e) Complétez le tableau en ajoutant les colonnes de pourcentage et de pourcentage cumulé.

f) Présentez vos données sous forme de graphique.

g) Trouvez le mode de cette distribution et donnez sa signification en tenant compte du contexte.

h) Trouvez la médiane de cette distribution et donnez sa signification en tenant compte du contexte.

i) Trouvez la moyenne de cette distribution et donnez sa signification en tenant compte du contexte.

j) Que diriez-vous au sujet de la symétrie de cette distribution ?

k) Trouvez l'écart type de cette distribution et donnez sa signification en tenant compte du contexte.

l) Déterminez le coefficient de variation et indiquez ce qu'il représente en tenant compte du contexte.

m) Précisez la valeur correspondant à C_{40}, à D_8 et à Q_1 et interprétez chacune de ces valeurs en tenant compte du contexte.

n) Quelle serait la cote Z d'un ménage possédant 2 téléphones ?

o) Combien de téléphones a un ménage dont la cote Z est 2,88 ?

4.2 LES VARIABLES QUANTITATIVES CONTINUES

En ce qui concerne les données provenant d'une variable quantitative continue, on constate le très grand nombre de valeurs différentes (à l'exclusion, ici, du cas où les choix de réponses sont sous forme de classes). Nous verrons comment procéder pour décrire les données et pour interpréter les différentes mesures à partir de la série de données brutes.

4.2.1 La présentation sous forme de tableau

EXEMPLE 4.24

Les données suivantes ont été recueillies lors d'un sondage sur le revenu des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps en 1990. Pour ce sondage, 160 hommes canadiens ont été interrogés, et on a noté le revenu de chacun.

Variable	Le revenu
Population	Tous les hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps en 1990
Unité statistique	Un homme canadien âgé de 15 ans et plus ayant travaillé à plein temps en 1990
Taille de l'échantillon	$n = 160$

Voici donc le revenu, exprimé en milliers de dollars, de ces 160 hommes canadiens interrogés[10] :

Note : Les données sont exprimées en milliers de dollars. Si l'on avait pris le revenu jusqu'au cent près, les 160 données seraient toutes différentes ou presque.

10. Échantillon fictif simulé à partir des informations contenues dans *Les gains des Canadiens*, Ottawa, Statistique Canada, Catalogue 96-317F, 1994, p. 46, tableau 4.3.

35,9	27,8	41,1	39,1	49,4	41,0	45,4	31,8	47,2	42,3
36,8	26,9	39,6	28,6	30,9	47,9	39,4	34,5	41,6	45,7
33,8	41,4	33,2	37,3	39,9	29,8	37,1	42,2	36,7	38,7
26,5	40,4	43,1	43,0	27,1	52,6	45,3	35,9	33,1	36,9
29,0	47,7	41,6	48,7	39,7	39,3	35,5	40,2	36,7	34,7
36,3	24,5	37,7	36,5	46,0	33,8	33,1	47,3	37,1	39,5
38,6	45,7	44,2	32,1	30,1	40,3	38,7	36,9	38,1	34,9
44,1	38,4	42,6	39,1	33,7	37,0	30,3	37,1	31,2	34,0
46,1	35,8	41,0	34,3	39,9	39,1	39,9	37,2	45,2	36,6
39,7	43,7	34,4	36,0	40,8	27,2	30,1	50,6	39,2	37,5
43,5	43,2	41,4	39,5	36,0	35,7	38,0	40,6	36,7	25,3
44,0	37,6	46,0	38,9	34,8	31,5	35,4	38,3	33,7	47,1
47,1	39,1	28,7	37,9	39,4	41,4	39,3	49,3	45,4	40,6
22,6	32,7	51,7	39,5	32,8	42,7	35,7	36,6	55,7	39,7
40,2	41,1	36,8	49,6	46,7	34,3	39,6	49,0	30,7	34,2
39,0	41,3	37,8	34,6	46,0	35,1	38,0	41,1	37,0	46,4

Il y a un trop grand nombre de valeurs différentes pour procéder comme dans le cas d'une variable quantitative discrète de la section 4.1.

Pour présenter les données sous forme de tableau, il faut les regrouper en classes. De telles données sont appelées **données groupées** et sont présentées au tableau 4.14.

TABLEAU 4.14
Répartition des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps en 1990, en fonction de leur revenu

Revenu (milliers de dollars)	Point milieu (milliers de dollars)	Nombre de Canadiens	Pourcentage des Canadiens	Pourcentage cumulé des Canadiens
De 20 à moins de 25	22,5	2	1,25	1,25
De 25 à moins de 30	27,5	10	6,25	7,50
De 30 à moins de 35	32,5	28	17,50	25,00
De 35 à moins de 40	37,5	63	39,38	64,38
De 40 à moins de 45	42,5	30	18,75	83,13
De 45 à moins de 50	47,5	23	14,38	97,51
De 50 à moins de 55	52,5	3	1,88	99,39
De 55 à moins de 60	57,5	1	0,63	100,00
Total		160	100,00	

Source : Adapté du tableau 4.3 de Statistique Canada. *Op. cit.*, p. 46.

La démarche à suivre est la suivante :

- **Première étape : Titrer le tableau**

  La formulation générale du titre est toujours **Répartition des unités statistiques en fonction de la variable**. Il faut évidemment adapter ce titre à chacune des situations.

- **Deuxième étape : Délimiter les classes**

  Les classes sont délimitées dans la première colonne. Puisque la variable est continue, cela signifie que les données pourraient prendre n'importe quelle valeur, avec autant de décimales imaginables, entre un minimum et un maximum donnés. On ne peut pas regrouper ces données de la même façon que les données discrètes,

car il y a trop de valeurs différentes. De plus, un regroupement par valeurs différentes ne laisse pas transparaître la continuité de la variable. Dans un tel cas, il faut réunir les données par classes contiguës.

- Déterminer le nombre de classes. Il faut d'abord choisir le nombre de classes le plus approprié pour regrouper les données. Il s'agit d'opter pour un nombre de classes qui ne soit ni trop petit ni trop grand. En 1926, W.H. Sturges a présenté une formule pour déterminer le nombre de classes qui était basée sur un modèle qui revient souvent :

  Nombre de classes = $1 + 3{,}322 \cdot \log n$,

  où n représente la taille de l'échantillon.

  Le tableau 4.15 présente le résultat de la formule de Sturges.

TABLEAU 4.15
Application de la formule de Sturges

Taille n	Nombre de classes souhaité
$23 \leq n \leq 45$	6
$46 \leq n \leq 90$	7
$91 \leq n \leq 180$	8
$181 \leq n \leq 361$	9
$362 \leq n \leq 723$	10
$724 \leq n \leq 1\,447$	11
$1\,448 \leq n \leq 2\,895$	12
$2\,896 \leq n \leq 5\,791$	13
$5\,792 \leq n \leq 11\,582$	14
$11\,583 \leq n \leq 23\,165$	15

La première colonne représente la taille de l'échantillon et la deuxième, le nombre de classes souhaité.

Note : Plusieurs tailles nécessitent le même nombre de classes. Ainsi, tous les échantillons dont la taille se situe de 91 à 180 nécessitent 8 classes.

Dans cet exemple, on a 160 données. Comme cette taille d'échantillon se situe dans la catégorie 91 à 180, le nombre de classes souhaité est donc 8.

- Évaluer l'étendue des données. L'étendue des données d'un échantillon est l'écart entre la plus petite donnée et la plus grande donnée. Ainsi, dans cet échantillon, la plus grande donnée est 55,7, c'est-à-dire 55 700 $, et la plus petite donnée est 22,6, c'est-à-dire 22 600 $. L'étendue de ces données est donc :

  Étendue = $55{,}7 - 22{,}6 = 33{,}1$, c'est-à-dire 33 100 $.

- Déterminer la largeur des classes. La technique présentée concerne l'élaboration de classes de largeurs égales. Cependant, on remarque que dans certains sondages les classes ne sont pas toutes de même largeur. Les raisons qui font que certaines classes ont des largeurs différentes varient d'un sondage à l'autre. Dans le présent ouvrage, de telles classes ne seront pas construites même si, à l'occasion, on travaillera avec des tableaux qui en comportent.

  La largeur des classes s'obtient en divisant l'étendue par le nombre de classes :

  $$\text{Largeur} = \frac{\text{Étendue}}{\text{Nombre de classes}}.$$

  À partir de cette valeur, on choisit la largeur à utiliser pour les classes (un multiple de 5 est souvent choisi, car il facilite les calculs et la lecture des graphiques). Pour l'échantillon de l'exemple :

$$\text{Largeur} = \frac{\text{Étendue}}{\text{Nombre de classes}} = \frac{33,1}{8} = 4,1375.$$

Pour choisir la largeur des classes, il faut tenir compte de l'ordre de grandeur des données Ainsi, dans l'exemple, la largeur des classes sera de 5 milliers de dollars (5 000 $).

- Former les classes. Il s'agit maintenant de choisir le point de départ de la première classe, c'est-à-dire la borne inférieure de la première classe à partir de laquelle les autres seront déterminées. Ce choix peut entraîner la modification du nombre de classes ou de la largeur des classes. Il faut retenir que ce choix doit faciliter la représentation graphique ainsi que le calcul des différentes mesures.

  Chaque valeur des données doit entrer dans une classe ; cette propriété s'appelle l'**exhaustivité**. Il faut aussi que chaque valeur puisse entrer dans une seule classe ; cette propriété s'appelle l'**exclusivité.** Pour respecter cette dernière propriété, la convention suivante sera utilisée : la borne inférieure est incluse dans la classe, tandis que la borne supérieure en est exclue. Par conséquent, les classes auront la forme « De... à moins de... ».

Dans cet exemple, avec une largeur de 5 000 $ pour couvrir une étendue de 33 100 $, 7 classes seraient suffisantes, mais cela dépend aussi de la borne inférieure de la première classe. La plus petite donnée étant 22 600 $, 20 000 $ serait un bon choix pour commencer les classes. Avec ce choix, la septième classe sera « De 50 000 $ à moins de 55 000 $ ». Où faudra-t-il placer les données dont les valeurs vont de 55 000 $ à 55 700 $? Dans une huitième classe.

– **Troisième étape : Déterminer le point milieu**

Dans la deuxième colonne, on indique le point milieu de chacune des classes. Le point milieu d'une classe s'obtient en additionnant la borne inférieure et la borne supérieure de la classe et en divisant le résultat par deux :

$$\text{Point milieu} = \frac{\text{Borne inférieure} + \text{Borne supérieure}}{2}.$$

Le point milieu de la première classe est :

$$\frac{20+25}{2} = 22,5 \text{ milliers de dollars.}$$

Les points milieux des classes seront utilisés pour tracer un type de graphique et pour calculer la moyenne et l'écart type à partir d'un tableau où les données sont regroupées en classes.

– **Quatrième étape : Indiquer le nombre d'unités statistiques par classe**

Dans la troisième colonne, on indique le nombre d'unités compilées dans chacune des classes. Ce nombre est appelé **fréquence** de la classe. Le nombre d'unités se trouve par le dépouillement des données, comme dans la section 4.1 relative aux variables quantitatives discrètes.

Ainsi, il y a 2 hommes canadiens ayant travaillé à temps plein en 1990 dont le revenu se situe de 20 à moins de 25 milliers de dollars, 10 dont le revenu se situe de 25 à moins de 30 milliers de dollars...

– **Cinquième étape : Établir le pourcentage des unités statistiques par classe**

La quatrième colonne exprime le nombre d'unités ou la fréquence sous forme de pourcentage.

Ainsi, il y a 1,25 % des hommes canadiens ayant travaillé à temps plein en 1990 dont le revenu se situe de 20 à moins de 25 milliers de dollars, 6,25 % dont le revenu se situe de 25 à moins de 30 milliers de dollars...

– **Sixième étape : Déterminer le pourcentage cumulé des unités statistiques par classe**

Dans la cinquième colonne, on indique le pourcentage cumulé d'une classe. Ce dernier correspond au pourcentage de données qui appartiennent à cette classe ou aux classes précédentes, ce qui veut dire toutes les données dont les valeurs sont inférieures ou égales à la borne supérieure de la classe. Le pourcentage cumulé s'interprète donc à l'aide de la borne supérieure de la classe.

Il y a 1,25 % des hommes canadiens ayant travaillé à temps plein en 1990 dont le revenu est inférieur à 25 milliers de dollars, 7,50 % dont le revenu est inférieur à 30 milliers de dollars... et 100 % dont le revenu est inférieur à 60 000 $. On dit inférieur, car la borne supérieure de chaque classe ne fait pas partie de la classe : « De... à moins de... ».

4.2.2 La présentation sous forme de graphique

Plusieurs formes de graphiques permettent de représenter les données. Dans le cas des données quantitatives continues, les graphiques utilisés sont l'histogramme, le polygone des pourcentages et la courbe des pourcentages cumulés ou ogive.

EXEMPLE 4.25

Reprenons l'exemple 4.24 sur le revenu des hommes canadiens.

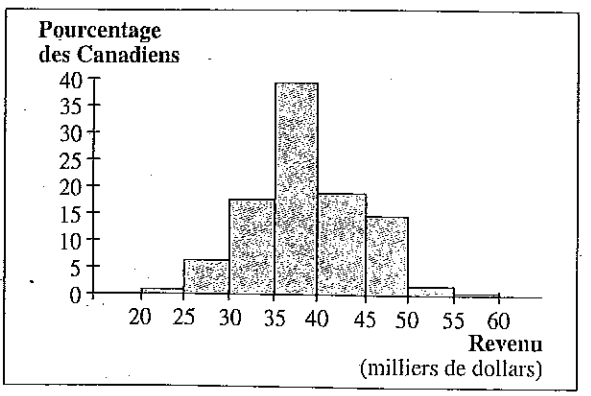
a) L'histogramme

La figure 4.8 illustre l'histogramme construit à partir des données du tableau 4.14.

TABLEAU 4.14

Revenu (milliers de dollars)	Point milieu (milliers de dollars)	Nombre de Canadiens	Pourcentage des Canadiens	Pourcentage cumulé des Canadiens
De 20 à moins de 25	22,5	2	1,25	1,25
De 25 à moins de 30	27,5	10	6,25	7,50
De 30 à moins de 35	32,5	28	17,50	25,00
De 35 à moins de 40	37,5	63	39,38	64,38
De 40 à moins de 45	42,5	30	18,75	83,13
De 45 à moins de 50	47,5	23	14,38	97,51
De 50 à moins de 55	52,5	3	1,88	99,39
De 55 à moins de 60	57,5	1	0,63	100,00
Total		160	100,00	

FIGURE 4.8
Répartition des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps en 1990, en fonction de leur revenu



– **Première étape : Titrer l'histogramme**

Le titre peut être le même que celui du tableau, puisqu'il représente la même répartition mais sous forme visuelle.

- **Deuxième étape : Placer les valeurs de la variable sur l'axe horizontal**

  On situe d'abord les bornes des classes (première colonne du tableau) sur l'axe horizontal, puis on identifie l'axe en indiquant les unités de mesure utilisées.

- **Troisième étape : Placer le pourcentage d'unités statistiques sur l'axe vertical**

  Sur l'axe vertical, on trace une échelle pour les pourcentages (quatrième colonne du tableau 4.14). Puisque le pourcentage le plus élevé est 39,38 %, il est inutile d'aller plus haut que 40 %. (On pourrait également choisir d'utiliser le nombre d'unités dans chaque classe.) Il faut identifier l'axe vertical en indiquant les unités de mesure de l'échelle choisie.

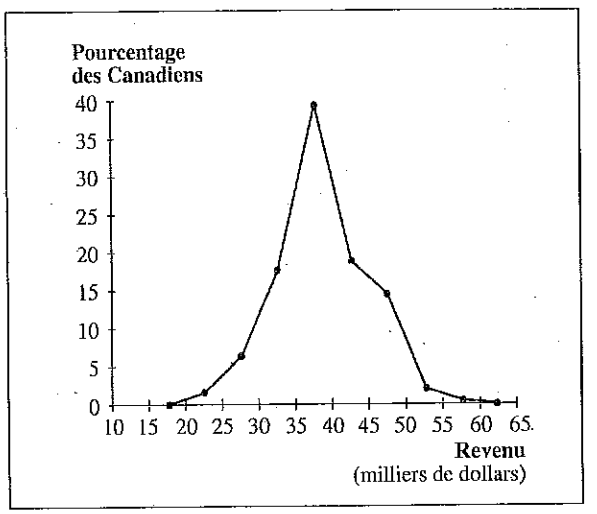
- **Quatrième étape : Tracer les rectangles**

  Pour chaque classe, on trace un rectangle dont la base est la largeur de la classe et la hauteur, le pourcentage d'unités dans la classe (si l'échelle est faite à partir de la quatrième colonne du tableau 4.14) ou le nombre d'unités (si l'échelle est faite à partir de la troisième colonne du même tableau).

b) Le polygone des pourcentages

La figure 4.9 illustre le polygone des pourcentages construit en se basant sur les données du tableau 4.14.

FIGURE 4.9
Répartition des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps en 1990, en fonction de leur revenu



Le polygone donne l'allure générale de la distribution de la variable étudiée. On peut placer plusieurs polygones sur un même graphique, ce qui n'est pas possible avec les histogrammes. (Pour faciliter les comparaisons, on peut placer sur le même graphique le polygone de la répartition du revenu des hommes et le polygone de la répartition du revenu des femmes.)

La démarche à suivre pour bâtir un polygone est la suivante :

- **Première étape : Titrer la figure**

  Le titre peut être le même que celui de l'histogramme, puisqu'il représente la même répartition.

- **Deuxième étape : Placer les valeurs de la variable sur l'axe horizontal**

  Pour que la figure soit un polygone fermé aux deux extrémités, à chacune des extrémités, il faut prévoir une classe de même largeur que les autres ayant 0 % d'unités, puisqu'il n'y a pas de données dont les valeurs se situent dans ces

classes. Ici les classes ayant 0 % d'unités sont « De 15 à moins de 20 » (le point milieu est 17,5) et « De 60 à moins de 65 » (le point milieu est 62,5).

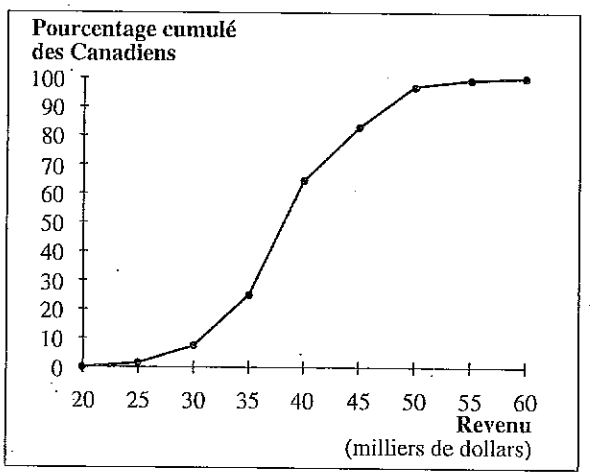
- **Troisième étape : Placer le pourcentage d'unités statistiques sur l'axe vertical**
  Sur l'axe vertical, on trace une échelle pour les pourcentages (quatrième colonne du tableau 4.14). Puisque le pourcentage le plus élevé est 39,38 %, il est inutile d'aller plus haut que 40 %. (On pourrait également choisir d'utiliser le nombre d'unités dans chaque classe.) Il faut identifier l'axe vertical en indiquant les unités de mesure de l'échelle choisie.
- **Quatrième étape : Tracer les segments**
  Vis-à-vis du point milieu de chacune des classes, on place un point à une hauteur égale au pourcentage (ou au nombre) d'unités dans la classe. On fait de même pour les deux classes ayant 0 % d'unités. Ensuite, on joint les points par des segments de droite.

Note : Parfois, il se peut qu'ajouter une classe au début ou à la fin d'une distribution de données groupées ne soit pas logiquement possible, comme ajouter une classe comprenant des valeurs négatives pour la taille ou des résultats scolaires. Dans ce cas, il faut tronquer le segment concerné.

c) La courbe des pourcentages cumulés (ogive)

La figure 4.10 illustre la courbe des pourcentages cumulés construite à partir des données du tableau 4.14.

FIGURE 4.10
Courbe des pourcentages cumulés des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps en 1990, en fonction de leur revenu



Cette courbe donne comme information le pourcentage cumulé des données depuis le début de la première classe. Autrement dit, elle donne le pourcentage des données ayant des valeurs inférieures ou égales à la valeur mentionnée.

La démarche à suivre pour bâtir une courbe des pourcentages cumulés (ogive) est :

- **Première étape : Titrer la figure**
  Le titre aura la forme suivante : Courbe des pourcentages cumulés des « unités statistiques » en fonction de la **variable**.
- **Deuxième étape : Placer les valeurs de la variable sur l'axe horizontal**
  On place d'abord les bornes des classes sur l'axe horizontal, puis on identifie l'axe en indiquant les unités de mesure utilisées.
- **Troisième étape : Placer les pourcentages cumulés sur l'axe vertical**
  Sur l'axe vertical, on trace une échelle pour les pourcentages cumulés. Dans ce cas-ci, l'échelle doit aller de 0 % à 100 %. On identifie l'axe en indiquant les unités de mesure utilisées.

– **Quatrième étape : Tracer les segments**

Vis-à-vis de la valeur de la **borne supérieure** de chacune des classes, on place un point dont la hauteur égale le pourcentage cumulé indiqué dans la dernière colonne du tableau 4.14. Vis-à-vis de la borne inférieure de la première classe, on place un point de hauteur 0 % qui est le point de départ. Avec les données groupées en classes, on ne connaît pas le pourcentage cumulé des données pour une valeur qui se situe entre les bornes d'une classe. Cependant, une répartition uniforme des données à l'intérieur d'une classe correspond à une représentation graphique sous forme linéaire (sous la forme d'une droite) pour l'accumulation des données de la classe. On suppose donc que les données sont réparties uniformément à l'intérieur de chacune des classes, ce qui permet de relier les points déjà tracés par des segments de droite.

4.2.3 Les mesures de tendance centrale

Les mesures de tendance centrale qui ont déjà été présentées dans la sous-section 4.1.3 sur les variables quantitatives discrètes s'appliquent aussi aux variables quantitatives continues. Les calculs et l'interprétation de ces mesures seront adaptés en fonction des classes plutôt que des valeurs précises.

A. Le mode (*Mo*) pour les données groupées en classes

Pour la variable quantitative discrète, le mode a été défini comme étant la valeur de la variable étudiée qui a la plus grande fréquence (le plus grand nombre d'unités statistiques) dans l'échantillon ou la population. En ce qui concerne la variable quantitative continue, une telle définition ne s'applique pas, car la variable quantitative continue peut prendre toutes les valeurs possibles entre deux valeurs. De ce fait, une donnée obtenue pour une telle variable revient rarement plus de quelques fois ; elle n'apparaît souvent qu'une seule fois. Ainsi, le mode calculé selon la définition précédente n'aurait alors aucune signification. Le fait de regrouper les données d'une variable quantitative continue en classes permet de donner une définition du mode en fonction des classes. On définira d'abord une classe modale. Ensuite, le mode sera défini sous la forme d'une valeur à l'intérieur de cette classe modale.

– **La classe modale**

Dans le cas d'une distribution où les classes sont de largeurs égales, la classe modale est celle qui a le plus grand pourcentage de données. (Le cas des distributions ayant des classes de largeurs inégales sera étudié à l'exemple 4.35.)

– **Le mode brut**

C'est la valeur centrale de la classe modale.

– **Le mode, une valeur approximative basée sur la répartition**

Il s'agit d'une valeur située dans la classe modale qui tient compte du nombre de données dans la classe précédant la classe modale et dans la classe suivant la classe modale. Le mode est défini comme suit :

$$Mo = B_i + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \cdot a.$$

a) La classe modale est d'abord repérée ;

b) B_i est la borne inférieure de la classe modale ;

c) a est la largeur de la classe modale ;

d) Δ_1 est la différence de hauteur entre le rectangle de la classe modale et le rectangle de la classe précédente ;

e) Δ_2 est la différence de hauteur entre le rectangle de la classe modale et le rectangle de la classe suivante.

EXEMPLE 4.26 Reprenons l'exemple 4.24 portant sur le revenu des hommes canadiens.

TABLEAU 4.14

Revenu (milliers de dollars)	Point milieu (milliers de dollars)	Nombre de Canadiens	Pourcentage des Canadiens	Pourcentage cumulé des Canadiens
De 20 à moins de 25	22,5	2	1,25	1,25
De 25 à moins de 30	27,5	10	6,25	7,50
De 30 à moins de 35	32,5	28	17,50	25,00
De 35 à moins de 40	37,5	63	39,38	64,38
De 40 à moins de 45	42,5	30	18,75	83,13
De 45 à moins de 50	47,5	23	14,38	97,51
De 50 à moins de 55	52,5	3	1,88	99,39
De 55 à moins de 60	57,5	1	0,63	100,00
Total		160	100,00	

Les classes du tableau 4.14 étant toutes de même largeur (5 000 $), la classe modale est « De 35 à moins de 40 milliers de dollars ». C'est dans cette classe de revenu qu'on trouve le plus de personnes, soit 63, ce qui représente une proportion de 39,38 %.

Le mode brut est 37,5 milliers de dollars, c'est la valeur centrale de la classe modale. Le revenu autour duquel il y a une plus forte concentration (densité) de données est d'environ 37 500 $.

Selon la formule, on calcule le mode de la façon suivante :

$$Mo = B_i + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \cdot a.$$

a) La classe modale est « De 35 à moins de 40 milliers de dollars » ;

b) B_i (la borne inférieure de la classe modale) égale 35 milliers de dollars ;

c) a (la largeur de la classe modale) vaut 5 milliers de dollars ;

d) La différence de hauteur entre le rectangle de la classe modale et le rectangle de la classe précédente est :

$$\Delta_1 = 39{,}38 - 17{,}50 = 21{,}88 ;$$

e) La différence de hauteur entre le rectangle de la classe modale et le rectangle de la classe suivante est :

$$\Delta_2 = 39{,}38 - 18{,}75 = 19{,}63.$$

On obtient donc le mode des données groupées :

$$Mo = B_i + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \cdot a$$

$$= 35 + \left(\frac{21{,}88}{21{,}88 + 19{,}63}\right) \cdot 5 = 35 + (0{,}5271 \cdot 5) = 35 + 2{,}636$$

$$= 37{,}636 \text{ milliers de dollars } (37\,636\,\$).$$

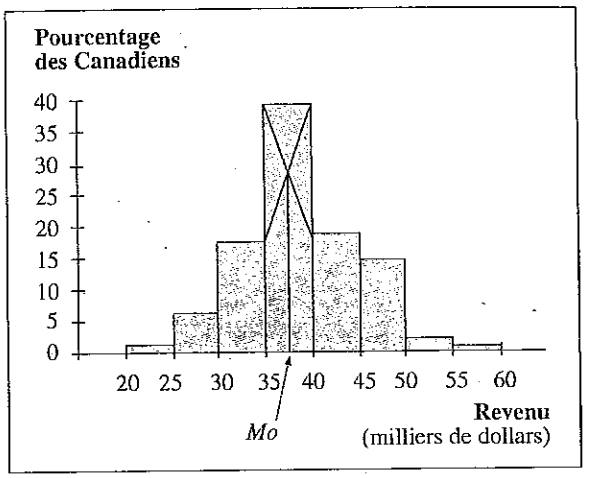
Cela signifie que le revenu autour duquel il y a la plus forte concentration (densité) de données est d'environ 37 636 $.

Note : Le mode, tel qu'il est calculé à l'aide de la formule, se situe vis-à-vis de l'intersection des deux diagonales dans le rectangle de la classe modale ;

plus le rectangle (précédent ou suivant) est élevé, plus il attire le mode dans sa direction.

La figure 4.11 illustre l'exemple que nous venons d'étudier.

FIGURE 4.11
Répartition des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps en 1990, en fonction de leur revenu



B. La médiane (*Md*) pour les données groupées en classes

La médiane est la valeur de la variable étudiée qui sépare le nombre de données ordonnées en deux groupes égaux. Lorsque les données sont groupées en classes, il est impossible de trouver la médiane des données brutes. Il est seulement possible d'en obtenir une valeur approximative. Pour ce faire, on utilise une formule d'interpolation linéaire basée sur la supposition que les données sont réparties uniformément dans chacune des classes. Cette valeur s'obtient aussi par une lecture de la courbe des pourcentages cumulés. La valeur obtenue est la médiane des données groupées.

- **La classe médiane**
  La classe médiane est celle qui accumule au moins 50 % des données. C'est dans cette classe que se situe la médiane.

- **La lecture de la courbe des pourcentages cumulés (ogive)**
  Il s'agit de déterminer la valeur (sur l'axe horizontal de l'ogive) pour laquelle le pourcentage cumulé est de 50 %. Ainsi, afin de déterminer la médiane des données groupées, on peut remplacer les calculs effectués à l'aide de la formule d'interpolation linéaire par une simple lecture de l'ogive. Cependant, comme un graphique reste imprécis, la lecture ne correspondra pas exactement à la valeur calculée à l'aide de la formule de l'interpolation linéaire.

- **La formule d'interpolation linéaire**
  La formule d'interpolation linéaire détermine avec précision la valeur lue sur l'ogive. La médiane des données groupées est obtenue à l'aide de la formule suivante :

$$Md = B_i + \frac{(50 - F)}{\%_{Md}} \cdot a.$$

a) La classe médiane est d'abord repérée ;

b) B_i est la borne inférieure de la classe médiane ;

c) a est la largeur de la classe médiane ;

d) F est le pourcentage de données cumulées dans les classes précédant la classe médiane ;

e) $\%_{Md}$ est le pourcentage des données dans la classe médiane.

EXEMPLE 4.27 Reprenons l'exemple 4.24 portant sur le revenu des hommes canadiens.

TABLEAU 4.14

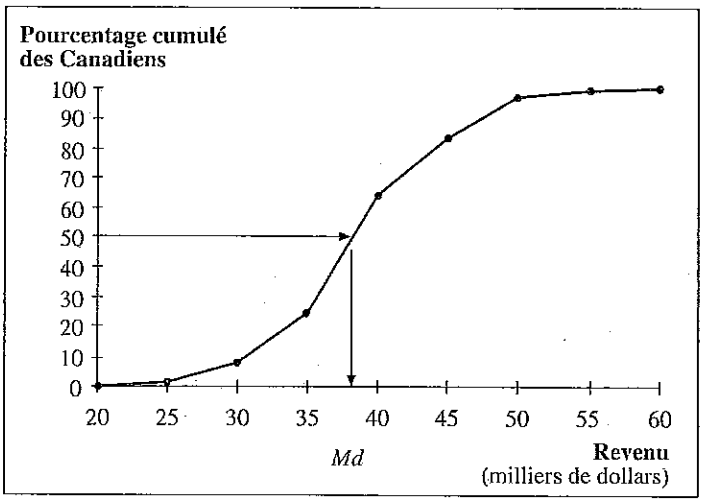
Revenu (milliers de dollars)	Point milieu (milliers de dollars)	Nombre de Canadiens	Pourcentage des Canadiens	Pourcentage cumulé des Canadiens
De 20 à moins de 25	22,5	2	1,25	1,25
De 25 à moins de 30	27,5	10	6,25	7,50
De 30 à moins de 35	32,5	28	17,50	25,00
De 35 à moins de 40	37,5	63	39,38	64,38
De 40 à moins de 45	42,5	30	18,75	83,13
De 45 à moins de 50	47,5	23	14,38	97,51
De 50 à moins de 55	52,5	3	1,88	99,39
De 55 à moins de 60	57,5	1	0,63	100,00
Total		160	100,00	

Si l'on examine le tableau 4.14, on constate que la classe médiane est « De 35 à moins de 40 milliers de dollars ». La valeur de la médiane se situe entre ces deux bornes.

a) La détermination de la médiane à l'aide de l'ogive

Il s'agit de trouver à quel endroit se situe le point qui cumule 50 % des données. La figure 4.12 illustre cette démarche.

FIGURE 4.12
Courbe des pourcentages cumulés des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps en 1990, en fonction de leur revenu



On peut donc lire sur ce graphique que la médiane est approximativement égale à 38 000 $.

b) La détermination de la médiane à l'aide de la formule

Il s'agit de trouver à quel endroit se situe le point qui accumule 50 % des données. La formule pour calculer la médiane est :

$$Md = B_i + \frac{(50 - F)}{\%_{Md}} \cdot a.$$

a) La classe médiane est « De 35 à moins de 40 milliers de dollars » ;

b) B_i (la borne inférieure de la classe médiane) est de 35 milliers de dollars ;

c) a (la largeur de la classe médiane) égale 5 milliers de dollars ;

d) F (le pourcentage de données cumulées dans les classes précédant la classe médiane) égale 25,00 % ;

e) $\%_{Md}$ (le pourcentage de données dans la classe médiane) vaut 39,38 %.

La médiane des données groupées est donc :

$$Md = B_i + \frac{(50 - F)}{\%_{Md}} \cdot a$$

$$= 35 + \frac{(50 - 25{,}00)}{39{,}38} \cdot 5 = 35 + 0{,}6348 \cdot 5$$

$$= 35 + 3{,}174$$

$$= 38{,}174 \text{ milliers de dollars (38 174 \$)}.$$

Cela signifie qu'**environ 50 % des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps, interrogés en 1990, ont un revenu d'au plus 38 174 $ (ou 38 000 $** si l'ogive a été utilisée).

Ici on emploie l'expression « environ » au lieu de l'expression « au moins », qui est utilisée dans le cas des variables quantitatives discrètes, car la médiane des données groupées est calculée en supposant que les données sont réparties uniformément à l'intérieur de chacune des classes. Notons que dans cet échantillon il y a 76 données sur 160 dont la valeur est d'au plus 38 174 $, soit 47,5 % des données. C'est pour cette raison qu'on ne peut garantir « au moins 50 % » mais seulement « environ 50 % ».

C. La moyenne ($\bar{x}$)

Comme dans le cas de la variable quantitative discrète, la moyenne correspond à la valeur où se situe le point d'appui qui tiendrait le système en équilibre. Cependant, dans le cas des données groupées en classes, on ne pourra trouver qu'une valeur approximative pour la moyenne des données brutes de l'échantillon ($\bar{x}$) ou de la population μ. Puisqu'on suppose toujours que les données sont réparties uniformément à l'intérieur de chacune des classes, alors le point milieu de chacune des classes correspond à la moyenne des données de sa classe. Ce point milieu sera utilisé pour représenter les données de sa classe, c'est-à-dire que toutes les données d'une classe seront considérées comme égales au point milieu de la classe. La moyenne se calcule en utilisant le point milieu et la fréquence de chacune des classes. La moyenne obtenue est approximative, mais sa valeur n'est pas très loin de la valeur qu'on pourrait obtenir en utilisant la série de données brutes.

EXEMPLE 4.28

Reprenons l'exemple 4.24 portant sur le revenu des hommes canadiens. La moyenne s'obtient toujours en divisant la somme de toutes les données par le nombre de données dans l'échantillon ou la population. Le tableau 4.16 montre comment procéder pour obtenir la somme approximative de toutes les données. Cette façon de faire est aussi celle qui est utilisée par la calculatrice en mode statistique.

Dans la dernière colonne, les valeurs représentent les sommes approximatives des données de chacune des classes :

TABLEAU 4.16
Détail du calcul pour obtenir la moyenne

Revenu (milliers de dollars)	Point milieu (milliers de dollars)	Nombre de Canadiens	Milieu • Nombre
De 20 à moins de 25	22,5	2	45,0
De 25 à moins de 30	27,5	10	275,0
De 30 à moins de 35	32,5	28	910,0
De 35 à moins de 40	37,5	63	2 362,5
De 40 à moins de 45	42,5	30	1 275,0
De 45 à moins de 50	47,5	23	1 092,5
De 50 à moins de 55	52,5	3	157,5
De 55 à moins de 60	57,5	1	57,5
Total		160	$\sum x = 6\,175{,}0$

- Pour les hommes canadiens interrogés dont le revenu est de 20 à moins de 25 milliers de dollars :
  Le point milieu est 22,5 milliers de dollars ;
  Il y a 2 données dans cette classe ;
  22,5 • 2 = 45 milliers de dollars.

- Pour ceux dont le revenu est de 25 à moins de 30 milliers de dollars :
  Le point milieu est 27,5 milliers de dollars ;
  Il y a 10 données dans cette classe ;
  27,5 • 10 = 275 milliers de dollars.

- Pour ceux dont le revenu est de 30 à moins de 35 milliers de dollars :
  Le point milieu est 32,5 milliers de dollars ;
  Il y a 28 données dans cette classe ;
  32,5 • 28 = 910 milliers de dollars.

- Pour ceux dont le revenu est de 35 à moins de 40 milliers de dollars :
  Le point milieu est 37,5 milliers de dollars ;
  Il y a 63 données dans cette classe ;
  37,5 • 63 = 2 362,5 milliers de dollars.

- Pour ceux dont le revenu est de 40 à moins de 45 milliers de dollars :
  Le point milieu est 42,5 milliers de dollars ;
  Il y a 30 données dans cette classe ;
  42,5 • 30 = 1 275 milliers de dollars.

- Pour ceux dont le revenu est de 45 à moins de 50 milliers de dollars :
  Le point milieu est 47,5 milliers de dollars ;
  Il y a 23 données dans cette classe ;
  47,5 • 23 = 1 092,5 milliers de dollars.

- Pour ceux dont le revenu est de 50 à moins de 55 milliers de dollars :
  Le point milieu est 52,5 milliers de dollars ;
  Il y a 3 données dans cette classe ;
  52,5 • 3 = 157,5 milliers de dollars.

- Pour ceux dont le revenu est de 55 à moins de 60 milliers de dollars :
  Le point milieu est 57,5 milliers de dollars ;
  Il y a 1 donnée dans cette classe ;
  57,5 • 1 = 57,5 milliers de dollars.

Le revenu moyen des 160 hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps, interrogés en 1990, est donc :

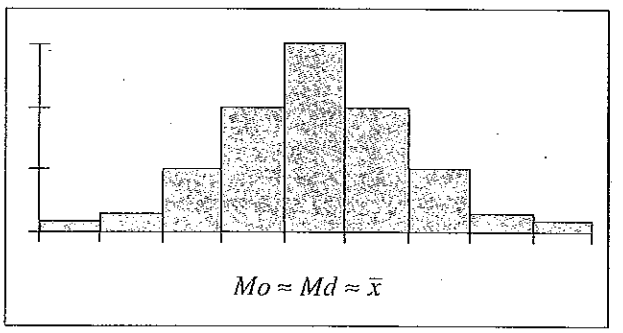
$$\bar{x} = \frac{\text{Revenu total}}{\text{Nombre de Canadiens}}$$
$$= \frac{45 + 275 + 910 + \ldots + 57,5}{160} = \frac{6\,175,0}{160}$$
$$= 38,594 \text{ milliers de dollars } (38\,594\,\$).$$

Les 160 hommes canadiens âgés de 15 et plus ayant travaillé à plein temps en 1990 ont donc un revenu moyen d'environ 38 594 $. Si l'on répartissait tous les revenus à parts égales entre les 160 hommes, chacun aurait un revenu d'environ 38 594 $.

D. Le choix de la mesure de tendance centrale

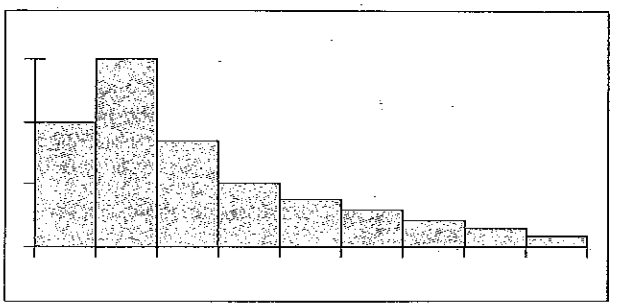
Comme dans le cas des variables quantitatives discrètes, on peut s'intéresser à la forme de représentation graphique d'une distribution de données d'une variable quantitative continue. Les propriétés de symétrie, d'asymétrie à gauche et d'asymétrie à droite sont définies de façon identique dans les deux cas.

FIGURE 4.13
Distribution symétrique



Dans ce cas, la moyenne est la mesure de tendance centrale la plus appropriée.

FIGURE 4.14
Distribution asymétrique à droite

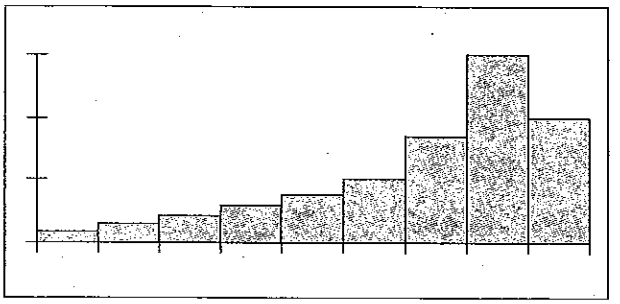


En général, la relation entre les trois mesures de tendance centrale est la suivante :

$Mo < Md < \bar{x}$.

Dans ce cas la médiane, qui n'est pas influencée par ces valeurs extrêmes, est la mesure de tendance centrale la plus appropriée.

FIGURE 4.15
Distribution asymétrique à gauche



En général, la relation entre les trois mesures de tendance centrale est la suivante :

$\bar{x} < Md < Mo.$

Dans ce cas la médiane, qui n'est pas influencée par ces valeurs extrêmes, est la mesure de tendance centrale la plus appropriée.

Comme pour les variables quantitatives discrètes, afin de déterminer si une distribution est symétrique ou asymétrique, il faut baser sa décision sur la représentation graphique et sur la relation entre les trois mesures de tendance centrale. Parfois, ce sont les trois mesures de tendance centrale qui influeront sur le choix et, en d'autres occasions, ce sera la représentation graphique.

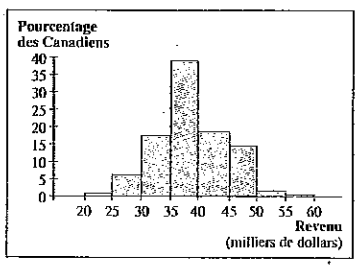
EXEMPLE 4.29

Reprenons l'exemple 4.24 portant sur le revenu des hommes canadiens. On a les données suivantes :

- Le revenu modal est de 37 636 $ (le mode brut est de 37 500 $) ;
- Le revenu médian est de 38 174 $;
- Le revenu moyen est de 38 594 $.

Les trois mesures de tendance centrale sont rapprochées. De plus, le graphique montre une symétrie. La moyenne est le bon choix comme mesure de tendance centrale.

FIGURE 4.8



4.2.4 Les mesures de dispersion

A. L'écart type (s) pour les données groupées par classes

Toujours en supposant que les données sont réparties uniformément à l'intérieur de chacune des classes, l'écart type se calcule en utilisant le point milieu de chacune

des classes. Là aussi, on a une valeur approximative de la valeur de l'écart type des données brutes de l'échantillon s ou de la population σ, mais elle ne sera pas loin de la valeur qu'on obtiendrait avec la série de données brutes. L'écart type tient compte des écarts entre la valeur de chacune des données et la valeur de la moyenne mais, dans ce cas-ci, ce sont les écarts entre les points milieux des classes et la valeur de la moyenne qui sont employés. Pour le calcul de l'écart type, on se sert de la même formule qui est utilisée par la calculatrice :

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}.$$

EXEMPLE 4.30

Reprenons l'exemple 4.24 sur le revenu des hommes canadiens et calculons l'écart type de la distribution des données groupées.

Le tableau 4.17 complète le tableau 4.16 présenté pour la calcul de la moyenne. Les deux dernières colonnes servent à calculer la somme de toutes les données et la somme de tous les carrés des données, et ce classe par classe.

TABLEAU 4.17
Détail du calcul pour obtenir l'écart type

Revenu (milliers de dollars)	Point milieu (milliers de dollars)	Nombre de Canadiens	Milieu • Nombre	Milieu^2 • Nombre
De 20 à moins de 25	22,5	2	45,0	1 012,50
De 25 à moins de 30	27,5	10	275,0	7 562,50
De 30 à moins de 35	32,5	28	910,0	29 575,00
De 35 à moins de 40	37,5	63	2 362,5	88 593,75
De 40 à moins de 45	42,5	30	1 275,0	54 187,50
De 45 à moins de 50	47,5	23	1 092,5	51 893,75
De 50 à moins de 55	52,5	3	157,5	8 268,75
De 55 à moins de 60	57,5	1	57,5	3 306,25
		160	$\sum x = 6\,175{,}0$	$\sum x^2 = 244\,400$

Pour trouver les deux sommes, on procède de façon similaire à la manière utilisée pour déterminer la moyenne. $\sum x$ représente la somme de toutes les données. On a déjà calculé cette somme à la sous-section 4.2.3 C sur la moyenne :

$\sum x = 6\,175$.

Si l'on recherche $\sum x^2$, le processus est le même. On procède classe par classe :

- Pour les hommes canadiens dont le revenu est de 20 à moins de 25 milliers de dollars :
  Le point milieu est de 22,5 milliers de dollars ;
  Il y a 2 données dans cette classe.
  Puisque le point milieu est :
  $x = 22{,}5$;
  $x^2 = 22{,}5^2 = 506{,}25$.
  Alors, pour les 2 données :
  $x^2 \cdot 2 = 506{,}25 \cdot 2 = 1\,012{,}5$.
- Pour ceux dont le revenu est de 25 à moins de 30 milliers de dollars :
  Le point milieu est de 27,5 milliers de dollars ;
  Il y a 10 données dans cette classe.
  Puisque le point milieu est :

$x = 27{,}5$;
$x^2 = 27{,}5^2 = 756{,}25$.
Alors, pour les 10 données :
$x^2 \cdot 10 = 756{,}25 \cdot 10 = 7\,562{,}5$.

- Pour ceux dont le revenu est de 30 à moins de 35 milliers de dollars :
Le point milieu est de 32,5 milliers de dollars ;
Il y a 28 données dans cette classe.
Puisque le point milieu est :
$x = 32{,}5$;
$x^2 = 32{,}5^2 = 1\,056{,}25$.
Alors, pour les 28 données :
$x^2 \cdot 28 = 1\,056{,}25 \cdot 28 = 29\,575{,}0$.

- Pour ceux dont le revenu est de 35 à moins de 40 milliers de dollars :
Le point milieu est de 37,5 milliers de dollars ;
Il y a 63 données dans cette classe.
Puisque le point milieu est :
$x = 37{,}5$;
$x^2 = 37{,}5^2 = 1\,406{,}25$.
Alors, pour les 63 données :
$x^2 \cdot 63 = 1\,406{,}25 \cdot 63 = 88\,593{,}75$.

- Pour ceux dont le revenu est de 40 à moins de 45 milliers de dollars :
Le point milieu est de 42,5 milliers de dollars ;
Il y a 30 données dans cette classe.
Puisque le point milieu est :
$x = 42{,}5$;
$x^2 = 42{,}5^2 = 1\,806{,}25$.
Alors, pour les 30 données :
$x^2 \cdot 30 = 1\,806{,}25 \cdot 30 = 54\,187{,}5$.

- Pour ceux dont le revenu est de 45 à moins de 50 milliers de dollars :
Le point milieu est de 47,5 milliers de dollars ;
Il y a 23 données dans cette classe.
Puisque le point milieu est :
$x = 47{,}5$;
$x^2 = 47{,}5^2 = 2\,256{,}25$.
Alors, pour les 23 données :
$x^2 \cdot 23 = 2\,256{,}25 \cdot 23 = 51\,893{,}75$.

- Pour ceux dont le revenu est de 50 à moins de 55 milliers de dollars :
Le point milieu est de 52,5 milliers de dollars ;
Il y a 3 données dans cette classe.
Puisque le point milieu est :
$x = 52{,}5$;
$x^2 = 52{,}5^2 = 2\,756{,}25$.
Alors, pour les 3 données :
$x^2 \cdot 3 = 2\,756{,}25 \cdot 3 = 8\,268{,}75$.

- Pour ceux dont le revenu est de 55 à moins de 60 milliers de dollars :
Le point milieu est de 57,5 milliers de dollars ;
Il y a 1 donnée dans cette classe.

Puisque le point milieu est :

$x = 57{,}5$;

$x^2 = 57{,}5^2 = 3\,306{,}25$.

Alors, pour la donnée :

$x^2 \cdot 1 = 3\,306{,}25 \cdot 1 = 3\,306{,}25$.

Pour l'ensemble des 160 hommes canadiens :

$$\begin{aligned}\sum x^2 &= 1\,012{,}5 + 7\,562{,}5 + 29\,575{,}0 + 88\,593{,}75 + 54\,187{,}5 + 51\,893{,}75 \\ &\quad + 8\,268{,}75 + 3\,306{,}25 \\ &= 244\,400{,}00 ;\end{aligned}$$

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$$

$$= \sqrt{\frac{244\,400 - \frac{(6\,175)^2}{160}}{160-1}}$$

$= 6{,}186$ milliers de dollars (6 186 $).

Note : Cette valeur est obtenue beaucoup plus rapidement avec la calculatrice.

La mesure $s = 6\,186$ $ informe sur la dispersion des 160 données autour de la moyenne $\bar{x} = 38\,594$ $ par homme canadien interrogé.

La même année, un échantillon de 125 Canadiennes âgées de 15 ans et plus ayant travaillé toute l'année a donné un revenu moyen de 25 900 $ avec un écart type de 5 646 $. La première constatation est que le revenu moyen des Canadiennes interrogées est inférieur au revenu moyen des Canadiens interrogés. L'autre constatation est que les Canadiennes interrogées ont un revenu moins dispersé autour de leur revenu moyen que celui des Canadiens interrogés autour de leur revenu moyen ; l'écart type du revenu des Canadiennes interrogées, soit 5 646 $, est inférieur à celui du revenu des Canadiens interrogés, soit 6 186 $.

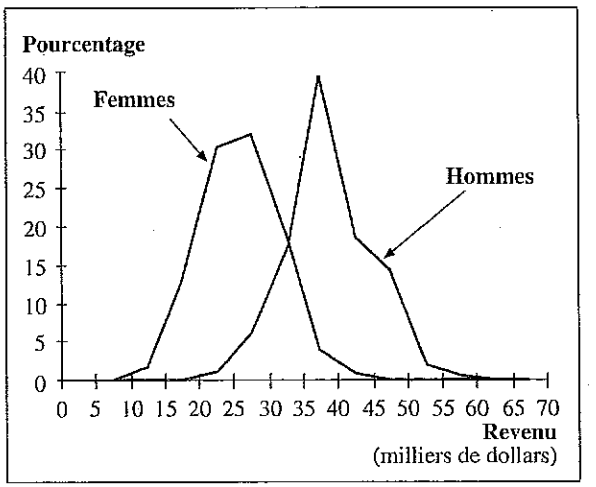
Le tableau 4.18 compare la distribution du revenu chez les hommes et les femmes des deux échantillons.

TABLEAU 4.18
Répartition des Canadiens et des Canadiennes âgés de 15 ans et plus ayant travaillé à plein temps en 1990, en fonction de leur revenu

Revenu (milliers de dollars)	Point milieu (milliers de dollars)	Nombre de Canadiens	Nombre de Canadiennes
De 10 à moins de 15	12,5	0	2
De 15 à moins de 20	17,5	0	16
De 20 à moins de 25	22,5	2	38
De 25 à moins de 30	27,5	10	40
De 30 à moins de 35	32,5	28	23
De 35 à moins de 40	37,5	63	5
De 40 à moins de 45	42,5	30	1
De 45 à moins de 50	47,5	23	0
De 50 à moins de 55	52,5	3	0
De 55 à moins de 60	57,5	1	0
Total		160	125

La figure 4.16 représente les deux polygones superposés de pourcentage des hommes et des femmes, ce qui permet de comparer les deux distributions.

FIGURE 4.16
Polygones superposés du pourcentage des hommes et des femmes



Sur ce graphique, on peut voir que le revenu des hommes a une distribution ayant des valeurs supérieures au revenu des femmes.

B. Le coefficient de variation (*CV*)

Le coefficient de variation se calcule et s'interprète de la même façon, qu'il s'agisse de données pour une variable quantitative discrète ou de données pour une variable quantitative continue. Le critère d'homogénéité ou de précision des données est toujours le même, c'est-à-dire que les données sont considérées comme homogènes si le coefficient de variation ne dépasse pas 15 %. Toutefois, ce critère n'est pas absolu et peut varier. En effet, Statistique Canada et les laboratoires utilisent d'autres pourcentages pour déterminer l'homogénéité de leurs données.

EXEMPLE 4.31

Reprenons l'exemple 4.30 sur le revenu des hommes canadiens et des femmes canadiennes.

Variables	Le revenu des hommes canadiens et le revenu des femmes canadiennes
Échelle de mesure	Échelle de rapport
Coefficient de variation pour le revenu des hommes interrogés	$CV = \frac{6\,186}{38\,594} \cdot 100 = 16,03\,\%$
Coefficient de variation pour le revenu des femmes interrogées	$CV = \frac{5\,646}{25\,900} \cdot 100 = 21,80\,\%$

Comme on l'a dit précédemment, les hommes ont un revenu moyen supérieur à celui des femmes, et la distribution du revenu des femmes est moins dispersée que

celle du revenu des hommes. Cependant, si l'on considère les valeurs des revenus moyens des hommes et des femmes, on s'aperçoit que la dispersion du revenu des hommes représente seulement 16,03 % de la valeur de leur revenu moyen, tandis que la dispersion du revenu des femmes représente 21,80 % de la valeur de leur revenu moyen. Ainsi, même si la dispersion du revenu des hommes est plus grande que celle du revenu des femmes, elle est moins importante si l'on examine leurs revenus moyens. La distribution du revenu des hommes est donc plus homogène que celle du revenu des femmes.

4.2.5 Les mesures de position

Les mesures de position pour des données quantitatives continues sont les mêmes que pour des données quantitatives discrètes. Ces mesures sont les quantiles et la cote *Z*.

A. Les quantiles

Dans le cas de données continues groupées en classes, les quantiles sont des valeurs qui subdivisent la surface totale de l'histogramme en tranches correspondant aux pourcentages désirés. Les valeurs obtenues seront approximatives, car on suppose encore que les données sont réparties uniformément à l'intérieur des classes. Pour obtenir les centiles, il faut diviser la surface en tranches de 1 %, les déciles en tranches de 10 % et les quartiles en tranches de 25 %.

Les notations utilisées sont C_1, C_2, C_3, ..., C_{99} pour les centiles, D_1, D_2, ..., D_9 pour les déciles et Q_1, Q_2 et Q_3 pour les quartiles, en gardant en mémoire que C_{50}, D_5 et Q_2 correspondent à la médiane. Les quantiles se trouvent de la même façon que la médiane, c'est-à-dire à l'aide soit de la courbe des pourcentages cumulés ou de la formule d'interpolation linéaire adaptée au pourcentage recherché. Comme dans le cas de la médiane, le centile C_p d'une distribution de données groupées est, sur l'axe horizontal, la valeur qui correspond à un pourcentage cumulé de p %.

Puisque les déciles et les quartiles peuvent s'exprimer en centiles, la formule d'interpolation linéaire ne sera présentée que pour les centiles. Alors C_p, le p-ième centile de données groupées, est obtenu par la formule suivante :

$$C_p = B_i + \frac{(p - F)}{\%_{C_p}} \cdot a.$$

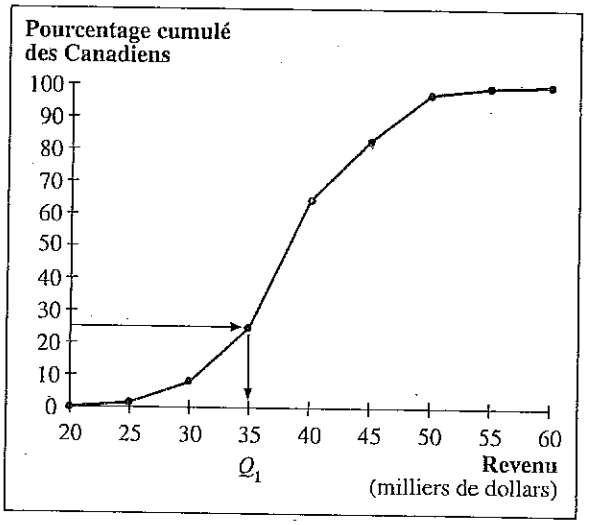
a) La classe du p-ième centile est d'abord repérée ;
b) B_i est la borne inférieure de la classe du p-ième centile ;
c) a est la largeur de la classe du p-ième centile ;
d) F est le pourcentage des données cumulées dans les classes précédant la classe du p-ième centile ;
e) $\%_{C_p}$ est le pourcentage de données dans la classe du p-ième centile.

EXEMPLE 4.32

Reprenons l'exemple 4.24 sur le revenu des hommes canadiens. Si l'on se base sur le tableau 4.14, on trouve la valeur du premier quartile ($Q_1 = C_{25}$) à l'aide de l'ogive :

Q_1 ou C_{25} est, sur l'ogive de cette distribution (figure 4.17), la valeur sur l'axe horizontal qui correspond à un pourcentage cumulé de 25 %.

FIGURE 4.17
Courbe des pourcentages cumulés des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps en 1990, en fonction de leur revenu



On s'aperçoit que la valeur de Q_1 correspond à un revenu d'environ 35 000 $. On dira donc qu'**environ 25 % des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps, interrogés en 1990, ont un revenu d'au plus 35 000 $.**

À l'aide de la formule d'interpolation linéaire, on recherche la valeur du premier quartile ($Q_1 = C_{25}$) :

TABLEAU 4.14

Revenu (milliers de dollars)	Point milieu (milliers de dollars)	Nombre de Canadiens	Pourcentage des Canadiens	Pourcentage cumulé des Canadiens
De 20 à moins de 25	22,5	2	1,25	1,25
De 25 à moins de 30	27,5	10	6,25	7,50
De 30 à moins de 35	32,5	28	17,50	25,00
De 35 à moins de 40	37,5	63	39,38	64,38
De 40 à moins de 45	42,5	30	18,75	83,13
De 45 à moins de 50	47,5	23	14,38	97,51
De 50 à moins de 55	52,5	3	1,88	99,39
De 55 à moins de 60	57,5	1	0,63	100,00
Total		160	100,00	

a) La classe de C_{25} est « De 30 à moins de 35 milliers de dollars » ;
b) 30 milliers de dollars est la borne inférieure de la classe de C_{25} ;
c) 5 milliers de dollars est la largeur de la classe de C_{25} ;
d) 7,50 % est le pourcentage cumulé de données précédant la classe de C_{25} ;
e) 17,50 % est le pourcentage de données dans la classe de C_{25}.

Donc, la valeur du premier quartile est de :

$$C_p = B_i + \frac{(p - F)}{\%_{C_p}} \cdot a \ ;$$

$$C_{25} = 30 + \left(\frac{25 - 7{,}50}{17{,}50}\right) \cdot 5 = 30 + (1 \cdot 5) = 30 + 5$$

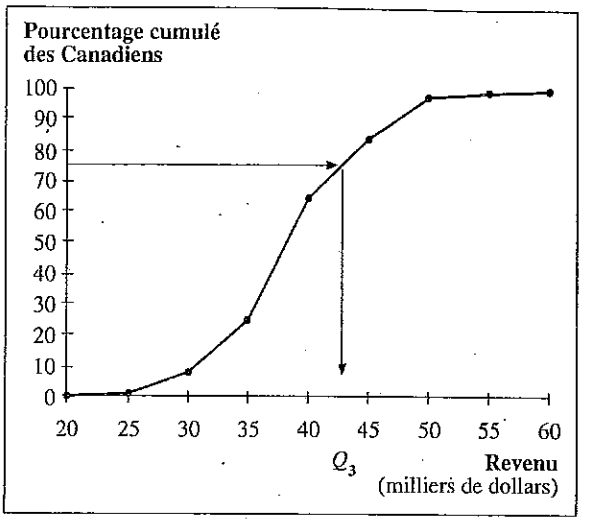
$$= 35 \text{ milliers de dollars } (35\,000\,\$).$$

Cela signifie qu'**environ 25 % des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps, interrogés en 1990, ont un revenu d'au plus 35 000 $.**

À l'aide de l'ogive, on recherche la valeur du troisième quartile ($Q_3 = C_{75}$) :

Sur l'ogive de cette distribution (figure 4.18), Q_3 ou C_{75} est la valeur sur l'axe horizontal qui correspond à un pourcentage cumulé de 75 %.

FIGURE 4.18
Courbe des pourcentages cumulés des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps en 1990, en fonction de leur revenu



On s'aperçoit que la valeur de Q_3 correspond à un revenu d'environ 43 000 $. On dira donc qu'**environ 75 % des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps, interrogés en 1990, ont un revenu d'au plus 43 000 $.**

À l'aide de la formule d'interpolation linéaire, on recherche la valeur du troisième quartile ($Q_3 = C_{75}$) :

a) La classe de C_{75} est « De 40 à moins de 45 milliers de dollars » ;
b) 40 milliers de dollars est la borne inférieure de la classe de C_{75} ;
c) 5 milliers de dollars est la largeur de la classe de C_{75} ;
d) 64,38 % est le pourcentage cumulé des données précédant la classe de C_{75} ;
e) 18,75 % est le pourcentage de données dans la classe de C_{75}.

Donc, la valeur du troisième quartile est de :

$$C_p = B_i + \frac{(p - F)}{\%_{C_p}} \cdot a \ ;$$

$$C_{75} = 40 + \left(\frac{75 - 64{,}38}{18{,}75}\right) \cdot 5$$

$$= 40 + (0{,}5664 \cdot 5) = 40 + 2{,}832$$

$$= 42{,}832 \text{ milliers de dollars } (42\,832\,\$).$$

Cela signifie qu'**environ 75 % des hommes canadiens âgés de 15 ans et plus ayant travaillé à plein temps, interrogés en 1990, ont un revenu d'au plus 42 832 $.**

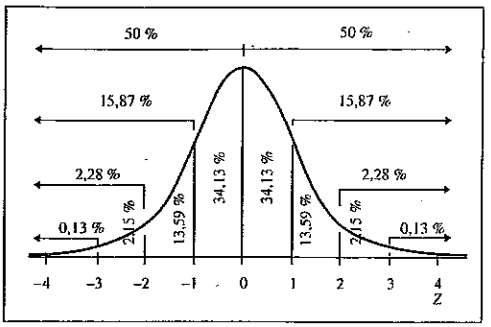
B. La cote *Z*

Quel que soit le type de variable quantitative (discrète ou continue), la cote Z se calcule toujours de la façon suivante :

$$Z = \frac{x - \bar{x}}{s} \text{ (pour un échantillon);}$$

$$Z = \frac{x - \mu}{\sigma} \text{ (pour une population);}$$

FIGURE 4.6



où x est la valeur d'une donnée, $\bar{x}$ ou μ est la moyenne de l'échantillon ou de la population et s ou σ est l'écart type de l'échantillon ou de la population. La cote Z est exprimée en longueurs d'écart type. Rappelons que l'interprétation de la cote Z est faite en se basant sur le modèle de la distribution normale (figure 4.6), modèle ayant une forme symétrique.

EXEMPLE 4.33

Reprenons l'exemple 4.24 sur le revenu des hommes canadiens. Le revenu moyen des hommes canadiens est de 38 594 $, et l'écart type est de 6 186 $. Ainsi, l'individu qui a un revenu de 45 000 $ a une cote Z de :

$$Z = \frac{45\,000 - 38\,594}{6\,186} = 1,04.$$ (Les cotes Z sont exprimées avec deux décimales.)

Le résultat précédent signifie que l'individu qui a un revenu de 45 000 $ se situe environ à une longueur d'écart type au-dessus du revenu moyen. La distribution du revenu des hommes étant symétrique, on interprète la valeur de la cote Z d'après le modèle de la distribution normale. Il y a environ 15,87 % des hommes ayant un revenu supérieur à 45 000 $ dans cet échantillon.

L'individu qui a un revenu de 25 000 $ a une cote Z de :

$$Z = \frac{25\,000 - 38\,594}{6\,186} = -2,20.$$

Ce résultat signifie que l'individu qui a un revenu de 25 000 $ se situe environ à deux longueurs d'écart type au-dessous du revenu moyen. Ainsi, d'après le modèle de la distribution normale, il y a un peu moins de 2,28 % des hommes ayant un revenu inférieur à 25 000 $ dans cet échantillon.

4.2.6 Quelques cas particuliers

Dans cette sous-section, le regroupement des données en classes se fera de manière différente. On verra en détail une application concernant un cas particulier d'asymétrie. Ensuite, on étudiera un cas de classes ouvertes et de classes de largeurs inégales et, pour terminer, une application avec des variables quantitatives discrètes.

A. Un cas particulier d'asymétrie

EXEMPLE 4.34

Lors de son « Enquête sociale générale » de 1990, Statistique Canada a recueilli des informations au sujet du revenu des couples canadiens sans enfant en 1990.

Variable	Le revenu total du couple
Type de variable	La variable est quantitative continue.
Population	Tous les couples canadiens sans enfant en 1990
Unité statistique	Un couple canadien sans enfant en 1990

Les données suivantes représentent le revenu total de 183 couples sans enfant[11], choisis au hasard, en 1990 :

3 600,88 $	16 598,10 $	35 562,91 $	79 154,33 $	79 748,53 $	49 020,66 $
5 789,06	21 944,49	48 750,57	81 579,64	74 149,91	38 750,88
9 731,44	26 649,98	52 345,04	76 622,82	62 721,03	49 565,42
12 092,65	15 576,80	49 785,15	83 858,46	71 379,44	57 726,68
14 090,85	25 125,58	54 549,70	72 235,48	79 753,11	47 032,99
12 305,52	18 142,19	32 742,09	79 266,03	69 931,03	33 526,72
11 519,97	26 673,33	58 170,72	78 616,90	80 306,10	52 328,56
945,77	21 260,57	47 820,37	68 337,05	73 254,49	67 644,89
15,11	29 480,42	56 363,41	69 894,41	70 656,15	62 012,39
580,46	22 048,86	38 920,26	74 764,24	70 018,92	61 043,73
12 990,81	15 510,88	37 325,36	74 236,88	85 998,11	83 083,90
1 921,29	16 516,16	40 013,43	69 250,77	83 216,65	85 957,82
2 803,43	20 353,71	44 087,65	73 631,70	42 245,55	63 397,63
41 275,06	27 673,12	40 205,69	63 832,51	31 508,84	72 849,82
34 481,64	29 192,48	51 315,96	84 825,28	47 623,52	44 671,77
45 251,32	23 903,78	58 215,58	81 444,14	31 855,83	35 342,27
59 955,14	22 126,22	56 920,07	78 071,23	33 117,47	36 483,05
37 042,45	17 326,88	48 669,09	65 722,22	56 086,92	36 875,82
43 682,06	24 439,37	47 192,30	69 535,51	34 489,88	37 912,23
33 590,81	17 642,29	36 890,47	84 470,96	38 354,44	43 522,75
54 851,83	28 646,81	44 735,86	64 971,47	59 263,89	57 487,72
42 187,87	20 021,36	42 136,60	88 200,93	40 672,63	54 479,20
38 875,39	23 321,02	49 656,97	83 038,12	30 420,24	54 797,81
35 937,38	16 560,56	47 553,03	73 932,00	57 735,83	33 057,04
50 230,11	21 951,81	57 830,13	82 507,10	36 856,59	47 891,78
35 609,61	15 257,27	44 963,84	85 262,92	54 893,03	33 244,73
82 044,74	26 545,15	56 634,42	84 874,72	80 159,61	40 997,65
88 086,49	24 088,26	40 008,85	73 753,47	71 589,10	52 103,34
60 477,92	28 455,46	49 634,08	78 851,28	73 470,56	33 946,04
81 350,75	21 131,93	54 516,74	75 325,48	60 282,91	33 917,66
55 680,41	21 177,71	38 608,97			

Taille de l'échantillon $n = 183$

11. Échantillon fictif simulé à partir des informations de *La famille et les amis, Enquête sociale générale, Série analytique*, Ottawa, Statistique Canada, Catalogue 11-612F, n° 9, ISBN 0-660-94435-9, 1994, p. 37-39.

a) Le tableau

Les données sont regroupées selon la démarche suivante :

– **Première étape : Titrer le tableau**

La formulation sera celle-ci : Répartition des couples canadiens, sans enfant, en fonction de leur revenu total en 1990.

– **Deuxième étape : Construire les classes**

Pour construire les classes, il convient de déterminer :

- le nombre de classes. La taille de l'échantillon est de 183. Le nombre de classes suggéré, d'après le tableau de Sturges pour les échantillons dont les tailles vont de 181 à 361, est de 9. On obtient la même valeur en utilisant la formule ci-après. Le nombre de classes est de :

  $1 + 3{,}322 \cdot \log 183 = 8{,}52 \cong 9$ classes.

- l'étendue des données. L'étendue des données est :

  88 200,93 $ – 15,11 $ = 88 185,82 $.

- la largeur des classes. La largeur des classes se calcule ainsi :

  $$\frac{88\,185{,}82}{9} = 9\,798{,}42\,\$;$$

  une largeur de 10 000 $ serait un bon choix.

- les classes. Comme la plus petite donnée a comme valeur 15,11 $, on commence la première classe à 0 $. Les classes sont : « De 0 $ à moins de 10 000 $ », « De 10 000 $ à moins de 20 000 $ », ..., « De 80 000 $ à moins de 90 000 $ ».

  Ces classes sont exhaustives et exclusives parce que chaque donnée entre dans une et une seule classe.

– **Troisième étape : Trouver le point milieu des classes**

Le point milieu de chacune des classes se place dans la deuxième colonne.

– **Quatrième étape : Dénombrer les unités statistiques**

Il s'agit de faire le dépouillement des données pour obtenir le nombre d'unités statistiques dans chaque classe. On peut voir qu'il y a 8 couples dont le revenu total est de 0 $ à moins de 10 000 $...

– **Cinquième étape : Établir les pourcentages**

Les pourcentages sont placés dans la quatrième colonne du tableau. Les 8 couples dont le revenu est de 0 $ à moins de 10 000 $ représentent 4,37 % des couples ; les 14 couples dont le revenu est de 10 000 $ à moins de 20 000 $ représentent 7,65 % des couples...

– **Sixième étape : Établir les pourcentages cumulés**

Les pourcentages cumulés sont placés dans la cinquième colonne du tableau. Il y a 12,02 % des couples dont le revenu total est inférieur à 20 000 $, 24,04 % des couples dont le revenu total est inférieur à 30 000 $...

Le tableau 4.19 résulte de la démarche précédente.

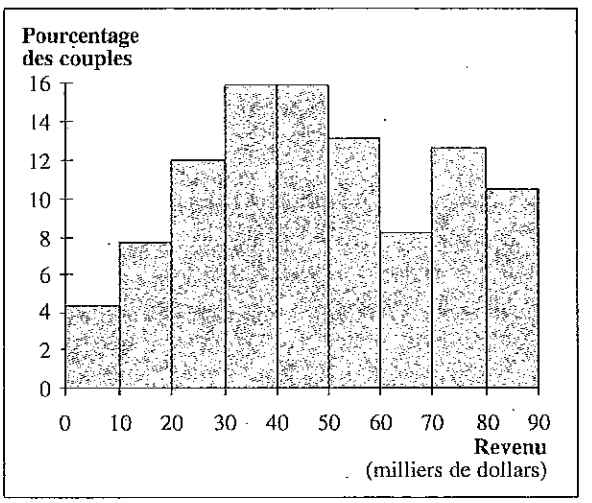
TABLEAU 4.19
Répartition des couples canadiens, sans enfant, en fonction de leur revenu total en 1990

Revenu (milliers de dollars)	Point milieu (milliers de dollars)	Nombre de couples	Pourcentage des couples	Pourcentage cumulé des couples
[0 ; 10[	5	8	4,37	4,37
[10 ; 20[	15	14	7,65	12,02
[20 ; 30[	25	22	12,02	24,04
[30 ; 40[	35	29	15,85	39,89
[40 ; 50[	45	29	15,85	55,74
[50 ; 60[	55	24	13,11	68,85
[60 ; 70[	65	15	8,20	77,05
[70 ; 80[	75	23	12,57	89,62
[80 ; 90[	85	19	10,38	100,00
Total		183	100,00	

b) Les graphiques

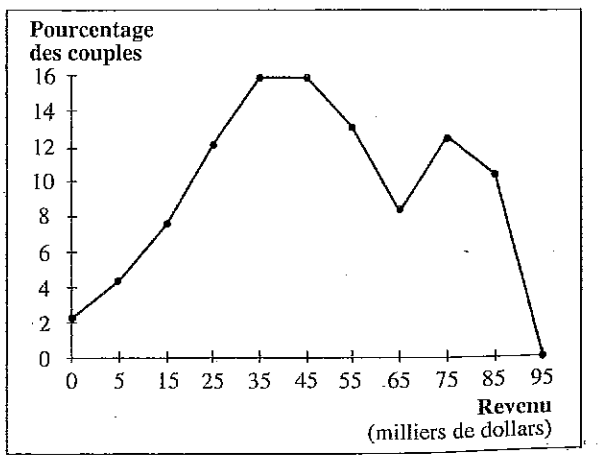
– La figure 4,19 présente l'**histogramme** construit à partir des données du tableau 4.19.

FIGURE 4.19
Répartition des couples canadiens, sans enfant, en fonction de leur revenu total en 1990



– La figure 4.20 illustre le **polygone** construit à partir des données du tableau 4.19.

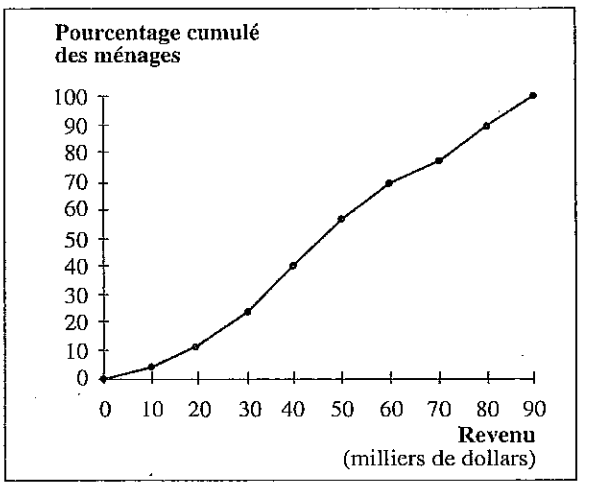
FIGURE 4.20
Répartition des couples canadiens, sans enfant, en fonction de leur revenu total en 1990



On a tronqué le polygone vis-à-vis de 0 $ pour ne pas avoir de revenu négatif.

– La figure 4.21 illustre l'**ogive** construite à partir des données du tableau 4.19.

FIGURE 4.21
Courbe des pourcentages cumulés des couples canadiens, sans enfant, en fonction de leur revenu total en 1990



c) Les mesures de tendance centrale

– **Le mode**

Il y a deux classes modales : « De 30 000 $ à moins de 40 000 $ » et « De 40 000 $ à moins de 50 000 $ ». La classe de revenu total qui revient le plus souvent chez les couples canadiens, sans enfant, interrogés, est « De 30 000 $ à moins de 50 000 $ ».

Le mode brut des données groupées est de 40 000 $, le point milieu des deux classes modales réunies. Le revenu total, chez les couples canadiens, sans enfant, interrogés, autour duquel il y a une plus forte concentration de données est d'environ 40 000 $.

Le mode des données groupées calculé à l'aide de la formule est de 40 000 $, qu'on prenne n'importe laquelle des deux classes modales pour effectuer le calcul. Le revenu total, chez les couples canadiens, sans enfant, interrogés, autour duquel il y a une plus forte concentration de données est d'environ 40 000 $.

– **La médiane**

La médiane des données groupées est de 46 379 $ (résultat obtenu à l'aide de la courbe des pourcentages cumulés ou de la formule). Environ 50 % des couples canadiens, sans enfant, interrogés, ont un revenu total d'au plus 46 379 $.

– **La moyenne**

La moyenne des données groupées est de 47 842 $. Le revenu total moyen des couples canadiens, sans enfant, interrogés, est d'environ 47 842 $.

– **Le choix de la mesure de tendance centrale**

Les trois mesures de tendance centrale sont dans l'ordre suivant :

Mode < Médiane < Moyenne
40 000 $ < 46 379 $ < 47 842 $

ce qui représente un signe d'asymétrie à droite. Cependant, l'histogramme (ou le polygone) n'a pas la forme traditionnelle d'asymétrie à droite. L'asymétrie est causée par une quantité importante de données qui se détachent à droite du bloc principal des données. Puisqu'une mesure de tendance centrale sert à localiser la valeur autour de laquelle les données sont dispersées, dans ce cas-ci, il est plus approprié de choisir la médiane comme mesure de tendance centrale ; car la valeur de la moyenne est plus touchée que la médiane par cette quantité importante de données à droite de la distribution.

d) Les mesures de dispersion

– **L'écart type**

L'écart type des données groupées est de 22 862 $. La dispersion du revenu total des couples canadiens, sans enfant, interrogés, donne un écart type d'environ 22 862 $.

– **Le coefficient de variation**

Le coefficient de variation des données groupées est de 47,79 %. La distribution du revenu total des couples canadiens, sans enfant, interrogés, n'est pas homogène puisque le coefficient de variation est supérieur à 15 %. Si l'on examine la valeur du revenu moyen, la dispersion du revenu est trop importante pour que l'on considère le revenu moyen comme mesure de tendance centrale des données.

e) Les mesures de position

– **Les quartiles des données groupées**

Q_1 est 30 606 $ (trouvé à l'aide de la courbe des pourcentages cumulés ou de la formule). Environ 25 % des couples canadiens, sans enfant, interrogés, ont un revenu total d'au plus 30 606 $.

Q_3 est 67 500 $ (trouvé à l'aide de la courbe des pourcentages cumulés ou de la formule). Environ 75 % des couples canadiens, sans enfant, interrogés, ont un revenu total d'au plus 67 500 $.

– **La cote Z**

La cote Z de 84 470,96 $ est 1,60 écart type. Un couple canadien, sans enfant, dont le revenu total est de 84 470,96 $, se situe à 1,60 écart type au-dessus du revenu total moyen de ces couples.

La cote Z de 17 642,29 $ est –1,32 écart type. Un couple canadien, sans enfant, dont le revenu total est de 17 642,29 $, se situe à 1,32 écart type au-dessous du revenu total moyen de ces couples.

La distribution du revenu n'étant pas symétrique, une interprétation des cotes Z à l'aide de la distribution normale n'est pas appropriée.

B. Des classes ouvertes et des classes de largeurs inégales

EXEMPLE 4.35

Les femmes d'avant le *baby-boom*

Le tableau 4.20 donne la répartition, en fonction du revenu de leur conjoint en 1990[12], de 135 femmes canadiennes âgées de 46 à moins de 56 ans, mariées ou vivant en union libre, choisies au hasard en 1991.

TABLEAU 4.20
Répartition des femmes mariées ou en union libre, âgées de 46 à moins de 56 ans, en 1991, en fonction du revenu de leur conjoint en 1990

Revenu du conjoint (milliers de dollars)	Point milieu (milliers de dollars)	Nombre de femmes	Pourcentage des femmes	Pourcentage cumulé des femmes
Moins de 10		33	24,44	24,44
De 10 à moins de 20	15	7	5,19	29,63
De 20 à moins de 30	25	25	18,52	48,15
De 30 à moins de 40	35	18	13,33	61,48
De 40 à moins de 50	45	19	14,07	75,55
De 50 à moins de 60	55	8	5,93	81,48
60 et plus		25	18,52	100,00
Total		135	100,00	

Source : Adapté du tableau 2.1 de Statistique Canada. *Op. cit.*, p. 10.

12. Échantillon fictif simulé à partir des données de *Les femmes du* baby-boom *: une génération au travail*, Ottawa, Statistique Canada, Catalogue 96-315F, 1994, p. 10.

Variable	Le revenu du conjoint
Type de variable	La variable est quantitative continue.
Population	Toutes les femmes canadiennes âgées de 46 à moins de 56 ans, mariées ou vivant en union libre en 1991
Unité statistique	Une femme canadienne âgée de 46 à moins de 56 ans, mariée ou vivant en union libre en 1991
Taille de l'échantillon	$n = 135$

a) Les classes ouvertes et les classes de largeurs inégales

Des classes de la forme « Moins de 10 » et « 60 et plus » sont des classes ouvertes, car une seule des deux bornes est précisée. La classe « 60 et plus » est utilisée parce que si l'on ajoutait d'autres classes de largeur 10, on risquerait d'avoir des classes avec peu ou pas de données. Supposons que la valeur de l'une des données soit 180. Faut-il changer la largeur des classes pour 1 seule donnée ? Par ailleurs, il ne faut pas faire de classes très larges qui regrouperaient trop de données dans ces classes et pas assez dans les autres ; l'information obtenue sur la répartition des données serait alors insatisfaisante.

Les graphiques et le calcul des différentes mesures pour les données groupées par classes nécessitent les deux bornes de classes. Il faudra donc choisir la borne manquante des classes ouvertes. Pour ce qui est de la première classe, on peut facilement concevoir que 0 $ serait un choix approprié pour la borne inférieure. Cependant, pour la dernière classe, on ignore jusqu'où vont les revenus de ces conjoints (70 000 $, 90 000 $...). On suppose ici que les valeurs individuelles des unités nous sont inconnues, ce qui est le cas pour les sondages où les choix de réponses sont des classes. L'une des façons de procéder dans une situation comme celle-ci consiste à prendre une largeur de classe qui est au moins le double de la largeur de la classe précédente. Toutefois, cette décision dépend aussi des valeurs plausibles qu'on pourrait avoir dans l'échantillon. Dans cet exemple, la borne supérieure sera fixée à 100 000 $.

Au fins de travail, les données prendraient la forme du tableau 4.21.

TABLEAU 4.21
Répartition des femmes mariées ou en union libre, âgées de 46 à moins de 56 ans, en 1991, en fonction du revenu de leur conjoint en 1990

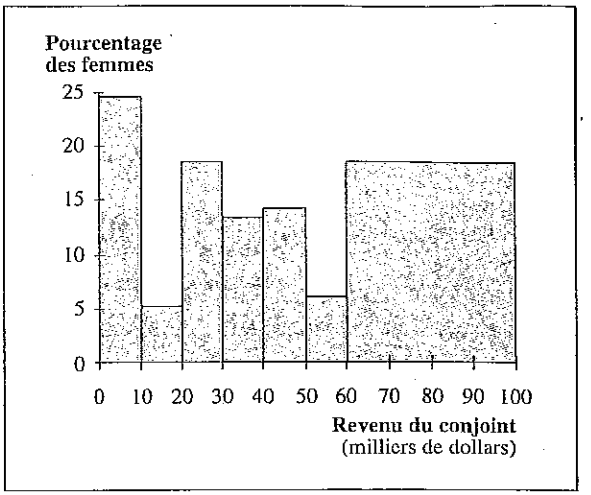
Revenu du conjoint (milliers de dollars)	Point milieu (milliers de dollars)	Nombre de femmes	Pourcentage des femmes	Pourcentage cumulé des femmes
De 0 à moins de 10	5	33	24,44	24,44
De 10 à moins de 20	15	7	5,19	29,63
De 20 à moins de 30	25	25	18,52	48,15
De 30 à moins de 40	35	18	13,33	61,48
De 40 à moins de 50	45	19	14,07	75,55
De 50 à moins de 60	55	8	5,93	81,48
De 60 à moins de 100	80	25	18,52	100,00
Total		135	100,00	

Le tableau 4.21 comprend une dernière classe qui est quatre fois plus large que les autres.

b) L'histogramme pour des données groupées dans des classes de largeurs inégales

La figure 4.22 présente l'histogramme construit de façon conventionnelle.

FIGURE 4.22
Répartition des femmes mariées ou en union libre, âgées de 46 à moins de 56 ans, en 1991, en fonction du revenu de leur conjoint en 1990



La lecture de l'histogramme indique qu'il y a :

- environ 18,5 % de ces femmes dont le conjoint a un revenu de 20 000 $ à moins de 30 000 $, ce qui est vrai ;
- environ 18,5 % de ces femmes dont le conjoint a un revenu de 60 000 $ à moins de 70 000 $, **ce qui est faux** ;
- environ 18,5 % de ces femmes dont le conjoint a un revenu de 70 000 $ à moins de 80 000 $, **ce qui est faux** ;
- environ 18,5 % de ces femmes dont le conjoint a un revenu de 80 000 $ à moins de 90 000 $, **ce qui est faux** ;
- environ 18,5 % de ces femmes dont le conjoint a un revenu de 90 000 $ à moins de 100 000 $, **ce qui est faux**.

Les classes « De 20 à moins de 30 » et « De 60 à moins de 100 » contiennent toutes les deux 25 % des données. Cependant, l'une des deux classes est quatre fois plus large que l'autre ; c'est ce qui fait dire que la densité (la concentration) des données est plus grande dans la classe « De 20 à moins de 30 » que dans celle « De 60 à moins de 100 ».

Dans un histogramme, la hauteur du rectangle doit être proportionnelle à la densité de données dans la classe. Dans le cas des classes de largeurs égales, cette propriété est respectée, ce qui n'est pas le cas lorsqu'au moins une classe a une largeur différente de celle des autres. Pour respecter cette propriété, il faut que la surface des rectangles soit égale au pourcentage de données dans la classe, de telle sorte que la somme de toutes les surfaces donne 100 %. Pour y arriver, il faudrait construire des rectangles dont la hauteur est égale au pourcentage divisé par la largeur de la classe :

$$\text{Hauteur} = \frac{\text{Pourcentage}}{\text{Largeur}}.$$

Pour la première classe, la hauteur est :

$$\frac{24,44}{10} = 2,44.$$

Cette valeur signifie que si l'on répartissait uniformément les données de la classe en 10 classes d'une largeur de 1 millier de dollars, il y aurait 2,44 % des données

dans chacune de ces 10 classes de même largeur. On peut donc dire que dans cette classe la densité est de 2,44 % des données par tranche de 1 millier de dollars.

Pour la deuxième classe, la hauteur est :

$$\frac{5,19}{10} = 0,52.$$

Cette valeur signifie encore que si l'on répartissait uniformément les données de la classe en 10 classes d'une largeur de 1 millier de dollars, il y aurait 0,52 % (la densité) de données dans chacune de ces 10 classes de même largeur.

Pour la dernière classe, la hauteur est :

$$\frac{18,52}{40} = 0,46.$$

Cette valeur signifie encore une fois que si l'on répartissait uniformément les données de la classe en 40 classes d'une largeur de 1 millier de dollars, il y aurait 0,46 % (la densité) des données dans chacune de ces 40 classes de même largeur. Une telle façon de procéder permet de comparer les classes, car la densité est calculée pour des tranches de 1 millier de dollars dans toutes les classes.

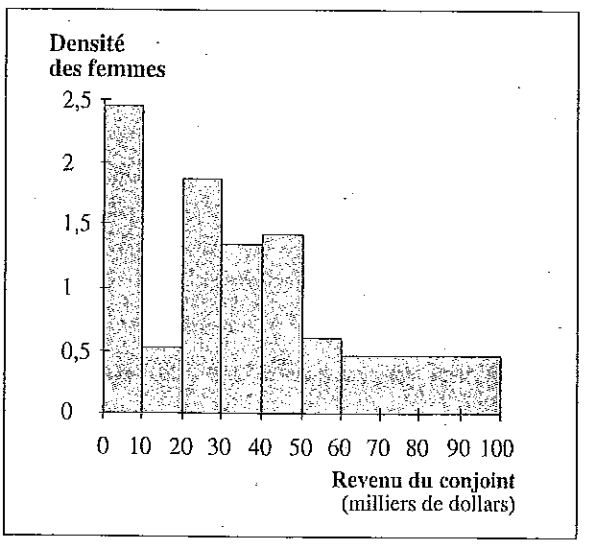
Si l'on reprend le tableau 4.21 et qu'on change la colonne de pourcentage par la colonne des densités, on obtient le tableau 4.22.

TABLEAU 4.22
Répartition des femmes mariées ou en union libre, âgées de 46 à moins de 56 ans, en 1991, en fonction du revenu de leur conjoint en 1990

Revenu du conjoint (milliers de dollars)	Point milieu (milliers de dollars)	Nombre de femmes	Densité des femmes	Pourcentage cumulé des femmes
De 0 à moins de 10	5	33	2,44	24,44
De 10 à moins de 20	15	7	0,52	29,63
De 20 à moins de 30	25	25	1,85	48,15
De 30 à moins de 40	35	18	1,33	61,48
De 40 à moins de 50	45	19	1,41	75,55
De 50 à moins de 60	55	8	0,59	81,48
De 60 à moins de 100	80	25	0,46	100,00
Total		135		

Si l'on construit le nouvel histogramme en tenant compte des densités, on obtient la figure 4.23.

FIGURE 4.23
Répartition des femmes mariées ou en union libre, âgées de 46 à moins de 56 ans, en 1991, en fonction du revenu de leur conjoint en 1990



La construction du polygone nécessite aussi l'utilisation de la densité des classes. Cependant, cette façon de procéder ne s'applique pas à la construction de la courbe des pourcentages cumulés, puisqu'elle tient compte seulement du pourcentage accumulé à la fin de chacune des classes ; dans la classe « De 60 à moins de 100 », on a accumulé le même pourcentage de données, soit 18,52 %, que cette classe soit subdivisée ou non.

c) Le calcul des mesures avec des classes de largeurs inégales

Quelle que soit la largeur de la classe, on utilise le point milieu pour effectuer les calculs. Pour des données groupées en classes, le calcul du mode est basé sur la hauteur des rectangles de l'histogramme ; il faudra donc utiliser les densités dans la formule.

C. Une variable quantitative discrète

EXEMPLE 4.36

Une recherche auprès d'un échantillon aléatoire de 178 compagnies québécoises, composées de 200 employés salariés ou moins et régis par une convention collective, a permis de noter le nombre de syndiqués dans chacune des compagnies.

Variable	Le nombre d'employés syndiqués
Type de variable	La variable est quantitative discrète.
Population	Toutes les compagnies québécoises de 200 employés salariés ou moins régis par une convention collective
Unité statistique	Une compagnie québécoise de 200 employés salariés ou moins régis par une convention collective
Taille de l'échantillon	$n = 178$

Voici la liste des données :

16	34	73	154	18	27	63	5	20	71
14	27	73	132	10	37	99	11	38	72
19	27	61	174	13	41	96	6	27	84
15	22	57	103	18	20	61	9	33	50
15	20	74	148	14	37	60	11	38	84
6	24	57	183	16	44	36	7	27	67
12	41	56	198	16	35	6	47	9	30
12	34	63	167	7	37	12	38	12	27
8	40	76	159	13	36	15	27	8	45
18	48	63	170	17	20	14	37	19	28
17	28	67	132	8	48	9	23	12	45
19	34	65	121	18	25	10	35	10	46
12	46	80	156	11	34	8	38	7	47
6	49	90	166	15	7	8	24	15	40
13	26	96	170	16	6	17	6	8	47
5	29	62	173	8	17	9	12	18	12
16	17	6	5	12	11	19	19	17	16
12	11	19	18	7	16	19	5		

Comme on peut le remarquer, ce cas est traité comme celui d'une variable quantitative continue parce qu'il y a un très grand nombre de valeurs différentes dans la série de données brutes. Il faudra faire des regroupements en classes pour présenter la distribution sous forme de tableau ou de graphique. On utilisera le même procédé que l'on a utilisé pour des données quantitatives continues, mais il ne faut pas oublier qu'il s'agit d'une variable quantitative discrète.

a) Le tableau

Les données sont regroupées selon la démarche suivante :

– **Première étape : Titrer le tableau**

La formulation sera celle-ci : Répartition des compagnies québécoises, de 200 employés salariés ou moins, en fonction du nombre de syndiqués.

– **Deuxième étape : Construire les classes**

Pour construire les classes, il convient de déterminer :

- le nombre de classes. La taille de l'échantillon est de 178. Le nombre de classes suggéré, d'après le tableau de Sturges pour les échantillons dont les tailles vont de 91 à 180, est de 8. On obtient la même valeur en utilisant la formule :

  $1 + 3{,}322 \cdot \log 178 = 8{,}48 \cong 8$ classes.

- l'étendue des données. L'étendue des données est :

  $198 - 5 = 193$ syndiqués.

- la largeur des classes. La largeur des classes se calcule ainsi :

  $$\frac{193}{8} = 24{,}13 \text{ syndiqués};$$

  le choix de regroupements par tranches de 25 serait approprié.

- les classes. Dans l'échantillon, puisque les valeurs vont de 5 à 198, on pourrait faire des regroupements à partir de 1, ce qui mènerait jusqu'à 200. Les regroupements seraient « De 1 à 25 », « De 26 à 50 », « De 51 à 75 », ..., « De 176 à 200 ».

  Ces classes sont exhaustives et exclusives parce que chaque donnée entre dans une et une seule classe. Il ne faut pas oublier que la variable est quantitative discrète.

– **Troisième étape : Trouver le point milieu des classes**

Dans la deuxième colonne du tableau 4.23, on note le point milieu de chacune des classes.

– **Quatrième étape : Dénombrer les unités statistiques**

Il s'agit de faire le dépouillement des données pour obtenir le nombre de données dans chaque classe et d'indiquer ce nombre dans la troisième colonne du tableau 4.23.

– **Cinquième étape : Établir les pourcentages**

Les données de la quatrième colonne représentent le nombre d'unités dans chaque classe sous forme de pourcentage.

– **Sixième étape : Établir les pourcentages cumulés**

Il y a 51,12 % des compagnies qui comptent au plus 25 syndiqués ; il y a 76,40 % des compagnies qui comptent au plus 50 syndiqués, etc.

Le tableau 4.23 résulte de la démarche précédente.

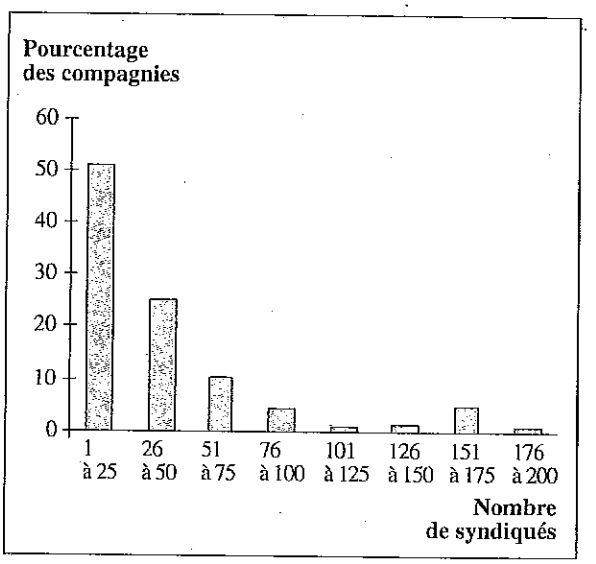
TABLEAU 4.23
Répartition des compagnies québécoises, de 200 employés salariés ou moins, en fonction du nombre de syndiqués

Nombre de syndiqués	Point milieu	Nombre de compagnies	Pourcentage des compagnies	Pourcentage cumulé des compagnies
De 1 à 25	13	91	51,12	51,12
De 26 à 50	38	45	25,28	76,40
De 51 à 75	63	18	10,11	86,51
De 76 à 100	88	8	4,49	91,00
De 101 à 125	113	2	1,12	92,12
De 126 à 150	138	3	1,69	93,81
De 151 à 175	163	9	5,06	98,87
De 176 à 200	188	2	1,12	100,00
Total		178	100,00	

b) Le graphique

Puisque la variable est quantitative discrète, on construit un diagramme en bâtons (figure 4.24) ; un bâton correspond à chaque regroupement de valeurs.

FIGURE 4.24
Répartition des compagnies québécoises, de 200 employés ou moins salariés, en fonction du nombre de syndiqués



c) Les mesures de tendance centrale

- **Le mode**

  La classe modale est de 1 à 25 syndiqués. Le nombre de salariés syndiqués qu'on trouve le plus souvent parmi les compagnies étudiées, ayant 200 employés salariés ou moins, est de 1 à 25 syndiqués.

  Le mode brut est le centre de la classe modale, soit 13 syndiqués. Le nombre de syndiqués autour duquel il y a une plus forte concentration de compagnies ayant 200 employés salariés ou moins, parmi les compagnies étudiées, est d'environ 13 syndiqués.

  Pour trouver la valeur du mode à l'aide de la formule, il faut considérer la variable comme une variable quantitative continue.

- **La médiane**

  La classe médiane est « De 1 à 25 ». Si l'on désire trouver une valeur à l'aide d'une ogive, il faut aussi considérer la variable comme une variable quantitative continue. On se contentera de remarquer que la valeur de la médiane des données

groupées devrait être d'environ 25 employés, étant donné que la classe « De 1 à 25 » cumule 51,12 % des données ; environ 50 % des compagnies québécoises étudiées, ayant 200 employés salariés ou moins, ont au plus 25 syndiqués.

– **La moyenne**

Le calcul de la moyenne se fait à l'aide du point milieu et du nombre de données de chacune des classes. La moyenne des données groupées donne 40,5 ; on peut dire que le nombre moyen d'employés syndiqués dans les compagnies québécoises étudiées, ayant 200 employés salariés ou moins, est d'environ 40,5 syndiqués.

– **Le choix de la mesure de tendance centrale**

Les trois mesures de tendance centrale sont dans l'ordre suivant :

Mode < Médiane < Moyenne

13 < 25 < 40,5

ce qui représente un signe d'asymétrie à droite. De plus, si l'on examine le graphique, on remarque qu'il y a une asymétrie à droite. Il est donc approprié de choisir la médiane comme mesure de tendance centrale.

d) Les mesures de dispersion

– **L'écart type**

L'écart type des données groupées est de 42,1 syndiqués. La dispersion du nombre d'employés syndiqués dans les compagnies québécoises étudiées, ayant 200 employés salariés ou moins, correspond à une dispersion dont l'écart type est d'environ 42,1 syndiqués.

– **Le coefficient de variation**

Le coefficient de variation des données groupées est de 103,95 %. La distribution du nombre de syndiqués dans les compagnies québécoises étudiées, ayant 200 employés salariés ou moins, n'est pas homogène puisque le coefficient de variation est nettement supérieur à 15 %. La dispersion relative à la valeur de la moyenne est trop importante pour considérer cette dernière comme mesure de tendance centrale des données.

EXERCICES

4.8 Le *baby-boom* désigne « l'explosion démographique qu'ont connue certains pays, dont les États-Unis, le Canada, la Nouvelle-Zélande et l'Australie, après la Seconde Guerre mondiale[13] ». Statistique Canada a fait une étude sur les femmes nées durant cette période ; les données suivantes représentent le revenu annuel, en milliers de dollars, de 120 femmes canadiennes, nées entre 1946 et 1965, ayant travaillé à temps plein toute l'année en 1990[14].

22,9	23,8	20,8	13,5	16,0	23,3	25,6	31,1	33,1	26,7	30,5	31,0
34,1	9,6	27,2	14,8	13,9	28,4	27,4	34,7	11,5	32,7	25,4	22,8
30,7	30,2	24,6	26,1	19,7	28,1	28,1	26,6	20,2	30,1	28,6	24,7
31,2	29,6	28,0	21,4	28,4	33,0	15,7	38,7	23,8	26,1	15,6	24,3
30,0	35,7	23,5	19,1	26,1	21,3	23,0	26,5	25,9	24,9	26,5	36,6
36,4	17,9	26,9	14,1	27,0	33,0	17,1	20,7	26,9	27,8	23,0	21,8
26,4	19,8	33,7	28,0	34,9	16,8	39,6	26,3	20,8	27,3	24,0	26,1
14,2	22,7	18,5	28,4	29,7	21,3	32,4	26,4	24,1	32,3	22,5	20,1
25,7	17,3	26,9	17,1	25,0	22,4	26,3	22,7	26,6	18,8	21,5	29,2
22,9	27,2	34,3	23,9	24,1	24,2	37,6	27,6	25,3	36,3	23,6	36,0

13. *Les femmes du* baby-boom *: une génération au travail*, Ottawa, Statistique Canada, Catalogue 96-315F, 1994, p. 5.

14. Échantillon fictif simulé à partir des informations de Statistique Canada. *Op. cit.*, p. 36.