ECOLE PREPARATOIRE EN SCIENCES ECONOMIQUES COMMERCIALES ET DES SCIENCES DE GESTION DE CONSTANTINE

Introduction à la Statistique Descriptive

DAKHMOUCHE Meghlaoui

Ecole Préparatoire en Sciences Economiques
Commerciales et des Sciences de Gestion
de Constantine

Introduction à la Statistique Descriptive

Dr. Meghlaoui Dakhmouche

Année Universitaire 2010/2011

Table des matières

INTRODUCTION

D'un point de vue pédagogique, il nous apparaît nécessaire de distinguer trois étapes naturelles pour l'enseignement des probabilités et des statistiques : la statistique descriptive, le calcul des probabilités élémentaires et théoriques, et la statistique théorique ou inférencielle. La statistique descriptive vise à résumer quantitativement et graphiquement l'information recueillie sur un ensemble concret au moyen d'une investigation exhaustive. Son but n'est pas d'expliquer mais de décrire et de dégager l'essentiel de l'information véhiculée par les données. Elle synthétise numériquement et graphiquement cette information. Le calcul de probabilité, quant à lui, a pour objet l'étude des phénomènes aléatoires. Il est fondé sur une axiomatique appropriée et se développe suivant une logique mathématique étrangère à toute préoccupation concrète immédiate. Enfin, la statistique théorique se rapporte à l'étude de l'induction statistique, c'est à dire l'analyse de l'information obtenue à partir d'un mécanisme aléatoire. Tandis que la statistique descriptive "constate" à l'aide d'une analyse exhaustive, en général coûteuse et parfois impossible à entreprendre, la statistique mathématique vise à cerner les caractéristiques de la population mère sur la base de l'étude d'échantillons aléatoires. Le développement historique de la connaissance dans ce domaine a plus ou moins respecté ces trois étapes. Souvent, on introduit la notion de probabilité comme une fréquence relative avant même la définition de la notion élémentaire de fréquence. Les éléments du langage des probabilités tels que, ensemble fondamental, évènement, probabilité, sont des généralisations naturelles des notions de population, caractère, fréquence. De même, la variable aléatoire est un prolongement naturel de la variable statistique. Comme le cheminement de la pensée va de l'observation des faits vers leur idéalisation abstraite, la statistique descriptive apparaît, par les problèmes qu'elle pose et les limites de ses possibilités, comme une introduction heuristique pour aborder le calcul des probabilités.

La statistique descriptive est, comme son nom l'indique, une méthode descriptive basée sur les observations recueillies à propos de l'étude de certains phénomènes d'ordre économique, sociologique ou expérimental. L'analyse des données se fait essentiellement dans deux directions principales. La première, d'essence géométrique, consiste à les classer et à les disposer de la manière la plus explicite possible, sous forme de tableaux, de graphiques ou de courbes.

La seconde a pour but de résumer l'information contenue dans les données à l'aide de certaines caractéristiques numériques. Ces deux axes ne sont pas exclusifs et sont souvent utilisés simultanément.

Le premier chapitre est consacré aux définitions des notions et des concepts fondamentaux de la statistique descriptive. Dans le deuxième chapitre nous proposons une méthode générale pour l'étude des distributions statistiques à une dimension. Nous y verrons les différentes façons de présenter des données statistiques et de les visualiser graphiquement. Au troisième chapitre on s'intéresse très sommairement à l'étude des distributions à deux caractères. On y définit aussi les notions de distributions marginales et conditionnelles. Le quatrième chapitre de ce cours est consacré à l'étude des caractéristiques de tendance centrale. Et on insistera plus spécialement sur le calcul pratique de la moyenne arithmétique et on introduira d'autres types de moyennes. Quant aux caractéristiques de dispersion, elles seront abordées dans l'avant-dernière partie de cet exposé où on définira les notions fondamentales de variance et d'écart-type. De même, il y sera fait allusion aux caractéristiques de forme. Enfin on termine cet exposé par la définition de la notion de série chronologique et par la définition des indices statistiques et leurs calculs pratiques.

Chapitre 1

Notions fondamentales de la statistique descriptive

La statistique est une méthode d'analyse des ensembles comportant un grand nombre d'éléments. C'est une science qui permet de traiter et d'analyser les résultats des mesures effectuées sur les individus d'une population relativement à un certain nombre de caractères. Les résultats des mesures sont, en général, appelés observations. Pour extraire l'information contenue dans ces observations il est nécessaire d'utiliser un certain nombre d'opérations logiques qui caractérisent les méthodes statistiques. Les éléments soumis à l'analyse doivent appartenir à un ensemble homogène et être délimités avec précision. Par la suite, ces éléments sont ordonnés et classés relativement à leurs mesures.

Pour être efficace, les méthodes statistiques doivent formaliser simplement le problème posé en utilisant des concepts mathématiques abstraits. Par exemple, tous les éléments classés dans le même sous-groupe sont considérés comme équivalents.

1.1 Concepts de base de la statistique descriptive

Les observations constituent la source principale de l'information statistique. Le statisticien doit définir avec précision l'ensemble étudié et les critères qui permettent sa description chiffrée. De ses origines historiques, la statistique a conservé en partie la terminologie de la démographie. On y

parle, par exemple, de population pour désigner un ensemble, et d'individus pour nommer les éléments de cet ensemble.

1.1.1 Population - Individu

Definition 1 *On appelle population l'ensemble des unités statistiques ou individus étudiés par le statisticien.*

Remarque 2 *Chaque observation porte sur un individu. On emploiera les termes de population et d'individu aussi bien lorsqu'il s'agit d'un ensemble d'êtres humains (population algérienne à la date du recensement, élèves d'un établissement scolaire, etc) ou d'un ensemble d'objets inanimés (production de pièces d'une usine, stocks de marchandises, etc) ou même d'un ensemble plus ou moins abstrait (ensemble des accidents de la route survenus au cours d'un mois de l'année, ensemble des jours ouvrables de l'année, etc). Les individus d'une population peuvent donc être, selon les cas, des êtres humains, des objets ou des évènements.*

1.1.2 Caractère - Modalité

Pour décrire une population on classe les individus qui la composent en un certain nombre de sous-ensembles. Le classement peut se faire relativement à un ou plusieurs caractères. Par exemple, pour décrire la population algérienne on pourra retenir les caractères sexe, âge, état matrimonial, catégorie socioprofessionnelle, etc. S'il s'agit du personnel d'une entreprise, le sexe et l'âge restent des caractères intéressants et on pourra y rajouter la profession, la qualification, etc.
Le choix d'un caractère détermine le critère qui servira à classer les individus de la population en deux ou plusieurs sous-ensembles. Le nombre de ces derniers correspond aux différentes situations possibles ou modalités du caractère. Afin que le classement d'un individu soit toujours possible sans ambiguïté, les différentes modalités d'un caractère doivent être à la fois exhaustives et incompatibles. Un individu ne doit appartenir qu'à un et un seulement des sous-ensembles obtenus. Ainsi, le caractère sexe a deux modalités qui déterminent dans une population le sous-ensemble des individus masculins et le sous-ensemble des individus féminins. Le nombre de modalités selon lesquelles on considère un caractère est fixé plus ou moins conventionnellement.

1.1.3 Tableau statistique

L'étude d'une population suivant un seul caractère est résumée dans un tableau statistique à une seule dimension ou à simple entrée, dont chaque case correspond à l'une des modalités du caractère. Dans chacune de ces dernières on y inscrit le nombre d'individus présentant cette modalité. Mais une population peut aussi être étudiée simultanément suivant deux ou plusieurs caractères. Le nombre de cases, donc de sous-ensembles incompatibles et exhaustifs, est alors égal au produit du nombres de modalités des différents caractères. Ainsi, le croisement du caractère sexe avec le caractère état matrimonial (en deux modalités) nous donne le tableau suivant :

Etat Matr/Sexe	Homme	Femme
Marié	H. mariées	F. mariées
Non Marié	H.n. mariées	F.n. mariées

Il est possible de croiser trois caractères, quatre caractères ou plus. Ainsi, on obtient des tableaux statistiques à trois, quatre dimensions ou plus. Mais en pratique, on ne peut croiser un trop grand nombre de caractères, car le nombre de cases du tableau augmente très vite et son utilisation devient fastidieuse.

1.2 Les différents types de caractères

Un caractère peut être qualitatif ou quantitatif. Les méthodes d'analyse d'une population diffèrent suivant la nature du caractère étudié.

1.2.1 Caractère qualitatif

Definition 3 *Un caractère qualitatif est un caractère dont les modalités échappent à la mesure.*

Remarque 4 *On ne peut pas quantifier numériquement les caractères qualitatifs, on ne peut que les constater. Par exemple, le sexe, la nationalité, la profession, etc.*

Example 5 *Considérons la répartition par nationalité des étrangers vivants en France (en Milliers) :*

Nat.	*All*	*Bene*	*Esp*	*Ita*	*Pol*	*Port*	*Autres Eu*	*Alg*	*Mar*	*Tun*	*Autres Etr*
Nb	25	60	120	80	100	210	140	650	310	60	420

1.2.2 Caractère quantitatif

Definition 6 *Un caractère est qualifié de quantitatif lorsqu'il est mesurable ou repérable.*

Definition 7 *A chaque unité statistique ou individu correspond un nombre représentant la mesure ou la valeur du caractère. Cette mesure est alors appelée variable statistique et est notée en général x.*

Remarque 8 *Les modalités du caractère sont les valeurs possibles ou ensemble de variation de la variable statistique.*

Une variable statistique peut être discrète ou continue.

Variable statistique discrète

Definition 9 *Une variable statistique est dite discrète lorsqu'elle ne peut prendre que des valeurs isolées dans son intervalle de variation.*

Remarque 10 *Les valeurs prises par une variable discrète sont en général des valeurs entières, par exemple le nombre d'enfants à charge dans une famille.*

Example 11 *Considérons la répartition du nombre de ventes d'un certain type d'appareil sur les jours ouvrables de l'année. Soit x la variable statistique "le nombre de ventes par jour ouvrable" :*

Nombre de ventes x	0	1	2	3	4	5	6
Nombre de jours n_i	24	57	75	53	33	7	4

Variable statistique continue

Definition 12 *Une variable statistique est dite continue lorsqu'elle peut prendre toutes les valeurs à l'intérieur de son intervalle de variation.*

Remarque 13 *Le nombre des valeurs possibles d'une variable statistique continue est toujours infini. Ainsi, on prendra pour modalités du caractère des classes de valeurs.*

Example 14 *La taille, le poids, l'âge d'une personne. La durée de vie d'une lampe, la distance séparant deux points.*

Remarque 15 *Souvent, la distinction entre variable statistique continue et variable statistique discrète est difficile. Par exemple, toute mesure est discrète du fait de sa précision limitée, alors que la nature intrinsèque de la variable statistique est continue (par exemple le diamètre d'une pièce usinée).*
Réciproquement, une variable de nature discrète pouvant prendre un très grand nombre de valeurs possibles, est considérée comme une variable statistique continue et ainsi ses valeurs sont regroupées en classes.

Example 16 *Le salaire d'un ouvrier, les bénéfices annuels des entreprises, la date de naissance d'une personne, la note d'un étudiant, le diamètre d'une pièce usinée.*

1.2.3 Notion de classe

Pour étudier une variable statistique continue on divise son ensemble de variation en intervalles ou classes de valeurs ayant une amplitude constante ou variable.

Example 17 *La variable "âge" est souvent découpée en classes quinquennales : 0 à moins de 5 ans, de 5 ans à moins de 10 ans, etc.*

Remarque 18 *Le choix du nombre de classes et de leur amplitude se fait en fonction de l'effectif de la population et de la précision des mesures. Les effectifs des classes doivent être significatifs pour éliminer les variations accidentelles qui apparaissent lorsqu'on considère de trop faibles effectifs. Par ailleurs, le nombre de classes doit aussi être suffisant et leurs amplitudes pas trop grandes pour ne pas masquer certaines particularités de la distribution statistique. Toute diminution inconsidérée du nombre de classes ou toute augmentation exagérée des amplitudes de celles-ci, induit une perte d'information.*

En conclusion, on remarque que la statistique descriptive porte sur une population sur laquelle aucun modèle statistique n'est défini a priori. On ne dispose que d'un ensemble de mesures sans structuration a priori.

Chapitre 2

Les distributions statistiques à une dimension

Après la définition du caractère à étudier sur la population, les observations obtenues sont ordonnées et forment ainsi une distribution statistique. Les distributions les plus simples sont naturellement celles relatives à un seul caractère. Elles sont généralement présentées sous forme de tableaux statistiques à simple entrée. L'information synthétisée dans un tableau statistique n'est pas souvent facile à obtenir par simple lecture. Alors la représentation de cette distribution statistique sous forme de diagramme rend plus simple l'accès à cette information. Selon que le caractère étudié soit qualitatif ou quantitatif, et suivant qu'il soit de nature discrète ou continue, on est amené à utiliser des représentations graphiques de différents types.

2.1 Présentation générale des tableaux statistiques

Considérons une population P composée de n individus. Sur chacun de ces individus on effectue une observation concernant un caractère C. Supposons que le caractère C admet k modalités :

$$M_1, M_2, ..., M_k$$

L'opération préliminaire est la mise en ordre des observations. Cela consiste à classer chacun des n individus de la population dans les k sous-ensembles

définis par les diverses modalités du caractère C. Pour chaque modalité M_i, $i = 1, 2, ..., k$, on inscrira dans le tableau statistique le nombre d'éléments (i.e. le cardinal) du sous-ensemble de la population correspondant.

Definition 19 *Le cardinal du sous-ensemble de la population correspondant à la modalité M_i est appelé* ***effectif*** *ou* ***fréquence absolue*** *et est noté n_i.*

Remarque 20 *Tous les individus de la population présentant la modalité M_i sont considérés comme équivalents relativement au caractère C. On ne retient alors que leur nombre.*

Definition 21 *La fréquence relative f_i de la modalité M_i est définie par le rapport :*

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

Remarque 22 *La fréquence f_i est la proportion des individus de la population présentant la modalité M_i. Alors, les fréquences permettent de comparer les structures des populations d'effectifs différents relativement à un caractère commun. Les modalités sont incompatibles, i.e. deux modalités distinctes d'un caractère donné ne peuvent pas être présentes chez un même individu en même temps . Elles sont aussi exhaustives, i.e. chaque individu de la population est classé dans un et un seul sous-groupe correspondant à une modalité. Donc chaque observation figure dans un et un seul sous-groupe de la population.*

Conséquence

La somme des effectifs n_i est égale à l'effectif total n de la population , i.e.

$$n = n_1 + n_2 + ... + n_k = \sum_{i=1}^{k} n_i$$

Alors, il en résulte que la somme des fréquences relatives f_i est égale à 1 , i.e.

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

Un tableau statistique décrivant une population P relativement à un caractère C, sera présenté sous la forme générale suivante :

Modalités du caract. C	Effectifs n_i
M_1	n_1
M_2	n_2
...	...
M_i	n_i
...	...
M_k	n_k

2.2 Les distributions à caractère qualitatif

La présentation d'un tableau statistique concernant un caractère qualitatif suit les règles générales. Une première synthèse de l'information contenue dans un tableau statistique est fournie par un graphique. Le principe de la représentation graphique des caractères qualitatifs est la proportionnalité des surfaces représentatives aux effectifs (ou aux fréquences) représentés. Il existe deux types de représentations fréquemment utilisées.

2.2.1 Représentation par tuyaux d'orgue

Cette représentation fait figurer les différentes modalités du caractère sous forme de rectangle ou de cylindres dont la base est constante et dont la hauteur est proportionnelle à l'effectif (ou à la fréquence).

Remarque 23 *Généralement, les différentes modalités sont ordonnées sur le graphique dans le sens des effectifs croissants ou décroissants.*

Example 24 *Reprenons l'exemple* 5 *et affichons la représentation par tuyaux d'orgue de la distribution des étrangers en France.*

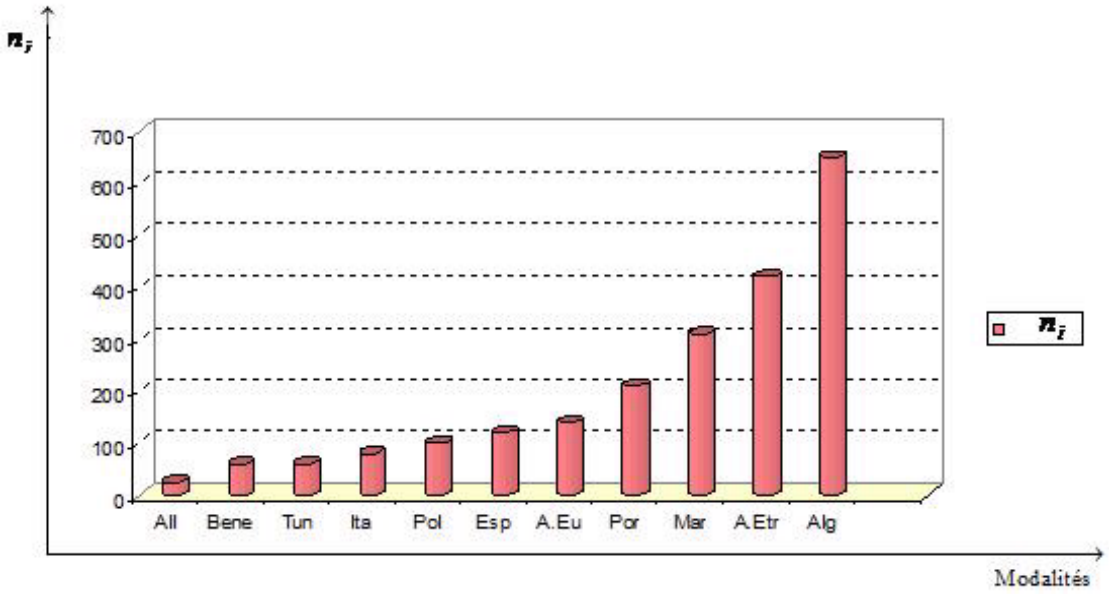


FIG. 2.1 – Répartition par tuyaux d'orgues des étrangers en France

2.2.2 Représentation par secteur

Dans cette représentation les aires et par conséquent les angles au centre sont proportionnels aux effectifs (ou aux fréquences) des différentes modalités. En effet,

$$\theta_i = 360° \frac{n_i}{n} = 360° f_i$$

Example 25 *Mieux que les tuyaux d'orgue, ce mode de figuration permet de visualiser l'importance relative de chaque modalité dans l'ensemble de la population. Pour des comparaisons dans l'espace, la représentation par secteur permet de mieux faire apparaître les différences entre les classes d'individus en valeurs absolues et en valeurs relatives.*

Example 26 *Reprenons l'exemple* 5 *et affichons la représentation par secteur de la distribution des étrangers en France.*

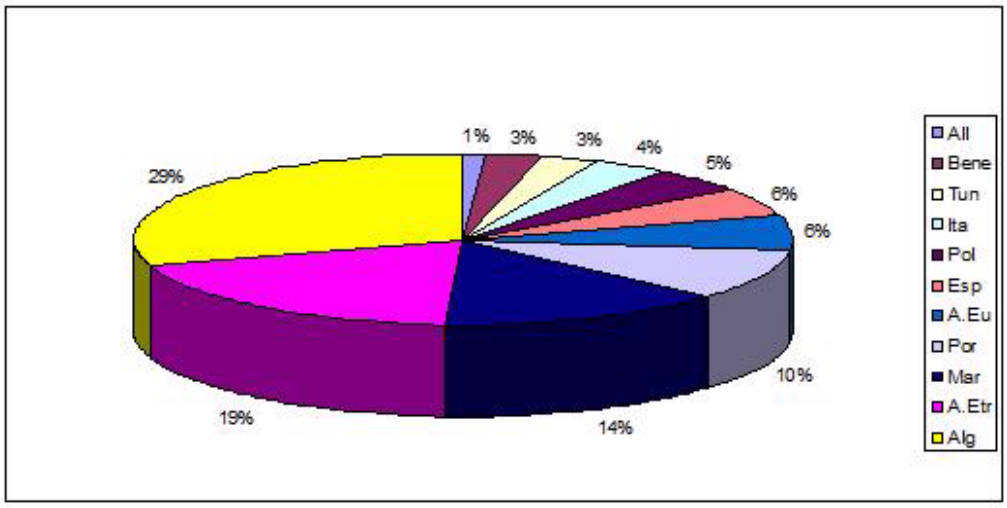


FIG. 2.2 – Représentation par secteur des étrangers en France

2.3 Les distributions à caractère quantitatif

2.3.1 Variable discrète

Tableau statistique

Les différentes modalités sont constituées par les valeurs possibles x_i de la variable statistique x. En face de chacune de ces valeurs on inscrit l'effectif n_i correspondant. Pour permettre les comparaisons entre populations d'effectifs différents, le tableau est complété par l'indication de la fréquence relative f_i correspondant à chaque valeur x_i.

Definition 27 *La fréquence cumulée croissante, notée F_i, est la somme des fréquences correspondantes aux valeurs de la variable statistique inférieures ou égales à x_i, i.e.*

$$F_i = f_1 + f_2 + ... + f_i = \sum_{h=1}^{i} f_h$$

La distribution statistique d'une variable quantitative discrète est en général présentée dans un tableau statistique tel que :

$Modalités$	$Effectifs$	$Fréquences$	$Fréquences\ cumulées$
x_1	n_1	f_1	$F_1 = f_1$
x_2	n_2	f_2	$F_2 = f_1 + f_2$
...	...	...	...
x_i	n_i	f_i	$F_i = f_1 + f_2 + ... + f_i$
...	...	...	...
x_k	n_k	f_k	$F_k = 1$
$Total$	n	1	

Tableau statistique : variable discrète

Remarque 28 *La fréquence cumulée croissante F_i indique la fréquence ou la proportion des individus de la population pour lesquels la variable statistique x est inférieure ou égale à x_i.*

Definition 29 *L'effectif cumulé croissant, noté N_i, est défini, similairement à la fréquence cumulée croissante, par la formule suivante :*

$$N_i = n_1 + n_2 + ... + n_i = \sum_{h=1}^{i} n_h$$

Proposition 30 *La fréquence cumulée croissante et l'effectif cumulé croissant sont liés par la relation :*

$$F_i = \frac{N_i}{n}$$

Démonstration : En effet,

$$N_i = \sum_{h=1}^{i} n_h \Longrightarrow \frac{N_i}{n} = \sum_{h=1}^{i} \frac{n_h}{n} = \sum_{h=1}^{i} f_h = F_i$$

■

Definition 31 *Il est parfois utile de définir les effectifs cumulés décroisants, notés $N_i^{'}$, tels que :*

$$N_i^{'} = n_k + n_{k-1} + ... + n_i$$

Remarque 32 *L'effectifs cumulé décroissant est le résultat de l'addition, de proche en proche, des effectifs d'une distribution observée en commençant par le dernier effectif, i.e.*

$$N_k^{'} = n_k \ ; \ N_{k-1}^{'} = n_k + n_{k-1} \ ; ... ; \ N_i^{'} = n_k + n_{k-1} + ... + n_i$$

En d'autres termes, $N_i^{'}$ est le nombre d'individus présentant une mesure du caractère x inférieure ou égale à x_i, i.e.

$$N_i^{'} = \text{ nombre de valeurs de } x \geq x_i$$

Il est clair que l'on peut définir les fréquences cumulées décroissantes, notées G_i, telles que :

$$G_i = \frac{N_i^{'}}{n} = \frac{n_k + n_{k-1} + ... + n_i}{n} = f_k + f_{k-1} + ... + f_i$$

Example 33 *Considérons la distribution des jours d'ouverture d'un magasin suivant le nombre de ventes d'un certain appareil A.*

Nombre de ventes x_i	*nombre de jours*	f_i	N_i	F_i
0	24	$0,096$	24	$0,096$
1	57	$0,228$	81	$0,324$
2	75	$0,300$	156	$0,624$
3	53	$0,212$	209	$0,836$
4	33	$0,132$	242	$0,968$
5	4	$0,016$	246	$0,984$
6	3	$0,012$	249	$0,996$
7	1	$0,004$	250	$1,00$
Totaux	250	1		

Représentation graphique Dans le cas des séries statistiques discrètes il existe deux types de représentations graphiques.

La représentation en diagramme en bâtons

Definition 34 *La représentation en diagramme en bâtons est la représentation de la distribution des fréquences ou des effectifs d'une variable discrète. A chaque valeur x_i portée en abscisse on fait correspondre un segment vertical de longueur proportionnelle à l'effectif n_i ou à la fréquence f_i de cette valeur.*

Example 35 *Reprenons l'exemple précédent et représentons la distribution des jours de l'année en fonction du nombre de ventes.*

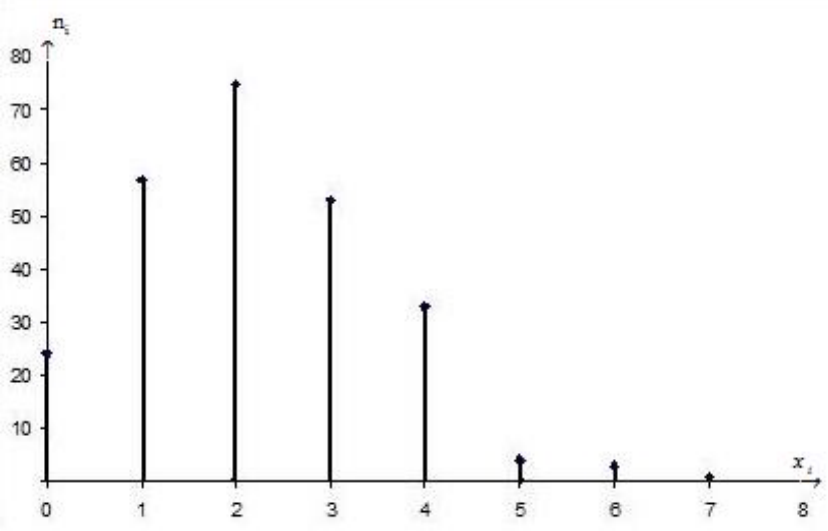


FIG. 2.3 – Représentation en diagramme en batôns

Courbe cumulative

Definition 36 *La courbe cumulative est la représentation graphique des effectifs cumulés ou des fréquences cumulées. C'est un graphique en escalier dont les paliers horizontaux ont pour ordonnées respectivement F_i ou N_i. Les marches de l'escalier correspondent aux valeurs possibles x_i de la variable statistique x et sont à des hauteurs proportionnelles aux effectifs cumulés ou aux fréquences cumulées.*

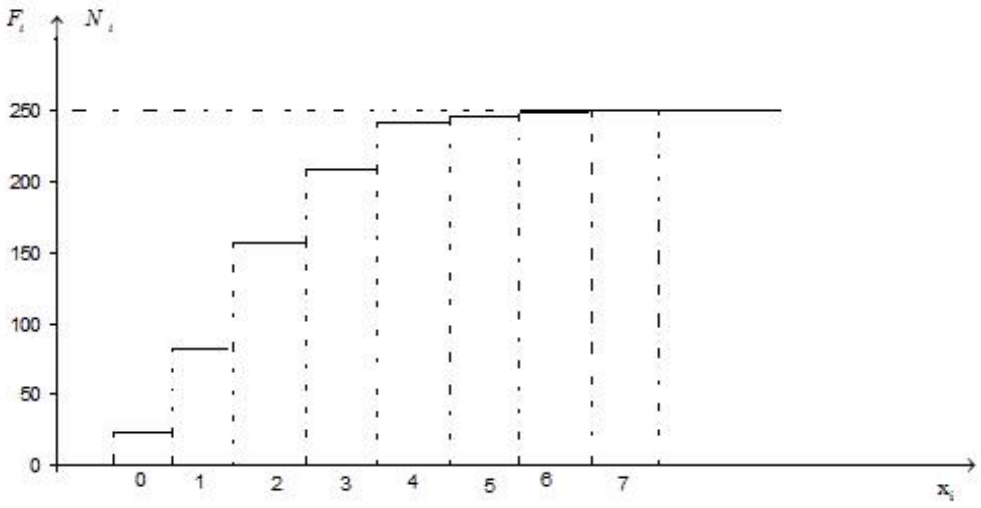


FIG. 2.4 – Courbe cumulative

Example 37 *Reprenons l'exemple* 31 *et traçons la courbe cumulative de la distribution des jours de l'année en fonction du nombre de ventes.*

Remarque 38 *La courbe cumulative est la représentation graphique de la proportion $F(x)$ des individus de la population pour lesquels la valeur de la variable statistique est inférieure ou égale à x . Cette fonction, définie pour toute valeur de x , est appelée fonction cumulative ou fonction de répartition. Elle est constante dans chaque intervalle séparant deux valeurs de la variable statistique, i.e. $F(x) = F_i$, $x_i \leq x < x_{i+1}$, elle est nulle pour toutes les valeurs de x inférieures à la plus petite valeur des x_i et est égale à* 1 *pour toutes les valeurs de x supérieures à la plus grande valeur des x_i . On peut aussi définir la fonction de répartition de la variable statistique x, notée aussi $F(x)$, comme la ligne brisée qui joint les milieux des paliers de la courbe cumulative.*

Example 39 *On reprend l'exemple précédent et on trace la fonction de répartition sur le graphe de la courbe cumulative.*

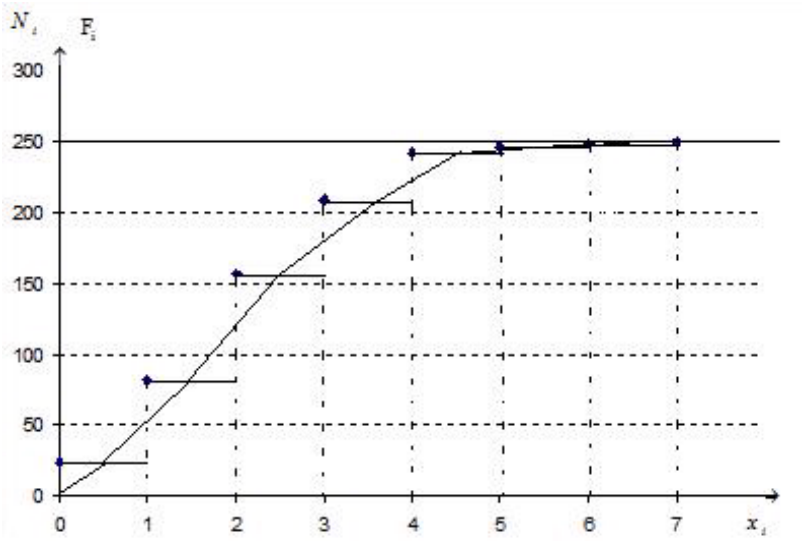


FIG. 2.5 – Fonction de répartition

2.3.2 Variable continue

Tableau statistique

Dans le cas d'un caractère quantitatif continu x, l'établissement du tableau de fréquences implique d'effectuer au préalable une répartition en classes des données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou intervalle de classe. En règle générale, on choisit des classes de même amplitude. Pour que la distribution des fréquences ait un sens, il faut que chaque classe comprenne un nombre (n_i) suffisant de valeurs. Diverses formules empiriques permettent d'établir le nombre de classes pour un échantillon de taille n. Les plus fréquemment utilisées sont :

- La règle de Sturge : $Nombre\ de\ classes = 1 + (3,3 \ln n)$
- La règle de Yule : $Nombre\ de\ classes = 2,54\sqrt[4]{n}$

L'amplitude a des classes est obtenue de la manière suivante :

$$a = \frac{x_{\max} - x_{\min}}{Nombre\ de\ classes}$$

avec $x_{\max}$ et $x_{\min}$ respectivement la plus grande et la plus petite valeur de x dans la série statistique.
Les modalités du caractère sont représentées par les différentes classes. Si l'on désigne respectivement par e_{i-1} et e_i les extrémités inférieure et supérieure de la classe $n°i$, on définit cette dernière comme suit :

$$e_{i-1} \leq x < e_i$$

Remarque 40 *Les fréquences et les fréquences cumulées sont définies de la même façon que dans le cas discret, ainsi que les effectifs et les effectifs cumulés.*

En général, les résultats des observations d'une variable statistique continue x sont disposés dans un tableau statistique tel que :

Classe $n°i$	*Lim des classes*	n_i	f_i	N_i	F_i
1	$e_0 \leq x < e_1$	n_1	f_1	N_1	F_1
2	$e_1 \leq x < e_2$	n_2	f_2	N_2	F_2
3	$e_2 \leq x < e_3$	n_3	f_3	N_3	F_3
...	...	...	...	...	...
i	$e_{i-1} \leq x < e_i$	n_i	f_i	N_i	F_i
...	...	...	...	...	...
k	$e_{k-1} \leq x < e_k$	n_k	f_k	$N_k = n$	$F_k = 1$
$Total$		$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} f_i = 1$		

Tableau statistique : Cas continu

Représentation graphique

Comme pour les variables discrètes il existe pour les variables statistiques contines deux types de représentations graphiques utilisés fréquemment.

Histogramme

Definition 41 *L'histogramme est la représentation graphique de la distribution des effectifs ou des fréquences d'une variable statistique continue. A chaque classe de valeurs de la variable statistique portée en abscisse, on fait correspondre un rectangle basé sur cette classe. Alors chaque modalité est représentée par un rectangle dont l'aire (et non la hauteur) est proportionnelle à la fréquence ou à l'effectif de cette classe.*

Remarque 42 *En général les classes de valeurs ont la même amplitude. Mais dans le cas contraire, on prendra pour unité d'amplitude u le P.G.C.D des différentes amplitudes $a_i = e_i - e_{i-1}$. Ensuite, on exprime l'amplitude des classes dans la nouvelle unité telle que :*

$$A_i = \frac{a_i}{u} = \frac{e_i - e_{i-1}}{u}$$

Par suite, la hauteur du rectangle représentatif de chaque classe sera égale à $h_i = \frac{f_i}{A_i}$ de telle sorte que la surface du rectangle soit égale à la fréquence de la classe correspondante, i.e.

$$S = A_i \left(\frac{f_i}{A_i} \right) = f_i$$

Example 43 *Considérons la répartition des ouvriers d'une entreprise suivant leur salaire mensuel net :*

Classe de Salaire(DA)	n_i	f_i	N_i	F_i
$12000 \leq x < 14000$	26	$0,186$	26	$0,186$
$14000 \leq x < 16000$	33	$0,235$	59	$0,421$
$16000 \leq x < 20000$	64	$0,458$	123	$0,879$
$20000 \leq x < 24000$	7	$0,050$	130	$0,929$
$24000 \leq x < 30000$	10	$0,071$	140	$1,000$
Total	140	$1,000$		

Traçons l'histogramme des fréquences de cette distribution.

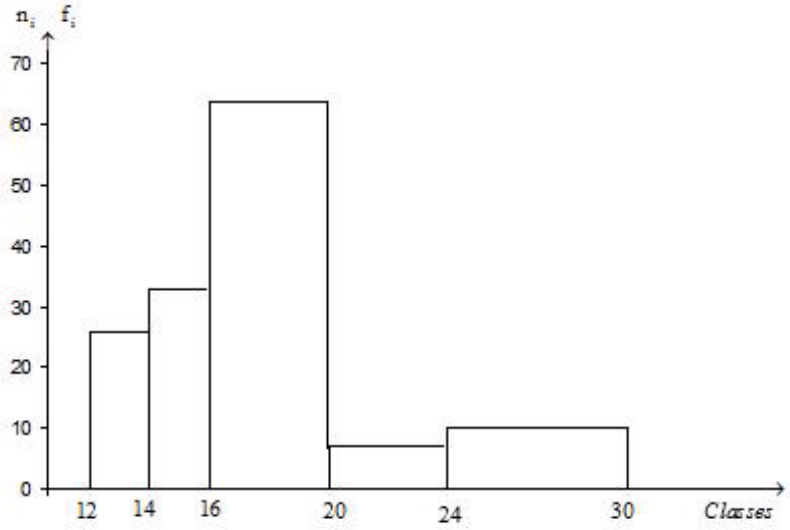


FIG. 2.6 – Histogramme des fréquences

Definition 44 *La courbe des fréquences est la fonction en escalier dont les paliers sont constitués par les bases supérieures des rectangles formant l'histogramme des fréquences.*

Definition 45 *Le polygône des fréquences est la ligne brisée qui relie les milieux des cotés supérieurs des rectangles de l'histogramme des fréquences.*

Example 46 *Reprenons l'exemple de la répartition des ouvriers d'une entreprise suivant leur salaire mensuel net et traçons la courbe des fréquences et le polygône des fréquences de cette distribution.*

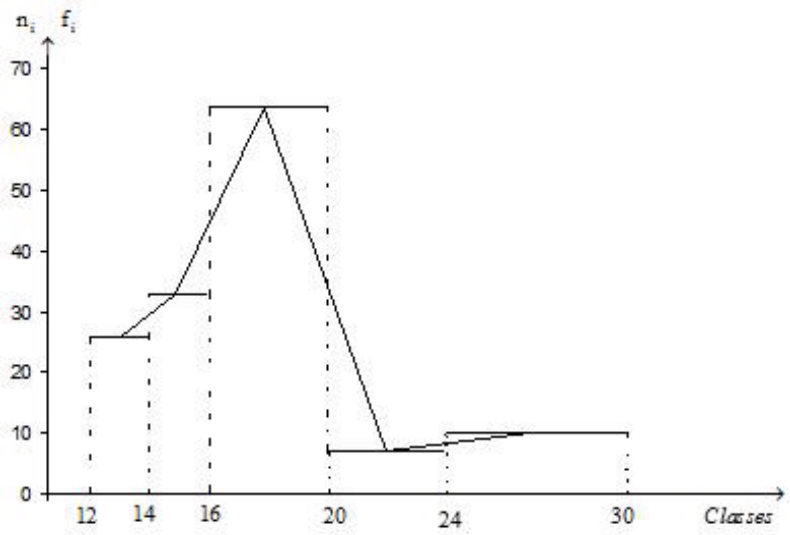


FIG. 2.7 – Courbe des fréquences et polygône des fréquences

Au moment de l'étude des lois de probabilité on comprendra mieux la signification de cette courbe des fréquences. Elle représente une approximation (ou estimation) de la distribution de probabilité théorique de la population relativement au caractère étudié.

Courbe cumulative

Definition 47 *Comme pour les variables discrètes, la coube cumulative ou histogramme des fréquences cumulées, est la représentation graphique de la fonction cumulative ou fonction de répartition* $F(x)$.

Les observations étant regroupées en classes, on ne connait de cette fonction que les valeurs correspondant aux extrémités supérieures des classes, i.e.

$$F(e_i) = F_i \qquad i = 1, 2, ..., k$$

Elle est estimée par le *polygône des fréquences cumulées* qui est la ligne brisée joignant les milieux des cotés supérieurs des rectangles de l'*histogramme des fréquences cumulées*.

Remarque 48 *Dans une certaine littérature on parle de fréquence "cumulée descendante" G_i et de fréquence "cumulée ascendante" F_i. Cette dénomination implique une confusion. En effet, on a tendance à admettre implicitement que G_i est égale à $1 - F_i$, ce qui n'est le cas. Par contre, quand on parle de fonction cumulative $F(x)$ qui est définie sur l'ensemble $\mathbb{R}$ en entier et telle que $\lim_{x \longrightarrow +\infty} F(x) = 1$ et $\lim_{x \longrightarrow -\infty} F(x) = 0$, on peut définir la fonction $G(x)$ telle que :*

$$G(x) = 1 - F(x)$$

Ainsi, sachant que $F(M_e) = \frac{1}{2}$ alors $G(M_e) = \frac{1}{2}$, i.e. l'intersection des fonctions $F(x)$ et $G(x)$ a lieu au point d'abscisse $x = M_e$. L'utilité de la fonction $G(x)$ intervient dans la détermination graphique de la médiane à condition que les tracés de $F(x)$ et $G(x)$ soient très précis.

Example 49 *Repenons l'exemple* 41 *et traçons l'histogramme des fréquences cumulées et le polygône des fréquences cumulées.*

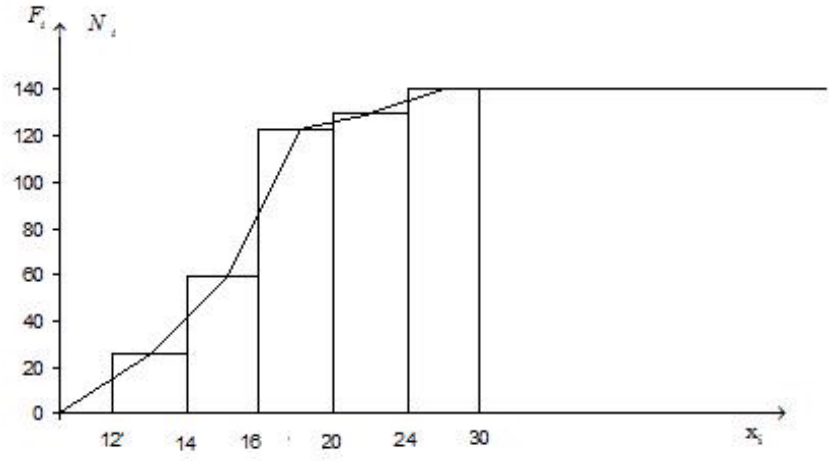


FIG. 2.8 – Histogramme et polygône des fréquences cumulées

Conclusion 50 *La notion de courbe des fréquences (resp. la courbe des fréquences cumulées) découle de l'idée suivante : si les amplitudes des classes diminuent et si le nombre des observations est suffisamment grand pour éviter les irrégularités dues à la faiblesse des effectifs, alors l'histogramme des*

fréquences (resp. l'histogramme des fréquences cumulées) tend, en tant que fonction en escalier, vers une courbe continue appelée courbe des fréquences (resp. courbe des fréquences cumulées) et qui, à la limite, converge vers la densité de la distribution théorique (resp. la fonction de répartition théorique) de la population.

Example 51 *On a mesuré la taille en centimètres d'une population de* 8585 *hommes. Les résultats sont résumés dans le tableau suivant :*

x	$x < 145$	$[145, 148[$	$[148, 151[$	$[151, 154[$	$[154, 157[$	$[157, 160[$	$[160, 163[$
n_i	2	4	14	41	83	169	394
N_i	2	6	20	61	144	313	707

$[163, 166[$	$[166, 169[$	$[169, 172[$	$[172, 175[$	$[175, 178[$	$[178, 181[$	$[181, 184[$
669	990	1223	1329	1230	1063	640
1376	2366	3589	4918	6148	7211	7851

$[184, 187[$	$[187, 190[$	$[190, 193[$	$[193, 196[$	$[196, 199[$	$[199, 202[$	$202 \leq x$
392	202	84	33	16	5	2
8243	8445	8529	8562	8578	8583	8585

En traçant la courbe des fréquences de cette distribution statistique, on peut remarquer que l'allure de cette courbe a une forme qui se raproche très nettement de celle d'une courbe normale. De même, si on trace la courbe des fréquences cumulées de cette distribution, on remarque que son allure est très voisine de celle de la fonction de répartition d'une loi normale.

.

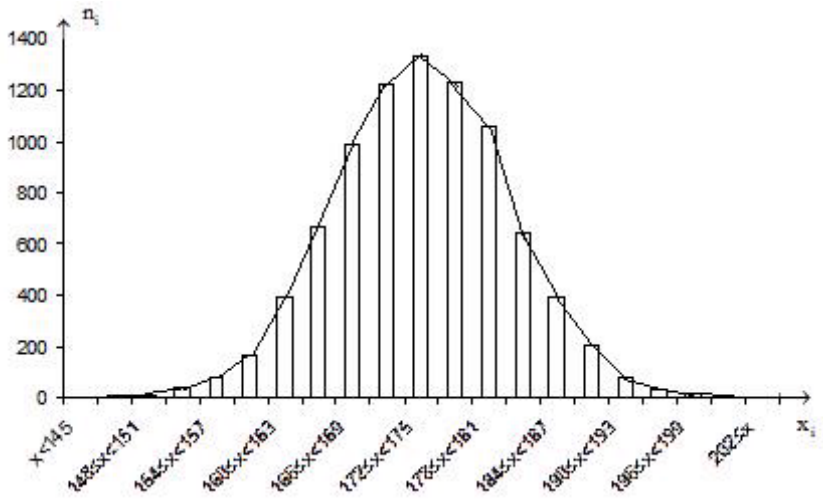


FIG. 2.9 – Courbe des fréquences

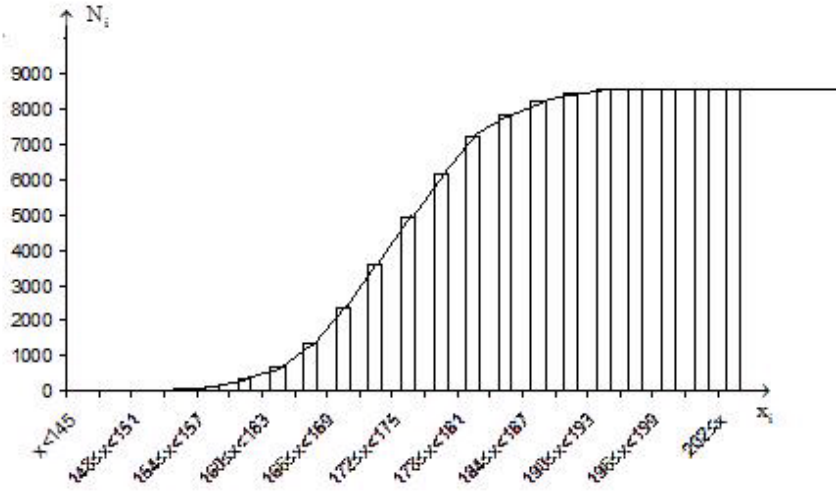


FIG. 2.10 – Courbe des fréquences cumulées

Chapitre 3

Caractéristiques de tendance centrale

La représentation graphique des distributions statistiques a permis une première synthèse de l'information contenue dans les tableaux statistiques. On peut comparer les différentes séries statistiques par simple comparaison de leurs représentations graphiques. Cette comparaison reste toutefois incommode et devient quasi-impossible si elle doit porter sur un grand nombre de distributions statistiques. Il est inconcevable de classer 1500 candidats à un concours de grandes écoles au vu de la représentation graphique des 1500 séries de 25 notes obtenues aux diverses épreuves par chaque candidat. Il est évidemment plus commode de calculer une note moyenne pour chacun des candidats et ensuite faire un classement. La tendance centrale caractérise l'ordre de grandeur de la variable statistique. Quant à la notion de dispersion, elle mesure la fluctuation des observations autour de cette tendance centrale. Le statisticien Yule a précisé les propriétés souhaitables que doit satisfaire une caractéristique de tendance centrale ou de dispersion : elle doit être définie de façon objective ; elle doit dépendre de toutes les observations ; elle doit avoir une signification concrète ; elle ne doit pas être sensible aux fluctuations d'échantillonnage ; elle doit être simple à calculer et doit se prêter aisément au calcul algébrique.
Trois caractéristiques de tendance centrale sont couremment utilisées : le mode, la médiane et la moyenne arithmétique. Dans certains cas,'usage d'autres caractéristiques de tendance centrale telles que la moyenne géométrique ou la moyenne harmonique, s'impose. Mais la caractéristique de tendance centrale la plus couramment utilisée est la moyenne arithmétique.

3.1 Les différentes caractéristiques de tendance centrale

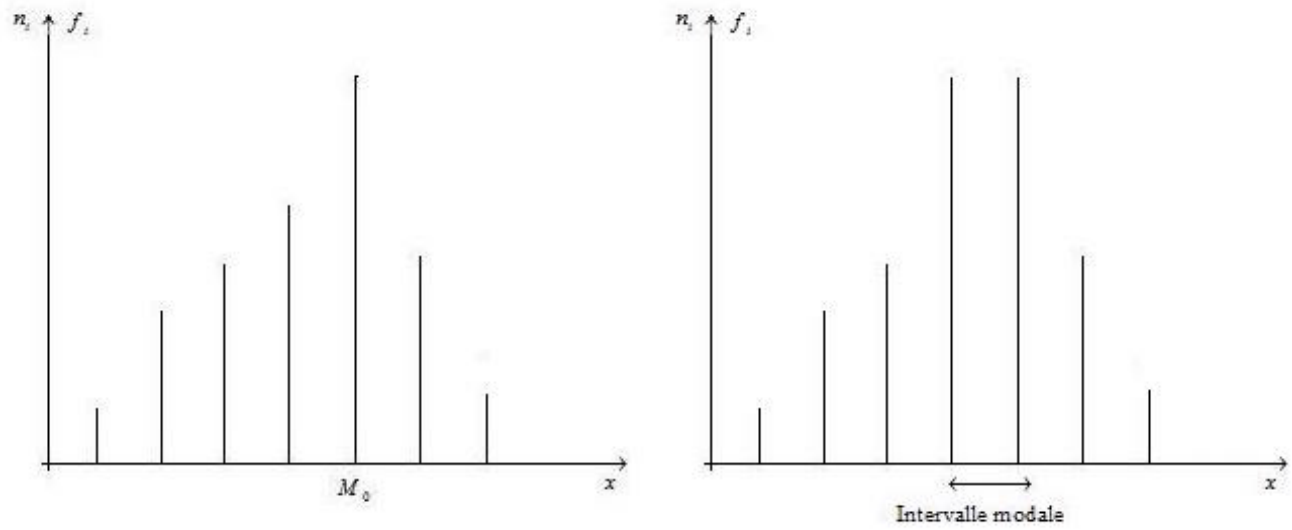
3.1.1 Le mode

Definition 52 *Le mode d'une distribution statistique, qu'on notera M_o, est la valeur de la variable statistique pour laquelle la fréquence est la plus grande.*

Remarque 53 *Le mode est donc la valeur de la variable statistique la plus fréquente.*

Détermination graphique

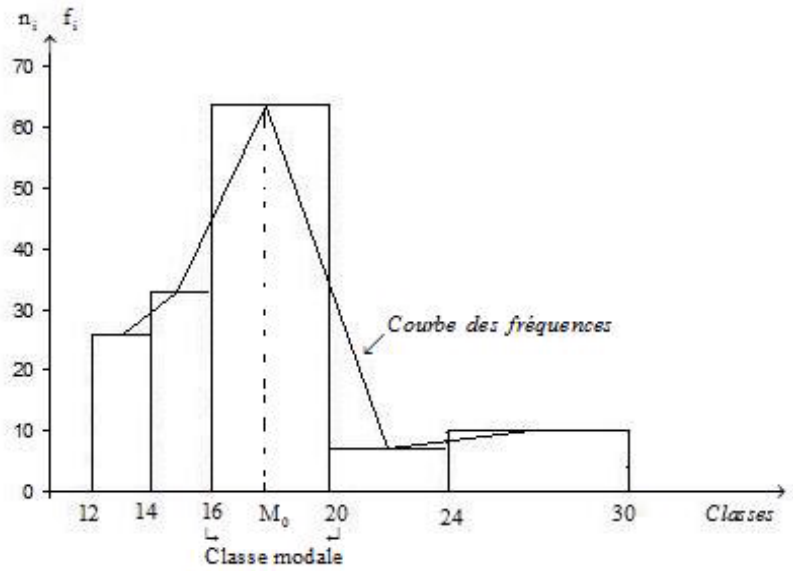
Lorsque la variable est discrète le mode est défini avec précision. Si, par exemple, deux valeurs successives de la variable statistique ont la même fréquence maximum, on dit qu'il y a un intervalle modal dont les extrémités correspondent à ces deux valeurs.



Détermination du mode : variable discrète

Lorsque la variable est continue, la détermination du mode est beaucoup moins précise car les fréquences dépendent du découpage en classe. L'utilisation de la courbe des fréquences ajustée sur l'histogramme, bien que peu

précise, conduit à une bonne estimation du mode dans le cas où les classes sont d'égale amplitude.



Détermination du mode : variable continue

Il est possible d'établir une formule d'interpolation linéaire pour le calcul exacte du mode dans le cas d'une répartition en classes d'amplitude quelconque.

3.1.2 Calcul du mode pour une distribution en classes d'innégales amplitudes

Considérons une série statistique continue regroupée en classes d'amplitudes inégales. Le mode est alors déterminé à l'intérieure de la classe modale (correspondant à la fréquence ou à l'effectif le plus grand). On peut identifier le mode comme la valeur médiane de la classe modale ou bien effectuer une interpolation linaire pour obtenir la valeur exacte du mode comme suit :

$$M_o = e_{i-1} + \frac{a_i\,(n_i - n_{i-1})}{(n_i - n_{i+1})\,(n_i - n_{i-1})} \tag{3.1}$$

où
e_{i-1} est la limite inférieure de la classe modale

a_i est l'amplitude de la classe modale
n_i est l'effectif de la classe modale
n_{i-1} est l'effectif de la classe inférieure la plus proche de la classe modale
n_{i+1} est l'effectif de la classe supérieure la plus proche de la classe modale

En adoptant les notations suivantes :

$$\Delta_m = n_i - n_{i-1} \quad \text{et} \quad \Delta_s = n_i - n_{i+1}$$

la relation (3.1) peut être présentée telle que :

$$M_o = e_{i-1} + \frac{a_i \Delta_m}{\Delta_s \Delta_m}$$

Example 54 *Soit x la variable statistique « taille d'une exploitation (en ha) ». Les résultats d'observations sont résumés dans le tableau statistique suivant :*

x	n_i	N_i
$]0;2[$	2	2
$[2;6[$	20	22
$[6;21[$	80	102
$[21;41[$	50	152
$[41;81[$	98	250
$[81;121[$	30	280
$\sum$	280	

La valeur du mode est calculée telle que :
*- **Valeur approchée** :*
La classe modale $[41;81[$ est d'effectif $n_i = 98$, d'où

$$M_o = 61 \text{ hectares}$$

*- **Valeur exacte :***
On utilise la formule d'interpolation linaire :

$$M_o = 41 + \frac{40 \times 48}{48 + 68} = 41 + \frac{1920}{116} = 57,55 \text{ hectares}$$

avec $e_{i-1} = 41$, $a_i = 40$, $\Delta_m = 98 - 50 = 48$ et $\Delta_s = 98 - 30 = 68$

Remarque 55 *Une distribution de fréquences peut présenter un seul mode (distribution unimodale) ou plusieurs modes (distribution bi ou trimodale).*

Propriétés

Les principaux avantages du mode font qu'il est facile à déterminer et qu'il a une signification immédiate. Par contre sa détermination n'est pas assez précise dans le cas continu. Elle dépend en partie du découpage en classes. Ainsi, il est sensible aux fluctuations d'échantillonnage et se prête très mal au calcul algébrique.

3.1.3 La médiane

Definition 56 *La médiane d'une distribution statistique, notée M_e, est la valeur de la variable statistique telle que le nombre des observations qui présentent une valeur inférieure à M_e soit égal au nombre des observations qui présentent une valeur supérieure à M_e.*

Remarque 57 *La médiane partage en deux effectifs égaux les observations rangées par ordre croissant ou décroissant. La médiane est la valeur M_e de la variable statistique pour laquelle la fréquence cumulée est égale à $\frac{1}{2}$, i.e.*

$$F(M_e) = \frac{1}{2}$$

Détermination pratique

Cas d'une variable discrète

Dans une série statistique composée de $2k+1$ observations et disposée par ordre croissant ou décroissant, la valeur de la $(k+1)^{ième}$ observation correspond à la médiane.

Example 58 *Considérons une série statistique composée de 9 mesures :* $18; 17; 13; 9; 8; 24; 19; 23; 28$. *Alors, la série disposée par ordre croissant donne :* $8; 9; 13; 17; 18; 19; 23; 24; 28$. *Donc la médiane est* $M_e = 18$.

Dans le cas d'une série statistique comportant $2k$ observations, il n'y a pas à proprement parler de médiane. Ainsi, on introduit la notion d'intervalle médian dont les extrémités correspondent aux valeurs de la k^{ieme} et de la $(k+1)^{ième}$ observations.

Example 59 *Supposons que la série statistique soit :* $8; 9; 13; 15; 17; 18; 19; 23; 24; 28$. *On convient de retenir pour valeur médiane la valeur* M_e *telle que :*

$$F(x_{i^-}) < \frac{1}{2} < F(x_{i^+})$$

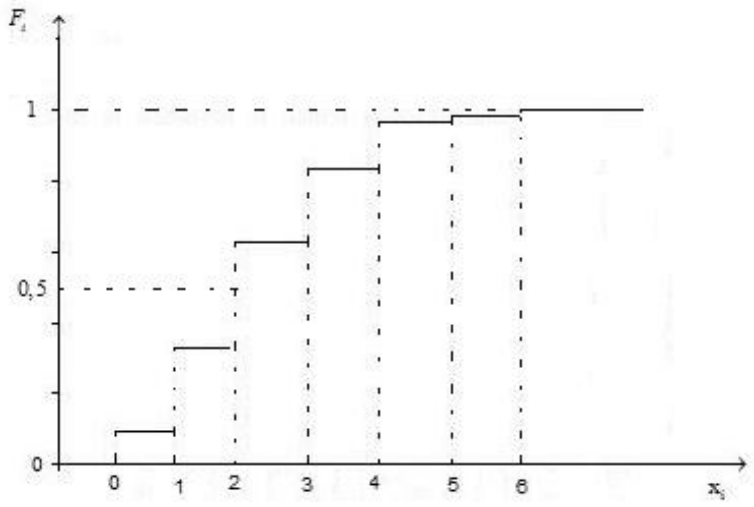
i.e telle que :

$$f_1 + f_2 + ... + f_{i-1} < \frac{1}{2} < f_1 + f_2 + ... + f_i$$

Graphiquement cette détermination est simple à partir de la courbe des effectifs cumulés ou celle des fréquences cumulées.

Example 60 *Considérons une distribution statistique représentée par le tableau suivant :*

x_i	0	1	2	3	4	5	6
N_i	24	81	156	205	240	248	250
F_i	$0,096$	$0,324$	$0,624$	$0,820$	$0,960$	$0,992$	$1,00$



Détermination graphique de la médiane : variable discrète

Cas d'une variable continue

Dans le cas d'une variable statistique continue la médiane est définie avec exactitude. Mais, en raison du regroupement par classe on ne peut généralement que la situer à l'intérieur d'une classe qu'on qualifiera de classe médiane.

Definition 61 *La classe n°i est une classe médiane si :*

$$F_{i-1} < \frac{1}{2} < F_i$$

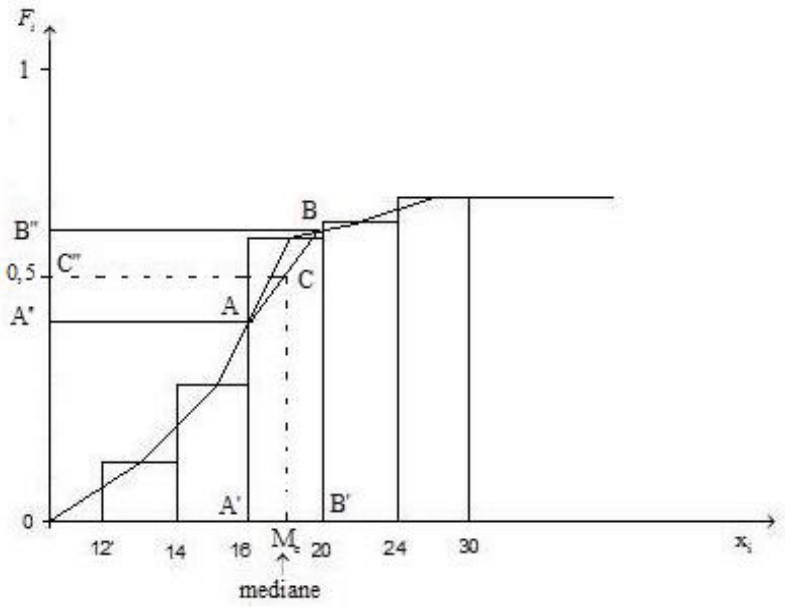
Détermination exacte de la médiane La solution de l'équation $F(M) = \frac{1}{2}$ est très simple graphiquement. On va montrer que la valeur de M_e est la même que celle obtenue par le calcul algébrique.

Proposition 62 *L'estimation de la valeur exacte de la médiane peut être obtenue par interpolation linéaire à l'intérieur de la classe médiane à l'aide de la relation suivante :*

$$M_e = e_{i-1} + a_i \frac{\frac{n}{2} - N_{i-1}}{n_i}$$

où e_{i-1} *est la borne inférieure de la classe médiane*
a_i *est l'amplitude de la classe médiane*
N_{i-1} *est l'effectif cumulé de la classe précédant la classe médiane*
n_i *est l'effectif de la classe médiane*
n *est l'effectif total de la population étudiée*

Démonstration : Considérons la figure suivante :



Détermination graphique de la médiane : variable continue

Alors, en utilisant quelques propriétés de géométrie euclidienne on remarque que :

$$\overline{OM} = \overline{OA'} + \overline{A'M}$$

Mais en vertu du théorème de Thalès :

$$\frac{\overline{A'M}}{\overline{A'B'}} = \frac{\overline{AC}}{\overline{AB}} = \frac{\overline{A''C''}}{\overline{A''B''}}$$

d'où

$$\overline{A'M} = \overline{A'B'}\frac{\overline{A''C''}}{\overline{A''B''}}$$

et par suite

$$\overline{OM} = \overline{OA'} + \overline{A'B'}\frac{\overline{A''C''}}{\overline{A''B''}}$$

En remplaçant ces quantités par leurs mesures algébriques, on obtient :

$$\overline{OM} = e_{i-1} + a_i\frac{\frac{n}{2} - nF\left(e_{i-1}\right)}{nF\left(e_i\right) - nF\left(e_{i-1}\right)}$$

D'où

$$M_e = e_{i-1} + a_i\frac{\frac{n}{2} - N_{i-1}}{n_i}$$

■

Propriétés La médiane répond assez bien à la plupart des conditions de Yule. Elle s'interprète aisément et se calcule facilement. Elle dépend de l'ensemble des observations. La valeur de la médiane n'est pas influencée par les observations aberrantes. Mais elle est sensible aux fluctuations d'échantillonnage et elle se prête mal au calcul algébrique en tant que solution de $F(M) = \frac{1}{2}$.

3.1.4 La médiale

Definition 63 *La médiale est la valeur de la variable statistique qui divise en deux la somme des valeurs de la variable.*

Example 64 *Considérons la répartition des employés d'une entreprise selon leur salaire mensuel net.*

Salaires (euros)	n_i	F_i	*Somme des salaires*	FQ_i *: Part cumu.des salaires*
$[800; 900[$	*25*	*0,212*	*21250*	*0,164*
$[900; 1000[$	*30*	*0,466*	*28500*	*0,385*
$[1000; 1100[$	*28*	*0,703*	*29400*	*0,613*
$[1100; 1500[$	*25*	*0,915*	*32500*	*0,865*
$[1500; 2000[$	*10*	*1*	*17500*	*1*
$\sum$	*118*		*129150*	

Alors, la médiale est à déterminer par interpolation dans la classe $[1000; 1100[$*, i.e.*

$$Médiale = 1000 + (1100 - 1000)\,\frac{0,5 - 0,385}{0,613 - 0,385} = 1050,4 \text{ euros}$$

Par comparaison, la médiane est déterminée par interpolation telle que :

$$Médiale = 1000 + (1100 - 1000)\,\frac{0,5 - 0,466}{0,703 - 0,466} = 1014,3 \text{ euros}$$

Remarque 65 *La médiale ne peut être inférieure à la médiane. La médiale est d'autant supérieure à la médiane que la distribution est plus concentrée. Dans l'exemple, l'écart médiale* $-$ *médiane* $= 1050,4 - 1014,3 = 36,1$ *euros. D'où, le ratio*

$$\frac{médiale - médiane}{étendue} = \frac{1050,4 - 1014,3}{2000 - 800} = 0,03$$

3.2 La moyenne arithmétique

Definition 66 *La moyenne arithmétique d'une variable statistique* x*, notée* $\overline{x}$*, est égale à la somme des valeurs prises par cette variable divisée par le nombre des observations.*

Example 67 *Les* 8 *ouvriers d'une petite entreprise ont perçu en janvier* 1990 *les salaires suivants :* $7500; 8300; 9100; 9600; 10700; 11300; 12000; 12500$ $D.A.$ *Le salaire moyen des ouvriers de cette entreprise en janvier* 1990 *est alors :*

$$\frac{7500 + 8300 + 9100 + 9600 + 10700 + 11300 + 12000 + 12500}{8} = 10125 \ D.A$$

3.2.1 Moyenne arithmétique simple

Definition 68 *Considérons une série statistique comportant n observations $x_1, x_2, ..., x_i, ..., x_n$. Alors la moyenne arithmétique simple est calculée à l'aide de l'expression suivante :*

$$\overline{x} = \frac{x_1 + x_2 + ... + x_i + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Remarque 69 *A chaque valeur prise par la variable statistique correspond un seul individu. Cette moyenne arithmétique est dite simple par opposition à la moyenne arithmétique pondérée.*

3.2.2 Moyenne arithmétique pondérée

Example 70 *Reprenons l'exemple précédent et supposons que les* 8 *ouvriers de l'entreprise aient perçu les salaires suivants :* 8300; 8300; 9600; 9600; 9600; 10800; 10800; 12500 *D.A. Le calcul du salaire moyen peut être effectué comme précédemment. Cependant, on aurait pu présenter les observations dans un tableau statistique tel que :*

Salaire x	*Effectif* n_i
8300	2
9600	3
10800	2
12500	1
$Total$	8

Il serait bien entendu erroné de dire que le salaire moyen des ouvriers est :

$$\frac{8300 + 9600 + 10800 + 12500}{4} = 10300 \ D.A$$

Les salaires doivent être pondérés par les effectifs correspondants, et donc :

$$\overline{x} = \frac{2 \times 8300 + 3 \times 9600 + 2 \times 10800 + 12500}{8} = 9937,50 \ D.A$$

La moyenne ainsi calculée est appelée moyenne arithmétique pondérée. Les coefficients de pondération sont les fréquences absolues des différentes valeurs de la variable statistique. Ce type de calcul de la moyenne est naturellement utilisé dans le cas d'observations regroupées en classe.

Definition 71 *Soit x une variable statistique pouvant prendre les k valeurs $x_1, x_2, ..., x_i, ..., x_k$ auquelles correspondent respectivement les k fréquences absolues ou effectifs $n_1, n_2, ..., n_i, ..., n_k$. Alors la moyenne arithmétique pondérée de cette variable a pour expression :*

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

Remarque 72 *$\overline{x}$ peut aussi s'écrire :*

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

où f_i représente la fréquence relative des différentes valeurs de la variable statistique. Les f_i, $i = 1, 2, ..., k$ sont appelés coefficients de pondération.

3.3 Calcul pratique de la moyenne arithmétique

3.3.1 Cas d'une variable discrète

Example 73 *Reprenons l'exemple du nombre de ventes d'un certain type d'appareil par jour ouvrable. Le nombre moyen de ventes n'est autre que la moyenne arithmétique des ventes. Pour calculer cette moyenne, il est toujours possible d'utiliser directement la formule de la définition. Mais souvent on utilise le tableau statistique où l'on dispose les différentes étapes des calculs tel que :*

x_i	n_i	$n_i x_i$
0	24	0
1	57	57
2	75	150
3	53	159
4	33	132
5	7	35
6	4	24
$Totaux$	$\sum_{i=1}^{k} n_i = 253$	$\sum_{i=1}^{k} n_i x_i = 557$

Alors,

$$\overline{x} = \frac{1}{253}\sum_{i=1}^{k} n_i x_i = \frac{557}{253} \approx 2,20$$

Le nombre moyen de ventes par jour ouvrable est donc $2,20$.

Souvent les calculs qui résultent de l'application de la formule de définition sur les valeurs brutes, peuvent s'avérer assez fastidieux. Il est alors possible d'alléger ces calculs en procédant à une transformation des données brutes, par exemple en choisissant une nouvelle origine x_0 pour la variable statistique x. On définit ainsi une nouvelle variable x'_i appelée variable auxiliaire telle que :

$$x'_i = x_i - x_0 \tag{3.2}$$

Théorème 74 *En adoptant le changement d'origine* $x'_i = x_i - x_0$*, on obtient la même relation entre* $\overline{x'}$ *et* $\overline{x}$*, i.e :*

$$\overline{x'} = \overline{x} - x_0$$

Démonstration : En effet, à chaque valeur x_i correspond une nouvelle valeur x'_i. Et d'après (3.2) on a :

$$n_i x'_i = n_i x_i - n_i x_0 \quad i = 1, 2, ..., k \tag{3.3}$$

Et en sommant les k équations (3.3), on obtient :

$$\sum_{i=1}^{k} n_i x'_i = \sum_{i=1}^{k} n_i x_i - \sum_{i=1}^{k} n_i x_0$$

Comme $\sum_{i=1}^{k} n_i = n$, alors en divisant par n les deux membres de l'égalité ci-dessus, il vient :

$$\frac{1}{n}\sum_{i=1}^{k} n_i x'_i = \frac{1}{n}\sum_{i=1}^{k} n_i x_i - x_0$$

D'où

$$\overline{x'} = \overline{x} - x_0$$

■

Remarque 75 *Ainsi on pourra calculer* $\overline{x'}$ *et en déduire* $\overline{x}$.

Example 76 *Reprenons l'exemple* 84 *ci-dessus et prenons pour nouvelle origine* $x_0 = 2$. *La variable auxiliaire est alors définie par :*

$$x"_i = x_i - 2$$

On obtient ainsi le tableau statistique suivant :

x_i	n_i	x'_i	$n_i x'_i$
0	24	-2	-48
1	57	-1	-57
2	75	0	0
3	53	1	53
4	33	2	66
5	7	3	21
6	4	4	16
$Totaux$	$\sum_{i=1}^{k} n_i = 253$		$\sum_{i=1}^{k} n_i x'_i = 51$

Alors, $\overline{x'} = \frac{1}{n}\sum_{i=1}^{k} n_i x'_i = \frac{51}{253} \approx 0,20$.
D'où

$$\overline{x} = \overline{x'} + 2 = 2,20$$

3.3.2 Cas d'une variable continue

La distribution d'une variable statistique continue est présentée, en général, sous forme de classes. La formule de définition de la moyenne ne peut être appliquée directement car on ne connait pas les valeurs exactes prises par la variable statistique, mais seulement le nombre d'observations à l'intérieur de chaque classe. On supposera alors que les observations sont réparties uniformément dans chaque classe. C'est à dire n'importe quelle valeur à l'intérieur de la classe peut représenter cette dernière. Par convention et sans trop de perte d'information, on prendra le centre de la classe comme représentant. Cette convention implique un biais systématique dans le calcul de la moyenne. Le centre de la classe $n°i$ sera noté en général X_i, et il est donné par la relation suivante :

$$X_i = \frac{e_i + e_{i-1}}{2}$$

où e_i et e_{i-1} désignent respectivement la borne supérieure et la borne inférieure de la classe $n°i$.

Ainsi on est ramené au calcul de la moyenne arithmétique dans le cas d'une variable discrète que l'on peut effectuer directement à partir de la définition ou en utilisant une variable auxiliaire.

Example 77 *Reprenons la distribution des ouvriers d'une entreprise suivant leur salaire mensuel.*
On prendra pour origine le centre de la classe modale, i.e. $X_{M_o} = 18000$. *On remarque aussi que les nombres* $\xi_i = X_i - 18000$, $i = 1, 2, ..., k$ *sont divisibles par* 1000. *Donc on prendra pour variable auxiliaire telle que :*

$$x'_i = \frac{X_i - 18000}{1000}$$

Les calculs seront toujours disposés dans un tableau du genre ci-dessous :

Classe de Salaire	n_i	X_i	X'_i	$n_i X'_i$
$12000 \leq x < 14000$	26	13000	-2	-52
$14000 \leq x < 16000$	33	15000	-1	-33
$16000 \leq x < 20000$	64	18000	0	0
$20000 \leq x < 24000$	7	22000	1	7
$24000 \leq x < 30000$	10	27000	2	20
Total	140			$\sum_{i=1}^{k} n_i x'_i = -58$

D'où

$$\overline{X'} = \frac{1}{148} \sum_{i=1}^{5} n_i X'_i = -\frac{58}{140} \approx -0,414$$

Et par conséquent

$$\overline{X} = 1000\overline{X'} + 18000 \approx 17586 \ D.A$$

D'une façon générale, le choix d'une nouvelle origine X_{M_o} et d'une nouvelle échelle de mesure u va permettre de réduire le volume des calculs.
On définit une variable auxiliaire X'_i par la transformation linéaire :

$$X'_i = \frac{X_i - X_{M_o}}{u} \tag{3.4}$$

où X_{M_o} est, en général, le centre de la classe modale et u est le $PGCD$ des amplitudes de classes.
En suivant le même raisonnement que pour le cas discret, on remarque que

si l'on adopte le changement de variable (3.4) il existe la même relation entre $\overline{X}$ et $\overline{X'}$, i.e.

$$\overline{X} = u\overline{X'} + X_{M_o}$$

3.3.3 Propriétés de la moyenne arithmétique

- La moyenne arithmétique répond assez bien à l'ensemble des conditions de Yule. Elle se prête facilement au calcul algébrique et a une signification concrète. Mais elle est sensible aux fluctuations d'échantillonnage.
- La somme algébrique des écarts des observations à la moyenne est nulle, i.e.

  $$\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right) = 0$$

  En effet

  $$\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right) = \sum_{i=1}^{k} n_i x_i - \overline{x} \sum_{i=1}^{k} n_i = n\overline{x} - n\overline{x} = 0$$

- La somme des carrés des écarts des observations à la moyenne est inférieure à la somme des carrés des écarts par rapport à toute autre valeur. En effet, soit :

  $$S(b) = \sum_{i=1}^{k} n_i \left(x_i - b\right)^2$$

  où $S(b)$ est un polynôme du second degré en b.
  Le polynôme $S(b)$ est minimum au point où sa dérivée par rapport à b est nulle, i.e.

  $$\frac{dS(b)}{db} = S'(b) = -2 \sum_{i=1}^{k} n_i \left(x_i - b\right) = 0$$

D'où

$$\sum_{i=1}^{k} n_i x_i - b \sum_{i=1}^{k} n_i = nb$$

Et par conséquent

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

– La moyenne $\overline{x}$ d'une population composée de deux sous-populations P_1 de moyenne $\overline{x_1}$ et P_2 de moyenne $\overline{x_2}$, s'exprime simplement en fonction de $\overline{x_1}$ et $\overline{x_2}$.
Supposons que la population P possède un effectif de n individus, et que les sous-populations P_1 et P_2 ont des effectifs respectifs n_1 et n_2 tels que $n_1+n_2 = n$. Soit n_{1i} le nombre d'individus de la sous-population P_1 présentant la modalité x_i du caractère et soit n_{2i} le nombre d'individus présentant la même modalité dans la sous-population P_2. Donc, dans la population P l'effectif des individus présentant la modalité x_i est $n_i = n_{1i} + n_{2i}$.
D'autre part

$$\overline{x_1} = \frac{1}{n_1}\sum_{i=1}^{k} n_{1i}x_i \quad \text{et} \quad \overline{x_2} = \frac{1}{n_2}\sum_{i=1}^{k} n_{2i}x_i$$

Alors, la moyenne $\overline{x}$ de la population est :

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{1}{n_1+n_2}\sum_{i=1}^{k}\left(n_{1i}+n_{2i}\right)x_i \\
&= \frac{1}{n_1+n_2}\left\{\sum_{i=1}^{k} n_{1i}x_i + \sum_{i=1}^{k} n_{2i}x_i\right\} \\
&= \frac{n_1}{n_1+n_2}\frac{1}{n_1}\sum_{i=1}^{k} n_{1i}x_i + \frac{n_2}{n_1+n_2}\frac{1}{n_2}\sum_{i=1}^{k} n_{2i}x_i
\end{aligned}$$

D'où

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

Donc la moyenne de la population totale apparait comme la moyenne pondérée des moyennes des sous populations.
Plus généralement, pour h populations on a :

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

Positions respectives du mode, de la médiane et de la moyenne arithmétique

Pour les distributions symétriques on a :

$$\overline{x} = M_e = M_o$$

Pour les distributions asymétriques on a deux situations selon que la distribution est plus plate à gauche qu'à droite et vice versa :

$$\overline{x} < M_e < M_o \quad \text{ou bien} \quad M_o < M_e < \overline{x}$$

3.4 Autres types de moyennes

En plus de la moyenne arithmétique, il existe d'autres types de moyennes. On les rencontre beaucoup moins fréquemment, mais leur utilisation est cependant recommandée dans certains cas.

3.4.1 Moyenne géométrique

Definition 78 *La moyenne géométrique simple d'une série de valeurs $x_1, x_2, ..., x_n$, notée G, est définie par :*

$$G = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} \tag{3.5}$$

Remarque 79 *En prenant le logarithme de la relation* (3.5)*, on obtient une autre formule pour la moyenne géométrique simple :*

$$\ln G = \frac{1}{n} \sum_{i=1}^{n} \ln x_i \tag{3.6}$$

Example 80 *Considérons une série de neuf nombres :* $18; 17; 13; 9; 8; 24; 19; 23; 28$. *Leur moyenne géométrique est alors :*

$$G = (18 \times 17 \times 13 \times 9 \times 8 \times 24 \times 19 \times 23 \times 28)^{\frac{1}{9}} = 16,36$$

On peut aussi calculer la moyenne géométrique en utilisant la formule (3.6)*. En effet*

$$\ln G = \frac{1}{9} \ln (18 \times 17 \times 13 \times 9 \times 8 \times 24 \times 19 \times 23 \times 28) = 2,795$$

Alors

$$e^{\ln G} = e^{2,795} = 16,36$$

Definition 81 *Soit x une variable statistique pouvant prendre les k valeurs $x_1, x_2, ..., x_n$. On dispose d'une série statistique de taille n comportant n_1 fois x_1, n_2 fois x_2,..., n_k fois x_k. Alors la moyenne géométrique pondérée est donnée par l'expression :*

$$G = \left(\prod_{i=1}^{k} x_i^{n_i} \right)^{\frac{1}{n}} \tag{3.7}$$

Remarque 82 *En prenant le logarithme dans la formule* (3.7) *on obtient une autre expression pour la moyenne géométrique pondérée :*

$$\ln G = \frac{1}{n} \sum_{i=1}^{k} n_i \ln x_i \tag{3.8}$$

Par ailleurs, la formule (3.7) *peut aussi s'écrire :*

$$G = \left(\prod_{i=1}^{k} x_i^{\frac{n_i}{n}} \right) = \left(\prod_{i=1}^{k} x_i^{f_i} \right)$$

où $f_i = \frac{n_i}{n}$ est la fréquence de la modalité x_i.

3.4.2 Propriétés de la moyenne géométrique

Considérons deux séries statistiques de même taille n, de deux variables statistiques x et y :

$$x_1, x_2, ..., x_n \quad \text{et} \quad y_1, y_2, ..., y_n$$

– Formons les produits $z_i = x_i y_i$, $i = 1, 2, ..., n$ et calculons moyenne géométrique $G(z)$ de ces produits :

$$G(z) = \left(\prod_{i=1}^{n} z_i \right)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} x_i y_i \right)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} \left(\prod_{i=1}^{n} y_i \right)^{\frac{1}{n}} = G(x)G(y)$$

Donc la moyenne géométrique du produit xy est égale au produit de moyennes géométriques de x et de y.

– Formons les rapports $q_i = \frac{x_i}{y_i}$ et calculons leur moyenne géométrique $G(q)$ telle que :

$$G(q) = \left(\prod_{i=1}^{n} q_i\right)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} \left(\frac{x_i}{y_i}\right)\right)^{\frac{1}{n}} = \frac{\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}}{\left(\prod_{i=1}^{n} y_i\right)^{\frac{1}{n}}} = \frac{G(x)}{G(y)}$$

Donc la moyenne géométrique du rapport $\frac{x}{y}$ est le rapport des moyennes géométriques de x et de y.

3.4.3 Moyenne harmonique

Definition 83 *La moyenne harmonique d'une série de valeurs $x_1, x_2, ..., x_n$, notée H, est définie par l'expression :*

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ...\frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Example 84 *Reprenons l'exemple* 88 *et calculons la moyenne harmonique :*

$$H = \frac{9}{\frac{1}{18} + \frac{1}{17} + \frac{1}{13} + \frac{1}{9} + \frac{1}{8} + \frac{1}{24} + \frac{1}{19} + \frac{1}{23} + \frac{1}{28}} = 14,97$$

Example 85 *Un spéculateur a consacré pendant* 4 *années la même somme S à l'achat de lingots d'or aux prix respectifs* 5400 *;* 5500 *;* 5800 *et* 6400 *U le kg. Le prix moyen d'achat du kilogramme d'or par le spéculateur n'est pas la moyenne arithmétique. En effet, la dépense totale effectuée par le spéculateur est $4S$. La première année il a acheté $q_1 = \frac{S}{5400}$ kg d'or, la deuxième année $q_2 = \frac{S}{5500}$ kg d'or, etc. Au total il a acheté la quantité d'or suivante :*

$$q_1 + q_2 + q_3 + q_4 = S\left[\frac{1}{5400} + \frac{1}{5500} + \frac{1}{5800} + \frac{1}{6400}\right]$$

Le prix d'achat moyen du kg d'or est donc :

$$P = \frac{4S}{q_1 + q_2 + q_3 + q_4} = \frac{4}{\frac{1}{5400} + \frac{1}{5500} + \frac{1}{5800} + \frac{1}{6400}} = 5750,6$$

Definition 86 *Considérons une variable statistique pouvant prendre les valeurs $x_1, x_2, ..., x_k$. Et supposons que l'on a obtenu n réalisations de cette variable avec les effectifs respectifs $n_1, n_2, ..., n_k$ tels que $\sum_{i=1}^{k} n_i = n$. Alors la moyenne harmonique pondérée est donnée par l'expression suivante :*

$$H = \frac{n}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + ...\frac{n_k}{x_n}} = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$

Example 87 *Une entreprise de transport possède* 10 *camions qui font la rotation entre un endroit A et un endroit B. Au cours d'une de ces rotations le trajet AB (distance D) a été couvert par ces véhicules aux vitesses moyennes suivantes :*

Vitesse Moy (Km/h)	40	60	70
Nombre de camions	3	5	2

Au total les camions ont parcouru une distance $10D$. *Pour couvrir le trajet AB,* 3 *camions ont mis un temps* $T_1 = \frac{D}{40}$, 5 *camions un temps* $T_2 = \frac{D}{60}$ *et les deux autres un temps* $T_3 = \frac{D}{70}$. *Au total le temps T mis par l'ensemble des camions pour parcourir la distance* $10D$ *a été :*

$$T = 3T_1 + 5T_2 + 2T_3 = D\left(\frac{3}{40} + \frac{5}{60} + \frac{2}{70}\right)$$

Donc, pour l'ensemble des camions la vitesse moyenne V a été :

$$V = \frac{10D}{3T_1 + 5T_2 + 2T_3} = \frac{10}{\frac{3}{40} + \frac{5}{60} + \frac{2}{70}} = 53,5\ Km/h$$

3.4.4 Généralisation de la notion de moyenne

Toutes les moyennes étudiées ont été définies suivant un principe commun. En effet, pour le calcul de chaque type de moyenne les observations ont été introduites sous une forme particulière. Par exemple, pour la moyenne harmonique ce fut l'inverse des observations, pour la moyenne géométrique ce fut leur logarithme. D'une manière générale, la définition d'une moyenne fait intervenir une fonction f des observations.

Definition 88 *Soit $\varphi(x)$ une fonction monotone de la variable statistique x. On appelle $\varphi-$moyenne le nombre C défini tel que :*

$$\varphi(C) = \frac{1}{n}\sum_{i=1}^{k} n_i \varphi(x_i) \tag{3.9}$$

Remarque 89 *A partir de cette définition générale on retrouve facilement les formules des différentes moyennes :*
1. Si on considère la fonction $\varphi(x) = \frac{1}{x}$. *D'après la relation* (3.9), $\varphi(H) = \frac{1}{n}\sum_{i=1}^{k} n_i \varphi(x_i)$ *où* H *est la moyenne harmonique. Alors*

$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{k}\frac{n_i}{x_i} \quad \Longrightarrow \quad H = \frac{n}{\sum_{i=1}^{k}\frac{n_i}{x_i}}$$

2. Si on prend $\varphi(x) = \ln x$, *alors on retrouve la moyenne géométrique :*

$$\ln G = \frac{1}{n}\sum_{i=1}^{k} n_i \ln x_i \quad \Longrightarrow \quad G = \left(\prod_{i=1}^{k} x_i^{n_i}\right)^{\frac{1}{n}}$$

3. La fonction Identité $f(x) = x$ *redonne, bien entendu, la moyenne arithmétique* $\overline{x}$ *:*

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

4. A partir de cette formule générale on peut construire de nouvelles moyennes. Par exemple, si on considère la fonction $\varphi(x) = x^2$, *on définit ainsi la moyenne quadratique* MQ *:*

$$(MQ)^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 \quad \Longrightarrow \quad MQ = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2}$$

3.4.5 Propriétés comparées des différentes moyennes

De toutes les moyennes c'est certainement la moyenne arithmétique qui répond le mieux aux conditions de Yule. En particulier, elle est d'un calcul aisé et rapide, et sa signification est facile à concevoir. Les moyennes arithmétique et quadratique sont influencées par les éléments le plus fréquents de la série. Au contraire, les moyennes hamonique et géométrique tendent à réduire l'influence des valeurs les plus fréquentes en faveur des valeurs les plus rares. Les moyennes harmonique H, géométrique G, arithmétique $\overline{x}$ et quadratique MQ d'une même série statistique sont, en général, classées dans l'ordre suivant :

$$H \leq G \leq \overline{x} \leq MQ$$

Example 90 *Pour la série des nombres* $18; 17; 13; 9; 8; 24; 19; 23$ *et* 28*, les différents types de moyennes sont telles que :*

$$H = 14,97 \ \leq \ G = 16,36 \quad \leq \quad \overline{x} = 17,66 \quad \leq \quad MQ = 18,78$$

Chapitre 4

Les caractéristiques de dispersion

Les caractéristiques de dispersion les plus fréquemment utilisées sont l'étendue, l'intervalle interquartile, l'écart absolu moyen, la variance et l'écart-type. Ces deux dernières caractéristiques sont les plus couramment utilisées. Le calcul de l'indice de concentration peut être recommandé dans certains cas. L'étendue et l'intervalle interquartile sont, dans leur principe, du type de la médiane. Les observations y interviennent par leurs rangs et non par leurs valeurs. L'écart absolu moyen et l'écart-type, au contraire, font intervenir l'écart à la moyenne arithmétique de chacune des observations. Ceux sont des moyennes d'écart à la moyenne. L'indice de concentration repose, quant à lui, sur un principe tout à fait différent.

4.1 Les différentes caractéristiques de dispersion

4.1.1 L'étendue

Definition 91 *L'étendue d'une distribution statistique, notée w, est la différence entre la plus grande et la plus petite des valeurs observées, i.e.*

$$w = x_{(n)} - x_{(1)}$$

où $x_{(n)} = \max_i (x_i)$ *et* $x_{(1)} = \min_i (x_i)$.

Propriétés

La signification de l'étendue est évidente et son calcul est immédiat. Mais cette caractéristique présente des inconvénients. Elle ne dépend que des termes extrèmes de la série et elle est donc très sensibles aux fluctuations d'échantillonnage. La forme de la distribution entre les extrèmes n'est pas prise en compte. Donc, l'étendue est une caractéristique de dispersion imparfaite.

4.1.2 Les quartiles et l'intervalle interquartile

Pour remédier aux inconvénients de l'étendue, on a pensé à minimiser l'influence des termes extrèmes de la série sur le calcul de la caractéristique de dispersion. Pour cela, on définit les quartiles Q_1, Q_2 et Q_3. Ces derniers sont les valeurs de la variable statistique telles que, les observations étant rangées par ordre croissant, un quart de celles-ci est inférieur à Q_1, un quart est compris entre Q_1 et Q_2, un quart compris entre Q_2 et Q_3, et le dernier quart est supérieur à Q_3. En d'autres termes Q_1, Q_2 et Q_3 sont les valeurs de la variable statistique pour lesquelles la fonction cumulative est respectivement est telle que :

$$F(Q_1) = 0,25 \;\; ; \;\; F(Q_2) = 0,50 \;\; et \;\; F(Q_3) = 0,75$$

Remarque 92 *Le deuxième quartile Q_2 est donc égal à la médiane.*

Definition 93 *On appelle intervalle interquartile, noté IQ, la différence entre les valeurs du troisième et du premier quartile, i.e.*

$$IQ = Q_3 - Q_1$$

Remarque 94 *L'intervalle interquartile est donc l'intervalle qui contient* 50% *des observations tout en laissant* 25% *à sa droite et* 25% *à sa gauche.*

Détermination pratique des quartiles

Le quartile se détermine de la même manière que la médiane. Soit il est déterminé graphiquement à partir de la courbe des effectifs cumulés ou celle des fréquences cumulées, Soit il est calculé par interpolation linéaire. La

formule de détermination des quartiles est la même que celle utilisée pour la détermination de la médiane $M_e = Q_2$. En effet,

$$Q_h = e_{h-1} + a_h \frac{\frac{hn}{4} - N_{h-1}}{n_h}$$

où e_{h-1} est la borne inférieure de la classe contenant le quartile $n°h$, $h = 1, 2, 3, 4$

a_h est l'amplitude de la classe contenant le quartile $n°h$, $h = 1, 2, 3, 4$

N_{h-1} est l'effectif cumulé de la classe précédant celle contenant le quartile $n°h$, $h = 1, 2, 3, 4$

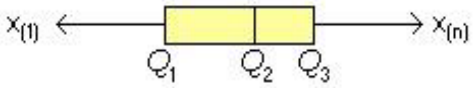
n_h est l'effectif de la classe contenant le quartile $n°h$, $h = 1, 2, 3, 4$

n est l'effectif total de la population étudiée

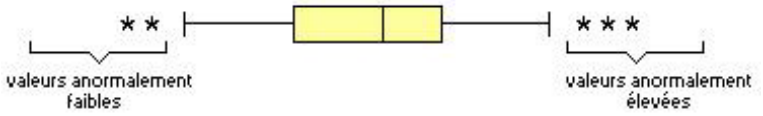
Propriétés

Les avantages de l'intervalle interquartile sont la rapidité de son calcul et sa signification immédiate. Sa détermination n'est pas précise et il se prête mal au calcul algébrique.

Diagramme en boîte (ou boîte à moustaches) Il s'agit d'un diagramme permettant de positionner les quartiles Q_1, Q_2, Q_3, au moyen de rectangles de largeur arbitraire, prolongés par des "moustaches" de part et d'autre, de longueur au plus égale à une fois et demie $Q_3 - Q_1$.



Si la plus petite ou la plus grande valeur observée se trouvent à l'intérieur, on raccourcit les moustaches correspondantes ; si elles se trouvent à l'extérieur, on positionne à part les valeurs "aberrantes" qui dépassent des moustaches :



Ces diagrammes sont surtout utiles pour comparer rapidement l'allure générale de plusieurs distributions.

4.1.3 Généralisation de la notion de quartile

Les déciles

Pour obtenir les quartiles on a divisé en quatre parties égales l'effectif de la série statistique préalablement ordonnée par ordre croissant. Les déciles, au nombre de 9, séparent l'effectif de la population étudiée en 10 parties égales. Le premier décile D_1 est tel que $\frac{1}{10}$ des observations lui est inférieur et d'une façon générale $\frac{1}{10}$ des observations est compris entre deux déciles successifs et on a :

$$F(D_1) = 0,1 \ ; \ F(D_2) = 0,2 \ ; \ F(D_3) = 0,3 \ ; ...; \ F(D_9) = 0,9$$

Remarque 95 *Les déciles sont déterminés de la même manière que les quartiles. La formule de détermination des déciles est la même que celle utilisée pour la détermination de la médiane $M_e = Q_2$. En effet,*

$$Q_h = e_{h-1} + a_h \frac{\frac{hn}{10} - N_{h-1}}{n_h}$$

où e_{h-1} est la borne inférieure de la classe contenant le décile $n°h$, $h = 1, 2, ...$
a_h est l'amplitude de la classe contenant le quartile $n°h$, $h = 1, 2, ...$
N_{h-1} est l'effectif cumulé de la classe précédant celle contenant le décile $n°h$, $h = 1, 2, ...$
n_h est l'effectif de la classe contenant le quartile $n°h$, $h = 1, 2, ...$
n est l'effectif total de la population étudiée

Les percentiles

Pour des séries comportant suffisamment d'observations on peut définir les percentiles tels que 1% des observations est compris entre deux percentiles successifs, i.e.

$$F(P_1) = 0,01 \ ; \ F(P_2) = 0,02 \ ; ...; \ F(P_{99}) = 0,99$$

Les quantiles

Plus généralement, on peut définir les quantiles.

Definition 96 *Le quantile d'ordre α $(0 \leq \alpha \leq 1)$, noté q_α, est la solution de l'équation $F(x) = \alpha$. Ainsi, en désignant par F^{-1} la fonction inverse de la fonction F on a alors :*

$$q_\alpha = F^{-1}(\alpha)$$

Remarque 97 *Une proportion α des individus de la population possède un caractère C de mesure inférieure à q_α.*

4.1.4 L'écart absolu moyen

Definition 98 *Soit x une variable statistique pouvant prendre les k valeurs $x_1, x_2, ..., x_k$ auxquelles correspondent les effectifs respectifs $n_1, n_2, ..., n_k$. L'écart absolu moyen, noté $\overline{e}$, est alors la moyenne arithmétique des valeurs absolues des écarts à la moyenne arithmétique, i.e.*

$$\overline{e} = \frac{1}{n} \sum_{i=1}^{n} n_i \left| x_i - \overline{x} \right|$$

Propriétés

L'écart absolu moyen satisfait assez bien aux premières conditions de Yule. Mais il se prète très mal au calcul algébrique. L'écart absolu moyen est minimum lorsqu'on prend les écarts par rapport à la médiane.

4.1.5 La variance et l'écart-type

L'écart-type sera défini à partir des carrés des écarts des observations à leur moyenne arithmétique. On déterminera de cette façon une sorte de distance moyenne des observations à la moyenne arithmétique. Cette distance, au sens mathématique du terme, servira comme mesure de dispersion de la variable statistique autour de sa caractéristique de tendance centrale.

Definition 99 *Considérons une variable statistique x pouvant prendre k valeurs $x_1, x_2, ..., x_k$ auxquelles correspondent les effectifs $n_1, n_2, ..., n_k$ tels que $\sum_{i=1}^{k} n_i = n$. Alors la variance de la variable statistique x , notée $Var(x)$ ou bien σ_x^2, est la moyenne arithmétique des carrés des écarts à la moyenne arithmétique :*

$$Var\left(x\right) = \sigma_x^2 = \frac{1}{n} \sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^2$$

Definition 100 *L'écart-type, noté σ_x, est égal à la racine carrée de la variance :*

$$\sigma_x = \sqrt{Var\left(x\right)} = \sqrt{\frac{1}{n} \sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^2}$$

Remarque 101 *L'écart-type est appelé parfois écart quadratique moyen* (EQM).

Example 102 *Considérons la série des salaires horaires de sept ouvriers d'une entreprise :* $30; 45; 51; 62; 70; 78; 84$ *D.A. Alors*

$$\overline{x} = 60 \ D.A$$

Les écarts à la moyenne arithmétique $(x_i - \overline{x})$ *sont :* $-30; -15; -9; 2; 10; 18; 24$.
Leurs carrés $(x_i - \overline{x})^2$ *sont :* $900; 225; 81; 4; 100; 324; 576$.
D'où

$$Var\,(x) = \frac{2210}{7} = 315,71 \quad et \quad \sigma_x = \sqrt{Var\,(x)} = \sqrt{315,71} = 17,76 \ DA$$

4.2 Calcul pratique de la variance et de l'écart-type

Les calculs de la moyenne arithmétique et de l'écart-type vont généralement de pair. On conservera le tableau déjà utilisé dans le calcul de la moyenne.

4.2.1 Cas d'une variable discrète

Calcul au moyen de la formule brute

Example 103 *Considérons le tableau statistique suivant et calculons l'écart-type.*

x_i	n_i	$n_i x_i$	$x_i - \overline{x}$	$n_i\,(x_i - \overline{x})^2$
1	25	25	$-2,18$	$136,81$
2	55	110	$-1,18$	$76,582$
3	75	225	$-0,18$	$2,43$
4	50	200	$0,82$	$33,62$
5	35	175	$1,82$	$115,934$
6	10	60	$2,82$	$79,524$
$Totaux$	$\sum_{i=1}^{k} n_i = 250$	$\sum_{i=1}^{k} n_i x_i = 795$		$\sum_{i=1}^{k} n_i\,(x_i - \overline{x})^2 = 444,9$

L'effectif $n = 250$ *et* $\overline{x} = 3,18$, *alors*

$$Var(x) = \sigma_x^2 = 1,78 \quad et \quad \sigma_x = \sqrt{1,78} = 1,335$$

Notons que même pour une distribution statistique aussi simple que celle étudiée, le calcul de l'écart-type est assez long et fastidieux

Calcul au moyen de la formule développée

Il est possible de développer la formule de définition de la variance telle que :

$$\begin{aligned} Var(x) &= \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i^2 - 2\overline{x}x_i + \overline{x}^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \frac{2\overline{x}}{n}\sum_{i=1}^{k} n_i x_i + \frac{\overline{x}^2}{n}\sum_{i=1}^{k} n_i \end{aligned}$$

D'où

$$Var(x) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2 \tag{4.1}$$

Remarque 104 *La relation* (4.1) *est appelée formule développée de la variance.*

Remarque 105 *Reprenons l'exemple* 111 *et calculons l'écart-type :*

x_i	n_i	$n_i x_i$	$n_i x_i^2$
1	25	25	25
2	55	110	220
3	75	225	675
4	50	200	800
5	35	175	875
6	10	60	360
$Totaux$	$\sum_{i=1}^{k} n_i = 250$	$\sum_{i=1}^{k} n_i x_i = 795$	$\sum_{i=1}^{k} n_i \overline{x}_i^2 = 2955$

Alors , $\overline{x} = 3,18$ *et* $Var(x) = \sigma_x^2 = 11,82 - (3,18)^2 = 1,71$.
D'où

$$\sigma_x = \sqrt{1,71} = 1,31$$

Remarque 106 *Le résultat ainsi obtenu est plus précis que celui obtenu par la méthode précédente, car l'approximation n'intervient qu'à travers le terme* $\overline{x}$. *Il est encore possible de simplifier les calculs en utilisant une translation d'origine.*

Calcul avec changement de variable

Considérons le changement de variable (changement d'origine) suivant :

$$x'_i = x_i - x_0 \tag{4.2}$$

Nous avons déjà établi qu'il existe la même relation entre $\overline{x'}$ et $\overline{x}$, i.e.

$$\overline{x'} = \overline{x} - x_0 \tag{4.3}$$

D'où, en retranchant les relations (4.2) et (4.3) membre à membre, on obtient :

$$x'_i - \overline{x'} = x_i - \overline{x}$$

Par suite, et d'après la définition de la variance :

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(x'_i - \overline{x'}\right)^2 = \sigma_{x'}$$

Théorème 107 *La valeur de l'écart-type est invariante par translation sur l'origine de la variable statistique.*

Example 108 *On considère toujours l'exemple* 111. *Prenons pour nouvelle origine de la variable statistique* $x_0 = 3$ *et considérons le changement de variable* $x'_i = x_i - 3$. *On obtient alors le tableau statistique suivant :*

x_i	n_i	x'_i	$n_i x'_i$	$n_i x_i'^2$
1	25	-2	-50	100
2	55	-1	-55	55
3	75	0	0	0
4	50	1	50	50
5	35	2	70	140
6	10	3	30	90
$Totaux$	$\sum_{i=1}^{k} n_i = 250$		$\sum_{i=1}^{k} n_i x'_i = 45$	$\sum_{i=1}^{k} n_i \overline{x}^2 = 435$

Alors, $\overline{x'} = 0,18$ *d'où* $\overline{x} = 0,18 + 3 = 3,18$
Et

$$Var\,(x") = \sigma_{x"}^2 = \frac{435}{250} - (0,18)^2 = 1,70$$

D'où

$$\sigma_x = \sigma_{x'} = 1,3$$

4.2.2 Cas d'une variable continue

Les observations à l'intérieur d'un même intervalle sont représentées par la valeur médium X_i (centre de la classe $n^\circ i$) définie telle que :

$$X_i = \frac{e_i + e_{i-1}}{2}$$

où e_i et e_{i-1} désignent respectivement les extrémités supérieure et inférieure de la classe $n^\circ i$.
Ainsi la variable X_i joue le même rôle qu'une variable discrète affectée d'un effectif égale au nombre des valeurs de la variable statistique appartenant à l'intervalle $[e_{i-1}, e_i[$. On ramène ainsi le calcul de la moyenne et de l'écart-type dans le cas continu à celui utilisé dans le cas discret. La moyenne est alors $\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_i X_i$ et la variance est $\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(X_i - \overline{X}\right)^2$. Dans la suite, on n'étudiera que la méthode de changement de variable.

Calcul avec changement de variable

Il est souvent intéressant de considérer le changement de variable suivant :

$$X_i' = \frac{X_i - X_{M_o}}{u} \tag{4.4}$$

où u est le $PGCD$ des amplitudes des classes et X_{M_o} est généralement le centre de la classe modale.
D'après les résultats précédents, il existe entre les moyennes $\overline{X}$ et $\overline{X'}$ la même relation que celle entre X_i et X_i', i.e.

$$\overline{X'} = \frac{\overline{X} - X_{M_o}}{u} \tag{4.5}$$

En retranchant les relations (4.4) et (4.5) membre à membre, il vient :

$$X_i' - \overline{X'} = \frac{X_i - X_{M_o}}{u} - \frac{\overline{X} - X_{M_o}}{u} = \frac{X_i - \overline{X}}{u}$$

Par suite, en remplaçant $(X_i - \overline{X})$ par $u\left(X_i' - \overline{X'}\right)$ dans la définition de σ_X^2, on obtient :

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \left(X_i - \overline{X}\right)^2 = \frac{u^2}{n}\sum_{i=1}^{k} n_i \left(X_i' - \overline{X'}\right)^2 = u^2\sigma_{X'}^2$$

D'où

$$\sigma_X^2 = u^2\sigma_{X'}^2$$

Théorème 109 *Si X_i et X'_i sont en relation fonctionnelle telle que $X_i = uX'_i + X_{M_o}$, alors σ_X et $\sigma_{X'}$ sont tels que :*

$$\sigma_X = u\sigma_{X'}$$

Example 110 *Reprenons l'exemple* 85 *de la distribution des ouvriers d'une entreprise suivant leur salaire mensuel.*

Classe de Salaire	n_i	X_i	X'_i	$n_iX'_i$	$n_iX'^2_i$
$12000 \leq x < 14000$	26	13000	-2	-52	104
$14000 \leq x < 16000$	33	15000	-1	-33	33
$16000 \leq x < 20000$	64	18000	0	0	0
$20000 \leq x < 24000$	7	22000	1	7	7
$24000 \leq x < 30000$	10	27000	2	20	40
Total	140			$\sum_{i=1}^{k} n_i x'_i = -58$	184

Le centre de la classe modale est $X_{M_o} = 18000$. Alors le changement de variable effectué est $X'_i = \frac{X_i - 18000}{1000}$. Ainsi, $\overline{X'} = -0,414$ d'où $\overline{X} = 17586$. Par ailleurs, $Var\,(X') = \sigma^2_{X'} = 1,143$, d'où

$$\sigma_X = 10^3\sigma_{X'} = 1069$$

Correction de Sheppard

Lorsque les observations sont regroupées par classe, l'hypothèse de la concentration au centre de la classe des observations se situant dans le même intervalle (i.e. quand le centre de la classe est substitué aux différentes valeurs observées) implique une approximation dans le calcul de l'écart-type. Pour le calcul de la moyenne arithmétique en général les erreurs se compensent, alors que pour celui de l'écart-type elles se rajoutent . Si la distribution statistique est unimodale et à support compact (i.e. si la courbe de la distribution est tangente à l'axe des abscisses aux extrémités), alors on peut corriger la valeur de l'ecart-type calulée à partir des observations regroupées en classe, avec la formule proposée par W.F. Sheppard :

$$\sigma_{corrigé} = \sqrt{\sigma^2_X - \frac{u^2}{12}}$$

où u représente le $PGCD$ de l'amplitude des classes.

Propriétés de l'écart-type

L'écart-type satisfait assez bien à l'ensemble des conditions de Yule. Il tient compte de toutes les observations. Il se prête facilement au calcul algébrique. C'est la caractéristique de dispersion la moins sensible aux fluctuations d'échantillonnage.

Propriétés de la variance

La variance d'une population P composée de deux sous-populations P_1 et P_2 de moyennes respectives $\overline{x_1}$ et $\overline{x_2}$, et de variances respectives σ_1^2 et σ_2^2, peut s'exprimer simplement en fonction de $\overline{x_1}$, $\overline{x_2}$, σ_1^2 et σ_2^2. Supposons que l'effectif de la population P soit n et que les effectifs des sous-populations sont respectivement n_1 et n_2 tels que :

$$n = n_1 + n_2 \quad \text{et} \quad n_i = n_{1i} + n_{2i} \quad i = 1, 2, ..., k$$

Par définition, la variance de la sous-population P_1 a pour expression :

$$\sigma_1^2 = \frac{1}{n_1} \sum_{i=1}^{k} n_{1i} \left(x_{1i} - \overline{x_1}\right)^2$$

Que l'on peut mettre sous la forme :

$$\sigma_1^2 = \frac{1}{n_1} \sum_{i=1}^{k} n_{1i} \left(x_{1i} - \overline{x}\right)^2 - \left(\overline{x_1} - \overline{x}\right)^2 \tag{4.6}$$

De même, la variance de la sous-population P_2 a pour expression :

$$\sigma_2^2 = \frac{1}{n_2} \sum_{i=1}^{k} n_{2i} \left(x_{2i} - \overline{x_2}\right)^2$$

que l'on peut mettre sous la forme :

$$\sigma_2^2 = \frac{1}{n_2} \sum_{i=1}^{k} n_{2i} \left(x_{2i} - \overline{x}\right)^2 - \left(\overline{x_2} - \overline{x}\right)^2 \tag{4.7}$$

Par ailleurs, la variance de la population P est définie telle que :

$$\begin{aligned}\sigma^2 &= \frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{k} \left(n_{1i} + n_{2i}\right)\left(x_i - \overline{x}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{k} n_{1i} \left(x_i - \overline{x}\right)^2 + \frac{1}{n}\sum_{i=1}^{k} n_{2i} \left(x_i - \overline{x}\right)^2 \\ &= \frac{n_1}{n}\left(\frac{1}{n_1}\sum_{i=1}^{k} n_{1i} \left(x_{1i} - \overline{x}\right)^2\right) + \frac{n_2}{n}\left(\frac{1}{n_2}\sum_{i=1}^{k} n_{2i} \left(x_{2i} - \overline{x}\right)^2\right)\end{aligned}$$

Et d'après les relations (4.6) et (4.7), la vriance σ^2 peut être écrite telle que :

$$\sigma^2 = \frac{n_1}{n}\left\{\sigma_1^2 + \left(\overline{x_1} - \overline{x}\right)^2\right\} + \frac{n_2}{n}\left\{\sigma_2^2 + \left(\overline{x_2} - \overline{x}\right)^2\right\}$$

Finalement

$$\sigma^2 = \frac{1}{n}\left(n_1\sigma_1^2 + n_2\sigma_2^2\right) + \frac{1}{n}\left(n_1\left(\overline{x_1} - \overline{x}\right)^2 + n_2\left(\overline{x_2} - \overline{x}\right)^2\right) \qquad (4.8)$$

La relation (4.8) se généralise aisément à une population constituée d'un nombre fini h quelconque de sous-populations. En effet, en désignant par n_i, $i = 1, 2, ..., h$, les effectifs des sous-populations P_i tels que :

$$n = \sum_{i=1}^{h} n_i \quad \text{et} \quad n_i = \sum_{j=1}^{k} n_{ij}$$

La variance totale de la population P est alors :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{h} n_i\sigma_i^2 + \frac{1}{n}\sum_{i=1}^{h} n_i \left(\overline{x_i} - \overline{x}\right)^2$$

Ainsi, on vient de démontrer le théorème suivant :

Théorème 111 *La variance totale dans une population constituée d'un nombre fini quelconque de sous-populations, est une somme de la moyenne des variances dans les sous-populations et de la variance entre les sous-populations, i.e*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{h} n_i\sigma_i^2 + \frac{1}{n}\sum_{i=1}^{h} n_i \left(\overline{x_i} - \overline{x}\right)^2$$

où $\overline{x_i}$, $i = 1, 2, ..., h$ est la moyenne arithmétique dans la sous-population $n°i$, $\overline{x}$ est la moyenne arithmétique de la population totale et σ_i^2, $i = 1, 2, ..., h$ est la variance de la sous-population $n°i$

Definition 112 *Le terme $\frac{1}{n}\sum_{i=1}^{h} n_i\sigma_i^2$ est appelé variance **intra-groupe** et est notée σ_{intra}^2. Le terme $\frac{1}{n}\sum_{i=1}^{h} n_i\left(\overline{x_i} - \overline{x}\right)^2$ est appelé variance **inter-groupe** et est noté σ_{inter}^2.*

Remarque 113 *La variance totale σ^2 peut être exprimée alors telle que :*

$$\sigma^2 = \sigma_{intra}^2 + \sigma_{inter}^2$$

D'une manière générale, la variance d'une population composée de plusieurs sous-populations résulte de deux facteurs : la variabilité interne à chaque sous-population et la variabilité entre les différentes sous-populations.

4.3 Autres caractéristiques d'une distribution statistique

4.3.1 Coefficient de variation

En général, l'écart-type et la moyenne s'expriment dans la même unité de mesure que la variable statistique. Or, on peut avoir à comparer des dispersions de distributions qui ne sont pas exprimées dans la même unité de mesure ou bien qui diffèrent par leurs moyennes. Alors, on introduit une caractéristique de dispersion relative.

Definition 114 *On appelle coefficient de variation, et on note CV, le rapport de l'écart-type à la moyenne arithmétique, i.e.*

$$CV = \frac{\sigma}{\overline{x}}$$

Remarque 115 *Le coefficient de variation est un nombre sans dimension. Il est par conséquent indépendant des unités de mesure choisies.*

Example 116 *Les distributions des salaires dans deux entreprises semblables E_1 et E_2 ont les caractéristiques suivantes :*

$$\overline{x}_1 = 19600\ DA \quad et \quad \sigma_1 = 2500\ DA$$

$$\overline{x}_2 = 18000\ DA \quad et \quad \sigma_2 = 1400\ DA$$

Les coefficients de variation sont donc :

$$CV_1 = \frac{2500}{19600} = 0,1275 \quad et \quad CV_2 = \frac{1400}{18000} = 0,0778$$

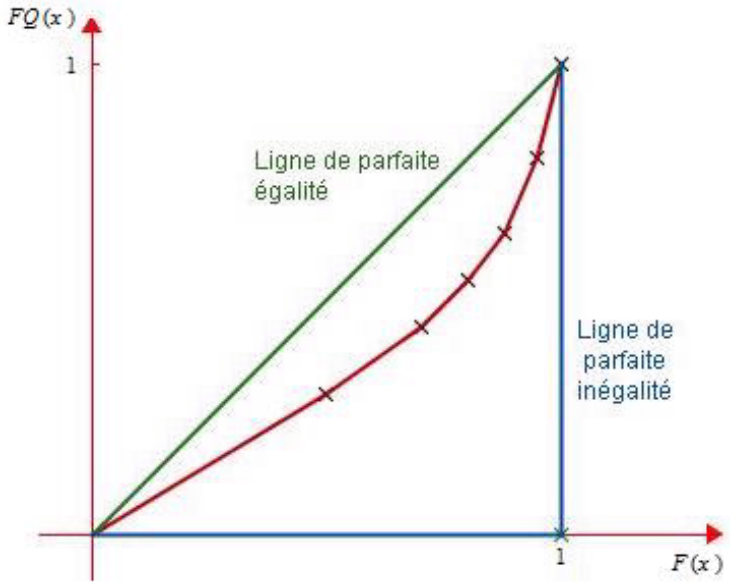
Supposons que l'on désire comparer ces distributions avec celle observée à propos d'une entreprise américaine comparable, avec les caractéristiques :

$$\overline{x} = 2800\ \$\ ;\ \sigma = 450\ \$$$

$$CV = \frac{450}{2800} = 0,0143$$

4.3.2 Courbe de concentration

La courbe de concentration exige comme pour la médiale, la connaissance pour chaque classe du nombre d'observations et de la somme des valeurs correspondantes.
Soit F_i la fréquence cumulée des observations et FQ_i le pourcentage cumulé de la somme des valeurs. Alors, la courbe de concentration est obtenue en traçant le graphe de FQ_i (ordonnée) en fonction de F_i (abscisse). On obtient alors la courbe suivante :



Courbe de concentration ou courbe de Lorenz

Remarque 117 *La courbe de concentration ou courbe de Lorenz est notamment utilisée en économie pour mesurer les inégalités de possession de richesse (on supposera donc que x représente un certain bien possédé par les individus de la population). Elle est fabriquée de la façon suivante. Soit x_i une valeur prise par x. On note $F(x)$ la proportion de la population pour laquelle $x < x_i$ (F est donc la courbe cumulative (fonction de répartition) de x). On note $FQ(x_i)$ la proportion du bien possédé par ces individus par rapport au bien total. Alors la courbe de Lorenz est la courbe joignant tous les points $(F(x_i), FQ(x_i))$. La courbe de Lorenz joint donc toujours le point $(0, 0)$ au point $(1, 1)$. Elle est située sous le segment joignant ces deux points.*

Definition 118 *La diagonale du carré circonscrit à la courbe de Lorenz s'appelle droite* ***d'équi-répartition****.*

Remarque 119 *La diagonale principale du graphique (droite d'équi–répartition) représente une distribution parfaitement égalitaire. Plus la courbe de concentration s'écarte de la droite d'équi–répartition, plus la distribution est inégalitaire. D'autre part, plus la dispersion est faible plus la courbe de concentration s'applatit sur la diagonale.*

4.3.3 Indice de concentration ou indice de Gini

C'est une mesure de dispersion proposée par le statisticien italien Corrado Gini. L'indice de concentration ou indice de Gini, noté G, est une mesure de dispersion relative d'une série statistique. Cette caractéristique ne s'applique qu'aux variables statistiques continues et à valeurs positives. Son calcul exige la connaissance pour chaque classe du nombre d'observations et de la somme des valeurs correspondantes. L'indice de concentration est défini en général à partir de la courbe de Lorenz d'une variable statistique positive x.

Definition 120 *L'indice de Gini d'une distribution statistique est le double de l'aire de la surface délimitée par la courbe de Lorenz et la première diagonale du carré unité.*

Remarque 121 *Du fait que $F(x)$ et $FQ(x)$ varient dans l'intervalle $[0, 1]$ et qu'ils sont nuls ou égaux à 1 en même temps, la courbe de concentration s'inscrit dans un carré unitaire. Elle se situe en dessous de la diagonale du carré car, en général, $F(x)$ est supérieur à $FQ(x)$. L'indice de Gini est*

toujours compris entre 0 *et* 1.
L'indice de Gini est très utilisé en économie comme mesure des inégalités dans une population. Supposons par exemple que la variable x correspond aux revenus dans une population. Si l'indice de Gini est proche de 0*, ceci signifie que les différences relatives sont en moyenne faible par rapport à la moyenne des revenus, i.e. les inégalités dans la population sont faibles. Si l'indice de Gini est proche de* 1*, au contraire il y a de fortes différences relatives en moyenne, i.e. les inégalités sont fortes.*

4.3.4 Calcul pratique de l'indice de Gini

L'aire comprise entre la courbe de Lorenz et la diagonale du carré est calculée par approximation. Pour rappel, l'aire d'un trapèze est telle que :

$$\frac{\text{hauteur} \times (\text{petite base} + \text{grande base})}{2}$$

Pour obtenir l'aire entre la courbe de Lorenz et la diagonale du carré, il faut soustraire l'aire des trapèzes en dessous de la courbe de concentration à $0,5$. Alors en posant $F_0 = FQ_0 = 0$ et $F_k = FQ_k = 1$ où k est le nombre de classes, l'idice de Gini est donné par la formule suivante :

$$G = 2\left(0,5 - \sum_{i=1}^{k} \frac{(F_i - F_{i-1})(FQ_{i-1} + FQ_i)}{2}\right)$$

Que l'on peut écrire aussi sous la forme :

$$G = 1 - \sum_{i=1}^{k} (F_i - F_{i-1})(FQ_{i-1} + FQ_i)$$

Example 122 *Reprenons l'exemple de la répartition des employés d'une entreprise selon leur salaire mensuel net.*

$Salaires\ (euros)$	n_i	F_i	FQ_i	$F_i - F_{i-1}$	$FQ_{i-1} + FQ_i$	$(F_i - F_{i-1})(FQ_{i-1} + FQ_i)$
$[800; 900[$	25	$0,212$	$0,164$	$0,212$	$0,164$	$0,034768$
$[900; 1000[$	30	$0,466$	$0,385$	$0,254$	$0,549$	$0,139446$
$[1000; 1100[$	28	$0,703$	$0,613$	$0,237$	$0,998$	$0,236526$
$[1100; 1500[$	25	$0,915$	$0,865$	$0,212$	$1,478$	$0,313336$
$[1500; 2000[$	10	1	1	$0,085$	$1,865$	$0,158525$
$\sum$	118					$0,882601$

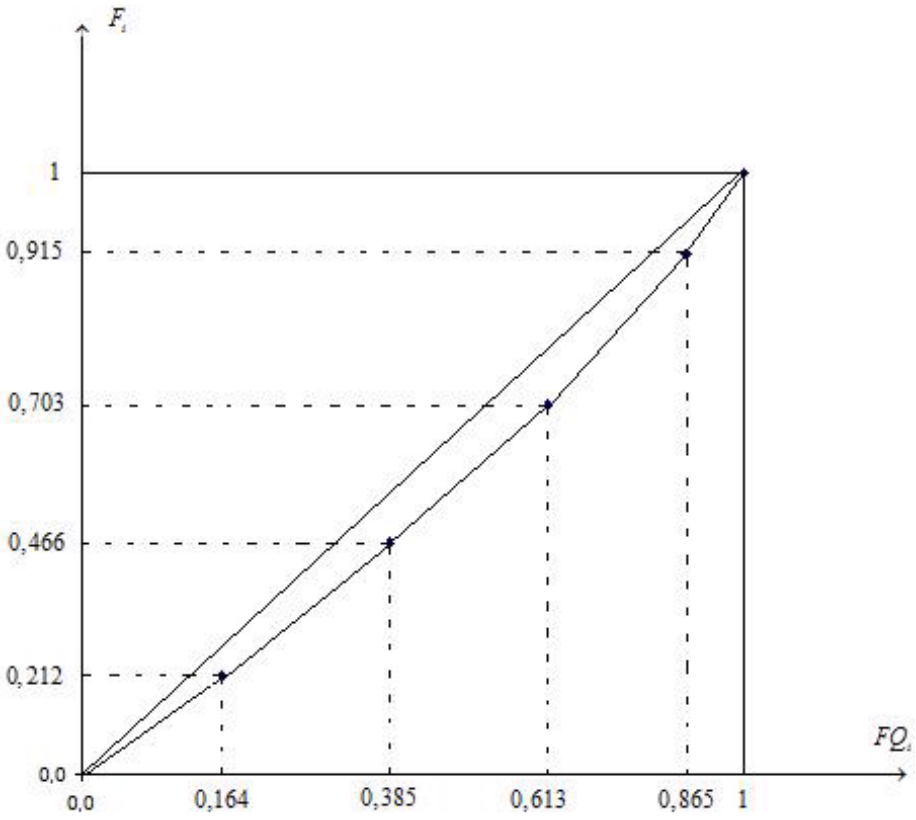


FIG. 4.1 – Calcul de l'indice de Gini

Alors, l'indice de Gini est tel que :

$$G = 1 - 0,8826 \approx 0,117$$

4.4 Les caractéristiques de forme

En plus des caractéristiques de tendance centrale et de dispersion, il serait instructif de définir des indices pour résumer l'information véhiculée par les données, sur l'allure et la forme de la distribution d'une série statistique. Pour une distribution statistique symétrique la moyenne, le mode et la médiane coïncident. Il est donc naturel de considérer la déviation de la moyenne par rapport au mode ou bien par rapport à la médiane, comme mesure d'asymétrie de la distribution statistique. K. Pearson a proposé comme mesure de l'asymétrie une quantité fonction du mode. Mais cette quantité est sujette à l'inconvénient à déterminer le mode. Cependant, pour une large classe de distributions de fréquences, la mesure d'asymétrie peut être déterminée exactement à l'aide des quatre premiers moments de la distribution.

Definition 123 *Soit $x_1, x_2, ..., x_n$ une distribution statistique d'une variable x. On appelle moment centré d'ordre r de la variable statistique x, noté μ_r, la quantité définie telle que :*

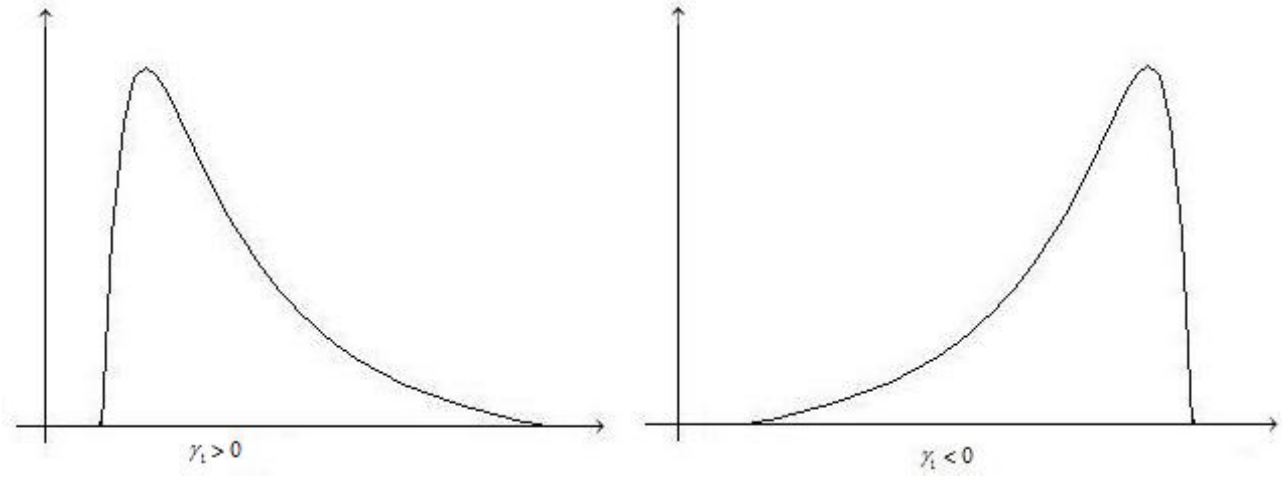
$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^r$$

4.4.1 Coefficient d'asymétrie (skewness)

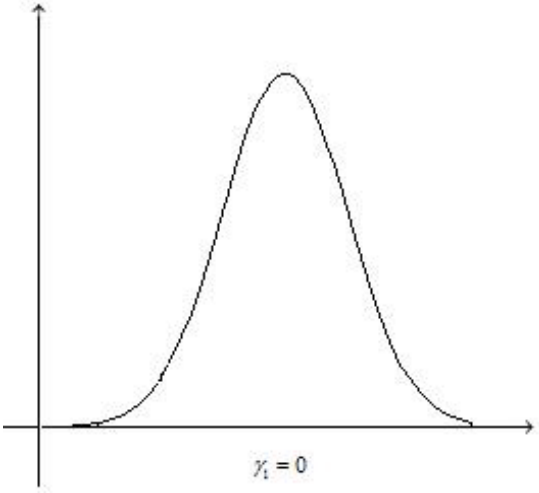
Une distribution statistique symétrique a tous ses moments centrés impairs nuls et a fortiori la moyenne μ_1.

Definition 124 *Soit μ_3 et μ_2 les moments centrés d'ordre respectifs 3 et 2 de la distribution statistiques. On appelle caractéristique d'asymétrie le coefficient γ_1 défini tel que :*

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}$$



Distributions asymétriques



Distribution symétrique

Remarque 125 *Le coefficient γ_1 est sans dimension, invariant par changement d'origine et d'échelle. Il est nul pour les distributions symétriques.*
On utilise également comme indice d'asymétrie le rapport :

$$d = \frac{Q_1 + Q_3 - 2M_e}{2M_e}$$

où Q_1 et Q_3 sont les quartiles, et M_e la médiane.
Pour les distributions unimodales γ_1 et d sont de même signe et ils s'annulent pour les distributions symétriques.

4.4.2 Coefficient d'aplatissement (Kurtosis)

Definition 126 *On appelle caractéristique d'aplatissement le coefficient* γ_2 *défini tel que :*

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\mu_4}{\sigma^4} - 3$$

où μ_4 *et* μ_2 *sont les moments centrés d'ordre respectifs* 4 *et* 2 *de* x*.*

Remarque 127 *Le coefficient* γ_2 *est sans dimension. Il est invariant par changement d'origine et d'échelle. La constante* 3 *est choisie de telle sorte que le coefficient* γ_2 *soit nul pour les distributions normales.*
Le coefficient γ_2 *est positif si la distribution est moins aplatie que la distribution normale et il est négatif dans le cas contraire.*
Les courbes pour lesquelles $\gamma_2 = 0$ *sont dites* ***mésokurtiques****, celles pour lesquelles* $\gamma_2 > 0$ *sont dites* ***leptokurtiques*** *et celles pour lesquelles* $\gamma_2 < 0$ *sont dites* ***platicurtiques****.*
Du fait de l'inégalité $\mu_4 \geq \mu_2^2$*, le coefficient d'aplatissement est toujours supérieur à* -2*.*

Chapitre 5

Distributions statistiques à deux dimensions

Pour l'étude de certains phénomènes complexes, il s'avère insuffisant de prendre en compte un seul caractère. Alors il en faut considérer deux caractères ou plus. L'analyse et la représentation des tableaux statistiques obtenus deviennent évidemment plus complexes. La représentation graphique, par exemple, n'est possible que dans un espace à trois dimensions au plus. En définissant les distributions marginales et conditionnelles, on peut ramener la représentation d'une distribution à plusieurs dimensions à quelques représentations unidimensionnelles. Dans la suite, on ne considérera que les séries statistiques à deux dimensions.

5.1 Présentation générale d'un tableau à double entrée

Considérons une population de n individus. Chacun de ces derniers est identifié par deux caractères A et B. Le caractère A comporte k modalités $A_1, A_2, ..., A_k$ et le caractère B en comporte m , $B_1, B_2, ..., B_m$. L'opération préliminaire consiste à classer les n individus dans $k \times m$ cases d'un tableau où figurent en ligne les modalités de A et en colonne les modalités de B. Dans chaque case $(i, j), i = 1, 2, ..., k$ et $j = 1, 2, ..., m$, on inscrira le nombre n_{ij} des éléments du sous-ensemble de la population contenant les individus présentant simultanément la modalité A_i du caractère A et la modalité B_j du caractère B.

Pour alléger les notations on indiquera par un 'point (.)' la sommation effectuée suivant l'indice 'i' ou l'indice 'j', i.e.

$$\sum_{j=1}^{m} n_{ij} = n_{i.} \quad ; \quad i = 1, 2, ..., k$$

$$\sum_{i=1}^{k} n_{ij} = n_{.j} \quad ; \quad j = 1, 2, ..., m$$

$$\sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{m} n_{.j} = \sum_{i,j} n_{ij} = n_{..} = n$$

Nous donnons ci-après la forme générale d'un tableau statistique à double entrée, appelé aussi tableau de contingence :

A/B	B_1	B_2	...	B_j	...	B_m	$Total$
A_1	n_{11}	n_{12}		n_{1j}		n_{1m}	$n_{1.}$
A_2	n_{21}	n_{22}		n_{2j}		n_{2m}	$n_{2.}$
...							
A_i	n_{i1}	n_{i2}		n_{ij}		n_{im}	$n_{i.}$
...							
A_k	n_{k1}	n_{k2}		n_{kj}		n_{km}	$n_{k.}$
$Total$	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.m}$	$n_{..} = n$

Tableau statistique à double entrée

Definition 128 *On appelle fréquence de l'évènement (A_i, B_j) la proportion des observations qui présentent simultanément les modalités A_i et B_j. Elle est notée f_{ij} et est définie telle que :*

$$f_{ij} = \frac{n_{ij}}{n}$$

Remarque 129 *Si on adopte les mêmes conventions d'écriture que pour les effectifs, en indiquant par un 'point' les sommations effectuées par rapport à l'indice 'i' ou par rapport à l'indice 'j', alors $f_{i.}$ est la somme des fréquences de la ligne n°i.*

Proposition 130 *Il est évident que d'après la définition de la fréquence f_i, on a :*

$$f_{i.} = \frac{n_{i.}}{n} \quad i = 1, 2, ..., k \qquad et \qquad f_{.j} = \frac{n_{.j}}{n} \quad j = 1, 2, ..., m$$

Démonstration :

$$f_{i.} = \sum_{j=1}^{m} f_{ij} = \sum_{j=1}^{m} \frac{n_{ij}}{n} = \frac{n_{i.}}{n}$$

et

$$f_{.j} = \sum_{i=1}^{k} f_{ij} = \sum_{i=1}^{k} \frac{n_{ij}}{n} = \frac{n_{.j}}{n}$$

■

Remarque 131 *Comme pour les distributions à un caractère la somme des fréquences est égale à l'unité. En effet ,*

$$\sum_{i=1}^{k} \sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{k} f_{i.} = \sum_{j=1}^{m} f_{.j} = 1$$

5.2 Distributions marginales

Definition 132 *La sommation suivant les lignes ou les colonnes des effectifs ou des fréquences, définit la distribution marginale du caractère A ou celle de B respectivement.*

Remarque 133 *La distribution marginale est la distribution statistique de l'un des caractères indépendamment de l'autre. Elle est lue sur l'une ou l'autre des marges du tableau, d'où son nom. Par exemple, la distribution marginale associée au caractère A est :*

$$n_{1.}, n_{2.}, ..., n_{k.} \quad ou\ bien \quad f_{1.}, f_{2.}, ..., f_{k.}$$

et la distribution marginale associée au caractère B est :

$$n_{.1}, n_{.2}, ..., n_{.m} \quad ou\ bien \quad f_{.1}, f_{.2}, ..., f_{.m}$$

Example 134 *L'étude d'une population de* 50 *individus suivant le poids (caractère B) et la taille (caractère A), a donné les résultats suivants :*

A/B	60	70	80	90	$Marge$
160	2	5	4	1	12
170	2	8	9	4	23
180	0	4	6	5	15
$Marge$	4	17	19	10	50

A/B	60	70	80	90	$Marge$
160	$0,04$	$0,10$	$0,08$	$0,02$	$0,24$
170	$0,04$	$0,16$	$0,18$	$0,08$	$0,46$
180	$0,00$	$0,08$	$0,12$	$0,10$	$0,30$
$Marge$	$0,08$	$0,34$	$0,38$	$0,20$	$1,00$

Les résultats peuvent être résumés dans un tableau statistique à double entrée en fonction des éffectifs ou des fréquences relatives.

5.3 Distributions conditionnelles

Definition 135 *Considérons la sous population des individus présentant la modalité B_j. Sur cette sous-population la distribution du caractère A est appelée distribution conditionnelle de A sachant B_j réalisé.*

Remarque 136 *Considérons les $n_{.j}$ individus présentant la modalité B_j. Parmi ceux-ci, il y a une proportion $\frac{n_{ij}}{n_{.j}}$ d'individus qui présentent en même temps la modalité A_i.*

Definition 137 *On dit que la fréquence conditionnelle de la modalité A_i liée par la modalité B_i est :*

$$f_{i/j} = f\left(A_i/B_j\right) = \frac{n_{ij}}{n_{.j}} \qquad j = 1, 2, ..., m$$

Remarque 138 *L'ensemble des fréquences conditionnelles du caractère A liées à la même modalité B_j du caractère B constitue la distribution conditionnelle de A sachant B_j réalisé. C'est une distribution à un seul caractère et il y a autant de distributions conditionnelles de A qu'il y a de modalités de B, i.e. autant que de colonnes du tableau statistique.*

Definition 139 *On définit de façon analogue la distribution conditionnelle de B liée par les modalités A_i tel que :*

$$f_{j/i} = f\left(B_j/A_i\right) = \frac{n_{ij}}{n_{i.}} \qquad i = 1, 2, ..., k$$

Remarque 140 *Les distributions conditionnelles conservent les mêmes propriétés que les distributions statistiques normales, i.e.*

$$\sum_{i=1}^{k} f_{i/j} = 1 \ , \quad j = 1, 2, ..., m$$

$$\sum_{j=1}^{m} f_{j/i} = 1 \quad , \quad i = 1, 2, ..., k$$

Example 141 *On reprend l'exemple* 55 *et on détermine les distributions conditionnelles.*

$A/\,B$	60	70	80	90
160	0,50	0,30	0,211	0,10
170	0,50	0,47	0,474	0,40
180	0,00	0,23	0,315	0,50
$Marge$	1,00	1,00	1,00	1,00

Distributions conditionnelles de A sachant $B_j, j = 1, 2, 3, 4$

$B/\,A$	160	170	180
60	0,167	0,087	0,00
70	0,416	0,350	0,267
80	0,334	0,391	0,400
90	0,083	0,172	0,333
$Marge$	1,00	1,00	1,00

Distributions conditionnelles de B sachant $A_i, i = 1, 2, 3, 4$

5.3.1 Propriétés des fréquences marginales et conditionnelles

Il est facile d'établir que :

$$f_{ij} = f\left(B_j/\,A_i\right) f_{i.} = f\left(A_i/\,B_j\right) f_{.j} = f_{j/i} f_{i.} = f_{i/j} f_{.j}$$

En effet,

$$\frac{n_{ij}}{n} = \frac{n_{ij}}{n_{i.}}\frac{n_{i.}}{n} = \frac{n_{ij}}{n_{.j}}\frac{n_{.j}}{n}$$

On retrouvera cette formule plus tard en calcul de probabilité sous le nom d'axiome des probabilités conditionnelles.

Moyennes conditionnelles

Il nous est possible de définir plus tard la notion de moyenne conditionnelle. Les moyennes conditionnelles sont les moyennes des distributions conditionnelles. On peut parler, par exemple, de la moyenne du caractère B chez les individus présentant la modalité A_i, $i = 1, 2, ..., k$ du caractère A ou bien de la moyenne du caractère A présentant la modalité B_j, $j = 1, 2, ..., m$ du caractère B.

5.4 Représentations graphiques des distributions à deux caractères

Le mode de représentation graphique d'une distribution à deux caractères n'est strictement possible que dans un espace à trois dimensions. Chacun des caractères est porté sur une dimension et la troisième est affectée aux effectifs ou aux fréquences.

5.4.1 Cas des caractères qualitatifs

Il n'est pas toujours possible de représenter les deux caractères de façon absolument symétrique. Cependant, on peut représenter la famille des distributions conditionnelles $A/\,B_j$, $j = 1, 2, ..., m$ (ou bien $B/\,A_i$, $i = 1, 2, ..., k$), de telle sorte que n_{ij} soit représenté par un rectangle de base $n_{.j}$ (ou $n_{i.}$) et que la hauteur soit proportionnelle à la fréquence conditionnelle $f_{i/j}$ (ou $f_{j/i}$).

5.4.2 Cas des caractères quantitatifs

Dans ce cas aussi on peut utiliser le mode de représentation énoncé plus haut. De plus, soient x et y les deux variables statistiques quantitatives discrètes. Soit n_{ij} l'effectif correspondant à la modalité (x_i, y_j), on peut repré-

senter cet effectif par un cercle centré au point (x_i, y_j) et de surface proportionnelle à n_{ij}.

Definition 142 *La représentation graphique d'une distribution à deux variables continues regroupées par classes est appelée stéréogramme. C'est un solide constitué par un ensemble de parallélépipèdes rectangles dont la base est formée par les couples d'intervalles de classe et dont les volumes sont proportionnels aux fréquences f_{ij} ou aux effectifs n_{ij}.*

Remarque 143 *Le parallélépipède relatif à la classe $n°i$ d'amplitude a_i de x, et à la classe $n°j$ d'amplitude b_i de y, a pour hauteur :*

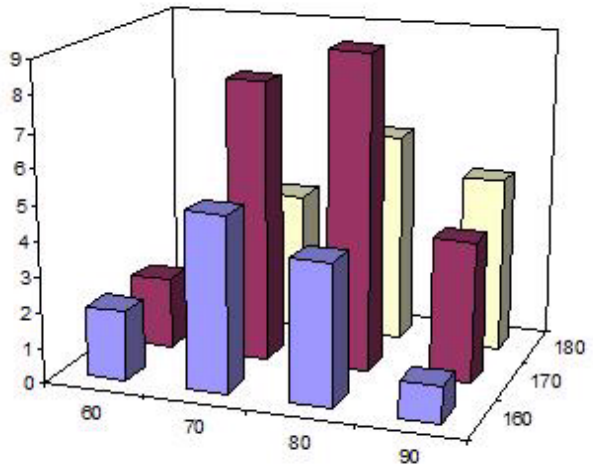
$$h_{ij} = \frac{f_{ij}}{a_i b_j}$$

de telle sorte que le volume de ce parallélépipède soit :

$$V_{ij} = a_i b_j \left(\frac{f_{ij}}{a_i b_j} \right) = f_{ij}$$

Ainsi le stéréogramme apparaît comme la généralisation de l'histogramme.

Example 144 *Reprenons l'exemple* 55 *et représentons par un stéréogramme la distribution du poids (caractère B) et de la taille (caractère A) des individus de la population.*



Représentation par stéréogramme

5.5 Covariance entre deux variables statistiques

Dans le cas des variables statistiques à deux dimensions, il est intéressant de pouvoir quantifier la variabilité de la population due à l'effet conjugué des variables considérées simultanément. Pour cela on introduit la notion de covariance.

5.5.1 Covariance

En général, la distribution des observations d'une population suivant deux caractères (x, y) sont disposées dans un tableau de contingence, alors la covariance est définie telle que :

Definition 145 *Soit (x, y) un couple de variables statistiques pouvant prendre les valeurs (x_i, y_j), $i = 1, 2, ..., k$ et $j = 1, 2, ..., m$ avec les effectifs respectifs (n_{ij}), $i = 1, 2, ..., k$ et $j = 1, 2, ..., m$. On appelle covariance des variables statistiques x et y , notée $Cov(x, y)$, la quantité définie telle que :*

$$Cov\,(x, y) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij}\,(x_i - \overline{x})\,(y_j - \overline{y})$$

Remarque 146 *Pour le calcul pratique, on utilisera souvent la formule développée de la covariance définie telle que :*

$$Cov\,(x, y) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} x_i y_j - \overline{xy}$$

Dans certaines situations il arrive que que les observations d'une population suivant deux caractères (x, y) soient appariées, i.e. les observations sont disponibles sous forme d'une suite (x_i, y_i), $i = 1, 2, ..., n$, alors dans cette situation la covariance est définie telle que :

Definition 147 *Soit (x_i, y_i), $i = 1, 2, ..., n$ une série d'observation d'un couple de variables statistiques (x, y). On appelle covariance des variables statistiques x et y , notée $Cov(x, y)$, la quantité définie telle que :*

$$Cov\,(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})$$

Remarque 148 *Pour le calcul pratique, on utilisera souvent la formule développée de la covariance définie telle que :*

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

5.5.2 Coefficient de corrélation

Il arrive souvent de vouloir comparer la variation d'une variable statistique par rapport à une autre définie sur les mêmes individus d'une quelconque population. Mais ces variables ne s'expriment pas souvent dans la même unité. Pour cela on définit le **coefficient de corrélation** qui est un coefficient normalisé sans dimension.

Definition 149 *On appelle coefficient de corrélation de deux variables statistiques x et y, et on le note $Corr(x,y)$ ou ρ , la quantité définie telle que :*

$$\rho = Corr(x,y) = \frac{Cov\,(x,y)}{\sigma_x \sigma_y}$$

Remarque 150 *Le coefficient de corrélation ρ est, à une constante près, le cosinus de l'angle entre les vecteurs $\overrightarrow{x}$ et $\overrightarrow{y}$.*

Propriété

Quelque soit le couple de variables statistiques (x,y) leur coefficient de corrélation $\rho = Corr\,(x,y)$ vérifie l'inégalité suivante :

$$-1 \leq \rho = Corr\,(x,y) \leq +1 \tag{5.1}$$

Les égalités ont lieu si et seulement si il existe deux constantes $a \neq 0$ et b telles que $y = ax + b$ ou bien $x = ay + b$.

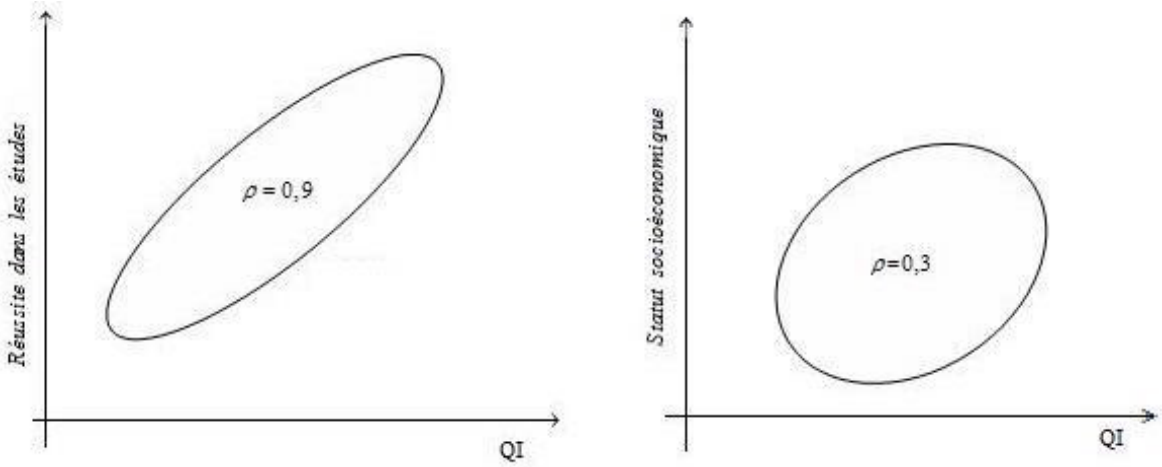
Example 151 *Reprenons l'exemple* 55 *et calculons la covariance et le coefficient de corrélation entre les caractères Aet B que l'on noteras x et y respectivement.*
Posons $(A_i, B_j) = (x_i, y_j)$, $i = 1,2,3$ et $j = 1,2,3,4$. Alors : $\overline{x} = 170,6$; $\overline{y} = 77$; $\sigma_x = 7,32$ et $\sigma_y = 8,77$ Par ailleurs

$$\frac{1}{n}\sum_{i=1}^{3}\sum_{j=1}^{4} n_{ij} x_i y_j = 13156$$

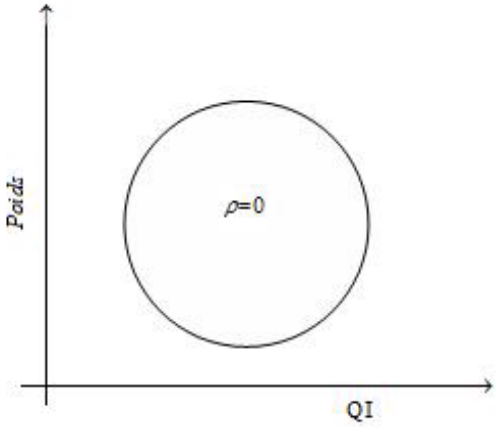
D'où

$$Cov\,(x,y) = 19,8 \quad et \quad \rho = Corr\,(x,y) = 0,3$$

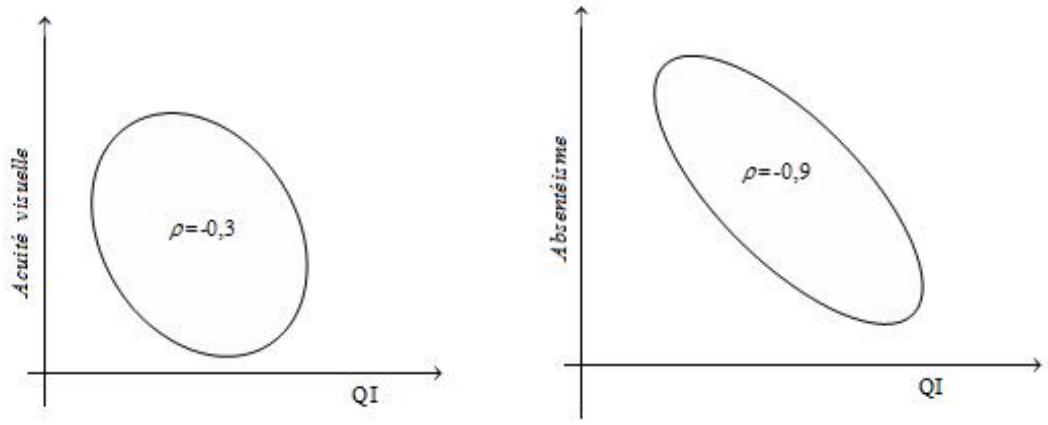
5.5.3 Différents genres de corrélation



Corrélations positives



Corrélation nulle



Corrélations négatives

5.6 Ajustement linéaire ou droite des moindres carrés

Soient x et y deux variables statistiques définies sur la même population. Les observations du couple (x, y) peuvent être présentées sous forme d'une série brute (x_i, y_i), $i = 1, 2, ..., n$ ou bien sous forme d'un tableau de contingence. On sait (5.1) que si le coefficient de corrélation entre x et y est voisin de $+1$ ou -1, il existe deux nombres réels $a \neq 0$ et b tels que $y = ax + b$ ou bien $x = ay + b$.

Definition 152 *Soient x et y deux variables statistiques définies sur la même population. L'équation $y = ax + b$ (rep. $x = a'y + b'$) est appelée droite de régression ou ajustement linéaire de y en x (resp. de x en y).*

Sachant que les constantes a et b existent, comment peut-on les déterminer ?
Les observations sur une population par rapport à deux caractères ou variables statistiques x et y nous fournissent une suite de couples (x_i, y_i), $i = 1, 2, ..., n$. En général, en raison des erreurs de mesure, les points (x_i, y_i) ne sont pas alignés, mais sont "presque" sur une même droite. Il faut alors choisir a et b de sorte que la droite soit la meilleure possible. Pour cela, il faut choisir une mesure de l'écart entre une droite $y = ax + b$ et le nuage de

points expérimentaux (x_i, y_i). On choisit en général le carré de la différence entre le point théorique et le point expérimental, c'est-à-dire $(y_i - (ax_i + b))^2$. L'écart total est donc :

$$f(a,b) = \sum_{i=1}^{n} (y_j - ax_i - b)^2 \tag{5.2}$$

Effectuer une régression linéaire, c'est trouver la droite qui minimise l'écart total, i.e. la somme des carrés des différences. On parle alors de droite des moindres carrés.

Proposition 153 *Soient x et y deux variables statistiques définies sur la même population. La fonction numérique définie sur $\mathbb{R}^2$ par l'équation (5.2) admet un minimum au point (α, β) tel que :*

$$\alpha = \frac{Cov(x,y)}{Var(x)} = Corr(x,y)\frac{\sigma_y}{\sigma_x}$$

$$\beta = \overline{y} - \alpha\overline{x}$$

Démonstration : Le minimum de la fonction $f(a,b)$ est obtenu au point (α, β) solution du système d'équations :

$$\begin{cases} \frac{\partial f(a,b)}{\partial a} = 0 \\ \frac{\partial f(a,b)}{\partial b} = 0 \end{cases}$$

Nous allons considérer deux situations.
A. Les observations sont présentées sous forme d'une série statistique brute (i.e. elle n'a pas été ordonnée dans un tableau). On dit, en général, que c'est une série d'observations couplées.
Alors, $f(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$ et donc

$$\begin{aligned} \frac{\partial f(a,b)}{\partial a} &= -2\sum_{i=1}^{n} x_i (y_i - ax_i - b) \\ &= -2\sum_{i=1}^{n} x_i y_i + 2a\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i = 0 \end{aligned}$$

Sachant que

$$\sum_{i=1}^{n} x_i y_i = n\{Cov(x,y) + \overline{x}.\overline{y}\}$$

et

$$\sum_{i=1}^{n} x_i^2 = n\left\{Var\left(x\right) + \overline{x}^2\right\} \quad \text{et} \quad \sum_{i=1}^{n} x_i = n\overline{x}$$

On déduit

$$Cov\left(x,y\right) + \overline{x}.\overline{y} - a\left\{Var\left(x\right) + \overline{x}^2\right\} - b\overline{x} = 0 \tag{5.3}$$

D'autre part

$$\begin{aligned}\frac{\partial f(a,b)}{\partial b} &= -2\sum_{i=1}^{n}\left(y_i - ax_i - b\right) \\ &= -2\sum_{i=1}^{n} y_i + 2a\sum_{i=1}^{n} x_i + 2\sum_{i=1}^{n} b = 0\end{aligned}$$

Sachant que

$$\sum_{j=1}^{n} y_j = n\overline{y} \quad \text{et} \quad \sum_{i=1}^{n} x_i = n\overline{x}$$

On déduit

$$\overline{y} - a\overline{x} - b = 0 \tag{5.4}$$

B. Les observations sont présentées dans un tableau de contingence.
Alors

$$\begin{aligned}\frac{\partial f(a,b)}{\partial a} &= -2\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}x_i\left(y_j - ax_i - b\right) \\ &= -2\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}x_iy_j + 2a\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}x_i^2 + 2b\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}x_i \\ &= -2\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}x_iy_j + 2a\sum_{i=1}^{k} n_{i.}x_i^2 + 2b\sum_{i=1}^{k} n_{i.}x_i = 0\end{aligned}$$

Sachant que

$$\sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}x_iy_j = n\left\{Cov\left(x,y\right) + \overline{x}.\overline{y}\right\}$$

et

$$\sum_{i=1}^{k} n_{i.}x_i^2 = n\left\{Var\left(x\right) + \overline{x}^2\right\} \quad \text{et} \quad \sum_{i=1}^{k} n_{i.}x_i = n\overline{x}$$

On déduit

$$Cov\left(x,y\right)+\overline{x}.\overline{y}-a\left\{Var\left(x\right)+\overline{x}^{2}\right\}-b\overline{x}=0 \tag{5.5}$$

D'autre part

$$\begin{aligned}\frac{\partial f(a,b)}{\partial b} &= -2\sum_{i=1}^{k}\sum_{j=1}^{m}n_{ij}\left(y_j-ax_i-b\right)\\ &= -2\sum_{i=1}^{k}\sum_{j=1}^{m}n_{ij}y_j+2a\sum_{i=1}^{k}\sum_{j=1}^{m}n_{ij}x_i+2b\sum_{i=1}^{k}\sum_{j=1}^{m}n_{ij}\\ &= -2\sum_{j=1}^{m}n_{.j}y_j+2a\sum_{i=1}^{k}n_{i.}x_i+2nb=0\end{aligned}$$

Sachant que

$$\sum_{j=1}^{m}n_{.j}y_j=n\overline{y}\quad\text{et}\quad\sum_{i=1}^{k}n_{i.}x_i=n\overline{x}$$

On déduit

$$\overline{y}-a\overline{x}-b=0 \tag{5.6}$$

La solution du système d'équations $\{(5.5);(5.6)\}$ est le point (α,β) tel que :

$$\alpha=\frac{Cov\left(x,y\right)}{Var\left(x\right)}$$

$$\beta=\overline{y}-\alpha\overline{x}$$

■

Remarque 154 *Il est évident que les couples d'équations* $\{(5.3);(5.4)\}$ *et* $\{(5.5);(5.6)\}$ *sont les mêmes. Donc, que l'on utilise les données brutes ou les données disposées dans un tableau de contingence, le minimum de la fonction* $f(a,b)$ *est le même.*
La quantité $\alpha=\frac{Cov(x,y)}{Var(x)}$ *peut être exprimée telle que :*

$$\alpha=\frac{Cov\left(x,y\right)}{Var\left(x\right)}=Corr\left(x,y\right)\frac{\sigma_y}{\sigma_x}$$

En effet

$$\alpha = \frac{Cov\left(x,y\right)}{Var\left(x\right)} = \frac{Cov\left(x,y\right)}{\sigma_x\sigma_x}\frac{\sigma_y}{\sigma_y} = \left\{\frac{Cov\left(x,y\right)}{\sigma_x\sigma_y}\right\}\frac{\sigma_y}{\sigma_x} = Corr\left(x,y\right)\frac{\sigma_y}{\sigma_x}$$

On peut constater que la pente de la droite de régression est proportionnelle au cefficient de corrélation et est de même signe. Le point de coordonnées $(\overline{x},\overline{y})$ *appartient toujours à la droite de régression.*

Chapitre 6

Les séries chronologiques

6.1 Généralités

Definition 155 *On appelle série chronologique ou temporelle une suite Y_t, $t = 1, 2, 3, ...$, d'observations chiffrées et ordonnées dans le temps d'un même phénomène.*

Example 156 *Nombre mensuel de vente de voitures neuves.*
Nombre annuel de naissance en Algérie.

Remarque 157 *Les dates d'observations sont généralement ordonnées de manière régulière dans le temps : on manipule des séries journalières, mensuelles, trimestrielles, annuelles. Plus généralement, pour les séries statistiques à deux dimensions, lorsque l'un des caractères est le temps, la série statistique est alors appelée série chronologique. Le deuxième caractère est quelconque.*

Représentation graphique

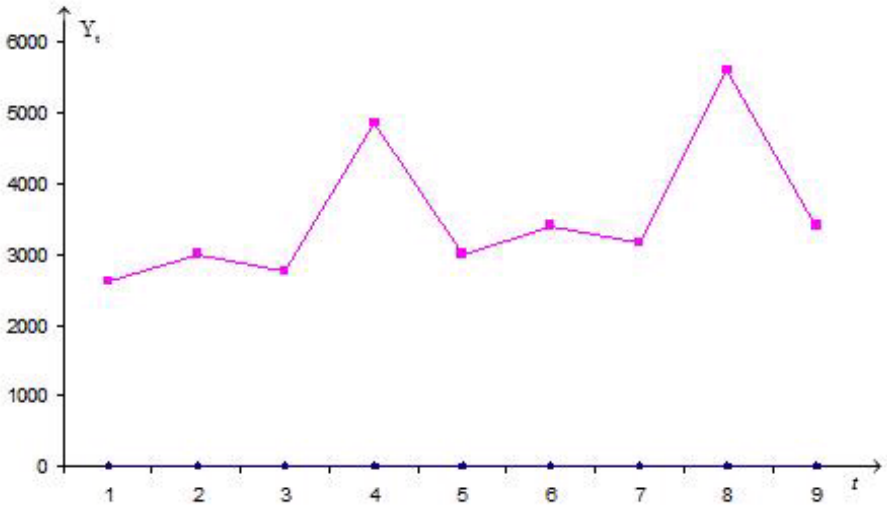
Pour la représentation graphique des séries chronologiques un certain nombre de précautions doivent être prises :

- S'il s'agit d'un stock de l'effectif de la population à une date déterminée, le point représentatif se place exactement à l'aplomb de la date de référence.
- S'il s'agit d'un flux comme la production mensuelle d'énergie électrique par exemple, ou d'une moyenne comme le prix moyen mensuel du kilogramme de pomme de terre par exemple, le point représentatif sera alors placé à la verticale du milieu de la plage.

– Le temps est toujours porté en abscisse et le deuxième caractère en ordonnée.
– On représente les points $(t; Y_t)$, que l'on relie par des segments de droites. On représente l'évolution de la grandeur considérée sur l'ensemble de la période observée.

Example 158 *Considérons la série trimestrielle du chiffre d'affaires en milliers de DA des ventes d'un magasin de* 1978 *à* 1982.

t	Y_t
1	2614
2	3010
3	2765
4	4856
5	3010
6	3397
7	3168
8	5624
9	3406



Représentation graphique de Y(t)

6.2 Analyse empirique d'une série chronologique

6.2.1 Décomposition d'une série chronologique

Le but de la décomposition d'une série chronologique est de distinguer dans l'évolution de la série, une tendance «générale», des variations saisonnières qui se répètent chaque année, et des variations accidentelles imprévisibles. L'intérêt de ceci est d'une part de mieux comprendre, de mieux décrire l'évolution de la série, et d'autre part de prévoir son évolution (à partir de la tendance et des variations saisonnières).

La tendance ou trend

Definition 159 *La tendance d'une série chronologique, notée C_t, est l'évolution à long terme de la série ou l'évolution fondamentale de la série.*

Example 160 *L'augmentation du chiffre d'affaire de* 1978 *à* 2005*.*

Les variations saisonnières ou saisonnalités

Definition 161 *Les variations saisonnières ou saisonnalités, notés S_t, sont des fluctuations périodiques à l'intérieur d'une année, et qui se reproduisent de façon plus ou moins permanente d'une année sur l'autre.*

Example 162 *Ces variations sont dues au rythme des saisons : climat, matières premières, congés, ...*

Propriétés

Les variations saisonnières se caractérisent par deux principes :

- Principe de répétition à l'identique : Les variations saisonnières sont périodiques de période p (nombre de mois) :

$$S_{t+p} = S_t$$

- Principe de conservation des aires : Par an, l'influence des variations saisonnières est nulle. Cela sera traduit à l'aide de la moyenne des S_t.

Les variations accidentelles ou résiduelles

Definition 163 *Les variations accidentelles ou résiduelles, notées ε_t, sont des fluctuations irrégulières et imprévisibles. Elles sont supposées en général de faible amplitude. C'est la composante aléatoire d'une série chronologique.*

Example 164 *Les variations résiduelles proviennent de circonstances non prévisibles : catastrophes naturelles, crise boursière, grèves ...*

6.2.2 Les modèles de composition des trois composantes

Le modèle additif

Definition 165 *Un modèle additif suppose que les trois composantes : tendance, variations saisonnières et variations accidentelles sont indépendantes les unes des autres. La série Y_t s'écrit comme la somme de ces 3 composantes :*

$$Y_t = C_t + S_t + \varepsilon_t$$

Remarque 166 *Graphiquement, l'amplitude des variations est constante autour de la tendance. En effet, si on joint les minima et les maxima de la série chronologique on obtient deux droites parallèles à la tendance.*

Le modèle multiplicatif

Il ya deux forme de modèles multiplicatifs.

Definition 167 *(1ère forme) Les variations saisonnières sont supposées dépendre de la tendance. Alors, on considère que Y_t s'écrit de la manière suivante :*

$$Y_t = C_t \times S_t + \varepsilon_t$$

Remarque 168 *Graphiquement, l'amplitude des variations (saisonnières) varie. En effet, en joignant les minima et les maxima de la série chronologique on constate que les deux droites ne sont pas parallèles entre elles.*

Definition 169 *(2ème forme) On suppose que les variations saisonnières et les variations accidentelles dépendent de la tendance. Alors, on considère que Y_t s'écrit de la manière suivante :*

$$Y_t = C_t \times S_t \times \varepsilon_t$$

Remarque 170 *1) Dans le cas d'une série chronologique Y_t à valeurs positives, le deuxième modèle multiplicatif se ramène à un modèle additif en considérant la série des logarithmes de Y_t :*

$$\ln(Y_t) = \ln(C_t) + \ln(S_t) + \ln(\varepsilon_t)$$

2) La seule différence entre les deux modèles multiplicatifs réside dans l'estimation des ε_t, qui n'a pas une grande importance pour l'instant.

6.2.3 Choix du modèle

Méthode de la bande

On utilise le graphe de la série et la droite passant par les minima et celle passant par les maxima.

- Si ces 2 droites sont à peu près parallèles : le modèle est additif.
- Si ces 2 droites ne sont pas parallèles : le modèle est multiplicatif.

Méthode du tableau de Buys et Ballot

On calcule, pour chacune des années, la moyenne et l'écart type. On représente les points d'abscisse la moyenne et d'ordonnée l'écart type de la même année sur un plan. On trace la droite des moindres carrés de ces points.

- Si l'écart type est indépendant de la moyenne le modèle est additif. La pente (a) de la droite des moindres carrés est très proche de 0.
- Si l'écart type est fonction de la moyenne le modèle est multiplicatif. La pente (a) de la droite des moindres carrés n'est pas nulle.

En conclusion pour décomposer une série chronologique on doit commencer par tracer son graphique, choisir un modèle de composition (additif ou multiplicatif), estimer la tendance C_t, estimer les variations saisonnières.

6.3 Les indices statistiques

Pour l'étude de certains phénomènes économiques et sociaux, on est souvent amené à décrire ou à comparer les variations de grandeurs simples telles que le prix du blé, la production d'acier ou le taux de fécondité d'une certaine population, etc. Pour les comparaisons dans le temps et dans l'espace de ces grandeurs, on introduit la notion d'indice statistique élémentaire. Ceux sont généralement des rapports de ces grandeurs. Mais il est plus instructif de

pouvoir suivre les évolutions de grandeurs plus complexes telles que le niveau général des prix, la production industrielle, le volume des importations, etc. Ces évolutions sont résumées par l'une ou l'autre des caractéristiques de tendance centrale de la série des indices élémentaires correspondants. On parle dans ce cas d'indices synthétiques.

6.3.1 Les indices élémentaires

Example 171 *a) Le prix du kilogramme d'un certain produit a été de* $15DA$ *en moyenne en* 1980 *et il est de* $32DA$ *en Octobre* 1998*. L'indice élémentaire du prix de ce produit en Octobre* 1998*, base* 100 *en* 1980*, est le rapport des deux prix exprimé en pourcentage :*

$$\mathcal{I}_{Oct98/Moy80} = \frac{32}{15}100 = 213,33$$

b) La consommation d'électricité a été de 16500 *Millions de Kwh en* 1988 *et de* 6200 *Millions de Kwh en* 1973*. L'indice élémentaire de la consommation d'électricité en* 1988*, base* 100 *en* 1973*, est le rapport des consommations des deux années exprimé en* % *:*

$$\mathcal{I}_{1988/1973} = \frac{16500}{6200}100 = 266,13$$

Plus généralement, considérons la variation dans le temps d'une grandeur simple X, prenant les valeurs $X_0, X_1, ..., X_t, ...$, aux dates (ou périodes) successives $0, 1, 2, ..., t, ...$

Definition 172 *On appelle indice élémentaire de la grandeur* X *à la date (ou période)* t *par rapport à la date (ou période)* 0*, le rapport :*

$$\mathcal{I}_{t/0} = \frac{X_t}{X_0}$$

Remarque 173 *La date ou période* 0 *est appelée* ***date de référence*** *ou* ***base de l'indice****. La date ou période* t *est appelée* ***date courante****. En général, ce rapport est exprimé en* % *tel que :*

$$\mathcal{I}_{t/0} = \frac{X_t}{X_0}100$$

On dit alors que l'indice à la date t *est exprimé base* 100 *à la date de référence* 0*.*

Les indices statistiques élémentaires sont utilisés surtout pour retracer l'évolution des grandeurs simples dans le temps. Mais ils peuvent aussi servir à des comparaisons dans l'espace.

Example 174 *La densité de la population algérienne a été de* $14,6\ h/Km^2$ *en* 1996*, alors que pour la région algéroise elle a été de* $1540\ h/Km^2$*. L'indice de densité de la région algéroise, l'ensemble de l'Algérie étant choisi comme base, est :*

$$\mathcal{I}_{RA/Al} = \frac{1540}{14,6}100 = 10580$$

L'indice de densité du sud algérien dont la densité de la population est de $0,5\ h/Km^2$*, par rapport à celle du pays, est alors :*

$$\mathcal{I}_{SA/Al} = \frac{0,5}{14,6}100 = 3,4$$

Propriétés

Les indices élémentaires possèdent deux propriétés fondamentales, la **circularité** et la **réversibilité**.

La circularité On dit qu'un indice statistique $\mathcal{I}$ est circulaire si $\forall\ t,\ t\prime$on a :

$$\mathcal{I}_{t/0} = \mathcal{I}_{t/t'} \times \mathcal{I}_{t'/0}$$

En effet,

$$\frac{X_t}{X_0} = \frac{X_t}{X_{t'}} \times \frac{X_{t'}}{X_0}$$

Remarque 175 *On peut comparer les grandeurs aux dates* t *et* $t\prime$*en prenant le quotient des indices* $\mathcal{I}_{t/0}$ *et* $\mathcal{I}_{t'/0}$*. On obtient ainsi un changement de base (la date de référence* $t\prime$*a été substituée à la date* 0*) . La propriété de circularité peut être généralisée à une suite d'indices, i.e.*

$$\mathcal{I}_{t/0} = \mathcal{I}_{t/t-1} \times \mathcal{I}_{t-1/t-2} \times ... \times \mathcal{I}_{2/1} \times \mathcal{I}_{1/0}$$

Réversibilité On dit qu'un indice statistique $\mathcal{I}$ est réversible si $\forall\ t$, on a :

$$\mathcal{I}_{0/t} = \frac{1}{\mathcal{I}_{t/0}}$$

En effet,

$$\frac{X_0}{X_t} = \frac{1}{\frac{X_t}{X_0}}$$

Remarque 176 *L'évolution d'un phénomène est souvent présentée sous forme d'une augmentation ou d'une diminution en pourcentage à l'aide de la formule suivante :*

$$\frac{Valeur\ nouvelle\ -\ Valeur\ primitive}{Valeur\ primitive} \times 100$$

Le pourcentage de variation ne possède pas les propriétés de circularité et de réversibilité des indices, et est donc moins maniable. Les pourcentages de variation ne se rajoutent pas.

6.3.2 Les indices synthétiques

Les grandeurs complexes sont fonction de quelques grandeurs simples. Ainsi le niveau général des prix est constitué des prix des divers aliments et boissons, du logement, de l'équipement ménager, de l'habillement, des services médicaux, des transports, des loisirs, etc. La construction d'un indice synthétique relatif à la variation d'une grandeur complexe consiste à résumer une série d'indices élémentaires.

Position du problème

Soit X une grandeur complexe composée des éléments $X^1, X^2, ..., X^j, ..., X^h$. La variable complexe X est, par exemple, le niveau général des prix, et $X^1, X^2, ..., X^j, ..., X^h$ représentent les prix des différents produits ou services offerts au public . Les indices élémentaires des constituants X^j, $j = 1, 2, ..., h$, de X sont calculés par la formule $\mathcal{I}^j_{t/0} = \frac{X^j_t}{X^j_0}$, $j = 1, 2, ..., h$. Mais cette suite d'indices n'apporte aucune information sur l'évolution du niveau général des prix. Il serait judicieux de les résumer ou de les synthétiser par un seul indice qu'on appellera indice synthétique de la grandeur complexe X.

Les différentes formules d'indices synthétiques

Trois formules d'indices synthétiques sont utilisées en pratique. Ceux sont les formules de Laspeyres, de Paasche et de Fisher.

Soit a_0^j le poids ou l'importance relative du constituants $n^\circ j$ dans la grandeur complexe X à la date 0, et par a_t^j son poids à la date t. Si X représente le niveau général des prix, a_t^j peut représenter, par exemple, la proportion des dépenses dans l'habillement ou dans l'achat des viandes, par rapport à la dépense totale des ménages à la date t. Ces importances relatives ou poids sont soumis à la contrainte suivante :

$$\sum_j a_0^j = \sum_j a_t^j = 1$$

Remarque 177 *Les coefficients a_0^j et a_t^j sont appelés coefficients de pondération.*

Indice de Laspeyres

Definition 178 *L'indice de Laspeyres, noté $\mathcal{L}$, est la moyenne arithmétique des indices élémentaires pondérés par les coefficients a_0^j à la date de référence :*

$$\mathcal{L}_{t/0} = \sum_j a_0^j \mathcal{I}_{t/0} = \sum_j a_0^j \frac{X_t^j}{X_0^j}$$

Indice de Paasche

Definition 179 *L'indice de Paasche, noté $\mathcal{P}$, est la moyenne harmonique des indices élémentaires pondérés par les coefficients a_t^j à la date courante :*

$$\mathcal{P}_{t/0} = \frac{1}{\sum_j \frac{a_t^j}{\mathcal{I}_{t/0}}} = \frac{1}{\sum_j a_t^j \frac{X_0^j}{X_t^j}}$$

Indice de Fisher

Definition 180 *L'indice de Fisher, noté $\mathcal{F}$, est la moyenne géométrique simple des indices de Laspeyres et de Paasche :*

$$\mathcal{F}_{t/0} = \sqrt{\mathcal{L}_{t/0} \times \mathcal{P}_{t/0}}$$

6.3.3 Les différents types d'indices statistiques

Désignons par p_0^j, p_t^j et q_0^j, q_t^j respectivement les prix et les quantités (volumes) correspondant au constituant j entrant dans le calcul d'indice.

Indice de valeur

Definition 181 *La valeur, pour un constituant j, est le produit du prix par la quantité correspondante.*

Definition 182 *L'indice de valeur, noté $\mathcal{V}$, est le rapport de la somme des valeurs relatives à la période courante, à la somme des valeurs relatives à la période de base :*

$$\mathcal{V}_{t/0} = \frac{\sum_j p_0^j q_t^j}{\sum_j p_0^j q_0^j}$$

Indice des prix

L'indice des prix comme l'indice de quantité peut être calculé selon l'une des formules de Laspeyres , de Paasche ou de Fisher .

Definition 183 *L'indice de Laspeyres des prix est donné par la formule suivante :*

$$\mathcal{L}_{t/0}(p) = \frac{\sum_j q_0^j p_0^j \times \frac{p_t^j}{p_0^j}}{\sum_j q_0^j p_0^j}$$

Remarque 184 *Les coefficients de pondération sont constitués par la part de la dépense totale des familles consacrée à la consommation des différents constituants pendant la période de base :*

$$a_0^j = \frac{q_0^j p_0^j}{\sum_j q_0^j p_0^j}$$

Dans le cas d'un indice de prix de détail, les coefficients de pondération sont appelés coefficients budgétaires.

L'indice de Laspeyres des prix peut aussi être défini tel que :

$$\mathcal{L}_{t/0}(p) = \frac{\textit{Dépense totale de la période de base évaluée au prix courant}}{\textit{Dépense totale de la période de base}}$$

Definition 185 *L'indice de Paasche des prix est donné par la formule suivante :*

$$\mathcal{P}_{t/0}(p) = \frac{\sum_j q_t^j p_t^j}{\sum_j q_t^j p_t^j \times \frac{p_0^j}{p_t^j}}$$

Remarque 186 *Les coefficients de pondération sont constitués par la part de la dépense totale des familles consacrée à la consommation des différents constituants pendant la période courante :*

$$a_t^j = \frac{q_t^j p_t^j}{\sum_j q_t^j p_t^j}$$

L'indice de Paasche des prix peut aussi être défini tel que :

$$\mathcal{P}_{t/0}(p) = \frac{\textit{Dépense totale de la période courante}}{\textit{Dépense totale de la période courante évaluée au prix de l'année de base}}$$

Indice de quantité ou de volume

Definition 187 *L'indice de Laspeyres de volume est défini tel que :*

$$\mathcal{L}_{t/0}(q) = \frac{\sum_j p_0^j q_t^j}{\sum_j p_0^j q_0^j}$$

Definition 188 *L'indice de Paasche de volume est défini tel que :*

$$\mathcal{P}_{t/0}(q) = \frac{\sum_j p_t^j q_t^j}{\sum_j p_t^j q_0^j}$$

Propriétés

Les indice de Laspeyres et de Paasche n'ont pas les propriétés de circularité et de réversibilité. L'indice de Fisher n'a pas la propriété de circularité, mais il est réversible :

$$\mathcal{F}_{0/t} = \sqrt{\mathcal{L}_{0/t} \times \mathcal{P}_{0/t}} = \frac{1}{\sqrt{\mathcal{L}_{t/0} \times \mathcal{P}_{t/0}}} = \frac{1}{\mathcal{F}_{t/0}}$$

Les trois types d'indices sont ordonnés de la façon suivante :

$$\mathcal{L}_{t/0} \leq \mathcal{F}_{t/0} \leq \mathcal{P}_{t/0}$$

Bibliographie

[1] Calot, G. (1969) Cours de statistique descriptive, Dunod.

[2] Delmas, B. (2009) Statistique descriptive pour l'économie et la gestion, Presses universitaires du Septentrion, 978-2-7574-0074-6

[3] Delmas, J. F. (2010) Introduction au calcul des probabilités et à la statistique, ENSTA, 978-2-7225-0922-1

[4] Duthil, G. (1998) Initiation à la statistique descriptive , Ellipse Marketing

[5] Grais, B. (2003) Statistique descriptive : Techniques statistiques , Dunod.

[6] Lejeune, M. (2010) Statistique : la théorie et ses applications, Springer, 978-2-8178-0156-8

[7] Olivier, E. (2008) L'essentiel de statistique descriptive, Gualino, 978-2-297-01103-7

[8] Mazerolle, F. (2005) Statistique descriptive : séries statistiques à une et deux variables, séries chronologiques, indices, Gualino, 2-84200-891-X

[9] Moore, D. and McCABE G. P. (2002) Introduction to the Practice of Statistics, 4ème édition, W.H. Freeman & Company.

[10] Morgenthaler, S. (2007) Introduction à la statistique, Presses polytechniques et universitaires romandes, 978-2-88074-734-3

[11] Spiegel, M. et Stephens, L. Statistique : Cours et problèmes, 3ème édition, Série Schaum/McGraw Hill

[12] Tassi, P. (2004) Méthodes statistiques, Economica, 2-7178-4859-2