Eliza (Melbourne)	350	13	16 200	22 500	30	30	Oui	Oui	Oui	52 500
Institut asiatique de management (Bangkok)	300	10	18 200	18 200	29	90	Non	Oui	Oui	25 000
Université d'Adelaïde	20	19	16 426	23 100	30	10	Non	Non	Oui	66 000
Université Massey (Palmerston North, Nouvelle Zélande)	30	15	13 106	21 625	37	35	Non	Oui	Oui	41 400
Institut royal de technologie de Melbourne	30	7	13 880	17 765	32	30	Non	Oui	Oui	48 900
Institut des études de management Jamnalal Bajaj (Mumbai)	240	9	1 000	1 000	24	0	Non	Non	Oui	7 000
Institut de technologie Curtin (Perth)	98	15	9 475	19 097	29	43	Oui	Non	Oui	55 000
Université des sciences managériales de Lahore	70	14	11 250	26 300	23	2,5	Non	Non	Non	7 500
Université Sains Malaisie (Penang)	30	5	2 260	2 260	32	15	Non	Oui	Oui	16 000
Université De La Salle (Manille)	44	17	3 300	3 600	28	3,5	Oui	Non	Oui	13 100

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Rapport

Utiliser les méthodes de statistiques descriptives pour résumer les données du tableau 3.11. Discuter vos résultats.

1. Résumer chaque variable de l'ensemble de données. Commenter et interpréter les valeurs minimales et maximales, ainsi que les moyennes et les proportions appropriées. Quelles nouvelles informations ces statistiques descriptives fournissent-elles concernant les écoles de commerce du Pacifique asiatique ?
2. Résumer les données pour comparer :
   a. Les différences entre les frais de scolarité pour étudiants locaux et étrangers.
   b. Les différences entre les salaires de départ des écoles qui exigent et qui n'exigent pas une expérience professionnelle.
   c. Les différences entre les salaires de départ des écoles qui effectuent et qui n'effectuent pas de test d'anglais.
3. Les salaires initiaux apparaissent-ils liés aux frais de scolarité ?
4. Présenter tout résumé graphique ou numérique supplémentaire pouvant aider à communiquer les données du tableau 3.11 à d'autres personnes.

PROBLÈME 4 *Les transactions en ligne de Heavenly Chocolates*

Heavenly Chocolates produit et vend du chocolat de qualité dans son usine et ses magasins de vente situés à Saratoga Springs, dans l'État de New York. Il y a deux ans, la société a développé un site Internet et a commencé à vendre ses produits en ligne. Les ventes par Internet ont dépassé toutes les attentes de la société et les responsables élaborent désormais des stratégies pour accroître encore davantage les ventes en ligne. Pour mieux connaître les clients en ligne, un échantillon de 50 transactions a été sélectionné à partir des ventes réalisées le mois dernier. Les données indiquant le jour de la semaine auquel la transaction a eu lieu, le portail d'accès à Internet que les clients ont utilisé, le temps passé sur le site Internet, le nombre de pages web visitées et le montant dépensé par chacun des 50 clients sont regroupées dans le fichier intitulé Clients. Une partie de cet ensemble de données est reproduit dans le tableau 3.12.

Heavenly Chocolates souhaiterait utiliser les données d'échantillon pour déterminer si les clients en ligne qui passent plus de temps sur le site et visitent plus de pages, dépensent également davantage durant leur visite sur le site Internet. La société souhaiterait également connaître l'impact du jour de la transaction et du navigateur Internet sur les ventes.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Tableau 3.12 ***Un échantillon de 50 transactions sur le site Internet de Heavenly Chocolates***

Client	Jour	Navigateur Internet	Temps (mn)	Nombre de pages visitées	Montant dépensé ($)
1	Lundi	Internet Explorer	12,0	4	54,52
2	Mercredi	Autre	19,5	6	94,90
3	Lundi	Internet Explorer	8,5	4	26,68
4	Mardi	Firefox	11,4	2	44,73
5	Mercredi	Internet Explorer	11,3	4	66,27
6	Samedi	Firefox	10,5	6	67,80
7	Dimanche	Internet Explorer	11,4	2	36,04
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
48	Vendredi	Internet Explorer	9,7	5	103,15
49	Lundi	Autre	7,3	6	52,15
50	Vendredi	Internet Explorer	13,4	3	98,75

Rapport

Utiliser les méthodes de statistiques descriptives pour mieux connaître les clients qui visitent le site Internet de Heavenly Chocolates. Inclure dans votre rapport les éléments suivants.

1. Des résumés graphiques et numériques du temps passé par les clients sur le site Internet, du nombre de pages visitées et du montant moyen dépensé par transaction. Discuter de ce que vous apprenez sur les clients en ligne de Heavenly Chocolates à partir de ces résumés numériques.
2. Résumer la fréquence, le montant total (en dollars) dépensé et le montant moyen dépensé par transaction pour chaque jour de la semaine. Quelles observations pouvez-vous faire quant à l'influence des jours de la semaine sur l'activité commerciale de Heavenly Chocolates ? Discuter.
3. Résumer la fréquence, le montant total (en dollars) dépensé et le montant moyen dépensé par transaction pour chaque type de navigateur Internet. Quelles observations pouvez-vous faire quant à l'influence du navigateur Internet sur l'activité commerciale de Heavenly Chocolates ? Discuter.
4. Représenter un nuage de points et calculer le coefficient de corrélation de l'échantillon pour déterminer la relation entre le temps passé sur le site Internet et le montant (en dollars) dépensé. Utiliser l'axe horizontal pour le temps passé sur le site Internet. Discuter.
5. Représenter un nuage de points et calculer le coefficient de corrélation de l'échantillon pour déterminer la relation entre le nombre de pages visitées et le

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

montant (en dollars) dépensé. Utiliser l'axe horizontal pour le nombre de pages visitées. Discuter.

6. Représenter un nuage de points et calculer le coefficient de corrélation de l'échantillon pour déterminer la relation entre le temps passé sur le site Internet et le nombre de pages visitées. Utiliser l'axe horizontal pour le nombre de pages visitées. Discuter.

PROBLÈME 5 *Les populations d'éléphants africains*

Alors que des millions d'éléphants erraient à travers l'Afrique, à partir du milieu des années 1980, le braconnage a décimé les populations d'éléphants sur le continent africain. Les éléphants sont importants dans les écosystèmes africains. Dans les forêts tropicales, les éléphants créent des passages dans la canopée qui participent à la croissance de nouveaux arbres. Dans la savane, les éléphants réduisent l'expansion des arbustes pour créer un environnement favorable aux animaux de pâturage. De plus, de nombreuses espèces de plantes doivent passer par le système digestif de l'éléphant pour entamer leur processus de germination.

Le statut actuel de l'éléphant est variable selon les pays ; dans certains pays, des mesures fortes ont été prises pour protéger efficacement les populations d'éléphants alors que dans d'autres pays, les populations d'éléphants restent soumises au braconnage (pour la viande et l'ivoire), sont confrontées à la dégradation de leur habitat et aux conflits avec les hommes. Le tableau 3.13 fournit les chiffres relatifs aux populations d'éléphants recensées dans plusieurs pays d'Afrique en 1979, 1989 et 2007 (Lemieux et Clarke, « The International Ban on Ivory Sales and Its Effects on Elephant Poaching in Africa », *British Journal of Criminology*, 49(4), 2009).

L'organisation à but non lucratif David Sheldrick Wildlife a été créée en 1977 en mémoire du naturaliste David Leslie William Sheldrick, qui a fondé le parc national de Tsavo East au Kenya et dirigé l'unité de planification du département de conservation et de gestion de la faune dans ce pays. Les responsables de l'organisation Sheldrick voudraient savoir ce que ces données indiquent quant à l'évolution des populations d'éléphants dans les différents pays d'Afrique depuis 1979.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Tableau 3.13 ***Les populations d'éléphants dans plusieurs pays d'Afrique en 1979, 1989 et 2007***

	Population d'éléphants		
Pays	1979	1989	2007
Angola	12 400	12 400	2 530
Botswana	20 000	51 000	175 487
Cameroun	16 200	21 200	15 387
République de Centre Afrique	63 000	19 000	3 334
Chad	15 000	3 100	6 435
Congo	10 800	70 000	22 102
République démocratique du Congo	377 700	85 000	23 714
Gabon	13 400	76 000	70 637
Kenya	65 000	19 000	31 636
Mozambique	54 800	18 600	26 088
Somalie	24 300	6 000	70
Soudan	134 000	4 000	300
Tanzanie	316 300	80 000	167 003
Zambie	150 000	41 000	29 231
Zimbabwe	30 000	43 000	99 107

Éléphants d'Afrique

Rapport

Utiliser les statistiques descriptives pour résumer les données et commenter l'évolution des populations d'éléphants dans les pays d'Afrique depuis 1979. A minima, votre rapport doit inclure les éléments suivants.

1. L'évolution annuelle moyenne des populations d'éléphants pour chaque pays entre 1979 et 1989 et une discussion relative aux pays qui ont vu les plus grands changements dans la population des éléphants sur cette période de 10 ans.
2. L'évolution annuelle moyenne dans les populations d'éléphants pour chaque pays entre 1989 et 2007 et une discussion relative aux pays qui ont vu les plus grands changements dans la population des éléphants sur cette période de 18 ans.
3. Une comparaison des résultats obtenus aux questions 1 et 2, et une discussion sur les conclusions que vous pouvez tirer de cette comparaison.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

ANNEXE 3.1 STATISTIQUES DESCRIPTIVES AVEC MINITAB

Dans cette annexe, nous décrivons comment utiliser Minitab pour développer des statistiques descriptives et construire des boîtes-à-pattes. Nous montrons ensuite comment utiliser Minitab pour obtenir les mesures de covariance et de corrélation entre deux variables.

A3.1.1 Statistiques descriptives

Le tableau 3.1 regroupe les données sur les salaires initiaux de douze jeunes diplômés d'une école de commerce. Ces données sont disponibles dans la colonne C2 du fichier Salaires de départ 2012. Les étapes suivantes génèrent les statistiques descriptives évoquées.

Étape 1. Sélectionner le menu **Stat**
Étape 2. Sélectionner le menu **Basic Statistics**
Étape 3. Sélectionner l'option **Display Descriptive Statistics**
Étape 4. Lorsque la boîte de dialogue Display Descriptive Statistics apparaît :
Entrer C2 dans la boîte **Variables**
Cliquer sur **OK**

La figure 3.15 représente les statistiques descriptives pour les données sur les salaires obtenues en utilisant Minitab. La définition des en-têtes est indiquée ci-dessous.

N	Nombre d'observations
N*	Nombre de données manquantes
Mean	Moyenne
SE Mean	Erreur quadratique moyenne
StDev	Écart type
Minimum	Valeur de l'observation la plus petite
Q1	Premier quartile
Median	Médiane
Q3	Troisième quartile
Maximum	Valeur de l'observation la plus grande

L'erreur quadratique moyenne, notée SEMean, est calculée en divisant l'écart type par la racine carrée de *N*. L'interprétation de cette mesure sera explicitée au

N	N*	Mean	SE Mean	StDev
12	0	3 540,0	47,8	165,7
Minimum	**Q1**	**Median**	**Q3**	**Maximum**
3 310,0	3 457,5	3 505,0	3 625,0	3 925,0

Figure 3.15 ***Statistiques descriptives fournies par Minitab***

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

chapitre 7, lorsque seront introduits les concepts d'échantillonnage et de distributions d'échantillonnage.

Les 10 statistiques descriptives qui apparaissent à la figure 3.15 sont les statistiques descriptives par défaut, sélectionnées automatiquement par Minitab. Ces statistiques descriptives intéressent la majorité des utilisateurs. Toutefois, Minitab fournit 15 statistiques descriptives supplémentaires qui peuvent être sélectionnées par l'utilisateur. La variance, le coefficient de variation, l'étendue, l'étendue interquartile, le mode et le degré d'asymétrie font partie des statistiques descriptives supplémentaires disponibles. Ces statistiques descriptives supplémentaires peuvent être obtenues en modifiant l'étape 4 comme suit :

Étape 4. Lorsque la boîte de dialogue Display Descriptive Statistics apparaît :
Sélectionner **Statistics**
Lorsque la boîte de dialogue Display Descriptive Statistics – Statistics apparaît :
Sélectionner la statistique descriptive souhaitée ou choisir **All** pour obtenir les 25 statistiques descriptives
Cliquer sur **OK**
Cliquer sur **OK**

Notez pour finir que les quartiles obtenus par Minitab $Q_1 = 3\,857{,}5$ et $Q_3 = 4\,025{,}0$ sont légèrement différents de ceux obtenus dans la section 3.1 ($Q_1 = 3\,865$ et $Q_3 = 4\,000$). Ceci est dû aux différentes conventions[2] utilisées pour identifier les quartiles. Par conséquent, les valeurs de Q_1 et de Q_3 fournies par une certaine convention ne sont pas forcément identiques aux valeurs fournies par une autre convention. Cependant, les différences sont négligeables, et les résultats fournis ne doivent pas fausser l'interprétation des quartiles.

A3.1.2 Boîte-à-pattes

Les étapes suivantes permettre de construire une boîte-à-pattes à partir des données sur les salaires initiaux.

Étape 1. Sélectionner le menu **Graph**
Étape 2. Sélectionner **Boxplot**
Étape 3. Sélectionner **Simple** et cliquer sur **OK**
Étape 4. Lorsque la boîte de dialogue Boxplot – One Y, Simple apparaît :
Entrer C2 dans la boîte **Graph variables**
Cliquer sur **OK**

[2] Lorsque les n observations sont classées en ordre croissant, Minitab utilise les positions données par $(n+1)/4$ et $3(n+1)/4$ pour localiser Q_1 et Q_3, respectivement. Lorsque la position obtenue est un chiffre décimal, Minitab extrapole entre les valeurs des deux observations adjacentes pour déterminer le quartile correspondant.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

A3.1.3 Covariance et corrélation

Le tableau 3.6 regroupe les données sur le nombre de spots publicitaires et le volume des ventes d'un magasin d'équipement hi-fi. Ces données sont disponibles dans le fichier en ligne Hi-fi, avec le nombre de spots publicitaires enregistré dans la colonne C2 et le volume des ventes dans la colonne C3. Les étapes suivantes illustrent comment calculer la covariance pour deux variables avec Minitab.

Étape 1. Sélectionner le menu **Stat**
Étape 2. Sélectionner le menu **Basic Statistics**
Étape 3. Sélectionner l'option **Covariance**
Étape 4. Lorsque la boîte de dialogue Covariance apparaît :
Entrer C2 C3 dans la boîte **Variables**
Cliquer sur **OK**

La feuille de résultats de Minitab fournit la variance pour chaque variable en plus de la covariance.

Pour obtenir le coefficient de corrélation pour le nombre de spots publicitaires et le volume des ventes, une seule modification est nécessaire dans la procédure précédente. À l'étape 3, choisir l'option **Correlation**.

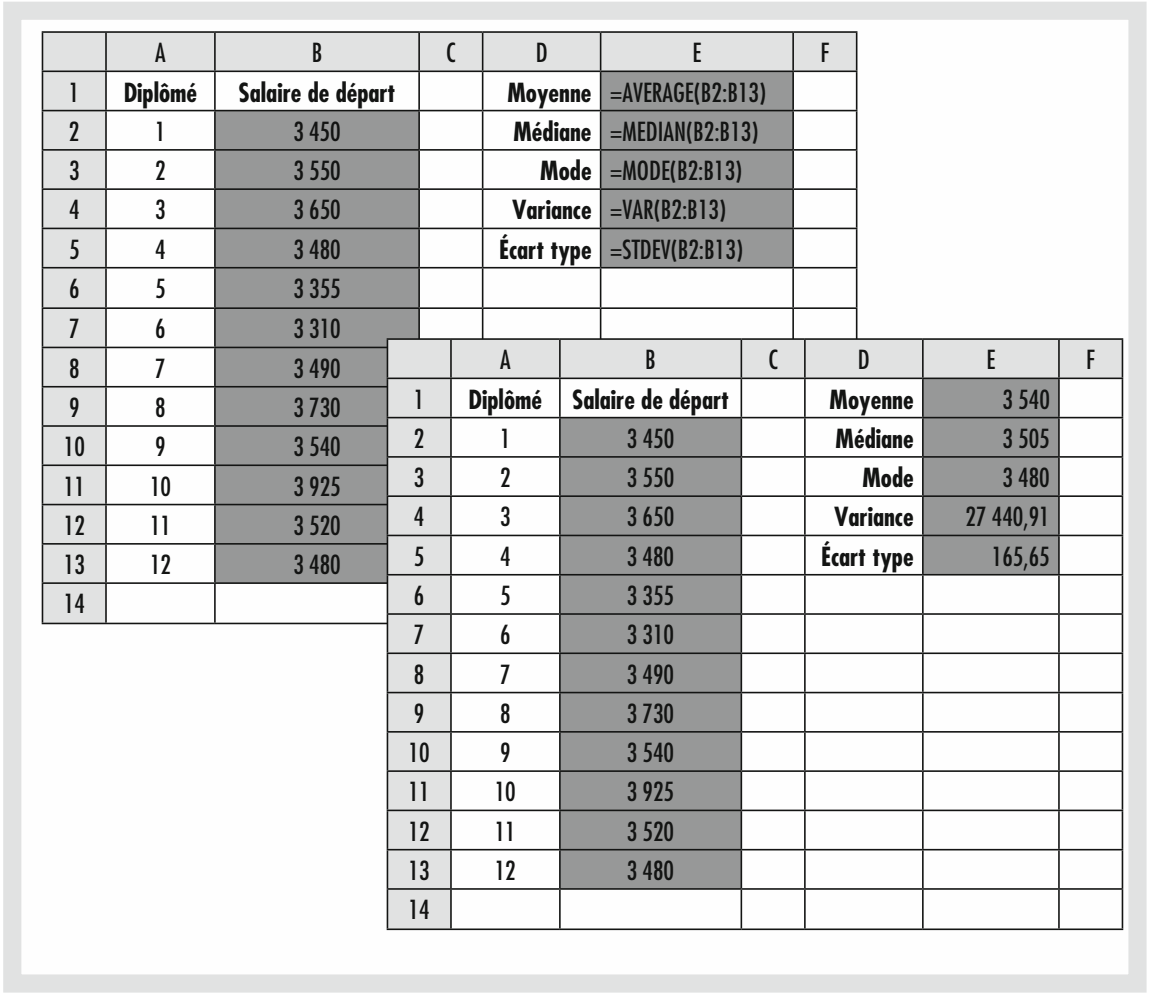
ANNEXE 3.2 STATISTIQUES DESCRIPTIVES AVEC EXCEL

Excel peut être utilisé pour générer les statistiques descriptives discutées dans ce chapitre. Dans cette annexe, nous montrons comment utiliser Excel pour obtenir les mesures de tendance centrale et de dispersion pour une seule variable, ainsi que la covariance et le coefficient de corrélation, mesures de la relation entre deux variables.

A3.2.1 Utiliser les fonctions Excel

Excel propose des fonctions pour calculer la moyenne, la médiane, le mode, la variance et l'écart type d'échantillon. Nous illustrons l'utilisation de ces fonctions en calculant ces différentes statistiques descriptives pour les données relatives aux salaires initiaux des jeunes diplômés d'une école de commerce, présentées dans le tableau 3.1 (fichier en ligne Salaire de départ 2012). Référez-vous à la figure 3.16 pour suivre les procédures. Les données sont enregistrées dans la colonne B.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.



	A	B	C	D	E	F
1	**Diplômé**	**Salaire de départ**		**Moyenne**	=AVERAGE(B2:B13)	
2	1	3 450		**Médiane**	=MEDIAN(B2:B13)	
3	2	3 550		**Mode**	=MODE(B2:B13)	
4	3	3 650		**Variance**	=VAR(B2:B13)	
5	4	3 480		**Écart type**	=STDEV(B2:B13)	
6	5	3 355				
7	6	3 310				
8	7	3 490				
9	8	3 730				
10	9	3 540				
11	10	3 925				
12	11	3 520				
13	12	3 480				
14						

	A	B	C	D	E	F
1	**Diplômé**	**Salaire de départ**		**Moyenne**	3 540	
2	1	3 450		**Médiane**	3 505	
3	2	3 550		**Mode**	3 480	
4	3	3 650		**Variance**	27 440,91	
5	4	3 480		**Écart type**	165,65	
6	5	3 355				
7	6	3 310				
8	7	3 490				
9	8	3 730				
10	9	3 540				
11	10	3 925				
12	11	3 520				
13	12	3 480				
14						

Figure 3.16 ***Utiliser les fonctions Excel pour calculer la moyenne, la médiane, le mode, la variance et l'écart type***

La fonction AVERAGE d'Excel peut être utilisée pour calculer la moyenne en entrant la formule suivante dans la cellule E1 :

= AVERAGE (B2 : B13)

De façon similaire, les fonctions = MEDIAN (B2 : B13), = MODE.SNGL (B2 : B13), = VAR (B2 : B13) et = STDEV (B2 : B13) sont entrées dans les cellules E2 : E5 pour calculer respectivement la médiane, le mode, la variance et l'écart type. La feuille de résultats au premier plan de la figure 3.16 présente les valeurs obtenues en utilisant les fonctions Excel, similaires à celles obtenues auparavant dans ce chapitre.

Pour trouver la variance, l'écart type et la covariance pour des données relatives à une population, suivre les mêmes étapes mais utiliser les fonctions VAR.P, STDEV.P et COV.P.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

	A	B	C	D	E	F
1	Semaine	Nombre de spots publicitaires	Volume des ventes		Covariance de la population	=COVAR(B2:B11,C2:C11)
2	1	2	50		Corrélation de l'échantillon	=CORREL(B2:B11,C2:C11)
3	2	5	57			
4	3	1	41			
5	4	3	54			
6	5	4	54			
7	6	1	38			
8	7	5	63			
9	8	3	48			
10	9	4	59			
11	10	2	46			
12						

	A	B	C	D	E	F
1	Semaine	Nombre de spots publicitaires	Volume des ventes		Covariance de la population	9,90
2	1	2	50		Corrélation de l'échantillon	0,93
3	2	5	57			
4	3	1	41			
5	4	3	54			
6	5	4	54			
7	6	1	38			
8	7	5	63			
9	8	3	48			
10	9	4	59			
11	10	2	46			
12						

Figure 3.17 *Utiliser les fonctions Excel pour calculer la covariance et la corrélation*

Excel propose également des fonctions qui peuvent être utilisées pour calculer les coefficients de covariance et de corrélation. Nous montrons ici comment ces fonctions peuvent être utilisées pour calculer la covariance d'échantillon et le coefficient de corrélation d'échantillon pour les données relatives au magasin d'équipement hi-fi, figurant dans le tableau 3.6 (fichier en ligne Hi-fi). Référez-vous à la figure 3.17 pour suivre les étapes de la procédure.

La fonction Covariance d'Excel, COVARIANCE.S, peut être utilisée pour calculer la covariance de l'échantillon en entrant la formule suivante dans la cellule F1 :

$$= \text{COVARIANCE.S (B2 : B11, C2 : C11)}$$

De façon similaire, la formule = CORREL (B2 : B11, C2 : C11) est entrée dans la cellule F2 pour calculer le coefficient de corrélation de l'échantillon. La feuille de calcul au premier plan de la figure 3.17 fournit les valeurs calculées par les fonctions Excel. Notez que la valeur de la covariance d'échantillon (11) est identique à celle obtenue en utilisant l'équation (3.12). De même, la valeur du coefficient de corrélation de l'échantillon (0,93) est la même que celle obtenue en utilisant l'équation (3.14).

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

A3.2.2 Utiliser les outils de statistiques descriptives d'Excel

Comme nous l'avons déjà montré, Excel fournit des fonctions statistiques pour calculer des statistiques descriptives d'un ensemble de données. Ces fonctions peuvent être utilisées pour calculer une à une les statistiques (par exemple, la moyenne, la variance, etc.). Excel propose également une variété d'outils d'analyse des données. L'un de ces outils, appelé Statistiques Descriptives, permet à un utilisateur de calculer une variété de statistiques descriptives simultanément. Nous montrons ici comment cet outil peut être utilisé pour calculer les statistiques descriptives des données sur les salaires initiaux des jeunes diplômés du tableau 3.1 (fichier en ligne Salaire de départ 2012).

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

Étape 2. Dans le groupe **Analysis**, cliquer sur **Data Analysis**

Étape 3. Lorsque la boîte de dialogue Data Analysis apparaît :
Choisir **Descriptive Statistics**

Étape 4. Lorsque la boîte de dialogue Descriptive Statistics apparaît :
Entrer B1:B13 dans la boîte **Input Range**
Sélectionner **Grouped By Columns**
Sélectionner **Labels in First Row**
Sélectionner **Output Range**
Entrer D1 dans la boîte **Output Range** (Ceci permet d'identifier le coin supérieur gauche de la feuille de calcul où les statistiques descriptives apparaîtront)

	A	B	C	D	E	F
1	**Diplômé**	**Salaire de départ**		**Salaire de départ**		
2	1	3 450				
3	2	3 550		**Moyenne**	3 540	
4	3	3 650		Erreur quadratique moyenne	47,82	
5	4	3 480		**Médiane**	3 505	
6	5	3 355		**Mode**	3 480	
7	6	3 310		**Écart type**	165,65	
8	7	3 490		**Variance**	27 440,91	
9	8	3 730		Kurtosis	1,7189	
10	9	3 540		**Asymétrie**	1,0911	
11	10	3 925		**Étendue**	615	
12	11	3 520		**Minimum**	3 310	
13	12	3 480		**Maximum**	3 925	
14				**Somme**	42 480	
				Observation	12	

Figure 3.18 *Feuille de résultats de l'outil Statistiques Descriptives d'Excel*

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

Sélectionner **Summary Statistics**
Cliquer sur **OK**

Les statistiques descriptives fournies par Excel apparaissent dans les cellules D1 : E15 de la figure 3.18. Celles traitées dans ce chapitre apparaissent en gras. Les autres seront étudiées ultérieurement dans cet ouvrage ou dans d'autres ouvrages plus avancés.

REMARQUES

Si la fonction **Analysis** n'apparaît pas dans votre barre des tâches ou si l'option **Data Analysis** n'apparaît pas, vous devez activer le pack d'outils Analysis en suivant les trois étapes suivantes :

1. Cliquer sur l'onglet **Fichier**, puis sur **Options** et ensuite sur la catégorie **Add-Ins**.
2. Dans la boîte **Manage**, cliquer sur **Excel Add-ins** et alors cliquer sur **Go**. La boîte de dialogue Add-Ins apparaîtra.
3. Dans la boîte **Add-Ins available**, sélectionner le complément **Data Analysis ToolPak** et cliquer sur **OK**.

Le groupe **Analysis** et l'option **Data Analysis** sont maintenant disponibles.

ANNEXE 3.3 STATISTIQUES DESCRIPTIVES AVEC STATTOOLS

Dans cette annexe, nous décrivons comment utiliser StatTools pour obtenir différentes statistiques descriptives et construire des boîtes-à-pattes. Nous montrons ensuite comment utiliser StatTools pour obtenir les mesures de covariance et de corrélation entre deux variables.

A3.3.1 Statistiques descriptives

Nous utilisons les données sur les salaires initiaux du tableau 3.1 pour illustrer la démarche (fichier en ligne Salaires de départ 2012). Commencez par utiliser Data Set Manager pour créer un ensemble de données StatTools pour ces données en utilisant la procédure décrite dans l'annexe du chapitre 1. Les étapes suivantes génèreront de nombreuses statistiques descriptives.

Étape 1. Cliquer sur le bouton **StatTools** dans la barre des tâches
Étape 2. Dans le groupe **Analyses**, cliquer sur **Summary Statistics**
Étape 3. Choisir l'option **One-Variable Summary**
Étape 4. Lorsque la boîte de dialogue apparaît :
Dans la section **Variables**, sélectionner **Salaires initiaux**
Cliquer sur **OK**

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.

De nombreuses statistiques descriptives apparaîtront, comme celles figurant dans la figure 3.18.

A3.3.2 Boîte-à-pattes

Nous utilisons les données sur les salaires initiaux du tableau 3.1 pour illustrer la démarche (fichier en ligne Salaires de départ 2012). Commencez par utiliser Data Set Manager pour créer un ensemble de données StatTools pour ces données en utilisant la procédure décrite dans l'annexe du chapitre 1. Les étapes suivantes créeront une boîte-à-pattes pour ces données.

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 **Box-Whisker Plot**
Étape 4. Lorsque la boîte de dialogue apparaît :
Dans la section **Variables**, sélectionner **Salaires initiaux**
Cliquer sur **OK**

Le symbole ¨ identifie une valeur aberrante et le symbole x la moyenne.

A3.3.3 Covariance et corrélation

Nous utilisons les données sur le magasin de hi-fi du tableau 3.6 pour illustrer le calcul de la covariance d'échantillon et du coefficient de corrélation d'échantillon (fichier en ligne Hi-fi). Commencez par utiliser Data Set Manager pour créer un ensemble de données StatTools pour ces données en utilisant la procédure décrite dans l'annexe du chapitre 1. Les étapes suivantes fourniront la covariance et le coefficient de corrélation d'échantillon.

Étape 1. Cliquer sur le bouton **StatTools** dans la barre des tâches
Étape 2. Dans le groupe **Analyses**, cliquer sur **Summary Statistics**
Étape 3. Choisir l'option **Correlation and Covariance**
Étape 4. Lorsque la boîte de dialogue apparaît :
Dans la section **Variables**,
Sélectionner **Nombre de spots publicitaires**
Sélectionner **Volume des ventes**
Dans la section **Tables to Create**
Sélectionner **Table of Correlations**
Sélectionner **Table of Covariances**
Dans la section **Table Structure** sélectionner **Symmetric**
Cliquer sur **OK**

Un tableau contenant le coefficient de corrélation et la covariance apparaîtra.

Statistiques pour l'économie et la gestion © DeBoeck. Reproduction interdite.