Cours de statistique descriptive

1. Analyse univariée

Support de cours destiné aux étudiants de la licence MOMR :

Université Charles-de-Gaulle Lille 3
UFR MSES

O. Torrès

Année universitaire 2007-8

Version du 2 octobre 2007, 09:34

Préface

Public. Étudiants de 1re année de licence MIASHS (Mathématiques et informatique appliquées aux sciences humaines et sociales)

But visé. Donner aux étudiants quelques outils de statistique descriptive leur permettant un premier contact avec les faits relevant du domaine des sciences humaines et sociales. L'accent est mis sur la compréhension de la fonction de chacun de ces outils. Cette compréhension est pour partie acquise en démontrant les propriétés attachées aux outils qui sont présentés.

Bibliographie.

- *Statistique descriptive*, Bernard Delmas, Nathan Université, 1996, `519.53 DEL`
- *Statistiques descriptives*, Gérard Chauvat et Jean-Philippe Réau, Armand Colin, coll. Flash U, 1992, `519.53 CHA`
- *Cours de statistique descriptive*, Gérard Calot, Dunod, coll. Décision, deuxième édition, 1973, `519.53 CAL`

Table des matières

Préface **i**

I Introduction et concepts de base **1**

1 Introduction **3**

1.1 Objet de la statistique descriptive 3

1.2 Statistique descriptive et sciences humaines, économiques et sociales 3

2 Principaux concepts **5**

2.1 Population, individu 5

2.2 Variable statistique ou caractère, modalité ou valeur 6

2.2.1 Variable ou caractère statistique 6

2.2.2 Modalité ou valeur 6

2.2.3 L'ensemble des modalités d'une variable 7

2.2.3.1 Définition 7

2.2.3.2 La constitution de l'ensemble des modalités 7

2.3 Classification des variables 8

2.3.1 Les classifications traditionnelles 9

2.3.1.1 Variables continues, variables discrètes 9

2.3.1.2 Variables qualitatives, variables quantitatives 9

2.3.1.3 Limites des classifications usuelles 10

2.3.2 Classifications pertinentes des variables 11

2.3.2.1 Variable nominale 11

2.3.2.2 Variable ordinale 12

2.3.2.3 Variable numérique 12

2.3.2.4 Commentaires 13

2.4 Transformation des variables 13

3 Les données et leur organisation **15**
3.1 Les données brutes . . . 15
3.2 Tableau élémentaire . . . 15
3.3 Un premier traitement statistique . . . 16
3.3.1 Petit nombre de modalités observées . . . 16
3.3.1.1 Tri à plat et effectifs des modalités . . . 16
3.3.1.2 Le tableau statistique . . . 17
3.3.2 Grand nombre de modalités observées . . . 17
3.3.2.1 Le regroupement par classes . . . 18
3.3.2.2 Tri à plat et effectifs de classes . . . 19
3.3.2.3 Tableau statistique . . . 20

II Analyse statistique univariée **23**

4 Effectifs, fréquences **27**
4.1 Effectifs . . . 27
4.2 Fréquences . . . 28
4.2.1 Définition . . . 28
4.2.2 Interprétation . . . 28
4.2.3 Propriétés . . . 28
4.2.4 Remarque . . . 28
4.3 Les densités : effectifs ou fréquences unitaires . . . 29
4.4 Effectifs cumulés . . . 30
4.4.1 Défintion . . . 30
4.4.2 Interprétation . . . 30
4.4.3 Propriétés . . . 31
4.4.4 Remarque . . . 31
4.5 Fréquences cumulées . . . 31
4.5.1 Définition . . . 31
4.5.2 Interprétation . . . 31
4.5.3 Propriétés . . . 32
4.5.4 Remarque . . . 33

5 Représentations graphiques **35**
5.1 Généralités . . . 35
5.2 Quelques représentations graphiques usuelles . . . 37
5.2.1 Diagrammes en secteurs . . . 38
5.2.1.1 Principe . . . 38

5.2.1.2 Réalisation . . . 38
5.2.1.3 Lecture . . . 39
5.2.1.4 Exemple . . . 39
5.2.2 Diagrammes en bâtons . . . 39
5.2.2.1 Principe . . . 39
5.2.2.2 Réalisation . . . 40
5.2.2.3 Lecture . . . 40
5.2.2.4 Exemples . . . 41
5.2.2.4.1 Variable numérique sans regroupement par classes 41
5.2.2.4.2 Variable numérique et regroupement par classes . 41
5.2.3 Histogramme des fréquences . . . 42
5.2.3.1 Le principe . . . 42
5.2.3.2 Réalisation . . . 42
5.2.3.3 Lecture . . . 43
5.2.3.4 Exemple . . . 44
5.2.4 Graphiques représentant les fréquences cumulées croissantes . . . 44
5.2.4.1 Graphe de la fonction de répartition . . . 45
5.2.4.2 Le polygone des fréquences cumulées . . . 46

6 Description numérique des données 49
6.1 Généralités . . . 49
6.2 Indicateurs de position (ou de tendance centrale) . . . 49
6.2.1 Le mode . . . 50
6.2.1.1 Définition . . . 50
6.2.1.2 Interprétation . . . 50
6.2.1.3 Propriétés . . . 50
6.2.2 La médiane . . . 51
6.2.2.1 Définitions . . . 51
6.2.2.2 Exemple . . . 52
6.2.2.3 Interprétation . . . 53
6.2.2.4 Propriétés . . . 53
6.2.2.5 Remarques . . . 55
6.2.2.6 Caractérisation de la médiane . . . 56
6.2.2.7 Définition et détermination de la médiane cas des regroupements par classes . . . 56
6.2.3 La moyenne arithmétique . . . 59
6.2.3.1 Définition . . . 60
6.2.3.2 Propriétés . . . 60

6.2.3.3 Interprétation 62
6.2.4 Comparaison 62
6.2.4.1 Le mode 63
6.2.4.2 La médiane 64
6.2.4.3 La moyenne 65
6.3 Indicateurs de dispersion 66
6.3.1 Les indicateurs de dispersion absolue : l'étendue et l'étendue interquartile 67
6.3.1.1 L'étendue 67
6.3.1.2 L'étendue interquatile 69
6.3.2 Les indicateurs de dispersion autour d'une tendance centrale 71
6.3.2.1 Le principe 71
6.3.2.2 L'écart absolu moyen 72
6.3.2.3 La variance et l'écart-type 74
6.3.3 Remarques sur les indicateurs de dispersion 77
6.4 Indicateurs de forme : asymétrie, aplatissement 79
6.4.1 Indicateurs d'asymétrie 80
6.4.1.1 (A)symétrie : définition, interprétation et propriétés 80
6.4.1.2 Mesures d'asymétrie 88
6.4.1.2.1 Les coefficients d'asymétrie de Pearson 88
6.4.1.2.2 Le coefficient d'asymétrie γ_1 88
6.4.1.2.3 Le coefficient d'asymétrie de Yule-Bowley 89
6.4.2 Indicateurs d'aplatissement 89
6.4.2.1 Définition 89
6.4.2.2 Propriétés et interprétation 91
6.5 Indicateurs de concentration 95

Annexe **97**

Première partie

Introduction et concepts de base

Chapitre 1

Introduction

1.1 Objet de la statistique descriptive

La statistique descriptive sert à décrire une population [un (gros) ensemble d'unités statistiques élémentaires] à l'aide d'indicateurs numériques ou de techniques graphiques.

1.2 Statistique descriptive et sciences humaines, économiques et sociales

Divers problèmes trouvent leur origine à l'intersection des statistiques et des phénomèmes relevant des sciences humaines et sociales. (Lire Delmas (1996), Introduction, Section 1.)

Chapitre 2

Principaux concepts

On définit dans cette section les notions de base de la statistique descriptive, telles qu'elles seront utilisées dans la suite du cours.

2.1 Population, individu

Définition 2.1

- *L'individu est l'unité statistique à laquelle on s'intéresse.*
- *La population est l'ensemble fini des individus statistiques que l'on s'apprète à décrire.*

1. Les individus statistiques peuvent être des humains, des êtres vivants (animaux, végétaux), mais aussi des objets (voitures, livres, ...), des entités juridiques (entreprises, départements, pays, ...)
2. La population, au moment de l'étude statistique, doit être définie de façon précise et sans ambiguité. Notemment, à propos de n'importe quoi, on doit pouvoir dire s'il fait ou nom partie de la population (*cf.* Chauvat et Réau (1992), Chapitre 1, Exercices).
3. Exemple : population composée des livres catalogués au 01/10/2002 à la bibliothèque centrale de l'Université Lille 3.
4. Notation : On désigne par N le nombre d'individus (distincts) de la population. Ce nombre est appelé *taille* de la population.
5. Les individus faisant partie de la population peuvent être distingués les uns des autres. Chaque individu de la population est répéré de façon unique par son numéro, qui est un entier compris entre 1 et N, inclus. L'ordre de numérotation n'a pas d'importance.

Exemple : chacun des livres de la bibliothèque est repéré de façon unique par son code-barre. En statistique, on préfèrera attribuer à chacun un nouvel identifiant, qui est un numéro d'ordre, attribué selon des méthodes qui peuvent varier (classement préalable par ordre croissant de code-barre, par exemple).

6. Notation : La population sera notée $\mathscr{P}$ et identifiée avec l'ensemble des numéros attribués aux individus qui la composent. On aura $\mathscr{P} = \{1, 2, \ldots, N\}$.

   Lorsqu'on veut désigner un individu quelconque dans la population, on utilise i qui désigne un entier quelconque compris entre 1 et N. On parlera alors de l'individu i.

2.2 Variable statistique ou caractère, modalité ou valeur

2.2.1 Variable ou caractère statistique

Définition 2.2 *Une* variable *statistique est un* moyen *de décrire chacun des individus de la population. Caractère est synonyme de variable.*

1. Une même population peut être décrite à l'aide de plusieurs variables.
2. Exemple : chacun des livres catalogués à la bibliothèque peut être décrit par son année de publication, la couleur de sa couverture, son nombre de pages, la discipline dont il traite, ses dimensions, son prix à l'achat, ...
3. Notation : On désigne les variables statistiques par des lettres latines majuscules. On parlera des variables X, Y, ...

2.2.2 Modalité ou valeur

Définition 2.3 *Une modalité d'une variable est une des façons possibles d'effectuer la description d'un individu au moyen de cette variable. Valeur est synonyme de modalité.*

1. On évitera la terminologie « valeur » et on lui préfèrera « modalité » car une valeur d'une variable statistique n'est pas nécessairement une valeur numérique.
2. Exemple : si on décrit les livres de la bibliothèque à l'aide de la variable « couleur de la couverture », une modalité (valeur) possible est « bleu ». Si on décrit ces livres à l'aide du nombre de pages, une modalité (valeur) possible est 436.

2.2.3 L'ensemble des modalités d'une variable

2.2.3.1 Définition

Définition 2.4 *L'ensemble des modalités d'une variable est l'ensemble des* différentes *façon possibles de décrire les individus de la population avec la variable.*

1. La définition précédente implique évidemment que les éléments de l'ensemble des modalités sont deux à deux distincts.
2. Notation : Si X dénote une variable statistique, l'ensemble de ses modalités est noté $\mathscr{M}_X$. Un élément typique de cet ensemble est noté x.
3. On dit qu'un individu de la population présente la modalité $x \in \mathscr{M}_X$ de la variable X si cet individu est décrit par x au moyen de la variable X.
4. Exemple : Si on décrit les livres de la bibliothèque par leur dimension (variable X), une modalité x de X sera un triplet de nombres positifs $x = (x_h, x_l, x_e)$ dont les composantes désignent dans l'ordre la hauteur, la largeur et l'épaisseur du livre. Dans ce cas $\mathscr{M}_X$ sera $\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_+$. Si un livre donné de la bibliothèque mesure 21 cm de haut, 15,2 cm de large et a une épaisseur de 4,7 cm, on dira que ce livre présente la modalité (21, 15,2, 4,7) de la variable « dimensions ».

   Si on utilise les disciplines dont ils traitent pour décrire les livres (variable Y), une modalité y de Y sera le nom d'une discipline (par exemple « statistique ») et $\mathscr{M}_Y$ sera l'ensemble de toutes les disciplines représentées dans le fonds de la bibliothèque : $\mathscr{M}_Y = \{$philosophie, histoire, psychologie, économie, statistique, littérature anglaise, $\ldots\}$. Si le livre auquel on s'intéresse est *Théorie générale de l'emploi, de l'intérêt et de la monnaie* (de J. M. Keynes), il présente la modalité « économie » de la variable « discipline traitée ».
5. Mathématiquement, une variable statistique X est une fonction définie sur $\mathscr{P}$ (ensemble de départ) à valeurs dans $\mathscr{M}_X$, (ensemble d'arrivée) réprésentée par

$$\begin{aligned} X : \mathscr{P} &\longrightarrow \mathscr{M}_X \\ i &\longmapsto X(i) \end{aligned}$$

   où $X(i)$ est l'élément de $\mathscr{M}_X$ qui désigne la modalité de la variable X présentée par l'individu i de $\mathscr{P}$. On aura $X(i) = x$ où $x \in \mathscr{M}_X$ si et seulement si l'individu i de la population $\mathscr{P}$ présente la modalité x de la variable X.

2.2.3.2 La constitution de l'ensemble des modalités

1. Il faut apporter un soin particulier à la constitution cet ensemble. Pour chaque variable servant à décrire la population, l'ensemble de ses modalités doit être constitué

de sorte qu'à chaque individu de la population on puisse assigner une modalité et une seule de la variable.

Cela signifie deux choses : (1) il ne doit pas y avoir d'individu dans la population ne présentant aucune modalité de l'une des variables ; (2) il ne doit pas y avoir d'individu dans la population présentant plusieurs modalités d'une même variable.

2. Exemple : On considère la population de livres de la bibliothèque que l'on souhaite décrire par la variable « discipline traitée ». Il se peut qu'à certains livres on puisse rattacher plusieurs disciplines. Ainsi l'ouvrage *Le capital* (de K. Marx) peut être considéré comme un ouvrage de philosophie, de sociologie ou d'économie. Pour de tels livres, il faudra choisir la discipline la plus pertinente, de sorte que ce livre ne présente pas plusieurs modalités de la même variable.

   Pour d'autres livres l'identification de la (ou des) discipline(s) dont ils traitent peut s'avérer plus difficile et il peut être judicieux de rajouter « autre » ou « divers » comme modalité de la variable.

3. Au début d'un traitement statistique, il est souvent important de considérer la nature de l'ensemble des modalités. Les différentes natures que nous aurons à considérer sont les suivantes.
   - L'ensemble $\mathscr{M}_X$ est dit *dénombrable* s'il est possible d'assigner un numéro à chacun de ses éléments, sans en oublier aucun. Sinon cet ensemble est dit non-dénombrable.

     Par exemple, l'ensemble des nombres entiers naturels $\mathbb{N} = \{0, 1, \ldots\}$ est dénombrable, tandis que l'intervalle $[0, 1]$ ne l'est pas.
   - L'ensemble $\mathscr{M}_X$ est dit *ordonné* s'il existe une relation, notée $\preccurlyeq$ dans le cas général, permettant de comparer tous les éléments de $\mathscr{M}_X$ et satisfaisant

     (a) $x \preccurlyeq x \ \forall x \in \mathscr{M}_X$ (la relation $\preccurlyeq$ est reflexive)

     (b) $x \preccurlyeq x'$ et $x' \preccurlyeq x''$ impliquent $x \preccurlyeq x''$ (la relation $\preccurlyeq$ est transitive)

     (c) $x \preccurlyeq x'$ et $x' \preccurlyeq x$ impliquent $x = x'$ (la relation $\preccurlyeq$ est antisymétrique)

     Une relation satisfaisant ces trois points est appelée une relation d'ordre. Par exemple, l'ensemble des réels $\mathbb{R}$ est ordonné par la relation usuelle $\leq$ « *inférieur ou égal à* ».

2.3 Classification des variables

Selon les propriétés de X ou de $\mathscr{M}_X$, il peut être utile d'introduire une classification des variables statistiques.

2.3.1 Les classifications traditionnelles

2.3.1.1 Variables continues, variables discrètes

Une première classification usuelle relève l'aspect continu ou discret d'une variable variable statistique.

1. Une variable X est dite discrète si $\mathscr{M}_X$ est un ensemble dénombrable, c'est à dire si X possède un nombre dénombrable de modalités. On note alors $\mathscr{M}_X$ en faisant la liste de ses éléments, une fois leur numérotation effectuée. Par ordre croissant de numéro, on note x_k le $k^{\,\mathrm{e}}$ élément de $\mathscr{M}_X$.

   Cette situation couvre le cas où $\mathscr{M}_X$ est un ensemble fini et le cas où $\mathscr{M}_X$ est un ensemble infini dénombrable.

2. Exemples :
   - les livres décrits selon leur nombre de pages.
   - les livres décrits selon la discipline à laquelle ils sont rattachés.

3. Les variables qui ne sont pas discrètes sont continues. Une variable X est une variable continue si $\mathscr{M}_X$ est un ensemble non-dénombrable. Pour de telles variables, si on choisit deux modalités, alors toutes les valeurs comprises entre ces deux modalités sont aussi des modalités de la variable.[1]

   Cette propriété n'est pas vérifiée pour les variables discrètes. Ainsi, pour les livres décrits selon la variable « nombre de pages », deux modalités possibles sont 12 et 21. Mais 14.7 qui est une valeurs comprise entre 12 et 21 n'est pas une des modalités de la variable.

4. Les variables qui décrivent des grandeurs liées au temps (nombre d'unités temporelles écoulées entre deux instants caractérisant la vie de l'individu), à l'espace (distances, surfaces, ...), à la masse (poids, ...), *etc*, ou des rapports entre de telles quantités (km/h, kg/cm^2, l/km, *etc*) sont considérées comme des variables continues.

2.3.1.2 Variables qualitatives, variables quantitatives

Une seconde classification repose sur la nature qualitative ou quantitative d'une variable statistique.

1. Une variable est dite qualitative si ses modalités ne sont pas quantifiables à l'aide d'une échelle quelconque.

   Typiquement ces variables décrivent une caractéristique non-quantifiable de façon naturelle et non-ambiguë, assimilée à une *qualité* (*i.e.*, une notion servant à *qualifier*).

[1] Ceci n'est pas tout à fait ce qui définit un ensemble non-dénombrable, mais dans notre contexte, on pourra toujours se ramener à des ensembles non-dénombrables ayant cette propriété.

Exemple : les modalités de la variable « discipline traitée » utilisée pour décrire les livres de la bibliothèque.

2. Une variable est dite quantitative si ses modalités peuvent être considérées comme des quantités exprimées dans une échelle de valeurs.

   Exemple : les modalités de la variable « nombre de pages » utilisée pour décrire les livres de la bibliothèque sont repérables dans l'échelle temporelle usuelle.

2.3.1.3 Limites des classifications usuelles

Ces deux types de classification ont des limites qui sont rencontrées beaucoup d'applications. De plus, telles qu'instaurées, ces classifications sont arbitraires du point de vue statistique et les critères qui les définissent masquent les caractéristiques des variables qui importent véritablement dans une étude statistique.

1. En pratique, lorsqu'on est en présence d'une variable répondant à la définition d'une variable statistique continue, la précision avec laquelle les modalités de cette variable sont relevées rend la variable discrète. Par exemple, si on s'intéresse à l'épaisseur des livres de la bibliothèque, l'instrument avec lequel on évalue cette épaisseur aura un degré de précision qui limitera les modalités possibles de cette variable. L'ensemble des modalités (épaisseurs) que l'on peut observer (mersurer) est alors discret.

   On considère également souvent que les variable exprimées en unités monétaires sont continues. Mais le prix à l'achat de n'importe quel livre de la bibliothèque sera exprimé avec une précision maximale égale au centime.

   De ce point de vue, il n'existe en pratique aucune variable continue.

2. Par ailleurs, ce n'est pas parce que les modalités d'une variable ne sont pas exprimées dans une échelle de valeurs numériques que la variable n'est pas quantitative.

   Par exemple, si on décrit les livres de la bibliothèque par leur prix à l'achat, on peut considérer comme ensemble de modalités l'ensemble de tous les nombres réels positifs. La variable est alors clairement quantitative puisqu'une modalité représente une quantité d'argent. Mais on peut également adopter pour cette même variable l'ensemble de modalités {bon marché, cher, très cher}. Il n'est pas évident que ces modalités ne puissent pas être interprétées comme des qualités, auquel cas la variable serait maintenant qualitative.

3. Réciproquement, ce n'est pas parce que les modalités sont mesurables que la variable présente les propriétés d'une variable quantitative.

   Par exemple, la variable qui décrit les dimensions des livres de la bibliothèque a des modalités $x = (x_h, x_l, x_e)$ qui ne présentent pas les propriétés usuelles des variables quantitatives : il n'y a pas d'ordre naturel sur ces modalités. Ainsi peut-on dire

qu'un livre présentant la modalité $(21, 15, 2)$ est « plus quelque chose » (plus grand, par exemple) qu'un livre présentant la modalité $(14, 22{,}5, 2)$?

4. Les points précèdents permettent de constater que les classifications traditionnellement utilisées deviennent floues et/ou perdent de leur intérêt en pratique.
5. De plus, ces classifications, telles qu'elles apparaissent, ne sont pas établies sur des critères statistiques et par conséquent sont arbitraires au yeux du statisticien. En vertu de quoi est-il intéressant d'un point de vue statistique de savoir qu'une variable est discrète ?
6. On constate que dans les classifications continu/discret et quantitatif/qualitatif qui sont présentées ci-dessus, la distinction est basée sur les propriétés de l'ensemble des modalités. C'est précisément ces propriétés, organisées d'une autre façon que celle que nous venons de voir qui vont permettre de dégager une classification des variables utile du point de vue de l'analyse statistique.

2.3.2 Classifications pertinentes des variables

Cette section reprend une grande partie de la section 2.3 du chapitre introductif de Delmas (1996). La classification des variables qui sera adoptée par la suite met en évidence les opérations statistiques qu'il est permis d'effectuer sur les modalités de ces variables. Il faut garder à l'esprit que bien souvent, ces modalités prennent la forme de valeurs numériques. Par conséquent, elles se prêtent aux calculs usuels. Cependant, il faut toujours se poser la question du sens qu'il est possible (ou pas) de donner aux résultats de ces calculs.

2.3.2.1 Variable nominale

Définition 2.5 *On dit que X est une variable nominale s'il n'est pas possible de définir de façon naturelle un ordre sur l'ensemble $\mathscr{M}_X$ de ses modalités.*

Caractéristiques. Pour de telles variables, on pourra toujours selectionner deux éléments x et x' de $\mathscr{M}_X$ pour lesquels aucune comparaison ayant un sens n'est possible.[2]

Exemples.

- Pour la variable « discipline » servant à décrire les livres de la bibliothèque, il n'y a aucune façon naturelle de comparer « sociologie » et « littérature anglaise ».
- On peut décrire les résidents français nés en France par le numéro minéralogique de leur département de naissance. On ne peut pas comparer les modalités « 59 » et

[2]Ce n'est pas tout à fait exact puisqu'on sait que x et x' sont des modalités distinctes.

« 62 », même si elles prennent une forme numérique, car une telle comparaison est équivalente à comparer « Nord » et « Pas-de-Calais », ce qu'il n'est pas possible de faire de façon naturelle.

– Typiquement toute variable dont les modalités sont formées sur la base d'un *système de codage* sont des variables nominales : sexe (codage 0,1), catégories socio-professionnelles (nomenclature de l'INSEE), ...

Traitements statistiques possibles. Les seuls traitements dont les résultats ont un sens sont ceux qui consistent en des opérations de dénombrement (comptage), ou qui reposent sur de telles opérations. On peut donc calculer des effectifs, des fréquences des modalités de la variable (voir chapitre 4) et utiliser le mode de la variable (voir section 6.2.1).

Traitements statistiques impossibles. L'absence de relation d'ordre implique qu'il est impossible d'utiliser tout outil statistique construit à partir de l'existence d'une relation d'ordre (cumuls, quantiles, ...) De plus, les opérations usuelles sur les nombres réels ($+$, $-$, $\times$, $\div$) ne peuvent pas être définies (par exemple dans le cas des livres décrits selon leur discipline), ou bien produiront des résultats sans signification (par exemple dans le cas des peronnes décrites par le numéro minéralogique de leur département de naissance).

2.3.2.2 Variable ordinale

Définition 2.6 *On dit que X est une variable ordinale si*

1. *il est possible de définir un ordre sur l'ensemble $\mathscr{M}_X$ des ses modalités ;*
2. *les écarts et les relations de proportionalité entre modalités de X n'ont aucune signification.*

Exemples.

– Notes attribuées à des objets en fonction d'un classement.

– Toute variable mesurant des préférences.

Traitements statistiques possibles. Il est possible d'utiliser les mêmes outils que pour des variables nominales. De plus, les objets statistiques établis à partir d'une relation d'ordre sur l'ensemble des modalités de la variable ont un sens.

Traitements statistiques impossibles. On ne peut pas effecuter des traitements qui exploitent autre chose que la relation d'ordre sur l'ensemble des modalités (écarts, ...)

2.3.2.3 Variable numérique

Définition 2.7 *On dit que X est une variable numérique si*

1. *il est possible de définir un ordre sur l'ensemble $\mathscr{M}_X$ des ses modalités ;*
2. *les écarts et les relations de proportionalité entre modalités de X ont une interprétation.*

Exemples. La variable « prix d'achat en € » utilisée pour décrire les livres appartient à l'échelle proportionnelle.

De façon générale, toute variable dont les modalités sont exprimées dans une unité de mesure numérique, possédant un zéro traduisant l'absence de phénomène (quantité nulle), est numérique.

Les variables dont les modalités sont des nombre réels servant à exprimer des quantités sont donc des variables numériques.

Traitements statistiques possibles. Tous.

Traitements statistiques impossibles. Aucun.

2.3.2.4 Commentaires

1. On remarquera que les différents types de variables statistiques sont emboîtés en ce qui concerne le traitement statistique des populations décrites. Par exemple, tout ce qu'il est possible de faire avec une variable nominale peut aussi être fait avec une variable ordinale ou numérique.
2. Ce type de classification des variables est moins arbitraire que ceux qui sont habituellement adoptés, puisqu'il répartit les diverses variables en plusieurs catégories, caractérisées par les traitements statistiques applicables aux variables.
3. Par conséquent, avant de commencer toute étude statistique, il est important de déterminer pour chaque variable la catégorie à laquelle elle appartient. Cela est d'autant plus important que souvent, les modalités des variables sont exprimées à l'aide de valeurs numériques. Par conséquent, un calcul est souvent possible, dans le sens où un résultat numérique apparaîtra à l'issue de ce calcul. Cependant, le principe de la classification des variables statistiques en échelles montre clairement que ces résultats n'ont parfois aucune signification. Il faut donc identifier le type des variables afin d'être en mesure d'interpréter les calculs et de se garder de donner un sens à un nombre qui n'en a aucun.

2.4 Transformation des variables

Dans beaucoup de problèmes, on peut être amené à transformer les variables initialement utilisées.

Définition 2.8 *Soit X une variable dont l'ensemble des modalités est $\mathscr{M}_X$ et g une application de $\mathscr{M}_X$ vers un ensemble $\mathscr{G}$. La transformation de X par g est la variable Y définie par $Y = g \circ X$.*

On note que Y est une application définie sur $\mathscr{P}$ et à valeurs dans $\mathscr{G}$. Pour tout $i \in \mathscr{P}$ on a $Y(i) = g\big(X(i)\big)$. L'ensemble des modalités de Y est $\mathscr{M}_Y = \{y_1, \ldots, y_J\}$. Celui-ci est caractérisé par $\mathscr{M}_Y = \{y \in \mathscr{G} \,|\, \exists x \in \mathscr{M}_X, y = g(x)\}$. Notons que cette caractérisation de $\mathscr{M}_Y$ équivaut à dire que l'application $g : \mathscr{M}_X \to \mathscr{M}_Y$ est surjective : $\forall y \in \mathscr{M}_Y$, $\exists x \in \mathscr{M}_X$, $y = g(x)$. Cela montre également que Y ne peut posséder plus de modalités que X : $J \leq K$.

Si g est aussi bijective, alors X et Y ont le même nombre de modalités ($J = K$) et si on note $\mathscr{M}_X = \{x_1, \ldots, x_K\}$, on peut définir $\mathscr{M}_Y = \{y_1, \ldots, y_K\}$, avec $y_k = g(x_k)$, $k = 1, \ldots, K$. Dans ce cas $X(i) = x_k \;\Leftrightarrow\; Y(i) = y_k$ et donc

$$\{i \in \mathscr{P} \,|\, X(i) = x_k\} = \{i \in \mathscr{P} \,|\, Y(i) = y_k\}.$$

Chapitre 3

Les données et leur organisation

3.1 Les données brutes

Définition 3.1 *Les données sont définies par l'ensemble des modalités observées pour chacune des variables et pour chacun des individus de la population.*

1. Exemple. Soit une population $\mathscr{P}$ de N individus est décrite par trois variables X, Y et Z. On observera pour chaque individu i de $\mathscr{P}$, trois modalités, notées $X(i)$, $Y(i)$ et $Z(i)$. Les données sont donc formées de la liste de ces modalités écrites pour tous les individus :

$$X(1), Y(1), Z(1), X(2), Y(2), Z(2), X(3), \ldots, Z(N-1), X(N), Y(N), Z(N).$$

2. Les données représentées sous cette forme sont appelées *données brutes*. Il est plus commode de présenter les données brutes sous forme de tableau plutôt que sous forme de liste.

3.2 Tableau élémentaire

Définition 3.2 *On appelle* tableau élémentaire *le tableau qui à chaque individu de la population reporté en première colonne associe la modalité que présente cet individu pour chacune des variables.*

1. Un tableau élémentaire aura donc la forme :

2. Lorsque la population est peu nombreuse (N petit), on peut construire l'étude statistique à partir de ce tableau.
3. Lorsque la population est nombreuse, il faut avoir recours à une transformation du tableau élémentaire afin de faciliter le traitement statistique des données.

TABLEAU 3.1 – *Tableau élémentaire*

Individu	Modalité de X	Modalité de Y	Modalité de Z
1	$X(1)$	$Y(1)$	$Z(1)$
2	$X(2)$	$Y(2)$	$Z(2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$X(i)$	$Y(i)$	$Z(i)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$X(N)$	$Y(N)$	$Z(N)$

4. **Pour simplifier l'exposition, et jusqu'à la fin du chapitre suivant, nous allons considérer une population décrite à l'aide d'une seule variable.**

3.3 Un premier traitement statistique : réorganisation des données et opérations de comptage

Lorsque la population est nombreuse, il est pratique de réorganiser les données avant de commencer l'analyse statistique. Deux cas peuvent se présenter.

3.3.1 Petit nombre de modalités observées

Lorsque la variable X présente peu de modalités, la plupart des modalités de X sont présentées par plusieurs individus de la population. C'est typiquement le cas lorsque $\mathscr{M}_X$ est fini et de cardinal très inférieur à N. On aura donc $\mathscr{M}_X = \{x_1, x_2, \ldots, x_K\}$ et K très petit par rapport à N.

Pour simplifier l'organisation des données, on effectuera une première opération élémentaire appelée *tri à plat*.

3.3.1.1 Tri à plat et effectifs des modalités

Définition 3.3 *Le tri à plat est l'opération qui consiste dénombrer, pour chaque modalité x_k observée de la variable X, les individus de la population qui présentent la modalité x_k de la variable.*

Ce dénombrement permet de constuire les *effectifs* des modalités de la variable.

Définition 3.4 *On appelle* effectif de la modalité x_k *le nombre noté n_k, défini par*

$$n_k = \#\{i \in \mathscr{P} | X(i) = x_k\}.$$

L'effectif n_k de la modalité x_k est le cardinal de l'ensemble (ou le nombre) des individus de la population présentant la modalité x_k de X.

Propriété 3.1 *Les effectifs satisfont $n_1 + \cdots + n_K = N$.*

Démonstration : Comme chaque individu de la population présente une et une seule modalité de X (voir le point 1 de la section 2.2.3.2), les ensembles $\{i \in \mathscr{P} | X(i) = x_1\}, \ldots, \{i \in \mathscr{P} | X(i) = x_K\}$ sont disjoints et leur union est $\mathscr{P}$. On a alors

$$\begin{aligned} N = \#\mathscr{P} &= \#\big(\{i \in \mathscr{P} \,|\, X(i) = x_1\} \cup \cdots \cup \{i \in \mathscr{P} \,|\, X(i) = x_K\}\big) \\ &= \#\{i \in \mathscr{P} \,|\, X(i) = x_1\} + \cdots + \#\{i \in \mathscr{P} \,|\, X(i) = x_K\} \\ &= n_1 + \cdots + n_K. \end{aligned}$$

□

3.3.1.2 Le tableau statistique

En effectuant ce calcul pour toutes les modalités de la variable, on peut représenter les données en associant à chaque modalité son effectif. On obtient le tableau 3.2 appelé *tableau statistique*. C'est sur lui qu'est basée l'analyse statistique.

TABLEAU 3.2 – *Tableau statistique*

Modalité de X	Effectif de la modalité
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_k	n_k
$\vdots$	$\vdots$
x_K	n_K

3.3.2 Grand nombre de modalités observées

Lorsque la variable X présente beaucoup de modalités différentes, presque tous les individus de la population présentent des modalités différentes de la variable. C'est typiquement le cas lorsque X est une variable *a priori* continue et appartient à l'échelle numérique. Dans ce cas, on procède parfois à une opération appellée *regroupement par classes* des modalités observées de la variable.

3.3.2.1 Le regroupement par classes

1. L'opération consiste à constituer J classes de modalités de X, notées C_1, C_2, ... , C_J, telles que

$$\text{(a)} \quad \forall j, k, j \neq k \Rightarrow C_j \cap C_k = \varnothing,$$
$$\text{(b)} \quad C_1 \cup C_2 \cup \cdots \cup C_J = \mathscr{M}_X$$

   On dit que C_1, ... , C_J forment une *partition* de $\mathscr{M}_X$. Par conséquent pour tout $x \in \mathscr{M}_X$, il existe une et une seule classe C_j contenant x.

2. Au lieu de décrire les individus par une modalité de la variable X, on les décrira par des classes de modalité. Autrement dit, au lieu d'associer à chaque individu i une modalité de X, notée $X(i)$, on lui associe un classe de modalité, notée $C(i)$, en utilisant la règle suivante : $C(i) = C_j \Leftrightarrow X(i) \in C_j$.

   $C(i)$ est la classe de modalités de l'individu i. Cette classe est C_j si et seulement si la modalité $X(i)$ de la variable X présentée par l'individu i appartient à la j^{e} classe.

3. En utilisant ce procédé, le tableau élémentaire se présente sous la forme suivante :

TABLEAU 3.3 – *Tableau élémentaire après regroupement par classes*

Individu	Classe de modalités de X
1	$C(1)$
2	$C(2)$
$\vdots$	$\vdots$
i	$C(i)$
$\vdots$	$\vdots$
N	$C(N)$

4. Comme nous l'avons mentionné, dans beaucoup de cas où le regroupement par classes est effecuté, la variable statistique est numérique. Pour de telles variables, les classes sont des intervalles de nombres réels :

$$C_k = [e_{k-1}\,;\, e_k[\;\text{ ou }\; C_k =]e_{k-1}\,;\, e_k] \;\text{ ou }\; C_k =]e_{k-1}\,;\, e_k[\;\text{ ou }\; C_k = [e_{k-1}\,;\, e_k],$$

   où e_{k-1} et e_k sont des nombres appelés respectivement *extrêmité inférieure* et *extrêmité supérieure* de la classe C_k.

La condition de non-chevauchement des classes (voir le point 1, (a)) et la convention de numérotation par ordre croissant de modalités se traduisent ici par

$$e_0 < e_1 < \cdots < e_K. \tag{3.1}$$

Tout élément de C_k est strictement inférieur à tout élément de C_l si et seulement si $k < l$. On traduit cette propriété en disant que C_k est *plus petite* que C_l lorsque $k < l$.

5. Dans le cas où on a choisi $C_k = [e_{k-1}\,;\, e_k[$ on aura pour tout individu i, $C(i) = C_k \;\Leftrightarrow\; e_{k-1} \leq X(i) < e_k$.

6. Remarque : on peut voir le regroupement par classes comme une opération qui consiste à transformer l'ensemble $\mathscr{M}_X$ des modalités de X en un ensemble noté $\mathscr{M}_X^C$ dont les éléments sont les classes $C_1, \ldots, C_J$ consituées lors du regroupement. On aura donc $\mathscr{M}_X^C = \{C_1, \ldots, C_J\}$.

   L'avantage, si on retient comme ensemble de modalités $\mathscr{M}_X^C$ au lieu $\mathscr{M}_X$, est que le premier est fini et contient toujours moins d'éléments que le second.

   L'inconvénient est que la description de la population à l'aide de X prennant des modalités dans $\mathscr{M}_X^C$ est moins fine (moins informative) qu'une description à l'aide de la même variable X prennant des modalités dans $\mathscr{M}_X$. En effet, pour un individu i, si on connaît $X(i)$, on connaît $C(i)$. Mais si on connaît seulement $C(i)$ alors en général on ne peut connaître $X(i)$.

7. Une fois le regroupement par classes effectué, on se retrouve dans la situation où la population est nombreuse et où le nombre de modalités dans $\mathscr{M}_X^C$ est faible, c'est-à-dire dans le cas décrit à la section 3.3.1.1. Un tri à plat est alors effectué à partir du tableau précédent.

3.3.2.2 Tri à plat et effectifs de classes

Le tri à plat consiste dans ce cas à dénombrer, pour chaque classe de modalités C_j de la variable X, les individus de la population présentant une modalité de la variable appartenant à la classe C_j. On peut alors calculer les effectifs de chaque classe.

Définition 3.5 *On appelle effectif de la classe C_j le nombre, noté n_j, défini par*

$$n_j = \#\{i \in \mathscr{P} | X(i) \in C_j\}.$$

L'effectif n_j de la classe de modalités C_j est le cardinal de l'ensemble (ou le nombre) des individus de la population présentant une modalité de X appartenant à la classe C_j. Les effectifs ainsi définis satisfont la propriété 3.1.

TABLEAU 3.4 – *Tableau statistique après regroupement par classes*

Classe de modalités de X	Effectif de la classe
C_1	n_1
C_2	n_2
$\vdots$	$\vdots$
C_j	n_j
$\vdots$	$\vdots$
C_J	n_J

TABLEAU 3.5 – *Tableau statistique après regroupement par classes*

Classe de modalités de X	Centre de classe	Effectif de la classe
$[e_0\,;\,e_1[$	x_1	n_1
$[e_1\,;\,e_2[$	x_2	n_2
$\vdots$	$\vdots$	$\vdots$
$[e_{j-1}\,;\,e_j[$	x_j	n_j
$\vdots$	$\vdots$	$\vdots$
$[e_{J-1}\,;\,e_J]$	x_J	n_J

3.3.2.3 Tableau statistique

1. Le résultat du tri à plat est présenté dans le tableau statistique qui prend ici la forme du tableau 3.4.

2. Lorsque les classes de modalités sont des intervalles d'extrêmités $e_0, e_1, \ldots, e_{j-1}, e_j, \ldots, e_J$, on peut faire figurer dans le tableau statistique les centres de ces classes, que l'on note alors x_j, définis par

$$x_j = \frac{e_j + e_{j-1}}{2}\,, \quad j = 1, \ldots, J.$$

   Si par exemple on a $C_j = [e_{j-1}\,;\,e_j[$ pour $j = 1, \ldots, J-1$ et $C_J = [e_{J-1}\,;\,e_J]$, le tableau statistique se présente sous la forme du Tableau 3.5 ou celle du Tableau 3.6.

TABLEAU 3.6 – *Tableau statistique après regroupement par classes (forme alternative)*

Extrêmité de classe	Centre de classe	Effectif de la classe
e_0	x_1	n_1
e_1	x_2	n_2
e_2		
$\vdots$	$\vdots$	$\vdots$
e_{j-1}	x_j	n_j
e_j		
$\vdots$	$\vdots$	$\vdots$
e_{J-1}	x_J	n_J
e_J		

Deuxième partie

Analyse statistique univariée

Dans cette partie, on s'intéresse à une population $\mathscr{P}$ de N individus décrits selon une seule variable statistique X.

1. On suppose acquises toutes les notions du chapitre précédent. Notamment, on suppose que le regroupement par classes et/ou le tri à plat on été effectués si besoin est. Dans tous les cas, on supposera que $\mathscr{M}_X = \{x_1, \ldots, x_K\}$. Si aucun regroupement par classes n'a été nécessaire, alors x_k désigne la k^{e} modalité de X, sinon, x_k désigne le centre de la k^{e} classe de modalités de X et K désigne le nombre de classes formées selon le principe présenté au point 1 de la section 3.3.2.1.
2. D'autre part, s'il existe un ordre sur $\mathscr{M}_X$, on suppose que la numérotation des modalités se fait par ordre croissant de ces dernières de sorte que $j < k \Leftrightarrow x_j < x_k$ et donc
$$x_1 < x_2 < \cdots < x_K.$$
3. Les données peuvent se présenter soit sous la forme du tableau élémentaire suivant :

TABLEAU 3.7 – *Tableau élémentaire d'analyse univariée*

Individu	Modalité
1	$X(1)$
2	$X(2)$
$\vdots$	$\vdots$
i	$X(i)$
$\vdots$	$\vdots$
N	$X(N)$

soit sous la forme d'un tableau statistique tel que le Tableau 3.2 ou du Tableau 3.5.

Chapitre 4

Effectifs, fréquences

Dans cette section, on présente les opérations de comptage élémentaires, qui servent à dénombrer les individus de la population selon des critères de modalités de X.

4.1 Effectifs

1. L'une de ces opérations est celle qui est effectuée lors du tri à plat décrit aux sections 3.3.1.1 et 3.3.2.2 du chapitre I et qui conduit au calcul des effectifs. Pour une population de taille N donnée, plus l'effectif n_k d'une modalité x_k est élevé, plus cette modalité est représentée au sein de la population. On peut alors interpréter n_k comme le poids mesurant la représentation de la modalité x_k dans la population, ou encore comme le poids de la sous-population constituée des individus présentant la modalité x_k dans la population totale.

2. Cependant, la connaissance seule de l'effectif d'une modalité ne permet pas de connaître l'importance de son poids dans la population. Pour cela, il faut aussi connaître la taille N de la population. En effet, un effectif de 10 pour une modalité n'indique pas un même poids de cette modalité dans la population, selon que la taille N de cette dernière est 20 ou 1000.

   Pour cette raison, on définit une mesure *relative* de la représentation d'une modalité dans une population.

4.2 Fréquences

4.2.1 Définition

Définition 4.1 *On appelle fréquence de (ou associée à) la modalité x_k de X le nombre noté f_k et défini par*

$$f_k = \frac{n_k}{N}.$$

4.2.2 Interprétation

1. La fréquence f_k est la proportion d'individus présentant la modalité x_k de X. Formellement,
$$f_k = \frac{1}{N}\#\{i \in \mathscr{P} | X(i) = x_k\}.$$
2. On constate que la fréquence f_k mesure l'importance de la sous-population composée des individus présentant la modalité x_k de la variable X. Tandis que l'effectif n_k mesure cette importance de façon absolue, la fréquence f_k la mesure relativement à la taille N de la population, et par conséquent de façon indépendante d'elle.

4.2.3 Propriétés

Propriété 4.1

1. $n_k = N \times f_k, \quad k = 1, \ldots, K.$

2. $0 \leq f_k \leq 1, \quad k = 1, \ldots, K.$

3. $\sum_{k=1}^{K} f_k = 1.$

Démonstration : Le point 1 résulte directement de la définition de f_k. Les inégalités $0 \leq n_k \leq N$ (voir la propriété 3.1) impliquent le 2$^{\text{e}}$ point. Pour montrer le 3$^{\text{e}}$, on note que

$$\sum_{k=1}^{K} f_k = \sum_{k=1}^{K} \frac{n_k}{N} = \frac{1}{N} \sum_{k=1}^{K} n_k = \frac{1}{N} N,$$

où la deuxième égalité résulte de la distributivité de la multiplication par rapport à l'addition et la dernière de la propriété 3.1. □

4.2.4 Remarque

On appelle la donnée des K couples (x_k, f_k), $k = 1, \ldots, K$, la *distribution statistique de la variable X dans la population $\mathscr{P}$*. Ces couples indiquent avec quelle importance chacune des modalités de X est représentée au sein de la population.

La distribution statistique d'une variable constitue typiquement le point de départ de toute analyse statistique.

4.3 Les densités : effectifs ou fréquences unitaires

1. Notons que lorsqu'un regroupement par classes a été effectué, n_k ou f_k mesurent l'importance de la classe de modalités $C_k = [e_{k-1}\,;\, e_k[$. Lors d'un regroupement par classes, il se produit alors un effect qui n'est pas souhaitable du point de vue de l'importance d'une classe de modalité. En effet, l'effectif n_k et la fréquence f_k de C_k ont tendance à être élevés si la taille de C_k est grande. Plus précisément, considérons les deux classes C_k et $\tilde{C}_k$ telles que $C_k \subset \tilde{C}_k$. On a alors $X(i) \in C_k \Rightarrow X(i) \in \tilde{C}_k$ et donc

$$\#\{i \in \mathscr{P} | X(i) \in C_k\} \leq \#\{i \in \mathscr{P} | X(i) \in \tilde{C}_k\}.$$

   Autrement dit, la plus grande classe $\tilde{C}_k$ a un effectif plus élevé que la plus petite C_k. Lorsque les classes sont des intervalles $C_k = [e_{k-1}\,;\, e_k[$, la taille de C_k est mesurée par son amplitude, définie de la manière suivante.

Définition 4.2 *L'amplitude d'une classe de modalités $C_k = [e_{k-1}\,;\, e_k[$, est le nombre noté a_k défini par $a_k = e_k - e_{k-1}$. L'amplitude est invariante par rapport à la forme choisie pour C_k : les classes $[e_{k-1}\,;\, e_k[$, $[e_{k-1}\,;\, e_k]$, $]e_{k-1}\,;\, e_k]$ et $]e_{k-1}\,;\, e_k[$ ont la même amplitude.*

2. Lors d'un regroupement par classes, il donc est facile de constituer les classes de façon que la fréquence la plus élevée soit obtenue pour une classe de modalités dont le centre, noté x, est choisi à l'avance. En effet, on construit une classe de centre x et dont on choisit l'amplitude suffisamment grande pour que la fréquence de cette classe soit plus élevée que celle des autres classes. Il est donc possible, lors du regroupement par classes, de constituer une classe dont l'effectif dépasse la moitié de N. Dans un tel cas, les autres classes auraient nécessairement des effectifs plus faibles.

3. Les remarques qui précèdent montrent que la fréquence ou l'effectif d'une classe C_k ne mesurent pas de façon adéquate le poids de C_k dans la population. On utilise pour cela les densités d'effectif ou de fréquence définies comme suit.

Définition 4.3

- *On appelle densité d'effectif (ou effectif unitaire) de la classe $C_k = [e_{k-1}\,;\, e_k[$ le nombre noté d_k défini par $d_k = \dfrac{n_k}{a_k}$.*
- *On appelle densité de fréquence (ou fréquence unitaire) de la classe $C_k = [e_{k-1}\,;\, e_k[$ le nombre noté δ_k défini par $\delta_k = \dfrac{f_k}{a_k}$.*

4. On déduit de la définition de la densité d'effectif d_k que celui-ci représente l'effectif de C_k pour une unité de la variable X. Cette interprétation justifie l'appellation

effectif unitaire. Celui-ci mesure la concentration des individus au sein de la classe C_k.

Pour illustrer ce qui précède, considérons deux classes C_j et C_k d'amplitudes différentes avec $a_j < a_k$, mais ayant le même effectif. Un même nombre de modalités sont réparties dans une amplitude de classe plus faible pour C_j que pour C_k. Autrement dit, la concentration des modalités observées sera plus élevée dans la classe C_j que dans C_k, ce qui revient à dire que la densité des modalités est plus grande dans C_j que dans C_k. Si maintenant C_j et C_k n'ont pas le même effectif, pour comparer la concentration des sous-populations de C_j et de C_k, on utilise les densités. Il se peut que $n_k > n_j$ mais que le classe C_j soit de plus forte densité que C_k si C_j est d'amplitude suffisamment petite par rapport à celle de C_k.

La notion de densité définie ci-dessus coïncide avec la notion usuelle de densité de population. La population à considérer ici est celle constituée des individus présentant une modalité dans une classe donnée.

4.4 Effectifs cumulés

4.4.1 Défintion

Définition 4.4 *On appelle effectif cumulé croissant de (ou associé à) la modalité x_k de X, le nombre noté N_k et défini par*

$$N_k = \sum_{j=1}^{k} n_j = n_1 + n_2 + \cdots + n_k.$$

4.4.2 Interprétation

1. D'après sa définition, il est évident que la valeur prise par N_k dépend des modalités auxquelles on a donné les numéros $1, 2, \ldots, k$. Par conséquent, cette numérotation, et donc la valeur de N_k, est arbitraire, sauf s'il existe un ordre naturel sur $\mathscr{M}_X$. L'interprétation des effectifs cumulés croissant n'est donc possible que s'il existe un ordre sur l'ensemble des modalités, autrement dit si X est une variable ordinale ou proportionnelle.
2. Si X est une variable ordinale ou numérique, N_k est le nombre d'individus présentant une modalité de X inférieure ou égale à x_k. Formellement, $N_k = \#\{i \in \mathscr{P} | X(i) \leq x_k\}$.
3. Il est possible de raffiner l'interprétation de F_k dans les cas où x_k désigne le centre de la $k^{\,\mathrm{e}}$ classe de modalités C_k.

(a) Si $C_k = [e_{k-1}\,;\, e_k[$ ou $C_k =]e_{k-1}\,;\, e_k[$, alors N_k est le nombre d'individus présentant une modalité de X strictement inférieure à e_k, ou encore $N_k = \#\{i \in \mathscr{P} | X(i) < e_k\}$.

(b) Si $C_k = [e_{k-1}\,;\, e_k]$ ou $C_k =]e_{k-1}\,;\, e_k]$, alors N_k est le nombre d'individus présentant une modalité de X inférieure ou égale à e_k, ou encore $N_k = \#\{i \in \mathscr{P} | X(i) \leq e_k\}$.

4.4.3 Propriétés

Propriété 4.2

1. $N_1 = n_1$ *et* $N_K = N$.
2. $N_k = N_{k-1} + n_k,\ \ k = 2, \ldots, K$.
3. $0 < N_1 < N_2 < \cdots < N_K$.

Démonstration : Les égalités du premier point résultent de la définition de N_1 et de N_K et de la propriété 3.1. Le deuxième point est obtenu en formant la différence entre N_k et N_{k-1}. Le troisième point résulte du second et de la positivité de $n_1, \ldots, n_K$. □

4.4.4 Remarque

Il est possible de définir les effectifs cumulés décroissants définis par $N_k^d = \sum_{j=k}^{K} n_j = n_k + n_{k+1} + \cdots + n_K,\ \ k = 1, \ldots, K$. L'interprétation et les propriétés de N_k^d se déduisent de celles de N_k.

4.5 Fréquences cumulées

4.5.1 Définition

Définition 4.5 *On appelle fréquence cumulée croissante de (ou associée à) la modalité x_k de X, le nombre noté F_k et défini par*

$$F_k = \sum_{j=1}^{k} f_j = f_1 + f_2 + \cdots + f_k.$$

4.5.2 Interprétation

1. On fera ici la même remarque que celle du point 1 de l'interprétation des effectifs cumulés croissants. L'interprétation de la valeur de F_k ne peut se faire que s'il existe un ordre sur $\mathscr{M}_X$, autrement dit si X est une variable ordinale ou proportionnelle.

Dans ce cas, F_k est la proportion d'individus présentant une modalité de X inférieure ou égale à x_k. Formellement,

$$F_k = \frac{\#\{i \in \mathscr{P} | X(i) \leq x_k\}}{N}.$$

2. On note que si X est une variable ordinale ou numérique dont les modalités sont $x_1, \ldots, x_K$, et si x est une valeur pour laquelle l'inégalité $x \leq x_k$ a un sens[1] pour tout $k = 1, \ldots, K$, alors on constate que

$$\{i \in \mathscr{P} \,|\, X(i) \leq x\} = \{i \in \mathscr{P} \,|\, X(i) \leq x_k\}$$

pour toute valeur x telle que $x_k \leq x < x_{k+1}$. On peut alors définir le nombre $F(x) = \frac{\{i \in \mathscr{P} \,|\, X(i) \leq x\}}{N}$, et l'égalité ci-dessus montre que $F(x) = F(x_k) = F_k$, pour tout $x \in [x_k; x_{k+1}[$.

Définition 4.6 *Si X est une variable numérique à modalités réelles, alors la fonction notée F_X définie par*

$$\begin{aligned} F_X : \mathbb{R} &\longrightarrow [0;1] \\ x &\longmapsto F_X(x) = F_k \;\Leftrightarrow\; x_k \leq x < x_{k+1} \end{aligned}$$

est appelée fonction de répartition de X.

3. Il est possible de raffiner l'interprétation de N_k dans les cas où x_k est le centre de la $k^{\,\mathrm{e}}$ classe de modalités C_k.
   (a) Si $C_k = [e_{k-1}\,;\, e_k[$ ou $C_k =]e_{k-1}\,;\, e_k[$, alors F_k est la proportion d'individus présentant une modalité de X strictement inférieure à e_k, ou encore $F_k = \frac{1}{N}\#\{i \in \mathscr{P} | X(i) < e_k\}$.
   (b) Si $C_k = [e_{k-1}\,;\, e_k]$ ou $C_k =]e_{k-1}\,;\, e_k]$, alors F_k est la proportion d'individus présentant une modalité de X inférieure ou égale à e_k, ou encore $F_k = \frac{1}{N}\#\{i \in \mathscr{P} | X(i) \leq e_k\}$.

4.5.3 Propriétés

Propriété 4.3

1. $F_1 = f_1$ et $F_K = 1$.

[1] Cela signifie que l'ordre défini sur $\mathscr{M}_X$ permet de comparer un élément x_k de $\mathscr{M}_X$ et une valeur x qui n'appartient pas nécessairement à $\mathscr{M}_X$. Ceci se produit typiquement lorsque la variable X est une variable numérique, dont les modalités représentent des quantités exprimées par des nombres réels. Si x_k est une modalité et si x est un nombre réel, alors on peut toujours comparer x et x_k à l'aide de la relation d'ordre usuelle sur les nombres réels.

2. $F_k = F_{k-1} + f_k, \;\; k = 2, \ldots, K.$

3. $0 < F_1 < F_2 < \cdots < F_K.$

4. $F_k = \frac{N_k}{N}, \quad k = 1, \ldots, K.$

Démonstration : Le premier point résulte de la définition de F_1 et de F_K, et de la propriété 4.1. Le second point s'obtient en formant la difference $K_k - F_{k-1}$. Le troisième point résulte du précédent et de la positivité de $f_1, \ldots, f_K$ (voir la propriété 4.1). Le quatrième s'obtient en notant que

$$F_k = \sum_{j=1}^{k} f_j = \sum_{j=1}^{k} \frac{n_j}{N} = \frac{1}{N} \sum_{j=1}^{k} n_j = \frac{N_k}{N},$$

où la deuxième égalité provient de la définition des fréquence, la troisième provient de la distributivité de la multiplication par rapport à l'addition, et la dernière provient de la définition de N_k. □

4.5.4 Remarque

Il est possible de définir les fréquences cumulées décroissantes définies par $F_k^d = \sum_{j=k}^{K} f_j = f_k + f_{k+1} + \cdots + f_K, \;\; k = 1, \ldots, K$. L'interprétation et les propriétés de F_k^d se déduisent de celles de F_k.

Chapitre 5

Représentations graphiques

5.1 Généralités

1. Les représentations graphiques visent à résumer les principales caractéristiques d'une population décrite selon une variable statistique en utilisant des éléments graphiques.
2. Typiquement (mais pas exclusivement), les graphiques utilisés représentent la façon dont les modalités de la variable sont réparties au sein de la population statistique. Autrmement dit, ils constistuent une façon de représenter la distribution statistique (voir la remarque de la section 4.2) d'une variable.

   Pour cette raison, la plupart de ces graphiques consistent en une représentation des couples (x_k, f_k), ou (x_k, n_k), ou encore (x_k, F_k), $k = 1, \ldots, K$.
3. Plusieurs types de graphiques peuvent être utilisés pour représenter une même population décrite au moyen d'une même variable. Quel que soit le type de représentation graphique adopté, certains critères intervenant dans la construction du graphique doivent être impérativement retenus.

   (a) Le graphique doit se suffir à lui-même. Il doit comporter toutes les informations nécessaires pour pouvoir le lire et l'interpréter. Cela implique en particulier qu'il doit comporter un titre et une légende. Ces deux composantes du graphiques doivent également pouvoir être lues et interprétées sans avoir besoin de recourrir à une autre source d'information. Il faudra donc éviter les symboles et notations mathématiques désignant des objets statistiques (X pour la variable, f_k pour les fréquences, x_k pour les modalités, *etc*).

   (b) D'autre part, le graphique ne doit pas avoir de composantes n'ayant aucune interprétation statistique. Tout ajout au graphique qui n'apporte aucune information de nature statistique doit être proscrit.

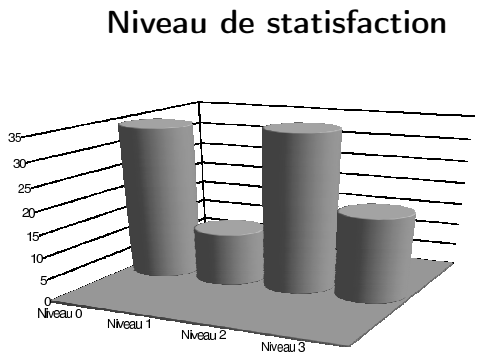
   Parmi les éléments qu'il faut éviter, on peut mentioner la couleur et la pers-

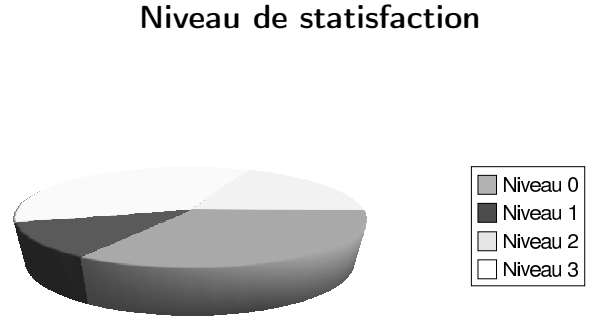
pective. Au mieux, ces ajouts n'apportent aucune information statistiquement interprétable au graphique, et au pire, ils peuvent déformer la perception que l'on a de la représentation graphique. Cela est particulièrement vrai pour les éléments de perspective, comme l'illustre l'exemple suivant.

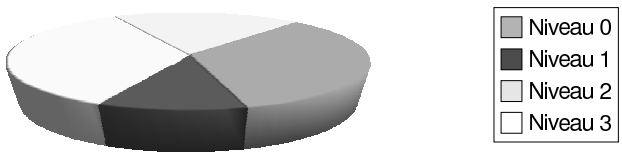
On considère une population d'individus décrits par le niveau de statisfaction (variable X) qu'ils expriment vis à vis de leur gouvernement. L'ensemble des modalités considéré est $\mathscr{M}_X = \{0, 1, 2, 3\}$, de sorte que plus la modalité d'un individu est élevé, plus l'individu est satisfait. Le tableau statistique est

Niveau de satisfaction	0	1	2	3
Fréquence	0,34	0,11	0,35	0,20

On peut proposer les (mauvaises) représentations graphiques suivantes.







Sur le premier graphique, il est difficile de déterminer la hauteur des « bâtons » correspondant aux niveaux 1 et 3. Il est également difficile de voir que le « bâton » du niveau 3 est quasiment deux fois plus haut que le « bâton » du niveau 1. Sur le deuxième graphique, il n'est pas évident que les secteurs correspondant aux niveaux 0 et 2 sont quasiment de même taille. Sur le troisième graphique, il n'est pas évident que la secteur corresondant au niveau 1 (fréquence de 11%) est quasiment deux fois plus petit que le secteur représentant le niveau 3 (fréquence de 20%)

(c) Il existe une exception à la règle mentionnée ci-dessus. On peut faire figurer sur le graphique des éléments dont le seul but est d'aider à la lecture du graphique, à condition que ces éléments ne nuisent pas à l'interprétation statistique du graphique. Par exemple, sur des graphiques à secteurs (« pie charts » ou « camemberts ») comprenant un grand nombre de secteurs, il peut être utile d'introduire de la couleur pour aider l'œil à différencier les secteurs les uns des autres. De même, en analyse bivariée, la couleur et la perspective peuvent être utilisées dans le but d'apporter de l'information statistiquement interprétable au graphique.

5.2 Quelques représentations graphiques usuelles

1. Les graphiques usuels représentent la repartition des modalités d'une variable X au sein d'une population $\mathscr{P}$. À chaque modalité on associe un élément du graphique dont la taille (longueur, hauteur, surface, ...) est d'autant plus importante que la modalité est représentée au sein de la population.
2. On peut mentionner deux principes utilisés pour construire des représentations graphiques.
   (a) Selon l'un, à chaque modalité on associe un élément du graphique d'une taille égale à la valeur de la fréquence ou de l'effectif de la modalité. C'est ce principe qui est adopté pour les diagrammes en bâtons.
   (b) Selon l'autre, la « taille » (en général la surface ou la longueur) totale du graphique est normalisé à 1 unité. Sur ce graphique, on associe à chaque modalité de la variable un élément dont la taille est égale à la fréquence de la modalité. Ce principe conduit aux diagrammes en secteurs dans lesquels les élements graphiques associés aux modalités sont les secteurs d'un disque (ou d'un demi-disque) et aux histogrammes dans lesquels ces éléments sont des rectangles.

5.2.1 Diagrammes en secteurs

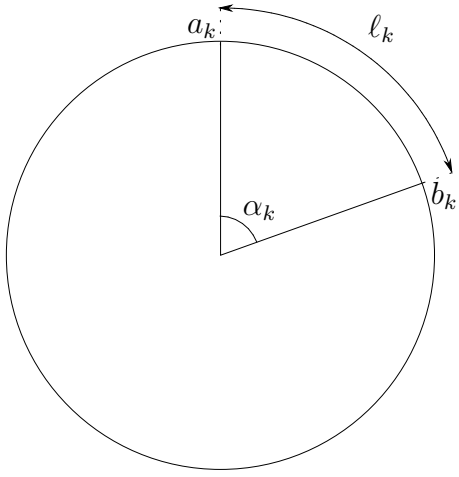
5.2.1.1 Principe

1. Le graphique consiste en un disque dont la surface est normalisée à 1 unité de surface. Ce disque est divisé en plusieurs secteurs (parts), chacun étant associé à une et seule modalité de la variable X. La proportion de la surface totale du disque occupée par le secteur associé à la modalité x_k est égale à f_k.
2. Au lieu de normaliser la surface du disque à 1, on peut normaliser son périmètre à 1 unité de longueur. Dans ce cas, la longueur de l'arc de cercle décrit par un secteur associé à une modalité x_k constitue une proportion du périmètre égale à f_k. C'est ce principe qui est utilisé pour la construction du diagramme.

5.2.1.2 Réalisation

1. Ce type de représentation graphique peut s'utiliser pour tous les types de variables et se réalise toujours de la même manière.
2. On trace un cercle de diamètre quelconque et on normalise sa circonférence à 1 unité de longueur. Considérons la modalité x_k dont la fréquence est f_k. À cette modalité sera associé un secteur décrivant un arc de cercle $\overset{\frown}{a_k b_k}$ de longueur ℓ_k unité (voir figure 5.1). Comme la circonférence du cercle est de 1 unité, pour que la longueur de l'arc $\overset{\frown}{a_k b_k}$ représente une proportion de la circonférence totale égale à f_k, il faut $\ell_k = f_k$.
3. Cependant, il est plus usuel de normaliser la circonférence du cercle à 360°. Une « règle de trois » permet de déterminer que la longueur en degrés de l'arc $\overset{\frown}{a_k b_k}$ est égale à $(360 \times f_k)°$. Ceci correspond à l'angle en degrés α_k formé par le secteur associé à x_k. Le principe de cette construction est illustré par la figure 5.1.

FIG. 5.1 – *Construction d'un diagramme en secteurs*



$\ell_k = f_k$ unité $= (360 \times f_k)°$
$\alpha_k = (360 \times f_k)°$

5.2.1.3 Lecture

Le diagramme en secteurs permet donc de représenter la fréquence avec laquelle chaque modalité de la variable est présente chez les individus d'une population. Le secteur (ou la part) associé(e) à une modalité est d'autant plus grand(e) sur le graphique que la modalité est représentée au sein de la population.

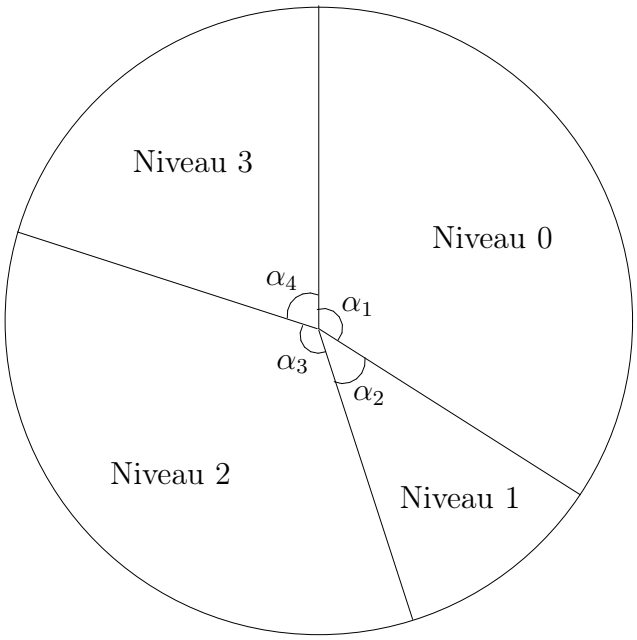
5.2.1.4 Exemple

On reprend les données du point 3b de la section 5.1. On détermine les angles des secteurs α_k en degrés pour chacune des modalités x_1 à x_4 :

$$\alpha_1 = (0{,}34 \times 360)^\circ = 122{,}4^\circ \qquad \alpha_2 = (0{,}11 \times 360)^\circ = 39{,}6^\circ$$
$$\alpha_3 = (0{,}35 \times 360)^\circ = 126^\circ \qquad \alpha_4 = (0{,}2 \times 360)^\circ = 72^\circ$$

Le tracé effctué sur la base de ces calculs produit la figure 5.2 suivante.

FIG. 5.2 – *Répartition des niveaux de statisfaction envers le gouvernement*



5.2.2 Diagrammes en bâtons

5.2.2.1 Principe

Les diagrammes en bâtons consistent en des représentations des couples (x_k, f_k) (ou (x_k, n_k)), $k = 1, \ldots, K$, où pour chaque modalité repérée sur l'axe horizontal, on représente

la fréquence (ou l'effectif) par un bâton, dont la hauteur repérée sur l'axe vertical est égale à la valeur de la fréquence (ou de l'effectif).

5.2.2.2 Réalisation

Ce type de diagramme peut s'utiliser pour toute variable. Cependant, on peut adapter le graphique au type de la variable.

1. *L'ordre des modalités.*
   - Si la variable est nominale, l'ordre dans lequel on place ses modalités le long de l'axe horizontal n'a aucune importance. On peut alors choisir par exemple l'ordre croissant ou décroissant des fréquences.
   - Si la variable est ordinale ou numérique, il existe un ordre sur les modalités et on suppose alors que $x_1 < x_2 < \cdots < x_K$. Cet ordre doit être utilisé pour représenter les modalités sur l'axe horizontal. Ainsi, de gauche à droite sur l'axe horizontal, on placera d'abord x_1, puis x_2, *etc* puis x_K.
2. *La largeur des bâtons.*
   - Si la variable est nominale ou ordinale, la largeur à la base des bâtons n'a aucune interprétation statistique. Par conséquent, on utilisera des bâtons d'une largeur nulle : ces bâtons auront l'épaisseur d'un trait, et cette épaisseur n'a aucune interprétation statistique.
   - Si la variable est numérique, et qu'il n'y a pas eu de regroupement par classes, on utilise également des bâtons de largeur nulle. Si un regroupement par classes a eu lieu, le diagramme en bâtons consiste en la représentation des couples (C_k, f_k) (ou (C_k, n_k)), $k = 1, \ldots, K$. Chaque classe $C_k = [e_{k-1}\,;\, e_k[$ est repérée sur l'axe horizontal par un segment de longueur égale à l'amplitude a_k de C_k, et d'extrémités e_{k-1} et e_k. Pour chacune de des classes ainsi repérées, on associe un bâton d'une hauteur égale à la fréquence (ou à l'effectif) de cette classe, et d'une largeur égale à l'amplitude de la classe.

     La largeur du bâton associé à une classe s'interprète donc directement comme l'amplitude de la classe. Cela n'est évidemment possible que si les différences de modalités peuvent s'interpréter.

5.2.2.3 Lecture

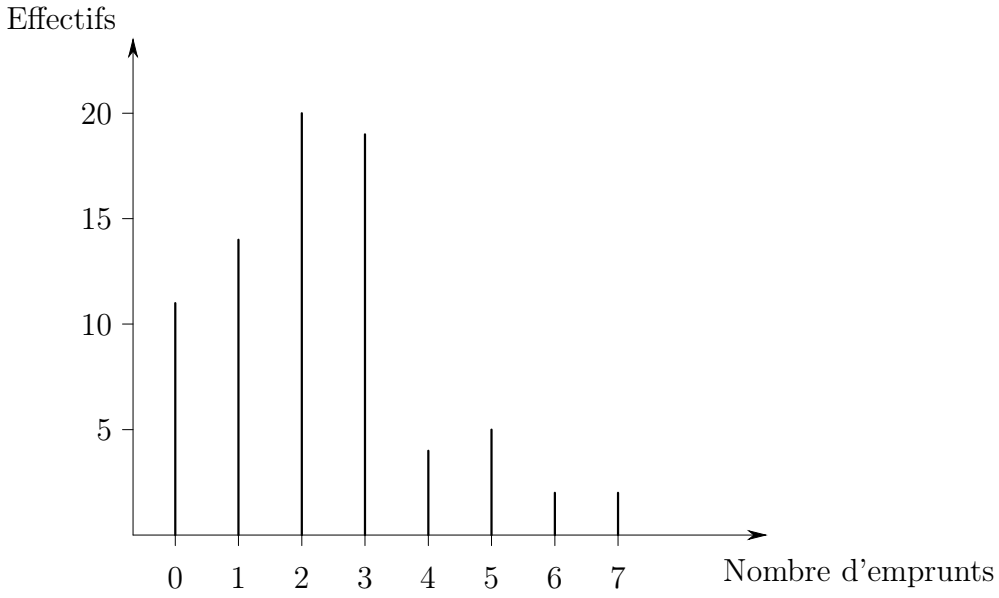
La hauteur du bâton associé à une modalité est d'autant élevée que cette dernière est représentée dans la population.

5.2.2.4 Exemples

5.2.2.4.1 Variable numérique sans regroupement par classes On décrit une population composée des livres du rayon « nouveautés » d'une bibliothèque en fonction de la variable « Nombre d'emprunts au cours du dernier trimestre ». On obtient le tableau suivant :

Nombre d'emprunts	0	1	2	3	4	5	6	7
Nombre de livres	11	14	20	19	4	5	2	2

Les modalités de la variable figurent sur la première ligne du tableau et les effectifs sur la seconde. On peut effectuer un diagramme en bâtons en représentant les couples (x_k, n_k), $k = 1, \ldots, 8$. Ce diagramme est le suivant :



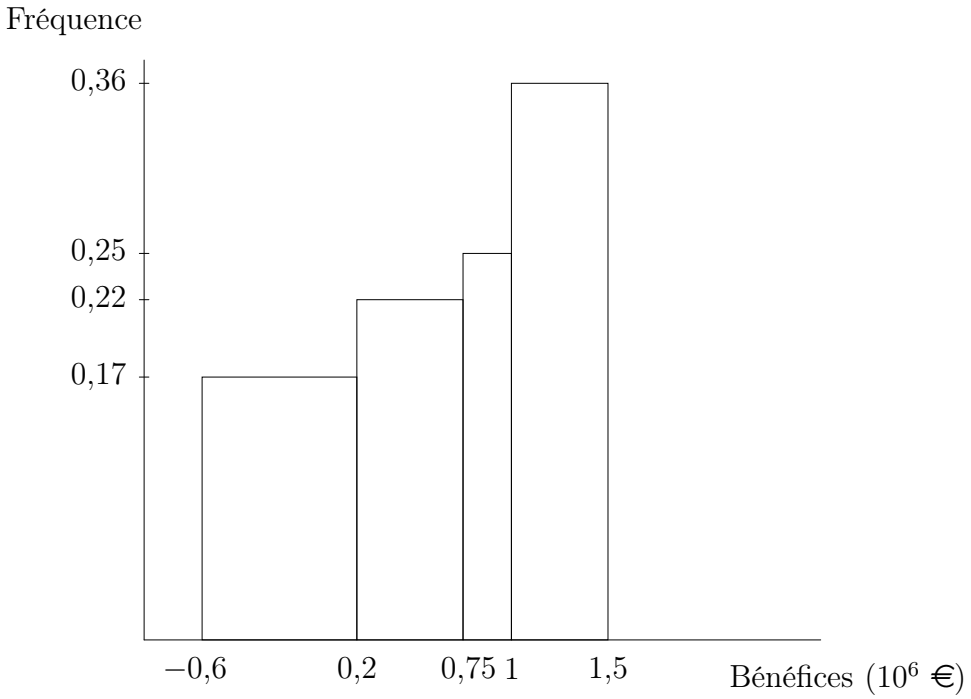
5.2.2.4.2 Variable numérique et regroupement par classes Une enquête sur les bénéfices annuels des entreprises de deux secteurs d'activité économique a produit les résultats suivants.

Valeurs des bénéfices (en millions d'euros)	Fréquences
$[-0{,}6\,;\,0{,}2\,[$	0,17
$[\,0{,}2\,;\,0{,}75\,[$	0,22
$[\,0{,}75\,;\,1\,[$	0,25
$[\,1\,;\,1{,}5\,]$	0,36

La variable (montant des bénéfices en millions d'euros) est numérique. Les amplitudes des classes ont donc une interprétation et on représentera donc pour chaque classe C_k un

bâton d'une lageur a_k et d'une hauteur f_k. On obtient donc le diagramme en bâtons de la figure 5.3.

FIG. 5.3 – *Répartition des bénéfices des entreprises du secteur d'activité*



5.2.3 Histogramme des fréquences

5.2.3.1 Le principe

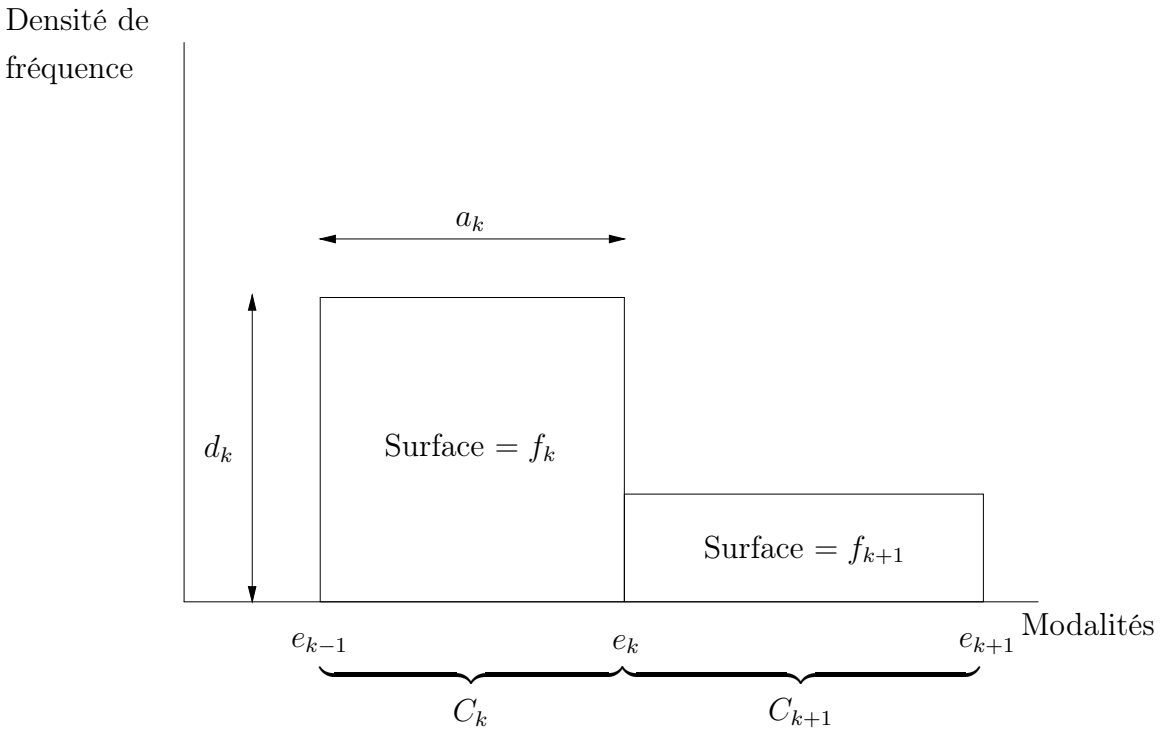
1. L'histogramme des fréquences ne s'utilise que pour une variable numérique. Ce type de représentation n'a d'intérêt que si les modalités de la variable ont été regroupées par classes. De plus, la situation dans laquelle l'histogramme présente un intérêt particulier par rapport au diagramme en bâtons est celle où les classes sont d'amplitudes inégales.
2. Le principe de l'histogramme consiste à normaliser la surface totale du graphique à 1 unité. Cette surface est découpée en rectangles. Chaque rectangle est associé à une classe de modalités de sorte que le rectangle associé à la classe C_k représente une proportion de la surface totale égale à la fréquence f_k de C_k.

5.2.3.2 Réalisation

1. On considère donc une variable numérique pour laquelle un regroupement par classes a été effectué. Chaque classe de modalités $C_k = [e_{k-1}\,;\, e_k[$ est repérée sur l'axe horizontal par un segment de longueur égale à l'amplitude $a_k = e_k - e_{k-1}$ de C_k,

et d'extrêmités e_{k-1} et e_k. Pour chaque classe ainsi repérée, on associe un rectangle d'une hauteur égale à la densité de fréquence d_k de C_k, et d'une largeur égale à l'amplitude de C_k. Cette propriété est illustrée par la figure 5.4.

FIG. 5.4 – *Construction de l'histogramme des fréquences*



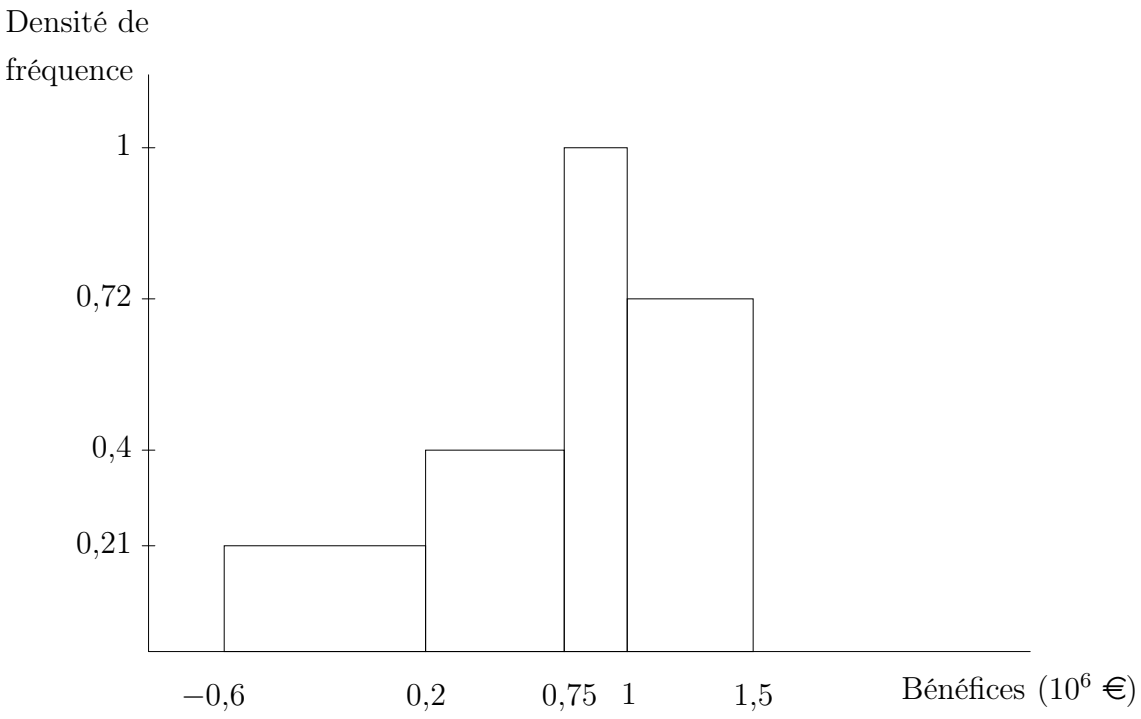
On vérifie aisément que la somme des surfaces des rectangles vaut 1, c'est à dire à la surface totale du graphique.

2. Dans le cas d'une variable numérique dont les modalités sont $x_1, \ldots, x_K$, pour laquelle aucun regroupement par classe n'a été effectué, on peut tout de même considérer qu'un tel regroupement est présent. Il suffit pour cela de considérer que pour $k = 1, \ldots, K$, la k^{e} classe C_k est $[x_k; x_k] = \{x_k\}$. Si on cherche à appliquer le principe de construction de l'histogramme présenté ci-dessus, on se heurte au problème du calcul des densités. En effet, toutes les classes sont d'amplitude nulle et le densités ne sont pas définies. On considère cependant que dans ce cas, l'histogramme coïncide avec le diagramme en bâtons des fréquences.

5.2.3.3 Lecture

1. L'histogramme permet de lire la valeur du mode (voir section 6.2.1). La classe modale est la classe correspondant au rectangle de plus grande hauteur. Le mode est le centre de cette classe.

FIG. 5.5 – *Répartition des bénéfices des entreprises du secteur d'activité*



2. L'histogramme donne également des indications sur d'autres caractéristiques de la variable X (voir section 6.4).

5.2.3.4 Exemple

On peut reprendre l'illustration de la section précédente. On calcule les amplitudes et les densités et on présente les résultats dans le tableau suivant.

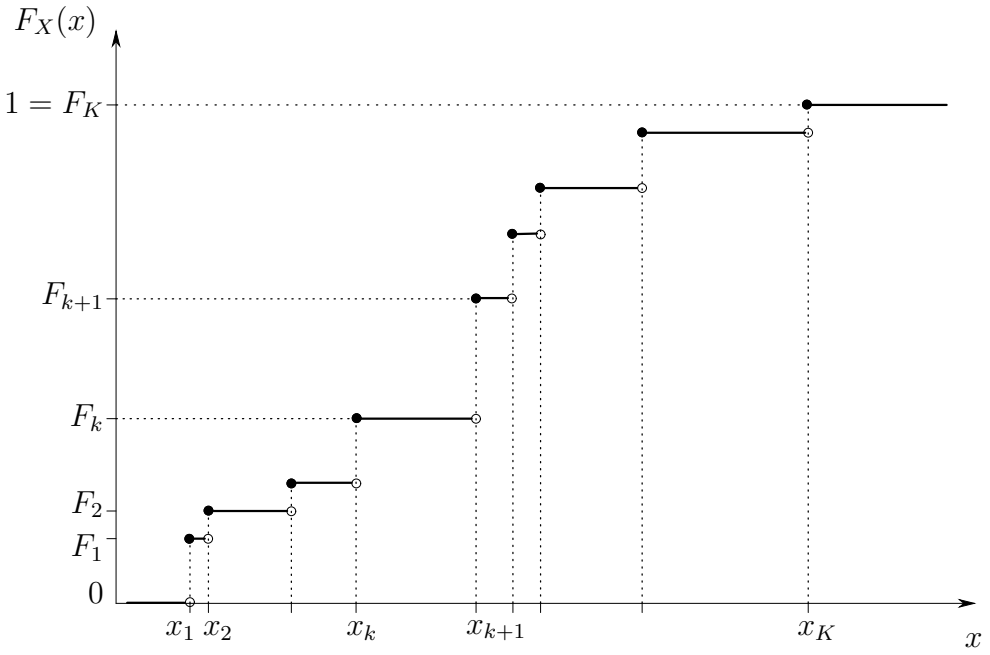
Valeurs des bénéfices (en millions d'euros)	Fréquences	Amplitudes	Densités
$[-0{,}6\,;\,0{,}2\,[$	0,17	0,8	0,21
$[\,0{,}2\,;\,0{,}75\,[$	0,22	0,55	0,4
$[\,0{,}75\,;\,1\,[$	0,25	0,25	1
$[\,1\,;\,1{,}5\,]$	0,36	0,5	0,72

L'histogramme des fréquences correspondant à ces données est la figure 5.5.

5.2.4 Graphiques représentant les fréquences cumulées croissantes

Les graphiques des sections précédentes sont construits en utilisant les fréquences ou les effectifs. Il est souvent intéressant d'utiliser également des graphiques fournissant une

FIG. 5.6 – Allure de la fonction de répartition d'une variable numérique



représentation des fréquences cumulées croissantes.

On sait grâce à la propriété 4.3 que la séquence $F_1, F_2, \ldots, F_K$ est croissante. Les graphiques présentés dans cette section permettent de visualiser la plus ou moins forte croissance de cette séquence.

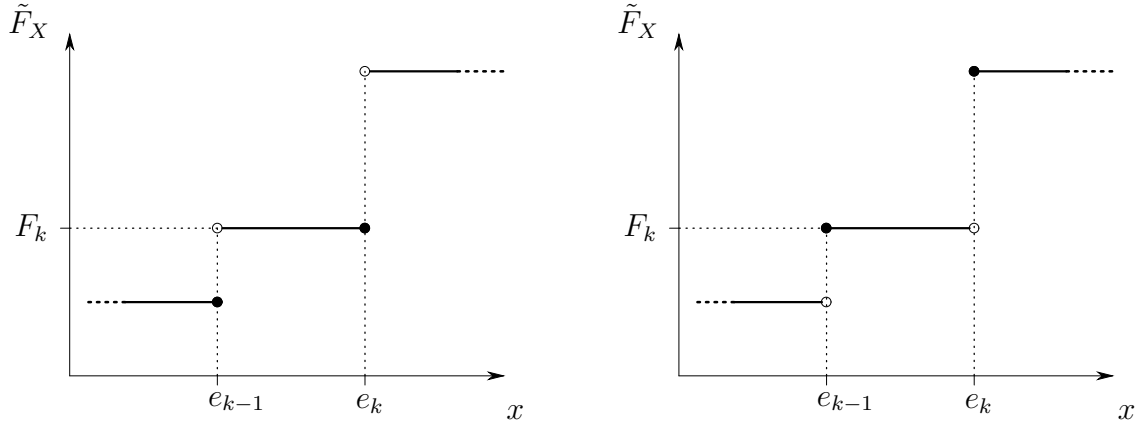
5.2.4.1 Graphe de la fonction de répartition

On rappelle que la fonction de répartion d'une variable numérique X est la fonction F_X définie à la section 4.5 (définition 4.6). L'allure typique de son graphe est représentée par la figure 5.6. Il permet de voir au voisinage de quelles modalités la séquence des fréquences cumulées croissantes croît le plus rapidement.

Si un regroupement par classes a eu lieu, la fonction de répartition est plus délicate à définir.[1] Dans ce cas, on peut effectuer le graphe de la fonction $\tilde{F}_X$ définie par $\tilde{F}_X(x) = F_k \Leftrightarrow x \in C_k$. Le graphe a la même allure que celui de la figure 5.6, à condition de faire figurer les extrêmités de classes $e_0, e_1, \ldots, e_K$ à la place des modalités $x_1, \ldots, x_K$ sur l'axe horizontal. De plus, il faut adapter les extrêmités des segments représentant $\tilde{F}_X$ selon que $C_k =]e_{k-1}; e_k]$ ou $C_k = [e_{k-1}; e_k[$. Au voisinage de la classe C_k, la fonction $\tilde{F}_X$ a donc l'allure représentée par l'un des graphiques de la figure 5.7.

[1] Si on ne dispose pas des données brutes, F_X est typiquement impossible à définir.

FIG. 5.7 – Allure de la fonction $\tilde{F}_X$



5.2.4.2 Le polygone des fréquences cumulées

1. On considère une variable numérique X pour laquelle on a effectué un regroupement par classes. Pour les classes $C_{k-1} = [e_{k-2}; e_{k-1}[$ et $C_k = [e_{k-1}; e_k[$, les fréquences cumulées croissantes sont F_{k-1} et F_k. La première donne la proportion d'individus présentant une modalité de X strictement inférieure à e_{k-1} et la seconde la proportion d'individus ayant une modalité strictement inférieure à e_k. Pour un nombre $x \in C_k$, on ne peut en général pas connaître $F_X(x)$, c'est à dire la proportion d'individus ayant une modalité de X inférieure ou égale à x. On sait simplement que

$$\{i \in \mathscr{P} \,|\, X(i) < e_{k-1}\} \subseteq \{i \in \mathscr{P} \,|\, X(i) \leq x\} \subseteq \{i \in \mathscr{P} \,|\, X(i) < e_k\},$$

   d'où on déduit $F_{k-1} \leq F_X(x) \leq F_k$.

2. On peut approximer la fonction F_X sur l'intervalle $[e_{k-1}; e_k]$ en supposant qu'elle est linéaire sur cette intervalle, c'est à dire que pour des nombres α_k et β_k, on a

$$\begin{cases} F_X(x) \simeq \hat{F}_X(x) = \alpha_k x + \beta_k, & \forall x \in [e_{k-1}; e_k] \\ \hat{F}_X(e_k) = F_k \quad \text{et} \quad \hat{F}_X(e_{k-1}) = F_{k-1} \end{cases} \tag{5.1}$$

   avec la convention que $F_0 = 0$. Pour déterminer les nombres α_k et β_k, on peut utiliser le fait que la variation relative de $\hat{F}_X$ est constante sur tout intervalle de $[e_{k-1}; e_k]$. Cela revient à dire qu'il existe un nombre Δ_k tel que :

$$\frac{\hat{F}_X(x) - \hat{F}_X(y)}{x - y} = \Delta_k, \qquad \forall x \in [e_{k-1}; e_k],\ \forall y \in [e_{k-1}; e_k]. \tag{5.2}$$

   C'est en particulier vrai pour $x = e_k$ et $y = e_{k-1}$ et donc

$$\Delta_k = \frac{\hat{F}_X(e_k) - \hat{F}_X(e_{k-1})}{e_k - e_{k-1}} = \frac{F_k - F_{k-1}}{e_k - e_{k-1}}\,.$$

D'après le point 2 de la propriété 4.3 et les définitions de l'amplitude de classe (définition 4.2) et de la densité de fréquence (définition 4.3), on constate que Δ_k est égal à la densité de fréquence δ_k de la classe C_k, De plus, pour tout nombre $x \in [e_{k-1}; e_k]$, la condition (5.2) implique également

$$\delta_k = \frac{\hat{F}_X(x) - \hat{F}_X(e_{k-1})}{x - e_{k-1}} = \frac{\hat{F}_X(x) - F_{k-1}}{x - e_{k-1}},$$

ou encore $\hat{F}_X(x) = \delta_k x + F_{k-1} - \delta_k e_{k-1}$. Si on pose

$$\alpha_k = \delta_k = \frac{F_k - F_{k-1}}{e_k - e_{k-1}} \qquad \text{et} \qquad \beta_k = F_{k-1} - \delta_k e_{k-1},$$

on obtient la caractérisation (5.1) de $\hat{F}_X$.

Propriété 5.1 *Sur l'intervalle $[e_0; e_K]$, la fonction $\hat{F}_X$ est continue et strictement croissante.*

Démonstration : À l'intérieur de chaque intervalle $[e_{k-1}; e_k]$ est linéaire. Elle est par conséquent continue en tout point $x \in]e_{k-1}; e_k[$. D'autre part en tout point e_k, $k = 1, \ldots, K-1$, les limites à gauche et à droite de $\hat{F}_X$ sont respectivement données par

$$\lim_{\substack{x \to e_k \\ x < e_k}} \hat{F}_X = \lim_{x \to e_k} (\alpha_k x + \beta_k) = \alpha_k e_k + \beta_k$$

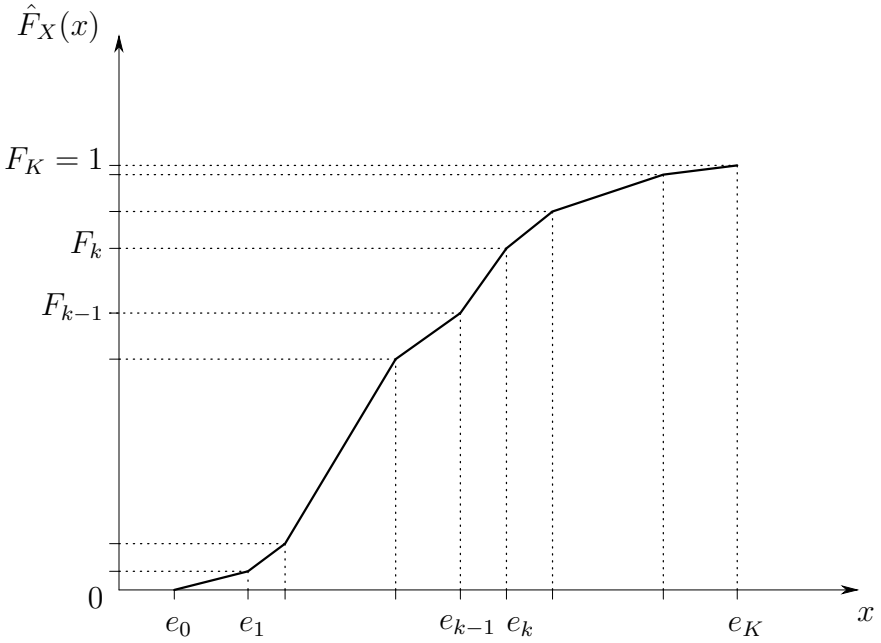
$$\text{et} \qquad \lim_{\substack{x \to e_k \\ x > e_k}} \hat{F}_X = \lim_{x \to e_k} (\alpha_{k+1} x + \beta_{k+1}) = \alpha_{k+1} e_k + \beta_{k+1}.$$

En utilisant les expressions de α_k, α_{k-1} β_k et β_{k-1} données ci-dessus, ces limites sont toutes deux égales à F_k. La fonction $\hat{F}_X$ est donc continue sur $]e_0; e_K[$. Comme $\hat{F}_X$ n'est pas définie à gauche de e_0, il suffit pour vérifier la continuité de $\hat{F}_X$ en e_0 de vérifier que $\lim_{\substack{x \to e_0 \\ x > e_0}} \hat{F}_X(x) = \hat{F}_X(e_0)$. Cette limite étant égale à $\lim_{x \to e_0}(\alpha_1 x + \beta_1) = \alpha_1 e_0 + \beta_1 = \hat{F}_X(e_0)$, on a le résultat voulu. Par un raisonnement similaire, on calcule la limite de $\hat{F}_X$ à gauche de e_K et on établit que cette fonction est continue en e_K.

Soient x et y deux nombres quelconques tels que $x < y$. On montre que $\hat{F}_X(y) > \hat{F}_X(x)$. Si x et y appartiennent au même intervalle $[e_{k-1}; e_k]$, on aura $\hat{F}_X(y) - \hat{F}_X(x) = \alpha_k(y - x)$. Comme $F_{k-1} < F_k$ (voir propriété 4.3), on a $\alpha_k > 0$, et donc $\hat{F}_X(y) - \hat{F}_X(x) > 0$. Autrement dit, $\hat{F}_X$ est croissante sur chacun des intervalles $[e_{k-1}; e_k]$, $k = 1, \ldots, K$. Si x et y appartiennent à des intervalles différents, $x \in [e_{k-1}; e_k]$ et $y \in [e_{l-1}; e_l]$, on doit avoir $e_k < e_{l-1}$ (sinon cela contredirait $y > x$). Comme $\hat{F}_X$ est croissante sur chacun de ces intervalles, on doit avoir

$$F_{k-1} \leq \hat{F}_X(x) \leq F_k \qquad \text{et} \qquad F_{l-1} \leq \hat{F}_X(y) \leq F_l. \tag{5.3}$$

FIG. 5.8 – Allure du polygone des fréquences cumulées



La croissance de la séquence $F_1, \ldots, F_K$ (voir la propriété 4.3) entraîne l'implication $e_k < e_{l-1} \Rightarrow F_k < F_{l-1}$. Ceci et les inégalités de (5.3) entraînent à leur tour $\hat{F}_X(x) < \hat{F}_X(y)$. □

Définition 5.1 *On appelle polygone des fréquences cumulées le graphe de la fonction* $\hat{F}_X$.

3. L'allure typique du polygone des fréquence cumulées est représenté à la figure 5.8. On vérifie que $\hat{F}_X$ est strictement croissante et continue. On vérifie qu'elle est linéaire sur chaque intervalle $[e_k; e_{k-1}]$, $k = 1, \ldots, K$ (on dit que $\hat{F}_X$ est linéaire par morceaux). On constate également que la courbe de $\hat{F}_X$ présente des coudes aux points d'abscisse e_k, $k = 1, \ldots, K-1$.

4. Tout comme le précédent, ce type de graphique permet également de visualiser immédiatement les taux de croissance des fréquences cumulées. En effet, sur la classe C_k, l'accroissement des fréquences est $F_k - F_{k-1} = f_k$ (d'après le point 2 de la propriété 4.3). Le taux d'accroissement est égal à la densité de fréqunce δ_k. Il est représenté par la pente du polygone entre les points de coordonnées (e_{k-1}, F_{k-1}) et (e_k, F_k). Plus cette pente est élevée, plus le taux de croissance d'une fréquence cumulée à la suivante est grand.

5. L'histogramme des fréquences et le polygone des fréquences cumulées sont fortement liés l'un à l'autre. En effet, le rectangle de l'histogramme associé à une classe est d'autant plus haut que la densité de cette classe est élevé, c'est à dire d'autant plus haut que la pente du polygone sur cette classe est élevée.

Chapitre 6

Description numérique des données

6.1 Généralités

1. Dans un contexte univarié, la description numérique des données consiste à résumer à l'aide de modalités d'une variable X certaines des caractéristiques de la population décrite par X (ou encore de la distribution statistique de X, telle que définie à section 4.2).
2. Ces descriptions sont obtenues à l'aide d'indicateurs. Chacun de ces indicateurs est construit de façon à pouvoir capturer et résumer *une et une seule* des caractéristiques de la distribution de X. Par conséquent, si on s'intéresse à plusieurs des propriétés de cette distribution, on est amené à utiliser simultanément plusieurs indicateurs.
3. En revanche, pour décrire l'une des caractéristiques d'une distribution, on peut utiliser plusieurs indicateurs. Chacun aura des propriétés propres qui constitueront des avantages ou des inconvénients par rapport aux autres. Quelles que soient ces propriétés, tout indicateur doit être construit de façon à effectivement capturer la caractéristique souhaitée. Un indicateur ne peut donc s'utiliser de façon interchangeable pour capturer plusieurs caractéristiques possibles.
4. Les caractéristiques d'une distribution auxquelles une analyse statistique est typiquement amenée à s'intéresser sont usuellement regroupées en quatre catégories.

6.2 Indicateurs de position (ou de tendance centrale)

Les indicateurs de position — ou indicateurs de tendance centrale — permettent de résumer par une seule modalité la valeur de toutes les modalités observées de X. Ils sont construits pour pouvoir décrire l'endroit de l'ensemble des modalités de X autour duquel se positionnent les données $X(1), \ldots, X(N)$. De façon générale, cet endroit est appelé *tendance centrale*.

Les indicateurs de position les plus couramment utilisés sont le mode, la médiane et la moyenne.

6.2.1 Le mode

6.2.1.1 Définition

Définition 6.1 *Le mode de la variable X est la modalité la plus fréquemment observée dans la population $\mathscr{P}$. On note* $\mathrm{Mo}(X)$ *le mode de X.*

6.2.1.2 Interprétation

Cet indicateur mesure la position des données en utilisant un principe de représentation (relativement) majoritaire : la position des données est décrite par la modalité la plus souvent obervée.

6.2.1.3 Propriétés

1. Le mode se calcule à partir des fréquences. Par conséquent il se calcule pour tous les types de variables.
2. La modalité x_k sera le mode de X si $f_k \geq f_j$, $\forall j = 1, \ldots, K$.
3. Il résulte du point précédent que le mode n'est pas nécessairement unique. Il est en effet possible que deux modalités de la variable soient d'une part aussi fréquemment observées l'une que l'autre, et d'autre part toutes les deux plus fréquemment observées que les autres modalités. Dans ce cas, le mode consiste en le couple de ces deux modalités et on dit que la distribution statistique de X est *bimodale*. Bien entendu, il est aussi possible que le mode consiste en un m-uplet de modalités, avec $2 \leq m \leq K$, auquel cas la distribution statistique de X est dite *multimodale*.
4. D'après sa définition, pour calculer le mode il faut pouvoir calculer la fréquence de chacune des modalités. Si un regroupement par classes a été effectué, ceci n'est pas possible puisque dans ce cas on n'observe pas les fréquences de chacune des modalités. On se contente alors de déterminer la *classe modale* de X. Celle-ci est la classe de modalités la plus représentée dans la population. On sait d'après le point 1 de la section 4.3 qu'il faut pour cela utiliser les densités. On définit alors la classe modale. Il est courant dans ce cas d'approximer le mode par le centre de la classe modale (lorsque celle-ci prend la forme d'un intervalle). Cette approximation du mode est souvent assimilée au mode lui-même (voir la définition 6.2 ci-dessous).
5. Le mode se repère aisément sur un diagramme en bâtons : $\mathrm{Mo}(X)$ est la modalité associée au bâton le plus haut.

Propriété 6.1 *Soit Y une transformation de X par g. Si $g : \mathscr{M}_X \to \mathscr{M}_Y$ est bijective alors* $\mathrm{Mo}(Y) = g\big(\mathrm{Mo}(X)\big)$.

Démonstration : Il suffit de noter que d'après la section 2.4 du chapitre I, on a $X(i) = \mathrm{Mo}(X)$ si et seulement si $Y(i) = g\big(\mathrm{Mo}(X)\big)$). □

6. En cas de regroupement par classes, on ne peut pas déterminer le mode en utilisant la définition qui en a été donnée. On utilisera par conséquent la définition suivante.

Définition 6.2

1. *Si un regroupement par classes des modalités a été effectué, on définit la classe modale comme étant la classe associée à la plus forte densité (de fréquence ou d'effectif).*
2. *Par convention, le mode de X est défini comme le centre de la classe modale.*

7. Avec cette définition, on peut aisément repérer le mode à partir de l'histogramme. Le mode est le centre de la classe à laquelle est associé le rectangle de l'histogramme le plus haut.

6.2.2 La médiane

La médiane d'une variable X est un indicateur déterminé à partir d'un classement des individus selon un critère d'ordre des modalités de X. Ceci n'est évidemment possible que s'il existe un ordre sur $\mathscr{M}_X$. Par conséquent la médiane est un indicateur qui ne s'utilise que pour des variables ordinales ou numériques.

6.2.2.1 Définitions

Définition 6.3 *Soit X une variable statistique et soit i un individu de la population. Le rang de i, noté $R(i)$, est sa position dans un classement des individus de la population par ordre croissant des modalités de X, une fois les* ex æquo *départagés.*

1. Les rangs sont définis de façon formelle dans l'Annexe. Il suffit de noter que l'individu de rang 1 présente une modalité qui ne peut être plus grande que celle de n'importe quel autre individu. L'individu de rang N présente une modalité au moins aussi grande que celle de n'importe quel autre individu. Plus généralement, pour deux individus i et j de la population on a $R(i) < R(j) \Rightarrow X(i) \leq X(j)$.
2. Puisque les rangs sont attribués après départage des *ex æquo*, deux individus distincts ne peuvent porter le même rang ; de plus, pour tout entier $m \in \{1, \ldots, N\}$, il existe un individu dont le rang est m. Autrement dit

$$\forall m \in \{1, \ldots, N\},\ \exists! i,\ R(i) = m.$$

Définition 6.4 *Pour deux individus i et j de la population, on dit que i précède j si $R(i) < R(j)$. On dit également dans ce cas que j suit i.*

3. On déduit des remarques précédentes que pour deux individus distincts i et j, soit i précède j et donc j suit i, soit j précède i et donc i suit j
4. La détermination des rangs permet d'identifier un individu « central », appelé *individu médian*, qui à son tour permet de déterminer la médiane de X.

Définition 6.5 *L'individu médian est un individu désigné par i_{Me} tel que $R(i_{\text{Me}}) = \left\lceil \frac{N}{2} \right\rceil$, où pour tout nombre réel y, $\lceil y \rceil$ désigne le plus petit nombre entier supérieur ou égal à y.*[1] *Ainsi on a*

1. $R(i_{\text{Me}}) = \frac{N+1}{2}$ *si N est* impair *;*
2. $R(i_{\text{Me}}) = \frac{N}{2}$ *si N est* pair.

On déduit immédiatement de la définition que

a) $\frac{N-1}{2}$ individus précèdent i_{Me} et $\frac{N-1}{2}$ suivent i_{Me}, si N est *impair*;

b) $\frac{N}{2} - 1$ individus précèdent i_{Me} et $\frac{N}{2}$ individus suivent i_{Me} si N est *pair*.

Définition 6.6 *La médiane de X, que l'on note* $\text{Me}(X)$, *est la modalité de X présentée par l'individu médian :* $\text{Me}(X) = X(i_{\text{Me}})$.

5. Lorsque N est pair, la médiane de X est assez souvent définie comme le milieu de l'intervalle qui sépare les modalités des individus de rangs $\frac{N}{2}$ et $\frac{N}{2}+1$. Ce milieu est obtenu en faisant la moyenne de ces deux modalités.

   Cependant, la médiane est un indicateur de position qui peut s'utiliser pour des variables ordinales. Pour de telles variables, une moyenne de deux modalités n'a aucun sens. Par conséquent, cette définition particulière de la médiane sera réservée aux variables numériques. Il en sera fait usage en particulier dans la section 6.4.1.

6.2.2.2 Exemple

Si les données (brutes) sont les suivantes

i	1	2	3	4	5	6	7
$X(i)$	17	19	25	16	17	20	11

alors un classement par ordre croissant des modalités est

Rang	1	2	3	4	5	6	7
i	7	4	1	5	2	6	3
$X(i)$	11	16	17	17	19	20	25

[1] Par exemple, si $y = \sqrt{2}$, $\lceil y \rceil = 2$. Pour tout nombre $y \in]1;2]$, on a $\lceil y \rceil = 2$.

Comme $N = 7$ est impair, i_{Me} est l'individu de rang $\frac{N}{2} + 1 = 4$. Donc $i_{\text{Me}} = 5$. On a alors $\text{Me}(X) = X(5) = 17$.

6.2.2.3 Interprétation

1. Dans le cas où N est impair, la valeur de la médiane permet de former deux sous-populations $\mathscr{P}_1$ et $\mathscr{P}_2$, contenant chacune $\frac{N-1}{2}$ individus de $\mathscr{P}$, de sorte que tout individu de $\mathscr{P}_1$ présente une modalité de X qui ne peut excéder la modalité de X présentée par n'importe quel individu de $\mathscr{P}_2$. $\mathscr{P}_1$ contient les individus de rangs 1 à $\frac{N-1}{2}$ et $\mathscr{P}_2$ contient les individus de rangs $\frac{N+3}{2}$ à N. En termes de rang, i_{Me} se trouve entre ces deux sous-populations

   Le rang de l'individu médian est donc central puisqu'il y a autant d'individus qui le précèdent que d'individus qui le suivent. La médiane $\text{Me}(X)$, puisqu'elle coïncide avec la modalité présentée par l'individu médian, se retrouve ainsi au centre de la distribution statistique de X, dans le sens où on peut trouver à la fois $\frac{N-1}{2}$ individus ayant une modalité inférieure ou égale à $\text{Me}(X)$ et $\frac{N-1}{2}$ individus ayant une modalité supérieure ou égale à $\text{Me}(X)$.

2. Si N est pair, la médiane agit de façon semblable. La sous-population $\mathscr{P}_1$ contient les individus de rangs 1 à $\frac{N}{2}$ et $\mathscr{P}_2$ contient les individus de rangs $\frac{N}{2} + 1$ à N. Ici, les deux sous-populations $\mathscr{P}_2$ sont toutes deux d'effectif $\frac{N}{2}$ et forment une partition de $\mathscr{P}$. L'individu médian fait partie de la première sous-population $\mathscr{P}_1$.

   Cependant, son rang n'est pas central au sens donné ci-dessus. On peut cependant considérer dans le cas où N est pair que les rangs $\frac{N}{2}$ et $\frac{N}{2} + 1$ sont centraux. L'un de ces deux rangs est celui de l'individu médian et, à 1 individu près, il y a autant d'individu qui précèdent l'individu médian que d'individus qui le suivent. Pour la même raison que précédemment, la médiane se situe donc aussi dans ce cas au centre de la distribution statistique de X.

3. Les commentaires qui précèdent montrent que la médiane est un indicateur qui décrit la position des données par leur « milieu ».

6.2.2.4 Propriétés

Propriété 6.2

1. *La médiane* $\text{Me}(X)$ *est l'une des modalités de* X *:* $\text{Me}(X) \in \mathscr{M}_X$.
2. *Si* Y *est une transformation de* X *par* g*, où* g *est une fonction strictement croissante, alors* $\text{Me}(Y) = g\big(\text{Me}(X)\big)$.

3. On a le tableau d'inégalités suivant :

N impair	*N pair*
$\#\{i \in \mathscr{P} \mid X(i) \leq \mathrm{Me}(X)\} \geq \frac{N+1}{2}$	$\#\{i \in \mathscr{P} \mid X(i) \leq \mathrm{Me}(X)\} \geq \frac{N}{2}$
$\#\{i \in \mathscr{P} \mid X(i) > \mathrm{Me}(X)\} < \frac{N+1}{2}$	$\#\{i \in \mathscr{P} \mid X(i) > \mathrm{Me}(X)\} \leq \frac{N}{2}$
$\#\{i \in \mathscr{P} \mid X(i) < \mathrm{Me}(X)\} < \frac{N+1}{2}$	$\#\{i \in \mathscr{P} \mid X(i) < \mathrm{Me}(X)\} < \frac{N}{2}$
$\#\{i \in \mathscr{P} \mid X(i) \geq \mathrm{Me}(X)\} \geq \frac{N+1}{2}$	$\#\{i \in \mathscr{P} \mid X(i) \geq \mathrm{Me}(X)\} \geq \frac{N}{2} + 1$

4. Soit x_{k^*} la modalité de X coïncidant avec $\mathrm{Me}(X)$*. Si on modifie la modalité de la variable pour un individu, alors la médiane n'est pas nécessairement modifiée. Si elle l'est, elle devient égale à x_{k^*-1} ou x_{k^*+1}.*

Démonstration :

1. Cela découle immédiatement de la définition de la médiane.

2. Si on classe les individus par ordre croissant de modalités de Y, g étant strictement croissante, le classement est le même que celui obtenu en utilisant les modalités de X. L'individu médian reste donc le même, que le classement soit fait selon X ou selon Y. La modalité de Y présentée par cet individu est donc la médiane de Y. Elle est égale à $Y(i_{\mathrm{Me}}) = g\big(X(i_{\mathrm{Me}})\big) = g\big(\mathrm{Me}(X)\big)$.

3. On commence par supposer N impair.

   1re ligne. On a $\#\{i \in \mathscr{P} \mid X(i) \leq \mathrm{Me}(X)\} = \tilde{R}(i_{\mathrm{Me}}) \geq R(i_{\mathrm{Me}}) = \frac{N+1}{2}$, où l'égalité résulte de la définition de $\tilde{R}$ (voir l'Annexe) et l'inégalité résulte du point 2 de la propriété en Annexe.

   2e ligne. Comme $\{i \in \mathscr{P} \mid X(i) > \mathrm{Me}(X)\} = \mathscr{P} \setminus \{i \in \mathscr{P} \mid X(i) \leq \mathrm{Me}(X)\}$, on déduit du premier résultat que $\{i \in \mathscr{P} \mid X(i) > \mathrm{Me}(X)\} \leq \frac{N-1}{2}$.

   3e ligne. D'après le point 4 de la propriété en Annexe, on a $X(i) < \mathrm{Me}(X) \Rightarrow R(i) < R(i_{\mathrm{Me}})$ et donc $\{i \in \mathscr{P} \mid X(i) < \mathrm{Me}(X)\} \subseteq \{i \in \mathscr{P} \mid R(i) < R(i_{\mathrm{Me}})\}$. D'où

   $$\#\{i \in \mathscr{P} \mid X(i) < \mathrm{Me}(X)\} \leq \#\{i \in \mathscr{P} \mid R(i) < R(i_{\mathrm{Me}})\} = R(i_{\mathrm{Me}}) - 1 = \tfrac{N-1}{2},$$

   où la première des égalités résulte du point 5 de la propriété en Annexe.

   4e ligne. $\#\{i \in \mathscr{P} \mid X(i) \geq \mathrm{Me}(X)\} = \mathscr{P} \setminus \#\{i \in \mathscr{P} \mid X(i) < \mathrm{Me}(X)\}$. Par le même argument que pour montrer l'inégalité de la 2e ligne, on obtient le résultat voulu.

   Lorsque N est pair, les inégalités se montrent exactement de la même manière, mais en utilisant $R(i_{\mathrm{Me}}) = \frac{N}{2}$ (voir la définition 6.5 de l'individu médian dans le cas N pair).

4. Preuve détaillée omise. On se contente de remarquer que si pour $i \in \mathscr{P}$ on modifie $X(i)$ alors on peut changer le rang de i. Ce rang, s'il était inférieur à celui de i_{Me} avant la modification, peut devenir supérieur après, et réciproquement. La seule conséquence est de décaler le rang de i_{Me} d'une unité. La modalité du nouvel individu médian (la nouvelle médiane) est donc proche celle de l'ancien. La nouvelle médiane est donc égale soit à x_k^* (auquel cas la médiane est inchangée), soit à x_{k^*-1} soit à x_{k^*+1}, qui sont les deux modalités adjacentes à x_k^*. □

6.2.2.5 Remarques

1. Lorsque N est pair et que la variable X est numérique, la médiane est souvent définie comme le milieu du segment qui sépare la modalité de l'individu médian de celle de l'individu qui se situe rang $\frac{N}{2}+1$ (c'est l'individu qui suit immédiatement l'individu médian).

   Dans l'exemple ci dessus, supposons qu'un huitième individu soit observé de sorte que les données sont

i	1	2	3	4	5	6	7	8
$X(i)$	17	19	25	16	17	20	11	27

   Un classement par ordre croissant des modalités est

Rang	1	2	3	(4)	5	6	7	8
i	7	4	1	(5)	2	6	3	8
$X(i)$	11	16	17	(17)	19	20	25	27

   Si on applique la définition 6.5 de la médiane avec $N = 8$ pair, l'individu médian est de rang $\frac{N}{2} = (4)$. Donc $i_{\text{Me}} = (5)$ et la médiane est $\text{Me}(X) = X((5)) = (17)$.

   Si on utilise maintenant la définition de la médiane qui consiste à prendre le milieu du segment qui sépare les modalités des individus de rangs $\frac{N}{2}$ et $\frac{N}{2}+1$, la médiane est alors égale à la moyenne des modalités des individus 5 et 2, c'est à dire $\frac{X(5)+X(2)}{2} = \frac{1}{2}(17+19) = 18$.

2. La médiane est définie comme la modalité de X présentée par l'individu médian. Pour pouvoir déterminer cet individu et connaître sa modalité, il faut connaître $X(1), \ldots, X(N)$. Autrement dit, l'utilisation directe de la définition de la médiane pour effectuer son calcul nécessite de disposer des données brutes. Lorsque ce n'est pas le cas, il faut un critère caractérisant la médiane. On montre ci-dessous que celui-ci peut être établi à partir de des fréquences cumulées croissantes ou des effectifs cumulés croissants.

6.2.2.6 Caractérisation de la médiane

On rappelle que l'effectif (la fréquence) cumulé(e) croissant(e) jusqu'à la modalité x_k s'interprète comme le nombre (la proportion) d'individus présentant une modalité inférieure ou égale à x_k. Formellement,

$$N_k = \#\{i \in \mathscr{P} \,|\, X(i) \leq x_k\}$$

Cet effectif (ou cette fréquence) peut alors servir à déterminer la médiane.

Propriété 6.3 *La médiane est donnée par* $\mathrm{Me}(X) = \min\{x_k \in \mathscr{M}_X \,|\, N_k \geq \frac{N}{2}\} = \min\{x_k \in \mathscr{M}_X \,|\, F_k \geq \frac{1}{2}\}$

Démonstration : Définissons $\mathscr{M}_X^* = \{x_k \in \mathscr{M}_X \,|\, N_k \geq \frac{N}{2}\}$ et $x_{k^*} = \min \mathscr{M}_X^*$. Notons que N_k est entier $\forall k = 1, \ldots, K$. Par conséquent, $\frac{N}{2}$ n'étant pas entier lorsque N est impair, l'ensemble $\mathscr{M}_X^*$ coïncide dans ce cas avec $\{x_k \in \mathscr{M}_X \,|\, N_k \geq \frac{N+1}{2}\}$.

N pair. On a $x_k < x_k^* \Rightarrow x_k \notin \mathscr{M}_X^* \Rightarrow \#\{i \in \mathscr{P} \,|\, X(i) \leq x_k\} = N_k < \frac{N}{2}$. Ceci montre que si on avait $x_k = \mathrm{Me}(X)$ avec $x_k < x_{k^*}$, l'inégalité de la première ligne du tableau (avec N pair) de la page 54 serait violée. Donc $x_k < x_{k^*} \Rightarrow x_k \neq \mathrm{Me}(X)$. Considérons maintenant le cas $x_k > x_{k^*}$, ou encore $x_{k-1} \geq x_{k^*}$. Notons que $x_{k-1} \geq x_{k^*} \Rightarrow N_{k-1} \geq N_{k^*} \Rightarrow N_{k-1} \geq \frac{N}{2}$. On a alors

$$\#\{i \in \mathscr{P} \,|\, X(i) < x_k\} = \#\{i \in \mathscr{P} \,|\, X(i) \leq x_{k-1}\} = N_{k-1} \geq \frac{N}{2}\,.$$

Cette inégalité montre que si $x_k = \mathrm{Me}(X)$, alors l'inégalité de la troisième ligne du tableau de la page 54 serait violée. Donc $x_k > x_{k^*} \Rightarrow x_k \neq \mathrm{Me}(X)$.

On a montré que $x_k \neq x_{k^*} \Rightarrow x_k \neq \mathrm{Me}(X)$. On a donc forcément $\mathrm{Me}(X) = x_{k^*} = \min\{x_k \in \mathscr{M}_X \,|\, N_k \geq \frac{N}{2}\}$.

Finalement, puisque $N_k \geq \frac{N}{2} \Leftrightarrow \frac{N_k}{N} \geq \frac{1}{2} \Leftrightarrow F_k \geq \frac{1}{2}$, l'ensemble $\mathscr{M}_X^*$ s'écrit aussi $\mathscr{M}_X^* = \{x_k \in \mathscr{M}_X \,|\, F_k \geq \frac{1}{2}\}$.

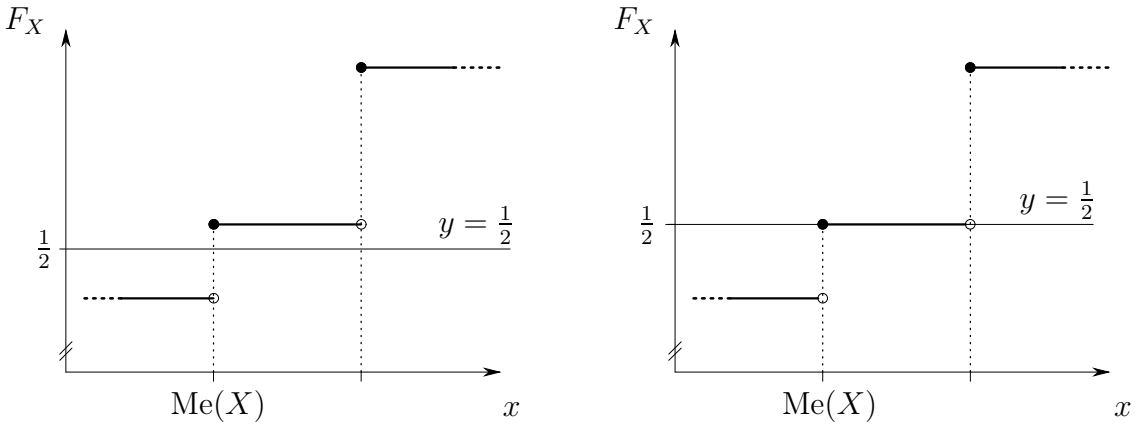
· N impair. Le résultat se montre exactement de la même façon que pour N pair. □

La propriété 6.3 permet de repérer facilement la médiane de X sur le graphe de sa fonction de répartition F_X. En effet, il suffit de repérer l'endroit où la « courbe » de cette fonction coupe la droite d'équation $y = \frac{1}{2}$. La médiane coïncide avec la plus petite modalité telle que la courbe de F_X est située au dessus de cette droite. Cette démarche est illustrée par les graphiques de la figure 6.1

6.2.2.7 Définition et détermination de la médiane cas des regroupements par classes

1. On se situe dans le cas où les classes sont $C_k = [e_{k-1}; e_k[$, $k = 1, \ldots, K-1$ et $C_K = [e_{K-1}; e_K]$.

FIG. 6.1 – Détermination de la médiane au moyen du graphe de la fonction de répartition



Définition 6.7 *La classe médiane est la classe contenant la modalité de l'individu médian.*

2. On peut montrer que la classe médiane se détermine de la même manière que la médiane. Plus précisément, on a le résultat suivant.

Propriété 6.4 *Soit* $k^* = \min\{k \in \mathcal{K} \,|\, N_k \geq \frac{N}{2}\}$, *où* $\mathcal{K} = \{1, \ldots, K\}$. *La classe médiane est* C_{k^*}, *la classe d'indice* k^*.

Démonstration : Définissons $\mathcal{K}^* = \{k \in \mathcal{K} \,|\, N_k \geq \frac{N}{2}\}$, ce qui permet d'écrire $k^* = \min \mathcal{K}^*$. Soit C_k une classe quelconque. Supposons que $k < k^*$, *i.e.*, $N_k < \frac{N}{2}$ et que $X(i_{\text{Me}}) \in C_k$. On aura

$$\#\{i \in \mathscr{P} \,|\, X(i) \leq \text{Me}(X)\} \leq \#\{i \in \mathscr{P} \,|\, X(i) < e_k\} = N_k < \frac{N}{2},$$

ce qui contredit la première ligne du tableau de la page 54. Par conséquent, pour qu'une classe C_k contienne la médiane, il faut que son indice k soit supérieur ou égal à k^*. Considérons alors une classe C_k avec $k > k^*$. On a nécessairement

$$N_k > N_{k-1} \geq N_{k^*} \geq \frac{N}{2},$$

puisque $N_1, \ldots, N_K$ forment une séquence croissante (voir propriété 4.2). Supposons que $X(i_{\text{Me}}) \in C_k$. On aura alors

$$\#\{i \in \mathscr{P} \,|\, X(i) < \text{Me}(X)\} \geq \#\{i \in \mathscr{P} \,|\, X(i) < e_{k-1}\} = N_{k-1} \geq \frac{N}{2},$$

ce qui contredit la troisième ligne du tableau de la page 54. On vient de montrer pour toute classe C_k avec $k \neq k^*$, $\text{Me}(X) \notin C_k$. Donc la classe médiane est nécessairement C_{k^*}.

Lorsque N est impair, le résultat se montre de la même façon, en utilisant la première colonne du tableau de la page 54. □

3. D'après la propriété 6.3, on a aussi $k^* = \min\{k \in \mathcal{K} \,|\, F_k \geq \frac{1}{2}\}$. Par conséquent, la propriété 6.4 permet de caractériser la classe médiane comme étant la plus petite classe[2] ayant une fréquence cumulée croissante supérieure ou égale à $\frac{1}{2}$.
4. Si la propriété 6.4 permet de déterminer dans quelle classe se situe la médiane, il n'est en général pas possible de connaître la valeur de la médiane. On définit alors la médiane de la manière suivante.

Définition 6.8 *Dans le cas d'un regroupement par classes, la médiane est définie comme le centre de la classe médiane.*

5. Avec cette définition, on remarque que la proposition 6.3, dans laquelle $x_1, \ldots, x_K$ désignent les centres de classes, peut s'utiliser pour déterminer la médiane.
6. Si les classes sont des intervalles prenant d'autres formes que celle introduite au début de cette section, cela ne change pas ses résultats et commentaires.
7. Lors d'un regroupement par classes, on a établit au point 3 que la classe médiane est la plus petite classe pour laquelle la fréquence cumulée croissante est supérieure ou égale à $\frac{1}{2}$. On peut donner une autre caractérisation de cette classe médiane en utilisant le polygone des fréquences cumulées. On utilise pour cela la propriété suivante.

Propriété 6.5

1. *Il existe une unique valeur $x^* \in [e_0; e_k]$ telle que $\hat{F}_X(x^*) = \frac{1}{2}$.*
2. *Soit k^* l'entier défini comme dans la propriété 6.4. On a $k^* = \min\{k \in \mathcal{K} \,|\, e_k \geq x^*\}$.*

Démonstration : Un théorème établit que si une fonction monotone g définie sur un intervalle I et à valeurs dans un intervalle L est continue sur I alors pour toute valeur $y \in L$ il existe une unique valeur $x \in I$ telle que $y = g(x)$. On note que la fonction $\hat{F}_X$, à valeurs dans $[0; 1]$, est continue et monotone sur $[e_0; e_K]$ (voir l'équation (5.1) et la propriété 5.1). On utilise le théorème pour déduire qu'il existe une unique valeur x^* telle que $\hat{F}_X(x^*) = \frac{1}{2}$. Pour montrer le second point, on utilise le fait que $\hat{F}_X$ est strictement croissante, ce qui permet d'écrire

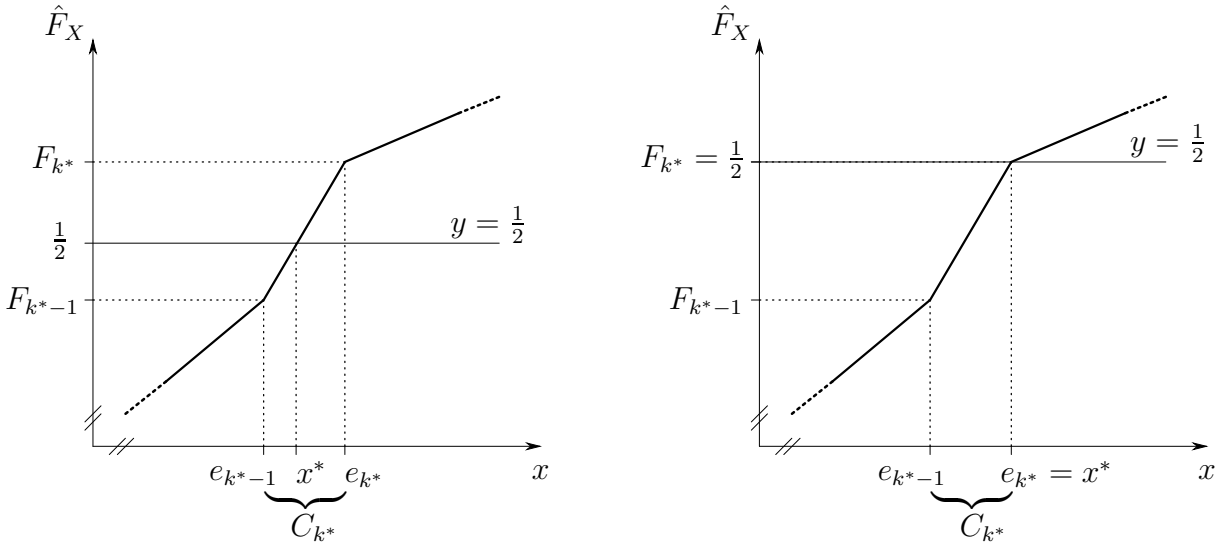
$$\min\{k \in \mathcal{K} \,|\, e_k \geq x^*\} = \min\{k \in \mathcal{K} \,|\, \hat{F}_X(e_k) \geq \hat{F}_X(x^*)\} = \min\{k \in \mathcal{K} \,|\, F_k \geq \tfrac{1}{2}\} = k^*,$$

où l'avant-dernière résulte de l'équation (5.1) et de la définition de x^*. □

[2]Voir le point 4 de la section 3.3.2.1.

8. En rapprochant l'expression de k^* donnée par la propriété 6.5 et la caractérisation de $\mathrm{Me}(X)$ fournie par la propriété 6.4, on conclut que la classe médiane est la classe dont l'extrêmité supérieure est e_{k^*}. D'après la propriété 6.5 et puisque $e_0, e_1, \ldots, e_K$ forment un séquence croissante (voir les inégalités (3.1) au point 4 de la section 3.3.2.1), on a $e_k < x^* \leq e_{k^*} < e_l$, $\forall k < k^* < l$. On peut donc écrire $e_{k^*} = \min\{e_0, \ldots, e_K \,|\, e_k \geq x^*\}$. Autrement dit, la classe médiane est la plus petite classe ayant une borne supérieure ou égale à x^*. Sur polygone des fréquences cumulées, x^* est le point l'axe des abscisses pour lequel la courbe de $\hat{F}_X$ coupe la droite horizontale d'équation $y = \frac{1}{2}$. On cherche alors la plus petite extrêmité e_{k^*} supérieure ou égale à x^* et la classe médiane est la classe dont les extrêmités inférieure et supérieure sont e_{k^*-1} et e_{k^*}.

FIG. 6.2 – Détermination de la classe médiane à partir du graphe de $\hat{F}_X$



9. L'illustration de cette méthode est donnée par les graphiques de la figure 6.2. L'intersection de la courbe de $\hat{F}_X$ et de la droite horizontale d'ordonnée $\frac{1}{2}$ permet de trouver x^* en abscisse. On trouve ensuite l'extremité supérieure e_{k^*} de la classe médiane en se déplaçant vers le haut le long de la courbe, à partir de son intersection avec la droite horizontale. Le premier coude trouvé correspond à un point d'abscisse e_{k^*}. Sur le graphique de droite, ce coude se situe au même endroit que l'intersection de la courbe avec la droite horizontale : e_{k^*} et x^* sont confondus.

6.2.3 La moyenne arithmétique

La moyenne arithmétique est un indicateur de position qui est déterminée sur la base d'une répartition égalitaire.

6.2.3.1 Définition

Définition 6.9 *La moyenne arithmétique (moyenne par la suite) de la variable X est le nombre noté $\overline{X}$ et défini par*

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X(i).$$

La moyenne, pour pouvoir être interprétée comme un indicateur de position, requiert de pouvoir additionner plusieurs modalités de la variable X. La moyenne ne s'utilise donc que pour des variables numériques.

6.2.3.2 Propriétés

Propriété 6.6

1. $\overline{X} = \frac{1}{N} \sum_{k=1}^{K} n_k x_k = \sum_{k=1}^{K} f_k x_k.$
2. $X(i) = X(j)\ \forall i, j = 1, \ldots, N \quad \Leftrightarrow \quad X(i) = \overline{X}\ \forall i = 1, \ldots, N.$
3. *Soit Y une transformation de X par g. On a* $\overline{Y} = \frac{1}{N} \sum_{k=1}^{K} n_k g(x_k).$

Démonstration :

1. Pour $k = 1, \ldots, K$, on définit $\mathscr{P}_{x_k} = \{i \in \mathscr{P} \,|\, X(i) = x_k\}$. On rappelle que $\#\mathscr{P}_{x_k} = n_k$ (voir définition 3.4 du chapitre 3). De plus, les K ensembles $\mathscr{P}_{x_1}, \ldots, \mathscr{P}_{x_K}$ forment évidemment une partition de $\mathscr{P}$. En utilisant les propriétés usuelles de l'addition (commutativité et associativité) on peut alors écrire

   $$\overline{X} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathscr{P}_{x_k}} X(i) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathscr{P}_{x_k}} x_k = \frac{1}{N} \sum_{k=1}^{K} n_k x_k.$$

   En utilisant la distributivité de la multiplication par rapport à l'addition, on a

   $$\frac{1}{N} \sum_{k=1}^{K} n_k x_k = \sum_{k=1}^{K} \frac{n_k}{N} x_k = \sum_{k=1}^{K} f_k x_k.$$

2. Si $X(i) = \overline{X}\ \forall i = 1, \ldots, N$, il est évident que $X(1) = \cdots = X(N)$. Réciproquement, si $X(1) = \cdots = X(N)$, alors

   $$\overline{X} = \frac{1}{N} \sum_{j=1}^{N} X(j) = \frac{1}{N} \sum_{j=1}^{N} X(i) = \frac{1}{N} N X(i) = X(i).$$

   Ceci est vrai $\forall i = 1, \ldots, N$.

3. On rappelle que le nombre de modalités de Y est noté J et que $\mathscr{M}_Y = \{y_1, \ldots, y_J\}$ (voir section 2.4 du chapitre 2). On forme les ensembles $\mathcal{K}_j = \{k \in \mathcal{K} \,|\, g(x_k) = y_j\}$, $j = 1, \ldots, J$, où $\mathcal{K} = \{1, \ldots, K\}$. Comme $g : \mathscr{M}_X \to \mathscr{M}_Y$ est surjective, $\mathcal{K}_j \neq \varnothing$, $j = 1, \ldots, J$. On note m_j l'effectif de la modalité y_j de Y : $m_j = \#\{i \in \mathscr{P} \,|\, Y(i) = y_j\}$. Notons que $Y(i) = y_j \;\Leftrightarrow\; \exists k \in \mathcal{K}_j, X(i) = x_k$. Autrement dit, $\{i \in \mathscr{P} \,|\, Y(i) = y_j\} = \bigcup_{k \in \mathcal{K}_j} \mathscr{P}_{x_k}$, où les ensembles $\mathscr{P}_{x_1}, \ldots \mathscr{P}_{x_K}$ sont définis comme dans la preuve du point précédent. Ces ensembles sont disjoints et par conséquent

$$m_j = \#\{i \in \mathscr{P} \,|\, Y(i) = y_j\} = \# \bigcup_{k \in \mathcal{K}_j} \mathscr{P}_{x_k} = \sum_{k \in \mathcal{K}_j} \#\mathscr{P}_{x_k} = \sum_{k \in \mathcal{K}_j} n_k. \tag{6.1}$$

On a alors :

$$\begin{aligned} \overline{Y} &= \frac{1}{N} \sum_{j=1}^{J} m_j y_j = \frac{1}{N} \sum_{j=1}^{J} \big(\sum_{k \in \mathcal{K}_j} n_k \big) y_j = \frac{1}{N} \sum_{j=1}^{J} \big(\sum_{k \in \mathcal{K}_j} n_k y_j \big) \\ &= \frac{1}{N} \sum_{j=1}^{J} \big(\sum_{k \in \mathcal{K}_j} n_k g(x_k) \big), \end{aligned} \tag{6.2}$$

où la première égalité vient du point 1 de la propriété, la deuxième de la relation (6.1), la troisième de la distributivité de la multiplication par rapport à l'addition, et la quatrième du fait que $k \in \mathcal{K}_j \;\Leftrightarrow\; y_j = g(x_k)$.

D'autre part, comme $\mathcal{K}_1, \ldots, \mathcal{K}_J$ forment une partition de $\mathcal{K}$, la commutativité et l'associativité de l'addition permettent d'écrire

$$\frac{1}{N} \sum_{k=1}^{K} n_k g(x_k) = \frac{1}{N} \sum_{j=1}^{J} \sum_{k \in \mathcal{K}_j} n_k g(x_k). \tag{6.3}$$

En combinant (6.2) et (6.3), on obtient le résultat voulu. □

Remarque. L'utilisation la plus courante du résultat précédent concerne le cas où $g(x) = ax + b$, où a et b sont des nombres réels quelconques. La propriété précédente permet de montrer que dans ce cas, $\overline{Y} = a\overline{X} + b$. En effet :

$$\begin{aligned} \overline{Y} &= \frac{1}{N} \sum_{k=1}^{K} n_k (a x_k + b) = \frac{1}{N} \sum_{k=1}^{K} a n_k x_k + \frac{1}{N} \sum_{k=1}^{K} n_k b \\ &= a \frac{1}{N} \sum_{k=1}^{K} n_k x_k + b \frac{1}{N} \sum_{k=1}^{K} n_k = a\overline{X} + b, \end{aligned}$$

puisque $\sum_{k=1}^{K} n_k = N_K = N$.

6.2.3.3 Interprétation

Pour interpréter la moyenne comme mesure de position, on peut avoir recours au raisonnement suivant.

Puisque l'utilisation de $\overline{X}$ nécessite que X soit numérique, on peut voir les modalités de X comme des montants (ou quantités) exprimés dans une certaine unité de mesure. On interprète alors $X(i)$ comme le montant de l'individu i et sur l'ensemble de la population $\mathscr{P}$, le montant total est $\sum_{i=1}^{N} X(i)$. On observe que la répartition du montant total $\sum_{i=1}^{N} X(i)$ sur la population $\mathscr{P}$ est $X(1), \ldots, X(N)$, dans le sens où on observe que sur le montant total, un montant $X(1)$ est attribué à l'individu 1, ..., un montant $X(N)$ est attribué à l'individu N.

On peut alors envisager une répartition autre que celle qu'on observe. Cela revient à re-répartir intégralement le montant total $\sum_{i=1}^{N} X(i)$ sur $\mathscr{P}$ de sorte qu'un montant $\tilde{X}(1)$ soit attribué à l'individu 1, ..., un montant $\tilde{X}(N)$ soit attribué à l'individu N, et que $\sum_{i=1}^{N} \tilde{X}(i) = \sum_{i=1}^{N} X(i)$ Il existe beaucoup de telles re-répartitions possibles. Parmi celles-ci, il existe une *répartition égalitaire* du montant total $\sum_{i=1}^{N} X(i)$ sur $\mathscr{P}$, qui est telle que (a) l'intégralité du montant est attribuée et (b) chaque individu se voit attribuer le même montant. Formellement, si on note $\overline{X}(1), \ldots, \overline{X}(N)$ cette répartition égalitaire de $\sum_{i=1}^{N} X(i)$ sur $\mathscr{P}$, ces deux conditions s'écrivent

$$(a) \quad \sum_{i=1}^{N} \overline{X}(i) = \sum_{i=1}^{N} X(i)$$

$$(b) \quad \overline{X}(1) = \overline{X}(2) = \cdots = \overline{X}(N)$$

On vérifie aisément que ces deux conditions sont équivalentes à

$$\overline{X}(i) = \overline{X}, \quad i = 1, \ldots, N.$$

Autrement dit, la re-répartition égalitaire sur $\mathscr{P}$ du montant total observé $\sum_{i=1}^{N} X(i)$ consiste à attribuer à chaque individu un montant égal à la moyenne arithmétique de X.

Ainsi, pour décrire la position des données $X(1), \ldots, X(N)$, la moyenne opère une redistribution égalitaire dans laquelle chaque individu se voit attribuer le même montant égal au montant total par individu $\frac{\sum_{i=1}^{N} X(i)}{N}$. La moyenne mesure la position par la valeur de ce montant.

6.2.4 Comparaison

Le mode, la médiane et la moyenne sont des indicateurs de position construits sur des principes distincts. Par conséquent, même s'ils sont destinés à mesurer la même caractéristique de la distribution de X, ils peuvent avoir (et ont) des propriétés distinctes.

On peut alors effectuer des comparaisons de ces propriétés et éventuellement dégager des situations dans lesquelles l'un de ces indicateurs est préférable à l'autre.

On commence par rappeler que pour certaines variables, on peut ne pas avoir le choix dans l'utilisation d'indicateurs de positions. Ainsi pour les variables nominales, seul le mode peut s'utiliser. Pour les variables ordinales, la moyenne n'est pas envisageable. Considérons le cas d'une variable numérique, pour laquelle il est possible d'utiliser chacun des indicateurs.

6.2.4.1 Le mode

1. Le mode a l'avantage d'être robuste à des erreurs d'observations. En effet, le mode est calculé en exploitant les effectifs (ou les fréquences) *seulement*. Par conséquent, si pour quelques individus de la population on se trompe en observant les modalités de la variable, il est fort possible que ces erreurs ne conduisent pas à une modification du mode. Pour illustrer cette remarque supposons par exemple que les données brutes soient

i	1	2	3	4	5	6	7	8	9	10
$X(i)$	17	1	10	11	18	10	9	18	19	10

   On vérifie que le mode de X est $\text{Mo}(X) = 10$. Supposons maintenant que pour l'individu 2, on ait mal saisi la modalité et qu'au lieu de 1, on ait $X(2) = 3$. Les données corrigées sont

i	1	2	3	4	5	6	7	8	9	10
$X(i)$	17	3	10	11	18	10	9	18	19	10

   Le mode corrigé est le même que celui qu'on avait calculé auparavant.

2. On a mentionné que le mode peut ne pas être unique. Cette caractéristique peut être un inconvénient. Cependant, si la bi- ou multimodalité est une caractéristique prononcée de la distribution de X, il peut être intéressant dans une étude statistique de le relever, ce que seul le mode peut faire en tant qu'indicateur de position.

   Si on reprend l'exemple ci-dessus en ajoutant deux individus de sorte que les données sont

i	1	2	3	4	5	6	7	8	9	10	11	12
$X(i)$	17	3	10	11	18	10	9	18	19	10	9	18

   on constate qu'il y a deux modes : 10 et 18. Si ces données sont des notes (sur 20) obtenues par des étudiants, il est interessant de constater qu'il y a deux modes et donc deux groupes : celui constitué des étudiants dont la note se situe autour du premier mode (10) et celui constitué des étudiants dont la note se situe autour du second (18). La médiane et la moyenne ne peuvent capturer ce type de propriété.

3. Un inconvénient du mode lorsqu'il est utilisé pour des variables numériques est que sa définition (et donc sa détermination) ne repose que sur une utilisation limitée des propriétés de la distribution de X. On rappelle que celle-ci est constituée des couples $(n_1, x_1), \ldots, (n_k, x_k)$. Le mode est déterminé en négligeant la donnée de la deuxième composante de chacun de ces couples, *i.e.*, les modalités de la variable X.[3] Il est par conséquent courant de considérer que le mode fournit une indication assez pauvre de la position de la distribution de X. Il faut cependant se rappeler que le mode fournit une certaine mesure de cette position et que les autres indicateurs ne peuvent fournir (voir par exemple le point précédent sur la multi-modalité).

6.2.4.2 La médiane

1. Par rapport au mode, la médiane est déterminée en utilisant une caractéristique supplémentaire de la distribution de X qu'est un classement par ordre croissant des modalités.
2. Elle fournit une position en indiquant un « milieu » dans la distribution de X. Pour un individu i, connaissant sa modalité $X(i)$, et la valeur de la médiane de X, on peut dire si cet individu présente une modalité plutôt faible ($X(i) < \text{Me}(X)$) ou plutôt élevée ($X(i) > \text{Me}(X)$). « Plutôt faible » signifie que la modalité présentée par i se situe parmi la moitié des plus faibles modalités observées (et « plutôt fort » signifie le contraire)

   Par exemple, au cours des échographies précédant la naissance d'un enfant, on effectue des mesures sur la taille du fœtus : longueur d'un tibia (variable X), périmètre crânien (variable Y), ... Connaissant les médianes $\text{Me}(X)$ et $\text{Me}(Y)$ calculées sur une population donnée, on est capable de dire si pour chacune des variables le fœtus présente une modalité plutôt faible ou plutôt grande.
3. Le principal avantage de la médiane est la propriété qu'elle partage avec le mode d'être robuste à des erreurs d'observations. Si on reprend l'exemple du point 1 de la section précédente, la médiane est égale à 10 lorsqu'une erreur est commise sur la modalité de l'individu 2 et elle est également égale à 10 lorsque cette erreur est corrigée.

   De manière plus intéressante, la médiane n'est en général pas affectée lorsque des erreurs d'observation sont commises pour des individus présentant de fortes ou de faibles modalités de la variable, c'est à dire pour des individus qui sont loin du rang central ($\frac{N}{2}$ ou $\frac{N+1}{2}$). Dans ces cas, l'erreur commise ne change en général pas

[3] On rappelle que dans le cas de variables nominales, on ne peut définir de structure sur l'ensemble des modalités qui en permettrait leur exploitation. Comme la détermination du mode n'est pas basée sur la structure de cet ensemble, il est normal que le mode puisse s'utiliser pour de telles variables.

l'individu médian et laisse donc la médiane inchangée. C'est le cas par exemple pour l'individu 2 de l'illustration précédente, dont la modalité ($X(2) = 3$) est la plus petite parmi celles qui sont observée. Si on se trompe en observant 1 à la place de 3 pour cet individu, la médiane n'est pas modifiée.

Cette remarque prend son importance si on ajoute que dans beaucoup de situations, les modalités extrêmes (très grandes et très petites) d'une variable sont pour de nombreuses raisons mal observées. On peut penser à la variable richesse d'un ménage. Il est connu qu'il est difficile de connaître la richesse des ménages les moins riches ainsi que celle des ménages les plus riches. La médiane, en tant qu'indicateur de position, se montre insensible à ce genre de difficultés.

6.2.4.3 La moyenne

1. Contrairement au mode et à la médiane, la moyenne est systématiquement affectée par des erreurs d'observation. Elle l'est d'autant plus que ces erreurs sont importantes. Pour illustrer cette caractéristique, considérons le cas des 10 individus de la section 6.2.4.1 pour lequels, lorsqu'il y a une erreur concernant l'individu 2 on a

i	1	2	3	4	5	6	7	8	9	10
$X(i)$	17	1	10	11	18	10	9	18	19	10

   La moyenne est $\overline{X} = 12{,}3$. On calcule également $\text{Mo}(X) = \text{Me}(X) = 10$. Supposons maintenant que pour l'individu 2, la modalité correcte soit $X(2) = 10$ (on avait par exemple oublié de saisir le chiffre 0 du nombre 10). Une fois la correction effectuée, la moyenne devient 13,2 alors que le mode et la médiane sont inchangés.

2. La moyenne est un indicateur de position que l'on peut voir comme basé, comme on l'a noté dans la section 6.2.3.3, sur une re-répartition de $\sum_{i=1}^{N} X(i)$. Cette re-répartition est fictive (non-observée) et, même si on sait qu'elle correspond à une répartition égalitaire, on ne sait quelle rôle lui faire jouer dans une analyse statistique.

   Pour illustrer ce point, considérons le cas d'une population constituée de tous les habitants d'un pays, décrits par la variable X désignant le revenu. Pour avoir une mesure de position de la distribution de X dans le pays, on peut calculer la moyenne de X. Pour cela, on calcule le revenu total du pays et on divise par le nombre d'habitants. Cela équivaut à calculer ce qu'on appelle le revenu par habitant. Ce revenu est celui que possèderait chaque habitant si chacun avait le même niveau de revenu (si on redistribuait le revenu du pays de manière égalitaire). On peut se demander quel est l'intérêt de considérer ce type de re-répartition et de mesurer la position de X de cette manière. Pour cette même variable, la médiane est la

valeur du revenu telle qu'on peut trouver $\frac{N}{2}$ habitants du pays ayant un revenu inférieur ou égal à cette valeur, et $\frac{N}{2}$ habitants, distincts des précédents, ayant un revenu supérieur ou égal à cette valeur (pour N pair). Cette mesure de position est directement interprétable en terme de répartition observée du revenu dans le pays. La médiane de X est la valeur du revenu permettant de partitionner la population en deux sous-populations $\mathscr{P}_1$ et $\mathscr{P}_2$ d'effectifs égaux, telles que tout individu de $\mathscr{P}_2$ est au moins aussi riche que tout individu de $\mathscr{P}_1$. Cette dernière sous-population contient les individus les moins riches et la première contient les individus les plus riches (étant entendu qu'un individu est d'autant plus riche que son revenu est élevé). La médiane permet par conséquent de marquer une séparation entre le groupe des individus les moins riches et le groupe des individus les plus riches. Ces deux groupes ayant le même effectif, la médiane marque ainsi le milieu de la distibution du revenu dans le pays.

3. Notons finalement que la moyenne et la médiane peuvent être éloignées l'une de l'autre (cet écart est cependant borné). C'est ce qu'illustre l'exemple du point 1 ci-dessus. Dans ce cas, l'utilisation conjointe de ces deux indicateurs est conseillée. Pour reprendre le cas de la variable revenu, le revenu moyen est typiquement plus élevé que le revenu médian. Par exemple, en 2000, le revenu individuel moyen des habitants de Montréal (Canada) était de 24735 € et le revenu médian de 17551 € (conversion faite sur la base du taux de change moyen €/CAD en 2000). Cet écart entre les deux indicateurs présente un intérêt en soi. Ce point sera discuté dans la section 6.4.

6.3 Indicateurs de dispersion

Les indicateurs de position étudiés dans la section précédente servent à désigner l'endroit de la distribution de X autour duquel se positionnent les modalités observées de X. Utilisé séparément des autres, aucun ne fournit de renseignement sur la façon dont les modalités se positionnent autour de cet endroit. Des caractéristiques intéressantes de la distribution de X sont

1. la plus ou moins grande dispersion des modalités observées dans leur ensemble ;

2. la plus ou moins grande dispersion des modalités dans leur regroupement autour d'une tendance centrale.

Les indicateurs de dispersion présentés dans cette section permettent de décrire ces caractéristiques. On distingue deux familles d'indicateurs de dispersion. La première contient des indicateurs de dispersion absolue dans le sens où on s'interesse à la première forme de dispersion. La seconde contient des indicateurs qui mesurent la dispersion des modalités

observées lorsqu'une mesure de position est prise comme point de référence.

Il faut noter que les indicateurs présentés ici utilisent tous une notion de distance, qui est elle même fondée sur le calcul d'une différence entre modalités. Pour cette raison, les indicateurs de dispersion qui seront étudiés *ne s'utilisent que pour des variables numériques.*

6.3.1 Les indicateurs de dispersion absolue : l'étendue et l'étendue interquartile

6.3.1.1 L'étendue

Définition 6.10 *On appelle étendue de (la distribution de) la variable X le nombre noté $ETD(X)$ et défini par $ETD(X) = X^M - X_m$ avec*

$$X^M = \max\{X(i), i \in \mathscr{P}\},$$
$$X_m = \min\{X(i), i \in \mathscr{P}\}.$$

Interprétation. L'étendue mesure la dispersion de la distribution de X par la distance maximale qu'on observe entre deux modalités de X. Plus l'étendue est grande, plus la dispersion est grande.

Propriété 6.7 *Soit Y la variable obtenue en transformant X par g.*

1. *Si g est une application croissante, alors* $\mathrm{ETD}(Y) = g(X^M) - g(X_m)$.
2. *Si g est une application décroissante, alors* $\mathrm{ETD}(Y) = g(X_m) - g(X^M)$.
3. *Si $g(x) = ax + b$, où a et b sont des nombres quelconques, on a* $\mathrm{ETD}(Y) = |a|\mathrm{ETD}(X)$.

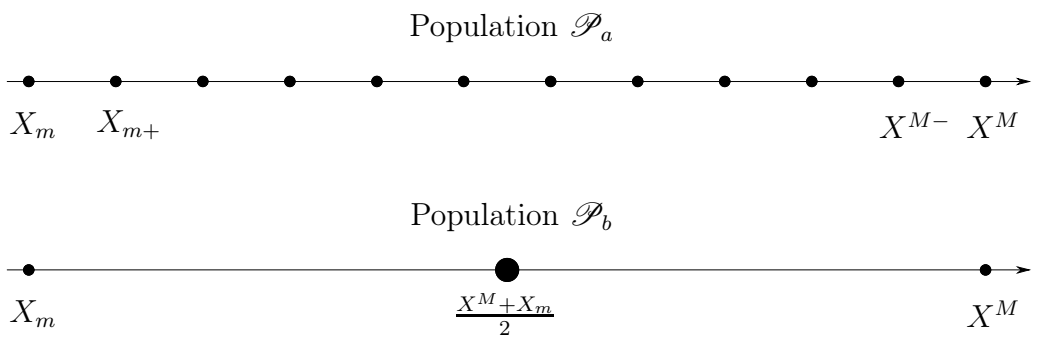
Démonstration :

1. Notons d'abord que $g(X^M)$ et $g(X_m)$ sont tous deux éléments de $\mathscr{M}_Y$. Si g est croissante, alors pour tout $x_k \in \mathscr{M}_X$, on a $g(X^M) \geq g(x_k) \geq g(X_m)$, ou encore $g(X^M) \geq y_j \geq g(X_m)$, pour tout $y_j \in \mathscr{M}_Y$.
2. Le raisonnement est le même qu'avant en notant cette fois que si g est décroissante, $g(X_m) \geq x_k \geq g(X^M)$, pour tout $x_k \in \mathscr{M}_X$.
3. On peut utiliser directement les deux points précédents. On peut aussi noter que si $a > 0$, $Y_m = aX_m + b$ et $Y^M = aX^M + b$, d'où $Y^M - Y_m = a(X^M - X_m)$. Si $a > 0$, $Y_m = -aX^M + b$ et $Y^M = -aX^m + b$, d'où $Y^M - Y_m = -a(X^M - X_m)$.□

Remarques.

1. La seule façon dont l'étendue utilise les valeurs des modalités autres que X^M et X_m réside dans le classement des modalités nécessaire à la détermination de ces deux

FIG. 6.3 – Répartition et étendue de X dans les populations $\mathscr{P}_a$ et $\mathscr{P}_b$



valeurs extrêmes.[4] Autrement dit, les modalités autres que X^M et X_m n'affectent l'étendue que par leur caractéristique d'être placées entre X^M et X_m. Une fois le classement fait, la valeur prise par l'étendue ne dépend plus des valeurs des modalités intermédiaires.

Pour illustrer cette propriété de l'étendue, considérons deux populations $\mathscr{P}_a$ et $\mathscr{P}_b$ toutes deux de taille N, décrites par une même variable X. On se place dans le cas où

$$\begin{aligned} & X^M = \max\{X(i), i \in \mathscr{P}_a\} = \max\{X(i), i \in \mathscr{P}_b\} \\ \text{et} \quad & X_m = \min\{X(i), i \in \mathscr{P}_a\} = \min\{X(i), i \in \mathscr{P}_b\}. \end{aligned}$$

On suppose aussi que dans $\mathscr{P}_a$, la distribution de X est telle que chaque individu présente une modalité différente des autres et que les modalités observées sont équidistantes dans l'intervalle $[X_m; X^M]$. On suppose également que dans $\mathscr{P}_b$, un seul individu présente la modalité X_m, un seul individu présente la modalité X^M et les $N-2$ autres individus présentent la même modalité $(X^M + X_m)/2$. Schématiquement, la distribution de X dans chacune des populations est représentée à la figure 6.3. Un axe horizontal orienté gauche-droite représente l'échelle de valeurs dans laquelle s'expriment les modalités de la variable X. Une modalité observée est représentée par un disque noir le long de cette échelle. Sur l'axe représentant la population $\mathscr{P}_b$, on a représenté la modalité $(X^M + X_m)/2$ par un disque de diamètre plus important, indiquant que plusieurs individus présentent cette modalité.

Dans ces deux populations, la dispersion de X, si elle est mesurée par l'étendue, sera la même. Cependant, l'étendue ignore les propriétés des modalités situées entre

[4] Il est faux d'affirmer, comme cela est fait dans de nombreux manuels de statistique, que l'étendue ne dépend pas des modalités intermédiaires, c'est à dire celles qui ne sont ni égales à X^M, ni égales à X_m. Pour s'en convaincre, il suffit d'exprimer $\mathrm{ETD}(X)$ de la manière suivante : $\mathrm{ETD}(X) = \max\{X(1), \ldots, X(N)\} - \min\{X(1), \ldots, X(N)\}$. Sous cette expression, l'étendue de X apparaît clairement comme une fonction de *toutes* les modalités de X.

X_m et X^M. Que celles-ci soient équidistantes comme dans $\mathscr{P}_a$ ou concentrées au même endroit comme dans $\mathscr{P}_b$, l'étendue reste la même. Or, après examen des graphiques de la figure 6.3, il peut être naturel de vouloir conclure que la dispersion des modalités de X n'est pas la même dans les deux populations. L'étendue ne permet pas de rendre compte de cette différence.

2. Notons que si on enlève des populations $\mathscr{P}_a$ et $\mathscr{P}_b$ les individus ayant les modalités X_m et X^M, alors le calcul de l'étendue de X dans ces populations retreintes donne des valeurs très différentes. En effet, pour la population $\mathscr{P}_a$, l'étendue devient égale à $X^{M-} - X_{m+}$, tandis que pour la population $\mathscr{P}_b$, celle-ci devient nulle. On voit par cette expérience que la valeur de l'étendue dépend de façon immédiate et déterminante des modalités extrêmes X^M et X_m.

   Or on a remarqué que ces modalités sont typiquement difficiles à mesurer sans erreur. Par conséquent, l'étendue est une mesure de dispersion qui souffre d'un manque de fiablilité. Pour éviter ce problème de sensibilité, on peut calculer une étendue sur une partie $\tilde{\mathscr{P}}$ de la population $\mathscr{P}$, $\tilde{\mathscr{P}}$ étant formée à partir de $\mathscr{P}$ en éliminant les individus présentant des valeurs extrêmes. Une méthode particulière de constitution de $\tilde{\mathscr{P}}$ conduit à l'*étendue interquartile*.

6.3.1.2 L'étendue interquatile

On rappelle que pour toute population $\mathscr{P}$ et toute variable (ordinale ou numérique) X, la médiane permet de constituer les parties $\mathscr{P}_1$ et $\mathscr{P}_2$ de $\mathscr{P}$, comme on l'a mentionné dans la section 6.2.2.3.

Définition 6.11 *Soit $\mathscr{P}$ une population et X une variable (numérique ou ordinale). On peut définir les trois quartiles de X de la façon suivante.*

- *Le premier quartile de X est la modalité de X, notée $Q_1(X)$, définie comme la médiane de X pour la sous-population $\mathscr{P}_1$.*
- *Le troisième quartile de X est la modalité de X, notée $Q_3(X)$, définie comme la médiane de X pour la sous-population $\mathscr{P}_2$.*
- *Le deuxième quartile de X est la modalité de X, notée $Q_2(X)$, définie par $Q_2(X) = \mathrm{Me}(X)$.*

Propriétés/interprétation des quartiles.

1. Il résulte des propriétés de $\mathscr{P}_1$ et de $\mathscr{P}_2$ que $Q_1(X) \leq Q_2(X) \leq Q_3(X)$.
2. $Q_1(X)$ permet de former les sous-populations $\mathscr{P}_{11}$ et $\mathscr{P}_{12}$ ayant le même effectif, et telles que pour tout individu i de $\mathscr{P}_{11}$, on a $X(i) \leq Q_1(X)$ et pour tout individu i de $\mathscr{P}_{12}$, on a $Q_1(X) \leq X(i) \leq Q_2(X)$.

De manière semblable $Q_3(X)$ permet de former les sous-populations $\mathscr{P}_{21}$ et $\mathscr{P}_{22}$ ayant le même effectif, et de sorte que pour tout individu i de $\mathscr{P}_{21}$, on a $Q_2(X) \leq X(i) \leq Q_3(X)$ et pour tout individu i de $\mathscr{P}_{22}$, on a $Q_3(X) \leq X(i)$.

3. De plus, $\mathscr{P}_{11}, \mathscr{P}_{12}, \mathscr{P}_{21}$, et $\mathscr{P}_{22}$ on le même effectif. Par conséquent, les trois quartiles permettent de former quatre sous-populations, d'effectifs égaux, constituée chacune d'environ $1/4$ des individus de $\mathscr{P}$. La particularité de ces sous-population est que

$$\forall i_1 \in \mathscr{P}_{11},\ \forall i_2 \in \mathscr{P}_{12},\ \forall i_3 \in \mathscr{P}_{21},\ \forall i_4 \in \mathscr{P}_{22},\ X(i_1) \leq X(i_2) \leq X(i_3) \leq X(i_4). \tag{6.4}$$

4. Comme les quartiles $Q_1(X)$ et $Q_3(X)$ sont des médianes de X dans les sous-populations $\mathscr{P}_1$ et $\mathscr{P}_2$, le contenu de la section 6.2.2 s'applique. En particulier, en cas de regroupement par classes, on peut déterminer les quartiles $Q_1(X)$ et $Q_3(X)$ en utilisant le polygone des fréquences cumulées, en suivant la méthode exposée au point 8 de la section 6.2.2.7.

   Pour déterminer la classe contenant le premier quartile $Q_1(X)$, on cherche le point P_{Q_1} situé à l'intersection du polygone des fréquences cumulées avec la droite horizontale d'équation $y = \frac{1}{4}$. L'abscisse de ce point est notée $x^*_{Q_1}$ et la classe contenant le quartile $Q_1(X)$ est la classe dont l'extrêmité supérieure est la plus petite extrêmité parmi celles qui sont au moins aussi grande que $x^*_{Q_1}$. Cela revient à chercher, en se déplaçant le long du polygone des fréquences cumulées vers le haut et à partir de P_{Q_1}, le coude du polygone le plus proche de P_{Q_1}. L'abscisse du point où se produit le coude est l'extrêmité supérieure de la classe contenant $Q_1(X)$.

   On procède de manière identique pour déterminer la classe contenant $Q_3(X)$, mais à partir du point d'intersection du polygone avec la droite d'équation $y = \frac{3}{4}$.

5. Les inégalités (6.4) montrent que les individus présentant des modalités extrêmes sont nécessairement situés dans les sous-populations $\mathscr{P}_{11}$ et $\mathscr{P}_{22}$. Par conséquent, si on veut construire un indicateur de dispersion semblable à l'étendue mais qui ne soit pas sensible à la présence de modalités extrêmes dans les données, on peut écarter pour le calcul les individus des sous-populations $\mathscr{P}_{11}$ et $\mathscr{P}_{22}$. Ce principe conduit à l'étendue interquartile.

Définition 6.12 *On appelle étendue interquartile de X, et on note* $\mathrm{EIQ}(X)$, *la valeur définie par* $\mathrm{EIQ}(X) = Q_3(X) - Q_1(X)$.

Interprétation et remarques.

1. L'étendue interquartile est l'étendue calculée sur la sous-population $\mathscr{P}_{12} \cup \mathscr{P}_{21}$, à laquelle on a ajouté, s'ils n'en faisaient pas déjà partie, l'individu médian et les individus i_{Q_1} et i_{Q_3} pour lesquels on a $X(i_{Q_1}) = Q_1(X)$ et $X(i_{Q_3}) = Q_3(X)$. On peut

également voir cette sous-population comme constituée (à éventuellement un individu près) de la moitié des individus $\mathscr{P}$, les individus étant exclus étant pour moitié des individus de modalité inférieure ou égale à $Q_1(X)$ et pour moitié des individus de modalité supérieure ou égale à $Q_3(X)$. La sous-population ainsi conservée ne contient pas d'individus présentant des modalités extrêmes de X. Par conséquent, en calculant l'étendue sur cette population "élaguée", cette mesure de dispersion n'est pas affectée par de possibles erreurs d'observations sur les modalités extrêmes.

2. L'étendue interquartile étant une étendue, elle s'utilise de la même manière pour mesurer la dispersion.
3. Le choix des quartiles pour élaguer la population est arbitraire. On peut utiliser à la place tout système de *quantiles* définis comme suit.

Définition 6.13 *Soit q un nombre compris entre 0 et 1. Le quantile d'ordre q de X est la modalité présentée par l'individu de rang $\lceil qN \rceil$, où pour tout nombre réel y, $\lceil y \rceil$ désigne le plus petit nombre entier supérieur ou égal à y.*

4. On constate que les quartiles $Q_1(X)$, $Q_2(X)$ et $Q_3(X)$ sont respectivement les quantiles d'ordre 0,25, 0,5, et 0,75 de X. Les *déciles* sont définis comme les quantiles d'ordre 0,1, 0,2, ...,0,9. Les *percentiles* sont les quantiles d'ordre 0,01, 0,02, ...,0,99.
5. L'étendue interquartile est donc l'étendue de X dans la sous-population constituée des individus dont les modalités sont comprises entre les quantiles d'ordre 0,25 et 0,75.

   En conservant ce principe, on peut choisir n'importe quelle valeur de $q < 0{,}5$ et calculer l'étendue de X dans la sous-population formée par les individus dont les modalités sont comprises entre les quantiles d'ordre q et $1 - q$. Comme l'objectif est ici d'élaguer la population en enlevant les individus présentant des modalités extrêmes, on choisit en général q petit ($q \leq 0{,}25$). Si on choisit $q = 0{,}1$, on est amené à calculer l'étendue de la sous-population composée des individus ayant une modalité comprise entre le 1$^{\text{er}}$ décile et le 9$^{\text{e}}$ décile. C'est un indicateur de dispersion formé sur le même principe que l'étendue interquartile, qu'on peut appeler *étendue interdécile*.

6.3.2 Les indicateurs de dispersion autour d'une tendance centrale

6.3.2.1 Le principe

Les indicateurs qui seront présentés ici seront tous construits d'après le principe suivant. On décrit une population $\mathscr{P}$ en utilisant une variable X, et on obtient les données

brutes $X(1), \ldots, X(N)$. On dispose aussi d'une mesure tendance centrale, obtenue à l'aide d'un des indicateurs de position (mode, médiane, moyenne). On note $\mathrm{TC}(X)$ cette mesure. Pour construire un indicateur de dispersion autour de $\mathrm{TC}(X)$, on commence par choisir une distance[5] sur $\mathscr{M}_X$, ou bien une fonction strictement croissante de cette distance, qu'on notera d. À l'aide de d, on peut pour tout $i \in \mathscr{P}$ calculer $D(i) = d\big(X(i), \mathrm{TC}(X)\big)$, qui mesure la distance entre la modalité de X présentée par i et la tendance centrale de X. On peut alors considérer $D(1), \ldots, D(N)$ comme les modalités observées de la variable $D : i \longmapsto d\big(X(i), \mathrm{TC}(X)\big)$. Ces modalités décrivent la dispersion de $X(1), \ldots, X(N)$ autour de $\mathrm{TC}(X)$. Un indicateur de dispersion de X autour de $\mathrm{TC}(X)$ est un résumé des distances $D(1), \ldots, D(N)$, *i.e.*, un résumé des valeurs de la variable D. Celui-ci est donné par l'un des indicateurs de position de la variable D. *On peut donc mesurer la dispersion de X par* $\mathrm{Mo}(D)$ *ou* $\mathrm{Me}(D)$ *ou* $\overline{D}$.

Si tous les indicateurs de dispersion présentés dans cette section obéissent à ce principe, ils se distinguent les uns des autres par les choix qui sont faits quant

1. au choix de d : on utilise soit la distance $d(a, b) = |a - b|$, soit son carré $d'(a, b) = d(a, b)^2 = (a - b)^2$;
2. au choix de $\mathrm{TC}(X)$: $\mathrm{Mo}(X)$ ou $\mathrm{Me}(X)$ ou $\overline{X}$;
3. au choix de la mesure de la position de D : $\mathrm{Mo}(D)$ ou $\mathrm{Me}(D)$ ou $\overline{D}$.

Par exemple, si $\mathrm{TC}(X) = \mathrm{Me}(X)$, si $d(a, b) = |a - b|$, et si la position de D est mesurée par $\overline{D}$, un indicateur de dispersion de X sera

$$\frac{1}{N} \sum_{i=1}^{N} |X(i) - \mathrm{Me}(X)|.$$

Si on utilise maintenant $\mathrm{Me}(D)$ comme indicateur de position de D, la mesure de dispersion de X sera $\mathrm{Me}(|X - \mathrm{Me}(X)|)$.

On présente deux mesures de dispersion couramment employées.

6.3.2.2 L'écart absolu moyen

Définition 6.14 *Pour une variable statistique X, l'écart absolu moyen (EAM) est la quantité définie par*

$$\mathrm{EAM}(X) = \frac{1}{N} \sum_{i=1}^{N} |X(i) - \overline{X}|.$$

[5]Une distance d sur un ensemble A est toute application à valeurs réelles définie sur $A \times A$ telle que pour n'importe quels éléments a, b, c de A on a (1) $d(a, b) \geq 0$ et $d(a, b) = 0 \Leftrightarrow a = b$, (2) $d(a, b) = d(b, a)$ et (3) $d(a, b) \leq d(a, c) + d(c, b)$.

Interprétation. Pour chaque $i \in \mathscr{P}$, $|X(i) - \overline{X}|$ est la distance entre usuelle entre $X(i)$ et $\overline{X}$.[6] L'EAM de X s'obtient directement comme la moyenne de ces distances. L'EAM de X est donc la moyenne des distances entre les modalités observées de X et leur moyenne. Plus $\text{EAM}(X)$ est élevé, plus cette dispersion est grande.

Propriété 6.8

1. $\text{EAM}(X) = \frac{1}{N}\sum_{k=1}^{K} n_k|x_k - \overline{X}| = \sum_{k=1}^{K} f_k|x_k - \overline{X}|$.
2. *Si Y est une transformation de X donnée par $Y = aX + b$, où a et b sont des nombres réels quelconques, alors* $\text{EAM}(Y) = |a|\text{EAM}(X)$.
3. $\text{EAM}(X) \geq 0$. *De plus,* $\text{EAM}(X) = 0 \Leftrightarrow X(1) = X(2) = \cdots = X(N)$.

Démonstration :

1. On obtient le résultat en utilisant la même démarche que dans la preuve de la propriété de la section 6.2.3.2.
2. Pour tout $i \in \mathscr{P}$, on a $Y(i) = aX(i) + b$, et en utilisant la remarque de la section 6.2.3.2, on a $\overline{Y} = a\overline{X} + b$. Par conséquent, $Y(i) - \overline{Y} = a\big(X(i) - \overline{X}\big)$ et $\text{EAM}(Y) = \frac{1}{N}\sum_{i=1} |a\big(X(i) - \overline{X}\big)| = \frac{1}{N}\sum_{i=1} |a|\,|X(i) - \overline{X}| = \frac{1}{N}|a|\sum_{i=1} |X(i) - \overline{X}| = |a|\text{EAM}(X)$.
3. Comme $\text{EAM}(X)$ est une somme de valeurs absolues divisée par $N > 0$, l'EAM est un rapport de deux nombres non-négatifs. Elle est donc non-négative. $\text{EAM}(X) = 0 \Leftrightarrow \sum_{i=1}^{N} |X(i) - \overline{X}| = 0 \Leftrightarrow |X(i) - \overline{X}| = 0$, $i = 1, \ldots, N$, où la dernière équivalence exprime le fait que la somme, à termes positifs ou nuls, est nulle si et seulement si tous ses termes sont nuls. La dernière condition est équivalente à $X(1) = X(2) = \cdots = X(N)$. □

Remarque.

1. Le point 3 de la propriété ci-dessus illustre que l'EAM satisfait une condition souhaitable d'un indicateur qui mesure la dispersion des modalités d'une variable autour de leur moyenne.

   En effet, $\text{EAM}(X)$ est toujours positive ou nulle. Elle atteint son minimum 0 si et seulement si les modalités de X sont toutes égales entre elles, c'est à dire si et seulement si elles sont toutes égales à leurs moyenne (voir le point 2 de la propriété de la section 6.2.3.2). Il est évident que dans ce cas, les modalités de X étant confondues avec leur moyenne, tout mesure de dispersion autour de la moyenne doit être à son minimum. C'est bien le cas avec l'EAM.

[6] La distance « usuelle » entre deux nombres réels a et b est celle qui est mesurée par $|a - b|$.

2. L'EAM de X est un nombre qui s'exprime dans la même unité de mesure que les modalités de X. Ainsi, si X est une variable dont les modalités sont exprimées en €, l'EAM de X sera exprimée en €. Si on calcule $\text{EAM}(X) = 2{,}7$ alors on pourra dire qu'en moyenne les distances entre les modalités de X et $\overline{X}$ sont de 2,7 €.

6.3.2.3 La variance et l'écart-type

La variance et l'écart-type diffèrent de l'EAM par le choix de la mesure de la distance entre $X(i)$ et $\overline{X}$. Au lieu d'utiliser la distance usuelle $|X(i) - \overline{X}|$, on utilise son carré $|X(i) - \overline{X}|^2 = (X(i) - \overline{X})^2$. Celui-ci n'est pas une distance au sens de la note 5 (le carré de la différence ne vérifie pas l'inégalité triangulaire $d(a,b) \leq d(a,c) + d(b,c)$). Cependant, $(X(i) - \overline{X})^2$ est une mesure de la distance $|X(i) - \overline{X}|$ entre $X(i)$ et $\overline{X}$, dans la mesure ou cette dernière est d'autant plus grande que le carré $(X(i) - \overline{X})^2$ est grand, et réciproquement.

S'il s'agit de mesurer la dispersion des modalités de X autour de leur moyenne, il faut disposer d'un indicateur qui signale si les distances entre les $X(i)$ et $\overline{X}$ sont élevées ou faibles dans leur ensemble. On voit donc que ce qui importe n'est pas d'utiliser la distance $|X(i) - \overline{X}|$ elle même, mais tout indicateur permettant d'établir si celle-ci est grande ou petite. Le carré $(X(i) - \overline{X})^2$ en est un. Sur ce principe est construite la variance de X.

Définition 6.15 *Pour une variable statistique X, la variance est la quantité notée* $\text{V}(X)$ *et définie par*

$$\text{V}(X) = \frac{1}{N} \sum_{i=1}^{N} \left(X(i) - \overline{X}\right)^2.$$

Interprétation. Comme on l'a noté ci-dessus, $\left(X(1) - \overline{X}\right)^2, \ldots \left(X(N) - \overline{X}\right)^2$ sont des mesures des distances qui séparent les modalités de leur moyenne. On constate que $\text{V}(X)$ est définie comme la moyenne de ces mesures. Elle mesure donc la dispersion des modalités de X autour de leur moyenne.

Propriété 6.9

1. $\text{V}(X) = \frac{1}{N} \sum_{k=1}^{K} n_k (x_k - \overline{X})^2 = \sum_{k=1}^{K} f_k (x_k - \overline{X})^2$.
2. *Si Y est une transformation de X donnée par $Y = aX + b$, où a et b sont des nombres réels quelconques, alors* $\text{V}(Y) = a^2 \text{V}(X)$.
3. $\text{V}(X) \geq 0$. *On a l'égalité* $\text{V}(X) = 0$ *si et seulement si* $X(1) = X(2) = \cdots = X(N)$.

Démonstration :

1. On obtient le résultat en utilisant la même démarche que dans la preuve de la propriété de la section 6.2.3.2.

2. Pour tout $i \in \mathscr{P}$, on a $Y(i) = aX(i) + b$, et en utilisant la remarque de la section 6.2.3.2, on a $\overline{Y} = a\overline{X} + b$. Par conséquent, $Y(i) - \overline{Y} = a\big(X(i) - \overline{X}\big)$ et $\mathrm{V}(Y) = \frac{1}{N}\sum_{i=1}\big[a\big(X(i) - \overline{X}\big)\big]^2 = \frac{1}{N}\sum_{i=1} a^2\big(X(i) - \overline{X}\big)^2 = \frac{1}{N}a^2\sum_{i=1}\big(X(i) - \overline{X}\big)^2 = a^2\mathrm{V}(X)$.

3. Comme $\mathrm{V}(X)$ est une somme de carré divisée par $N > 0$, la variance est un rapport de deux nombres non-négatifs. Elle est donc non-négative. $\mathrm{V}(X) = 0 \Leftrightarrow \sum_{i=1}^{N}\big(X(i) - \overline{X}\big)^2 = 0 \Leftrightarrow (X(i) - \overline{X}\,)^2 = 0$, $i = 1, \ldots, N$, où la dernière équivalence exprime le fait que la somme, à termes positifs ou nuls, est nulle si et seulement si tous ses termes sont nuls. La dernière condition est équivalente à $X(1) = X(2) = \cdots = X(N)$. □

La formule suivante donne une autre expression de la variance.

Propriété 6.10 (Formule de König-Huygens[7])

$$\mathrm{V}(X) = \frac{1}{N}\sum_{i=1}^{N} X(i)^2 - \overline{X}^2\,.$$

On en déduit

$$\mathrm{V}(X) = \frac{1}{N}\sum_{k=1}^{K} n_k x_k^2 - \overline{X}^2 = \sum_{k=1}^{K} f_k x_k^2 - \overline{X}^2.$$

Démonstration : Pour tout $i = 1, \ldots, N$, on a $\big(X(i) - \overline{X}\big)^2 = X(i)^2 + \overline{X}^2 - 2X(i)\overline{X}$. Par conséquent

$$\begin{aligned}
\sum_{i=1}^{N}\big(X(i) - \overline{X}\big)^2 &= \sum_{i=1}^{N}\big(X(i)^2 + \overline{X}^2 - 2X(i)\overline{X}\big) \\
&= \sum_{i=1}^{N} X(i)^2 + \sum_{i=1}^{N}\overline{X}^2 - \sum_{i=1}^{N} 2X(i)\overline{X} && \text{(par associativité)} \\
&= \sum_{i=1}^{N} X(i)^2 + N\overline{X}^2 - 2\overline{X}\sum_{i=1}^{N} X(i) && \text{(par distributivité)} \\
&= \sum_{i=1}^{N} X(i)^2 + N\overline{X}^2 - 2N\overline{X}^2 = \sum_{i=1}^{N} X(i)^2 - N\overline{X}^2.
\end{aligned}$$

En divisant par N les deux membres des égalités ci-dessus, on obtient la première égalité du point de la propriété. Le deux autres égalités de ce point sont obtenues en appliquant à $\frac{1}{N}\sum_{i=1}^{N} X(i)^2$ la même démarche que dans la preuve de la propriété de la section 6.2.3.2. □

[7] Christiaan Huyghens (1629-1695), mathématicien néerlandais et Johann Samuel König (1712-1757), mathématicien allemand.

Remarques.

1. La formule de König-Huygens est surtout utile d'un point de vue pratique. Si on doit faire les calculs sur une calculette, alors il est plus aisé de calculer la somme de $X(1)^2, \ldots, X(N)^2$, puis de lui soustraire $\overline{X}^2$ que d'effectuer d'abord le carré des différences $X(1) - \overline{X}, \ldots, X(N) - \overline{X}$ puis d'en faire la somme, comme on le ferait si on utilisait la définition de $\mathrm{V}(X)$. De plus la formule de König-Huygens donne une expression de la variance dont il est facile de se souvenir. En effet la formule exprime la variance comme une différence de deux termes. Le second est la carré de la moyenne. Le premier est la moyenne des valeurs $X(1)^2, \ldots, X(N)^2$. Ces valeurs sont des modalités observées de la variable Y définie par $Y = X^2$. Autrement dit

$$\frac{1}{N}\sum_{i=1}^{N} X(i)^2 = \overline{Y} = \overline{X^2}.$$

   La formule de König-Huygens s'écrit alors

$$\mathrm{V}(X) = \overline{X^2} - \overline{X}^2$$

   et s'énonce souvent : « *La variance de X est la moyenne des carrés moins le carré de la moyenne* ».

2. On peut faire à propos de la variance la même remarque que pour l'EAM (voir la remarque de la section 6.3.2.2).

3. La variance est un nombre qui s'exprime à l'aide du carré de l'unité de mesure des modalités de X. Si X est une variable dont les modalités sont exprimées en mètre, alors la variance de X s'exprime en mètre carré. Cependant, même si X décrit une longueur, la quantité $\mathrm{V}(X)$ ne peut s'interpréter aisément et naturellement comme une surface. Plus encore, si X s'exprime en €, il est impossible de donner un sens à un nombre exprimé en €^{2}. Pour cette raison, on peut utiliser l'écart-type à la place de la variance.

Définition 6.16 *Pour une variable statistique X, l'écart-type est la quantité notée $\sigma(X)$ et définie par*

$$\sigma(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X(i) - \overline{X}\right)^2}.$$

Remarques et propriétés.

1. L'écart-type mesure la dispersion des modalités de X autour de leur moyenne de la même manière que la variance : cette dispersion est d'autant plus petite que $\sigma(X)$ est petit. On peut également faire à propos de $\sigma(X)$ les mêmes

remarques que pour l'EAM et la variance (voir la remarque de la section 6.3.2.2 et le point 2 de la remarque ci-dessus). Par conséquent, on peut dire que la dispersion de X est d'autant plus petite que son écart-type est proche de 0.

2. De son lien avec $\mathrm{V}(X)$, l'écart-type tire toutes ses propriétés de la variance. Par exemple, les points 2 et 3 de la propriété 6.9 ci-dessus donnent

$$\begin{aligned} &i)\ \ Y = aX + b \Rightarrow \sigma(Y) = |a|\sigma(X) \\ &ii)\ \ \sigma(X) = 0 \ \Leftrightarrow\ \mathrm{V}(X) = 0 \ \Leftrightarrow\ X(1) = X(2) = \cdots = X(N) \end{aligned}$$

L'utilsation la plus courante de cette propriété consiste à prendre $a = \frac{1}{\sigma(X)}$ et $b = -\frac{\overline{X}}{\sigma(X)}$. La variable Y est dans ce cas égale à

$$Y = \frac{X - \overline{X}}{\sigma(X)}.$$

La remarque qui suit la propriété 6.6 permet de montrer que $\overline{Y} = 0$ et le point $i)$ ci-dessus entraîne $\sigma(Y) = 1$. Réciproquement, si X est une variable telle que $\overline{X} = 0$ et $\sigma(X) = 1$, alors quels que soient les réels $a > 0$ et b, a variable $Y = aX + b$ a une moyenne $\overline{Y} = b$ et un écart-type $\sigma(Y) = a$.

6.3.3 Remarques sur les indicateurs de dispersion

1. Supposons que pour une variable X, on ait pour une population $\mathscr{P}$ obtenu une dispersion (obtenue avec l'EAM ou l'écart-type) égale à 4. La dispersion est-elle élevée ou faible ? Sans rien connaître d'autre, on ne peut répondre. Si maintenant on sait que la moyenne est égale à 1, on peut juger cette dispersion assez élevée. Tout jugement sera alors établit sur la base d'une comparaison de la valeur de l'indice de dispersion avec la valeur de la moyenne.
2. Supposons maintenant que pour une population différente $\tilde{\mathscr{P}}$ décrite avec la même variable X, on ait une dispersion égale à 4, c'est à dire égale à celle de $\mathscr{P}$, mais que la moyenne de X pour $\tilde{\mathscr{P}}$ soit égale à 7. On aura tendance à juger que la dispersion est moins importante pour $\tilde{\mathscr{P}}$ que pour $\mathscr{P}$. Cette conclusion est une fois de plus établie sur la base d'une comparaison entre dispersion et moyenne.
3. Les deux points précédents montrent qu'il peut être utile de diposer d'indicateur qui mesurent la dispersion relativement à la valeur de la moyenne. Le points qui suivent consituent d'autres arguments dans cette direction.
4. Les propriétés 6.1, 6.2 et 6.6 du mode, de la médiane et de la moyenne montrent que si les modalités $X(1), \ldots, X(N)$ sont toutes multpliées par la même constante positive, alors ces indicateurs de position sont tous multipliés par cette constante.

Cette propriété est parfaitement souhaitable pour de tels indicateurs, puisqu'ils sont supposés mesurer la position des modalités. Si ces modalités sont toutes modifiées de la même façon, alors tout bon indicateur de position doit en rendre compte (si ce n'était pas le cas, un tel indicateur serait insensible à la position des données).

5. Par exemple si X est un prix en dollars, on peut vouloir le convertir en euros. Cette conversion consiste à multiplier toutes les modalités (tous les prix) par une constante a, qui désigne dans ce cas le taux de change \$/€. Il est normal que tout indicateur de position soit affecté par cette transformation des valeurs de X et la répercute en indiquant une position des prix en euros égale à a fois cette position lorsque les prix sont exprimés en dollars.
6. Les indicateurs de dispersion partagent cette propriété comme le montrent les propriétés 6.7, 6.8 et 6.9. Ainsi, si toutes les modalités sont multipliées par un nombre positif a (par exemple lors d'un conversion \$/€ de prix), alors la distance moyenne entre les modalités et leurs moyennes (EAM) est multipliée par a. Autrement dit, la dispersion devient a fois plus élevée. On peut trouver inopportune cette propriété. C'est la raison pour laquelle on utilise des mesures de dispersion *relative*.
7. De telles mesures sont construites selon le principe de la section 6.3.2.1, mais pour lequelles les distances entre les modalités et une tendance centrale sont mesurées relativement à cette tendance centrale. Cette méthode permet d'obtenir un écart absolu relatif moyen défini par

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{X(i)-\overline{X}}{\overline{X}}\right|$$

dans lequel les distances $|X(i)-\overline{X}|$ sont rapportées à $|\overline{X}|$. Ce même principe permet également de construire le coefficient de variation, plus répandu que la mesure précédente.

Définition 6.17 *On appelle coefficient de variation de X, et on note* $\mathrm{CV}(X)$, *la quantité définie par*

$$\mathrm{CV}(X)=\frac{\sigma(X)}{|\overline{X}|}\,.$$

8. Pour calculer la variance de X, on utilise pour mesurer la distance entre $X(i)$ et $\overline{X}$ la quantité $\big(X(i)-\overline{X}\,\big)^2$. Si pour les raisons exposées ci-dessus on utilise une mesure relative, on calculera $\left(\frac{X(i)-\overline{X}}{\overline{X}}\right)^2$. Avec une telle distance, la mesure de dispersion équivalente à la variance est

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{X(i)-\overline{X}}{\overline{X}}\right)^2=\frac{1}{\overline{X}^2}\,\frac{1}{N}\sum_{i=1}^{N}\big(X(i)-\overline{X}\,\big)^2=\frac{\mathrm{V}(X)}{\overline{X}^2}\,.$$

Le coefficient de variation est donc la racine carrée de cette « variance ».

Propriété 6.11

1. *L'écart absolu relatif moyen et le coefficient de variation sont respectivement l'écart absolu moyen et l'écart-type de la variable $\tilde{X}$ définie par $\tilde{X} = \frac{X}{\overline{X}}$.*
2. *Si Y est la variable définie par $Y = aX$, où a est un nombre réel quelconque non nul, alors l'écart absolu relatif moyen et le coefficient de variation de X sont respectivement égaux à ceux de Y.*

Démonstration : Cette propriété découle directement d'une application du point 2 des propriétés 6.8 et 6.9. □

Remarques.

1. Le premier point de la propriété 6.11 montre que l'écart absolu relatif moyen et le coefficient de variation mesurent la dispersion de X de la même façon que l'EAM et l'écart-type.
2. Le deuxième point de cette propriété montre que ces indicateurs ne sont pas affectés par l'unité dans laquelle sont exprimées les modalités de X, pourvu que le passage d'une unité à l'autre se fasse de manière affine.
3. Le coefficient de variation est souvent défini comme le rapport entre l'écart-type et la moyenne $\sigma(X)/\overline{X}$. Cependant, cette définition comporte deux inconvénients. Le premier est que le coefficient de variation défini ainsi est sensible au changement d'unité de mesure si celle-ci affecte le signe des modalités. Le second réside dans le fait qu'avec une telle définition, le coefficient de variation ne partage pas les propriétés de positivité des mesures des dispersion (EAM, variance et écart-type) présentées jusqu'ici.

6.4 Indicateurs de forme : asymétrie, aplatissement

1. Les indicateurs de forme permettent de décrire par des valeurs numériques l'allure (la forme) de la distribution statistique d'une variable X. On entend par forme de la distribution de X celle de son histogramme. Celle-ci peut être caractérisée par plusieurs de ses aspects. Les plus étudiés sont la multimodalité, l'aplatissement et la symétrie.
2. Il est à noter que les aspects auxquels on s'intéresse dans cette section sont définis à partir des deux dimensions de l'histogramme : sa hauteur et sa largeur. Cette dernière ne peut avoir d'interprétation que si la variable dont on étudie l'histogramme est numérique. Le contenu de cette section ne s'applique par conséquent qu'à de telles variables.

6.4.1 Indicateurs d'asymétrie

6.4.1.1 (A)symétrie : définition, interprétation et propriétés

En statistique, la notion de symétrie (et son contraire l'asymétrie) réfère à la carctéristique suivante de la distribution d'une variable.

Définition 6.18 *Soit X une variable aléatoire numérique ayant K modalités réelles distinctes $x_1 < x_2 < \cdots < x_K$. On dit que la distribution de X sur $\mathscr{P}$ est symétrique si*

1. $\exists x^* \in \mathbb{R}$, $\frac{1}{2}(x_k + x_{K-k+1}) = x^*$ $k = 1, \ldots, \left\lceil \frac{K}{2} \right\rceil$,
2. $n_k = n_{K-k+1}$, $k = 1, \ldots, \left\lceil \frac{K}{2} \right\rceil$,

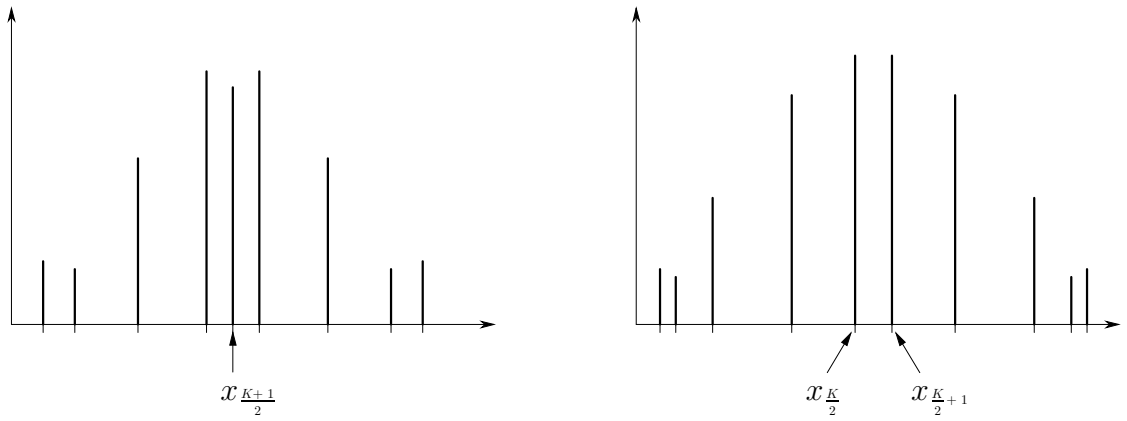
où on rappelle que pour tout nombre réel y, $\lceil y \rceil$ est le plus petit nombre entier supérieur ou égal à y.

1. On note immédiatement que si K est pair, alors $\left\lceil \frac{K}{2} \right\rceil = \frac{K}{2}$, tandis que si K est impair, on a $\left\lceil \frac{K}{2} \right\rceil = \frac{K+1}{2}$.
2. On remarque aussi que la convention de numérotation par ordre croissant des modalités de X entraîne $x_k < x_{K-k+1}$, pour $k = 1, \ldots, \left\lceil \frac{K}{2} \right\rceil$. Par conséquent, la condition 1 établit qu'on peut trouver un nombre x^* qui est le centre de *tous* les intervalles $[x_k; x_{K-k+1}]$, $k = 1, \ldots, \left\lceil \frac{K}{2} \right\rceil$. Les modalités du couple (x_k, x_{K-k+1}) sont donc équidistantes de x^*, avec $x_k < x^* < x_{K-k+1}$.
3. Si K est pair, la condition 1 est en particulier vraie pour $k = \frac{K}{2}$ pour lequel elle s'écrit $x^* = \frac{1}{2}(x_{\frac{K}{2}} + x_{\frac{K}{2}+1})$. De la même façon, si K est impair, pour $k = \left\lceil \frac{K}{2} \right\rceil = \frac{K+1}{2}$, on a $K - k + 1 = \frac{K+1}{2}$ et la condition 1 s'écrit $x^* = x_{\frac{K+1}{2}}$. Ces égalités donnent la valeur centrale x^* dans les cas K pair et K impair.
4. Pour interpréter la condition 2, on considère d'abord le cas où K est impair. Dans ce cas, les points qui précèdent permettent d'établir que la modalité $x_{\frac{K+1}{2}}$ se trouve au centre de l'intervalle $[x_1; x_K]$. On considère deux modalités x_k et x_l à égale distance de $x_{\frac{K+1}{2}}$, telles que $x_k < x_{\frac{K+1}{2}} < x_l$. La condition 2 établit alors que ces deux modalités ont le même effectif. Pour une telle variable, l'histogramme a la forme représentée par le graphique de gauche de la figure 6.4.

   Si K est pair, les deux modalités centrales sont $x_{\frac{K}{2}}$ et $x_{\frac{K}{2}+1}$. La condition 2 établit que si on choisit deux modalités x_k et x_l telles que $x_k \leq x_{\frac{K}{2}} < x_{\frac{K}{2}+1} \leq x_l$, et telles que la distance entre x_k et $x_{\frac{K}{2}}$ est la même que celle entre $x_{\frac{K}{2}+1}$ et x_l, alors $n_k = n_l$. Dans ce cas, l'histogramme a l'allure de celui de droite sur la figure 6.4.

5. On constate qu'une distribution est symétrique si l'histogramme des fréquences qui lui correspond est symétrique autour de l'axe vertical d'équation $x = x^*$, avec $x^* =$

FIG. 6.4 – Allure des histogrammes pour des distributions symétriques



$\frac{1}{2}\left(x_{\frac{K}{2}} + x_{\frac{K}{2}+1}\right)$ si K est pair, et $x^* = x_{\frac{K+1}{2}}$ si K est impair. La distribution d'une variable X est donc symétrique lorsque les modalités de X se répartissent dans la population de manière identique à droite et à gauche de la (des) valeur(s) centrale(s).

5. Un propriété importante des distributions symétriques est la suivante.

Propriété 6.12 *Si X est une variable statistique dont la distribution est symétrique, alors $\overline{X} = \mathrm{Me}(X) = x^*$, lorsque la médiane est définie comme au point 5 de la section 6.2.2 dans le cas où N est pair. Si de plus le mode de X est unique, alors $\overline{X} = \mathrm{Me}(X) = \mathrm{Mo}(X) = x^*$.*

Démonstration :

(a) K pair. On commence par montrer que $\frac{N}{2} = N_{\frac{K}{2}}$. En utilisant l'associativité et commutativité de l'addition, on a

$$N = \sum_{k=1}^{K} n_k = \sum_{k=1}^{\frac{K}{2}} n_k + \sum_{k=\frac{K}{2}+1}^{K} n_k = 2\sum_{k=1}^{\frac{K}{2}} n_k = 2N_{\frac{K}{2}} \tag{6.5}$$

où la première égalité provient de la propriété 3.1, la troisième égalité résulte de la condition 2 de la définition 6.18, et la dernière résulte de la définition 4.4 des effectifs cumulés croissants. Notons que ce résultat permet d'exprimer N comme le double du nombre entier $N_{\frac{K}{2}}$. N est donc pair. Par conséquent, si on applique la définition de la médiane donnée au point 5 de la section 6.2.2, la médiane est la moyenne des modalités des individus de rangs $\frac{N}{2}$ et $\frac{N}{2}+1$. On déduit de l'égalité $\frac{N}{2} = N_{\frac{K}{2}}$ que l'individu de rang $\frac{N}{2}$ présente la modalité $x_{\frac{K}{2}}$. La modalité présentée par l'individu de rang $\frac{N}{2}+1$ est soit $x_{\frac{K}{2}}$, soit $x_{\frac{K}{2}+1}$. Si cette modalité était $x_{\frac{K}{2}}$, alors on aurait $N_{\frac{K}{2}} \geq \frac{N}{2}+1$, ce qui contredit l'égalité $\frac{N}{2} = N_{\frac{K}{2}}$. L'individu de

rang $\frac{N}{2}+1$ présente donc la modalité $x_{\frac{K}{2}+1}$ et la médiane est

$$\mathrm{Me}(X) = \frac{1}{2}(x_{\frac{K}{2}} + x_{\frac{K}{2}+1}) = x^*,$$

d'après le point 3. On calcule ensuite la moyenne en utilisant le point 1 de la définition 6.18 :

$$\begin{aligned}\overline{X} &= \frac{1}{N}\sum_{k=1}^{K} n_k x_k = \frac{1}{N}\left[\sum_{k=1}^{\frac{K}{2}} n_k x_k + \sum_{k=\frac{K}{2}+1}^{K} n_k x_k\right] = \frac{1}{N}\sum_{k=1}^{\frac{K}{2}} n_k (x_k + x_{K-k+1}) \\ &= \frac{1}{N}\sum_{k=1}^{\frac{K}{2}} n_k 2x^* = \frac{2x^*}{N}\sum_{k=1}^{\frac{K}{2}} n_k = x^* \frac{2N_{\frac{K}{2}}}{N}\end{aligned}$$

où la troisième égalité provient de la condition 2 de la définition 6.18, la quatrième de la condition 1 de cette même définition, et la dernière de la définition des effectif cumulés croissants, les autre égalités étant obtenues par associativité et commutativité de l'addition, et par distributivite de la multiplication par rapport à l'addition. En utilisant l'équation (6.5), on obtient $\overline{X} = x^*$. Notons enfin que le mode ne peut être unique. En effet, si on se donne n'importe quel nombre y dans $[0;1]$, si une modalité a un effectif plus élevé que y, alors la condition 1 implique qu'il existe nécessairement une autre modalité avec la même propriété. Il ne peut donc y avoir une seule modalité ayant un effectif plus élevé que les autres.

(b) K impair. On note que dans ce cas $\left\lfloor \frac{K}{2} \right\rfloor = \frac{K-1}{2}$. On commence par déterminer $\mathrm{Me}(X)$. Pour cela, on calcule

$$N = \sum_{k=1}^{K} n_k = \sum_{k=1}^{\frac{K-1}{2}} n_k + n_{\frac{K+1}{2}} + \sum_{k=\frac{K+3}{2}}^{K} n_k = 2\sum_{k=1}^{\frac{K-1}{2}} n_k + n_{\frac{K+1}{2}} = 2N_{\frac{K-1}{2}} + n_{\frac{K+1}{2}} \tag{6.6}$$

où la première égalité résulte de la propriété 3.1, la troisième de la condition 2 de la définition 6.18, et les autres proviennent de l'associativité et commutativité de l'addition. Cette relation s'écrit

$$N_{\frac{K-1}{2}} = \frac{N - n_{\frac{K+1}{2}}}{2} \tag{6.7}$$

ce qui implique $N_{\frac{K-1}{2}} < \frac{N}{2}$. D'autre part, en utilisant le point 2 de la propriété 4.2, on a

$$N_{\frac{K+1}{2}} = N_{\frac{K-1}{2}} + n_{\frac{K+1}{2}} = \frac{N - n_{\frac{K+1}{2}}}{2} + n_{\frac{K+1}{2}} = \frac{N}{2} + \frac{n_{\frac{K+1}{2}}}{2} \tag{6.8}$$

où la deuxième égalité provient de (6.7). Ceci permet de déduire que $N_{\frac{K+1}{2}} > \frac{N}{2}$. Si N est impair, les deux inégalités obtenues $N_{\frac{K-1}{2}} < \frac{N}{2}$ et $N_{\frac{K+1}{2}} > \frac{N}{2}$ impliquent

que l'individu de rang $\frac{N+1}{2}$ présente la modalité $x_{\frac{K+1}{2}}$ de X. Dans ce cas, $\text{Me}(X) = x_{\frac{K+1}{2}}$. Lorsque N est pair, la médiane est la moyenne des modalités des individus de rangs $\frac{N}{2}$ et $\frac{N}{2}+1$. Les inégalités $N_{\frac{K-1}{2}} < \frac{N}{2} < N_{\frac{K+1}{2}}$ s'écrivent

$$N_{\frac{K-1}{2}} < \frac{N}{2} \tag{6.9}$$

$$N_{\frac{K+1}{2}} \geq \frac{N}{2} + 1. \tag{6.10}$$

L'inégalité (6.9) implique que les individus de rang $\frac{N}{2}$ et $\frac{N}{2}+1$ ont une modalité strictement supérieure à $x_{\frac{K-1}{2}}$. L'inégalité (6.10) implique que les individus de rang $\frac{N}{2}$ et $\frac{N}{2}+1$ ont une modalité inférieure ou égale à $x_{\frac{K+1}{2}}$. On conclut que ces deux individus présentent la modalité $x_{\frac{K+1}{2}}$. La médiane de X est donc $\text{Me}(X) = x_{\frac{K+1}{2}} = x^*$, d'après le point 3 ci-dessus. On calcule maintenant la moyenne de X :

$$\begin{aligned}
\overline{X} &= \frac{1}{N}\sum_{k=1}^{K} n_k x_k = \frac{1}{N}\left[\sum_{k=1}^{\frac{K-1}{2}} n_k x_k + n_{\frac{K+1}{2}} x_{\frac{K+1}{2}} + \sum_{k=\frac{K+1}{3}}^{K} n_k x_k\right] \\
&= \frac{1}{N}\left[n_{\frac{K+1}{2}} x_{\frac{K+1}{2}} + \sum_{k=1}^{\frac{K-1}{2}} n_k (x_k + x_{K-k+1})\right] = \frac{1}{N}\left[n_{\frac{K+1}{2}} x^* + \sum_{k=1}^{\frac{K-1}{2}} n_k 2x^*\right] \\
&= x^* \frac{1}{N}\big(n_{\frac{K+1}{2}} + 2N_{\frac{K-1}{2}}\big)
\end{aligned}$$

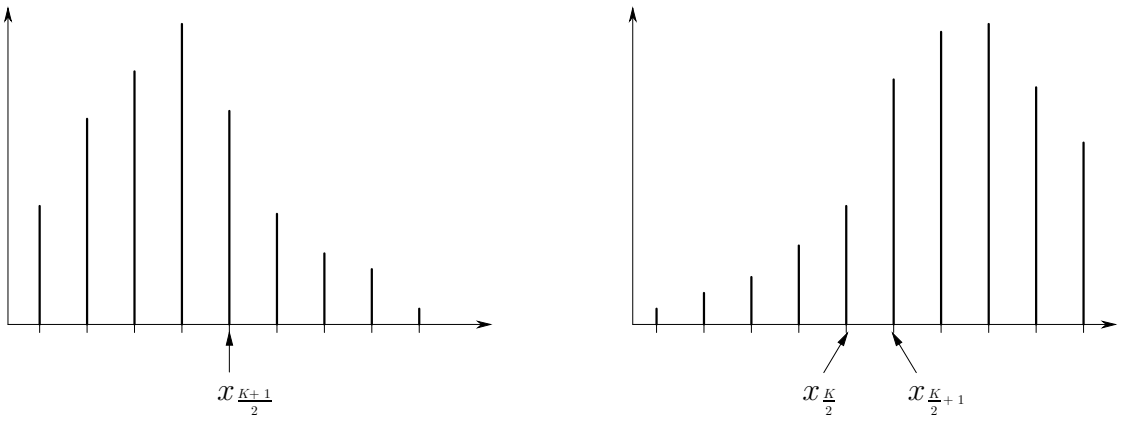
où la troisième égalité résulte de la condition 2 de la définition 6.18, la quatrième proovient de l'utilisation de la condition 1 de cette définition et du point 3, et la dernière de la définition des effectifs cumulés croissants. En utilisant l'équation (6.6), on déduit immédiatement que $\overline{X} = x^*$.

On montre finalement que si $\text{Mo}(X)$ est unique, alors $\text{Mo}(X) = x_{\frac{K+1}{2}}$. Supposer le mode unique revient à supposer qu'il existe une et une seule modalité de X, notée x_{k^*}, telle que $n_{k^*} > n_k$, $\forall k \neq k^*$. S'il se trouvait que $k^* \neq \frac{K+1}{2}$, la condition 2 de la définition 6.18 impliquant que $n_{k^*} = n_{K-k^*+1}$, on aurait une contradiction avec la condition $n_{k^*} > n_k$, $\forall k \neq k^*$ définissant x_{k^*} comme l'unique mode de X. Le mode existant toujours, on doit donc forcément avoir dans ce cas $k^* = \frac{K+1}{2}$ et donc $\text{Mo}(X) = x_{\frac{K+1}{2}} = x^*$. □

6. La propriété 6.12 montre qu'on peut baser un indicateur d'asymétrie de la distribution de X sur la distance entre sa moyenne et sa médiane, ou, lorsque le mode est unique, sur la distance entre la médiane et le mode ou le mode et la moyenne (voir la section 6.4.1.2).

7. Si la distribution n'est pas symétrique, elle peut présenter une « régularité » dans son asymétrie. Une distribution présente une asymétrie à gauche (resp. droite) si les

FIG. 6.5 – Allure des histogrammes pour des distributions asymétriques



densités de fréquence sont dans l'ensemble plus élevées pour les modalités inférieures (resp. supérieures) aux valeurs centrales que pour mes modalités supérieures (resp. inférieures). Les graphiques de la figure 6.5 présentent les deux types d'asymétrie.

8. En cas d'asymétrie, la propriété 6.12 indique simplement la médiane et la moyenne sont différentes. Cependant, cette propriété ne donne pas d'information sur la forme d'asymétrie (à droite ou à gauche). Il faut une étude un peu plus détaillée de la façon dont la présence d'asymétrie à gauche ou à droite affecte les positions relatives des indicateurs de tendance centrale en cas de d'asymétrie.

9. Pour cela, il est utile de se référer au cas simple d'une variable X n'ayant que trois modalités équidistantes x_1, x_2 et x_3, avec la convention $x_1 < x_2 < x_3$. Ceci équivaut à dire que si on définit $a = x_2 - x_1$, on doit avoir $x_2 = x_1 + a$ et $x_3 = x_1 + 2a$. La distribution de X est symétrique et unimodale si $n_1 = n_3 < n_2$ (ou encore $f_1 = f_3 < f_2$). On a dans ce cas $\text{Me}(X) = \overline{X} = \text{Mo}(X) = x_2$. Cette situation sera prise comme point de référence.

   Supposons qu'on modifie la distribution de X de manière à la rendre asymétrique à gauche. Pour cela, on change la valeur des fréquences de sorte que $f_1 > f_2 > f_3$ (voir le point 7 ci-dessus). Ceci implique que $1 = f_1 + f_2 + f_3 > 3f_3$, ou encore que $f_3 < \frac{1}{3}$. On montre que cette modification par rapport à la distribution symétrique rapproche le mode, la médiane et la moyenne de x_1.

   (a) Le mode. Puisque f_1 est la plus grande fréquence, la distribution de X reste unimodale, de mode x_1.

   (b) La médiane. Pour que x_3 soit la médiane, il faudrait que $f_1+f_2+f_3 \geq 0{,}5$, ce qui est le cas, mais il faudrait aussi que $f_1 + f_2 < 0{,}5$. Or l'identité $f_1 + f_2 + f_3 = 1$ et l'inégalité $f_3 < \frac{1}{3}$ impliquent que cette dernière condition est impossible. Pour que la médiance soit égale à x_1, il suffit que $f_1 > 0{,}5$, ce qui se produit

pour des distributions avec une asymétrie à gauche assez marquée.

La conséquence de l'asymétrie à gauche sur la médiane est donc de rapprocher celle-ci de x_1.

(c) La moyenne. En utilisant l'équidistance entre les trois modalités de X et l'identité $f_1 + f_2 + f_3 = 1$, on calcule

$$\overline{X} = f_1x_1 + f_2x_2 + f_3x_3 = f_1x_1 + f_2(x_1 + a) + f_3(x_1 + 2a) = x_1 + af_2 + 2af_3$$
$$= x_1 + a(f_2 + 2f_3).$$

On va montrer qu'avec l'asymétrie à gauche traduite par la condition $f_1 > f_2 > f_3 > 0$, cette moyenne est stritement inférieure à x_2. Une façon de ré-écrire les inégalités $f_1 > f_2 > f_3$ consiste à dire que pour des nombres $\epsilon > 0$ et $\eta > 0$, on peut écrire $f_1 = f_2 + \epsilon$ et $f_3 = f_2 - \eta$. Dans ce cas, l'identité $f_1 + f_2 + f_3 = 1$ devient équivalente à $3f_2 + \epsilon - \eta = 1$, et on doit donc avoir

$$f_2 = \frac{1 - \epsilon + \eta}{3}, \quad f_3 = \frac{1 - \epsilon - 2\eta}{3} \quad \text{et} \quad \overline{X} = x_1 + a(f_2 + 2f_3) = x_1 + a(1 - \epsilon - \eta).$$

La condition $f_3 > 0$ s'écrit $\frac{1-\epsilon}{2} > \eta$. On constate immédiatement que quelles que soient les valeurs $\epsilon > 0$ et $\eta > 0$ pour lesquelles $\frac{1-\epsilon}{2} > \eta$, on aura

$$x_1 + a(1 - \epsilon - \eta) < x_1 + a = x_2.$$

Ceci montre que si on modifie les fréquences d'une distribution symétrique de manière à la rendre asymétrique à gauche, alors on diminue sa moyenne.

En résumé, si on change la distribution symétrique et unimodale de X de manière à la rendre asymétrique à gauche, alors le mode passe de x_2 à x_1. La médiane est déplacée vers x_1 dès que $f_1 \geq 0{,}5$. La moyenne devient strictement inférieure à x_2.

Dans le cas général, le déplacement de la médiane est typiquement plus important que celui de la moyenne. Dans l'exemple ci-dessus, pour que la médiane passe de x_2 à x_1, il suffit que $f_1 \geq 0{,}5$. En revanche, pour que la moyenne soit égale à x_1 il faut que $f_1 = 1$ et $f_2 = f_3 = 0$. Ceci correspond à une distribution extrême dans laquelle tous les individus de la population présentent la modalité x_1.[8]

10. Si on voulait maintenant déformer la distribution de X pour que de symétrique, elle devienne asymétrique à droite, on choisirait $f_1 < f_2 < f_3$. La même démarche que dans le point précédent montrerait que le mode passe de x_2 à x_3, la moyenne et la médiane se déplacent de x_2 vers x_3.

[8] On ne peut plus dans ce cas parler de symétrie ou d'asymétrie.

11. Lorsqu'on considère des variables quelconques, les positions relatives du mode, de la médiane et de la moyenne sont

$$\begin{aligned} \mathrm{Mo}(X) < \mathrm{Me}(X) < \overline{X} \quad & \text{si la distribution de } X \text{ est asymétrie à gauche} \\ \mathrm{Mo}(X) > \mathrm{Me}(X) > \overline{X} \quad & \text{si la distribution de } X \text{ est asymétrie à droite} \end{aligned}$$

Cette caractéristique permettra de construire des indicateurs d'asymétrie.

12. On notera que l'asymétrie a aussi une influence sur le signe de la variable $(X - \overline{X})^p$, pour tout nombre entier positif impair p. Pour un individu i quelconque, on a évidemment $\big(X(i) - \overline{X}\big)^p > 0 \;\Leftrightarrow\; X(i) > \overline{X}$.

Dans le cas où X est symétrique, on observe avec une même fréquence des valeurs positives et négatives de $\tilde{X}$. En effet, si K est pair, alors $N_{\frac{K}{2}} = \frac{N}{2}$ (voir le début de la démonstration de la propriété 6.12) et $\overline{X} = \frac{1}{2}(x_{\frac{K}{2}} + x_{\frac{K}{2}+1})$. Par conséquent

$$\begin{aligned} & \#\{i \in \mathscr{P} \,|\, X(i) < \overline{X}\} = \#\{i \in \mathscr{P} \,|\, X(i) \leq x_{\frac{K}{2}}\} = N_{\frac{K}{2}} = \tfrac{N}{2}\,, \\ \text{et} \quad & \#\{i \in \mathscr{P} \,|\, X(i) > \overline{X}\} = \#\{i \in \mathscr{P} \,|\, X(i) \geq x_{\frac{K}{2}+1}\} \\ & \qquad\qquad = N - \#\{i \in \mathscr{P} \,|\, X(i) \leq x_{\frac{K}{2}}\} = N_{\frac{K}{2}} = \tfrac{N}{2}\,. \end{aligned}$$

Autrement dit, le nombre d'individus i pour lesquels $\tilde{X}(i) = X(i) - \overline{X}$ est strictement positif est égal au nombre d'individus j pour lesquels $\tilde{X}(j)$ est strictement négatif. Si K est impair, $\overline{X} = x_{\frac{K+1}{2}}$, et on obtient

$$\begin{aligned} \#\{i \in \mathscr{P} \,|\, X(i) < \overline{X}\} &= \#\{i \in \mathscr{P} \,|\, X(i) < x_{\frac{K+1}{2}}\} = \#\{i \in \mathscr{P} \,|\, X(i) \leq x_{\frac{K-1}{2}}\} \\ &= N_{\frac{K-1}{2}} = \frac{N - n_{\frac{K+1}{2}}}{2} \qquad \text{[d'après (6.7)]} \\ \text{et} \quad \#\{i \in \mathscr{P} \,|\, X(i) > \overline{X}\} &= \#\{i \in \mathscr{P} \,|\, X(i) \geq x_{\frac{K+3}{2}}\} \\ &= N - \#\{i \in \mathscr{P} \,|\, X(i) \leq x_{\frac{K+1}{2}}\} = N - N_{\frac{K+1}{2}} \\ &= N - \left(\frac{N}{2} + \frac{n_{\frac{K+1}{2}}}{2}\right) \qquad \text{[d'après (6.8)]} \\ &= \frac{N - n_{\frac{K+1}{2}}}{2}\,, \end{aligned}$$

ce qui conduit à la même conclusion qu'avec K pair.

Si on considère maintenant le cas d'une distribution asymétrique à gauche, on observera relativement plus souvent de petites modalités de X que de grandes (voir le point 7). Autrement dit, si on choisit deux modalités x_k et x_l de X telles que $x_k < \overline{X} < x_l$, on aura $f_k > f_l$ et on observera plus souvent x_k que x_l, et donc plus souvent la valeur négative $x_k - \overline{X}$ que la valeur positive $x_l - \overline{X}$. Si de plus x_k et x_l sont à égale distance a de $\overline{X}$, on aura $(x_k - \overline{X}) = -a$ et $(x_l - \overline{X}) = a$, d'où

$$f_k(x_k - \overline{X})^3 + f_l(x_l - \overline{X})^3 = f_k(-a)^3 + f_l a^3 = a^3(f_l - f_k) < 0.$$

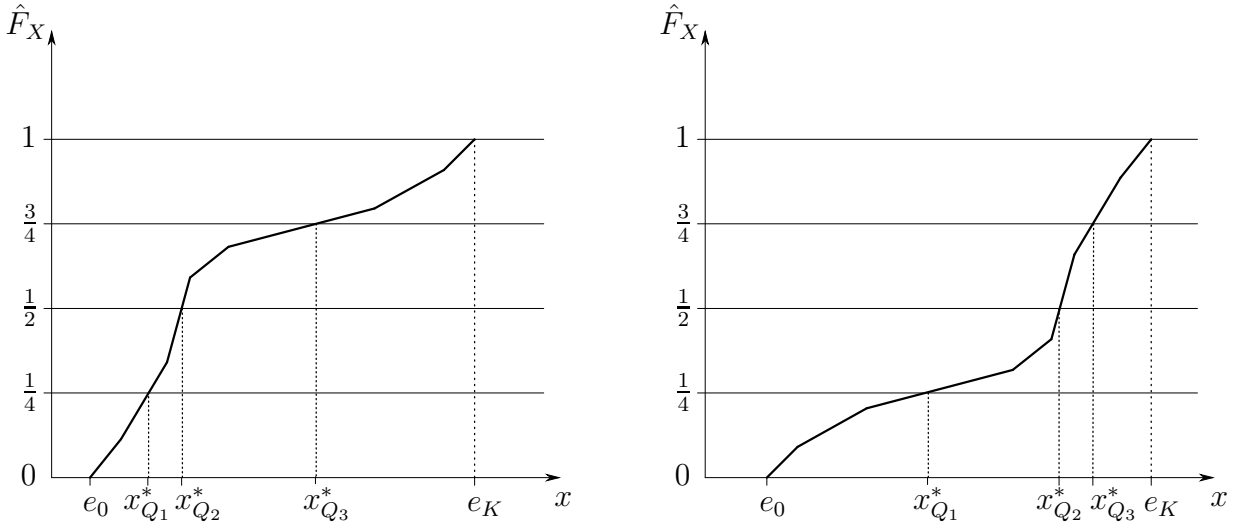
Dans le cas d'une asymétrie à droite, le même raisonnement conduit à une valeur positive de $f_k(x_k - \overline{X})^3 + f_l(x_l - \overline{X})^3$.

13. On peut finalement mentionner une autre conséquence de l'asymétrie. On a mentionné au point 5 de la section 5.2.4.2 que l'histogramme a des propriétés qui sont reliées à celles du polygone des fréquences cumulées. Par conséquent, l'asymétrie a aussi un impact sur la forme du polygone.

    Si la distribution de X est asymétrique à gauche, les densités de fréquences sont relativement plus élevées pour les petites modalités de X. Par conséquent, pour de telles modalités le taux de croissance de la fonction $\hat{F}_X$ plus rapide que pour des modalités élevées (voir le point 4 de la section 5.2.4.2). Le polygone des fréquences cumulées a dans ce cas l'allure représentée par le grahique de gauche de la figure 6.6. En cas d'asymétrie à droite, un raisonnement similaire montre que le polygone des fréquences cumulées a une pente plus élevée pour les fortes modalités de X que pour les petites. Son allure est alors représentée par le graphique de droite de la figure 6.6.

    Ces graphiques permettent d'illustrer une conséquence importante de ce qui vient d'être observé. Les fréquences cumulées $\frac{1}{4}$ et $\frac{1}{2}$ sont atteintes pour des modalités très proches l'une de l'autre en cas d'asymétrie à gauche (voir le graphique de gauche, figure 6.6). D'après le point 4 de la section 6.3.1.2, il s'ensuit que les valeurs $x^*_{Q_1}$ et $x^*_{Q_2}$ sont proches l'une de l'autre, et qu'il en est de même pour les quartiles $Q_1(X)$ et $Q_2(X)$. Ceci s'oppose à ce qui se passe en cas d'asymétrie à droite, où ce sont les fréquences cumulées $\frac{1}{2}$ et $\frac{3}{4}$ qui sont atteintes pour des modalités très proches. Par un raisonnement semblable, on conclut que les quartiles $Q_2(X)$ et $Q_3(X)$ sont rapprochés dans ce cas.

FIG. 6.6 – Allure du polygone des fréquences cumulées en cas d'asymétrie



Les propriétés et remarques qui précèdent suggèrent des principes sur lesquels peuvent être construits des indicateurs d'asymétrie.

6.4.1.2 Mesures d'asymétrie

Les indicateurs d'asymétrie permettent de décrire numériquement les propriétés de symétrie (ou d'asymétrie) d'une distribution. En utilisant les remarques de la section précédente, on peut proposer plusieurs indicateurs d'asymétrie.

6.4.1.2.1 Les coefficients d'asymétrie de Pearson

1. Ces coefficients sont suggérés par la propriété 6.12 et par le point 11 de la section précédente : ces coefficients sont basés sur les différence entre la moyenne et les deux autres indicateurs de tendance centrale.

Définition 6.19 *On appelle coefficients d'asymétrie de Pearson les deux coefficients définis par*

$$P_1(X) = \frac{\overline{X} - \text{Mo}(X)}{\sigma(X)} \qquad et \qquad P_2(X) = \frac{\overline{X} - \text{Me}(X)}{\sigma(X)}.$$

2. À la place de $P_1(X)$ et $P_2(X)$, on utilise parfois $3P_1(X)$ et $3P_2(X)$, respectivement.
3. Comme $\sigma(X) > 0$, les propriétés 6.1, 6.2 (point 2) et 6.6 (point 3) permettent d'interpréter $P_1(X)$ comme la différence entre la moyenne et le mode de la variable $\hat{X} = \frac{X}{\sigma(X)}$. De même, $P_2(X)$ s'interprète comme la différence entre la médiane et la moyenne de $\hat{X}$.
4. D'après la propriété 6.12 et les commentaires faits à la section 6.4.1.1, $P_2(X)$ est nul si la distribution de X est symétrique ; $P_1(X)$ est également nul dans ce cas, à condition que la distribution de X soit aussi unimodale. Observer $P_2(X) < 0$, c'est à dire $\overline{X} < \text{Me}(X)$, indique une asymétrie à droite, et $P_2(X) > 0$ une asymétrie à gauche.
5. On peut montrer que $-1 \leq P_2(X) \leq 1$.

6.4.1.2.2 Le coefficient d'asymétrie γ_1

1. Ce coefficient est construit à partir des résultats du point 12 de la section 6.4.1.1.

Définition 6.20 *Le coefficient d'asymétrie γ_1 est défini par*

$$\gamma_1(X) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{X(i) - \overline{X}}{\sigma(X)} \right)^3.$$

2. Notons qu'en utilisant la même méthode que pour la propriété 6.6 (point 1), on peut écrire

$$\gamma_1(X) = \frac{1}{\sigma(X)^3}\frac{1}{N}\sum_{i=1}^{N}\big(X(i)-\overline{X}\big)^3 = \frac{1}{\sigma(X)^3}\sum_{k=1}^{K} f_k(x_k-\overline{X})^3$$

D'après les remarques du point 12 de la section 6.4.1.1, si X a une distribution symétrique alors la somme est nulle. Lorsqu'elle est négative (positive), cela indique que la distribution est asymétrique à gauche (droite). Le coefficient $\gamma_1(X)$ ayant le même signe que cette somme, on a l'interprétation suivante :
– si $\gamma_1(X)$ est nul, la distribution de X est symétrique ;
– si $\gamma_1(X)$ est négatif (positif), la distribution de X est asymétrique à gauche (à droite).

6.4.1.2.3 Le coefficient d'asymétrie de Yule-Bowley

1. La construction du dernier indicateur repose sur les observations faites au point 13 de la section 6.4.1.1. Cet indicateur est défini comme suit :

Définition 6.21 *On appelle coefficient d'asymétrie de Yule-Bowley*[9] *le nombre noté* $B(X)$ *defini par*

$$B(X) = \frac{\big[Q_3(X)-Q_2(X)\big]-\big[Q_2(X)-Q_1(X)\big]}{Q_3(X)-Q_1(X)},$$

où $Q_1(X)$, $Q_2(X)$ *et* $Q_3(X)$ *désignent les quartiles de* X *(voir la définition 6.11).*

2. Les commentaires du point 13 de la section 6.4.1.1 permettent d'affirmer qu'en cas d'asymétrie à gauche, $Q_3(X)-Q_2(X) \geq Q_2(X)-Q_1(X)$ et donc $B(X) \geq 0$. Ce coefficient sera négatif en cas d'asymétrie à droite.

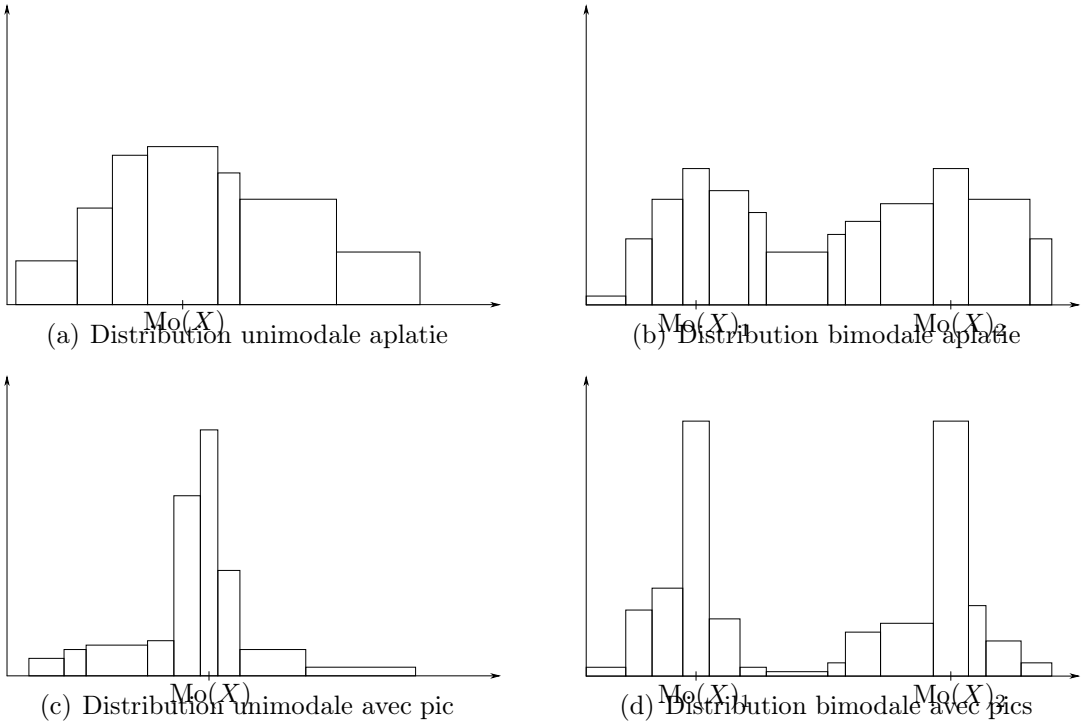
6.4.2 Indicateurs d'aplatissement

6.4.2.1 Définition

Lorsqu'on trace l'histogramme des fréquences d'une variable X, il peut apparaître un ou plusieurs modes. Ceux-ci peuvent être plus ou moins prononcés. Un mode est d'autant plus prononcé que sa densité de fréquence est élevée par rapport à celle des autres modalités. Si un mode est très marqué, on distinguera clairement à cet endroit un pic dans l'histogramme, ce pic étant d'autant plus haut que ce mode est marqué. Au contraire, si un mode est peu marqué, l'histogramme aura une allure plutôt aplatie.

[9] George Udny Yule (1871-1951), mathématicien britannique et Arthur L. Bowley (1869-1957), économiste britannique.

FIG. 6.7 – Illustration des caractéristiques d'aplatissement



(a) Distribution unimodale aplatie (b) Distribution bimodale aplatie

(c) Distribution unimodale avec pic (d) Distribution bimodale avec pics

Les graphiques du haut de la figure 6.7 correspondent à des distributions statistiques ayant des modes peu prononcés et pour lesquelles les histogrammes sont plutôt aplatis autour des modes. La distribution de droite comporte deux modes. Les histogrammes du bas de la figure présentent au contraire un pic très marqué autour des modes, indiquant que ces derniers sont très prononcés.

L'aplatissement de la distribution de X est une caractéristique qui désigne la forme plus ou moins aplatie de l'histogramme des fréquences de X. Les indicateurs d'aplatissement ont pour but de fournir une description numérique de cette caractéristique. Le plus répandu de ces indicateur est défini comme suit.

Définition 6.22 *On appelle indice d'aplatissement (*kurtosis *en anglais) de X le nombre noté $\kappa(X)$ et défini par*

$$\kappa(X) = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(X(i) - \overline{X}\right)^4}{\sigma(X)^4}.$$

Notons immédiatement que pour les mêmes raisons que dans les propriétés 6.8 et 6.9, on

peut écrire

$$\kappa(X) = \frac{\frac{1}{N}\sum_{k=1}^{K} n_k(x_k - \overline{X}\,)^4}{\sigma(X)^4} = \frac{\sum_{k=1}^{K} f_k(x_k - \overline{X}\,)^4}{\sigma(X)^4}\,. \tag{6.11}$$

6.4.2.2 Propriétés et interprétation

Il est très souvent écrit dans les manuels de statistique que ce coefficient est d'autant plus faible que la distribution de X est plate autour du mode. Cette affirmation est fausse en l'état, comme le montre ce qui suit.

On commence par donner une propriété qui établit une borne inférieure pour $\kappa(X)$ et qui permet en même temps de comprendre ce qui peut affecter la valeur de $\kappa(X)$.

Propriété 6.13 *Soit Y la variable définie par $Y = \frac{(X-\overline{X})^2}{\mathrm{V}(X)}$.*

1. *$\overline{Y} = 1$ et $\kappa(X) = \mathrm{V}(Y) - 1$.*
2. *$\kappa(X) \geq 1$.*

Démonstration : D'après la définition de Y, on a $Y(i) = \frac{(X(i)-\overline{X})^2}{\mathrm{V}(X)}$, $i = 1, \ldots, N$. Puisque $\sigma(X)^2 = \mathrm{V}(X)$, ou encore $\sigma(X)^4 = \mathrm{V}(X)^2$, on a

$$Y(i) = \frac{(X(i) - \overline{X}\,)^2}{\sigma(X)^2} \qquad \text{et} \qquad Y(i)^2 = \frac{(X(i) - \overline{X}\,)^4}{\sigma(X)^4}\,.$$

On peut donc écire

$$\kappa(X) = \frac{\frac{1}{N}\sum_{i=1}^{N} \big(X(i) - \overline{X}\,\big)^4}{\sigma(X)^4} = \frac{1}{N}\sum_{i=1}^{N} \frac{\big(X(i) - \overline{X}\,\big)^4}{\sigma(X)^4} = \frac{1}{N}\sum_{i=1}^{N} Y(i)^2, \tag{6.12}$$

où la deuxième égalité s'obtient en distribuant $\frac{1}{\sigma(X)^4}$ sur la somme, et la dernière de la définition de Y. Pour calculer $\overline{Y}$, on applique la formule de la moyenne et on a

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} \frac{\big(X(i) - \overline{X}\,\big)^2}{\mathrm{V}(X)} = \frac{1}{\mathrm{V}(X)}\,\frac{1}{N}\sum_{i=1}^{N} \big(X(i) - \overline{X}\,\big)^2 = \frac{1}{\mathrm{V}(X)}\mathrm{V}(X) = 1, \tag{6.13}$$

où la deuxième égalité est obtenue en factorisant $\frac{1}{\mathrm{V}(X)}$ dans l'expression de $\overline{X}$. On peut alors calculer l'expression de la variance de Y en utilisant la formule de König-Huygens :

$$\mathrm{V}(Y) = \frac{1}{N}\sum_{i=1}^{N} Y(i)^2 - \overline{Y}^2 = \kappa(X) - 1,$$

où la dernière égalité résulte de l'utilisation de (6.12) et (6.13). Le deuxième point de la propriété résulte de cette égalité et du fait que $\mathrm{V}(Y) \geq 0$ (voir le point 3 de la propriété 6.9). □

Cette propriété montre que $\kappa(X)$ sera d'autant plus grand que Y est dispersé autour de sa moyenne, égale à 1. L'indice $\kappa(X)$ atteindra donc sa valeur la plus faible, $\kappa(X) = 1$, si et seulement $V(Y) = 0$, c'est à dire (d'après le point 3 de la propriété 6.9) si et seulement si $Y(1) = \cdots = Y(N) = \overline{Y} = 1$. Notons que

$$Y(i) = 1 \;\Leftrightarrow\; \frac{|X(i) - \overline{X}\,|}{\sigma(X)} = 1 \;\Leftrightarrow\; X(i) - \overline{X} = \pm\sigma(X) \;\Leftrightarrow\; X(i) = \overline{X} \pm \sigma(X).$$

Ceci est la démonstration d'un résultat qu'on peut énoncer par la propriété suivante.

Propriété 6.14 *L'indice d'aplatissement $\kappa(X)$ est d'autant plus grand que les modalités de la variable X sont dispersées autour des valeurs $\overline{X} - \sigma(X)$ et $\overline{X} + \sigma(X)$. L'indice $\kappa(X)$ atteint sa valeur minimale 1 si et seulement si pour tout individu $i = 1, \ldots, N$, on a $X(i) = \overline{X} + \sigma(X)$ ou $X(i) = \overline{X} - \sigma(X)$.*

Cette propriété est fondamentale dans l'interprétation de $\kappa(X)$, puisqu'elle montre que $\kappa(X)$ est un indicateur de la dispersion des modalités $X(1), \ldots, X(N)$ de X autour des deux valeurs $\overline{X} - \sigma(X)$ et $\overline{X} + \sigma(X)$. Ceci conduit à deux remarques :

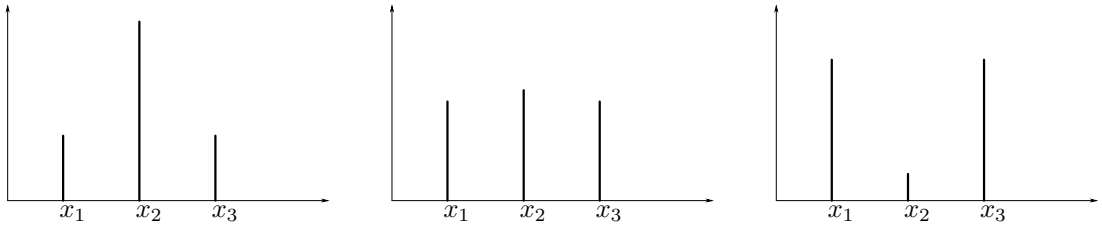
- D'après ce qui précède, $\kappa(X)$ n'apparaît pas directement comme une mesure de l'aplatissement de la distribution de X autour de son mode. Il est en tout cas faux d'affirmer que $\kappa(X)$ est d'autant plus petit que la distribution de X autour de son mode est aplatie.
- Une interprétation et une utilisation correctes de $\kappa(X)$ doivent tenir compte du résultat de la propriété 6.14

Pour illustrer ces deux points, il peut être utile d'effectuer le calcul de la valeur de $\kappa(X)$ dans le cas simple où la variable X n'a que trois modalités équidistantes x_1, x_2 et x_3, pour lesquelles on observe $f_1 = f_3$. L'équidistance signifie que la distance entre x_1 et x_2 est la même que celle qui sépare x_3 de x_2. Notons a cette distance. On doit alors avoir, en utilisant la convention qui consiste à numéroter les modalités par ordre croissant : $x_1 = x_2 - a$ et $x_3 = x_2 + a$. L'égalité $f_1 = f_3$ revient à dire qu'il y a autant d'individus qui présentent la première modalité que d'individus qui présentent la troisième. Comme la somme des fréquences vaut 1, on a

$$1 = f_1 + f_2 + f_3 = 2f_1 + f_2 \;\Leftrightarrow\; f_1 = f_3 = \frac{1 - f_2}{2}\,.$$

L'histogramme des fréquences pour cette variable a une allure qui ressemble à celle de l'un des histogrammes de la figure 6.8.

FIG. 6.8 – Histogrammes d'une variable symétrique avec 3 modalités équidistantes



Pour pouvoir calculer $\kappa(X)$, on commence par calculer la moyenne de X. On a

$$\begin{aligned}\overline{X} &= \sum_{k=1}^{3} f_k x_k = \frac{1-f_2}{2}(x_2-a) + f_2 x_2 + \frac{1-f_2}{2}(x_2+a) \\ &= \frac{1-f_2}{2}(x_2 - a + x_2 + a) + f_2 x_2 \\ &= x_2.\end{aligned}$$

On peut ensuite utiliser ce résultat pour calculer la variance de X :

$$\begin{aligned}\mathrm{V}(X) &= \sum_{k=1}^{3} f_k (x_k - \overline{X})^2 = \frac{1-f_2}{2}(x_1-x_2)^2 + f_2(x_2-x_2)^2 + \frac{1-f_2}{2}(x_3-x_2)^2 \\ &= \frac{1-f_2}{2}a^2 + \frac{1-f_2}{2}a^2 \\ &= (1-f_2)a^2.\end{aligned}$$

Puis on calcule le numérateur de $\kappa(X)$ en utilisant l'expression (6.11) :

$$\begin{aligned}\sum_{k=1}^{3} f_k (x_k - \overline{X})^4 &= \frac{1-f_2}{2}(x_1-x_2)^4 + f_2(x_2-x_2)^4 + \frac{1-f_2}{2}(x_3-x_2)^4 \\ &= \frac{1-f_2}{2}a^4 + \frac{1-f_2}{2}a^4 \\ &= (1-f_2)a^4.\end{aligned}$$

On obtient finalement :

$$\kappa(X) = \frac{(1-f_2)a^4}{\big((1-f_2)a^2\big)^2} = \frac{(1-f_2)a^4}{(1-f_2)^2 a^4} = \frac{1}{(1-f_2)^2}\,. \tag{6.14}$$

On peut maintenant utiliser ces résultats pour commenter les deux remarques faites ci-dessus.

– Si l'affirmation selon laquelle « plus $\kappa(X)$ est petit plus l'histogramme de X est plat » était vraie, alors le calcul de $\kappa(X)$ devrait donner les plus petites valeurs pour

des histogrammes ressemblant à celui du centre de la figure 6.8. Plus précisément, la valeur de $\kappa(X)$ pour l'histogramme totalement plat devrait être la plus petite qu'on puisse obtenir. Or l'histogramme plat correspond à $f_1 = f_2 = f_3 = \frac{1}{3}$. Dans ce cas, l'expression (6.14) permet de caluler que $\kappa(X) = \frac{1}{(2/3)^2} = \frac{9}{4}$. Il est clairement possible d'obtenir des valeurs plus petites pour des distributions dans lesquelles $f_2 < \frac{1}{3}$. En effet, il est facile de vérifier que $\kappa(X)$ tel qu'exprimé par (6.14) croît avec f_2. Par conséquent, plus la valeur de f_2 est petite, plus l'indice d'aplatissement est petit. Autrement dit, les valeurs les plus petites de $\kappa(X)$ ne sont pas obtenues pour des histogrammes plats, mais pour des histogrammes qui ont l'allure de celui de gauche sur la figure 6.8. Ceci illustre le premier point.

On peut aussi noter que l'indice $\kappa(X)$ d'autant plus grand que f_2 s'approche de 1. En ce qui concerne l'allure de l'histogramme des fréquences de X, $\kappa(X)$ est d'autant plus grand que le pic en $\overline{X} = x_2$ est marqué. Sur la figure 6.8, $\kappa(X)$ diminue lorsqu'on se déplace de l'histogramme de gauche vers l'histogramme de droite.

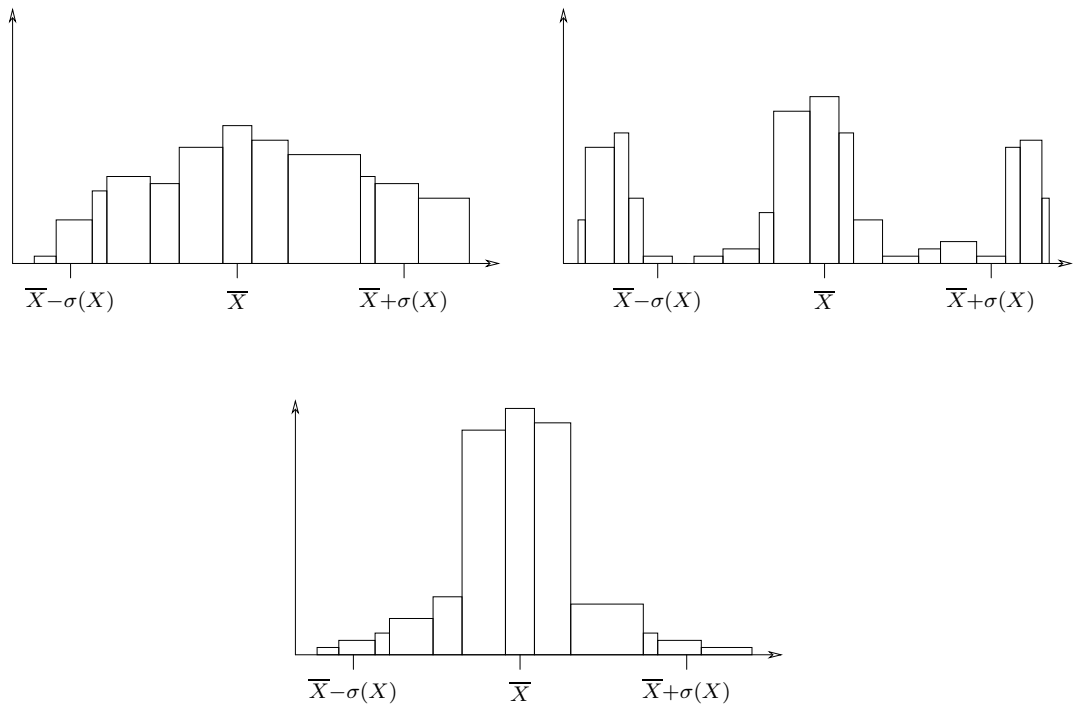
– La propriété 6.14 indique clairement que $\kappa(X)$ mesure la dispersion des modalités de X autour des deux valeurs $\overline{X} - \sigma(X)$ et $\overline{X} + \sigma(X)$. Plus précisément, $\kappa(X)$ est d'autant plus élevé que cette dispersion est grande. Par conséquent, toute interprétation de l'indice d'aplatissement $\kappa(X)$ doit être reliée à cette caractéristique de la distribution de X.

  Une dispersion élevée des modalités de X autour de $\overline{X} \pm \sigma(X)$ peut être obtenue de deux manières. La première se produit lorsque les densités de fréquence de X sont relativement élevées pour des valeurs proches du milieu de l'intervalle $[\overline{X} - \sigma(X); \overline{X} + \sigma(X)]$, c'est à dire proches de $\overline{X}$. Ce cas est illustré par l'histogramme du bas de la figure 6.9. Celui-ci illustre le cas d'une distribution ayant un pic très marqué autour de la moyenne $\overline{X}$.

  La seconde survient lorsque les densités de fréquence sont relativement élevées pour des modalités éloignées de l'intervalle $[\overline{X} - \sigma(X); \overline{X} + \sigma(X)]$. Cette situation est illustrée par les histogrammes du haut de la figure 6.9. On voit qu'elle correspond à deux types de distributions. Le premier est celui des distributions unimodales et pour lequelles les densités de fréquence restent relativement élevées pour des modalités loin du mode (histogramme en haut à gauche). L'autre type désigne les distributions dont les histogrammes possèdent des pics pour des modalités extrêmes de la variable (histogramme en haut à droite). Dans ce cas, les distributions sont multimodales ou possèdent des modes locaux à des modalités éloignées de l'intervalle $[\overline{X} - \sigma(X); \overline{X} + \sigma(X)]$.

On retiendra de cette section que l'indicateur d'aplatissement le plus répandu ne convient pas pour mesurer cette caractéristique. De faibles valeur de l'indice $\kappa(X)$ n'indiquent pas nécessairement une distribution aplatie. Il est donc recommandé de ne pas

FIG. 6.9 – Dispersion de X autour de $\overline{X} \pm \sigma(X)$



utiliser cet indicateur et de s'en tenir à une évaluation graphique (au moyen de l'histogramme) des propriétés d'aplatissement d'une distribution.

6.5 Indicateurs de concentration

`À compléter...[2 octobre 2007]`

Annexe

Les rangs

Définition

On considère une population $\mathscr{P}$ de N individus et une variable ordinale ou numérique X. L'ensemble des modalités de X est $\mathscr{M}_X = \{x_1, \ldots, x_K\}$.

Définition 6.23 *On appelle le rang (croissant) de l'individu i et on note $\tilde{R}(i)$ le nombre défini par*

$$\tilde{R}(i) = \#\{j \in \mathscr{P} \,|\, X(j) \leq X(i)\}$$

$\tilde{R}(i)$ est donc le nombre d'individus de la population ayant une modalité de X qui n'est pas plus grande que celle de i. Par conséquent, $\tilde{R}(i) < \tilde{R}(j) \Leftrightarrow X(i) < X(j)$ et $X(i) = X(j) \Leftrightarrow \tilde{R}(i) = \tilde{R}(j)$. Cette dernière équivalence montre qu'il peut y avoir dans la population des individus *ex æquo* en termes de rang.

On note $r_1, r_2, \ldots, r_M$ les rangs *distincts* calculés. Dans cette notation, on convient que $r_1 < r_2 < \cdots < r_M$, avec $M \leq N$. On construit alors les ensembles d'individus ayant un même rang donné

$$\mathscr{R}_m = \{i \in \mathscr{P} \,|\, \tilde{R}(i) = r_m\}, \quad m = 1, \ldots, M$$

et on note $d_m = \#\mathscr{R}_m$ le cardinal de cet ensemble. Pour effectuer un départage des *ex æquo*, on procède de la manière suivante :

1. on affecte un et un seul rang $R(i)$ dans $\{1, \ldots, d_1\}$ à chaque individu i de $\mathscr{R}_1$:

$$\forall i \in \mathscr{R}_1, R(i) \in \{1, \ldots, d_1\} \ ;$$
$$\forall i, j \in \mathscr{R}_1, i \neq j \Rightarrow R(i) \neq R(j) \ ;$$

2. pour $m = 2, \ldots, M$ on affecte un et un seul rang $R(i)$ dans $\{\sum_{\ell=1}^{m-1} d_\ell + 1, \ldots, \sum_{\ell=1}^{m-1} d_\ell +$

$d_m\}$ à chaque individu i de $\mathscr{R}_m$:

$$\forall i \in \mathscr{R}_m, R(i) \in \{\sum_{\ell=1}^{m-1} d_\ell + 1, \ldots, \sum_{\ell=1}^{m-1} d_\ell + d_m\} ;$$
$$\forall i, j \in \mathscr{R}_m, i \neq j \Rightarrow R(i) \neq R(j).$$

Cette méthode revient à départager arbitrairement les individus de même rang $\tilde{R}$ de sorte qu'une fois le départage effectué, les nouveaux rangs R satisfassent les conditions suivantes :

1. un individu i a un (et un seul) rang $R(i)$ dans $\{1, \ldots, N\}$;
2. pour tout entier n de $\{1, \ldots, N\}$ on peut trouver un individu i tel que $R(i) = n$;
3. $R(i) \leq R(j) \Rightarrow X(i) \leq X(j)$.

Exemple

Considérons une population de $N = 12$ individus pour lesquels les modalités sont données par le tableau suivant :

i	1	2	3	4	5	6	7	8	9	10	11	12
$X(i)$	7	5	7	8	3	6	7	4	5	9	0	1

Pour $i = 1$, on a $X(i) = 7$ et $\{j \in \mathscr{P} \,|\, X(j) \leq X(1)\} = \{1, 2, 3, 5, 6, 7, 8, 9, 11, 12\}$ d'où $\tilde{R}(1) = \#\{j \in \mathscr{P} \,|\, X(j) \leq X(1)\} = 10$. On détermine de la même manière $\tilde{R}(i)$ pour $i = 2, \ldots, 12$ et on a

i	1	2	3	4	5	6	7	8	9	10	11	12
$X(i)$	7	5	7	8	3	6	7	4	5	9	0	1
$\tilde{R}(i)$	10	6	10	11	3	7	10	4	6	12	1	2

On constate alors que les rangs distincts qu'on a obtenu sont 1, 2, 3, 4, 6, 7, 10, 11, 12, d'où $M = 9$. On posera alors

$$r_1 = 1, \quad r_2 = 2, \quad r_3 = 3,$$
$$r_4 = 4, \quad r_5 = 6, \quad r_6 = 7,$$
$$r_7 = 10, \quad r_8 = 11, \quad r_9 = 12.$$

Pour départager les *ex æquo*, on détermine alors les ensembles $\mathscr{R}_m$, $m = 1, \ldots, 9$. Pour le premier on a

$$\mathscr{R}_1 = \{i \in \mathscr{P} \,|\, \tilde{R}(i) = 1\} = \{11\} \quad \text{et} \quad d_1 = 1.$$

On a de la même manière

$$\begin{array}{ll} \mathscr{R}_2 = \{12\} \text{ et } d_2 = 1 & \mathscr{R}_3 = \{5\} \text{ et } d_3 = 1 \\ \mathscr{R}_4 = \{8\} \text{ et } d_4 = 1 & \mathscr{R}_5 = \{2,9\} \text{ et } d_5 = 2 \\ \mathscr{R}_6 = \{6\} \text{ et } d_6 = 1 & \mathscr{R}_7 = \{1,3,7\} \text{ et } d_7 = 3 \\ \mathscr{R}_8 = \{4\} \text{ et } d_8 = 1 & \mathscr{R}_9 = \{10\} \text{ et } d_9 = 1 \end{array}$$

On peut donc construire les rangs $R(i)$, $i = 1, \ldots, 12$, permettant de départager les *ex æquo*. En suivant la méthode décrite précédemment, on a

$$\begin{array}{lll} R(11) = d_1 = 1, & R(12) = d_1 + 1 = 2, & R(5) = d_1 + d_2 + 1 = 3, \\ R(8) = \sum_{\ell=1}^{3} d_\ell + 1 = 4, & R(2) = \sum_{\ell=1}^{4} d_\ell + 1 = 5, & R(9) = \sum_{\ell=1}^{4} d_\ell + d_5 = 6, \\ R(6) = \sum_{\ell=1}^{5} d_\ell + 1 = 7, & R(1) = \sum_{\ell=1}^{5} d_\ell + 1 = 8, & R(3) = \sum_{\ell=1}^{5} d_\ell + 2 = 9, \\ R(7) = \sum_{\ell=1}^{5} d_\ell + d_6 = 10, & R(4) = \sum_{\ell=1}^{6} d_\ell + 1 = 11, & R(10) = \sum_{\ell=1}^{7} d_\ell + 1 = 12. \end{array}$$

On peut alors présenter le classement dans un tableau :

i	1	2	3	4	5	6	7	8	9	10	11	12
$X(i)$	7	5	7	8	3	6	7	4	5	9	0	1
$\tilde{R}(i)$	10	6	10	11	3	7	10	4	6	12	1	2
$R(i)$	8	5	9	11	3	7	10	4	6	12	1	2

Propriétés

i et j désignent des individus quelconques de $\mathscr{P}$. On a les propriétés suivantes.

1. Pour tout $m = 1, \ldots, M$, $r_m = \sum_{\ell=1}^{m} d_\ell$;
2. $\tilde{R}(i) \geq R(i)$;
3. $R(i) \leq R(j) \Rightarrow X(i) \leq X(j)$;
4. $X(i) < X(j) \Rightarrow R(i) < R(j)$;
5. $\#\{i \in \mathscr{P} \,|\, R(i) < R(j)\} = R(j) - 1$;

Démonstration. On rappelle d'abord que $\tilde{R}(i) < \tilde{R}(j) \Leftrightarrow X(i) < X(j)$ et $X(i) = X(j) \Leftrightarrow \tilde{R}(i) = \tilde{R}(j)$.

1. Soit $m \in \{1, \ldots, M\}$ et un individu $i \in \mathscr{P}$ tel que $\tilde{R}(i) = r_m$. Par définition de $\tilde{R}(i)$, cette égalité s'écrit $r_m = \#\{j \in \mathscr{P} \,|\, X(j) \leq X(i)\}$. Or $\{j \in \mathscr{P} \,|\, X(j) \leq X(i)\} = \{j \in \mathscr{P} \,|\, X(j) = X(i)\} \cup \{j \in \mathscr{P} \,|\, X(j) < X(i)\}$. Dans cette union, les deux ensembles sont évidemment disjoints. Par conséquent

   $$\#\{j \in \mathscr{P} \,|\, X(j) \leq X(i)\} = \#\{j \in \mathscr{P} \,|\, X(j) = X(i)\} + \#\{j \in \mathscr{P} \,|\, X(j) < X(i)\},$$

   ou encore

   $$r_m = \#\{j \in \mathscr{P} \,|\, X(j) = X(i)\} + \#\{j \in \mathscr{P} \,|\, X(j) < X(i)\}. \tag{6.15}$$

Le premier ensemble dans l'union est celui de tous les individus ayant la même modalité que i, et donc (d'après ce qu'on vient de rappeler ci-dessus) le même rang que i. Comme $\tilde{R}(i) = r_m$, cet ensemble s'écrit $\{j \in \mathscr{P} \,|\, \tilde{R}(j) = r_m\}$, qui est aussi l'ensemble $\mathscr{R}_m$. Son cardinal est d_m. Donc l'égalité (6.15) s'écrit

$$r_m = d_m + \#\{j \in \mathscr{P} \,|\, X(j) < X(i)\}. \tag{6.16}$$

Le second ensemble dans l'union est aussi (encore d'après le rappel ci-dessus) l'ensemble de tous les individus j de $\mathscr{P}$ ayant un rang $\tilde{R}(j)$ strictement inférieur à celui de i. C'est donc l'ensemble des individus dont le rang $\tilde{R}(j)$ est r_1 ou ... ou r_{m-1}. On peut donc écrire

$$\begin{aligned}\{j \in \mathscr{P} \,|\, X(j) < X(i)\} &= \{j \in \mathscr{P} \,|\, \tilde{R}(j) = r_1\} \cup \cdots \cup \{j \in \mathscr{P} \,|\, \tilde{R}(j) = r_{m-1}\} \\ &= \mathscr{R}_1 \cup \cdots \cup \mathscr{R}_{m-1}.\end{aligned}$$

D'après les rappels faits au début de cette démonstration, les ensembles $\mathscr{R}_1, \ldots, \mathscr{R}_{m-1}$ sont disjoints. Par conséquent $\#\{j \in \mathscr{P} \,|\, X(j) < X(i)\} = \sum_{\ell=1}^{m-1} \#\mathscr{R}_\ell = \sum_{\ell=1}^{m-1} d_\ell$. On peut donc finalement réécrire l'égalité (6.16) :

$$r_m = d_m + \sum_{\ell=1}^{m-1} d_\ell = \sum_{\ell=1}^{m} d_\ell.$$

2. C'est une conséquence immédiate du point précédent. En effet, pour un individu quelconque i, soit m tel que $\tilde{R}(i) = r_m$. Le point précédent donne $\tilde{R}(i) = \sum_{\ell=1}^{m} d_\ell$. Or $\tilde{R}(i) = r_m \;\Leftrightarrow\; i \in \mathscr{R}_m$, ce qui implique $R(i) \in \{\sum_{\ell=1}^{m-1} d_\ell + 1, \ldots, \sum_{\ell=1}^{m} d_\ell\} = \{\sum_{\ell=1}^{m-1} d_\ell + 1, \ldots, \tilde{R}(i)\}$. D'où $R(i) \leq \tilde{R}(i)$.
3. Soient i et j deux individus tels que $R(i) \leq R(j)$. Soient m et n tels que $\tilde{R}(i) = r_m$ et $\tilde{R}(j) = r_n$. On a alors $i \in \mathscr{R}_m$ et $j \in \mathscr{R}_n$. D'après la méthode de construction de $R(i)$ et de $R(j)$ on doit avoir $R(i) \leq \sum_{\ell=1}^{m} d_\ell$ et $R(j) \geq \sum_{\ell=1}^{n-1} d_\ell + 1$. Pour que l'inégalité $R(i) \leq R(j)$ soit toujours possible, il faut que $\sum_{\ell=1}^{m} d_\ell \leq \sum_{\ell=1}^{n-1} d_\ell + 1$. Comme $\forall \ell = 1, \ldots, M$, $d_\ell \geq 1$, voit aisément que cette dernière inégalité n'est possible que si $n \geq m$. Or ceci équivaut à $r_n \geq r_m$, ou encore $\tilde{R}(j) \geq \tilde{R}(i)$, ou encore $X(j) \geq X(i)$.
4. Soient i et j deux individus tels que $X(i) < X(j)$. Soient m et n tels que $\tilde{R}(i) = r_m$ et $\tilde{R}(j) = r_n$. On a $X(i) < X(j) \;\Leftrightarrow\; \tilde{R}(i) < \tilde{R}(j) \;\Leftrightarrow\; r_m < r_n \;\Leftrightarrow\; m < n$. Comme $R(i) \in \{\sum_{\ell=1}^{m-1} d_\ell + 1, \ldots, \sum_{\ell=1}^{m} d_\ell\}$ et $R(j) \in \{\sum_{\ell=1}^{n-1} d_\ell + 1, \ldots, \sum_{\ell=1}^{n} d_\ell\}$, $m < n$ implique $R(i) < R(j)$.
5. C'est évident ! ... Mais ça sert ailleurs.