ΣX_i
$\Sigma(X_i - \overline{X})$ et $\Sigma(X_i - \mu)$
$\Sigma(X_i - \overline{X})^2$ et $\Sigma(X_i - \mu)^2$

Formules

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

$$\overline{X} = \frac{\Sigma X_i}{N}$$

CHAPITRE 4

Les mesures de variation

On raconte l'histoire d'un homme qui se noya dans un ruisseau qui avait en moyenne 10 centimètres de profondeur. Ce vieux conte devrait nous rappeler qu'il faut considérer bien plus que la tendance centrale dans la description des variables. Les distributions peuvent être caractérisées par leur variation en plus de leur tendance centrale, et quelquefois cette variation est plus importante que la tendance centrale. Nous allons maintenant présenter la variance et l'écart-type qui permettent de mesurer la variation d'une distribution. Nous apprendrons à évaluer le degré de dispersion ou de concentration de la distribution d'une variable, c'est-à-dire dans quelle mesure les scores sont semblables ou différents l'un de l'autre. Nous verrons également les différentes formes de distribution et les courbes normales. Nous utiliserons ce que nous savons des mesures de tendance centrale et des écarts-types pour transformer les scores en ce que l'on nomme les scores standardisés ou scores-Z, qui permettent de comparer la position des scores dans une distribution. Finalement, nous apprendrons comment utiliser l'information se rapportant à la tendance centrale et à l'écart-type d'un échantillon pour estimer la tendance centrale d'une population.

Après avoir lu ce chapitre vous pourrez :

1. Définir et calculer l'écart-type et la variance pour un échantillon et une population.
2. Reconnaître les conditions d'utilisation de l'écart-type et de la variance.
3. Mesurer l'asymétrie d'une variable.
4. Expliquer dans des termes généraux ce qu'est une courbe nor-

6. Expliquer ce qu'est une distribution d'échantillonnage.
7. Calculer et interpréter les intervalles de confiance autour de la moyenne.
8. Faire preuve de prudence dans l'interprétation des mesures de tendance centrale et des mesures de variation dans le cas des données agrégées.

4.1 Les mesures de variations : la variance et l'écart-type

Maintenant que nous savons ce que sont les mesures de tendance centrale, jetons un coup d'œil aux ***mesures de variation***[1], des mesures qui résument comment les scores sont agglomérés ou dispersés. Il est souvent utile de savoir dans quelle mesure les scores sont similaires ou différents les uns des autres. Les scores sont-ils si similaires qu'ils se massent tous autour de quelques valeurs ? Ou sont-ils plutôt étalés sur une grande surface ? Si les scores tendent à s'agglomérer, nous disons qu'ils sont homogènes. S'ils ont tendance à être dispersés, nous disons qu'ils sont hétérogènes.

Voici trois groupes de scores, les premiers étant relativement homogènes, les derniers relativement hétérogènes :

	Groupe A Relativement homogènes	*Groupe B* Entre les deux	*Groupe C* Relativement hétérogènes
	64	44	34
	68	63	58
	70	80	90
	71	91	101
	69	74	79
	66	56	46
Moyenne	*68*	*68*	*68*

Les trois groupes ont la même moyenne : 68. Mais les scores du groupe A sont relativement similaires, tournant autour de la moyenne de 68. En revanche les scores du groupe B sont beaucoup plus dispersés et ceux du groupe C le sont encore plus.

Décrire la variation consiste à mesurer la divergence des scores par rapport à un score typique, le score moyen. Si les scores divergent beaucoup, la distribution se présentera de façon assez étendue, le gros des scores se situant loin du score typique. Si les scores ne divergent pas beaucoup, ils se masseront beaucoup plus près du score typique et s'agloméreront ensemble. Nous utilisons habituellement la moyenne (plutôt que le mode ou la médiane) comme point de référence à partir duquel on mesure les écarts. La moyenne d'un ensemble de scores est un point de référence utile pour décrire la variation car elle tient directement compte de tous les scores. De plus, la moyenne possède la propriété de minimiser la somme des écarts entre elle-même et chacun des scores. Rappelez-vous de ce que nous disions de la moyenne à la section précédente : la somme des écarts des scores par rapport à la moyenne est zéro. Voilà qui est drôlement minimisé ! Rappelez-vous également que la somme du carré des écarts par rapport à la moyenne est un minimum. C'est-à-dire que la somme du carré des différences entre chaque score et la moyenne est moindre que la somme du carré des différences entre chaque score et n'importe quelle autre mesure.

La variance et l'écart-type sont deux mesures de variation très apparentées qui résument dans quelle mesure les scores sont concentrés autour de la moyenne. Considérons-les à tour de rôle.

En premier lieu, ***la variance***. Je présenterai deux façons légèrement différentes d'obtenir une variance. La première concerne les données de population, l'autre les données d'échantillon. La variance d'un ensemble de scores décrivant une *population* entière se calcule comme suit :

1. Calculez la moyenne.
2. Soustrayez la moyenne de chacun des scores. (Certaines de ces différences seront négatives. C'est normal.)
3. Mettez au carré chacune de ces différences. (C'est-à-dire multipliez chaque différence par elle-même. Rappelez-vous que la multiplication de deux nombres négatifs donne un nombre positif.)
4. Additionnez toutes ces différences élevées au carré.
5. Divisez cette somme par le nombre total de scores.

La réponse sera la variance des scores.

J'ai signalé dans la section 3.4 que la somme des écarts entre les scores et la moyenne est toujours zéro. En d'autres termes, $\Sigma(X - \mu) = 0$. C'est pour cette raison qu'à l'étape 2 du calcul nous

Autrement, nous aboutirions invariablement à une somme nulle, ce qui ne nous renseignerait guère sur la variation[2].

Le calcul est exprimé de façon plus concise par la formule suivante :

$$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N}$$

lorsque σ^2 = la variance de données de population

X_i = le score du i^e cas

μ = la moyenne de la population

N = le nombre total de cas dans la population

σ est la lettre grecque sigma ; σ^2 se prononce « sigma carré ». Dans quelques paragraphes, j'aurai l'occasion d'expliquer pourquoi nous employons σ^2 comme symbole de la variance. Pour le moment, je mentionnerai uniquement que nous utilisons un caractère grec parce qu'il s'agit de données qui décrivent une population (et non un échantillon). Rappelez-vous que μ est le symbole de la moyenne d'une population (et non celle d'un échantillon). Nous avons déjà appris que Σ signifie que nous devons additionner tout ce qui suit. Dans cette formule Σ signifie donc qu'il faut additionner toutes les différences élevées au carré entre chaque score et la moyenne. Nous avons déjà vu ce total (section 3.4) : c'est la somme des déviations au carré ou, plus couramment, la somme des carrés.

Vous ne pouvez trouver un meilleur point de référence que la moyenne si vous souhaitez minimiser la somme des écarts au carré. Cette propriété ne devrait pas vous surprendre car nous avons dit plus haut (section 3.4) que la moyenne minimise la somme des déviations « simples » (entendre non élevées au carré). Cette somme est toujours nulle. Notez que nous obtenons la variance en divisant la somme des carrés par N. Autrement dit, la variance est la moyenne des écarts au carré des scores par rapport à la moyenne. C'est si important que je veux que vous le répétiez : *la variance est la moyenne des écarts au carré des scores par rapport à la moyenne.*

Vous pouvez donc sans peine comprendre en quoi la variance s'avère une mesure utile de la variation d'une distribution. Si les scores se distribuent de façon étendue et lâche autour de la moyenne, les écarts seront grands et, ainsi, la somme des carrés et la variance seront élevées. Si au contraire les scores s'agglomèrent près de la

2. Il est également possible de mesurer la variation en faisant la somme des valeurs absolues des écarts. Toutefois, cette mesure, appelée déviation moyenne, a une utilité statistique moindre que la variance basée sur la somme des carrés et, par conséquent, elle est [illegible]

moyenne, les écarts, la somme des carrés et la variance seront faibles. Aussi, une forte variance témoigne d'une grande variation, une faible variance d'une variation plus faible.

Toutefois, un problème subtil apparaît lorsque nous travaillons avec des données *d'échantillon*. Bien que cette $\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N}$ définisse la variance des scores, cela ne convient pas aux données d'échantillon. Pour des données d'échantillon, la formule donne une estimation biaisée de la variance de la population. Par « biaisée » il faut entendre que, si nous calculions la variance de tous les échantillons de taille N possibles choisis aléatoirement parmi une population, et qu'ensuite nous calculions la moyenne de ces variances d'échantillon, cette variance moyenne des échantillons serait différente de la variance de la population. En fait, avec cette formule, la moyenne des variances de tous les échantillons sous-estime toujours la variance de la population. Cela pose un problème car nous n'aimons pas les biais en statistique.

Heureusement, les statisticiens ont imaginé une manière de corriger ce biais dans le calcul de la variance de données d'échantillon. Il suffit de diviser la somme des carrés par N – 1 plutôt que par N. Ainsi, la variance d'une population estimée à partir de données d'échantillon se calcule de cette façon :

$$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{N - 1}$$

lorsque s^2 = la variance de l'échantillon

X_i = le score du i^e cas

$\overline{X}$ = la moyenne de l'échantillon

N = le nombre total de cas dans l'échantillon

Parce qu'en sciences sociales nous travaillons très souvent avec des données d'échantillon, comme celles du General Social Survey, ce sera cette formule que j'utiliserai tout au long du livre quand je calculerai la variance. En pratique, l'emploi de N ou N – 1 comme dénominateur importe seulement lorsque N est très faible. Pour de grands échantillons, à l'instar ceux du General Social Survey, diviser par 2 903 ou par 2 904 fait bien peu de différence. Néanmoins, nous emploierons en général N – 1 lorsque nous aurons affaire à des données d'échantillon.

Les quelques éléments de terminologie que nous verrons maintenant s'avéreront par la suite d'une grande utilité. Le dénominateur N – 1 se nomme degrés de liberté de la variance. La raison

examinions tout à l'heure : la somme des écarts par rapport à la moyenne est toujours 0. Ainsi, si nous connaissons tous les écarts excepté un seul, il est aisé de calculer ce dernier écart. Il s'agit du nombre qui annulera la somme des déviations. Par exemple, si, sur trois scores, nous en connaissons deux, soit 4 et 5, et que la moyenne des trois scores est de 5, alors le score inconnu (appelons-le X) ne pourra être que 6. En effet :

$$(4-5) + (5-5) + (X-5) = 0$$
$$-1 + 0 + X - 5 = 0$$
$$X = 1 + 5$$
$$X = 6$$

En d'autres termes, les écarts par rapport à la moyenne sont quelque peu limités. Pour une variance donnée, seuls N – 1 écarts peuvent librement varier ; une fois qu'ils sont connus, le dernier est fatalement déterminé. Nous disons donc qu'il existe N – 1 degrés de liberté. Cette notion de degrés de liberté reviendra souvent dans ce livre en référence à plusieurs statistiques.

Bien que les logiciels de statistiques puissent trouver pour nous les variances, en calculer quelques-unes nous-mêmes nous aidera à mieux comprendre en quoi elles consistent. Voici un exemple simple de calcul de la variance. Soit les scores 64, 66, 68, 69, 70, 71 d'un échantillon de taille N = 6. Le score moyen est 68. Calculons la variance en dressant d'abord un tableau comme celui-ci :

X	$X_i - \overline{X}$	$(X_i - \overline{X})^2$
64	64 – 68 = –4	16
68	68 – 68 = 0	0
70	70 – 68 = 2	4
71	71 – 68 = 3	9
69	69 – 68 = 1	1
66	66 – 68 = –2	4
Somme	0	34

$$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{N-1}$$
$$= \frac{34}{6-1}$$
$$= \frac{34}{5}$$
$$= 6{,}80$$

La variance de cet ensemble de scores est 6,80. (Remarquez, une fois de plus, que si vous faites la somme de la colonne des différences $X_i - \overline{X}$ le résultat sera toujours 0. Si cela n'est pas le cas, vous avez commis une erreur de calcul quelque part.)

À propos, voici les variances pour les trois groupes que nous avons vus précédemment.

	Groupe A Relativement homogènes	*Groupe B* Entre les deux	*Groupe C* Relativement hétérogènes
	64	44	34
	68	63	58
	70	80	90
	71	91	101
	69	74	79
	66	56	46
Moyenne	*68*	*68*	*68*
s^2	*6,80*	*290,80*	*686,80*

Comme on peut s'en apercevoir du premier coup d'œil, les scores du groupe A sont plus semblables entre eux que les scores du groupe B qui à leur tour sont plus semblables que ceux du groupe C. Cependant, nous avons maintenant une mesure quantitative de ces variations – les variances sont de 6,80, 290,80 et 686,80. Ces variances mesurent la dispersion des scores autour de la moyenne.

Avant de continuer, je tiens à souligner que dans le cas des variables dichotomiques telles que le sexe, codé 0 pour les hommes et 1 pour les femmes, la variance est P(1 – P), lorsque P est la proportion de cas codés 1. En d'autres mots, la variance est la proportion de cas codés 1 multipliée par la proportion de cas codés 0. Nous avons vu dans le chapitre 3 que la moyenne d'un variable dichotomique codée 0 et 1 est la proportion de cas codés 1. Ainsi, bien que nous utilisions habituellement les moyennes et les variances pour les variables d'intervalles ou de proportions, elles sont également utiles pour les variables dichotomiques, même nominales. Je ne m'étendrai pas sur les raisons pour lesquelles la variance d'une variable dichotomique est P(1 – P).

L'écart-type, pour sa part, est la racine carrée de la variance. (Cela signifie évidemment que la variance équivaut au carré de l'écart-type.) Pour obtenir l'écart-type d'un ensemble de scores :

1. Calculez la variance.
2. Trouvez la racine carrée de cette variance.

Cette racine carrée sera l'écart-type.

Voici deux formules équivalentes pour calculer l'écart-type de données de population :

$$\sigma = \sqrt{\text{Variance}} \quad \text{et } \sigma = \sqrt{\frac{\Sigma(X_i - \mu)^2}{N}}$$

Et voici les formules pour des données d'échantillon :

$$s = \sqrt{\text{Variance}} \quad \text{et } s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N - 1}}$$

Pour les scores susmentionnés, l'écart-type du groupe A est $\sqrt{6,80} = 2,61$, les écarts-types des groupes B et C sont respectivement de $\sqrt{290,80} = 17,05$ et $\sqrt{686,80} = 26,21$.

L'écart-type (*standard deviation* en anglais) est représenté par la lettre s. Voilà pourquoi la variance, qui est le carré de l'écart-type, est notée s^2. De plus, de la même façon que les statisticiens se servent du $\bar{X}$ romain pour la moyenne d'un échantillon et du μ grec pour la moyenne d'une population, ils emploient s pour indiquer l'écart-type de données d'échantillon et σ pour l'écart-type de données de population. Par conséquent, la variance d'une population est notée σ^2.

Pour des données exprimées en entiers, on a coutume d'arrondir l'écart-type et la variance à la seconde décimale. C'est de cette façon que je procéderai tout au long du livre. Cette règle générale peut bien sûr souffrir des exceptions. Il est néanmoins impératif que vous *réfléchissiez* à ce que signifient les décimales, et cela pour chaque situation particulière.

L'interprétation d'une variance ou d'un écart-type est facile. Les formules que nous avons examinées plus haut nous ont bien montré que moins il y a de variation entre les scores, plus petite sera la somme des carrés, donc, par le fait même, la variance et l'écart-type. S'il n'y a aucune variation parmi les scores, la variance et l'écart-type seront nuls. Dans un pareil cas, tous les scores seront identiques (tous les scores seront égaux à la moyenne) et, en fait, cette « variable » sera une constante. Plus il y a de variation dans les scores, plus l'écart-type et la variance sont grands. Bien que la variance ou l'écart-type le plus faible possible soit 0, il n'existe théoriquement pas de limite supérieure à ces mesures de variation.

Étant donné que la variance n'est en réalité que le carré de l'écart-type, pourquoi avoir deux mesures différentes ? Pourquoi ne pas se limiter soit à la variance, soit à l'écart-type ? La raison est que la variance et l'écart-type ont chacun des propriétés particulières. La variance possède plusieurs propriétés mathématiques qui sont très importantes pour certaines méthodes statistiques plus avancées. Les statisticiens peuvent faire bien des choses à l'aide du concept de variance. Nous en verrons des exemples lorsque nous traiterons de l'analyse de variance et des techniques de régression aux chapitres 8 à 12.

La variance a néanmoins un inconvénient. Parce qu'elle met à la puissance 2 les écarts par rapport à la moyenne sans ensuite les remettre en base 1, elle s'exprime dans une échelle différente de celle des scores. Règle générale, à cause de cette différence d'échelle, la variance ne semble pas « correcte ». En guise de mesure de la variation entre les scores, l'écart-type, au contraire, paraît « correct ». Bien sûr, il se doit d'être « correct » étant donné qu'il se calcule en trouvant la racine carrée de l'écart au carré moyen. Il remet en base 1 un nombre préalablement élevé à la puissance 2. Cette opération fait en sorte que la mesure de variation est ramenée sur la même échelle que les scores originaux.

Considérons les trois groupes de scores que nous avons vus précédemment. Nous avons déterminé que leurs variances sont respectivement de 6,80, 290,80 et 686,80. Ces variances sont beaucoup plus importantes que les groupes de scores qu'elles décrivent. Même la plus petite, 6,80, semble trop grande pour exprimer la variation de scores qui vont de 64 à 71. En revanche, les écarts-types des groupes A, B et C (2,61, 17,05 et 26,21) sont du même ordre de grandeur que les scores à partir desquels ils ont été calculés.

Prenons un exemple : l'âge des répondants au General Social Survey. Les scores s'étendent de 18 à 89, avec une moyenne de 44,8 (en excluant bien sûr les données manquantes). La variance de l'âge est un colossal 284,5. Cependant, l'écart-type semble plus raisonnable avec 16,9, car il est exprimé dans la même échelle d'années que l'âge des répondants. Par conséquent, lorsque que nous faisons rapport des résultats d'analyses statistiques, nous employons le plus souvent l'écart-type, et non la variance, la première étant davantage à la mesure de nos intelligences bien humaines que la seconde. La plupart du temps donc, servez-vous de la variance lorsque vous avez à faire des raisonnements statistiques plus avancés et choisissez l'écart-type lorsque vient le temps de présenter vos résultats.

Puisque la variance et l'écart-type reposent sur les écarts par rapport à la moyenne, et parce que (strictement parlant) la moyenne

ne peut être calculée que pour des variables d'intervalles/ratio, la variance et l'écart-type ne sont appropriés que pour des variables mesurées au niveau d'intervalles/ratio. Strictement parlant donc, ne calculez pas de variance ou d'écart-type pour des variables de niveau nominal ou ordinal. Mais, comme nous l'avons indiqué dans le cas des moyennes, il est parfois bénéfique de violer cette règle et de calculer des variances ainsi que des écarts-types lorsque nous avons affaire à des variables ordinales, à condition que cela puisse nous fournir des informations utiles à propos de la variable en question.

Faites attention aux scores extrêmes, qui peuvent affecter indûment la variance et l'écart-type, et tout spécialement aux cas déviants. Nous avons déjà vu que des scores anormalement bas ou anormalement hauts peuvent rendre la moyenne bien peu représentative de la distribution. L'effet de ces scores extrêmes est encore plus grand en ce qui a trait aux mesures de variation. La variance et l'écart-type ne font pas que reposer sur les écarts par rapport à la moyenne, ils élèvent au carré ces écarts, accroissant ainsi leur effet. La mise au carré de tout nombre supérieur à 1 l'augmente de façon exponentielle – c'est-à-dire l'augmente énormément. L'écart d'un score de 4 par rapport à une moyenne de 8 n'est que de 4. Seulement, le *carré* de cet écart est de 16.

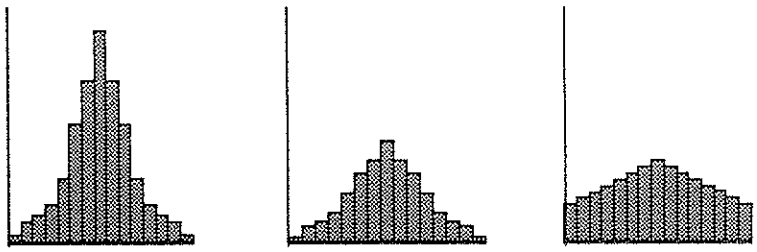
Dans certaines situations extrêmes, une variance ou un écart-type peut se voir littéralement « gonflé » par l'action d'un seul cas déviant. Voici un exemple : le pourcentage de la population de chaque État américain qui se déclare de souche asiatique a un écart-type de 8,65 lorsque Hawaii est compris, et de seulement 1,54 lorsque que l'on exclut Hawaii. (J'ai mentionné dans la section 3.4. que Hawaii, avec ses 61,8 %, constituait la quintessence du cas déviant.) En de telles situations, vous devez d'abord chercher d'où provient le score extrême. Ensuite, songez à l'exclure de votre analyse, du moins lorsque vous calculerez la moyenne, la variance ainsi que l'écart-type. Rappelez-vous également d'exclure les données manquantes lorsque vous calculez les variances et les écarts-types.

4.2 De la forme des distributions

À l'instar des gens, les distributions ont des silhouettes. Nous avons déjà remarqué que certaines distributions étaient unimodales, d'autres bimodales, et d'autres trop plates pour être décrites en termes de modes. Certaines distributions sont grandes et étroites, certaines sont petites et larges, d'autres encore sont entre les deux. Certaines distributions présentent une longue « queue » à gauche, d'autres à droite, d'autres dans les deux directions et certaines n'en [illegible]

Les statisticiens emploient le terme ***kurtose*** lorsqu'ils décrivent l'escarpement des distributions de variables d'intervalles/ratio. Les distributions ***leptokurtiques*** sont hautes et minces ; les distributions ***platykurtiques*** sont basses et aplaties ; et les distributions ***mésokurtiques*** se situent quelque part au milieu. La figure 4.1 présente des exemples de ces trois formes générales que peuvent prendre les distributions.

Figure 4.1. Les formes générales des distributions



L'asymétrie constitue également une caractéristique importante des distributions. Comme nous l'avons observé dans la section 3.6, certaines variables sont distribuées de façon symétrique alors que d'autres présentent, à gauche ou à droite, une forme asymétrique. Les statisticiens ont mis au point une mesure décrivant l'asymétrie, mesure qui repose sur les effets de l'asymétrie sur la moyenne et la médiane. Rappelons que, à la section 3.5, nous disions que les scores extrêmes « tirent » littéralement vers eux la moyenne, alors que la médiane ne dépend que du score se trouvant au milieu de la distribution. Ainsi, la moyenne est toujours davantage rapproché du lieu de l'asymétrie que peut l'être la médiane. La formule suivante mesure l'asymétrie :

$$\text{Asymétrie} = \frac{3(\overline{X} - \text{Md})}{s}$$

Il faut que l'on divise la formule par l'écart-type pour la mettre à l'échelle de la variable. (La raison pour laquelle on multiplie $(\overline{X} - \text{Md})$ par 3 est plus complexe et il n'est pas nécessaire d'en parler pour le moment.)

Remarquez que dans la formule la différence entre la moyenne et la médiane permet d'évaluer l'importance de l'asymétrie. Plus une distribution est asymétrique, plus la différence entre la moyenne et la médiane est importante et plus le numérateur est important. L'asymétrie est positive lorsqu'elle se situe à droite, et négative lorsqu'elle se situe à gauche. Vous comprendrez pourquoi en comparant les positions relatives des moyennes et des médianes dans les diagrammes des distributions que l'on retrouve à la section 3.5. Une asymétrie nulle signifie que la distribution est symétrique, puisque dans ce cas la moyenne équivaut à la médiane, et le numérateur est nul.

Par exemple, voici les moyennes, médianes et écarts-types du nombre d'années d'instruction des femmes et des hommes dans le General Social Survey :

Statistiques	Hommes	Femmes
Moyenne	13,56	13,21
Médiane	13,12	12,66
Écart-type	2,95	2,91

Et voici les coefficients d'asymétrie pour ces deux distributions :

Hommes

$$\text{Asymétrie} = \frac{3(\overline{X} - \text{Md})}{s} = \frac{3(13,56 - 13,12)}{2,95} = \frac{3(0,44)}{2,95} = \frac{1,32}{2,95} = 0,45$$

Femmes

$$\text{Asymétrie} = \frac{3(\overline{X} - \text{Md})}{s} = \frac{3(13,21 - 12,66)}{2,91} = \frac{3(0,55)}{2,91} = \frac{1,65}{2,91} = 0,57$$

Le coefficient d'asymétrie indique que les distributions du niveau d'instruction des femmes et des hommes ont une asymétrie positive, avec cependant une asymétrie beaucoup plus marquée chez les femmes que chez les hommes (0,57 chez les femmes et 0,45 chez les hommes).

Mise en garde : Il ne faut pas évaluer l'asymétrie en comparant seulement les moyennes et les médianes ou en calculant les coefficients d'asymétrie. Les comparaisons moyennes-médianes et les coefficients d'asymétrie peuvent être trompeurs dans le cas des distributions bizarres. Il faut toujours vérifier visuellement la distribution à l'aide d'un graphique. Dans l'exemple précédent, on considérerait ordinairement à la fois les graphiques des distributions et les coefficients d'asymétrie dans le niveau d'instruction.

4.3 Les scores standardisés ou scores-Z

Un score standardisé – appelé aussi ***score-Z***[3] – mesure à combien d'écarts-types de la moyenne se situe un score donné. Les scores-Z sont particulièrement utiles lorsque l'on compare des scores provenant de distributions dont les moyennes et les écarts-types sont différents. Par exemple, qui de ces deux étudiants de l'Université Yale a la meilleure note par rapport à sa classe : Bill avec une note de 87 dans une classe avec une moyenne de 81 et un écart-type de 6 ; ou Hillary avec une note de 83 dans une classe avec une moyenne de 76 et un écart-type de 4 ? Nous pouvons répondre à cette question en standardisant les notes et en comparent les scores-Z.

Pour convertir un score en un score-Z :

1. Soustrayez la moyenne du score.
2. Divisez cette différence par l'écart-type.

Le résultat obtenu est le score standardisé.

En voici la formule :

$$Z_i = \frac{X_i - \overline{X}}{s}$$

lorsque Z_i = le score standardisé du i^e cas

X_i = le score du i^e cas

$\overline{X}$ = la moyenne

s = l'écart-type

Évidemment, parce que les scores peuvent être plus grands ou moins grands que la moyenne, les scores-Z peuvent être soit positifs, soit négatifs. Une valeur positive signifie que le score est supérieur à la moyenne ; à l'inverse, une valeur négative signifie que le score est

3. Pour être plus précis, il faudrait mentionner qu'en fait les scores-Z concernent les scores des seules variables distribuées normalement, alors que les scores standardisés réfèrent aux scores de n'importe quelle variable. Toutefois, les termes score-Z et score standardisé

inférieur à la moyenne. Voici les scores-Z correspondant aux résultats de Bill et d'Hillary, respectivement 87 et 83 :

Note de Bill	Note d'Hillary
$Z_i = \frac{X_i - \overline{X}}{s}$	$Z_i = \frac{X_i - \overline{X}}{s}$
$= \frac{87 - 81}{6}$	$= \frac{83 - 76}{4}$
$= \frac{6}{6}$	$= \frac{7}{4}$
$= 1{,}00$	$= 1{,}75$

La note de Bill est respectable, car elle se situe 1,00 écart-type au-dessus de la moyenne de sa classe. Cependant la note de 83 d'Hillary est plus impressionnante puisqu'elle se trouve 1,75 écart-type au-dessus de la moyenne de sa classe. La note de 83 d'Hillary est meilleure que la note de 87 de Bill car elle est plus éloignée de la moyenne lorsque cette distance est mesurée en écart-type.

Une variable standardisée est une variable dont les scores ont tous été convertis en scores standardisés. C'est-à-dire que chaque score a été transformé pour correspondre au nombre précis d'écarts-types qui le séparent de la moyenne. Toutes les variables standardisées ont donc la même échelle, avec une moyenne de 0 et un écart-type de 1,00. C'est-à-dire que toutes les distributions de scores-Z ont une moyenne de 0 et un écart-type de 1,00. Les scores-Z ne changent pas la position relative des scores. Les scores élevés restent relativement élevés et les scores faibles restent relativement faibles.

Voici le calcul des scores-Z pour six cas ($\overline{X} = 68$ et $s = 2{,}61$) que nous avons vus plus tôt dans ce chapitre :

X_i	$X_i - \overline{X}$	$(X_i - \overline{X})/s = Z_i$
64	$64 - 68 = -4$	$-4/2{,}61 = -1{,}532$
68	$68 - 68 = 0$	$0/2{,}61 = 0{,}000$
70	$70 - 68 = 2$	$2/2{,}61 = 0{,}766$
71	$71 - 68 = 3$	$3/2{,}61 = 1{,}149$
69	$69 - 68 = 1$	$1/2{,}61 = 0{,}383$
66	$66 - 68 = -2$	$-2/2{,}61 = -0{,}766$
Total	0,000	

Notez que la somme de ces scores standardisés est de 0. Notez également que, puisque $\frac{\Sigma Z_i}{N} = \frac{0}{6} = 0$, la moyenne de ces scores standardisés est aussi de 0. Si vous le désirez, je vous laisse calculer l'écart-type de ces scores-Z mais je peux vous assurer que l'écart-type est de 1,00. Toutes les distributions de scores-Z ont une moyenne de 0 et un écart-type de 1,00.

Les scores-Z possèdent une autre caractéristique intéressante. La somme du carré des scores-Z est égale à N-1 ou N_1, les degrés de liberté de la variance, selon la formule utilisée pour calculer l'écart-type. Dans notre exemple, $\Sigma Z^2 = N\text{-}1$. Nous utiliserons cette propriété lorsque, dans le chapitre 10, nous étudierons les coefficients de corrélation.

Un avertissement (assez évident) concernant les scores-Z : dans la mesure où les scores-Z reposent sur la moyenne et l'écart-type, ils n'ont de sens que pour des variables d'intervalles/ratio. Ne calculez ni n'employez de scores-Z pour des variables nominales ou ordinales.

Un dernier avertissement : rappelez-vous que convertir des scores en scores standardisés *ne transforme pas* pour autant une distribution « non normale » en une distribution normale. Si une variable n'est pas distribuée normalement, la distribution de ses scores standardisés ne sera pas plus « normale ».

4.4 La distribution normale

Il existe un type particulier de distribution qui est à ce point important que nous lui devons une attention toute spéciale : la distribution normale. Sans doute savez-vous déjà qu'une distribution normale est symétrique et qu'elle se présente sous la forme d'une cloche. Il est possible toutefois que vous ignoriez qu'il existe plusieurs distributions normales et que ce ne sont pas toutes les distributions en forme de cloche, symétriques, qui peuvent être appelées normales – mais seulement celles qui respectent la formule compliquée qui suit :

$$Y = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$

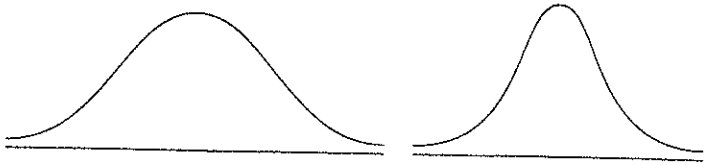
Heureusement nous ne sommes pas obligés d'utiliser cette formule pour comprendre ce que sont les distributions normales. Nous n'aurons donc pas à en connaître beaucoup sur cette formule. (Vous étudierez plus en profondeur cette formule dans un cours ou dans un manuel portant sur les statistiques avancées.) Je veux cependant

que nous observions cette formule afin que vous puissiez constater que les distributions normales ne dépendent que de la moyenne μ et de l'écart-type σ. (Notez que j'use ici des notations se rapportant aux populations plutôt qu'aux échantillons.) Les autres termes, e (la base des logarithmes naturels, ou 2,71828...) et π (pi, ou 3,14159...), sont des constantes. En fait, il existe un nombre infini de distributions normales, une pour chaque combinaison possible d'une moyenne et d'un écart-type.

Une distribution normale avec une moyenne de μ et un écart-type de σ est notée $N(\mu, \sigma)$. Si l'ensemble des scores à un examen a une distribution $N(85,12)$, ces scores formeront une distribution normale avec une moyenne de 85 et un écart-type de 12. Pour une moyenne donnée (μ), une distribution normale peut être haute et étroite (lorsque σ est petit) ou basse et large (si σ est grand).

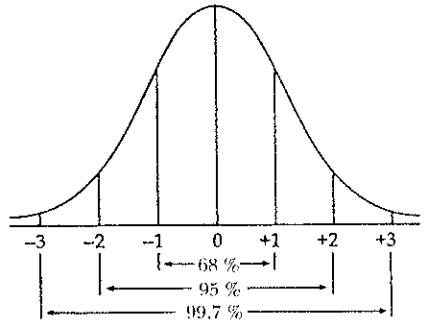
La figure 4.2 donne l'exemple de deux distributions, chacune normale mais avec des écarts-types différents. La distribution normale de gauche affiche, bien sûr, un écart-type plus grand que la distribution de droite. La valeur de la moyenne μ déplace ces distributions vers la gauche (pour une μ plus faible) ou vers la droite (pour une μ plus grande). Dans la mesure où toutes les distributions normales sont symétriques, la moyenne se trouve en tout temps au centre de la distribution et équivaut toujours à la médiane et au mode. Peu importe la moyenne et l'écart-type, les points d'inflexion où la courbe normale passe d'une forme convexe à une forme concave se situent toujours à un écart-type de la moyenne.

Figure 4.2. Distributions normales avec des écarts-types différents



Bien que les distributions normales ne soient pas identiques, elles partagent toutes une importante caractéristique : à un nombre d'écarts-types donné se trouve toujours associée une proportion égale de scores. Ainsi, par exemple, l'intervalle d'un écart-type de part et d'autre de la moyenne comprend 68 % des cas ; à l'intérieur de 2 écarts-types on trouve un peu plus que 95 % des cas ; et à l'intérieur de 3 écarts-types on trouve 99,7 % des cas. La figure 4.3 présente ces proportions. Vous les utiliserez assez fréquemment pour vouloir les mémoriser : 68–95–99,7.

Figure 4.3. Aires sous une courbe normale



Il est quelquefois plus utile de commencer avec un pourcentage donné puis de décrire plus précisément le nombre d'écarts-types qui délimitent l'intervalle comprenant ce pourcentage. Ainsi par exemple, dans une distribution normale, 95 % des cas se trouvent à l'intérieur de 1,96 écart-type de la moyenne et 99 % des cas se trouvent à l'intérieur de 2,58 écarts-types de la moyenne. Nous reverrons ces distances particulières de la moyenne – 1,96 et 2,58 – lorsque nous parlerons des intervalles de confiance à la section 4.6.

Malgré son nom, une distribution normale n'est pas plus « juste » ou « convenable » que n'importe quelle autre distribution. Un nombre étonnant de variables que l'on peut mesurer dans le monde « naturel » suivent des distributions qui sont à peu près normales : le nombre de grains dans les épis de maïs, le nombre de cheveux sur la tête d'hommes et de femmes, le nombre de fourmis dans les fourmilières, le nombre d'étoiles dans les galaxies... et ainsi de suite. Les distributions des résultats à plusieurs examens normalisés sont également « normaux », bien que cela ne soit guère surprenant étant donné que les notes sont attribuées de façon à obtenir une distribution normale. Hélas, les distributions normales sont plus rares dans le domaine social qu'elles peuvent l'être dans les domaines de la biologie et de la physique. Plusieurs des variables dont nous usons tels le revenu et l'instruction ont une forme plutôt asymétrique.

Néanmoins, la distribution normale joue, comme nous aurons l'occasion de le voir à plusieurs reprises dans ce livre, un rôle central dans le raisonnement statistique en sciences sociales. Même les variables qui n'ont pas une distribution normale peuvent produire des statistiques qui sont distribuées presque normalement. Nous utiliserons cette importante particularité dans la section suivante lorsque nous aborderons les méthodes permettant de généraliser des données d'échantillon à la population dont elles sont tirées.

4.5 Les distributions d'échantillonnage

Il est possible d'utiliser des distributions de données d'échantillon afin de décrire la population de laquelle fut tiré l'échantillon. Mais, dans un premier temps, il nous faut savoir ce qu'est une distribution d'échantillonnage. Pour cela, un exemple sera d'un grand secours. Considérez la variable QI mesurée par les scores à un test normalisé de quotient intellectuel. J'ai choisi comme exemple le quotient intellectuel, pas tant parce que je « crois » au test de QI, mais seulement parce qu'il convient particulièrement à mon explication. Contrairement à la plupart des variables avec lesquelles nous travaillons lorsque nous disposons de données d'échantillon, nous savons comment se distribuent les scores de QI dans la population. Les tests de QI sont notés de façon à ce que, si nous les administrons à une vaste population, tel l'ensemble des Américains adultes, la distribution des scores prendra la forme d'une distribution normale semblable à celles que nous avons observées dans la section précédente. De plus, le score moyen pour une population de cette taille sera 100. En notation mathématique : $\mu = 100$.

Livrons-nous à ce que les Allemands nomment une expérience *gedanken* – une expérience en pensée, une expérience imaginaire. Supposez que nous administrions un test de QI aux 200 millions d'Américains adultes. Bien sûr, cela serait guère pratique. En fait ce serait virtuellement impossible. Mais supposons qu'il s'agisse là d'une étude gigantesque. Comme je l'ai mentionné au paragraphe précédent, nous découvrirons alors que ces 200 millions de scores forment une distribution normale avec une moyenne de 100.

Supposons maintenant que nous fassions quelque chose de possible. Supposons que nous choisissions au hasard un échantillon de, disons, 1 500 Américains adultes et que nous mesurions le QI de chacun d'eux. Comment seraient distribués ces 1 500 scores ? Quelle serait la moyenne de l'échantillon $\overline{X}$? Réponse : nous ne pouvons en être certains tant que nous n'examinons pas la véritable distribution des 1 500 scores qui composent l'échantillon.

Il serait tentant – mais fallacieux – de penser qu'un tel échantillon de 1 500 cas, à la manière de la population de laquelle il aurait été extrait, se distribuerait de façon normale et afficherait une moyenne de 100. Bien qu'il soit possible de trouver une distribution normale où $\overline{X} = 100$, cela est peu probable, du moins exactement. Après tout, puisque l'échantillon est sélectionné aléatoirement, il se peut que – uniquement par l'effet du hasard – cet échantillon soit composé de 1 500 adultes dont les QI se situeraient légèrement au-dessus de la moyenne de la population (100). Ou peut-être pourrions-

nous, par le simple hasard, obtenir un échantillon dont la moyenne des QI se situe quelque peu en deçà de la moyenne de la population. Il est même possible de retrouver dans un même échantillon les 1 500 QI les plus élevés des États-Unis (et ainsi $\overline{X}$ très fort). C'est extraordinairement peu probable, mais cela pourrait survenir. Et, de la même façon, l'échantillon pourrait être composé des 1 500 adultes les moins intelligents (ce qui nous donnerait un $\overline{X}$ très faible). Cela aussi serait extraordinairement improbable, mais cela aussi pourrait arriver. Il existe des trillions et des trillions d'autres échantillons qui seraient également possibles, chacun présentant sa propre distribution des QI et sa propre moyenne. Bien que la plupart des échantillons seront assez semblables à la population d'où ils proviennent, certains seront cependant fort peu représentatifs de cette population. Un échantillon pourrait, par exemple, ne contenir que des nonnes, ou que des gauchers, ou que des nonnes gauchères. Oui, il existe des légions d'échantillons possibles.

Étant donné que l'échantillonnage se fait de façon aléatoire, nous ne pouvons savoir comment se présentera la distribution d'un échantillon avant d'avoir analysé les scores de cet échantillon. Certes la plupart des échantillons possibles auront des formes plutôt normales et des moyennes se situant quelque part aux alentours de 100, la moyenne de la population. Certains échantillons cependant nous laisseront voir des moyennes élevées, d'autres des moyennes basses. Certains seront même assez différents de la population.

Supposez maintenant que nous fassions quelque chose de totalement impossible (rappelez-vous : ceci est une expérience imaginaire). Supposez que nous choisissions tous les échantillons de 1 500 personnes qu'il soit possible de tirer. Il y a des trillions et des trillions d'échantillons différents. Vous-même seriez dans certains de ces échantillons si vous êtes Américains (pas dans la plupart cependant). Je serais moi-même dans certains échantillons (mais, ici aussi, pas dans la plupart). Nous pourrions même nous retrouver ensemble dans quelques échantillons. (Il est effrayant de penser que, dans certains échantillons, je pourrais me retrouver aux côtés de Madonna et de Michael Jackson.)

Supposez, de plus, que nous calculions la moyenne de chacun de ces trillions d'échantillons et qu'alors nous dressions la distribution de ces trillions de moyennes d'échantillon. Ce que nous obtiendrions serait ***une distribution d'échantillonnage***. D'un point de vue plus général, une distribution d'échantillonnage est une distribution de statistiques (dans notre exemple, de moyennes) provenant de tous les échantillons possibles d'une taille donnée (ici, 1 500 cas) que l'on peut tirer d'une population précise (ici, les 200 millions d'Américains

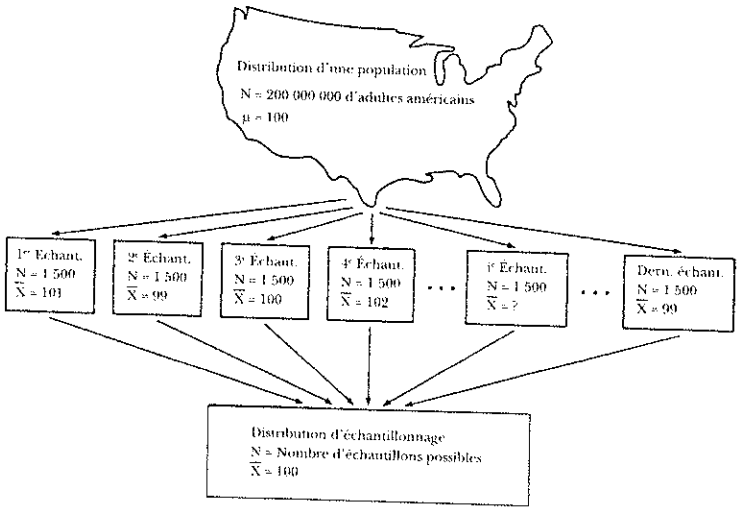
adultes). Une distribution d'échantillonnage de la moyenne comprend les moyennes de tous les échantillons possibles de taille N.

Remarquez qu'il existe trois sortes de distributions qu'il faut distinguer :

La distribution d'une population :	La distribution des scores dans une population.
La distribution d'un échantillon :	La distribution des scores à l'intérieur d'un échantillon d'une taille donnée.
La distribution d'échantillonnage :	La distribution d'une statistique quelconque (par exemple, la moyenne) de tous les échantillons possibles d'une taille donnée.

Bien que les deux dernières aient des noms qui peuvent se ressembler, elles sont bien différentes. La figure 4.4 montre un schéma expliquant ces trois distributions.

Figure 4.4. Distribution d'une population, distributions d'échantillons et distribution d'échantillonnage



Même si, en réalité, nous ne pouvons tirer d'une population aussi importante que 200 millions tous les échantillons possibles d'une taille donnée (il y a beaucoup trop d'échantillons possibles), les statisticiens se sont servis des mathématiques pour projeter ce que serait la distribution d'échantillonnage de certaines statistiques importantes. Si nous tenons pour acquis quelques postulats à propos de la population et si nous tirons aléatoirement de celle-ci un échantillon, nous serons en mesure de connaître la distribution, pour tous les échantillons qu'il est possible de tirer, de certaines statistiques comme la moyenne. Exprimé autrement, nous connaîtrons leur distribution d'échantillonnage.

La distribution d'échantillonnage de la moyenne d'échantillons aléatoires possède des caractéristiques extrêmement importantes. À mesure qu'augmente la taille N de l'échantillon, la distribution d'échantillonnage de la moyenne s'apparente de plus en plus à une distribution normale, dont la moyenne est semblable à celle de la population et dont l'écart-type est de $\frac{\sigma}{\sqrt{N}}$. On peut décrire cette distribution symbolique par $N(\mu, \frac{\sigma}{\sqrt{N}})$. Les statisticiens nomment cette tendance le théorème de la limite centrale, un des concepts les plus importants des statistiques.

En fait la distribution d'échantillonnage de la moyenne est assez semblable, dans les cas où la taille des échantillons est de 30 cas ou plus, à une distribution normale. Cela est vrai peu importe la forme que prend, à l'intérieur de la population, la distribution de la variable. Aussi, même si une variable n'est pas distribuée normalement à l'intérieur de la population, la moyenne de toutes les moyennes d'échantillons possibles sera identique à celle de la population, et l'écart-type de la distribution des moyennes de tous les échantillons possibles équivaudra à $\frac{\sigma}{\sqrt{N}}$.

Rappelez-vous que dans la section précédente nous avons vu que, dans une distribution normale, une proportion donnée de cas sont inclus dans un intervalle délimité par un nombre précis d'écarts-types par rapport à la moyenne. Mettons maintenant cela en perspective avec le théorème de la limite centrale. Le théorème de la limite centrale nous permet de connaître le nombre de « cas » (ici, le nombre de moyennes d'échantillons) se retrouvant dans un intervalle délimité par un nombre donné d'écarts-types par rapport à la moyenne de la distribution d'échantillonnage. C'est précisément ce que nous ferons dans la prochaine section.

Mais avant, voyons quelques termes et formules. L'écart-type d'une distribution d'échantillonnage revêt une importance si grande que nous lui donnons un nom particulier – ***l'erreur-type*** désignée par le symbole $\sigma_{\bar{x}}$. Comme nous venons précisément de le voir il y a

quelques lignes, l'erreur-type de la moyenne – c'est-à-dire l'écart-type de la distribution d'échantillonnage des moyennes de tous les échantillons d'une taille précise qu'il est possible d'extraire aléatoirement d'une population – nous est donnée par cette formule :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

Par exemple, la variable mesurant le temps passé quotidiennement devant la télévision a un écart-type de 2,14, basé sur les 1940 cas du General Social Survey. En utilisant s pour estimer σ, nous trouvons son erreur-type :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

$$= \frac{2,14}{\sqrt{1940}}$$

$$= \frac{2,14}{44,045}$$

$$= 0,049$$

L'erreur-type – l'écart-type de la distribution d'échantillonnage – est donc de 0,049. Nous userons abondamment de l'erreur-type dans la section suivante. En terminant, sachez que j'ai conservé trois décimales à l'erreur-type étant donné que, dans la prochaine section, nous nous en servirons pour calculer d'autres statistiques.

4.6 Les intervalles de confiance

La meilleure estimation que nous puissions avoir de la moyenne de la population est la moyenne d'un échantillon. Cependant, parce que le hasard préside au choix de l'échantillon et que des échantillons choisis de cette façon varient entre eux, nous savons fort bien que cette estimation a bien peu de chances d'être rigoureusement exacte. La moyenne de la population peut, en fait, se trouver en dessous ou au-dessus de la moyenne de notre échantillon. Il paraît donc extrêmement utile de connaître l'intervalle, de part et d'autre de la moyenne, à l'intérieur duquel il est probable, croyons-nous, de trouver la moyenne de la population. Cet intervalle est appelé un intervalle de confiance.

C'est l'erreur-type qui nous permet de trouver l'intervalle de confiance. Dans le cas de notre expérience « imaginaire » sur les QI, nous savions quelle était la moyenne de la population : $\mu = 100$. Cela est inhabituel toutefois. En recherche, nous ne connaissons pas, la plupart du temps, où se trouve la moyenne de la population. Pour l'estimer, nous utilisons la moyenne de l'échantillon. Règle générale donc, nous nous servons d'une statistique d'échantillon afin d'apprécier le paramètre d'une population. Après tout, si nous savions d'avance la moyenne de la population, il ne serait d'aucun intérêt de manipuler comme nous le faisons des données d'échantillons.

Le théorème de la limite centrale démontre que plus N est grand, plus la distribution d'échantillonnage de la moyenne ressemble à une distribution normale avec une moyenne égale à la moyenne de la population μ et un écart-type $\frac{\sigma}{\sqrt{N}}$ (c'est-à-dire $\sigma_{\bar{x}}$).

Rappelez-vous notre discussion de la section 4.4 à propos des distributions normales. Nous avions alors mentionné que 95 % des scores sont compris à l'intérieur de 1,96 écart-type de part et d'autre de la moyenne. Afin d'obtenir cet ***intervalle de confiance à 95 %***, il suffit simplement de soustraire 1,96 écart-type de la moyenne de l'échantillon (pour connaître la limite inférieure de l'intervalle) et d'additionner 1,96 écart-type à la moyenne de l'échantillon (pour connaître sa limite supérieure). La formule serait la suivante :

$$\text{L'intervalle de confiance à 95 \%} = \bar{X} \pm 1,96\sigma_{\bar{x}}$$

Nous pouvons alors affirmer avec une certitude de 95 % que la moyenne de la population se retrouve dans cet intervalle.

Les chercheurs préfèrent parfois travailler avec un intervalle de confiance plus large, avec ***un intervalle de confiance à 99 %***. Souvenez-vous que nous avons dit, à la section 4.7, que 99 % des scores se situent à l'intérieur de 2,58 écarts-types de chaque côté de la moyenne. Aussi nous est-il possible, grâce à la formule qui suit, d'obtenir l'intervalle de confiance à 99 %.

$$\text{L'intervalle de confiance à 99 \%} = \bar{X} \pm 2,58\sigma_{\bar{x}}$$

Nous sommes certains à 99 % que la moyenne de la population se situera à l'intérieur de cet intervalle.

Pour un exemple d'intervalles de confiance, prenez une variable comme le nombre d'heures quotidiennes passées à regarder la télévision. Dans le General Social Survey, cette variable affichait une moyenne de 2,90 et un écart-type de 2,14. Dans la section précédente, nous avons déterminé l'erreur-type de la moyenne : $\sigma_{\bar{x}} = 0,049$.

Trouvons maintenant l'intervalle de confiance de 95 % :

$$\text{L'intervalle de confiance de 95 \%} = \overline{X} \pm 1{,}96\sigma_{\bar{x}}$$
$$= 2{,}90 \pm 1{,}96(0{,}049)$$
$$= 2{,}90 \pm 0{,}096$$
$$= 2{,}80 \text{ à } 3{,}00$$

Ainsi, l'intervalle de confiance de 95 % est celui qui se situe entre 2,80 et 3,00. Quatre-vingt-quinze pour cent de toutes les moyennes d'échantillons possibles seront donc comprises dans cet intervalle.

De la même façon il est possible de trouver l'intervalle de confiance de 99 % :

$$\text{L'intervalle de confiance de 99 \%} = \overline{X} \pm 2{,}58\sigma_{\bar{x}}$$
$$= 2{,}90 \pm 2{,}58(0{,}049)$$
$$= 2{,}90 \pm 0{,}126$$
$$= 2{,}77 \text{ à } 3{,}03$$

Nous pouvons être sûrs à 99 % de retrouver la moyenne de la population entre 2,77 et 3,03.

Sans même vous en rendre compte, vous voilà en train de travailler avec des statistiques inférentielles. Vous inférez, à l'aide de données d'échantillons, des caractéristiques de la population de laquelle proviennent ces échantillons.

4.7 Avertissement concernant les statistiques univariées

Calculer des mesures de tendance centrale, des variances, des écarts-types, des scores-Z et des intervalles de confiance est, en fait, très simple. Toutefois trois problèmes peuvent survenir : 1. les niveaux de mesure impropres ; 2. les catégories de grandeur inégale (cela comprend les catégories dites « ouvertes ») ; 3. les données manquantes. Cette section traite, tour à tour, de chacun de ces trois cas problématiques.

En premier lieu, il faut savoir que les logiciels de statistiques ne tiennent généralement pas compte, dans leurs analyses, des niveaux de mesure. La moyenne, l'écart-type et la variance ne conviennent pas, règle générale, à des variables nominales ou ordinales. De la même façon, la médiane n'est guère appropriée pour des variables nominales. De plus, parce que la moyenne est habituellement inadéquate lorsque nous travaillons avec des variables nominales ou ordinales, les intervalles de confiance par rapport à la moyenne ne sont pas appropriés pour ces types de variables. L'ordinateur calculera allègrement tout ce qui lui sera demandé même si cela est insensé – comme des moyennes, des écarts-types et des variances pour des variables ordinales, ou encore des médianes pour des variables nominales.

Vous devez donc faire preuve d'une grande vigilance lorsque vient le temps d'interpréter les statistiques fournies par un logiciel. Vous devez porter une attention particulière pour ne vous servir que des mesures de tendance centrale, des écarts-types, des variances et des intervalles de confiance qui sont appropriées au niveau de mesure de vos variables. Règle générale, ces statistiques réclament des niveaux de mesure d'intervalles ou de ratio. Comme toujours le moment le plus important dans l'analyse statistique est celui de la réflexion. Ne laissez donc pas l'ordinateur réfléchir pour vous.

Un second problème survient lorsque nous tentons d'obtenir la moyenne, l'écart-type ou la variance pour des variables dotées de catégories de grandeur inégale. Regardez, par exemple, le tableau 4.1 décrivant la distribution de fréquences du nombre d'enfants des répondants au sondage. Vous remarquerez que l'encodage de la variable met dans une même catégorie, portant le score 8, les 25 cas qui ont répondu avoir plus de 7 enfants. Cette catégorie « composite » peut très bien contenir des répondants ayant 9, 10, 11 enfants, et peut-être même plus. Comme des scores aussi élevés sont tous réduits à 8, les mettre dans la même catégorie a pour effet de diminuer la moyenne et de réduire l'écart-type et la variance. Par bonheur, étant donné que le nombre de répondants ayant 8 enfants ou plus est plutôt bas (seulement 25 cas, donc 0,9 % du nombre total de cas), cet effet est, ici, probablement bien faible. Nous devons néanmoins être attentifs à de tels effets, aussi petits soient-ils. Lorsque la proportion des cas se retrouvant dans une catégorie ouverte ou composite est très grande, nous devons user de la plus grande prudence.

Tableau 4.1. Nombre d'enfants (en fréquences)

Nombre d'enfants	f
0	882
1	461
2	770
3	420
4	222
5	94
6	48
7	27
8 et plus	25
(N)	(2 889)

Un tel problème se pose dès que les catégories d'une variable sont de grandeur inégale. Typiquement, cela survient lorsque nous avons affaire à des catégories composites ou encore (comme c'était le cas pour le nombre d'enfants, notre exemple précédent) à des catégories ouvertes. Bien qu'il n'y ait pas de solution simple à ce problème, il doit scrupuleusement être pris en considération lors de l'interprétation d'une moyenne ou d'une mesure de variation. Encore une fois, la réflexion (la vôtre, pas celle de votre ordinateur) est essentielle lors d'analyses statistiques. Réfléchissez aux variables que vous analysez ainsi qu'à ce qu'impliquent pour les résultats de votre analyse leur niveau de mesure et leurs valeurs.

Un dernier problème : les données manquantes. Règle générale, de la même façon que vous procédiez lors du calcul des pourcentages, excluez les valeurs manquantes lorsque vous calculez des mesures de tendance centrale et des mesures de variation. L'existence de valeurs telles que « Ne sait pas » et « Pas de réponse » brouille toutes les statistiques univariées. Habituellement codées avec des nombres élevés (7, 8 ou 9 pour les codes sans décimales et 97, 98 ou 99 pour les codes décimaux), la présence de ces valeurs gonflerait indûment la moyenne, l'écart-type et la variance.

Résumé du chapitre 4

Voici ce que nous avons appris dans ce chapitre :

- La variance d'une variable dichotomique codée 0 et 1 correspond à la proportion de cas codés 0 multipliée par la proportion de cas codés 1.
- La variance et l'écart-type mesurent la dispersion des scores par rapport à la moyenne. Il convient de les calculer seulement pour des variables d'intervalles/ratio.
- Les scores extrêmes produisent des effets considérables sur la variance et sur l'écart-type.
- Les scores de variables d'intervalles/ratio peuvent être convertis en scores-Z (ou scores standardisés). Il suffit pour ce faire de les soustraire de la moyenne et de diviser cette différence par l'écart-type. Les scores standardisés nous permettent de comparer l'emplacement relatif des scores à l'intérieur des distributions.
- Une variable standardisée est une variable dont les scores ont été transformés en scores standardisés.
- Les variables standardisées affichent invariablement une moyenne de 0 et un écart-type de 1,00.

- La somme des carrés d'une variable standardisée est égale à N.
- Une distribution normale est symétrique et se présente sous la forme d'une cloche. Toutes les distributions symétriques et en forme de cloche ne sont cependant pas des distributions normales.
- Pour toutes les distributions normales, on trouve toujours la même proportion de scores à l'intérieur d'un intervalle délimité par un nombre donné d'écarts-types par rapport à la moyenne.
- Une distribution d'échantillonnage est la distribution d'une statistique d'échantillon (par exemple, la moyenne) pour tous les échantillons d'une taille donnée qu'il est possible de tirer d'une population précise.
- L'erreur-type est l'écart-type d'une distribution d'échantillonnage.
- Un intervalle de confiance indique les chances qu'un paramètre de la population, comme la moyenne, se situe à l'intérieur d'un espace précis.
- Les données manquantes doivent être exclues de l'analyse
- Nous devons être prudents lorsque nous interprétons des mesures de tendance centrale, des mesures de variation et des intervalles de confiance que nous fournit un ordinateur. Il faut se garder d'employer des statistiques qui ne conviennent pas au niveau de mesure des variables analysées. Nous devons également déceler les effets que peuvent causer des catégories composites ou ouvertes. Enfin, il est important d'exclure les données manquantes.

Principaux concepts et procédures

Termes et idées

mesure de variation
variance
écart par rapport à la moyenne
degré de liberté
écart-type
kurtose
leptokurtique
platykurtique
mésokurtique
asymétrie
score standardisé
score-Z
variable standardisée
distribution normale
distribution d'échantillonnage
théorème de la limite centrale
erreur-type
intervalle de confiance

Symboles

σ et σ^2
s et s^2
P et (1 – P)
Z
$\sigma_{\bar{x}}$

Formules

$$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N}$$

$$s^2 = \frac{\Sigma(X_i - \bar{X})^2}{N - 1}$$

$$\sigma = \sqrt{\text{Variance}}$$

$$\text{Asymétrie} = \frac{3(\bar{X} - \text{Md})}{s}$$

$$Z_i = \frac{X_i - \bar{X}}{s}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

et

$$\text{intervalle de confiance de 95 \%} = \bar{X} \pm 1{,}96\sigma_{\bar{x}}$$

$$\sigma = \sqrt{\frac{\Sigma(X_i - \mu)^2}{N}}$$

$$\text{intervalle de confiance de 99 \%} = \bar{X} \pm 2{,}58\sigma_{\bar{x}}$$

$$s = \sqrt{\text{Variance}}$$

et

$$s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N - 1}}$$

RAPPORT D'ANALYSE N°2

MESURES DE TENDANCE CENTRALE ET DE VARIATION

La plupart du temps on peut indiquer les médianes, moyennes et écarts-types des variables directement dans le texte du rapport sans recourir à un tableau. Toutefois les tableaux sont plus efficaces lorsqu'il s'agit de présenter les moyennes et les écarts-types d'un grand nombre de variables.

Le tableau 1 présente les mesures de tendance centrale et les écarts-types des variables « niveaux d'instruction » et « prestige de l'emploi » pour les hommes et pour leurs parents. Les données proviennent du General Social Survey. La moyenne et la médiane du niveau d'instruction montrent toutes deux une augmentation substantielle dans le niveau d'instruction moyen. Les hommes ont un niveau d'instruction médian d'un an de plus que leurs parents, et le niveau d'instruction moyen est plus élevé de deux ans. Toutefois, la variation du niveau d'instruction, mesurée par l'écart-type, a considérablement décru, surtout quand elle est comparée avec celle des pères. Les répondants indiquent en moyenne que le prestige lié à leur emploi est légèrement plus bas que celui de leurs pères, bien que la variation des scores des répondants soit beaucoup plus grande. Les scores de prestige d'emploi légèrement plus bas des répondants du GSS comparés à ceux de leurs pères s'expliquent peut être par les différences d'âge et de situation dans le cycle de vie.

TABLEAU 1 ICI

Mettre le tableau à la fin du rapport :