Faculté des Sciences
Département de Mathématique

Statistique descriptive

Année Académique
2020-2021

Section:
Bachelier en Sciences Informatiques
Titulaire:
Arnout Van Messem
Assistant:
Carole Baum

Introduction

Comment échapper aux "statistiques"? A la radio ou à la télévision, dans les journaux, abondent sondages et référendums, tableaux et graphiques. Même si c'est dans l'information que les effets sont les plus visibles, d'autres aspects de notre vie de tous les jours sont influencés par les statistiques. Dans la société de consommation dans laquelle nous vivons, la plus grande partie des marchandises et services qui nous sont proposés ont des coûts fixés après étude statistique.

Exemple 1 *Les compagnies d'assurance basent les tarifs de leurs différents contrats d'assurance vie sur les tables de mortalité établies par les statisticiens. Ces tables de mortalité évoluent au cours du temps puisque, notamment, l'espérance de vie à la naissance des hommes et des femmes augmente au cours du temps.*

Mais, "Les chiffres disent toujours ce que veut l'homme habile qui sait en jouer" (Le baron Macaulay). Il est donc primordial pour chacun d'être suffisamment formé pour être un lecteur critique, un consommateur avisé et un citoyen responsable.

La Fédération Wallonie-Bruxelles a tenu compte de la nécéssité de faire évoluer les connaissances en statistique des élèves de l'enseignement secondaire en Belgique francophone. Depuis le début des années 2000, une place plus importante est réservée aux concepts essentiels de la statistique. Dans la plupart des filières d'études, cette formation initiale en statistique est ré-activée et étoffée dans le cadre du bachelier.

Notice historique

A l'origine, la statistique était la comptabilité de l'état, et c'est de là qu'elle tire son nom. Le mot statistique, utilisé pour la première fois en 1672, dérive étymologiquement du mot latin *status* (état). On peut dire qu'il y a eu des statistiques dès que les états ont commencé à estimer nécessaire de rassembler des données sur la composition de leur population, les impôts payés par les citoyens, les besoins de l'armée, les productions agricoles, les effets d'une maladie contagieuse,... Un recensement célèbre est celui qu'Auguste ordonna déjà à l'époque de la naissance du Christ. Les premières études de données statistiques se trouvent

surtout en Angleterre. De John Graunt (1620–1674) à Thomas Robert Malthus (1766–1834), l'attention se portait principalement sur les problèmes démographiques et la description de divers aspects d'une population (mortalité, fécondité, proportions des deux sexes,...). On trouve la première exploitation systématique, rationnelle et orientée vers des conclusions, de données scientifiques dans l'ouvrage "Sur l'homme et le développement de ses facultés. Essai d'une physique sociale" (1835) de notre compatriote Adolphe Quetelet (1796–1874). D'après ses recherches anthropologiques, les grandeurs biologiques suivent toutes ce que l'on appelle une distribution normale. Quetelet contribua de façon importante à la fondation de la *Royal Statistical Society* de Londres (1835) et de l'*American Statistical Society* (1839) et parvint à faire se tenir à Bruxelles le premier congrès international de statistique. Le but principal de ce congrès, et de nombreux autres congrès postérieurs, fut la définition de méthodes tendant à rendre comparables internationalement les résultats statistiques des divers pays participants. Quetelet est appelé le fondateur de la statistique moderne.

Depuis le 17ème siècle déjà, le calcul des probabilités était appliqué à l'analyse des données statistiques et à la formulation d'hypothèses statistiques, entre autres par les Anglais Halley et de Moivre, les Suisses Bernoulli et Euler, les Français Laplace et Poisson.

La statistique, dite mathématique, issue de ces travaux, joue un rôle important d'auxiliaire dans presque toutes les branches de l'activité scientifique.

Définition

La statistique est la science du dénombrement. Elle mesure et analyse des phénomènes qui se produisent un grand nombre de fois.

Exemple 2 *Les statistiques nationales font état de 44319 mariages célébrés en 2017 en Belgique, 23059 divorces ayant été enregistrés la même année.(Source: Office belge de statistique, site web* `http://statbel.fgov.be/`*).*

Exemple 3 *"La Côte a connu un été exceptionnel : pas moins de 6.8 millions de touristes d'un jour s'y sont déplacés lors des mois de juillet et août." (Source: RTBF-Info du 31 aooût 2018).*

Nous adopterons cependant la définition plus complète suivante:

La statistique est la science qui rassemble, organise, résume et analyse des données et qui permet d'interpréter les résultats et de tirer des conclusions afin d'aider à la prise de décisions.

Contenu du cours de statistique descriptive

Ces notes de cours sont consacrées à la statistique descriptive dont le but essentiel est de présenter l'information disponible d'une façon compréhensible et condensée. Comme précisé ci-dessus, les cours de mathématiques de l'enseignement secondaire introduisent déjà de nombreux concepts et techniques exploités en statistique descriptive (et en probabilité). Ceux-ci sont revus dans le cadre de ce cours tout en adoptant une démarche à la fois plus théorique (les propriétés mathématiques des différentes approches sont développées et démontrées) mais également plus pratique (en parallèle au cours théorique, l'apprentissage du logiciel gratuit `R` est prévu).

De très nombreux ouvrages présentent cette matière. Ce sont surtout les livres de Dehon, Droesbeke et Vermandele (2008), Bragard et Alexandre (1995) et Dodge (1999) qui sont utilisés comme référence.

Notons enfin que les notices historiques proviennent principalement du dictionnaire encyclopédique rédigé par Y. Dodge (2007).

Je tiens à remercier G. Haesbroeck pour la crèation de ces notes de cours.

Table des matières

Chapitre 1

Notions de base

La statistique repose sur les concepts de base suivants: la population, les variables, les observations et les données. Ce chapitre, inspiré de Dodge (1999), développe leurs définitions. Par ailleurs, les analyses statistiques (notamment celles développées dans les médias) reposent souvent sur le calcul de pourcentages (ou taux ou proportions/probabilités). Les définitions de ces notions de base ainsi que du pourcentage de variation (concept très fréquemment utilisé dans la vie de tous les jours) sont rappelées à la fin de ce chapitre.

1.1 La population

La population est l'ensemble de toutes les unités (personnes, produits, pays,...) d'intérêt pour l'étude statistique menée.

Exemple 4 *Dans une étude sur l'emploi, la population est l'ensemble des personnes en âge de travailler. Dans un sondage d'opinion pour les élections communales, la population est constituée de l'ensemble des électeurs d'une commune. Dans l'étude de l'efficacité d'une initiative pédagogique, la population est l'ensemble des étudiants concernés par le cours. Dans une étude sur la durée de vie d'une ampoule d'un type donné, la population est l'ensemble des ampoules de ce type sortant de la chaîne de production.*

Les éléments d'une population statistique n'étant pas nécessairement des êtres humains mais pouvant être des choses, des événements,..., il faut généraliser l'acception habituelle du terme population.

Les éléments de la population sont appelés *individus* ou *unités statistiques*. Le nombre total d'individus de la population est appelé *effectif* de la population.

1.2 Les variables

Les éléments d'une population possèdent la caractéristique commune d'appartenir à la même population mais ils peuvent différer selon d'autres critères. Par exemple, les étudiants de premier bachelier en sciences informatiques ont choisi les mêmes études mais pratiquent des sports différents, habitent dans des communes différentes, passent, en moyenne sur une semaine, un temps variable sur les réseaux sociaux... En statistique, ces caractéristiques sont appelées des *caractères* ou des *variables*. Elles servent à décrire la population en question. Elles sont directement mesurées sur chaque unité de la population.

Une variable est désignée par un nom et est souvent notée par une lettre majuscule (par exemple X). Ses valeurs possibles sont appelées *modalités*, *catégories* ou simplement *valeurs*.

Exemple 5 *La variable précisant les préférences politiques des habitants d'une commune peut être appelée simplement* `Parti` *et a, par exemple, comme modalités les catégories* `CdH`*,* `Ecolo`*,* `MR`*,* `PS` *et* `Autre`*. La variable qui décrit la taille des individus de la population peut s'intituler naturellement* `Taille` *et prend les valeurs 1,7 m, 1,84 m...*

Les différentes valeurs prises par les individus pour les variables permettent de les classer dans certains sous-ensembles. Le classement ne pourra se faire sans ambiguité que si les modalités des variables sont *mutuellement exclusives* (un individu ne peut appartenir à deux catégories à la fois) et *exhaustives* (tout individu se trouve dans au moins une catégorie). Nous verrons dans la suite que les classes construites sur les valeurs prises par les variables statistiques pourront être ordonnées sur des échelles de qualité différente selon le *type* de la variable. En effet, on distingue habituellement deux types de variables:

- Les variables *quantitatives* qui peuvent être mesurées ou énumérées;
- Les variables *qualitatives* qui ne peuvent être ni mesurées ni dénombrées mais seulement constatées.

Ces deux types de variables, ainsi que les échelles permettant de comparer leurs modalités, sont décrits ci-dessous.

1.2.1 Variables qualitatives

Une variable qualitative est une variable dont les modalités ne sont pas les résultats d'une mesure, et sont dès lors appelées *catégories* ou *attributs*. Par exemple, la variable `Genre` est qualitative, de même que la variable `Parti`. Une variable qualitative qui ne possède que deux catégories est *dichotomique*. La variable `Genre` est une telle variable ainsi que la variable `Fumeur` dont les catégories sont `oui` ou `non`.

Les modalités d'une variable qualitative peuvent être classées à partir d'une *échelle nominale* ou d'une *échelle ordinale*:

- Echelle nominale: lorsque les catégories d'une variable ne sont pas naturellement ordonnées, cette variable est définie sur une échelle nominale. C'est le cas des variables `Genre` et `Parti`.
- Echelle ordinale: lorsque les catégories peuvent être ordonnées, la variable est définie sur une échelle ordinale. Habituellement, cet ordre ne permet pas de déterminer la magnitude des différences entre les groupes. Ce type d'échelle est particulièrement utilisé lorsqu'il s'agit d'évaluer une situation, une performance, une satisfaction. Par exemple, une préparation à un examen peut être jugée `excellente`, `bonne`, `suffisante`, `insuffisante` ou `mauvaise`.

1.2.2 Variables quantitatives

Une variable quantitative est une variable dont les modalités ont des valeurs numériques. Citons, par exemple, l'âge, le revenu, la taille, le nombre d'enfants dans une famille,... Les modalités représentent l'ensemble des valeurs possibles de la variable.

Exemple 6 *La variable donnant le nombre d'enfants d'une famille pourrait avoir comme modalités les nombres entiers 0,1,2,3,... tandis que la variable* `Revenu` *peut prendre n'importe quelle valeur réelle positive ou nulle.*

Les variables quantitatives dépendent de l'unité dans laquelle elles sont exprimées. Par exemple, la taille peut être mesurée en mètres ou en centimètres, le revenu en dollars ou en euros. L'unité choisie dépend souvent de la précision de l'appareil de mesure utilisé. Lorsque la précision est grande, les valeurs ou modalités d'une variable quantitative peuvent être très nombreuses. Souvent, de telles mesures donnent lieu à des groupements en classes.

Exemple 7 *Les statistiques financières donnent le revenu net imposable de tous les déclarants du Royaume. Les statistiques officielles publiées par le SPF Economie, PME, Classes Moyennes et Energie regroupent les revenus en classes en précisant le nombre de contribuables se trouvant dans chaque tranche de revenus et en indiquant la masse totale de leurs revenus.*

Une distinction importante parmi les variables quantitatives concerne leur caractère discret ou continu:

- Variable *discrète*: une variable quantitative est discrète si l'ensemble de ses valeurs possibles est dénombrable. Typiquement, les variables discrètes s'obtiennent par un procédé de comptage.

Exemple 8 *Le nombre d'enfants d'une famille est 0, 1, 2, 3,..., mais ne peut pas être 2,5 ou 3,7.*

- Variable *continue*: une telle variable peut prendre n'importe quelle valeur dans un intervalle qui lui est propre. Typiquement, les variables continues s'obtiennent par une mesure.

  Exemple 9 *Le revenu d'un contribuable belge peut être n'importe quelle valeur positive; le poids d'un nouveau né peut varier de 1 kg à 5 kg, le poids pouvant être 3,3 kg, 3,32 kg ou 3,321 kg selon la précision de la balance.*

De plus, les variables quantitatives, continues ou discrètes, sont principalement mesurées sur des échelles *d'intervalle* ou *de rapport*.

- Echelle d'intervalle: elle inclut toutes les caractéristiques de l'échelle ordinale mais de plus, elle permet de tenir compte de la différence entre deux valeurs d'une variable. Par contre, les rapports entre les valeurs d'une telle échelle n'ont pas de sens. En outre, la valeur zéro comme donnée ne signifie pas l'absence de la caractéristique étudiée. Autrement dit, le zéro est un zéro arbitraire.

  Exemple 10 *Dans le cadre d'un recensement de la population, on demande l'année de naissance. Avec une telle variable, il n'est pas possible d'établir de rapports entre les données. Cependant, l'intervalle de 20 ans entre les dates de naissance de deux personnes signifie qu'ils ont 20 ans de différence. De plus, l'année zéro fait référence à la naissance de Jésus-Christ mais cette année zéro aurait pu être fixée à n'importe quel autre moment. Notons de plus qu'être né lors de l'année zéro n'implique pas l'absence de date de naissance.*

- Echelle de rapport: il s'agit de l'échelle la plus riche en propriétés. Elle possède un zéro naturel qui indique l'absence du phénomène étudié. Différences et rapports entre valeurs y ont un sens (sauf quand on divise par la donnée zéro).

  Exemple 11 *Les variables* `Revenu`, `Taille`, `Nombre d'enfants`,... *utilisent une telle échelle pour exprimer leurs valeurs. Avoir un revenu égal à zéro Euro signifie qu'on n'a pas de revenu. De plus, une personne peut posséder un revenu deux fois supérieur à celui de son voisin.*

Les variables quantitatives peuvent aussi s'exprimer sur les échelles plus pauvres présentées pour les variables qualitatives, mais cela entraîne une perte d'informations.

1.3 Les observations et les données

Les résultats observés d'une ou plusieurs variables sur une population constituent les observations. Celles-ci étant propres à chacun des individus de la population, elles ont des valeurs fixes, ce qui n'est pas le cas de la variable qui change d'un élément à l'autre.

Dans certaines études, des contraintes de temps ou budgétaires ou encore des impossibilités matérielles ne permettent pas d'observer chaque individu de la population. Une partie des unités statistiques est alors sélectionnée pour constituer un *échantillon* auquel l'analyse statistique est appliquée. Lorsque l'échantillon respecte certaines propriétés, les résultats obtenus à partir de ses éléments peuvent être élargis à la population complète. La crédibilité à accorder à ces résultats dépend de l'accord plus ou moins étroit des éléments de l'échantillon avec les éléments de la population. On parle alors du problème de la *représentativité* de l'échantillon. Les notions d'échantillon et de sa représentativité par rapport à la population totale seront appronfondies dans des cours ultérieurs traitant de statistique inférentielle.

Si une étude statistique porte sur p variables et n individus, l'ensemble des observations récoltées peut se présenter sous la forme du Tableau 1.1 à n lignes et p colonnes (sans compter celle contenant les indices $1, \ldots, n$) appelé *tableau individus* $\times$ *caractères*.

Tableau 1.1: Tableau individus $\times$ caractères

Individus	Variables				
	1	...	j	...	p
1	x_{11}	...	x_{1j}	...	x_{1p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_{i1}	...	x_{ij}	...	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	...	x_{nj}	...	x_{np}

Une colonne correspond à un caractère étudié (ou à une variable) tandis que chaque ligne décrit un individu de la population (ou de l'échantillon). L'élément x_{ij} à l'intersection de la $i$ème ligne et de la $j$ème colonne est la valeur observée de la $j$ème variable sur le $i$ème individu. Cette valeur x_{ij} est soit un nombre (la $j$ème variable est quantitative), soit une expression (la $j$ème variable est qualitative). Pour faciliter le traitement informatique des grands ensembles de données par des logiciels statistiques, les catégories des variables qualitatives sont souvent codées par les chiffres 1,2,3,... Evidemment, cela doit rester clair, dans l'esprit de l'utilisateur des données, que ces valeurs numériques associées aux variables qualitatives ne sont que des codes et ne peuvent donc pas être manipulées algébriquement comme les valeurs des variables quantitatives. Par ailleurs, notons que lorsqu'une seule

variable est considérée ($p = 1$), on parle de série statistique *univariée* et on utilise souvent une notation simplifiée pour la décrire. Si la variable est notée X, on représente la valeur prise par la variable pour l'individu i par x_i ($i = 1, \ldots, n$) et on note cette série $S = \{x_1, x_2, \ldots, x_n\}$.

Dans les chapitres suivants, les méthodes statistiques présentées seront illustrées à partir de données mesurées sur l'ensemble des communes wallonnes. Ces données ont été collectées à partir des sites `www.lavenir.net` et `http://statbel.fgov.be/fr/statistiques` (site du SPF Economie). En particulier, les variables suivantes ont été enregistrées:

- `PartiBourgmestre`: variable qualitative nominale précisant le parti du Bourgmestre de la commune (période 2013-2018), modalités `CdH`, `Ecolo`, `MR`, `PS` et `Autre`.
- `NombrePartis`: variable quantitative discrète précisant le nombre de partis se trouvant dans la majorité communale.
- `Chomage`: variable quantitative continue (exprimée en %) donnant le taux de chômage dans la communes (en 2010), ce taux correspondant à la proportion des personnes sans travail et disponibles sur le marché de l'emploi, par rapport au total de la population active.
- `PrixMaison`: variable quantitative coninue précisant le prix moyen des maisons vendues sur le sol de la commune lors de l'année 2012.
- `IndiceRichesse`: indice (variable quantitative continue, discrétisée à l'unité) caractérisant la richesse d'une commune par rapport à une valeur de référence égale à 100 (et correspondant à la richesse moyenne en Belgique).

Par ailleurs, la variable qualitative nominale `Provinces` distingue les communes des cinq provinces wallonnes. Notons que le statisticien travaille habituellement *en aveugle*, les individus de la base de données étant de simples numéros. Dans ce cas-ci cependant, les noms des communes sont disponibles.

Une deuxième base de données, basée sur les déclarations des revenus des contribuables belges en 2002, est également illustrée dans les notes.

1.4 Pourcentages, taux, proportions et pourcentages de variation

De nombreuses informations données dans le presse le sont sous la forme de pourcentages. Le mathématicien travaillera plus volontiers avec une proportion tandis que l'économiste utilisera assez souvent des taux. Il est utile de se rendre compte que ces concepts diffèrent uniquement par le choix d'une base. Par ailleurs, lorsque l'on souhaite quantifier l'évolution d'une

quantité entre deux situations ou deux périodes (augmentation d'une part de marché, diminution d'un prix...), il convient de le faire en termes relatifs et non absolus. Les définitions utiles pour manipuler convenablement toutes ces notions sont listées ci-dessous.

- La *proportion* indique quelle partie de la population correspond à une des catégories de la variable étudiée. Une proportion est un nombre entre 0 et 1 et s'obtient en divisant le nombre d'unités de la population possédant la caractéristique voulue par l'effectif total de la population.

- Le *pourcentage* indique, sur une base de 100, quelle partie de la population correspond à une des catégories de la variable étudiée. Il s'obtient en multipliant la proportion par 100.

- Le *taux* indique, sur une base de 1, 10, 100, 1000,... quelle partie de la population correspond à une des catégories de la variable étudiée. Il est obtenu en divisant le nombre d'unités possédant la caractéristique voulue par l'effectif total et en multipliant le résultat par la base 1, 10, 100,... Lorsque la base est égale à 100, le taux coïncide avec le pourcentage et se note avec le signe %. Le choix de la base dépend soit d'une convention, soit de la fréquence d'occurrence de la caractéristique.

- Un *pourcentage de variation* dans le temps mesure le pourcentage d'augmentation ou de diminution qu'une variable a subi dans le temps. Il se calcule de la manière suivante:

$$\% \text{ de variation} = \frac{\text{valeur finale} - \text{valeur initiale}}{\text{valeur initiale}} \times 100, \tag{1.1}$$

  où `valeur initiale` et `valeur finale` représentent respectivement les valeurs (ou les pourcentages d'une modalité) de la variable à l'instant initial et à l'instant final. Si le pourcentage de variation est positif (resp. négatif), cela signifie qu'il y a eu une augmentation (resp. diminution) de la valeur entre les deux périodes.

1.5 Exercices

1. Trouver des variables (au moins 3) susceptibles de caractériser les concepts suivants:

   (a) la santé économique d'une PME.

   (b) la performance d'un ouvrier.

   (c) la gestion du travail d'un étudiant.

   Préciser pour chacune des variables trouvées son type (qualitatif ou quantitatif) ainsi que ses modalités.

2. Le tableau 1.2 donne quelques informations numériques sur la population belge au 1er janvier 2017. Les nombres d'hommes, de femmes et de personnes de 65 ans et plus sont indiqués pour les trois régions (Wallonie, Flandre et Bruxelles).

Tableau 1.2: Quelques données démographiques *(Source: Statistics Belgium,* `bestat.statbel.fgov.be`*)*

Régions	Hommes	Femmes	65 ans et plus
Wallonie	1764335	1850138	651573
Flandre	3221295	3294716	1287035
Bruxelles	582375	609229	156489

   (a) Quelle est la région enregistrant le plus haut taux de féminité?

   (b) Laquelle des trois régions a le plus haut taux de personnes âgées de 65 ans et plus par 1.000 habitants ?

3. Le tableau 1.3 donne les nombres d'habitants des trois régions (Wallonie, Flandre et Bruxelles) au premier janvier 2007.

Tableau 1.3: Quelques données démographiques *(Source: Statistics Belgium,* `bestat.statbel.fgov.be`*)*

Régions	Nombres d'habitants
Wallonie	3435879
Flandre	6117440
Bruxelles	1031215

   (a) Déterminer les pourcentages de variation des nombres d'habitants des trois régions entre le premier janvier 2007 et le premier janvier 2017 (données du Tableau 1.2).

   (b) Chaque région dispose de son gouvernement. En 2007, les gouvernements wallon et flamand étaient chacun constitués de 9 ministres, tandis que le gouvernement de la Région Bruxelles-Capitale n'en comportait que 5. Puisque la taille de la population varie d'une région à l'autre, il est difficile de comparer les nombres de ministres. Calculer dès lors un "taux de ministres" dans chaque région afin d'établir une base commune de comparaison.

4. (a) Soit une population P de n individus divisée en deux sous-populations P_1 et P_2 d'effectifs n_1 et n_2 avec $n_1 + n_2 = n$. Si le pourcentage d'individus de P_1 vérifiant

une caractéristique donnée est $p_1\%$ tandis que le pourcentage d'individus de P_2 possédant la même caractéristique est $p_2\%$, quel est le pourcentage d'individus de la population totale P ayant la même caractéristique?

(b) Le tableau 1.4 contient une partie des résultats d'un sondage sur la monarchie réalisé par le quotidien Le Soir en septembre 2017. Sachant que le sondage est basé sur les réponses de 1000 Belges, déterminer approximativement le nombre de Flamands, de Wallons et de Bruxellois interrogés pour ce sondage.

Tableau 1.4: La monarchie en Belgique doit être ramenée à un rôle purement protocolaire, sans la moindre forme de pouvoir?

	Flamands	Wallons	Bruxellois	Royaume
D'accord	$61,3\%$	$34,5\%$	39%	50,6%
Pas d'accord	$29,8\%$	$54,5\%$	$54,9\%$	40,2%
Sans avis	$8,9\%$	11%	$6,1\%$	9,2%

(c) Cette répartition vous semble-t-elle judicieuse?

5. Lors d'un sondage, 1690 personnes ont été interrogées sur leurs habitudes tabagiques. On a obtenu les résultats du tableau 1.5 Quelle est le nombre de femmes interrogées

Tableau 1.5: Habitudes tabagiques de 1690 personnes

	Hommes	Femmes	Total
Fumeur	46%	61%	52%
Non fumeur	54%	39%	48%

lors de ce sondage?

6. Le tableau 1.6 donne le nombre d'employés de sexe féminin et masculin d'une firme pour les années 2010 et 2015:

   (a) Calculer le pourcentage de variation entre 2010 et 2015 du nombre d'employés masculins.

   (b) Sachant que le nombre de femmes employées par cette firme a diminué de 15% entre 2010 et 2015, déterminer le nombre total d'employés de la firme en 2015.

7. Dans une école, 40% des élèves ont une mauvaise vue; 70% des élèves ayant une mauvaise vue portent des lunettes; les 30% restant portent des lentilles de contact. Dans cette école, on compte 21 paires de lunettes. Quelle affirmation est vraie?

Tableau 1.6: Nombre d'employés de sexe féminin et masculin d'une firme pour les années 2010 et 2015

	Année	
Sexe	2010	2015
M	45	59
F	55	x

(a) 45 élèves ont une mauvaise vue;

(b) 30 élèves ont une bonne vue;

(c) on compte 70 élèves dans l'école;

(d) aucune des affirmations précédente n'est vraie.

8. Un bien d'une valeur de 1795 euros a subi une baisse de 12% puis une hausse de 56%. Quel est son nouveau prix ? Si par contre, ce bien a subi une baisse de 72% puis une baisse de 59%. Quel est son nouveau prix ?

9. Les affirmations suivantes sont-elles vraies ou fausses? Justifier.

   (a) Si le prix d'un article augmente de 5%, le prix de trois articles augmente de 15%;

   (b) Un pourcentage est toujours inférieur à 100;

   (c) Avant les soldes, une robe coûtait 110 euros. Pendant les soldes, elle coûte 100 euros. La réduction est donc de 10%;

   (d) Si une grandeur est multipliée par 3 alors elle augmente de 300%;

   (e) En France en 1954, les agriculteurs représentaient $34,8\%$ de la population active, 20 ans après, ils représentent $17,9\%$ de la population active. Le nombre d'agriculteurs a donc été divisé par 2.

   (f) Un bien qui subit une baisse de 62% puis une hausse de 62% ne change pas de prix.

Chapitre 2

Organisation et représentation des données

Dans ce chapitre, on ne considère que des séries univariées obtenues par l'observation d'une seule variable sur une population de taille n. Des tableaux et graphiques, spécifiques aux variables qualitatives ou quantitatives, sont présentés afin de donner un premier aperçu des caractéristiques propres aux données. Les notations et la présentation suivent de près l'ouvrage de Dehon, Droesbeke et Vermandele (2008).

2.1 Variables qualitatives

Cette section concerne des variables qualitatives mais s'applique aussi aux variables quantitatives mesurées sur une échelle ordinale. Considérons une variable qualitative X avec J modalités $m_1, m_2, \ldots, m_J$ observée sur une population de n individus.

2.1.1 Tableau des effectifs

La première opération de mise en ordre des données consiste à dénombrer le nombre d'individus se trouvant dans chacune des catégories de la variable et de transcrire ces informations dans un tableau qui révèle rapidement et clairement la structure des données. Notons n_1 le nombre d'unités statistiques ayant pour modalité m_1, n_2 le nombre d'unités ayant pour modalité m_2, et ainsi de suite pour $n_3, \ldots, n_J$. Les nombres $n_1, \ldots, n_J$ sont les *effectifs* des modalités de la variable. Ils définissent la répartition de la population selon la variable X et sont souvent communiqués via le Tableau 2.1 appelé *tableau des effectifs*.

Comme chaque individu de la population se trouve dans une et une seule catégorie, la

Tableau 2.1: Tableau des effectifs

Modalités de la variable	Effectifs
m_1	n_1
m_2	n_2
$\vdots$	$\vdots$
m_J	n_J

somme des effectifs correspond au nombre total d'unités statistiques dans la population:

$$\sum_{j=1}^{J} n_j = n.$$

Lorsque la variable est ordinale, le tableau des effectifs présente habituellement les modalités dans leur ordre naturel.

Exemple 12 *Considérons la série suivante correspondant au parti du bourgmestre des 27 communes de la province du Brabant wallon*[1] *(classées dans l'ordre alphabétique):*
`PS - MR - Autre - MR - MR - MR - MR - MR - CdH - PS - CdH - MR - MR - MR - Ecolo - MR - Autre - Ecolo - CdH - PS - PS - MR - PS - MR - PS - MR - MR`

Ainsi que précisé au chapitre précédent, cette variable a cinq modalités: `CdH`, `Ecolo`, `MR`, `PS` *et* `Autre`. *Le Tableau 2.2 décrit la distribution des effectifs.*

Tableau 2.2: Tableau des effectifs pour la variable `PartiBourgmestre` observée sur l'ensemble des communes de la province du Brabant Wallon

Catégories	Effectifs
`CdH`	3
`Ecolo`	2
`MR`	14
`PS`	6
`Autre`	2
Total	27

[1]Seule la province du Brabant wallon est considérée dans cette partie afin de traiter un petit nombre d'observations (27 communes seulement sont reprises dans cette province).

2.1.2 Distribution de fréquences

Le tableau des effectifs décrit la répartition de la population en termes absolus. En remplaçant les effectifs par des proportions, on obtient une description relative donnant lieu à la distribution des fréquences.

La *fréquence* f_i de la modalité $m_i, i = 1, \ldots, J$, est définie par la proportion relative à la modalité m_i, c'est-à-dire le rapport

$$f_i = \frac{n_i}{n}.$$

L'ensemble des ratios $f_1, \ldots, f_J$ calculés pour les différentes modalités de la variable fournit la *distribution de fréquences*. Les fréquences vérifient la relation $\sum_{j=1}^{J} f_j = 1$ et peuvent être ajoutées dans une nouvelle colonne dans le tableau des effectifs qui sera ensuite simplement appelé *tableau statistique*.

Remarque: Comme introduit ci-dessus, les fréquences sont habituellement exprimées sous la forme de *proportions* (nombres entre 0 et 1). Fréquemment également, les fréquences sont traduites en pourcentages en multipliant les proportions f_i par 100. Les fréquences peuvent aussi être exprimées en *taux*. Un taux indique, sur une base de 1, 10, 100, 1000,... quelle partie de la population correspond à une des catégories de la variable étudiée. Il est obtenu en multipliant la fréquence par la base 1, 10, 100,... Lorsque la base est égale à 1 (resp. 100), le taux coïncide avec la proportion (resp. le pourcentage). Le choix de la base dépend soit d'une convention, soit de la fréquence d'occurrence de la caractéristique.

2.1.3 Diagramme en barres

La répartition de la population et sa distribution de fréquences peuvent être visualisées sur un *diagramme en barres*.

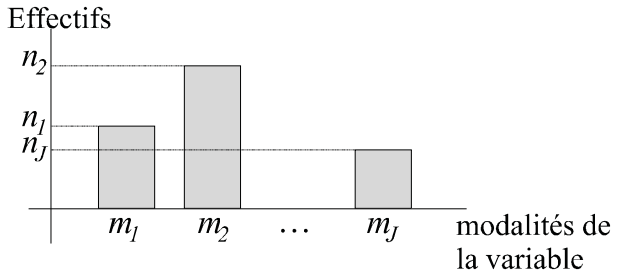


Figure 2.1: Diagramme en barres pour une variable qualitative

Pour construire un diagramme en barres, on associe à chaque modalité distincte observée une colonne verticale (ou horizontale) dont la base n'a pas de signification mais dont la

hauteur (ou longueur) représente l'effectif ou la fréquence de la modalité. Les différentes colonnes sont représentées légèrement espacées comme le montre la Figure 2.1.

Si l'échelle de mesure est ordinale, l'ordre de présentation des modalités est évident. Dans le cas nominal, le choix est plus arbitraire. Le logiciel exploité pour construire les graphiques illustrant ces notes (à savoir le logiciel `R`, `http://cran.r-project.org/`), place par défaut les modalités dans l'ordre alphabétique mais d'autres logiciels les placent dans l'ordre croissant des effectifs.

Exemple 13 *La Figure 2.2 représente, par un diagramme en barres basé sur les effectifs donnés au Tableau 2.2, la répartition de la variable* `PartiBourgmestre` *observée sur l'ensemble des communes de la province du Brabant wallon.*

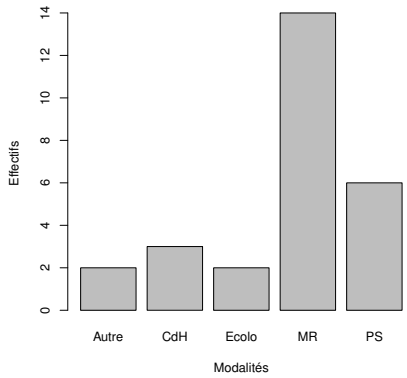
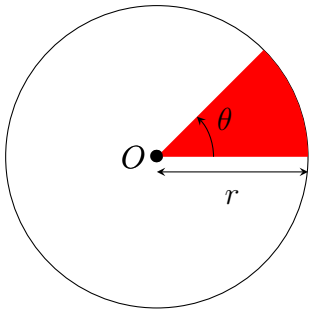


Figure 2.2: Diag. en barres pour `PartiBourgmestre` observée sur les communes du Brabant wallon

2.1.4 Diagramme en secteurs

La répartition de la population et sa distribution de fréquences sont parfois plus expressives lorsqu'on les représente à l'aide d'un *diagramme en secteurs* (ou *camembert*, ou, en anglais, *pie-chart*). Le diagramme en secteurs consiste à représenter la population totale par un disque et à associer à chaque modalité un secteur circulaire (ou part de camembert) dont l'aire est proportionnelle à son effectif ou sa fréquence. Si le disque est de rayon r, l'aire du secteur circulaire d'angle θ (voir dessin) est égale à $\pi r^2 \frac{\theta}{360^\circ}$ et est donc proportionnelle à l'angle au centre θ.



Pour construire un diagramme en secteurs, on doit convertir les fréquences du tableau statistique en un angle, exprimé en degrés (ou en radians). Le disque ayant un angle au centre de 360 degrés et l'aire d'un secteur circulaire étant proportionnelle à son angle au centre, on a, par simple règle de trois,

$$\begin{array}{rcccl} \text{Population totale} & \longleftrightarrow & \text{fréquence} = 1 & \longleftrightarrow & \theta = 360 \text{ degrés} \\ \text{Modalité } m_i & \longleftrightarrow & \text{fréquence} = f_i & \longleftrightarrow & \theta_i = 360 \times f_i \text{ degrés} \end{array}$$

Exemple 14 *La Figure 14 représente, par un diagramme en secteurs, la répartition de la variable* `PartiBourgmestre` *observée sur l'ensemble des communes de la province du Brabant Wallon. On peut vérifier que l'angle au centre du secteur associé à la modalité* `MR` *vaut* $\frac{14}{27} \times 360$ *degrés* ≈ 187 *degrés.*

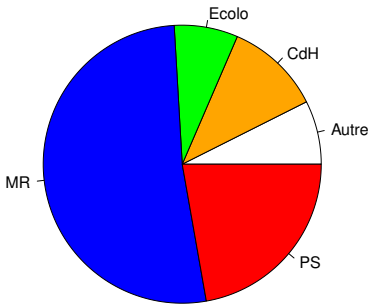


Figure 2.3: Diagramme en secteurs pour la variable `PartiBourgmestre` observée sur l'ensemble des communes de la province du Brabant Wallon

Les représentations des diagrammes en secteurs sont fort variées. Certains secteurs peuvent être détachés par rapport aux autres, le camembert peut être dessiné en trois dimensions... Il faut néanmoins s'assurer de respecter la contrainte de proportionalité des secteurs.

2.2 Variables quantitatives discrètes

2.2.1 Distributions des effectifs et des fréquences

Les modalités d'une variable quantitative discrète sont des valeurs numériques, souvent exprimées en nombres entiers. Les modalités sont donc discontinues tout en respectant un ordre naturel comme une variable qualitative ordinale. Notons les différentes valeurs prises par la variable $x_1, \ldots, x_J$ avec, par convention, $x_1 < \ldots < x_J$. En triant les unités statistiques d'une population selon ces valeurs, on obtient les effectifs $n_1, \ldots, n_J$ décrivant la répartition de la population. Les effectifs divisés par l'effectif total n donnent les fréquences $f_1, \ldots, f_J$ et tous ces résultats peuvent être résumés dans un tableau statistique.

Exemple 15 *La variable* `NombrePartis` *est quantitative discrète et les valeurs observées, sur l'ensemble des communes, sont 1, 2, 3 et 4. Le tableau 2.3 résume les informations statistiques disponibles.*

Tableau 2.3: Tableau statistique pour la variable `NombrePartis` sur l'ensemble des communes

Valeurs	Effectifs	Fréquences
1	181	0.691
2	65	0.248
3	15	0.057
4	1	0.004
Total	262	1

2.2.2 Diagramme en bâtons

Les effectifs et fréquences peuvent aussi être reportés sur un diagramme. Cependant, à la différence des variables qualitatives, les valeurs des variables quantitatives sont des quantités numériques qui peuvent être placées sur l'axe des réels. Dans ce contexte, le diagramme en barres présenté dans le cas qualitatif se transforme en un *diagramme en bâtons*. Ce diagramme consiste à construire, dans un systèmes d'axes orthogonaux, des segments de droite (des bâtons) parallèles à l'axe des ordonnées, élevés en les abscisses $x_j, j = 1, \ldots, J$, et dont les hauteurs sont égales à l'effectif ou la fréquence des valeurs correspondantes.

Exemple 16 *La Figure 2.4 représente le diagramme en bâtons construits à partir des effectifs de la variable* `NombrePartis` *calculés dans le tableau 2.3.*

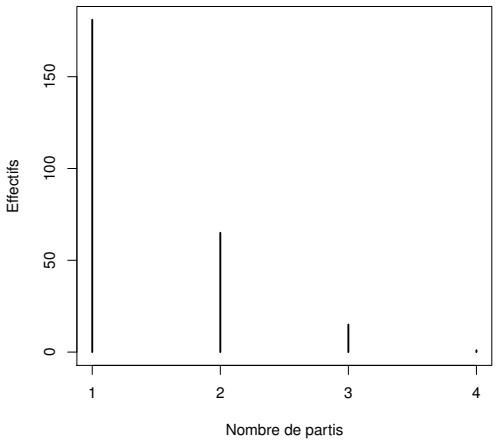


Figure 2.4: Diagramme en bâtons pour la variable `NombrePartis` sur l'ensemble des communes

2.2.3 Effectifs cumulés et courbe cumulative

Au lieu de s'intéresser à une valeur particulière x_j de la variable, on peut s'intéresser à l'ensemble des valeurs de la variable inférieures ou égales à x_j.

Exemple 17 *Dans l'exemple 15, on constate que 181 communes comptent exactement un parti dans la majorité, 65 communes en ont 2,... Le nombre de communes ayant au plus 2 partis dans leurs majorités est donc* $181 + 65 = 246$.

L'effectif de cet ensemble de valeurs est appelé *effectif cumulé* de x_j et est défini par

$$N_j = \sum_{i=1}^{j} n_i.$$

On vérifie aisément que

$$N_1 = n_1, N_j = N_{j-1} + n_j, j = 2, \ldots, J-1, N_J = n.$$

Les effectifs cumulés peuvent être visualisés sur une courbe en escalier appelée *courbe cumulative*. Une telle courbe est illustrée en toute généralité à la Figure 2.5.

Les marches de l'escalier correspondent en abscisse aux valeurs observées x_j de la variable. La hauteur du palier relatif à x_j vaut N_j tandis que la hauteur de la contremarche qui précède vaut n_j. La courbe cumulative est définie pour toute valeur de x observée ou non et peut donc être décrite par une équation du type $y = N(x)$. Pour tout x, la fonction $N(x)$ indique

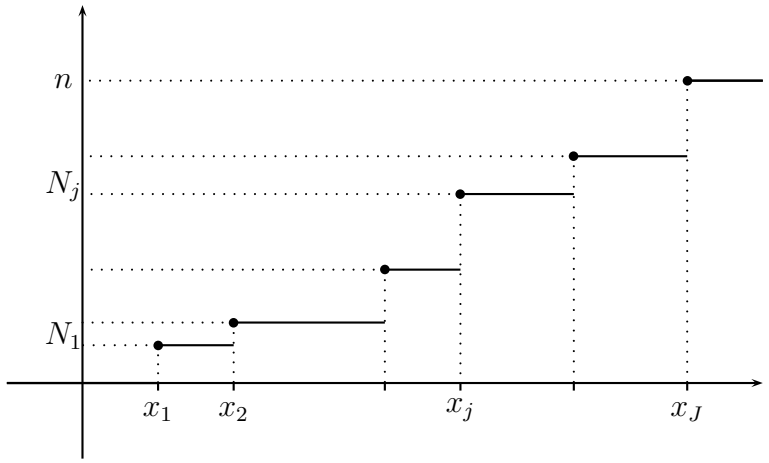


Figure 2.5: Courbe cumulative

le nombre d'observations inférieures ou égales à x. Elle est constante dans chaque intervalle séparant deux valeurs consécutives de la variable:

$$N(x) = N_j \text{ pour } x_j \leq x < x_{j+1}.$$

2.2.4 Fréquences cumulées et fonction de répartition

Au lieu de parler en terme d'effectifs, on peut exprimer l'importance relative du nombre d'observations inférieures ou égales à x_j par rapport au nombre total d'observations en utilisant la fréquence cumulée F_j de x_j. Par définition,

$$F_j = \frac{N_j}{n}, j = 1, \ldots, J.$$

Ces fréquences cumulées vérifient les relations

$$F_1 = f_1; F_j = \sum_{i=1}^{j} f_i = F_{j-1} + f_j; F_J = 1,$$

et peuvent se visualiser sur une courbe cumulative en changeant dans la Figure 2.5 les unités sur l'axe des ordonnées. La courbe cumulative est la représentation graphique de la proportion $F(x)$ des individus de la population pour lesquels la valeur de la variable est inférieure ou égale à x (pour toute valeur de x). Cette fonction $y = F(x)$ est appelée *fonction de répartition* et vérifie

$$F(x) = F_j \text{ pour } x_j \leq x < x_{j+1}.$$

Par définition, $F(x)$ est nulle pour les valeurs de x inférieures à la plus petite valeur observée de la variable et est égale à 1 pour les valeurs de x supérieures à la plus grande valeur observée.

Habituellement, les effectifs et fréquences cumulés sont aussi reportés dans le tableau statistique décrivant la répartition de la population.

Exemple 18 *Pour la variable* `NombrePartis` *décrite dans les exemples 15 et 16, le tableau statistique 2.4 regroupe toutes les informations décrivant numériquement la répartition (ou la distribution) de l'ensemble des communes wallonnes selon cette variable.*

Tableau 2.4: Tableau statistique complet pour la variable `NombrePartis` observée sur l'ensemble des communes

Valeurs	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
1	181	0.691	181	0.691
2	65	0.248	246	0.939
3	15	0.057	261	0.996
4	1	0.004	262	1.000
Total	262	1	x	x

De plus, la Figure 2.6 représente la courbe cumulative des fréquences ou fonction de répartition.

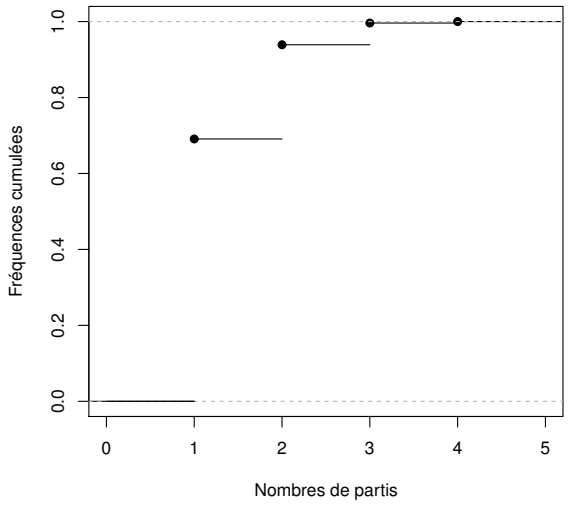


Figure 2.6: Courbe cumulative des fréquences pour la variable `NombrePartis` observée sur l'ensemble des communes

Les fréquences et effectifs cumulés peuvent être également exploités pour caractériser la répartition d'une population suivant une variable qualitative ordinale.

Pour terminer cette section, notons encore que Dehon, Droesbeke et Vermandele (2008) définissent aussi des effectifs et fréquences cumulés *à droite* correspondant aux nombres et proportions d'observations *supérieures ou égales* aux valeurs observées. Nous n'utiliserons pas ces notions ici mais elles apportent une information complémentaire par rapport aux effectifs et fréquences cumulés (*à gauche*) N_j et F_j.

2.3 Variables quantitatives continues

Une variable quantitative continue peut en théorie prendre n'importe quelle valeur à l'intérieur d'un intervalle. En pratique, la précision de l'instrument de mesure n'est jamais infinitésimale et les données obtenues sont en quelque sorte des valeurs arrondies ou discrètes. Par exemple, un poids sera exprimé en milligrammes, grammes, kilos ou tonnes selon l'objectif et les moyens de l'étude, et si les résultats sont exprimés en kilo, les valeurs seront très souvent fournies arrondies à l'unité.

On pourrait donc envisager de décrire la répartition d'une population suivant une variable continue comme on vient de le faire pour une variable discrète. Suivons une démarche similaire à celle présentée par Dehon, Droesbeke et Vermandele (2008):

Exemple 19 *Les observations relatives à la variable* `IndiceRichesse` *observée sur les 84 communes de la province de Liège sont résumées dans le tableau 2.5. Pour rappel, les indices ont été fournis arrondis à l'unité.*

Tableau 2.5: Valeurs des observations de la variable `IndiceRichesse` observée sur les 84 communes de la province de Liège

93	85	96	110	103	107	95	99	98	103	91	106	104	85
106	84	83	124	104	90	109	110	73	117	85	118	98	114
98	108	91	94	108	87	93	106	101	83	100	94	106	107
86	85	92	92	107	96	95	100	106	127	128	119	108	98
97	92	90	102	108	94	79	93	82	98	96	111	92	99
104	102	106	90	97	105	82	102	92	90	102	106	96	96

Pour déterminer les valeurs distinctes prises par la variable, il suffit d'ordonner les valeurs observées, habituellement de la plus petite (ou la moins bonne,...) à la plus grande (ou la meilleure,...). On obtient ainsi ce que l'on appelle la série *ordonnée* notée $\tilde{S} = \{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ où $x_{(i)} \leq x_{(j)}$ si $i \leq j$. Pour mieux comprendre le passage de la série initiale à la série ordonnée, considérons l'exemple discret suivant (la définition de la série ordonnée est la même dans les cas discret et continu):

Exemple 20 *Une enquête menée dans un quartier comptant dix maisons tentait de déterminer le flux de trafic local. Pour cela, le nombre de voitures (variable X) possédées par chacune des familles a été enregistré. Le tableau suivant reprend les données observées où l'indice i correspond aux numéros des maisons:*

i	1	2	3	4	5	6	7	8	9	10
x_i	0	3	1	2	1	0	2	2	2	3

Comme les données x_i sont numériques, elles peuvent être rangées dans l'ordre croissant: 2 données sont nulles (les familles habitant les maisons des numéros 1 et 6 ne possèdent pas de voiture), 2 données sont égales à 1 (les familles d'indices 3 et 5 ont une seule voiture), 4 données sont égales à 2 (les familles d'indices 4, 7, 8, 9 possèdent deux voitures), et deux familles (indices 2 et 10) ont trois voitures. La série ordonnée est donc $\tilde{S} = \{0, 0, 1, 1, 2, 2, 2, 2, 3, 3\}$. Ainsi, $x_{(2)} = x_1, x_{(10)} = x_{10}, \ldots$

Revenons maintenant à l'exemple de l'indice de richesse.

Exemple 21 *En triant les données dans l'ordre croissant comme dans le tableau 2.6, on repère effectivement plus facilement les différentes valeurs prises par la variable ainsi que le nombre de fois que chacune d'elles a été observée.*

Tableau 2.6: Observations ordonnées de la variable `IndiceRichesse` observée sur les 84 communes de la province de Liège

73	79	82	82	83	83	84	85	85	85	85	86	87	90
90	90	90	91	91	92	92	92	92	92	93	93	93	94
94	94	95	95	96	96	96	96	96	97	97	98	98	98
98	98	99	99	100	100	101	102	102	102	102	103	103	104
104	104	105	106	106	106	106	106	106	106	107	107	107	108
108	108	108	109	110	110	111	114	117	118	119	124	127	128

Transcrire les effectifs des valeurs observées dans un tableau des effectifs conduirait à un grand nombre de lignes tandis que de nombreux effectifs seraient de faible amplitude. De même, utiliser un diagramme en bâtons pour représenter la répartition de la population mène à la Figure 2.7 qui est fort chahutée même si certaines tendances s'en dégagent.

Pour une meilleure vision de la répartition de la population, il est courant d'effectuer des groupements. Plusieurs valeurs proches de la variable sont regroupées dans une même catégorie (appelée *classe*). On parle alors de *données groupées*.

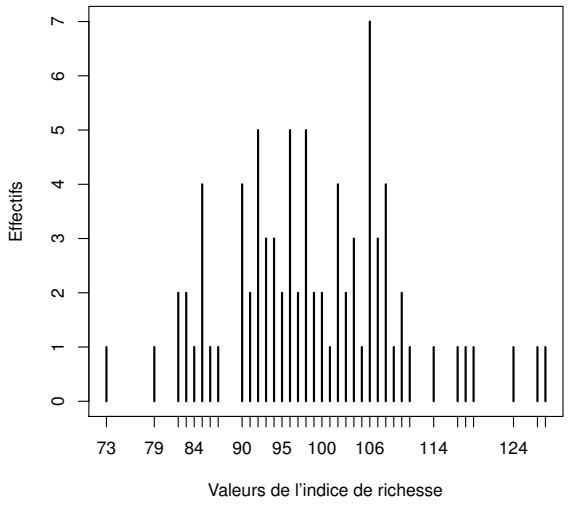
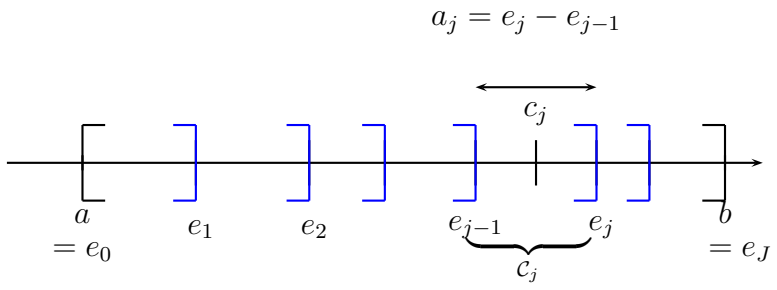


Figure 2.7: Diagramme en bâtons pour la variable `IndiceRichesse` observée sur les 84 communes de la province de Liège

2.3.1 Groupement des données

Considérons une variable continue X dont les valeurs se situent dans l'intervalle de variation $[a, b]$. On divise cet intervalle en J classes $\mathcal{C}_1, \ldots, \mathcal{C}_J$: de $a = e_0$ à e_1, de e_1 à $e_2, \ldots$, et de e_{J-1} à $e_J = b$. La $j$ème classe $\mathcal{C}_j$ est caractérisée par ses deux *bornes* e_{j-1} et e_j, par son *amplitude* $a_j = e_j - e_{j-1}$, par son *centre* $c_j = \frac{e_j + e_{j-1}}{2}$ et aussi par son *effectif* n_j (qui est le nombre d'observations se trouvant dans la classe).



De manière à obtenir des classes mutuellement exclusives, il faut préciser si les bornes inférieure et supérieure de la classe appartiennent ou non à la classe. Par convention (et sauf mention explicite d'une autre règle), nous considérerons des classes $\mathcal{C}_j =]e_{j-1}, e_j]$ pour $j = 2, \ldots, J$ et $\mathcal{C}_1 = [e_0, e_1]$. Une classe correspondant à un intervalle dont la borne inférieure ou supérieure n'existe pas est dite *ouverte* (ou *non bornée*). De telles classes ont un inconvénient majeur: elles n'ont pas de centre.

On définit aussi

- La fréquence de la classe $\mathcal{C}_j$: $f_j = \frac{n_j}{n}$.
- L'effectif cumulé de la classe $\mathcal{C}_j$ (c'est-à-dire le nombre d'observations appartenant aux j premières classes, ou encore, inférieures ou égales à e_j): $N_j = \sum_{i=1}^{j} n_i$.
- La fréquence cumulée de la classe $\mathcal{C}_j$: $F_j = \frac{N_j}{n}$.

Il n'y a pas de méthode universelle pour grouper des données. Souvent, le statisticien recourt à des règles plus ou moins empiriques pour répondre aux questions suivantes:

- Combien de classes faut-il considérer?

  → Si on forme trop peu de classes, on risque de perdre trop d'informations par rapport à la série de départ. Par contre, considérer trop de classes entraîne les mêmes inconvénients que pour la série non groupée (diminution de la clarté dans la présentation des résultats). Comme règle générale, il est recommandé de ne pas utiliser moins de 5 classes ou plus de 15 classes dans le groupement des données. Souvent, le nombre de classes est fixé par la nature du problème. Si ce n'est pas le cas, on peut toujours suivre la *formule de Sturges* qui préconise un nombre J de classes égal au plus petit nombre entier supérieur ou égal à k avec

  $$k = 1 + \frac{10}{3} \log_{10} n$$

  où n est l'effectif de la population.

- Les classes ont-elles toutes la même amplitude?

  → C'est le cas le plus fréquent et le plus simple puisque des classes de longueurs différentes entrainent des difficultés supplémentaires dans les représentations graphiques.

- Comment choisir les bornes des classes?

  → Souvent, la première borne inférieure coïncide avec la donnée la plus petite et la dernière borne supérieure avec la donnée la plus grande. Ensuite, suivant les décisions prises pour les deux questions précédentes, on peut déterminer les bornes intermédiaires.

Un outil assez performant pour déterminer des groupements adéquats est le *diagramme en tiges et feuilles* (stem and leaf display, en anglais) proposé par Tukey (1977). Ce graphique décrit la distribution des effectifs de la série tout en présentant l'ensemble des valeurs distinctes observées. Pour construire un tel diagramme, chaque valeur de la série est décomposée

en deux parties: la partie principale appelée *tige* et la partie secondaire appelée *feuille*. Ensuite, les tiges sont alignées les unes en dessous des autres (dans l'ordre croissant) et les feuilles sont écrites à côté, de nouveau dans l'ordre croissant. La définition des tiges et des feuilles dépend des données. Les deux exemples suivants vont illustrer cela.

Exemple 22 *Chaque observation de la série statistique* $S = \{6.4, 7.6, 8.8, 8.8, 9.1, 9.4, 9.5, 9.8, 10.0, 10.4, 10.5, 10.6, 11.3, 11.3, 11.9, 12.0, 12.1, 12.1, 12.4, 12.4\}$ *peut être décomposée en sa partie entière et sa partie décimale. Un diagramme en tiges et feuilles pour cet ensemble de données peut être alors construit en les deux étapes suivantes:*

1. *isoler dans une colonne la partie entière des données, les tiges*
2. *inscrire, pour toute observation, la partie décimale, la feuille, dans une deuxième colonne sur la même ligne que la partie entière correspondante*

6	4
7	6
8	88
9	1458
10	0456
11	339
12	01144

Exemple 23 *En reprenant l'exemple de l'indice de richesse, les différentes tiges peuvent être définies par les valeurs 70, 80, 90, 100, 110 et 120 tandis que les feuilles correspondent aux unités. Le diagramme en tiges et feuilles relatif à ce choix de tiges et de feuilles est représenté ci-dessous.*

7	39
8	22334555567
9	000011222223334445566666778888899
10	00122223344456666666777888889
11	0014789
12	478

Cette représentation montre qu'une répartition en six classes d'amplitude 10 ($\mathcal{C}_1 = [70, 80]$; $\mathcal{C}_2 =]80, 90], \ldots$*) est envisageable (attention: vu la convention choisie pour les bornes des classes, les feuilles 0 sont classées avec les feuilles de la tige précédente). Néanmoins, en procédant de la sorte, 59 valeurs prises par la variable sont confinées dans les deux classes centrales (sur 84 valeurs). Il est possible d'affiner le diagramme afin de décomposer les tiges en deux parties; la première partie reprenant les feuilles de 0 à 4; la deuxième, les feuilles de 5 à 9. Dans ce cas, le diagramme prend la forme suivante*

7	3
7	9
8	22334
8	555567
9	00001122222333444
9	5566666778888899
10	001222233444
10	5666666677788889
11	0014
11	789
12	4
12	78

Vu ce diagramme, une décomposition en huit classes d'amplitudes variables (5 ou 10 unités) semble convenir. Notons que ce nombre de classes est aussi conseillé par la formule de Sturges (k vaut 7.4). Le tableau 2.7 résume l'information dont on dispose après le groupement en classes.

Tableau 2.7: Tableau statistique pour la variable pour la variable `IndiceRichesse` observée sur les 84 communes de la province de Liège

Classes	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
$[70; 80]$	2	0.02	2	0.02
$]80; 90]$	15	0.18	17	0.20
$]90; 95]$	15	0.18	32	0.38
$]95; 100]$	16	0.19	48	0.57
$]100; 105]$	11	0.13	59	0.70
$]105; 110]$	17	0.20	76	0.90
$]110; 120]$	5	0.06	81	0.96
$]120; 130]$	3	0.04	84	1
Total	84	1	x	x

Exemple 24 *Le Service Public Fédéral "Economie" publie dans la revue "Statistiques financières" les données relatives à la statistique fiscale des revenus soumis à l'impôt des personnes physiques. Ces données sont habituellement fournies déjà groupées dans des classes précises. Ces classes sont au nombre de 101, les 100 premières d'amplitude 1.000 Euro (définies selon la convention contraire par rapport à celle que nous avons privilégiée), tandis que la dernière est non bornée. Vu le nombre de déclarations concernées (de l'ordre de 5 millions), la formule de Sturges préconise une répartition en 23 classes. Le Tableau 2.8 propose une telle répartition pour la distribution des revenus correspondant à l'exercice 2002 en considérant des classes d'amplitude 1, 2, 5, 10, 15 ou 30 (la dernière classe restant non bornée).*

Tableau 2.8: Répartition du revenu total net imposable en Belgique pour l'exercice 2002 (revenus de 2001)

Classes (unité: 1000 Euro)		Nombres de déclarations	Pourcentages
1	$[0, 2[$	152.812	3,11
2	$[2, 3[$	47.179	0,96
3	$[3, 4[$	50.414	1,03
4	$[4, 5[$	52.955	1,08
5	$[5, 6[$	58.689	1,20
6	$[6, 7[$	75.877	1,55
7	$[7, 8[$	99.202	2,02
8	$[8, 9[$	103.927	2,12
9	$[9, 10[$	143.691	2,93
10	$[10, 11[$	203.232	4,14
11	$[11, 12[$	201.342	4,10
12	$[12, 13[$	203.870	4,16
13	$[13, 14[$	182.769	3,73
14	$[14, 15[$	191.786	3,91
15	$[15, 16[$	182.325	3,72
16	$[16, 17[$	172.746	3,52
17	$[17, 18[$	169.881	3,46
18	$[18, 19[$	166.977	3,40
19	$[19, 20[$	162.825	3,32
20	$[20, 25[$	639.957	13,05
21	$[25, 30[$	405.970	8,28
22	$[30, 45[$	697.043	14,20
23	$[45, 75[$	420.780	8,56
24	$[75, +\infty[$	119.374	2,42
Total		4.905.623	100

Notons que la convention choisie par le SPF Economie pour la construction des classes n'est pas la même que celle proposée dans ces notes: les bornes supérieures des classes n'appartiennent pas aux classes.

2.3.2 Histogramme

La répartition de la population en classes peut être visualisée sur un *histogramme*. Dans un système d'axes orthogonaux, chaque classe correspond à un rectangle dont la base coïncide avec l'intervalle de la classe (segment de longueur a_j) et dont la *surface* est proportionnelle à son effectif ou à sa fréquence. Dans un histogramme, c'est la façon dont la masse se répartit et la position que celle-ci occupe sur l'axe réel qu'il faut analyser. De même, la présence d'un pic (ou mode) ou de plusieurs pics permet de tirer des informations utiles sur la série statistique.

- Lorsque toutes les classes ont la même amplitude $a_1 = \ldots = a_J = a$, il est équivalent de construire des rectangles dont la hauteur est proportionnelle aux effectifs ou fréquences des classes. Il est même souvent commode de prendre la hauteur d'un rectangle exactement égale à l'effectif ou la fréquence de la classe correspondante. Dans ce cas, la surface totale de l'histogramme vaut an ou a.

  Exemple 25 *Un histogramme caractérisant la distribution de l'indice de richesse dans les communes liégeoises lorsque celui-ci est réparti en 6 classes d'amplitude 10 est représenté à la Figure 2.8. La hauteur des rectangles est prise simplement égale à l'effectif des classes.*

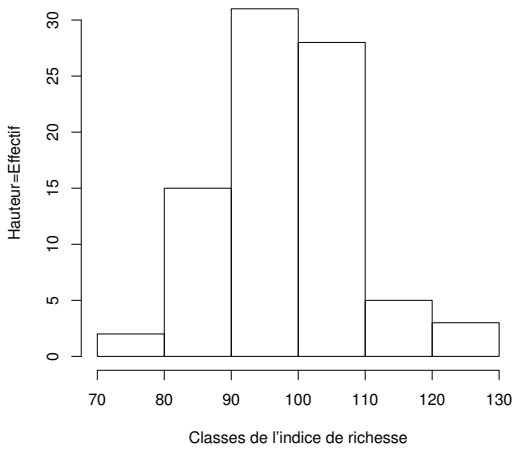


Figure 2.8: Histogrammes pour la répartition en 6 classes d'amplitude constante pour l'indice de richesse dans les communes liégeoises

  On constate que l'impression qui ressort de l'analyse de l'histogramme est fort similaire à celle obtenue à partir du diagramme en tiges et feuilles (à une rotation près) représenté à l'exemple 23.

  L'histogramme montre une distribution assez symétrique de la distribution de l'indice de richesse, avec la plus grosse partie de sa masse placée au centre de la distribution. Il présente un seul pic.

- Lorsque les amplitudes des classes sont différentes, prendre comme hauteur l'effectif ou la fréquence de la classe ne mène plus à des aires caractérisées par le même facteur de proportionalité, ce qui donne un poids trop important aux grandes classes. Le plus simple est alors de construire des *histogrammes d'aire unitaire* (aire totale égale à 1),

dont les rectangles ont des aires exactement égales à la fréquence de la classe. Notons h_j la hauteur du $j$ème rectangle, la base du rectangle correspondant, par définition, à l'amplitude de la classe a_j. Pour construire un histogramme d'aire unitaire, on doit donc trouver h_j tel que

$$h_j a_j = \text{ aire du } j\text{ème rectangle } = f_j \Longleftrightarrow h_j = \frac{f_j}{a_j}.$$

Exemple 26 *Considérons à nouveau la distribution des indices de richesse des communes de la province de Liège. En exploitant cette fois-ci la répartition en huit classes d'amplitude 5 ou 10, on obtient l'histogramme de la Figure 2.9.*

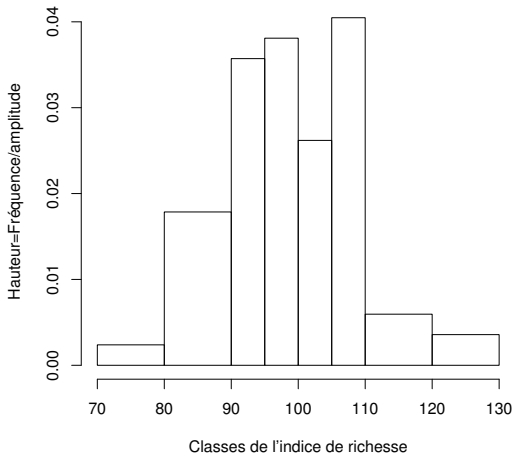


Figure 2.9: Histogrammes pour la répartition en 8 classes de l'indice de richesse dans les communes liégeoises

L'histogramme est ici aussi fort similaire à l'image de la distribution procurée par le diagramme en tiges et feuilles correspondant.

Exemple 27 *Comme deuxième exemple, revenons à la répartition des nombres de déclarants Belges classés selon les catégories de revenus considérée à l'exemple 24. Les amplitudes des classes bornées sont égales à 1, 2, 5, 15 et 30, tandis que la dernière classe a une amplitude infinie. Une classe non bornée est ignorée lors de la construction de l'histogramme d'aire unitaire (la hauteur qui lui serait associée est nulle puisque la fréquence, valeur comprise entre 0 et 1, est divisée par une amplitude de classe infinie).*

L'histogramme d'aire unitaire est représenté à gauche à la Figure 2.10.

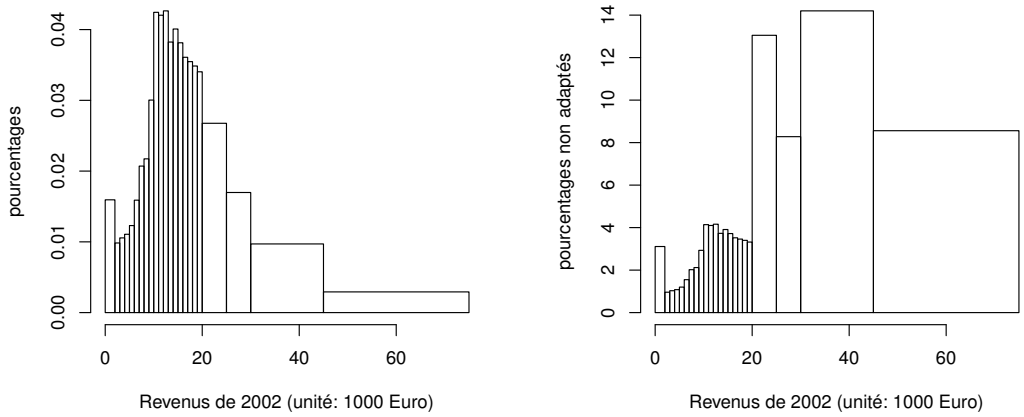


Figure 2.10: Histogrammes de la répartition des revenus en Belgique pour l'exercice 2002: en ajustant les hauteurs en fonction des amplitudes (à gauche) sans ajuster les hauteurs (à droite)

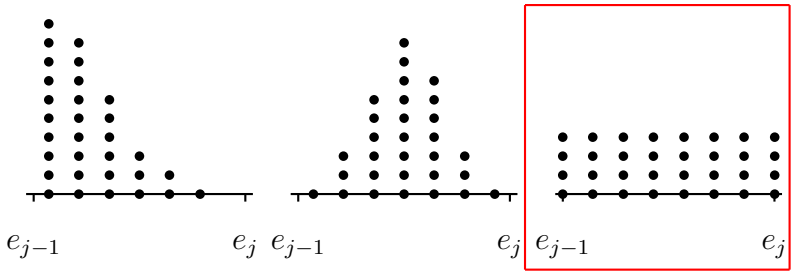
Ne pas ajuster la hauteur des rectangles en fonction de l'amplitude des classes introduit une distorsion pour les classes les plus étendues. Dans l'histogramme de droite de la Figure 2.10, les hauteurs de tous les rectangles sont données par les fréquences des classes. Les plus grandes classes sont surévaluées par rapport aux classes des déclarants aux revenus plus modestes.

Remarques:

- Si l'on souhaite comparer deux séries de données à l'aide d'histogrammes, il faut prendre certaines précautions. Si les effectifs totaux des deux séries sont équivalents, toute construction est adéquate. Par contre, si les effectifs totaux sont différents, il est impératif d'exploiter des histogrammes d'aire unitaire.

- L'histogramme décrit la répartition de la masse totale au sein des différentes classes tout en supposant qu'au sein de chacune des classes, la répartition est *uniforme*. Plus précisément, considérons la $j$ème classe comptant n_j observations. Les valeurs précises des observations ont été "perdues" lors du groupement mais, initialement, celles-ci se répartissaient d'une certaine façon à l'intérieur de l'intervalle de classe, par exemple selon les schémas suivants[2].

[2]Chaque point est une observation ayant une valeur comprise entre les deux bornes des classes. Lorsque deux observations ont la même valeur, elles sont représentées l'une au-dessus de l'autre.

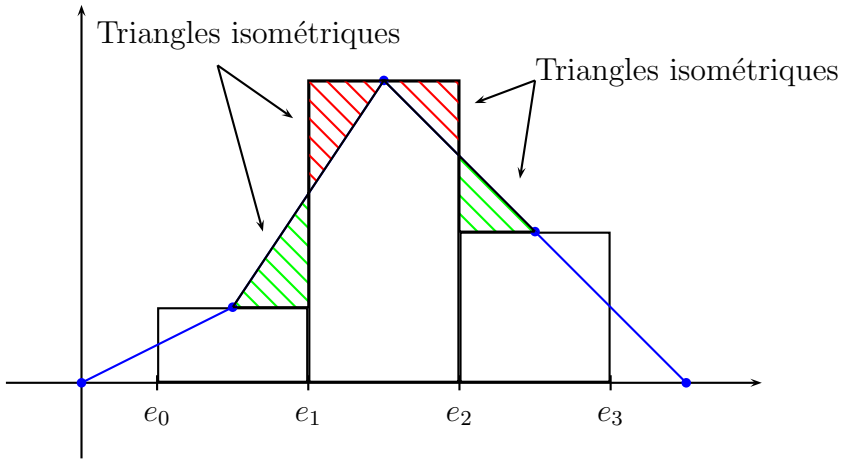


L'histogramme est construit sous l'hypothèse d'une répartition uniforme (3ème schéma). Sous cette hypothèse, quelle que soit la classe (par exemple $\mathcal{C}_j$) et pour toute unité d'amplitude Δ, il existe $k_j \in I\!R^+$ tel que $k_j\Delta$ observations se trouvent dans l'intervalle $]x, x+\Delta]$, pour tout $x \in \mathcal{C}_j$. Il s'agit d'une hypothèse simplificatrice.

2.3.3 Polygone des effectifs ou des fréquences

Le *polygone des effectifs ou des fréquences* est une autre représentation graphique qui donne de la distribution une image plus continue.

- Lorsque toutes les classes ont la même amplitude a, le polygone des effectifs (resp. des fréquences) s'obtient en reliant par une ligne brisée les points qui se trouvent au milieu des côtés supérieurs des rectangles de l'histogramme des effectifs (resp. des fréquences), comme suit:



Par définition, la somme des surfaces des rectangles de l'histogramme vaut an ou a. Le segment reliant deux milieux consécutifs délimite deux triangles isométriques (donc de surfaces égales), un se trouvant sous la ligne brisée et l'autre au-dessus. La surface de l'histogramme tronquée par le segment est donc récupérée par une surface identique sous la ligne. Pour que cette propriété reste vraie pour l'entièreté de la surface délimitée par le polygone et l'axe des abscisses, on complète la courbe avec les deux points $(e_0 - \frac{a}{2}, 0)$ et $(e_J + \frac{a}{2}, 0)$.

Exemple 28 *Le polygone construit sur l'histogramme représenté à la Figure 2.8 est tracé à la Figure 2.11.*

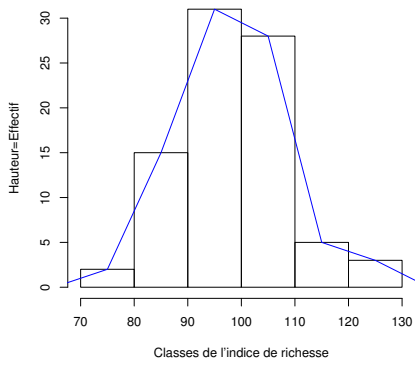
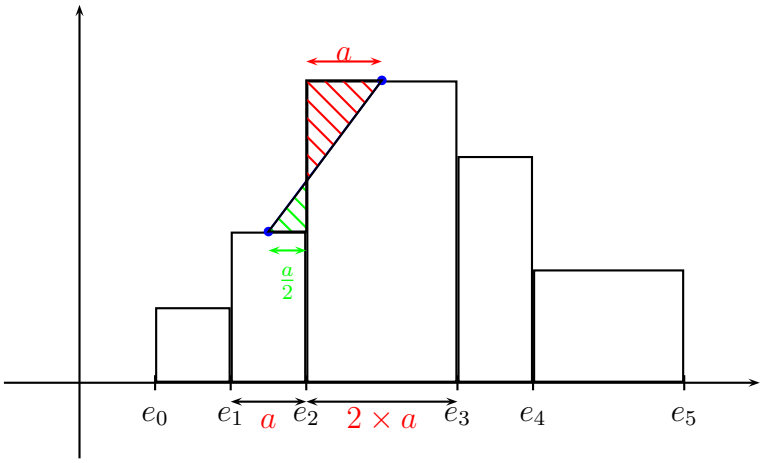


Figure 2.11: Polygone caractérisant la distribution de l'indice de richesse dans la province de Liège

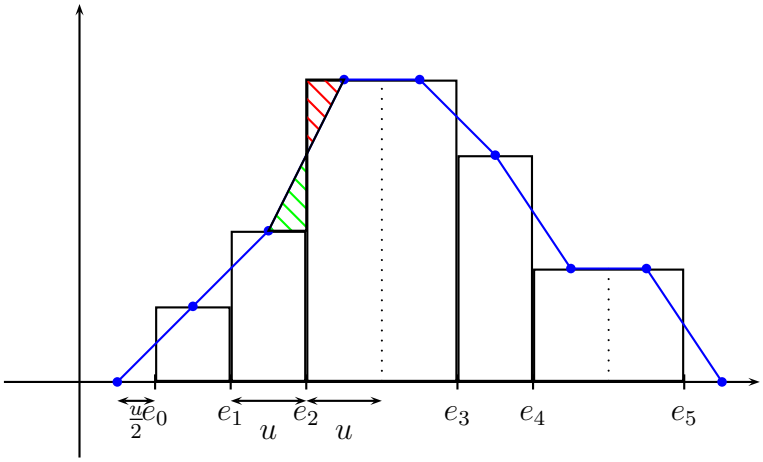
Constatons que l'on peut, à propos du polygone, dresser les mêmes constats que ceux déjà mis en évidence pour l'histogramme: le polygone est plus ou moins symétrique avec la plus grosse partie de la masse au centre de la distribution et il présente un unique pic.

- Lorsque les amplitudes des classes varient, relier les milieux des côtés supérieurs des rectangles de l'histogramme ne permet pas de respecter la propriété de conservation de la surface.



En effet, comme illustré ci-dessus, les segments joignant les milieux de deux rectangles consécutifs de bases différentes ne délimitent plus des triangles isométriques.

Pour assurer le caractère isométrique des triangles, il faudrait relier les milieux de rectangles de même largeur. Pour obtenir de tels rectangles, nous allons artificiellement décomposer les rectangles en "sous-rectangles" de même base u. Cette base u s'appelle l'*unité d'amplitude de classe* et correspond au plus grand commun diviseur des diverses amplitudes de classe. Ce sont les milieux des côtés supérieurs des nouveaux rectangles ainsi construits qui, reliés par une ligne brisée, vont définir le polygone, les grandes classes correspondant à des paliers horizontaux.



En complétant le polygone par les deux points $(e_0 - \frac{u}{2}, 0)$ et $(e_J + \frac{u}{2}, 0)$, la surface délimitée par celui-ci et l'axe des abscisses est égale à la surface de l'histogramme.

Exemple 29 *Le polygone de la Figure 2.12 correspond aux indices de richesse décomposés en les huit classes d'amplitudes variables (pour lesquelles l'unité d'amplitude vaut 5).*

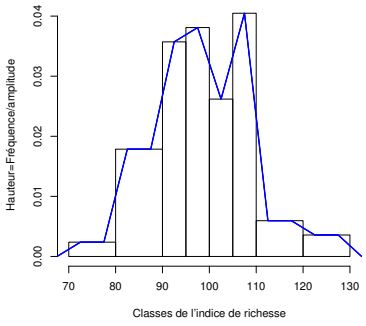


Figure 2.12: Polygone caractérisant la distribution de l'indice de richesse dans la province de Liège lorsque les classes sont d'amplitude variable

Comme déjà souligné, l'histogramme et le polygone permettent d'avoir une image rapide des caractéristiques principales des données (les pics, les baisses, les points de concentration,...). Le polygone a un avantage sur l'histogramme: il permet de comparer directement plusieurs distributions de fréquences.

Exemple 30 *Une comparaison entre les répartitions des revenus en Belgique pour deux années différentes peut être obtenue en superposant les polygones construits à partir des histogrammes unitaires.*

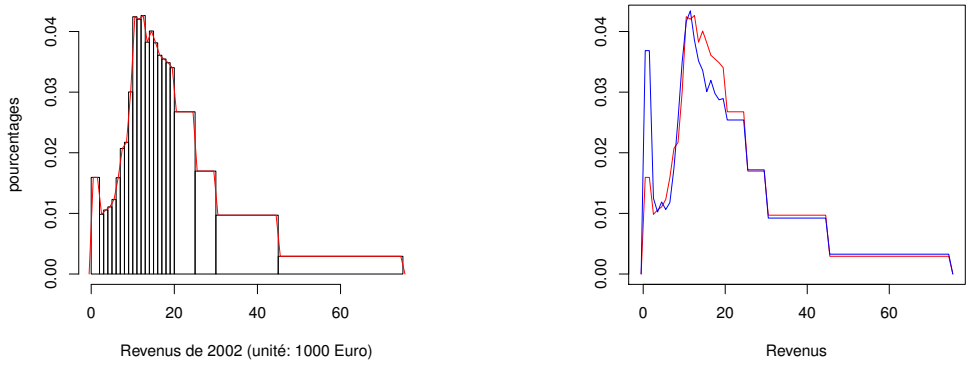


Figure 2.13: Polygone de la répartition des revenus de 2002 et comparaison des répartitions des revenus pour les année 2002 et 2007.

Le graphique de gauche de la Figure 2.13 correspond au polygone construit sur l'histogramme "unitaire" de la répartitions des revenus de l'année 2002. La surface de l'histogramme (et donc également, la surface sous son polygone) est un peu inférieure à 1 (il ne faut pas oublier que la classe non bornée, représentant une masse non nulle, est ignorée dans cette représentation). Le graphique de droite de la Figure 2.13 superpose ce polygone à celui construit sur l'histogramme unitaire décrivant la répartition des revenus en 2007. On constate que les polygones des fréquences des deux années considérées sont fort similaires mis à part le pic plus important en 2007 qu'en 2002 pour les petits revenus.

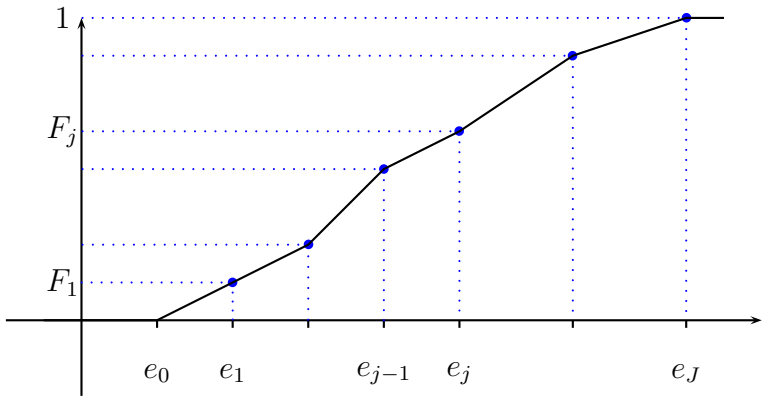
2.3.4 Ogive des fréquences cumulées

Par définition, la fréquence cumulée F_j de la classe $\mathcal{C}_j$ est la proportion d'observations se trouvant dans les j premières classes ou inférieures ou égales à la borne supérieure e_j de la $j$ème classe. Comme dans le cas d'une variable quantitative discrète, on voudrait obtenir une représentation graphique de la fonction de répartition $y = F(x)$ qui, pour toute valeur de x, donne la fréquence des observations inférieures à x.

Une première esquisse de cette fonction peut être obtenue en reportant dans un système d'axes orthogonaux les couples de points $(e_j, F_j), j = 1, \ldots, J$. Pour "boucher les trous", c'est-à-dire pour définir la valeur de la fonction entre deux bornes de classes, il faut imposer une condition sur la façon dont les observations se répartissent au sein de chaque classe. A nouveau, l'hypothèse la plus simple est celle de répartition uniforme, qui suppose que les observations remplissent équitablement toute la classe. Plus concrètement, si $N(x)$ désigne le nombre d'observations de la série inférieures ou égales à x, alors, en exploitant les mêmes notations que précédemment, pour toute classe $\mathcal{C}_j$ et pour une unité d'amplitude Δ, il existe une constante $k_j \in \mathbb{R}^+$ telle que, pour tout $x \in \mathcal{C}_j$,

$$N(x + \Delta) - N(x) = k_j \Delta \Leftrightarrow \frac{N(x + \Delta) - N(x)}{\Delta} = k_j.$$

En passant à la limite pour $\Delta \to 0$, on obtient $N'(x) = k_j$ ou encore $N(x) = k_j x + b_j$. L'hypothèse de répartition uniforme permet donc de représenter la fonction de répartition dans chaque classe par une ligne brisée joignant les couples de points (e_j, F_j). En notant finalement que la fonction doit être nulle pour toute valeur de x inférieure à la borne inférieure e_0 et égale à 1 pour toute valeur de x supérieure à la dernière borne supérieure e_J, on obtient la ligne brisée représentée ci-dessous et qui est appelée *ogive des fréquences cumulées*.



La même démarche peut être suivie pour représenter l'évolution des effectifs cumulés mais l'exploitation des fréquences cumulées est plus fréquente en pratique.

L'ogive des fréquences cumulées peut être exploitée de plusieurs façons pour obtenir des informations sur la distribution des fréquences. Par définition, pour une valeur x^* donnée, $F(x^*)$ indique la proportion des observations dont la valeur est inférieure ou égale à x^*. Réciproquement, pour toute proportion $y^* \in]0, 1[$, la fonction inverse de F donne la valeur de la variable en dessous de laquelle une proportion y^* d'observations se trouve.

Chaque application implique simplement le calcul de l'équation d'une droite, afin de pouvoir déterminer l'ordonnée y^* correspondant à une abscisse donnée ou, réciproquement, retrouver l'abscisse x^* correspondant à une ordonnée donnée.

Exemple 31 *Les ogives des fréquences cumulées pour les répartitions des revenus en 2002 et 2007 sont presque des courbes lisses (voir la Figure de gauche de la Figure 2.14). Le graphique de droite caractérise lui les années 1986 et 1996 (F_1 correspond à l'année 1986 et F_2 à l'année 1996, les revenus étant exprimés en unités égales à 1000 francs belges). Pour un revenu x donné, $F_1(x)$ et $F_2(x)$ indiquent les proportions des personnes déclarant un revenu inférieur ou égal à x lors de l'année considérée. Réciproquement, à partir d'une proportion $0 < y < 1$, les fonctions inverses permettent d'estimer le revenu maximal déclaré par la proportion y des personnes de revenus les plus bas. On constate ainsi que le pourcentage de personnes ayant un revenu inférieur à $x^* = 1200 \times 1000$ francs est supérieur en 1986 qu'en 1996. On voit aussi que le revenu maximal déclaré par la moitié la plus pauvre des déclarants est passé de 500 à 700 (en unité égale à 1 000 francs) en dix ans. Par contre, la comparaison des courbes des années 2002 et 2007 ne montre pas de changement significatif en 5 ans avec des revenus exprimés en Euro.*

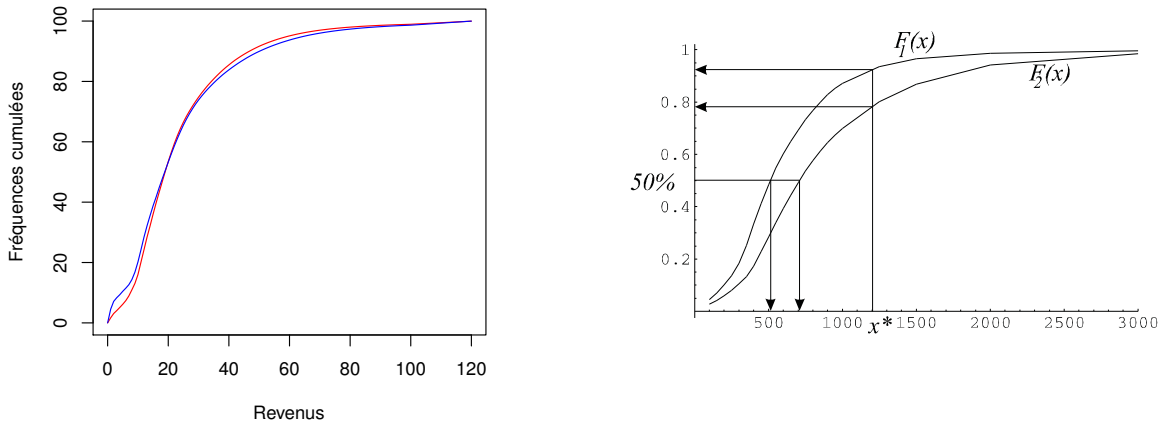


Figure 2.14: Ogives des fréquences cumulées pour les répartitions des revenus en Belgique pour les années 2002 et 2007 (à gauche) et, à titre d'information, pour les année 1986 et 1996 (à droite)

On constate que l'ogive des fréquences cumulées décrit, comme l'histogramme, la façon dont la masse se répartit sur l'intervalle de définition $[a, b]$. Elle permet notamment de déterminer la proportion d'individus dont la valeur observée se trouve entre deux valeurs x_1

et x_2 en exploitant la différence $F(x_2) - F(x_1)$. Les deux graphiques contiennent donc des informations équivalentes, même si celles-ci s'obtiennent à l'aide d'un calcul d'aire à partir de l'histogramme et à l'aide d'une interpolation linéaire à partir de l'ogive.

La Propriété 1 formalise le lien entre ces deux représentations graphiques.

Propriété 1 *Soit X une variable continue dont les valeurs sont groupées en J classes et dont la distribution des fréquences est décrite par un histogramme de surface unitaire ainsi que par l'ogive des fréquences cumulées $y = F(x)$. Pour toute valeur x^*, la surface délimitée par l'histogramme et l'axe des abscisses et située à gauche de x^* est égale à l'ordonnée $F(x^*)$.*

Preuve:
Soient les classes $[e_0, e_1],\ ]e_1, e_2], \ldots,\ ]e_{J-1}, e_J]$. Notons $A(x^*)$ la surface à gauche de x^*,

- si $x^* < e_0$, $A(x^*) = 0$ et $F(x^*) = 0$.
- si $x^* \geq e_J$, $A(x^*) = 1$ et $F(x^*) = 1$.
- si $x^* \in]e_{j-1}, e_j]$,

$$A(x^*) = \sum_{k=1}^{j-1} a_k \frac{f_k}{a_k} + (x^* - e_{j-1})\frac{f_j}{a_j} = F_{j-1} + \frac{f_j}{a_j}(x^* - e_{j-1}).$$

La fonction $F(x)$ est une droite entre les points (e_{j-1}, F_{j-1}) et (e_j, F_j) dont l'équation s'écrit

$$y - F_{j-1} = \frac{F_j - F_{j-1}}{e_j - e_{j-1}}(x - e_{j-1}).$$

Le point $(x^*, F(x^*))$ appartenant à cette droite, on obtient

$$F(x^*) = F_{j-1} + \frac{f_j}{a_j}(x^* - e_{j-1}).$$

La conclusion $F(x^*) = A(x^*)$ est immédiate. □

Pour conclure cette section sur les variables statistiques quantitatives continues, notons que la fonction de répartition (ou ogive) telle que définie ci-dessus dépend de la répartition en classes effectuée. Une autre construction de cette fonction est également classique lorsque l'on dispose des données brutes $x_1, \ldots, x_n$. Cette autre version s'appelle la *fonction de répartition empirique* et se définit comme suit:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_{(i)} \leq x)$$

où la fonction indicatrice $I(x_{(i)} \leq x)$ prend la valeur 1 ou 0 selon que $x \geq x_{(i)}$ ou non. Par définition, $F(x_{(i)}) = i/n$.

Cette fonction est identique à la fonction cumulative des fréquences cumulées introduite pour décrire la distribution des fréquences cumulées d'une variable quantitative discrète. La différence principale provient du fait que la plupart des "contre-marches" de la fonction en escalier sont de hauteur assez réduite (puisqu'il y a peu de répétitions parmi les valeurs observées d'une variable quantitative continue).

La fonction de répartition et l'ogive sont définies dans la même optique: déterminer la masse d'individus dont la valeur observée pour la variable est inférieure ou égale à une valeur donnée. Elles correspondent à deux stratégies différentes: estimer cette proportion à partir des données brutes ou approximer cette proportion à l'aide d'une répartition en classe. Dans la plupart des cas (sauf lorsque la répartition en classes est peu adéquate), les deux fonctions seront proches l'une de l'autre, ainsi qu'illustré dans le cas de l'indice de richesse décomposé en classes d'amplitude 10 à la Figure 2.15.

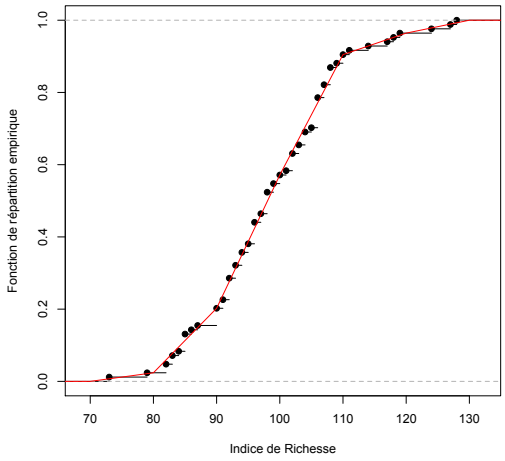


Figure 2.15: Ogives des fréquences cumulées pour les indices de richesse répartis en 6 classes et fonction de répartition empirique obtenue sur les données brutes

2.4 Exercices

1. Le tableau 2.9 décrit, en pourcentages, les résultats d'une enquête menée par le Service Public Fédéral Economie (Statistics Belgium) sur les activités sportives des hommes et femmes en Belgique.

   (a) Quelle est la variable étudiée, son type et ses modalités?

Tableau 2.9: Comparaison de l'activité sportive des belges. *(Source: Statistics Belgium)*

	Entraînement intensif (+ 4 h /semaine)	Activités légères (- 4 h /semaine)	Aucune activité sportive
Hommes	27,6%	44%	28,4%
Femmes	8,6%	54,2%	37,2%

(b) Construire deux diagrammes en secteurs afin de décrire la répartition des hommes et des femmes en fonction des modalités de la variable.

(c) En associant à chaque modalité de la variable deux barres côte à côte, une pour les hommes et une pour les femmes, construire un diagramme en barres permettant de comparer sur le même graphique les activités sportives des hommes et des femmes.

2. Le tableau 2.10 décrit les principales caractéristiques d'activité [3] de la population belge pour l'année 1999.

Tableau 2.10: Activités de la population belge. *(Source: Statistics Belgium)*

Modalités	Effectifs (en milliers)
Actif occupé	4007
Chômeur	375
Enfants (de moins de 15 ans)	1805
Non actif de 15 ans à 64 ans	2356
Non actif de 65 ans et plus	1670

(a) Représenter la distribution des fréquences à l'aide d'un diagramme en secteurs en indiquant les pourcentages sur les secteurs.

(b) Représenter la série à l'aide d'un diagramme en barres.

3. Lors d'une enquête, on a interrogé 100 travailleurs afin de connaître le nombre de personnes qu'ils avaient à charge. Les données recueillies sont les suivantes:

[3]Selon le Bureau International du Travail, est chômeuse toute personne de 15 ans et plus qui est sans travail, disponible pour travailler et à la recherche d'un emploi.

0	3	1	2	4	3	3	3	4	1	4	1	3	2	2
1	5	2	5	2	4	2	0	3	3	2	4	1	5	2
5	2	5	1	3	2	2	2	4	4	5	1	2	4	2
1	2	2	1	3	5	1	0	3	2	3	1	4	2	4
5	4	1	2	0	2	1	2	2	3	5	2	1	4	2
2	4	0	3	3	2	2	2	0	4	1	2	1	3	1
1	5	3	1	3	2	2	0	2	0					

(a) Etablir un tableau statistique de 5 colonnes: Valeurs – Effectifs – Effectifs cumulés – Fréquences – Fréquences cumulées.

(b) Représenter la répartition de la population étudiée par un diagramme adéquat basé sur les effectifs.

(c) Dessiner la courbe cumulative des fréquences cumulées.

4. Le tableau 2.11 donne la répartition des ménages d'un village en fonction du nombre de personnes constituant le ménage.

Tableau 2.11: Répartition des ménages en fonction du nombre de personnes

Valeurs	Effectifs
1	361
2	453
3	227
4	209
5	83
6	59

(a) Quelle est la variable étudiée? En déterminer son type et ses modalités. Quelle est la population considérée?

(b) Que représente l'effectif total n de la population étudiée?

(c) Combien d'habitants y a-t-il dans la ville considérée?

(d) Compléter le tableau des effectifs avec les fréquences, les effectifs et fréquences cumulés.

(e) A partir du tableau, déterminer le pourcentage de ménages constitués d'au plus 2 personnes et le pourcentage de ménages constitués d'au moins 4 personnes.

(f) Représenter la distribution des fréquences à l'aide d'un diagramme en bâtons.

(g) Dessiner la courbe cumulative des fréquences cumulées.

5. Les observations relatives à la variable "Poids" obtenues sur une population d'étudiants masculins de premier bachelier sont transcrites dans le tableau suivant.

68	70	67	75	72	71	67	65	60	60	65	65
77	95	85	70	70	72	66	75	90	65	62	70
52	60	59	65	68	71	97	65	57	75	77	75
85	56	77	67	62	52	67	72	79	60	72	69
58	55	75	75	78	65	95	65	90	72	72	60

(a) Répartir les 60 données correspondant aux étudiants masculins en les 5 classes d'amplitude constante (égale à 10 kg)

$$\mathcal{C}_1 = [50, 60], \mathcal{C}_2 =]60, 70], \mathcal{C}_3 =]70, 80],$$

$$\mathcal{C}_4 =]80, 90] \text{ et } \mathcal{C}_5 =]90, 100].$$

Etablir pour ces données groupées un tableau statistique de 5 colonnes: Classes – Effectifs – Effectifs cumulés – Fréquences – Fréquences cumulées et représenter la distribution des fréquences par un histogramme et le polygone des fréquences. Le groupement effectué vous semble-t-il adéquat?

(b) Améliorer le groupement décrit en (a) en divisant certaines classes et en en regroupant d'autres (*suggestion: un tel groupement comprend, par exemple, 6 classes d'amplitudes variables et d'unité d'amplitude égale à 5 kg).* Refaire le tableau statistique, l'histogramme et le polygone des fréquences pour cette nouvelle répartition en classes.

(c) Construire l'ogive des fréquences cumulées pour les données groupées obtenues en (b). Estimer le nombre d'étudiants dont le poids se situe entre 67.5 kg et 77.5 kg. Comparer l'estimation obtenue avec la vraie valeur. Si les étudiants sont classés par poids croissants, estimer à partir de l'ogive le poids du 30ème étudiant. Comparer cette estimation avec le vrai poids.

6. Le tableau 2.12 décrit la répartition des habitants d'une région en fonction du salaire annuel net.

(a) Compléter le tableau statistique en ajoutant les fréquences, les effectifs cumulés et les fréquences cumulées des classes.

(b) Représenter un histogramme d'aire totale unitaire et représenter le polygone correspondant.

Tableau 2.12: Répartition des salaires dans la population active d'une région

Classes (en milliers d'UM)	Effectifs
$[0, 10]$	181
$]10, 15]$	116
$]15, 20]$	89
$]20, 25]$	39
$]25, 30]$	28
$]30, 35]$	60
$]35, 50]$	91
$]50, 60]$	58
$]60, 100]$	65

(c) A partir de l'histogramme, estimer le nombre de personnes dont le salaire annuel net est inférieur à 17500 UM et estimer le nombre de personnes dont le salaire annuel net est compris entre 22500 et 37500 UM.

(d) Représenter l'ogive des fréquences cumulées. Exploiter l'ogive pour recalculer les estimations demandées au point (c).

(e) Quel salaire annuel net s vérifie $F(s) = 0.5$ où F est l'ogive représentée au point (d)? Que représente intuitivement ce salaire annuel net s?

Chapitre 3

Réduction des données

Ce chapitre décrit comment résumer à l'aide d'un petit nombre de valeurs numériques l'information contenue dans une série statistique univariée. Après une éventuelle mise en forme telle que proposée dans le Chapitre 2, nous supposerons que la série est disponible dans l'un des trois formats suivants:

- Série *brute*:
$$S = \{x_1, x_2, \ldots, x_n\}$$
où x_i est la valeur observée de la variable sur le $i$ème individu.

- Série *recensée*:
$$S = \{(x_j, n_j), j = 1, \ldots, J\}$$
avec $x_1 < \ldots < x_j$ qui désignent les valeurs distinctes observées d'effectifs $n_1, \ldots, n_J$.

- Série *groupée*:
$$S = \{(c_j, n_j) : j = 1, \ldots, J\}$$
où $c_1, \ldots, c_J$ représentent les centres des classes d'effectifs $n_1, \ldots, n_J$.

Les résumés statistiques sont des paramètres qui permettent de caractériser soit la position de l'ensemble de données sur la droite réelle (et dans certains cas, sa tendance centrale), soit la dispersion des données, soit la forme de la distribution (lorsque celle-ci est représentée par un histogramme ou un diagramme en bâtons, par exemple). De nouveau, de très nombreux ouvrages traitent de ce sujet mais le fil conducteur de ce chapitre est calqué sur celui suivi par Dehon, Droesbeke et Vermandele (2008).

3.1 Paramètres de position

3.1.1 La moyenne arithmétique

La *moyenne arithmétique* d'une série statistique univariée se note $\bar{x}$ et est égale à la somme des observations divisée par l'effectif total n de la série.

La définition de la moyenne entraîne quelques remarques:

1. Une moyenne arithmétique ne peut se calculer que si les valeurs observées sont numériques. Une série correspondant à l'étude d'une variable qualitative ne possède donc pas de moyenne arithmétique.

2. La moyenne arithmétique est indépendante de l'ordre des observations dans la série.

3. La moyenne arithmétique est rarement égale à une valeur observée. Dans le cas d'une variable discrète, la moyenne arithmétique peut même ne pas être associée à une valeur *observable* de la variable.

   Exemple 32 *Considérons un groupe de cinq individus de taille 172cm, 176cm, 179cm, 186cm et 188cm. La moyenne de ces tailles vaut 180.2 cm. Cette taille moyenne n'a pas été observée et est la taille d'un individu fictif (individu moyen) dont la seule raison d'être est de représenter un milieu.*

   Exemple 33 *Le nombre moyen de partis dans la majorité dans les communes wallones est de 1.4, ce qui n'est évidemment pas une valeur observable pour cette variable. Il faut donc interpréter ce résultat avec prudence. On peut dire que les communes concernées comprennent en moyenne plus de 1 parti, mais moins de 2.*

4. Le calcul de la moyenne arithmétique doit s'adapter à la forme dans laquelle la série statistique est fournie:

   - Si on dispose de toutes les observations particulières $S = \{x_1, \ldots, x_n\}$, $\bar{x}$ est calculée par la formule
   $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{3.1}$$
   - Si la série est donnée via le tableau des effectifs, on dispose des couples $(x_j, n_j), j = 1, \ldots, J$ et chaque valeur observée x_j apparaît dans la définition (3.1) un nombre de fois égal à son effectif n_j:
   $$\bar{x} = \frac{\sum_{j=1}^{J} n_j x_j}{n}, \tag{3.2}$$

avec $n = \sum_{j=1}^{J} n_j$. Par définition, $\frac{n_j}{n}$ est la fréquence f_j de x_j et l'égalité (3.2) peut s'écrire

$$\bar{x} = \sum_{j=1}^{J} \frac{n_j}{n} x_j = \sum_{j=1}^{J} f_j x_j.$$

- Lorsque la série est groupée en J classes d'effectifs $n_1, \ldots, n_J$, la moyenne arithmétique des données groupées ne peut se calculer exactement que si la moyenne des observations de chaque classe est connue. Notons $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_J$ les moyennes des classes. Comme, par définition, la moyenne arithmétique est égale à la somme de toutes les observations divisée par leur nombre total, il suffit de connaître la somme des observations de chaque classe pour pouvoir calculer $\bar{x}$. Or, la somme des observations de la $j$ème classe vaut $n_j \bar{x}_j$ et la somme de toutes les observations groupées est donnée par $n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_J\bar{x}_J$. En divisant cette somme par l'effectif total $n = n_1 + \ldots + n_J$, on obtient la moyenne arithmétique

$$\bar{x} = \frac{\sum_{j=1}^{J} n_j \bar{x}_j}{n}.$$

Exemple 34 *Le Tableau 2.8 de l'exemple 24 du chapitre 2 présente seulement une partie de la publication officielle du SPF Economie pour décrire les statistiques fiscales des revenus soumis à l'impôt des personnes physiques. En effet, en plus de la répartition des déclarants selon les différentes classes de revenus, les statistiques indiquent également la masse totale (c'est-à-dire la somme) des revenus déclarés par les individus de chacune des classes. Le Tableau 3.1 est le tableau complet de la statistique fiscale du Royaume pour l'exercice de 2002.*

On dispose donc directement des sommes des observations de chaque classe. Notons la masse totale des revenus de la classe c_j par m_j. Le revenu moyen $\bar{m}$ est donné par $\bar{m} = \frac{\sum_{j=1}^{24} m_j}{n} = 24691.76$.

Lorsque l'on ne dispose que des limites des classes avec les effectifs correspondants sans posséder ni les données initiales ni les moyennes des classes, il n'est plus possible de déterminer exactement la moyenne arithmétique de la série. On peut cependant en obtenir une valeur approchée, que l'on notera toujours $\bar{x}$ par abus de notation. Pour parvenir à cette approximation, on suppose à nouveau que les observations de la classe $\mathcal{C}_j$ sont uniformément réparties dans la classe. Cela permet de considérer le centre de la classe c_j comme la moyenne des valeurs observées dans cette classe. La moyenne arithmétique est donc

$$\bar{x} = \frac{\sum_{j=1}^{J} n_j c_j}{n}.$$

Tableau 3.1: Montant et répartition en pourcents du revenu total net imposable en Belgique pour l'exercice 2002

Classes		Nombres de déclarations	%	Montant des revenus	%
1	$[0,2[$	152 812	3.11	143 477 244	0.11
2	$[2,3[$	47 179	0.96	117 727 370	0.10
3	$[3,4[$	50 414	1.03	177 581 318	0.15
4	$[4,5[$	52 955	1.08	238 124 842	0.20
5	$[5,6[$	58 689	1.20	323 655 391	0.27
6	$[6,7[$	75 877	1.55	496 101 333	0.41
7	$[7,8[$	99 202	2.02	745 566 907	0.62
8	$[8,9[$	103 927	2.12	884 436 706	0.73
9	$[9,10[$	143 691	2.93	1 373 325 220	1.13
10	$[10,11[$	203 232	4.14	2 145 593 991	1.77
11	$[11,12[$	201 342	4.10	2 313 775 629	1.91
12	$[12,13[$	203 870	4.16	2 546 638 256	2.10
13	$[13,14[$	182 769	3.73	2 468 146 816	2.04
14	$[14,15[$	191 786	3.91	2 781 779 980	2.30
15	$[15,16[$	182 325	3.72	2 824 336 576	2.33
16	$[16,17[$	172 746	3.52	2 849 732 579	2.35
17	$[17,18[$	169 881	3.46	2 972 416 927	2.45
18	$[18,19[$	166 977	3.40	3 088 535 702	2.55
19	$[19,20[$	162 825	3.32	3 174 523 620	2.62
20	$[20,25[$	639 957	13.05	14 279 109 275	11.79
21	$[25,30[$	405 970	8.28	11 103 299 769	9.16
22	$[30,45[$	697 043	14.20	25 435 294 910	20.98
23	$[45,75[$	420 780	8.56	23 449 489 989	19.38
24	$[75,+\infty[$	119 374	2.42	15 195 811 336	12.54
Total		4 905 623	100	121 128 481 686	100

Exemple 35 *Revenons à l'exemple 21 du Chapitre 2 dans lequel les valeurs des indices de richesse des 84 communes liégeoises étaient explicitées. A partir de la série brute, la moyenne exacte peut être calculée. Celle-ci vaut 98.9. Cependant, au lieu d'avoir toutes les données individuelles, cette série pourrait avoir été fournie déjà groupée via le tableau statistique construit lors de l'exemple 23. Les centres des classes sont respectivement 75, 85, 92.5, 97.5, 102.5, 107.5, 115 et 125 avec les effectifs 2, 15, 15, 16, 11, 17, 5, et 3. Une valeur approchée de la moyenne arithmétique de la série est donc* $\bar{x} = \frac{2\times 75+15\times 85+\ldots+3\times 125}{84} = \frac{8277.5}{84} = 98.5$ *au lieu de 98.9 qui est la valeur précise.*

La moyenne arithmétique est le paramètre le plus utilisé pour mesurer la tendance centrale d'une série statistique. Sa popularité est principalement basée sur ses propriétés

mathématiques, détaillées ci-dessous, pour une série sous forme brute (mais les propriétés sont également valables pour les autres types de séries):

1. Si un changement d'échelle et d'origine est effectué sur les observations $x_1, \ldots, x_n$ pour obtenir la nouvelle série $S' = \{x'_1, \ldots, x'_n\}$ avec $x'_i = ax_i + b$ où a et b sont des constantes réelles, alors la moyenne arithmétique $\bar{x}'$ de la série S' est donnée par

$$\bar{x}' = a\bar{x} + b,$$

où $\bar{x}$ est la moyenne arithmétique de $S = \{x_1, \ldots, x_n\}$.

Preuve: $\bar{x}' = \frac{\sum_{i=1}^n x'_i}{n} = \frac{1}{n}\sum_{i=1}^n (ax_i + b) = \frac{1}{n}\left(a\sum_{i=1}^n x_i + nb\right) = a\bar{x} + b.$ □

2. Définissons la *série des valeurs centrées* S_c en prenant les différences entre les valeurs observées de la série $S = \{x_1, \ldots, x_n\}$ et la moyenne arithmétique de la série S. Cette nouvelle série $S_c = \{x_1 - \bar{x}, \ldots, x_n - \bar{x}\}$ a une moyenne nulle.

Preuve: Il s'agit d'un cas particulier de la propriété 1 en prenant $a = 1$ et $b = -\bar{x}$. On obtient directement $\bar{x}_c = \bar{x} - \bar{x}$. □

Une autre façon d'interpréter cette propriété est de dire que la somme des valeurs centrées positives (c'est-à-dire $x_i - \bar{x} \geq 0$) est compensée par la somme des valeurs centrées négatives (c'est-à-dire $x_i - \bar{x} \leq 0$).

3. La moyenne arithmétique a une signification concrète. Par la propriété précédente, on sait que les éléments de la série situés à gauche de $\bar{x}$ sont compensés par les éléments situés à droite de $\bar{x}$.

Cette situation peut être représentée physiquement par une droite graduée sur laquelle les différentes valeurs rencontrées dans la série sont reportées. Des masses proportionnelles aux effectifs de ces valeurs sont ensuite placés sur la droite. La moyenne arithmétique coïncide avec le point d'équilibre des masses ainsi placées. On dit que la moyenne est le *centre de gravité de la série.*

Exemple 36 *Considérons la série suivante:* $S = \{2, 3, 4, 4, 4, 5, 5, 6, 7, 7\}$ *(voir Figure 3.1) dont la moyenne arithmétique est égale à* $\bar{x} = 4.7$.

Lorque la balance est placée à l'emplacement de la moyenne, la balance est en équilibre.

4. La somme des carrés des écarts des éléments d'une série par rapport à la moyenne arithmétique de la série est inférieure ou égale à la somme des carrés des écarts par rapport à toute autre valeur $a \in \mathbb{R}$. Cette propriété s'écrit aussi

$$\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - a)^2 \ \forall\ a \in \mathbb{R}.$$

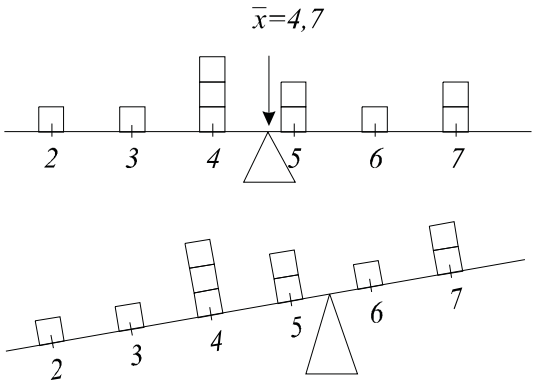


Figure 3.1: La moyenne arithmétique comme centre de gravité d'une série

Preuve: Il suffit de rechercher le minimum sur $\mathbb{R}$ de la fonction $f(x) = \sum_{i=1}^{n}(x_i - x)^2$.

□

5. Si les individus d'une population P sont répartis en k sous-populations $P_1, \ldots, P_k$ d'effectifs $n_1, \ldots, n_k$ (avec $n_1 + \ldots + n_k = n$), alors la moyenne arithmétique globale d'une variable X étudiée sur la population complète peut être calculée à partir des moyennes $\bar{x}_1, \ldots, \bar{x}_k$ des sous-populations par la formule suivante:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i} = \sum_{i=1}^{k} p_i \bar{x}_i,$$

où p_i est la proportion d'individus de la population P appartenant à la sous-population P_i.

Calculer la moyenne arithmétique n'est pas toujours la meilleure stratégie. En effet, par définition de la moyenne arithmétique, on attribue à chaque observation un poids égal à $\frac{1}{n}$:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n} 1 \times x_i}{\sum_{i=1}^{n} 1}.$$

Comme chaque observation a le même poids sur le calcul de $\bar{x}$, une valeur beaucoup plus petite ou beaucoup plus grande que toutes les autres observations va fortement influencer la moyenne arithmétique. On dit que la moyenne est influencée par les *valeurs extrêmes* de la variable car elles ont tendance à attirer la moyenne vers elles. A la limite, une seule observation tendant vers l'infini peut emmener la moyenne avec elle.

Les valeurs extrêmes peuvent être dues à des erreurs de mesure ou d'encodage (on parle alors d'*erreurs grossières* ou de valeurs aberrantes) ou sont le reflet de faits exceptionnels intéressants à analyser plus en détails. Le traitement des valeurs aberrantes est le sujet de la *statistique robuste*. Pour diminuer l'effet de telles observations sur la moyenne arithmétique, des versions un peu modifiées de la moyenne ont été introduites:

- La moyenne arithmétique pondérée: elle permet d'associer aux différentes valeurs de la série statistique des poids différents. A chaque observation x_i de la série $S = \{x_1, \ldots, x_n\}$, on attribue un poids w_i, positif ou nul, permettant d'indiquer son importance relative par rapport aux autres observations. La *moyenne arithmétique pondérée* par les poids w_i est définie par

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}. \tag{3.3}$$

  En plus de leur attrait lorsque des valeurs aberrantes sont présentes dans un ensemble de données, les moyennes arithmétiques pondérées sont utiles dans beaucoup de domaines où il s'avère important de relativiser l'importance des observations. Citons le cas bien connu du calcul de la moyenne des résultats d'examens, calcul dans lequel le résultat de chaque cours est pondéré par le poids (c'est-à-dire le nombre de crédits) qui lui est associé. De même, ce type de moyenne est très souvent exploité en finance pour calculer des taux moyens, des échéances moyennes, ...

  Exemple 37 *Une personne désire calculer le taux de change moyen correspondant à la situation suivante: elle a d'abord acheté q_1 dollars au taux de t_1 euros le dollar; ensuite q_2 dollars au taux de t_2 euros le dollar;...; et enfin, q_n dollars au taux de t_n euros le dollar. En tout, cette personne s'est procurée $Q = \sum_{i=1}^n q_i$ dollars pour une dépense totale égale à $D = \sum_{i=1}^n q_i t_i$ euros. Le taux de change moyen t est le coût unitaire du dollar tel qu'une quantité Q de dollars puisse est obtenue en dépensant D euros:*

$$D = t \times Q \Rightarrow t = \frac{D}{Q} = \frac{\sum_{i=1}^n q_i t_i}{\sum_{i=1}^n q_i}.$$

  Le taux moyen est donc la moyenne arithmétique pondérée des taux t_i où les poids sont donnés par les quantités achetées.

- Les moyennes tronquées: la *moyenne tronquée au seuil* α, notée $\bar{x}_\alpha$, avec $0 \leq \alpha < \frac{1}{2}$ est un cas particulier de moyenne arithmétique pondérée. Le calcul de $\bar{x}_\alpha$ consiste à attribuer un poids nul aux αn plus petites observations ainsi qu'aux αn plus grandes observations, où n est la taille de la population ou de l'échantillon. Le paramètre α étant une proportion, le produit αn n'est pas nécessairement un nombre entier. Dans ce cas, on attribuera un poids nul aux $[\alpha n]$ plus petites et plus grandes observations, où $[k]$ désigne le plus grand entier inférieur ou égal à k.

  Si la série ordonnée est constituée des observations $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, la moyenne tronquée au seuil α est donnée par

$$\bar{x}_\alpha = \frac{\sum_{i=[\alpha n]+1}^{n-[\alpha n]} x_{(i)}}{n - 2[\alpha n]}.$$

Remarque: lorsque le paramètre α est égal à 0, $\bar{x}_\alpha$ coïncide avec la moyenne arithmétique simple.

3.1.2 La médiane

L'introduction du concept de médiane a pour objectif essentiel de caractériser une série par une valeur qui se trouve au milieu des observations lorsque celles-ci sont rangées par valeurs croissantes. En d'autres termes, la médiane est une valeur telle que le nombre d'observations lui étant inférieures soit à peu près égal au nombre d'observations qui lui sont supérieures. La médiane ne s'applique que lorsque les observations peuvent être ordonnées de la plus petite à la plus grande. Elle concerne donc des variables qui sont mesurées sur une échelle au moins ordinale et ne convient pas pour des variables qualitatives mesurées sur une échelle nominale.

La définition intuitive qui vient d'être donnée, c'est-à-dire que la médiane partage les observations de la série ordonnée en deux groupes d'effectifs (à peu près) égaux à $\frac{n}{2}$ (où n désigne l'effectif de la série), doit être précisée afin de pouvoir calculer la médiane d'une série statistique (médiane notée $\tilde{x}$ dans la suite). Pour cela, il faut tenir compte de la forme dans laquelle la série est fournie:

1. Si on dispose des données brutes et que celles-ci sont distinctes deux à deux, la première opération consiste à trier les observations pour obtenir $\tilde{S} = \{x_{(1)}, \ldots, x_{(n)}\}$. La détermination de la médiane dépend alors de la parité de l'effectif n de la série:

   (a) Si n est impair ($n = 2k + 1$), alors la médiane correspond à l'observation de rang $k + 1$: $\tilde{x} = x_{(k+1)}$.

   (b) Si n est pair ($n = 2k$), alors toute valeur située entre l'observation de rang $\frac{n}{2} = k$ et l'observation de rang $k+1$ vérifie la propriété caractérisant la médiane. On dit que ces deux observations définissent un *intervalle médian*. Cependant, un intervalle est moins aisé à manipuler qu'une valeur. C'est pourquoi, dans le cas d'une variable quantitative, on suit généralement la convention de définir la médiane par la moyenne arithmétique des deux observations qui délimitent l'intervalle médian:
   $$\tilde{x} = \frac{x_{(k)} + x_{(k+1)}}{2}.$$
   Avec cette convention, la médiane est unique.

Exemple 38 *La médiane de la série des cinq tailles considérée à l'exemple 32 est donnée par l'observation du milieu, à savoir* $\tilde{x} = 179cm$.

2. Si la série est donnée par l'ensemble des couples $(x_i, n_i), i = 1, \ldots, J$ où $x_1 < x_2 < \ldots < x_J$ sont les valeurs distinctes observées avec les effectifs $n_1, \ldots, n_J$, la détermination de la médiane peut se faire à partir des distributions des effectifs cumulés ou des fréquences cumulées ou à partir des courbes cumulatives correspondantes.

   (a) S'il existe une valeur x_j telle que $N_{j-1} < \frac{n}{2} < N_j$ (resp. $F_{j-1} < \frac{1}{2} < F_j$), alors $\tilde{x} = x_j$.

   (b) S'il existe une valeur x_j telle que $N_j = \frac{n}{2}$ (resp. $F_j = \frac{1}{2}$), alors $\tilde{x} = \frac{x_j + x_{j+1}}{2}$.

   Ces deux cas sont représentés sur respectivement le premier et le deuxième dessin de la Figure 3.2.

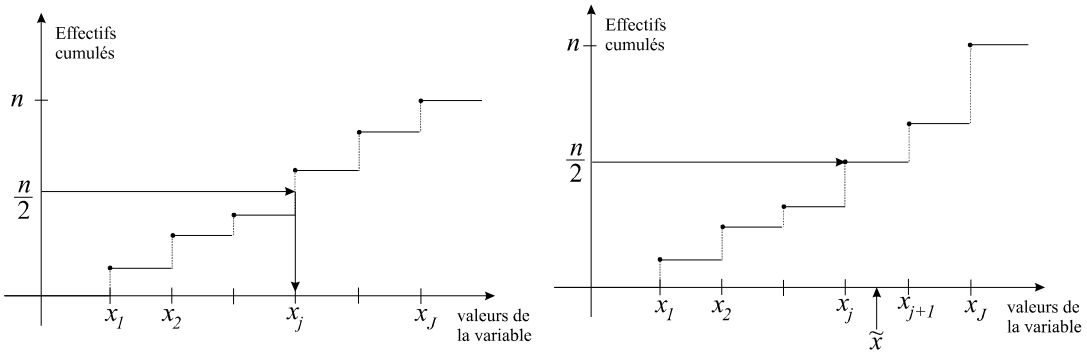


Figure 3.2: Détermination de la médiane à partir de la courbe cumulative $y = N(x)$

Exemple 39 *A partir du tableau statistique établi lors de l'exemple 15 ou à partir de la courbe cumulative des fréquences représentée à l'exemple 18, on constate que le nombre médian de partis dans les majorités communales en Wallonie vaut 1. En effet,* $N_1 = 181 > \frac{262}{2}$ *(ou* $F_1 = 0.69 > 0.5$*).*

Exemple 40 *Si on retourne aux indices de richesse calculés sur les communes de la province de Liège, le tableau ordonné décrit à l'exemple 21 permet de déterminer la médiane. Il y a un nombre pair d'observations (*$n = 84$*) et les observations de rang 42 et 43 sont toutes les deux égales à 98, ce qui donne* $\tilde{x} = 98$.

3. Si l'on dispose d'une série groupée en J classes $\mathcal{C}_1, \ldots, \mathcal{C}_J$ sans plus avoir la série ordonnée qui a guidé la répartition en classes, alors la médiane ne peut pas être déterminée avec précision. On peut cependant obtenir une valeur approchée de ce paramètre à partir de l'ogive des effectifs ou fréquences cumulés. En effet, en supposant que les observations sont uniformément réparties au sein de chaque classe, l'ogive des effectifs cumulés (resp. des fréquences cumulées), d'équation $y = N(x)$

(resp. $y = F(x)$) permet d'estimer le nombre (resp. la proportion) d'observations de la série inférieures ou égales à x. Par construction, dans la plupart des cas, l'ogive est continue et strictement croissante[1] sur l'intervalle $[e_0, e_J]$, ce qui permet de définir la médiane $\tilde{x}$ comme l'unique solution de l'équation:

$$N(x) = \frac{n}{2} \text{ ou } F(x) = \frac{1}{2} \tag{3.4}$$

Plus précisément, notons $\mathcal{C}_m$ la classe médiane (c'est-à-dire la classe qui contient la médiane, ou encore, la première classe dont l'effectif cumulé est supérieur ou égal[2] à $\frac{n}{2}$) et prenons les notations, assez naturelles suivantes:

e_{m-1} : borne inférieure de $\mathcal{C}_m$
e_m : borne supérieure de $\mathcal{C}_m$
a_m : amplitude de $\mathcal{C}_m$
n_m : effectif de $\mathcal{C}_m$
N_{m-1} : effectif cumulé de la classe précédant $\mathcal{C}_m$

La médiane est définie par l'abscisse $\tilde{x}$ correspondant à l'ordonnée $\frac{n}{2}$ sur l'ogive $y = N(x)$. Elle s'obtient donc en remplaçant, dans l'équation de l'ogive, y par $\frac{n}{2}$. Or, pour tout $x \in \mathcal{C}_m$, l'ogive est estimée par le segment de droite joignant les deux points (e_{m-1}, N_{m-1}) et (e_m, N_m). D'où,

$$N(x) = N_{m-1} + \frac{n_m}{a_m}(x - e_{m-1}),\ \forall\ x \in c_m.$$

En isolant x et en remplaçant $N(x)$ par $\frac{n}{2}$, on obtient

$$\tilde{x} = e_{m-1} + a_m \frac{\frac{n}{2} - N_{m-1}}{n_m}. \tag{3.5}$$

Les effectifs et effectifs cumulés peuvent être remplacés par les fréquences correspondantes dans l'expression 3.5.

Exemple 41 *Si on ne dispose pas des indices de richesse individuels pour les communes de la province de Liège, la médiane peut être estimée à partir de la distribution des effectifs cumulés correspondant à la répartition en les 8 classes d'amplitude 10 ou*

[1] L'ogive est strictement croissante uniquement si toutes les classes ont des effectifs non nuls. En cas de séries bimodales (voir la notion de mode plus loin), il se pourrait que certaines classes situées entre les deux modes soient vides et, au cas où le palier horizontal correspondrait exactement à la masse 1/2, le calcul de la médiane tel qu'expliqué devrait être adapté.

[2] Si les bornes supérieures des classes sont exclues des classes, la classe médiane est la première classe dont l'effectif cumulé dépasse strictement $n/2$.

5 décrite à l'exemple 23. La classe médiane est la classe $\mathcal{C}_4 =]95; 100]$ *puisque son effectif cumulé vaut 48 (*$> n/2 = 42$*) alors que l'effectif cumulé de la classe précédente est 32. La valeur approchée de la médiane à partir de la formule (3.5) est donc*

$$\tilde{x} = 95 + 5 \times \frac{48 - 32}{16} = 100.$$

Exemple 42 *Le revenu médian des déclarants belges peut être facilement estimé à partir du Tableau 3.1 de l'exemple 34 si l'on ajoute une colonne pour les effectifs ou pour les pourcentages cumulés. En se basant sur les pourcentages cumulés, on constate que* $F_{17} \times 100 = 46.74 < 50 < 50.15 = F_{18} \times 100$*. La classe médiane est donc* $c_{18} = [18, 19[$*. Par interpolation linéaire à l'intérieur de cette classe, on obtient une valeur approchée de la médiane:*

$$\tilde{x} = 18 + 1 \times \frac{0.5 - 0.467}{0,034} = 18.958.$$

Sachant que l'unité exploitée dans la définition des classes est égale à 1 000 Euro, le revenu médian vaut 18 958 Euro et diffère donc significativement du revenu moyen.

Propriétés de la médiane

1. La médiane a une interprétation simple: environ la moitié des observations la précède et l'autre moitié la suit. Dans le cas des variables discrètes, la décomposition de la série en deux parties (avant et après la médiane) n'est pas toujours bien équilibrée puisqu'il y a des sauts dans la distribution des effectifs ou des fréquences. Comme certaines observations peuvent coïncider (notamment au centre de la distribution), il est plus correct de dire qu'**au moins** 50% des observations ont une valeur inférieure ou égale à la médiane et **au moins** 50% des observations ont une valeur supérieure ou égale.

   Exemple 43 *Le nombre médian de partis dans les majorités communales vaut 1. Seules 30% des communes ont un nombre de partis supérieur à 1, tandis que 69% ont un nombre de partis égal à la médiane.*

   Dans le cas continu, la médiane a une interprétation intéressante à partir de l'histogramme. Rappelons que lorsque l'histogramme recouvre une aire unitaire, l'ordonnée $F(x)$ de l'ogive des fréquences cumulées est égale à l'aire située sous l'histogramme à gauche de l'abscisse x. Comme la médiane vérife $F(\tilde{x}) = \frac{1}{2}$, l'abscisse $\tilde{x}$ coupe l'histogramme en deux parties d'aires égales.

2. Si un changement d'échelle et d'origine est effectué sur les observations $x_1, \ldots, x_n$ pour obtenir la nouvelle série $S' = \{x'_1, \ldots, x'_n\}$ avec $x'_i = ax_i + b$ où a et b sont des constantes réelles, alors la médiane $\tilde{x}'$ de la série S' est donnée par

$$\tilde{x}' = a\tilde{x} + b,$$

où $\bar{x}$ est la médiane de $S = \{x_1, \ldots, x_n\}$.

Preuve: La transformation va soit conserver soit inverser l'ordre des observations de départ. La médiane étant basée sur les observations centrales sera définie à partir des versions transformées de ces observations centrales. □

3. La médiane ne dépend pas directement des valeurs de tous les éléments de la série puisque, dans son calcul, les différents éléments n'interviennent que par leur ordre et non par leur valeur. Si on diminue la valeur d'une observation inférieure à la médiane ou si on augmente la valeur d'une observation qui lui est supérieure, la médiane ne change pas. Cette propriété explique la robustesse de la médiane vis-à-vis des observations aberrantes. Ce paramètre est même le paramètre de tendance centrale le plus robuste face à la contamination.

4. La somme des écarts absolus des éléments d'une série par rapport à la médiane de la série est inférieure ou égale à la somme des écarts absolus par rapport à toute autre valeur $a \in \mathbb{R}$. Cette propriété s'écrit aussi

$$\sum_{i=1}^{n} |x_i - \tilde{x}| \leq \sum_{i=1}^{n} |x_i - a| \ \forall \ a \in \mathbb{R}. \tag{3.6}$$

Preuve: Il suffit de rechercher le minimum sur $\mathbb{R}$ de la fonction $f(x) = \sum_{i=1}^{n} |x_i - x|$.

□

3.1.3 Les quantiles

La médiane est un paramètre de tendance centrale. Elle peut cependant être aussi vue comme un cas particulier d'un paramètre de position plus général appelé *quantile*. Alors que la médiane consiste à diviser une série ordonnée en deux parties d'effectifs à peu près égaux à $\frac{n}{2}$, les quantiles d'ordre p partagent cette série en p sous-ensembles contenant approximativement un nombre d'observations égal à $\frac{n}{p}$.

Les quantiles les plus fréquemment utilisés en pratique sont

- les *quartiles* qui divisent une série en quatre parties égales. Le deuxième quartile coïncide avec la médiane. Les deux autres quartiles sont habituellement notés Q_1 et Q_3.

- les *déciles* qui divisent une série en 10 parties égales.
- les *centiles* qui divisent une série en 100 parties égales.

Les quantiles ne sont utiles que pour une variable quantitative. Comme l'idée consiste à diviser une série d'observations en un nombre quelconque de parties, une division fine n'est envisageable que lorsque le nombre d'observations est relativement élevé. Les quartiles peuvent être généralement utilisés pour toutes les séries. Le calcul des quantiles se fait exactement de la même manière que ce qui a déjà été décrit pour la médiane à part qu'il faut raisonner, pour chaque quantile, avec une fraction α au lieu de $\frac{1}{2}$, α correspondant à la masse située à gauche du quantile dans la série ordonnée.

Exemple 44 *A partir de la distribution des fréquences de la série des revenus déclarés en 2002 en Belgique, on peut déterminer que la classe* $\mathcal{C}_{12} = [12; 13[$ *est celle contenant le premier quartile (première classe dont la fréquence cumulée est supérieure ou égale à 0.25) tandis que le troisième quartile se trouve dans la classe* $\mathcal{C}_{22} = [30; 45[$*. D'où, par interpolation linéaire au sein de chacune de ces classes, on obtient*

$$Q_1 = 12 + 1 \times \frac{0.25 - 0.242}{0.0416} = 12.182 \text{ et } Q_3 = 30 + 15 \times \frac{0.75 - 0.748}{0.142} = 30.222,$$

exprimés en unité égale à 1 000 Euro. Ces informations sur la distribution des revenus indiquent, par exemple, qu'approximativement 25% des contribuables déclarent un revenu inférieur à 12 182 Euro, tandis que les 25% des contribuables les mieux rémunérés déclarent un revenu supérieur à 30 222 Euro.

Comme autre exemple, intéressons-nous aux premier et dernier déciles de la distribution des revenus. Dans la pratique, on parle aussi des quantiles 10% et 90% (le pourcentage faisant référence au pourcentage d'individus ayant une valeur de la variable inférieure au quantile en question). Le pourcentage cumulé de la classe $[7, 8[$ *est égal à 10.95 (celui de la classe précédente vaut 8.93%). On a donc*

$$D_1 = 7 + 1 \times \frac{10 - 8.93}{2.02} = 7.52$$

ce qui permet de dire que le revenu maximal des individus se trouvant parmi les 10% les plus pauvres est de 7 520 Euro. Le 9ème décile se trouve dans la classe $[45, 75[$ *(dont le pourcentage cumulé vaut 97.6%). On a donc*

$$D_9 = 45 + 30 \times \frac{90 - 89.02}{8.56} = 48.43$$

Les 10% déclarants les plus riches déclarent un revenu de minimum 48 430 Euro.

Savoir de plus que les plus riches déclarants ont un revenu supérieur à 75 000 Euro (qui est la borme inférieure de la dernière classe) permet de constater que la dispersion des revenus parmi les pauvres est beaucoup moins forte que la dispersion des revenus parmi les riches. On reviendra sur ces constatations dans les sections suivantes.

L'ensemble des valeurs $x_{(1)}, Q_1, \tilde{x}, Q_3$ et $x_{(n)}$ donne un bon résumé de la série statistique étudiée, comme on le verra plus en détail dans le prochain chapitre.

3.1.4 Le mode

Ce paramètre de position est, à première vue, le plus simple. Il est défini directement à partir de la série de données et n'implique pas d'opérations algébriques. Il s'applique à tous les types de variables et répond à un objectif bien précis: déterminer la modalité ou la valeur observée de la variable qui apparaît le plus souvent dans la série.

Cette valeur ou modalité la plus fréquente définit le *mode* et est notée x_M. Ce paramètre est souvent simple à obtenir dès qu'on dispose d'un diagrammes en bâtons ou en barres représentant la série: x_M correspond à l'effectif le plus élevé et donc au plus haut bâton ou la plus haute barre.

Exemple 45 *La Figure 3.3 reprend les diagrammes en barres et en bâtons représentés lors des exemples 13 et 16 du chapitre 2. Le premier graphique caractérise la distribution des partis des bourgmestres du Brabant wallon alors que le second graphique décrit les nombres de partis dans les majorités communales en Wallonie.*

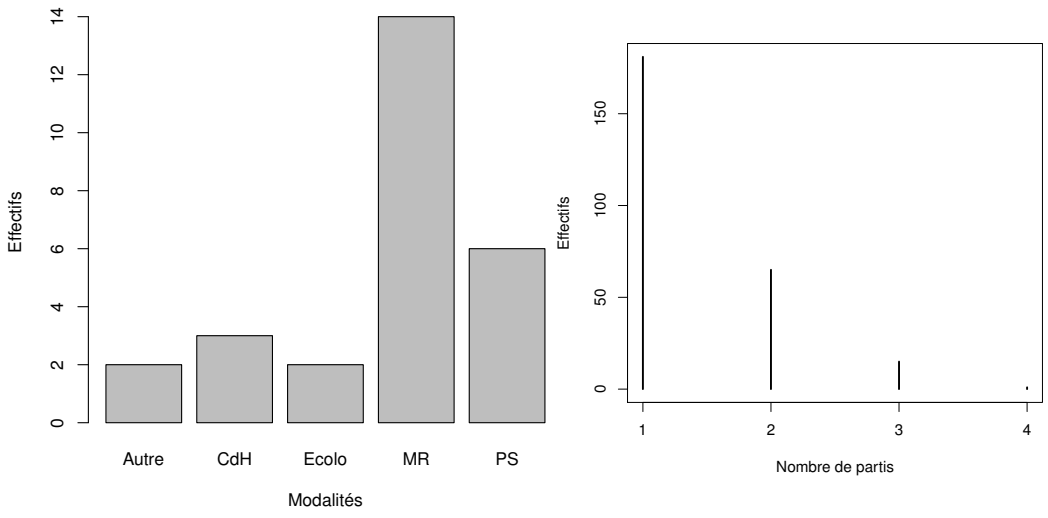


Figure 3.3: Diagramme en barres pour la variable `PartiBourgmestre` observée sur l'ensemble des communes de la province du Brabant Wallon et diagramme en bâtons pour la variable `NombrePartis` mesurée à l'échelle de la Wallonie.

On voit directement à partir de ces diagrammes que le mode de la variable qualitative `PartiBourgmestre`*, sur la province du Brabant wallon, est la modalité* `MR` *et le mode de la variable discrète* `NombrePartis` *est la valeur 1.*

Dehon, Droesbeke et Vermandele (2008) signalent cependant que l'utilisation du mode n'est pas toujours évidente. Ils émettent à ce propos les remarques suivantes:

1. Les diagrammes en bâtons de la Figure 3.4 illustrent le fait que ce paramètre n'est pas nécessairement unique et qu'il peut même ne pas exister.

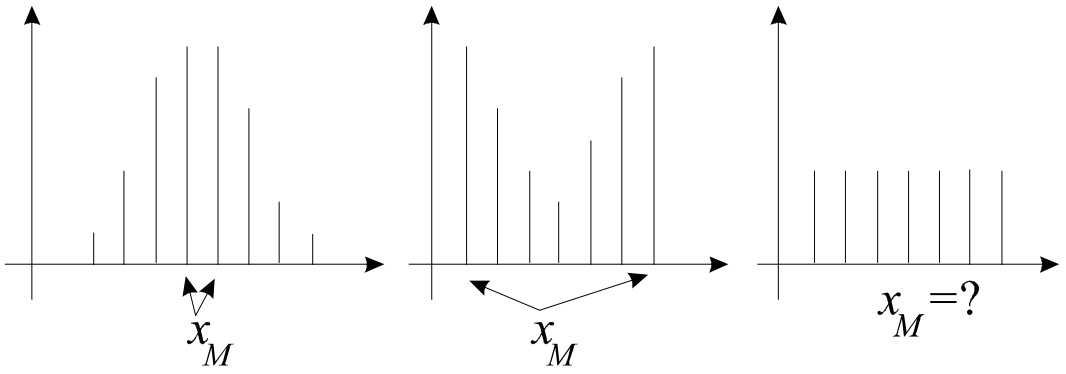


Figure 3.4: Où se trouve le mode?

2. Il se peut aussi que la série présente deux valeurs x_j et x_k telles que $n_j > n_{j-1}$ et $n_j > n_{j+1}$ ainsi que $n_k > n_{k-1}$ et $n_k > n_{k+1}$. On appelle ces valeurs des *modes relatifs* et celui correspondant à l'effectif le plus élevé est dit *mode absolu*. On distingue alors:

   - Les séries *unimodales* qui ne possèdent qu'un seul mode. Si le diagramme en bâtons est en "forme de cloche", ce mode unique peut constituer une valeur centrale de la série.
   - Les séries *plurimodales* qui possèdent plusieurs modes. Ce type de séries s'obtient notamment lorsque des séries unimodales (de tendances centrales différentes) sont mélangées.

3. Lorsque les données sont groupées, le concept et la détermination du mode deviennent encore moins précis.

   - Si toutes les classes ont la même amplitude, on peut définir une classe modale comme étant celle dont l'effectif associé est le plus élevé. De nouveau, l'existence et l'unicité d'une telle notion ne sont pas garanties.
   - Si les classes ont des amplitudes différentes, on ne peut plus simplement comparer les effectifs car les grandes classes sont favorisées. Il faut plutôt comparer les hauteurs des rectangles de l'histogramme d'aire unitaire puisque ceux-ci sont construits en ajustant les fréquences en fonction des amplitudes des classes.

L'inconvénient de cette définition est que le mode dépend très fort de la répartition en classes effectuée.

Exemple 46 *Pour la distribution des indices de richesse des communes de la province de Liège, la classe modale correspondant à la répartition en 6 classes illustrée par l'histogramme de la Figure 2.8 est la classe* $]90; 100]$ *alors qu'en exploitant l'autre répartition (histogramme de la Figure 2.9), on obtient* $]105; 110]$ *comme classe modale.*

Exemple 47 *L'histogramme 2.10 obtenu dans l'exemple 27 permet de comparer les fréquences ajustées en fonction des amplitudes des classes de la répartition des revenus en 2002. La classe modale est la classe* $\mathcal{C}_{12} = [12, 13[$.

Essayer de définir une valeur approchée du mode à partir de données groupées comme on l'a fait pour les autres paramètres de position n'est pas très satisfaisant. Une approximation acceptable pour des séries relatives à l'observation d'une variable statistique continue dont le polygone des effectifs est en forme de cloche (plus ou moins symétrique) a été introduite par Yule et Kendall. Pour ce type de séries statistiques, ces deux statisticiens ont constaté une relation empirique entre la moyenne $\bar{x}$, la médiane $\tilde{x}$ et le mode x_M :

$$\bar{x} - x_M \approx 3(\bar{x} - \tilde{x}).$$

Cette relation peut être exploitée pour estimer le mode lorsqu'on connaît (même approximativement) $\bar{x}$ et $\tilde{x}$:

$$x_M \approx 3\tilde{x} - 2\bar{x}. \tag{3.7}$$

Exemple 48 *La répartition des indices de richesse de la province de Liège en six classes d'amplitude constante mène à un polygone plus ou moins symétrique et en forme de cloche. Dans les exemples 35 et 41, les estimations* $\bar{x} = 98.5$ *et* $\tilde{x} = 100$ *ont été calculées. Une estimation du mode est donc* $x_M = 103$*, valeur qui n'appartient pas à la classe modale* $]90, 100]$. *Avec les données individuelles à notre disposition, on peut vérifier que la valeur la plus souvent observée est en réalité 107.*

Exemple 49 *Le polygone des fréquences décrivant la distribution des revenus est de nouveau une courbe en forme de cloche, mais celle-ci s'écarte fortement de la symétrie. Si on utilise l'approximation (3.7) à partir des estimations* $\bar{x} = 24691.76$ *(exemple 34) et* $\tilde{x} = 18958$ *(exemple 42), on obtient* $x_M = 7492.95$*. Cette valeur est très éloignée de la classe modale* $[12000; 13000[$.

3.1.5 Choix d'un paramètre de tendance centrale

Dans certains cas, l'utilisation de l'un ou l'autre paramètre est suggérée par le contexte de l'étude statistique. Comparons, par exemple, les valeurs des différents paramètres $\bar{x}, \tilde{x}$ et x_M obtenus pour la variable `NombrePartis` mesurée sur l'ensemble de la Wallonie, la variable `IndiceRichesse` considérée sur la province de Liège et le revenu des déclarants belges pour l'exercice 2002.

Exemple 50 *Pour la variable discrète* `NombrePartis`*, tous les paramètres sont fort proches:* $\bar{x} = 1.4$ *et* $\tilde{x} = x_M = 1$*. Cependant, les interprétations diffèrent. Dans le cas de la médiane, l'interprétation n'est pas très naturelle puisqu'on observe un saut important dans la distribution des effectifs à la valeur 1. Dans le cas de la moyenne, on obtient une valeur non observable de la variable. Par contre, le mode est univoquement déterminé et caractérise le type de coalition le plus fréquent en Wallonie, à savoir un seul parti dans la majorité.*

Exemple 51 *Pour la variable* `IndiceRichesse`*, les trois paramètres (calculés sur la série brute) sont donnés par* $\bar{x} = 98.9$*,* $\tilde{x} = 98$ *et* $x_M = 107$*. Le mode est moins adéquat dans ce contexte continu car le diagramme en bâtons présente plusieurs modes. Par contre, si on reporte les valeurs de la moyenne et de la médiane sur l'un ou l'autre des histogrammes construits, on voit qu'elles décrivent bien la tendance centrale de la distribution. Vu les belles propriétés mathématiques de la moyenne arithmétique, ce paramètre sera dans ce cas préféré.*

Exemple 52 *Pour la distribution des revenus, le mode estimé n'est pas très cohérent par rapport à la classe modale. Ce paramètre est donc à éviter pour ce type de distributions. La moyenne arithmétique est elle fortement influencée par le petit nombre de déclarants dont le revenu est très élevé. La mesure de tendance centrale la plus fiable est donc sans conteste la médiane.*

En toute généralité, on peut dire que chacun de ces indicateurs de tendance centrale est sensible à certains aspects de la distribution. Leurs valeurs sont donc souvent différentes. Par définition,

- La moyenne prend en compte la valeur de chaque observation d'une série.
- Le mode indique une seule valeur de la série, celle qui est observée le plus souvent.
- La médiane indique simplement un rang.

Empiriquement, on constate les relations suivantes entre les trois paramètres:

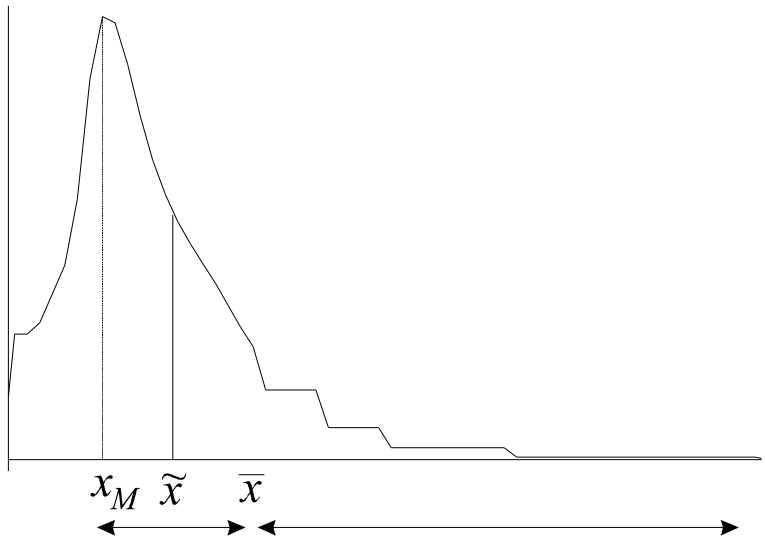


Figure 3.5: Etalement vers la droite

- Si la série est parfaitement symétrique autour d'un paramètre de tendance centrale, $\bar{x} = \tilde{x} = x_M$. Le centre de gravité de la série est situé là où l'effectif est maximum. De plus, en cet endroit, on trouve autant d'observations à gauche qu'à droite.

- Si la série est unimodale et modérément asymétrique, on observe le plus souvent soit $x_M \leq \tilde{x} \leq \bar{x}$ (on parle alors d'étalement vers la droite) ou $\bar{x} \leq \tilde{x} \leq x_M$ (étalement vers la gauche). Par exemple, le polygone des fréquences d'une série présentant un étalement vers la droite est tracé à la Figure 3.5. Le mode x_M est la valeur la plus à gauche et correspond au pic du polygone. Comme $\tilde{x}$ se trouve après x_M, cela signifie qu'au pic, la moitié des observations n'a pas encore été rencontrée. Finalement, pour compenser l'ensemble des écarts négatifs entre x_M et $\bar{x}$, des valeurs relativement grandes de la variable doivent être observées. Cette dernière condition explique l'étalement vers la droite.

Ce qu'il faut surtout garder à l'esprit est que les paramètres de tendance centrale donnent des valeurs représentatives de la série mais ne suffisent pas à la caractériser complètement. D'autres informations utiles vont être introduites dans la section suivante.

3.2 Les paramètres de dispersion

Dans cette section, seules les variables quantitatives sont considérées (car la notion d'écart entre observations est nécessaire pour mesurer la dispersion d'un ensemble de données). Une première manière de cerner le concept de dispersion consiste à ordonner les observations en construisant $\tilde{S} = \{x_{(1)}, \ldots, x_{(n)}\}$.

3.2.1 L'étendue

L'étendue de la série $S = \{x_1, \ldots, x_n\}$ est la différence entre la plus grande et la plus petite des valeurs de la série:

$$E = x_{(n)} - x_{(1)}.$$

L'étendue est le paramètre de dispersion le plus simple mais il présente l'inconvénient de ne pas tenir compte de toutes les observations et d'être particulièrement sensible à la présence de valeurs extrêmes. De plus, ce paramètre ne peut pas être calculé avec exactitude lorsqu'on ne dispose que de données groupées.

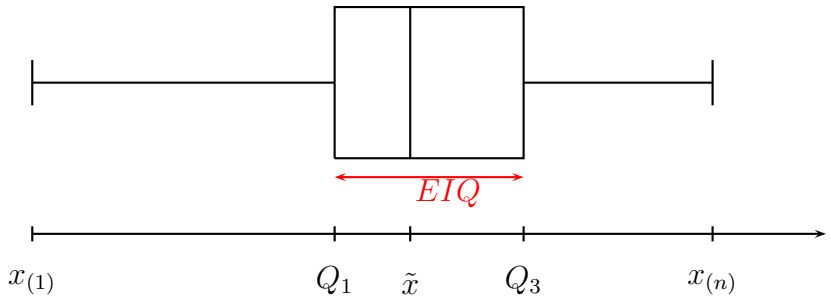
3.2.2 L'écart interquartile et les boîtes à moustaches

1. L'écart-interquartile: les quartiles $Q_1, \tilde{x}$ et Q_3 ont été définis comme les valeurs de la série ordonnée partageant l'effectif total en quatre parties à peu près égales. Si on considère l'intervalle $[Q_1, Q_3]$, on devrait y retrouver approximativement 50% des observations. L'écart-interquartile EIQ est défini par la longueur de cet intervalle:

   $$EIQ = Q_3 - Q_1.$$

   Sa définition est donc similaire à celle de l'étendue mais permet de réduire l'influence des valeurs extrêmes prises par la variable puisqu'il coïncide avec l'étendue de la série tronquée des 25% des observations les plus petites et les plus grandes.

2. Les boîtes à moustaches: Il est possible de résumer l'information fournie par les quartiles et les intervalles qui les séparent en construisant une *boîte à moustaches* ou diagramme en boîte (en anglais, *box-and-wisker plot*, Tukey, 1977). Son mode de construction est le suivant: on englobe les 50% d'observations centrales dans une boîte dont la longueur est égale à l'écart-interquartile et coupée en deux parties (de longueurs habituellement inégales) par la médiane $\tilde{x}$ de la série. Cette boîte est ensuite prolongée à sa gauche et à sa droite par deux moustaches jusqu'à $x_{(1)}$ et $x_{(n)}$:



Une boîte à moustaches indique de façon claire quelques traits marquants de la série observée:

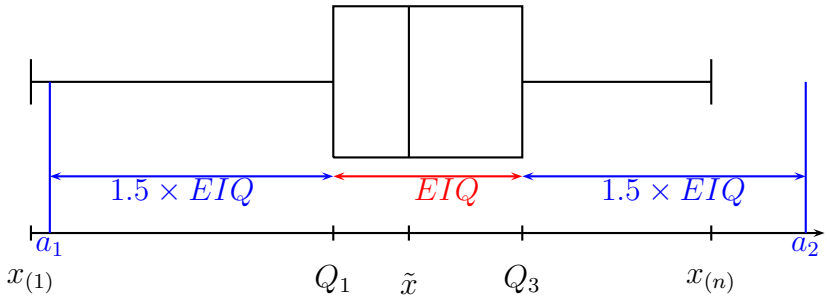
- La médiane renseigne sur la valeur centrale de la série.
- Les longueurs des deux parties de la boîte rendent compte de la dispersion et de la symétrie des valeurs situées au centre de la série.
- Les longueurs des moustaches indiquent la dispersion et la symétrie présentes parmi les plus petites et les plus grandes observations (chaque moustache représente le comportement d'environ 25% des observations).

Lorsque la série présente des valeurs extrêmes, les moustaches risquent de devenir très grandes, ce qui nuit à leur interprétation. C'est pourquoi une version un peu modifiée de la boîte à moustaches, appelée *Schematic Plot*, a été introduite. Dehon, Droesbeke et Vermandele (2008) introduisent pour ce faire les concepts de *valeurs pivots, valeurs adjacentes* et *valeurs extérieures.*

(a) Les valeurs pivots a_1 et a_2 sont définies par les relations suivantes:

$$\begin{aligned} a_1 &= Q_1 - 1,5 \times EIQ \\ a_2 &= Q_3 + 1,5 \times EIQ \end{aligned}$$

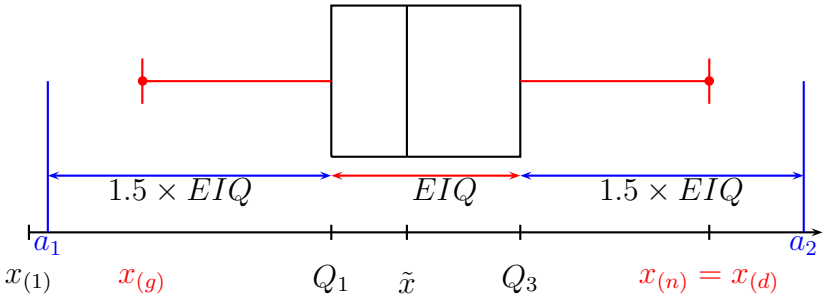
Ces valeurs sont situées de part et d'autre de la boîte à une distance égale à une fois et demi la longueur de la boîte. La plupart des séries ne contenant pas de valeurs aberrantes ont leurs observations comprises dans l'intervalle $[a_1, a_2]$.



(b) Les valeurs adjacentes sont données par les observations de rang g et d telles que

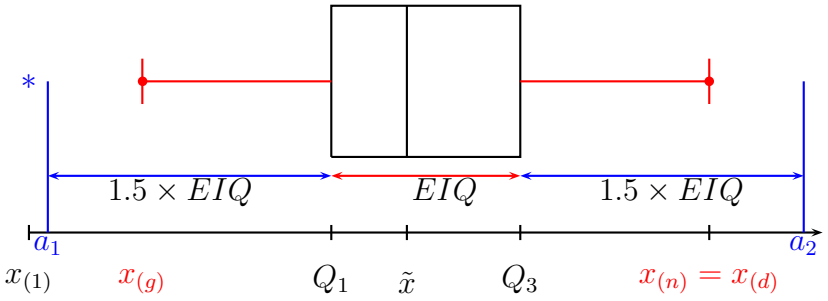
$$\begin{aligned} x_{(g)} &= \text{la plus petite observation supérieure ou égale à } a_1 \\ x_{(d)} &= \text{la plus grande observation inférieure ou égale à } a_2 \end{aligned}$$

Les moustaches de la boîte à moustaches modifiée relient les quartiles à ces valeurs qui correspondent aux valeurs observées les plus proches des valeurs pivots tout en appartenant à l'intervalle $[a_1, a_2]$. La boîte à moustaches devient donc



Si toutes les observations de la série sont comprises entre a_1 et a_2, alors $x_{(g)} = x_{(1)}$ et $x_{(d)} = x_{(n)}$ et la boîte représentée au départ est obtenue de nouveau. Dans le cas contraire, une attention particulière doit être consacrée aux observations se trouvant à l'extérieur de cette représentation.

(c) Les valeurs extérieures sont les observations situées en dehors de $[a_1, a_2]$. Elles sont généralement représentées par des points ou des étoiles comme le montre le graphique ci-dessous. Une observation extérieure n'est pas nécessairement une donnée aberrante. Par contre, toute observation aberrante sera sûrement parmi les valeurs extérieures. Il faut donc toujours interpréter les résultats avec prudence et discernement.



Les boîtes à moustaches sont très utiles pour visualiser les concepts de centralité et de dispersion mais aussi des concepts de symétrie et d'asymétrie comme nous le verrons dans la suite. Elles permettent de plus de comparer aisément plusieurs séries entre elles.

Exemple 53 *Considérons la variable* `IndiceRichesse` *et décomposons ses valeurs en fonction des cinq provinces. La Figure 3.6 caractérise, à l'aide de boîtes à moustaches,*

les distributions de la variable dans chaque province. La différence entre les tendances centrales est mise en évidence par les positions relatives des boîtes. Visiblement, la province du Brabant wallon est privilégiée avec une médiane fortement décalée par rapport aux autres, la province de Liège étant elle aussi légèrement mieux placée que les trois dernières provinces. On constate même d'un seul coup d'oeil que le premier quartile de la série du Brabant wallon est supérieur à l'ensemble des troisièmes quartiles: il y a donc plus ou moins 75% des communes du Brabant wallons dont l'indice de richesse est supérieur à l'indice de richesse de plus ou moins 75% communes des autres provinces. La variabilié de `IndiceRichesse` *au centre des séries semble similaire dans quatre des cinq provinces, la province du Luxembourg montrant une forte concentration au centre de sa distribution (ce qui explique aussi le nombre important de communes extérieures). En tenant compte également des moustaches, les provinces les plus dispersées en termes d'indice de richesse sont les provinces de Liège et du Hainaut.*

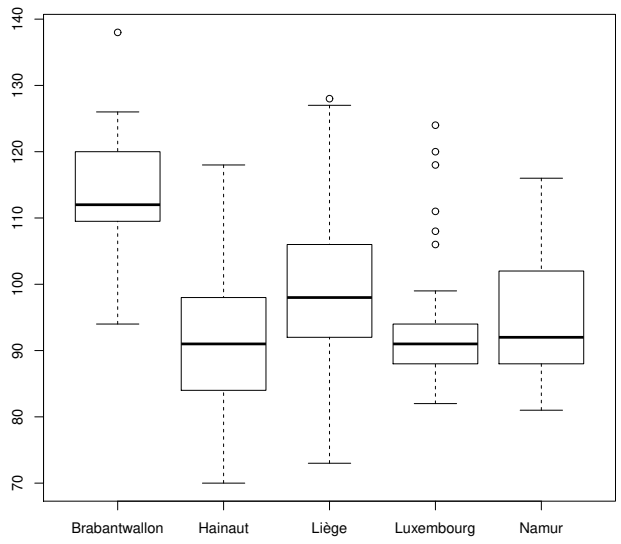


Figure 3.6: Boîte à moustaches de la variable `IndiceRichesse` dans les différentes provinces

Lorsque les nombres d'observations différent assez fortement d'une série à l'autre, la largeur des boîtes à moustaches peut être modifiée de manière à transmettre l'information via le graphique. La convention la plus classique est de prendre la largeur proportionnelle à la racine carrée de l'effectif. Dans le cas des provinces (dont les effectifs seront donnés dans quelques pages dans le tableau 3.2), les boîtes à moustaches ajustées en largeur sont représentées à la Figure 3.7.

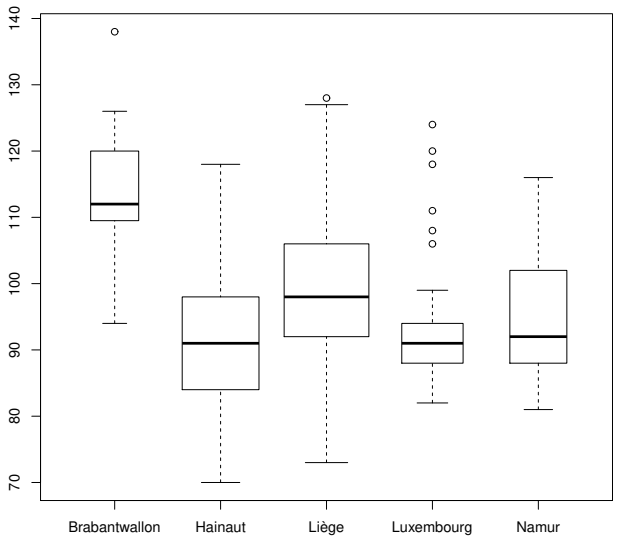


Figure 3.7: Boîte à moustaches de la variable `IndiceRichesse` dans les différentes provinces, en tenant compte des effectifs de chaque province

Dans le cas d'une série de données groupées, les observations individuelles sont généralement manquantes, ce qui ne permet pas de représenter exactement les moustaches de la boîte. On se contente alors de dessiner des moustaches allant jusqu'aux valeurs pivots sachant que les paramètres $(Q_1, \tilde{x}, Q_3)$ sont estimés par interpolation linéaire.

3.2.3 La variance et l'écart-type

L'étendue et l'écart interquartile exploitent essentiellement les observations occupant certains rangs particuliers dans les séries ordonnées. D'autres paramètres de dispersion utilisent quant à eux chaque observation. Ceux-ci dépendent habituellement du calcul préalable d'un paramètre de tendance centrale et mesurent ensuite la dispersion en se basant sur les écarts entre ce paramètre et chaque observation.

Considérons une valeur centrale estimée par un paramètre de position a. Une forte concentration des observations autour de ce paramètre se traduit par des écarts $x_i - a$ petits en amplitude. Inversement, une grande dispersion entraîne des écarts importants. La moyenne de tous ces écarts pourraient dès lors servir à mesurer la dispersion des valeurs autour du paramètre de position. Cependant, pour éviter que des écarts positifs ne soient compensés par des écarts négatifs, il faut considérer les différences $x_i - a$ sans leur signe, par exemple en utilisant la valeur absolue $|x_i - a|, i = 1, \ldots, n$ ou en considérant les écarts au

carré.

L'utilisation d'une valeur absolue mène au calcul de l'*écart absolu moyen* ou de l'*écart absolu médian* selon le paramètre de tendance centrale exploité. Nous ne considérerons pas ces paramètres dans ce cours car, en pratique, c'est la moyenne des écarts élevés au carré que l'on privilégie (les écarts étant construits à partir de la moyenne arithmétique).

Soit $S = \{x_1, \ldots, x_n\}$ une série quantitative quelconque. La moyenne des carrés des écarts entre les observations de S et la moyenne arithmétique $\bar{x}$ de S est la *variance* de la série. On la note habituellement s^2 avec

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

On constate à partir de cette définition que l'unité dans laquelle la variance s'exprime vaut le carré de l'unité utilisée pour les valeurs observées. Cette caractéristique rend difficile l'interprétation de ce paramètre. C'est pourquoi on lui préfère généralement l'écart-type défini comme étant la racine carrée de la variance

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Exemple 54 *Pour l'ensemble des cinq tailles considérées dans l'exemple 32, la variance vaut* s^2 *=36.16 cm*^{2}*. L'écart-type quant à lui vaut 6 cm.*

Exemple 55 *Revenons à l'indice de richesse décomposé en les différentes provinces. Le tableau ci-dessous donne la moyenne et l'écart-type de cette variable dans les différentes provinces.*

Provinces	Moyennes	Ecarts-types	Nombres
Brabant wallon	114.1	8.63	27
Hainaut	91.7	10.56	69
Liège	98.9	10.66	84
Luxembourg	93.4	9.55	44
Namur	94.2	8.87	38

Tableau 3.2: Moyennes et écarts-types de la variable `IndiceRichesse` décomposée en fonction des provinces

A partir de ce tableau, des commentaires similaires à ceux déjà mis en évidence à l'exemple 53 peuvent être à nouveau soulignés: c'est la province du Brabant wallon qui est la mieux

placée en termes d'indice de richesse (sa moyenne est supérieure aux autres et sa dispersion est la plus faible). La province de Liège est également privilégiée en moyenne mais présente aussi la plus fort variabilité.

Lorsque la série est recensée ou groupée, la définition de la variance se fait comme suit:

- Si $S = \{(x_j, n_j), j = 1, \ldots, J\}$ où $x_1 < x_2 < \ldots < x_J$ sont les valeurs distinctes observées avec les effectifs $n_1, \ldots, n_J$, la variance de la série vaut

$$s^2 = \frac{1}{n} \sum_{j=1}^{J} n_j (x_j - \bar{x})^2,$$

avec $\bar{x} = \frac{1}{n} \sum_{j=1}^{J} n_j x_j$.

- Si la série est groupée en J classes de centres $c_1, \ldots c_J$ et d'effectifs $n_1, \ldots, n_J$, il n'est pas possible de calculer exactement la variance lorsque les données individuelles ne sont pas disponibles. On peut calculer une valeur approximative de s^2 en remplaçant les observations par les centres des classes dans lesquelles elles tombent:

$$s^2 = \frac{1}{n} \sum_{j=1}^{J} n_j (c_j - \bar{x})^2, \tag{3.8}$$

avec $\bar{x} = \frac{1}{n} \sum_{j=1}^{J} n_j c_j$. Cependant, cette façon de procéder pour obtenir une approximation est plus controversée dans le cas d'un paramètre de dispersion que pour un paramètre de position. En effet, on ne peut plus supposer que les erreurs commises dans les classes se corrigent d'elles-mêmes par compensation entre les erreurs positives et les erreurs négatives (puisque les erreurs sont élevées au carré). Une abondante littérature existe sur le sujet et beaucoup de suggestions ont été faites afin d'apporter une correction à la formule (3.8) pour améliorer l'approximation. Par exemple, dans le cas d'une répartition en classes d'amplitude constante (égale à a), la correction de Sheppard consiste à soustraire $\frac{a^2}{12}$ de la formule (3.8).

Propriétés de la variance

La plupart des propriétés de l'écart-type découlent des propriétés de la variance (conservées par passage à la racine carrée) qui est, quant à elle, plus facile à manipuler mathématiquement.

1. Le choix de $\bar{x}$ comme paramètre de position dans la définition de s^2 n'est pas arbitraire puisque, comme nous l'avons démontré dans une section précédente, la moyenne minimise la somme des carrés des écarts entre les éléments d'une série et une constante

$a \in \mathbb{R}$. La moyenne est donc le "meilleur" paramètre de tendance centrale à insérer dans

$$s^2(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 \tag{3.9}$$

dans la mesure où il donne la plus petite valeur possible à l'expression (3.9). Notons que cette propriété s'obtient directement à partir du théorème suivant:

Théorème de König-Huygens: La moyenne des carrés des écarts entre les observations d'une série et un paramètre a se décompose de la façon suivante:

$$s^2(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 = s^2 + (\bar{x} - a)^2 \tag{3.10}$$

où s^2 est la variance de la série.

Preuve: Il suffit de développer $(x_i - a)^2$. □

2. Le calcul de la variance d'une série peut être long et fastidieux. En pratique, on utilise la formule équivalente suivante:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

Preuve: En prenant $a = 0$ dans la relation (3.10), il vient

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = s^2 + \bar{x}^2 \Rightarrow s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

□

3. Comparer chaque observation à la moyenne revient à comparer toutes les observations entre elles comme le montre la formule équivalente suivante de la variance:

$$s^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2.$$

Preuve: Il s'agit d'une simple manipulation des signes sommatoires. □

4. Si on effectue un changement d'échelle sur les observations initiales $x_1, \ldots, x_n$ pour obtenir la série $S' = \{x'_1, \ldots, x'_n\}$, avec $x'_i = ax_i + b$, $a, b \in \mathbb{R}$, alors la variance s'^2 de S' est donnée par $s'^2 = a^2 s^2$ où s^2 est la variance de S. Pour les écarts-types, la relation devient $s' = |a|s$. Les changements d'origine n'ont donc aucun effet sur la variance et l'écart-type.

Preuve: Une propriété similaire a établi que la moyenne de la nouvelle série suit la même transformation que les données: $\bar{x}' = a\bar{x} + b$. Dès lors, par définition de la variance, on a

$$s'^2 = \frac{1}{n}\sum_{i=1}^{n}(x'_i - \bar{x}')^2 = \frac{1}{n}\sum_{i=1}^{n}(ax_i - b - a\bar{x} + b)^2 = a^2 s^2.$$

□

5. Décomposition de la variance: Si une population P de n individus est partagée en k sous-populations $P_1, \ldots, P_k$ d'effectifs $n_1, \ldots, n_k$ (avec $\sum_{i=1}^{k} n_i = n$), de moyennes $\bar{x}_1, \ldots, \bar{x}_k$ et de variances $s_1^2, \ldots, s_k^2$, alors la variance s^2 de la population globale peut être déduite des paramètres des sous-populations par la relation

$$s^2 = \frac{\sum_{i=1}^{k} n_i s_i^2}{n} + \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{n},$$

où $\bar{x}$ est la moyenne globale de P.

Preuve: La série globale $S = \{x_1, \ldots, x_n\}$ est constituée des éléments des k sous-séries. Notons I_j le sous-ensemble d'indices parmi $1, \ldots, n$ des individus appartenant à la sous-population P_j. Par définition,

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i\in I_j}(x_i - \bar{x})^2. \tag{3.11}$$

La formule finale s'obtient en appliquant le théorème de König-Huygens dans chaque sous-population avec $a = \bar{x}$. □

L'interprétation de cette relation est importante: le premier terme est une moyenne pondérée des variances des sous-populations. On l'appelle "variance *dans* les groupes". Le second terme porte le nom de "variance *entre* les groupes" puisqu'il peut être considéré comme une variance des moyennes $\bar{x}_i$. On a alors la décomposition suivante de la variance globale:

Variance globale	=	Variance *dans* les groupes	+	Variance *entre* les groupes

6. Propriété de Tchebychev: Soit $S = \{x_1, \ldots, x_n\}$ une série de moyenne $\bar{x}$ et d'écart-type s. La proportion d'observations s'écartant d'au moins t écarts-types de la moyenne est inférieure ou égale à $\frac{1}{t^2}$. En d'autres termes, le nombre d'observations situées dans l'intervalle $]\bar{x} - t\ s, \bar{x} + t\ s[$ vaut au moins $n - \frac{n}{t^2}$.

Preuve: Par définition, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Séparons les indices $i = 1, \ldots, n$ en les deux sous-ensembles

$$\begin{aligned} I_1 &= \{i : |x_i - \bar{x}| < ts\} \\ I_2 &= \{i : |x_i - \bar{x}| \geq ts\} \end{aligned}$$

ce qui permet de scinder la somme en deux parties dans la définition de la variance: $s^2 = \frac{1}{n}\sum_{i\in I_1}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i\in I_2}(x_i - \bar{x})^2$. Chaque terme étant positif ou nul, on obtient les inégalités suivantes en exploitant de plus la définition des ensembles I_1 et I_2:

$$s^2 \geq \frac{1}{n}\sum_{i\in I_2}(x_i - \bar{x})^2 \geq \frac{1}{n}\sum_{i\in I_2} t^2 s^2 = \frac{k}{n} t^2 s^2, \tag{3.12}$$

où k est le nombre d'indices appartenant à I_2 ou encore le nombre d'observations s'écartant d'au moins t écarts-types de la moyenne. En réarrangeant les termes de (3.12), on obtient $\frac{k}{n} \leq \frac{1}{t^2}$, où k/n est la proportion cherchée. □

3.2.4 Le coefficient de variation

Le coefficient de variation constitue une mesure relative de dispersion puisqu'il est défini par

$$CV = \frac{s}{\bar{x}}.$$

Il s'agit d'un nombre "pur" (c'est-à-dire sans unité) que l'on exprime généralement en %. Il permet de comparer plusieurs séries dépendant d'unités différentes.

Exemple 56 *Intéressons-nous à la province de Liège. Dans les exemples précédents, la variable* `IndiceRichesse` *a été considérée en détail. Via le tableau 3.2, on sait notamment que la moyenne vaut 98.9 et que l'écart-type est égal à 10.66. Si maintenant on analyse également le taux de chômage, on obtient une moyenne des taux égale à 12.08 avec un écart-type égal à 4.85. Les variables étant exprimées en des unités différentes, il est difficile en comparant les écarts-types de déterminer quelle variable présente la plus grande dispersion entre les communes de la province de Liège. Les coefficients de variation sont par contre utilisables: on obtient 0.108 (ou 11%) pour la variable* `IndiceRichesse` *et 0.402 (ou 40%) pour la variable* `Chomage`*. La dispersion est donc nettement plus forte dans la série des taux de chômage.*

3.2.5 Choix d'un paramètre de dispersion

La mesure de la dispersion dans un ensemble de données est primordiale puisqu'elle est liée au caractère d'homogénéité ou d'hétérogénéité d'un groupe.

La variance, l'écart-type et le coefficient de variation sont les paramètres les plus courants alors que les écarts absolus moyen ou médian ne sont que rarement exploités.

La boîte à moustaches est habituellement très utile et facile à construire. Elle offre des informations complémentaires par rapport aux autres paramètres.

3.3 Les paramètres de forme

Les paramètres de forme regroupent deux sortes de valeurs typiques d'une série: les paramètres de dissymétrie et les paramètres d'aplatissement. La plupart de ces paramètres sont définis à partir des moments centrés de la série.

On appelle *moment centré d'ordre k* d'une série $S = \{x_1, \ldots, x_n\}$ l'expression

$$m_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k.$$

Propriétés et remarques:

1. Les moments d'ordre 1 et 2 correspondent respectivement à $m_1 = 0$ et $m_2 = s^2$.

2. Si la série est donnée via les couples $(x_j, n_j),\ j = 1, \ldots, J$, le moment centré d'ordre k devient $m_k = \frac{1}{n} \sum_{j=1}^{J} n_j (x_j - \bar{x})^k$. Lorsque la série est groupée, une approximation de m_k peut être obtenue en remplaçant chaque observation x_i par le centre de la classe dans laquelle elle tombe. Notons que cette approximation n'est valide que lorsque les observations sont uniformément réparties au sein des classes. De plus, les erreurs positives et négatives commises dans les classes ont tendance à se compenser lorsque k est impair alors qu'elles s'additionnent lorsque k est pair.

3. Les moments centrés sont invariables par rapport aux changements d'origine. Par contre, tout changement d'unité sur les observations entraîne une modification des moments: si $x'_i = ax_i, i = 1, \ldots, n$, alors $m'_k = a^k m_k$.

4. On appelle par ailleurs moments non centrés d'ordre k, la quantité μ_k définie par

$$\mu_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k.$$

Dans la suite de ce chapitre, nous utiliserons principalement m_3 et m_4.

3.3.1 Les paramètres de dissymétrie

D'une manière générale, les moments centrés d'ordre pair $m_4, m_6, \ldots$ sont, comme la variance m_2, des paramètres de dispersion. Par contre, les moments centrés d'ordre impair, $m_3, m_5, \ldots$, sont des indices de dissymétrie ou d'obliquité. Ils sont nuls pour les distributions symétriques et différents de 0 pour les distributions dissymétriques, et d'autant plus grands en valeur absolue que la dissymétrie est accentuée.

La Figure 3.8 représente à partir de polygones des fréquences (obtenus pour n assez grand et une amplitude de classe petite) les trois types de dissymétrie que l'on peut rencontrer.

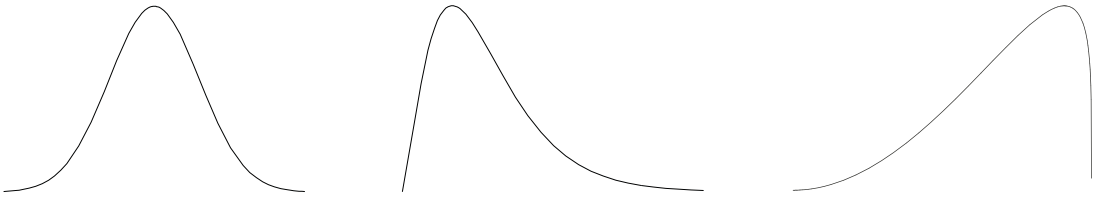


Figure 3.8: Cas typiques de dissymétrie ou symétrie: le polygone des fréquences du premier graphique est symétrique, le second dissymétrique à gauche et le troisième dissymétrique à droite.

On parle de symétrie pour le premier dessin, de dissymétrie à gauche pour le deuxième et de dissymétrie à droite dans le dernier. Le caractère non nul des moments centrés d'ordre impair indique la présence de dissymétrie dans les données tandis que les signes renseignent sur le type de dissymétrie. En effet, on sait que la moyenne correspond au centre de gravité de la série: en cas de dissymétrie à gauche (et donc d'étalement sur la droite), il y aura de très grands écarts positifs entre des observations x_i et $\bar{x}$. Ces grands écarts élevés au cube (ou à une puissance impaire supérieure) vont avoir une importance prépondérante dans le calcul des moments d'ordre impair. Ceux-ci seront dès lors positifs pour ce genre de dissymétrie. Le même raisonnement peut être suivi pour les autres types de dissymétrie.

Les principaux paramètres de dissymétrie sont les suivants:

1. Le coefficient de dissymétrie de Fisher:

   Ce coefficient est basé sur le moment d'ordre 3. Pour éviter que le coefficient construit ne soit affecté par un changement d'unité (voir remarque), Fisher a proposé de diviser m_3 par le cube de l'écart-type, ce qui donne lieu au coefficient de dissymétrie de Fisher:

   $$\gamma_1 = \frac{m_3}{s^3}.$$

   Le signe de γ_1 est celui de m_3.

   Exemple 57 *On obtient pour des séries caractérisées par les polygones de la Figure 3.8 respectivement les valeurs 0, 1.25 et -0.69 pour ce paramètre.*

2. Coefficients de dissymétrie empiriques:

   D'autres coefficients de dissymétrie sont plus rapides à calculer mais leurs propriétés résultent uniquement de constatations empiriques:

   - Le coefficient empirique de dissymétrie de Pearson:

     $$S_k = \frac{\bar{x} - x_M}{s}.$$

Ce coefficient n'est pas exploitable lorsque la série n'a pas de mode ou en a plusieurs.

Exemple 58 *Pour les séries caractérisées par les polygones de la Figure 3.8, les valeurs du coefficient empirique de Pearson sont 0, 0.63 et -0.98.*

- Le coefficient empirique de dissymétrie de Yule et Kendall:

$$Y_k = \frac{Q_1 + Q_3 - 2\tilde{x}}{Q_3 - Q_1}.$$

Il mesure en réalité la dissymétrie à l'intérieur de la partie centrale de la boîte à moustaches. En effet, Y_k peut se récrire

$$Y_k = \frac{Q_1 - \tilde{x} + Q_3 - \tilde{x}}{EIQ}.$$

Donc, $Y_k \in [-1, 1]$ et

  - Y_k est négatif si la médiane est plus proche du troisième que du premier quartile.
  - Y_k est nul si la médiane est à mi-chemin entre le premier et le troisième quartile.
  - Y_k est positif si la médiane est plus proche du premier que du troisième quartile.

Exemple 59 *Pour les séries caractérisées par les polygones de la Figure 3.8, les valeurs du coefficient empirique de Yule et Kendall sont 0, 0.15 et -0.14.*

L'interprétation des coefficients empiriques est la même que celle du coefficient de Fisher. Cependant, ils ne peuvent être considérés que comme des outils d'appréciation simples à obtenir mais pouvant être parfois contradictoires.

Exemple 60 *L'ordre de grandeur des paramètres des deux distributions non symétriques représentées dans la Figure 3.8 est différent. A partir de γ_1, on obtient une plus grande dissymétrie dans le deuxième graphique tandis que S_k donne le résultat contraire et que Y_k conduit à des valeurs approximativement égales en valeur absolue. Notons que les conclusions tirées des tous les paramètres concernant le type de dissymétrie sont les mêmes, ce qui n'est malheureusement pas toujours le cas.*

Remarques: Les coefficients de Fisher et de Pearson sont sensibles aux valeurs extrêmes puisqu'ils dépendent de la moyenne et de l'écart-type de la série. Par contre, le coefficient de Yule et Kendall Y_k est complètement insensible aux extrêmes puisqu'il n'exploite que la partie centrale de la boîte à moustaches.

3.3.2 Les paramètres d'aplatissement

Ces paramètres tentent de caractériser l'aplatissement d'une série par rapport à une courbe de référence, appelée la *courbe de Gauss*. La courbe de Gauss est définie par

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Elle est représentée sur chacun des deux graphiques de la Figure 3.9 avec une autre courbe décrivant le polygone des fréquences d'une population d'intérêt.

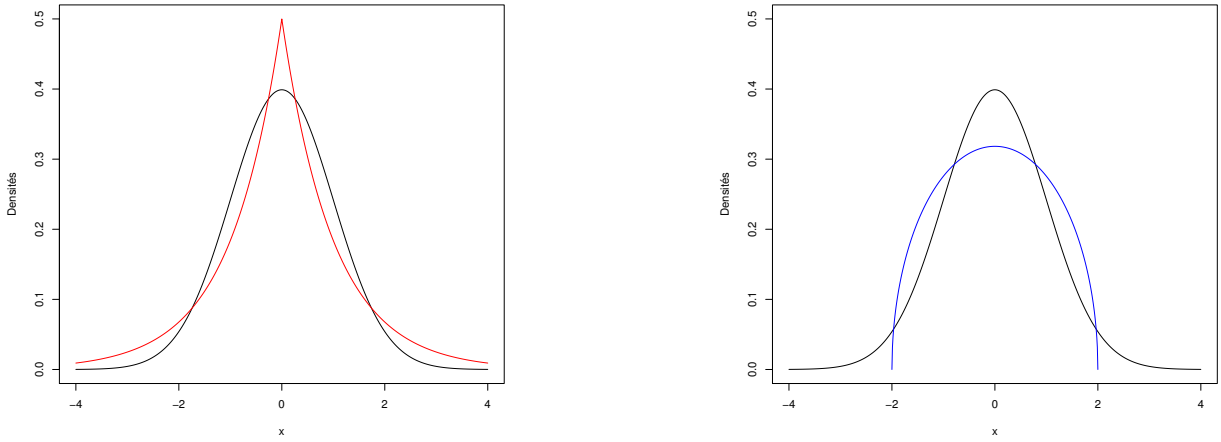


Figure 3.9: Polygones de fréquences pour deux distributions (en supposant qu'il y a beaucoup d'observations groupées dans des petites classes d'amplitude constante), dont un correspond à la courbe de Gauss (présente sur les deux graphiques).

On constate que cet autre polygone peut présenter une forme plus effilée au centre que la courbe de Gauss et cumuler également plus de masse dans les queues de la distribution (graphique de gauche). Une telle distribution est dite *leptokurtique*. Le polygone peut aussi être plus aplati au centre que la courbe de référence et avoir encore moins de masse que celle-ci dans les queues (graphique de droite). Il s'agit d'une distribution *platikurtique*.

Pour déterminer de quel type est une distribution, on calcule le moment centré d'ordre 4. Plus précisément, on dispose des deux paramètres suivants:

1. Le coefficient d'aplatissement de Pearson: $b_2 = \frac{m_4}{s^4}$.
2. Le coefficient d'aplatissement de Fisher: $\gamma_2 = \frac{m_4}{s^4} - 3 = b_2 - 3$. Le coefficient de Fisher est donc le coefficient de Pearson duquel on a soustrait 3, l'idée étant d'annuler ce coefficient pour la courbe de référence. Dès lors, les distributions leptokurtiques (resp. platikurtiques) seront caractérisées par une valeur de γ_2 positive (resp. négative).

3.4 Exercices

1. (a) Calculer la moyenne arithmétique, la médiane, le mode et les premier et troisième quartiles des séries suivantes:

$$S_1 = \{3, 5, 2, 6, 5, 9, 5, 2, 8, 6\}$$

$$S_2 = \{84, 91, 72, 68, 87, 78, 52, 92, 71, 85, 62, 82, 99\}.$$

   (b) Construire des séries possédant la même moyenne arithmétique, le même mode et la même médiane mais différant notamment par les effectifs et/ou les quartiles.

2. Un grossiste dispose d'un stock important de pommes. Celles-ci sont réparties dans des caisses contenant chacune 12 pommes. La distribution du nombre de pommes de qualité supérieure par caisse est décrite dans le tableau suivant:

Nombre de pommes	0	1	2	3	4	5	6	7	8	9	10	11	12
Nombre de caisses	1	0	1	3	5	7	14	33	54	66	72	78	66

   (a) En moyenne, combien y a-t-il de pommes de qualité supérieure par caisse?

   (b) Combien de pommes de qualité supérieure un client trouvera-t-il le plus fréquemment s'il achète, sans être suffisamment attentif, une caisse de pommes chez ce grossiste?

   (c) Si le grossiste décide de ne conserver que la moitié des caisses (évidemment celles contenant le plus possible de pommes de qualité supérieure), quel est le nombre minimum de pommes de qualité supérieure contenues dans ces caisses privilégiées?

   (d) Sachant que les amis du grossiste reçoivent en cadeau 10% des caisses (choisies parmi les meilleures), tandis que les 10% des caisses les moins bonnes sont gardées pour confectionner de la compote, déterminer le nombre moyen de pommes de qualité supérieure dans les caisses restantes.

3. Répondant à une offre d'emploi, une personne s'interroge sur le montant de ses rémunérations futures en cas d'embauche. Le directeur de l'entreprise de vente à domicile lui répond que le salaire moyen de la firme est supérieur à 1600 euros par mois, mais que, pendant la période de formation, l'employé ne gagnera que 500 euros chaque mois, puis sera augmenté dans la suite.

   Avant de signer le contrat d'embauche, la personne a mené son enquête et a obtenu les renseignements suivants:

   - Le directeur gagne 12500 euros par mois.

- Le sous-directeur gagne 6000 euros par mois.
- Chacun des 4 chefs de secteur gagne 1200 euros par mois.
- Chacun des 5 techniciens gagne 950 euros par mois.
- Chacun des 10 démarcheurs gagne 600 euros par mois.

(a) Le directeur a-t-il dit la vérité au candidat?

(b) Quelle question le candidat aurait-il dû poser au patron pour avoir une estimation plus réaliste de son salaire futur?

4. Un village se compose de 4 quartiers. On connaît le nombre d'habitants par quartier et le nombre de véhicules par habitant.

Quartiers	Nbre d'habitants	Nbre de véhicules/habitant
A	1835	0,4583
B	1624	0,4883
C	729	0,5048
D	974	0,4774

Déterminer le nombre moyen de véhicules par habitant dans le village. De quelle moyenne s'agit-il?

5. La série groupée des poids de 60 étudiants masculins de premier bachelier est décrite dans le tableau suivant:

Classes	Effectifs
$[50; 60]$	12
$]60; 65]$	10
$]65; 70]$	12
$]70; 75]$	14
$]75; 80]$	5
$]80; 100]$	7
Total	60

(a) Pour cette répartition en classes, estimer la moyenne arithmétique $\bar{x}$ et la médiane $\widetilde{x}$ de la série des poids. Déterminer la classe modale et calculer une valeur approchée du mode x_M.

(b) Comment ces valeurs se comparent-elles par rapport aux paramètres $\bar{x}, \widetilde{x}$ et x_M calculés directement à partir des données brutes de la série sachant que celles-ci valent $\bar{x} = 70,02$, $\widetilde{x} = 69,5$ et $x_M = 65$? Donner un ordre de grandeur de l'erreur commise dans le calcul de $\bar{x}$ en remplaçant les observations des classes c_1 et c_4 par les centres des classes. Pour rappel, dans la classe $]50;60]$, les observations distinctes sont 52, 52, 55, 56, 57, 58, 59, 60, 60, 60, 60 et 60 tandis que dans la classe $]70;75]$, les observations sont 71, 71, 72, 72, 72, 72, 72, 72, 75, 75, 75, 75, 75 et 75.

(c) Donner une approximation des quartiles Q_1 et Q_3. A peu près 15 observations de la série initiale doivent avoir une valeur inférieure à Q_1 et le même nombre, une valeur supérieure à Q_3. En pratique, est-ce le cas?

6. Un institut de sondage a enquêté sur les frais d'entretien déclarés par les ménages possédant une résidence secondaire. Les dépenses (exprimées en une certaine unité monétaire et groupées en 7 classes) ainsi que les effectifs sont repris dans le tableau suivant, mais certaines données fournies par l'enquêteur sont restées indéchiffrables.

Classes pour les frais déclarés	Effectifs
$[0,4]$	6
$]4,8]$	n_2
$]8,12]$	n_3
$]12,e_4]$	17
$]e_4,22]$	14
$]22,30]$	11
$]30,42]$	3
Total	100

(a) Retrouver les valeurs manquantes n_2 et n_3 sachant que le premier quartile vaut $Q_1 = 7$.

(b) Sachant que $\bar{x} = 13$, déterminer la borne e_4 de la classe.

7. On a interrogé 92 représentants de commerce sur le nombre de kilomètres qu'ils effectuaient par jour pour leur travail. Les résultats sont repris dans le tableau ci-dessous, duquel certaines données ont disparu.

Trajets en km	Nombres de représentants
$[10, 20]$	x_1
$]20, 40]$	26
$]40, x_2]$	19
$]x_2, x_3]$	24
$]x_3, 100]$	14

(a) Retrouver les valeurs manquantes x_1, x_2 et x_3 sachant que le trajet médian est égal à 45,79 km et que le trajet moyen est égal à 49,89 km.

(b) Construire l'ogive des fréquences cumulées et vérifier graphiquement la valeur de la médiane.

8. Soit la série statistique $S = \{-2, 2, -1, 1, x\}$, où x désigne un nombre réel arbitraire.

   Calculer la médiane de S lorsque le paramètre x varie. Représenter graphiquement la fonction qui, à tout réel x, associe la médiane de S.

   Même problème en remplaçant la médiane par la moyenne arithmétique de S.

   Commenter les résultats et comparer les deux situations.

9. Soit $S = \{x_1, \ldots, x_n\}$ une série statistique univariée de moyenne $\bar{x}$ et de variance s_x^2. Calculer la moyenne et la variance de la série des valeurs centrées et réduites $Z = \{z_1, \ldots, z_n\}$ où $z_i = \frac{x_i - \bar{x}}{s_x}$.

10. On dispose de la série suivante groupée en 5 classes dont seuls les effectifs des deuxième et troisième classes sont connus.

Classes	Effectifs n_i
$[0, 2]$	n_1
$]2, 4]$	5
$]4, 6]$	6
$]6, 8]$	n_4
$]8, 10]$	n_5
Total	20

    Calculer la médiane de cette série sachant que la moyenne vaut 4,7 et la variance 5,71.

11. Le tableau suivant donne la répartition en 5 classes des salaires (en unités monétaires) des employés d'une entreprise.

Classes de salaire c_i	Effectifs n_i	Fréquences cumulées F_i
$[2, 6]$	n_1	0,1
$]6, 10]$	n_2	0,33
$]10, 12]$	n_3	0,6
$]12, 14]$	n_4	0,8
$]14, 18]$	n_5	1

Sachant que $s^2 = 8$, $\sum_{i=1}^{5} f_i \widetilde{x}_i^2 = 133,8$ et que $\sum_{i=1}^{5} n_i \widetilde{x}_i = 673$, où $\widetilde{x}_i$ est le centre et f_i la fréquence de la i-ème classe, calculer les effectifs n_i de chaque classe et l'effectif total n.

12. Une enquête auprès de 100 travailleurs a fourni une série de données relative au nombre d'emplois antérieurs occupés au cours des dix dernières années. Les deux premières colonnes du tableau statistique sont reprises dans le tableau suivant:

Valeurs x_i	Effectifs n_i
0	8
1	19
2	32
3	17
4	14
5	10
Total	100

(a) Calculer la moyenne arithmétique, la médiane et le mode de la série. Interpréter les résultats obtenus. Quel est le paramètre le plus indiqué pour cet ensemble de données?

(b) Déterminer les quartiles Q_1 et Q_3. Représenter ensuite la boite à moustaches correspondante.

(c) Calculer la variance et l'écart-type.

(d) Le diagramme en bâtons associé à ces données discrètes semble-t-il symétrique? Calculer les coefficients de dissymétrie de Fisher et Pearson.

13. Les poids de 22 étudiantes de premier bachelier sont donnés par la série ordonnée suivante:

$$S = \{47, 48, 49, 50, 53, 55, 55, 55, 56, 56,$$

$$58, 59, 61, 62, 62, 63, 63, 63, 64, 65, 65, 66\}.$$

(a) Calculer la moyenne arithmétique, la médiane et le mode de cette série et interpréter ces paramètres.

(b) Calculer l'écart-type pour la série des poids des filles. Estimer le nombre et le pourcentage d'étudiantes dont le poids se trouve entre $\bar{x}-2s$ et $\bar{x}+2s$ et comparer avec les nombres atteints réellement.

(c) Calculer les quartiles Q_1 et Q_3. Sachant que la série des poids des étudiants masculins de la même section correspond aux paramètres suivants:

$$Q_1 = 65; \widetilde{x} = 69,5; Q_3 = 75,$$

tandis que les observations individuelles sont reprises ci-dessous, comparer les deux distributions de poids à l'aide de boîtes à moustaches.

68	70	67	75	72	71	67	65	60	60	65	65
77	95	85	70	70	72	66	75	90	65	62	70
52	60	59	65	68	71	97	65	57	75	77	75
85	56	77	67	62	52	67	72	79	60	72	69
58	55	75	75	78	65	95	65	90	72	72	60

Quelle distribution donne la boîte la plus symétrique? Calculer le coefficient de Yule et Kendall pour les deux boîtes. Pour la série des poids des filles, comparer ce coefficient avec le coefficient de dissymétrie de Fisher.

(d) Dans une autre section comportant 25 étudiantes, les poids ont été aussi mesurés et en voici un résumé succinct : $\bar{x} = 62$, $s = 3,334$. Pour une répétition, les étudiantes des deux sections sont regroupées. A partir des paramètres individuels calculés sur les deux séries, calculer la moyenne arithmétique et la variance des poids de ce groupe de 47 étudiantes. Commenter la décomposition de la variance.

14. Une firme F comprend deux filiales A et B. Les salaires moyens (en euros) par catégorie socio-professionnelle sont donnés dans le tableau suivant :

	A		B	
Catégories	Sal. moyens	Eff.	Sal. moyens	Eff.
Ouvriers	1500	50	1200	10
Employés	1750	85	1600	20
Cadres	3000	5	2500	30

Comparer la dispersion des salaires moyens dans les deux établissements en calculant les variances s_1^2 et s_2^2. Calculer la variance globale des salaires de l'entreprise complète. Discuter la décomposition de la variance.

15. Pendant 14 semaines, on a relevé la recette, en milliers d'euros, d'un supermarché le lundi et le samedi. Les résultats sont repris dans le tableau ci-dessous.

Sem.	Recette lundi (X)	Recette samedi (Y)
1	57	83
2	60	93
3	52	77
4	49	69
5	56	81
6	46	70
7	51	71
8	63	91
9	49	70
10	57	82
11	40	55
12	45	65
13	65	105
14	55	80

Décrire les deux séries univariées correspondant respectivement aux recettes du lundi et à celles du samedi par des boîtes à moustaches (la construction des boîtes, c'est-à-dire le calcul des quartiles, des valeurs pivots,..., doit être expliquée).

16. Pour une série statistique $S = \{x_1, x_2, ..., x_n\}$, on connaît la moyenne arithmétique $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et la variance $s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$.

    On ajoute à S un élément y de manière à former la nouvelle série $S' = S \cup \{y\}$.

    (a) Calculer la variance $(s')^2$ de S' en fonction de y et des paramètres connus de S.

    (b) Etudier le comportement de $(s')^2$ lorsque y varie.

    (c) Etudier le comportement de $(s')^2$ lorsque l'effectif n de S grandit indéfiniment.

Chapitre 4

Série statistique bivariée

Ce chapitre décrit quelques méthodes permettant de présenter et d'analyser des séries statistiques obtenues par l'observation de deux variables sur les individus d'une même population. Une partie de l'analyse proposée s'inspire des méthodes vues dans les deux chapitres précédents pour les séries statistiques univariées (distributions des effectifs, représentations graphiques,...). Cependant, l'analyse des séries doubles doit aussi explorer les liens éventuels existant entre les deux variables considérées. Cette étude fait appel à la théorie de la corrélation et à la régression que nous aborderons dans ce chapitre. Notons que seules les méthodes descriptives nous intéressent dans cette partie du cours. La référence principale est celle de Bragard et Alexandre (1995), les autres ouvrages présentant, pour la plupart, ce sujet après avoir introduit des notions plus avancées de statistique.

4.1 Tableau de contingences

Appelons X la première variable et Y la seconde et considérons une population de n individus. Les données brutes peuvent être fournies via un tableau individus $\times$ caractères de n lignes et 2 colonnes:

	Variables	
Individus	X	Y
1	x_1	y_1
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

Souvent, la série est plus simplement décrite par l'ensemble des n couples (x_i, y_i) observés:

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

Lorsque l'effectif n de la population est grand, il peut être utile de condenser les données en une distribution des effectifs. Notons $x_1, x_2, \ldots, x_J$ les valeurs distinctes observées de la variable X (avec $x_1 < x_2 < \ldots < x_J$) et $y_1, y_2, \ldots, y_K$ les valeurs distinctes observées de la variable Y (avec $y_1 < y_2 < \ldots < y_K$). Désignons par n_{jk} l'effectif associé au couple (x_j, y_k) pour $1 \leq j \leq J$ et $1 \leq k \leq K$. La distribution des effectifs est habituellement décrite dans un tableau statistique à double entrée, appelé *tableau de contingence*. Une ligne est réservée à chaque valeur observée de X et une colonne à chaque valeur observée de Y. A l'intersection de la j ème ligne et de la k ème colonne se trouve l'effectif n_{jk}.

	Valeurs de Y				
Valeurs de X	y_1	$\ldots$	y_k	$\ldots$	y_K
x_1	n_{11}	$\ldots$	n_{1k}	$\ldots$	n_{1K}
$\vdots$					
x_j	n_{j1}	$\ldots$	n_{jk}	$\ldots$	n_{jK}
$\vdots$					
x_J	n_{J1}	$\ldots$	n_{Jj}	$\ldots$	n_{JK}

Tableau 4.1: Tableau de contingence pour une série double

La somme des éléments du tableau doit rendre le nombre total d'individus dans la population:

$$\sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk} = n.$$

Exemple 61 *Considérons l'ensemble des communes de Wallonie et intéressons-nous à leur appartenance à l'une des cinq provinces ainsi qu'au parti du bourgmestre. Le tableau 4.2 décrit la distribution des effectifs conjoints.*

Partis	Provinces				
	Brabantwallon	Hainaut	Liège	Luxembourg	Namur
Autre	2	0	7	3	1
CdH	3	13	16	23	12
Ecolo	2	1	1	1	0
MR	14	16	32	10	17
PS	6	39	28	7	8

Tableau 4.2: Tableau de contingence `PartiBourgmestre` × `Provinces` pour l'ensemble des communes de Wallonie

Lorsque le nombre de lignes et/ou de colonnes du tableau de contingence est trop grand, il est possible de recourir à des groupements en classes pour l'une des variables ou pour les deux. Les lignes et colonnes du tableau correspondent alors aux classes définies pour la variable X et pour la variable Y.

Au lieu d'indiquer les effectifs dans le tableau de contingence, on peut calculer les fréquences des couples observés:

$$f_{jk} = \frac{n_{jk}}{n}.$$

Des représentations graphiques décrivant le tableau de contingence peuvent être obtenues à 3 dimensions en généralisant certains des diagrammes décrits au Chapitre 2.

Exemple 62 *Afin de visualiser les effectifs indiqués dans le tableau de contingences 4.2, un diagramme en barres en 3 dimensions peut être représenté. En plaçant les modalités de chacune des variables dans le plan (de façon arbitraire puisque les variables sont qualitatives nominales), on associe à chaque couple observé un parallélipipède rectangle dont la base (a priori carrée) est la même pour tous mais dont la hauteur est égale à l'effectif ou à la fréquence du couple en question. Ce diagramme est illustré à la Figure 4.1.*

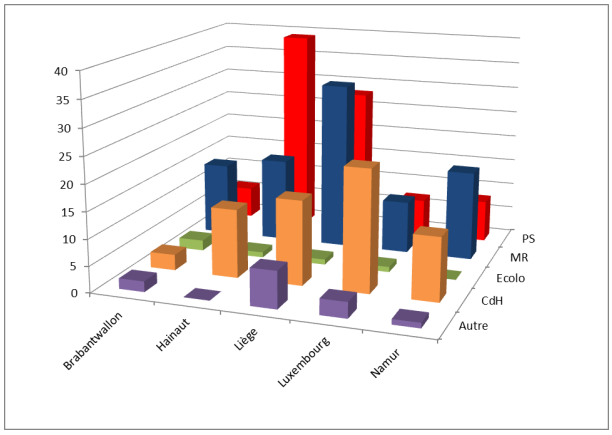


Figure 4.1: Diagramme en barres à 3 dimensions pour décrire la distribution des effectifs du couple `PartiBourgmestre`, `Provinces`.

Si deux variables quantitatives discrètes étaient étudiées conjointement, un diagramme en bâtons en 3 dimensions pourrait être représenté en associant un segment de longueur égale à l'effectif ou la fréquence à chaque couple observé et situé dans le plan de base. Dans le cas de deux variables quantitatives continues, un histogramme devrait être construit en associant à chaque couple de classes un parallélipipède rectangle dont la base correspondrait aux intervalles de classes et dont le volume serait proportionnel à l'effectif ou la fréquence du couple. Cependant, au lieu de travailler en 3 dimensions, lorsque les deux variables sont

quantitatives, la représentation graphique la plus classique est le *diagramme de dispersion*. Il s'agit d'une manière naturelle et simple de visualiser les données: on représente, dans un système d'axes orthogonaux, chaque individu i de la population par un point d'abscisse x_i et d'ordonnée y_i. Les observations constituent un nuage de points qui donne déjà une idée précise de la façon dont les individus se dispersent dans le plan.

Exemple 63 *Considérons, pour l'ensemble des communes de Wallonie, les couples de variables (*`Chomage`,`IndiceRichesse`*) et (*`PrixMaison`,`IndiceRichesse`*). Les nuages de points sont représentés à la Figure 4.2.*

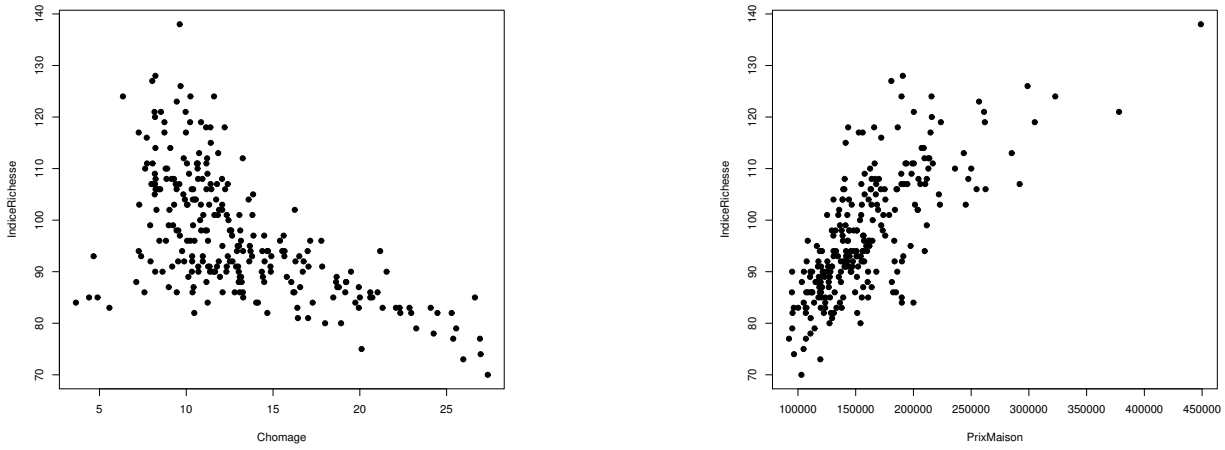


Figure 4.2: Nuages de points basés sur les 262 communes wallonnes pour les couples de variables (`Chomage`,`IndiceRichesse`), à gauche, et (`PrixMaison`,`IndiceRichesse`), à droite.

4.2 Distributions marginales

L'étude d'une série bivariée $S = \{(x_i, y_i), 1 \leq i \leq n\}$ doit d'abord commencer par l'étude des séries marginales univariées obtenues en ne considérant qu'une variable à la fois dans le tableau individus $\times$ caractères:

- Série marginale en X : $S_X = \{x_i, 1 \leq i \leq n\}$.
- Série marginale en Y : $S_Y = \{y_i, 1 \leq i \leq n\}$.

Le tableau de contingence étant une version condensée de la série double, il permet de décrire non seulement la distribution des effectifs de S mais aussi les distributions marginales de S_X et S_Y.

En calculant la somme des effectifs des couples d'abscisse x_j, on définit l'effectif marginal de x_j $(1 \leq j \leq J)$. Celui-ci se note $n_{j\bullet}$ et s'obtient par la formule

$$n_{j\bullet} = \sum_{k=1}^{K} n_{jk}.$$

De même, en additionnant les effectifs de tous les couples d'ordonnée y_k, on obtient l'effectif marginal de y_k : $n_{\bullet k} = \sum_{j=1}^{J} n_{jk}$.

Les distributions marginales de X et Y sont données par

$$S_X = \{(x_j, n_{j\bullet}), 1 \leq j \leq J\} \text{ et } S_Y = \{(y_k, n_{\bullet k}), 1 \leq k \leq K\},$$

et sont habituellement reprises dans une ligne ou colonne supplémentaire dans le tableau de contingence.

Exemple 64 *Le tableau 4.3 est le tableau de contingence complet correspondant à la distribution jointe des variables* `Provinces` *et* `PartiBourgmestre`*.*

Partis	Provinces					
	Brabantwallon	Hainaut	Liège	Luxembourg	Namur	Dist Marg
Autre	2	0	7	3	1	13
CdH	3	13	16	23	12	67
Ecolo	2	1	1	1	0	5
MR	14	16	32	10	17	89
PS	6	39	28	7	8	88
Dist Marg	27	69	84	44	38	262

Tableau 4.3: Tableau de contingence `PartiBourgmestre` × `Provinces` pour l'ensemble des communes de Wallonie

Des fréquences marginales peuvent être obtenues en divisant les effectifs marginaux par n: $f_{j\bullet} = \frac{n_{j\bullet}}{n}$ et $f_{\bullet k} = \frac{n_{\bullet k}}{n}$. De même, on peut calculer des effectifs ou fréquences cumulés marginaux.

Remarques:

1. $\sum_{j=1}^{J} n_{j\bullet} = \sum_{k=1}^{K} n_{\bullet k} = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk} = n$.
2. L'étude de la série double permet de disposer de deux séries statistiques simples. Cependant, l'inverse n'est pas vrai. A partir des distributions des effectifs des deux séries simples S_X et S_Y, on ne peut pas construire un tableau de contingence relatif à l'observation simultanée des deux variables X et Y.

4.3 Distributions conditionnelles

Une distribution conditionnelle consiste à fixer a priori la valeur d'une variable et à examiner les variations de l'autre variable compte tenu de cette contrainte. Considérons par exemple une distribution conditionnelle de Y en X. Fixons une valeur de la variable X: prenons par exemple x_j. Examinons alors l'ensemble des couples observés (x_j, y_k) avec $1 \leq k \leq K$. Les effectifs n_{jk} de ces couples définissent une distribution univariée appelée distribution conditionnelle de Y en X avec $X = x_j$: on la notera

$$S_{Y|x_j} = \{(y_k, n_{jk}), 1 \leq k \leq K\}.$$

Cette série comporte $n_{j\bullet}$ observations. Les fréquences conditionnelles sont donc données par

$$f_{k|x_j} = \frac{n_{jk}}{n_{j\bullet}}.$$

On procède de la même manière pour définir une distribution conditionnelle de X en Y.

Exemple 65 *Au Chapitre 3 (voir exemple 53), l'indice de richesse des communes a été étudié dans chacune des provinces. Les boîtes à moustaches ont notamment permis de visualiser les distributions conditionnelles de cet indice tout en imposant chacune des modalités possibles de la variable qualitative* `Provinces`*.*

4.4 Réduction des données

Les paramètres utilisés pour caractériser les séries statistiques bivariées sont de deux types:

- Certains ne concernent qu'une variable à la fois: ils servent essentiellement à caractériser individuellement les diverses distributions marginales et conditionnelles en recourant aux définitions introduites pour les séries univariées.
- D'autres, au contraire, servent à décrire les relations existant entre les deux séries d'observations.

4.4.1 Tendance centrale et dispersion des séries marginales

Touts les paramètres vus dans le Chapitre 3 peuvent être exploités de façon marginale. Dans la suite, nous utiliserons essentiellement les moyennes et variances marginales, dont les définitions, adaptées aux notations introduites dans ce contexte bivariée, sont reprises ci-dessous.

La moyenne marginale $\bar{x}$ de la série $S_X = \{x_1, \ldots, x_n\}$ peut être calculée par

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

ou, à partir du tableau de contingence, par

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{J} n_{j\bullet} x_j.$$

De même, la variance marginale de la série S_X est donnée par

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ ou } s_x^2 = \frac{1}{n} \sum_{j=1}^{J} n_{j\bullet} (x_j - \bar{x})^2.$$

Les moyenne et variance marginales de la variable Y, $\bar{y}$ et s_y^2, se définissent de façon analogue. Comme dans le cas univarié, les moyennes marginales ont une signification concrète intéressante. Dans le plan $\mathbb{R}^2$, le point de coordonnées $(\bar{x}, \bar{y})$ représente le centre de gravité du nuage de points.

4.4.2 Covariance et corrélation

Le tableau de contingence et, dans le cas quantitatif, le diagramme de dispersion permettent d'avoir une première idée du lien éventuel entre les deux variables. C'est le diagramme de dispersion qui est le plus riche en informations conjointes, mais aussi marginales. Dès sa construction, il est important d'effectuer certaines constatations à propos des points suivants:

- Le nuage de points est-il concentré ou au contraire dispersé?
- Y a-t-il une structure dans sa composition (la relation sous-jacente est-elle linéaire, quadratique, logarithmique,...)?
- Certaines valeurs individuelles semblent-elles aberrantes?
- ...

Ces quelques questions permettent de commencer l'analyse bivariée des données. Dans certains cas, les variables peuvent agir indépendamment l'une de l'autre. Dans d'autres cas, on peut s'attendre à observer une association entre les valeurs des deux séries.

Dans l'hypothèse d'une association, on tentera de mesurer le degré d'association existant entre les variables considérées. Une association entre deux variables n'implique pas nécessairement un lien de causalité. Elle peut être due à une coïncidence, à une évolution temporelle dans le même sens ou à une cause externe. Cependant, dans certains cas, cette

association traduira une relation de dépendance entre les deux variables. Une telle relation permet d'expliquer (au moins partiellement) le comportement d'une des variables en fonction de l'autre.

On distingue deux types de dépendance:

- La dépendance fonctionnelle: cela signifie qu'il existe une fonction f telle que $y = f(x)$.
- La dépendance statistique: elle traduit une association moins nette que la dépendance fonctionnelle et correspond au lien le plus simple, à savoir le lien linéaire (augmentation simultanée des deux variables ou comportement antagoniste entre elles)

Dans la suite de ce chapitre, nous considérons uniquement les couples de variables quantitatives. Une première mesure d'association linéaire est donnée par la covariance de la série statistique double $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Elle est définie par l'expression

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

où $\bar{x}$ et $\bar{y}$ désignent les moyennes marginales. A partir de la distribution des effectifs, on obtient

$$s_{xy} = \frac{1}{n} \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk}(x_j - \bar{x})(y_k - \bar{y}). \tag{4.1}$$

De même, si les variables sont groupées, une estimation de la covariance peut être obtenue en remplaçant les valeurs x_j et y_k de la définition (4.1) par les centres des classes.

Ce coefficient est positif ou négatif suivant la position des observations par rapport au centre de gravité $(\bar{x}, \bar{y})$ de la série.

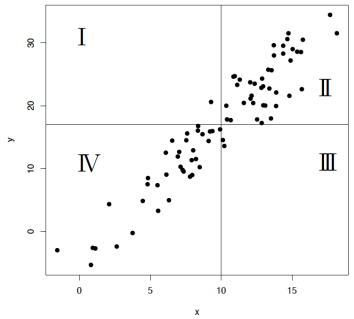


Figure 4.3: Interprétation du signe de la covariance à partir d'une série double (X, Y).

En effet, à partir de la Figure 4.3 basée sur deux variables quantitatives quelconques (avec le plan découpé verticalement par $\bar{x}$ et horizontalement par $\bar{y}$), on constate que les

points situés dans les quadrants I et III apportent une contribution négative à la covariance puisque $x_i - \bar{x} \leq 0$ et $y_i - \bar{y} \geq 0$ dans I et $x_i - \bar{x} \geq 0$ et $y_i - \bar{y} \leq 0$ dans III, ce qui donne un produit $(x_i - \bar{x})(y_i - \bar{y})$ négatif ou nul. Par contre, les points des quadrants II et IV contribuent positivement à la covariance. Si la covariance est positive (resp. négative), cela signifie qu'il y a prédominance d'observations dans les parties II et IV (resp. I et III). On voit donc que la covariance est positive ou négative selon que la relation entre les deux séries de données est croissante ou décroissante. De plus, la covariance est nulle ou presque nulle quand il y a compensation entre les deux catégories de points.

Exemple 66 *A partir des diagrammes de dispersion présentés à la Figure 4.2, on voit directement que la relation entre le chômage et l'indice de richesse est plutôt décroissante (une commune dont le taux de chômage est élevé sera plutôt caractérisée par un faible indice de richesse). La covariance est égale à -33.5. Par contre, les prix moyens des maisons et l'indice de richesse évoluent dans le même sens sur l'ensemble des communes. Une commune d'indice élevé aura tendance à avoir un prix moyen plutôt supérieur à la moyenne et une autre commune dont l'indice de richesse est inférieur à la moyenne correspondra à un prix moyen faible. La covariance est égale à 408 248.7.*

Propriétés de la covariance:

1. La covariance est souvent calculée en exploitant la formule équivalente suivante:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \text{ ou } s_{xy} = \frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} x_j y_k - \bar{x}\bar{y}.$$

2. Effet du changement d'échelle et d'origine sur la covariance: si on effectue une transformation affine sur les observations pour obtenir la nouvelle série double $S' = \{(x'_i, y'_i) : 1 \leq i \leq n\}$ avec $x'_i = ax_i + b$ et $y'_i = cy_i + d$, avec $a, b, c, d \in \mathbb{R}$, alors la covariance de la nouvelle série vérifie $s'_{xy} = acs_{xy}$.

   Interprétation: le signe de la covariance renseigne sur le caractère croissant ou décroissant de la relation entre les deux variables. Cependant, comme la covariance est influencée par les changements d'échelle, la valeur absolue de s_{xy} ne permet pas de se faire une idée de l'intensité de la relation.

3. La covariance de la série double S est toujours, en valeur absolue, inférieure ou égale au produit des écarts-types marginaux s_x, s_y : $|s_{xy}| \leq s_x s_y$. L'égalité n'est possible que si et seulement si tous les points observés sont situés sur une même droite.

   Preuve: Il suffit d'appliquer l'inégalité de Cauchy-Schwarz aux vecteurs centrés $(x_1 - \bar{x}, \ldots, x_n - \bar{x})^t$ et $(y_1 - \bar{y}, \ldots, y_n - \bar{y})^t$. □

4. Pour une série bivariée, il est commode de présenter les variances marginales et la covariance dans une matrice carrée de dimension 2, appelée la *matrice de variances-covariances*:
$$S = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}.$$
Cette matrice est toujours symétrique (car $s_{xy} = s_{yx}$ par définition) et de déterminant positif ou nul (det $S = s_x^2 s_y^2 - s_{xy}^2 \geq 0$ par la propriété précédente). De plus, elle est le résultat du produit matriciel suivant:
$$S = \frac{1}{n} M^t M$$
où M est la matrice de n lignes et 2 colonnes contenant les valeurs centrées:
$$M = \begin{pmatrix} x_1 - \bar{x} & y_1 - \bar{y} \\ \vdots & \vdots \\ x_n - \bar{x} & y_n - \bar{y} \end{pmatrix}.$$

La propriété 3 stipule que la covariance d'une série statistique double est toujours (en valeur absolue) inférieure ou égale au produit des écarts-types marginaux. Comme on ne peut pas juger de l'intensité de la relation entre les deux variables rien qu'à l'aide de la valeur de s_{xy}, il semble logique de prendre le produit des écarts-types comme élément de référence. On définit alors le coefficient de corrélation par le rapport suivant:
$$r_{xy} = \frac{s_{xy}}{s_x s_y},$$
si les variances marginales s_x et s_y diffèrent de 0. Si la même transformation affine que celle effectuée à la propriété 2 est à nouveau appliquée, on obtient
$$r'_{xy} = \frac{a\ c}{|a\ c|} r_{xy}.$$
Ce coefficient est donc indépendant des changements d'origine et, dans une certaine mesure, des changements d'échelle effectués sur les observations. La seule perturbation possible est un changement de signe lorsque le signe des valeurs prises par une des deux variables est modifié lors du changement d'échelle. Cette perturbation est cependant naturelle et attendue puisqu'une relation croissante entre deux variables devient nécessairement décroissante si une des deux variables change de signe.

En parlant du signe de r_{xy}, il correspond, par définition, à celui de s_{xy}. De plus, par définition à nouveau, $-1 \leq r_{xy} \leq 1$ et, par Cauchy-Schwarz, r_{xy} est égal à 1 ou -1 uniquement lorsque les points (x_i, y_i) sont tous situés sur une même droite. Le coefficient de corrélation mesure en fait l'intensité de la relation existant entre deux séries d'observations, pour autant que cette liaison soit linéaire ou approximativement linéaire.

Exemple 67 *Pour les deux couples de variables considérés précédemment, on obtient les corrélations suivantes: -0.61 pour le couple* `(Chomage,IndiceRichesse)` *et 0.73 pour le couple* `(PrixMaison,IndiceRichesse)`.

Remarques:

1. Dans la littérature statistique, plusieurs noms sont associés à ce coefficient de corrélation: on parle du coefficient de corrélation *linéaire* ou encore du coefficient de corrélation de Bravais-Pearson.

2. Malgré la présence d'une corrélation importante entre deux séries, il faut éviter toute interprétation abusive. En effet, cette corrélation n'implique pas nécessairement une relation directe de cause à effet entre les deux variables. Cette corrélation est souvent due au fait que les variables étudiées sont soumises à des influences communes.

4.5 Régression

4.5.1 Introduction à la régression linéaire

L'analyse de régression tente de déterminer s'il est possible d'expliquer les valeurs prises par une des deux variables lorsque les valeurs prises par l'autre sont connues. Elle cherche donc à déterminer une relation mathématique entre les deux variables. Lorsque la relation entre les deux variables est linéaire, on parle de *régression linéaire*. C'est ce type de régression que nous envisagerons ici.

L'analyse de régression consiste donc à estimer la relation liant les valeurs x et y des deux variables en recherchant l'équation d'une droite qui s'ajuste au mieux aux valeurs observées.

On peut définir une droite de régression de Y en X (d'équation $y = ax + b$) ou une droite de régression de X en Y (d'équation $x = a'y + b'$), de sorte que les variables X et Y ne jouent plus des rôles symétriques. En effet, dans le cas de la droite de régression de Y en X, on dit que X est la variable explicative et Y la variable dépendante (ou expliquée).

Cette analyse est utile lorsque le diagramme de dispersion a une forme générale linéaire ou approximativement linéaire. Dans certains cas, la relation entre X et Y n'est pas vraiment linéaire mais une simple transformation de l'une ou l'autre (ou les deux) variable peut transformer la relation existante en une relation linéaire. Citons par exemple les relations suivantes:

$$y = b\ a^x \text{ ou } y = b\ x^a$$

Dans le premier cas, en passant au logarithme sur l'ensemble des valeurs de Y (pour autant que celles-ci soient strictement positives), on obtient la relation linéaire $\ln y = \ln b + x\ \ln a$

entre $\ln Y$ et X (relation dont les coefficients sont maintenant $\ln b$ et $\ln a$). Dans le deuxième cas, il faut passer au logarithme sur X et sur Y (pour autant que ce soit possible) pour arriver à une relation linéaire.

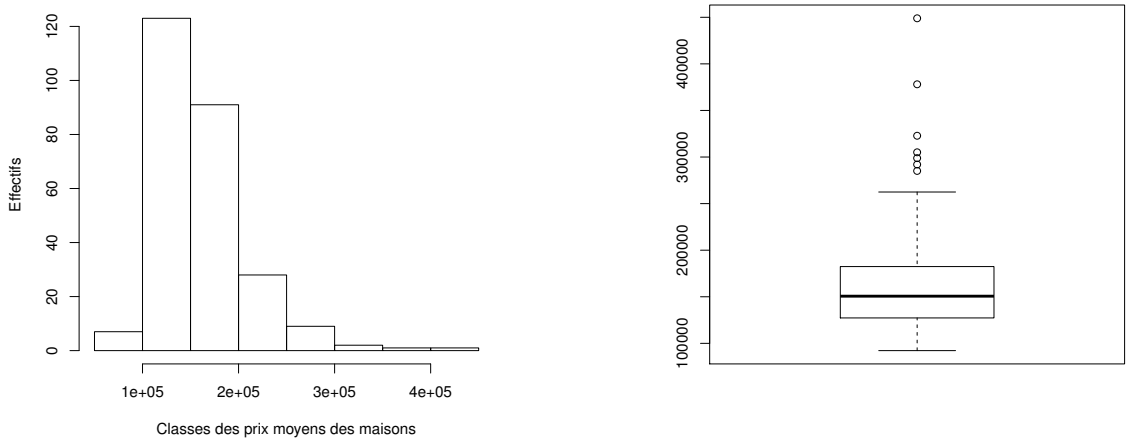


Figure 4.4: Histogramme et boîte à moustaches de la variable `PrixMaison`

Exemple 68 *La transformation logarithmique est très fréquemment exploitée dans les séries économiques pour lesquelles les valeurs observées sont souvent positives et très variables. Prenons par exemple la variable* `PrixMaison`*. L'histogramme et la boîte à moustaches de la Figure 4.4 montrent que certaines valeurs de cette série sont fort grandes par rapport aux autres, sans être aberrantes. Cela traduit simplement un étalement de la distribution vers la droite.*

Néanmoins, cet étalement perturbe la relation linéaire attendue entre l'indice de richesse et cette variable car l'indice de richesse ne varie pas aussi intensément. En passant au logarithme sur les valeurs des prix moyens des maisons, la relation linéaire avec l'indice de richesse est plus nette ainsi qu'illustré à la Figure 4.5. Notons que la corrélation entre l'indice de richesse et le logarithme du prix moyen des maisons est égale à 0.75.

4.5.2 Droite des moindres carrés

Il existe de nombreuses méthodes pour déterminer des valeurs adéquates pour estimer les paramètres a et b de la droite $y = ax + b$ qui pourrait résumer la relation entre Y et X. Afin de dé/Users/gentiane/Desktop/Capture decran 2018-01-26 a 18.03.08.pngterminer la "meilleure droite", il faut préciser le critère à atteindre.

Le critère le plus populaire en statistique est le critère des moindres carrés qui préconise de déterminer a et b de manière à minimiser la somme des carrés des écarts entre les valeurs

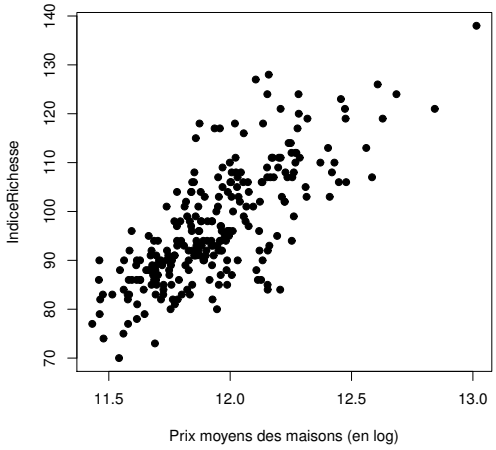


Figure 4.5: Diagramme de dispersion de l'indice de richesse en fonction du logarithme des prix moyens des maisons en Wallonie

observées y_i et les valeurs correspondant aux abscisses x_i et situées sur la droite. Ces valeurs sont appelées valeurs estimées ou ajustées par la régression et sont notées $\hat{y}_i$. Pour tout x_i, l'écart existant entre la valeur observée et la valeur estimée est appelé résidu de la régression et se note $\delta_i = y_i - \hat{y}_i$. Une illustration des ces quantités est donnée à la Figure 4.6.

Les estimations de a et b (notées, par convention, $\hat{a}$ et $\hat{b}$) basées sur ce critère des moindres carrés sont définies à partir de paramètres statistiques déjà rencontrés auparavant ainsi que le développe la Propriété 2.

Propriété 2 *Soit $S = \{(x_i, y_i), 1 \leq i \leq n\}$ une série statistique (quantitative) bivariée. L'équation de la droite de régression obtenue par la méthode des moindres carrés est donnée par*

$$y = \hat{a}x + \hat{b} \text{ où } \hat{a} = \frac{s_{xy}}{s_x^2} \text{ et } \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

Preuve: La démonstration s'obtient en cherchant les valeurs de a et b minimisant la fonction

$$S(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2. \tag{4.2}$$

□

Exemple 69 *Intéressons-nous à nouveau à l'indice de richesse et au logarithme du prix moyen des maisons (voir exemple 68). La droite de régression calculée par la technique des moindres carrés est donnée par*

$$\texttt{IndiceRichesse} = -308.06 + 33.92 \times \ln \texttt{PrixMaison}$$

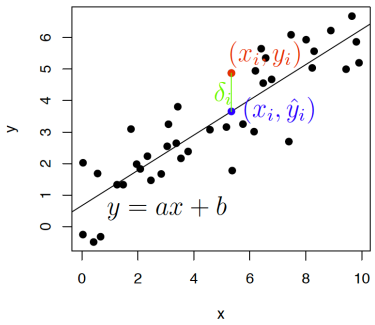


Figure 4.6: Valeurs observées, valeurs ajustées et résidus d'une droite de régression

et est représentée à la Figure 4.7.

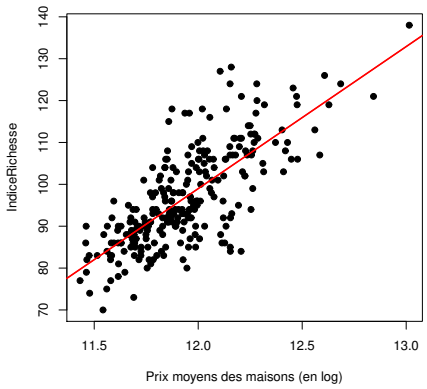


Figure 4.7: Droite de régression par la technique des moindres carrées

Remarques:

1. La droite de régression de X en Y s'obtient de la même manière en minimisant la somme des carrés des écarts horizontaux entre les valeurs observées x_i et les valeurs ajustées $\hat{x}_i$. L'équation de la droite des moindres carrés devient $x = \hat{a}'y + \hat{b}'$ où $\hat{a}' = \frac{s_{xy}}{s_y^2}$ et $\hat{b}' = \bar{x} - \hat{a}'\bar{y}$.

2. Les deux droites de régression se coupent au point $(\bar{x}, \bar{y})$.

3. Notons $m = \hat{a}$ (resp. $m' = \frac{1}{\hat{a}'}$) le coefficient angulaire de la droite de régression de Y en X (resp. de X en Y). On a $\frac{m}{m'} = \hat{a}\hat{a}'$. De plus, à partir des expressions de $\hat{a}$ et $\hat{a}'$, on obtient la relation suivante: $\hat{a}\hat{a}' = r_{xy}^2$. Si $r_{xy}^2 = 1$, alors $\frac{m}{m'} = \hat{a}\hat{a}' = 1 \Leftrightarrow m = m'$.

Les deux droites de régression coïncident lorsque le coefficient de corrélation vaut 1 ou -1.

4.5.3 Analyse des résidus obtenus par la technique des moindres carrés

Après l'ajustement de la droite de régression par la technique des moindres carrés, on dispose de deux nouvelles séries statistiques, à savoir

1. La série des résidus: $S_r = \{\delta_1, \ldots, \delta_n\}$ où $\delta_i = y_i - \hat{y}_i$.
2. La série des valeurs ajustées (ou estimées): $S_{\text{est}} = \{\hat{y}_1, \ldots, \hat{y}_n\}$ où $\hat{y}_i = \hat{a}x_i + \hat{b}$.

La méthode des moindres carrés consiste à minimiser la somme des carrés des résidus (qui sont les écarts entre les valeurs observées et les valeurs ajustées). Ces résidus sont donc censés être petits en valeur absolue. Il est important d'examiner attentivement les résidus obtenus à partir de la droite de régression pour pouvoir juger de l'adéquation de la relation linéaire. La présence d'une structure particulière dans les résidus lorsque ceux-ci sont représentés en fonction des valeurs de la variable explicative doit mettre en doute la relation linéaire.

Exemple 70 *Les résidus correspondant à la droite de régression ajustée à l'exemple 69 sont représentés à gauche dans la Figure 4.8. Aucune structure n'est visible. Par contre, les résidus du second graphique présentent clairement une structure. Ces résidus ont été construits à partir de l'estimation par moindres carrés de la droite de régression expliquant l'indice de richesse en fonction du chômage (voir graphique de gauche de la Figure 4.2). Une relation linéaire n'est pas adéquate car la connaissance du taux de chômage ne semble pas donner la même information selon que le chômage est faible ou plutôt élevé. Et, en effet, les communes dont le taux de chômage est faible ont des résidus plus variables et plus grands que les communes dont le taux de chômage est plus élevé.*

Les paramètres moyenne et variance des séries des résidus et des valeurs ajustées permettent de juger de la qualité de l'ajustement effectué à l'aide de la droite $y = \hat{a}x + \hat{b}$.

Propriété 3 *Si les paramètres a et b de la droite de régression de Y en X sont estimés par la méthode des moindres carrés, alors la moyenne de la série des résidus est nulle et la variance, notée s_δ^2 et appelée variance résiduelle de Y par rapport à X, vaut $s_\delta^2 = s_y^2(1 - r^2)$ où $r = r_{xy}$ est le coefficient de corrélation entre les deux variables.*

De plus, la moyenne de la série des valeurs ajustées est égale à la moyenne des valeurs observées, tandis que sa variance, notée $s_{\hat{y}}^2$, vaut $r^2 s_y^2$.

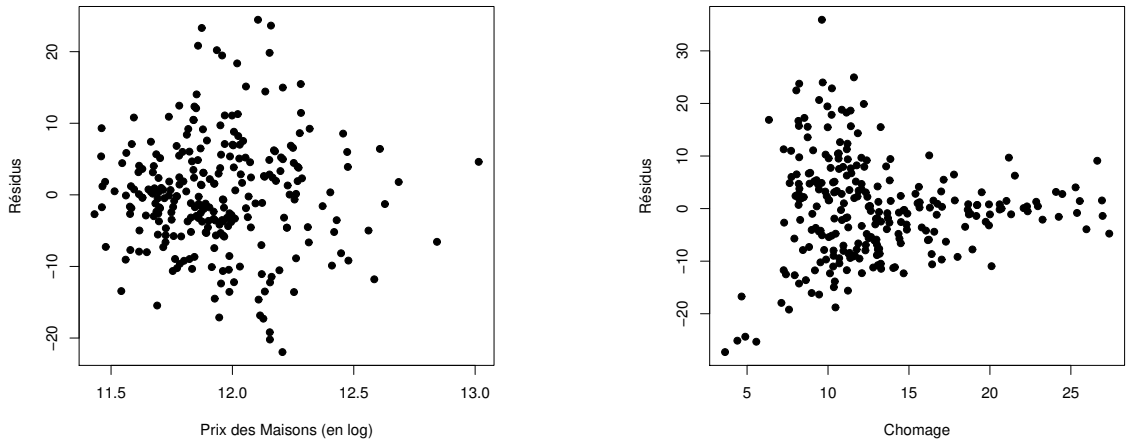


Figure 4.8: Résidus des droites de régression des moindres carrés estimées sur les nuages de points de la Figure 4.2

Preuve: Pour la moyenne de la série des résidus, on a, par définition,

$$\bar{\delta} = \frac{1}{n}\sum_{i=1}^{n}\delta_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b}) = 0$$

en exploitant les définitions de $\hat{a}$ et $\hat{b}$. Pour la variance, on a $s_\delta^2 = \frac{1}{n}\sum_{i=1}^{n}\delta_i^2$ puisque $\bar{\delta} = 0$. De là,

$$\begin{aligned} s_\delta^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left((y_i - \bar{y}) - \hat{a}(x_i - \bar{x})\right)^2 \text{ en exploitant la définition de } \hat{b} \end{aligned}$$

On développe ensuite le carré de la différence pour obtenir, après distribution du facteur $1/n$:

$$s_\delta^2 = s_y^2 - 2\hat{a}s_{xy} + \hat{a}^2 s_x^2.$$

En remplaçant $\hat{a}$ par son expression, on obtient

$$s_\delta^2 = s_y^2 - \frac{s_{xy}^2}{s_x^2} = s_y^2(1 - r^2).$$

Pour la série des valeurs ajustées, on obtient une moyenne égale à $\bar{y}$ en exploitant la propriété $\bar{\delta} = 0$. Pour la variance, en notant que $\hat{b} = \bar{y} - \hat{a}\bar{x}$, il vient

$$s_{\hat{y}}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{a}x_i + \bar{y} - \hat{a}\bar{x} - \bar{y})^2 = \hat{a}^2 s_x^2 = r^2 s_y^2.$$

□

Conséquence: on obtient une décomposition de la variance totale de la série S_Y:

$$s_y^2 = s_{\hat{y}}^2 + s_\delta^2.$$

Le premier terme de la décomposition est la variance expliquée par la régression linéaire (part de la variabilité totale de Y expliquée par la variable X) alors que le second terme rend compte de la variabilité résiduelle (non expliquée par la variable X). On appelle *coefficient de détermination* le paramètre

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

qui correspond au pourcentage de la variance totale expliquée par la régression linéaire de Y en X. Dans le cas particulier de cette régression linéaire simple (une seule variable explicative), R^2 correspond au carré du coefficient de corrélation.

Exemple 71 *Le coefficient de détermination de la droite de régression de l'indice de richesse en fonction du prix moyen des maisons (en logarithme) est égal à 0.55. Dès lors, de l'ordre de 55% de la variabilité observée dans la série des indices de richesse s'explique par la simple connaissance du prix moyen des maisons.*

4.5.4 Manque de robustesse de la droite des moindres carrés

Soient $y_1, \ldots, y_n$ n observations dont on aimerait déterminer une valeur "représentative". Sans information complémentaire, la valeur représentative la plus classique est la moyenne qui, pour rappel, est le paramètre de tendance centrale solution du problème d'optimisation des moindres carrés défini par

$$\min_{b \in \mathbb{R}} \sum_{i=1}^{n} (y_i - b)^2$$

Si on dispose maintenant d'information complémentaire par l'intermédiaire des valeurs d'une variable explicative X, la valeur ajustée de Y sachant que $X = x_i$ est donnée par $\hat{a}x_i + \hat{b}$ avec $\hat{a}$ et $\hat{b}$ solutions du problème d'optimisation des moindres carrés défini par

$$\min_{(a,b) \in \mathbb{R}^2} \sum_{i=1}^{n} (y_i - b - ax_i)^2$$

On a déjà constaté l'effet qu'une observation extrême peut avoir sur la moyenne arithmétique. Etant donné que la droite de régression des moindres carrés est définie à partir du même type de critères, il est naturel d'imaginer un impact non négligable sur celle-ci lorsque les données contiennent des observations atypiques.

Dans le contexte de la régression, on distingue différents types d'observations atypiques:

- les points aberrants verticaux: ils correspondent à une valeur de la variable dépendante qui s'écarte fortement des autres valeurs de Y observées pour des valeurs similaires de X;
- les *bons* points de levier: ils correspondent à une valeur extrême de la variable explicative, tout en vérifiant la relation linéaire globale;
- les *mauvais* points de levier: ils correspondent à une valeur extrême de la variable explicative, tout en ne vérifiant pas la relation linéaire globale;

Ceux-ci sont illustrés à la Figure 4.9

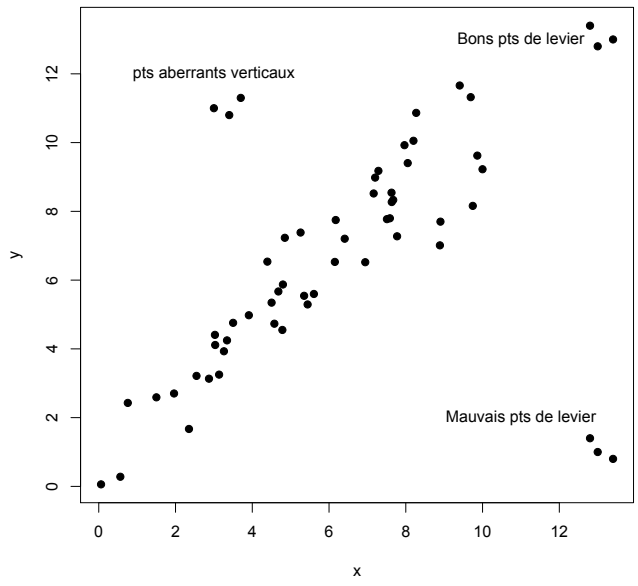


Figure 4.9: Différents points atypiques en régression linéaire

Si on ajuste une droite de régression par la technique des moindres carrés lorsque les données contiennent des points aberrants verticaux ou des mauvais points de levier (même en nombre très limité), l'estimation peut être fortement perturbée, comme par exemple l'estimation par moindres carrés illustrée à gauche à la Figure 4.10. Si les trois points de levier sont supprimés, la droite de régression se trouve à la position attendue, au centre du nuage principal des données (voir l'illustration de droite de la même figure).

Vu la présence des carrés dans la fonction objective de la technique des moindres carrées, cette procédure doit, pour minimiser la fonction, empêcher les résidus trop importants, quitte à effectuer un effet de levier sur la droite pour réduire les résidus potentiellement trop grands.

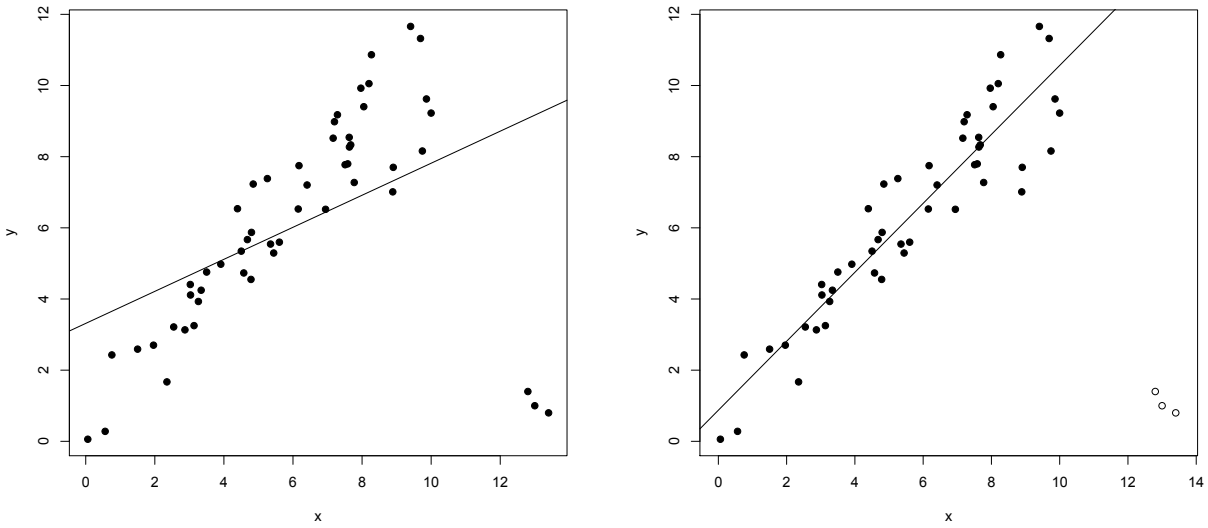


Figure 4.10: Droites de régression estimées par la technique des moindres carrés sur toutes les données (à gauche), sur les données sans les points de levier (à droite)

Plusieurs propositions de "robustifications' de la technique des moindres carrés (L_2) ont été proposées dans la littérature. Celles-ci se décomposent en deux approches principales:

1. Modification de la fonction des résidus:

   Par définition, les moindres carrés considèrent la fonction

   $$\min_{a,b\in\mathbb{R}} \sum_{i=1}^{n} r_i^2(a,b) \Leftrightarrow \min_{a,b\in\mathbb{R}} \sum_{i=1}^{n} \rho(r_i(a,b))$$

   avec $\rho : t \to t^2$. En changeant la fonction ρ, la technique peut être robustifiée. Notamment, $\rho(t) = |t|$ donne la technique L_1. Beaucoup d'autres propositions existent dans la littérature!

2. Elimination des (trop) grands résidus:

   La non résistance de la technique des moindres carrés s'explique aussi par le fait qu'elle prend en compte tous les résidus. Il serait possible de "tronquer" la somme comme suit:

   $$\min_{a,b\in\mathbb{R}} \sum_{i=1}^{h} r_{(i)}^2(a,b)$$

   pour un nombre h (supérieur à $n/2$) à fixer. On obtient ainsi le critère LTS (Least Trimmed Squares).

4.6 Exercices

1. Le Tableau de contingence 4.4 décrit la répartition d'une population constituée de 535 ménages selon les deux variables suivantes : X représente le nombre de pièces de l'habitation et Y correspond au nombre d'enfants du ménage.

Tableau 4.4: Tableau de contingence pour la série double "Nombre de pièces - Nombre d'enfants".

	Valeurs de Y				
Valeurs de X	0	1	2	3	4
1	7	3	2	1	0
2	24	32	21	2	1
3	16	35	54	26	4
4	9	28	74	55	12
5	4	12	46	13	12
6	2	6	16	11	7

   (a) Que représentent les effectifs n_{23} et n_{54} ? Calculer les fréquences correspondantes.

   (b) Déterminer les distributions marginales des variables X et Y. Calculer les moyennes, médianes et variances marginales ainsi que les modes marginaux.

   (c) Déterminer la distribution conditionnelle de la variable Y sachant que le nombre de pièces du logement est égal à 4. Calculer la moyenne et la variance de cette distribution conditionnelle.

   (d) Calculer la covariance de la série double.

2. Le directeur d'une entreprise vinicole a l'habitude d'offrir à ses employés une prime de fin d'année de 10, 15, 20 ou 25 unités monétaires. Le tableau 4.5 représente le tableau de contingence mettant en rapport la variable Y = "montant de la prime" avec la variable X = "taille de la cave à vin exprimée en nombre de bouteilles".

   (a) Quelle est la prime le plus souvent distribuée ? Que représente cette valeur dans la distribution marginale de Y?

   (b) Quelle est la somme totale dépensée par le directeur pour offrir les primes de fin d'année?

   (c) Si le directeur voulait être équitable par rapport à tous ses employés, quel montant donnerait-il à chacun? Cette valeur est-elle proche des moyenne et médiane marginales de Y?

Tableau 4.5: Tableau de contingence pour la série double "Taille de la cave - Montant de la prime".

	Y			
X	10	15	20	25
$[0, 100]$	11	13	5	3
$]100, 200]$	14	21	15	6
$]200, 300]$	5	8	9	7
$]300, 400]$	3	6	8	9
$]400, 500]$	2	3	5	4

(d) Comparer les nombres moyens de bouteilles détenues par les employés en tenant compte des différentes primes perçues.

3. Le tableau 4.6 reprend les âges de l'époux (variable X) et de l'épouse (variable Y) pour les 100 derniers mariages enregistrés dans une ville.

Tableau 4.6: Ages des époux pour les 100 derniers mariages enregistrés dans une ville

	Classes de Y			
Classes de X	$[18, 22]$	$]22, 26]$	$]26, 30]$	$]30, 34]$
$[20, 24]$	14	5	2	0
$]24, 28]$	15	19	7	3
$]28, 32]$	5	12	5	2
$]32, 36]$	0	1	2	2
$]36, 40]$	0	1	3	2

(a) Calculer l'âge moyen des époux et des épouses au moment du mariage.

(b) Comment peut-on obtenir à partir du tableau de contingence les informations suivantes: (1) 38% des femmes se marient entre 22 et 26 ans; (2) 19% des mariages concernent des hommes âgés de 24 à 28 ans et des femmes âgées de 22 à 26 ans; (3) 31,6% des femmes âgées de 22 à 26 ans se sont mariées avec des hommes de 28 à 32 ans.

4. Le gérant d'un magasin d'appareils électroménagers a enregistré chaque semaine le nombre x_i de centaines d'appels téléphoniques reçus de l'extérieur et le chiffre d'affaires

y_i réalisé en unités monétaires. Les résultats sont repris dans le tableau 4.7 où n_{ij} indique le nombre de semaines où le magasin a reçu x_i centaines d'appels téléphoniques et a fait y_j UM comme chiffre d'affaires.

Tableau 4.7: Tableau de contingence pour la série double "Nombre d'appels téléphoniques - Chiffre d'affaires".

Nombre d'appels téléphoniques	Chiffre d'affaires 1	2	3	4	5
2	9	5	3	1	0
3	4	5	7	3	1
4	0	6	9	6	3
5	0	5	14	14	5

(a) Combien de semaines cette enquête a-t-elle duré?

(b) Combien d'appels téléphoniques le magasin reçoit-il en moyenne par semaine?

(c) Déterminer les distributions marginales ainsi que les moyennes et variances marginales.

(d) Comparer les chiffres d'affaires moyens réalisés lorsque le magasin reçoit 2 ou 5 appels téléphoniques sur la semaine.

(e) Calculer la covariance et le coefficient de corrélation linéaire entre les deux variables.

5. Une étude statistique a été réalisée auprès des 500 entreprises d'un même secteur industriel. Deux variables étaient considérées: la variable X correspond à la taille de l'entreprise (en nombre de salariés) et la variable Y au niveau de leur salaire mensuel (en unités monétaires). L'étude a permis d'obtenir les renseignements du tableau 4.8.

   (a) Déterminer les salaires mensuels moyen et médian des employés des entreprises considérées.

   (b) Déterminer, pour chaque classe de taille d'entreprise, le salaire mensuel moyen.

   (c) Quelle relation existe-t-il entre le salaire moyen global calculé en (a) et les salaires moyens calculés en (b).

6. Soit S la série double suivante, due à Anscombe (1973) :

x_i	10	8	13	9	11	14
y_i	8,04	6,95	7,58	8,81	8,33	9,96

Tableau 4.8: Répartition de 500 entreprises en fonction du nombre de salariés et du salaire mensuel

Nombre de salariés (X)	Niveau de salaire mensuel (Y) 5000	10 000	30 000
$[0, 500]$	13	5	2
$]500, 1500]$	60	25	15
$]1500, 3000]$	90	30	30
$]3000, 5000]$	137	65	28

x_i	6	4	12	7	5
y_i	$7,24$	$4,26$	$10,84$	$4,82$	$5,68$

(a) Représenter le nuage de points de cette série bivariée.

(b) Rechercher l'équation de la droite des moindres carrés. Calculer les résidus.

7. Au Grand-Duché de Luxembourg, on a relevé les taux de chômage suivants:

Année	Taux de chômage en %
75	4
76	4,4
77	5
78	5,2
79	5,9
80	6,3
81	7,4
82	8,1
83	8,3
84	9,7
85	10,2
86	10,4
87	10,5

(a) A partir de ces données, déterminer quel type de liaison fonctionnelle $y = f(x)$ existe entre le taux de chômage y et l'année x.

(b) Rechercher l'équation de la droite des moindres carrés exprimant y en fonction de x et calculer le coefficient de corrélation linéaire. Interpréter le coefficient de détermination.

(c) Estimer les taux de chômage livrés par la droite des moindres carrés et calculer les résidus. Commenter la situation des années 81, 86 et 87. Estimer le taux de chômage pour les années 2150 et 1965; commenter ces résultats et en tirer des conclusions pratiques.

8. Dans les années 1950, beaucoup de recherches statistiques ont tenté d'établir un lien entre la consommation de cigarettes et le cancer du poumon. Une des enquêtes s'est concentrée sur 11 pays et sur les variables suivantes:

   X est le nombre de cigarettes par habitant du pays en 1930

   Y est le taux (calculé sur une base fixée) de mortalité due au cancer du poumon parmi la population du pays lors de l'année 1950.

   Les données sont indiquées dans le tableau ci-dessous.

Pays	X	Y
Australie	455	170
Canada	510	150
Danemark	380	165
Etats-Unis	1280	190
Finlande	1115	350
Grande-Bretagne	1145	465
Hollande	460	245
Islande	220	58
Norvège	250	90
Suède	310	115
Suisse	530	250

(a) Considérer la série univariée S_Y. Construire une boîte à moustaches et commenter ses caractéristiques principales.

(b) Considérer la série double $S = \{(x_i, y_i); 1 \leq i \leq 11\}$. On donne les résumés numériques suivants:

$$\sum_{i=1}^{11} x_i = 6\ 655,\ \sum_{i=1}^{11} y_i = 2\ 248,\ \sum_{i=1}^{11} x_i^2 = 5\ 503\ 675,$$

$$\sum_{i=1}^{11} y_i^2 = 600\ 664, \sum_{i=1}^{11} x_i y_i = 1\ 698\ 535.$$

i. Une relation linéaire entre les deux variables semble-t-elle plausible? Construire le nuage de points et justifier.

ii. Déterminer la droite de régression de Y en X obtenue par la méthode des moindres carrés. Représenter cette droite sur le nuage de points et commenter.

iii. Calculer le pourcentage de la variance de Y expliquée par la régression linéaire et commenter.

iv. Si, en Belgique, le nombre de cigarettes consommées par habitant en 1930 s'élève à 530, quel est le taux de mortalité due au cancer du poumon attendu en 1950?

v. Les Etats-Unis et la Grande-Bretagne se distinguent quelque peu des autres observations. Construire deux nouvelles séries de données S_{EU} et S_{GB} contenant les observations initiales sauf respectivement les Etats-Unis ou la Grande-Bretagne. Calculer, pour chaque série, les coefficients de la droite de régression de Y en X obtenue par la méthode des moindres carrés.

vi. Calculer le coefficient de détermination de chaque série et commenter.

vii. Comparer le résidu de la Grande-Bretagne calculé à partir de ces nouvelles droites avec le résidu obtenu pour la série complète.

Bibliographie

Bragard, L., et Alexandre, P., *Statistique descriptive à l'usage des sciences humaines*, Editions Derouaux, 1995.

Dagnelie, P., *Statistique théorique et appliquée. Tome 1: statistique descriptive et base de l'inférence statistique*, De Boeck, 1998.

Dehon, C., Droesbeke, J.J., et Vermandele, C., *Eléments de statistique*, Editions Ellipses (4ème édition), 2008.

Dodge, Y., *Premiers pas en statistique*, Springer-Verlag, France, 1999.

Dodge, Y., *Statistique: Dictionnaire encyclopédique*, Springer-Verlag, 2007.

Grimmett, G.R., et Stirzaker, D.R., *Probability and random processes*, Oxford University Press, 1992.

Hogg, R.V., McKean, J.W., et Craig, A.T., *Introduction to Mathematical Statistics*, Pearson Education (6ème édition), 2005.

Leboeuf, Ch., Roque, J.L., et Guegand, J, *Cours de probabilité et de statistique*, Editions Ellipses (2ème édition), 1987.

Mood, A.M., Graybill, F.A., et Boes, D.C., *Introduction to the theory of statistics*, McGraw-Hill (3ème édition), 1974.

Tukey, J. W., *Exploratory Data Analysis*, Broché, 1977