CHAPITRE 2

Les distributions de fréquences et de pourcentages

Dans ce chapitre, je décris la façon de résumer des informations concernant des variables prises une par une. En d'autres mots, nous nous attarderons à des situations univariées. Nous examinerons ici les distributions de fréquences et de pourcentages ainsi que les diagrammes circulaires, les diagrammes en bâtons et les projections cartographiques.

Une fois ce chapitre lu, vous pourrez :

1. Transformer des données brutes en distributions de fréquences univariées.
2. Transformer ces distributions de fréquences en distributions de pourcentages.
3. Interpréter les distributions de pourcentages univariées.
4. Comprendre que les pourcentages sont des fréquences standardisées.
5. Produire et expliquer des distributions de pourcentages cumulatifs.
6. Construire des tableaux univariés de pourcentages présentés de façon convenable et les interpréter.
7. Appliquer les règles générales relatives à la fusion des catégories.
8. Reconnaître ce que sont les données manquantes à exclure d'une analyse.
9. Comprendre qu'une analyse peut être basée sur un sous-ensemble de cas.
10. Produire des diagrammes circulaires et des diagrammes en bâtons et les interpréter.
11. Expliquer en quoi les niveaux de mesure affectent les distributions cumulatives et le choix d'un diagramme

12. Expliquer ce qu'est un cas déviant et le reconnaître dans une distribution.
13. Produire et interpréter des cartes décrivant la distribution de variables écologiques.

2.1 Les distributions de fréquences

Une façon simple et directe de résumer des informations concernant une variable est de compter le nombre de cas pour chaque valeur. Ce résumé de la variation d'une variable est une ***distribution de fréquences*** qui attire notre attention sur le nombre de cas correspondant à chaque valeur plutôt qu'aux valeurs elles-mêmes.

Commençons par quelques exemples très simples. D'abord la variable « Désobéissance civile » dont on a parlé au premier chapitre. Je vous ai dit que, dans le General Social Survey des États-Unis, on avait demandé aux répondants si les gens devaient obéir aux lois sans exception ou s'il y avait des situations exceptionnelles où les gens devaient suivre leur conscience même si cela impliquait de violer la loi. Pour des raisons pratiques nous allons considérer que les réponses à cette question constituent la variable « Désobéissance civile ». Le tableau 2.1 rapporte les scores de 50 répondants réels du General Social Survey pour la variable « Désobéissance civile ».

Ces 50 cas constituent un échantillon de l'échantillon du General Social Survey, pour ainsi dire. Pour s'initier aux distributions de fréquences et de pourcentages, il sera beaucoup plus facile d'utiliser seulement ces 50 cas plutôt que les 2904 cas du sondage. Les deux premiers répondants croient qu'il faut suivre sa conscience, le troisième et le quatrième croient qu'il faut toujours obéir aux lois, le cinquième cas croit plutôt qu'il faut suivre sa conscience, et ainsi de suite. Ce sont des ***données brutes***, les scores tels qu'ils se présentent au départ.

Nous pouvons compiler le nombre de cas de chaque score, comme dans le tableau 2.2. Ce décompte est une distribution de fréquences nous montrant combien de répondants ont répondu « obéir aux lois », et combien ont répondu « suivre sa conscience ». L'initiale f qui se trouve au sommet de la colonne de droite est l'abréviation conventionnelle qu'on utilise en statistiques pour signifier « ***fréquence*** ». Nous avons réduit 50 fragments d'information à deux nombres – les décomptes pour chacune des deux valeurs de la variable « Désobéissance civile ». Nous pouvons sans peine lire que 21 répondants croient qu'il faut toujours obéir aux lois et que 29 répondants croient qu'il y a des occasions où les gens doivent suivre leur

Tableau 2.1. Scores de 50 répondants du General Social Survey américain de 1996 pour la variable « Désobéissance civile »

Numéro du cas	Désobéissance civile	Numéro du cas	Désobéissance civile
01	Suivre sa conscience	26	Suivre sa conscience
02	Suivre sa conscience	27	Suivre sa conscience
03	Obéir aux lois	28	Suivre sa conscience
04	Obéir aux lois	29	Obéir aux lois
05	Suivre sa conscience	30	Obéir aux lois
06	Obéir aux lois	31	Obéir aux lois
07	Obéir aux lois	32	Suivre sa conscience
08	Suivre sa conscience	33	Obéir aux lois
09	Suivre sa conscience	34	Obéir aux lois
10	Suivre sa conscience	35	Suivre sa conscience
11	Suivre sa conscience	36	Suivre sa conscience
12	Obéir aux lois	37	Suivre sa conscience
13	Obéir aux lois	38	Obéir aux lois
14	Obéir aux lois	39	Obéir aux lois
15	Obéir aux lois	40	Suivre sa conscience
16	Suivre sa conscience	41	Suivre sa conscience
17	Suivre sa conscience	42	Suivre sa conscience
18	Suivre sa conscience	43	Suivre sa conscience
19	Obéir aux lois	44	Suivre sa conscience
20	Obéir aux lois	45	Obéir aux lois
21	Obéir aux lois	46	Suivre sa conscience
22	Suivre sa conscience	47	Obéir aux lois
23	Suivre sa conscience	48	Suivre sa conscience
24	Suivre sa conscience	49	Obéir aux lois
25	Suivre sa conscience	50	Suivre sa conscience

Tableau 2.2. Décompte de la variable « Désobéissance civile »

Désobéissance civile	Décompte	f
Suivre sa conscience	//// //// //// //// //// //// ////	29
Obéir aux lois	//// //// //// //// ////	21

Il nous faut pouvoir communiquer l'information clairement et de façon succincte. Rien ne sert d'embêter les gens avec notre méthode de décompte. Le tableau 2.3 transforme notre décompte en un tableau de fréquences clair et attrayant.

Tableau 2.3. Opinions concernant la désobéissance civile (en fréquences)

Désobéissance civile	f
Suivre sa conscience	29
Obéir aux lois	21
Total	50

Considérez maintenant un second exemple. Supposez qu'on ait demandé aux répondants de dire le nombre d'années de scolarité qu'ils ont effectuées. Le tableau 2.4 présente les réponses des 50 mêmes répondants au General Social Survey, catégorisées en cinq niveaux d'instruction.

Tableau 2.4. Scores hypothétiques de 50 répondants pour la variable « Niveau d'instruction »

Numéro du cas	Niveau d'instruction	Numéro du cas	Niveau d'instruction
01	Secondaire	26	Universitaire avancé
02	Premier cycle universitaire	27	Pas de secondaire
03	Pas de secondaire	28	Premier cycle universitaire
04	Secondaire	29	Pas de secondaire
05	Secondaire	30	Secondaire
06	Secondaire	31	Pas de secondaire
07	Premier cycle universitaire	32	Secondaire
08	Secondaire	33	Secondaire
09	Secondaire	34	Collège
10	Secondaire	35	Secondaire
11	Universitaire avancé	36	Premier cycle universitaire
12	Secondaire	37	Universitaire avancé
13	Secondaire	38	Secondaire
14	Secondaire	39	Pas de secondaire
15	Pas de secondaire	40	Secondaire
16	Secondaire	41	Premier cycle universitaire
17	Secondaire	42	Secondaire
18	Collège	43	Secondaire
19	Universitaire avancé	44	Pas de secondaire
20	Pas de secondaire	45	Secondaire
21	Secondaire	46	Pas de secondaire
22	Premier cycle universitaire	47	Premier cycle universitaire
23	Collège	48	Premier cycle universitaire
24	Secondaire	49	Secondaire
25	Premier cycle universitaire	50	Pas de secondaire

Le tableau 2.5 nous donne les fréquences. Notez que cette fois j'ai laissé tomber les marques de décompte et que j'ai donné un titre au tableau, comme nous le ferions si nous avions à présenter ces informations à d'autres. Ce tableau de fréquences condense 50 parcelles d'information en seulement 5 nombres (en excluant le total). Nous découvrons que 4 répondants disent avoir fait des études supérieures, 9 ont fait des études de premier cycle à l'université, 3 ont un diplôme de collège, et ainsi de suite[1].

Tableau 2.5. Niveau d'instruction atteint (en fréquences)

Niveau d'instruction	f
Universitaire avancé	4
Premier cycle universitaire	9
Collège	3
Secondaire	24
Pas de secondaire	10
Total	50

Regrouper des cas pour en faire une distribution de fréquences, c'est aussi facile que... un, deux, trois ! Tout ce que vous avez à faire, c'est de compter. Certes, dans la mesure où il ne sera plus possible de connaître le score qu'avait chacun des cas individuels, nous perdons quelques détails que contenaient initialement les données brutes. Toutefois la distribution de fréquences aide grandement à comprendre nos données. Nous savons maintenant combien de cas se rattachent à chacun des scores et ainsi nous avons une meilleure idée de la façon dont sont distribués ces scores.

1. On fait référence ici au système d'éducation américain qui compte généralement cinq niveaux : primaire, secondaire, collège (« Junior College »), premier cycle universitaire (« Undergraduate Degree ») et universitaire avancé (« Graduate Degree »). Cette structure peut varier d'un État à l'autre et même d'une localité à l'autre (N.D.T.).

2.2 Les distributions de pourcentages

Si les distributions de fréquences s'avèrent utiles pour résumer l'information, elles sont parfois difficiles à interpréter. Une distribution de fréquences d'un grand nombre de cas produit souvent des fréquences très élevées qui peuvent être malaisées à comprendre. En effet nous pouvons assez facilement manipuler des fréquences peu élevées comme 4, 9 et 3. Mais lorsqu'il s'agit de centaines, de milliers ou de centaines de milliers, la plupart d'entre nous éprouvons des difficultés. Mais là n'est pas le seul problème.

En effet, la comparaison entre plusieurs distributions de fréquences ou plus se complique si ces distributions sont basées sur des nombres de cas différents. Il est difficile de comparer, par exemple, la distribution de fréquences du niveau d'instruction des répondants du Texas avec celle du Vermont puisque ces États ont une population de taille différente. Le recensement américain rapporte que 2 204 099 habitants du Texas ont un diplôme de collège, alors que c'est le cas de 91 522 habitants du Vermont. Que pouvons-nous cependant conclure à propos du niveau d'instruction relatif des Texans et des habitants du Vermont ? Après tout, le Texas a une population adulte supérieure en nombre à celle du Vermont – 16 986 335 contre seulement 562 758. Les Texans sont-ils plus susceptibles d'avoir un diplôme de collège ? Seulement à partir des fréquences, cela est bien difficile à déterminer.

Nous pouvons surmonter ces problèmes si nous standardisons le résumé de notre distribution en calculant quelle serait chacune des fréquences si le nombre total de cas était exactement 100. On appelle « ***pourcentage*** » le résultat de cette standardisation. Nous sommes en fait si familiers avec les pourcentages que rarement nous songeons à ce qu'ils sont réellement : les pourcentages sont ce que seraient les fréquences s'il y avait 100 cas au total. Convertir en pourcentage réduit les grands nombres à des nombres plus aisément manipulables (c'est-à-dire des pourcentages) qui vont de 0 à 100. En standardisant chaque fréquence selon la même base – 100 – nous pouvons facilement comparer les pourcentages. Nous pouvons par exemple découvrir qu'il y a relativement plus de diplômés de collège au Vermont qu'au Texas : 16 % des adultes du Vermont comparativement à 13 % des adultes du Texas.

Je sais, vous avez appris les pourcentages à l'école primaire. Voici quand même un rappel de la façon de procéder, juste pour vous rafraîchir la mémoire :

1. Diviser chaque fréquence par le nombre total de cas.
2. Multiplier ce résultat par 100.

Sous forme de formule :

$$\text{pourcentage} = \frac{f}{N}(100)$$

lorsque f = fréquence

N = nombre total de cas

Par exemple, des 50 cas mentionnés dans la section précédente, 29 croient qu'il faut suivre sa conscience même si cela implique de violer la loi, ou $\frac{f}{N}(100) = \frac{29}{50}(100) = 58\,\%$. De la même façon, 21 répondants considèrent qu'il faut toujours obéir aux lois, $\frac{21}{50}(100) = 42\,\%$. L'addition d'un ensemble de pourcentages doit toujours donner un total de 100 (sauf dans les cas où, en arrondissant les pourcentages, leur somme est légèrement au-dessus ou au-dessous de 100).

Quand on se contente de diviser une fréquence par N (sans multiplier le quotient par 100), nous obtenons une ***proportion***. Donc la proportion des cas qui ont répondu qu'il fallait suivre sa conscience est de 0,58 et la proportion de ceux qui ont répondu qu'il fallait toujours obéir aux lois est de 0,42. Tout comme les pourcentages varient de 0 à 100, les proportions varient de 0 à 1,00. Et tout comme la somme d'un ensemble de pourcentages doit toujours égaler 100, la somme d'un ensemble de proportions doit toujours égaler 1,00 (lorsqu'il n'y a pas d'erreur due à l'arrondissement). Les pourcentages ne sont que des proportions multipliées par 100. Pourcentages et proportions donnent la même information, mais sur une base différente (100 et 1, respectivement). Les conventions et les préférences en ce qui concerne l'utilisation soit de pourcentages soit de proportions varient d'une discipline à l'autre, et même d'une sous-discipline à l'autre. De façon générale les spécialistes des sciences sociales utilisent les pourcentages et c'est ce que nous ferons dans ce livre.

Le tableau 2.6 est un ***tableau de pourcentages*** donnant la distribution de nos 50 réponses à la question portant sur la désobéissance civile. Nous observons qu'un plus grand pourcentage des répondants approuvent la désobéissance civile et qu'un plus faible pourcentage croient qu'il faut toujours obéir aux lois (58 % contre 42 %).

Tableau 2.6. Opinions concernant la désobéissance civile (en pourcentages)

Désobéissance civile	Pourcentages
Suivre sa conscience	58
Obéir aux lois	42
Total	100
(N)	(50)

La lettre N, qui se trouve entre parenthèses au bas du tableau, donne le nombre total de cas sur lequel se basent les pourcentages. En statistique, la lettre majuscule « N » correspond habituellement au nombre total de cas.

Si nous exprimons en pourcentages la distribution de fréquences du niveau d'instruction, nous obtenons le tableau 2.7. Celui-ci montre que près d'un cinquième (18 %) des répondants au sondage ont un diplôme de collège. La moitié (48 %) des répondants sont des diplômés du secondaire qui n'ont jamais fréquenté le collège ni l'université. Un cinquième (20 %) ont un niveau d'instruction inférieur au secondaire.

Tableau 2.7. Niveau d'instruction atteint (en pourcentages)

Niveau d'instruction	Pourcentages
Universitaire avancé	8
Premier cycle universitaire	18
Collège	6
Secondaire	48
Pas de secondaire	20
Total	100
(N)	(50)

Un mot d'avertissement concernant les pourcentages : soyez prudents lorsque vous convertissez des fréquences en pourcentages et que le nombre total de cas est faible. Les pourcentages seront « instables » si N est peu élevé, si bien que nous ne pouvons leur faire confiance. Le déplacement d'un cas d'une valeur à une autre produira alors des changements considérables dans les pourcentages. Quand N égale 50, chaque cas représente 2 points de pourcentage. Dans les tableaux 2.6 et 2.7, par exemple, le déplacement d'un cas d'une valeur à une autre fera perdre à l'une des valeurs 2 points de pourcentage et augmentera l'autre d'autant. Les statisticiens divergent dans leur définition de ce qui constitue un petit N. Certains mettent en garde contre l'usage de pourcentages lorsque le nombre total de cas est inférieur à environ 30. D'autres considèrent plutôt des N de 50 et même de 100 comme des cas limites en dessous desquels il ne faut guère descendre. Tous cependant s'entendent pour nous exhorter à ne pas nous fier à des pourcentages qui reposeraient sur un faible nombre de cas. (Je reconnais qu'en voulant simplifier les choses, j'ai choisi un exemple bien près de la limite et qui la dépasse même, peut-être. Je promets que je ne recommencerai pas trop souvent.)

23 Les distributions cumulatives

Un pourcentage cumulatif est le pourcentage de tous les scores égaux ou inférieurs à une valeur donnée. Pour obtenir un pourcentage cumulatif :

1. Additionnez toutes les fréquences qui ont la valeur donnée ou qui sont de valeur inférieure.
2. Divisez cette somme par le nombre total de cas.
3. Multipliez ce résultat par 100.

Exprimé en formule :

$$\text{Pourcentage cumulatif} = \frac{F}{N}(100)$$

lorsque F = fréquence cumulative (c'est-à-dire la somme des fréquences inférieures ou égales à une valeur donnée)

N = nombre total de cas

De façon équivalente (bien qu'il y ait parfois des erreurs dues à l'arrondissement), nous pouvons trouver le pourcentage cumulatif en additionnant les pourcentages de la valeur donnée et des valeurs qui lui sont inférieures. Bien que l'initiale F majuscule soit utilisée par les statisticiens pour renvoyer à des choses différentes, nous l'utilisons ici pour représenter les fréquences cumulatives, c'est-à-dire la somme de toutes les fréquences inférieures ou égales à une valeur donnée.

Le tableau 2.8 est le tableau de pourcentages cumulatifs qui correspond à nos données hypothétiques sur le niveau d'instruction. Ainsi à partir des fréquences provenant du tableau 2.5, le pourcentage

cumulatif des cas qui ont un niveau d'instruction secondaire ou moins nous est donné comme suit :

$$\text{Pourcentage cumulatif} = \frac{F}{N}(100)$$

$$= \frac{10+24}{50}(100)$$

$$= \frac{34}{50}(100)$$

$$= 68$$

On peut calculer le pourcentage cumulatif des cas qui ont un diplôme de collège ou moins de la façon suivante :

$$= \frac{10+24+3}{50}(100)$$

$$= \frac{37}{50}(100)$$

$$= 74$$

Tableau 2.8. Niveau d'instruction atteint (en pourcentages cumulatifs)

Niveau d'instruction	Pourcentages cumulatifs
Universitaire avancé	100
Premier cycle universitaire	92
Collège	74
Secondaire	68
Pas de secondaire	20
(N)	(50)

Nous pouvons également obtenir ce pourcentage cumulatif en additionnant 20 (le pourcentage des répondants qui ont un niveau d'instruction inférieur au secondaire), 48 (le pourcentage des répondants de niveau secondaire) et 6 (le pourcentage de diplômés de collège). Ainsi, 20 + 48 + 6 = 74. Notez aussi que chacun des pourcentages cumulatifs correspond à la somme des pourcentages cumulatifs des valeurs précédentes plus le pourcentage de la valeur donnée (68 + 6 = 74 pour les diplômés de collège ou des niveaux inférieurs). Le pourcentage cumulatif de la valeur la plus élevée est toujours 100 % puisque que tous les cas doivent avoir ou bien cette valeur ou bien une valeur inférieure.

Nous pouvons imaginer sans peine l'utilité des distributions cumulatives. Elles répondent à des questions comme « Quel pourcentage des répondants ont 40 ans ou moins ? », « Quel est le pourcentage des gens qui ont des revenus de 25 000 $ ou moins ? », ou « Quel pourcentage des répondants regardent la télévision trois heures ou moins par jour[2] ? » De façon assez évidente les distributions cumulatives de pourcentages ne sont pas utiles pour les variables dichotomiques, comme, par exemple, la variable sur la désobéissance civile. Il faut noter aussi qu'une distribution cumulative n'est pertinente que dans la mesure où les valeurs d'une variable peuvent être ordonnées. Donc les fréquences et les pourcentages cumulatifs n'ont de signification de prime abord que pour des variables ordinales ou d'intervalles/ratio. En ce qui a trait aux variables nominales telles la religion ou la région géographique, il est rarement sensé de se servir d'une distribution cumulative parce que les valeurs n'ont pas entre elles d'ordre véritable. Nous ne pouvons pas dire « catholique ou moins » ou « provinces de l'Ouest ou moins ».

Après avoir fait vos premières armes dans l'analyse de données, vous verrez comment il est possible parfois d'utiliser les distributions cumulatives pour des variables nominales. Considérez, par exemple, la variable « statut marital », mesurée par les catégories « marié », « veuf », « divorcé », « séparé » et « jamais marié » (dont les codes sont les nombres de 1 à 5 respectivement). Nous pouvons nous servir d'une distribution cumulative pour découvrir le nombre ou le pourcentage des cas dont les scores se situent dans les quatre valeurs « les plus faibles » – c'est-à-dire « marié », « veuf », « divorcé » et « séparé ». Ces quatre valeurs constituent l'ensemble des cas « déjà mariés ». Le pourcentage cumulatif des répondants se disant séparés nous renseigne en fait sur le pourcentage des répondants qui furent déjà mariés, en opposition à ceux qui ne le furent jamais.

2. Les distributions cumulatives dont nous discutons ici sont toutes des distributions cumulatives de type « ou moins ». Les distributions cumulatives de type « et plus » indiquent la fréquence ou le pourcentage de scores qui ont une valeur donnée ou qui ont une valeur supérieure.

2.4 Produire des tableaux lisibles et bien présentés

Assurez-vous, dans vos tableaux de pourcentages – et, en fait, dans toutes vos analyses –, de ne pas prétendre à plus de précision que ne vous en fournissent vos données. Ne retenez que les décimales significatives – c'est-à-dire les décimales qui sont fiables et en lesquelles vous avez confiance. Pratiquement, cela signifie que vous devez habituellement arrondir les pourcentages soit au dernier nombre entier, soit à la première décimale. Cette règle informelle souffre bien sûr des exceptions, qui sont néanmoins assez peu nombreuses lorsque vous travaillez en sciences sociales avec des données scientifiques. Les pourcentages avec plus d'une décimale prétendent souvent à une fausse précision. Les calculatrices et les ordinateurs nous fournissent généralement un grand nombre de chiffres après la virgule (souvent jusqu'à huit chiffres). Ces décimales sont la plupart du temps non significatives.

À titre de règle générale en statistiques, il est préférable de conserver le nombre maximal de décimales que vous pouvez lorsque vous calculez ou lorsque vous utilisez des formules, etc., de façon à minimiser les erreurs qui pourraient être dues à l'arrondissement lors des calculs. Arrondissez ensuite votre résultat final en gardant le nombre de décimales que vous aviez initialement plus une. Si vos calculs portaient au départ sur des nombres entiers (comme pour des compilations de cas), arrondissez vos résultats (des pourcentages par exemple) à la première décimale. Parce qu'il s'agit là d'une règle bien générale, les exceptions sont tolérées. Toutefois, comme toujours, *pensez* à ce que vous êtes en train de faire et décidez jusqu'à quelle décimale vous avez confiance en vos résultats.

Les décimales superflues (celles qui habituellement se situent au-delà du premier chiffre après la virgule) devraient normalement être arrondies. Quelques exemples : arrondissez 21,32 à 21,3 ; arrondissez 62,81 à 62,8 ; arrondissez 15,66 à 15,7. Mais que faire pour arrondir un nombre se terminant par 5 comme 48,65 ou 17,35 ? Un usage courant consiste à arrondir les « 5 » au nombre *pair* le plus proche. Ainsi, par exemple, 48,65 sera arrondi à 48,6, alors que 17,35 sera arrondi à 17,4. C'est cette « règle du nombre pair » que je suivrai dans ce livre. Elle nous assure qu'à long terme environ la moitié des nombres se terminant par 5 auront été arrondis à l'inférieur, l'autre moitié au supérieur.

Quelques mots à propos de la forme des tableaux. Dans la section précédente, j'ai construit mes tableaux de manière aussi présentable que possible. Bien qu'il n'y ait pas un modèle « officiel » ou « convenable » concernant les tableaux de pourcentages, la forme des tableaux présentés dans ce livre respecte le format d'édition de l'*American Sociological Review* et des autres publications de l'American Sociological Association. Dans d'autres disciplines – la psychologie, la science politique, les sciences de l'éducation, etc. – on préfère des formats légèrement différents. Il y a même des variantes à l'intérieur des disciplines. Consultez un exemplaire récent d'une publication scientifique importante de votre discipline pour des exemples concernant la forme des tableaux ainsi que des exemples de tableaux présentés convenablement.

Voici quelques conseils d'ordre général qui vous permettront de construire des tableaux univariés de pourcentages bien présentés.

- Numérotez vos tableaux en chiffres arabes si vous en présentez plus d'un.
- Choisissez un titre direct qui énonce clairement mais succinctement les variables qui sont décrites dans le tableau. Indiquez aussi la source des données, à moins que cette source ne soit mentionnée dans le texte accompagnant le tableau.
- Intitulez la colonne de gauche du nom de la variable (Désobéissance civile, Instruction). Utilisez pour cela des noms clairs, descriptifs, plutôt que les noms abrégés et codés dont on fait usage dans les fichiers de données.
- Intitulez la colonne de droite : « Pourcentages » (ou « Fréquences », ou « Pourcentages cumulatifs », selon le cas).
- Assurez-vous que les catégories sont mutuellement exclusives et collectivement exhaustives (comme c'est décrit dans la section 1.6). Chacun des scores doit être compris dans une et une seule catégorie de valeurs.
- Ajoutez une rangée « Total » qui compile tous les pourcentages. Celle-ci guide la compréhension du lecteur.
- Ajoutez aussi une rangée (N) présentant le nombre de cas à partir desquels furent calculés les pourcentages. (Cette rangée est parfois intitulée « Nombre de cas ».) Elle permet au lecteur d'évaluer la stabilité des pourcentages et de calculer les fréquences individuelles sur lesquelles le calcul des pourcentages est basé.
- Soyez conséquents en ce qui concerne les décimales. Par exemple, n'arrondissez par certains pourcentages au nombre entier et d'autres à un chiffre après la virgule.
- À moins que vous ayez de bonnes raisons d'attirer l'attention du lecteur sur les fréquences, n'inscrivez pas les fréquences individuelles dans un tableau – seulement les pourcentages. Un

lecteur curieux pourra recalculer les fréquences en multipliant N par le pourcentage et en divisant le produit par 100.

- Disposez les pourcentages à droite, alignés les uns sur les autres. (La plupart des logiciels de traitement de texte vous permettent d'aligner ensemble les virgules décimales, ce qui est encore mieux.)
- Ne conservez pas les symboles % après chacun des pourcentages. Ils sont superflus, ils encombrent le tableau et dénotent un mauvais goût.
- Ne tracez pas de lignes verticales dans un tableau. Elles aussi s'avèrent encombrantes. Comme unique guide pour l'œil du lecteur et comme gage de clarté pour votre tableau, ne tracez qu'une ligne double horizontale entre le titre et les en-têtes de colonnes ainsi que de simples lignes horizontales sous ces en-têtes et au bas du tableau, comme je le fais dans ce livre.
- Soyez très propres. Alignez rigoureusement les entrées et les virgules, tracez des lignes horizontales de longueur identique, etc.

Si le style ne doit jamais empiéter sur la substance, vous devez cependant les deux aux lecteurs de vos tableaux. Comment savoir si vos tableaux sont présentés convenablement ? Une personne raisonnablement intelligente devrait être capable de les lire sans ambiguïté, avec un effort minimal. Alors demandez-vous en préparant vos tableaux : mon copain ou ma copine pourrait-il lire ce tableau correctement avec un effort minimal ? (Attention : ce test n'est valide que si votre copain ou votre copine est raisonnablement intelligent ou intelligente.)

Je crois que les tableaux que nous avons vus précédemment étaient bien présentés, bien que je ne prétende pas qu'ils fussent parfaits. (Rien n'est parfait, hormis bien sûr la musique de Mozart et le ragoût de maman.) Le tableau 2.7, par exemple, a probablement trop de catégories. Peut-être y a-t-il trop peu de diplômés de collège ou de programmes universitaires avancés pour justifier l'existence de catégories séparées. Peut-être aurions-nous dû combiner les répondants qui ont obtenu un diplôme de collège et ceux qui ont terminé des études universitaires en une seule catégorie, autrement dit, fusionner trois catégories de la variable en une seule.

2.5 Fusionner des catégories

Quelques variables possèdent tant de valeurs qu'il peut s'avérer très utile de fusionner certaines catégories de valeurs afin d'en obtenir un nombre se prêtant mieux à des tableaux de fréquences et de pourcentages. La variable du GSS décrivant l'âge des répondants a, par exemple, plus de 70 valeurs comprises entre 18 et 89. Plusieurs des valeurs n'ont que peu de cas (il n'y a que 5 répondants âgés de 85 ans). Il est difficile, lorsque l'on tente de résumer la distribution de cette variable, de déceler une cohérence parmi plus de 70 catégories. Avec un si grand nombre de valeurs, plus d'une page est nécessaire si l'on veut imprimer la distribution des pourcentages de la variable « âge ». Un des buts principaux que se proposent les statistiques descriptives est de condenser l'information afin de la rendre plus compréhensible. Il serait douteux de dire qu'un tableau de pourcentages s'étalant sur plusieurs pages serve cette cause. Aussi longtemps que nous pourrons nous passer de détails précis, la cohérence d'une distribution de l'âge réduite à, disons, cinq ou six catégories nous semblera plus évidente que celle d'une distribution de l'âge mesuré en 70 valeurs. Selon nos objectifs de recherche, il pourrait être utile de créer les catégories suivantes pour la variable « Âge » du GSS :

18-29
30-39
40-49
50-64
65 et plus

Prenez note que ces catégories sont mutuellement exclusives et collectivement exhaustives, et ce n'est pas par hasard !

Les variables d'intervalles/ratio ont souvent de nombreuses valeurs et ces valeurs, par conséquent, doivent être fusionnées. Il en va également de même de plusieurs variables nominales et ordinales. Par exemple, une variable indiquant l'État dans lequel réside le répondant (une variable nominale donc) est souvent d'une utilité plus grande lorsqu'elle est réduite aux huit ou neuf régions géographiques les plus importantes, comme la Nouvelle-Angleterre, le Midwest, et ainsi de suite. Même ces catégories assez larges s'avéreront dans certains cas plus utiles si elles sont fusionnées davantage, pour obtenir des catégories encore plus grandes telles le Nord, le Sud et l'Ouest. Certaines situations, bien sûr, commandent que nous conservions les détails que nous fournit un nombre plus élevé de valeurs. Mais il est souvent profitable de combiner ou de fusionner des valeurs, même s'il en résulte la perte de certaines informations.

Nous pouvons également fusionner des catégories dans lesquelles se rangent très peu de cas dans le but d'obtenir des fréquences plus élevées. Par exemple, quand on leur demande s'ils trouvent la vie stimulante, routinière ou morne, la plupart des répondants au GSS affirment que la vie de tous les jours est soit stimulante, soit plutôt routinière. Relativement peu d'entre eux – moins de 5 % –

disent de la vie qu'elle est morne. Peut-être voudrions-nous alors fondre la catégorie « morne » et la catégorie « plutôt routinière » en une seule catégorie « routinière/morne ».

Les chercheurs fusionnent souvent les catégories qui fournissent plus de détails que n'en réclame l'analyse. Pour mesurer les attitudes politiques, on a demandé aux répondants du GSS de se situer eux-mêmes sur une échelle comportant trois degrés de libéralisme, trois degrés de conservatisme et un point central étiqueté « Modéré ». Cette échelle peut être très utile pour certaines techniques statistiques ; mais pour les tableaux de pourcentages il vaudrait peut-être mieux combiner les trois catégories de libéralisme d'une part et les trois catégories de conservatisme, d'autre part, pour obtenir trois catégories : libéral, modéré, conservateur.

Il existe donc plusieurs bonnes raisons de regrouper certaines valeurs : pour obtenir un nombre de catégories qui se prête mieux à l'analyse, pour se débarrasser de détails superflus, et pour éviter de se retrouver avec des catégories dans lesquelles il n'y a que peu de cas. Mais comment décider de la façon de fusionner les catégories d'une variable ? Combien de catégories doit-on conserver ou créer ? Lesquelles ? Voilà des questions auxquelles on ne peut guère répondre définitivement. Des gens raisonnables et sensés pourront très bien être en désaccord sur la meilleure façon de regrouper les catégories d'une variable. Voici cependant quelques conseils, exposés *grosso modo* en ordre de préséance, du plus important au moins important, qu'il est bon de garder à l'esprit :

1. Fusionnez les valeurs d'une variable de façon à ce que cela soit cohérent à l'aune de vos objectifs ou de votre question de recherche. Que voulez-vous tirer de ces données ? Par exemple, si votre recherche porte sur les différences entre les chrétiens et les non-chrétiens, il peut sembler tout à fait sensé de combiner en une seule catégorie les catholiques et les protestants. Mais confondre ainsi catholiques et protestants n'a aucun sens si vous voulez comparer les protestants et les catholiques.
2. Créez des catégories qui sont collectivement exhaustives et mutuellement exclusives. En d'autres mots, assurez-vous que chacun des scores possibles ne figure que dans une seule des catégories fusionnées. Il existe cependant une exception (importante) à cette règle de l'exhaustivité collective : vous pouvez exclure des données manquantes telles que Ne sais pas, Pas d'opinion, Refus de répondre. Nous nous attarderons plus longuement sur ce point dans la prochaine section.
3. Regroupez les catégories d'une variable de façon à ne pas eclipser les structures importantes de la distribution. Ne conservez toutefois pas non plus un nombre trop élevé de catégories qui rendraient la distribution malaisée à comprendre. À moins que vous ne vouliez examiner des détails à l'intérieur d'une distribution, tentez en général de travailler avec seulement six ou sept catégories – avec un nombre moindre encore si cela ne cache pas d'informations essentielles.
4. Conservez des catégories homogènes et « culturellement intelligibles ». Choisissez autant que possible des catégories « naturelles ». Dans le cas de la variable « âge » dont nous nous sommes servi précédemment en guise d'exemple, il semble sensé d'avoir une catégorie « 65 et plus » puisque cela correspond plus ou moins, en Occident, au groupe des retraités. De même, dans le cas d'une variable « allégeance politique » aux États-Unis, dont les catégories seraient « fortement Républicain », « Républicain », « modérément Républicain » et les catégories homologues pour les Démocrates, il paraît cohérent de réduire les trois valeurs « Républicaines » en une valeur « Républicain » et les trois valeurs « Démocrates » en une seule valeur « Démocrate ».
5. Dans la limite des règles précédentes et pour autant que cela soit raisonnable, efforcez-vous de garder la même grandeur pour les catégories des variables d'intervalles/ratio. Le revenu, par exemple, peut être segmenté en catégories de 20 000 $ chacune.
6. Sans violer les règles précédentes, respectez les conventions culturelles en créant, dans le cas des variables d'intervalles/ratio, des catégories en base 5 ou en base 10. Des catégories pour la variable revenu qui respecteraient ces conventions seraient, par exemple, 0 à 19 999 $; 20 000 à 39 999 $; 40 000 à 59 999 $... Il paraîtrait un peu idiot d'user de catégories aussi bizarres que 0 à 14 332 $, 14 333 à 28 665 $, 28 666 à 42 997 $... Nous ne pensons pas, dans notre culture (ni dans les autres cultures que je connais d'ailleurs), en fonction de ce type de catégorie mathématique. Bien sûr, dans le cas de variables qui, comme le revenu, ont quelques valeurs extrêmes, vous devrez placer une catégorie « ouverte » à une des extrémités (ou aux deux extrémités) de la distribution dans le but d'inclure ces cas extrêmes. La variable « revenu », par exemple, pourrait contenir une catégorie telle « 100 000 $ et plus » qui embrasserait les revenus élevés qui se comptent en millions.
7. Dans le cas où cela n'entrerait pas en conflit avec ce qui est dit plus haut, créez des catégories contenant un nombre sensiblement égal de cas si vous avez l'intention de procéder à des

analyses de tableau bivarié ou multivarié. (Cela permet d'obtenir un nombre suffisant de cas pour que soit possible la mise en pourcentage, un point dont nous reparlerons abondamment au chapitre 5.)

Que devez-vous donc faire lorsque vous regroupez les valeurs d'une variable ? Réfléchir, voilà ce qu'il faut faire. En fait, réfléchissez sérieusement ! Quelles catégories semblent cohérentes ? Quelles catégories se prêtent le mieux à votre analyse ? Quelles catégories conservent le plus d'information tout en rendant les données plus facilement manipulables ? Pensez-y sérieusement.

Par exemple, songez à la variable du General Social Survey qui indique les années de scolarité des répondants. Cette variable a 21 valeurs, de 0 (pas de scolarité) à 8 ans d'université (c'est-à-dire 20 années de scolarité). Comment regrouper les valeurs de cette variable ? Nous pourrions créer les catégories 0-6 ans, 7-13 ans, 14-20 ans de scolarité. Il s'agirait certes de catégories d'égale grandeur. Le problème réside dans le fait qu'une classification en trois catégories comme celle-ci n'appréhende pas très bien la distribution de la variable et ne respecte guère la signification culturelle et sociale du niveau de scolarité aux États-Unis. D'abord, cela ne tient pas compte de la très grande proportion des Américains titulaires d'un diplôme d'études secondaires qui n'ont cependant jamais fréquenté le collège. Cela englobe aussi dans une même catégorie les décrocheurs de la 7e année et ceux qui ont terminé une année universitaire. Ceux qui ont décroché au collège sont inclus avec ceux qui viennent de terminer leur première année d'études avancées. Ces catégories forment un véritable salmigondis qui ne peut guère nous éclairer.

Il est beaucoup mieux, je crois, de regrouper de façon suivante les années de scolarité : 0-11 ans, 12 ans, 13-15 ans, 16 ans, 17-20 ans. Ainsi nous obtenons des catégories significatives qui distinguent entre les répondants qui n'ont pas terminé leurs études secondaires, ceux qui sont diplômés du secondaire, ceux qui ont étudié au collège, ceux qui sont titulaires d'un diplôme d'études universitaires de premier cycle et ceux qui ont fait des études avancées. Selon la question de recherche posée ou la méthode d'analyse employée, il existe certes d'autres manières également bonnes de regrouper les années de scolarité. Ce qui est important, une fois de plus, c'est de réfléchir à ce que vous voulez trouver et à ce qui servira le mieux votre analyse de données.

26 L'exclusion de données manquantes

Nous avons vu à la section 1.8 qu'il arrive, dans les sondages, que des questions ne s'appliquent pas à certains répondants. Il n'y a pas lieu, par exemple, de demander s'ils sont heureux en mariage aux répondants qui ne sont pas mariés. Nous avons vu aussi que, dans le General Social Survey, on pose certaines questions à un sous-ensemble des répondants de telle sorte que toutes les questions ne sont pas posées à tous les répondants. Cette technique permet d'inclure plus de questions dans un sondage, bien qu'elle limite le nombre de répondants à certaines questions. Le code 9999 est utilisé comme score pour les cas où une question n'est pas applicable ou bien où l'information n'a pas été colligée. Il arrive aussi que les répondants choisis pour un sondage ne veuillent pas répondre ou ne soient pas en mesure de répondre à certaines questions, ou encore n'aient aucune opinion sur la question posée. Ces réponses sont notées « Ne sais pas », « Pas de réponse », « Refus de répondre » et « Pas d'opinion ». De plus, il est possible que, pour certaines variables et pour certains cas d'une banque de données agrégées, l'information ne soit pas disponible. Par exemple, il est possible que les informations concernant les dépenses publiques dans le domaine de la défense ne soient pas disponibles pour la Russie ou pour l'Iran. La donnée est alors « Non disponible » ou « Incertaine ».

La majorité des chercheurs excluent de leur analyse les valeurs telles que « Pas de réponse », « Pas d'opinion », et « Non disponible » dans la mesure où elles ne fournissent aucune information utile concernant les scores d'un cas. Il y a bien peu de raisons qui justifient de les inclure dans le calcul des pourcentages, par exemple. Ce type de valeur que nous devons rejeter au moment de l'analyse est appelé ***données manquantes***. Il y a plusieurs valeurs que nous traitons habituellement (pas toujours cependant) comme des données manquantes : Ne sais pas, Pas d'opinion, Pas de réponse, Refus de répondre, Non applicable et Incertain.

Parfois, pas toujours cependant, la valeur « Autre », dans la mesure où elle se présente souvent comme une catégorie résiduelle regroupant un ensemble bigarré de cas qui ont peu en commun, est aussi considérée comme une donnée manquante. Dans le cas d'une variable comme la religion, la catégorie « Autre », par exemple, comprend des religions aussi diverses que l'islam, l'hindouisme et le bouddhisme. La décision d'inclure ou d'exclure cette catégorie dans une analyse dépend de la question de recherche qui est posée. C'est la même chose pour des catégories comme « Cela dépend » ou « Ne peut choisir ». Selon les objectifs de la recherche, de telles catégories

peuvent être utiles car elles peuvent correspondre à une opinion intermédiaire entre deux extrêmes ou bien elles peuvent être inutiles parce que l'information qu'elles fournissent n'est pas utile pour la recherche. Par exemple, dans la question du General Social Survey concernant l'avortement on demande au répondant s'il pense que l'avortement devrait être permis à toute femme qui le désire, quelle que soit la raison. On pourrait penser que la réponse « Ne sait pas » exprime une incertitude qui se situerait entre le « Oui » et le « Non ». Dans ce cas, nous ne considérerions pas la réponse « Ne sait pas » comme une donnée manquante et nous ne l'exclurions pas de l'analyse. Par contre, nous continuerions d'exclure les « Refus de répondre ». La réponse « Pas d'opinion » est ambiguë elle aussi et peut parfois être incluse dans une analyse pour signifier une position intermédiaire concernant un enjeu. Les décisions que vous prenez concernant l'inclusion ou l'exclusion de valeurs comme « Ne sait pas » ou « Pas d'opinion » dépendent de votre question de recherche et de l'interprétation sémantique que vous faites des réponses à une question. Il n'y a pas de recette miracle. Comme toujours, réfléchissez sérieusement lorsque ce choix se présente.

Rejeter une donnée manquante n'est pas indifférent puisque cela change le nombre total de cas – N – à partir duquel on calcule le pourcentage. Le tableau 2.9, par exemple, montre la distribution d'une variable du General Social Survey mesurant l'attitude envers l'avortement selon que l'on conserve ou que l'on exclut les catégories des données manquantes. Bien que la distance relative entre les pourcentages reste la même, les pourcentages, eux, augmentent lorsque sont exclues les données manquantes. Le pourcentage de ceux et celles affirmant que l'avortement doit être légal en toutes circonstances augmente de 42,6 à 45,3. Le pourcentage de ceux qui préconisent l'inverse passe de 51,3 à 54,7.

Idéalement il ne devrait y avoir aucune donnée manquante. Nous préférons que chaque cas affiche une information pour chacune des variables de l'analyse. Or cette situation est rarement celle dans laquelle nous nous trouvons. À l'instar des chaussettes orphelines dans la lessive, les données manquantes sont une preuve de plus que nous vivons dans un monde qui est loin d'être parfait.

Tableau 2.9. Attitude envers la légalisation de l'avortement, selon que les valeurs manquantes sont incluses ou exclues (en pourcentages)

L'avortement devrait-il être légal ?	Données manquantes incluses	Données manquantes exclues
Oui	42,6	45,3
Non	51,3	54,7
Ne sait pas	5,5	—
Pas de réponse	0,6	—
Total	100,0	100,0
(N)	(1075)	(1010)

2.7 Les sous-ensembles de cas

Il nous paraît souvent utile de restreindre une analyse à certains groupes de cas uniquement. Si nous nous intéressons au niveau d'instruction atteint par les Américains par exemple, nous préférerons peut-être restreindre notre analyse, dans la mesure où une proportion importante d'Américains dans le début de la vingtaine fréquentent encore l'école, aux seuls répondants qui ont plus de 25 ans. Ou si nous étudions les clivages de classes parmi les Afro-Américains, nous pouvons décider de limiter notre analyse uniquement aux Afro-Américains. Si nous nous intéressons à la variation, d'un pays à l'autre, des revenus provenant de l'industrie légale du jeu, il est possible que nous voulions circonscrire l'analyse aux seuls pays qui ont légalisé le jeu. Un ***sous-ensemble*** est un ensemble de cas choisis pour une analyse en fonction du score qu'ils affichent à certaines variables particulières. Des exemples de sous-ensembles : les répondants qui ont plus de 25 ans, les répondants afro-américains, les pays où le jeu a été légalisé.

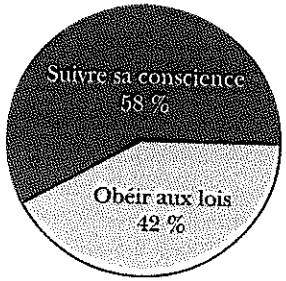
2.8 Les diagrammes circulaires et les diagrammes en bâtons

La plupart des recherches réitèrent le lieu commun selon lequel les êtres humains se saisissent plus rapidement de l'information et la retiennent plus longtemps lorsqu'elle est présentée visuellement, par des graphiques ou des diagrammes. Cette préférence pour l'information visuelle est probablement à la fois biologique et culturelle. Même des personnes que les statistiques laissent froides apprécient généralement les présentations à l'aide de diagrammes. Songez au

succès de certains magazines d'information et de leurs diagrammes tape-à-l'œil.

Il existe de nombreux types de diagrammes, les deux plus courants qui conviennent à l'analyse univariée étant les ***diagrammes circulaires*** et les ***diagrammes en bâtons***. La figure 2.1 montre un diagramme circulaire qui illustre la distribution de la variable « Désobéissance civile » parmi nos 50 répondants. Lire un graphique comme celui-ci, c'est de la tarte. (Désolé, je n'ai pu résister.) Le cercle complet représente le nombre total de répondants. Les tranches peuvent représenter soit des pourcentages, soit des fréquences. Dans les deux cas, la taille de chacune des pointes de la tarte est proportionnelle au pourcentage ou à la fréquence d'une valeur donnée. Plus grand est le pourcentage ou le nombre de cas d'une valeur précise, plus grande sera la portion du cercle. Parce que 58 % des répondants croient que les gens devraient suivre leur conscience, la tranche se rapportant à cette valeur occupera 58 % de la tarte. Les 42 % des cas qui croient que les gens devraient toujours obéir aux lois accaparent 42 % de la tarte.

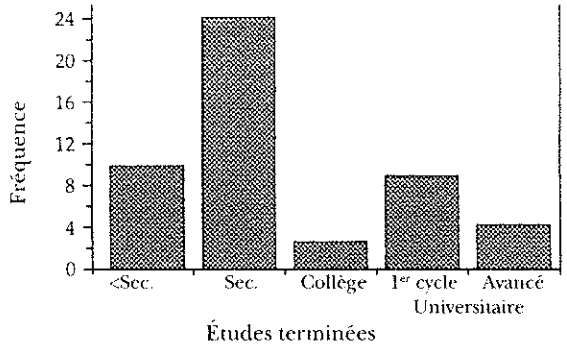
Figure 2.1. Distribution des attitudes concernant la désobéissance civile



Bien qu'il n'y ait pas de règle solide et inviolable qui le dicte, les diagrammes circulaires sont utiles seulement dans le cas de variables ayant moins de, disons, 8 ou 9 valeurs. Avec un plus grand nombre de valeurs, certaines portions de la tarte seraient trop petites pour illustrer clairement une information. Il vaut mieux alors utiliser un diagramme en bâtons.

De plus le diagramme circulaire sert le plus souvent pour des variables nominales ou, comme dans l'exemple précédent, les variables dichotomiques. Les variables ordinales et d'intervalles/ratio sont, règle générale, mieux présentées par des diagrammes en bâtons, lesquels illustrent avec une plus grande clarté l'ordre existant parmi les valeurs. Nous constatons donc, pour la première fois, l'influence qu'exerce le niveau de mesure sur le choix des techniques statistiques. Si une variable est de type nominal et qu'elle possède peu de valeurs, songez à user d'un diagramme circulaire. Si la variable est ordinale ou d'intervalles/ratio ou si elle possède de nombreuses valeurs, vous devriez vous servir plutôt d'un diagramme en bâtons. La figure 2.2 montre un diagramme en bâtons qui illustre la distribution du niveau d'instruction pour les 50 cas de notre exemple.

Figure 2.2. Niveau d'instruction atteint



À l'instar des diagrammes circulaires, les diagrammes en bâtons peuvent dépeindre soit des pourcentages, soit des fréquences. L'affaire est simple : plus il y a de cas d'une valeur donnée, plus haute sera la bande. La hauteur de chaque bande est proportionnelle au pourcentage ou au nombre de cas de cette valeur. Dans le diagramme de la figure 2.2 par exemple, il y a trois fois plus de diplômés de premier cycle universitaire que de diplômés de collège. Aussi le bâton représentant les premiers est-il trois fois plus haut que celui des seconds. La fréquence ou le pourcentage des cas pour chacune des valeurs se lit sur l'axe vertical situé à la gauche du tableau.

Soit dit en passant, lorsque la variable est une variable continue et d'intervalles/ratio, ce type de diagramme en bâtons est nommé ***histogramme***. Dans ce cas précis, les bandes devront se toucher afin d'indiquer que la variable est continue. Dans la figure 2.2, j'ai traité la variable « Niveau d'instruction » comme une variable discrète. C'est pourquoi les bandes ne se touchent pas.

Voici certaines conventions que vous devriez suivre lorsque vous construisez un diagramme en bâtons.

- Lorsqu'il y a plus d'un diagramme, nommez chacun d'eux par le titre « Figure » et numérotez-les en chiffres arabes.
- Donnez au graphique un titre aussi clair que concis, un titre qui précise la variable illustrée.

- L'axe vertical devrait correspondre approximativement à 60 % ou 75 % de la longueur totale de l'axe horizontal. Cette proportion confère une certaine uniformité aux graphiques et facilite par le fait même leur comparaison. Pour des variables qui ont de nombreuses valeurs, tel l'âge, vous devrez parfois faire exception à cette règle.
- Si vous présentez deux diagrammes ou plus concernant des variables similaires (par exemple, l'instruction du répondant et l'instruction du conjoint du répondant), conservez la même échelle pour tous ces diagrammes. L'utilisation de la même échelle facilite la comparaison des diagrammes.
- Dans le cas de variables ordinales ou d'intervalles/ratio, faites la liste des valeurs de la plus basse à la plus élevée au fur et à mesure que vous vous déplacez de la gauche vers la droite de l'axe horizontal.
- Les bâtons doivent être habituellement d'égale largeur, de façon à ce que l'aire occupée par une bande soit proportionnelle à leur fréquence ou à leur pourcentage.
- Nous devrions retrouver des espaces entre les bâtons d'un diagramme illustrant une variable discrète (le niveau d'instruction tel que nous l'avons mesuré auparavant, par exemple). Ne conservez cependant pas d'espace entre les bâtons lorsqu'il s'agit de variables continues (comme l'instruction mesurée en années de scolarité terminées).
- Étiquetez l'axe vertical des noms « Fréquence » ou « Pourcentage » afin que le lecteur sache ce que représente cet axe.
- Afin d'éviter que la perception de la surface des bâtons ne soit faussée, l'axe vertical doit débuter à la valeur zéro. Autrement, les bandes plus grandes sembleront indûment grandes au regard des bâtons plus petits et la différence entre les hauteurs respectives des bâtons aura tendance à être exagérée. (Vous verrez souvent cette règle informelle violée dans les médias, et, à l'occasion, vous devrez vous-mêmes y faire exception. Soyez attentifs à cette façon de faire dans les médias, et soyez très prudents lorsque vous ferez entorse à cette règle.)

Ces normes concernent uniquement quelques-unes des qualités d'un graphique bien conçu. Heureusement il existe d'excellents bouquins qui décrivent comment créer de bons graphiques.

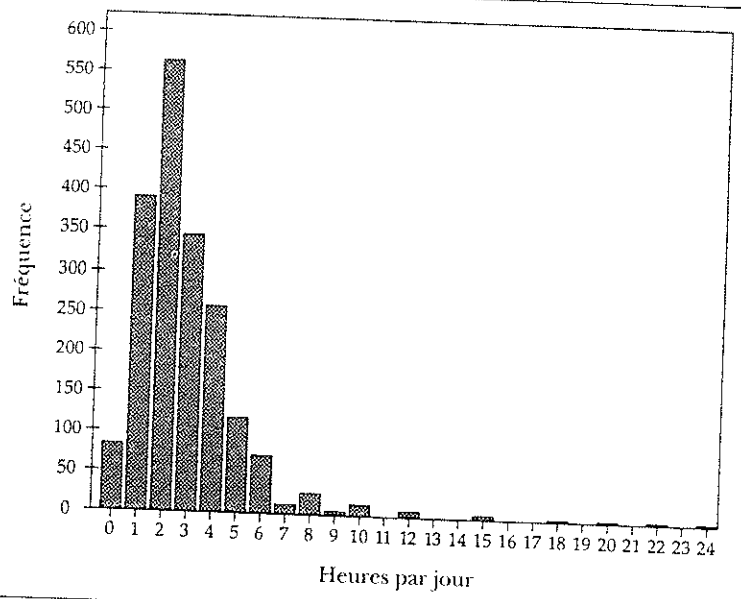
2.9 Les cas déviants

Les diagrammes en bâtons sont d'une grande aide pour déceler les ***cas déviants*** – les cas dont le score d'une variable d'intervalles/ratio (et de certaines variables ordinales qui ont beaucoup de valeurs) est isolé, et paraît anormalement haut ou anormalement bas. Ces scores ne sont pas simplement des scores élevés ou faibles, mais surtout ils se retrouvent éloignés aux extrémités d'une distribution, détachés de la plupart des autres scores. Plusieurs des méthodes statistiques que nous examinerons plus tard sont faussées par la présence de cas déviants. Ainsi, il nous sera nécessaire de les prendre en considération lors de l'analyse des données.

La figure 2.3 montre un diagramme en bâtons (plus précisément, un histogramme) illustrant la variable « Écoute quotidienne de la télévision » telle qu'on la trouve dans le General Social Survey. Voyez-vous des cas déviants ? Bien sûr, ces petites bandes à l'extrême droite, représentant des scores visiblement détachés du reste de la distribution. Ce sont des répondants qui affirment écouter la télévision 16, 20, 22 et même, ô prodige !, 24 heures à chaque jour. Peut-être, au moment même où vous lisez ce livre, ces types sont-ils collés à leur téléviseur, regardant le canal météo ou le réseau des sports. Plus probablement, ce sont des répondants qui exagèrent le temps qu'ils passent devant leur téléviseur. Nous pouvons aussi découvrir les cas déviants à l'aide de tableaux de fréquences et de pourcentages, bien qu'il soit généralement plus facile de le faire à l'aide d'un diagramme en bâtons. Ils ressortent aux extrémités gauche et droite.

Restez toujours sur le qui-vive car les cas déviants peuvent affecter, souvent négativement, une analyse. Nous verrons des exemples de ces effets dans les prochains chapitres. Lorsque vous rencontrez des cas déviants, tentez d'abord de comprendre la raison de leur existence. Découlent-ils d'erreurs dans la mesure ou dans la compilation des données ? Le cas échéant, nous devrons corriger ces erreurs. Ou ne témoignent-ils pas plutôt de quelque processus inhabituel ? En d'autres mots, constituent-ils des anomalies – des événements auxquels nos théories ne nous permettaient pas de songer ? Si tel est le cas, il nous faudra tenter de les comprendre, peut-être en révisant nos concepts théoriques. Ou encore ces cas déviants ne sont-ils pas simplement d'heureux hasards, des événements singuliers qui n'ont aucune implication théorique même s'ils brouillent nos analyses ? Dans le cas de l'écoute de la télévision, je soupçonne que nous ayons affaire à des télévores ou à des répondants qui exagèrent. Dans ce cas, il est possible d'exclure ces cas de notre analyse.

igure 2.3. Écoute quotidienne de la télévision



Souvent nous excluons les cas déviants des analyses de données. Nous examinerons, tout au long de ce livre, les raisons qui peuvent justifier ce choix. Mais nous ne devons jamais exclure les cas déviants de façon mécanique. Il est impératif d'abord de déterminer leur signification et ce qu'ils impliquent pour notre analyse, et ensuite de décider si leur exclusion aiderait ou nuirait à l'atteinte de nos objectifs de recherche. Nous pouvons produire deux analyses, une avec les cas déviants, l'autre sans, et ensuite choisir laquelle des deux répond le mieux à la question de recherche que nous nous sommes posée. Mon leitmotiv : réfléchir est la chose la plus importante en statistiques. Nous apprendrons à gérer les cas déviants au moment opportun.

0 La cartographie des variables écologiques

Les données écologiques proviennent souvent de variables continues d'intervalles/ratio, tels les moyennes, les taux et les pourcentages, qui peuvent prendre tellement de valeurs qu'il est difficile de les décrire à l'aide de diagrammes circulaires, de diagrammes en bâtons, de distributions de fréquences ou de distributions de pourcentages. Examinez, par exemple, le taux de fertilité des 50 pays les plus peuplés, c'est-à-dire le nombre moyen de naissances par femme. Les taux vont de 1,37 en Italie à 7,15 en Ouganda, peu de pays ayant des taux identiques. Par conséquent, un diagramme ou un tableau de distribution ne serait guère éloquent. Un diagramme en bâtons montrerait des bâtons sensiblement de même hauteur pour la plupart, chacun indiquant le seul pays auquel correspond ce score précis. Un tableau consisterait en une liste des scores des 50 pays avec un ou peut-être quelques cas par valeur. Bien sûr, nous pourrions fusionner des catégories et parfois cela s'avère très utile. La fusion de catégories, cependant, implique une perte d'information.

Il existe d'autres types de distributions que les distributions de fréquences et de pourcentages. On peut mieux illustrer les distributions spatiales d'aires géographiques à l'aide de cartes. Les cartes n'étaient-elles pas l'outil principal utilisé par les sociologues de la première heure, comme le Belge Adolphe Quetelet ou le Français André Guerry dans les années 1820-1830 ? Des variables comme le taux de fertilité peuvent être représentées à l'aide d'une carte muette sur laquelle les pays sont ombragés ou colorés en fonction de leur score. Nous pouvons ainsi observer comment se distribue géographiquement une variable.

La figure 2.4 présente une carte qui illustre le taux de fertilité des 50 pays les plus peuplés. Cette carte montre que les hauts taux de fertilité se concentrent en Afrique et au Moyen-Orient, et que l'on trouve des faibles taux en Europe et dans certains pays de l'Asie de l'Est. Ces structures peuvent être observées sans peine à l'aide d'une seule carte. Nous aurions bien des difficultés à tenter de les déceler à partir d'une simple liste des pays et des taux de fertilité. (Notez que les aires blanches sur la carte indiquent des pays qui ne font pas partie des 50 pays les plus peuplés ou pour lesquels aucune information n'est disponible.)

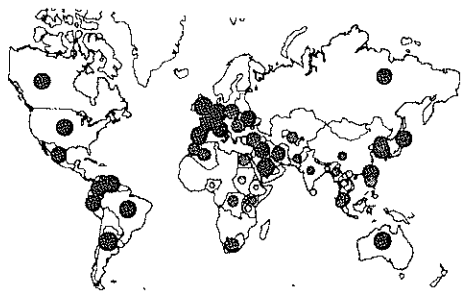
Figure 2.4 Taux de fertilité des 50 pays les plus peuplés

Comparez la carte illustrant les taux de fertilité avec celle qui montre, à la figure 2.5, le pourcentage de la population vivant dans des zones urbaines. Hum... étrange ! La distribution que nous observons ici ressemble passablement à celle du taux de fertilité. Bien sûr la coloration est inversée, les pays les plus sombres sur la carte du taux de fertilité tendent à avoir une coloration plus claire sur la carte de l'urbanisation. Mais les distributions sont semblables. L'Europe et certains pays d'Asie de l'Est sont parmi les plus urbanisés et ont les taux de fertilité les plus faibles. L'Afrique et une partie de l'Asie sont moins urbanisées tout en ayant les taux de fertilité les plus élevés. À n'en pas douter, ces schémas, s'ils ne sont pas identiques, sont néanmoins similaires.

Figure 2.5 Taux d'urbanisation des 50 pays les plus peuplés

Les représentations cartographiques comme les figures 2.4 et 2.5 peuvent parfois être trompeuses parce que les unités les plus grandes occupent plus d'espace que les plus petites. Nos yeux (et notre cerveau) accordent naturellement plus d'attention à de grands pays comme les États-Unis ou la Russie qu'à des petits pays comme le Japon ou l'Allemagne. Nous obtenons parfois une meilleure représentation visuelle de la distribution spatiale d'une variable à l'aide d'une carte à tache comme celle de la figure 2.6 qui montre le niveau d'urbanisation. Cette carte présente la même information que la carte de la figure 2.5 mais représente le niveau d'urbanisation de chaque pays par une tache circulaire. La dimension de la tache est proportionnelle au taux d'urbanisation du pays. Plus la tache est grande, plus le taux d'urbanisation est élevé. Remarquez que la dimension de la tache ne dépend pas de la superficie du pays. Bien qu'ils soient minuscules sur la carte, le Japon et les pays d'Europe sont représentés par une grande tache, car ils sont très urbanisés.

Figure 2.6 Taux d'urbanisation des 50 pays les plus peuplés



Nous verrons, au chapitre 10, comment analyser plus efficacement les liens entre ces variables. Mais observer ces cartes nous donne déjà une bonne idée des similarités qui existent entre ces schémas. La projection cartographique est un outil important pour décrire une variable écologique et pour comparer sa distribution géographique avec celle d'autres variables écologiques.

2.11 Résumé du chapitre 2

Voici ce qui a été examiné dans ce chapitre :

- Les distributions de fréquences et de pourcentages résument les scores d'une variable.
- Les pourcentages standardisent les fréquences en base 100, rendant ainsi les données plus faciles à interpréter et à comparer.
- Les pourcentages doivent être employés avec prudence lorsque le nombre total de scores est faible (inférieur à 30).
- Les distributions cumulatives nous renseignent sur le nombre ou le pourcentage de tous les scores égaux ou inférieurs à une valeur donnée.
- Les distributions cumulatives ne sont pertinentes que pour des variables ordinales ou d'intervalles/ratio.
- Les tableaux que l'on présente à d'autres doivent être lisibles et doivent respecter un modèle conventionnel.
- Quand on fusionne des catégories, il faut réfléchir à la façon la plus utile de combiner les valeurs d'une variable, eu égard à la question de recherche et à la façon dont on veut mener l'analyse.
- Les valeurs comme « Non applicable », « Ne sait pas », « Sans d'opinion », « Incapable de choisir », « Pas de réponse », « Refus de répondre », « Incertain » devraient habituellement être

considérées comme des données manquantes et être exclues de l'analyse. Parfois on traite la valeur « Autre » comme une donnée manquante.

- Quand des valeurs comme « Ne sait pas » ou « Sans opinion » ont une signification conceptuelle ou théorique, on ne devrait pas les exclure de l'analyse.
- Il arrive parfois que l'on doive sélectionner un sous-ensemble de cas sur la base de leur score dans certaines variables.
- Les distributions de fréquences et de pourcentages peuvent être illustrées visuellement par des diagrammes circulaires ou en bâtons.
- Lorsque nous avons affaire à des variables ordinales, d'intervalles/ratio, ou encore à des variables nominales qui ont de nombreuses valeurs, les diagrammes en bâtons sont généralement préférables aux diagrammes circulaires.
- Il faut être attentif aux cas déviants, ceux dont les scores extrêmes peuvent biaiser l'analyse statistique.
- La distribution d'une variable écologique peut être illustrée visuellement à l'aide d'une projection cartographique.

Principaux concepts et procédures

Termes et idées

istribution de fréquences
onnées brutes
équence
ableau de fréquences
ourcentage
roportion
bleau de pourcentages
ourcentages cumulatifs
écimales significatives
bleaux bien présentés
sion de catégories
nnées manquantes
us-ensemble
agramme circulaire
agramme en bâtons
stogramme
déviant
tographie

Symboles

f
N
F

Formules

$$\text{Pourcentage} = \frac{f}{N}(100)$$

$$\text{Pourcentage cumulatif} = \frac{F}{N}(100)$$