

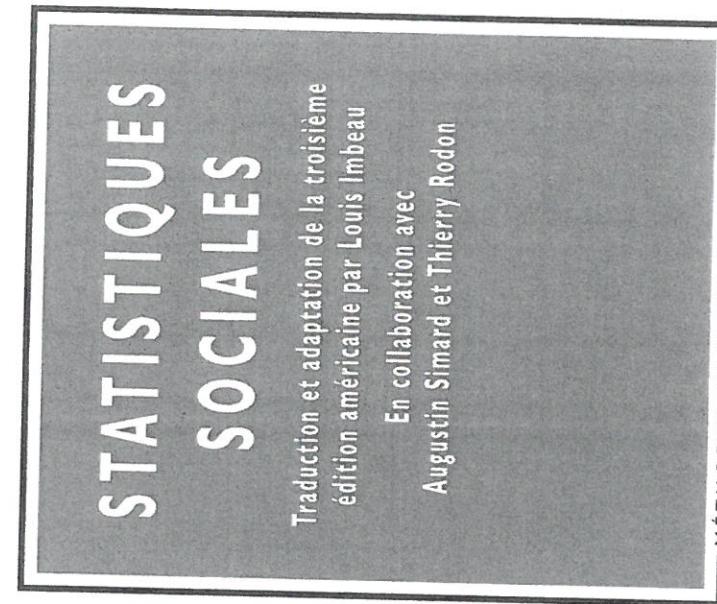
1999

William FOX

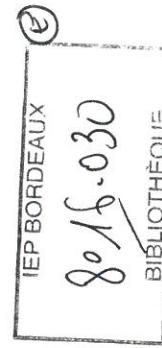
MÉTHODES EN SCIENCES HUMAINES

Collection dirigée par Jean-Marie De Ketelaer,
Jean-Marie Van der Maren et Marie Duris Bellat

- ALBARELLO L., Apprendre à chercher
ANIS J., Texte et ordinateur. L'écriture réinventée ?
ARCAND R., BOURBEAU N., La communication efficace
COLSON J., Le discours
COSNERFROY L., Méthodes de travail et démarches de pensée
CRÈTE J., IMBEAU L. M., Comprendre et communiquer la science
DE KETELE J.-M., ROEGIERS X., Méthodologie du recueil d'informations (3^e éd.)
ENGLEBERT A., Le mémoire sur ordinateur
FOX W., Statistiques sociales
HOWELL D. C., Méthodes statistiques en sciences humaines
HUBERMAN A. M., MILES M. B., Analyse des données qualitatives
JONES R. A., Méthodes de recherche en sciences humaines
JUCQUOIS G., Rédiger, présenter, composer (2^e éd.)
Jucquois G., VIELLE C., Le comparatisme dans les sciences de l'homme
LAVEAULT D., Grégoire J., Introduction aux théories des tests en sciences humaines
LENOBLE-PINSON M., La rédaction scientifique
LESSARD-HÉBERT M., GOYETTE G., BOUTIN G., La recherche qualitative. Fondements et pratiques
MACE G., Guide d'élaboration d'un projet de recherche (2^e éd.)
PIRET A., NISET J., BOURGEOS E., L'analyse structurale
THIRY P., Notions de logique (3^e éd.)
VAN DER MAREN J.-M., Méthodes de recherche pour l'éducation (2^e éd.)
VAN DER MAREN J.-M., La recherche appliquée en pédagogie



MÉTHODES EN SCIENCES HUMAINES



 De Boeck Université

Figure 2. Ratio homme-femme dans les 50 États américains
Nombre d'hommes par 100 femmes



CHAPITRE 3

Les mesures de tendance centrale

Les distributions de pourcentages et les méthodes d'analyse visuelle, tels les diagrammes et les cartes, sont certes d'une grande aide lorsque vient le temps de résumer des informations portant sur une variable. Ces techniques statistiques rendent plus aisée la manipulation de l'information. Toutefois nous pouvons résumer d'une manière bien plus concise des informations univariées en calculant des **mesures de tendance centrale**. Une mesure de tendance centrale est une valeur typique ou représentative d'un ensemble de scores. À l'instar des pourcentages, vous avez appris à calculer certaines de ces mesures à la petite école. Cependant, au cas où vous seriez un peu rouillé, nous y jetterons un rapide coup d'œil.

À la suite de ce chapitre, vous pourrez :

1. Définir ce que sont des modes, des médianes et des moyennes, et être capable de les calculer.
2. Expliquer les caractéristiques importantes du mode, de la médiane et de la moyenne.
3. Expliquer et reconnaître les conditions dans lesquelles chaque type de mesure de tendance centrale est approprié.
4. Expliquer les effets que produisent les scores extrêmes sur les moyennes.
5. Expliquer pourquoi ces scores extrêmes n'influent guère sur les modes et les médianes.
6. Expliquer en quoi consiste une somme de carrés et la calculer.
7. Reconnaître des distributions unimodales et bimodales.
8. Reconnaître l'asymétrie négative et positive dans la distribution d'une variable et expliquer son effet sur la médiane et la moyenne.

3.1 Le mode

Il existe de nombreux types de mesures de tendance centrale, certains étant d'ailleurs assez ésotériques. (N'avez-vous jamais entendu parler de moyenne géométrique ou de moyenne harmonique ?) Trois types de mesure de tendance centrale néanmoins s'avèrent particulièrement utiles : le mode, la médiane et la moyenne. Considérons tour à tour chacune d'elles.

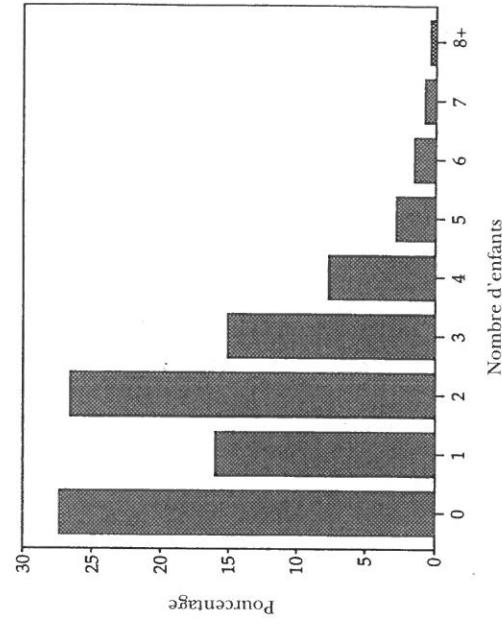
Un **mode** (parfois abrégé *Mo*) est le score qui apparaît le plus fréquemment pour une variable donnée. Aux États-Unis (comme dans plusieurs autres pays), le sexe modal est le féminin car il y a plus de femmes que d'hommes. Le mode des années de scolarité est 12 parce qu'il y a plus de gens qui ont terminé exactement douze années de scolarité que toute autre valeur. En jetant un coup d'œil furtif sur le diagramme illustrant l'écoute de la télévision que l'on retrouve à la figure 2.3 de la section 2.9, nous remarquons que le mode de cette variable est 2 – plus de répondants affirment regarder la télévision deux heures par jour que toute autre valeur.

Lorsqu'un seul des scores apparaît clairement comme le plus commun, nous disons de la variable qu'elle est *unimodale*. Un diagramme en bâtons décrivant une variable unimodale ne laissera voir qu'une seule bosse prononcée. La variable mesurant les heures d'écoute quotidienne de la télévision des répondants du General Social Survey (voir la figure 2.3) est bel et bien unimodale – le nombre de scores de 2 est beaucoup plus élevé que n'importe quel autre score. La conception que les répondants de la même enquête se font de la famille idéale est, elle aussi, fortement unimodale. Plus de la moitié (55 %) des répondants considèrent que deux enfants est le nombre idéal, et la barre qui y correspond domine toutes les autres. Réflétant les tendances démographiques, la variable « Âge », quant à elle, n'a pas un score unique qui domine tous les autres, mais sa distribution montre une croissance graduelle jusqu'à la fin de la trentaine pour ensuite décroître graduellement vers les âges avancés. L'âge aussi est distribué de façon unimodale.

Il arrive parfois qu'un diagramme en bâtons illustrant une variable ordinaire ou d'intervalle/ratio présente une distribution à deux bosses, un peu à la manière du dos d'un chameau. Il conviendra alors de décrire cette variable comme une variable *bimodale*. Les deux bosses n'ont nullement besoin d'être de la même hauteur. Il suffit qu'elles soient sensiblement égales entre elles et plus grandes que les bâtons des autres valeurs pour qu'on puisse dire que la variable est bimodale. La figure 3.1 montre un graphique qui décrit une variable bimodale du General Social Survey rapportant le nombre d'enfants

de chaque répondant. De façon assez nette, les modes se situent aux valeurs 0 et 2, chacune d'elles comprenant à peu près 25 % de tous les cas. Les bandes se rapprotant à ces valeurs dominent visuellement les autres.

Figure 3.1. Nombre d'enfants



Parfois la distribution d'une variable sera plutôt plate, sans score particulier concentrant une forte proportion des cas. La variable du General Social Survey mesurant le fondamentalisme ou le libéralisme de la religion du répondant présente ce genre de distribution aplatie. Les scores se répartissent à peu près également d'une catégorie à l'autre de l'échelle de libéralisme et aucune des catégories n'attire un nombre particulièrement élevé de cas. Dans un cas comme celui-là, on dit que la variable n'a pas de mode. Si on veut caractériser un mode qui n'est pas très prononcé, on parle alors d'une distribution faiblement modale.

Il n'existe aucune formule qui permette de calculer le mode. Nous trouvons le mode uniquement en discernant la valeur qui apparaît le plus souvent parmi l'ensemble des cas. Nous pouvons faire cela sans grande peine à l'aide d'un tableau de pourcentages ou de fréquences ou encore à l'aide d'un diagramme en bâtons.

Le mode dépend des différences de fréquences des scores. Il n'implique ni l'ordre des valeurs ni les unités de mesure. Ainsi nous

poumons obtenir un mode pour toutes les variables, peu importe leur niveau de mesure – nominal, ordinal ou d'intervalles/ratio. Le mode est la seule mesure de tendance centrale qui convient aux variables nominales.

Néanmoins le mode est de peu d'utilité dans le cas de variables qui ont de nombreuses valeurs car souvent aucune valeur ne peut se démarquer des autres par une fréquence plus élevée. Tel est souvent le cas de variables d'intervalles et de ratio. Par exemple, les variables continues d'intervalles/ratio (le taux de meurtres par 100 000 habitants, par exemple), que l'on retrouve dans des banques de données écologiques, ont peu de cas avec des scores identiques et, ainsi, n'ont pas de mode qui puisse nous intéresser. Pour de telles variables, le mode que l'on peut obtenir nous renseigne bien piètement sur le score typique de la distribution.

3.2 La médiane

La **médiane** (que l'on abrège parfois *Md*) est la valeur qui divise en deux parties égales un ensemble ordonné de scores. L'expression « un ensemble ordonné de scores » implique que les scores puissent être disposés en ordre du plus petit au plus grand. La médiane est le point en dessous duquel se trouve une moitié des scores et au-dessus duquel se situe l'autre moitié des scores. C'est en quelque sorte le milieu ou le point central des scores ordonnés. Soyez vigilants : la médiane est le point milieu de l'ensemble des scores – c'est-à-dire des mesures réelles – et non des valeurs possibles.

Dénicher la médiane est particulièrement facile lorsque que nous avons affaire à un nombre impair de cas :

1. Disposez les scores en ordre, du plus petit au plus grand.
2. Trouvez le score qui se trouve au milieu.

La valeur du score situé exactement au milieu est la médiane. Nous pouvons obtenir le score médian, celui qui se trouve au centre, grâce à cette formule $\frac{N+1}{2}$, dans laquelle N représente le nombre de cas.

Voici un exemple :

| Scores originaux | Scores ordonnés |
|------------------|-----------------|
| 6 | 1 |
| 2 | 2 |
| 5 | 4 |
| 9 | 5 <= Médiane |
| 1 | 6 |
| 6 | 6 |
| 4 | 9 |

La médiane est 5. Ce nombre est le quatrième score $\frac{N+1}{2} = \frac{7+1}{2} = 4$ de l'ensemble ordonné des scores. Il y a autant de scores inférieurs à 5 que de scores supérieurs à 5 (3 dans les deux cas).

Dans le cas où il y aurait un nombre pair de scores, trouver la médiane est un peu plus difficile... mais pas vraiment. Voici ce qu'il faut faire :

1. Disposez les scores en ordre, du plus petit au plus grand.
2. Trouvez les deux scores qui se trouvent au milieu.
3. Faites la moyenne de ces deux scores en les additionnant et en divisant la somme par deux.

Il en résulte la médiane. Voici un exemple à partir d'un ensemble de six scores (déjà ordonnés du plus petit au plus grand) :

$$\begin{array}{ccccccc} & & & 4 & & & \\ & & & \textcircled{6} & \xrightarrow{\quad\quad\quad} & \frac{8+9}{2} & = 8,5 \\ & & & \textcircled{8} & & & \\ & & & 9 & & & \\ & & & & 10 & & \\ & & & & & 15 & \end{array}$$

La médiane est 8,5. La moitié des scores sont inférieurs à 8,5, la moitié lui sont supérieurs.

Il importe peu que les deux scores qui se trouvent au milieu de la distribution aient la même valeur. Lorsque cette situation se présente et que les deux scores centraux ont la même valeur, cette valeur est la médiane. Si, par exemple, il y avait trois scores 8 additionnels dans la dernière distribution que nous avons examinée (auquel cas les scores seraient 4, 6, 8, 8, 8, 9, 10, 15), la médiane serait 8.

Qu'une variable ordinaire ait des valeurs « alphabétiques » plutôt que numériques n'importe guère plus. Ces scores peuvent s'ordonner et nous pouvons trouver la médiane. Voilà, en guise d'exemple, comment on trouve la médiane pour la variable « classe sociale », « supérieure ».

| Scores originaux | Scores ordonnés |
|------------------|-------------------|
| Moyenne | |
| Ouvrière | Inférieure |
| Inférieure | Ouvrière |
| Superérieure | Ouvrière |
| Moyenne | Moyenne < Médiane |
| Moyenne | Moyenne |
| Ouvrière | Supérieure |

Certes trouver la médiane est très facile. Seulement je compliquerai maintenant cette tâche en traitant d'un problème que l'on a souvent à affronter étant donné le type de données avec lesquelles nous travaillons. Les spécialistes des sciences sociales se servent souvent de variables qui sont conceptuellement continues, dans la mesure où elles peuvent théoriquement prendre n'importe quelle valeur, mais qui sont mesurées par des scores discrets, entiers (comme des nombres entiers). Tel est le cas de plusieurs des variables décrivant des attitudes – des variables que l'on voit fréquemment dans les sondages de sciences sociales. Prenons par exemple la variable du General Social Survey qui mesure l'attitude des répondants envers les immigrants. On a demandé aux répondants de l'enquête dans quelle mesure ils étaient d'accord avec l'affirmation suivante : « Les immigrants volent les emplois de ceux qui sont nés ici ». Les choix de réponses qu'on leur a donnés étaient les suivants :

- 1 Fortement d'accord
- 2 D'accord
- 3 Ni d'accord, ni en désaccord
- 4 En désaccord
- 5 Fortement en désaccord

Les répondants doivent en fait « arrondir » leur score au nombre entier le plus près. Même si un répondant soutenait l'affirmation légèrement plus (ou légèrement moins) que ne le suggère la valeur « D'accord », on lui attribuait le score 2. D'un point de vue conceptuel, les scores de cette variable vont de 0,5 à 5,5. La « véritable » réponse d'un répondant peut se situer n'importe où à l'intérieur de ce continuum. Théoriquement, un répondant pourrait avoir un score de 3,2, de 4,8, ou encore de n'importe quelle valeur comprise entre 0,5 et 5,5. Les concepteurs de l'enquête ont créé les questions de telle façon que, dans les faits, les répondants doivent « arrondir » leur « véritable » réponse au nombre entier le plus près. Aussi les « véritables » scores des répondants qui ont répondu 2, par exemple,

peuvent se trouver n'importe où entre 1,5 et 2,5, l'intervalle à l'intérieur duquel les réponses sont arrondies à 2.

Règle générale, nous calculons la médiane en assumant, pour simplifier, que les scores à l'intérieur de la catégorie contenant la médiane sont disposés de façon à ce qu'un espace de même grandeur se trouve entre chacun d'eux, et cela pour tout l'intervalle de cette catégorie. À l'intérieur de cet intervalle, nous trouverons le score qui divise en moitiés la distribution. Cette procédure permet d'obtenir une médiane plus raffinée, plus précise que celle que nous obtiendrions en prenant bêtement la valeur entière de la catégorie qui contient la médiane.

À titre d'exemple, voici la distribution de fréquences et de fréquences cumulatives d'une variable mesurant l'attitude des répondants envers les immigrants et les emplois :

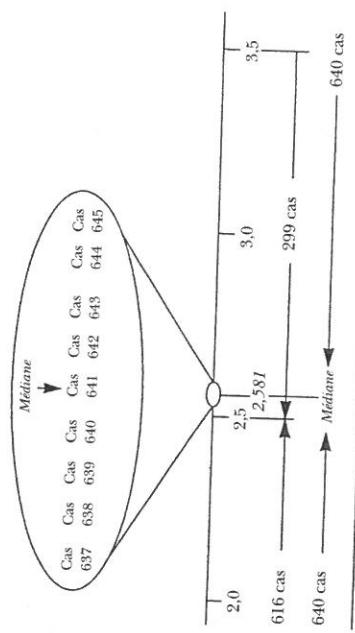
| Valeur | f | F |
|--------|-----|------------|
| 1 | 170 | 170 |
| 2 | 446 | 616 |
| 3 | 299 | 915 |
| 4 | 301 | 1 216 |
| 5 | 65 | 1 281 |
| (N) | | 299 scores |

Rappelez-vous, comme nous le disions à la section 2.3, que F est le symbole signifiant « fréquences cumulatives ». La distribution des fréquences cumulatives aide à déceler où se trouve la médiane. Parce qu'il y a 1 281 scores au total, la médiane est la valeur du 641^e score. (Il y a 640 scores qui lui sont inférieurs et 640 qui lui sont supérieurs.) Nous pouvons donc dire, grâce à cette distribution cumulative, que le 641^e score se situe quelque part parmi les 299 scores 3.

Afin de calculer la médiane, nous assumons qu'il y a la même distance entre chacun des 299 scores compris dans l'intervalle allant de 2,5 à 3,5, c'est-à-dire l'intervalle de la valeur 3. Nous interpolons pour découvrir à quel endroit se situe la médiane dans cet intervalle. « Interpoler » signifie trouver le point, à l'intérieur de l'intervalle allant de 2,5 à 3,5, qui correspond au 641^e score.

L'endroit où se situe la médiane est représenté graphiquement par la figure 3.2. Imaginez que les cas soient distribués sur l'ensemble du continuum et qu'ils y soient espacés également l'un de l'autre. Les 299 cas dont le score est 3 seraient distribués également entre les scores 2,5 et 3,5. L'ovale montre un agrandissement de la portion de cet intervalle qui contient la médiane. Voyez ?... La médiane est là ! Le 641^e score.

Figure 3.2. Interpolation de la médiane



C'est bien joli, me direz-vous, mais comment interpolate-t-on ? Simplement en employant la formule ci-dessous :

$$Md = L + \left(\frac{\frac{N}{2} - F}{f} \right) i$$

lorsque Md = la médiane

L = la limite inférieure de l'intervalle contenant la médiane

N = le nombre total de scores

F = la fréquence cumulative des scores inférieurs à l'intervalle contenant la médiane

f = le nombre de scores que comprend l'intervalle contenant la médiane

i = la largeur de l'intervalle contenant la médiane (c'est-à-dire la limite supérieure de l'intervalle moins sa limite inférieure)

Concernant l'attitude envers les immigrants et les emplois :

$$\begin{aligned} Md &= L + \left(\frac{\frac{N}{2} - F}{f} \right) i \\ &= 2,5 + \left(\frac{\frac{1281}{2} - 616}{299} \right) 1 \\ &= 2,5 + \frac{24,5}{299} \\ &= 2,5 + 0,081 \\ &= 2,581 \\ &= 2,6 \end{aligned}$$

En d'autres termes, si les 299 scores 3 étaient disposés à distance égale les uns des autres dans l'intervalle 2,5 à 3,5, la médiane serait 2,581 ou, une fois arrondie, 2,6. La médiane se situe donc près de l'extrémité inférieure de l'intervalle de la valeur 3 (Ni d'accord, ni en désaccord). Vous conviendrez avec moi qu'il s'agit là d'une estimation bien plus précise de la médiane que le serait la simple valeur 3.

Puisque le calcul de la médiane exige que les scores soient ordinés, on ne peut trouver une médiane que pour des variables ordinaires ou d'intervalles/ratio. Il n'y a pas de médiane pour les variables nominales parce qu'il n'y a pas d'ordre entre les scores d'une variable nominale. Une médiane pour des variables nominales comme le sexe, la race ou la religion n'aurait pas de sens. Alors je répète : une variable nominale n'a pas de médiane !

Remarquez que les scores extrêmes n'affectent pas la médiane parce que celle-ci décrivit seulement le score qui se situe au centre d'une distribution ordonnée. Par exemple, les ensembles suivants de scores ont la même médiane, 5,5 :

| | <i>Ensemble A</i> | <i>Ensemble B</i> | <i>Ensemble C</i> |
|----|-------------------|-------------------|-------------------|
| 51 | 1 | 51 | |
| 52 | 52 | 52 | |
| 54 | 54 | 54 | |
| 55 | 55 | 55 | < Médiane |
| 56 | 56 | 56 | |
| 56 | 56 | 56 | |
| 59 | 59 | 590 | |

Nous verrons dans la prochaine section que cette indépendance par rapport aux valeurs extrêmes est un avantage important de la médiane.

Pour des variables comme celles du General Social Survey, il peut sembler sage d'arrondir la médiane à la première décimale. Il n'y a cependant aucune règle proprement dite à cet effet. Vous devrez donc bien réfléchir, pour chacune des médianes, à la décimale à laquelle il est approprié d'arrondir.

Assurez-vous d'avoir exclu les données manquantes (« Ne sait pas », « Pas de réponse », etc.) avant de calculer une médiane. De telles valeurs fournissent peu d'information. C'est pourquoi il vaut mieux les exclure. D'ailleurs, l'inclusion de ces données transformerait une variable ordinaire ou d'intervalles/ratio en variable nominale puisqu'il n'y a pas d'ordre entre les valeurs des données manquantes.

Vous aurez probablement remarqué qu'il importe peu que la médiane n'indique pas un score réel ou même un score possible. La médiane reste tout de même le point milieu d'un ensemble ordonné de scores. Les profanes des statistiques sont amusés de voir une médiane « impossible ». Ils se bidonnent lorsqu'on leur dit que le nombre médian de personnes par ménage aux États-Unis est de 2,7. « Comment est-ce possible ? » s'esclaffent-ils. « Une personne est enceinte ? » Mais maintenant que vous savez ce qu'est une médiane et qu'il n'est nul besoin qu'elle corresponde à une valeur réelle, vous ne vous perturerez plus à cet humour gras.

3.3 La moyenne

Enfin définissons ce qu'est une moyenne. La *moyenne* est la mesure de tendance centrale que l'on obtient en additionnant tous les scores et en divisant ensuite cette somme par le nombre de scores. C'est vraiment simple :

1. Additionnez tous les scores.

2. Divisez la somme par le nombre de scores.

Voici un exemple à partir de cet ensemble de six scores : 4, 8, 10, 11, 9, 6. Pour obtenir la moyenne, additionnez tous les scores : $4 + 8 + 10 + 11 + 9 + 6 = 48$. Ensuite, divisez cette somme par le nombre de scores : $48 \div 6 = 8$. La moyenne est 8.

Oui, je sais que vous calculez déjà des moyennes depuis des années. Examinez quand même la formule de la moyenne pour comprendre comment elle se calcule :

$$\bar{X} = \frac{\sum X_i}{N}$$

lorsque \bar{X} = la moyenne

X_i = le score du i^{e} cas

N = le nombre de scores

Quelques mots sur la notation. \bar{X} est prononcé « X-barre ». Les statisticiens utilisent la notation \bar{X} pour indiquer la moyenne d'un échantillon. Pour signifier la moyenne d'une *population*, ils emploient le symbole μ (prononcé « mu »). En général, les lettres romaines renvoient aux statistiques portant sur un échantillon et les caractères grecs aux paramètres d'une population. (Rappelez que nous disions, à la section 1.3 qu'une information qui concerne seulement un échantillon est appelée une statistique, qu'une information portant sur l'ensemble d'une population est un paramètre.) Donc, pour des données de population, la formule de la moyenne devient : $\mu = \frac{\sum X_i}{N}$.

L'indice i dans la notation X_i désigne les scores individuels. C'est en quelque sorte un score générique. X_1 est le premier score, X_2 le second, etc. Ainsi, X_i est le score d'un éventuel i^{e} cas. Cela n'implique pas que les scores doivent être disposés dans un ordre particulier. Σ est la lettre majuscule grecque sigma. Nous la verrons souvent tout au long de ce livre. Les statisticiens utilisent Σ pour indiquer la somme de tout ce qui suit ce caractère. Ainsi, ΣX_i signifie la somme de tous les scores individuels.

Dans l'exemple bien simple qui précède, la moyenne – 8 – se trouvait parmi notre ensemble de valeurs. Mais à l'instar de la médiane, il n'est pas nécessaire que la moyenne soit une valeur « réelle » ou même possible. La moyenne d'un groupe à un test « vrai ou faux », par exemple, peut très bien se chiffrer à 77,8, même si personne ne peut obtenir un tel score.

Il est habituellement de bon ton, pour la plupart des données dont nous nous servons en sciences sociales, d'arrondir la moyenne à la première décimale. Mais peu importe les données avec lesquelles vous travaillez, *réfléchissez à ce que signifient les décimales dans vos calculs*. Ne présentez pas de décimales superflues qui suggéreraient que la moyenne a un degré de précision que ne lui confèrent pas les données sur lesquelles elle repose. Et pendant qu'on y est, n'oubliez pas d'exclure les données manquantes (« Ne sait pas », « Pas de réponse », etc.) avant de calculer une moyenne. L'information que confèrent ces données n'est pas utilisable dans le calcul de la moyenne.

Attention à l'interprétation des moyennes de variables écologiques. De telles moyennes traitent chaque unité écologique comme un cas unique et décrit la tendance centrale des unités et non de la population totale des individus qui forment ces unités. Par exemple, dans le calcul du nombre moyen d'années d'instruction pour les 50 États américains, on traite sur un pied d'égalité les États de la Californie et du Wyoming en dépit de la grande différence dans la taille de leurs populations. Le nombre moyen d'années d'instruction réfère à la tendance centrale pour les 50 États et non pour l'ensemble de la population américaine. Cela est vrai aussi pour les médianes, mais les problèmes d'interprétation des moyennes de variables écologiques sont plus fréquents. Alors faites en sorte de les éviter.

3.4 Les propriétés de la moyenne

Lorsque vous calculez une moyenne, faites attention aux scores qui sont particulièrement bas ou élevés, et tout spécialement aux cas déviants. Un ou deux scores extrêmes peuvent rendre une moyenne bien peu représentative de la plupart des scores, surtout quand le nombre de cas est petit. Imaginez que l'on remplace le score 6, dans l'exemple précédent, par 600. Ce serait sans aucun doute un cas déviant. Avec les scores 4, 8, 10, 11, 9 et 600, la moyenne grimperait à 107, ce qui n'est guère typique ou représentatif de ce que sont « réellement » ces six scores. Dans ce cas, la médiane, 10,5, serait une meilleure mesure de la tendance centrale.

Ou considérez le revenu moyen dans l'île de Gilligan. Le revenu de M. Howell, le millionnaire, gonflerait exagérément le revenu moyen des habitants de l'île. La médiane serait beaucoup plus représentative des revenus de Gilligan et de ses compères. Dans le calcul de la moyenne, le revenu de un million de dollars de M. Howell a le même poids que le revenu de 100 Gilligan qui gagneraient chacun 10 000 \$ par année. Et c'est la même chose sur le continent. Il faudrait 100 000 programmeurs gagnant chacun 50 000 \$ par an pour

égaler le revenu de 5 milliards de dollars du président de Microsoft, Bill Gates. Il n'est pas surprenant que les organismes nationaux de recensement rapportent le revenu médian plutôt que le revenu moyen !

Un autre exemple, véritable celui-ci. Parmi les 50 États américains, le pourcentage de la population se disant de souche asiatique présente un cas déviant bien visible : Hawaii avec 61,8 %. Le score le plus élevé suivant est celui de la Californie, 9,6 %. La moyenne de la variable est 2,8 %, Hawaii comprise ; elle chute à 1,6 % dès que l'on exclut ce cas déviant. Cette dernière moyenne est beaucoup plus représentative du score typique parmi les 50 États américains.

J'ai signalé à la section 2.9 que, lorsque vous rencontrez un cas déviant, il faut commencer par en chercher l'origine. Ensuite observez-en l'effet sur votre analyse – ici, sur la moyenne. Si l'effet est important (comme c'est le cas de nos exemples), songez soit à exclure le cas déviant, soit à vous servir de la médiane plutôt que de la moyenne, cette première n'étant pas affectée par la présence de cas déviants.

Alors attention aux cas déviants quand vous travaillez avec des moyennes. Plus précisément, surveillez les distributions qui ont des queues prononcées vers les valeurs élevées ou les valeurs faibles, même si l'y a pas de cas déviant comme tel. Comme la moyenne tient compte de tous les scores, ces queues peuvent en faire un représentant bien douteux du score typique. Mieux vaut utiliser la médiane dans une telle situation.

J'en ai assez dit des problèmes liés à l'utilisation de la moyenne. Passons à ses nombreux avantages. La moyenne a une propriété à la fois intéressante et importante : lorsque nous la soustrayons de chaque score et que nous additionnons toutes ces différences, le résultat est invariablement zéro. De façon plus succincte, la somme des écarts entre les scores et la moyenne est nulle. Algébriquement : $\sum(X_i - \mu) = 0$ pour des données de population et $\Sigma(X_i - \bar{X}) = 0$ pour des données d'échantillon. Certaines différences seront positives, d'autres négatives. Additionnées, elles s'annulent exactement.

Cette propriété de la moyenne présente un grand intérêt tant théorique que pratique. La moyenne « équilibré » pour ainsi dire une distribution. Si les cas étaient des poids (des barres d'un gramme, disons) disposés à des points correspondant à leur score, la moyenne serait le point où les cas se trouveraient en parfait équilibre. La balance du haut dans la figure 3.3 montre graphiquement l'équilibre des six scores que nous avons utilisés précédemment comme exemples. Les trois poids de droite sont équilibrés par les deux poids plus loin à gauche. Essayez d'équilibrer la distribution au point médian et vous la verrez basculer, comme nous le voyons dans la balance du

égalier le revenu de 5 milliards de dollars du président de Microsoft, Bill Gates. Il n'est pas surprenant que les organismes nationaux de recensement rapportent le revenu médian plutôt que le revenu moyen !

Un autre exemple, véritable celui-ci. Parmi les 50 États américains, le pourcentage de la population se disant de souche asiatique présente un cas déviant bien visible : Hawaii avec 61,8 %. Le score le plus élevé suivant est celui de la Californie, 9,6 %. La moyenne de la variable est 2,8 %, Hawaii comprise ; elle chute à 1,6 % dès que l'on exclut ce cas déviant. Cette dernière moyenne est beaucoup plus représentative du score typique parmi les 50 États américains.

J'ai signalé à la section 2.9 que, lorsque vous rencontrez un cas déviant, il faut commencer par en chercher l'origine. Ensuite observez-en l'effet sur votre analyse – ici, sur la moyenne. Si l'effet est important (comme c'est le cas de nos exemples), songez soit à exclure le cas déviant, soit à vous servir de la médiane plutôt que de la moyenne, cette première n'étant pas affectée par la présence de cas déviants.

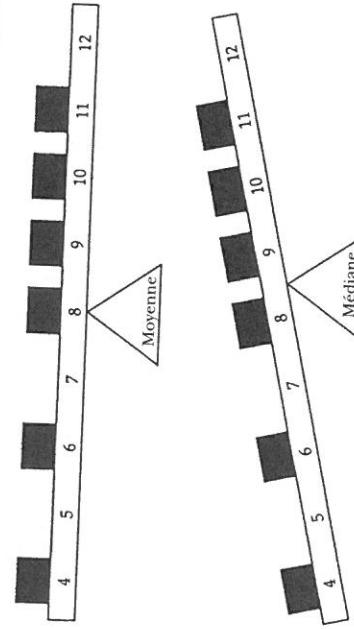
Alors attention aux cas déviants quand vous travaillez avec des moyennes. Plus précisément, surveillez les distributions qui ont des queues prononcées vers les valeurs élevées ou les valeurs faibles, même si l'y a pas de cas déviant comme tel. Comme la moyenne tient compte de tous les scores, ces queues peuvent en faire un représentant bien douteux du score typique. Mieux vaut utiliser la médiane dans une telle situation.

J'en ai assez dit des problèmes liés à l'utilisation de la moyenne. Passons à ses nombreux avantages. La moyenne a une propriété à la fois intéressante et importante : lorsque nous la soustrayons de chaque score et que nous additionnons toutes ces différences, le résultat est invariablement zéro. De façon plus succincte, la somme des écarts entre les scores et la moyenne est nulle. Algébriquement : $\sum(X_i - \mu) = 0$ pour des données de population et $\Sigma(X_i - \bar{X}) = 0$ pour des données d'échantillon. Certaines différences seront positives, d'autres négatives. Additionnées, elles s'annulent exactement.

Cette propriété de la moyenne présente un grand intérêt tant théorique que pratique. La moyenne « équilibré » pour ainsi dire une distribution. Si les cas étaient des poids (des barres d'un gramme, disons) disposés à des points correspondant à leur score, la moyenne serait le point où les cas se trouveraient en parfait équilibre. La balance du haut dans la figure 3.3 montre graphiquement l'équilibre des six scores que nous avons utilisés précédemment comme exemples. Les trois poids de droite sont équilibrés par les deux poids plus loin à gauche. Essayez d'équilibrer la distribution au point médian et vous la verrez basculer, comme nous le voyons dans la balance du

bas. Les scores les moins élevés sont alors trop éloignés et déséquilibrent la balance vers la gauche.

Figure 3.3. La moyenne équilibre un ensemble de scores



| X | $X_i - \bar{X}$ | $(X_i - \bar{X})^2$ |
|--------------|-----------------|---------------------|
| 4 | 4 - 8 = -4 | 16 |
| 8 | 8 - 8 = 0 | 0 |
| 10 | 10 - 8 = 2 | 4 |
| 11 | 11 - 8 = 3 | 9 |
| 9 | 9 - 8 = 1 | 1 |
| 6 | 6 - 6 = 0 | 0 |
| <i>Somme</i> | | 34 |

La somme des carrés est 34. Tout autre nombre qu'on pourrait substituer à la moyenne (8) donnerait une somme plus élevée.

Plus tard nous utiliserons ce concept de la somme des carrés quand nous traiterons de l'analyse de la variance et des techniques de régression et de corrélation. Rappelez-vous-en !

3.5 La moyenne d'une variable dichotomique

Notez que, strictement parlant, la moyenne ne peut être calculée que pour des variables mesurées au niveau d'intervales/ratio. Strictement parlant, il est insensé de calculer, par exemple, la religion moyenne ou la classe sociale moyenne. La religion est une variable nominale ; elle n'est pas mesurée par une unité de mesure standard. Ses valeurs – protestant, catholique, juif – ne peuvent pas être additionnées. De la même façon, si on mesure la classe au niveau ordinal (inférieure, ouvrière, moyenne et supérieure), il ne convient pas, strictement parlant, de calculer une moyenne. La raison est que (strictement parlant) les valeurs d'une variable ordinaire, parce qu'il n'existe pas nécessairement entre elles d'intervalles de même grandeur, ne peuvent s'additionner de façon raisonnable. Nous ne pouvons faire la somme des valeurs « classe inférieure », « classe ouvrière », « classe moyenne » et « classe supérieure ».

J'ai employé à plusieurs reprises l'expression « strictement parlant » parce qu'il peut parfois s'avérer utile d'agir comme si les nombres représentant les valeurs d'une variable ordinaire avaient une signification arithmétique, et ainsi de calculer une moyenne à partir de ces nombres. Dans la section précédente, nous avons vu que c'était le cas pour la médiane, et cela est vrai aussi pour la moyenne. En ce qui concerne les classes sociales, par exemple, nous pourrions décider d'assigner le nombre 1 à la classe inférieure, 2 à la classe ouvrière, 3 à la classe moyenne et 4 à la classe supérieure, un peu comme nous l'avons fait lorsque que nous avons calculé, à la section 3.2, une médiane interpolée. Nous pourrions alors calculer une moyenne en nous basant sur ces nombres, en les utilisant comme s'il s'agissait là de

Voici une caractéristique encore plus importante de la moyenne, caractéristique qui est reliée à la précédente : la moyenne minimise la somme des déviations au carré de chaque score par rapport à la moyenne. Le mot **déivation** renvoie à la différence entre un score et élevons chaque différence au carré et additionnons ces carrés, nous obtenons une somme qui est plus petite que celle que nous obtiendrions en utilisant tout autre nombre que la moyenne. Algébriquement, cela revient à dire que, pour des données d'échantillon, $\Sigma(X_i - \bar{X})^2$ est un minimum ; et que, pour des données de population, $\Sigma(X_i - \mu)^2$ est un minimum. Aucun autre nombre que la moyenne ne peut produire une somme plus petite.

Ces deux expressions – $\Sigma(X_i - \bar{X})^2$ et $\Sigma(X_i - \mu)^2$ – sont tellement importantes en statistiques qu'elles ont un nom qui leur est propre :

somme des carrés. Et ce concept est tellement utile que je veux vous répéter ce qu'il signifie : la somme des carrés est la somme des déviations par rapport à la moyenne, élevées au carré. Prenons, par exemple, les six cas que nous avons utilisés dans la section précédente, et dont la moyenne est égale à $\bar{X} = 8$:

scores. Il serait donc possible de parler de la classe sociale moyenne. Après tout, il n'est guère difficile d'interpréter intelligemment cette moyenne. Par exemple, une moyenne de la classe sociale se situant à 2,7 indiquerait que la classe typique des répondants se situe entre la classe ouvrière et la classe moyenne, un peu plus près de la seconde que de la première.

Cette façon de traiter les variables ordinaires comme s'il s'agissait de variables d'intervalles/ratio est condamnable aux yeux de statisticiens puristes. Mais, de la même manière qu'on ne fait pas d'omelette sans casser les œufs, vous ne pourrez vous livrer à certaines analyses sans violer quelques règles... *pourvu que vous soyez conscient de ce que vous êtes en train de faire.* Nous pouvons parfois faire meilleur usage des données si nous nous donnons la peine de calculer la moyenne d'une variable ordinaire. Nous devrons cependant toujours observer la plus grande prudence lorsque viendra le temps d'interpréter ces moyennes. En qualité de novice des statistiques, ne sachant donc pas toujours très bien ce que vous faites, vous ne devriez pas calculer la moyenne de variables ordinaires. Vous devriez vous en tenir plutôt à la médiane. Sachez toutefois que plusieurs sociologues utilisent la moyenne lorsqu'ils ont affaire à des données ordinaires. Au fur et mesure que vous deviendrez plus familiers avec les statistiques et l'analyse de données, vous pourrez trouver utile d'agir de cette façon vous aussi.

Que dire des variables nominales ? Voit-on parfois des chercheurs assigner des nombres aux valeurs d'une variable nominale comme la religion pour ensuite en calculer la moyenne ? La réponse est « *parfois* » si la variable est dichotomique ; elle est un nébranlable « *NON* » lorsque la variable a trois valeurs ou plus. Même les statisticiens les plus iconoclastes respectent cette règle. Mais lorsque les valeurs d'une variable *dichotomique*, même si elle est nominale, sont codées 0 et 1, la moyenne de la variable est égale à la proportion des cas codés 1. Supposons, par exemple, que les hommes sont codés 0 et les femmes 1 pour la variable « Sexe » et que nous avons les cinq cas suivants :

| Sexe | Code |
|-------|------|
| Femme | 1 |
| Homme | 0 |
| Femme | 1 |
| Femme | 1 |
| Homme | 0 |

$$\Sigma X_i = 3$$

Remarquez que la proportion de femmes est 0,60 (trois sur cinq).

Remarquez aussi que $\bar{X} = \frac{\Sigma X_i}{N} = \frac{3}{5} = 0,60$. Eh oui ! La moyenne est égale à la proportion de femmes. Cela fonctionne chaque fois que les scores d'une variable dichotomique sont codés 0 et 1. En fait, bien que certains correctifs arithmétiques soient nécessaires, le même principe s'applique si les valeurs d'une variable dichotomique sont codées à l'aide de n'importe quelles nombres consécutifs. Si la variable est codée 1 et 2, par exemple, la moyenne moins 1 sera égale à la proportion des cas codés 2. Plus généralement, la moyenne moins le code le plus petit est égale à la proportion des cas désignés par le code le plus élevé.

Cette utilisation des variables dichotomiques, même si elles sont nominales, sera de toile de fond aux statistiques. Au chapitre 10, nous nous pencherons sur les relations entre les variables nominales dichotomiques dans les techniques de régression et de corrélation. Au chapitre 12, nous apprendrons à transformer des variables nominales en variables « factices » codées 0 et 1 de manière à pouvoir les utiliser. Mais pour l'instant, ne calculons pas les moyennes des variables nominales autres que dichotomiques, et même avec ces dernières, nous serons très prudents.

3.6 Lequel employer - le mode, la médiane ou la moyenne ?

Nous avons vu trois mesures de tendance centrale. Mais quelle mesure utiliser : le mode, la médiane ou la moyenne ? La réponse à cette question dépend du niveau de mesure de la variable, de sa distribution ainsi que de ce que l'on tente de découvrir à l'aide de cette variable. Je vous donnerai ici quelques conseils généraux afin de vous guider dans le choix de la mesure de tendance centrale appropriée.

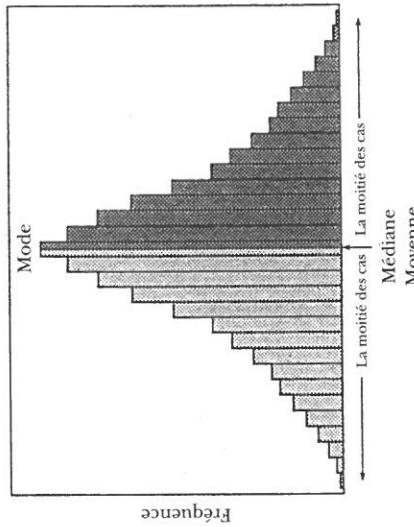
Comme nous l'avons mentionné dans la section précédente, nous ne pouvons (strictement parlant) calculer une moyenne que pour des variables d'intervalle ou de ratio. Bien sûr nous pouvons aussi calculer le mode et la médiane pour ce type de variable. Néanmoins, en écartant certaines considérations (sur lesquelles je reviendrai dans quelques instants), nous devons employer la moyenne lorsque nous travaillons avec des variables d'intervalle/ratio. Tout compte fait, la moyenne est la seule mesure dont le calcul incorpore la totalité des scores. Ainsi la moyenne comporte plus d'informations que le mode et la médiane. Comme il est préférable de faire usage de la totalité de l'information dont nous disposons, la moyenne, en oubliant pour le moment certaines autres considérations, est habituellement la mesure la plus indiquée pour les variables d'intervalle/ratio.

La médiane est la mesure à choisir lorsque que nous avons à faire à des variables ordinaires pour la simple raison que nous ne pouvons régler généralement, calculer la moyenne pour de telles variables (bien que cela soit parfois possible comme nous l'avons vu à la section précédente). La médiane incorpore au moins des informations concernant le ou les scores centraux de la distribution, ce qui est déjà plus que ne le fait le mode. Donc, toutes choses égales par ailleurs, servez-vous de la médiane lorsque vous travaillez avec des variables ordinaires. Cela ne laisse que le mode comme mesure privilégiée pour des variables nominales. En effet, pour des variables nominales, nous n'avons guère le choix puisque leur faible niveau de mesure ne nous permet pas de nous servir de la médiane ni de la moyenne. Bien sûr, comme nous venons de le voir, il est possible de calculer la moyenne des variables nominales dichotomiques codées 0 et 1 en considérant la proportion de cas codés 1. Néanmoins, cela n'est d'aucune utilité dans le cas des variables nominales non dichotomiques. On est généralement limité au mode pour les variables nominales.

Il ne faut toutefois pas dénigrer le mode. Le mode s'avère parfois utile, particulièrement lorsqu'il est nécessaire de connaître le score qui apparaît le plus souvent. Par définition, il y a plus de scores identiques au mode que de scores identiques à toute autre valeur, ce qui n'est pas nécessairement vrai dans le cas de la médiane ou de la moyenne. Son utilité est donc grande lorsqu'il s'agit de prédire un score réel. Si vous avez à deviner précisément le score d'un cas connu, le mode est ce que vous pouvez tenter raisonnablement de mieux. Cette qualité du mode sera particulièrement utile lorsque nous examinerons, au chapitre 7, les mesures d'association. De plus, le mode fournit parfois, dans le cas de variables ordinaires et d'intervalles/ratio, des informations concernant la forme que prend la distribution des scores, des informations qui s'ajoutent à celles que l'on obtient de la médiane et de la moyenne.

Puisque nous parlons de la forme de la distribution, mentionnons que, dans une distribution symétrique et unimodale, le mode, la médiane et la moyenne affichent tous la même valeur. La valeur qui apparaît le plus souvent dans la distribution est également à la fois la valeur qui la divise en deux parties égales et la moyenne arithmétique. Vous pouvez observer cela dans la distribution symétrique que de la figure 3.4.

Figure 3.4. Distribution symétrique

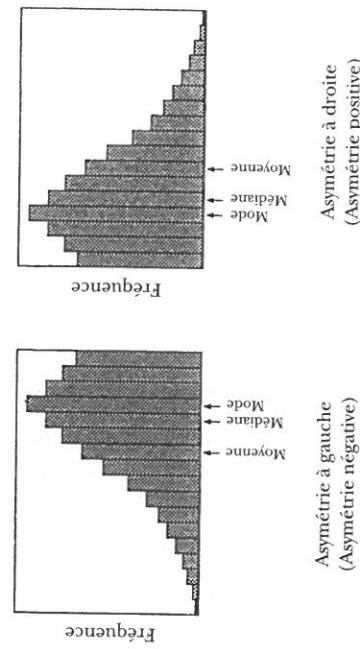


À l'inverse, lorsque la forme de la distribution n'est pas symétrique, les trois mesures de tendance centrale ont des valeurs différentes. La médiane et la moyenne sont affectées à des degrés différents par l'*asymétrie* d'une variable – c'est-à-dire la mesure dans laquelle la distribution d'une variable présente une longue queue s'étendant dans une direction ou dans l'autre. Une distribution peut être asymétrique à gauche à cause de la présence de valeurs normalement faibles. Les scores faibles « tirent », pour ainsi dire, la moyenne vers eux. La moyenne est donc plus petite que la médiane lorsque l'asymétrie de la distribution se trouve à gauche. Inversément, une distribution dans laquelle l'asymétrie est à droite verra sa moyenne plus grande que sa médiane parce que les scores élevés « tireront » la moyenne dans leur direction. La figure 3.5 montre graphiquement l'effet de l'asymétrie sur la moyenne et sur la médiane. Vous pouvez remarquer que, dans une distribution unimodale asymétrique, la médiane se situe toujours entre le mode et la moyenne.

Il faut donc redoubler de prudence lorsqu'une distribution est fortement asymétrique. Comme je l'ai mentionné précédemment, la moyenne ne fournit guère, dans de telles situations, une valeur typique ou représentative. Lorsque que nous avons affaire à des distributions fortement asymétriques, il est généralement préférable d'employer la médiane plutôt que la moyenne, même pour des variables d'intervalle/ratio, et cela parce que la médiane « résiste » beaucoup

mieux aux scores extrêmes. Si vous découvrez des cas déviants, vous avez deux choix : ou bien vous les excluez et utilisez la moyenne, ou bien vous les conservez et utilisez la médiane.

Figure 3.5. Effet de l'asymétrie (à gauche et à droite) sur la moyenne et la médiane



- La somme des écarts par rapport à la moyenne ($(X_i - \bar{X})$) est 0.
 - La moyenne minimise la somme du carré des écarts $\Sigma(X_i - \bar{X})^2$.
 - La moyenne d'une variable dichotomique codée 0 et 1 correspond à la proportion de cas codés 1.
 - La médiane et la moyenne ont la même valeur lorsque les distributions sont symétriques.
 - Le mode, la médiane et la moyenne ont la même valeur dans les distributions unimodales symétriques.
 - Les scores extrêmes (y compris les cas déviants) affectent la moyenne, mais non la médiane ou le mode.
 - L'asymétrie d'une distribution affecte la moyenne, et n'affecte pas la médiane. La médiane est donc préférable à la moyenne dans le cas de distributions hautement asymétriques.
- Les rapports de recherche décrivant les mesures de tendance centrale sont d'habitude accompagnés de descriptions des mesures de variance, le sujet du prochain chapitre. J'attendrai donc la fin du chapitre 4 avant de vous présenter des exemples de rapport pour les mesures de tendance centrale.

3.7 Résumé du chapitre 3

Voici ce que nous avons appris dans ce chapitre :

- Il existe trois types de mesures de tendance centrale : le mode, la médiane et la moyenne.
- Le mode est la valeur qui apparaît le plus fréquemment. Il peut être calculé pour des variables de tous les niveaux. Il est, de fait, la seule mesure de tendance centrale qui convient à des variables nominales.
- Il est possible de distinguer une distribution selon le nombre de modes qu'elle possède – unimodale ou bimodale.
- La médiane est le point milieu d'un ensemble ordonné de scores. La moitié des scores lui sont inférieurs, l'autre moitié supérieure. La médiane peut être calculée pour des variables ordinaires et d'intervalles/ratio.
- La moyenne est une mesure de tendance centrale arithmétique. Il convient de la calculer pour des variables d'intervalles/ratio et parfois pour des variables ordinaires et des variables nominales dichotomiques.

mesure de tendance centrale

mode

distribution unimodale

distribution bimodale

médiane

moyenne

déviation (écart)

somme des carrés

asymétrie

asymétrie négative (à gauche)

asymétrie positive (à droite)

Symboles

Mo

Md

\bar{X}

μ

X^-