Introduction à la méthode statistique

Statistique et probabilités

Cours et exercices corrigés

Bernard Goldfarb
Catherine Pardoux

7e édition

DUNOD

ISBN 978-2-10-059167-1

Table des matières

Avant-propos

Tout le monde sait et dit que celui qui observe sans idée, observe en vain.
Éléments de philosophie, Alain (1868 – 1951)

Le recueil, le traitement et l'analyse de l'information sont au cœur de tous les processus de gestion et de décision. L'enrichissement et le développement des méthodes de description, de prévision et de décision ont ainsi contribué à positionner la statistique appliquée[1] au carrefour de l'observation et de la modélisation.

L'utilisation des méthodes statistiques s'est généralisée avec le développement et l'interprétation des résultats fournis par les logiciels et progiciels (généralistes ou spécialisés) assurant la gestion des données, les calculs, les représentations graphiques…

Depuis plus de vingt ans, les logiciels statistiques[2] ont considérablement modifié l'analyse statistique des données, et maintenant l'enseignement de la statistique. Sous peine d'être noyé, non plus dans les calculs mais dans les résultats, l'utilisateur doit disposer d'idées précises sur les outils, leurs fonctions et leurs champs d'application.

Nous avons voulu guider les futurs utilisateurs de données vers les descriptions statistiques et les représentations courantes rencontrées dans tous les domaines de l'activité humaine.

Une brève introduction illustrée à la pratique et à l'usage de la simulation, donne une première vue sur un outil incontournable dans des secteurs tels que la logistique, la stratégie, ou encore l'analyse financière…

La visualisation par tableaux et graphiques[3] est une clef indispensable pour traiter et comprendre efficacement les multiples ensembles de données statistiques ; l'usage généralisé qui en est fait pour tous les publics et par de très nombreux médias confirme son importance.

1. À laquelle les programmes, tant de l'enseignement secondaire que de l'enseignement supérieur, accordent une place de plus en plus importante.
2. Sans compter les langages de programmation et bibliothèques de programmes, comme le logiciel libre R, qui mettent la pratique de très nombreux traitements à la portée du plus grand nombre.
3. La représentation visuelle est remarquablement mise en valeur dans le très bel ouvrage de Edward R. Tufte, *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut, 1991.

Dans cette 7e édition, nous avons maintenu toute notre attention sur les visualisations, et nous avons accentué les illustrations de la pratique du tableur Excel® largement répandu.

La théorie reste volontairement limitée pour donner toute son importance à l'approche interprétative des données. Le lecteur, selon ses connaissances préalables et son intérêt pour la formalisation, pourra sauter en première lecture la présentation de certains supports théoriques. Ce livre n'est qu'une introduction à la méthode statistique, et nous donnons quelques références d'ouvrages pour élargir idées et connaissances.

Les données de nombreux exemples ont été remises à jour. Les exercices ont été, pour la plupart, renouvelés. Des exercices complémentaires (avec leurs corrigés) sont aussi disponibles en ligne sur le site des éditions Dunod (www.dunod.com).

Ce livre est issu de nombreuses expériences d'enseignement en formation initiale comme en formation continue notamment pour des étudiants en sciences économiques, en sciences de gestion, et en informatique de gestion ; il tient compte de leurs besoins et des dernières évolutions.

Nous remercions par avance les lectrices et les lecteurs qui voudront bien nous faire part de leurs remarques ou suggestions.

Bernard Goldfarb
Catherine Pardoux

1. Distributions statistiques à un caractère

Le savant doit ordonner ; on fait la science avec des faits
comme une maison avec des pierres ;
mais une accumulation de faits n'est pas plus une science
qu'un tas de pierres n'est une maison.

La Science et l'hypothèse, Henri Poincaré (1854-1912)

La statistique descriptive est un ensemble de méthodes permettant de décrire, présenter, résumer des données souvent très nombreuses. Ces méthodes peuvent être numériques (tris, élaboration de tableaux, calcul de moyennes...) et/ou mener à des représentations graphiques.

I. Définitions

A. Population, individu, échantillon

Une *population* est l'ensemble des éléments auxquels se rapportent les *données* étudiées. En statistique, le terme « population » s'applique à des ensembles de toute nature : étudiants d'une académie, production d'une usine, poissons d'une rivière, entreprises d'un secteur donné…

Des enquêtes de l'Office statistique des communautés européennes donnent la durée hebdomadaire moyenne du travail des salariés à temps complet pour 15 pays membres. Les résultats de ces enquêtes ne donnent pas d'information « atomisée » à un niveau plus bas que le pays ; la population de référence n'est donc pas ici l'ensemble (plusieurs millions) de tous les salariés des 15 pays. L'étude de ces 15 observations concerne un ensemble

de 15 *unités* (*statistiques*), les 15 pays sélectionnés qui constituent la *population* de l'étude.

Dans une population donnée, chaque élément est appelé « individu » ou « unité statistique ».

La collecte d'informations sur une population peut être effectuée sur la totalité des individus ; on parle alors d'enquêtes *exhaustives*. Lorsque la taille de la population étudiée est élevée, de telles enquêtes sont fort coûteuses ou impossibles, et le cas échéant, leurs résultats souvent très longs à rassembler peuvent être dépassés avant même la fin de l'enquête. C'est la raison pour laquelle on a souvent recours aux enquêtes par *sondage* qui portent sur une partie de la population appelée *échantillon.* Les observations obtenues sur une population ou sur un échantillon constituent un ensemble de données auxquelles s'appliquent les méthodes de la statistique descriptive dont le but est de décrire le plus complètement et le plus simplement l'ensemble des observations qu'elles soient relatives à toute la population ou seulement à un sous-ensemble.

B. Variables

Chaque individu d'une population peut être décrit selon une ou plusieurs *variables* qui peuvent être des caractéristiques qualitatives ou prendre des valeurs numériques.

Une variable est dite *qualitative* si ses différentes réalisations (modalités) ne sont pas numériques. Ainsi : le sexe, la situation matrimoniale, la catégorie socioprofessionnelle… sont des variables qualitatives. On peut toujours rendre numérique une telle variable en associant un nombre à chaque modalité ; on dit alors que les modalités sont codées. Bien entendu, les valeurs numériques n'ont dans ce cas aucune signification particulière, et effectuer des opérations algébriques sur ces valeurs numériques n'a pas de sens.

Une variable est dite *quantitative* lorsqu'elle est intrinsèquement numérique : effectuer des opérations algébriques (addition, multiplication…) sur une telle variable a alors un sens. Une variable quantitative peut être une variable statistique discrète ou continue.

Les *variables statistiques discrètes* sont des variables qui ne peuvent prendre que des valeurs isolées, discrètes. Le nombre d'enfants d'une famille, le nombre de pétales d'une fleur, le nombre de buts marqués lors d'une rencontre de football… sont des variables quantitatives discrètes. Le plus fréquemment, les valeurs possibles sont des nombres entiers.

Les *variables statistiques continues* peuvent prendre toutes les valeurs numériques possibles d'un ensemble inclus dans $\mathbb{R}$: le revenu, la taille, le taux de natalité sont des variables continues.

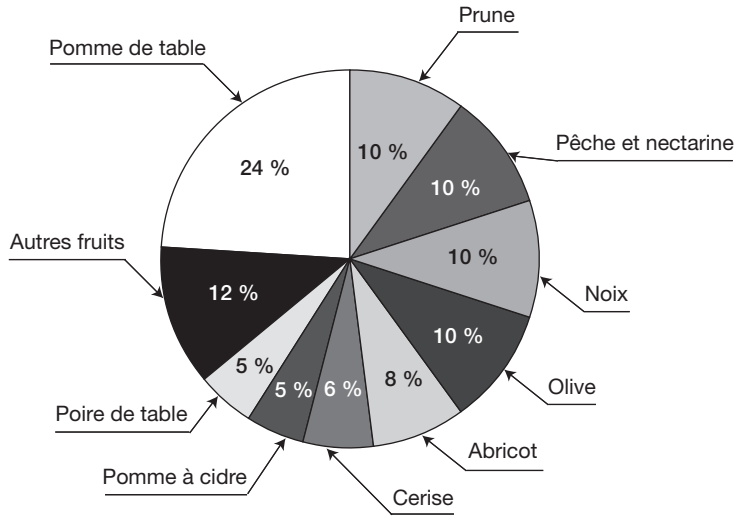
La distinction entre variables quantitatives discrètes et continues peut paraître factice, car toute mesure est discrète en raison d'une précision toujours limitée ; et inversement, lorsqu'une variable discrète peut prendre un grand nombre de valeurs et que la taille de la population (ou de l'échantillon) étudiée est élevée, on regroupera des valeurs voisines et la variable sera, par extension, traitée comme une variable continue. En pratique, lorsque les valeurs d'une variable sont regroupées en k classes, la variable est traitée comme une variable quantitative continue, mais elle peut aussi être envisagée comme une variable qualitative à k modalités.

Les données dont on dispose sont les modalités ou valeurs prises par plusieurs variables qualitatives ou quantitatives sur les individus d'une population ou d'un échantillon ; pour une population d'entreprises, on peut disposer, par exemple, de données sur le chiffre d'affaire, le bénéfice net, le nombre d'employés, la masse salariale annuelle, le secteur d'activité principale…

On peut, dans un premier temps, décrire chaque variable séparément, puis ensuite, étudier les relations ou liaisons existantes entre elles. Ainsi, dans ce livre, nous envisagerons d'abord les populations statistiques décrites selon une seule variable, puis selon deux variables. L'étude des populations caractérisées par plus de deux variables n'est pas abordée dans cet ouvrage.

II. Représentations graphiques

Deux méthodes de représentation des données vont être exposées. On commencera par celles adaptées aux données nombreuses et/ou anonymes, c'est-à-dire pour lesquelles l'identité des individus n'a pas été relevée ou ne présente pas d'intérêt à être conservée pour l'interprétation. Ceci n'est pas le cas lorsque les individus sont peu nombreux (régions, pays…), où on définira un nouveau mode de représentation graphique dû à J.W. Tukey (§ II.B.). L'étude d'une population selon une variable sera restreinte au cas des variables quantitatives, car la description d'une population selon une variable qualitative est totalement résumée dans un tableau de pourcentages ou dans un diagramme circulaire, appelé aussi diagramme en « camembert » (*cf.* figure 1.1).



Extrait de Agreste, *GraphAgri 2006*,
Ministère de l'Agriculture et de la Pêche.

Figure 1.1 – Surface du verger français en 2005

A. Distributions statistiques et représentations graphiques

Considérons une variable observée sur une population $\mathbb{P}$ de n individus. Si la variable X prend k valeurs ou ensembles de valeurs (appelés dans ce qui suit, modalités), le premier traitement des données brutes consiste à compter le nombre n_i d'individus qui présentent la i^e modalité ($i = 1, 2, \ldots, k$).

1) Variables statistiques discrètes

Les résultats concernant les observations de la variable X dont l'ensemble des valeurs est $\{x_i\ ,\ i = 1,\ldots\ ,\ k\}$, sont présentés dans le tableau des effectifs $(x_i\ ,\ n_i)$ ou dans le tableau des fréquences $(x_i\ ,\ f_i)$ avec $f_i = n_i/n$ (on utilise souvent le pourcentage $100 \cdot f_i$). Il est préférable de calculer les fréquences à partir des effectifs cumulés (§ II.A.3) afin que des erreurs successives d'arrondis ne donnent pas une somme totale de fréquences différente de 1.

Tableau des effectifs

Modalité	Effectif
x_1	n_1
$\vdots$	$\vdots$
x_i	n_i
$\vdots$	$\vdots$
x_k	n_k
	$\sum_{i=1}^{k} n_i = n$

Tableau des fréquences

Modalité	Fréquence
x_1	$f_1 = n_1/n$
$\vdots$	$\vdots$
x_i	$f_i = n_i/n$
$\vdots$	$\vdots$
x_k	$f_k = n_k/n$
	$\sum_{i=1}^{k} f_i = 1$

On présente logiquement les modalités numériques en ordre croissant. On peut associer à ces tableaux une représentation graphique appelée « diagramme en bâtons ».

Un *diagramme en bâtons* (*cf.* figure 1.2) est construit dans un système d'axes rectangulaires ; les valeurs de la variable statistique X sont portées en abscisse ; à partir de chaque valeur x_i, on trace un segment de droite vertical et dont la hauteur est proportionnelle à l'effectif correspondant. On peut retenir indifféremment une échelle qui explicite les effectifs n_i, ou une échelle qui explicite les fréquences f_i. Pour les distributions du tableau 1.1, on pourrait représenter sur le même graphique les diagrammes en bâtons de plusieurs pays avec des couleurs différentes, chaque couleur correspondant à un pays, ce qui permettrait de comparer les distributions du nombre de personnes par ménage.

Tableau 1.1 – Ménages suivant le nombre de personnes du ménage dans quelques pays en 1995 (%)

	Allemagne	Espagne	Finlande	France	Grèce	Irlande	Italie	Pays-Bas	Portugal
Ménages de :									
– 1 personne	34,4	12,7	37,4	29,2	20,7	22,8	22,7	30,6	13,7
– 2 personnes	32,3	24,5	31,0	31,8	28,9	23,1	23,1	34,0	26,4
– 3 personnes	16,0	21,8	14,4	16,8	19,8	15,6	15,6	13,4	24,7
– 4 personnes	12,6	24,0	11,9	14,2	21,7	17,1	17,1	15,9	22,8
– 5 personnes et plus	4,7	17,0	5,3	8,0	8,9	21,4	21,4	6,2	12,4
Ensemble (en milliers)	34 413	12 112	2 222	23 126	3 756	1 146	1 146	6 425	3 275

Source : Tableaux de l'Économie Française 1999-2000, INSEE.

Nombre de personnes	f_i (%)
1	29,2
2	31,8
3	16,8
4	14,2
5 ou plus	8,0
	100 %

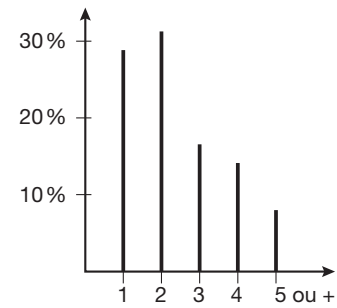


Figure 1.2 – Diagramme en bâtons – *Nombre de personnes par ménage en France en 1995*

2) Variables statistiques continues

L'infinité des valeurs observables ne rend pas possible la généralisation du diagramme en bâtons. *Le domaine de variation* d'une *variable statistique continue X* est partagé en *k parties*. L'intervalle $[\, x_{i-1}\, ,\, x_i\, [$ fermé à gauche, ouvert à droite, est appelé i^{e} classe ($i = 1, 2, \ldots, k$) ; son amplitude est égale à :

$$a_i = x_i - x_{i-1}$$

Il arrive que l'amplitude des classes extrêmes soit indéterminée : la première classe étant définie par « *moins de…* », et la dernière par « *plus de…* » (*cf.* tableau 1.2).

Le choix des *extrémités* des classes se fait à partir des données brutes ; le nombre *k* de classes doit être modéré (usuellement entre 4 et 10). Le découpage en classes est assez souvent choisi tel que l'amplitude des classes soit constante, ou tel que les effectifs des classes soient constants (par exemple, 10 % de la population dans chaque classe, *cf.* tableau 1.6).

Le classement d'une série statistique correspond à une perte d'information par rapport aux données initiales puisque seuls les effectifs des classes sont retenus. Le travail sur une telle série impose alors l'hypothèse que les données sont réparties *uniformément* à l'intérieur de chacune des classes. On parle aussi d'*équirépartition* des individus ou encore d'*homogénéité* dans chacune des classes. Chaque partie de la classe correspond alors à un effectif proportionnel à sa longueur. L'idée est, bien sûr, que chaque classe représente une entité qui doit se distinguer par rapport aux autres classes. Comme précédemment, les résultats sont présentés dans un tableau d'effectifs ou de fréquences. On associe à un tel tableau un *histogramme* qui est une représentation graphique très répandue. L'histogramme est constitué de la juxtaposition de rectangles (pour respecter l'hypothèse d'équirépartition) dont les bases représentent les différentes classes et dont *les surfaces sont proportionnelles aux effectifs des classes* (*cf.* figure 1.3).

On verra par la suite qu'une difficulté du travail avec des séries classées est le choix des limites pour les classes extrêmes, indispensable aussi pour le tracé de l'histogramme.

À la i^e classe, correspond un rectangle dont la base est l'intervalle $[x_{i-1}, x_i[$ et dont la surface est proportionnelle à la fréquence f_i (ou à l'effectif n_i). Si les classes ont toutes la même amplitude, les hauteurs des rectangles sont proportionnelles aux fréquences. Dans le cas où les classes sont d'amplitudes inégales, la hauteur du rectangle correspondant à la i^e classe d'amplitude a_i sera $h_i = f_i / a_i$. La surface du rectangle représentant la i^e classe sera ainsi égale à f_i

Pour une série d'observations relatives à une variable statistique X discrète ou continue classée, la donnée des modalités et de leurs fréquences est appelée « *distribution statistique* » de la variable X.

Tableau 1.2 – Chômeurs BIT selon le sexe et l'ancienneté de chômage en septembre 2006

	Distribution en milliers		**Distribution en pourcentage**	
Ancienneté d'inscription	Hommes	Femmes	Hommes	Femmes
Moins d'un mois	180,3	181,0	16,5	16,8
D'un à moins de trois mois	203,9	204,9	18,6	19,0
De trois à moins de six mois	169,3	163,1	15,5	15,1
De six mois à moins d'un an	202,1	191,1	18,5	17,7
D'un à moins de deux ans	197,3	199,3	18,0	18,5
De deux à moins de trois ans	74,5	75,4	6,8	7,0
Trois ans ou plus	67,1	62,9	6,1	5,8
Ensemble	1 094,5	1 077,7	100	100
Ancienneté moyenne en jours	341	334		

Source : Bulletin Mensuel des Statistiques du Travail, www.travail.gouv.fr, octobre 2006.

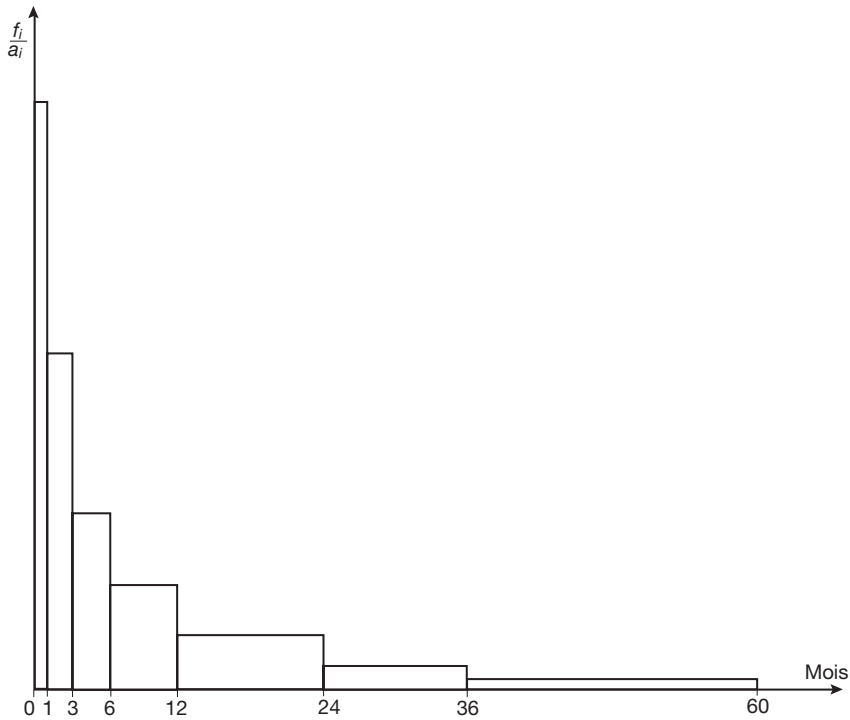


Figure 1.3 – Histogramme de la distribution des chômeurs « Femmes » selon l'ancienneté (voir tableau 1.2)

La classe « Trois ans ou plus » est supposée bornée supérieurement par 5 ans (60 mois).

3) Fréquences cumulées et courbe cumulative

a) Tableau des fréquences cumulées

Les tableaux de fréquences (ou d'effectifs) qui viennent d'être définis peuvent être modifiés de facon à présenter un résumé des données sous une forme différente.

On appelle *effectif cumulé* de la i^e classe, le nombre d'individus N_i pour lesquels la variable prend une valeur inférieure à x_i :

$$N_i = \sum_{j \leq i} n_j \text{ pour } i = 1, 2, \ldots, k$$

On définit de même F_i, la *fréquence cumulée* de la i^e classe : $F_i = N_i/n$

Les tableaux d'effectifs cumulés ou de fréquences cumulées se déduisent des tableaux d'effectifs ou de fréquences (non cumulés) en substituant aux effectifs ou fréquences non cumulés les effectifs ou fréquences cumulés. Les deux types de tableaux sont donc équivalents (cf. figures 1.2 et 1.4).

b) Fonction cumulative et courbe cumulative

La *courbe cumulative* ou courbe des fréquences cumulées est la représentation graphique des fréquences cumulées. Plus précisément, la courbe cumulative est la représentation graphique de la proportion *F(t)* des individus de la population dont le caractère prend une valeur inférieure à *t*. Cette fonction, appelée *fonction cumulative* ou *fonction de répartition*, est :

1. définie pour tout $t \in \mathbb{R}$

2. croissante (mais non strictement croissante)

3. nulle pour t inférieur à $\min_{1 \le i \le n} x_i$

4. égale à 1 pour t au moins égal à $\max_{1 \le i \le n} x_i$

Pour une variable statistique *discrète*, cette fonction est une *fonction en escalier*, présentant en chacune des valeurs possibles x_i, un saut égal à la fréquence correspondante f_i (*cf.* figure 1.4).

Dans le cas d'une variable statistique *continue*, la fonction cumulative n'est connue que pour les valeurs de *X* égales aux extrémités des classes. L'hypothèse d'équirépartition (§ II.A.2) implique que la fonction *F* est linéaire entre ces valeurs (*cf.* figure 1.5). *Cette fonction est donc continue et linéaire par morceaux.* Ici encore, il est nécessaire de choisir des limites pour les classes extrêmes.

t	*F(t)* (%)
< 1	0
[1 ; 2[	29,2
[2 ; 3[	61,0
[3 ; 4[	77,8
[4 ; 5[	92,0
≥ 5	100

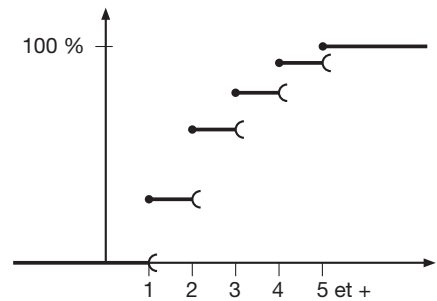


Figure 1.4 – Graphe des fréquences cumulées de la distribution représentée à la figure 1.2

Ces fréquences cumulées sont des fréquences cumulées *ascendantes*, car elles ont été obtenues en calculant les fréquences F_i d'individus pour lesquelles le caractère étudié *X* est *au plus* égal à x_i ; on peut aussi définir les fré-

t	*F(t)* (%)
0	0
1	16,8
3	35,8
6	50,9
12	68,7
24	87,2
36	94,2
60	100

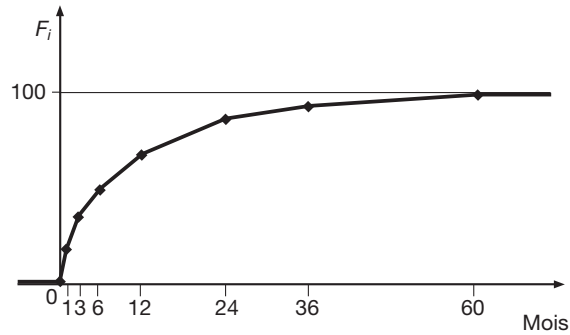


Figure 1.5 – Courbe cumulative de la distribution représentée à la figure 1.3

quences cumulées *descendantes*, c'est-à-dire les fréquences pour lesquelles le caractère étudié X est supérieur à x_i. Quand on ne spécifie pas le type de fréquences cumulées, on sous-entend qu'il s'agit des fréquences cumulées ascendantes.

B. Le diagramme « branche et feuille »

Lorsque la taille de la population étudiée n'est pas trop élevée (inférieure à la centaine), il est intéressant d'utiliser la représentation en ***diagramme « branche et feuille »*** due à J. W. Tukey[1]. Ce diagramme tient à la fois du *tableau* et de la *représentation graphique* et donne une vision d'ensemble des données *sans perdre* l'information numérique valeur par valeur.

1) Profondeur d'une observation

Selon qu'on range les valeurs observées de la variable statistique X de la plus faible à la plus élevée ou de la plus élevée à la plus faible, on associe à chaque observation x_i deux rangs, croissant et décroissant. On dit alors que la distribution est ordonnée.

On appelle ***profondeur*** de x_i le nombre égal au *plus petit des deux rangs*.

Le tableau 1.3 présente pour les années 1990, 2000 et 2010 les durées hebdomadaires du travail des salariés à temps complet dans 15 pays de l'Union Européenne.

1. J. W. Tukey, *Exploratory Data Analysis* (EDA), Addison-Wesley, 1977.

Tableau 1.3 – Durée hebdomadaire du travail des salariés à temps complet dans 15 pays de l'Union européenne (heures)

Pays	1990	2000	2010
Allemagne	39,9	40,1	40,6
Autriche	40,1	40,1	42
Belgique	38	38,5	39,2
Danemark	39	39,3	37,7
Espagne	40,7	40,6	40,4
Finlande	38,4	39,3	39,1
France	39,6	38,9	39,4
Grèce	40,2	40,9	40,5
Irlande	40,4	39,9	38,4
Italie	38,6	38,6	39
Luxembourg	39,9	39,8	40
Pays-Bas	39	39	38,9
Portugal	41,9	40,3	40,2
Royaume-Uni	43,7	43,6	42,2
Suède	40,7	40	39,9

Source : Tableaux de l'Économie Française, INSEE.

Pour l'année 2000, le tableau 1.4 donne la profondeur de chaque valeur. Le nombre de pays étant impair et égal à 15, il y a deux valeurs de profondeur 1, 2, 3, 4, 5, 6, 7 et une seule valeur de profondeur 8.

Tableau 1.4 – Pays ordonnés selon la durée hebdomadaire du travail des salariés à temps complet en 2000

Rang croissant	Rang décroissant	Profondeur	Durée (heures)	Pays
1	15	1	38,5	Belgique
2	14	2	38,6	Italie
3	13	3	38,9	France
4	12	4	39,0	Pays-Bas
5	11	5	39,3	Danemark
6	10	6	39,3	Finlande
7	9	7	39,8	Luxembourg
8	8	**8**	**39,9**	**Irlande**
9	7	7	40,0	Suède
10	6	6	40,1	Allemagne
11	5	5	40,1	Autriche
12	4	4	40,3	Portugal
13	3	3	40,6	Espagne
14	2	2	40,9	Grèce
15	1	1	43,6	Royaume-Uni

2) La représentation en diagramme « branche et feuille »

Son principe consiste à distinguer, pour tout nombre, deux parties : le chiffre de plus « faible poids », la *feuille*, et le chiffre (ou nombre) de plus « haut poids », la *branche*.

Diagramme branche et feuille Année 1990 The decimal point is at the |	Diagramme branche et feuille Année 2000 The decimal point is at the |	Diagramme branche et feuille Année 2010 The decimal point is at the |
38 | 046	38 | 569	37 | 7
39 | 00699	39 | 03389	38 | 49
40 | 12477	40 | 011369	39 | 01249
41 | 9	41 |	40 | 02456
42 |	42 |	41 |
43 | 7	43 | 6	42 | 02

Figure 1.6 – Diagramme « Branche et feuille » (logiciel R) pour les séries du tableau 1.3

Pour le diagramme de l'année 2000 (*cf.* figure 1.6) :

– la valeur 38,5 est représentée par la branche 38 et la feuille 5 ;

– la valeur 38,6 est représentée par la branche 38 et la feuille 6 ;

– la valeur 38,9 est représentée par la branche 38 et la feuille 9.

Ces trois observations conduisent à l'écriture : 38 | 569

Un histogramme à classes égales d'amplitude 1 donne une représentation similaire, mais un avantage du diagramme branche et feuille est de conserver l'information donnée par le premier chiffre décimal, donc de conserver la répartition à l'intérieur des classes. Cette propriété est intéressante lorsque le nombre d'observations n'est pas trop élevé.

On peut compléter ce diagramme en indiquant symétriquement l'identité de chaque feuille (*cf.* figure 1.7). On pourrait aussi représenter *dos à dos* les distributions correspondant à deux années différentes pour suivre l'évolution de la durée hebdomadaire du travail.

France Italie Belgique	38	569
Irlande Luxembourg Finlande Danemark Pays-Bas	39	03389
Grèce Espagne Portugal Autriche Allemagne Suède	40	011369
	41	
	42	
Royaume-Uni	43	6

Figure 1.7 – Diagramme « Branche et feuille » complété par l'identité des pays (année 2000)

III. Les indicateurs statistiques

Le tableau de distribution d'une variable statistique présente l'information recueillie sur cette variable. Une représentation graphique en fournit un portrait pour appréhender plus facilement la globalité de l'information. On peut désirer aller plus loin en cherchant à caractériser la représentation visuelle par des éléments synthétiques sur :

- la valeur de la variable située au « centre » de la distribution : la *tendance centrale* et, plus généralement, un *indicateur de position* non nécessairement centrale, liée à un rang donné ;
- la variation des valeurs : la *dispersion* ;
- la *forme* de la distribution ;
- les aspects particuliers : valeurs *extrêmes*, *groupes* de valeurs…

Ces indicateurs étant exprimés dans les unités de la variable étudiée, on verra qu'il peut être intéressant pour comparer plusieurs distributions entre elles de calculer des *caractéristiques de dispersion relative.*

A. Conditions de Yule

Le statisticien britannique Yule[1] a énoncé un certain nombre de *propriétés* souhaitées pour les indicateurs des séries statistiques ; ceux-ci doivent être d'une part, des résumés « maniables » et d'autre part, les plus exhaustifs possibles relativement à l'information contenue dans les données.

1. G. Udny Yule et M. G. Kendall, *An Introduction to the Theory of Statistics*, Charles Griffin & Co, 14e édition, 1950.

Dans son schéma, une caractéristique statistique doit être une valeur-type :
1. définie de façon objective et donc indépendante de l'observateur ;
2. dépendante de toutes les observations ;
3. de signification concrète pour être comprise par des non-spécialistes ;
4. simple à calculer ;
5. peu sensible aux fluctuations d'échantillonnage ;
6. se prêtant aisément aux opérateurs mathématiques classiques.

En réalité, on ne dispose pas de caractéristiques répondant simultanément à ces six conditions. Le choix d'un indicateur sera l'objet d'un compromis guidé par la spécificité de l'étude en cours.

B. Les indicateurs de tendance centrale et de position

Selon l'usage courant, toutes les mesures de tendance centrale méritent le nom de « moyenne ». Lorsqu'on parle de moyenne, on pense à la moyenne arithmétique ; mais il existe d'autres types de moyennes, chacune d'entre elles ayant la propriété de *conserver une caractéristique* de l'ensemble quand on remplace chaque élément de l'ensemble par cette valeur unique ; chaque moyenne n'a donc d'intérêt que pour autant que cette propriété soit utile[1].

Les « moyennes » sont des valeurs abstraites qui, sauf par hasard, ne correspondent à aucune réalisation concrète.

1) La moyenne arithmétique

On appelle *moyenne arithmétique* la somme de toutes les données statistiques divisée par le nombre de ces données. La moyenne arithmétique *conserve la somme totale des valeurs* observées : si on modifie les valeurs de deux observations d'une série statistique tout en conservant leur somme, la moyenne de la série sera inchangée.

Soit la série statistique de données brutes : $x_1, \ldots, x_i, \ldots, x_n$, sa moyenne arithmétique a pour expression :

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

Bien entendu, si une valeur x_i de X est observée n_i fois, comme $\underbrace{x_i + x_i + \ldots + x_i}_{n_i \text{ fois}} = n_i x_i$, la formule précédente devient :

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

1. Ch. Antoine, « Les moyennes au quotidien », dans *Les Moyennes*, Que Sais-je, PUF, n° 3383, 1998, p. 107.

où k désigne le nombre de valeurs *distinctes* de X et $f_i = \frac{n_i}{n}$

Lorsqu'on a une variable statistique continue, on ne connaît pas les valeurs exactes prises par la variable, mais seulement le nombre d'observations à l'intérieur de chaque classe. Pour calculer la moyenne arithmétique d'une telle variable, on ramène *chaque observation au centre de sa classe,* ceci en raison de l'hypothèse d'équirépartition à l'intérieur des classes, et celà revient à considérer la moyenne des individus de la i^e classe égale à $(x_{i-1} + x_i)/2$.

Dans le cas des classes extrêmes non limitées, le choix des limites de ces classes influe évidemment sur la valeur de la moyenne arithmétique. Ces limites devront être choisies en fonction des connaissances sur les données et en n'oubliant pas l'hypothèse de base : l'homogénéité à l'intérieur des classes. Pour une classe extrême dans laquelle on sait qu'il n'y a pas équirépartition, les observations étant vraisemblablement en majorité regroupées sur une partie de la classe, il conviendra de choisir la borne extrême :

- moins faible que la borne réelle (supposée) s'il s'agit de la première classe ;
- plus faible que la borne réelle (supposée) s'il s'agit de la dernière classe.

C'est ce qui a été fait pour la série présentée au tableau 1.2 et à la figure 1.3, l'ancienneté moyenne du chômage a été considérée égale à 48 mois pour ceux dont l'ancienneté était au moins égale à 36 mois et la borne supérieure de la dernière classe a été de ce fait fixée à 60 mois (l'hypothèse d'équirépartition amène à considérer que la moyenne des observations d'une classe est égale au centre de la classe).

Propriétés

1. La moyenne est une caractéristique qui satisfait à toutes les conditions de Yule, sauf à la conditions 5 : une observation « extrême » (exceptionnellement élevée ou faible) peut avoir une forte incidence sur sa valeur.

2. La somme algébrique des écarts des valeurs d'une variable statistique à sa moyenne arithmétique est nulle : $\sum_{i=1}^{k} f_i(x_i - \bar{x}) = 0$

3. Lorsqu'on fait subir à une variable statistique X une transformation affine, c'est-à-dire un changement d'origine et d'unité $\{Y = aX + x_0\}$, sa moyenne arithmétique subit la même transformation : $\bar{y} = a\bar{x} + x_0$

4. Soit une population $\mathbb{P}$ de taille n partagée en deux sous-populations $\mathbb{P}_1$ de taille n_1 et $\mathbb{P}_2$ de taille n_2.

Soit X, une variable statistique observée sur la population $\mathbb{P}$, on peut exprimer sa moyenne $\bar{x}$ en fonction de ses moyennes $\overline{x_1}$ sur $\mathbb{P}_1$ et $\overline{x_2}$ sur

$\mathbb{P}_2$ en remarquant que la somme totale $n\bar{x}$ s'obtient en additionnant $n_1\overline{x_1}$ et $n_2\overline{x_2}$:

$$\bar{x} = \frac{1}{n}(n_1\overline{x_1} + n_2\overline{x_2})$$

Ce résultat se généralise à une partition en k sous-populations ($k \geq 2$) :

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

▶ **Exemple**

L'ancienneté moyenne d'inscription au chômage pour hommes et femmes réunis en septembre 2006 est égale à (*cf.* tableau 1.2 pour les données) :

$$\bar{x} = \frac{1}{2\,172{,}2}\,(1\,094{,}5 \cdot 341 + 1\,077{,}7 \cdot 334) \;.\; 338 \text{ jours}$$

2) D'autres moyennes

a) La moyenne géométrique

C'est la moyenne applicable à des mesures de grandeurs dont la croissance est géométrique ou exponentielle.

La *moyenne géométrique conserve le produit des* x_i : si on modifie les valeurs de deux observations tout en conservant leur produit, la moyenne géométrique sera inchangée.

La moyenne géométrique G de la série de valeurs $x_1, \ldots, x_i, \ldots, x_n$ supposées toutes positives (strictement), est définie ainsi :

$$G = \sqrt[n]{\prod_{i=1}^{n} x_i} \quad \Rightarrow \quad \ln(G) = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i)$$

Lorsque la distribution de la variable statistique est donnée par les k couples (x_i, n_i), les x_i étant tous positifs ; la moyenne géométrique a pour expression :

$$G = \sqrt[n]{\prod_{i=1}^{k} x_i^{\,n_i}} = \prod_{i=1}^{k} x_i^{\,f_i} \quad \Rightarrow \quad \ln(G) = \sum_{i=1}^{k} f_i \ln(x_i)$$

▶ **Exemple**

Supposons que pendant une décennie, les salaires aient été multipliés par 2 et que pendant la décennie suivante, ils aient été multipliés par 4 ; le coefficient multiplicateur moyen par décennie est égal à :

$$\sqrt{2 \cdot 4} = \sqrt{8} \approx 2{,}83$$

La moyenne arithmétique (= 3) n'est pas égale au coefficient demandé.

Prenons, par exemple, un salaire de 300 € au début de la première décennie, il sera de $300 \cdot 2 \cdot 4 = 2\,400$ € au bout des vingt ans, ce qui équivaut à $300 \cdot (2{,}83)^2$, soit un coefficient multiplicateur moyen de 2,83 par décennie.

b) La moyenne harmonique

La *moyenne harmonique* est l'inverse de la moyenne arithmétique des inverses des valeurs. L'*inverse de la moyenne harmonique conserve ainsi la somme des inverses des* x_i : si on modifie les valeurs de deux observations tout en conservant la somme de leurs inverses, la moyenne harmonique sera inchangée.

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \quad \text{ou} \quad H = \frac{1}{\sum_{i=1}^{k} \frac{f_i}{x_i}}$$

La moyenne harmonique peut être utilisée lorqu'il est possible d'attribuer un sens réel aux inverses des données en particulier pour les taux de change, les taux d'équipement, le pouvoir d'achat, les vitesses. Elle est notamment utilisée dans les calculs d'*indices*.

▶ **Exemple**

On achète des dollars une première fois pour 100 € au cours de 1,23 € le dollar, une seconde fois pour 100 € au cours de 0,97 € le dollar.

Le cours moyen du dollar pour l'ensemble de ces deux opérations est égal à :

$$\frac{200}{\frac{100}{1{,}23} + \frac{100}{0{,}97}} \approx 1{,}085 \text{ €}$$

La moyenne arithmétique (= 1,1) ne représente pas le cours moyen du dollar.

Comparaison des 3 moyennes étudiées

On montre que si les x_i sont tous positifs :

$$\min_{1 \le i \le n} x_i \le H \le G \le \bar{x} \le \max_{1 \le i \le n} x_i$$

L'égalité de deux de ces moyennes entre elles entraîne leur égalité dans leur ensemble, et dans ce cas, toutes les valeurs x_i sont égales.

3) Le mode

Pour obtenir une mesure de la tendance centrale non influencée par les valeurs extrêmes de la distribution, on peut prendre la valeur – ou la classe de valeurs – du caractère pour laquelle le diagramme en bâtons – respectivement l'histogramme – présente son *maximum* : c'est le *mode* – respectivement l'*intervalle modal* – de la distribution ; dans le cas où le diagramme en bâtons – ou l'histogramme – présente aussi un maximum local, il y a deux modes – respectivement deux classes modales.

Lorsque la variable statistique est discrète, le mode se définit donc à l'aide du tableau de distribution ou du diagramme en bâtons. Pour la distribution présentée à la figure 1.2, le mode est égal à 2. Si la fréquence maximum correspond à deux valeurs successives de la variable, il y a un *intervalle modal.*

Lorsqu'une distribution présente plusieurs modes auxquels correspondent (généralement) des fréquences différentes, c'est souvent l'indice du mélange de deux ou plusieurs populations ayant chacune leur mode propre (*cf*. figure 1.8). Un exemple peut en être la distribution des pointures de chaussures des hommes et femmes réunies.

Lorsque la variable statistique est continue, la *classe modale* est la classe dont la fréquence par unité d'amplitude est la plus élevée. Pour la distribution présentée à la figure 1.3, la classe modale est la classe [1, 3[. Mais cette détermination n'est absolument pas précise, car elle dépend du découpage en classes retenu ; son intérêt est limité par cette imprécision.

Dans le cas d'une distribution discrète, le mode satisfait aux conditions 1, 3, 4 et 5 de Yule. Dans le cas de la distribution du nombre d'enfants par famille, le mode est réellement une valeur typique et paraît mieux correspondre à la réalité que la moyenne arithmétique qui est rarement un nombre entier et qui est sensiblement influencée par un nombre relativement petit de familles très nombreuses. À l'inverse de la moyenne arithmétique, le mode néglige délibérement la précision numérique au profit de la représentativité. Dans un tel cas, il est souvent souhaitable de disposer de ces deux mesures de la tendance centrale.

Le mode, historiquement l'un des premiers paramètres de position utilisés, est un peu moins employé aujourd'hui.

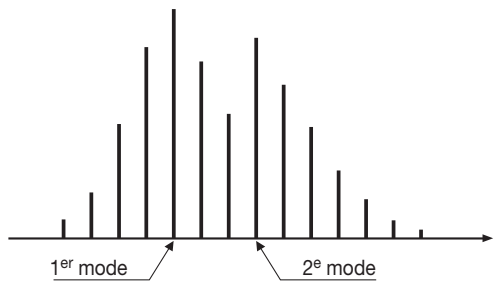


Figure 1.8 – Exemple de distribution bimodale d'une variable discrète

4) La médiane et les quantiles

Bien qu'homogènes dans leur composition, de nombreuses distributions présentent de très grands écarts entre les valeurs extrêmes de leurs éléments.

De plus, elles ont souvent un manque de symétrie prononcé, les éléments ayant tendance à s'agglomérer plus près d'un extrême que de l'autre. Les

distributions de salaires ou de revenus en donnent des exemples typiques. Il est évident que, dans de tels cas, nous avons besoin d'une mesure de la tendance centrale qui ne soit pas influencée par un nombre relativement petit de valeurs extrêmes se situant en « queue » de la distribution.

a) La médiane

La *médiane* est la valeur de la variable statistique telle qu'il y ait autant d'observations supérieures et d'observations inférieures à cette valeur. Elle partage la série statistique en deux parties d'égal effectif. Elle se détermine soit à partir de la série des valeurs ordonnées, soit à partir de la fonction cumulative (§ II.A.3).

Pour les ***variables statistiques discrètes***, la médiane est déterminée à l'aide de la « profondeur ».

Dans le cas où la série comporte un nombre impair n d'observations, la médiane est égale à la valeur de profondeur maximum $(n + 1)/2$: pour la série des 15 valeurs du tableau 4, la médiane est égale à la valeur de profondeur 8, soit 39,9 *h*.

Dans le cas où la série comporte un nombre pair n d'observations, la médiane est la moyenne arithmétique des deux valeurs de profondeur $n/2$ et est ainsi définie comme la valeur de profondeur $(n + 1)/2$.

La *médiane* est ainsi dans tous les cas *la valeur de* ***profondeur*** $(n + 1)/2$.

Lorsque les données d'une variable statistique discrète sont classées, il n'existe généralement pas une valeur médiane *Me* pour laquelle la fonction cumulative vaut 50 %. Il faut dans ce cas utiliser d'autres valeurs typiques pour caractériser la tendance centrale de la série : ceci est le cas pour la distribution du nombre de personnes par ménage dont la fonction cumulative est représentée à la figure 1.4.

Pour les ***variables statistiques continues***, la valeur médiane *Me* est telle que $\boldsymbol{F(Me) = 50\ \%}$. On commence par chercher la *classe médiane* à l'aide des fréquences cumulées, la classe médiane $[x_{i-1}, x_i[$ étant telle que $F_{i-1} < 50\ \%$ et $F_i > 50\ \%$. La valeur de la médiane s'obtient ensuite par *interpolation linéaire* en raison de l'hypothèse d'équirépartition à l'intérieur des classes. Cette détermination peut se faire par le calcul ou graphiquement (*cf.* figure 1.9) :

$$\frac{Me - x_{i-1}}{x_i - x_{i-1}} = \frac{0,5 - F_{i-1}}{f_i} \quad \Rightarrow \quad Me = x_{i-1} + (x_i - x_{i-1}) \cdot \frac{0,5 - F_{i-1}}{f_i}$$

Pour la distribution de l'ancienneté du chômage des femmes (tableau 1.2 et figure 1.5), la médiane appartient à la classe [3 ; 6[:

$$Me = 3 + 3 \cdot \frac{50 - 35,8}{15,1} \approx 5,8 \text{ mois}$$

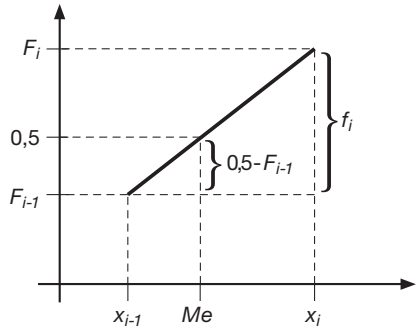


Figure 1.9 – Détermination graphique de la médiane pour une variable continue

La médiane peut aussi être déterminée à partir de la courbe des fréquences cumulées comme l'abscisse du point d'ordonnée 50 %.

Une *seule* observation très élevée (ou très faible) peut influencer fortement la moyenne, alors que la médiane peut supporter sans être modifiée qu'une moitié des observations soit très élevée (ou très faible) : on dit que la médiane est *résistante.* La médiane satisfait aux conditions 1, 3, 4 et 5 de Yule.

Dans le cas de distribution unimodale, la médiane est fréquemment comprise entre la moyenne arithmétique et le mode, et plus près de la moyenne que du mode. Si la distribution est symétrique, ces *trois caractéristiques* de tendance centrale sont *confondues* (*cf.* figure 1.10).

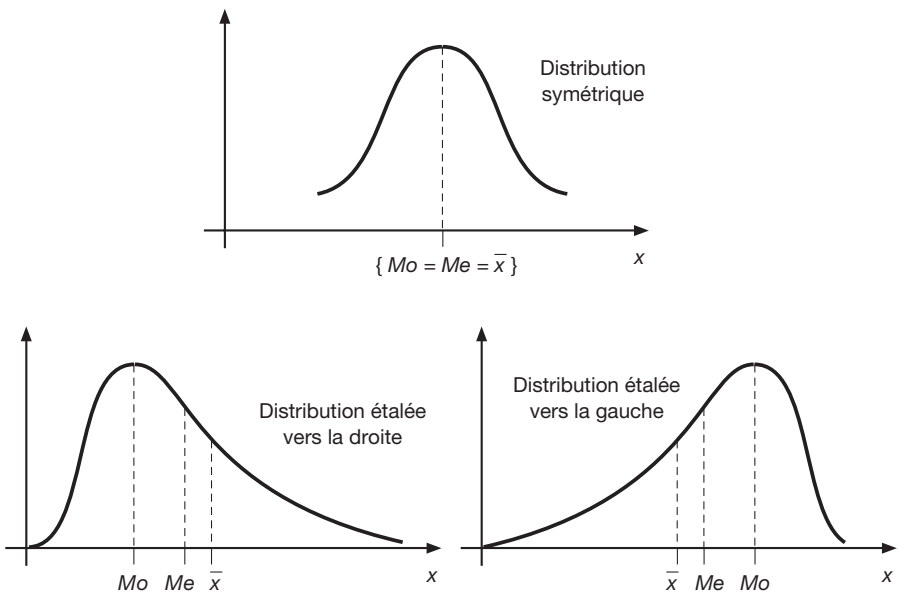


Figure 1.10 – Positions respectives du mode, de la médiane et de la moyenne

b) Les quantiles

Les *quantiles* sont des *indicateurs de position*.

Le *quantile d'ordre* α ($0 \leq \alpha \leq 1$), noté x_α, est tel qu'une proportion α des individus ait une valeur du caractère X inférieure ou égale à x_α

Le quantile $x_{0,5}$ est égal à la médiane.

On utilise couramment les quantiles d'ordre 1/4, 1/2 et 3/4. Ils sont ainsi notés et nommés :

Q_1 = *premier quartile* = $x_{0,25}$
Q_2 = *deuxième quartile* = *médiane* = $x_{0,5}$
Q_3 = *troisième quartile* = $x_{0,75}$

Les quartiles se déterminent, comme la médiane, à l'aide de la profondeur (variable discrète), ou à l'aide des fréquences cumulées (variable continue).

a) Cas d'une variable statistique *discrète*

Le premier quartile Q_1 et le troisième quartile Q_3 sont des éléments de *même profondeur égale à* $(m + 1)/2$ où m désigne la *partie entière* de la profondeur de la médiane. Ainsi par exemple :

– pour une série de 39 observations, la profondeur de la médiane est égale à 20, et celle des quartiles est égale à 10,5 ;

– pour une série de 50 observations, la médiane a une profondeur égale à 25,5, et les quartiles Q_1 et Q_3 sont de profondeur 13 puisque la partie entière de la profondeur de la médiane est égale à 25.

Pour la distribution de la durée hebdomadaire du travail dans 15 pays de l'Union européenne en 2000 (tableau 1.4), la profondeur de la médiane est égale à 8, et celle des premier et troisième quartiles égale à 4,5 :

$$Q_1 = 39{,}15\ h \quad \text{et} \quad Q_3 = 40{,}2\ h$$

La pratique de la détermination des quartiles ne respecte pas toujours la définition précédente due à Tukey. La détermination des premier et troisième quartiles n'est pas standardisée.

Ainsi les calculatrices de poche (TI, Casio...) déterminent le 1er quartile (resp. le 3e quartile) comme la médiane des valeurs de profondeur inférieure (resp. supérieure) à la profondeur de la médiane. Le résultat diffère de celui calculé avec la définition de Tukey dans le cas d'un nombre impair d'observations.

Les dernières versions d'Excel® proposent deux fonctions « QUARTILE.EXCLURE » et « QUARTILE.INCLURE ». La fonction QUARTILE.INCLURE fournit les trois quartiles calculés selon Tukey, ainsi que le minimum et le maximum :

– Minimum = QUARTILE.INCLURE(plage de valeurs ; 0) ;
– Q_1 = QUARTILE.INCLURE(plage de valeurs ; 1) ;
– Q_2 = QUARTILE.INCLURE(plage de valeurs ; 2) ;
– Q_3 = QUARTILE.INCLURE(plage de valeurs ; 3) ;
– Maximum = QUARTILE.INCLURE(plage de valeurs ; 4).

b) Cas d'une variable statistique *continue*

Compte tenu de l'hypothèse d'équirépartition et étant donné que

$$F(Q_1) = 0{,}25 \text{ et } F(Q_3) = 0{,}75$$

on calcule les quartiles par *interpolation linéaire.* Pour la distribution de l'ancienneté du chômage des femmes (*cf.* figure 1.5) :

$$Q_1 = 1 + 2 \cdot \frac{25 - 16{,}8}{19} \approx 1{,}9 \text{ mois}$$

$$Q_3 = 12 + 12 \cdot \frac{75 - 68{,}7}{18{,}5} \approx 16{,}1 \text{ mois}$$

On peut définir à partir des quartiles Q_1 et Q_3 le paramètre de tendance centrale $(Q_1 + Q_3)/2$, égal à la médiane dans le cas d'une distribution symétrique, ainsi que l'intervalle interquartile $[Q_1, Q_3]$ qui contient 50 % des observations.

Plus généralement, deux quantiles d'ordres complémentaires x_α et $x_{1-\alpha}$ définissent un intervalle dont le milieu peut être considéré comme un paramètre de tendance centrale.

De la même façon, on définit les *déciles* D_1, D_2,… , D_9 qui sont les quantiles $x_{i/10}$ (i = 1 à 9), les *vingtiles*, quantiles $x_{i/20}$ (i = 1 à 19), les *centiles,* etc.

Les classes d'une variable statistique continue sont souvent définies à l'aide des déciles. Dans ce cas, on a 10 classes contenant chacune 10 % de l'effectif total (*cf.* tableau 1.5 et figure 1.11).

Tableau 1.5 – Distribution des salaires annuels nets de tous prélèvements pour les salariés à temps complet du secteur privé et semi-public

Déciles* (en euros courants)	**Ensemble**		**Hommes**		**Femmes**	
	2000	**2006**	**2000**	**2006**	**2000**	**2006**
D_1	10 790	12 718	11 230	13 181	10 190	12 075
D_2	12 220	14 219	12 760	14 776	11 420	13 431
D_3	13 520	15 545	14 140	16 209	12 500	14 531
D_4	14 910	16 977	15 580	17 729	13 710	15 715
Médiane	**16 500**	**18 631**	**17 270**	**19 466**	**15 130**	**17 141**
D_6	18 410	20 685	19 330	21 657	16 810	18 924
D_7	20 890	23 430	22 170	24 734	18 850	21 300
D_8	24 780	27 826	26 660	29 787	21 620	24 590
D_9	32 810	36 941	35 020	40 305	26 950	30 962
D_9/D_1	3	2,9	3,2	3,1	2,6	2,6
Salaire moyen	20 400	23 292	21 890	24 912	17 510	20 232

* En 2006, 10 % des salariés à temps complet du secteur privé et semi-public gagnent un salaire annuel net inférieur à 12 718 euros, 20 % inférieur à 14 219 euros…

Source : INSEE.