4

IICT – BAS
eISSN: 2367-8666
Lecture Notes in Computer Science and Technologies

Statistique descriptive

Vera Angelova

eISBN: 978-619-7320-01-5

The series **Lectures Notes in Computer Science and Technologies of the Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences** presents in an electronic format textbooks for undergraduate, graduate and PhD students studied various programs related to Informatics, Computational Mathematics, Mathematical Modeling, Communication Technologies, etc., as well as for all readers interested in these scientific disciplines. The Lecture Notes are based on courses taught by scientists of the Institute of Information and Communication Technologies - BAS in various Bulgarian universities and the Center for Doctoral Training in BAS. The published materials are with open access - they are freely available without any charge.

Editorial board

eISSN: 2367-8666

Avec la collaboration de madame Viviane Baligand et monsieur François Mimiague - Professeur à l'Université de Bordeaux IV, qui ont posé les basses de l'enseignement en Statistique au programme français de la Faculté de gestion et d'économie à l'Université de Sofia.

Table des matières

Introduction

Définition 1 **La Statistique**, c'est l'étude des variations observables. C'est une méthode qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à les analyser et à les interpréter.

Méthodes statistiques

• I-ère étape : **Rassembler** - On collecte des données :
— **soit de manière exhaustive** ou recensement où chaque individu de la population est étudié selon le ou les caractères étudiés.
— **soit par sondage** ou **échantillonnage** qui conduit à n'examiner qu'une fraction de la population, un **échantillon**.
• II-ème étape : **Organiser** - On présente les résultats à l'aide de tableaux, de diagrammes, de graphiques
• III-ème étape : **Analyser** - On résume un tableau de données à l'aide d'un petit nombre de paramètres.
Ces trois étapes constituent la **Statistique descriptive** - l'objectif de nos études.

Dans la phase "analyser" on remplace en fait des données nombreuses résultant de mesures par quelques paramètres porteurs d'information.

En apparence il y a perte d'information, mais si on choisit bien ces paramètres on pourra comparer, interpréter, prévoir.
• IV-ème étape : **Interpréter** - aspect de la ***statistique inductive*** ou ***inférative*** - On interprète les résultats : on les compare avec ceux déduits de la théorie des probabilités.
On pourra donc :
— évaluer une grandeur statistique comme la moyenne ou la variance (estimateurs, intervalles de confiance) ;
— savoir si deux populations sont comparables (tests d'hypothèses) ;
— déterminer si deux grandeurs sont liées et de quelle façon (corrélation, ajustement analytique).
La perte d'information dans la phase "analyser" pose le problème de déterminer quels paramètres et quels modèles mathématiques choisir pour résoudre les problèmes posés dans des domaines si différents.

Il ne faut cependant jamais oublier que l'interprétation dans un domaine concret reste affaire de spécialiste dans le domaine considéré.

Chapitre 1

Vocabulaire et concepts de base

Observation : On désigne par **observation** toute constatation concernant une situation ou un phénomène.

On peut observer le temps qu'il fait, le nombre de personnes qui monte à un arrêt le bus, la couleur des yeux du voisin.

Exemple 1.0.1 /Feuille 1/ : Le principal d'un Université étudie les notes du dernier examen de mathématiques de 20 étudiants d'un groupe. Voici la liste des notes obtenues par les étudiants : 10 7 5 9 13 11 16 17 14 13 16 8 6 10 8 11 10 12 7 9.

Exemple 1.0.2 [Groupes ethniques] /Feuille 1/ : La répartition des groupes ethniques dans une classe de 30 élèves donne le tableau suivant :

Groupes ethniques	Poular	Wolof	Sérère	Diola	Bambara
Effectifs	6	9	7	5	3

Exemple 1.0.3 [Taille] /Feuille 1/ : Voici les tailles (en cm) des 25 élèves d'une classe de 3ème :
165 145 150 150 166 165 160 158 162 165 158 165 162 154 158 160 162 154 165 160 160 158 154 158 160.

Population : L'ensemble soumis à observation est appelé **population**. Cet ensemble doit être bien défini. Le nombre d'individus dans la population est la **taille** de la population.

Exemple 1.0.1 : Les étudiants du groupe consistent la population. 1.0.2 : Les 30 élèves de la classe consistent la population. 1.0.3 : Les 25 élèves de la classe de 3ème consistent la population.

Individu : Tout élément de la population est appelé **individu** ou **élément statistique**.

Exemple 1.0.1 : Chaque étudiant du groupe est un individu. 1.0.2 et 1.0.3 Chaque élève de la classe est un individus.

Les individus peuvent être de nature fort différents :
- *êtres humains* : population d'un pays, élèves d'une classe, salariés d'une entreprise,...
- *objets* : voitures d'un parc automobile, pièces mécaniques fabriquées,...
- *faits ou actes* : appels téléphoniques, arrivées de bateaux, fautes commises,...
- *unités composées* : tonnes/km transportés, ...

Les éléments (individus) d'une population donnée peuvent être caractérisés de nombreuses façons. Par exemple chaque individu de la population d'un pays peut être caractérisé par son sexe, son état civil, son poids, sa taille, le nombre d'enfants à charge, le nombre de diplômes,... Ces caractéristiques sont appelées **caractères statistiques**.

Caractère : C'est la propriété, la caractéristique ou l'aspect singulier que l'on se propose d'observer dans la population ou l'échantillon. Un caractère qui fait le sujet d'une étude porte aussi le nom de **variable statistique.**

Exemple : 1.0.1 : Les notes des étudiants constituent le caractère (ou variable statistique). 1.0.2 Les groupes ethniques constituent le caractère. 1.0.3 La taille des élèves constitue le caractère.

Il existe 2 types de variables :

Les variables qualitatives : sont des variables dont les résultats possibles sont des qualités. Ces résultats sont appelés **modalités.**

Exemple : Le sexe, le programme d'études ou l'état civil sont des exemples de variables qualitatives.

Les variables quantitatives : sont des variables dont les résultats possibles sont des valeurs numériques. Ces résultats sont appelés **modalités ou valeurs.**

Exemple : L'âge, le poids et la taille sont des exemples de variables quantitatives.

Les variables qualitatives peuvent être séparées en deux catégories.

Les variables qualitatives nominales : sont des variables qualitatives dont les modalités ne possèdent aucun ordre naturel.

Exemple : Le caractère considéré en Exemple ethnique [Groupe ethnique] et l'Exercice 3 /Feuille 1/.

Les variables qualitatives ordinales : sont des variables qualitatives dont les modalités peuvent être ordonnées.

Exemple : Le caractère considéré en Exercices 4 et 5 /Feuille 1/.

Les modalités des caractères qualitatifs **nominales** sont exprimables par des **noms** et **ne sont pas hiérarchisées** - dans les tables des fréquences ne figurent pas les effectifs et fréquences cumulées.

Un caractère nominal peut être **dichotomique** s'il ne peut prendre que deux modalités.

Exemple : Sexe - Exercice 1 /Feuille 1/, la présence ou l'absence d'un caractère, etc.

Les modalités des caractères qualitatifs **ordinales** traduisent le **degré** d'un état caractérisant un individu sans que ce degré ne puisse être défini par un nombre qui résulte d'une mesure. Les modalités sont alors **hiérarchisées** - les tables de fréquences se construisent de manière analogue à celle d'une variable quantitative discrète.
Exemple : le stade d'une maladie

Les variables quantitatives peuvent être également séparées en deux catégories.

Un caractère quantitatif est **discret** si les modalités (valeurs) possibles sont isolées, souvent entières. En général il résulte d'un comptage ou dénombrement.

Exemple : Notes des étudiants - Exemple 1.0.1 /Feuille 1/, nombre d'enfants d'une famille - Exercice 1 /Feuille 1/.

Un caractère quantitatif est **continue**, lorsqu'il peut prendre n'importe quelles valeurs d'un intervalle.

Exemple : Poids, taille, diamètre de pièces, salaires...

Remarque 2 *:* En réalité le nombre de valeurs possibles pour un caractère donné dépend de la précision de la mesure. On peut considérer comme continu un caractère discret qui peut prendre un grand nombre de valeurs.

Le diagramme 1.1 résume bien le tout

Les échelles de mesure. L'étude des variables doit se faire avec des outils de mesure. Il n'est pas possible de mesurer le degré de satisfaction de la même façon qu'on peut mesurer la taille d'un individu. Voilà pourquoi, on utilise différentes échelles de mesures.

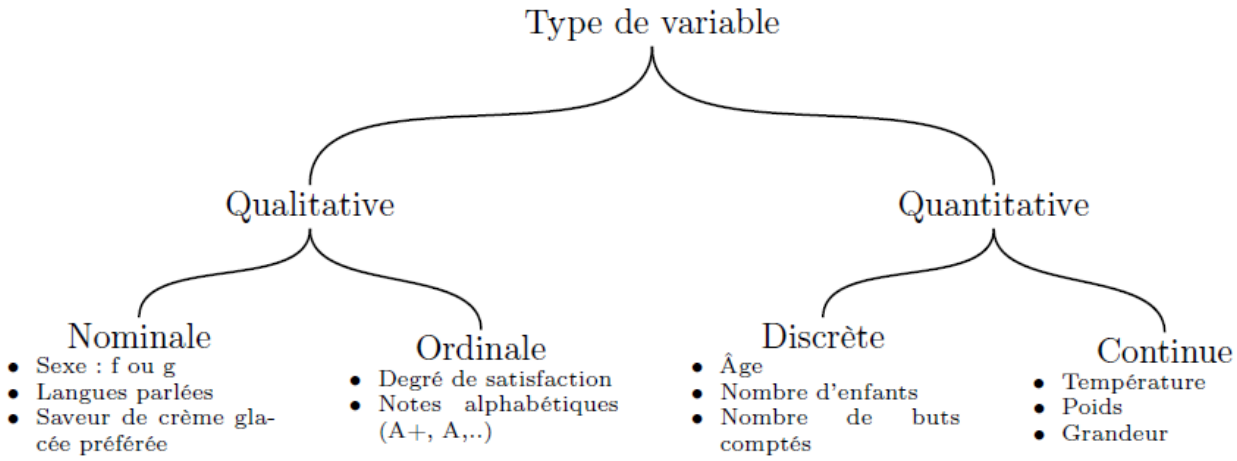


FIGURE 1.1 : Diagramme des différents types de variables

Il existe 4 types d'échelles de mesure. *Le mot clé pour les retenir est NOIR.*

(1) L'échelle de mesure est dite **N**ominale si les résultats sont des modalités ne possédant pas d'ordre naturel.

(2) L'échelle de mesure est dite **O**rdinale si les résultats possèdent un ordre naturel.

(3) L'échelle de mesure est dite d'**I**ntervalles si les résultats sont des valeurs ayant un zéro arbitraire.

(4) L'échelle de mesure est dite de **R**apports si les résultats sont des valeurs ayant un zéro absolu.

Exemple (Échelle d'intervalles). Lorsque la température augmente de 10 degrés Celsius à 20 degrés Celsius, on ne peut pas dire qu'il fait 2 fois plus chaud, parce que 10 degrés Celsius correspond à 50 degrés Farenheit et 20 degrés Celsius correspond à 68 degrés Farenheit, ce qui ne correspond pas au double de la température. Alors on peut seulement dire qu'il fait 10 degrés de plus.

Exemple (Échelle de rapports). Par contre, dans le cas d'une variable mesurée à l'aide d'une échelle de rapports, on peut affirmer qu'un élève qui a 2 emplois a deux fois plus d'emplois qu'un élève qui n'a qu'un seul emploi.

Les quatre échelles de mesure possédant des propriétés fort différentes permettent différents niveaux de mesure d'un phénomène exprimée par un caractère. Le niveau (l'échelle) de mesure utilisé détermine le type d'analyses statistiques qu'il est permis de faire sur des données. Le tableau suivant reflète ce point de vue :

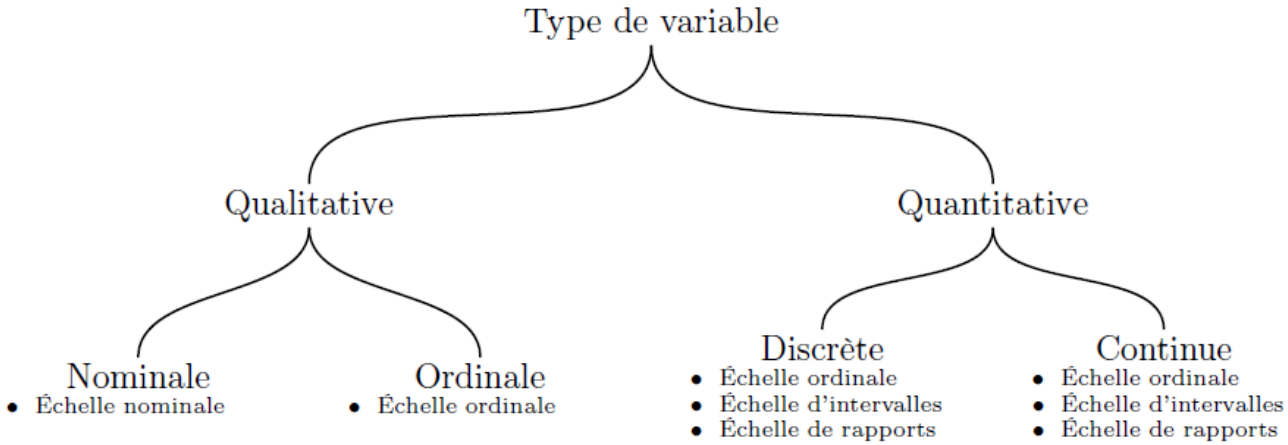


FIGURE 1.2 : Diagramme des différentes échelles de mesure

TYPE D'ÉCHELLES

	NOMINALE	ORDINALE	INTERVALLES ÉGAUX	DE PROPORTIONS
Vérifications permises	Déterminer si deux valeurs sont égales	Déterminer si une valeur est plus petite ou plus grande qu'une autre	Déterminer si des intervalles sont égaux	Déterminer si des proportions sont égales
Opérations empiriques permises	Identifier et classer	Mettre en rang	Évaluer les différences entre les distances des scores individuels	Déterminer des proportions, des fractions ou des multiples
Exemples de variables	Sexe, programme d'études	Ordre de remise des examens	Scores aux tests et échelles standardisés "zéro" arbitraire	Longueur, poids, température en degrés kelvin "zéro" absolu
Analyses statistiques permises	Tableau de fréquences ; mode ; coefficient de contingence	+ médiane ; centiles ; Khi carré	+ moyennes ; r de Pearson ; anova ; F ; test t	+ moyenne géométrique ; coefficient de variation

Paramètre : Quantité numérique qui résume un aspect de la population. (valeur moyenne d' une variable).

Pour recueillir des informations sur une population statistique, l'on dispose de deux méthodes :
— la **méthode exhaustive** ou recensement où chaque individu de la population est étudié selon le ou les caractères étudiés.
— la **méthode des sondages** ou **échantillonnage** qui conduit à n'examiner qu'une fraction de la population, un **échantillon**.

L'échantillonnage représente l'ensemble des opérations qui ont pour objet de prélever un certain nombre d'individus dans une population donnée.

Echantillon : C'est un sous ensemble de la population considérée. Le nombre d'individus dans l'échantillon est la **taille** de l'échantillon.

Statistique : Quantité numérique qui résume un aspect d'un échantillon. (valeur moyenne observée sur un échantillon).

Données statistiques

Il existe deux **sortes de données statistiques :**

— celles qui concernent une population et précisent comment les éléments de cette population se répartissent en classes - ***les séries statistiques***. Il y a déplacement dans l'espace.

L'ordre dans lequel les observations se présent dans la série statistique - autrement dit l'ordre dans lequel des individus se présentent dans le tableau individus x caractères - est souvent **arbitraire** (ordre de prise en considération, ordre alphabétique ; ...) Modifier cet ordre n'a généralement aucune influence sur le traitement de la série statistique. /Exemple : Exercice 2 de Feuille 1/

— celles qui concernent une mesure ou une quantité et précisent comment cette mesure ou cette quantité évolue dans le temps. Ce sont les ***séries chronologiques ou chronique***. Il y a déplacement dans le temps.

L'ordre dans lequel les observations se présentent dans la série chronologique est fixé par le contexte. Généralement on désigne les série chronologiques par $(x_t; t = 1; \ldots, n)$ où le choix de la lettre t pour l'indice fait référence au "temps" - Exemple : Exercice 4 de Feuille 1. Dans ce cas l'ordre des observations successives est toujours déterminé par l'écoulement naturel du temps.

Exemple 1.0.4 On considère les 14 étudiants en Bases de la statistique 2 en 2015. On s'intéresse au nombre d'absences aux cours et au nombre de participation au tableau pendant les cours.

On choisit au hasard 5 étudiants parmi les 14 et analyse leurs résultats.
Analyse :
Population : les étudiants en Base de statistique 2 en 2015.
Taille de la population : $N = 14$.
Individu ou élément statistique : un étudiant du groupe en Gestion ou du groupe en Économie.
Caractères : a/ nombre d'absences aux cours, b/ nombre de participations au tableau pendant les cours
Paramètres : moyenne μ, écart-type σ.
Échantillon : 5 étudiants pris des 14 au hasard.
Taille de l'échantillon : $n = 5$.
Caractères : a/ nombre d'absences aux cours et b/ nombre de participations au tableau pendant les cours.
Statistiques : moyenne $\bar{x}$, écart-type s.

Supposons que l'on étudie p caractères sur une population de N individus. Si les individus et les caractères sont représentés respectivement par les identificateurs $1, 2, \ldots, i, \ldots, N$ et $1, 2, \ldots, j, \ldots, p$ on obtient le tableau suivant :

Individus	Variables (caractères)					
	1	2		j		p
1	x_{11}	x_{12}	$\ldots$	x_{1j}	$\ldots$	x_{1p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	x_{i1}	x_{i2}	$\ldots$	x_{ij}	$\ldots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
N	x_{N1}	x_{N2}	$\ldots$	x_{Nj}	$\ldots$	x_{Np}

Chaque ligne de ce tableau correspond à un individu pour lequel on dispose de p valeurs observées.
Chaque colonne correspond à une variable dont on effectue N mesures.

La valeur x_{ij} indique le résultat de la mesure du caractère j observée pour l'individu i.
L'ensemble des lignes d'un tableau I x C à p colonnes définit une *série statistique p-variée.*

Si $p = 1$, la série est dite univariée et dans ce cas on utilise une notation simplifiée :

i	1	2	3	$\ldots$	i	$\ldots$	N
x	x_1	x_2	x_3	$\ldots$	x_i	$\ldots$	x_n

Chapitre 2

Organisation d'une série statistique univariée

Il existe plusieurs niveaux de description statistique : la présentation brute des données, des présentations par tableaux numériques, des représentations graphiques et des résumés numériques fournis par un petit nombre de paramètres caractéristiques.

2.1 Séries statistiques

Définition 3 Une **série statistique** correspond aux différentes modalités d'un caractère sur un échantillon d'individus appartenant à une population donnée.

Une série statistique est obtenue par l'observation d'une variable chez n individus (ou éléments, unités...) ; elle correspond à la liste des valeurs ou modalités prises par la variable chez chacun des n individus.

2.1.1 Série observée, données brutes

La saisie des données nous a fourni une suite de n valeurs observées de la variable X : $\{x_1, x_2, \ldots, x_n\}$. C'est la **série observée, données brutes.**

Exemple 2.1.1 [Age] /Feuille 2/ : Ages de 100 employés d'une entreprise (échantillon)

60	39	23	30	29	26	29	41	40	32
63	22	32	52	46	35	25	28	33	33
20	25	42	34	29	43	41	31	30	36
58	21	24	55	51	28	18	40	44	38
32	21	30	31	25	49	31	26	33	36
43	34	35	22	33	38	34	34	33	34
23	26	57	23	26	36	39	31	35	34
34	51	40	50	35	45	28	36	32	39
26	48	17	45	45	25	25	30	36	30
43	25	27	21	53	25	38	33	37	33

$x_1 = 60\,;\ x_2 = 39\,;\ \ldots$

Exemple 2.1.2 [Notes] /Feuille 2/ : Relève des notes d'un groupe de 30 étudiants à l'examen en Statistique. On observe la variable aléatoire X ="relève des notes d'un étudiant". On a les résultats les suivants :

$$3; 2; 4; 6; 5; 2; 3; 6; 4; 4; 2; 3; 3; 4; 4$$
$$5; 6; 4; 4; 3; 3; 3; 3; 4; 4; 4; 4; 5; 5; 5$$

$x_1 = 3\,;\ x_2 = 2\,;\ \ldots$
Évidemment les notes sont égales aux nombres 2,3,4,5 et 6. Le nombre des modalités du caractère est 5.

2.1.2 Série ordonnée

Ordonner une série d'observations, quand le caractère étudie est mesure sur une échelle ordinale ou quantitative, consiste en une première réorganisation naturelle des données qui permet de mieux apprécier la répartition des valeurs observées et de faciliter la construction des distributions observées.

Si X est mesuré sur une échelle ordinale ou quantitative, on peut organiser les données en ordonnant les valeurs observées. On obtient ainsi la **série ordonnée** que nous noterons :

$$x_{(1)},\ x_{(2)},\ x_{(3)},\ \ldots,\ x_{(n)} \quad \text{ou } x_{(i)};\ i = 1, \ldots, n \quad \text{avec } x_{(i)} \leq x_{(j)} \text{ si } i < j$$

On place l'indice de la valeur ordonné entre parenthèses pour distinguer de celui qui définit la donnée observée. Cet indice (i) est appelé le **rang** de l'observation correspondante. La plus petite observation, $x_{(1)}$, est donc l'observation de rang 1 (elle est parfois aussi notée x_{min}) ; la plus grande observation, $x_{(n)}$, est l'observation de rang n (autre notation possible : x_{max}).

Dans le cas d'une variable ordinale, le symbole $\leq$ représente la relation d'ordre choisie pour l'échelle de mesure : par exemple, dans le cas où l'échelle de mesure correspond à une échelle d'appréciation de la qualité d'un produit, $x_{(i)} \leq x_{(i)};\ \ (i \leq j)$ signifie que $x_{(i)}$ est moins bon ou de même qualité que $x_{(j)}$.

La série ordonnée des données de l'exemple 2.1.1

17	18	20	21	21	21	22	22	23	23
23	24	25	25	25	25	25	25	25	26
26	26	26	26	27	28	28	28	29	29
29	30	30	30	30	30	31	31	31	31
32	32	32	32	33	33	33	33	33	33
33	34	34	34	34	34	34	34	35	35
35	35	36	36	36	36	36	37	38	38
38	39	39	39	40	40	40	41	41	42
43	43	43	44	45	45	45	46	48	49
50	51	51	52	53	55	57	58	60	63

$x_{(1)} = 17\,;\ x_{(2)} = 18\,;\ \ldots$

La série ordonnée de l'exemple 2.1.2 [Notes] est :

$2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 4$
$4; 4; 4; 4; 4; 4; 4; 5; 5; 5; 5; 5; 6; 6; 6$ $\qquad x_{(1)} = 2;\ x_{(2)} = 2;\ x_{(3)} = 2;\ x_{(4)} = 3; \ldots$

2.1.3 Tableaux statistiques, D.O.1

Le **tableau de distribution des effectifs/fréquences, la distribution observée univariée /D.O.1/** est un mode synthétique de présentation des données. Sa constitution est immédiate dans le cas d'un caractère discret et d'un caractère qualitatif ordinal, mais nécessite en revanche une transformation des données dans le cas d'un caractère continu.

A. Fréquences absolues, relatives et cumulées

Supposons qu'il y ait k valeurs distinctes $x_1, x_2, x_3, \ldots, x_i, \ldots, x_k$ (k modalités) et notons n_i, le nombre de fois que la valeur x_i a été observée. Ce nombre est appelé **l'effectif** ou **fréquence absolue** de la modalité x_i.

Le nombre n d'individus est appelé **l'effectif total** de la population observée.

On obtient le **tableau des effectifs, distribution observée** - la distribution de la variable statistique /le caractère/ considérée suivant :

modalités	x_1	x_2	$\ldots$	x_i	$\ldots$	x_k	effectif total
effectifs	n_1	n_2		n_i		n_k	$n = n_1 + n_2 + \ldots + n_k$

Ce tableau définit une **distribution statistique observée** à une dimension (D.O.1).

Définition 4 **Distribution observée (D.O.1) :** Une distribution observée se construit à partir d'une série statistique en faisant la liste des valeurs distinctes qui apparaissent dans la série statistique et en associant à chaque valeur distincte son effectif, c'est-à-dire le nombre de fois que la valeur apparaît dans la série.

Exemple 2.1.2 [Notes] :

La note 2 se rencontre 3 fois, la note 3 - 8 fois, note 4 - 11 fois, note 5 - 5 fois et note 6 - 3 fois. L'effectif de la valeur 2 est 3, l'effectif de la valeur 3 est 8, etc. Les notes en ordre accèdent et les effectifs correspondants forment la **distribution des effectifs** :

Note x_i	Effectif n_i
2	3
3	8
4	11
5	5
6	3

Une **distribution statistique observée** est définie par les valeurs distinctes qui apparaissent dans la série observée ou ordonnée et le nombre de fois que chacune d'elles apparaît.

Tout caractère statistique observé sur une population définit une distribution statistique que l'on peut représenter par l'ensemble des couples (x_i, n_i).

Au lieu de l'effectif n_i il est parfois plus utile de considérer le rapport $f_i = n_i/n$. Ce nombre est appelé la **fréquence** ou **fréquence relative** de la valeur x_i.

On peut aussi dans certain cas utiliser le **pourcentage**. C'est la fréquence exprimée en pour cent. Il est égal à $100 \times f_i$.

Les effectifs cumulés : $N_i = n_1 + n_2 + \ldots + n_i$.
C'est le nombre d'observations inférieures ou égales à x_i.

Règles de calcul des effectifs cumulés croissants (la série est ordonnée suivant l'ordre croissant) :
— L'effectif cumulé croissant de la 1ère modalité est toujours égal à l'effectif partiel de la 1ère modalité.
— L'effectif cumulé croissant d'une modalité est la somme de l'effectif partiel de cette modalité et de l'effectif cumulé croissant de la modalité précédente.
— L'effectif cumulé croisant de la dernière modalité est toujours égal à l'effectif total.

La colonne des effectifs cumulés décroissants :
— L'effectif cumulé décroissant de la 1ère modalité est toujours égal à l'effectif total.
— L'effectif cumulé décroissant d'une modalité est la différence de l'effectif cumulé de la modalité précédente et de l'effectif partiel de cette modalité.
— L'effectif cumulé décroisant de la dernière modalité est toujours égal à l'effectif partiel de la dernière modalité.

Les fréquences cumulées : $g_i = f_1 + f_2 + \ldots + f_i$.

B. Caractères quantitatifs discrets

Dans le cas d'un caractère quantitatif discret, l'établissement de la distribution des données observées associées avec leurs fréquences est immédiate.
Exemple : Pour l'exemple 2.1.2 [Notes] on peut calculer aussi les effectifs cumulés, les fréquences et les fréquences cumulées pour la variable observée :

Note x_i	Effectif n_i	Effectif cumulé $N_i = \sum_{j=1}^{i} n_j$	Fréquence $f_i = n_i/n$	Fréquence cumulée $g_i = \sum_{j=1}^{i} f_j$
2	3	3	3/30 =0,1	3/30=0,1
3	8	11	8/30=0,2666	11/30=0,3666
4	11	22	11/30=0,3666	22/30=0,7333
5	5	27	5/30=0,1666	27/30=0,9
6	3	30	3/30=0,1	1
Total	30		1	

Distribution de fréquences observées

C. Caractères quantitatifs continues

Lorsque l'on observe un caractère continu ou un caractère discret avec un grand nombre de valeurs distinctes (comme dans l'exemple 2.1.1) 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. On procède à un **regroupement en classes**.

Définition 5 **Distribution groupée (D.G.1) :** Une distribution groupée se construit à partir d'une série statistique en regroupant les observations de la série dans un certain nombre de classes et en associant à chaque classe son effectif, c'est-à-dire le nombre d'observations de la série qu'elle contient.

- Si le caractère est qualitatif ou discontinu, une classe contient tous les individus ayant la même modalité ou la même valeur du caractère.
- Si le caractère est continu, une classe est un intervalle.

Pour construire ces intervalles, on respecte les règles suivantes :
— 1. Le nombre de classes est compris entre 5 et 20 (de préférence entre 6 et 12)
— 2. Les classes sont constituées d'intervalles semi-ouverts consécutifs en général $[a, b[$. Chaque classe (sauf la dernière) contient sa borne inférieure mais pas sa borne supérieure.
— 3. L'**amplitude d'une classe** est le nombre $|b - a|$. Les classes peuvent avoir toutes la même amplitude ou avoir des amplitudes différentes. Chaque fois que cela est possible, les amplitudes des classes sont égales.
— 4. Le nombre $x_i^* = (a + b)/2$ est appelé **indice** ou **centre de classe**. Dans les calculs, une classe sera représentée par son centre, qui est le milieu de l'intervalle.

Une fois la classe constituée, on considère les individus répartis uniformément entre les deux bornes (ce qui entraîne une perte d'informations par rapport aux données brutes).
Exemple : La distribution observée de l'exemple 2.1.1 est :
— Données regroupées en classes (intervalles).
— Effectif d'une classe = nombre d'observations

Age	Effectif
15-20	2
20-25	10
25-30	19
30-35	27
35-40	16
40-45	10
45-50	6
50-55	5
55-60	3
60-65	2

/FREQUENCY($\{x_i\}$;$\{19, 24, 29, \ldots, 59\}$), F2, Ctrl+Shift+Enter/
/FREQUENCY(J1 :J60 ;Q1 :Q6), F2, Ctrl+Shift+Enter/
/Tools/Data Analisys/Histogram(input range ;bin range)/

Problèmes liés aux classes : choix du nombre de classes, choix des amplitudes, valeurs extrêmes de la classe (arrondi ou troncature)

Choix du nombre des classes
Diverses formules empiriques permettent d'établir le nombre de classes pour un échantillon de taille n.

Règle empirique :	Nombre de classes k égale à l'entier le plus faible tel que $2^k \geq n$
La règle de STURGE :	Nombre de classes $k = 1 + (3, 3 \log n)$
La règle de YULE :	Nombre de classes $k = 2, 5\sqrt[4]{n}$

L'amplitude (la longueur) de chaque classe est obtenu ensuite de la manière suivante :

Amplitude de classe $h = (x_{max} - x_{min})/k$
avec x_{max} et x_{min}, respectivement la plus grande et la plus petite valeur de X dans la série statistique.

Le fait de construire des classes de même longueur peut simplifier l'analyse de la D.G.1. Il est toutefois parfois préférable d'utiliser des classes de longueurs différentes. Il en est ainsi, par exemple, de manière à pouvoir inclure dans la classe de plus grande longueur que les précédentes des observations sans devoir créer des classe intermédiaires vides.

A partir de x_{min} on obtient les limites de classes ou **bornes de classes** par addition successive de l'intervalle de classe. En règle général, on tente de faire coïncider **l'indice de classe** ou valeur centrale de la classe avec un nombre entier ou ayant peu de décimales.

Exemple 2.1.3 [Bovins] /Feuille 2/ : Nombre de bovins dans les fermes privées dans une région donnée. Taille de la population $N = 60$. Série observée :

$$4; 2; 3; 5; 0; 1; 6; 1; 5; 4; 7; 10; 13; 16; 19; 8; 4; 11; 14; 17; 2; 11$$
$$9; 8; 12; 15; 18; 10; 13; 18; 7; 11; 14; 17; 21; 8; 9; 11; 10; 16; 8; 10$$
$$15; 17; 19; 11; 9; 13; 11; 11; 12; 9; 10; 11; 10; 12; 10; 11; 12; 14$$

On forme la série ordonnée

$$0; 1; 1; 2; 2; 3; 4; 4; 4; 5; 5; 6; 7; 7; 8; 8; 8; 8; 9; 9; 9; 9$$
$$10; 10; 10; 10; 10; 10; 10; 11; 11; 11; 11; 11; 11; 11; 11; 11; 12; 12; 12; 12$$

$$13; 13; 13; 14; 14; 14; 15; 15; 16; 16; 17; 17; 17; 18; 18; 19; 19; 21$$

Les modalités sont 21. Il faut organiser les individus en classes entre 6 et 12. Comme la taille de la population (60) n'est pas trop grande on va grouper les individus en 6 - 7 classes. Si la taille de la population est grande, il faut choisir plus grand nombre des classes. Le groupage en classes s'effectue de la manière la suivante : On détermine la valeur minimale x_{min} et la valeur maximale x_{max} de la variable. Dans notre cas $x_{min} = 0$ et $x_{max} = 21$. La différence $x_{max} - x_{min}$ s'appelle **l'étendu** de la variable et on la note par Δ :

$$\Delta = x_{max} - x_{min}.$$

Dans notre cas $\Delta = 21 - 0 = 21$. L'étendu est la longueur de l'intervalle qui contient toutes les valeurs de la variable observée. Ordinairement on divise l'intervalle en sous-intervalles égaux - classes. Dans notre cas on doit choisir un petit nombre de classes (population de faible taille). On peut choisir le nombre des classes entre 6 et 7. Comme 21 est divisible en 7, on choisit $k = 7$ classes. L'amplitude de chaque classe h est le quotient de l'étendu Δ par le nombre des classes k :

$$h = \frac{\Delta}{k} = \frac{21}{7} = 3.$$

Si l'étendu n'est pas divisible au nombre des classes la dernière classe contient peu d'éléments que les autres classes.

Après avoir déterminer Δ, k et h on forme les classes en quelles l'étendu est groupé :

$$\Delta : [0; 3), [3; 6), [6; 9), \ldots, [18; 21]$$

Les classes sont fermées à gauche et ouvertes à droite - 6 appartient à la classe $[6; 9)$ et non pas à la classe $[3; 6)$.

La série groupée en classes est donné dans le **tableau distribution des effectifs, distribution groupée unitaire (D.G.1)** :

Données regroupées en classes et effectifs des classes. L'effectif d'une classe = nombre d'observation dans cette classe :

Nombre de bovins	Effectif
$[0; 3)$	5
$[3; 6)$	6
$[6; 9)$	7
$[9; 12)$	20
$[12; 15)$	10
$[15; 18)$	7
$[18; 21]$	5
Total	60

/FREQUENCY($\{x_i\}$;$\{19, 24, 29, \ldots, 59\}$), F2, Ctrl+Shift+Enter/
/FREQUENCY(A1 :A60 ;B1 :B7), F2, Ctrl+Shift+Enter/
/Tools/Data Analisys/Histogram(input range ;bin range)/
/Tools/Data Analisys/Histogram(A1 :A60 ;B1 :B7)/

On vas calculer les centres des classes, les fréquences, les effectifs cumulés et les fréquences cumulées. On obtient le tableau de distribution des fréquences :

Nombre de bovins classe $[a;b)$	Centre de la classe $x_i^* = \frac{a+b}{2}$	Effectif n_i	Effectif cumulé $N_i = \sum_{j=1}^{i} n_j$	Fréquence $f_i = \frac{n_i}{n}$	Fréquence cumulée $g_i = F_i$ $F_i = \frac{1}{n}\sum_{j=1}^{i} n_j$
$[0;3)$	1,5	5	5	5/60	5/60
$[3;6)$	4,5	6	11	6/60	11/60
$[6;9)$	7,5	7	18	7/60	18/60
$[9;12)$	10,5	20	38	20/60	38/60
$[12;15)$	13,5	10	48	10/60	48/60
$[15;18)$	16,5	7	55	7/60	55/60
$[18;21]$	19,5	5	60	5/60	1
Total :		60		1	

D. Variable qualitative

Table des fréquences

- Lorsque la variable est ordinale, elle est construite de manière analogue à celle d'une variable quantitative discrète.

- Lorsque la variable est nominale, n'y figurent pas les effectifs et fréquences cumulés.

2.1.4 Représentations graphiques

Elles servent à visualiser la répartition des individus. Les représentations graphiques ont l'avantage de renseigner immédiatement sur l'allure générale de la distribution. Elles facilitent l'interprétation des données recueillies.

A. Caractères qualitatifs

On utilise des **diagrammes à secteurs circulaires** (camembert), ou des **diagrammes en tuyaux d'orgue, diagrammes en barres** (analogue au diagramme en bâtons), des **diagrammes en bandes** - tous représentant la répartition en effectif ou fréquence des individus dans les différentes modalités de la série. Le principe est de représenter des aires proportionnelles aux fréquences de la variable statistique.

Exemple 2.1.4 /Feuille 3/ : Le tableau suivant donne la composition en acides gras insaturés en grammes pour 100 grammes d'huile d'olive vierge :

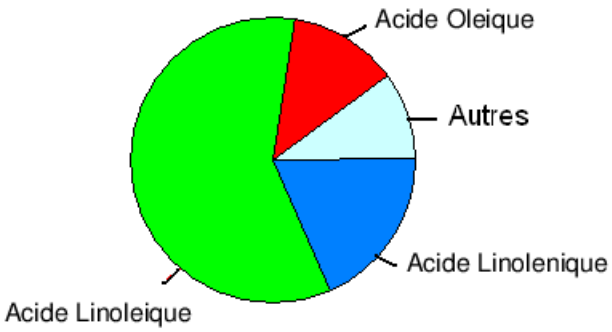
Modalité	Effectif n_i	$\alpha_i = 360 f_i$
Acide Oléique	18,6	$360 \times \frac{18,6}{100} = 66,96$
Acide Linoléique	58,6	$360 \times \frac{58,6}{100} = 210,96$
Acide Liolénique	12,7	$360 \times \frac{12,7}{100} = 45,72$
Autres composants	10,1	$360 \times \frac{10,1}{100} = 36,36$

A.1. Diagramme circulaire (Camembert)

Il s'agit de représenter par une portion sur un disque, la part d'une modalité d'une variable. Cette part correspond à la fréquence et associe à chacune des fréquences une mesure d'angle, la somme de ces mesures vaudra 360°, la mesure du disque complet. Ces mesures s'obtiennent tout simplement par une règle de trois.
A 1 (ou 100%) correspond 360°, à f_i correspond α_i, un nombre compris entre 0 et 360°. On a donc :

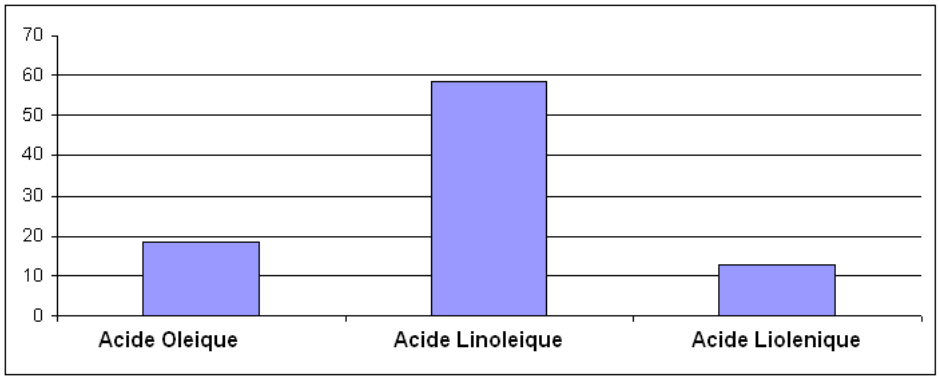
$$\alpha_i = f_i \times 360 = \frac{n_i}{n} \times 360.$$

Exemple : Diagrammes circulaires de la composition de l'huile d'olive de l'exemple 2.1.4 :



A.2. Tuyau d'orgue ou diagramme en bâtons

Il s'agit de représenter chaque modalité par un rectangle vertical dont la hauteur est proportionnelle à l'effectif (ou fréquence).
Exemple : Le diagramme en bâtons de l'exemple 2.1.4 :



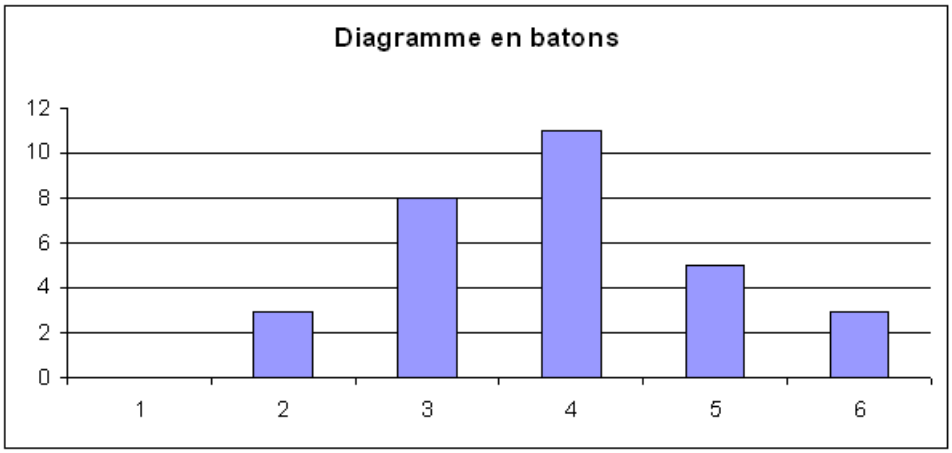
B. Caractères quantitatifs discrets

On utilise un **diagramme différentiel en bâtons**, complété du diagramme des fréquences cumulées appelé **diagramme cumulatif**.

B.1 Diagramme différentiel en bâtons

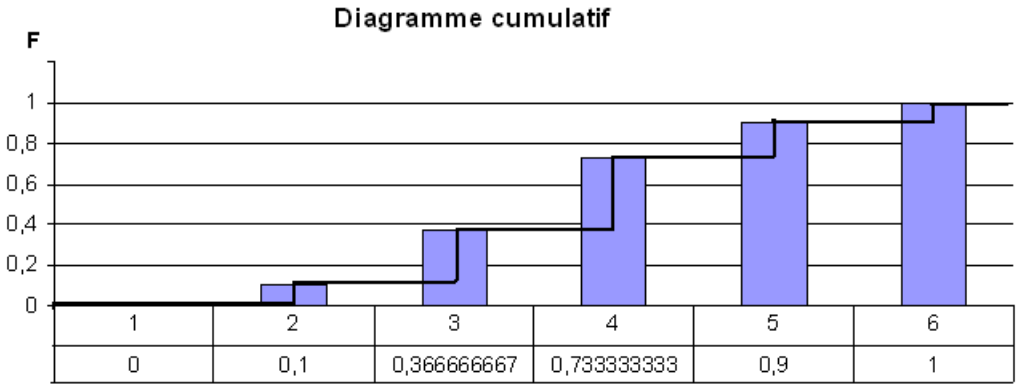
Au diagramme en bâtons la hauteur des bâtons correspond à l'effectif n_i associé à chaque modalité du caractère x_i. On a en ordonnée les effectifs n_i et en abscisse les différentes modalités de la variable étudiée.

Exemple : Pour l'exemple 2.1.2 [Notes] le diagramme en bâtons est :



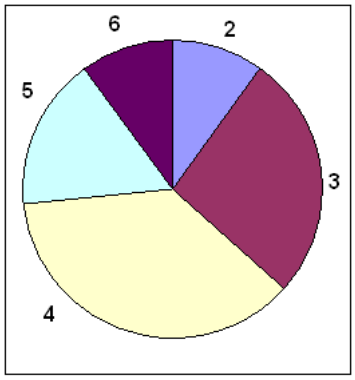
B.2 Diagramme cumulatif

Le diagramme cumulatif est la représentation graphique d'une fonction F, appelée fonction de répartition de la variable statistique.



B.3 Diagramme circulaire

Exemple : Le diagramme circulaire de l'exemple 2.1.2 [Notes] :



C. Caractère quantitatif continu et caractère quantitatif discret, groupé en classes

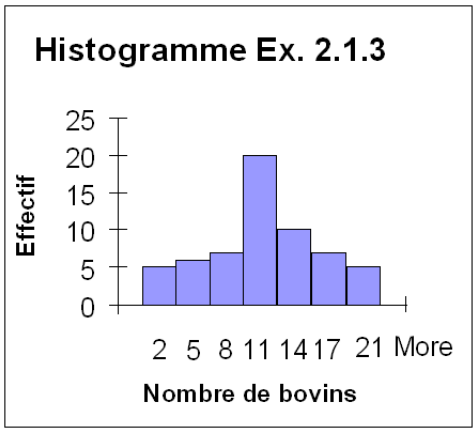
C.1. Histogramme

Pour les caractères quantitatifs continus (et les caractères quantitatifs discrets groupés en classes, aussi), la représentation graphique est l'histogramme : ce sont des rectangles juxtaposés dont chacune des bases est égale à l'intervalle de chaque classe et dont la hauteur est telle que **l'aire de chaque rectangle soit proportionnelle** aux effectifs (histogramme des effectifs) ou aux fréquences de la classe correspondante (histogramme des fréquences). En revanche lorsque les intervalles de classe sont inégaux, des modifications s'imposent pour conserver cette proportionnalité. Dans ce cas, en ordonnée, au lieu de porter l'effectif, on indique **l'effectif unitaire** qui est égal au **rapport de l'effectif sur l'amplitude de la classe**.

$$\text{Effectif unitaire } = n_i/h_i \quad \text{montre la densité des observations}$$

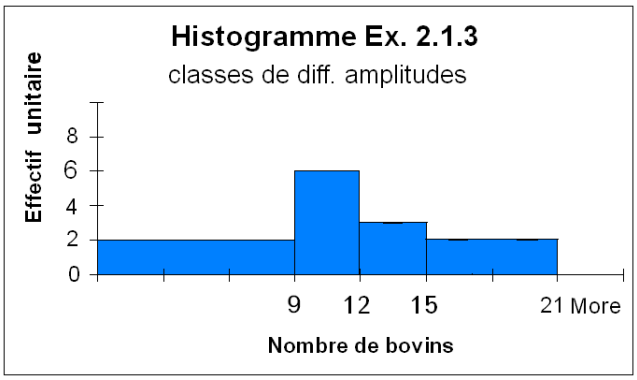
Ainsi la superficie de chaque rectangle représente alors l'effectif associé à chaque classe.

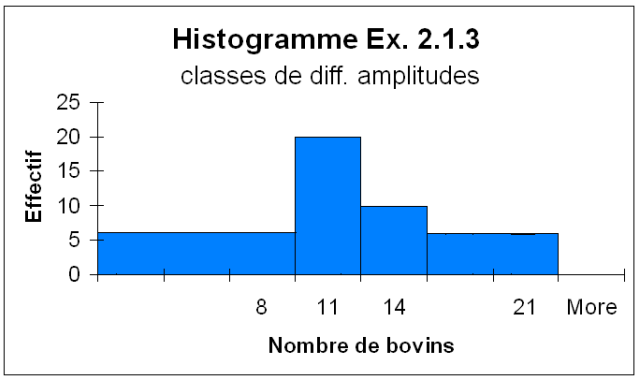
Exemple : Pour l'exemple 2.1.3 [Bovins] le histogramme des effectifs est :



/Tools/Data Analisys/
Histogram(input range ;bin range)
√ Chart Output/

Si les valeurs des premières trois classes sont regroupées en une classe [0 :9) d'amplitude $h_1 = 9$ et les valeurs observées des deux dernières classes sont regroupées en une classe [15 :21) d'amplitude $h_4 = 6$, les histogrammes des effectifs unitaires n_i/h_i et des effectifs n_i sont les suivants :

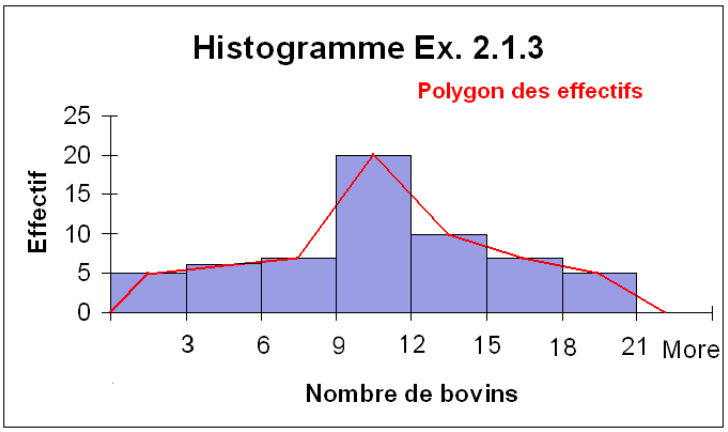




C.2. Polygone des effectifs (des fréquences).

On obtient le polygone des effectifs (ou des fréquences) en reliant les milieux des bases supérieures des rectangles (les centres des classes).

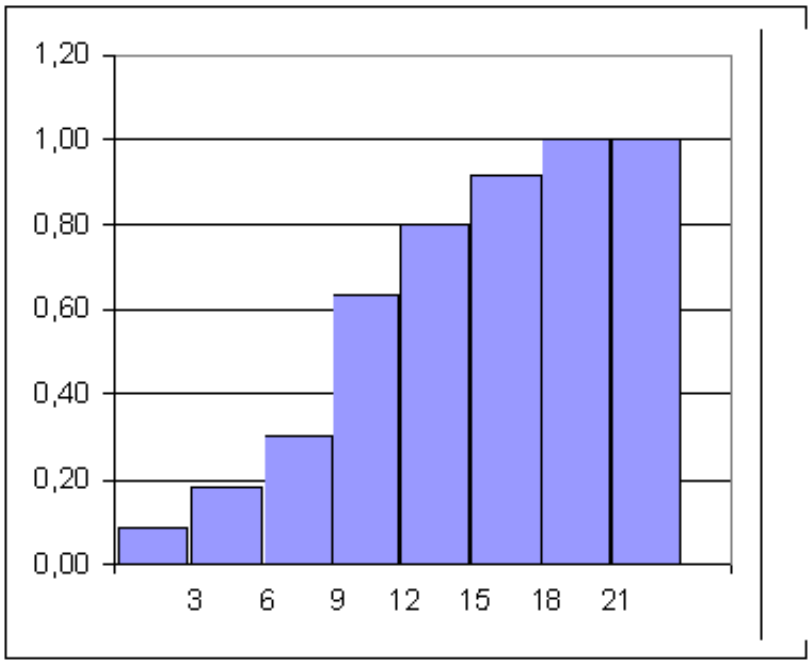
Exemple : Pour l'exemple 2.1.3 [Bovins] le polygone des effectifs est donné en rouge :



C.3. Histogramme des fréquences cumulées croissantes, décroissantes

L'histogramme des fréquences cumulées croissantes - ce sont des rectangles juxtaposés dont chacune des bases est égale à l'amplitude de chaque classe et dont la hauteur est telle que l'aire de chaque rectangle soit proportionnelle aux fréquences cumulées croissants de la classe correspondante.

Exemple : Pour l'exemple 2.1.3 [Bovins] l'histogramme des fréquences cumulées croissantes est donné ci-dessous :



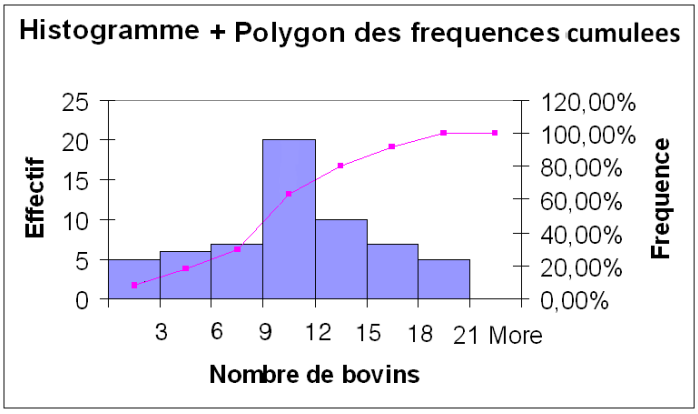
C.4. Courbe cumulative

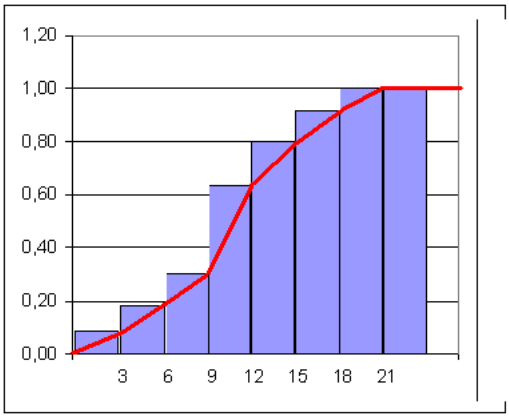
La courbe cumulative (ou polygone F des fréquences cumulées (N des effectifs cumulés)) est obtenue en portant les points dont les abscisses représentent la borne supérieure de chaque classe et les ordonnées des fréquences cumulées (les effectifs cumulés) correspondantes, puis en reliant ces points par des segments de droite. Son équivalent dans la théorie probabiliste est la fonction de répartition.

La courbe cumulative des effectifs est désignée par $N(.)$. Si on fait l'hypothèse que les observations sont réparties *uniformément* au sein de chaque classe, $N(x)$ fournit une approximation du **nombre d'observations** dans la série statistique initiale qui sont **inférieures ou égales à** x.

De manière analogue, la courbe cumulative des fréquences est désignée par $F(.)$. Si on fait l'hypothèse que les observations sont réparties *uniformément* au sein de chaque classe, $F(x)$ fournit une approximation de la **proportion d'observations** dans la série statistique initiale qui sont **inférieures ou égales à** x**.**

Exemple : Pour l'exemple 2.1.3 [Bovins] la courbe cumulative est donnée en rouge :





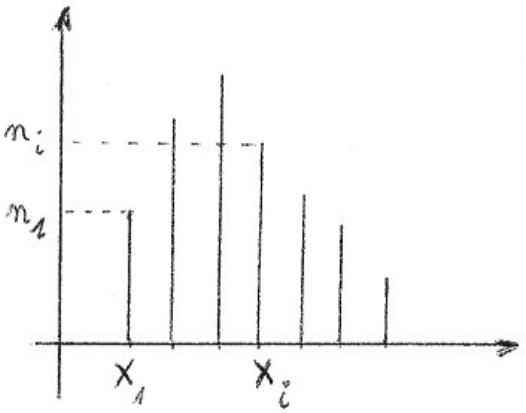
/Tools/Data Analisys/Histogram(input range;bin range)$\surd$ Cumulative Percentage, $\surd$ Chart Output/

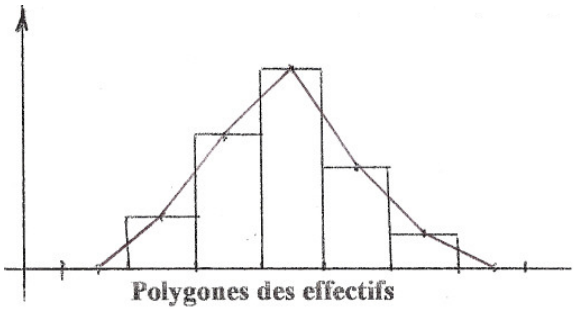
Représentation graphique de paramètres quantitatives :

Série statistique, D.O.1 D.G.1

a/ représentation des effectifs

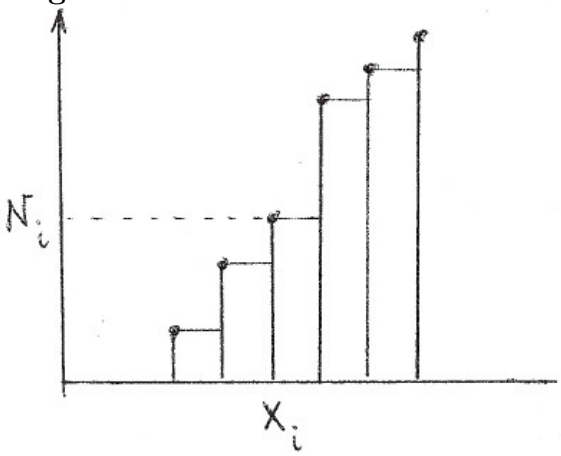
Diagramme en bâtons **Histogramme des effectifs** (même amplitude)

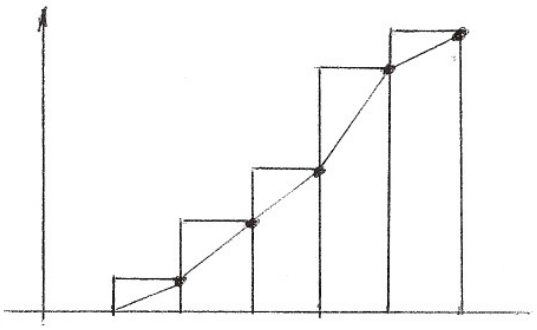




b/ représentation des effectifs cumulés

Diagramme des effectifs cumulés **Histogramme des effectifs cumulés**





Courbes cumulatives

Remarque 6 On peut remplacer dans les graphiques les effectifs par les fréquences.

2.1.5 Liens avec les concepts probabilistes

Les concepts qui viennent d'être présentés sont les homologues de concepts du calcul des probabilités et il est possible de disposer en regard les concepts homologues (voir la table ci-dessous).

Probabilités	Statistique
Espace fondamental	Population
Épreuve	Tirage (d'un individu), expérimentation
Évènement élémentaire	Individu, observation
Variable aléatoire	Caractère
Épreuves répétées	Échantillonnage
Nombre de répétitions d'une épreuve	Taille de l'échantillon, effectif total
Probabilité	Fréquence observée
Loi de probabilité	Distribution observée ou loi empirique
Espérance mathématique	Moyenne observée
Variance	Variance observée

2.2 Synthèse par des paramètres d'une série univariée. Indicateurs numériques.

Les représentations au moyen de tableaux et de graphiques constituent une mise en ordre et donne une possibilité de se faire une idée globale du problème étudié. Elles ne suffisent pas si l'on veut approfondir l'analyse.
On leur associe donc un certain nombre de valeurs caractéristiques, appelées **paramètres** ou valeurs typiques, qui ont pour but de résumer dans une certaine mesure les informations recueillies. Ces paramètres seront eux mêmes à la base de nouvelles représentations graphiques. Ils doivent aussi faciliter la comparaison entre des séries distinctes de même nature.

Caractères qualitatifs nominaux

Il n'existe pas, à part le mode de caractéristiques communément adaptées pour décrire une variable qualitative.

Caractères quantitatifs et caractères qualitatifs ordinaux

Remarque 7 Les paramètres représentent une transition entre la statistique purement descriptive et l'estimation des paramètres qui caractérisent les distributions de probabilité.

Les paramètres les plus courants peuvent être répartis en trois catégories :
les paramètres de *position*, les paramètres de *dispersion*, les paramètres de *forme*.

La détermination de ces paramètres nécessite de nombreux et longs calculs. Il y a donc nécessité d'utiliser la calculatrice, l'ordinateur et de rationaliser les calculs "à la main".

A. Paramètres de position

Caractéristiques de tendance centrale

Ces paramètres ont pour objectif dans le cas d'un caractère quantitatif ou bien caractère qualitatif ordinal, de caractériser **l'ordre de grandeur** des observations. Leurs valeur est comprise entre les valeurs extrêmes de la série.

2.2.1 Mode

Le mode noté x_m ou M_o est la valeur observée qui apparaît le plus souvent.

Le mode est facile à déterminer pour les diagrammes en bâtons. C'est la valeur qui a le plus grand effectif, elle correspond au bâtonnet le plus long.
Exemple : Le mode de la série de l'exemple 2.1.2 [Notes] est la modalité 4. (La modalité 4 a le plus grand effectif 11).

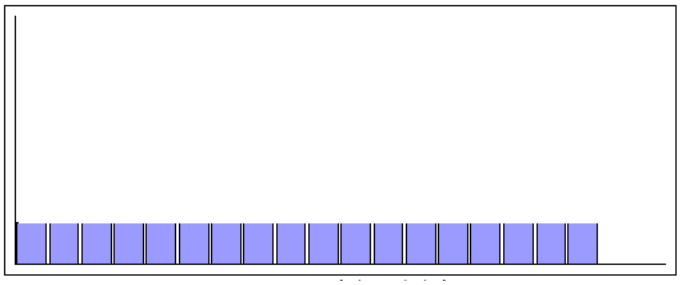
Pour une distribution groupée on parle de *classe modale*. Cette classe dépend souvent du groupement considéré.

Si toutes les classes ont la même longueur : la classe modale est la classe dont l'effectif est le plus élevé.

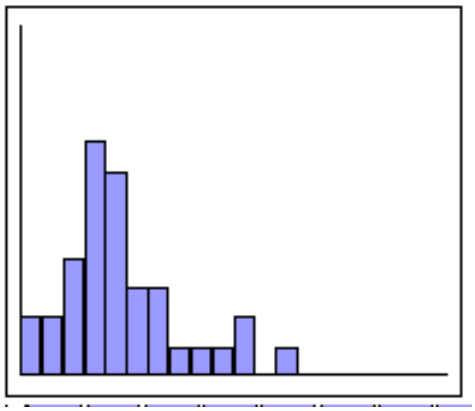
Si les classes n'ont pas toutes la même longueur : la classe modale est la classe dont l'effectif unitaire est le plus élevé.
Exemple 2.1.3 [Bovins] : La classe modale est la classe $[9 : 12)$ d'effectif maximal $n_i = 20$.

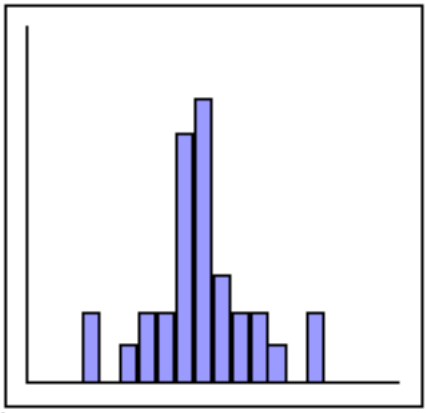
Distribution

sans mode (uniforme)

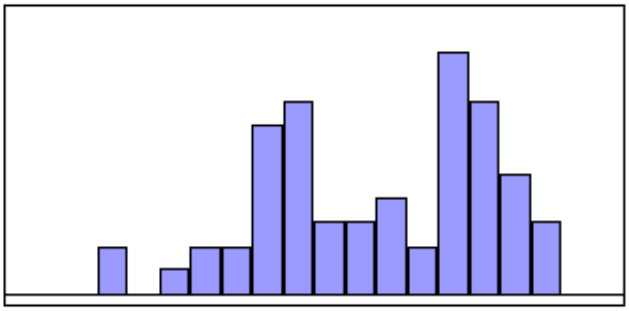


unimodale
(le mode est une valeur centrale ou non)

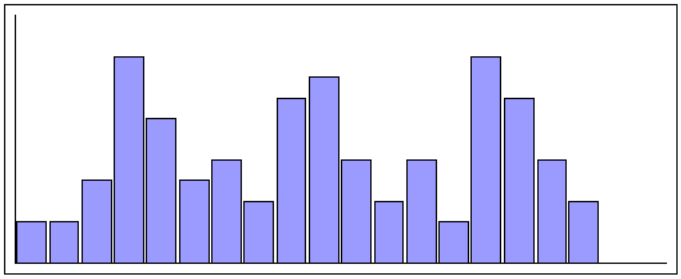




bimodale
(les modes au centr des valeurs observées ou non)



plurimodal



Exemple 2.2.1 /Feuille 4/ [6] : Considérons une étude de prix d'un même article en fonction de la marque qui le commercialise. Ce genre d'étude est fréquent.

Tableau des observations :

Intervalle de prix	Nombres de marques
$[165-170[$	6
$[170-175[$	10
$[175-180[$	5
$[180-185[$	4
$[185-190[$	3
$[190-195[$	2

Tableau de la distribution :

Classe $[a-b[$	Centre de classe $(a+b)/2$	Effectif n_i	Effectif cumulé N_i	Fréquence f_i
$[165-170[$	167,5	6	6	6/30
$[170-175[$	172,5	10	16	10/30
$[175-180[$	177,5	5	21	5/30
$[180-185[$	182,5	4	25	4/30
$[158-190[$	187,5	3	28	3/30
$[190-195[$	192,5	2	30	2/30

La classe modale de la série est la classe $[170-175[$. (La classe $[170-175)$ a le plus grand effectif 10).

On peut identifier le mode comme la valeur médiane de la classe de fréquence maximale ou bien effectuer une interpolation linéaire pour obtenir la valeur exacte du mode comme suit :

$$M_o = x_m + \frac{i\Delta_i}{\Delta_s + \Delta_i} \quad /\text{MODE}(\text{x}_\text{i})/$$

avec
$x_m = a$: limite inférieure de la classe d'effectif maximal
i : intervalle de classe $(x_{m+1}{-}x_m)$
Δ_i : Écart d'effectif entre la classe modale et la classe inférieure la plus proche
Δ_s : Écart d'effectif entre la classe modale et la classe supérieure la plus proche
Exemple : Dans le cas de l'exemple 2.2.1, la valeur du mode est :

- Valeur approchée :

La classe de fréquence maximale est $[170, 175)$ avec $n_i = 10$ d'où $Mo = \frac{170+175}{2} = \frac{345}{2} = 172,5$.

- Valeur exacte :

$$M_o = 170 + \frac{5*4}{5+4} = 170 + \frac{20}{9} = 170 + 2,222 = 172,222 \quad \text{d'où} \quad M_o = 172,222$$

avec $x_m = a = 170$, $\Delta_i = 10 - 6 = 4$, $\Delta_s = 10 - 5 = 5$ et $i = 5$.

Si la distribution des valeurs est symétrique, la valeur du mode est proche de la valeur de la moyenne arithmétique.

$$M_o \approx \bar{x}.$$

Il existe des distributions sans mode, unimodales (le mode est une valeur centrale ou non), bimodales et plurimodales (les modes au centre des valeurs observées ou non).

Remarque 8 Le mode n'est un paramètre de tendance centrale que dans le cas d'une distribution unimodale en forme de cloche ! C'est un paramètre indiquant les valeurs fréquentes.

On peut déterminer le mode d'une distribution observée relative à une variable qualitative nominale. Il correspond alors à la modalité la plus fréquemment observée. Cependant, la notion de centralité n'a pas de sens dans ce cas.

2.2.2 Médiane

La médiane notée M_e d'une série ordonnée est la valeur de la variable telle que l'on ait autant d'éléments qui ont une valeur supérieure ou égale à M_e que d'éléments qui ont une valeur inférieure ou égale à M_e.
La médiane donne l'ordre de grandeur de les observations. Elle a pour avantage d'être peu sensible aux valeurs numériques de la série ; elle ne dépend guère que de l'ordre des observations et est constante même si les premiers et dernier observations varient considérablement. Elle n'est pas toujours facile à calculer, et parfois même n'existe pas.

Si l'on range les éléments par ordre croissant des valeurs prises par la variable, la médiane est la valeur prise par l'élément qui partage l'ensemble de départ en deux sous-ensembles de même effectif. La médiane est la modalité qui sépare la population en deux parties de même effectif.
La médiane est une caractéristique de tendance centrale plus robuste que la moyenne (pas influencée par les valeurs extrêmes) mais elle est influencée par le nombre d'observations.

Remarque 9 : La médiane correspond à la valeur telle que la fréquence cumulée est égale 1/2.

Exemples de médiane :
Médiane d'une série statistique

- **Première convention**

Soit la série statistique $\{x_i; i = 1, \ldots, n\}$, donnant lieu à la série statistique ordonnée $\{x_{(i)}; i = 1, \ldots, n\}$. La médiane $x_{1/2}$ de cette série est l'observation de rang $[n/2]$:
En d'autres termes,

— si n est **pair** : $x_{1/2} = x_{(n/2)}$

— si n est **impair** : $x_{1/2} = x_{((n+1)/2)}$

- **Seconde convention**

Soit la série statistique $\{x_i; i = 1, \ldots, n\}$, donnant lieu à la série statistique ordonnée $\{x_{(i)}; i = 1, \ldots, n\}$.

— Si n est **pair** :

$$x_{1/2} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}.$$

— si n est **impair** :

$$x_{1/2} = x_{((n+1)/2)}.$$

Quelle convention choisir ?
Dans le cas où la taille de la série statistique est impair, les deux conventions conduisent à la même valeur pour la médiane. Dans le cas où la taille de la série est pair, la médiane obtenue selon la première convention est généralement peu différente de la médiane calculée selon la seconde convention. L'utilisateur peut donc opter pour la convention avec laquelle il se sent le plus à l'aise, son choix n'ayant finalement que peu d'impact sur le résultat obtenu.

La seconde convention est plus fréquemment utilisée par les statisticiens (les logiciels statistiques, par exemple, calculent généralement la médiane selon cette seconde convention ; c'est plus souvent cette seconde convention qui est présentée dans les manuels de statistique).

Exemple 2.2.2 /Feuille 4/ : Effectif total n impair.

Modalité x_i	7	8	9	10	12	13	14	15	17	18	Total n
Effectifs n_i	2	1	3	2	1	1	2	2	2	1	17

La série ordonnée est :

$$\underbrace{7;7;8;9;9;9;10;10}_{8;\text{ valeurs}}\ \underbrace{12}_{\text{la médiane}}\ ;\underbrace{13;14;14;15;15;17;17;18}_{8\text{ valeurs}}$$

$$n=17;\quad M_e=x_{((n+1)/2)}=x_{((17+1)/2)}=x_{(9)}=12.$$

La médiane de cette série est la modalité 12. Les 8 valeurs du 1er groupe sont des valeurs inférieures ou égales à 12 et les 8 valeurs du 2ème groupe sont des valeurs supérieures ou égales à 12.

Exemple 2.2.3 /Feuille 4/ :Effectif total n pair. Seconde convention.

Modalité x_i	8	9	10	11	13	14	15	16	17	Total
Effectifs n_i	2	2	3	1	1	1	1	2	1	14

La série ordonnée est :

$$\underbrace{8;8;9;9;10;10;10;}_{7\text{ valeurs}}\underbrace{11;13;14;15;16;16;17}_{7\text{ valeurs}}$$

$$n=14;\quad M_e=\frac{x_{(n/2)}+x_{(n/2+1)}}{2}=\frac{x_{(7)}+x_{(8)}}{2}=\frac{10+11}{2}=10,5.$$

La médiane est $M_1=x_{1/2}=\frac{10+11}{2}=\frac{21}{2}=10,5$.
On accepte pour médiane un nombre qui n'est pas une observation.

Médiane d'une distribution observée univariée (D.O.1)

La détermination de la médiane d'une D.O.1 se fait à partir des effectifs cumulés associés aux valeurs distinctes observées :

- S'il existe une valeur $x_{(j)}$ telle que $N_j=\frac{n}{2}$, alors :

$$\begin{aligned} M_e &= x_{1/2}=x_{(j)} \quad \text{I convention};\\ M_e &= x_{1/2}=\frac{x_{(j)}+x_{(j+1)}}{2} \quad \text{II convention};\end{aligned}$$

Exemple 2.2.3

x_i	8	9	10	11	13	14	15	16	17	Total
n_i	2	2	3	1	1	1	1	2	1	14
N_i	2	4	7	8	9	10	11	13	14	

$n=\frac{14}{2}=7=N_3$

I convention : $M_e = x_{(j)} = x_{(3)} = 10$.
Il convention : $M_e = \frac{x_{(j)}+x_{(j+1)}}{2} = \frac{10+11}{2} = 10,5$

• Si aucun des effectifs cumulés n'est égal à $\frac{n}{2}$, alors la médiane est égale à la plus petite valeur observée dont l'effectif cumulé est $> \frac{n}{2}$; en d'autres termes, la médiane est égale à la valeur $x_{(j)}$ telle que

$$N_{j-1} < \frac{n}{2} \leq N_j.$$

Ainsi, la médiane est la plus petite valeur observée dont l'effectif cumulé est **supérieur ou égal** à $n/2$.

Exemple 2.2.2

Modalité x_i	7	8	9	10	12	13	14	15	17	18	Total n
Effectifs n_i	2	1	3	2	1	1	2	2	2	1	17
Effectifs cumulés N_i	2	3	6	8	9	10	12	14	16	17	

$$n = 17; \quad \frac{n}{2} = \frac{17}{2} = 8,5$$

$$N_4 = 8 < \frac{n}{2} = 8,5 < N_5 = 9 \Longrightarrow M_e = x_{1/2} = x_{(5)} = 12.$$

Médiane d'une distribution groupée univariée (D.G.1)

Dans le cas où on a une D.G.1, mais qu'on ne dispose plus de la série statistique qui a permis sa construction, on ne peut plus déterminer la médiane de manière exacte. On peut cependant obtenir une valeur approchée de la médiane.

La détermination de la valeur de $x_{1/2}$ se fait en deux étapes :
1. On commence par déterminer la classe dans laquelle se trouve $x_{1/2}$;
2. On détermine ensuite la valeur de $x_{1/2}$ dans cette classe.

Détermination de la valeur (approchée) de la médiane (Etape 1)

La première étape consiste à déterminer la classe contenant $x_{1/2}$. Cette classe correspond à la classe telle que $C_j = [a_j; b_j[$

$$N_{j-1} < \frac{n}{2} \leq N_j$$

C_j est la première classe dont l'effectif cumulé est supérieur ou égal à $n/2$, ou encore, de manière équivalente, telle que

$$F_{j-1} < \frac{1}{2} \leq N_j$$

C_j est la première classe dont la fréquence cumulée est supérieure ou égale à $1/2$.

Détermination de la valeur (approchée) de la médiane (Etape 2)

Après avoir déterminé la classe contenant le $n^e/2$ individu de l'échantillon, en supposant que tous les individus de cette classe sont uniformément répartis à l'intérieur, la position exacte du $n^e/2$ individu est obtenue de la façon suivante par **interpolation linéaire :**

$$M_e = x_m + (x_{m+1} - x_m)\left(\frac{\frac{n}{2} - N_i}{n_i}\right), \quad /\text{MEDIAN}(\text{x}_\text{i})/$$

avec
$x_m = a$: limite inférieure de la classe dans laquelle se trouve le $n^e/2$ individu (classe médiane).
$x_{m+1} = b$: limite supérieure de la classe dans laquelle se trouve le $n^e/2$ individu (classe médiane).
n_i : effectif de la classe médiane
N_i : effectif cumulé inférieur à x_m
n : taille de l'échantillon

Exemple : Cas de D.G.1 - exemple 2.2.1

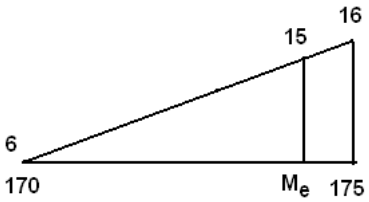
$n = 30$, la $n/2 = 30/2 = 15^{\text{ème}}$ valeur se situe dans la classe $[170, 175)$ qui contient les individus de 7 à 16. D'ici avec $x_m = 170$, $x_{m+1} = 175$,
$x_{m+1} - x_m = 175 - 170 = 5$, $N_i = 6$, $n_i = 10$

$$M_e = 170 + (175 - 170)\left(\frac{15-6}{10}\right) = 170 + 5\frac{9}{10} = 170 + 4,5 = 174,5$$

D'où la Médiane $M_e = 174,5$.

Classe $[a - b[$	Effectif cumulé N_i
$[165 - 170[$	6
$[170 - 175[$	16
$[175 - 180[$	21
$[180 - 185[$	25
$[158 - 190[$	28
$[190 - 195[$	30

On peut utiliser la règle des triangles semblables. Du polygone des effectifs pour la classe modale $[170, 175[$ on a :



D'après la règle des triangles semblables on peut écrire :

$$\frac{M_e - 170}{175 - 170} = \frac{15 - 6}{16 - 6}, \quad M_e = 170 + 5\frac{9}{10} = 170 + 4,5 = 174,5.$$

Remarque 10 : La médiane ne s'applique qu'aux échelles ordinales, d'intervalles et de rapport, car elle nécessite un ordre linéaire entre les variables.
Si la distribution des valeurs est symétrique, la valeur de la médiane est proche de la valeur de la moyenne arithmétique.

$$M_e \approx \bar{x}.$$

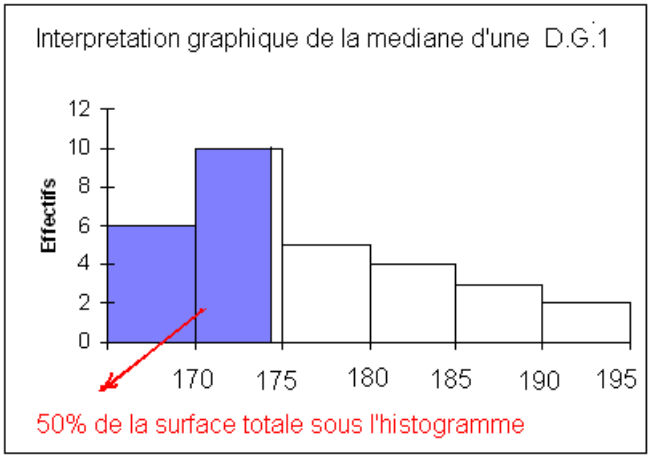
La médiane peut être considérée comme un cas particulier d'une valeur plus générale appelée quantile ou *fractile*.

Interprétation graphique de la médiane d'une D.G.1.

La surface totale sous l'histogramme des effectifs est égale à n, le nombre total d'observations. On a, à gauche de la médiane $x_{1/2} = M_e$ exactement 50% de la surface totale de l'histogramme, soit une surface égale à $n/2 = 30/2 = 15$.

De manière similaire, la surface totale sous l'histogramme des fréquences est égale à 1 et la surface sous cet histogramme à gauche de la médiane $x_{1/2}$ est exactement égale à 0.5 (50%).

Exemple : Pour l'exemple 2.2.1 l'interprétation graphique de la médiane nous donne :



2.2.3 Moyennes

La plus couramment utilisée, très facile à calculer et possède d'importantes propriétés théoriques, par ailleurs assez faciles à établir. La moyenne, toutefois, possède l'inconvénient d'être très sensible au retrait ou à l'ajout d'une observation "aberrante" : on dit que c'est une statistique peu *robuste*.

2.2.1.1 Moyenne arithmétique $\bar{x}$

Série statistique observée : on calcule la moyenne arithmétique de la série en additionnant toutes les modalités et en divisant le résultat obtenu par l'effectif total.

Si les nombres sont connus isolément : $x_1, x_2, x_3, ..., x_i, ..., x_n$, la moyenne arithmétique de ces n nombres est le nombre défini par :

$$\bar{x} = (x_1 + x_2 + x_3 + ... + x_n).\frac{1}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Distribution observée univariée (D.O.1) : on calcule la somme des produits (modalité x effectif partiel correspondant) puis on divise le résultat obtenu par l'effectif total.

Si les nombres sont affectés d'un poids (effectif) et si les données observées x_i sont **données par k modalités d'effectif** n_i (caractère discret, dont le nombre de modalités k est moins que 12) :

x_1	x_2	...	x_i	...	x_k
n_1	n_2	...	n_i	...	n_k

,
il faut les pondérer par les effectifs correspondants. La moyenne arithmétique est le nombre défini par

$$\bar{x} = (n_1x_1 + n_2x_2 + ... + n_ix_i + ... + n_kx_k).\frac{1}{n} = \frac{1}{n}\sum_{i=1}^{k} n_ix_i = \sum_{i=1}^{k} f_ix_i$$

où $n = n_1 + n_2 + \dots n_k$.

Distribution groupée univariée (D.G.1) : on calcule la somme des produits (centre de la classe x effectif partiel correspondant) puis on divise le résultat obtenu par l'effectif total.

$$\bar{x} = (n_1 x_1^* + n_2 x_2^* + ... + n_i x_i^* + ... + n_k x_k^*).\frac{1}{n} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^* = \sum_{i=1}^{k} f_i x_i^*$$

où $n = n_1 + n_2 + \dots n_k$ et x_i^* est le centre de la classe i.

Description	Moyenne	Excel
Statistique observée	$\bar{x} = E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$	AVERAGE(x_i)
D.O.1, k modalités	$\bar{x} = E(X) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i$	$\frac{\text{SUMPRODUCT}(n_i, x_i)}{\text{SUM}(n_i)}$
D.G.1, k classes	$\bar{x} = E(X) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^* = \sum_{i=1}^{k} f_i x_i^*$	$\frac{\text{SUMPRODUCT}(n_i, x_i^*)}{\text{SUM}(n_i)}$

TABLE 2.1 : Moyenne [6]

- **Signification statistique de la moyenne arithmétique :**

la moyenne arithmétique d'une série statistique observée est la valeur associée à chaque observation si la somme totale des valeurs était partagée également entre toutes les observations : $\sum n_i x_i = n\bar{x}$.

Remarques :
— la moyenne arithmétique est toujours calculable pour des observations numériques. Elle est unique, mais ne représente pas nécessairement une valeur réellement observée ni même observable.
— La moyenne arithmétique dépend de toutes les observations et est très sensible aux valeurs extrêmes. La stabilité est bonne lorsque n est grand.
— On considère parfois la *moyenne tronquée* : $\frac{x_2+x_3+...+x_{n-1}}{n-2}$

- **Propriétés de la moyenne arithmétique**

1. La somme algébrique des écarts d'un ensemble de nombres à leur moyenne arithmétique est nulle.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

2. La somme des carrés des écarts d'un ensemble de nombres à un nombre donné a est minimale ssi ce nombre est la moyenne arithmétique de ces nombres.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - a)^2, \text{ pour } a \neq \bar{x}.$$

3. Si on "réunit" deux ensembles E_1 et E_2 d'observations, le premier d'effectif n_1 et de moyenne arithmétique $\bar{x}_1$, le second d'effectif n_2 et de moyenne arithmétique $\bar{x}_2$, la moyenne arithmétique $\bar{x}$ de la série E obtenue d'effectif $n = n_1 + n_2$ s'exprime à partir des paramètres de E_1 et E_2 par la relation suivante :

$$\bar{x} = (n_1\bar{x}_1 + n_2\bar{x}_2)/(n_1 + n_2).$$

4. Influence d'un changement d'origine ou d'unité sur la moyenne arithmétique :
La moyenne arithmétique $\bar{y}$ des nombres y_i obtenus pour tout i par la relation $y_i = (x_i - x_0)/d$ et affectés des mêmes effectifs est donnée par la relation

$$\bar{y} = (\bar{x} - x_0)/d.$$

- **Exemples de calcul de la moyenne arithmétique**

— **Calcul de la moyenne d'une D.O.1**. Il est commode d'utiliser le tableau suivant :

x_i	n_i	$x_i n_i$
x_1	n_1	$x_1 n_1$
x_2	n_2	$x_2 n_2$
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$x_k n_k$
	$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} x_i n_i$

Exemple : Calcul de la moyenne arithmétique pour l'exemple 2.1.2 (Notes) :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i = \frac{2.3 + 3.8 + 4.11 + 5.5 + 6.3}{3 + 8 + 11 + 5 + 3} = \frac{117}{30} \approx 3,9.$$

Si dans le tableau des notes des étudiants on ajoute une colonne des produits $x_i n_i$, on obtient le tableau :

x_i	n_i	$x_i n_i$
2	3	6
3	8	24
4	11	44
5	5	25
6	3	18
	$\sum_{i=1}^{5} n_i = 30$	$\sum_{i=1}^{5} x_i n_i = 117$

et on obtient $\bar{x} = \frac{117}{30} \approx 3,9$.

— **Calcul de la moyenne d'une D.G.1**

Si la série statistique est groupée en classes, dans l'expression $\bar{x} = \frac{\sum_{i=1}^{k} x_i^* n_i}{\sum_{i=1}^{k} n_i}$ les valeurs de x_i^* sont les centres des classes. Le tableau pour calculer la moyenne d'une série en classe est le suivant :

Classes	Effectifs	Centres	Produits
$x_i; x_{i+1}$	n_i	x_i^*	$x_i^* n_i$
$x_1; x_2$	n_1	x_1^*	$x_1^* n_1$
$x_2; x_3$	n_2	x_2^*	$x_2^* n_2$
$\vdots$	$\vdots$	$\vdots$	
$x_k; x_{k+1}$	n_k	x_k^*	$x_k^* n_k$
	$\sum_{i=1}^{k} n_i$		$\sum_{i=1}^{k} x_i^* n_i$

Exemple : Calcul de la moyenne arithmétique pour l'exemple 2.1.3 (Bovins) :
Le tableau pour calculer la moyenne arithmétique est le suivant :

Classes	Effectifs	Centres	Produits
$x_i; x_{i+1}$	n_i	x_i^*	$x_i^* n_i$
moins de 3	5	1,5	7,5
3 ;6	6	4,5	27,0
6 ;9	7	7,5	52,5
9 ;12	20	10,5	210,0
12 ;15	10	13,5	135,0
15 ;18	7	16,5	115,5
18 ;21	5	19,5	97,5
	$\sum_{i=1}^{k} n_i = 60$		$\sum_{i=1}^{k} x_i^* n_i = 645,0$

On obtient pour la moyenne arithmétique

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i^* = \sum_{i=1}^{k} f_i x_i^* = \frac{645}{60} \approx 10,75 \text{ c.a.d. à peu près 11 bovins.}$$

La moyenne arithmétique peut être calculée directement des valeurs du caractère observé. Le rapport de la somme des valeurs observées et leur nombre est $\bar{x} = 11,17$. évidement cette valeur diffère la moyenne arithmétique calculée de la série groupée en classes. Cela est due au fait que les centres des classes sont calculés à la base de l'hypothèse que la distribution des valeurs dans chaque classe soit uniformément.

Remarque 11 La moyenne $\bar{x}$ est fort sensible à la présence de valeurs extrêmes (valeurs aberrantes), c'est-à-dire de valeurs nettement plus grandes ou nettement plus petites que les autres observations de la série.
Exemple :
$\{1, 1, 2, 2, 2, 4, 4, 4, 7\} \rightarrow \bar{x} = 3$: la moyenne est bien une valeur centrale.
$\{1, 1, 2, 2, 2, 4, 4, 4, 7, 77\} \rightarrow \bar{x} = 10,78$: la moyenne n'est plus du tout une valeur centrale puisqu'elle est supérieure à toutes les observations, sauf la dernière.
Une valeur extrême (ou aberrante) « attire à elle » la moyenne. Ceci a pour conséquence que la moyenne ne correspond plus alors à une valeur centrale de la série.

- **Usage et limite de la moyenne arithmétique**

Dans la pratique la moyenne arithmétique est souvent utilisée pour comparer des situations différentes. Mais l'utilisation de la moyenne arithmétique est-elle toujours justifiée ?
Exemple : Si l'on décrète que les locaux universitaires sont correctement chauffés si globalement la température est de 19 degrés, pourrait-on donner cours dans une salle où règne une température de 12 degrés car dans une autre salle il y a 26 degrés ?
Exemple : Dans une société de 7 personnes les salaires mensuels sont respectivement de

$$1000; \quad 1200; \quad 1300; \quad 1500; \quad 1700; \quad 1800 \quad \text{et} \quad 12500\,€.$$

Peut-on dire que la moyenne arithmétique est représentative de ces salaires ?

2.2.1.2 Moyenne géométrique.

Lorsque les valeurs d'une série statistique varient en gros selon une progression géométrique, il est préférable de substituer à la moyenne arithmétique la moyenne géométrique $\bar{x}_g$. On utilise la moyenne géométrique pour calculer la moyenne d'une série de nombres relatifs, positifs, par exemple pour calculer la moyenne de la variation en pourcentage de la valeur d'une variable.

Soit $x_{1,2}, ..., x_n$ n nombres réels strictement positifs. La moyenne géométrique est défini par :

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i} \quad \text{pour des séries statistiques}$$

$$\bar{x}_g = \sqrt[\sum_{i=1}^{k} n_i]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}} \quad \text{pour des données regroupées en } k \text{ modalités (D.O.1)}$$

$$\bar{x}_g = \sqrt[\sum_{i=1}^{k} n_i]{(x_1^*)^{n_1} (x_2^*)^{n_2} \dots (x_k^*)^{n_k}} \quad \text{pour des données regroupées en } k \text{ classes (D.G.1)}$$

Lors d'absence d'un informatique approprié, on utilise les expressions :

$$\log \bar{x}_g = \frac{1}{n}(\log x_1 + \log x_2 + \ldots + \log x_n), \quad \text{série statistique}$$

$$\log \bar{x}_g = \frac{1}{\sum_{i=1}^{k} n_i}(n_1 \log x_1 + n_2 \log x_2 + \ldots + n_k \log x_k), \quad \text{données regroupées en } k \text{ modalités}$$

$$\log \bar{x}_g = \frac{1}{\sum_{i=1}^{k} n_i}(n_1 \log x_1^* + n_2 \log x_2^* + \ldots + n_k \log x_k^*), \quad \text{données regroupées en } k\text{classes.}$$

La moyenne géométrique est l'antilogarithme du résultat.

Exemple 2.2.4 /Feuille 4/ :Les affaires mensuels d'une compagnie ont augmentés pendant les premier 5 mois de l'année de 12% par mois et pendant les 7 suivants mois de 8%. Calculer la moyenne des croissance des ventes mensuelles i_m.
Le taux de croissance pendant les premiers 5 mois est de 1,12 et pour chaque des mois suivants de 1,08.
Dans ce schémas pour 100 €, on obtient : janvier 100* 1,12 = 112 € ; février : $112 * 1,12 = 100 * 1,12 * 1,12 = 100 * 1,12^2 = 125,44$; mars ...
Le taux annulaire de croissance des vantes est

$$(1 + i_m)^{12} = (1 + 0,12)^5 * (1 + 0,08)^7 = 1,12^5 * 1,08^7 = 3,020343947.$$

Le taux mensuel de croissance des vantes est :

$$1 + i_m = \sqrt[12]{(1 + 0,12)^5 * (1 + 0,08)^7} = \sqrt[12]{1,12^5 * 1,08^7} \approx 1,0965.$$

$$/\exp(\frac{5}{12} * \ln(1,12) + \frac{7}{12} * \ln(1,08)) = 1,09649/$$

Alors la moyenne mensuelle de la croissance des vantes est $i_m = 0,0965$ ou de 9,65%.

Si on utilise la moyenne arithmétique, on obtient $\frac{0,12*5+0,08*7}{12} \approx 0,0967$, qui n'est pas le vrai résultat. Dans cet exemple on utilise la moyenne géométrique.

On peut donc dire que le calcul de la moyenne géométrique est fondé pour les taux (les quantités $(1 + croissance)$), et non pour les croissances elles-mêmes.

On peut généraliser le résultat.
Considérons n périodes égales pendant lesquelles un capital a été placé successivement aux croissances $i_1, i_2, \ldots, i_n$.
Le taux moyen est donné par la relation :

$$1 + i_m = \sqrt[n]{(1+i_1).(1+i_2).\ldots.(1+i_n)}.$$

2.2.1.3 Moyenne harmonique.

La moyenne harmonique est l'inverse de la moyenne arithmétique des inverses. On l'utilise lorsque les valeurs de la variable sont en rapport réciproque de la règle moyennant. Par exemple le calcul de la moyenne de la productivité du travail par les données de consommation de temps.

Soit $x_1, x_2, \ldots, x_n$ n nombres réels non nuls. La moyenne harmonique est définie par :

$$\frac{1}{\bar{x}_h} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i} \quad \Longrightarrow \bar{x}_h = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}, \text{ série statistique}$$

$$\bar{x}_h = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k}\frac{n_i}{x_i}} \quad \text{lors de données regroupées en } k \text{ modalités}$$

$$\bar{x}_h = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k}\frac{n_i}{x_i^*}} \quad \text{lors de données regroupées en } k \text{ classes.}$$

Exemple 2.2.5 /Feuille 4/ : Une voiture se déplace à mi-chemin avec une vitesse de $v_1 = 60km/h$, tandis que l'autre moitié - de vitesse de $v_2 = 70km/h$. Calculer la vitesse moyenne v_m de la voiture.
La vitesse moyenne est le rapport entre la distance totale et le temps total. On obtient :

$$v_m = \frac{S}{\frac{S/2}{v_1} + \frac{S/2}{v_2}}$$

d'où

$$v_m = \frac{1}{\frac{1/2}{v_1} + \frac{1/2}{v_2}} = \frac{2v_1v_2}{v_1 + v_2}.$$

$$v_m = \frac{2 * 60 * 70}{60 + 70} \approx 64,61km/h.$$

Il faut utiliser la moyenne harmonique lorsque la variable considérée intervient par l'intermédiaire de son *inverse*. Dans le cas de notre exemple la vitesse fait essentiellement intervenir le quotient :

$$\frac{\text{espace parcouru}}{\text{temps}}.$$

Lors de données positives la validité des inégalités suivants est prouvée :

$$\bar{x}_h \leq \bar{x}_g \leq \bar{x}.$$

Insistons sur le fait que le choix de la moyenne à utiliser est fonction du problème particulier à étudier et que la logique seule permet de déterminer *quelle* est la moyenne qui convient. Tous les paramètres présentés s'expriment dans les mêmes unités et par rapport à la même origine que les valeurs analysées.

2.2.4 Comparaison des mesures de tendance centrale

	Avantages	Inconvénients
Moyenne arithmétique	- Facile à calculer, - Répond au principe des moindres carrés.	- Fortement influencée par les valeurs extrêmes de la v.a., - Représente mal une population hétérogène (plurimodale).
Médiane	- Pas influencée par les valeurs extrêmes de la v.a., - Peu sensible aux variations d'amplitude des classes, - Calculable sur des caractères cycliques (saison, etc.) où la moyenne a peu de signification.	- Se prête mal aux calculs statistiques, - Suppose l'équi-répartition des données - Ne représente que la valeur qui sépare l'échantillon en 2 parties égales.
Mode	- Pas influencée par les valeurs extrêmes de la v.a., - Calculable sur des caractères cycliques (saison, etc.) où la moyenne a peu de signification, - Bon indicateur de population hétérogène.	- Se prête mal aux calculs statistiques, - Très sensible aux variations d'amplitude des classes, - Son calcul ne tient compte que des individus dont les valeurs se rapprochent de la classe modale

Le choix de l'indicateur convenable pour résumer une série et en donner une tendance centrale dépend de la "forme" générale de la série statistique étudiée : pluri-modale ; symétrique ou asymétrique.

Dans le cas d'une série pluri-modale la moyenne et la médiane ne représentent pas la série et les modes sont les seuls indicateurs de position informatifs.

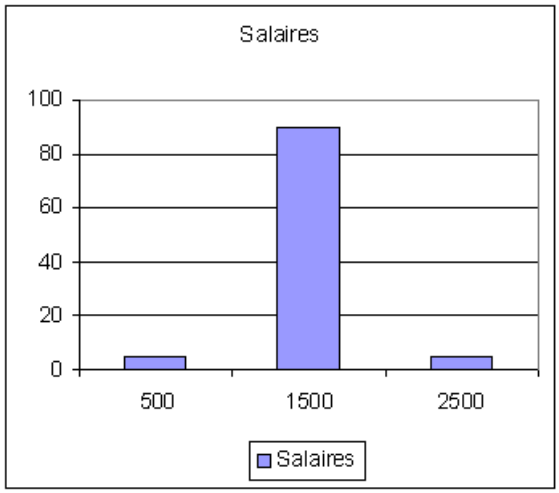
Soit la D.G.1 de forme pluri-modale :

Salaires x_i en €	x_i^*	n_i
[0,4000[	2000	45
[4000, 28000[	16000	10
[28000, 32000[	30000	45

de moyenne $\bar{x} = 16000$€, médiane $Me = 16000$€ et deux classes modales : [0, 4000[et [28000, 32000[. L'indicateur informatif sont les modes.

Soit la D.G.1 de forme symétrique :

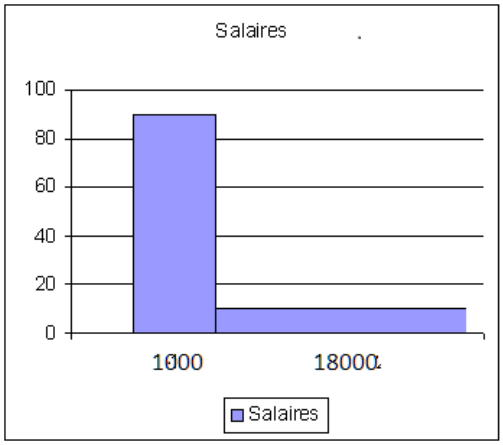
Salaires x_i en €	x_i^*	n_i
[0,1000[	500	5
[1000, 2000[	1500	90
[2000, 3000[	2500	5



de moyenne $\bar{x} = 1500$€, médiane $Me = 1500$€ et mode dans la classe modale : [1000, 2000[. Les trois indicateurs peuvent être utilisés, mais la moyenne est la préférable à cause de ses avantages - le calcul facile.

Soit la D.G.1 de forme asymétrique :

Salaires x_i en €	x_i^*	n_i
[0,2000[	1000	90
[2000, 38000[	18000	10



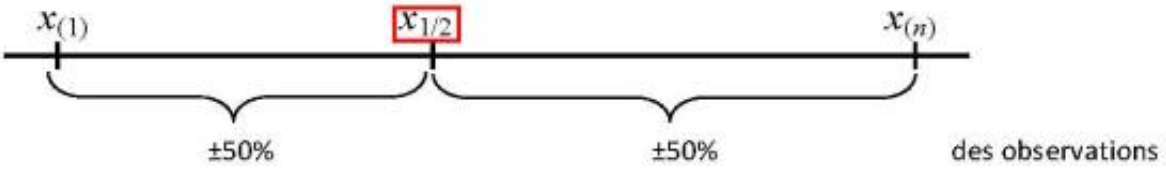
de moyenne $\bar{x} = 2900$€, médiane $Me = 1100$€ et mode dans la classe modale : [1000, 2000[. La moyenne est trop influencée par les gros salaires, c'est pourquoi l'indicateur le plus adaptée est la médiane.

2.2.5 Quantiles

On suppose que l'on dispose d'une série ordonnée. Soit p un nombre compris entre 0 et 1.

On appelle **quantile d'ordre** p la valeur x_p de la variable telle qu'il y ait au moins une proportion p des observations inférieures ou égales à x_p et la proportion complémentaire $(1-p)$ de valeurs qui sont supérieures ou égales à x_p. Avec cette notation $M_e = x_{1/2}$.

Médiane $p = 1/2$



La médiane partage la série statistique ordonnée en deux sous-ensembles qui contiennent chacun (environ) la moitié des observations.

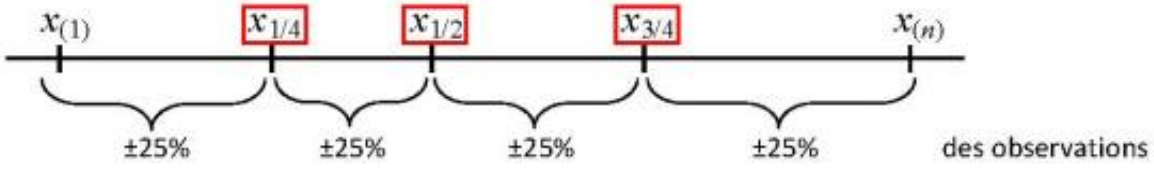
Quartiles $p = 1/4 \quad p = 1/2 \quad p = 3/4$

/QUARTILE(x_i,j), $j = 1, 2, 3$/

/Find the kth smallest member in the array of values, where

$k = p(n-1) + 1 \approx k_{(entier)} + f_{(fraction)}$

$x_p = x_{(k)} + f(x_{(k+1)} - x_{(k)})$/



Les 3 quartiles partagent la série statistique ordonnée en 4 sous-ensembles qui contiennent chacun (environ) un quart (25%) des observations.

Déciles $p = 1/10 \quad p = 2/10 \quad \ldots$, corresponds aux 9 fractiles $x_{0,1}$, $x_{0,2}$, …, $x_{0,9}$

Les 9 déciles partagent la série statistique ordonnée en 10 sous-ensembles qui contiennent chacun (environ) un dixième (10%) des observations.

Percentiles $p = 1/100 \quad p = 2/100 \quad \ldots$, corresponds aux 99 fractiles $x_{0,01}$, $x_{0,02}$, …, $x_{0,99}$

/PERCENTILE(x_i,j), $j = 0,01; 0,02; \ldots 0,99$/

Soit n l'effectif total.
Le 1er quartile Q_1 est l'abscisse du point qui a pour ordonnée $\frac{n}{4}$.
Le 2ème quartile Q_2 est l'abscisse du point qui a pour ordonnée $\frac{n}{2}$: c'est la médiane.
Le 3ème quartile Q_3 est l'abscisse du point qui a pour ordonnée $\frac{3n}{4}$.

Les quantiles d'une série statistique

Définition 12 Soit la série statistique $\{x_i; i = 1, \ldots, n\}$, donnant lieu à la série statistique ordonnée $\{x_{(j)}; j = 1, \ldots, n\}$. Considérons la proportion p $(0 < p < 1)$

- Si np n'est pas un nombre entier : $x_p = x_{(\lceil np \rceil)}$.

Le quantile x_p d'ordre p correspond à l'observation de rang $\lceil np \rceil$, où $\lceil np \rceil$ désigne le plus petit entier supérieur ou égal à np ($\lceil np \rceil$ est la valeur obtenue en arrondissant np à l'entier supérieur).

- Si np est un nombre entier :
  - Première convention : $x_p = x_{(np)}$
  - Seconde convention : $x_p = \frac{x_{(np)} + x_{(np+1)}}{2}$.

Exemple : Considérons la série statistique suivante $(n = 8)$:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	7	15	8	3	9	8	5

Recherchons les quantiles d'ordre 1/2, d'ordre 1/4 et d'ordre 1/5 :

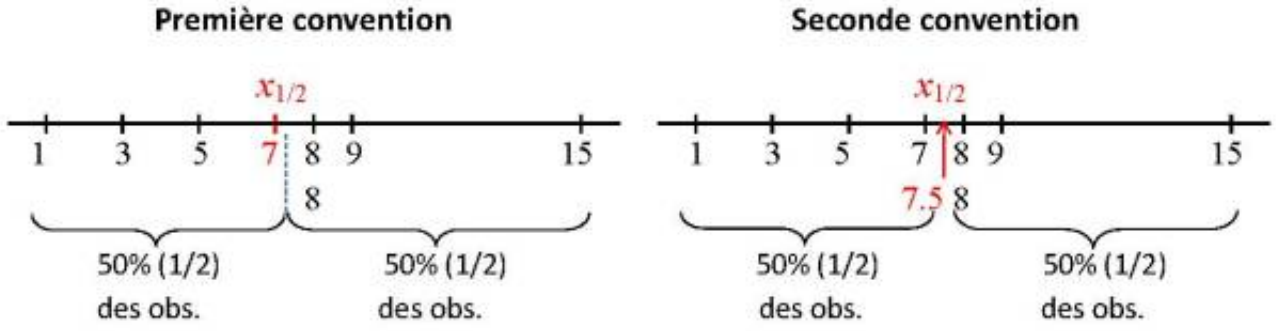
$$x_{1/2} = ? \quad x_{1/4} = ? \quad x_{1/5} = ?$$

La toute première étape consiste à ranger les observations par ordre croissant de manière à obtenir la série statistique ordonnée :

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$
1	3	5	7	8	8	9	15

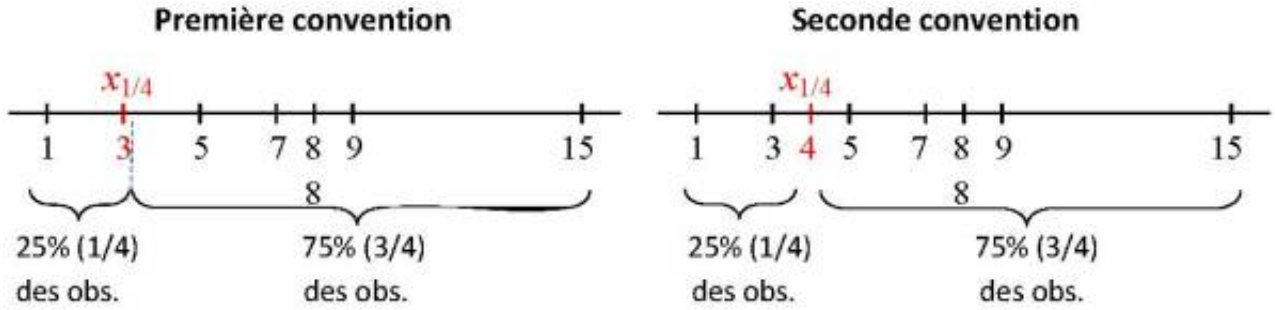
a) Quantile d'ordre 1/2 :

$$np = \frac{8}{2} = 4 \longrightarrow x_{1/2} = \begin{cases} x_{(4)} = 7 & \text{(première convention)} \\ \frac{x_{(4)} + x_{(5)}}{2} = \frac{7+8}{2} = 7,5 & \text{(seconde convention)} \end{cases}$$



b) Quantile d'ordre 1/4 :

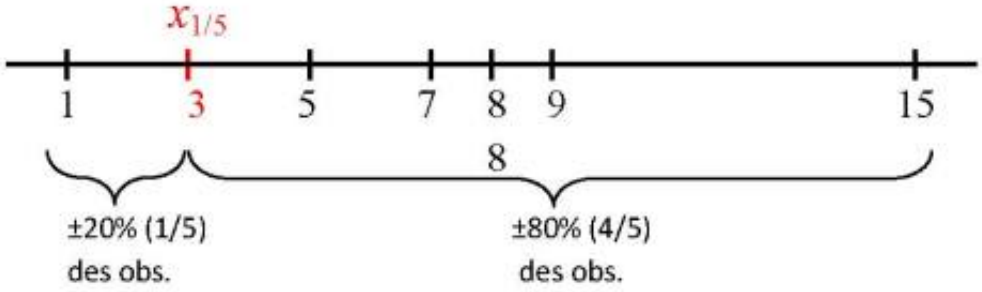
$$np = \frac{8}{4} = 2 \longrightarrow x_{1/4} = \begin{cases} x_{(2)} = 3 & \text{(première convention)} \\ \frac{x_{(2)} + x_{(3)}}{2} = \frac{3+5}{2} = 4 & \text{(seconde convention)} \end{cases}$$



Le quantile d'ordre 1/4 partage la série statistique ordonnée en deux sous-ensembles qui contiennent respectivement 1/4 et 3/4 des observations.

c) Quantile d'ordre 1/5 :

$$np = \frac{8}{5} = 1,6 \longrightarrow x_{1/5} = x_{\lceil 1,6 \rceil} = x_{(2)} = 3.$$



Le quantile d'ordre 1/5 partage la série statistique ordonnée en deux sous-ensembles qui contiennent respectivement à peu près 1/5 et des 4/5 observations.

Les quantiles d'une D.O.1

La détermination du quantile x_p d'ordre p $(0 < p < 1)$ d'une D.O.1 se fait à partir des effectifs (ou fréquences) cumulé(e)s associé(e)s aux valeurs distinctes observées.

• Si aucun des effectifs cumulés n'est égal à np – ou si aucune des fréquences cumulées n'est égale à p – alors le quantile x_p est égal à la valeur $x_{(j)}$ telle que

$$N_{j-1} < np < N_j$$

c'est-à-dire telle que

$$F_{j-1} < p < F_j$$

• S'il existe une valeur $x_{(j)}$ telle que $N_j = np$ $(F_j = p)$, alors

— Première convention : $x_p = x_{(j)}$

— Seconde convention : $x_p = \frac{x_{(j)} + x_{(j+1)}}{2}$

Exemple : Obtenir les quartiles Q_1 et Q_3 de la série statistique de l'exemple 2.1.2 [Notes].

Note	Effectif cumulée N_i
2	3
3	11
4	22
5	27
6	30

$$\begin{aligned} Q_1 = Q_{1/4} = x_{1/4}: \quad & n = 30;\ \ np = 30\frac{1}{4} = 7,5 \Longrightarrow \bar{\exists} N_i = np = 7,5 \\ & \text{comme } N_1 = 3 < np = 7,5 < N_2 = 11 \\ & \longrightarrow x_{1/4} = Q_1 = x_{(2)} = 3. \\ Q_3 = Q_{3/4} = x_{3/4} \quad & n = 30;\ \ np = 30\frac{3}{4} = 22,5 \Longrightarrow \bar{\exists} N_i = np = 22,5 \\ & \text{comme } N_3 = 22 < np = 22,5 < N_4 = 27 \\ & \longrightarrow x_{3/4} = Q_3 = x_{(4)} = 5. \end{aligned}$$

Les quantiles d'une D.G.1

Dans le cas où on a une D.G.1, mais qu'on ne dispose plus de la série statistique qui a permis sa construction, on ne peut plus déterminer les quantiles de manière exacte. On peut cependant obtenir des valeurs approchées à partir de la courbe cumulative des effectifs ou des fréquences.

On prend comme valeur approchée du quantile d'ordre p $(0 < p < 1)$ la valeur x_p telle que $N(x_p) = np$ ou $F(x_p) = p$.

Deux étapes sont nécessaires pour déterminer la valeur de x_p :

Étape 1 : détermination de la classe contenant x_p

La classe contenant x_p est la classe $C_j = [a_j; b_j]$ si et seulement si (en termes d'effectifs cumulés) C_j est la première classe dont l'effectif cumulé est supérieur ou égal à np :

$$N_{j-1} < np \leq N_j$$

c.-à-d. (en termes de fréquences cumulées)
C_j est la première classe dont la fréquence cumulée est supérieure ou égale à p :

$$F_{j-1} < p \leq F_j.$$

Étape 2 : Détermination de la valeur approchée du quantile dans la classe C_j

- A partir de la courbe cumulative des effectifs :

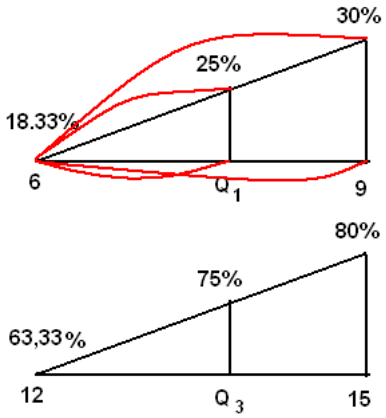
$$x_p = a_j + \frac{np - N_{j-1}}{N_j - N_{j-1}}(b_j - a_j)$$

- A partir de la courbe cumulative des fréquences :

$$x_p = a_j + \frac{p - F_{j-1}}{F_j - F_{j-1}}(b_j - a_j).$$

Exemple : Obtenir les quartiles Q_1 et Q_3 de la série statistique de l'exemple 2.1.3 [Bovins].

Nombre de bovins classe $[a : b]$	Fréquence cumulée $\sum_{j=1}^{i} f_j$	Fréquence cumulée en pourcentages
$[0; 3[$	0,08	8%
$[3; 6[$	0,1833	18,33%
$[6; 9[$	0,3	30%
$[9; 12[$	0,6333	63,33%
$[12; 15[$	0,8	80 %
$[15; 18[$	0,916	91,6 %
$[18; 21[$	1	100 %



$Q_1 = Q_{1/4} = Q_{25}$ est dans l'intervalle $[6; 9)$. D'après la règle des triangles semblables on a :

$$\frac{Q_1 - 6}{9 - 6} = \frac{25 - 18,33}{30 - 18,33}; \quad Q_1 = 6 + 3\frac{6,67}{11,67}; \quad Q_1 = 7,71.$$

$Q_3 = Q_{3/4} = Q_{75}$ est dans l'intervalle $[12; 15)$. D'après la règle des triangles semblables on a :

$$\frac{Q_3 - 12}{15 - 12} = \frac{75 - 63,33}{80 - 63,33}; \quad Q_3 = 12 + 3\frac{11,67}{16,67}; \quad Q_3 = 14,10.$$

B. Paramètres de dispersion

Deux séries statistiques peuvent avoir même valeur centrale, mais se différencier par la dispersion des valeurs observées autour de cette valeur. On appelle caractéristique de dispersion, une fonction des observations dont la valeur rend compte du plus ou moins grand *étalement* des valeurs observées autour de leur tendance centrale.

2.2.6 Étendue

Se définit comme étant égale à la différence entre la plus grande et la plus petite valeur observée :

$$\begin{aligned} \Delta = e = x_{max} - x_{min} = x_{(n)} - x_{(1)} &\quad \text{série statistique} \\ \Delta = e = x_{(k)} - x_{(1)} &\quad \text{D.O.1 en } k \text{ modalités} \\ \Delta = e = b_{(k)} - a_{(1)} &\quad \text{D.G.1 en } k \text{ classes.} \end{aligned}$$

Le principal avantage est sa simplicité de calcul, mais il est, par contre, extêmement peu robuste.

Exemple : L'étendue de la D.O.1 de l'exemple 2.1.2 est :

$$e = 6 - 2 = 4.$$

Exemple : Pour la D.G.1 (Exemple 2.1.3) [Bovins] l'étendue est :

$$e = 21 - 0 = 21.$$

2.2.7 Écarts interquantile ou interdécile

L'écart interquantile E_Q se définit comme étant la longueur de l'intervalle interquantile défini par les valeur x_p et $x_{(1-p)}$.

Cas particuliers :
Intervalle interquartile : $[Q_1, Q_3]$
Propriété de l'intervalle interquartile : Il contient 50% des observations de la distribution.
Ecart interquartile : $E_Q = Q_3 - Q_1 = x_{3/4} - x_{1/4}$ - caractéristique de dispersion extrêmement robuste

L'écart interquartile est une mesure de la dispersion des 50% d'observations centrales.

Intervalle interdécile : $[x_{1/10}, x_{9/10}]$.
Propriété de l'intervalle interdécile : Il contient 80% des observations de la distribution.
L'écart interdécile E_D se définit comme étant la longueur de l'intervalle interdécile défini par les valeur $x_{1/10}$ et $x_{9/10}$.

L'écart interdécile est une mesure de la dispersion des 80% d'observations centrales.

Exemple. Obtenir l'intervalle interdécile $[x_{1/10}, x_{9/10}]$ et l'écart interdécile E_D pour la série statistique de l'exemple 2.1.2 [Notes].
Du tableau de la distribution

Note	Fréquence cumulée $\sum_{j=1}^{i} f_j$
2	0,1
3	0,3666
4	0,7333
5	0,9
6	1

Le premier décile on obtient immédiatement : $x_{1/10} = 2$
Le neuvième décile est $x_{9/10} = 5$.

L'intervalle $[x_{1/10}, x_{9/10}] = [2; 5]$ est l'intervalle cherché. L'intervalle interdécile comporte 80% des observations de la série statistique.

L'écart interdécile E_D est $E_D = 3$.

Exemple. Obtenir l'intervalle interdécile $[x_{1/10}, x_{9/10}]$ et l'écart interdécile E_D pour la série statistique de l'exemple 2.1.3 [Bovins].
Du tableau de la distribution

Nombre de bovins classe $[a : b]$	Fréquence cumulée $\sum_{j=1}^{i} f_j$
$[0; 3[$	0,08
$[3; 6[$	0,1833
$[6; 9[$	0,3
$[9; 12[$	0,6333
$[12; 15[$	0,8
$[15; 18[$	0,916
$[18; 21[$	1

pour le calcul du premier décile on doit faire une interpolation entre les valeurs $(3; 0,08)$ et $(6; 0,18)$.
On trouve :

$$\frac{x_{1/10} - 3}{6 - 3} = \frac{1/10 - 0,08}{0,18 - 0,08}$$

ou encore :

$$\frac{x_{1/10} - 3}{3} = \frac{0,02}{0,1}; \quad x_{1/10} = 3 + 3 * \frac{0,02}{0,1} = 3,6.$$

Pour le calcul du neuvième on procède par interpolation linéaire entre les valeurs $(15; 0,8)$ et $(18; 0,916)$.
On trouve :

$$\frac{x_{9/10} - 15}{18 - 15} = \frac{9/10 - 0,8}{0,916 - 0,8}$$

ou encore :

$$\frac{x_{9/10} - 15}{3} = \frac{0,1}{0,116}; \quad x_{9/10} = 15 + 3 * \frac{0,1}{0,116} = 17,59.$$

L'intervalle $[x_{1/10}, x_{9/10}] = [3,6; 17,59]$ est l'intervalle cherché. L'intervalle interdécile comporte 80% des observations de la série statistique.

L'écart interdécile E_D est $E_D = 13,99$.

Tous les indicateurs n'ont pas le même intérêt. Selon la situation traitée les éléments à observer seront différents.

Exemple 2.2.6 /Feuille 4/ :Le tableau suivant décrit la répartition des accidents de la route selon les heures de la journée. On souhaite dégager les tendances essentielles de ces informations.

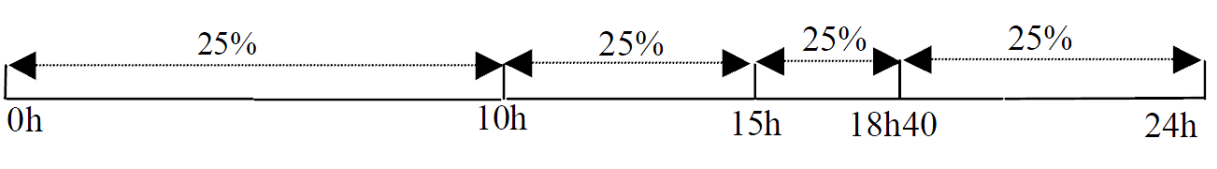
tranche horaire (en heures)	[0,3[	[3,6[	[6,9[	[9,12[	[12,15[	[15,18[	[18,21[	[21,24[
nombre d'accidents	8155	6258	15284	18006	23703	29759	29172	13022

On constate directement qu'un calcul de la moyenne serait ici sans intérêt :affirmer que les accidents de la route ont lieu en moyenne à 14h04 n'a pas de sens. Cependant, les renseignements relatifs à la répartition sont plus intéressants.

Un graphique des fréquences cumulées nous permet de le voir :

tranche horaire	fréquences	fréquences cumulées
[0,3[	5.7	5.7
[3,6[	4.4	10.1
[6,9[	10.7	20.8
[9,12[	12.5	33.3
[12,15[	16.5	49.8
[15,18[	20.8	70.6
[18,21[	20.3	90.9
[21,24[	9.1	100

La classe horaire [15 ;18[est la plus dangereuse (20% des accidents) : c'est le mode de la série.
Quelques points significatifs sur le graphique : les points d'ordonnées 50, 25 et 75.
Leurs abscisses (obtenues par lecture sur le graphique) nous permettent d'affirmer que les accidents se repartissent selon le schéma :



Et nous voyons ainsi tout l'intérêt des quartiles (10 h et 18h40) et de la médiane (15h).
On peut résumer cette courte étude par :
La période noire des accidents est 15h - 18h40.
Dans la journée, si un accident sur deux se produit entre 10h et 18h40, c'est entre 15h et 18h40 qu'a lieu le quart des accidents.

Pour caractériser la dispersion d'une série statistique de n observations, nous pouvons nous intéresser à la concentration des valeurs observées autour d'une valeur centrale, par exemple la moyenne arithmétique $\bar{x}$. Une forte concentration des observations autour de $\bar{x}$ se traduit par des écarts $(x_i - \bar{x})$ de faible amplitude, une grande dispersion implique qu'il existe des écarts importants.
Nous pouvons poser la question : "de combien s'écarte-t-on en moyenne de $\bar{x}$?".
Mais comme la somme des écarts à la moyenne est nulle, il faudra considérer ces écarts sans leur signe.

2.2.8 Ecart moyen

L'**écart moyen** ou écart moyen absolu se défini comme la moyenne arithmétique des valeurs absolues des écarts entre les observations et leur moyenne arithmétique.

L'intérêt de cette valeur est à cause de son calcul facile et l'interprétation simple et évidente.

Considérons une statistique $x_1, x_2, \ldots, x_n$ de moyenne $\bar{x}$.

Exemple Considérons l'exemple 2.1.2 [Notes]

Description	Ecart moyen	Excel
Série statistique	$EM = \frac{1}{n}\sum_{i=1}^{n} \lvert x_i - \bar{x}\rvert$	AVDEV(x_i)
D.O.1, k modalités	$EM = \frac{1}{n}\sum_{i=1}^{k} n_i.\lvert x_i - \bar{x}\rvert = \sum_{i=1}^{k} f_i.\lvert x_i - \bar{x}\rvert$	
D.G.1 en k classes	$EM = \frac{1}{n}\sum_{i=1}^{k} n_i.\lvert x_i^* - \bar{x}\rvert = \sum_{i=1}^{k} f_i.\lvert x_i^* - \bar{x}\rvert$	

TABLE 2.2 : Ecart-moyen [6]

Rappelons qu'on a trouvé $\bar{x} = 3,9$.

Note	Effectif n_i	Écarts en valeurs absolues $\lvert x_i - \bar{x}\rvert$	Produit $n_i.\lvert x_i - \bar{x}\rvert$
2	3	1,9	5,7
3	8	0,9	7,2
4	11	0,1	1,1
5	5	1,1	5,5
6	3	2,1	6,3
Total	30		25,8

Pour l'écart moyen on obtient :

$$EM = \frac{1}{n}\sum_{i} n_i|x_i - \bar{x}| = \frac{25,8}{30} = 0,86.$$

L'écart absolu moyen des notes est donc de 0.86, ce qui signifie que les notes s'écartent en moyenne de 0.86 de la moyenne. Il n'y a donc pas, en moyenne, de gros écarts à la moyenne.

Si on considère une deuxième série statistique Y de notes, pour laquelle on trouve un écart absolu moyen de 3.8, cela signifie que les notes de la série Y s'écartent généralement beaucoup plus de la moyenne que celles de la série X. On peut donc conclure que la dispersion des notes de la série Y est plus forte que celle de la série X.

Exemple Considérons l'exemple 2.1.3 [Bovins]

Rappelons qu'on a trouvé $\bar{x} = 10,75$.

Classe $[a-b[$	Centre de classe $x_i^* = (a+b)/2$	Effectif n_i	Écarts en valeurs absolues $\lvert x_i^* - \bar{x}\rvert$	Produit $n_i.\lvert x_i^* - \bar{x}\rvert$
$[0-3[$	1,5	5	9,25	46,25
$[3-6[$	4,5	6	6,25	37,5
$[6-9[$	7,5	7	3,25	22,75
$[9-12[$	10,5	20	0,25	5
$[12-15[$	13,5	10	2,75	27,5
$[15-18[$	16,5	7	5,75	40,25
$[18-21[$	19,5	5	8,75	43,75
Total		60		223

Pour l'écart moyen on obtient :

$$EM = \frac{1}{n}\sum_i n_i|x_i - \bar{x}| = \frac{223}{60} = 3,716667.$$

2.2.9 Variance

La **variance** est la moyenne des carrés des écarts à la moyenne arithmétique

$$\text{Variance de la série} \qquad s^2 = \frac{1}{n}\sum n_i\,(x_i - \bar{x})^2$$

Comme pour l'écart moyen, on a le tableau qui suit : Exemple 2.1.2 [Notes]

Description	Variance	Excel
Série statistique	$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$	VARP(input range)
D.O.1 de k modalités	$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i.(x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i.(x_i - \bar{x})^2.$	
D.G.1, k classes	$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i.(x_i^* - \bar{x})^2 = \sum_{i=1}^{k} f_i.(x_i^* - \bar{x})^2.$	

TABLE 2.3 : Variance [6]

On construit le tableau que voici :

Note	Effectif n_i	Écarts en valeurs absolues $\lvert x_i - \bar{x}\rvert$	(Écarts)^2 $(x_i - \bar{x})^2$	Produit $n_i.(x_i - \bar{x})^2$
2	3	1,9	3,61	10,83
3	8	0,9	0,81	6,48
4	11	0,1	0,01	0,11
5	5	1,1	1,21	6,05
6	3	2,1	4,41	13,23
Total	30			36,7

$$s^2 = \frac{1}{n}\sum_i n_i(x_i - \bar{x})^2 = \frac{36,7}{30} = 1,22333.$$

Exemple : Dans le cas de l'exemple 2.1.3 [Bovins] on construit le tableau que voici :

Classe $[a-b[$	Centre x_i^*	Effectif n_i	Écarts $\lvert x_i^* - \bar{x}\rvert$	(Écarts)^2 $(x_i^* - \bar{x})^2$	Produit $n_i.(x_i^* - \bar{x})^2$
$[0-3[$	1,5	5	9,25	85,56	427,81
$[3-6[$	4,5	6	6,25	39,06	234.375
$[6-9[$	7,5	7	3,25	10,56	73,9375
$[9-12[$	10,5	20	0,25	0,0625	1,25
$[12-15[$	13,5	10	2,75	7,56	75,625
$[15-18[$	16,5	7	5,75	33,06	231,44
$[18-21[$	19,5	5	8,75	76,56	382,81
Total		60			1427,25

$$s^2 = \frac{1}{n}\sum_i n_i(x_i^* - \bar{x})^2 = \frac{1427,25}{60} = 23,79.$$

Exemple : En utilisant les formules simplifiées (Table 2.4) pour l'exemple 2.1.2 [Notes] on

Description	variance
Série statistique	$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$
D.O.1, k modalités	$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i.{x_i}^2 - \bar{x}^2 = \sum_{i=1}^{k} f_i.{x_i}^2 - \bar{x}^2.$
D.G.1 k classes	$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i.{x_i^*}^2 - \bar{x}^2 = \sum_{i=1}^{k} f_i.{x_i^*}^2 - \bar{x}^2.$

TABLE 2.4 : Variance - formules simplifiées [6]

obtient :

Notes x_i	Notes^2 ${x_i}^2$	Produit $n_i.x_i^2$
2	4	12
3	9	75
4	16	176
5	25	125
6	36	108
Total		493

On obtient

$$s^2 = \frac{1}{n}\sum_i n_i x_i^2 - \bar{x}^2 = \frac{1}{30}493 - 3,9^2 = 1,22333.$$

Exemple En utilisant les formules simplifiées pour l'exemple 2.1.3 [Bovins] on obtient :

Classe $[a-b[$	Centre x_i^*	(Centres)^2 ${x_i^*}^2$	Produit $n_i.x_i^{*2}$
$[0-3[$	1,5	2,25	11,25
$[3-6[$	4,5	20,25	121,5
$[6-9[$	7,5	56,25	393,75
$[9-12[$	10,5	110,25	2205
$[12-15[$	13,5	182,25	1822,5
$[15-18[$	16,5	272,25	1905,75
$[18-21[$	19,5	380,25	1901,25
Total			8361

On obtient

$$s^2 = \frac{1}{n}\sum_i n_i x_i^{*2} - \bar{x}^2 = \frac{1}{60}8361 - 10,75^2 = 23,7875.$$

Propriétés de la variance

1. Si toutes les valeurs sont identiques la variance est nulle.

2. La variance fait intervenir tous les termes de la série et est sensible aux valeurs aberrantes (une valeur aberrante est toujours la plus petite ou la plus grande valeur de la série de mesures).

3. **Théorème de Koning-Huygens :** Soit c un paramètre de centralité et considérons la moyenne des carrés des écarts entre les observations et le paramètre c. Cette valeur est minimale lorsque c est la moyenne arithmétique, (voir propriété de la moyenne).

4. Influence d'un changement d'origine ou d'unité. La variance n'est pas un nombre sans dimension et l'unité dans laquelle s'exprime la variance est le carré des unités utilisées pour les valeurs observées.
Soit $S(x_i, n_i)$ (i varie de 1 à n) une distribution statistique observée de moyenne arithmétique

$\bar{x}$ et de variance $Var(x)$. La variance $Var(y)$ de la série $S(y_i, n_i)$ telle que pour tout i $y_i = (x_i - x_0)/d$ est donnée par la relation

$$Var(y) = Var(x)/d^2.$$

5. Variance de la "réunion" de deux séries statistiques.
Si on réunit deux ensembles E_1 et E_2 d'observations, le premier d'effectif n_1, de moyenne $\bar{x}_1$ et de variance s_1^2, le deuxième d'effectif n_2, de moyenne $\bar{x}_2$, et de variance s_2^2, la variance s^2 de la série E obtenue d'effectif $n = n_1 + n_2$ s'exprime à partir des paramètres de E_1 et E_2 par la relation suivante :

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2 + n_1(\bar{x}_1 - x)^2 + n_2(\bar{x}_2 - x)^2}{n_1 + n_2} = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} + \frac{n_1.n_2}{(n_1 + n_2)^2}(\bar{x}_1 - \bar{x}_2)^2.$$

2.2.10 Ecart-type

La variance n'est pas comparable directement à la moyenne, car l'unité de la variance est le carré de l'unité de la variable. Pour que l'indicateur de dispersion puisse être comparé aux paramètres de tendance centrale (moyenne, médiane et mode), il suffit d'en prendre la racine carrée. La racine carrée de la variance s'appelle **écart-type**.

L'**écart-type** d'une série statistique est égal à la racine carrée de la variance et est désigné par s.

$$s = \sqrt{Var(x)} \quad \text{/STDEVP(imput range)/.}$$

Une série peu dispersée (ayant des valeurs regroupées autour de la valeur moyenne) aura un écart-type plutôt faible.

Remarque 13 Pour une distribution symétrique, pratiquement toutes les observations sont situées entre $\bar{x} - 3s$ et $\bar{x} + 3s$.

2.2.11 Comparaison de séries statistiques

Très souvent l'analyse des séries est liée à la comparaison de deux ou plusieurs séries qui ne sont pas dans la même unité. Considérons un exemple :

- Soit x la série statistique de 4 produits en \$: 100\$, 200\$, 300\$ et 400\$
- Soit y la série statistique des 4 produits en € : 15€, 30€, 45€, 60€

Intuitivement les deux séries sont dispersées de la même manière, mais :

$$s_X = 111,8\$; \quad s_Y = 16,8€$$

Pour comparer des séries qui ne sont pas dans la même unité, il faut transformer les caractéristiques de dispersion

2.2.12 Coefficient de variation

La valeur de l'écart-type est bien sûr liée à la grandeur de la variable X. Elle est d'autant plus petite que la valeur mesurée est faible. Mais cela ne permet pas de comparer la dispersion de deux séries dont l'une présente des valeurs élevées et l'autre des valeurs petites. Pour obtenir un renseignement sur la dispersion relative des données on peut calculer le **coefficient de variation** qui exprime l'écart type par rapport à la moyenne des résultats. Il est généralement exprimé en pourcentage. La formule utilisée est la suivante :

$$CV\% = \left|\frac{s_x}{\bar{x}}\right| \times 100.$$

Le coefficient de variation est une mesure relative de dispersion. Le CV permet d'apprécier la représentativité de la moyenne par rapport à l'ensemble des observations. Il donne une bonne idée du degré d'homogénéité d'une série. Il faut qu'il soit le plus faible possible ($< 15\%$ en pratique).

Ce coefficient, parce que sans unité, permet la comparaison des distributions dont les moyennes sont trop différentes (de différents ordres de grandeur), ou des distributions de nature similaire mais correspondant à des observations faites en des lieux et/ou à des dates différentes. Ce paramètre constitue une mesure relative de dispersion, une mesure de la concentration.

Le coefficient de variation est indispensable pour l'étude de la comparaison du risque de deux actions.

Exemple 2.2.7 [6] Considérons A et B mesurées dans la même monnaie. On a par exemple :

$$\bar{x}_A = 150 \quad s_A = 5.$$
$$\bar{x}_B = 50 \quad s_B = 3.$$

$\bar{x}$ décrit la cote moyenne de l'action, s est une mesure de sa *variabilité absolue.*
Il apparaît que le coefficient de la variance décrit mieux que tout autre le risque "pur" d'une action. En effet :

— Dans le cas de l'action A, on trouve

$$CV_A = \frac{5}{150} = 0,033333.$$

— Pour l'action B :

$$CV_B = \frac{3}{50} = 0,06.$$

Le risque de l'action A est évidement inférieur au risque de l'action B, même si, en premier approche, on constate que son écart-type est plus grand.

Exemple 2.2.8 On désire comparer les distributions (groupées) des bénéfices nets hebdomadaires en euros de 2 magasins, sur 100 semaines comprenant toutes 6 jours d'ouverture. Les paramètres des deux distributions sont :
Magasin 1 : $\bar{x} = 2900$; $s = 1063$ Magasin 2 : $\bar{x} = 13000$; $s = 1077$.

Les 2 distributions (groupées) ayant pratiquement le même écart-type, on pourrait avoir tendance à penser qu'elles présentent la même dispersion. Mais, en y regardant de plus près, on se convainc aisément qu'une perte ou un gain de euros n'aura pas le même impact pour le premier magasin (pour lequel le bénéfice hebdomadaire moyen n'est que de 2 900 euros) et pour le second (pour lequel le bénéfice hebdomadaire moyen s'élève à 13 000 euros).

Dans ces conditions, on peut penser recourir à la mesure de dispersion relative qu'est le coefficient de variation.

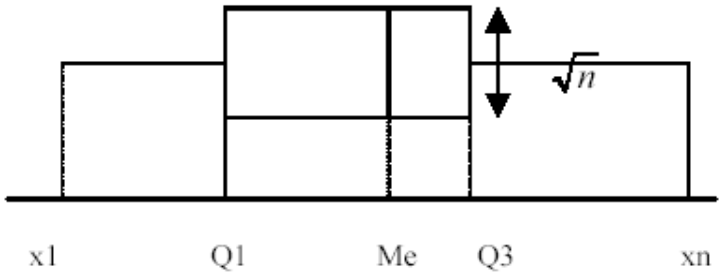
$$\text{Magasin 1 :} \qquad CV = \frac{1063}{2900} = 0,367 = 36\%$$
$$\text{Magasin 2 :} \qquad CV = \frac{1077}{13000} = 0,083 = 8,3\%$$

Ces 2 coefficients de variation montrent mieux l'influence réelle d'un gain ou d'une perte équivalent(e) à l'écart-type pour chacun des magasins.

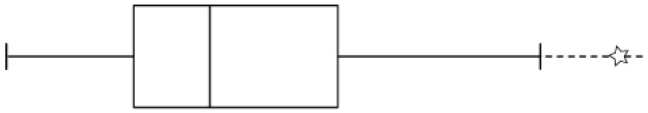
2.2.13 Boîte à moustaches

- Résume la série à partir de ses valeurs extrêmes, ses quartiles et sa médiane.
- Permet une comparaison visuelle immédiate de plusieurs séries.

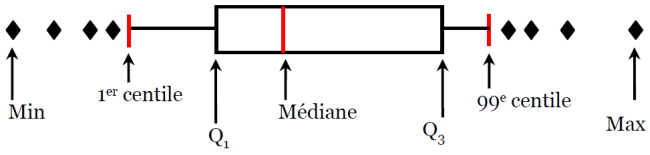
Construction :
Sur un axe horizontal, on place les valeurs extrêmes et les quartiles.
On trace un rectangle de longueur l'interquartile et la largeur proportionnelle à la racine carrée de la taille de la série.



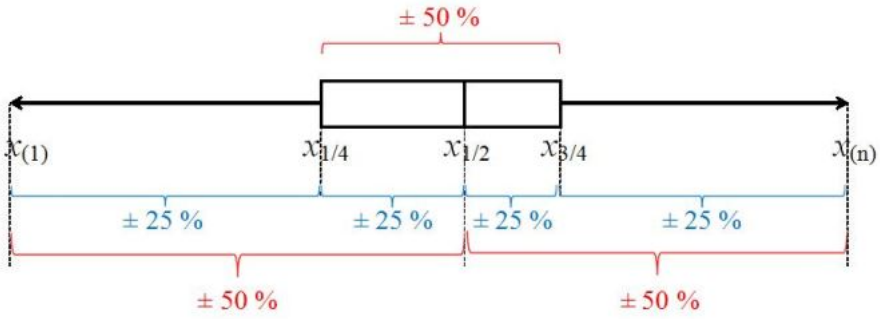
On partage le rectangle par un segment vertical au niveau de la médiane M_e.
De part et d'autre de la boîte, on définit deux « moustaches » de longueur égale à 1,5 fois l'étendue interquartile. Si une observation dépasse une des moustaches, elle est considérée comme " aberrante " et individualisée.



S'il n'y a pas d'observation aberrante, la moustache s'arrête à l'observation immédiatement supérieure x_n (ou inférieure x_1).



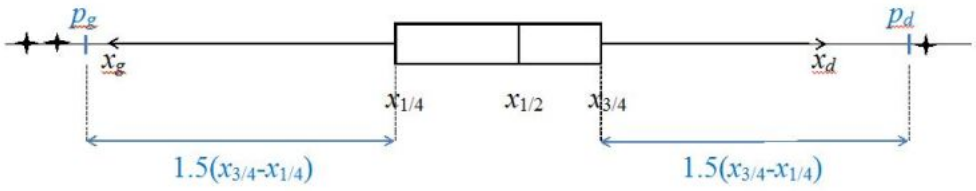
Une boîte à moustaches nous indique de façon simple et visuelle quelques traits marquants de la série observée :



- la médiane nous renseigne sur le milieu de la série. Un décalage de la médiane vers le Q_1 (à gauche ou vers le bas) signale l'existence d'une dissymétrie à gauche de la distribution ;
- les largeurs des deux parties de la boîte rendent compte de la dispersion des valeurs situées au centre de la série (la boîte contient 50% (environ) de l'ensemble des observations : 25% à gauche de la médiane et 25% à sa droite) ;
- la longueur des moustaches renseigne sur la dispersion des valeurs situées au début de la série ordonnée (les valeurs les plus petites correspondant à 25% des observations) ou à la fin de celle-ci (les valeurs les plus grandes correspondant aussi à 25% des observations) ;
- de façon générale, la boîte et les moustaches seront d'autant plus étendues que la dispersion de la série statistique est grande.

Pour certains auteurs, les moustaches s'étendent jusqu'aux valeurs extrêmes même si celles-ci dépassent l'intervalle défini plus haut. Dans ce cas les moustaches risquent de devenir très longues. Pour éviter cet inconvénient, on modifie la représentation en introduisant le concept de *valeur pivot*, de *valeur adjacente* et de *valeur extérieure*.

Construction de la version modifiée



La version modifiée de la boîte à moustaches se construit en 4 étapes :
1. construction de la boîte, comme dans la version de base ;
2. calcul des valeurs pivots gauche (p_g) et droite (p_d) ;

3. détermination des valeurs adjacentes gauche (x_g) et droite (x_d) : ces valeurs adjacentes correspondent aux extrémités des moustaches gauche et droite ;
4. détermination des valeurs extérieures éventuelles.

Les valeurs pivots
Les valeurs pivots sont définies par les relations suivantes :

$$\begin{cases} p_g = x_{1/4} - 1,5(x_{3/4} - x_{1/4}) & \text{pivot gauche} \\ p_d = x_{3/4} + 1,5(x_{3/4} - x_{1/4}) & \text{pivot droit} \end{cases}$$

Elles sont situées de part et d'autre de la boîte, à une distance valant 1,5 fois l'écart interquartile.

Remarque 14 La définition des valeurs pivots résulte d'une constatation : la plupart des séries statistiques qui ne contiennent pas de valeurs extrêmes ou aberrantes, ont leurs observations situées dans l'intervalle $[p_g, p_d]$.

Remarque 15 p_g et p_d ne coïncident généralement pas avec des valeurs observées. Il s'agit juste de valeurs **calculées** dans le but de déterminer, dans un deuxième temps, les valeurs adjacentes.

Les valeurs adjacentes (extrémités des moustaches)
Les valeurs adjacentes, contrairement aux valeurs pivots, doivent être des valeurs observées de la série statistique. Elles correspondront aux extrémités des moustaches gauche et droite du diagramme en boîte.

On définit les **valeurs adjacentes** par rapport aux valeurs pivots p_g et p_d comme suit :
• la valeur **adjacente gauche**, notée x_g, est la **plus petite valeur observée supérieure ou égale à** p_g ;
• la **valeur adjacente droite**, notée x_d, est la **plus grande valeur observée inférieure ou égale à** p_d**.**

Les valeurs extérieures
Si toutes les observations s_i sont comprises entre le pivot gauche p_g et le pivot droit p_d, alors $x_g = x_{(1)}$ et $x_d = x_{(n)}$. Dans le cas contraire, on isole les valeurs observées situées en dehors de l'intervalle $[p_g, p_d]$ pour en examiner les caractéristiques.

Toutes les observations situées en dehors de $[p_g, p_d]$ sont dites **extérieures**. Elles sont représentées par des symboles appropriés (étoiles, points, triangles, ...) de manière à être mises en évidence.

Remarque 16 Lorsque toutes les observations x_i sont comprises entre le pivot gauche p_g et le pivot droit p_d, $x_g = x_{(1)}$, $x_d = x_{(n)}$ et il n'y a pas de valeur extérieure. Dans ce cas, la version modifiée de la boîte à moustaches coïncide avec la version de base.

Remarque 17 Toute valeur extérieure n'est pas nécessairement extrême ou aberrante, mais une valeur extrême ou aberrante sera généralement une valeur extérieure.

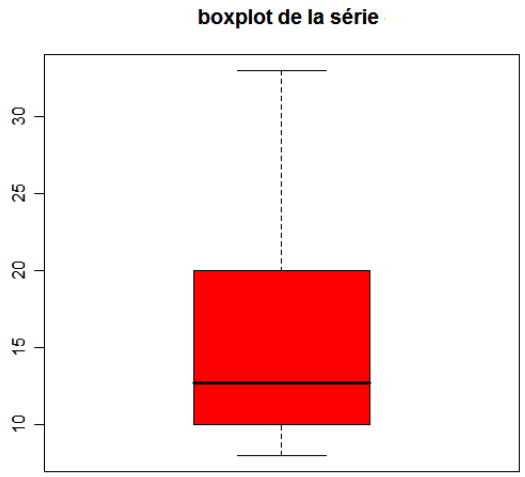
Remarque 18 Valeurs pivots du second type Afin de compléter l'analyse, on utilise parfois des valeurs pivots du second type définies par $p_g = x_{1/4} - 2(x_{3/4} - x_{1/4})$ et $p_d = x_{3/4} + 2(x_{3/4} - x_{1/4})$. Elles sont situées de part et d'autre de la boîte à une distance valant **deux fois** l'écart interquartile.

Le fait pour une valeur extérieure d'être en dehors de l'intervalle $[p_g, p_d]$ renforce la présomption « d'aberration ».

La plupart des logiciels statistiques distinguent les valeurs extérieures qui se trouvent en dehors de l'intervalle $[p_g, p_d]$ des autres valeurs extérieures en les représentant sur le diagramme en boîte avec des symboles différents.

Exemple 2.2.9 /Feuille 5/ : Série : $8; 8,5; 10; 11; 12,5; 13; 15; 20; 25; 33$

Min.	Q_1	M_e	Moyenne	Q_3	Max.
8.00	10,25	12,75	15,60	18,75	33,00



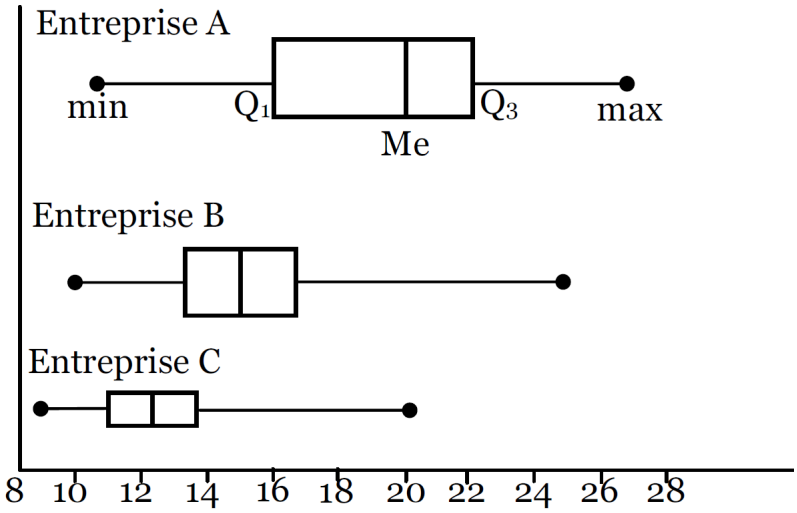
Min $= \text{Min}(x_1; x_{10})$; $Q_1 = \text{Quartile}(x_1; x_{10}; 1)$; $M_e = \text{Median}(x_1; x_{10})$; Moyenne= $\text{Average}(x_1; x_{10})$; Q_3=$\text{Quartile}(x_1; x_{10}; 3)$;Max = $\text{Quartile}(x_1; x_{10}; 4)$
Tools/Data Analysis/Descriptive Statistics/Summary Statisitcs

Cette représentation permet également de comparer facilement la distribution de différentes variables, ou encore de la même variable pour différentes modalités d'une variable qualitative.

Exemple 2.2.10 /Feuille 5/ : Comparer les salaires dans les trois entreprises suivantes d'un même secteur industriel.

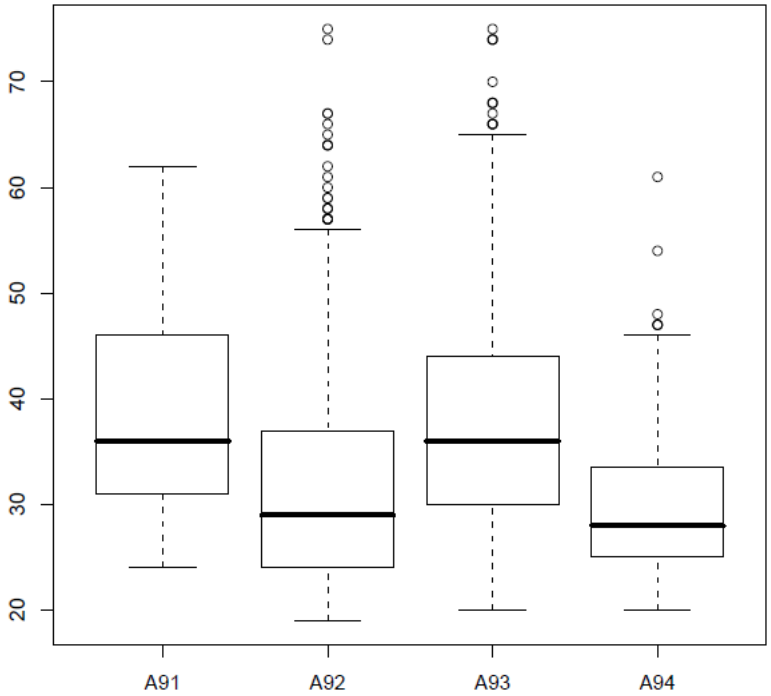
Entreprise	Taille	min	Q_1	M_e	Q_3	max
A	125	10 500	16 000	20 000	22 000	27 000
B	75	10 000	13 500	15 000	17 000	25 000
C	25	8 500	11 000	12 500	14 000	20 500

À partir des données, on obtient la représentation suivante :



Les salaires en B sont plus homogènes qu'en A, car $Q_{3_B} - Q_{1_B} < Q_{3_A} - Q_{1_A}$.
Les salaires en A sont plus élevés globalement qu'en B : $M_{eA} > M_{eB}$.

Exemple 2.2.11 /Feuille 5/ : La figure ci-dessous représente la distribution des 1000 clients d'une banque allemande, d'après leur état marital, ou on introduit le codage le suivant :
A91 : homme divorcé / séparé ;
A92 : femme divorcé / séparé / mariée ;
A93 : homme célibataire ;
A94 : homme marié / veuf .
On remarque ainsi que parmi les clients de la banque allemande les femmes divorcées, séparées ou mariées ainsi que les hommes mariés ou veufs sont généralement moins âgés que les hommes célibataires, divorcés ou séparés.



$A92$ et $A93$ ont presque la même homogénéité.
La boîte à moustaches permet d'illustrer la liaison entre une variable qualitative / état marital / et une variable quantitative / age / en représentant côte à côte des boîtes à moustaches pour chaque modalité de la variable qualitative.

Pour comparer des populations qui n'ont pas le même effectif, on trace la largeur du rectangle proportionnelle à la racine carrée de la population.

2.2.14 Inégalité de Bienaymé - Tchébycheff [3]

L'intérêt de l'écart-type comme paramètre de dispersion réside dans le fait que l'on démontre que, quelle que soit la distribution étudiée, on a au minimum $1-1/t^2$ des observations comprises entre la moyenne et plus ou moins t fois l'écart-type. Cette propriété est connue sous le nom d'inégalité de Bienaymé-Tchebicheff :

Soit $t \in \mathbb{R}^+$. La somme des fréquences des x_i extérieurs à l'intervalle $]\bar{x} - ts_x, \bar{x} + ts_x[$ est inférieure ou égale à $1/t^2$.

$$P(\bar{x} - ts < X < \bar{x} + ts) > 1 - \frac{1}{t^2}.$$

L'application de cette relation donne, pour $t = 2$, une probabilité minimale de $1 - 1/2^2 = 75\%$ d'observer des valeurs comprises $\bar{x} - 2s$ et $\bar{x} + 2s$.

2.2.15 Intervalles remarquables

La moyenne arithmétique $\bar{x}$ et l'écart-type s_x permettent de construire des intervalles du type

$$[\bar{x} - ts_x, \bar{x} + ts_x].$$

Suivant la valeur de t et la forme de la distribution, on peut parfois prévoir le pourcentage de la masse totale des observations incluses dans un tel intervalle. Par exemple pour une distribution en forme de cloche, pas trop dissymétrique, l'intervalle $[\bar{x} - 2s_x, \bar{x} + 2s_x]$ contient environ 95 % des observations.

2.2.16 Valeurs centrées-réduites z_j

Soit une série statistique $\{x_i; i$ varie de 1 à $n\}$, de moyenne $\bar{x}$ et d'écart-type s_x. Les valeurs centrées réduites z_i sont définies par :

$$z_i = (x_i - \bar{x})/s_x.$$

Ces valeurs sont sans dimension et ont une moyenne nulle et une variance égale à 1.

2.2.17 Effet d'une transformation linéaire sur les indicateurs

Transformation linéaire : $Y = aX + b$ $(a \neq 0)$

Moyenne $(Y) = a$ Moyenne$(X) + b$
Médiane $(Y) = a$ Médiane$(X) + b$
Mode $(Y) = a$ Mode$(X) + b$

Variance $(Y) = a^2$ Variance(X)
Écart-type $(Y) = |a|$ Écart-type (X)

C. Paramètres de forme

Considérons l'histogramme des fréquences d'une distribution groupée où l'effectif est réparti en classes de même amplitude. En joignant les milieux des bases supérieures et en considérant 2 intervalles, d'effectif nul, on obtient le polygone des fréquences (ou *courbe des fréquences*). L'aire comprise entre le polygone et l'axe des abscisses vaut 1.
Certaines courbes de fréquences peuvent avoir des formes particulières : courbe symétrique (en cloche), courbe asymétrique, courbe en U, courbe bimodale, ...

On peut construire des paramètre de forme à l'aide des **moments** de la distribution.

2.2.18 Moment non centré d'ordre r

On appelle moment non centré d'ordre r d'une série statistique $\{x_i\}$, $i = 1, \ldots, n$ ou d'une distribution groupée $S(x_i, n_i)$ le nombre

$$m_r = \frac{1}{n}\sum_{i}^{n} x_i^r \quad \text{Série statistique}$$

$$m_r = \frac{1}{n}\sum_{i}^{k} n_i x_i^r \quad \text{D.O.1 de } k \text{ modalités}$$

$$m_r = \frac{1}{n}\sum_{i}^{k} n_i x_i^{*r} \quad \text{D.G.1 en } k \text{ classes.}$$

2.2.19 Moment centré d'ordre r

On appelle moment centré en a d'ordre r d'une distribution groupée $S(x_i, n_i)$ le nombre

$$m_{ar} = 1/n \sum_{i}^{n} n_i (x_i - a)^r.$$

Cas particulier : moment centré en $\bar{x}$ d'ordre r

$$\begin{aligned} \mu_r &= 1/n \sum_{i}^{n} (x_i - \bar{x})^r \quad \text{Série statistique} \\ \mu_r &= 1/n \sum_{i}^{k} n_i (x_i - \bar{x})^r \quad \text{D.O.1 de } k \text{ modalités} \\ \mu_r &= 1/n \sum_{i}^{k} n_i (x_i^* - \bar{x})^r \quad \text{D.G.1 en } k \text{ classes} \end{aligned}$$

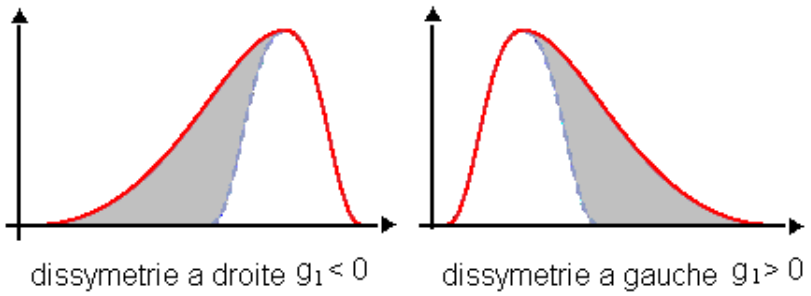
Le mathématicien Fischer a défini un coefficient qui permet de caractériser le degré d'asymétrie et un coefficient qui permet de caractériser l'aplatissement de la courbe de fréquences.

2.2.20 Symétrie

Une série a une distribution est symétrique si ses valeurs sont également dispersées de part et d'autre de la valeur centrale, c'est-à-dire si le graphe de la distribution - histogramme ou diagramme en bâton en fréquences - admet une axe de symétrie.

Coefficient de symétrie de Fischer : $g_1 = \mu_3 / s^3$

Si $g_1 = 0$ il y a symétrie, si $g_1 > 0$ il y a dissymétrie à gauche, si $g_1 < 0$ il y a dissymétrie à droite.



Un décalage de la médiane vers le bas (Q_1) signifie l'existence d'une dissymétrie à gauche de la distribution.

2.2.21 Aplatissement /Kurtosis/

Une distribution est plus ou moins aplatie selon que les fréquences des valeurs voisines des valeurs centrales diffèrent peu ou beaucoup les unes par rapport aux autres.

Coefficient d'aplatissement de Fisher : $g_2 = (\mu_4/s^4) - 3$.

Si $g_2 = 0$ la courbe est normale,
si $g_2 < 0$, la concentration des valeurs autour de la moyenne est faible, : la courbe est plus aplatie ;
si $g_2 > 0$, la concentration des valeurs de la série autour de la moyenne est forte : la courbe est plus pointue.

En pratique, ces coefficients servent à contrôler la proximité de l'histogramme et de la courbe en cloche :

- cas $g_1 \approx 0$ et $g_2 \approx 0$: la répartition des données est plus ou moins normale ;
- cas $g_1 \neq 0$ ou $g_2 \neq 0$: la répartition des données est différente de la loi normale.

D. Paramètres de concentration

2.2.22 Courbe de concentration. Courbe de Lorentz

Il existe une courbe particulière très utilisée en économie et gestion : c'est la courbe de concentration. Elle concerne les variables continues à valeurs positives et se construit à partir des fréquences relatives cumulées croissantes.

On procède de façon suivante pour l'établir :

Étudions la concentration de la distribution d'un caractère x_i^* réparti en classes d'effectifs n_i. On calculera pour tout i :

$$N_i = \sum_{j=1}^{i} n_j \quad \text{nombre d'individus inférieur à } x_i$$

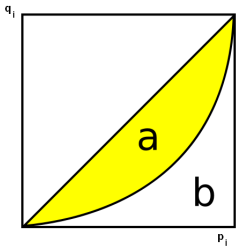
$$S_i = \sum_{j=1}^{i} n_j x_j^* \quad \text{valeur cumulée du caractère inférieur à } x_i^*$$

On rapporte ensuite ces chiffres à l'effectif total et à la valeur totale pour construire la courbe :

$$p_i = F_i = \frac{\sum_{j=1}^{i} n_j}{\sum_{j=1}^{n} n_j} = \frac{N_i}{n}$$

$$q_i = \frac{\sum_{j=1}^{i} n_j x_j^*}{\sum_{j=1}^{n} n_j x_j^*} = \frac{S_i}{S_n}$$

On trace la courbe de Lorentz $q_i = f(p_i)$, qui représente les fréquences cumulées de la masse de la variable q_i en fonction des fréquences cumulées p_i.



La courbe se réduirait à la première bissectrice (égalité parfaite) si toutes les valeurs de x_i étaient égales. Plus la courbe s'éloigne de celle-ci, plus la distribution est inégalement répartie.

Exemple 2.2.12 Calculons la courbe de concentration dans le cas des données suivantes :

$$x_1^* = 112, x_2^* = 151, x_3^* = 210, x_4^* = 225, x_5^* = 230, x_6^* = 354, x_7^* = 360, x_8^* = 450.$$

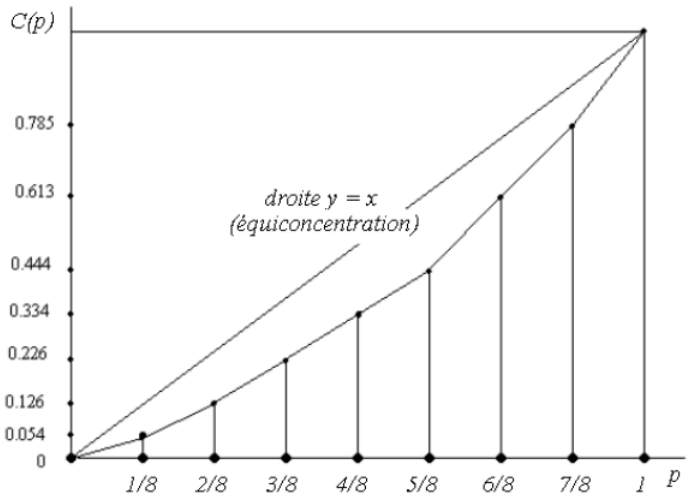
On calcule la somme de toutes les valeurs :

$$S_n = S_8 = \sum_{i=1}^{8} x_i^* = 2092$$

On a :

x_i^*	N_i	$p_i = F_i = N_i/n$	$S_j = \sum_{i=1}^{i} x_j^*$; $q_i = S_i/S_n$	
112	1	$p_1 = F_1 = 1/8$	$q_1 = S_1/S_8 = 112/2092$	$= 0.054$
151	2	$p_2 = F_2 = 2/8$	$q_2 = S_2/S_8 = (112 + 151)/2092$	$= 0.126$
210	3	$p_3 = F_3 = 3/8$	$q_3 = (112 + 151 + 210)/2092$	$= 0.226$
225	4	$p_4 = F_4 = 4/8$	$q_4 = (112 + 151 + 210 + 225)/2092$	$= 0.334$
230	5	$p_5 = F_5 = 5/8$	$q_5 = (112 + 151 + 210 + 225 + 230)/2092$	$= 0.444$
354	6	$p_6 = F_6 = 6/8$	$q_6 = (112 + 151 + 210 + 225 + 230 + 354)/2092$	$= 0.613$
360	7	$p_7 = F_7 = 7/8$	$q_7 = (112 + 151 + 210 + 225 + 230 + 354 + 360)/2092$	$= 0.785$
450	8	$p_8 = F_8 = 8/8$	$q_8 = (112 + 151 + 210 + 225 + 230 + 354 + 360 + 450)/2092$	$= 1$

D'où la courbe de concentration :



On peut résumer : 50% des observations donnent 33,4% de la masse de la valeur observée. 75% des observations représentent 61,3% de la valeur. A peu près 33% des éléments supérieurs donnent 50% de la masse de la valeur observée.

Nous admettons ici les propriétés suivantes :
- L'équiconcentration signifie que les observations sont constantes ;
- La fonction $q_i = f(p_i)$ est croissante et égale à 1 pour $p_i = p_n = 1$ (ou i = n) ;
- $q_i = f(p_i)$ est toujours inférieur à p_i, ce qui signifie que la courbe de concentration est toujours en dessous de la droite $y = x$.
- $q_i = f(p_i)$ augmente de plus en plus vite.

2.2.23 Coefficient (indice) de concentration de Gini

Comme on peut le constater sur la figure ci-dessus, l'aire comprise entre la droite $y = x$ et la courbe de concentration varie de 0 à 0.5. L'usage en statistique étant d'utiliser des paramètres variant de 0 à 1, on définit le coefficient de Gini par le double de cette aire :

On appelle **indice de concentration g de Gini** ou **coefficient de concentration de Gini** : 2 fois l'aire entre la courbe de Lorentz et la première bissectrice.

Son calcul n'est pas simple. On peut utiliser la formule suivante

Si X est obtenu par une série statistique, i.e. si dans une population de n individus, on a observé les valeurs $x_1, \dots x_n$, alors l'indice de Gini vaut le nombre sans dimension :

$$g = \frac{d}{\bar{x}},$$

où d est la différence moyenne entre les observations :

$$d = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} |x_i - x_j|}{n(n-1)}$$

et $\bar{x}$ est la moyenne arithmétique de $x_1, \dots, x_n$, c'est-à-dire

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

L'indice (ou coefficient) de Gini est un indicateur synthétique d'inégalités de salaires (de revenus, de niveaux de vie...). Il varie entre 0 et 1. Il est égal à 0 dans une situation d'égalité parfaite où tous les salaires, les revenus, les niveaux de vie... seraient égaux. A l'autre extrême, il est égal à 1 dans une situation la plus inégalitaire possible, celle où tous les salaires (les revenus, les niveaux de vie...) sauf un seraient nuls. Entre 0 et 1, l'inégalité est d'autant plus forte que l'indice de Gini est élevé.

Pour l'exemple 2.2.12 l'indice de Gini est

$$g = 2\left(0,5 - \frac{1}{8}\frac{0,054}{2} - \left(\frac{2}{8} - \frac{1}{8}\right)\frac{0,054 + 0,126}{2} - \frac{1}{8}\frac{0,126 + 0,226}{2} - \frac{1}{8}\frac{0,226 + 0,334}{2}\right.$$

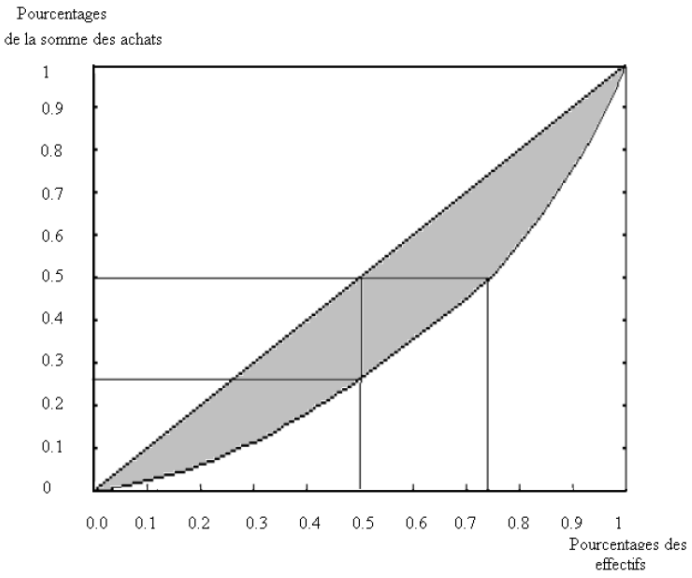
$$
\begin{aligned}
&- \frac{1}{8}\frac{0,334+0,444}{2} - \frac{1}{8}\frac{0,444+0,613}{2} - \frac{1}{8}\frac{0,785+0,613}{2} - \frac{1}{8}\frac{1+0,785}{2}\Big) \\
&= 2(0,5 - 0,00334608 - 0,011203394 - 0,021988528 - 0,034984465 - 0,048577916 \\
&\quad -0,066025335 - 0,087356597 - 0,111555927) \\
&= 0,229924 \Longrightarrow \text{ faible concentration.}
\end{aligned}
$$

Remarque 19 Une baisse de l'indice de Gini observée entre deux dates indique une diminution globale des inégalités. A l'inverse, une élévation de l'indice reflète une augmentation globale des inégalités.

Les **propriétés du coefficient de Gini** sont les suivantes :

- Plus le coefficient est proche de 1, plus la somme dépend des observations les plus grandes.
- Plus le coefficient est proche de 0, moins la somme dépend des observations les plus grandes.

Exemple 2.2.13 La courbe de concentration des achats de la clientèle d'Euromarket est donnée en figure en bas.



Courbe de concentration des achats de la clientèle d'Euromarket

Les 50% plus petits achats représentent à peu près 27% du total des ventes. Il faut considérer les 75% (environ) plus petits achats pour obtenir la moitié du chiffre d'affaire. On peut dire aussi que les 25% clients les plus importants réalisent la moitié du chiffre d'affaires ou encore que le montant de leurs achats est le double de la moyenne.

L'aire totale du carré est égale à 1, et le coefficient de concentration de Gini est le double de l'aire colorée en gris. Il est ici égal à 0.35 : la concentration des achats n'est pas très forte, et la perte de quelques gros clients n'aurait pas d'effet important sur le chiffre d'affaires total.

Chapitre 3

Organisation d'une série statistique bivariée

Il arrive souvent que l'étude statistique d'une population porte simultanément sur plusieurs caractères.
Dans le cas de 2 caractères, on obtient une série statistique bivariée.
Considérons une population d'effectif n sur laquelle on observe deux variables : X et Y.
Pour chaque individu on obtient un couple de valeurs (x_i, y_i).
Une série statistique bivariée est définie par l'ensemble des n couples $\{(x_i, y_i); i \text{ varie de 1 à } n\}$.
On peut envisager 2 méthodes d'organisation.

1. Répartir les couples en classes.
S'il existe des couples identiques, on notera n_{ij} l'effectif du couple (x_i, y_j). On obtient un tableau à double entrées appelé **tableau de contingence**.
Les caractères peuvent être qualitatifs ou quantitatifs.

2. Si les 2 caractères sont quantitatifs discrets et si tous les couples sont différents, on peut représenter chaque couple par un point dans un système d'axes orthogonaux.
On obtient un **nuage de points**.

3.1 Tableau de contingence

Considérons un ensemble de n éléments sur lequel on observe deux caractères :
— le caractère X à k modalités : x_i, i varie de 1 à k
— le caractère Y à r modalités : y_j, j varie de 1 à r.
Si nous désignons par n_{ij} l'effectif du couple (x_i, y_j) nous pouvons définir une distribution observée à deux dimensions : $S((x_i, y_j); n_{ij})$.
Cette distribution peut se représenter par un **tableau de contingence**

X/Y	y_1	y_2	$\dots$	y_j	$\dots$	y_r	
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1r}	n_{1*}
$\vdots$							
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ir}	n_{i*}
$\vdots$							
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kr}	n_{k*}
	n_{*1}	n_{*2}	$\dots$	n_{*j}	$\dots$	n_{*r}	n_{**}

La dernière ligne du tableau de contingence est appelée **marge horizontale** du tableau et par convention et noté n_{*j}, l'étoile indiquant que l'indice supprimé n'intervient pas. De même, la dernière colonne du tableau est appelée **marge verticale** et est notée n_{i*}. Notons enfin que la marge des marges, c'est-à-dire l'effectif total est noté n_{**}.

Un tableau de contingence permet de définir les concepts suivants :

— la **fréquence** du couple $(x_i y_j)$: $f_{ij} = n_{ij}/n$

— les **distributions marginales** : $S_X(x_i, n_{i*})$, $i = 1, \dots, k$ et $S_Y(y_j, n_{*j})$, $j = 1, \dots, r$

— les **distributions conditionnelles** en X pour j fixé ou en Y pour i fixé

— les **fréquences conditionnelles :** $f(x_i|y_j) = n_{ij}/n_{*j}$ et $f(y_j|x_i) = n_{ij}/n_{i*}$

Illustrons cette notation par quelques exemples.

Exemple 3.1.1 /Feuille 6/ :On a interrogé 300 personnes à la sortie d'une grande surface et on a obtenu les résultats suivants

	Achat du produit A	Non achat du produit A	
Homme	30	70	100
Femme	80	120	200
	110	190	300

Notation :
$n_{*2} = 190$, $n_{2*} = 200$, $n_{12} = 70$, $n_{**} = n = 300$.

Quel est le pourcentage de personnes qui ont acheté le produit A ?
Quel est le pourcentage de femmes qui n'ont pas acheté le produit A ?
Parmi ceux qui ont acheté le produit A quel est le pourcentage d'hommes ?
Parmi les hommes quel est le pourcentage des acheteurs ?

Solution :
On a les deux caractères observés : X - Sexe des clients et Y - acheteur ou non.

y_j / x_i	Achat du produit A $x_i\|y_1$	Non achat du produit A $x_i\|y_2$	n_{i*}
Homme $x_1\|y_j$	30	70	100
Femme $x_2\|y_j$	80	120	200
n_{*j}	110	190	300

— la **fréquence** du couple (Homme,acheteur) $= (x_1, y_1)$: $f_{11} = n_{11}/n = 30/300 = 0.1$

— les **distributions marginales** : $S_X(x_i, n_{i*})$, $i = 1, 2$: (Homme, 100) ;(Femme, 200)

$S_Y(y_j, n_{*j})$, $j = 1, 2$: (Acheteur,110) ;(Non acheteur, 190).

Quel est le pourcentage de personnes qui ont acheté le produit A ? : $n_{*1}/n = 110/300$.

Quel est le pourcentage de femmes qui n'ont pas acheté le produit A ? : $n_{22}/n = 120/300$

— les **distributions conditionnelles** en X pour j fixé ou en Y pour i fixé :
la distribution du sexe parmi les acheteurs : (Homme, 30) ;(Femme,80) ;
la distribution du sexe parmi les non acheteurs : (Homme, 70) ;(Femme, 120)
la distribution des acheteur parmi les hommes : (Acheteurs,30) ;(Non acheteurs, 70) ;
la distribution des acheteur parmi les femmes : (Acheteuses,80) ;(Non acheteuses, 120).

— les **fréquences conditionnelles :** $f(x_i|y_j) = n_{ij}/n_{*j}$ et $f(y_j|x_i) = n_{ij}/n_{i*}$

Parmi ceux qui ont acheté le produit A quel est le pourcentage d'hommes ? : $n_{11}/n_{*1} = 30/110$.
Parmi les hommes quel est le pourcentage des acheteurs ? : $n_{11}/n_{1*} = 30/100$.

Si les caractères sont quantitatifs, on peut définir

— les **moyennes** $\bar{x}$ et $\bar{y}$ et les **écarts type** s_X et s_Y des distributions marginales et de chaque distribution conditionnelle (Voir Table 3.1).

	Distribution marginale S_X	Distribution marginale S_Y
Moyenne	$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_{i*}x_i$	$\bar{y} = \frac{1}{n}\sum_{j=1}^{r} n_{*j}y_j$
Variance	$s_X^2 = \left(\frac{1}{n}\sum_{i=1}^{k} n_{i*}x_i^2\right) - \bar{x}^2$	$s_Y^2 = \left(\frac{1}{n}\sum_{j=1}^{r} n_{*j}y_j^2\right) - \bar{y}^2$

TABLE 3.1 : Moyenne d'une série bivariée

— la **covariance** de deux caractères $cov(X, Y)$ ou $\sigma_{X,Y}$, c'est la moyenne pondérée des produits des deux variables centrées : Table 3.2

Série statistique	$cov(X,Y) = \sigma_{X,Y} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$
Donnés groupées	$cov(X,Y) = \sigma_{X,Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}(x_i - \bar{x})(y_j - \bar{y})$.
Tableau de corrélation	$cov(X,Y) = \sigma_{X,Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}(x_i^* - \bar{x})(y_j^* - \bar{y})$.

Table 3.2 - Covariance

Le calcul de la covariance par la formule ci-dessus n'est guère commode : il faut d'abord calculer les moyennes, puis les différences, puis leur produit et enfin la moyenne des produits. On préfère utiliser une autre formule pour le calcul.

Propriété : la covariance est égale à la moyenne des produits moins le produit des moyennes.

Série statistique	$cov(X,Y) = \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}$
Donnés groupées	$cov(X,Y) = \sigma_{X,Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}x_i y_j - \bar{x}\bar{y}.$
Tableau de corrélation	$cov(X,Y) = \sigma_{X,Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}x_i^* y_j^* - \bar{x}\bar{y}.$

Table 3.3 - Covariance - formules simplifiées

Remarque 20 Si les variables sont indépendantes : $cov(X,Y) = 0$.

Remarque 21 La covariance d'une variable avec elle-même est égale à la variance de cette variable : $cov(X,X) = s_x^2$.

Exemple 3.1.2 /Feuille 6/ : Répartition des salaires mensuels par ancienneté et montant

		Répartition des salaires mensuels par ancienneté et montant (en $ liduriens) : Y				
		5000 $	7000 $	9000 $	12 000 $	Ensemble n_{*j}
Répartition des salaires mensuels par ancienneté du salarié (en années) : X	1 an	87	57	11	3	158
	3 ans	39	45	14	19	117
	5 ans	15	36	47	25	123
	8 ans	8	14	24	9	55
	Ensemble n_{i*}	149	152	96	56	453

Quel est le salaire mensuel moyen dans l'entreprise ?
Quelle est l'ancienneté moyenne dans l'entreprise ?
Quel est le salaire mensuel moyen des salariés ayant 3 ans d'ancienneté ?
Quelle est l'ancienneté moyenne des salariés qui ont un salaire mensuel de 9 000$ liduriens ?

Solution

Les caractères observés sont : X = ancienneté d'un salarié et Y = salaires des salariés. D'ici, les distributions marginales sont S_X et S_Y :

X : ancienneté d'un salarié calcul de $\bar{x}$				Y : salaires des salariés calcul de $\bar{y}$			
i	x_i	n_{i*}	$n_{i*}x_i$	j	y_j	n_{*j}	$n_{*j}y_j$
1	1	158	158	1	5000	149	745000
2	3	117	351	2	7000	152	1064000
3	5	123	615	3	9000	96	864000
4	8	55	440	4	12000	56	672000
Total	-	453	1564	Total	-	453	3345000
$\frac{Total}{n}$	-	1	$\bar{x} = 3,453$	$\frac{Total}{n}$	-	1	$\bar{y} = 7384$

Le salaire mensuel moyen dans l'entreprise est $\bar{y} = 7384$
L'ancienneté moyenne dans l'entreprise est $\bar{x} = 3.453$
Le salaire mensuel moyen des salariés ayant 3 ans d'ancienneté :

		Répartition des salaires mensuels par ancienneté et montant (en $ liduriens)				
		5000 $	7000 $	9000 $	12 000 $	Ensemble
Répartition des salaires mensuels par ancienneté du salarié (en années)	1 an	87	57	11	3	158
	3 ans	39	45	14	19	117
	5 ans	15	36	47	25	123
	8 ans	8	14	24	9	55
	Ensemble	149	152	96	56	453

$(39 * 5000 + 45 * 7000 + 14 * 9000 + 19 * 12000)/117 = 7384,62$
L'ancienneté moyenne des salariés qui ont un salaire mensuel de 9 000$ liduriens :

(11+3*14+5*47+8*24)/96=5.

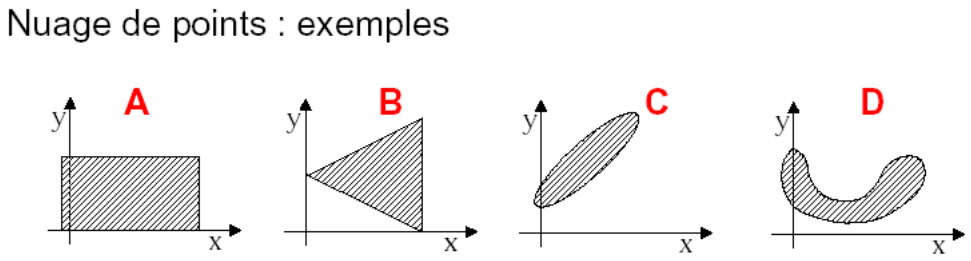
3.2 Nuage de points

Pour 2 caractères quantitatifs discrets, considérons individuellement les n couples (x_i, y_j) obtenus. On peut représenter cette série à l'aide du tableau :

individu i	1	2	...	i	n	
x	x_1	x_2	...	x_i	...	x_n
y	y_1	y_2	...	y_i	...	y_n

et représenter les couples par des points dans le plan.

Remarque 22 Choix des échelles : Dans le cas de deux variables homogènes (exprimées dans la même unité), on prend la même échelle sur les deux axes ; dans le cas de deux variables hétérogènes, il est préférable de représenter les points de la série centrée et réduite ou de choisir des échelles appropriées (automatique avec la plupart des logiciels).
En observant les nuages obtenus, nous pouvons remarquer que le nuage peut être quelconque ou avoir une certaine forme. Dans ce cas on peut supposer qu'il existe une relation entre X et Y, et éventuellement admettre un ajustement par une courbe.



- 1 : corrélation non linéaire
- 2 : absence de liaison en moyenne mais pas en dispersion
- 3 : corrélation linéaire
- 4 : absence de liaison

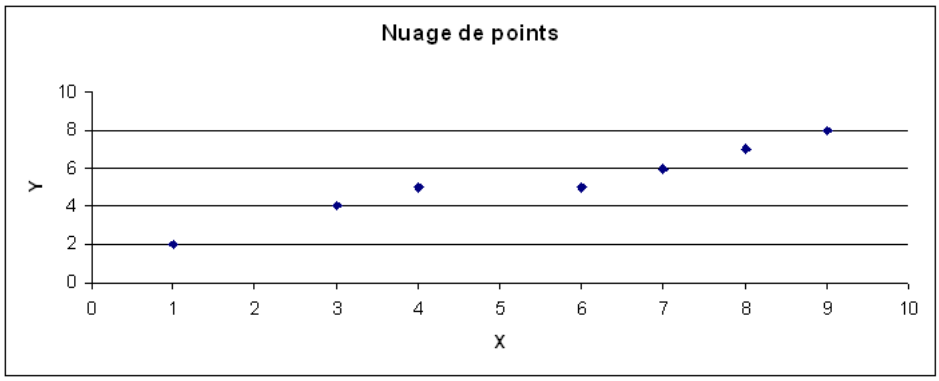
Si on obtient un nuage allongé, on peut supposer une **corrélation linéaire** et proposer un ajustement par une **droite**.

3.2.1 Ajustement linéaire

Le but de l'ajustement linéaire est de tracer une droite qui passe "au plus près" de tous les points du nuage.
Se posent dès lors 3 questions :
1. Comment choisir la droite ?
2. Quand peut - on admettre l'ajustement linéaire ?
3. La corrélation implique-t-elle un lien de cause à effet, c.à.d. la formulation mathématique a -t -elle une signification statistique ?

Exemple 3.2.1 /Feuille 6/ : Pour un ensemble de 7 individus on a obtenu le tableau suivant

individu i	1	2	3	4	5	6	7
X	1	3	4	6	7	8	9
Y	2	4	5	5	6	7	8

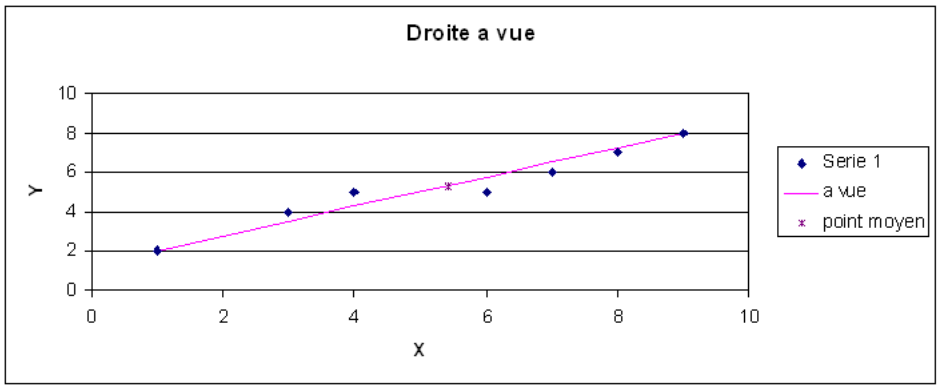


Pour ce nuage de points de forme allongée on peut estimer que l'ajustement linéaire est valable.

Première méthode : **Tracer une droite à vue**
On choisit 2 points du nuage, par exemple $A(1,2)$ et $B(9,8)$ et on trace la droite (AB).
Cette droite est un ajustement linéaire. Cette droite a pour équation $y = ax+b = 0,75x+1,25$.
Voila pourquoi :

$$y = ax + b \quad \left| \begin{array}{l} 2{=}1.\text{a} + \text{b} \\ 8 = 9.\text{a}{+}\text{b} \end{array} \right. \Longrightarrow \left| \begin{array}{ll} \text{b}{=}2\text{-a} & 6{=}8\text{a} \quad \text{a}{=}3/4 = 0{,}75 \\ 8{=}9\text{a}{+}2\text{-a} & \text{b}{=}2\text{-a}{=}2\text{-}3/4{=}5/4{=}1{,}25 \end{array} \right.$$

Cette droite ne passe pas par le point moyen $(\bar{x}, \bar{y}) = (5,43; 5,29)$: pour $x = \bar{x}$, on obtient $y = \bar{y}$.

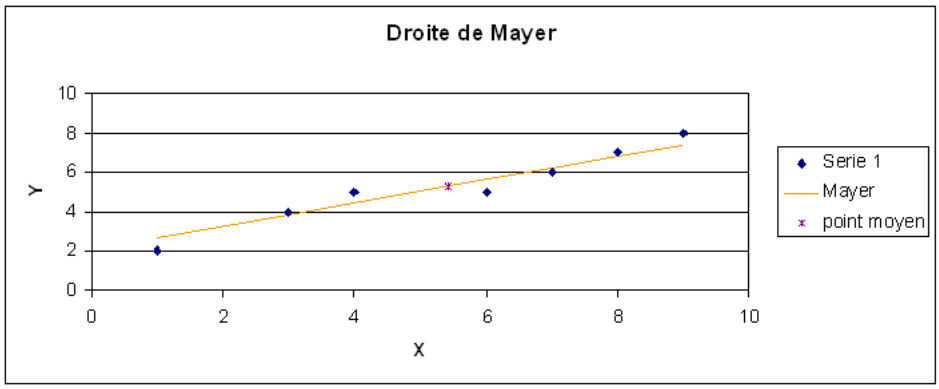


Deuxième méthode : **Droite de Mayer**
On partage l'ensemble des points en 2 sous-ensembles de même effectif si possible. Dans notre cas $E_1 = \{(1,2); (3,4); (4,5)\}$ et $E_2 = \{(6,5); (7,6); (8,7); (9,8)\}$ par exemple.

Pour chacun de ces sous-ensembles on calcule les coordonnées du point moyen.
Respectivement $G_1 = (\bar{x}_1, \bar{y}_1) = (2,67; 3,67)$ pour E_1 et $G_2 = (\bar{x}_2, \bar{y}_2) = (7,5; 6,5)$ pour E_2.
La droite (G_1, G_2) est aussi un ajustement linéaire. C'est la **droite de Mayer.**
L'équation de cette droite est : $y = \frac{17}{29}x + \frac{61}{29} = 0,59x + 2,1$. Voici pourquoi :

$$y = ax + b \quad \left| \begin{array}{l} 3{,}67{=}2{,}67.\text{a} + \text{b} \\ 6{,}5 = 7{,}5.\text{a}{+}\text{b} \end{array} \right. \Longrightarrow \left| \begin{array}{l} a = \frac{2{,}83}{4{,}83} = 0,586207 \\ b = 3,67 - 2,67\frac{2{,}83}{4{,}83} = 2,103448 \end{array} \right.$$

Cette droite passe toujours par le point moyen du nuage $(\bar{x}, \bar{y}) = (5,43; 5,29)$: pour $x = \bar{x}$, on obtient $y = \bar{y}$.

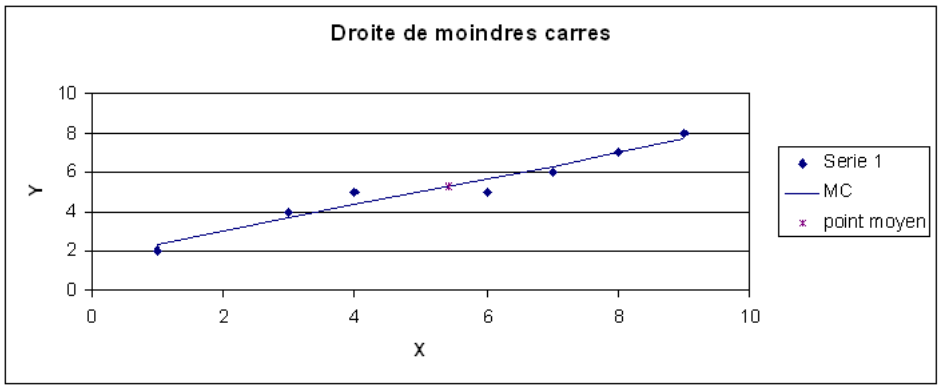


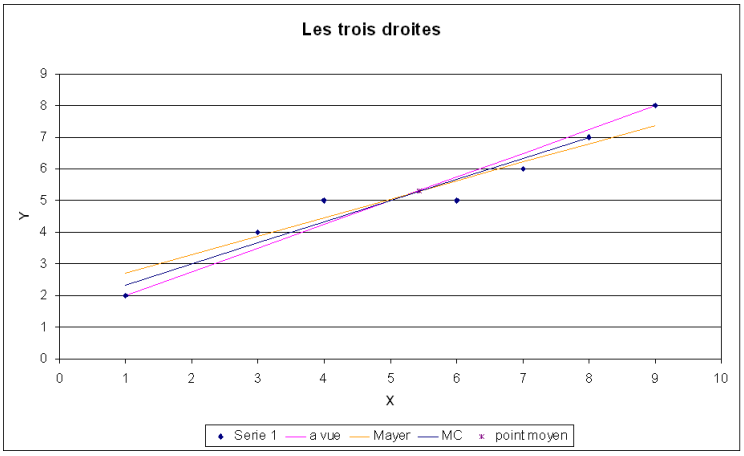
Laquelle de ces deux droites est la "plus proche" du nuage des points observés ?
Pour répondre à cette question nous allons "mesurer" la distance de chacun des points M_i (x_i, y_i) à chacune des droites suivant Y, c.à.d. pour une même valeur de X.
En comparant la somme des carrés des écarts, nous constatons que la droite de Mayer est "plus proche" du nuage.
Nouveau problème mathématique : peut-on trouver une droite telle que la somme des carrés des écarts soit minimale ? Cette droite est la **droite des moindres carrés**

Troisième méthode : **Droite des moindres carrés**

a/ L'équation de la droite des moindres carrés de Y en X est donnée par $D_Y : y = ax + b$ avec $a = \frac{cov(X,Y)}{Var(X)}$ et $b = \bar{y} - a\bar{x}$. D_y passe par le point moyen du nuage.
Cette droite est encore appelée **droite de régression linéaire de Y en X ou de Y par rapport à X**. Cette droite D_Y minimise la somme des carrés des écarts en ordonnée entre les représentations graphiques des valeurs et leurs projections sur D_Y selon (0y), pondérés de leur probabilité.

Pour l'exemple donné on obtient : $D_Y : y = \frac{2}{3}x + \frac{5}{3}$.





b/ On peut définir de la même manière la droite des moindres carrés de X en Y $D_X : x = a'y + b'$ avec $a' = \frac{cov(X,Y)}{Var(Y)}$ et $b' = \bar{x} - a'\bar{y}$. D_X est appelée la droite de régression linéaire de X par

rapport à Y ; cette droite D_X minimise la somme des carrés des écarts en abscisse entre les représentations graphiques des valeurs et leurs projections sur D_X selon $(0x)$, pondérés de leur probabilité. D_X passe par le point moyen du nuage.
On peut en déduire : $D_X : y = \frac{1}{a'}x - \frac{b'}{a'}$ et représenter les deux droites dans le même système d'axes.

Le point de coordonnées $(\bar{x}, \bar{y})$ est commun aux deux droites D_Y et D_X.
Les deux droites D_X et D_Y font entre elles un angle appelé **angle de régression**.

Les droites des moindres carrés donnent une réponse à la première question posée : Comment choisir la droite ? Pour répondre à la deuxième question : 'Quand peut - on admettre l'ajustement linéaire ?', nous allons introduire le **coefficient de corrélation linéaire**.

Coefficient de corrélation linéaire

Le coefficient de corrélation r est un indicateur de dépendance entre deux phénomènes. Ce concept est très utile dans la gestion et l'administration des entreprises. Il permet d'entrevoir, puis de vérifier l'existence d'un bien entre des phénomènes tel que les salaires et les prix, l'absentéisme et taux de primes, les accidents du travail et les heures supplémentaires. . . etc.

De façon graphique, le coefficient de corrélation indique le plus ou moins grand de degré de rapprochement des deux droites de régression.

Il est défini comme étant égale à la racine carrée du produit de la pente des deux droites de régression : $r^2 = a \,.\, a'$. Ce qui donne la formule :

$$r(X,Y) = \frac{cov(X,Y)}{\sqrt{Var(X) \,.\, Var(Y)}} = \frac{cov(X,Y)}{s_X \,.\, s_Y}$$

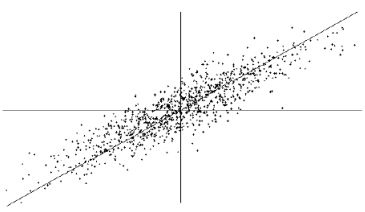
Exemple 3.2.1 : Le coefficient de corrélation de l'exemple est

$$r(X,Y) = \frac{cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{X,Y}}{s_X s_Y} = \frac{4,73}{\sqrt{7,10 * 3,35}} = 0,97.$$

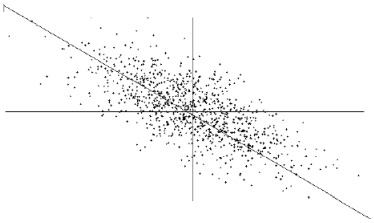
Propriété : Le coefficient de corrélation :

- Est un nombre sans dimension entre 0 et ± 1 ;
- il peut être positif, nul ou négatif.

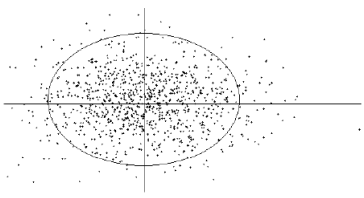
$$-1 \leq r(X,Y) \leq 1$$



Coefficient de corrélation proche de 1



Coefficient de corrélation proche de -1



Coefficient de corrélation proche de 0

En pratique on considère souvent qu'il y a corrélation linéaire si :

$$-1 \leq r(X,Y) \leq -0,87 \quad \text{ou} \quad 0,87 \leq r(X,Y) \leq 1$$

Interprétation de la corrélation et de la régression :

• Lorsque les points du nuage ne sont pas alignés, le coefficient de corrélation r est, en valeur absolue, inférieur à 1. Donc $-1 < r < 1$. Les deux droites de régressions sont alors distinctes ($aa' < 1$).

• La fidélité de la représentation du nuage des points par les droites de régression est fonction de la valeur des coefficients de corrélation. Plus cette dernière en valeur absolue s'approche de 1, plus cette fidélité est importante.

• si r est proche de 1, les deux phénomènes sont en relation étroite et leur sens de variation est identique : à un accroissement de x correspond un accroissement de y.

Comme par exemple l'évolution des salaires et des prix.

• Si r est proche de -1, les deux phénomènes sont en relation étroite, et leur sens de variation est inverse. Autrement dit un accroissement de x correspond à une diminution de y.

• Si r est compris entre -0,5 et 0,5, il n'y a pas de véritable relation linéaire entre x et y. Cela peut provenir d'une indépendance ou d'une relation non linéaire entre les deux phénomènes x et y.

Le nuage de points est dans ce cas très indicatif.

En règle générale, la corrélation :

— Est bonne si $|r| \geq 0,8$.

— Est moyenne si $0,5 < |r| < 0,8$.

— Est mauvaise si $|r| < 0,5$.

Le calcul, en premier, du coefficient de corrélation linéaire permet de justifier l'ajustement affine éventuel. Lorsque la corrélation linéaire est forte, les méthodes d'ajustement affine peuvent alors être utilisées. Les points sont alignés si et seulement si $|r(X,Y)| = 1$.
Les deux droites D_X et D_Y sont confondues si et seulement si $|r(X,Y)| = 1$. Lorsque le coefficient de corrélation est proche de 1 en valeur absolue, on parle de bon ajustement. Inversement, lorsque ce coefficient est proche de 0, on parle de mauvais ajustement, et dans ce cas, les variables X et Y sont presque non covariées (i.e. elles n'ont rien à voir entre-elles).
Ainsi, ce coefficient rend compte de la validité de la régression linéaire.

Remarque 23 Il convient d'interpréter avec prudence la corrélation qui existe entre deux phénomènes. L'existence d'un corrélation forte n'implique pas nécessairement un lien de causalité. Le lien peut être fortuit ou il peut exister une cause commune.
Important : La covariance et le coefficient de corrélation ne permettent de mettre en évidence qu'une relation linéaire entre X et Y.
Si deux variables sont statistiquement indépendantes (aucun lien), la corrélation est nulle, mais l'inverse est faux : il peut exister un lien autre que linéaire entre elles.

Résidus

On appelle **résidu** e_i le terme défini par la différence entre la valeur observée y_i et l'ordonnée du point de la droite de régression d'abscisse x_i, pour $i = 1, \ldots, n$.

$$e_i = y_i – (ax_i + b), \quad i = 1, \ldots, n$$

Ils mesurent la proximité entre la droite et les points, et ce sont les plus petites erreurs possibles suivant ce critère. Plus la moyenne de leurs carrés est faible, plus la droite est proche des points.

La série des résidus possède les propriétés suivantes :

- sa moyenne est nulle ;
- sa variance est égale à $s_e^2 = (1–r^2)s_y^2$, où r est le coefficient de corrélation des couples (x_i, y_i) $i = 1, \ldots, n$, et s_y^2 la variance des y_i, $i = 1, \ldots, n$;
- le coefficient de corrélation entre les x_i et les e_i est égal à 0.

Coefficient de détermination ou d'explication R^2

Définie par

$$R^2 = r^2(X, Y) \quad /\text{Fonction Excel – COEFFICIENT.DETERMINATION}/$$

le coefficient de détermination ou d'explication est un indice de mesure de la qualité d'ajustement du modèle. Il détermine à quel point l'équation de régression est adaptée pour décrire la distribution des points.

Le coefficient de détermination R^2 s'exprime en pourcentage. Il indique la part de variation (dispersion) expliquée par la liaison linéaire, c'est-à-dire par la droite de régression.

La proportion de la variation totale de Y inexpliquée par la droite de régression c'est-à-dire par la connaissance de la variable explicative X est : $1 - R^2$.

Si le R^2 est nul, cela signifie que l'équation de la droite de régression détermine 0% de la distribution des points. Cela signifie que le modèle mathématique utilisé n'explique absolument pas la distribution des points.

Si le R^2 vaut 1, cela signifie que l'équation de la droite de régression est capable de déterminer 100% de la distribution des points. Cela signifie, que le modèle mathématique utilisé, ainsi que les paramètres a et b calculés sont ceux qui déterminent la distribution des points.

Cela se traduit de manière graphique selon la relation suivante : plus le coefficient de détermination se rapproche de 0, plus le nuage de points est diffus autour de la droite de régression. Au contraire, plus le R^2 tend vers 1, plus le nuage de points se rapproche de la droite de régression. Quand le points sont exactement alignés sur la droite de régression, $R^2 = 1$.

Note : Le R^2 n'est le carré du r que dans le cas particulier de la régression linéaire. Dans les autres régressions (logarithmique, exponentielle, puissance, etc.) ce n'est pas le cas. C'est pour éviter cette confusion facile qu'on note habituellement le r du coefficient de corrélation en minuscule, et celui du coefficient de détermination en majuscule.

Comme le coefficient de détermination R^2 représente la fraction de la variance de y "expliquée" par la corrélation de y avec x, un coefficient de corrélation $r = 0,9$ correspond à un coefficient de détermination $R^2 = 0,81$ Sela signifie, que 81% de la variance de y est expliquée par la corrélation. La la corrélation est bonne.

Un coefficient de corrélation $r = 0,5$ correspond à un coefficient de détermination $R^2 = 0,25$. Dan ce cas, seulement 25% de la variance de y est expliquée par la corrélation. On a une corrélation est mauvaise.

Pour l'exemple 3.2.1 le coefficient de détermination R^2 est :

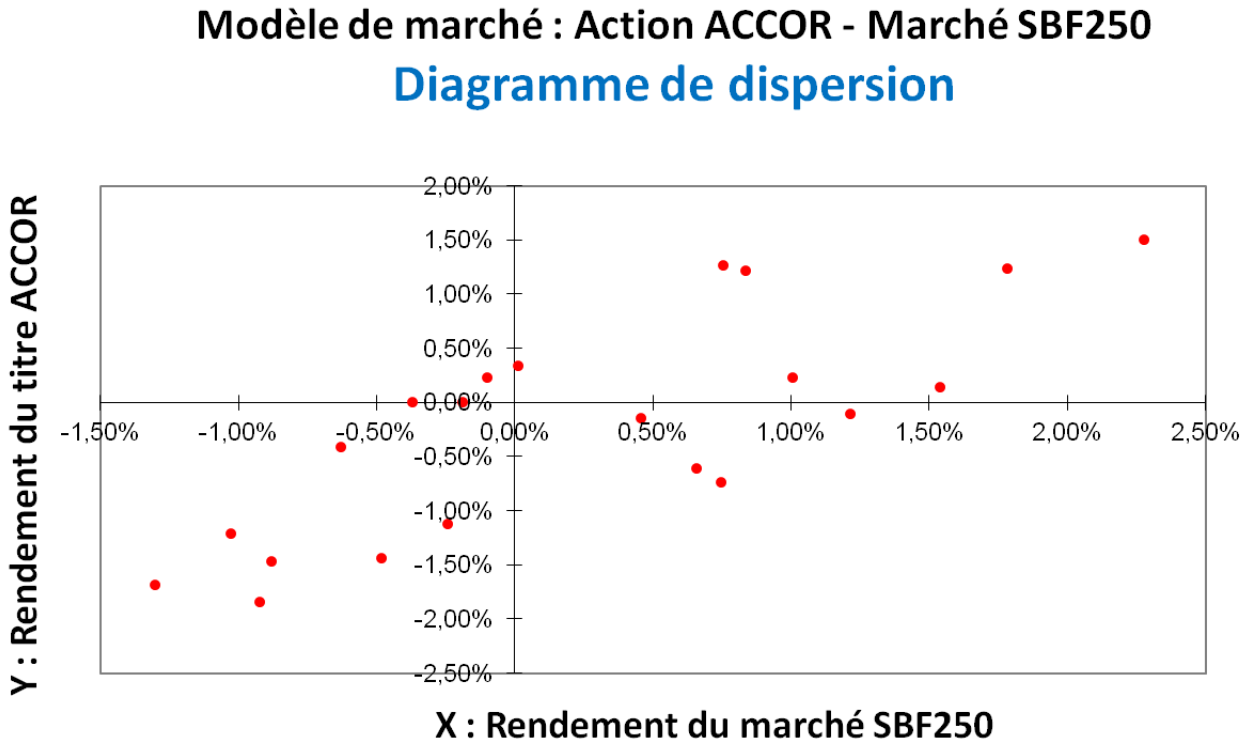
$$R^2 = 94\%$$

Cela signifie, que la part de la variabilité des points observés qui est expliquée linéairement par le modèle de régression est 94%. La mesure de la qualité du modèle d'ajustement de l'exemple 3.2.1 prouve une bonne corrélation.

Exemple 3.2.2 Exemple - Modèle de marché Comportement d'un titre relativement aux mouvements du marché

On a relevé 21 couples d'observations des mouvements du cours de l'action ACCOR et du marché SBF-250 (indice boursier) correspondant aux variations journalières pendant un mois.

X : marché SBF-250 $\Longleftarrow$ Variable explicative
Y : cours de l'action ACCOR $\Longleftarrow$ Variable à expliquer



Les points ont tendance à s'aligner selon une droite (pente positive), une liaison linéaire entre les variations du titre et celles du marché semble très plausible.

L'équation de la droite de régression (D) est :

$$(D) : y = ax + b$$

On détermine les coefficients a et b.Le coefficient a (pente de la droite (D)) est

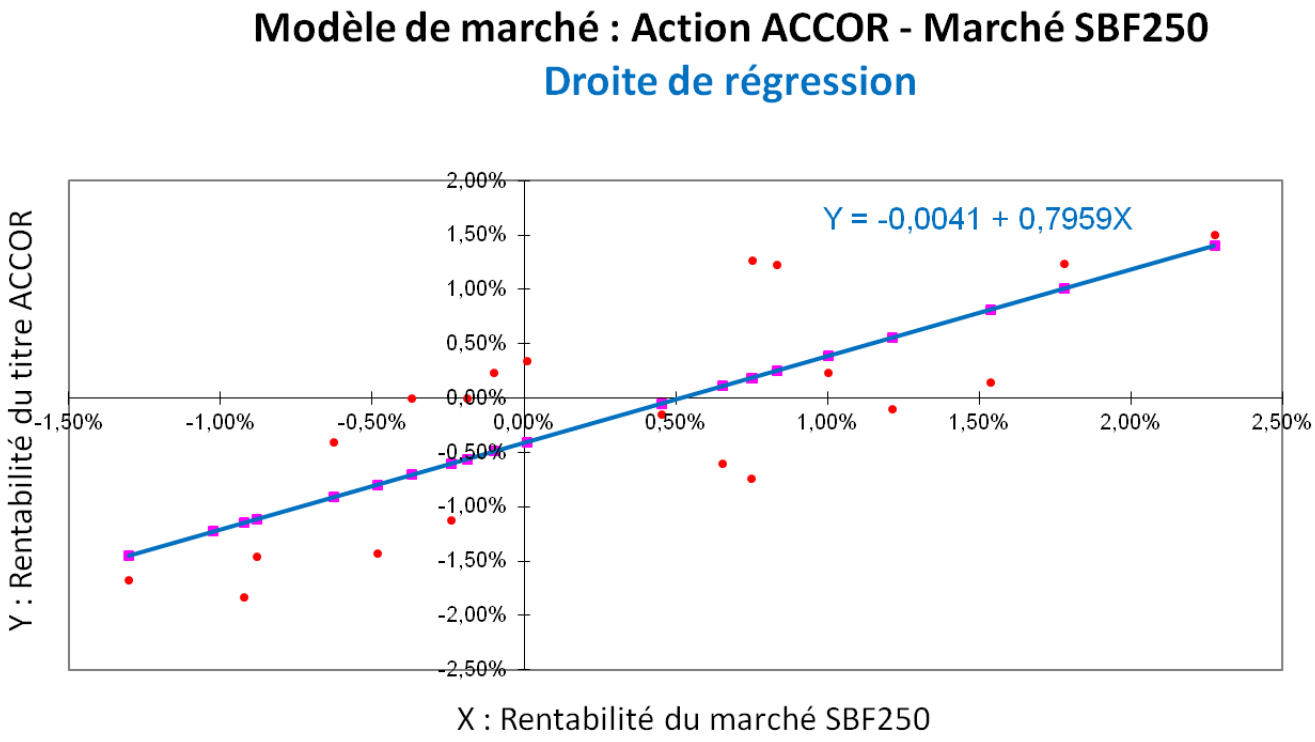
$$a = Cov(X, Y)/s_X^2 = -0,0041$$

Fonction Excel : $PENTE(y; x)$

Le coefficient b (ordonnée à l'origine de la droite (D)) est

$$b = \bar{y} - a\bar{x} = 0,7959$$

Fonction Excel : $ORDONNEE.ORIGINE(y; x)$



La droite d'ajustement (D) est

$$(D) : Y = -0,0041 + 0,7959X$$

On analyse l'ajustement linéaire. Le coefficient $b = -0,41\%$ est une estimation de la rentabilité fixe du titre. Le coefficient $a = 79,59\%$ estime la volatilité du titre / marché. Le coefficient de détermination $R^2 = 61,14\%$ mesure la qualité du modèle d'ajustement et donne la part de la variabilité du titre ACCOR qui est expliquée ''linéairement" par le modèle de régression : modèle de marché. La valeur de R^2 donne le risque systématique du titre. La valeur complémentaire $1 - R^2$ - le risque spécifique du titre.

Ajustement linéaire de distributions groupées

Les distributions des deux caractères observés se représentent par un tableau de contingence

X/Y		y_1	y_2	$\dots$	y_j	$\dots$	y_r	$\sum n_{i*}$
		y_1^*	y_2^*	$\dots$	y_j^*	$\dots$	y_r^*	
x_1	x_1^*	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1r}	n_{1*}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	x_i^*	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ir}	n_{i*}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	x_k^*	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kr}	n_{k*}
$\sum n_{*j}$		n_{*1}	n_{*2}	$\dots$	n_{*j}	$\dots$	n_{*r}	n_{**}

où

$$\begin{array}{ll} y_i^* & \text{est le centre de l'intervalle } i \text{ de } y \\ x_j^* & \text{est le centre de l'intervalle } j \text{ de } x \end{array}$$

Les paramètres de la droite de moindres carrées $D : y = ax + b$ sont calculés du système

$$\left| \begin{array}{l} \sum_{i=1}^{r} y_i n_{i*} = a \sum_{i=1}^{r} n_{i*} + b \sum_{j=1}^{k} x_j n_{*j} \\ \sum_{i,j=1}^{r,k} y_i x_j n_{ij} = a \sum_{j=1}^{k} x_j n_{*j} + b \sum_{j=1}^{k} x_j^2 n_{*j} \end{array} \right.$$

Les expressions des coefficients a et b de la droit D sont :

$$\begin{array}{rcl} a & = & \dfrac{\sum_{i,j=1}^{n_{**}} x_j y_i n_{ij} - n_{**} \overline{xy}}{\sum_{j=1}^{n_{**}} x_j^2 n_{*j} - n_{**} \overline{x}^2} \\ b & = & \bar{y} - a\bar{x} \end{array}$$

Exemple 3.2.3 Donner l'ajustement linéaire entre la production y et les immobilisations x des entreprises industrielles

Production en mille d'euros y	Capital en million d'euros x										
		10 - 16	16 - 22	22 - 28	28 - 34	34 - 40	40 - 46	46 - 52	52 - 58	58 - 64	
	y_i^*/x_j^*	13	19	25	31	37	43	49	55	61	$\sum n_{i*}$
300 - 320	310	1									1
320 - 340	320		3								3
340 - 360	350		2	3							5
360 - 380	370			5	3	2					10
380 - 400	390				8	10					18
400 - 420	410				2	5	3				10
420 - 440	430						4				4
440 - 460	450						2	7			9
460 - 480	470								9		9
480 - 500	490									2	2
$\sum n_{*j}$		1	5	8	13	17	9	7	9	2	71

Pour calculer les coefficients a et b de la droite d'ajustement on compose la table

i/j	y_i^*	n_{i*}	x_j^*	n_{j*}	$y_i n_{i*}$	$x_j n_{*j}$	x_j^2	$x_j^2 n_{*j}$	$y_i^2 n_{i*}$	$y_i x_j n_{ij}$
1	310	1	13	1	310	13	169	169	96100	4030
2	330	3	19	5	990	95	361	1805	326700	18810
3	350	5	25	8	1750	200	625	5000	612500	13300
4	370	10	31	13	3700	403	961	12493	1369000	26250
5	390	18	37	17	7020	629	1369	23273	2737800	46250
6	410	10	43	9	4100	387	1849	16641	1681000	34410
7	43	4	49	7	1720	343	2410	16807	739600	27380
8	450	9	55	9	4050	495	3025	27225	1822500	96720
9	470	9	61	2	4230	122	3721	7442	1988100	144300
10	490	2			980				480200	25420
11										75850
12										52890
13										73960
14										38700
15										154350
16										232650
17										59780
$\sum$	$\times$	71	$\times$	71	28850	2687	14481	110855	11853500	1125050

Les moyennes des distributions marginales sont

$$\begin{aligned} \bar{x} &= \frac{\sum_{j=1}^{k} x_j^* n_{*j}}{n} = \frac{2687}{71} = 37,845 \quad \text{million } € \\ \bar{y} &= \frac{\sum_{i=1}^{r} y_i^* n_{i*}}{n} = \frac{28850}{71} = 406,338 \quad \text{mille } € \end{aligned}$$

Les coefficients a et b de la droite de régression D sont

$$\begin{aligned} a &= \frac{\sum_{i,j=1}^{n_{**}} x_j y_i n_{ij} - n_{**}\overline{xy}}{\sum_{j=1}^{n_{**}} x_j^2 n_{*j} - n_{**}\overline{x}^2} = \frac{1125050 - 71 \times 37,845 \times 406,338}{110855 - 71 \times 37,845^2} = \frac{33221,9}{9165,68} = 3,625 \quad \text{million } € \\ b &= \bar{y} - a\bar{x} = 406,338 - 3,625 \times 37,845 = 269,150 \end{aligned}$$

La droit de régression a l'équation

$$y = 3,625x + 269,150$$

On voit qu'un accroissement des immobilisation d'un million euros donne un accroissement moyen de 3,625 million d'euros de la production.

Le coefficient de corrélation r_{xy} de Bravais qui caractérise la liaison de x et y est

$$\begin{aligned} r_{xy} &= \frac{n_{**} \sum_{i,j}^{n_{**}} y_i x_j n_{ij} - \sum_{j=1}^{k} x_j n_{*j} \sum_{i=1}^{r} y_i n_{i*}}{\sqrt{\left[n_{**} \sum_{j=1}^{n_{*j}} x_j^2 n_{*j} - \bar{x}^2\right] \left[n_{**} \sum_{i=1}^{n_{i*}} y_i^2 n_{i*} - \bar{y}^2\right]}} \\ &= \frac{71 \times 1125050 - 2687 \times 28850}{\sqrt{[71 \times 110588 - 2687^2]\,[71 \times 11853500 - 28850^2]}} \\ &= \frac{79,801}{\sqrt{6611,352}} = 0,981 \end{aligned}$$

Le coefficient de Bravais proche de 1 montre une liaison linéaire forte entre les immobilisations et la production pour les 71 entreprises industrielles étudiées.

Le coefficient de détermination $R^2 = 0,9632$ montre que 96,32% des variations de la production sont dues aux variations des immobilisations corporelles.

Corrélation - Causalité

Donnons maintenant la réponse de la question 3.

Il convient d'interpréter avec prudence la corrélation qui existe entre deux phénomènes ; l'existence d'une corrélation forte n'implique pas nécessairement un lien de causalité. Le lien peut être fortuit ou il peut exister une cause commune.

Par exemple, il y a une forte corrélation entre les ventes de glaces et celles de lunettes de soleil.

Il n'y a pas cependant de relation de cause à effet entre ces deux phénomènes.

La forte corrélation s'explique ici par l'existence d'une cause commune : le soleil qui donne chaleur et lumière.

Lorsque, on utilise le logiciel Excel, on travaille avec les données non-groupées (individuelles). Les commandes pour effectuer un ajustement linéaire sont :

Menu : Outils

➡ Utilitaire d'analyse

➡ Analyse de la covariance

On obtient ainsi la matrice des variances et covariances.

➡ Analyse de la corrélation

On obtient la matrice des corrélations.

Une représentation graphique on obtient par

Diagramme de dispersion

Fonction Excel – type de graphique : Nuage de points

Droite d'ajustement linéaire

Option du graphique : Ajouter une courbe de tendance

Equation de la droite d'ajustement - Possibilité de faire des prévisions

Ajustement linéaire par changement de variable

Si le nuage de points (x_i, y_i) ne permet pas d'envisager un ajustement par une droite, il est parfois possible d'effectuer un changement de variable sur X, sur Y ou sur les deux qui permet d'aboutir à un ajustement linéaire entre les nouvelles variables.

1. Modèle semi-logarithmique
a/On remplace les couples (x_i, y_i) par les couples (x_i, z_i) ou on pose $z_i = \ln y_i$.
Si les couples (x_i, z_i) forment un nuage allongé de points, on peut envisager une droite d'ajustement d'équation : $z = ax + b$ ce qui donne entre X et Y : $\ln y = ax + b$ et finalement

$$y = h\, e^{ax}, \quad h = e^b$$

ce qui correspond à une croissance exponentielle.

b/On remplace les couples (x_i, y_i) par les couples (u_i, y_i) ou on pose $u_i = \ln x_i$.
Si les couples (u_i, y_i) forment un nuage allongé de points, on peut envisager une droite d'ajustement d'équation : $y = au + b$ ce qui donne entre X et Y :

$$y = a \ln x + b.$$

2. Modèle doublement logarithmique
On remplace les couples (x_i, y_i) par les couples (u_i, z_i) ou on pose $u_i = \ln x_i$ et $z_i = \ln y_i$.
Si les couples (u_i, z_i) forment un nuage allongé de points, on peut envisager une droite d'ajustement d'équation : $z = au + b$ ce qui donne entre X et Y : $\ln y = a \ln x + b$ et finalement

$$y \;=\; e^b\, x^a = h\, x^a.$$

Exemple 3.2.4 /Feuille 6/ :

On considère le tableau de données ci-dessous :

x	1	2	3	4	5	6	7	8	9	10
y	1.65	2.72	4.48	7.39	12.18	20.09	33.12	54.6	90.02	148.41

1) Représenter graphiquement les couples (x_i, y_i) $i = 1, \ldots 10$.
2) Effectuer la régression linéaire de Y par X à l'aide des résultats fournis ci-dessous.

Sommes	
des observations x :	55
des observations y :	374.66
des carrés x^2 :	385
des carrés y^2 :	34843.99
des produits xy :	3194.45

3) On prendra pour valeurs $a = 13.74$ et $b = -38.12$ dans la droite de régression $y = ax + b$. Calculer la valeur estimée de Y pour $x = 5$ et $x = 12$. Représenter la droite sur le graphique. /Voir Fig 3.1/ Calculer le coefficient de corrélation.

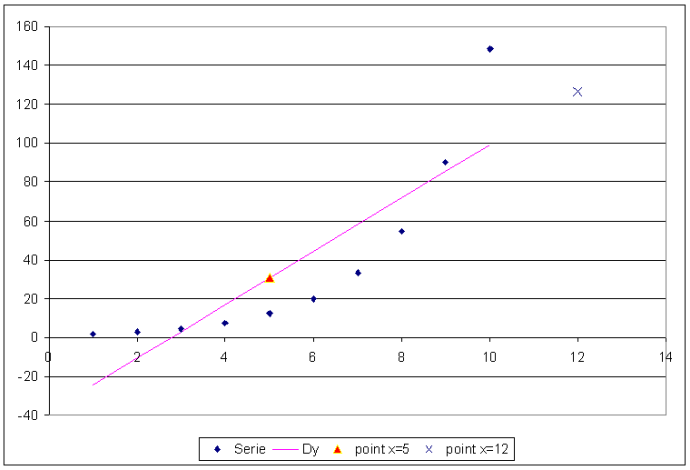


FIGURE 3.1 : $D_y : y = ax + b$, $a = 13.74$, $b = -38.12$

4) Effectuer un ajustement linéaire par changement de la variable Y en utilisant le modèle semi-logarithmique $z = \ln y$. Utiliser les résultats fournis ci-dessous

Sommes	
des observations z :	27.50
des carrés z^2 :	96.25304
des produits xz :	192.503

Présenter graphiquement les couples (x_i, z_i), $i = 1, \ldots, 10$.
5) On prendra pour valeurs $a = 0,499946$, $b = 0,000431$ dans la droite de régression $D_z : y = e^b\, e^{ax}$. Calculer la valeur estimée de y pour $x = 5$ et $x = 12$. Représenter la droite de régression sur le graphique. /Voir Fig. 3.2/

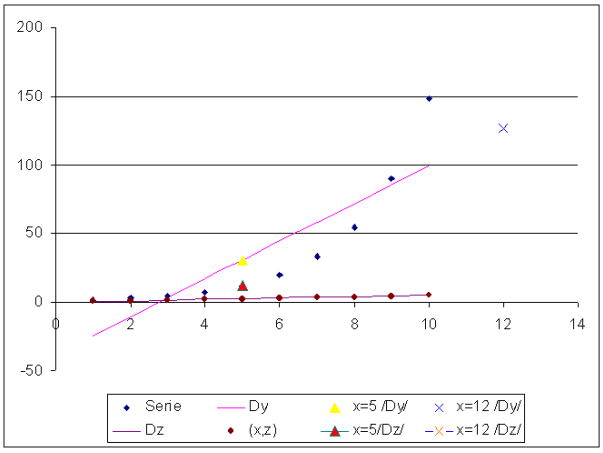


FIGURE 3.2 : $D_z : y = e^b\, e^{ax}$, $a = 0,499946$, $b = 0,000431$

Chapitre 4

Séries chronologiques

Les données statistiques particulières en sens que les observations sont régulièrement échelonnées dans le temps et constituent ce que l'on appelle des séries chronologiques. Ce genre de données est bien connu en économie : la quasi-totalité des indices de prix, de production etc. sont calculés régulièrement par l'INSEE ou d'autres établissements. Elles sont fréquentes aussi en gestion : surveillance du niveau des stocks, suivi des ratios d'une entreprise etc.... Leur particularité vient de l'introduction du temps dans l'analyse de ces données : on étudie une suite de couples de la forme (t, x_t), où x_t est l'observation de la variable à l'instant t.

Une série chronologique permet d'étudier l'évolution d'une variable dans le temps comme, par exemple, l'évolution du prix du baril de pétrole au cours des dix dernières années.

On peut présenter une série chronologique au moyen d'un tableau dans lequel la première colonne représente le temps et les autres colonnes la variable qui évoluent en fonction du temps. Voyons un exemple de tableau pour une variable qualitative :

Gagnant de la Coupe Grey, 2003-2007

Année	Équipe gagnante
2003	Eskimos
2004	Argonauts
2005	Eskimos
2006	Lions
2007	Roughriders

Dans le cas d'une série chronologique portant sur une variable mesurée selon une échelle d'intervalles ou de rapports, le tableau qui la représente porte habituellement un titre commençant par : « Évolution de ...». Le tableau ci-dessous montre bien cette particularité.

Table 4.1 : Évolution de l'assistance des spectateurs à la Coupe Grey, 2003-2007

Année	Assistance
2003	50 909
2004	51 242
2005	59 157
2006	44 786
2007	52 230

Il est aussi possible de tracer un graphique qui représente l'évolution d'une variable à travers le temps. Il existe deux types de graphique pour représenter les séries chronologiques : les histogrammes et les lignes brisées. Ce type de graphique permet de comparer facilement les différences entre les valeurs. Voici un exemple de ligne brisée :

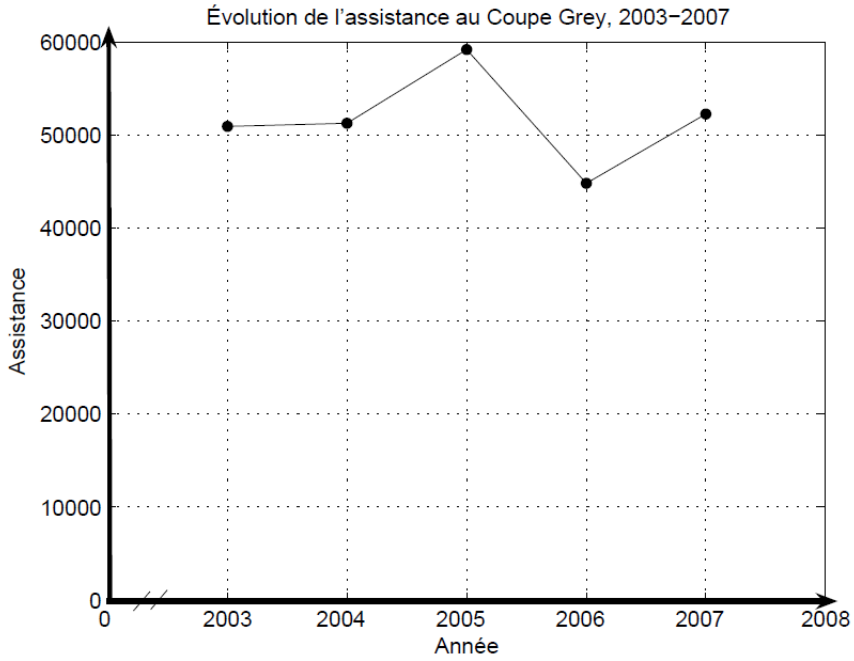


FIGURE 4.1 : Exemple de graphique à la linge brisée.

Les principales représentations utilisent :
- un diagramme polaire (notamment pour des données mensuelles)
- un diagramme triangulaire
- un repère cartésien (les temps figurent sur l'axe des abscisses).

Le traitement de séries chronologiques est mathématiquement identique à celui des séries statistiques. Il existe cependant des différences. Une différence essentielle en ce sens est que l'une des variables, à savoir *le temps* est *déterministe* et non aléatoire. Ce qui importe dans les séries chronologiques n'est pas l'ensemble des valeurs $\{x_1, x_2, \ldots, x_t\}$ de l'échantillon, mais la succession de ces valeurs. De plus, les séries chronologiques bénéficient de modèles particuliers fondés sur l'idée qu'il y a superposition de phénomènes temporels simples.

4.1 Analyse des séries chronologiques

4.1.1 Description d'une série chronologique

Exemple 4.1.1 La responsable « Transports - Livraisons » de l'entreprise Yopmilk produisant des produits laitiers frais (yaourts, fromages frais,...) dispose pour les trois années précédentes des statistiques d'expédition suivantes, concernant les yaourts aromatisés :

	1988	1989	1990
Janvier	2450	2525	2630
Février	2470	2530	2635
Mars	2550	2800	2700
Avril	2540	2600	2710
Mai	2800	2900	3000
Juin	2850	2950	3050
Juillet	3140	3250	2800
Août	3150	3300	3350
Septembre	2800	2900	3000
Octobre	2540	2660	2710
Novembre	2470	2530	2635
Décembre	2200	2300	2400

Afin d'améliorer la qualité des « transports » de la société, le responsable de ce service souhaite :
- connaître les caractéristiques de l'évolution des ventes au cours d'une année et ce, afin de maîtriser les phénomènes conjoncturels
- connaître la prévision des expéditions pour 1991.

Les données sont représentées dans le repère suivant (figure 4.2) par les points : $M_i(t_i, y_i)$ (les mois t_i étant numérotés de 1 à 36).

Les variations d'un phénomène (ou des données observées) résultent de la composition (par addition ou par multiplication) de 4 composantes (ou de 4 causes)

- la tendance générale ou mouvement de longue durée (appelé aussi trend) noté T : pour Yopmilk, d'après la figure 4.2, les expéditions semblent augmenter sur les 3 années
- les variations saisonnières notées S : augmentations au mois d'août et diminutions au mois de décembre pour Yopmilk
- les variations cycliques notées C (ces variations peuvent avoir pour période plusieurs années) : c'est ce qu'on appelle le cycle de Kondratiev, qui résulte du fait que, suivant la théorie de Kondratiev, à une période de prospérité économique succède mécaniquement une période de dépression.
- les variations accidentelles (dues, par exemple, à des grèves, des pannes, ...) notées A : pour Yopmilk diminution non expliquée au mois de juillet 1990.

4.1.2 Détermination de la tendance générale

a/ **Ajustement linéaire**

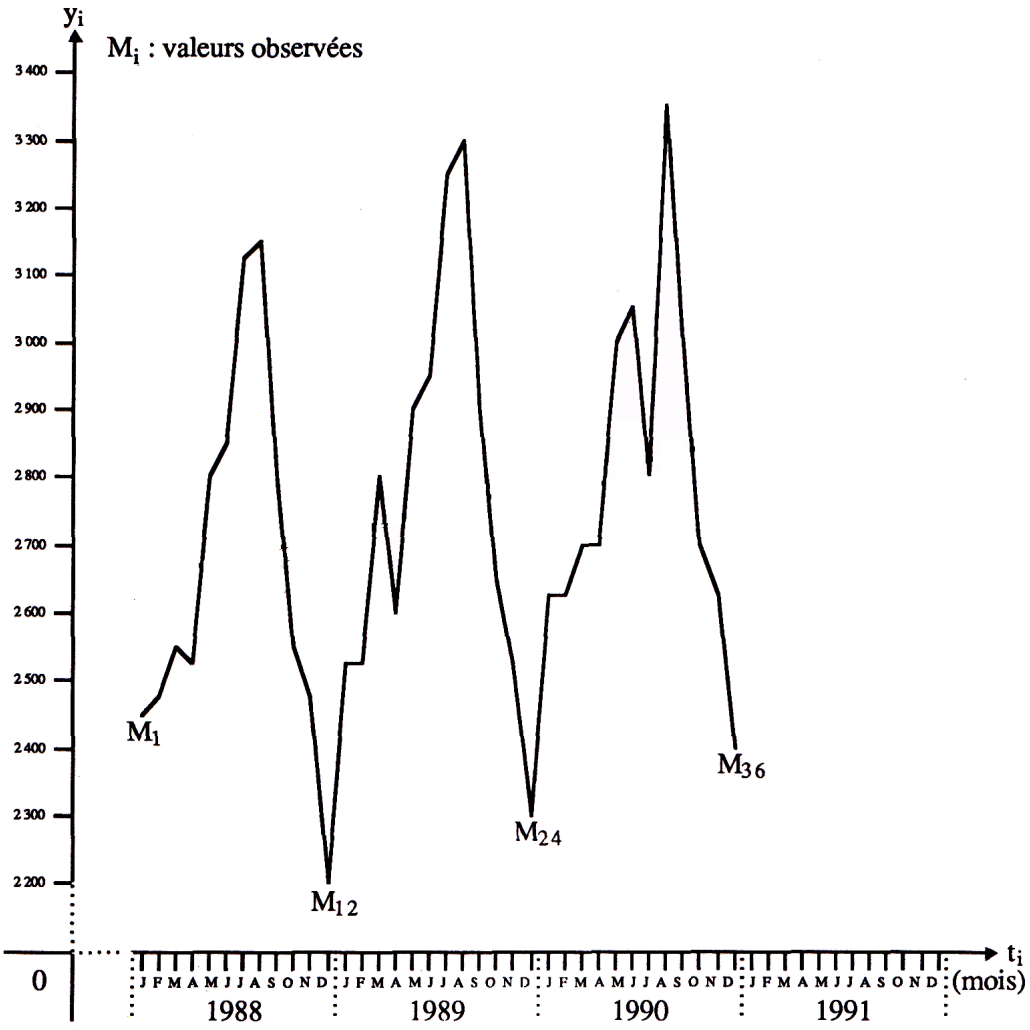
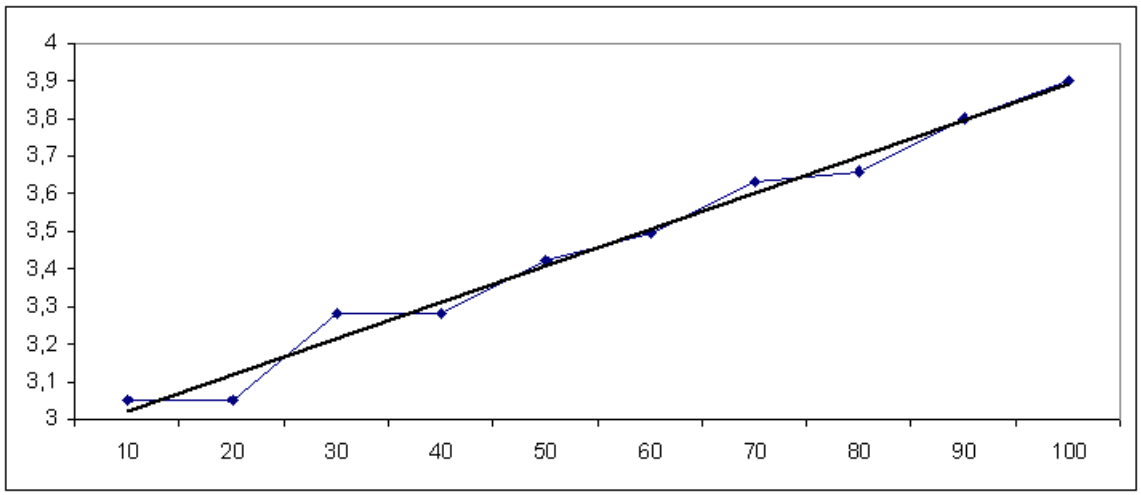


FIGURE 4.2 : Expédition des yaourts aromatisés

Lorsque les points M_i (t_i, y_i) pour $i = 1, \ldots, n$, qui représentent la série chronologique semblent alignés (c'est-à-dire qu'il n'existe pas ou peu de variations saisonnières), le trend peut être représenté par une droite obtenue par l'une des méthodes de l'ajustement affine : droite de Mayer ou droite des moindres carrés.



b/ **Méthode des moyennes mobiles**

Cette méthode est **fréquemment** utilisée pour déterminer le trend d'une série chronologique : la tendance générale, n'étant plus perturbée par les variations saisonnières, apparaît plus nettement.

Il faut pour cela former des groupes de longueur l de valeurs observées. Chaque groupe se déduit du précédent en supprimant la première donnée et en introduisant la donnée suivante de la série. Les points qui représentent alors la nouvelle série ont pour :

- abscisse : la date centrale (médiane) du groupe de valeurs considérées d'où l'intérêt de choisir, lorsque l'exemple s'y prête, un nombre impair $l = 2k + 1$ de dates par groupe)

- ordonnée : la moyenne arithmétique des y_i du groupe.

- Nombre impaire $l = 2k+1$ de dates du groupe : On appelle moyenne mobile centrée de longueur impaire $l = 2k + 1$ à l'instant i la valeur moyenne y'_i des observations $y_{i-k}, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_{i+k}$:

$$y'_i = (y_{i-k} + \ldots + y_{i-1} + y_i + y_{i+1} + \ldots + y_{i+k})/l$$

- Nombre paire $l = 2k$ de dates du groupe. Pour que i soit l'indice central (médian) lorsque le nombre l de valeurs est pair, un ajustement est nécessaire pour se ramener à un nombre impair d'observations : On appelle moyenne mobile centrée de longueur paire $l = 2k$ à l'instant i la valeur moyenne y'_i des observations $y_{i-k}, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_{i+k}$, la première et la dernière étant pondérées par 0,5 :

$$y'_i = (0,5y_{i-k} + y_{i-k+1} + \ldots + y_{i-1} + y_i + y_{i+1} + \ldots + y_{i+k-1} + 0,5y_{i+k})/l$$

Ainsi pour une moyenne mobile sur 12 mois, on utilise le calcul suivant :

$$y'_i = \frac{\left(\frac{y_{i-6}}{2}\right) + y_{i-5} + \ldots + y_i + \ldots + y_{i+5} + \left(\frac{y_{i+6}}{2}\right)}{12}$$

La première valeur d'une moyenne mobile de longueur 4 (= 2*2) ou 5 (= 2*2 + 1) que l'on

peut calculer, est à l'instant $t = 3$, puisque la première observation connue est x_1 :

$$\begin{aligned} y'_3 &= (0,5x_1 + x_2 + x_3 + x_4 + 0,5x_5)/4 \quad (l = 4) \\ y'_3 &= (x_1 + x_2 + x_3 + x_4 + x_5)/5 \quad (l = 5) \end{aligned}$$

De façon générale, ne peut calculer de moyenne mobile en $i = 1$, $i = 2$, ..., $i = k$ puisque les formules ne peuvent être appliquées si l'on connaît x_{i-k}. De même, si n est le nombre total d'observations, on ne peut calculer y'_n, ..., y'_{n-k+1} puisqu'il faut connaître x_{n+k}.

L'avantage des moyennes mobiles est d'atténuer la composante accidentelle tout en conservant les tendances linéaires : la série est dite "lissée", et est d'autant plus lissée que la longueur de la moyenne mobile est élevée comme on peut le constate sur les figures 4.3 et 4.4 des séries Alcatel sur lesquelles on a représenté respectivement les moyennes mobiles de longueur 4 et de longueur 14.

Dans les journaux financiers, les moyennes mobiles ne sont pas centrées : on utilise les moyennes des 50 ou 100 derniers observations avant l'instant t pou définir la tendance à l'instant t.

L'inconvénient des moyennes mobiles de grand longueur est qu'on ne dispose d'aucune information sur la tendance pour deux longs intervalles au début et à la fin de la période d'observation. Il faut donc choisir la longueur des moyennes mobiles suivant le nombre d'observations et l'objectif de l'analyse.

Les valeurs de y'_7 et y'_8, pour l'Exemple 4.1.1, s'obtiennent par

$$\begin{aligned} y'_7 &= \frac{\left(\frac{2450}{2}\right) + 2470 + 2550 + \ldots + 3140 + \ldots + 2470 + 2200 + \left(\frac{2525}{2}\right)}{12} \\ y'_8 &= \frac{\left(\frac{2470}{2}\right) + 2550 + 2540 + \ldots + 3150 + \ldots + 2200 + 2525 + \left(\frac{2530}{2}\right)}{12} \end{aligned}$$

Ainsi est obtenue une série de moyennes mobiles représentée par les 24 points : $M'_i(t_i, y'_i)$ pour $i = 7, \ldots, 30$.

	1988			1989			1990		
	t_i	y_i	y'_i	t_i	y_i	y'_i	t_i	y_i	y'_i
Janvier	1	2450	/	13	2525	2722	25	2630	2787
Février	2	2470	/	14	2530	2732	26	2635	2770
Mars	3	2550	/	15	2800	2743	27	2700	2776
Avril	4	2540	/	16	2600	2752	28	2710	2782
Mai	5	2800	/	17	2900	2760	29	3000	2789
Juin	6	2850	/	18	2950	2766	30	3050	2797
Juillet	7	3140	2666	19	3250	2775	31	2800	/
Août	8	3150	2672	20	3300	2784	32	3350	/
Septembre	9	2800	2685	21	2900	2784	33	3000	/
Octobre	10	2540	2698	22	2660	2792	34	2710	/
Novembre	11	2470	2705	23	2530	2793	35	2635	/
Décembre	12	2200	2713	24	2300	2801	36	2400	/

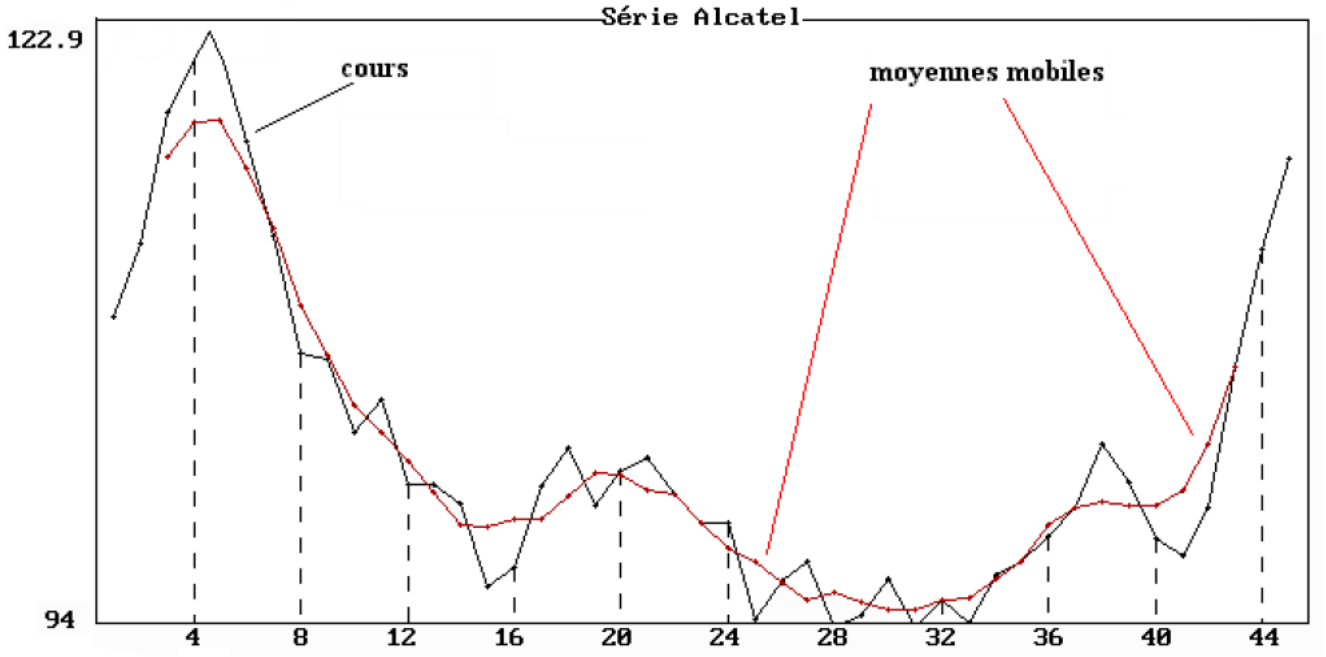


FIGURE 4.3 : Moyennes mobiles de longueur 5

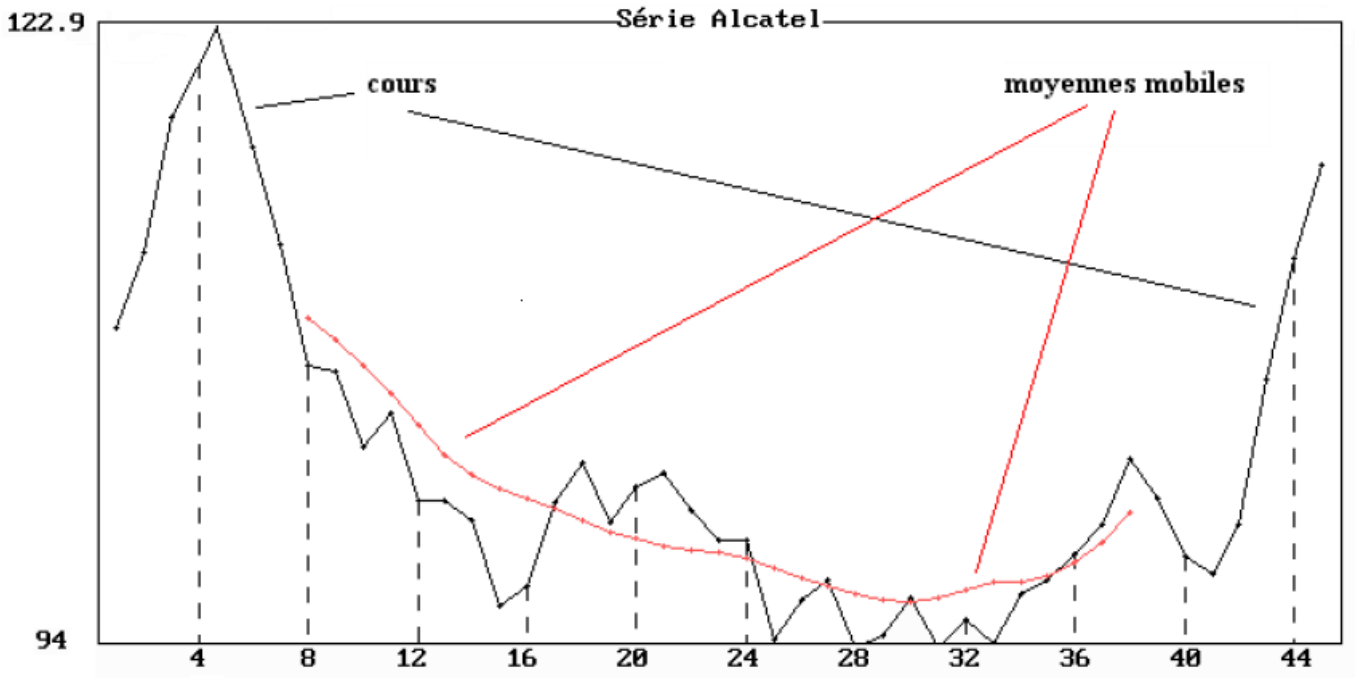
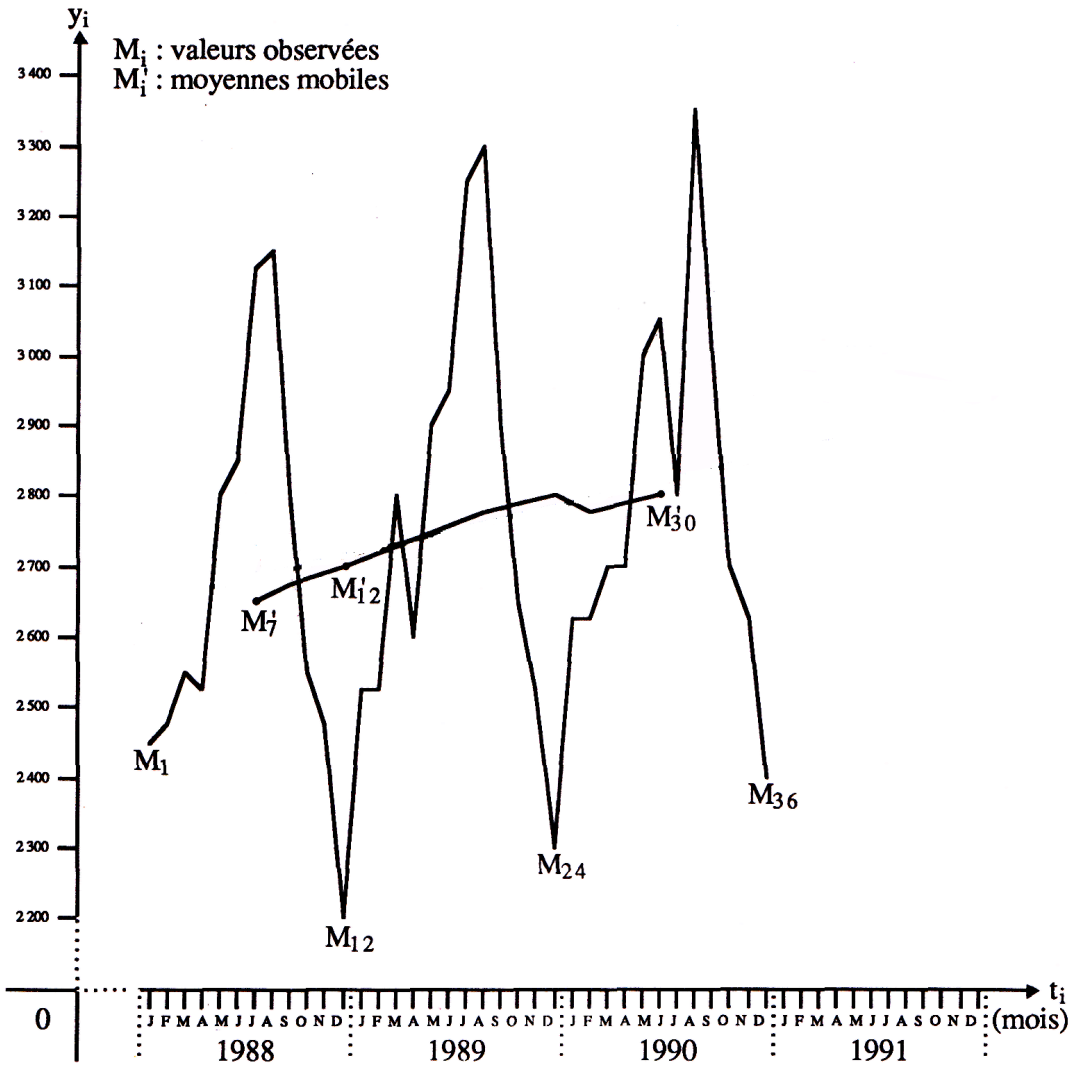


FIGURE 4.4 : Moyennes mobiles de longueur 14



La courbe obtenue montre que le phénomène est croissant. Cette courbe permet de mettre en évidence la tendance générale.

La longueur des moyennes mobiles dépend aussi de la période des variations saisonnières.

Une variation saisonnière est caractérisée par le fait qu'elle se produit à intervalles de temps réguliers, d'où d'ailleurs le terme saisonnier.

Exemple 4.1.2 Étudions la série chronologique x_1 observée trimestriellement pendant 6 ans :

	1^{er} trimestre	2^{e} trimestre	3^{e} trimestre	4^{e} trimestre
Année 1	89,658	97,593	108,906	114,157
Année 2	96,205	99,399	112,763	119,185
Année 3	99,602	105,192	116,556	121,911
Année 4	103,272	109,644	121,208	126,508
Année 5	105,637	113,428	125,641	131,147
Année 6	111,118	117,215	129,776	132,880

L'observation de chaque trimestre est soumis à un effet particulier qui revient tous les ans ; il y a donc 4 variations saisonnières correspondantes chacune à un trimestre. La période notée p des

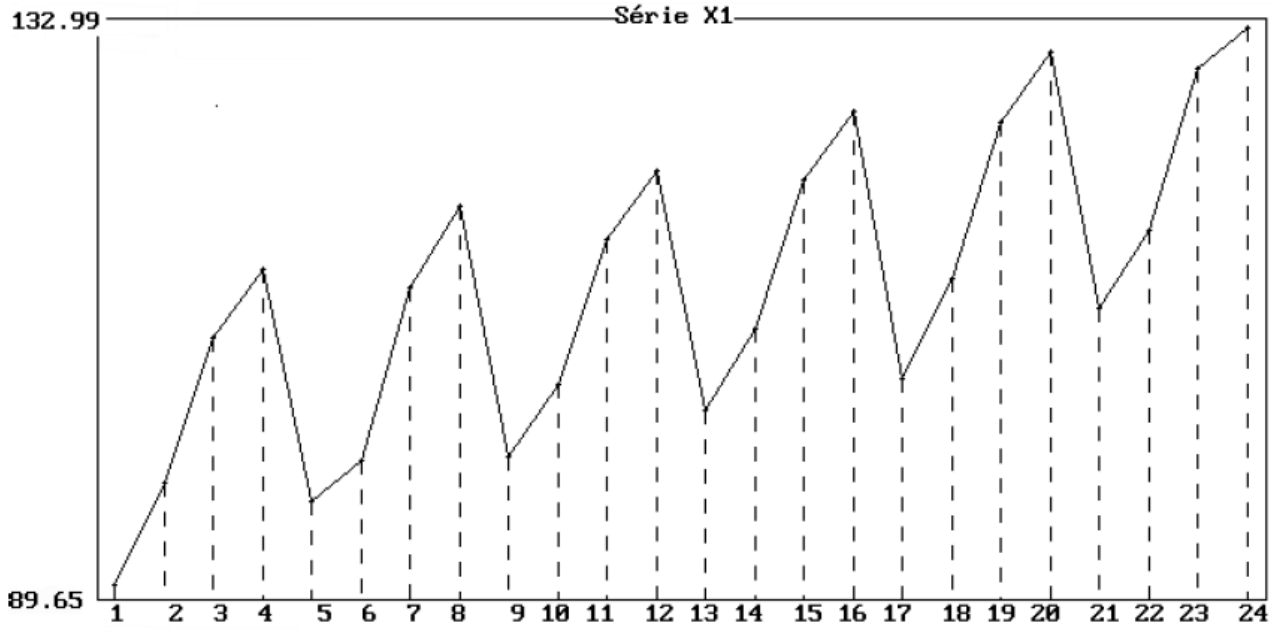
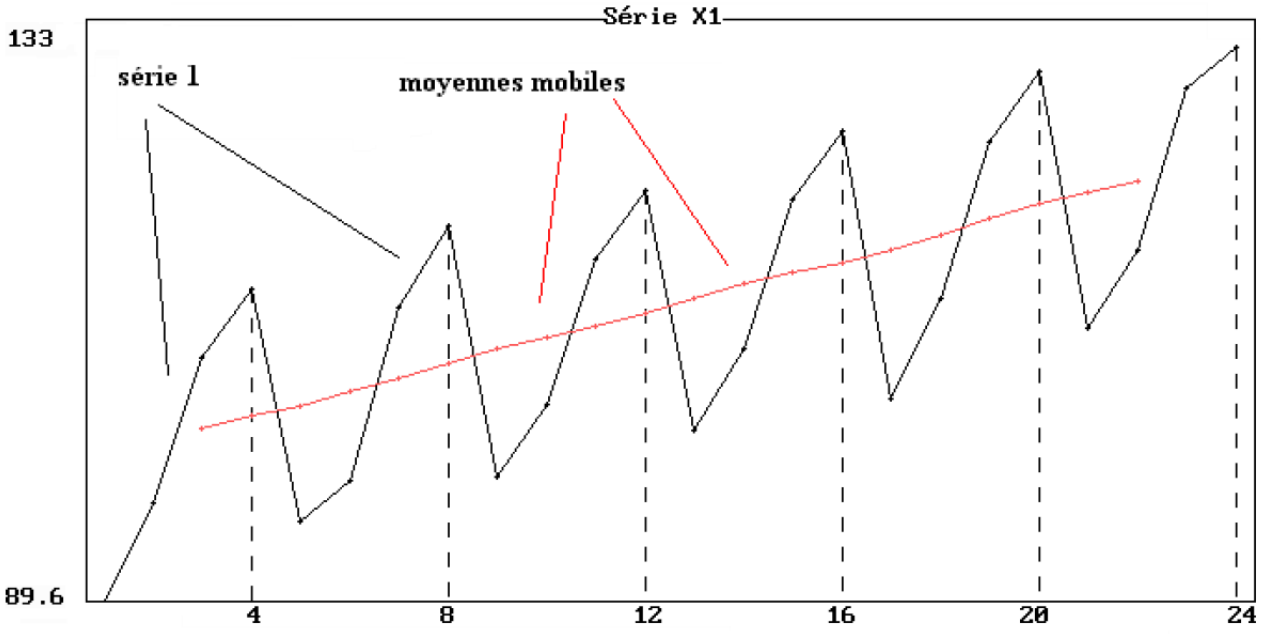


FIGURE 4.5 : Représentation graphique de la série - données observées trimestriellement pendant 6 ans

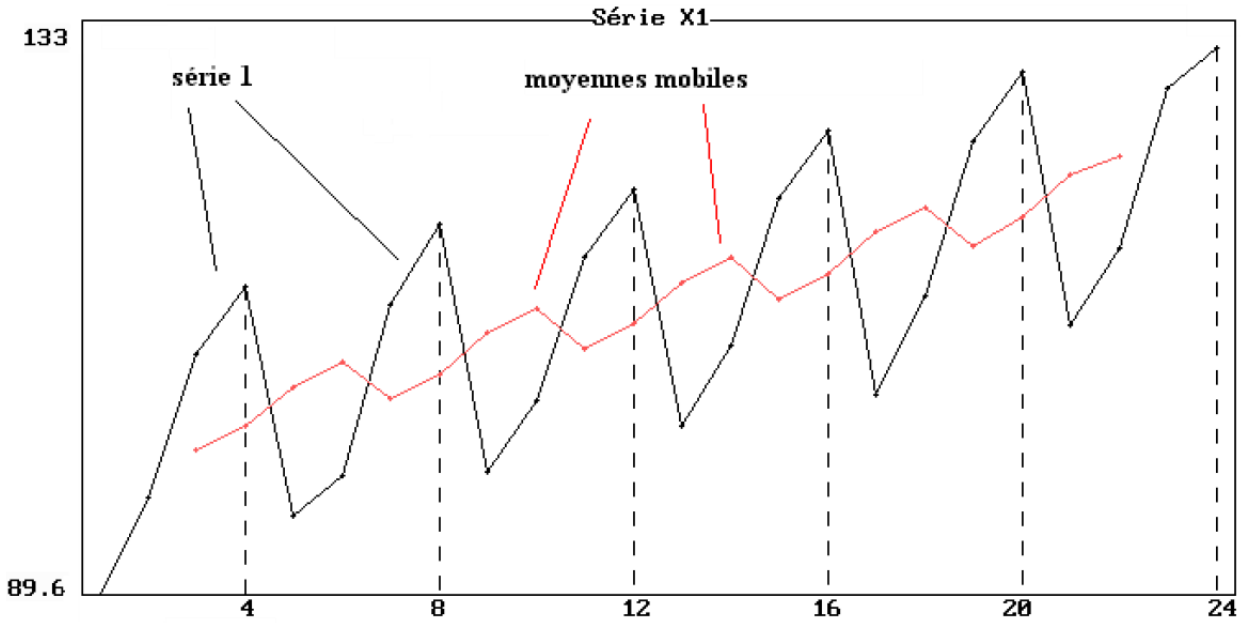
variations saisonnières est la longueur exprimée en unités de temps séparant deux variations saisonnières dues à un même phénomène.

Quand la série est soumise à des variations saisonnières de même période p, la période est alors le nombre de variations saisonnières. Ce n'est pas toujours réalisé : les ventes par tranches horaires d'un hypermarché sont soumises par exemple à une première variation saisonnière due à l'heure et à une seconde due à la journée. Ce cas est traité théoriquement en considérant une période égale au plus petit commun multiple des deux périodes : deux variations saisonnières de périodes 4 et 6 donnent une variation saisonnière de période $12(= 3 \times 4 = 2 \times 6)$.

Il n'est pas toujours facile de distinguer la tendance lorsque la série chronologique est soumise à des variations saisonnières. La méthode mathématique consiste à calculer les moyennes mobiles en choisissant comme longueur la période des variations saisonnières, de façon à les faire disparaitre. Si la moyenne mobile choisie est de longueur différente, les variations saisonnières ne sont pas toujours éliminées :



Représentation graphique de la série et des moyennes mobiles de longueur 4



Représentation graphique de la série et des moyennes mobiles de longueur 5

Contrairement aux moyennes mobiles de longueur 4, les moyennes mobiles de longueur 5 n'éliminent pas les variations saisonnières.

Les moyennes mobiles d'une série soumise à des variations saisonnières de période p ne sont pas soumises à des variations saisonnières si la longueur l est égale à la période p, et plus généralement si leur longueur est un multiple de la période.

Les moyennes mobiles d'une série chronologique dont la tendance est linéaire et les variations saisonnières sont de période p font apparaître la tendance et disparaitre les variations saisonnières si leur longueur l est égale à la période p.
Le moyennes mobiles ont l'avantage d'atténuer les variations accidentelles, mais l'inconvénient de n'être définies ni au début ni à la fin de la période observée.

4.1.3 Modélisation et désaisonnalisation

Un modèle de série chronologique est une équation précisant la façon dont les composantes s'articulent les unes par rapport aux autres pour constituer la série chronologique. Il existe de très nombreux modèles, et parmi eux deux modèles classiques simples : le modèle additif et le modèle multiplicatif.

Dans les deux modèles présentés, la longueur des moyennes mobiles doit être impérativement égale à la période des variations saisonnières.

En Exemple 4.1.2 les données sont représentées dans le tableau sous une forme particulière : en lignes, ce sont les années, et en colonnes les trimestres : le terme y_t correspondant à la t^e observation est alors noté $y_{i,j}$, i donnant l'année (la ligne) et j le trimestre (la colonne).

La relation entre les indices i et j d'une part et l'instant t d'autre part est la suivante :

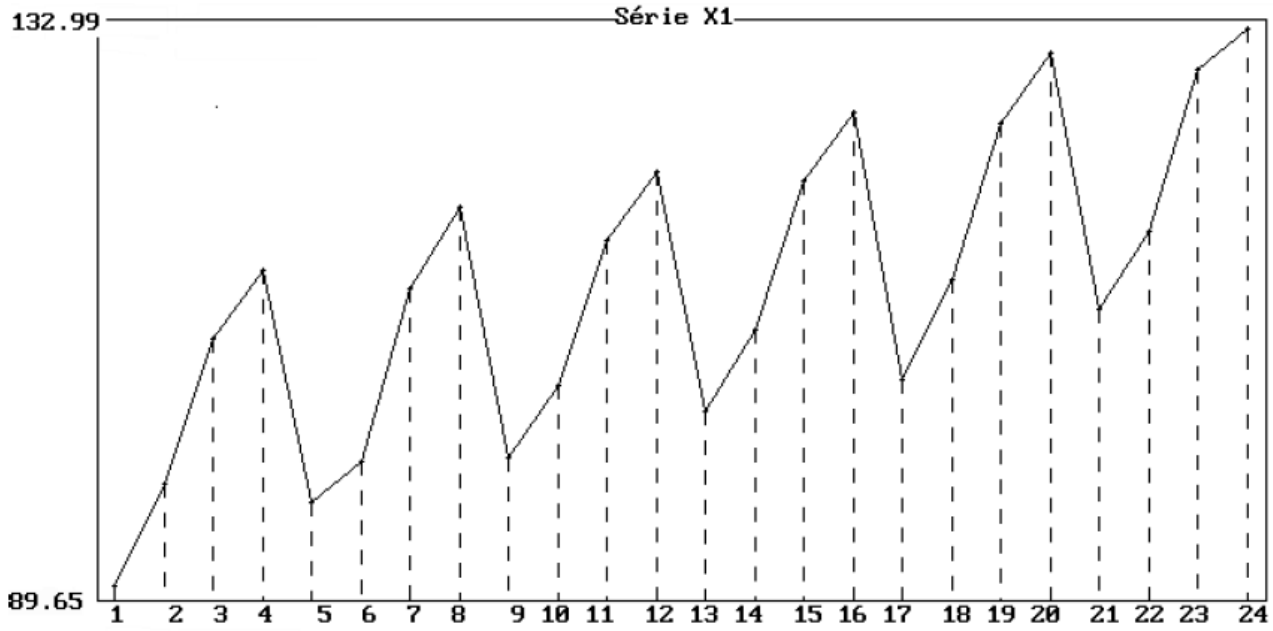
$$t = (i-1)p + j$$

Exemple :

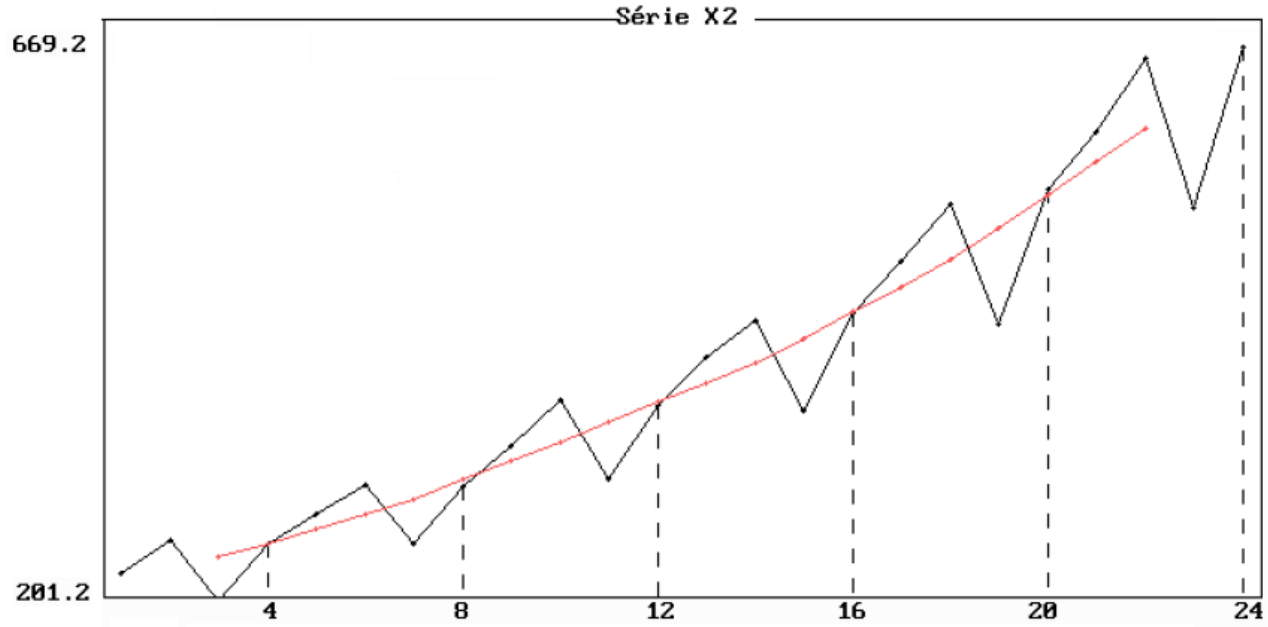
	$j=1$	$j=2$	$j=3$	$j=4$
$i=1$	$t=1$	$t=2$	$t=3$	$t=4$
$i=2$	$t=5$	$t=6$	$t=7$	$t=8$
$i=3$	$t=9$	$t=10$	$t=11$	$t=12$

Exemple pour $n=3$ et $p=4$.

Si y_t désigne la valeur observée à l'époque t, les schémas de décomposition additif et multiplicatif sont :
- **le schéma additif :** $y_t = T_t + S_t$ (si l'on ne tient pas compte de C_t et A_t). Les variations saisonnières s'ajoutent à la tendance générale (le mouvement saisonnier est d'amplitude constante)



- **le schéma multiplicatif :** les variations saisonnières ont une amplitude proportionnelles à la tendance générale (l'ampleur du mouvement saisonnier augmente lorsque le trend augmente).



C'est-à-dire :

$$y_t = T_t + T_t\alpha_t = T_t(1 + \alpha_t) = T_t S_t.$$

Les termes S_t sont appelés coefficients saisonniers.

Modèle additif de série chronologique.

La série chronologique y_t se décompose en une tendance notée T_t et des variations saisonnières S_t de période p (égales à S_1, S_2, S_3, ..., S_p).

Le modèle, dans lequel la variation saisonnière s'ajoute simplement à la tendance est le modèle le plus simple, :

pour tout $t = 1, \ldots, T \quad y_t = T_t + S_t$.

Le modèle additif s'exprime donc en général de la façon suivante :

pour tout $i = 1, \ldots, n$ pour tout $j = 1, \ldots, p \quad y_{i,j} = T_{i,j} + S_j$

Le terme S_j caractérise la variation saisonnière à l'instant j de chaque période i : du trimestre j dans le cas particulier des séries 1 et 2 ($p = 4$), du mois j dans des données mensuelles ($p = 12$) etc. . . . Les moyennes mobiles seront aussi notées $M'_{i,j}$.

Les termes S_j du modèle additif exprimé sous la forme précédente sont appelés coefficients saisonniers du modèle additif.

On peut calculer la différence entre l'observation et la tendance :

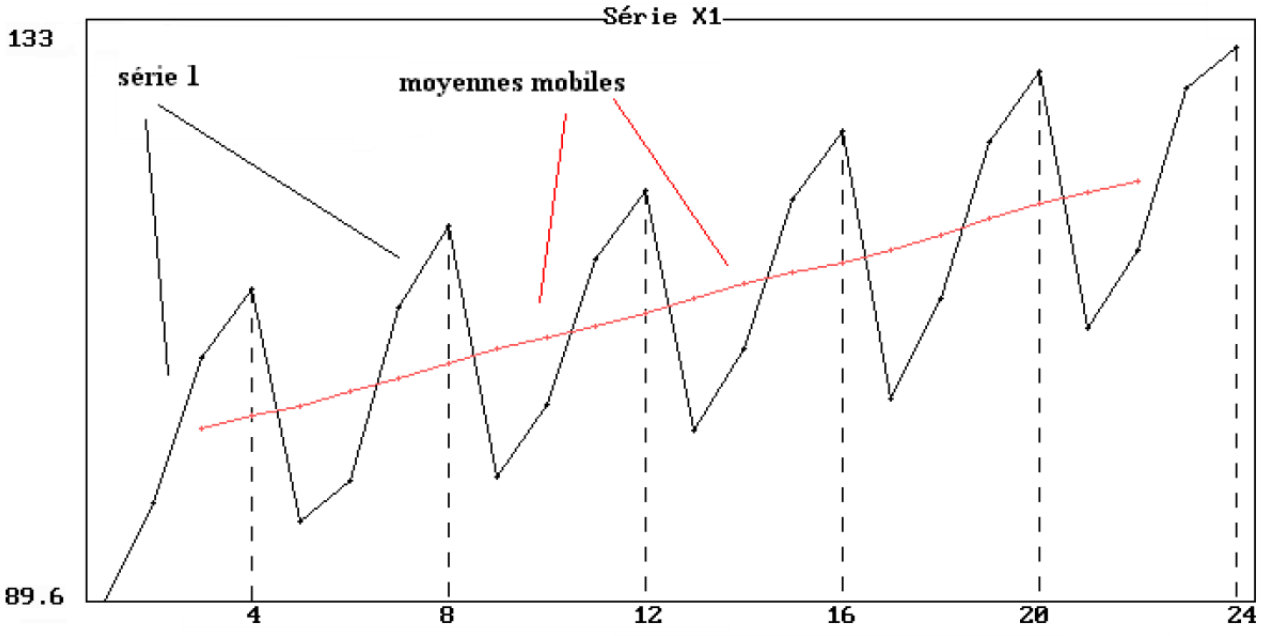
pour tout $i = 1, \ldots n$ pour tout $j = 1, \ldots, p \quad y_{i,j} - T_{i,j} = S_j$

Pour un même trimestre de l'exemple 4.1.2, la différence entre l'observation et la tendance est donc à peu près constant et égale à S_j.

Les moyennes mobiles de longueur l égale à la période des variations saisonnières sont des approximations de la tendance. On peut donc considérer que la différence entre une observation $y_{i,j}$ et la moyenne mobile $M'_{i,j}$ correspondante est à peu près constante pour j fixé :

pour tout $i = 1, \ldots, n$ pour tout $j = 1, \ldots, p \quad y_{i,j} - M'_{i,j} = S_j$

Cette propriété est recherchée sur la représentation graphique de la série x_t de l'exemple 4.1.2 pour déterminer si cette série suit un modèle additif ou non. Elle peut être observée sur la figure suivante dans laquelle la tendance est caractérisée par les moyennes mobiles de longueur 4 :



Les différences entre x_3 et M'_3, entre x_7 et M'_7, entre x_{11} et M'_{11} en sont à peu près constantes, de même les différences entre x_4 et M'_4, x_8 et M'_8, x_{12} et M'_{12} etc.

On peut en déduire les différences $x_{i,j} - M'_{i,j}$. Les moyennes mobiles donnant une première approximation de la tendance $T_{i,j}$, les colonnes du tableau des différences contiennent des approximations des coefficients S_j.

Exemple 4.1.2 : Les moyennes mobiles et par suite différences $x_{i,j} - M'_{i,j}$ ne sont pas définies aux premier et deuxième trimestres de la première année, ni aux troisième et quatrième trimestres de la dernière :

	$1^e r$ trimestre	2^e trimestre	3^e trimestre	4^e trimestre
Année 1			103.39678	104.44080
Année 2	105.14860	106.25917	107.31233	108.46116
Année 3	109.65950	110.47448	111.27404	112.28937
Année 4	113.42748	114.58360	115.45379	116.22236
Année 5	117.24943	118.38337	119.64839	120.80691
Année 6	121.79719	122.54573		

Moyennes mobiles de longueur 4 de la série x_1

	1^{er} trimestre	2^e trimestre	3^e trimestre	4^e trimestre
Année 1			5.50932	9.71580
Année 2	-8.94393	-6.86050	5.45047	10.72334
Année 3	-10.05746	-5.28258	5.28226	9.62153
Année 4	-10.15538	-4.93910	5.75471	10.28534
Année 5	-11.61263	-4.95497	5.99271	10.33979
Année 6	-10.67929	-5.33023		

Différences entre les observations et les moyennes mobiles de la série x_1

Les différences apparaissant dans une même colonne sont proches les uns des autres et caractérisent le modèle additif.

Les différences $x_{i,j} - M'_{i,j}$ sont donc des approximations des coefficients S_j. Leur moyenne (ou leur médiane), pour chaque colonne j, donne une première estimation $S'j$:

$$S'_j = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - M'_{i,j})$$

On obtiendra enfin les estimations définitives S_j en centrant ces termes S'_j :

- on calcule la moyenne des S'_j :

$$\bar{S}'_j = \frac{1}{p}(S'_1 + s'_2 + \ldots + S'_p)$$

- on centre en posant :

$$\text{pour tout} \quad j = 1, \ldots, p \quad S_j = S'_j - \bar{S}'_j$$

Exemple 4.1.2 : Du tableau des différences entre les observations et les moyennes mobiles on en déduit les moyennes suivantes :

$$S'_1 = -10.2897 \quad S'_2 = -5.4735 \quad S'_3 = 5.5979 \quad S'_4 = 10.1371$$

- On calcule la moyenne des S'_j : $\bar{S}'_j = -0.007039938$
- Les valeurs définitives sont obtenues en posant $S_j = S'_j - \bar{S}'_j$:

$$S_1 = -10.2827 \quad S_2 = -5.4664 \quad S_3 = 5.6049 \quad S_4 = 10.1442$$

Règle de calcul des estimations des coefficients saisonniers du modèle additif

- on calcule les différences entre les observations et les moyennes mobiles ;
- on calcule la moyenne ou la médiane S'_j des différences de chaque colonne du tableau ;
- on calcule la moyenne $\bar{S}'_j$ de ces valeurs S'_j ;
- on obtient les estimations S_j en centrant les valeurs S'_j : $S_j = S'_j - \bar{S}'_j$.

Modèle multiplicatif de série chronologique

Le modèle multiplicatif est le suivant :

$$\text{pour tout} \quad t = 1, \ldots, T \quad y_t = T_t(1 + \alpha_t)$$

En présentant les données comme dans le paragraphe précédent, le modèle multiplicatif s'exprime de la façon suivante :

$$\text{pour tout} \quad i = 1, \ldots, n \quad \text{pour tout} \quad j = 1, \ldots, p \quad y_{i,j} = T_{i,j}(1 + \alpha_j)$$

Le terme α_j caractérise la variation saisonnière du trimestre j dans des particulier, du mois j dans des données mensuelles etc.

On peut calculer la différence entre l'observation et la tendance :

$$\text{pour tout} \quad i = 1, \ldots, n \quad \text{pour tout} \quad j = 1, \ldots, p \quad y_{i,j} - T_{i,j} = T_{i,j}\alpha_j$$

Considérons le cas particulier $j = 1$ (1^{er} trimestre de l'année i).

$$\text{pour tout} \quad i = 1, \ldots, n \quad y_{i,1} - T_{i,1} = T_{i,1}\alpha_{i,1}$$

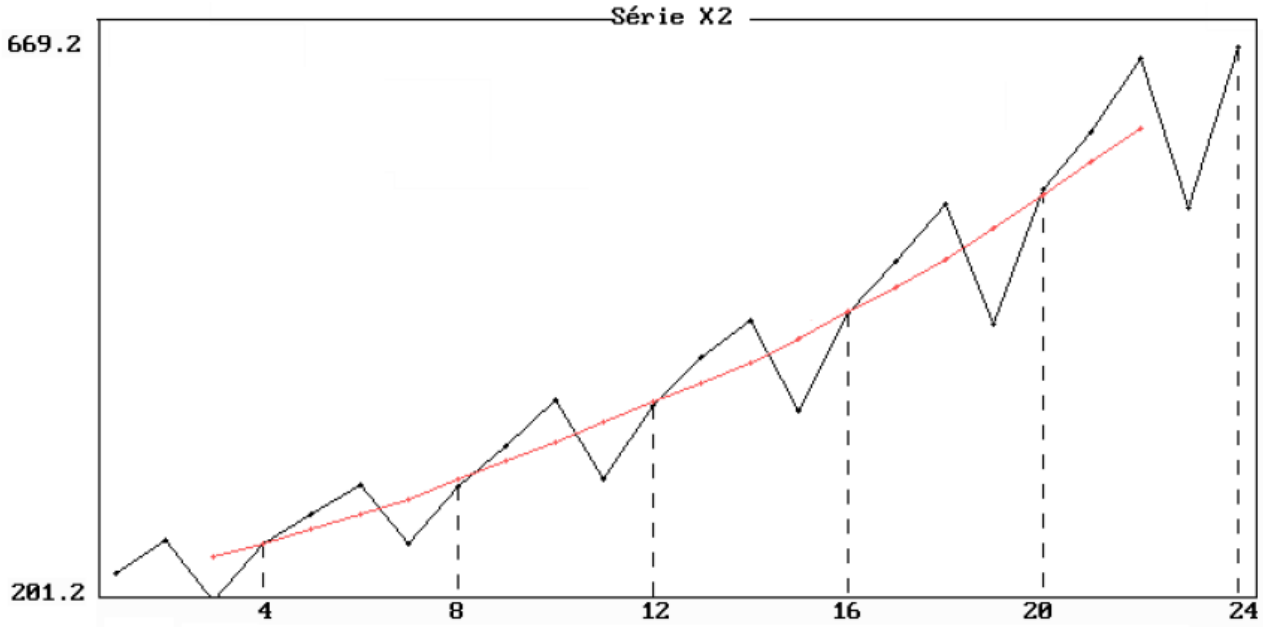
La différence $y_{i,1} - T_{i,1}$ entre l'observation et la tendance est proportionnelle à la tendance $T_{i,1}$: lorsque cette tendance est croissante, la différence augmente, lorsqu'elle est décroissante, il diminue.

Le même raisonnement peut évidemment être tenu pour j fixé quelconque. Les différences permettent ainsi de déterminer si la série chronologique étudiée suit un modèle multiplicatif.

Exemple 4.1.3 On considère la série chronologique x_2 donnée ci-dessous :

	1^{er} trimestre	2^{e} trimestre	3^{e} trimestre	4^{e} trimestre
Année 1	224.3705	253.2811	201.2421	248.9411
Année 2	274.3802	300.1641	248.9038	298.4386
Année 3	331.9657	371.4032	303.4313	365.9029
Année 4	406.6326	437.9967	361.5774	444.8447
Année 5	488.4166	536.5268	435.5698	549.3614
Année 6	598.0016	659.2896	533.2156	669.2675

Tableau de la série chronologique x_2
(modèle multiplicatif, période $p = 4$)



Série x_2 et moyennes mobiles de longueur 4

Cette série est soumise à des variations saisonnières de période 4 ; la tendance, caractérisée par les moyennes mobiles de longueur 4, est croissante, et la différence entre une observation x_t et la moyenne mobile M'_t a tendance à augmenter pour une même variation saisonnière : l. Les différences entre x_3 et M'_3, entre x_7 et M'_7, entre x_{11} et M'_{11} augmentent visiblement, de même que les différences entre x_4 et M'_4, x_8 et M'_8, x_{12} et M'_{12} etc.

Pour quantifier les variations saisonnières, on considère les rapports $y_{i,j}/T_{i,j}$:

$$\text{pour tout} \quad i = 1, \ldots, n \quad \text{pour tout} \quad j = 1, \ldots, p \quad \frac{y_{i,j}}{T_{i,j}} = 1 + \alpha_j$$

En utilisant l'approximation de la tendance par les moyennes mobiles, on constate donc que les rapports x_3/M'_3, x_7/M'_7, x_{11}/M'_{11}, ... sont à peu près constants et donnent une approximation de $1 + \alpha_3$, de même les rapports x_4/M'_4, x_8/M'_8, x_{12}/M'_{12} etc. donnent une approximation de $1 + \alpha_4$:

$$\text{pour tout} \quad j = 1, \ldots, p \quad \frac{y_{i,j}}{M'_{i,j}} = 1 + \alpha_j = S_j$$

Exemple 4.1.3 : les tableaux ci-dessous contiennent les moyennes mobiles de la série x_2 de l'exemple 4.1.3 et les rapports $x_{i,j}/M'_{i,j}$

	1^{er} trimestre	2^e trimestre	3^e trimestre	4^e trimestre
Année 1			238.210	250.322
Année 2	262.140	274.284	287.670	303.773
Année 3	319.494	334.743	352.509	370.167
Année 4	385.759	402.895	422.986	445.525
Année 5	467.090	489.404	516.167	545.210
Année 6	572.761	599.955		

Moyennes mobiles de longueur 4 de la série x_2

	^{1er} trimestre	2^e trimestre	3^e trimestre	4^e trimestre
Année 1			0.84481	0.99449
Année 2	1.04670	1.09435	0.86524	0.98244
Année 3	1.03904	1.10952	0.86078	0.98848
Année 4	1.05411	1.08712	0.85482	0.99847
Année 5	1.04566	1.09629	0.84385	1.00761
Année 6	1.04407	1.09890		

Rapports des observations aux moyennes mobiles de la série x_2

Les rapports dans chaque colonne du tableau ci-dessus sont à peu près constants.

Les rapports $y_{i,j}/M'_{i,j}$ sont donc des approximations des termes $1 + \alpha_j$ que l'on appelle coefficients saisonniers dans le cas du modèle multiplicatif.

Les termes $S_j = 1 + \alpha_j$ du modèle multiplicatif exprimé sous la forme précédente sont appelés coefficients saisonniers du modèle multiplicatif.

On obtient des premières estimations S'_j des coefficients saisonniers en calculant la moyenne (ou la médiane) des rapports figurant dans chaque colonne. Par analogie avec les coefficients saisonniers S_j du modèle additif, dont la moyenne est égale à 0, on cherche des estimations définitives S_j de moyenne 1 :

- on calcule la moyenne :

$$\bar{S}' = \frac{1}{p}(S'_1 + S'_2 + \ldots + S'_p)$$

- on pose :

$$\text{pour tout} \quad j = 1, \ldots, p \quad S_j = S'_j/\bar{S}'_j$$

Les coefficients saisonniers estimés S_j sont ainsi de somme p :

$$S_1 + S_2 + \ldots + S_p = \frac{1}{\bar{S}'}(S'_1 + S'_2 + \ldots + S'_p) = p$$

ce qui équivaut à une moyenne des α égale à 0 puisque l'on a $S_j = 1 + \alpha_j$.

Exemple 4.1.3 :

- Considérons le tableau des rapports des observations aux moyennes mobiles d'exemple 4.1.3
- on en déduit les moyennes suivantes :

$$S'_1 = 1.045913 \quad S'_2 = 1.097236 \quad S'_3 = 0.8539006 \quad S'_4 = 0.9942986$$

- on calcule la moyennes des S'_j : $\bar{S}' = 0,9978371$

- les valeurs définitives sont obtenues de façon que les S'_j soient de moyenne 1

$$S_1 = 1.04818 \quad S_2 = 1.099614 \quad S_3 = 0.8557515 \quad S4 = 0.9964539$$

Règle de calcul des estimations des coefficients saisonniers du modèle multiplicatif

- on calcule les rapports des observations aux moyennes mobiles ;
- on calcule la moyenne ou la médiane des rapports S'_j de chaque colonne du tableau ;
- on calcule la moyenne $\bar{S}'$ de ces valeurs ;
- on obtient les estimations S_j en posant $S_j = S'_j/\bar{S}'$.

Désaisonnalisation

Les coefficients saisonniers permettent d'éliminer d'une observation les effets de la variation saisonnière correspondante. On obtient ainsi les valeurs corrigées des variations saisonnières, ou encore les valeurs désaisonnalisées.

L'avantage de cette désaisonnalisation est de permettre la comparaison de deux observations soumises à des variations saisonnières différentes.

On appelle observation corrigée des variations saisonnières la valeur $y'_{i,j}$ obtenue en éliminant l'effet saisonnier sur la valeur $y_{i,j}$.

$$\begin{aligned} \text{modèle additif :} \quad & y'_{i,j} = y_{i,j} - S_j \\ \text{modèle multiplicatif :} \quad & y'_{i,j} = y_{i,j}/S_j \end{aligned}$$

Les valeurs corrigées des variations saisonnières (expression souvent abrégée par c.v.s.) caractérisent à la fois la tendance et la variation accidentelle.

Exemple : on donne ci-dessous les quatre dernières observations de la série x_2 (année 6) et les valeurs corrigées des variations saisonnières de l'exemple 4.1.3 :

	1^{er} trimestre	2^e trimestre	3^e trimestre	4^e trimestre
observations :	598.00160	659.28960	533.21560	669.26750
valeur c.v.s. :	570.51396	599.56452	623.09629	671.64924

L'observation du deuxième trimestre est largement supérieure à celle du troisième, mais c'est l'inverse pour les valeurs c.v.s. : la tendance est restée croissante au troisième trimestre.

Supposons que l'observation du premier trimestre de l'année 7 soit égale à 720.15. Pour savoir si la tendance est restée à la hausse, on calcule la valeur désaisonnalisée :

$$x'_{7,1} = 720.15/1.04818 = 687.04771$$

et on la compare à la valeur désaisonnalisée du quatrième trimestre de l'année précédente :

$$x'_{6,4} = 671.649$$

La valeur c.v.s. $x'_{7,1}$ est supérieur à $x'_{6,4}$. La tendance est restée à la hausse si la différence est supérieure à la variation accidentelle . Il faudrait donc comparer cette différence à l'écart type des variations accidentelles, calculé sur les données antérieures. Il semble que dans la pratique, cette comparaison ne soit guère effectuée.

4.1.4 Méthodologie de la prévision

Deux cas sont possibles.

a/ **La série ne comporte pas de composantes saisonnières**

La prévision est alors faite par extrapolation sur la droite de tendance obtenue par un ajustement affine. Les valeurs extrapolées sont lues directement sur le graphique ou obtenues en remplaçant, dans l'équation de la droite, la variable t par la date souhaitée.

b/ **La série des valeurs observées comporte des composantes saisonnières**

- **Utilisation des moyennes mobiles**

Les moyennes mobiles permettent d'éliminer, en partie, les variations saisonnières. La série des valeurs observées représentée par les points :

$$M_i(t_i, y_i) \quad \text{pour} \quad i = 1, \ldots, n$$

est remplacée par la nouvelle série représentée par les points :

$$M'_i(t_i; y'_i), \quad \text{pour} \quad i = \alpha, \ldots, p \quad (\alpha > 1; p < n).$$

Le calcul du coefficient de corrélation linéaire pour les séries : t_i et y'_i , $i = \alpha, \ldots, p$ permet de justifier l'ajustement affine éventuel (ou de connaître le degré de fiabilité de l'estimation lorsque la corrélation linéaire n'est pas forte).

Enfin, le nuage de points M'_i (t_i, y'_i) est ajusté à une droite (droite de Mayer, droites des moindres carrés). Il est donc possible à partir de l'équation de cette droite, de déterminer une estimation $\hat{y}'$ (valeur prévisionnelle) à une date future.

Par la suite, le symbole $\hat{}$ sera utilisé pour désigner une estimation.

La valeurs $\hat{y}'$ ne tenant pas compte, en totalité, de l'effet saisonnier, la composante saisonnière doit être « réintroduite », ce qui est obtenu par les **coefficients saisonniers** S_i.

L'estimation définitive $\hat{y}$ est alors :
- dans le cas d'un schéma additif : $\hat{y} = \hat{y}' + S_i$
- dans le cas d'un schéma multiplicatif : $\hat{y} = \hat{y}' \times S_i$

s_i étant le coefficient saisonnier qui correspond à la date de l'estimation.

Exemple : Détermination des niveaux futurs des expéditions de Yopmilk (exemple 4.1.1)

Le responsable du service « Transports-Livraisons » de Yopmilk souhaite estimer les expéditions pour 1991.

Les points $M'_i\ (t_i, y'_i)$ obtenus par la méthode des moyennes mobiles, permettent de calculer le coefficient de corrélation linéaire $r(t, y')$:

$$r(t, y') = 0,913.$$

La corrélation linéaire étant forte, l'estimation sera faite à partir de la première droite des moindres carrés :

$$y' - \bar{y'} = a(t - \bar{t}) \text{ avec } a = \frac{cov(t, y')}{Var(t)} \text{ soit } y' = 5,47t + 2650,7 \text{ (D)}$$

Cette droite est construite sur le graphique 4.2 à partir des points : A (7 ; 2 689) et B (30 ; 2 815).

Par exemple, le mois d'août 1991 correspondant à $t = 44$:

$$\hat{y}'_{44} = 5,47 \times 44 + 2650,7 = 2891.$$

La démarche est la même pour tous les mois de 1991 (cf. tableau 4.1).

Les valeurs du trend étant calculées, il reste à déterminer les influences saisonnières.

Le mouvement saisonnier semblant avoir une amplitude constante, le schéma additif sera choisi :

$$\hat{y}_{44} = \hat{y}'_{44} + S_8 \text{ avec } S_8 : \text{ coefficient saisonnier des mois d'août.}$$

Pour cet exemple, les « saisons » sont des mois. Le tableau de la série observée et des moyennes mobiles

	1988			1989			1990		
	t_i	y_i	y'_i	t_i	y_i	y'_i	t_i	y_i	y'_i
Janvier	1	2450	/	13	2525	2722	25	2630	2787
Février	2	2470	/	14	2530	2732	26	2635	2770
Mars	3	2550	/	15	2800	2743	27	2700	2776
Avril	4	2540	/	16	2600	2752	28	2710	2782
Mai	5	2800	/	17	2900	2760	29	3000	2789
Juin	6	2850	/	18	2950	2766	30	3050	2797
Juillet	7	3140	2666	19	3250	2775	31	2800	/
Août	8	3150	2672	20	3300	2784	32	3350	/
Septembre	9	2800	2685	21	2900	2784	33	3000	/
Octobre	10	2540	2698	22	2660	2792	34	2710	/
Novembre	11	2470	2705	23	2530	2793	35	2635	/
Décembre	12	2200	2713	24	2300	2801	36	2400	/

permet de calculer les différences des observations y_j et les moyennes mobiles y'_j :

	1988	1989	1990	S'_j	$S_j = S'_j - \bar{S}'_j$
Janvier	/	-513	-501	-507	-517
Février	/	-197	-157	-177	-187
Mars	/	-203	-135	-169	-179
Avril	/	57	-76	-10	-19
Mai	/	-152	-73	-113	-122
Juin	/	140	211	176	166
Juillet	184	253	/	219	209
Août	474	475	/	475	465
Septembre	475	516	/	497	487
Octobre	115	116	/	116	106
Novembre	-158	-124	/	-141	-151
Décembre	-235	-263	/	-249	-259

La colonne S'_j contient les moyennes des différences pour chaque saison.

La dernière colonne donne les coefficients saisonniers $S_j = S'_j - \bar{S}'_j$. La moyenne des S'_j est

$$\bar{S}'_j = \frac{1}{12}\sum_{i=1}^{12} S'_j = 9,6667$$

Le coefficient saisonnier S_8 du mout d'aout est

$$\begin{aligned} S_8 &= 465 \\ \hat{y}_{44} &= \hat{y}'_{44} + S_8 = 2891 + 465 = 3356. \end{aligned}$$

Les $\hat{y}'_i$ sont calcules à partir de l'equation de droite :

$$y' = 5,47t + 2650,7$$

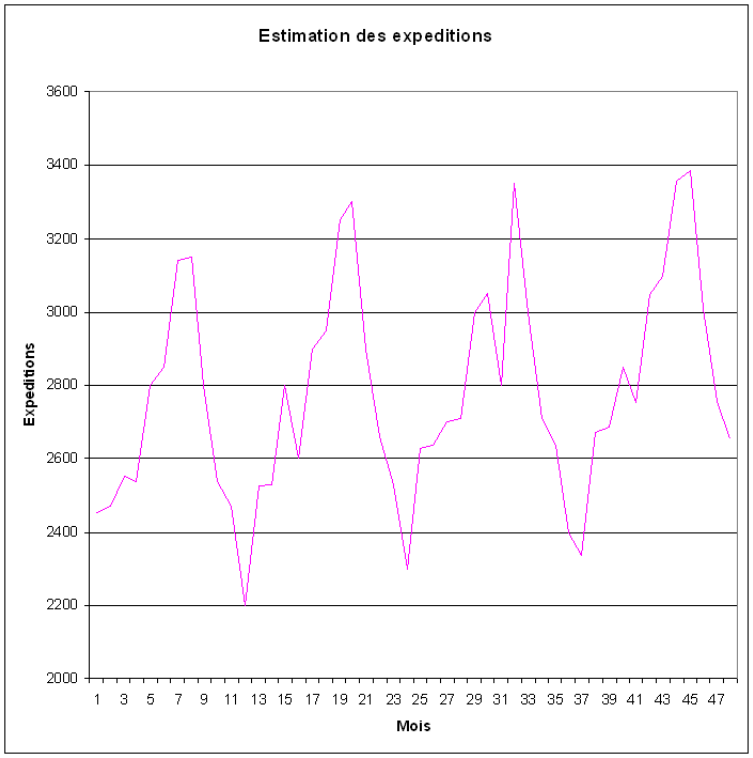
Les estimations des expéditions pour 1991 sont :

Les $\hat{y}'_i$ sont calculés à partir de l'équation de droite :

$$y' = 5,47t + 2650,7.$$

Mois de 1991	$\hat{y}'_i = 5,47t_i + 2650,7$	S_i $(i = 1, \ldots, 12)$	$\hat{y}_i$
Janvier (37)	2853	S_1 =-517	2336
Février (38)	2859	S_2 =-187	2672
Mars (39)	2864	S_3 =-179	2685
Avril (40)	2 870	S_4 =-19	2850
Mai (41)	2875	S_5 = -122	2753
Juin (42)	2880	S_6 = 166	3046
Juillet (43)	2886	S_7 =209	3095
Août (44)	2891	S_8 = 465	3356
Septembre (45)	2897	S_9 = 487	3384
Octobre (46)	2902	S_{10} =106	3008
Novembre (47)	2908	S_{11} =-151	2757
Décembre (48)	2913	S_{12} =-259	2655

TABLE 4.1 : Estimations des expéditions



Chapitre 5

Echantillonnage

5.1 Introduction

L'échantillonnage représente l'ensemble des opérations qui ont pour objet de prélever un certain nombre d'individus dans une population donnée.

Avantages de l'échantillonnage

L'analyse d'un échantillon au lieu de la population cout moindre, gain de temps et c'est la seule méthode qui donne des résultats dans le cas d'un test destructif.

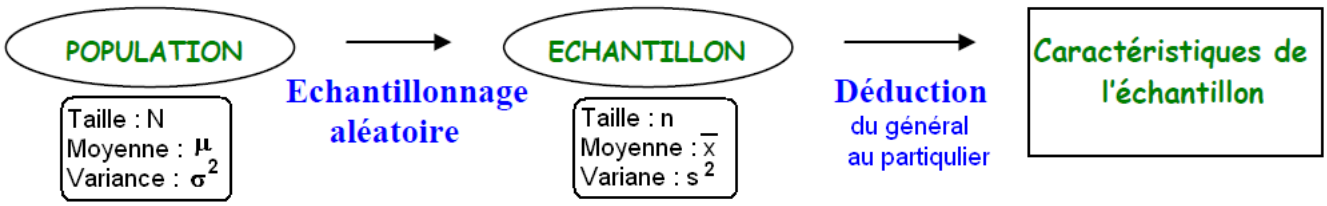


FIGURE 5.1 : Statistique descriptive

Inconvénients de l'échantillonnage

L'échantillonnage a pour but de fournir suffisamment d'informations pour pouvoir faire des déductions sur les caractéristiques de la population. Les résultats obtenus d'un échantillon à l'autre sont en général différents et différents également de la valeur de la caractéristique correspondante dans la population. Ces différences sont dues aux fluctuations d'échantillonnage. Pour pouvoir tirer des conclusions valables, il faut déterminer les lois de probabilités qui régissent ces fluctuations.

Pour que les résultats observés lors d'une étude soient généralisables à la population statistique, **l'échantillon doit être représentatif** de cette dernière, c'est à dire qu'il doit refléter fidèlement sa composition et sa complexité. Seul l'échantillonnage aléatoire assure la représentativité de l'échantillon.

Un échantillon est qualifié d'**aléatoire** lorsque chaque individu de la population a une probabilité connue et non nulle d'appartenir à l'échantillon.

Le cas particulier le plus connu est celui qui affecte à chaque individu la même probabilité d'appartenir à l'échantillon.

Il y a 2 grandes catégories de méthodes d'échantillonnage :
— l'échantillonnage non aléatoire : l'analyste utilise son expérience et son jugement pour constituer l'échantillon avec tous les risques de non représentativité de celui-ci. On identifie dans la population-mère, quelques critères de répartition significatifs puis on essaye de respecter cette répartition dans l'échantillon d'individus interrogés.
La méthode d'échantillonnage non-probabiliste est utilisée lorsqu'il n'est pas possible de constituer une liste exhaustive de toutes les unités du sondage.
— l'échantillonnage aléatoire ou probabiliste : il permet de calculer précisément l'erreur due à l'échantillonnage et par conséquent de juger de la valeur de l'information partielle obtenue (donc de la représentativité de l'échantillon).

Par la suite, nous ne parlerons que de l'échantillon aléatoire simple : c'est un échantillon choisi de telle sorte que chaque unité de la population ait la même probabilité d'être sélectionnée dans l'échantillon et que chaque échantillon de même taille tiré de la population ait la même probabilité d'être choisi. On laisse dans ce cas le hasard choisir l'échantillon en utilisant par exemple une table de nombres au hasard.

Un échantillon aléatoire simple peut être tiré avec ou sans remise.

Dans l'*échantillon aléatoire simple avec remise*, chaque unité est remise dans la population après avoir été observée et avant qu'une autre unité soit choisie. Il y a donc indépendance entre les résultats d'un tirage à l'autre et chaque unité conserve la même probabilité d'être sélectionnée.

Dans l'*échantillon aléatoire simple sans remise* (échantillonnage exhaustif), l'unité tirée n'est pas remise ce qui modifie, pour une unité particulière, la probabilité d'être choisie d'un tirage à l'autre (si l'échantillon est choisi dans une population finie de N unités, chaque unité a une probabilité $\frac{1}{N}$ d'être choisie au 1er tirage, chaque unité restante une probabilité $\frac{1}{N-1}$ d'être choisie au 2e tirage, etc...).

Dans ce cas, il n'y a pas indépendance d'un tirage à l'autre. Si l'on a affaire à une population infinie ou si n, taille de l'échantillon, est relativement petite par rapport à N, taille de la population mère, on peut supposer qu'il y a indépendance d'une épreuve à l'autre, même si les tirages sont effectués sans remise.

Dans le cas contraire, lorsque la population est finie et lorsque $n > 0,05N$, il faut tenir compte d'un facteur de correction ou d'exhaustivité (voir l'estimation de l'écart type).

Par la suite, nous distinguerons 2 catégories de problèmes :
— **les problèmes de distribution d'échantillonnage :** dans ce cas, on connaît la valeur de certains paramètres de la population mère et on cherche à induire des renseignements sur les valeurs que peuvent prendre ces paramètres dans l'échantillon.
— **les problèmes d'estimation :** on connaît la valeur de certains paramètres dans l'échantillon et on cherche à induire des renseignements sur les valeurs que peuvent prendre ces paramètres dans la population mère.

Dans la suite du cours on utilisera les symboles suivants :
— pour la population mère : taille : N, moyenne arithmétique de la variable étudiée : μ, variance : σ^2, écart type : σ.
— pour l'échantillon : taille : n, moyenne arithmétique mesurée sur l'échantillon : $\bar{x}$, variance : s^2, écart-type : s.

	Population	Échantillon
Définition	C'est l'ensemble des unités considérées par le statisticien.	C'est un sous-ensemble de la population choisie pour étude.
Caractéristiques	Ce sont les paramètres	Ce sont les statistiques
Notations	N = taille de la population (si elle est finie)	n = taille de l'échantillon
Si on étudie un caractère **quantitatif**	moyenne de la population $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	moyenne de l'échantillon $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
	écart-type de la population $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$	écart-type de l'échantillon $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ $s' = \sqrt{\frac{n}{n-1}}s$
Si on étudie un caractère **qualitatif**	proportion dans la population p	proportion dans l'échantillon f

5.2 Les problèmes de distribution d'échantillonnage

5.2.1 Distribution d'échantillonnage de la moyenne $\bar{X}$

Dans une population mère de taille N, on peut tirer plusieurs échantillons de taille n : $\left(C_N^n = \frac{N!}{n!(N-n)!}\right)$.

Pour chaque échantillon, on peut calculer une moyenne :

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

et une variance

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

La valeur de l'espérance mathématique $\bar{x}$ et de la variance s^2 varient d'un échantillon à l'autre. C'est cette variation qui donne naissance à la distribution des variables aléatoires :

- **échantillonnage de la moyenne ou moyenne d'échantillon** $\bar{X}$, caractérisée par :

$E(\bar{X})$: l'espérance mathématique des moyennes calculées sur tous les échantillons de taille n.

$s_{\bar{X}}$: l'écart type de la distribution d'échantillonnage, qui représente la dispersion de l'ensemble des moyennes d'échantillons de taille n autour de $E(\bar{X})$

- **variance d'échantillon** $S'^2_{\bar{X}}$ définie par

$$S'^2_{\bar{X}} = \frac{n}{n-1}S^2_{\bar{X}} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

L'espérance de $S'^2_{\bar{X}}$ est la variance de la population et $S'^2_{\bar{X}}$ $/E(S'^2_{\bar{X}}) = \sigma^2/$ est un estimateur **sans biais** de σ^2.

Avant de continuer, essayons de comprendre sur un exemple ce qui se passe.

Exemple 5.2.1 /Feuille 8/ Une population est constituée de 5 clients d'un magasin. Le propriétaire du magasin s'intéresse à la somme moyenne (en €) laissée par chaque client dans le magasin lors d'une journée. On a obtenu les résultats suivants.

Client	Somme de l'achat (en €)
A	7
B	3
C	6
D	10
E	4
Total	30

La moyenne de la population est $\mu = 30/5 = 6$.

Soit le propriétaire choisit un échantillon de taille 3. On peut se poser les questions les suivantes :
- Combien sont les différents échantillons possibles qu'il peut choisir parmi les 5 clients observés ?
- Quelles sont les différentes valeurs possibles pour la moyenne de l'échantillon choisi ?
- Quelle relation existe-t-elle entre cette moyenne d'échantillon et la véritable moyenne 6 de la population ?

Enumérons les possibilités dans un tableau :

Numéro de l'échantillon	Echantillon	Sommes des achats dans cet échantillon	Moyenne de l'échantillon
1	A, B, C	7,3,6	5.33
2	A, B, D	7,3,10	6.67
3	A, B, E	7,3,4	4.67
4	A, C, D	7,6,10	7.67
5	A, C, E	7,6,4	5.67
6	A, D, E	7,10,4	7.00
7	B, C, D	3,6,10	6.33
8	B, C, E	3,6,4	4.33
9	B, D, E	3,10,4	5.67
10	C, D, E	6,10,4	6.67
Total			60.00

Les réponses des questions posées sont :
– Il y a 10 échantillons ($C_5^3 = 10$).
– La moyenne des échantillons varie entre 4.33 et 7.67. On constate que la distribution des moyennes d'échantillon est moins dispersée que la distribution des valeurs d'achat des différents clients, située entre 3 et 10.
– Il est possible que deux échantillons aient la même moyenne. Dans cet exemple, aucun n'a la moyenne de la population ($\mu = 6$).
– La moyenne des moyennes d'échantillon est $E(\bar{X}) = 60/10 = 6$.

On peut conclure que la variable $\bar{X}$ - moyenne de l'échantillon et une variable aléatoire dont les paramètres de position, de distribution etc. diffèrent des paramètres de la variable observée dans la population et en plus dépendent du choix de l'échantillon. Pour pouvoir déduire la moyenne μ de la population à la base de la moyenne $\bar{X}$ de l'échantillon, il faut connaitre la distribution de la variable aléatoire $\bar{X}$ et l'écart-type de cette distribution.

I. Cas : moyenne μ et écart-type σ de la population connues :

A) Si la population est infinie ou si l'échantillonnage est non exhaustif (tirage avec remise) :

— l'espérance mathématique de $\bar{X}$ est égale à la moyenne de la population :

$$E(\bar{X}) = \mu$$

— la variance de $\bar{X}$ est égale à la variance de la population divisée par la taille n de l'échantillon :

$$s^2_{\bar{X}} = \frac{\sigma^2}{n} \rightarrow s_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

Soit E_1, E_2, ..., E_p : p échantillons de taille n issues d'une même population mère de moyenne μ et de variance σ^2.
Soit $\bar{x}_1$, $\bar{x}_2$, ..., $\bar{x}_p$: leurs moyennes respectives.
Soit $\bar{X}$: la variable aléatoire qui prend pour valeur ces moyennes :

$$\bar{X} = \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$$

Alors lorsque $n \geq 30$, $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ en vertu du théorème central limite.

Exemple 5.2.2 /Feuille 8/ Une machine effectue l'ensachage d'un produit.
On sait que les sacs ont un poids moyen de 250g avec un écart-type de 25g.
Quelles sont les caractéristiques de la moyenne des poids d'un échantillon de 100 sacs ?

Solution.
$(P) : \mu = 250, \quad \sigma = 2,5; \quad (E) : n = 100 > 30$
$\bar{X}$ suit la loi normale de paramètres $\mu = 250$ et $\frac{\sigma}{\sqrt{n}} = \frac{25}{10} = 2,5$.

Remarque 24 1. Nous venons de démontrer ce que nous avions constaté sur l'exemple 5.2.1 : la moyenne de la distribution d'échantillonnage des moyennes est égale à la moyenne de la population.
2. On constate que plus n croît, plus $Var(\bar{X})$ décroît.

Dans l'exemple 5.2.1, nous avions en effet constaté que la distribution des moyennes d'échantillon était moins dispersée que la distribution initiale. En effet, à mesure que la taille de l'échantillon augmente, nous avons accès à une plus grande quantité d'informations pour estimer la moyenne de la population. Par conséquent, la différence probable entre la vraie valeur de la moyenne de la population et la moyenne échantillonnale diminue. L'étendue des valeurs possibles de la moyenne échantillonnale diminue et le degré de dispersion de la distribution aussi.

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ est aussi appelé l'**erreur-type** de la moyenne.

B) Si l'échantillonnage est exhaustif (tirage sans remise) dans une population finie (avec $n > 0.05N$) : on doit tenir compte d'un facteur d'exhaustivité pour déterminer $s_{\bar{X}}$.

Celui-ci devient : $s_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$.

Échantillonnage exhaustif (tirage sans remise) dans une population finie avec $n > 0.05N$

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right)$$

Exemple 5.2.3 /Feuille 8/ Dans une usine textile, on utilise une machine automatique pour couper des morceaux de tissu. Lorsque la machine est correctement ajustée, la longueur des morceaux de tissu est en moyenne de 90 cm avec un écart type de 0.60 cm.

Pour contrôler la longueur des morceaux de tissu, on tire dans la production d'une journée un échantillon aléatoire de 200 morceaux.

a) Si l'on suppose que la longueur X des morceaux de tissu suit une loi normale, calculer la probabilité que la moyenne de l'échantillon soit au plus égale à 89.90 cm, ceci dans 2 cas :
— production de la journée : 10 000 morceaux
— production de la journée : 2 000 morceaux.

b) Déterminer la même probabilité sans faire l'hypothèse que X soit distribuée normalement.

c) Si la moyenne observée sur cet échantillon est de 90.30 cm, celui-ci est-il représentatif de la population mère en prenant un risque de 5 % de se tromper ? (avec $N = 10\ 000$).

Solution :

a) Production journalière $= N = 10\ 000$; Taille de l'échantillon $= n = 200$; $\frac{n}{N} = 0.02$

Même si l'échantillonnage est exhaustif, ce n'est pas la peine de tenir compte du coefficient d'exhaustivité.

Dans ce cas $E(\bar{X}) = 90$ cm et $s_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.6}{\sqrt{200}} = 0.042$.

Comme $X \sim \mathcal{N}(90, 0.6) \rightarrow \bar{X} \sim \mathcal{N}(90, 0.042)$

$$P(\bar{X} \leq 89.9) = P\left(T \leq \frac{89.9 - 90}{0.042}\right) = P(T \leq -2.38) = 1 - \pi(2.38) = 0.0087 \rightarrow 0.87\%$$

Production journalière $= N = 2\ 000 \rightarrow \frac{n}{N} = 0.1 \rightarrow$ on doit tenir compte du coefficient d'exhaustivité

$$s_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{0.6}{\sqrt{200}}\sqrt{\frac{2000-200}{2000-1}} = 0.04$$

$$\bar{X} \sim \mathcal{N}(90, 0.04)$$

$$P(\bar{X} \leq 89.9) = P\left(T \leq \frac{89.9 - 90}{0.04}\right) = P(T - 2.5) = 1 - \pi(2.5) = 0.0062 \rightarrow 0.62\%$$

b) Même si l'on ne fait plus l'hypothèse que X soit une variable normale, comme $n = 200 > 30$, le théorème central limite permet de dire quel $\bar{X} \sim \mathcal{N}(90, 0.042)$ pour $N = 10000$.

On trouvera donc la même probabilité $P(\bar{X} \leq 89.9) \rightarrow 0,87\%$.

Pour $N = 2000$ $\bar{X} \sim \mathcal{N}(90, 0.04)$ et on trouve donc la même probabilité $P(\bar{X} \leq 89.9) \rightarrow 0,62\%$.

c) L'échantillon est représentatif de la population mère avec un intervalle de confiance de 95 % lorsque :

$$P(\mu - t\, s_{\bar{X}} \leq \bar{x} \leq \mu + t\, s_{\bar{X}}) = 0.95$$

Lorsque la probabilité d'un intervalle symétrique est de 0.95, on a

$$t = 1.96 \quad \left(\pi(t) - \pi(-t) = 2\pi(t) - 1 = 0.95 \rightarrow \pi(t) = \frac{1.95}{2} = 0,975 \rightarrow t = 1,96\right).$$

$$P(90 - 1.96 \times 0.042 \leq \bar{x} \leq 90 + 1.96 \times 0.042) = 0.95$$

L'intervalle est donc $[89.917; 90.082]$. Comme $\bar{x} = 90.3$ cm, ne se situe pas dans cet intervalle de confiance, l'échantillon n'est pas jugé représentatif de la population mère (avec un risque de 5 % de se tromper).

C) Distribution de $\bar{X}_1 - \bar{X}_2$

Il peut arriver en statistique que l'on désire comparer 2 populations relativement à une certaine caractéristique X.
Population 1 de taille N_1 : caractéristique X_1, moyenne : μ_1, variance σ_1^2, écart-type σ_1
Population 2 de taille N_2 : caractéristique X_2, moyenne : μ_2, variance σ_2^2, écart-type σ_2

Pour comparer ces 2 populations, on tire indépendamment un échantillon aléatoire de taille n_1 dans la 1re et un échantillon aléatoire de taille n_2 dans la 2e et on considère la distribution de la différence $(\bar{X}_1 - \bar{X}_2)$.

D'après les propriétés de l'espérance mathématique et de la variance, on a :

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$$

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad \rightarrow \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{s^2_{\bar{X}_1} + s^2_{\bar{X}_2}}$$

Si $\quad X_1 \sim \mathcal{N}(\mu_1, \sigma_1),\ X_2 \sim \mathcal{N}(\mu_2, \sigma_2) \rightarrow \left(\bar{X}_1 - \bar{X}_2\right) \sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$

— Si n_1 et n_2 sont grands (supérieurs à 30), quelles que soient les distributions de X_1 et X_2, $(\bar{X}_1 - \bar{X}_2)$ suivra une loi normale de mêmes paramètres, en vertu du théorème central limite.
— On utilisera le facteur d'exhaustivité dans les mêmes conditions (tirages sans remise, populations finies avec $n_i > 0.05N_i$).

Exemple 5.2.4 /Feuille 8/ Deux sociétés fabriquent des piles électriques d'un certain format.

Les piles de la société 1 ont une durée d'utilisation moyenne de 230 heures avec un écart type de 30 heures. Les piles de la société 2 ont une durée d'utilisation moyenne de 210 heures

avec un écart type de 20 heures. Quelle est la probabilité que la durée d'utilisation moyenne d'un échantillon aléatoire simple de 100 piles de la société 1 soit d'au moins 30 heures de plus que la durée d'utilisation moyenne d'un échantillon aléatoire simple de 125 piles de la société 2 ?

Solution :
Soit :

X_1 la durée d'utilisation des piles de la société 1,

X_2 la durée d'utilisation des piles de la société 2.

On ne connaît pas les distributions de X_1 et X_2, mais comme les tailles $n_1 = 100$ et $n_2 = 125$ sont grandes (> 30), on peut dire que :

$$\begin{aligned}
(\bar{X}_1 - \bar{X}_2) &\sim \mathcal{N}(230 - 210, \sqrt{\frac{30^2}{100} + \frac{20^2}{125}} \\
(\bar{X}_1 - \bar{X}_2) &\sim \mathcal{N}(20; 3.493) \\
P(\bar{X}_1 - \bar{X}_2 \geq 30) &= P\left(Z > \frac{30 - 20}{3.493}\right) = P(Z > 2.86) \\
&= 1 - \pi(2.86) = 0.0021 = 0.21\%
\end{aligned}$$

II. Cas : variance σ^2 de la population inconnue

A. Un grand échantillon ($n \geq 30$) permet de déduire une valeur fiable pour σ^2 en calculant la variance de l'échantillon s^2 et en posant

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

Les remarques précédentes restent valables :

Un grand échantillon $n \geq 30$ de variance s

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{s}{\sqrt{n-1}}\right)$$

B. Cas des petits échantillons : $n < 30$
On considère exclusivement le cas où X suit une loi normale dans la population.

Lorsque l'échantillonnage s'effectue à partir d'une population normale de variance inconnue et que la taille de l'échantillon est petite ($n < 30$), l'estimation de la variance effectuée par la variance de l'échantillon n'est plus fiable. Comme s^2 varie trop d'échantillon en échantillon, on ne peut plus écrire que $\sigma^2 \approx \frac{n}{n-1}s^2$. L'écart-type de la distribution de $\bar{X}$ $\frac{\sigma}{\sqrt{n}}$, approximé par $\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n-1}}$ n'est plus une constante et sa valeur varie dans chaque échantillon.

La variable écart-type d'échantillon, notée S et définie par $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ est une variable aléatoire. Le dénominateur de la variable aléatoire $T = \frac{\bar{X}-\mu}{s/\sqrt{n-1}} = \frac{\sqrt{n-1}(\bar{X}-\mu)}{s}$ n'est pas une constante ; d'ici la variable T ne suit plus alors une loi normale.

En divisant numérateur et dénominateur par σ, on écrit T sous la forme

$$T = \frac{\sqrt{n-1}(\bar{X}-\mu)}{s} = \frac{\sqrt{n-1}\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2}},$$

dont le numérateur est composé par une variable aléatoire qui suit une loi $\mathcal{N}(0,1)$, multipliée par un facteur$\sqrt{n-1}$, et le dénominateur est une somme de carrés de variables suivant aussi la loi $\mathcal{N}(0,1)$. Le carré du dénominateur suit donc une loi du χ^2. Pour pouvoir utiliser correctement les tables du χ^2 il faut déterminer le nombre de degrés de liberté. Le nombre de degrés de liberté est toujours associée à une somme de carrés et représente le nombre de carrés indépendants dans cette somme. On peut calculer le nombre de degrés de liberté d'après deux règles :
- on effectue la différence entre le nombre total de carrés et le nombre de relations qui lient les différents éléments de la somme ;
- on effectue la différence entre le nombre total de carrés et le nombre de paramètres que l'on doit estimer pour effectuer le calcul.

Pour déterminer les degrés de liberté de la somme $\sum_{i=1}^{n}(\frac{X_i-\bar{X}}{\sigma})^2$, d'après la première règle le nombre de carrés dans la somme est n. Il y a une relation entre les variables $\sum_{i=1}^{n}(X_i-\bar{X})=0$. Le nombre de degrés de liberté est donc $n-1$.

D'après la deuxième règle le nombre de carrés dans la somme est n. Lorsqu'on dit que $\sum_{i=1}^{n}(\frac{X_i-\bar{X}}{\sigma})^2$ est une somme de carrés de variables normales centrées réduites, on remplace μ par $\bar{X}$. On a estimé un paramètre. donc le nombre de degrés de liberté est $n-1$.

Si $n<30$, et σ inconnu, la variable $T=\frac{\bar{X}-\mu}{s/\sqrt{n-1}}$ suit une loi de Student à $n-1$ degrés de liberté, notée T_{n-1}.

Exemple 5.2.5 /Feuille 8/ Le responsable d'une entreprise a accumulé depuis des années les résultats à un test d'aptitude à effectuer un certain travail. Il semble plausible de supposer que les résultats au test d'aptitude sont distribués suivant une loi normale de moyenne $\mu=150$ et de variance $\sigma^2=100$. On fait passer le test à 25 individus de l'entreprise. Quelle est la probabilité que la moyenne de l'échantillon soit entre 146 et 154 ?

Solution :

On considère la variable aléatoire $\bar{X}$ moyenne d'échantillon pour les échantillons de taille $n=25$. On cherche à déterminer $P(146<\bar{X}<154)$.

Pour cela, il nous faut connaître la loi suivie par $\bar{X}$. Examinons la situation. Nous sommes en présence d'un petit échantillon ($n<30$) et heureusement dans le cas où la variable X (résultat au test d'aptitude) suit une loi normale. De plus, σ est connu. Donc $\bar{X}$ suit $\mathcal{N}(\mu,\frac{\sigma}{\sqrt{n}})=\mathcal{N}(150,10/5)$. On en déduit que $T=\frac{\bar{X}-150}{2}$ suit $\mathcal{N}(0,1)$.

La table donne

$$\begin{aligned} P(146<\bar{X}<154) &= P\left(\frac{146-150}{2}<T<\frac{154-150}{2}\right)=P(-2<T<2)\\ &= 2P(0<T<2)=2\times(P(T<2)-P(T<0))=2\times(0,9772-0,5)\\ &= 2\times 0.4772=0.9544. \end{aligned}$$

5.2.2 Distribution de la variance d'échantillon $S'^2_{\bar{X}}$

Supposons que X suit une loi normale.

On considère la variable $Y = \frac{nS^2_{\bar{X}}}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2$.

Y est une somme d'écarts réduits relatifs à une variable normale, donc Y suit une loi du χ^2 à $n-1$ degrés de liberté (on perd un degré de liberté car on a estimé le paramètre μ par $\bar{X}$).

$$Y = \frac{(n-1)S'^2_{\bar{X}}}{\sigma^2} \sim \chi^2_{n-1}.$$

Approximation de la distribution de S'^2 dans le cas des grands échantillons : $n \geq 30$

Lorsque n est grand ($n \geq 30$), on peut approcher la loi χ^2_ν par la loi $\mathcal{N}(\nu, \sqrt{2\nu})$. Donc Y suit approximativement une loi normale, $E(Y) \approx n-1$ et $Var(Y) \approx 2(n-1)$.

$$\textbf{Si } n \geq 30,\ S'^2_{\bar{X}} \sim \mathcal{N}\left(\sigma^2, \sigma^2\sqrt{\frac{2}{n-1}}\right) \text{ en première approximation.}$$

,

La loi de $S'^2_{\bar{X}}$ est alors approximativement normale, son espérance vaut σ^2 et sa variance approximativement

$$Var(S'^2_{\bar{X}}) = Var\left(\frac{\sigma^2}{n-1}Y\right) = \frac{\sigma^4}{(n-1)^2}Var(Y) \approx \frac{2\sigma^4}{n-1}.$$

5.2.3 Distribution d'échantillonnage d'une proportion F

Dans certaines circonstances en gestion, on peut traiter les données sous forme de proportions (taux d'absentéisme, de rebuts, de réussite...).

Notations :

Population mère : p : proportion moyenne ; $q = 1 - p$ = proportion complémentaire

Echantillon : f : fréquence observée dans l'échantillon de taille n.

Soit F la fréquence d'apparition du caractère dans un échantillon de taille n. Donc $F = X/n$ où X est le nombre de fois où le caractère apparaît dans le n-échantillon.

Par définition X suit $\mathcal{B}(n, p)$. Donc $E(X) = np$ et $Var(X) = npq$.

A) Si la population est infinie ou si l'échantillonnage est non exhaustif (tirage avec remise), on montre que :

$$E(F) = p; \quad s^2_F = \frac{pq}{n}; \quad s_F = \sqrt{\frac{pq}{n}}$$

Si n est grand ($n \geq 30$) et $np \geq 15, nq \geq 15$, alors $\mathcal{B}(n, \frac{p}{n}) \to \mathcal{N}(p, \sqrt{\frac{pq}{n}})$ et d'ici $F \sim \mathcal{N}(p, \sqrt{\frac{pq}{n}})$

B) Si l'échantillonnage est exhaustif (tirage sans remise) dans une population finie (avec $n > 0.05N$) : on doit tenir compte du facteur d'exhaustivité.

$$F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}\right)$$

Échantillonnage exhaustif (tirage sans remise) dans une population finie (avec $n > 0.05N$)

$$F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}\right)$$

Exemple 5.2.6 /Feuille 8/ [5] Le directeur financier d'une société sait par expérience que 12 % des factures émises ne sont pas réglées dans les 10 jours ouvrables suivant l'échéance. Il fait prélever un échantillon aléatoire de 500 factures.
Quelle est la probabilité qu'au moins 70 factures ne sont pas réglées dans le délais, sachant que l'ensemble des factures pouvant être étudiées est de plusieurs dizaines de milliers.

Solution :

Soit F = "proportion d'échantillon dans un échantillon de taille 500". $P\left(F \geq \frac{70}{500}\right)$ =? - Distribution d'échantillonnage d'une proportion F ; échantillonnage exhaustif (tirage sans remise) dans une population finie, mais $n < 0,05N$, donc il ne faut pas tenir compte du facteur d'exhaustivité.
Ici $p = 0.12$, $q = 1 - p = 1 - 0.12 = 0.88$.
Comme $n = 500 > 30$, $np = 500 * 0,12 = 60 > 15$, $nq = 500 * 0,88 = 440 > 15 \Longrightarrow$ approximation de la loi binomiale par la loi normale :

$$F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right) = \mathcal{N}\left(0,12; \sqrt{\frac{0,12 * 0,88}{500}}\right) = \mathcal{N}(0,12;\ 0,015)$$

$$\begin{aligned} P\left(F \geq \frac{70}{500}\right) &= P\left(F > \frac{69,5}{500}\right) = P\left(Z > \frac{0,139 - 0,12}{0,015}\right) \\ &= 1 - P\left(Z < \frac{0,019}{0,015}\right) = 1 - P(Z < 1,27) = 1 - \pi(1,27) = 1 - 0,8997 \approx 0,1 \end{aligned}$$

$\approx 10\%$ de chances pour que plus de 70 factures dans un 500 échantillon soient non réglées dans le délais.

Exemple 5.2.7 /Feuille 8/ Selon une étude sur le comportement du consommateur, 25% d'entre eux sont influencés par la marque, lors de l'achat d'un bien. Si on interroge 100 consommateurs pris au hasard, quelle est la probabilité pour qu'au moins 35 d'entre eux se déclarent influencés par la marque ?

Solution :

Soit F = "proportion d'échantillon dans un échantillon de taille 100".

$P(F > 0,35) = ? \Longrightarrow$ il faut déterminer la loi de F. $n = 100 > 30$; $np = 100 \times 0,25 = 25 > 15$ et $nq = 100 \times 0,75 = 75 > 15$

$$\Longrightarrow F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right) = \mathcal{N}(0,25, 0,0433).$$

On utilise la variable $Z = \frac{F-0,25}{0,0433}$ qui suit la loi $\mathcal{N}(0,1)$.

$$\begin{aligned} P(F > 0,35) = P(Z > 2,31) &= 0,5 - P(0 < Z < 2,31) \\ &= 0,5 - 0,4896 = 0,0104. \end{aligned}$$

Conclusion :. Il y a environ une chance sur 100 pour que plus de 35 consommateurs dans un 100 - échantillon se disent influencés par la marque lorsque l'ensemble de la population contient 25% de tels consommateurs.

C) Distribution de $F_1 - F_2$

Lorsque n_1 et n_2 sont grands, alors :

$$(F_1 - F_2) \sim \mathcal{N}\left(p_1 - p_2; \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$

5.3 Synthèse sur les distributions d'échantillonnage

Variable aléatoire	Définition	Paramètres descriptifs	Loi
F Proportion d'échantillon	$F = X/n$, $X \sim \mathcal{B}(n,p)$ $E(X) = np$ $Var(X) = npq$	$E(F) = p$ $Var(F) = \frac{pq}{n}$	$n \geq 30$, $np > 15$, $nq > 15$ $\mathcal{B}(n, \frac{p}{n}) \to \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$
			tirage avec remise (sans remise et $n < 0,05N$) $F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right)$
			tirage sans remise et $n > 0,05N$ $F \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}\right)$
$F_1 - F_2$ $F_1 \sim \mathcal{N}\left(p_1, \sqrt{\frac{p_1q_1}{n_1}}\right)$ $F_2 \sim \mathcal{N}\left(p_2, \sqrt{\frac{p_2q_2}{n_2}}\right)$	$F_1 - F_2$	$E(F_1 - F_2) = p_1 - p_2$ $Var(F_1 - F_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$	$n_1 \geq 30$; $n_2 \geq 30$ $F_1 - F_2 \sim \mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}\right)$

TABLE 5.1 : Synthèse sur les distributions d'échantillonnage

<table>
<tr><th>Variable aléatoire</th><th>Définition</th><th>Paramètres descriptifs</th><th colspan="4">Loi</th></tr>
<tr><td rowspan="4">$\bar{X}$ Moyenne d'échantillon</td><td rowspan="4">$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$
$= \frac{1}{n}\sum_{i=1}^{n} X_i$</td><td rowspan="4">$E(\bar{X}) = \mu$
$Var(\bar{X}) = \frac{\sigma^2}{n}$</td><td colspan="2">$n \geq 30$</td><td colspan="2">$n < 30,\ X \sim \mathcal{N}(\mu, \sigma)$</td></tr>
<tr><td>σ connu</td><td>σ inconnu
estimation fiable
$\hat{\sigma}^2 = \frac{n}{n-1}s^2$</td><td>$\sigma$ connu</td><td>σ inconnu
estimation fiable
$\hat{\sigma}^2 = \frac{n}{n-1}s^2$</td></tr>
<tr><td colspan="2">tirage avec remise ;
tirage sans remise et $n < 0,05N$
$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$</td><td rowspan="2">$T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n-1}}}$
$T \sim T_{n-1}$</td></tr>
<tr><td colspan="2">tirage sans remise et $n > 0,05N$
$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right)$</td></tr>
<tr><td rowspan="3">$X_1 : n_1, \mu_1, \sigma_1$
$X_2 : n_2, \mu_2, \sigma_2$</td><td rowspan="3">$\bar{X}_1 - \bar{X}_2$</td><td rowspan="3">$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$;
$Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$</td><td colspan="2">$n_1, n_2 \geq 30$; $n_i < 0,05N$</td><td colspan="2">$n_1, n_2 < 30$ et
$X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$,
$X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$</td></tr>
<tr><td colspan="4">$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$</td></tr>
<tr><td colspan="4">$n_i > 0,05N \rightarrow$ facteur d'exhaustivité</td></tr>
<tr><td rowspan="2">$S'^2_{\bar{X}}$ Variance d'échantillon
- estimation de σ^2</td><td>$S^2_{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$</td><td>$E(S^2_{\bar{X}}) = \frac{n-1}{n}\sigma^2$,</td><td colspan="2">$n \geq 30$</td><td colspan="2">$n < 30$</td></tr>
<tr><td>$S'^2_{\bar{X}} = \frac{n}{n-1}S^2$
$= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$</td><td>$E(S'^2_{\bar{X}}) = \sigma^2$</td><td colspan="2">$S'^2_{\bar{X}} \sim \mathcal{N}\left(\sigma^2, \sigma^2\sqrt{\frac{2}{(n-1)}}\right)$</td><td colspan="2">$\frac{(n-1)S'^2_{\bar{X}}}{\sigma^2} \sim \chi^2_{n-1}$</td></tr>
</table>

TABLE 5.2 : Synthèse sur les distributions d'échantillonnage

Bibliographie

[1] Aragon, Y. *Séries temporelles avec R* Springer, 2011.

[2] Brockwell, P. J., Davis, R. A. *Time series, theory and methods*, Second edition, Springer, 2006

[3] Giard, V. *Statistique appliquée à la gestion avec exercices corrigés et utilisation d'Excel*, Economica, Paris, 1995

[4] Gouriéroux, C., Monfort, A. *Séries temporelles et modèles dynamiques*, Economica, 1995

[5] Dumoulin, D. *Mathématiques de gestion. Cours et applications* Collection D.E.C.S. dirigée par Th. Lamorlette, Economica, Paris, 1987

[6] Justens, D. *Statistique pour décideurs* De Boeck - Entreprise, Bruxelles, 1990

Annexe

Feuilles

Feuille 1 : Vocabulaire et concepts de base

Feuille 2 : Organisation d'une série univariée

Feuille 3 : Synthèse par l'image

Feuille 4 : Synthèse par des paramètres

Feuille 5 : Paramètres de dispersion : étendue, écarts interquartile ou interdécile, écart moyen, écart-type, variance. Paramètres de forme

Feuille 6 : Série statistique bivariée

Feuille 7 : Séries chronologiques

Feuille 8 : Échantillonnage

Exemples

Organisation d'une série statistique univariée. D.O.1

Organisation d'une série statistique univariée. D.G.1

Organisation d'une série statistique univariée. Exemple 2.2.1

Boite à moustaches. Exercice 29

Ajustement linéaire. Exercice 40

Séries chronologiques. Exercice 48

Séries chronologiques. Exercice 49

Echantillonnage. Exercice 61

Tables statistiques

Table de Loi Normale

Fractiles de la Loi Normale

Fractiles de la loi du χ^2_ν

Table de la loi de Student

Feuilles

Feuille 1 : Vocabulaire et concepts de base

Feuille 2 : Organisation d'une série univariée

Feuille 3 : Synthèse par l'image

Feuille 4 : Synthèse par des paramètres

Feuille 5 : Paramètres de dispersion : étendue, écarts interquartile ou interdécile, écart moyen, écart-type, variance. Paramètres de forme

Feuille 6 : Série statistique bivariée

Feuille 7 : Séries chronologiques

Feuille 8 : Échantillonnage

Feuille 1 : Vocabulaire et concepts de base

Ensemble observé (population), caractère observé, nature du caractère observé, modalités du caractère observé, effectif d'une modalité, effectif total, effectifs cumulés, fréquence, fréquences cumulées.

Exemple 1.0.1 Le principal d'un Université étudie les notes du dernier examen de mathématiques de 20 étudiants d'un groupe. Voici la liste des notes obtenues par les étudiants :

10 – 7 – 5 – 9 – 13 – 11 – 16 – 17 – 14 – 13 – 16 – 8 – 6 – 10 – 8 – 11 – 10 – 12 – 7 – 9.

Exemple 1.0.2 [Groupes ethniques] La répartition des groupes ethniques dans une classe de 30 élèves donne le tableau suivant :

Groupes ethniques	Poular	Wolof	Sérère	Diola	Bambara
Effectifs	6	9	7	5	3

Exemple 1.0.3 [Taille] Voici les tailles (en cm) des 25 élèves d'une classe de 3ème :

165 145 150 150 166 165 160 158 162 165 158 165 162 154 158 160 162 154 165 160 160 158 154 158 160.

Exemple 1.0.4 On considère les 14 étudiants en Bases de la statistique 2 en 2015. On s'intéresse au nombre d'absences aux cours et au nombre de participation au tableau pendant les cours. On choisit au hasard 5 étudiants parmi les 14 et analyse leurs résultats.

Exercices

1. Dans un ensemble de six amis A, B, C, D, E, F on relève les "mesures" suivantes :

Individus	sexe	nombre d'enfants	taille(en m)	poids(en kg)
A 1	m	5	1,80	82
B 2	m	2	1.75	91
C 3	f	4	1.67	57
D 4	m	1	1.90	83
E 5	m	0	1.77	77
F 6	f	2	1.75	67

2. Répartition de la population française par tranche d'âge (en millions d'individus) en 1987

Classes	effectifs
0-20 ans	17
20-40 ans	15
40-60 ans	12
60-80 ans	9
plus de 80 ans	2
ensemble	55

Ensemble observé :
Caractère observé
Nature du caractère observé
Modalités du caractère observé
Effectif total

3. Enquête concernant la couleur des yeux des personnes interrogées

modalités x_i	foncés	verts ou bleus	gris	ensemble
effectifs n_i	1015	595	140	1750
fréquences f_i				
fréquences en %				

4. Considérons la série relative à la population (en millions de personnes) d'un pays entre 1910 et 1980

Année	1910	1920	1930	1940	1950	1960	1970	1980
Population	1,1	1,3	2,2	4,0	6,6	8,3	9,3	9,6

5. Deux instituteurs corrigent une série de travaux de mathématique. Le premier attribue à chaque copie une note de 1 à 5 - caractère quantitatif. Le second note en utilisant les appréciations : "parfait"; "bon"; "suffisant"; "insuffisant" - caractère qualitatif ordinal.

6. Deux instituts de sondage estiment l'impact auprès du public de leur dernier enquête. Le premier demande à 1000 personnes de noter sur 5. Le second demande à 800 personnes de souligner l'appréciation qui leur semble la plus correcte : "très convainc"; " convainc"; "indifférant"; "sans impact"; "avec effet"; "négatif".

Différence entre série statistique et chronologique

7. Mesure du nombre de suicides selon le jour de la semaine - série chronologique.

8. Nombre d'objets défectueux dans une population. On peut mesurer ce nombre pour une durée déterminée, chaque jour pendant deux mois; par machine par mois; il faut déterminer dans quelle mesure ces modifications influencent la qualification.

- chaque jour pendant deux mois - série chronologique.
- par machine par mois - série statistique.

Devoir 1. Pour les séries des Exercices 1 et 2 de Feuille 1 déterminer : ensemble observé; caractère observé - nom, nature /type/ du caractère observé, modalités, effectif total.

Feuille 2 : Organisation d'une série univariée

Exemple 2.1.1 [Age] Ages de 100 employés d'une entreprise (échantillon) et La série ordonnée

60	39	23	30	29	26	29	41	40	32	17	18	20	21	21	21	22	22	23	23
63	22	32	52	46	35	25	28	33	33	23	24	25	25	25	25	25	25	25	26
20	25	42	34	29	43	41	31	30	36	26	26	26	26	27	28	28	28	29	29
58	21	24	55	51	28	18	40	44	38	29	30	30	30	30	30	31	31	31	31
32	21	30	31	25	49	31	26	33	36	32	32	32	32	33	33	33	33	33	33
43	34	35	22	33	38	34	34	33	34	33	34	34	34	34	34	34	34	35	35
23	26	57	23	26	36	39	31	35	34	35	35	36	36	36	36	36	37	38	38
34	51	40	50	35	45	28	36	32	39	38	39	39	39	40	40	40	41	41	42
26	48	17	45	45	25	25	30	36	30	43	43	43	44	45	45	45	46	48	49
43	25	27	21	53	25	38	33	37	33	50	51	51	52	53	55	57	58	60	63

Exemple 2.1.2 [Notes] Relève des notes d'un groupe de 30 étudiants à l'examen en Statistique. On observe la variable aléatoire X ="relève des notes d'un étudiant".

On a les résultats les suivants :

$3; 2; 4; 6; 5; 2; 3; 6; 4; 4; 2; 3; 3; 4; 4$

$5; 6; 4; 4; 3; 3; 3; 3; 4; 4; 4; 4; 5; 5; 5$

La série ordonnée est :

$2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 4$

$4; 4; 4; 4; 4; 4; 4; 5; 5; 5; 5; 5; 6; 6; 6$

Exemple 2.1.3 [Bovins] Nombre de bovins dans les fermes privées dans une région donnée. Taille de la population $N = 60$.

Série observée :

$4; 2; 3; 5; 0; 1; 6; 1; 5; 4; 7; 10; 13; 16; 19; 8; 4; 11; 14; 17; 2; 11$

$9; 8; 12; 15; 18; 10; 13; 18; 7; 11; 14; 17; 21; 8; 9; 11; 10; 16; 8; 10$

$15; 17; 19; 11; 9; 13; 11; 11; 12; 9; 10; 11; 10; 12; 10; 11; 12; 14$

Série ordonnée

$0; 1; 1; 2; 2; 3; 4; 4; 4; 5; 5; 6; 7; 7; 8; 8; 8; 8; 9; 9; 9; 9$

$10; 10; 10; 10; 10; 10; 10; 11; 11; 11; 11; 11; 11; 11; 11; 11; 12; 12; 12; 12$

$13; 13; 13; 14; 14; 14; 15; 15; 16; 16; 17; 17; 17; 18; 18; 19; 19; 21$

Exercices

9. Dans une entreprise de vente par correspondance on a relevé le nombre de commandes effectuées dans l'année par chacun des vingt-cinq clients dont les noms figurent sur une page du fichier. On a obtenu : 3 ; 1 ; 2 ; 1 ; 3 ; 3 ; 5 ; 4 ; 2 ; 3 ; 1 ; 2 ; 2 ; 2 ; 2 ; 1 ; 2 ; 1 ; 4 ; 1 ; 2 ; 2 ; 2 ; 1 ;1. (série observée). Écrire la série ordonnée. Donner la distribution statistique observée :

Nombre de commandes x_i					
Effectifs n_i					
Effectifs cumulés N_i					
Fréquences f_i					
Fréquences cumulées g_i					

Déterminer le pourcentage de clients ayant passé au plus 3 commandes.

10. Pour chacun des 25 clients de l'exercice 9., on a relevé le montant total en euros des commandes pour une année. Les résultats figurent dans le tableau suivant :

Montant en €	Effectifs (n_i)	centre de classe (x_i^*)	N_i	f_i	g_i	$n_i x_i^*$
[600, 800 [	6					
[800, 1000[	8					
[1000, 1200[	6					
[1200, 1400[	4					
[1400, 1600[	1					

Déterminer le pourcentage de commandes dont le montant est supérieur ou égal à 800 € et strictement inférieur à 1 400 €. Déterminer la fréquence pour un client d'avoir passé des commandes pour un total strictement inférieur à 1000 €. Quel est le montant total de toutes les commandes de ces 25 clients ?

Devoir 2. Exercice [Profil des lecteurs du Journal de la commune]
Reprenons le tableau individus x caractères présentant les réponses à 4 questions d'une enquête menée auprès d'un échantillon de 40 habitants d'une certaine commune afin d'étudier leurs habitudes de lecture du Journal trimestriel de la commune.

Individu i	Age x_i	Nombre de personnes dans le foyer y_i	Fréquence de lecture z_i	Sexe w_i	Individu i	Age x_i	Nombre de personnes dans le foyer y_i	Fréquence de lecture z_i	Sexe w_i
1	17	4	régulièrement	femme	21	10	3	jamais	homme
2	12	2	rarement	homme	22	40	5	régulièrement	femme
3	15	3	rarement	femme	23	54	5	rarement	femme
4	87	1	toujours	femme	24	25	3	régulièrement	homme
5	32	1	jamais	femme	25	53	4	rarement	femme
6	33	2	régulièrement	homme	26	27	3	rarement	femme
7	45	4	jamais	homme	27	57	4	régulièrement	homme
8	46	1	rarement	homme	28	59	2	régulièrement	femme
9	29	2	régulièrement	homme	29	13	5	rarement	femme
10	38	3	rarement	femme	30	53	3	régulièrement	homme
11	76	2	toujours	homme	31	67	3	toujours	femme
12	65	2	toujours	femme	32	16	5	rarement	homme
13	59	6	régulièrement	femme	33	55	4	rarement	homme
14	12	2	jamais	homme	34	49	6	régulièrement	femme
15	14	4	régulièrement	homme	35	58	2	jamais	femme
16	15	2	rarement	homme	36	21	2	jamais	homme
17	66	2	rarement	femme	37	95	2	rarement	femme
18	38	2	rarement	femme	38	28	3	régulièrement	homme
19	40	4	régulièrement	femme	39	65	2	régulièrement	femme
20	42	5	régulièrement	homme	40	89	1	toujours	homme

a) Construisez la D.O.1 des effectifs et des fréquences pour la variable « Sexe ».

b) Construisez la D.O.1 des effectifs et des fréquences pour la variable « Fréquence de lecture ».

c) Construisez la D.O.1 des effectifs et des fréquences pour la variable « Nombre de personnes dans le foyer » .

d) Construisez la D.G.1 des effectifs et des fréquences pour la variable « Age ».

Feuille 3 : Synthèse par l'image

Exemple 2.1.4
Le tableau suivant donne la composition en acides gras insaturés en grammes pour 100 grammes d'huile d'olive vierge :

Modalité	Effectif n_i	$\alpha_i = 360 f_i$
Acide Oléique	18,6	$360\frac{18,6}{100} = 66,96$
Acide Linoléique	58,6	$360\frac{58,6}{100} = 210,96$
Acide Liolénique	12,7	$360\frac{12,7}{100} = 45,72$

Exercices

11. Construire le diagramme en bâtons des effectifs, des effectifs cumulés croissants, des fréquences et des fréquences cumulées croissantes de la distribution statistique de l'Exercice 9.

12. Construire l'histogramme des effectifs, le polygone des effectifs, l'histogramme des effectifs cumulés croissants et le polygone des effectifs cumulés croissants du tableau statistique de l'Exercice 10.

13. Prenons comme exemple la population des rencontres de football d'un week-end dont les résultats sont publiés dans les pages sportives d'un journal donné. Comme variable statistique, nous prendrons le nombre total de buts marqués au cours de chacune des rencontres. Nous constituons ainsi un échantillon de 50 rencontres, ce qui nous donne les nombres suivants :

2 0 2 8 2 2 4 2 0 4 0 5 2 2 5 2 0 3 3 3 0 2 2 1 3
2 4 0 2 7 8 1 1 5 3 6 2 3 1 2 1 0 4 5 2 4 3 1 1 6

Construire le diagramme en bâtons des effectifs, des effectifs cumulés croissants, des fréquences et des fréquences cumulées croissantes

14. On a mesuré en cm la taille de 54 enfants ; Cela a donné les résultats suivants :

127 129 126 132 125 133 131 128 133 120 135 127 125 112
134 125 136 132 130 123 121 129 133 127 133 135 131 115
134 130 128 132 127 126 141 138 118 136 139 138 138 122
146 134 143 128 142 133 136 131 132 124 127 134

Construire l'histogramme des effectifs, le polygone des effectifs, l'histogramme des effectifs cumulés croissants, le polygone des effectifs cumulés croissants et le diagramme cumulatif du tableau statistique.

Devoir 3.
Considérez l'Exercice [Profil des lecteurs du Journal de la commune], Devoir 2. Faites une représentation graphique adéquate :

a) pour la variable « Sexe ».

b) pour la variable « Fréquence de lecture ».

c) pour la variable « Nombre de personnes dans le foyer ».

d) pour la variable « Age ». Construisez l'histogramme des fréquences cumulées et la courbe cumulative des fréquences.

Feuille 4 : Synthèse par des paramètres

Paramètres de position : Mode ; moyennes ; médiane ; quantiles

Exemple 2.2.1 Considérons une étude de prix d'un même article en fonction de la marque qui le commercialise. Ce genre d'étude est fréquent. On trouve le tableau suivant :

Intervalle de prix	$[165-170[$	$[170-175[$	$[175-180[$	$[180-185[$	$[185-190[$	$[190-195[$
Nombres de marques	6	10	5	4	3	2

Compléter le tableau de distribution. Calculer le mode, la médiane, la moyenne arithmétique.

Exemple 2.2.2

Modalité x_i	7	8	9	10	12	13	14	15	17	18	Total n
Effectifs n_i	2	1	3	2	1	1	2	2	2	1	17

Exemple 2.2.3

Modalité x_i	8	9	10	11	13	14	15	16	17	Total
Effectifs n_i	2	2	3	1	1	1	1	2	1	14

Exemple 2.2.4 Les affaires mensuels d'une compagnie ont augmentés pendant les premier 5 mois de l'année de 12% par mois et pendant les 7 suivants mois de 8%. Calculer la moyenne des croissance des ventes mensuelles i_m.

Exemple 2.2.5 Une voiture se déplace à mi-chemin avec une vitesse de $v_1 = 60km/h$, tandis que l'autre moitié - de vitesse de $v_2 = 70km/h$. Calculer la vitesse moyenne v_m de la voiture.

Exemple 2.2.6 Le tableau suivant décrit la répartition des accidents de la route selon les heures de la journée. On souhaite dégager les tendances essentielles de ces informations.

tranche horaire (en heures)	[0,3[	[3,6[	[6,9[	[9,12[	[12,15[	[15,18[	[18,21[	[21,24[
nombre d'accidents	8155	6258	15284	18006	23703	29759	29172	13022

Exercices

15. Un capital est placé un an à 5 %, puis à 15 % l'année suivante. Quel est le taux annuel moyen de placement ?

16. Un capital est placé un an à 7 %, puis à 5 % les trois années suivantes. Quel est le taux d'intérêt annuel moyen de placement pour la période de quartes ans ?

17. Une voiture va de A a B a 60 km/h et revient de B vers A a 100 km/h. Quelle est sa vitesse moyenne sur la totalité du parcours.

18. On donne les nombres : 3 ; 4 ; 6 ; 8 ; 10. Calculer la moyenne arithmétique, la moyenne géométrique et la moyenne harmonique de ces nombres.

19. Dans une société de 7 personnes les salaires mensuels (en EUR) sont respectivement de 1 000,1 200,1 300 , 1500, 1 700, 1 800 ,12 500.
Calculer le salaire moyen et le salaire médian dans cette société.

20. La moyenne de 4 nombres est 10, quel doit être le cinquième nombre pour que la moyenne soit 11,5 ?

21. Dans une matière, la note trimestrielle est la moyenne des notes obtenues à quatre contrôles. La moyenne des trois premières notes est 12. Pour quelles valeurs possibles de la quatrième note la moyenne trimestrielle sera-t-elle :
a/ supérieure à 12 ; b/ inférieure à 12 ; c/ égale à 12 ?

22. Dans une usine trois ouvriers fabriquent la même pièce. La productivité du travail du premier ouvrier est une pièce par 1 heure, du second - une pièce par 2 heures et du troisième - 1 pièce par 3 heures. Donner le temps moyen de production unitaire.

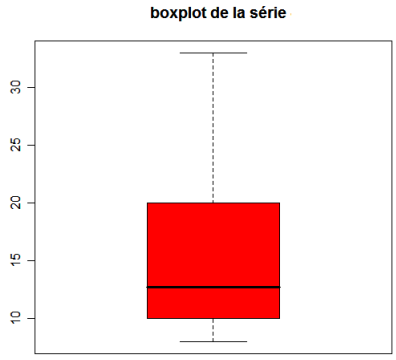
Devoir 4. : Question 1. : Déterminer les paramètres de position : mode, médiane, moyenne, les quartiles Q_1 et Q_3 pour les quatre caractères étudiés de l'Exercice [Profil des lecteurs du Journal de la commune], Devoir 2.
Question 2 : Exercice 16 ;
Question 3 : Exercice 22

Feuille 5 : Paramètres de dispersion : étendue, écarts interquartile ou interdécile, écart moyen, écart-type, variance. Paramètres de forme

Exemple 2.2.9

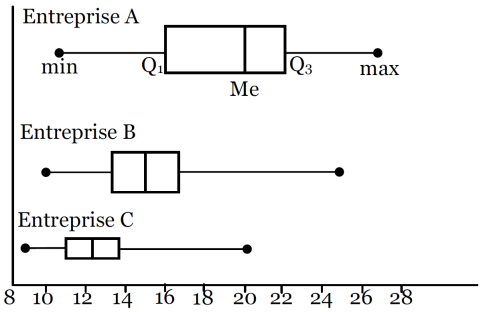
Série : 8; 8, 5; 10; 11; 12, 5; 13; 15; 20; 25; 33

Min.	Q_1	Me	Moyenne	Q_3	Max.
8.00	10,25	12,75	15,60	18,75	33,00



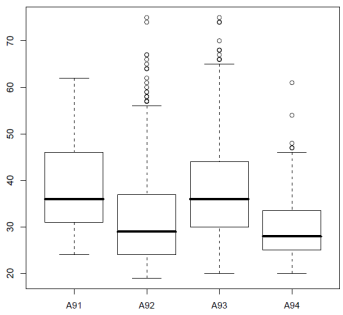
Exemple 2.2.10 Comparer les salaires dans les trois entreprises suivantes d'un même secteur industriel.

Entreprise	A	B	C
Taille	125	75	25
min	10 500	10 000	8 500
Q_1	16 000	13 500	11 000
M_e	20 000	15 000	12 500
Q_3	22 000	17 000	14 000
max	27 000	25 000	20 500



Exemple 2.2.11 La figure ci-dessous représente la distribution des 1000 clients d'une banque allemande, d'après leur état marital, ou on introduit le codage le suivant :

A91 : homme divorcé / séparé ;
A92 : femme divorcé / séparé / mariée ;
A93 : homme célibataire ;
A94 : homme marié / veuf .



Exemple 2.2.7 Considérons A et B mesurées dans la même monnaie. On a par exemple :

$$\bar{x}_A = 150 \quad s_A = 5.$$
$$\bar{x}_B = 50 \quad s_B = 3.$$

$\bar{x}$ décrit la cote moyenne de l'action, s est une mesure de sa *variabilité absolue.*
Commenter le risque des deux actions.

Exemple 2.2.8 On désire comparer les distributions (groupées) des bénéfices nets hebdomadaires en euros de 2 magasins, sur 100 semaines comprenant toutes 6 jours d'ouverture. Les parametres des deux distributions sont :
Magasin 1 : $\bar{x} = 2900$; $s = 1063$ Magasin 2 : $\bar{x} = 13000$; $s = 1077$.

Exercices

23. Déterminer la moyenne arithmétique, la médiane et les quartiles de chacune des distributions observées suivantes. Dessiner les boîtes « à moustaches »

x_i	n_i	$n_i x_i$	N_i
0	5		
1	8		
2	6		
3	3		
4	2		
5	1		

classes	n_i	x_i^*	$n_i x_i^*$	N_i	N_i^*
[0,20[	3				
[20 , 40 [	5				
[40 , 60 [	9				
[60 , 80 [	12				
[80 , 100[	5				
[100,120[	4				
[120 ,140 [	2				

24. On observe la série des notes (sur 20) obtenues par 2 étudiants A et B. Séries observées :
A - 10 ; 12 ; 7 ; 11 ; 8 ; 9 ; 13
B - 14 ; 5 ; 3 ; 6 ; 19 ; 6 ; 17
Déterminer la médiane, les quartiles, la moyenne arithmétique de ces deux séries. Quelles remarques pouvez-vous faire ? Calculer l'étendue, l'écart interquartile. De combien les notes obtenues s'écartent-elles de la moyenne ? (Calculer $x_i - \bar{x}$ pour chaque note). Calculer la somme des écarts à la moyenne pour chaque étudiant. Calculer la variance et l'écart type pour chaque étudiant.

25. Soit la distribution d'après le nombre de commandes de 25 clients (voir Exc. 9 Feuille 2.) Compléter le tableau ci-dessous. Déterminer les paramètres de dispersion de cette distribution statistique.

x_i	1	2	3	4	5	
n_i	8	10	4	2	1	25
N_i						
$n_i x_i$						
$(x_i - \bar{x})$						
$(x_i - \bar{x})^2$						
$n_i(x_i - \bar{x})^2$						
$n_i x_i^2$						

26. Un ouvrier a 2 chemins différents pour aller au travail. Les deux chemins exigent une attente au feux. L'ouvrier a enregistré le temps d'attente (en minutes) d'une série de 6 trajets pour chaque chemin. Les résultats sont données dans le tableau :

Chemin 1	15	15	11	17	14	12
Chemin 2	11	14	17	15	16	11

Calculer la moyenne et l'écart-type de la durée du chaque chemin et analyser leur fiabilité.

Paramètres de forme.

27. On donne les nombres suivants : 4, 5, 7, 8, 9. Calculer les moments d'ordre 1 , 2 et 3. Calculer les moments centrés d'ordre 1 , 2 , et 3.

28. On considère la distribution de 10 salariés d'après le salaire mensuel en milliers d'€. Déterminer le coefficient d'asymétrie et le coefficient d'aplatissement de cette distribution.

Classes	n_i	x_i	$n_i x_i$	x_i^2	$n_i x_i^2$	x_i^3	$n_i x_i^3$
[0, 10[	3						
[10, 20[	2						
[20, 30[	4						
[30, 40[	1						

29. Un jury de délibération désire analyser les résultats (notes sur 100 points) obtenus par 10 étudiants dans 7 matières distinctes. Le tableau ci-dessous est le tableau individus x caractères contenant ces résultats. Déterminer les « boîtes à moustaches » pour les 7 cours observés.

Étudiants	Matières						
	C_1	C_2	C_3	C_4	C_5	C_6	C_7
E_l	52	47	51	69	83	76	24
E_2	23	44	19	67	24	75	23
E_3	83	58	63	77	85	83	27
E_4	75	51	43	85	86	80	30
E_5	04	46	25	33	27	14	19
E_6	35	56	31	47	77	77	21
E_7	67	49	27	75	79	78	29
E_8	92	57	73	83	87	84	93
E_9	12	42	23	59	21	79	18
E_{10}	46	54	48	73	29	81	25

C_1	C_2	C_3	C_4	C_5	C_6	C_7
04	42	19	33	21	14	18
12	44	23	47	24	75	19
23	46	25	59	27	76	21
35	47	27	67	29	77	23
46	49	31	69	77	78	24
52	51	43	73	79	79	25
67	54	48	75	83	80	27
75	56	51	77	85	81	29
83	57	63	83	86	83	30
92	58	73	85	87	84	93

30. Paramètres de position et de dispersion

Considérerons la distribution observée des nombres de jours d'absence de 20 ouvriers d'une usine au cours d'une année :

Nombre de jours d'absence x_j	Effectif n_j	Eff. cumulé N_j
0	3	3
1	5	8
2	2	10
3	2	12
4	3	15
5	3	18
7	1	19
42	1	20
Total	$n = 20$	

a) Déterminez la valeur du 1^{er} quartile $x_{1/4}$
b) Déterminez la valeur du quantile d'ordre 1/3
c) Déterminez la valeur du 9^e décile
d) Quelle est l'étendue de cette D.O.1 ?
e) Déterminez l'écart interquartile de cette D.O.1.
f) Déterminez l'écart interdécile de cette D.O.1.
g) Nous avons déjà vu que $x_{1/4} = 1$, $x_{1/2} = 2$ et $x_{3/4} = 4$ (selon la 1^{re} convention). Construisez la boîte à moustaches associée à cette D.O.1.

31. Exercice 23 - modifié

Pour la série observée

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

a) Donner la distribution statistique (x_i, n_i), $i = 1, n$
b) Déterminer l'effectif total
c) Tracer le diagramme différentiel (en bâtons) des effectifs et le diagramme intégral (la courbe cumulative) des fréquences
d) Déterminer le mode et la médiane graphiquement et après par calcul
e) Calculer la valeur moyenne de la série
f) Déterminer le quartiles, le I et le 9^e déciles
g) Calculer la variance et l'intervalle interquartiles et l'intervalle interdécile. Quelles sont les propriétés de ces deux intervalles ?
h) Calculer l'étendue, l'écart-type et le coefficient de variation. Commenter.
i) Calculer le coefficient d'asymétrie et le coefficient d'aplatissement. Commenter.
j) Construire la boite à moustaches.
k) Pouvez vous construire la courbe de Lorentz ?

Pour la distribution observée suivante :

classes	[0,20 [	[20 , 40 [	[40 , 60 [	[60 , 80 [	[80 , 100[	[100,120[	[120 ,140 [
n_i	3	5	9	12	5	4	2

a) Déterminer l'effectif total
b) Donner une représentation graphique de la distribution observée et tracer la courbe cumulative des fréquences (polygone des fréquences)
c) Déterminer le mode et la médiane graphiquement et après par calcul
d) Calculer la valeur moyenne $\bar{x}$
e) Déterminer le quartiles, le I et le 9^e déciles
f) Calculer la variance, l'écart-type, l'étendue et le coefficient de variation, l'intervalle interquartiles et l'intervalle interdécile. Quelles sont les propriétés de ces deux intervalles ?
g) Calculer les coefficients d'asymétrie et d'aplatissement. Commenter.
h) Tracer la courbe de Lorentz de concentration. Commenter l'inégalité de la série.
i) Construire la boite à moustaches

Devoir 5. Question 1. : Pour les deux caractères quantitatifs étudiés de l'Exercice [Profil des lecteurs du Journal de la commune], Devoir 2. trouvez les paramètres de dispersion : l'étendue, l'intervalle interdécille, l'écart interdécile, l'écart moyen et la variance.
Question 2. : Courbe de Lorentz

LA COURBE DE LORENTZ ET LE COEFFICIENT DE GINI

Dans une entreprise X, de 354 salariés, on a la répartition des salaires suivante :

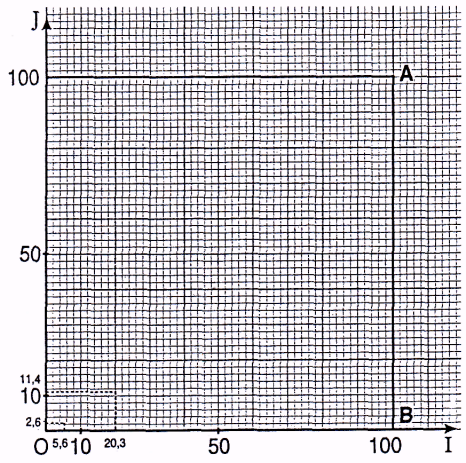
Salaire mensuel (en milliers de F)	[3 ; 4[	[4 ; 5[	[5 ; 8[	[8 ; 10[	[10 ; 12[	[12 ; 16[	[16 ; 20[	[20 ; 30[
Effectif	20	52	171	58	23	23	3	4

La masse salariale de cette entreprise est la somme de tous les salaires qu'elle doit verser chaque mois. Nous nous proposons d'étudier la répartition de la masse salariale dans cette entreprise. Pour cela, dressons le tableau suivant :

			I				J
Salaire mensuel (en milliers de francs)	Effectifs n_i	Fréq. (%)	Fréquences cumulées croissantes (%)	Masse salariale pour chaque classe : $n_i x_i$ où x_i est le centre de la classe	Masse salariale par classe, cumulée croissante	Fréq. (%) masse.	Masse salariale par classe, cumulée croissante (en %)
[3 ; 4[	20		5,6	70	70		2,6
[4 ; 5[	52		20,3	234	304		11,4
[5 ; 8[	171		68,6				
[8 ; 10[	58		85,0				
[10 ; 12[	23		91,5				
[12 ; 16[	23		98,0				
[16 ; 20[	3		98,9				
[20 ; 30[	4		100				
Masse salariale totale							

courbe de Lorentz

1. Placez, dans un repère orthogonal sur papier millimétré, les huit points de coordonnées (I, J), où I et J sont les nombres figurant dans la 3e et la 6e colonnes du tableau :
(5,6 ; 2,6), (20,3 ; 11,4), ...
Joignez ces points par une ligne courbe continue et régulière.
La courbe que vous avez tracée est **la courbe de Lorentz** de l'entreprise X.



Le rapport $\frac{\mathcal{A}}{\mathcal{B}}$, appelé **coefficient de Gini** est compris entre 0 et 1.

• **Le coefficient de Gini mesure la concentration des salaires ;** *plus sa valeur est proche de zéro, plus la répartition de la masse salariale est égalitaire.*

2. Calculez l'indice de Gini.

Feuille 6 : Série statistique bivariée

Tableau de contingence.

Exemple 3.1.1 On a interrogé 300 personnes à la sortie d'une grande surface et on a obtenu les résultats suivants :

	Achat du produit A	Non achat du produit A	
Homme	30	70	100
Femme	80	120	200
	110	190	300

Quel est le pourcentage de personnes qui ont acheté le produit A ?
Quel est le pourcentage de femmes qui n'ont pas acheté le produit A ?
Parmi ceux qui ont acheté le produit A quel est le pourcentage d'hommes ?

Exemple 3.1.2 Répartition des salaires mensuels par ancienneté et montant

		Répartition des salaires mensuels par ancienneté et montant (en $ liduriens)				
		5000 $	7000 $	9000 $	12 000 $	Ensemble
Répartition des salaires mensuels par ancienneté du salarié (en années)	1 an	87	57	11	3	158
	3 ans	39	45	14	19	117
	5 ans	15	36	47	25	123
	8 ans	8	14	24	9	55
	Ensemble	149	152	96	56	453

Quel est le salaire mensuel moyen dans l'entreprise ?
Quelle est l'ancienneté moyenne dans l'entreprise ?
Quel est le salaire mensuel moyen des salariés ayant 3 ans d'ancienneté ?
Quelle est l'ancienneté moyenne des salariés qui ont un salaire mensuel de 9 000$ liduriens ?

Exercices

32. En 1996, la Sécurité routière a dénombré 125406 accidents en France, les répertoriant dans un fichier de la façon suivante :

- accidents en ville ou à la campagne ;
- accidents à une intersection ou hors intersection.

On donne les renseignements suivants :

- 85324 accidents ont eu lieu en ville ;
- 42 927 accidents ont eu lieu à une intersection ;
- 35 051 accidents ont eu lieu en ville et à une inter section.

1. Recopier et compléter le tableau suivant :

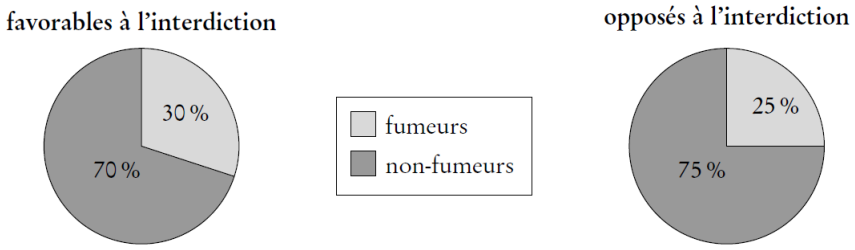
	En ville	À la campagne	Total
À une intersection			
Hors intersection			
Total			

33. Une étude sur les 1 000 employés de l'entreprise AU PRÉ travaillant dans quatre hypermarchés notés A, B, C, D a établi que :
— 400 employés travaillent dans le magasin C ;
— il y a respectivement 30 % et 20 % des employés de l'entreprise qui travaillent dans les magasins B et D ;
— il y a 45 % de femmes parmi les 1 000 employés de l'entreprise ;
— il y a 50 % de femmes parmi les employés du magasin A ;
— il y a 30 % de femmes parmi les employés du magasin B ;
— il y a 30 % de femmes parmi les employés du magasin C.
1. Compléter le tableau suivant :

Sexe / Magasin	Femmes	Hommes	Total
A			
B			
C			
D			
Total			

2. Quel est le pourcentage d'hommes parmi les employés du magasin D ?
3. Quel est le pourcentage d'employés du magasin D parmi les hommes travaillant dans les quartes hypermarchés ?

34. Voici un sondage concernant l'opinion des Français sur l'interdiction de la vente du tabac aux moins de 16 ans (sondage réalisé sur un échantillon représentatif de 500 personnes majeures).



1. Recopier et compléter le tableau suivant à l'aide des informations données par les diagrammes en secteurs circulaires. (Arrondir à l'unité la plus proche.)

Êtes-vous	favorable ?	opposé ?	sans opinion ?	total
fumeur ?			0	
non-fumeur ?				
total	316	160		500

2. a) Calculer la fréquence conditionnelle des personnes favorables par rapport aux fumeurs $f_{fum}(fav) =?$.
b) Calculer la fréquence conditionnelle des personnes favorables par rapport aux non-fumeurs $f_{non-fum}(fav) =?$.

35. On étudie la répartition en milliers de personnes, de la population active canadienne en avril 2014 selon le sexe (hommes / femmes) et le niveau d'emploi (Temps plein / Temps partiel) . La population active canadienne en 2014 compte 14000 milliers. On sait que la fréquence des hommes est de 0,56 et on connait les fréquences conditionnelles :

$$f_{hommes}(\text{plein temps}) = 0,30$$
$$f_{femmes}(\text{temps partiel}) = 0,25$$

En déduire la fréquence de la modalité "emploi en temps partiel".

Nuage de points. Ajustement linéaire.

Exemple 3.2.1. Sur un ensemble de 7 individus on a observé 2 caractères X et Y. On a obtenu le tableau suivant

Individu i	1	2	3	4	5	6	7
X	1	3	4	6	7	8	9
Y	2	4	5	5	6	7	8

a/ Représenter le nuage de points (x_i, y_i) dans un système orthonormé. Unités 2 cm.
b/ Calculer les coordonnées du point moyen $G(\bar{x}, \bar{y})$ de ce nuage de points, la variance de X, la variance de Y et la covariance de (X, Y).
c/ Tracer une droite à vue
On choisit 2 points du nuage, par exemple $A(1, 2)$ et $B(9, 8)$. Tracer la droite (AB) et déterminer l'équation réduite de cette droite. Cette droite passe-t-elle par G ?
d/ Droite de Mayer
On partage l'ensemble des points en 2 sous-ensembles de même effectif si possible.
$E_1 = \{(1, 2); (3, 4); (4, 5)\}$ et $E_2 = \{(6, 5); (7, 6); (8, 7); (9, 8)\}$. Pour chacun de ces sous-ensembles on calcule les coordonnées du point moyen. $G_1(\bar{x}_1, \bar{y}_1)$ point moyen de E_1 et $G_2(\bar{x}_2, \bar{y}_2)$ point moyen de E_2.
Dessiner la droite (G_1, G_2) et déterminer son équation réduite. Cette droite est appelée droite de Mayer. Cette droite passe-t-elle par le point moyen du nuage ?
e/ On appelle M_i le point de coordonnées (x_i, y_i) du tableau, M'_i le point d'abscisse x_i qui appartient à ta droite (AB) et M''_i le point d'abscisse x_i qui appartient à la droite (G_1, G_2).
Équation de (AB) : $y' = 0,75x + 1,25$.
Équation de (G_1, G_2) : $y'' = \frac{17}{29}x + \frac{61}{29}$
Compléter le tableau suivant :

x_i	1	3	4	6	7	8	9
y_i	2	4	5	5	6	7	8
y'_i							
y''_i							
$e'_i = y'_i - y_i$							
$e''_i = y''_i - y_i$							

f/ Déterminer la droite des moindres carrés de Y en X (droite de régression de Y en X) et la droite des moindres carrés de X en y (droite de régression de X en y).

Tracer les droites de moindres carrés dans le même système d'axes que le nuage de points.
g/ Calculer le coefficient de corrélation linéaire $r(x, y)$.

Exemple 3.2.4. On considère le tableau de données ci-dessous :

x	1	2	3	4	5	6	7	8	9	10
y	1.65	2.72	4.48	7.39	12.18	20.09	33.12	54.6	90.02	148.41

1) Représenter graphiquement les couples (x_i, y_i) $i = 1, \ldots 10$.
2) Effectuer la régression linéaire de Y par X à l'aide des résultats fournis ci-dessous.

Sommes	
des observations x :	55
des observations y :	374.66
des carrés x^2 :	385
des carrés y^2 :	34843.99
des produits xy :	3194.45

3) On prendra pour valeurs $a = 13.74$ et $b = -38.12$ dans la droite de régression $y = ax + b$. Calculer la valeur estimée de Y pour $x = 5$ et $x = 12$. Représenter la droite sur le graphique.

Exercices

36. On a administré un test de lecture à 12 enfants âgés de 7, 8 et 9 ans. Voici les résultats obtenus par ces sujets :

i	Variable X Âge	Variable Y Note au test
1	7	6
2	8	8
3	9	8
4	7	7
5	9	9
6	8	8
7	7	6
8	9	9
9	8	7
10	9	8
11	8	9
12	7	7

1. Représenter cette série statistique par un nuage de points.
2. Calculer la moyenne, l'écart–type et la variance de la variable X et de la variable Y
3. Déterminer l'équation d'une droite à vue.
4. Déterminer l'équation de la droite de Mayer
5. Déterminer l'équation de la droite de régression de Y en X.
6. Déterminer le coefficient de corrélation entre X et Y est le coefficient de détermination R^2.
7. Déterminer le résultat prédit au test pour un enfant âgé de 10 ans. Représenter la droite de régression sur le nuage de points.

8. Interpréter vos résultats à partir des mesures calculées et à partir du graphique que vous avez tracé.

37. Le tableau suivant donne la distance de freinage nécessaire à une automobile circulant sur une route humide pour s'arrêter.

Vitesse de l'automobile x_i en km/h	30	40	50	60	70	80	90	100	110	120
Distance de freinage d_i en mètres	18	26	40	58	76	98	120	148	180	212

a). Représenter le nuage de points (x_i, y_i) dans un repère orthogonal.
b) Trouver l'équation et tracer la droite D_Y des moindres carrées sur le graphique précédent.
c) En utilisant l'équation de la droite D_Y , déterminer une estimation de y si la vitesse de l'automobile était de 140 km/h.
d) À l'aide de la droite d'ajustement de la figure de l'introduction, estimer graphiquement la distance de freinage à 140 km/h.
e) Utiliser le modèle $z_i = \sqrt{y_i}$. Calculer la distribution (x_i, z_i) et représenter le nuage des points (x_i, z_i).
f) Trouver l'équation et tracer la droite D_Z des moindres carrées sur le graphique.
g) En utilisant l'équation de la droite D_Z , déterminer une estimation de y si la vitesse de l'automobile était de 140 km/h.

38. Les données dans le tableau qui suit concernent le geyser Old Faithful situé dans le parc national Yellowstone (Wyoming) aux États-Unis. Il est limité ici à 30 observations (parmi 272 à l'origine) qui indiquent la durée E des éruptions et le temps d'attente A entre les éruptions successives. Les temps sont mesurés en minutes.

Attente	79	54	74	62	85	55	88	85	51	85
Éruption	3.6	1.8	3.3	2.3	4.5	2.9	4.7	3.6	2.0	4.3
Attente	54	84	78	48	83	52	62	84	52	79
Éruption	1.8	3.9	4.2	1.8	4.7	2.2	1.8	4.8	1.6	4.2
Attente	51	48	78	69	74	83	55	76	78	79
Éruption	1.8	1.8	3.5	3.1	4.5	3.6	2.0	4.1	3.8	4.4

a) Calculer les moyennes et les variances des durées d'éruption et des temps d'attente.
b) Calculer la droite de régression $D_{E/A}$ des éruptions par rapport aux temps d'attente par la méthode des moindres carrés.
c) Donner une représentation graphique du diagramme de dispersion et de la droite de régression $D_{E/A}$.
d) Calculer les coefficients de corrélation et de détermination.

39. Les derniers recensements de la population de la ville de Carfain ont abouti aux données suivantes :

année	1993	1995	1999	2002	2004	2006	2009	2012
nombre d'années x_i depuis 1992	1	3	7	10	12	14	17	
population y_i ((en milliers d'habitants)	4,4	4,7	4,8	4,9	5,5	5,5	5,7	

1. Représenter le nuage des points associé à la série statistique $(x_i; y_i)$ dans le plan rapporté à un repère orthogonal.
2. Déterminer le coefficient de corrélation linéaire entre x et y. Que peut-on en déduire ? Un ajustement affine est-il indiqué dans cette situation ? Pourquoi ?
3. a) Donner une équation de la droite de régression de y en x (pour les coefficients on prendra les valeurs décimales arrondies à 10^{01} près). Tracer cette droite sur le schéma précédent.
b) Donner une prévision de la population de Carfain en 2012 par la méthode des moindres carrées.
c) Calculer le coefficient de détermination et commenter son importance.

40. A l'Université A une nouvelle méthode d'enseignement est introduite. Le tableau donne l'évolution du taux en % des notes parfaites à l'examen final à la fin des cours d'après la nouvelle méthode d'enseignement.

Année	2010	2011	2012	2013	2014
Rang de l'année X	1	2	3	4	5
Taux de notes en % Y	2	10	18	24	30

a) Représenter dans un repère le nuage des points
b) Existe-il une corrélation linéaire entre les variables X et Y ?
c) Déterminer l'équation de la droite des moindres carrées D_y (droite de y en x)
d) Tracer D_y
e) Donner une estimation pour la 6^{eme} année.

Ajustement exponentiel

Pour les phénomènes à croissance forte ou à décroissance rapide, il peut être pertinent d'approcher la forme générale du nuage par une fonction exponentielle de la forme $x \to e^{ax+b}$ ou $x \to \alpha e^x + \beta$. Pour déterminer les valeurs de a et b , ou de α et β, on effectue un ajustement affine (par exemple, par la méthode des moindres carrés) sur la série $(x_i; lny_i)$ ou la série $(e^{x_i}; y_i)$.

Exercices

41. Le tableau ci-dessous indique le salaire brut annuel, en euros, perçu par un cadre.

année	2005	2006	2007	2008	2009
rang x_i de l'année	1	2	3	4	5
salaire y_i (en €)	42 900	54 200	64 100	81 600	102 000
$z_i = ln(y_i)$	10,666	10,900	11,068	11,310	11,533

1. Représenter graphiquement la série $(x_i; y_i)$ dans un repère orthogonal et puis dans un repère semi logarithmique
a) Calculer le coefficient de corrélation linéaire de la série $(x_i; y_i)$
2. On pose $z_i = lny_i$.
a) Représenter graphiquement la série $(x_i; z_i)$ dans un repère orthogonal
b) Calculer le coefficient de corrélation linéaire de la série $(x_i; z_i)$
c) Déterminer une équation de la droite de régression de x en z par la méthode des moindres carrées
d) En déduire une relation entre y et x de la forme $y = \alpha e^{\beta x}$
e) En déduire une estimation du salaire des cadres en 2012.

Ajustement logarithmique À l'opposé des fonctions exponentielles, les fonctions logarithmes $x \to a \ln x + b$ ou $x \to \ln(\alpha x + \beta)$ sont tout à fait indiquées dans la modélisation des phénomènes à (dé)croissance lente. Pour cela, on effectue un ajustement affine sur la série $(\ln x_i; y_i)$ ou $(x_i; e^{y_i})$.

Exercices

42. Le tableau ci-dessous donne la production d'électricité d'origine nucléaire en France, exprimée en milliards de kWh, entre 1979 et 2004 (source : site web du Ministère de l'industrie).
Les rangs des années sont calculés par rapport à l'année 1975.

année	1979	1985	1990	1995	2000	2001	2002	2003	2004
rang x_i de l'année	4	10	15	20	25	26	27	28	29
production y_i	37,9	213,1	297,9	358,8	395,2	401,3	416,5	420,7	427,7

1. Représenter graphiquement la série $(x_i; y_i)$ dans un repère orthogonal.
2. On pose $z_i = lnx_i$. On s'intéresse à la série statistique $(z_i; y_i)$.
a) Calculer le coefficient de corrélation linéaire de la série $(z_i; y_i)$
b) Déterminer une équation de la droite de régression de z en y par la méthode des moindres carrées
c) En déduire une relation entre y et x de la forme $y = \alpha e^{\beta x}$
d) En déduire une estimation de la production en 2012.

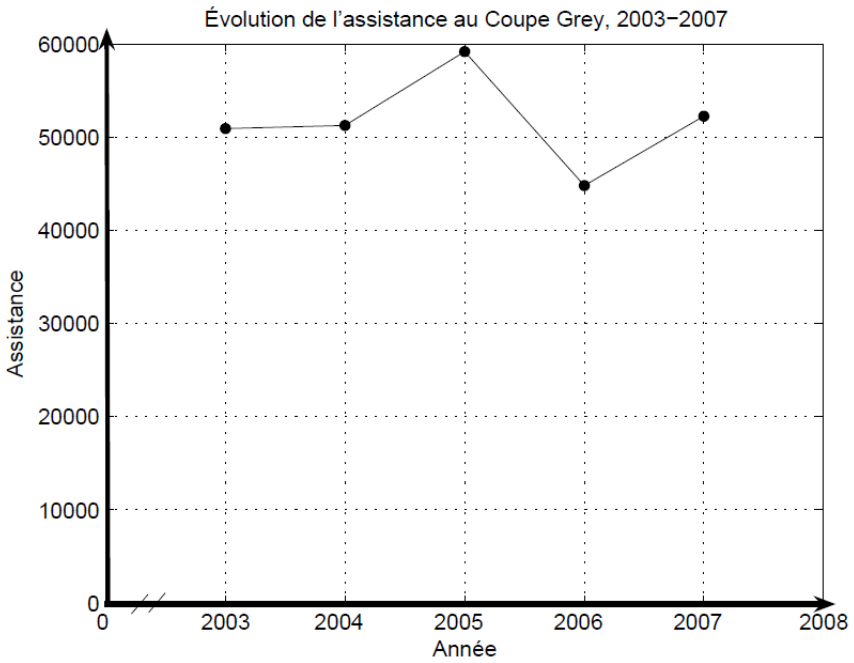
Feuille 7 : Séries chronologiques

Exemple Gagnant de la Coupe Grey, 2003-2007

Année	Équipe gagnante
2003	Eskimos
2004	Argonauts
2005	Eskimos
2006	Lions
2007	Roughriders

Exemple Évolution de l'assistance des spectateurs à la Coupe Grey, 2003-2007

Année	Assistance
2003	50 909
2004	51 242
2005	59 157
2006	44 786
2007	52 230



Exemple de graphique à la linge brisée.

Exemple 4.1.1 La responsable « Transports - Livraisons » de l'entreprise Yopmilk produisant des produits laitiers frais (yaourts, fromages frais,...) dispose pour les trois années précédentes des statistiques d'expédition suivantes, concernant les yaourts aromatisés :

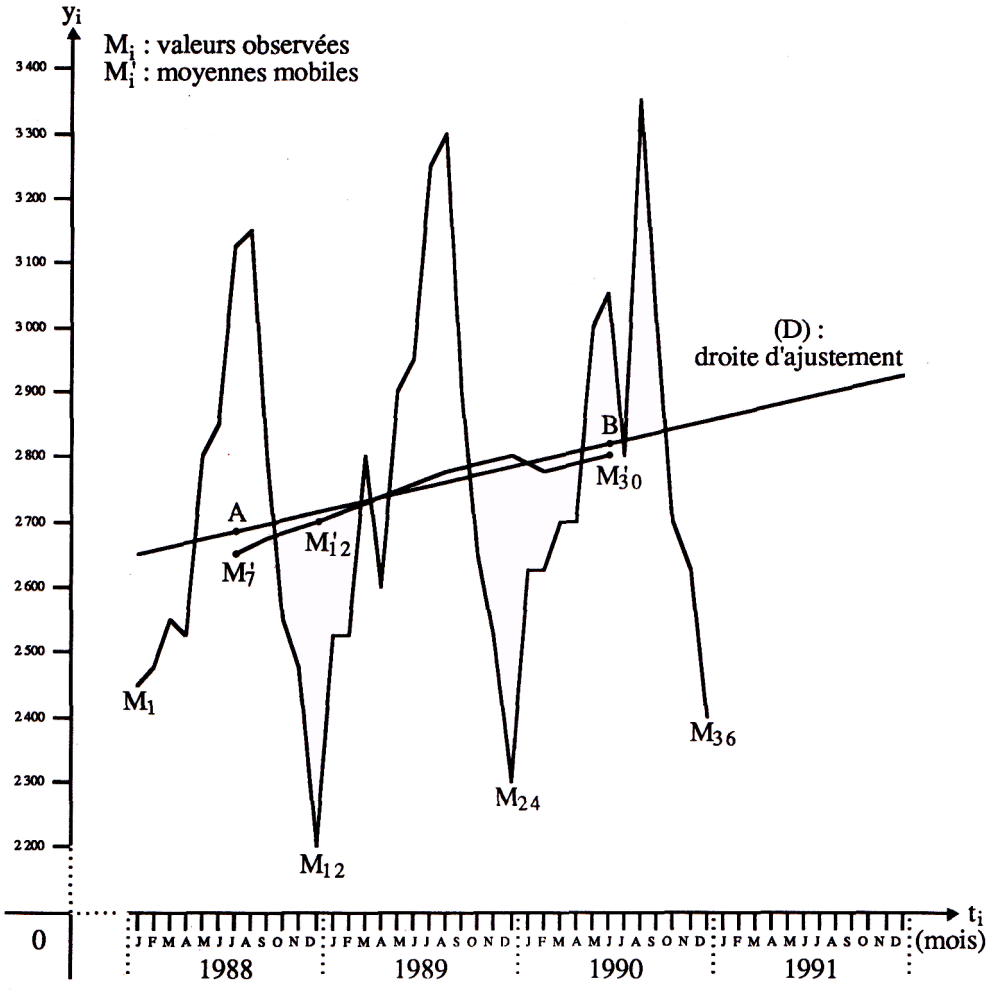
	1988	1989	1990
Janvier	2450	2525	2630
Février	2470	2530	2635
Mars	2550	2800	2700
Avril	2540	2600	2710
Mai	2800	2900	3000
Juin	2850	2950	3050
Juillet	3140	3250	2800
Août	3150	3300	3350
Septembre	2800	2900	3000
Octobre	2540	2660	2710
Novembre	2470	2530	2635
Décembre	2200	2300	2400

Afin d'améliorer la qualité des « transports » de la société, le responsable de ce service souhaite :
- connaître les caractéristiques de l'évolution des ventes au cours d'une année et ce, afin de maîtriser les phénomènes conjoncturels
- connaître la prévision des expéditions pour 1991.

Mois de 1991	$\hat{y}'_i$	s_i $(i = 1, \ldots, 12)$	$\hat{y}_i$
Janvier (37)	2853	s_1 =-210	2643
Février (38)	2859	s_2 =-200	2659
Mars (39)	2864	s_3 =-62	2802
Avril (40)	2 869,5	s_4 =-128,5	2741
Mai (41)	2875	s_5 = 155	3030
Juin (42)	2880	s_6 = 205	3085
Juillet (43)	2886	s_7 = 177	3063
Août (44)	2891	s_8 = 522	3413
Septembre (45)	2897	s_9 = 155	3052
Octobre (46)	2902	s_{10} =-108	2794
Novembre (47)	2908	s_{11} =-200	2708
Décembre (48)	2913	s_{12} =-445	2468

Table 6.1 - Estimations des expéditions

Expédition des yaourts aromatisés



Exemple 4.1.2 Étudions la série chronologique x_1 observée trimestriellement pendant 6 ans :

	1^{er} trimestre	2^{e} trimestre	3^{e} trimestre	4^{e} trimestre
Année 1	89,658	97,593	108,906	114,157
Année 2	96,205	99,399	112,763	119,185
Année 3	99,602	105,192	116,556	121,911
Année 4	103,272	109,644	121,208	126,508
Année 5	105,637	113,428	125,641	131,147
Année 6	111,118	117,215	129,776	132,880

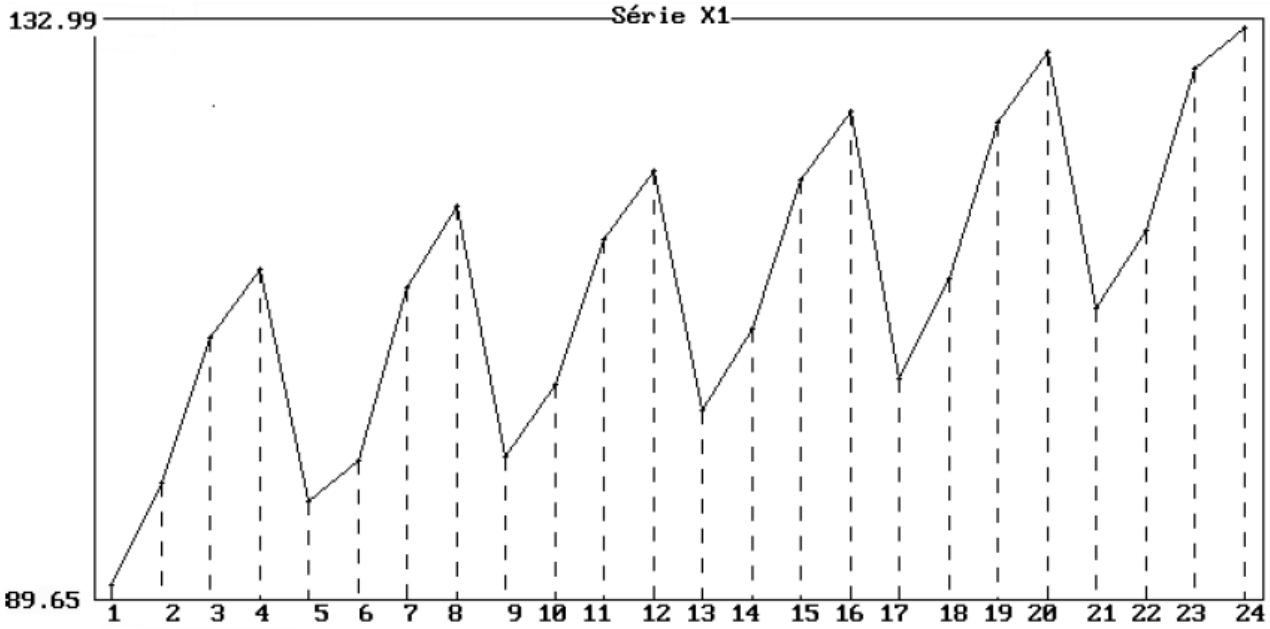
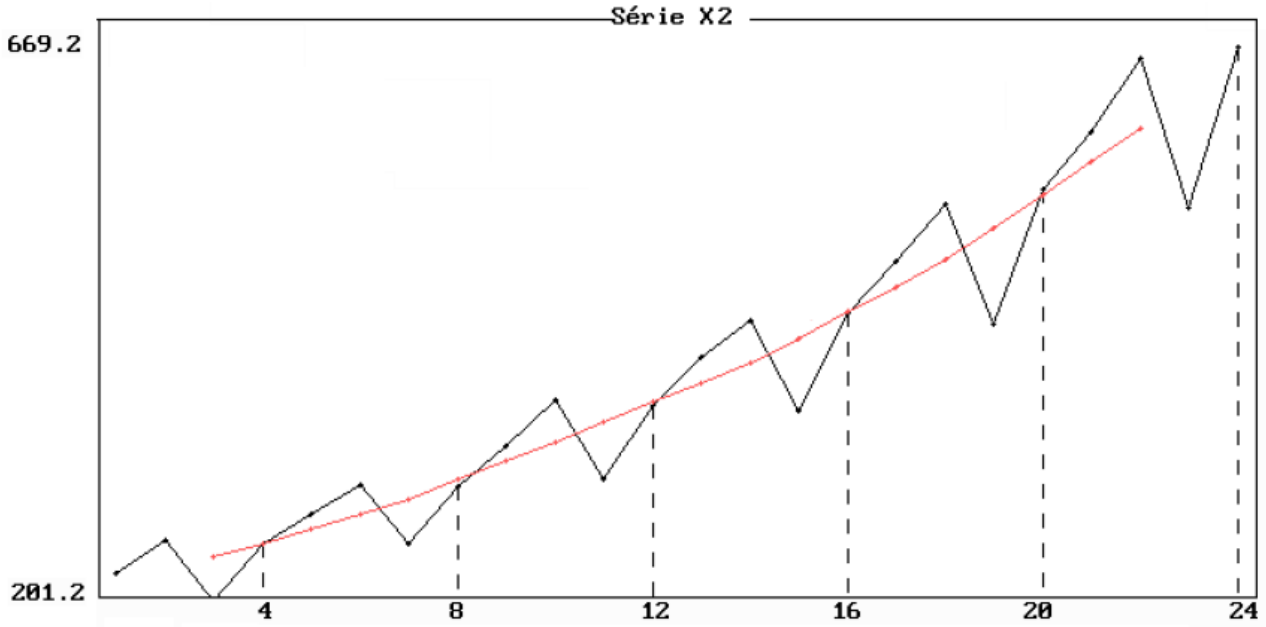


FIGURE 5.2 : Représentation graphique de la série - données observées trimestriellement pendant 6 ans

Exemple 4.1.3 On considère la série chronologique x_2 donnée ci-dessous :

	1^{er} trimestre	2^{e} trimestre	3^{e} trimestre	4^{e} trimestre
Année 1	224.3705	253.2811	201.2421	248.9411
Année 2	274.3802	300.1641	248.9038	298.4386
Année 3	331.9657	371.4032	303.4313	365.9029
Année 4	406.6326	437.9967	361.5774	444.8447
Année 5	488.4166	536.5268	435.5698	549.3614
Année 6	598.0016	659.2896	533.2156	669.2675

Tableau de la série chronologique x_2
(modèle multiplicatif, période $p = 4$)



Série x_2 et moyennes mobiles de longueur 4

Exercices

43. Bourse

On note le cours d'une action pendant 5 jours de suite :

jours	15 mars	16 mars	17 mars	18 mars	19 mars
cours	135	143	140	154	152

Déterminer l'équation de la tendance (droite) et prévoir le cours du jour suivant. Faire un graphique.

44. Bidules

Les ventes trimestrielles d'une entreprise (en milliers de bidules) ont suivi l'évolution suivante pendant 4 ans :

dates t	1	2	3	4	5	6	7	8
série Z_t	235	298	221	340	268	327	242	378
dates t	9	10	11	12	13	14	15	16
série Z_t	300	368	292	421	334	421	322	465

a) Représenter graphiquement l'évolution des ventes. Expliquer quelle méthode de prévision peut convenir. Rappeler le principe de la méthode.
b) Déterminer la tendance (ajustement linéaire).
c) Calculer les coefficients saisonniers.
d) Établir des prévisions sur 1 an.
e) Corriger la série des variations saisonnières.

N.B. Reporter tous les résultats de l'étude sur le même graphique et préciser la signification de chaque coefficient.

45. Un homme de lettres

L'impresario d'un chanteur célèbre a établi la statistique du nombre de lettres reçues chaque trimestre pour la période des 4 dernières années :

		années			
		1	2	3	4
trimestres	1	1000	1200	1500	1400
	2	1500	1800	2000	1800
	3	1400	1700	1800	1600
	4	800	1200	1300	1100

a) Faire un graphique de la série. Que remarque-t-on ?
b) Déterminer la tendance en utilisant la méthode des moyennes mobiles. Reporter les valeurs sur le graphique.
c) Calculer les coefficients saisonniers. Expliquer ce qu'ils signifient.
d) Etablir la série C.V.S. Reporter les valeurs sur le graphique. Commenter les résultats.

46. La série des indices trimestriels de ventes de marchandises d'une entreprise est fournie pour trois années par le tableau suivant

Année / Trimestre	Année 1	Année 2	Année 3
I	118,2	148,6	163,3
II	129,0	154,5	175,3
III	138,9	163,0	189,1
IV	157,1	184,0	217,9

a) Tracer la figure de la série.
b) Il y a-t-il un caractère saisonnier ?
c) En supposant que le modèle de composition des mouvements soit additif, calculer la série corrigée des variations saisonnières. La tendance sera estimée par le calcul des moyennes mobiles.

47. Correction des variations saisonnières du chômage

D. E. F. M données trimestrielles :

Année / Trimestre	1	2	3	4	5	6
I		2252	2482	2509	2702	2624
II		2184	2301	2386	2527	2437
III		2280	2335	2499	2579	2470
IV	2227	2522	2479	2677	2680	

Demandeurs d'emploi en fin de mois (D. E. F. M.)

a) Déterminer les coefficient saisonnier, en posant l'hypothèse d'une composition multiplicative des composants de la chronique, par la méthode des moyennes mobiles sur quatre trimestres.
b) Calculer alors la série CVS.
c) Représenter sur le même graphique la sérié brute et la sérié CVS.
d) Faire une prédiction pour III trimestre de l'année suivante (7).

48. L'exercice a pour objectif l'étude du chiffre d'affaires trimestriel (en milliers d'euros) de l'entreprise Peugeot présenté dans le tableau ci-dessous.

Année / Trimestre	2010	2011
I	13986	15414
II	14408	15721
III	12993	13450
IV	14674	15237

1. Représentez cette série chronologique sur un graphique. On adoptera un modèle additif pour modéliser cette série chronologique.

Date	Temps	$y(t)$	$y'(t)$	$T(t)$	C.V.S. $= y_i - S_i$	$S_i = S'_i - S$	S'_i
2010, I	1	13986	$\times$				
II	2	14408	$\times$				
III	3	12993	14193,8				
IV	4	14674	14536,4				
2011, I	5	15414	14757,6				
II	6	15721	??				
III	7	13450	$\times$				
IV	8	15237	$\times$				

2. Complétez la 3ème colonne du tableau précédent en calculant la moyenne mobile d'ordre 4 (ordre que vous justifierez) manquante. Représentez la série des moyennes mobiles sur le graphique.
3. Calculez la tendance T(t) par un modèle de régression linéaire (que vous justifierez en calculant le coefficient de corrélation) de la série y' en fonction du temps. Représentez la droite de régression sur le graphique.
4. Complétez les trois dernières colonnes du tableau visant à calculer les coefficients saisonniers.
5. Quel chiffre d'affaires l'entreprise Peugeot pouvait-elle espérer les trois premiers trimestres de 2012 ? Quels sont vos commentaires ?

49. Pendant deux semaines consécutives, on a observe le nombre de visiteurs d'un musée dont les jours de fermeture sont le samedi et le dimanche.

	Lundi	Mardi	Mercredi	Jeudi	Vendredi
Première semaine	7	5	35	5	6
Deuxième semaine	8	9	45	8	9

Considérons un modèle additif : $Y = T + S$.
1). Représentez graphiquement Y en fonction du temps. Pourquoi prend-on un modèle additif ?
2). Calculez les moyennes mobiles d'ordre 5, notées MM. Représentez graphiquement cette moyenne mobile. Pourquoi prend-on un ordre 5 ?
3). Effectuez un ajustement linéaire sur cette série chronologique Y. Justifier que le modèle est adéquate. Représentez graphiquement cet ajustement.
4). Déterminez les composantes saisonnières par la méthode de comparaison à la tendance.
5). Effectuer la désaisonnalisation (Calculer la série corrigée des variations saisonnières (c.v.s)). 6). Sur base du modèle additif et des résultats ci-dessus, donnez la prévision pour le lundi et le mardi de la 3-ième semaine.

Feuille 8 : Échantillonnage

Exemple 5.2.1 Une population est constituée de 5 clients d'un magasin. Le propriétaire du magasin s'intéresse à la somme moyenne (en €) laissée par chaque client dans le magasin lors d'une journée. On a obtenu les résultats suivants.

Etudiant	Temps d'étude (en heures)
A	7
B	3
C	6
D	10
E	4
Total	30

La moyenne de la population est $\mu = 30/5 = 6$.

Soit le propriétaire choisit un échantillon de taille 3. On peut se poser les questions les suivantes :
- Combien sont les différentes échantillons possibles qu'il peut choisir parmi les 5 clients observés ?
- Quelles sont les différentes valeurs possibles pour la moyenne de l'échantillon choisi ?
- Quelle relation existe-t-elle entre cette moyenne d'échantillon et la véritable moyenne 6 de la population ?

Exemple 5.2.2 Une machine effectue l'ensachage d'un produit.
On sait que les sacs ont un poids moyen de 250g avec un écart-type de 25g.
Quelles sont les caractéristiques de la moyenne des poids d'un échantillon de 100 sacs ?

Exemple 5.2.3 Dans une usine textile, on utilise une machine automatique pour couper des morceaux de tissu. Lorsque la machine est correctement ajustée, la longueur des morceaux de tissu est en moyenne de 90 cm avec un écart type de 0.60 cm.
Pour contrôler la longueur des morceaux de tissu, on tire dans la production d'une journée un échantillon aléatoire de 200 morceaux.
a) Si l'on suppose que la longueur X des morceaux de tissu suit une loi normale, calculer la probabilité que la moyenne de l'échantillon soit au plus égale à 89.90 cm, ceci dans 2 cas :
— production de la journée : 10 000 morceaux
— production de la journée : 2 000 morceaux.
b) Déterminer la même probabilité sans faire l'hypothèse que X soit distribuée normalement.
c) Si la moyenne observée sur cet échantillon est de 90.30 cm, celui-ci est-il représentatif de la population mère en prenant un risque de 5 % de se tromper ? (avec $N = 10\ 000$).

Exemple 5.2.4 Deux sociétés fabriquent des piles électriques d'un certain format.
Les piles de la société 1 ont une durée d'utilisation moyenne de 230 heures avec un écart type de 30 heures. Les piles de la société 2 ont une durée d'utilisation moyenne de 210 heures avec un écart type de 20 heures. Quelle est la probabilité que la durée d'utilisation moyenne d'un échantillon aléatoire simple de 100 piles de la société 1 soit d'au moins 30 heures de plus que la durée d'utilisation moyenne d'un échantillon aléatoire simple de 125 piles de la société 2 ?

Exemple 5.2.5 Le responsable d'une entreprise a accumulé depuis des années les résultats à un test d'aptitude à effectuer un certain travail. Il semble plausible de supposer que les résultats au test d'aptitude sont distribués suivant une loi normale de moyenne $\mu = 150$ et de variance

$\sigma^2 = 100$. On fait passer le test à 25 individus de l'entreprise. Quelle est la probabilité que la moyenne de l'échantillon soit entre 146 et 154 ?

Exemple 5.2.6 Le directeur financier d'une société sait par expérience que 12 % des factures émises ne sont pas réglées dans les 10 jours ouvrables suivant l'échéance. Il fait prélever un échantillon aléatoire de 500 factures.
Quelle est la probabilité qu'au moins 70 factures ne sont pas réglées dans le délais, sachant que l'ensemble des factures pouvant être étudiées est de plusieurs dizaines de milliers.

Exemple 5.2.7 Selon une étude sur le comportement du consommateur, 25% d'entre eux sont influencés par la marque, lors de l'achat d'un bien. Si on interroge 100 consommateurs pris au hasard, quelle est la probabilité pour qu'au moins 35 d'entre eux se déclarent influencés par la marque ?

Exercices

50. [5] La SGM souhaite mieux connaitre la répartition des impayés dans son portefeuille de clients. Sur l'ensemble des 20000 dossiers traités annuellement au service contentieux, un échantillon aléatoire de 30 dossiers a été prélevé aux fins d'étude, qui a permis d'obtenir un montants moyen empirique d'impayés de 2660,50 K€ et un écart-type empirique des impayés de 279,66 K€.
Quelle serait la probabilité pour que, sur l'ensemble des dossiers, le montant moyen d'impayés soit inférieur à 2300 K€ ?

51. Les statistiques des notes obtenues en mathématique au BAC - STI en France pour l'année 2013 sont : moyenne nationale $\mu = 10,44$ et écart-type $\sigma = 1,43$.
Une classe de BTS comporte 35 élèves en 2013/2014 issue d'un BAC STI en 2013.
Calculer la probabilité que la moyenne de cette classe soit supérieure à 10.

52. Après la correction d'une épreuve d'examen comportant 1600 candidats, on constate que les notes ont pour moyenne 12 et pour écart-type 3. On se propose de prélever un échantillon aléatoire exhaustif de 100 notes.
a) Quelle est la probabilité que la note moyenne d'un tel échantillon soit supérieure à 12,5 ?
b) Quelle est la probabilité d'avoir, dans un échantillon exhaustif de taille $n = 25$ une note moyenne supérieure à 12,5, si les notes des épreuves sont distribuées normalement ?
c) Quelle est la probabilité d'avoir une note moyenne supérieure à 12,5 dans un échantillon exhaustif de taille $n = 25$, si les notes des épreuves sont distribuées normalement et si l'épreuve d'examen comporte 400 candidats ?

53. [5] 96% des ménages français possèdent un réfrigérateur.
a) Quelle est la probabilité pour que, dans un échantillon de 1 200 ménages, la fréquence relative soit comprise entre 0,95 et 0,97.

54. [5] Le responsable des achats d'une grande surface décide de prendre un échantillon au hasard de 20 boites de haricots verts provenant d'une conserverie A, dont les poids nets égouttées sont distribués normalement de moyenne 478,10 et d'écart-type 17,507. Le responsable d'achats tire au hasard un échantillon de 22 boites de haricots verts provenant d'une conserverie B, dont les poids nets égouttées sont distribuées normalement de moyenne 478,10 et d'écart-type 13,306.

Quelle est la probabilité pour que la différence entre les poids nets égouttées moyens des haricots verts mis en boites dans les 2 conserveries sont entre -8,99 et 7,29 ?

55. Dans un restaurant d'entreprise, une employée sert du riz dans des assiettes individuelles. Le poids X (exprimé en gramme) de riz versé dans une assiette est une variable aléatoire qui suit une loi normale d'espérance mathématique 150g et d'écart-type 20g. Périodiquement, un contrôleur vient étudier un échantillon de 15 assiettes.
a) Quelle est la loi de probabilité et les paramètres de $\bar{X}$, le poids moyen de riz par assiette observé dans cet échantillon ? On prend 15 assiettes au hasard.
b) Quelle est la probabilité d'obtenir sur cet échantillon un poids moyen de riz inférieur à 140g ?
c) Quelle est la probabilité que le poids total de riz obtenu soit compris entre 2095g et 2405g ?
d) On prend une assiette au hasard, quelle est la probabilité qu'elle contienne moins de 140g de riz ?
e) Combien d'assiette faut-il prendre pour avoir 90 chances sur 100 d'obtenir un poids moyen de riz par assiette compris entre 144,5g et 155,5g ?

56. Les masses des colis reçus dans un grand magasin sont distribuées normalement avec une moyenne de 300 kg et un écart-type de 50 kg. Quelle est la probabilité qu'un groupe de 25 paquets reçus au hasard et chargés sur un monte-charge dépasse la limite de sécurité du monte-charge de 8 200 kg ?

57. Une usine fabrique des pièces circulaires dont le diamètre moyen μ doit être 5 cm avec un écart-type d'au plus $\sigma = 0,24$ cm. On réalise un test en utilisant un échantillon aléatoire de taille 36 prélevé au hasard dans la production ; le diamètre de chaque pièce de l'échantillon est mesuré.
La règle de décision du test est la suivante :

- si le diamètre moyen de l'échantillon est strictement inférieur à 4,92 cm ou strictement supérieur à 5,08 cm, le procédé de fabrication doit être arrêté, vérifié et réajusté à la valeur centrale requise, soit 5 cm ;
- si le diamètre moyen se situe à l'intérieur de l'intervalle [4,92 ; 5,08], on considère alors que le procédé fonctionne correctement et qu'il n'y a pas lieu d'intervenir.

On considère que le procédé de fabrication fonctionne selon la loi normale $\mathcal{N}(5; 0,24)$. Quelle est la probabilité d'arrêter la fabrication à la base des caractéristiques de l'échantillon ?

58. Une étude préalable a montré que dans une production en grand série, une machine fabrique des câbles électriques avec un pourcentage de câbles défectueux égal à 2%. Une entreprise commande 400 de ces câbles. Quelle est la probabilité pour que, dans cet envoi, on trouve plus de 3% de câbles défectueux ?

59. On suppose que l'age des élèves de terminale en France métropolitaine suit une loi normale de paramètre $\mu = 19$ ans et un écart-type $\sigma = 1,5$ ans. On considère un échantillon de taille $n = 50$. Quel loi suit l'age moyen des élèves dans l'échantillon ?

60. Une machine fabrique des disques pleins en grandes quantité. On suppose que la variable aléatoire X qui, à chaque disque tiré au hasard, associe son diamètre suit la loi normale

$\mathcal{N}(\mu;\sigma)$, où $\mu = 12,8$mm et $\sigma = 2,1$ mm.
Quelle loi suit la variable aléatoire $\bar{X}$, qui à chaque échantillon aléatoire non exhaustif de taille $n = 49$, associe la moyenne des diamètres des disques de cet échantillon ?

61. Une entreprise fournit des lots d'environ 10 000 pièces. Elle certifie que les lots ont une proportion de défectueux n'excédant pas 3 %.
Un client réceptionne chaque lot et effectue un test. Ce test conduit à la règle de décision suivante pour un échantillon aléatoire de 500 pièces issu d'un lot :

- le lot est accepté si l'échantillon contient au plus 21 pièces défectueuses,
- le lot est refusé si l'échantillon contient plus de 21 pièces défectueuses.

1) Si la proportion de défectueux du lot est 3 %, déterminer la probabilité que le lot testé soit refusé.
2) Quelle est la probabilité que le client accepte un lot dont la proportion de défectueux est 6 % ?

Exemples

Organisation d'une série statistique univariée. D.O.1

Organisation d'une série statistique univariée. D.G.1

Organisation d'une série statistique univariée. Exemple 2.2.1

Boite à moustaches. Exercice 29

Ajustement linéaire. Exercice 40

Séries chronologiques. Exercice 48

Séries chronologiques. Exercice 49

Echantillonnage. Exercice 61

Organisation d'une série statistique univariée. D.O.1

Exemple 2.1.2 [Notes] Relève des notes d'un groupe de 30 étudiants à l'examen en Statistique. On observe la variable aléatoire X ="relève des notes d'un étudiant". On a les résultats les suivants :

$$3; 2; 4; 6; 5; 2; 3; 6; 4; 4; 2; 3; 3; 4; 4$$
$$5; 6; 4; 4; 3; 3; 3; 3; 4; 4; 4; 4; 5; 5; 5$$

Évidemment les notes sont égales aux nombres 2,3,4,5 et 6. Le nombre des modalités du caractère est 5.
La série ordonnée est :

$$2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 34; 4; 4; 4$$
$$4; 4; 4; 4; 4; 4; 45; 5; 5; 5; 5; 6; 6; 6$$

Note x_i	Effectif n_i	Effectif cumulé $N_i = \sum_{j=1}^{i} n_j$	Fréquence $f_i = n_i/n$	Fréquence cumulée $g_i = \sum_{j=1}^{i} f_j$
2	3	3	3/30 =0,1	3/30=0,1
3	8	11	8/30=0,2666	11/30=0,3666
4	11	22	11/30=0,3666	22/30=0,7333
5	5	27	5/30=0,1666	27/30=0,9
6	3	30	3/30=0,1	1
Total	30		1	

Distribution de fréquences observées

Diagramme différentiel en bâtons

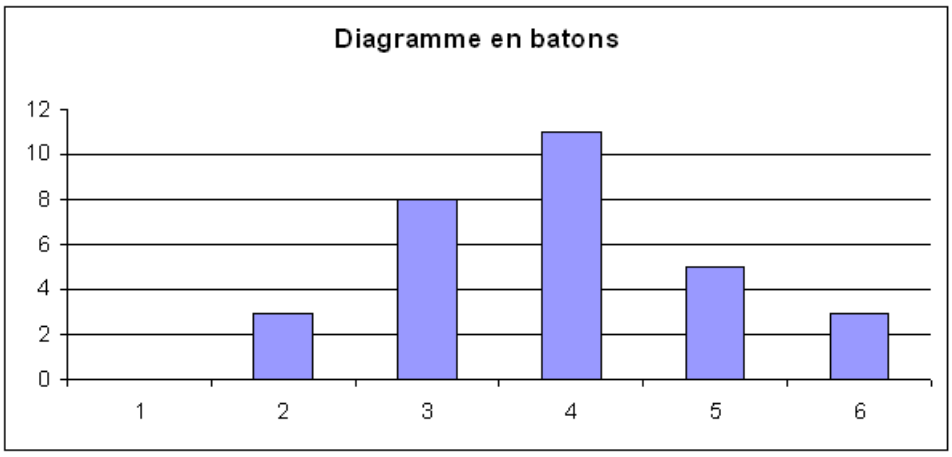


Diagramme cumulatif

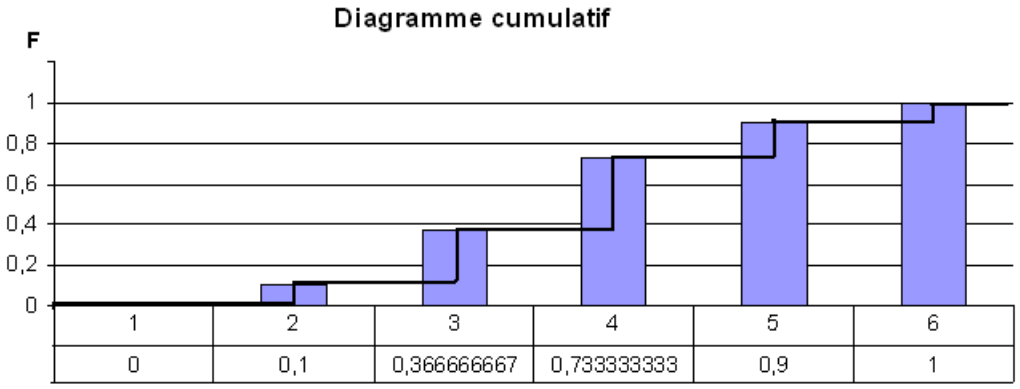
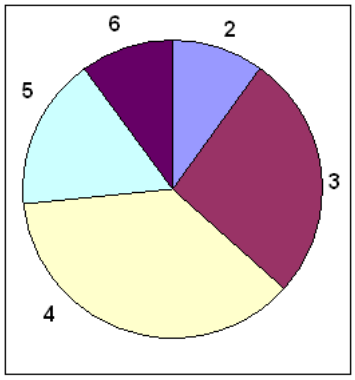


Diagramme circulaire



Mode Le mode de la série est la modalité 4. (La modalité 4 a le plus grand effectif 11).

Médiane $x_{1/2}$ $n = 30,\ \nexists N_j = \frac{30}{2} = 15 \Longrightarrow N_2 = 11 < \frac{30}{2} = 15 < N_3 = 22 \Longrightarrow x_{1/2} = x_{(3)} = 4$.

Moyenne

$$\bar{x} = \frac{2.3 + 3.8 + 4.11 + 5.5 + 6.3}{3 + 8 + 11 + 5 + 3} = \frac{117}{30} \approx 3,9.$$

Si dans le tableau des notes des étudiants on ajoute une colonne des produits $x_i n_i$, on obtient le tableau :

x_i	n_i	$x_i n_i$
2	3	6
3	8	24
4	11	44
5	5	25
6	3	18
	$\sum_{i=1}^{5} n_i = 30$	$\sum_{i=1}^{5} x_i n_i = 117$

et on obtient $\bar{x} = \frac{117}{30} \approx 3,9$.

Les quantilles Obtenir les quartiles Q_1 et Q_3 de la série statistique de l'exemple.

Note	Effectif cumulée N_i
2	3
3	11
4	22
5	27
6	30

$$
\begin{aligned}
Q_1 = Q_{1/4} = x_{1/4}: \quad & n = 30; \;\; np = 30\frac{1}{4} = 7,5 \Longrightarrow \bar{\exists} N_i = np = 7,5 \\
& \text{comme } N_1 = 3 < np = 7,5 < N_2 = 11 \\
& \longrightarrow x_{1/4} = Q_1 = x_{(2)} = 3. \\
Q_3 = Q_{3/4} = x_{3/4} \quad & n = 30; \;\; np = 30\frac{3}{4} = 22,5 \Longrightarrow \bar{\exists} N_i = np = 22,5 \\
& \text{comme } N_3 = 22 < np = 22,5 < N_4 = 27 \\
& \longrightarrow x_{3/4} = Q_3 = x_{(4)} = 5.
\end{aligned}
$$

paramètres de dispersion

Etendue $e = 6 - 2 = 4$.

Intervale interdecile Obtenir l'intervalle interdécile $[x_{1/10}, x_{9/10}]$ et l'écart interdécile E_D : Du tableau de la distribution

Note	Fréquence cumulée $\sum_{j=1}^{i} f_j$
2	0,1
3	0,3666
4	0,7333
5	0,9
6	1

Le premier décile on obtient immédiatement : $x_{1/10} = 2$
Le neuvième décile est $x_{9/10} = 5$.

L'intervalle $[x_{1/10}, x_{9/10}] = [2; 5]$ est l'intervalle cherché. L'intervalle interdécile comporte 80% des observations de la série statistique.

Lécart interdécile E_D est $E_D = 3$.

Ecart moyen Rappelons qu'on a trouvé $\bar{x} = 3,9$.

Note	Effectif n_i	Ecarts en valeurs abbsolues $x_i - \bar{x}$	Produit $n_i.\|x_i - \bar{x}\|$
2	3	1,9	5,7
3	8	0,9	7,2
4	11	0,1	1,1
5	5	1,1	5,5
6	3	2,1	6,3
Total	30		25,8

Pour l'écart moyen on obtient :

$$EM = \frac{1}{n}\sum_i n_i|x_i - \bar{x}| = \frac{25,8}{30} = 0,86.$$

Variance

Exemple On construit le tableau que voici :

Note	Effectif n_i	Ecarts en valeurs abbsolues $\|x_i - \bar{x}\|$	(Ecarts)^2 $(x_i - \bar{x})^2$	Produit $n_i.(x_i - \bar{x})^2$
2	3	1,9	3,61	10,83
3	8	0,9	0,81	6,48
4	11	0,1	0,01	0,11
5	5	1,1	1,21	6,05
6	3	2,1	4,41	13,23
Total		30		36,7

$$s^2 = \frac{36,7}{30} = 1,22333.$$

En utilisant les formules simplifiées on obtient :

Notes x_i	$Notes^2$ ${x_i}^2$	Produit $n_i.x_i^2$
2	4	12
3	9	75
4	16	176
5	25	125
6	36	108
Total		493

On obtient

$$s^2 = \frac{1}{30}493 - 3,9^2 = 1,22333.$$

Ecart-type

$$s = \sqrt{s^2} = 1,106044$$

Coefficient de variation

$$CV = \frac{s}{\bar{x}} * 100 = \frac{1,106}{3,9^2} * 100 = 28,36\% > 15\%$$

nonhomogène distribution

Moments centré

$$\begin{aligned} \mu_3 &= \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^3 = 3,05 \\ \mu_4 &= \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^4 = 3,6675 \end{aligned}$$

Coefficient d'asymétrie

$$g_1 = \frac{\mu_3}{s^3} = 0,198069 > 0$$

dissymétrie à gauche

Coefficient d'aplatissement

$$g_2 = \frac{\mu_4}{s^4} - 3 = 3,6679 > 0$$

plus pointue - concentration forte autour de $\bar{x}$.

Organisation d'une série statistique univariée. D.G.1

Exemple 2.1.3 [Bovins] Nombre de bovins dans les fermes privées dans une région donnée. Taille de la population $N = 60$. Série observée :

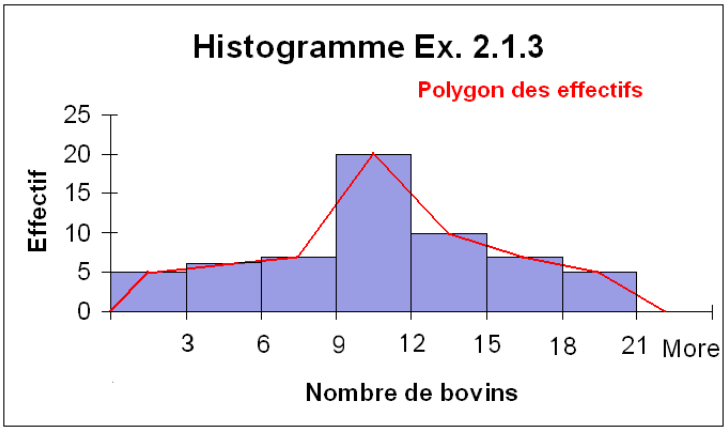
$$4; 2; 3; 5; 0; 1; 6; 1; 5; 4; 7; 10; 13; 16; 19; 8; 4; 11; 14; 17; 2; 11$$
$$9; 8; 12; 15; 18; 10; 13; 18; 7; 11; 14; 17; 21; 8; 9; 11; 10; 16; 8; 10$$
$$15; 17; 19; 11; 9; 13; 11; 11; 12; 9; 10; 11; 10; 12; 10; 11; 12; 14$$

On forme la série ordonnée

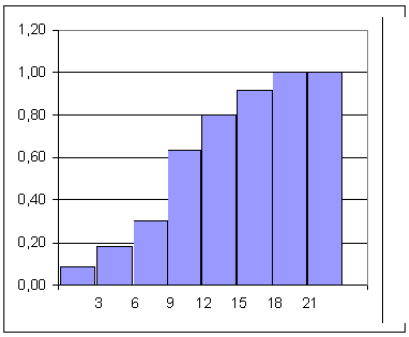
$$0; 1; 1; 2; 2; 3; 4; 4; 4; 5; 5; 6; 7; 7; 8; 8; 8; 8; 9; 9; 9; 9$$
$$10; 10; 10; 10; 10; 10; 10; 11; 11; 11; 11; 11; 11; 11; 11; 11; 12; 12; 12; 12$$
$$13; 13; 13; 14; 14; 14; 15; 15; 16; 16; 17; 17; 17; 18; 18; 19; 19; 21$$

Nombre de bovins classe $[a; b)$	Centre de la classe $x_i^* = (a+b)/2$	Effectif n_i	Effectif cumulé $N_i = \sum_{j=1}^{i} n_j$	fréquence $f_i = \frac{n_i}{n}$	Fréquence cumulée $\frac{1}{n}\sum_{j=1}^{i} n_j$
$[0; 3)$	1,5	5	5	5/60	5/60
$[3; 6)$	4,5	6	11	6/60	11/60
$[6; 9)$	7,5	7	18	7/60	18/60
$[9; 12)$	10,5	20	38	20/60	38/60
$[12; 15)$	13,5	10	48	10/60	48/60
$[15; 18)$	16,5	7	55	7/60	55/60
$[18; 21]$	19,5	5	60	5/60	1
		Total 60		1	

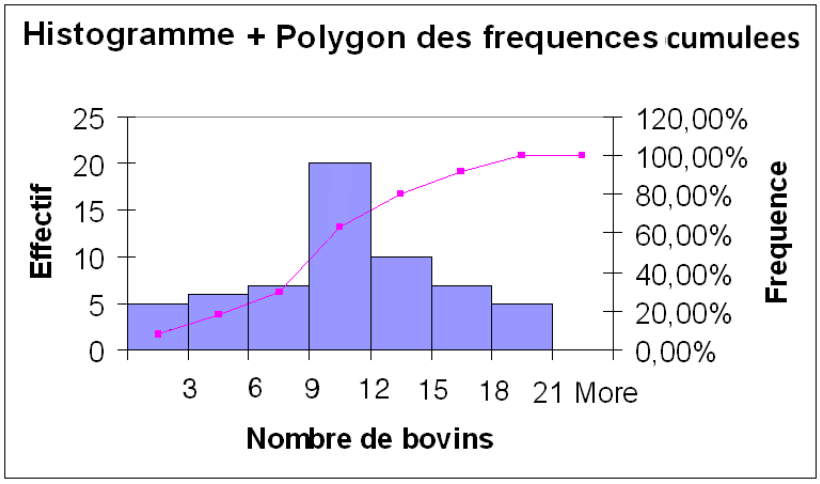
Histogramme. Le polygone des effectifs est donné en rouge :

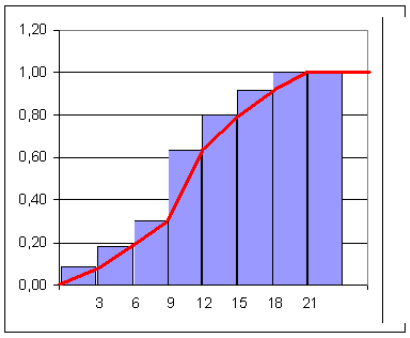


Histogramme des fréquences cumulées croissantes



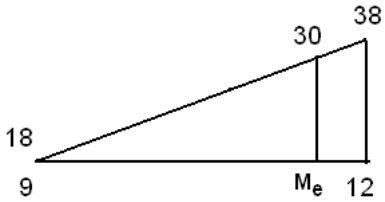
Courbe cumulative ou polygone des fréquences cumulées





Mode La classe modale de la série est la classe $[9;12)$ d'effectif maximal $n = 20$.

Médiane $n = 60$, $\frac{n}{2} = \frac{60}{2} = 30$, $\bar{\exists} N_j = \frac{n}{2} = 30$, $x_{1/2} \in [9;12[$.



D'après la règle des triangles semblables on peut écrire :

$$\frac{M_e - 9}{12 - 9} = \frac{30 - 18}{38 - 18}, \quad M_e = 9 + 3\frac{12}{20} = 9 + 1,8 = 10,8.$$

Moyenne Le tableau pour calculer la moyenne arithmétique est le suivant :

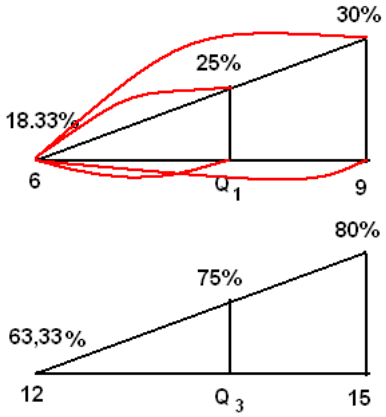
Classes $x_i; x_{i+1}$	Effectifs n_i	Centres x_i^*	Produits $x_i^* n_i$
moins de 3	5	1,5	7,5
3 ;6	6	4,5	27,0
6 ;9	7	7,5	52,5
9 ;12	20	10,5	210,0
12 ;15	10	13,5	135,0
15 ;18	7	16,5	115,5
18 ;21	5	19,5	97,5
	$\sum_{i=1}^{k} n_i = 60$		$\sum_{i=1}^{k} x_i^* n_i = 645,0$

On obtient pour la moyenne aruthmétique

$$\bar{x} = \frac{645}{60} \approx 10,75 \text{ c.a.d. à peu près 11 bovins.}$$

Les quantilles Obtenir les quartiles Q_1 et Q_3 de la série statistique de l'exemple 2.1.3.

Nombre de bovins classe $[a : b]$	Fréquence cumulée $\sum_{j=1}^{i} f_j$	Fréquence cumulée en pourcentages
$[0; 3[$	0,08	8%
$[3; 6[$	0,1833	18,33%
$[6; 9[$	0,3	30%
$[9; 12[$	0,6333	63,33%
$[12; 15[$	0,8	80 %
$[15; 18[$	0,916	91,6 %
$[18; 21[$	1	100 %



$Q_1 = Q_{1/4} = Q_{25}$ est dans l'intervalle $[6; 9)$. D'après la règle des triangles semblables on a :

$$\frac{Q_1 - 6}{9 - 6} = \frac{25 - 18,33}{30 - 18,33}; \quad Q_1 = 6 + 3\frac{6,67}{11,67}; \quad Q_1 = 7,71.$$

$Q_3 = Q_{3/4} = Q_{75}$ est dans l'intervalle $[12; 15)$. D'après la règle des triangles semblables on a :

$$\frac{Q_3 - 12}{15 - 12} = \frac{75 - 63,33}{80 - 63,33}; \quad Q_3 = 12 + 3\frac{11,67}{16,67}; \quad Q_3 = 14,10.$$

paramètres de dispersion

Etendue $e = 21 - 0 = 21$.

Intervale interdecile Obtenir l'intervalle interdécile $[x_{1/10}, x_{9/10}]$ et l'écart interdécile E_D pour la série statistique de l'exemple 2.1.3.
Du tableau de la distribution

Nombre de bovins classe $[a : b]$	Fréquence cumulée $\sum_{j=1}^{i} f_j$
$[0; 3[$	0,08
$[3; 6[$	0,1833
$[6; 9[$	0,3
$[9; 12[$	0,6333
$[12; 15[$	0,8
$[15; 18[$	0,916
$[18; 21[$	1

pour le calcul du premier décile on doit faire une interpolation entre les valeurs $(3; 0,08)$ et $(6; 0,18)$.
On trouve :

$$\frac{x_{1/10} - 3}{6 - 3} = \frac{1/10 - 0,08}{0,18 - 0,08}$$

ou encore :

$$\frac{x_{1/10} - 3}{3} = \frac{0,02}{0,1}; \quad x_{1/10} = 3 + 3 * \frac{0,02}{0,1} = 3,6.$$

Pour le calcul du neuvième on procède par interpolation linéaire entre les valeurs $(15; 0,8)$ et $(18; 0,916)$.
On trouve :

$$\frac{x_{9/10} - 15}{18 - 15} = \frac{9/10 - 0,8}{0,916 - 0,8}$$

ou encore :

$$\frac{x_{9/10} - 15}{3} = \frac{0,1}{0,116}; \quad x_{9/10} = 15 + 3 * \frac{0,1}{0,116} = 17,59.$$

L'intervalle $[x_{1/10}, x_{9/10}] = [3,6; 17,59]$ est l'intervalle cherché. L'intervalle interdécile comporte 80% des observations de la série statistique.

Lécart interdécile E_D est $E_D = 13,99$.

Ecart moyen Rappelons qu'on a trouvé $\bar{x} = 10,75$.

Classe $[a - b[$	Centre de classe $x_i^* = (a + b)/2$	Effectif n_i	Ecarts en valeurs abbsolues $x_i^* - \bar{x}$	Produit $n_i.\|x_i^* - \bar{x}\|$
$[0 - 3[$	1,5	5	9,25	46,25
$[3 - 6[$	4,5	6	6,25	37,5
$[6 - 9[$	7,5	7	3,25	22,75
$[9 - 12[$	10,5	20	0,25	5
$[12 - 15[$	13,5	10	2,75	27,5
$[15 - 18[$	16,5	7	5,75	40,25
$[18 - 21[$	19,5	5	8,75	43,75
Total		60		223

Pour l'écart moyen on obtient :

$$EM = \frac{223}{60} = 3,716667.$$

Variance on construit le tableau que voici :

Classe $[a-b[$	Centre x_i^*	Effectif n_i	Ecarts $x_i^* - \bar{x}$	(Ecarts)^2 $(x_i^* - \bar{x})^2$	Produit $n_i.(x_i^* - \bar{x})^2$
$[0-3[$	1,5	5	9,25	85,56	427,81
$[3-6[$	4,5	6	6,25	39,06	234.375
$[6-9[$	7,5	7	3,25	10,56	73,9375
$[9-12[$	10,5	20	0,25	0,0625	1,25
$[12-15[$	13,5	10	2,75	7,56	75,625
$[15-18[$	16,5	7	5,75	33,06	231,44
$[18-21[$	19,5	5	8,75	76,56	382,81
Total		60			1427,25

$$s^2 = \frac{1427,25}{60} = 23,79.$$

Exemple En utilisant les formules simplifiées pour l'exemple 2.1.3 on obtient :

Classe $[a-b[$	Centre x_i^*	(Centres)^2 x_i^{*2}	Produit $n_i.x_i^{*2}$
$[0-3[$	1,5	2,25	11,25
$[3-6[$	4,5	20,25	121,5
$[6-9[$	7,5	56,25	393,75
$[9-12[$	10,5	110,25	2205
$[12-15[$	13,5	182,25	1822,5
$[15-18[$	16,5	272,25	1905,75
$[18-21[$	19,5	380,25	1901,25
Total			8361

On obtient

$$s^2 = \frac{1}{60}8361 - 10,75^2 = 23,7875.$$

Ecart-type

$$s = \sqrt{s^2} = 4,877$$

Coefficient de variation

$$CV = \frac{s}{\bar{x}} * 100 = \frac{4,877}{10,75^2} * 100 = 45,3697\% > 15\%$$

nonhomogène distribution

Moments centré

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^3 = \frac{1}{60}(-774,379) = -12,90625$$

$$\mu_4 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^4 = \frac{84073,92}{60} = 1401,23$$

Coefficient d'asymétrie

$$g_1 = \frac{\mu_3}{s^3} = -0,111 < 0$$

dissymétrie à droite.

Coefficient d'aplatissement

$$g_2 = \frac{\mu_4}{s^4} - 3 = -0,52 < 0$$

plus aplatie - concentration faible autour de $\bar{x}$.

Organisation d'une série statistique univariée. Exemple 2.2.1

Considérons une étude de prix d'un même article en fonction de la marque qui le commercialise. Ce genre d'étude est fréquent.

Tableau des observations :

Intervalle de prix	Nombres de marques
$[165-170[$	6
$[170-175[$	10
$[175-180[$	5
$[180-185[$	4
$[185-190[$	3
$[190-195[$	2

Tableau de la distribution :

Classe $[a-b[$	Centre de classe $(a+b)/2$	Effectif n_i	Effectif cumulé N_i	Fréquence f_i
$[165-170[$	167,5	6	6	6/30
$[170-175[$	172,5	10	16	10/30
$[175-180[$	177,5	5	21	5/30
$[180-185[$	182,5	4	25	4/30
$[158-190[$	187,5	3	28	3/30
$[190-195[$	192,5	2	30	2/30

La classe modale de la série est la classe $[170-175[$. (La classe $[170-175)$ a le plus grand effectif 10).

On peut identifier le mode comme la valeur médiane de la classe de fréquence maximale ou bien effectuer une interpolation linaire pour obtenir la valeur exacte du mode comme suit :

$$M_o = x_m + \frac{i\Delta_i}{\Delta_s + \Delta_i} \quad /\text{MODE}(\text{x}_\text{i})/$$

avec
$x_m = a$: limite inférieure de la classe d'effectif maximal
i : intervalle de classe $(x_{m+1}{-}x_m)$
Δ_i : Ecart d'effectif entre la classe modale et la classe inférieure la plus proche
Δ_s : Ecart d'effectif entre la classe modale et la classe supérieure la plus proche.

Mode

- Valeur approchée :

La classe de fréquence maximale est $[170, 175)$ avec $n_i = 10$ d'où $Mo = \frac{170+175}{2} = \frac{345}{2} = 172,5$.

- Valeur exacte :

$$M_o = 170 + \frac{5*4}{5+4} = 170 + \frac{20}{9} = 170 + 2,222 = 172,222 \quad \text{d'où} \quad M_o = 172,222$$

avec $x_m = a = 170$, $\Delta_i = 10-6 = 4$, $\Delta_s = 10-5 = 5$ et $i = 5$.

Médiane

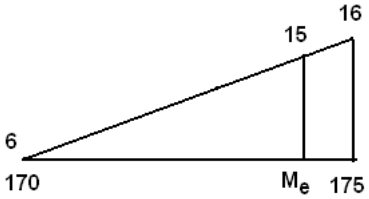
$n = 30$, la $n/2 = 30/2 = 15^{\text{ème}}$ valeur se situe dans la classe $[170, 175)$ qui contient les individus de 7 à 16. D'ici avec $x_m = 170$, $x_{m+1} = 175$,
$x_{m+1} - x_m = 175 - 170 = 5$, $N_i = 6$, $n_i = 10$

$$M_e = 170 + (175-170)\left(\frac{15-6}{10}\right) = 170 + 5\frac{9}{10} = 170 + 4,5 = 174,5$$

D'où la Médiane $M_e = 174,5$.

Classe $[a-b[$	Effectif cumulé N_i
$[165-170[$	6
$[175-170[$	16
$[175-180[$	21
$[180-185[$	25
$[158-190[$	28
$[190-195[$	30

On peut utiliser la règle des triangles semblables. Du polygone des effectifs pour la classe modale $[170, 175[$ on a :



D'après la règle des triangles semblables on peut écrire :

$$\frac{M_e - 170}{175 - 170} = \frac{15 - 6}{16 - 6}, \quad M_e = 170 + 5\frac{9}{10} = 170 + 4,5 = 174,5.$$

Moyenne

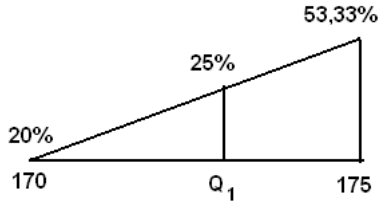
$$\bar{\mathrm{x}} = \frac{5295}{30} = 176,5.$$

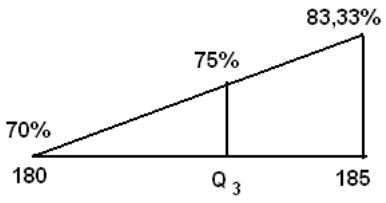
Classe $[a-b[$	Centre de classe $(a+b)/2$	Effectif n_i	Produit $x_i^* n_i$
$[165-170[$	167,5	6	1005
$[170-175[$	172,5	10	1725
$[175-180[$	177,5	5	887,5
$[180-185[$	182,5	4	730
$[158-190[$	187,5	3	562,5
$[190-195[$	192,5	2	385
Total		30	5295

Quantiles

Obtenir les quartiles Q_1 et Q_3

Classe classe $[a;b)$	Fréquence cumulée $\sum_{j=1}^{i} f_j$	Fréquence cumulée en pourcentages
$[165-170[$	0,2	20%
$[170-175[$	0,5333	53,33%
$[175-180[$	0,7	70%
$[180-185[$	0,8333	83,33%
$[158-190[$	0,9333	93,33 %
$[190-195[$	1	100%





$Q_1 = Q_{1/4} = Q_{25}$ est dans l'intervalle $[170; 175)$. D'après la règle des triangles semblables on a :

$$\frac{Q_1 - 170}{175 - 170} = \frac{25 - 20}{53,33 - 20}; \quad Q_1 = 170 + 5\frac{5}{38,33}; \quad Q_1 = 170,75.$$

$Q_3 = Q_{3/4} = Q_{75}$ est dans l'intervalle $[180; 185)$. D'après la règle des triangles semblables on a :

$$\frac{Q_3 - 180}{185 - 180} = \frac{75 - 70}{83,33 - 70}; \quad Q_3 = 180 + 5\frac{5}{13,33}; \quad Q_3 = 181,8755.$$

Etendue
On trouve immédiatement :

$$e = 195 - 165 = 30.$$

Intervale interdécile
Obtenir l'intervalle interdécile $[x_{1/10}, x_{9/10}]$:
Du tableau de la distribution

Classe $[a-b[$	Centre de classe $(a+b)/2$	Effectif n_i	Eff. cumul. N_i	Fréquence f_i	Fréq. cumul. g_i
$[165-170[$	167,5	6	6	$6/30 = 0{,}2$	0,2
$[170-175[$	172,5	10	16	$10/30 = 0{,}33$	0,5333
$[175-180[$	177,5	5	21	$5/30 = 0{,}167$	0,7
$[180-185[$	182,5	4	25	$4/30 = 0{,}133$	0,8333
$[185-190[$	187,5	3	28	$3/30 = 0{,}1$	0,9333
$[190-195[$	192,5	2	30	$2/30 = 0{,}067$	1

il est relativement aisé de calculer le premier décile, qui doit se situer au centre de la première classe.

$$\text{On a donc :} \quad x_{1/10} = 167,5.$$

Pour le calcul du neuvième on procède par interpolation linéaire entre les valeurs $(185; 0,83333)$ et $(190; 0,9333)$.
On trouve :

$$\frac{x_{9/10} - 185}{190 - 185} = \frac{9/10 - 0,8333}{0,9333 - 0,8333}$$

ou encore :

$$\frac{x_{9/10} - 185}{5} = \frac{0,0667}{0,1}; \quad x_{9/10} = 185 + 5 * \frac{0,0667}{0,1} = 188,33.$$

L'intervalle $[x_{1/10}, x_{9/10}] = [167,5; 188,33]$ est l'intervalle cherché. L'intervalle interdécile comporte 80% des observations de la série statistique.

Ecart interdécile E_D

$$E_D = x_{9/10} - x_{1/10} = 20,83.$$

Ecart moyen
Rappelons qu'on a trouvé $\bar{x} = 176,5$.

Classe $[a-b[$	Centre de classe $x_i^* = (a+b)/2$	Effectif n_i	Ecarts en valeurs abbsolues $x_i^* - \bar{x}$	Produit $n_i.\lvert x_i^* - \bar{x}\rvert$
$[165-170[$	167,5	6	9	54
$[170-175[$	172,5	10	4	40
$[175-180[$	177,5	5	1	5
$[180-185[$	182,5	4	6	24
$[158-190[$	187,5	3	11	33
$[190-195[$	192,5	2	16	32
Total		30		188

Pour l'écart moyen on obtient :

$$EM = \frac{188}{30} = 6,267.$$

Variance

Exemple Dans le cas de l'exemple 2.2.1 on construit le tableau que voici :

Classe $[a-b[$	Centre x_i^*	Effectif n_i	Ecarts $x_i^* - \bar{x}$	(Ecarts)^2 $(x_i^* - \bar{x})^2$	Produit $n_i.(x_i^* - \bar{x})^2$
$[165-170[$	167,5	6	9	81	486
$[170-175[$	172,5	10	4	16	160
$[175-180[$	177,5	5	1	1	5
$[180-185[$	182,5	4	6	36	144
$[158-190[$	187,5	3	11	121	363
$[190-195[$	192,5	2	16	256	512
Total		30			1670

$$s^2 = \frac{1670}{30} = 55,667.$$

En utilisant les formules simplifiées on obtient :

Classe $[a-b[$	Centre x_i^*	(Centres)^2 x_i^{*2}	Produit $n_i.x_i^{*2}$
$[165-170[$	167,5	28056,25	168337,5
$[170-175[$	172,5	29756,25	297562,5
$[175-180[$	177,5	31506,25	157531,25
$[180-185[$	182,5	33306,25	133225
$[158-190[$	187,5	35156,25	105468,75
$[190-195[$	192,5	37056,25	74112,5
Total			936237,5

On obtient

$$s^2 = \frac{1}{30}936237,5 - 176,5^2 = 55,667.$$

Ecart-type

$$s = \sqrt{s^2} = 7,76$$

Coefficient de variation

$$CV = \frac{s}{\bar{x}} * 100 = \frac{7,76}{176,5} * 100 = 4,227\% < 15\%$$

homogène distribution

Moments centré

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^3 = \frac{1}{30}(8040) = 268$$

$$\mu_4 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^4 = \frac{222110}{30} = 7403,667$$

Coefficient d'asymétrie

$$g_1 = \frac{\mu_3}{s^3} = 0,645 > 0$$

dissymétrie à gauche.

Coefficient d'aplatissement

$$g_2 = \frac{\mu_4}{s^4} - 3 = -0,61079 < 0$$

plus aplatie - concentration faible autour de $\bar{x}$.

Boite à moustaches. Exercice 29

Un jury de délibération désire analyser les résultats (notes sur 100 points) obtenus par 10 étudiants dans 7 matières distinctes. Le tableau ci-dessous est le tableau individus x caractères contenant ces résultats. Déterminer les « boîtes à moustaches » pour les 7 cours observés.

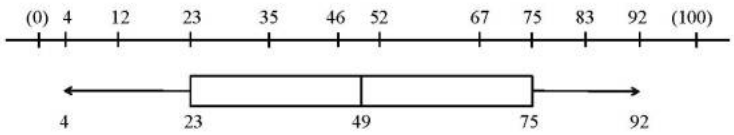
Étudiants	Matières						
	C_1	C_2	C_3	C_4	C_5	C_6	C_7
E_l	52	47	51	69	83	76	24
E_2	23	44	19	67	24	75	23
E_3	83	58	63	77	85	83	27
E_4	75	51	43	85	86	80	30
E_5	04	46	25	33	27	14	19
E_6	35	56	31	47	77	77	21
E_7	67	49	27	75	79	78	29
E_8	92	57	73	83	87	84	93
E_9	12	42	23	59	21	79	18
E_{10}	46	54	48	73	29	81	25

C_1	C_2	C_3	C_4	C_5	C_6	C_7
04	42	19	33	21	14	18
12	44	23	47	24	75	19
23	46	25	59	27	76	21
35	47	27	67	29	77	23
46	49	31	69	77	78	24
52	51	43	73	79	79	25
67	54	48	75	83	80	27
75	56	51	77	85	81	29
83	57	63	83	86	83	30
92	58	73	85	87	84	93

Solution Déterminons la boîte à moustaches pour les résultats obtenus dans le cours C_1 :

- $x_{1/2} = \dfrac{x_{(5)} + x_{(6)}}{2} = \dfrac{46 + 52}{2} = 49$
- $x_{1/4} = x_{([2,5])} = x_{(3)} = 23$
  $x_{3/4} = x_{([7,5])} = x_{(8)} = 75$
- $x_{3/4} - x_{1/4} = 75 - 23 = 52$
- $p_g = 23 - 1,5 \times 52 = -55$
  $p_d = 75 + 1,5 \times 52 = 153$

Tous les résultats obtenus dans le cours sont compris dans l'intervalle $[p_g, p_d]$; $x_g = x_{(1)} = 04$ et $x_d = x_{(10)} = 92$. La version modifiée de la boîte à moustaches coïncide avec la version de base.



Déterminons à présent la boîte à moustaches pour les résultats obtenus dans le cours C_7 :

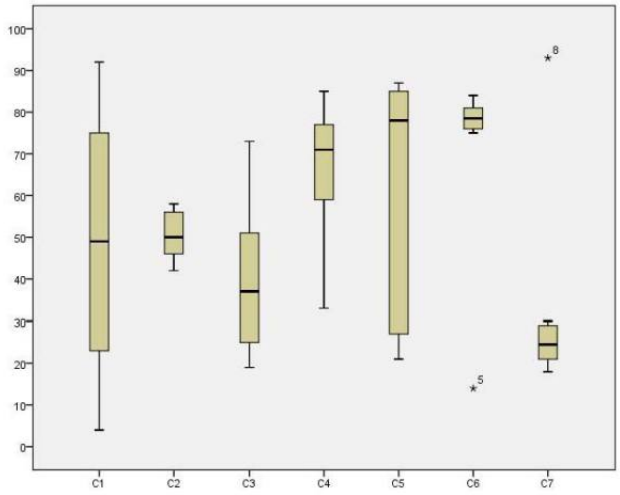
- $x_{1/2} = \dfrac{x_{(5)} + x_{(6)}}{2} = \dfrac{24 + 25}{2} = 24,5 \quad (2^e \text{ convention})$
- $x_{1/4} = x_{([2,5])} = x_{(3)} = 21$
  $x_{3/4} = x_{([7,5])} = x_{(8)} = 29$
- $x_{3/4} - x_{1/4} = 29 - 21 = 8$
- $p_g = 21 - 1,5 \times 8 = 9$
  $p_d = 29 + 1,5 \times 8 = 41$

Tous les résultats obtenus dans le cours C_7 sont supérieurs à p_g ; dès lors, $x_g = x_{(1)} = 18$;

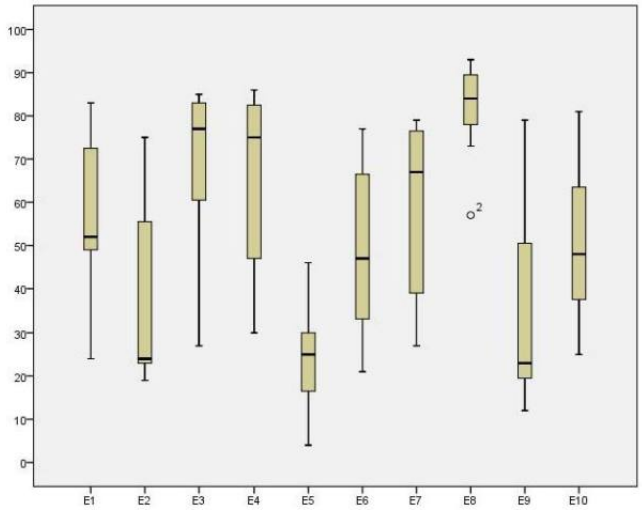
Le plus grand résultat inférieur ou égal à p_g est 30 ; on a donc $x_d = 30$ et $x_{(10)} = 93$ est une valeur extérieure, représentée par une étoile.

Manifestement, seul l'étudiant E_8 a trouvé grâce aux yeux de l'enseignant de C_7 .

Ces représentations graphiques sont simples à construire. Elles permettent de voir aisément la manière dont les observations se répartissent, soit par cours, soit par étudiant, et facilitent donc la comparaison entre cours et entre étudiants, comme on peut le constater dans les deux figures ci-dessous.



Boîtes à moustaches par cours



Boîtes à moustaches par étudiant

Ajustement linéaire. Exercice 40

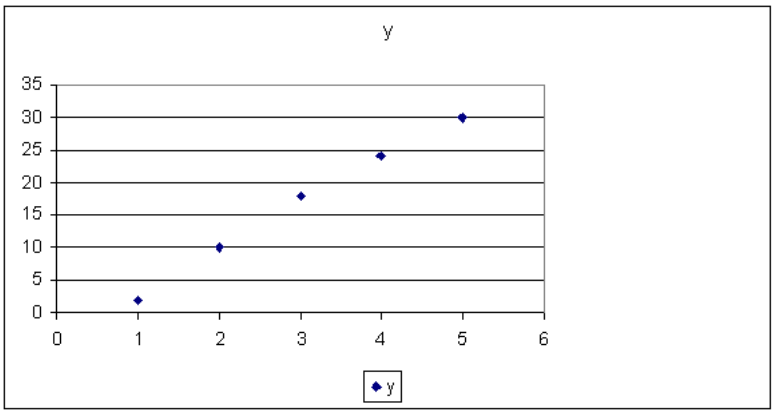
A l'Université A une nouvelle méthode d'enseignement est introduite. Le tableau donne l'évolution du taux en % des notes parfaites à l'examen final à la fin des cours d'après la nouvelle méthode d'enseignement.

Année	2010	2011	2012	2013	2014
Rang de l'année X	1	2	3	4	5
Taux de notes en % Y	2	10	18	24	30

a) Représenter dans un repère le nuage des points
b) Existe-il une corrélation linéaire entre les variables X et Y ?
c) Déterminer l'équation de la droite des moindres carrées D_y (droite de y en x)
d) Tracer D_y
e) Donner une estimation pour la 6^{eme} année.

Solution

a) Représenter dans un repère le nuage des points



b) Existe-il une corrélation linéaire entre les variables X et Y ?

$$r = \frac{cov(X,Y)}{s_X * s_Y} = \frac{\sum_{i=1}^{5} x_i y_i - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{5}\sum_{i=1}^{5} x_i^2 - \bar{x}^2\right)\left(\frac{1}{5}\sum_{i=1}^{5} y_i^2 - \bar{y}^2\right)}}$$
$$= \frac{\frac{1}{5}322 - 3 * 16.8}{\sqrt{\left(\frac{55}{5} - 3^2\right)\left(\frac{1904}{5} - 16.8^2\right)}} = \frac{14}{\sqrt{2 * 98.56}} = 0.99715$$

Forte liaison linéaire.

c) Déterminer l'équation de la droite des moindres carrées D_y (droite de y en x)

$$D_y : y = av + b \qquad a = \frac{cov(x,y)}{var(x)} = \frac{14}{2} = 7$$

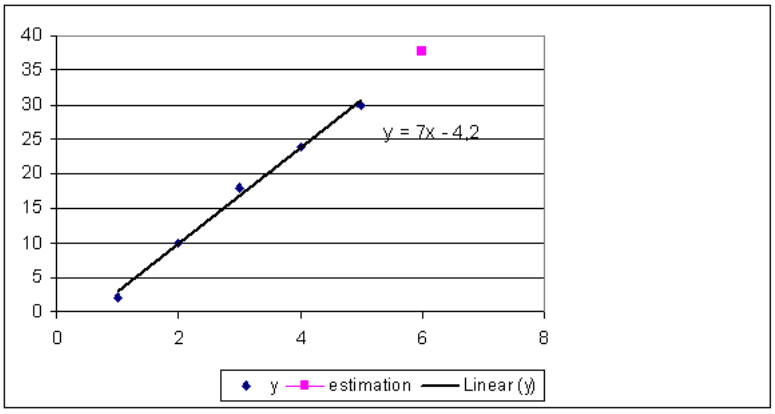
$$b = \bar{y} - a\bar{x} = 16.8 - 7 * 3 = -4.2$$

$$D_y : y = 73 - 4.2$$

d) Tracer D_y
$A(0;\ -4.2)\,;\ B(6;\ 37.8)\,;$

e) Donner une estimation pour la 6^{eme} année (2016).

$$y(6) = 7 * 6 - 4.2 = 42 - 4.2 = 37.8$$



Séries chronologiques. Exercice 48

L'exercice a pour objectif l'étude du chiffre d'affaires trimestriel (en milliers d'euros) de l'entreprise Peugeot présenté dans le tableau ci-dessous.

Année / Trimestre	2010	2011
I	13986	15414
II	14408	15721
III	12993	13450
IV	14674	15237

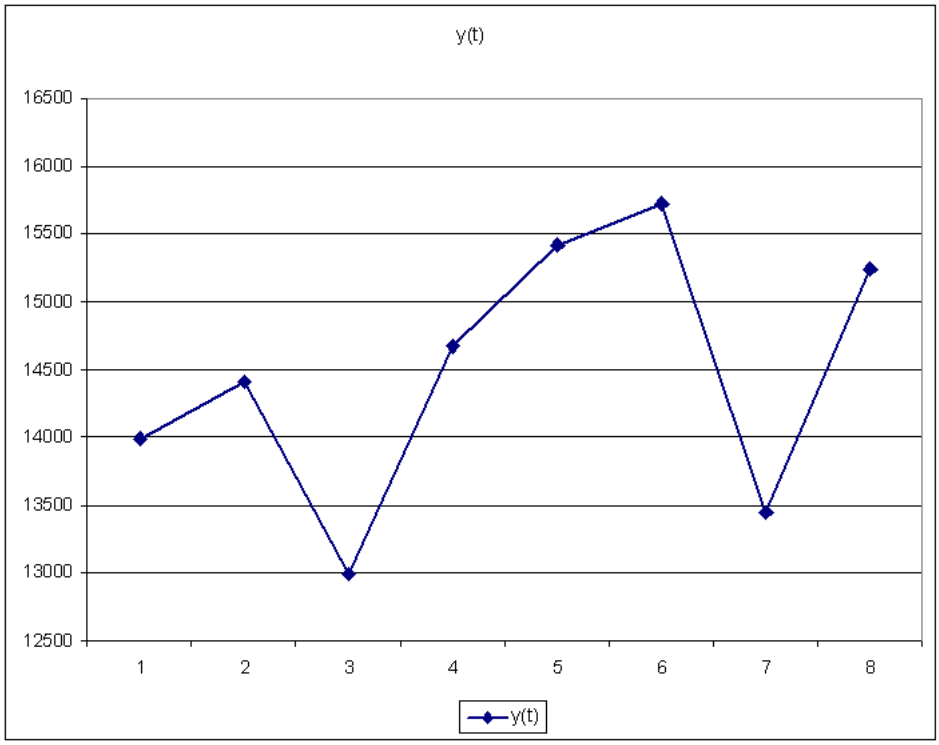
1. Représentez cette série chronologique sur un graphique. On adoptera un modèle additif pour modéliser cette série chronologique.

Date	Temps	$y(t)$	$y'(t)$	$T(t)$	C.V.S. $= y_i - S_i$	$S_i = S'_i - \bar{S}$	S'_i
2010, I	1	13986	×				
II	2	14408	×				
III	3	12993	14193,8				
IV	4	14674	14536,4				
2011, I	5	15414	14757,6				
II	6	15721	??				
III	7	13450	×				
IV	8	15237	×				

2. Complétez la 3ème colonne du tableau précédent en calculant la moyenne mobile d'ordre 4 (ordre que vous justifierez) manquante. Représentez la série des moyennes mobiles sur le graphique.
3. Calculez la tendance T(t) par un modèle de régression linéaire (que vous justifierez en calculant le coefficient de corrélation) de la série y' en fonction du temps. Représentez la droite de régression sur le graphique.
4. Complétez les trois dernières colonnes du tableau visant à calculer les coefficients saisonniers.
5. Quel chiffre d'affaires l'entreprise Peugeot pouvait-elle espérer les trois premiers trimestres de 2012 ? Quels sont vos commentaires ?

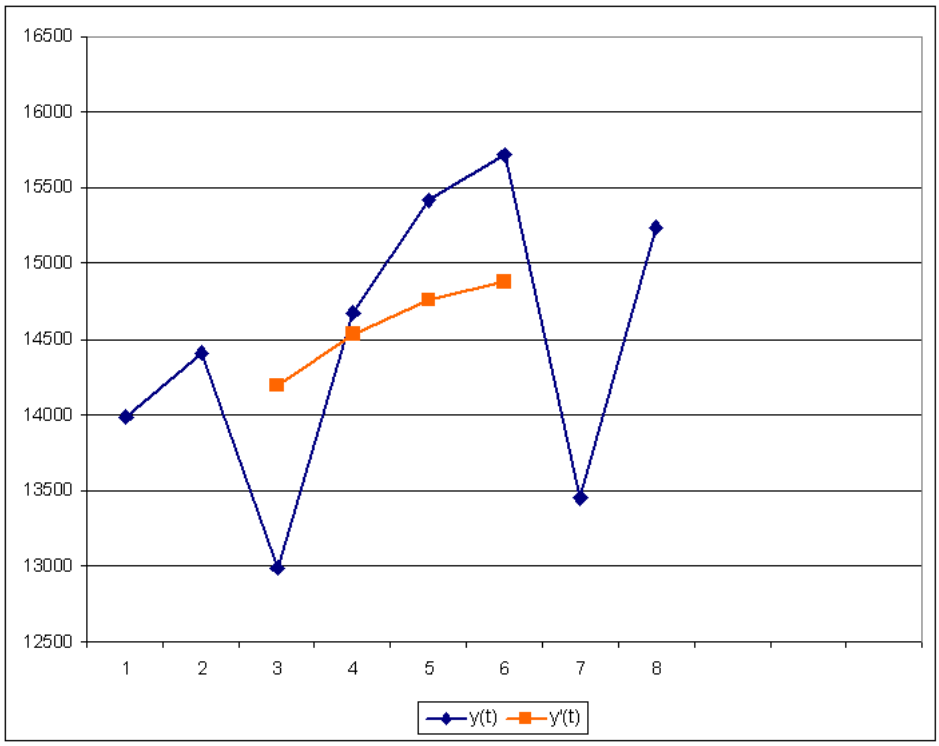
Solution

1. Représentation de la série chronologique sur un graphique.



2. Calcul de la moyenne mobile d'ordre 4 manquante dans la 4-ème colonne du tableau. Représentation de la série des moyennes mobiles sur le graphique.

$$y'(6) = \frac{y(4)/2 + y(3) + y(2) + y(1)/2}{4} = \frac{14674/2 + 15414 + 15724 + 13450 + 15237/2}{4} = 14885.13$$



3. Calcul de la tendance T(t) par un modèle de régression linéaire, (justifié en calculant le coefficient de corrélation) de la série y' en fonction du temps. Représentation la droite de régression sur le graphique.

$$\begin{aligned} r &= \frac{cov(t,y')}{s_t * s_{y'}} \\ cov(t,y') &= \frac{1}{4}\sum_{i=1}^{4} t_i * y'_i - \overline{t}\overline{y'} \end{aligned}$$

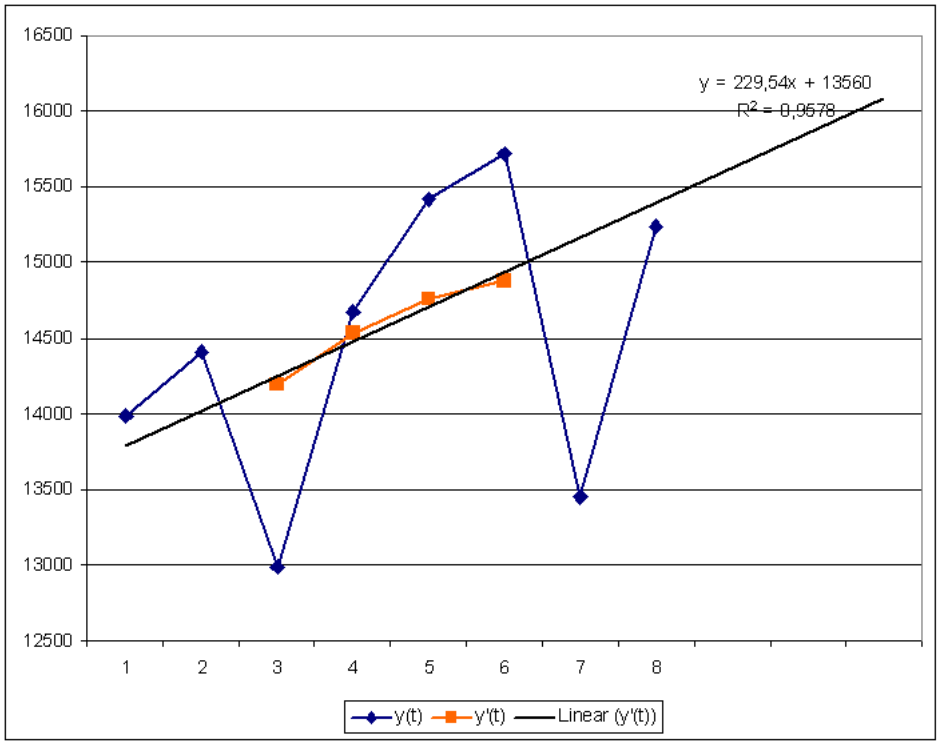
Le valeurs des sommes sont : $\sum t_i y'_i = 263826.6$; $\sum t_i = 15$; $\sum y'_i = 58373.125$; $\sum t_i^2 = 86$; $\sum {y'}_i^2 = 852130402$

$$\begin{aligned} \bar{t} &= \frac{1}{4}\sum_{i=1}^{4} t_i = 18/4 = 4.5 \\ \overline{y'} &= \frac{1}{4}\sum_{i=1}^{4} y'_i = 58373.125/4 = 14593.28125 \\ cov(t,y') &= \frac{1}{4}\sum_{i=1}^{4} t_i y'_i - \bar{t} * \overline{y'} \\ &= 263826.6/4 - 4.5 * 14593.28 = 286.87 \\ s_t &= \sqrt{\frac{1}{4}\sum_{i=1}^{4} t_i^2 - \bar{t}^2} = \sqrt{\frac{1}{4}86 - 435^2} = 1.18 \\ s_{y'} &= \sqrt{\frac{1}{4}\sum_{i=1}^{4} y_i'^2 - \overline{y'}^2} = \sqrt{\frac{1}{4}852130402 - 14593.28^2} = 262.1886 \\ r &= \frac{cov(t,y')}{s_t * s_{y'}} \\ cov(t,y') &= \frac{1}{4}\sum_{i=1}^{4} t_i * y'_i - \bar{t} * \overline{y'} = \frac{286.87}{1.18 * 262.19} = 0.9786 \Rightarrow \text{ liaison linéaire forte} \end{aligned}$$

Ligne de régression : $T(t) = at + b$, avec

$$\begin{aligned} a &= \frac{cov(t,y')}{var(t)} = \frac{286.87}{1.18^2} = 229.5 \\ b &= \overline{y'} - a\bar{t} = 14593.28 - 229.5 * 4.5 = 13561 \\ T(t) &= a * t + b = 229.5 * t + 13561 \end{aligned}$$

$T(1) = 229.5 + 13561 = 13790.5$; $T(10) = 2295 + 13561 = 15856$



4. Complétez les deux dernières colonnes du tableau visant à calculer les coefficients saisonniers.

Première estimation des coefficients saisonniers

$$
\begin{aligned}
S'_i &= y_i - y'_i, \quad i = 1,2,3,4 \\
S'_3 &= y_3 - y'_3 = 12993 - 14193.8 = -1200.8 \\
S'_4 &= y_4 - y'_4 = 14674 - 14536.4 = 135.6 \\
S'_1 &= y_1 - y'_1 = 656.4 \\
S'_2 &= y_2 - y'_2 = 835.88
\end{aligned}
$$

$$
\overline{S'} = \frac{1}{4}\sum_{i=1}^{4} S'_i = 107.27
$$

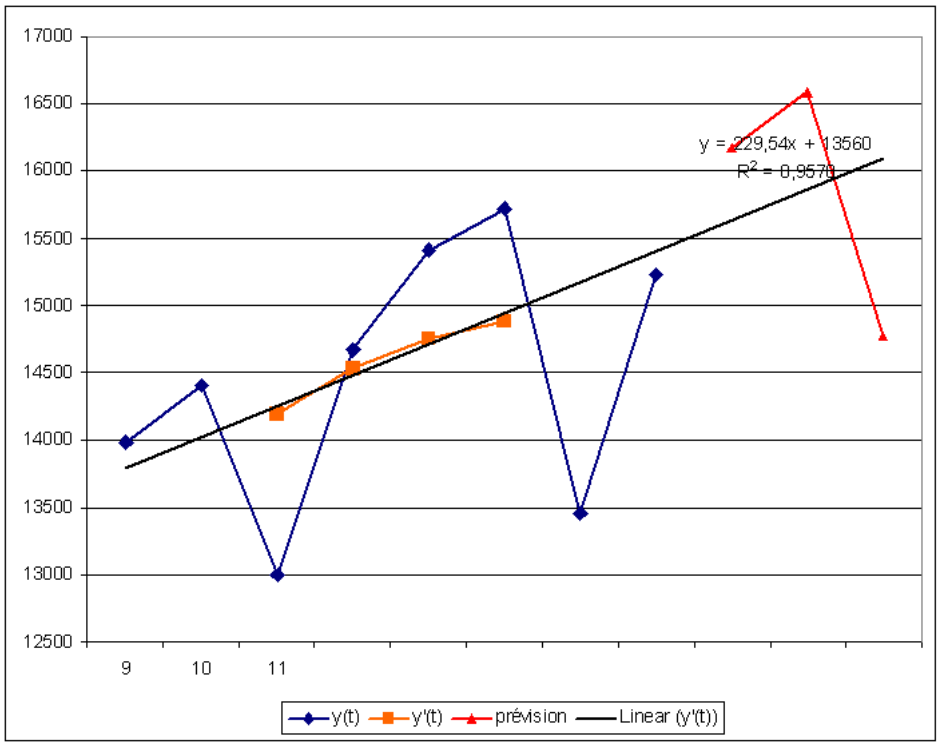
Coefficients saisonniers

$$
\begin{aligned}
S_1 &= S'_1 - \overline{S'} = 656.375 - 107.2813 = 549.09 \\
S_2 &= S'_2 - \overline{S'} = 835.88 - 107.28 = 758.59 \\
S_3 &= S'_3 - \overline{S'} = -1200.75 - 107.28 = -1308.09 \\
S_4 &= S'_4 - \overline{S'} = 137.63 - 107.28 = 30.34
\end{aligned}
$$

Date	Temps	$y(t)$	$y'(t)$	$T(t)$	C.V.S. $= y_i - S_i$	$S_i = S'_i - \bar{S}$	S'_i
2010, I	1	13986	×	13790			
II	2	14408	×	14019			
III	3	12993	14193,8	14249		-1200.8	-1308.09
IV	4	14674	14536,4	14479		135.6	30.34
2011, I	5	15414	14757,6	14708		656.4	549.09
II	6	15721	14885.13	14938		835.88	758.59
III	7	13450	×	15167			
IV	8	15237	×	15397			

5. Quel chiffre d'affaires l'entreprise Peugeot pouvait-elle espérer les trois premiers trimestres de 2012 ? Quels sont vos commentaires ?

$$\begin{aligned}
\hat{y}(9) &= a * 9 + b + S_1 = 229.54 * 9 + 13560.3 + 549.09 = 16175.23 \\
\hat{y}(10) &= a * 10 + b + S_2 = 229.54 * 10 + 13560.3 + 728.59 = 16584.27 \\
\hat{y}(11) &= a * 11 + b + S_3 = 229.54 * 11 + 13560.3 - 1308.03 = 14777.18
\end{aligned}$$



Séries chronologiques. Exercice 49

Pendant deux semaines consécutives, on a observe le nombre de visiteurs d'un musée dont les jours de fermeture sont le samedi et le dimanche.

	Lundi	Mardi	Mercredi	Jeudi	Vendredi
Première semaine	7	5	35	5	6
Deuxième semaine	8	9	45	8	9

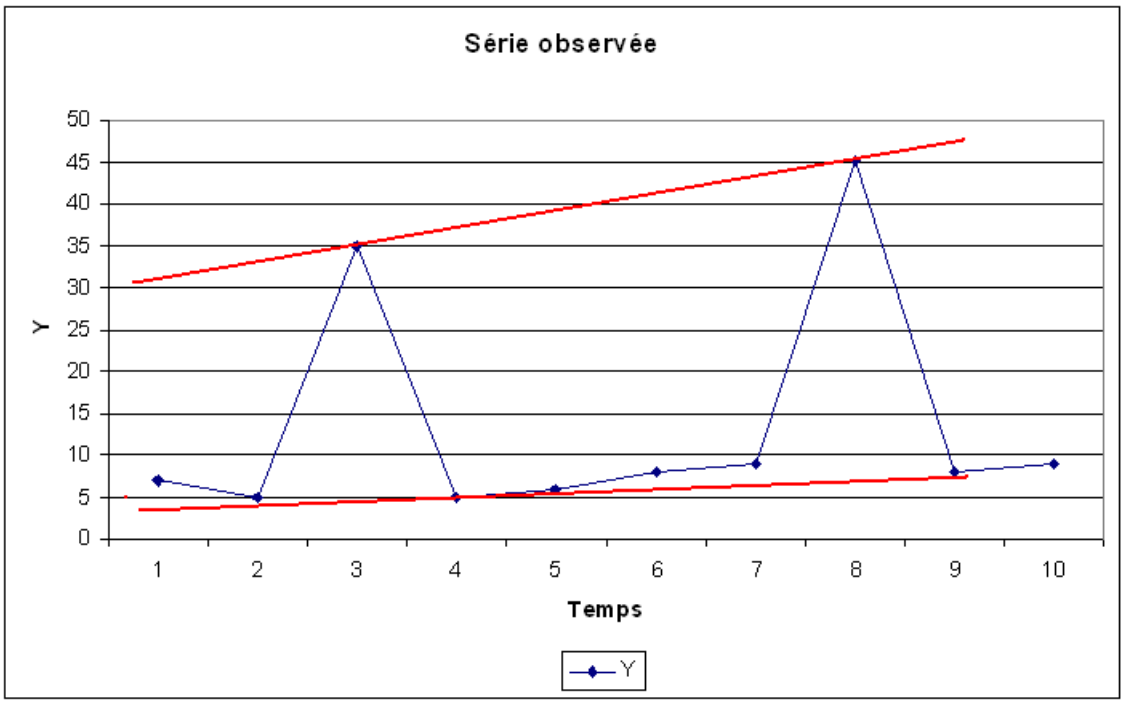
Considérons un modèle additif : $Y = T + S$.
1). Représentez graphiquement Y en fonction du temps. Pourquoi prend-on un modèle additif ?
2). Calculez les moyennes mobiles d'ordre 5, notées MM. Représentez graphiquement cette moyenne mobile. Pourquoi prend-on un ordre 5 ?
3). Effectuez un ajustement linéaire sur cette série chronologique Y. Justifier que le modèle est adéquate. Représentez graphiquement cet ajustement.
4). Déterminez les composantes saisonnières par la méthode de comparaison à la tendance.
5). Effectuer la désaisonnalisation (Calculer la série corrigée des variations saisonnières (c.v.s)).
6). Sur base du modèle additif et des résultats ci-dessus, donnez la prévision pour le lundi et le mardi de la 3-ième semaine.

Solution

Premièrement on énumère les jours des observations de 1 à 10.

t	y
1	7
2	5
3	35
4	5
5	6
6	8
7	9
8	45
9	8
10	9

1). Représentez graphiquement Y en fonction du temps. Pourquoi prend-on un modèle additif ?

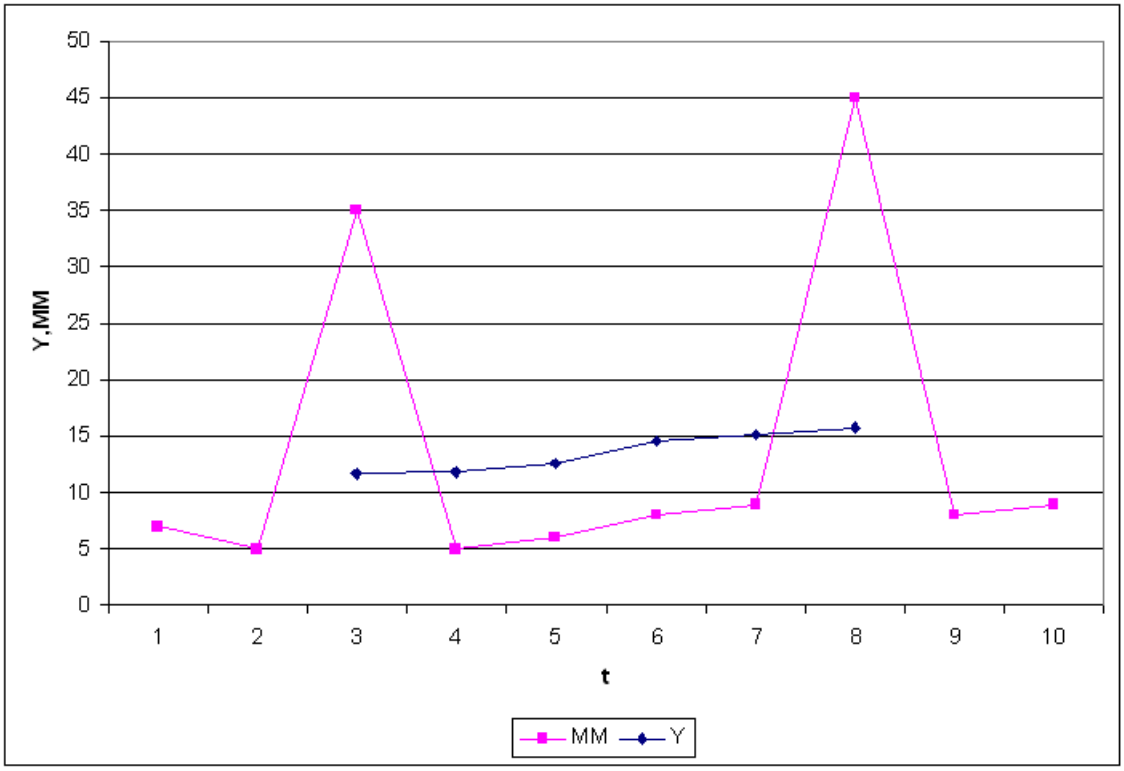


On choisit le modèle additif, parce que le graphique des observations se situe presque dans un tube.

2). Calculez les moyennes mobiles d'ordre 5, notées MM. Représentez graphiquement cette moyenne mobile. Pourquoi prend-on un ordre 5 ?
$l = 5$: L'ordre l des moyennes mobiles MM doit être multiple de la période p des variations saisonnières pour atténuer la composante saisonnière.

$$
\begin{aligned}
MM_3 &= \frac{y_1 + y_2 + y_3 + y_4 + y_5)}{5} = \frac{7 + 535 + 5 + 6}{5} = 11.6 \\
MM_4 = &= \frac{y_2 + y_3 + y_4 + y_5 + y_6)}{5} = \frac{5 + 35 + 5 + 6 + 8}{5} = 11.8 \\
MM_5 &= \frac{y_3 + y_4 + y_5 + y_6 + y_7}{5} = \frac{35 + 5 + 6 + 8 + 9}{5} = 12.6 \\
MM_6 &= \frac{y_4 + y_5 + y_6 + y_7 = y_8}{5} = \frac{5 + 6 + 8 + 9 + 45}{5} = 14.6 \\
MM_7 &= \frac{y_5 + y_6 + y_7 = y_8 + y_9}{5} = \frac{6 + 8 + 9 + 45 + 8}{5} = 15.2 \\
MM_8 &= \frac{y_6 + y_7 = y_8 + y_9 + y_{10}}{5} = \frac{8 + 9 + 45 + 8 + 9}{5} = 15.8
\end{aligned}
$$

t	y	MM
1	7	×
2	5	×
3	35	11.6
4	5	11.8
5	6	12.6
6	8	14.6
7	9	15.2
8	45	15.8
9	8	×
10	9	×



3). Effectuez un ajustement linéaire sur cette série chronologique Y. Justifier que le modèle est adéquate. Représentez graphiquement cet ajustement.
On effectue un ajustement linéaire des points des moyennes mobiles. La droite de régression, obtenue d'après la méthode des moindres carrées est

$$T = at + b,$$

avec $a = \frac{cov(t,MM)}{s_t^2}$ et $b = \overline{MM} - a * \bar{t}$.
Pour les sommes on a :

$$\sum_{i=3}^{8} t_i = 33; \quad \sum_{i=3}^{8} MM_i = 81.6; \quad \sum_{i=3}^{8} t_i^2 = 199; \quad \sum_{i=3}^{8} MM_i^2 = 1126; \quad \sum_{i=3}^{8} t_i * MM_i = 465.4$$

D'ici

$$\bar{t} = \frac{1}{6}\sum_{i=3}^{8} t_i = \frac{33}{6} = 5.5; \quad \overline{MM} = \frac{1}{6}\sum_{i=3}^{8} MM_i = \frac{81.6}{6} = 13.6$$

$$s_t^2 = \frac{1}{6}\sum_{i=3}^{8} t_i^2 - \bar{t}^2 = \frac{199}{6} - 5.5^2 = 2.91666; \quad s_t = \sqrt{s_t^2} = \sqrt{2.91666} = 1.7078;$$

$$s_{MM}^2 = \frac{1}{6}\sum_{i=3}^{8} MM_i^2 - \overline{MM}^2 = \frac{1126}{6} - 13.6^2 = 2.7733; \quad s_{MM} = \sqrt{s_{MM}^2} = \sqrt{2.7733} = 1.6653$$

$$cov(t, MM) = \frac{1}{6}\sum_{i=3}^{8} t_i * MM_i - \bar{t} * \overline{MM} = \frac{465.4}{6} - 5.5 * 13.6 = 2.7666.$$

Alors, pour le coefficient de corrélation $r = \frac{cov(t,MM)}{s_t * s_{MM}}$ on obtient la valeur

$$r \quad = \quad \frac{cov(t, MM)}{s_t * s_{MM}} = \frac{2.7666}{1.7078 * 2.7666} = 0.97 > 0.87 \text{ et } \approx 1$$

Le coefficient de corrélation $r = 0.97$ justifie un modèle adéquate à la distribution observée.

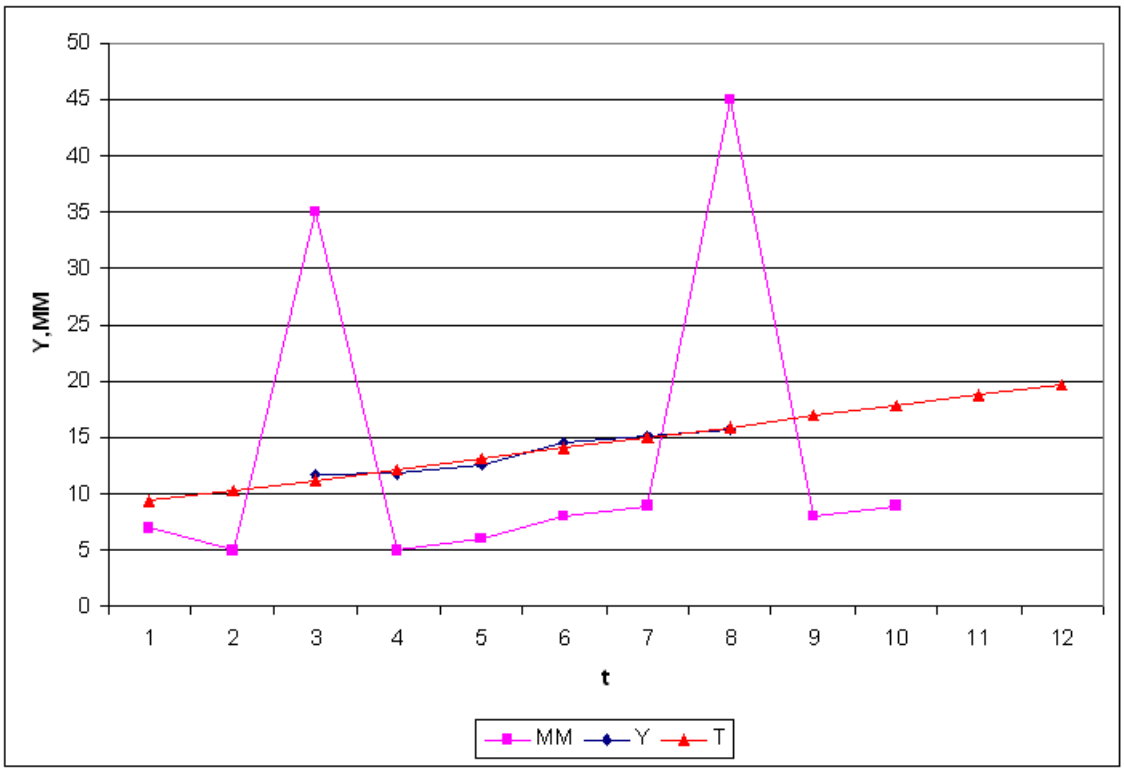
La droite de régression est :

$$a = \frac{cov(t, MM)}{s_t^2} = \frac{2.7666}{2.91666} = 0.95 \quad b = \overline{MM} - a * \bar{t} = 13.6 - 0.955.5 = 8.28$$
$$T : a * t + b = 0.95 * t + 8.38$$

La tendance de la série observée, obtenue du modèle de l'ajustement linéaire s'obtient de l'équation de la droite T pour les différents valeurs de t. Comme question 5 exige l'obtention de c.v.s et question 6 exige une prédiction pour le lundi et le mardi de la 3-ème semaine, on va calculer aussi les valeur de T pour $t = 11$ et $t = 12$. :

t	y	MM	T
1	7	×	9.33
2	5	×	10.28
3	35	11.6	11.23
4	5	11.8	12.18
5	6	12.6	13.13
6	8	14.6	14.07
7	9	15.2	15.02
8	45	15.8	15.97
9	8	×	16.97
10	9	×	17.87
11			18.82
12			19.77

Représentation de la tendance générale T sur le graphique :



4). Déterminez les composantes saisonnières par la méthode de comparaison à la tendance. On a choisi le modèle additif. Dans ce cas la première approximation du coefficient saisonnier est

$$\begin{aligned} aS_i &= y_i = T_i, \\ aS_1 &= y_1 - T_1 = 7 - 9.33 = -2.33 \\ \ldots &= \ldots \end{aligned}$$

t	y	MM	T	aS_i	S_i'	S_i	c.v.s	Y'
1	7	×	9.33	-2.33	-4.20	-4.30	11.3	
2	5	×	10.28	-5.28	-5.65	-5.75	10.75	
3	35	11.6	11.23	23.77	26.4	26.3	8.7	
4	5	11.8	12.18	-7.18	-8.05	-8.15	13.15	
5	6	12.6	13.13	-7.13	-7.997	-8.097	14.097	
6	8	14.6	14.07	-6.07			12.30	
7	9	15.2	15.02	-6.02			14.75	
8	45	15.8	15.97	29.03			18.7	
9	8	×	16.97	-8.92			8	
10	9	×	17.87	-8.87			17.097	
11			18.82					14.52
12			19.77					14.01
13			20.71					47.01
14			21.66					13.51
15			22.61					14.51

La première estimation des coefficients saisonniers est la moyenne des valeurs des premières

approximations des coefficients saisonniers pour chaque saison :

$$\begin{aligned} S'_1 &= (aS_1 + aS_6)/2 = (-2.33 - 6.07)/2 = -4.20 \\ S'_2 &= (aS_2 + aS_7)/2 = (-5.28 - 6.03)/2 = -5.65 \end{aligned}$$

La moyenne des premières estimation des 5 coefficients saisonniers est

$$\bar{S}' = (-4.30 - 5.65 + 26.4 - 8.05 - 7.998)/5 = 0.1$$

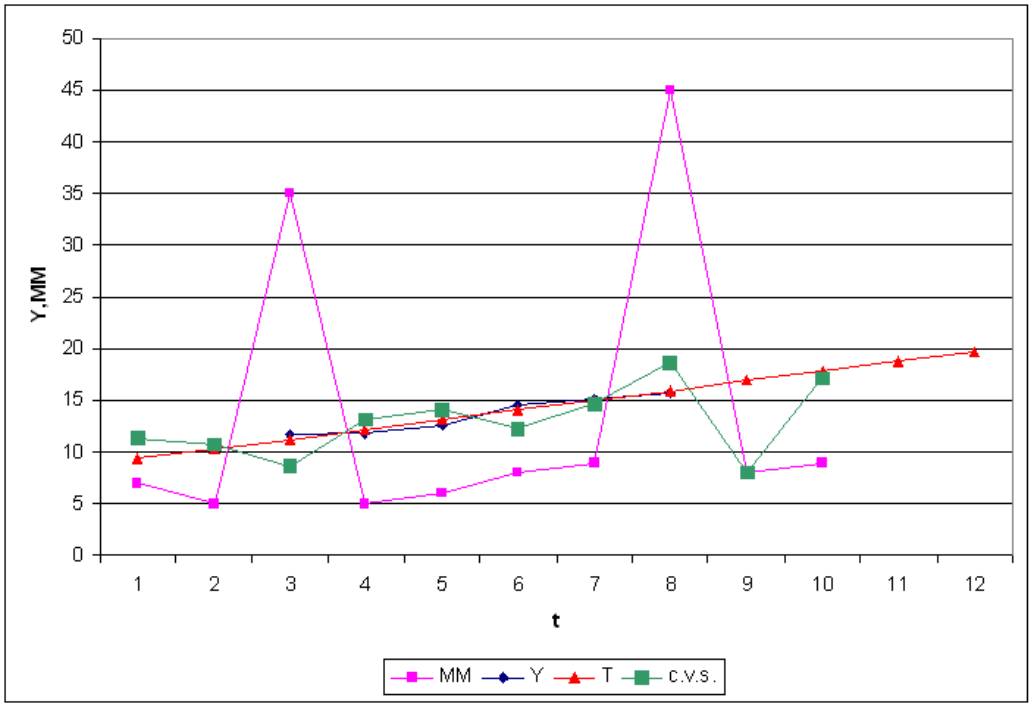
Les valeurs des coefficients saisonniers sont, respectivement

$$\begin{aligned} S_i &= S'_i - \bar{S} : \\ S_1 &= S'_1 - \bar{S} = -4.20 - 0.1 = -4.30 \\ S_2 &= S'_2 - \bar{S} = -5.65 - 0.1 = -5.75 \\ \dots &\quad \dots \end{aligned}$$

5). Effectuer la désaisonnalisation (Calculer la série corrigée des variations saisonnières (c.v.s)).

La desaisonnalisation s'effectue en enlevant de la série observée la composante saisonnière :

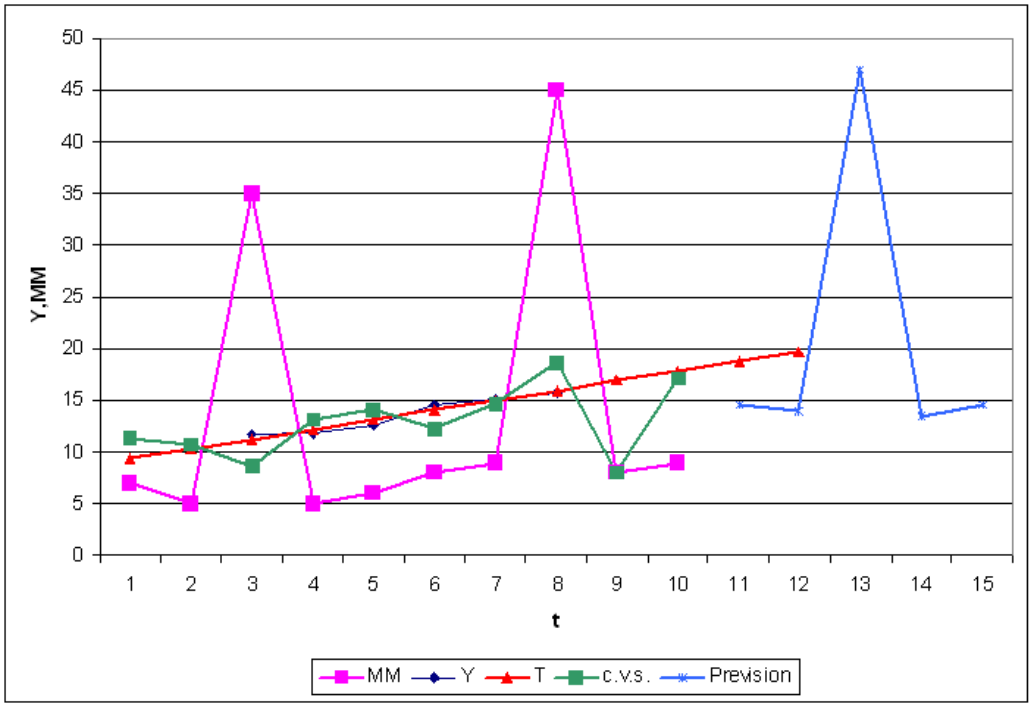
$$\begin{aligned} c.v.s._i &= y_i - S_i \\ c.v.s._1 &= y_1 - S_1 = 7 + 4.30 = 11.3 \\ c.v.s._2 &= y_2 - S_2 = 5 + 5.75 = 10.75 \\ \dots &\quad \dots \end{aligned}$$



6). Sur base du modèle additif et des résultats ci-dessus, donnez la prévision pour le lundi et le mardi de la 3-ième semaine.

On obtient les valeurs prévues y'_t à partir de de la tendance générale (la droite T) pour les valeurs de $t = 12$ et 13 en y ajoutant la composante saisonnière :

$$\begin{aligned} y'_1 &= T(1) + S_1 = a * 1 + b + S_1 = 0.95 * 11 + 8.38 - 4.30 = 18.82 - 4.30 = 14.52 \\ y'_2 &= T(2) + S_2 = \ldots = 14.01 \\ \ldots &\quad \ldots \end{aligned}$$



Echantillonnage. Exercice 61

Une entreprise fournit des lots d'environ 10 000 pièces. Elle certifie que les lots ont une proportion de défectueux n'excédant pas 3 %.
Un client réceptionne chaque lot et effectue un test. Ce test conduit à la règle de décision suivante pour un échantillon aléatoire de 500 pièces issu d'un lot :

- le lot est accepté si l'échantillon contient au plus 21 pièces défectueuses,
- le lot est refusé si l'échantillon contient plus de 21 pièces défectueuses.

1) Si la proportion de défectueux du lot est 3 %, déterminer la probabilité que le lot testé soit refusé.
2) Quelle est la probabilité que le client accepte un lot dont la proportion de défectueux est 6 % ?

Solution

$$\begin{aligned} P: &\quad N = 20000; \quad p = 0.03; \quad q = 1 - p = 0.97 \\ E: &\quad n = 500 \\ &\quad \text{taux d'échantillonnage } \frac{n}{N} = \frac{500}{20000} = 0.025 < 0.05 \end{aligned}$$

On ne fait compte du facteur d'exhaustivité.

$$n > 30 \quad np = 500 * 0.03 = 15 \ \ (\geq 15) \quad nq = 500 * 0.97 = 485 \ \ (\geq 15)$$

$$\Rightarrow F = \text{ proportion des pièces défectueuses :}$$

$$F \sim \mathcal{N}\left(p; \sqrt{\frac{pq}{n}}\right) = \mathcal{N}\left(0.03; \sqrt{\frac{0,03 * 0.97}{500}}\right) = \mathcal{N}\left(0.03; 0.00763\right)$$

Pour que le lot soit refusé il faut que $F \geq 22$. Loi Binomiale approximée par la loi normale - correction de continuité.

$$\begin{aligned} P\left(F > \frac{21.5}{500}\right) &= P(F > 0.043) = 1 - P\left(Z < \frac{0.043 - 0.03}{0.00763}\right) \\ &= 1 - P(Z < 1.79) = 1 - \pi(1.70) = 1 - 0.95543 = 0.04457 \rightarrow 4.46\% \end{aligned}$$

Tables statistiques

Table de Loi Normale

Fractiles de la Loi Normale

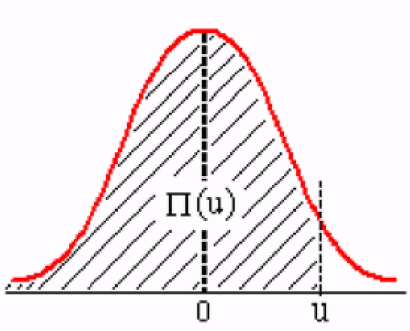
Fractiles de la loi du χ^2_ν

Table de la loi de Student

Table de la loi Normale

Fonction de répartition Π de la loi normale centrée réduite : $U \to \mathcal{N}(0,1)$
Probabilité de trouver une valeur inférieure à u

$$\Pi(u) = P(U \leq u); \Pi(-u) = P(U \leq -u) = 1 - \Pi(u)$$



u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	**0.89617**	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992

Exemple : $\Pi(1.26) = P(U \leq 1.26) = 0.89617 = 89.62\%$

Fractiles de la loi normale

$$U \to \mathcal{N}(0,1)$$

Pour $P < 0.5$ (colonne de gauche et ligne supérieure). Les fractiles sont négatifs.
Pour $P > 0.5$ (colonne de droite et ligne inférieure). Les fractiles sont positifs.

P	0	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01	
0	infini	3.0902	2.8782	2.7478	2.6521	2.5758	2.5121	2.4573	2.4089	2.3656	2.3263	**0.99**
0.01	2.3263	2.2904	2.2571	2.2262	2.1973	2.1701	2.1444	2.1201	2.0969	2.0748	2.0537	**0.98**
0.02	2.0537	2.0335	2.0141	1.9954	1.9774	1.9600	1.9431	1.9268	1.9110	1.8957	1.8808	**0.97**
0.03	1.8808	1.8663	1.8522	1.8384	1.8250	1.8119	1.7991	1.7866	1.7744	1.7624	1.7507	**0.96**
0.04	1.7507	1.7392	1.7279	1.7169	1.7060	1.6954	1.6849	1.6747	1.6646	1.6546	1.6449	**0.95**
0.05	1.6449	1.6352	1.6258	1.6164	1.6072	1.5982	1.5893	1.5805	1.5718	1.5632	1.5548	**0.94**
0.06	1.5548	1.5464	1.5382	1.5301	1.5220	1.5141	1.5063	1.4985	1.4909	1.4833	1.4758	**0.93**
0.07	1.4758	1.4684	1.4611	1.4538	1.4466	1.4395	1.4325	1.4255	1.4187	1.4118	1.4051	**0.92**
0.08	1.4051	1.3984	1.3917	1.3852	1.3787	1.3722	1.3658	1.3595	1.3532	1.3469	1.3408	**0.91**
0.09	1.3408	1.3346	1.3285	1.3225	1.3165	1.3106	1.3047	1.2988	1.2930	1.2873	1.2816	**0.90**
0.10	1.2816	1.2759	1.2702	1.2646	1.2591	1.2536	1.2481	1.2426	1.2372	1.2319	1.2265	**0.89**
0.11	1.2265	1.2212	1.2160	1.2107	1.2055	1.2004	1.1952	1.1901	1.1850	1.1800	1.1750	**0.88**
0.12	1.1750	1.1700	1.1650	1.1601	1.1552	1.1503	1.1455	1.1407	1.1359	1.1311	1.1264	**0.87**
0.13	1.1264	1.1217	1.1170	1.1123	1.1077	1.1031	1.0985	1.0939	1.0893	1.0848	1.0803	**0.86**
0.14	1.0803	1.0758	1.0714	1.0669	1.0625	1.0581	1.0537	1.0494	1.0451	1.0407	1.0364	**0.85**
0.15	1.0364	1.0322	1.0279	1.0237	1.0194	1.0152	1.0110	1.0069	1.0027	0.9986	0.9945	**0.84**
0.16	0.9945	0.9904	0.9863	0.9822	0.9782	0.9741	0.9701	0.9661	0.9621	0.9581	0.9542	**0.83**
0.17	0.9542	0.9502	0.9463	0.9424	0.9385	0.9346	0.9307	0.9269	0.9230	0.9192	0.9154	**0.82**
0.18	0.9154	0.9116	0.9078	0.9040	0.9002	0.8965	0.8927	0.8890	0.8853	0.8816	0.8779	**0.81**
0.19	0.8779	0.8742	0.8706	0.8669	0.8632	0.8596	0.8560	0.8524	0.8488	0.8452	0.8416	**0.80**
0.20	0.8416	0.8381	0.8345	0.8310	0.8274	0.8239	0.8204	0.8169	0.8134	0.8099	0.8064	**0.79**
0.21	0.8064	0.8030	0.7995	0.7961	0.7926	0.7892	0.7858	0.7824	0.7790	0.7756	0.7722	**0.78**
0.22	0.7722	0.7688	0.7655	0.7621	0.7588	0.7554	0.7521	0.7488	0.7454	0.7421	0.7388	**0.77**
0.23	0.7388	0.7356	0.7323	0.7290	0.7257	0.7225	0.7192	0.7160	0.7128	0.7095	0.7063	**0.76**
0.24	0.7063	0.7031	0.6999	0.6967	0.6935	0.6903	0.6871	0.6840	0.6808	0.6776	0.6745	**0.75**
0.25	0.6745	0.6713	0.6682	0.6651	0.6620	0.6588	0.6557	0.6526	0.6495	0.6464	0.6433	**0.74**
0.26	0.6433	0.6403	0.6372	0.6341	0.6311	0.6280	0.6250	0.6219	0.6189	0.6158	0.6128	**0.73**
0.27	0.6128	0.6098	0.6068	0.6038	0.6008	0.5978	0.5948	0.5918	0.5888	0.5858	0.5828	**0.72**
0.28	0.5828	0.5799	0.5769	0.5740	0.5710	0.5681	0.5651	0.5622	0.5592	0.5563	0.5534	**0.71**
0.29	0.5534	0.5505	0.5476	0.5446	0.5417	0.5388	0.5359	0.5330	0.5302	0.5273	0.5244	**0.70**
0.30	0.5244	0.5215	0.5187	0.5158	0.5129	0.5101	0.5072	0.5044	0.5015	0.4987	0.4958	**0.69**
0.31	0.4958	0.4930	0.4902	0.4874	0.4845	0.4817	0.4789	0.4761	0.4733	0.4705	0.4677	**0.68**
0.32	0.4677	0.4649	0.4621	0.4593	0.4565	0.4538	0.4510	0.4482	0.4454	0.4427	0.4399	**0.67**
0.33	0.4399	0.4372	0.4344	0.4316	0.4289	0.4261	0.4234	0.4207	0.4179	0.4152	0.4125	**0.66**
0.34	0.4125	0.4097	0.4070	0.4043	0.4016	0.3989	0.3961	0.3934	0.3907	0.3880	0.3853	**0.65**
0.35	0.3853	0.3826	0.3799	0.3772	0.3745	0.3719	0.3692	0.3665	0.3638	0.3611	0.3585	**0.64**
0.36	0.3585	0.3558	0.3531	0.3505	0.3478	0.3451	**0.3425**	0.3398	0.3372	0.3345	0.3319	**0.63**
0.37	0.3319	0.3292	0.3266	0.3239	0.3213	0.3186	0.3160	0.3134	0.3107	0.3081	0.3055	**0.62**
0.38	0.3055	0.3029	0.3002	0.2976	0.2950	0.2924	0.2898	0.2871	0.2845	0.2819	0.2793	**0.61**
0.39	0.2793	0.2767	0.2741	0.2715	0.2689	0.2663	0.2637	0.2611	0.2585	0.2559	0.2533	**0.60**
0.40	0.2533	0.2508	**0.2482**	0.2456	0.2430	0.2404	0.2378	0.2353	0.2327	0.2301	0.2275	**0.59**
0.41	0.2275	0.2250	0.2224	0.2198	0.2173	0.2147	0.2121	0.2096	0.2070	0.2045	0.2019	**0.58**
0.42	0.2019	0.1993	0.1968	0.1942	0.1917	0.1891	0.1866	0.1840	0.1815	0.1789	0.1764	**0.57**
0.43	0.1764	0.1738	0.1713	0.1687	0.1662	0.1637	0.1611	0.1586	0.1560	0.1535	0.1510	**0.56**
0.44	0.1510	0.1484	0.1459	0.1434	0.1408	0.1383	0.1358	0.1332	0.1307	0.1282	0.1257	**0.55**
0.45	0.1257	0.1231	0.1206	0.1181	0.1156	0.1130	0.1105	0.1080	0.1055	0.1030	0.1004	**0.54**
0.46	0.1004	0.0979	0.0954	0.0929	0.0904	0.0878	0.0853	0.0828	0.0803	0.0778	0.0753	**0.53**
0.47	0.0753	0.0728	0.0702	0.0677	0.0652	0.0627	0.0602	0.0577	0.0552	0.0527	0.0502	**0.52**
0.48	0.0502	0.0476	0.0451	0.0426	0.0401	0.0376	0.0351	0.0326	0.0301	0.0276	0.0251	**0.51**
0.49	0.0251	0.0226	0.0201	0.0175	0.0150	0.0125	0.0100	0.0075	0.0050	0.0025	0.0000	**0.50**
	0.01	**0.009**	**0.008**	**0.007**	**0.006**	**0.005**	**0.004**	**0.003**	**0.002**	**0.001**	**0**	**P**

Exemple : $\Pi(u) = P(U \leq u) = P = 0.6340 \Rightarrow u = 0.3425$;
$\Pi(u) = P(U \leq u) = P = 0.4020 \Rightarrow u = -0.2482$

Fractiles de la loi du χ^2_ν

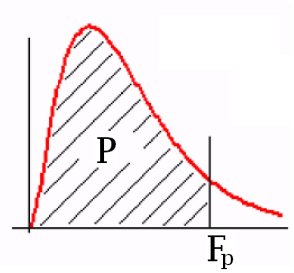
Pour $S \sim \chi^2_\nu$ à ν degrés de liberté le fractile χ^2_p d'ordre P est tel que :

$$P(X \leq \chi^2_p) = p$$

La table donne les fractiles χ^2_p, en fonction de ν, pour certaines valeurs de P.
Pour les valeurs de ν ne figurant pas dans la table, on pourra procéder par interpolation.
Par exemple, pour $\nu = 10$ et $P = 0,975$, on lit $\chi^2_p = 20,5$
et pour $P = 0,025$, on lit $\chi^2_p = 3,25$.
Pour $\nu = 75$ et $P = 0,975$, on lit $\chi^2_p = \frac{1}{2}(95,0 + 106,6) = 100,8$.

Fractiles de la loi de χ^2

Cette table donne les fractiles F_P de la loi de khi-deux à ν degrés de liberté : $P = P(\chi^2_\nu \leq F_P)$



ν P	0.010	0.020	0.025	0.050	0.100	0.150	0.200	0.800	0.900	0.950	0.975	0.980	0.990
1	0.000	0.001	0.001	0.004	0.016	0.036	0.064	1.642	2.706	3.841	5.024	5.412	6.64
2	0.020	0.040	0.051	0.103	0.211	0.325	0.446	3.219	4.605	5.991	7.378	7.824	9.21
3	0.115	0.185	0.216	0.352	0.584	0.798	1.005	4.642	6.251	7.815	9.348	9.837	11.35
4	0.297	0.429	0.484	0.711	1.064	1.366	1.649	5.989	7.779	9.488	11.143	11.668	13.28
5	0.554	0.752	0.831	1.145	1.610	1.994	2.343	7.289	9.236	11.070	12.833	13.388	15.09
6	0.872	1.134	1.237	1.635	2.204	2.661	3.070	8.558	10.645	12.592	14.449	15.033	16.81
7	1.239	1.564	1.690	2.167	2.833	3.358	3.822	9.803	12.017	14.067	16.013	16.622	18.48
8	1.646	2.032	2.180	2.733	3.490	4.078	4.594	11.030	13.362	15.507	17.535	18.168	20.09
9	2.088	2.532	2.700	3.325	4.168	4.817	5.380	12.242	14.684	16.919	19.023	19.679	21.67
10	2.558	3.059	3.247	3.940	4.865	5.570	6.179	13.442	15.987	**18.307**	20.483	21.161	23.21
11	3.053	3.609	3.816	4.575	5.578	6.336	6.989	14.631	17.275	19.675	21.920	22.618	24.73
12	3.571	4.178	4.404	5.226	6.304	7.114	7.807	15.812	18.549	21.026	23.337	24.054	26.22
13	4.107	4.765	5.009	5.892	7.042	7.901	8.634	16.985	19.812	22.362	24.736	25.472	27.69
14	4.660	5.368	5.629	6.571	7.790	8.696	9.467	18.151	21.064	23.685	26.119	26.873	29.14
15	5.229	5.985	6.262	7.261	8.547	9.499	10.307	19.311	22.307	24.996	27.488	28.259	30.58
16	5.812	6.614	6.908	7.962	9.312	10.309	11.152	20.465	23.542	26.296	28.845	29.633	32.00
17	6.408	7.255	7.564	8.672	10.085	11.125	12.002	21.615	24.769	27.587	30.191	30.995	33.41
18	7.015	7.906	8.231	9.390	10.865	11.946	12.857	22.760	25.989	28.869	31.526	32.346	34.81
19	7.633	8.567	8.907	10.117	11.651	12.773	13.716	23.900	27.204	30.144	32.852	33.687	36.19
20	8.260	9.237	9.591	10.851	12.443	13.604	14.578	25.038	28.412	31.410	34.170	35.020	37.57
21	8.897	9.915	10.283	11.591	13.240	14.439	15.445	26.171	29.615	32.671	35.479	36.343	38.93
22	9.542	10.600	10.982	12.338	14.041	15.279	16.314	27.301	30.813	33.924	36.781	37.659	40.29
23	10.196	11.293	11.689	13.091	14.848	16.122	17.187	28.429	32.007	35.172	38.076	38.968	41.64
24	10.856	11.992	12.401	13.848	15.659	16.969	18.062	29.553	33.196	36.415	39.364	40.270	42.98
25	11.524	12.697	13.120	14.611	16.473	17.818	18.940	30.675	34.382	37.652	40.646	41.566	44.31
26	12.198	13.409	13.844	15.379	17.292	18.671	19.820	31.795	35.563	38.885	41.923	42.856	45.64
27	12.879	14.125	14.573	16.151	18.114	19.527	20.703	32.912	36.741	40.113	43.195	44.140	46.96
28	13.565	14.847	15.308	16.928	18.939	20.386	21.588	34.027	37.916	41.337	44.461	45.419	48.28
29	14.256	15.574	16.047	17.708	19.768	21.247	22.475	35.139	39.087	42.557	45.722	46.693	49.59
30	14.953	16.306	16.791	18.493	20.599	22.110	23.364	36.250	40.256	43.773	46.979	47.962	50.89
40	22.164	23.838	24.433	26.509	29.051	30.856	32.345	47.269	51.805	55.758	59.342	60.436	63.69
50	29.707	31.664	32.357	34.764	37.689	39.754	41.449	58.164	63.167	67.505	71.420	72.613	76.15
60	37.485	39.699	40.482	43.188	46.459	48.759	50.641	68.972	74.397	79.082	83.298	84.580	88.38
70	45.442	47.893	48.758	51.739	55.329	57.844	59.898	79.715	85.527	90.531	95.023	96.388	100.42
80	53.540	56.213	57.153	60.391	64.278	66.994	69.207	90.405	96.578	101.88	106.63	108.07	112.33

Exemple : $\nu = 10 d.d.l.$ $P = P(\chi^2_{10} \leq F_P) = 0.95 \Rightarrow F_P = 18.307$
Approximation : Pour $\nu > 100 d.l.l.$ $\chi^2(\nu) \approx \mathcal{N}(\nu; \sqrt{2\nu})$ ou $\sqrt{2\chi^2} - \sqrt{2\nu - 1} \approx \mathcal{N}(0, 1)$

Table de la loi de Student

Soit une v.a. T ayant une densité de Student à ν degrés de liberté.
Le fractile t_p d'ordre P est tel que :

$$P(T \leq t_p) = \int_{-\infty}^{t_p} f(t)f(t)dt = P$$

Pour les valeurs de $P \leq 0,40$ on a $t_p = -t_{1-p}$.

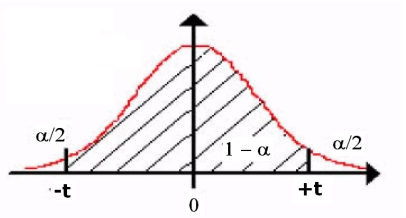
Pour les valeurs de ν ne figurant pas dans la table, on pourra :
- procéder par interpolation - utiliser l'approximation par la loi normale réduite ($\nu > 100$).

Par exemple, pour $\nu = 9$ et $P = 0,975$, on lit $t_p = 2,262$
et pour $P = 0,025$, on déduit $t_p = -2,262$.
Pour $\nu = 75$ et $P = 0,975$, on lit $t_p = \frac{1}{2}(1,994 + 1,990) = 1,992$.

Table de la loi de Student

Cette table donne les fractiles de la loi de Student à ν degrés de liberté : valeur t ayant la probabilité α d'être dépassée en valeur absolue :

$$P(|T_\nu| \leq t) = P(-t \leq T_\nu \leq t) = 1 - \alpha$$
$$P(|T_\nu| > t) = 1 - P(|T_\nu| \leq t) = \alpha$$



ν α	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.005	0.001
1	0.1584	0.3249	0.5095	0.7265	1	1.3764	1.9626	3.0777	6.3137	12.706	31.821	63.656	127.32	636.58
2	0.1421	0.2887	0.4447	0.6172	0.8165	1.0607	1.3862	1.8856	2.92	4.3027	6.9645	9.925	14.089	31.6
3	0.1366	0.2767	0.4242	0.5844	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8408	7.4532	12.924
4	0.1338	0.2707	0.4142	0.5686	0.7407	0.941	1.1896	1.5332	2.1318	2.7765	3.7469	4.6041	5.5975	8.6101
5	0.1322	0.2672	0.4082	0.5594	0.7267	0.9195	1.1558	1.4759	2.015	2.5706	3.3649	4.0321	4.7733	6.8685
6	0.1311	0.2648	0.4043	0.5534	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074	4.3168	5.9587
7	0.1303	0.2632	0.4015	0.5491	0.7111	0.896	1.1192	1.4149	1.8946	2.3646	2.9979	3.4995	4.0294	5.4081
8	0.1297	0.2619	0.3995	0.5459	0.7064	0.8889	1.1081	1.3968	1.8595	2.306	2.8965	3.3554	3.8325	5.0414
9	0.1293	0.261	0.3979	0.5435	0.7027	0.8834	1.0997	1.383	1.8331	2.2622	2.8214	3.2498	3.6896	4.7809
10	0.1289	0.2602	0.3966	0.5415	0.6998	0.8791	1.0931	1.3722	**1.8125**	**2.2281**	2.7638	3.1693	3.5814	4.5868
11	0.1286	0.2596	0.3956	0.5399	0.6974	0.8755	1.0877	1.3634	1.7959	2.201	2.7181	3.1058	3.4966	4.4369
12	0.1283	0.259	0.3947	0.5386	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.681	3.0545	3.4284	4.3178
13	0.1281	0.2586	0.394	0.5375	0.6938	0.8702	1.0795	1.3502	1.7709	2.1604	2.6503	3.0123	3.3725	4.2209
14	0.128	0.2582	0.3933	0.5366	0.6924	0.8681	1.0763	1.345	1.7613	2.1448	2.6245	2.9768	3.3257	4.1403
15	0.1278	0.2579	0.3928	0.5357	0.6912	0.8662	1.0735	1.3406	1.7531	2.1315	2.6025	2.9467	3.286	4.0728
16	0.1277	0.2576	0.3923	0.535	0.6901	0.8647	1.0711	1.3368	1.7459	2.1199	2.5835	2.9208	3.252	4.0149
17	0.1276	0.2573	0.3919	0.5344	0.6892	0.8633	1.069	1.3334	1.7396	2.1098	2.5669	2.8982	3.2224	3.9651
18	0.1274	0.2571	0.3915	0.5338	0.6884	0.862	1.0672	1.3304	1.7341	2.1009	2.5524	2.8784	3.1966	3.9217
19	0.1274	0.2569	0.3912	0.5333	0.6876	0.861	1.0655	1.3277	1.7291	2.093	2.5395	2.8609	3.1737	3.8833
20	0.1273	0.2567	0.3909	0.5329	0.687	0.86	1.064	1.3253	1.7247	2.086	2.528	2.8453	3.1534	3.8496
21	0.1272	0.2566	0.3906	0.5325	0.6864	0.8591	1.0627	1.3232	1.7207	2.0796	2.5176	2.8314	3.1352	3.8193
22	0.1271	0.2564	0.3904	0.5321	0.6858	0.8583	1.0614	1.3212	1.7171	2.0739	2.5083	2.8188	3.1188	3.7922
23	0.1271	0.2563	0.3902	0.5317	0.6853	0.8575	1.0603	1.3195	1.7139	2.0687	2.4999	2.8073	3.104	3.7676
24	0.127	0.2562	0.39	0.5314	0.6848	0.8569	1.0593	1.3178	1.7109	2.0639	2.4922	2.797	3.0905	3.7454
25	0.1269	0.2561	0.3898	0.5312	0.6844	0.8562	1.0584	1.3163	1.7081	2.0595	2.4851	2.7874	3.0782	3.7251
26	0.1269	0.256	0.3896	0.5309	0.684	0.8557	1.0575	1.315	1.7056	2.0555	2.4786	2.7787	3.0669	3.7067
27	0.1268	0.2559	0.3894	0.5306	0.6837	0.8551	1.0567	1.3137	1.7033	2.0518	2.4727	2.7707	3.0565	3.6895
28	0.1268	0.2558	0.3893	0.5304	0.6834	0.8546	1.056	1.3125	1.7011	2.0484	2.4671	2.7633	3.047	3.6739
29	0.1268	0.2557	0.3892	0.5302	0.683	0.8542	1.0553	1.3114	1.6991	2.0452	2.462	2.7564	3.038	3.6595
30	0.1267	0.2556	0.389	0.53	0.6828	0.8538	1.0547	1.3104	1.6973	2.0423	2.4573	2.75	3.0298	3.646
50	0.1263	0.2547	0.3875	0.5278	0.6794	0.8489	1.0473	1.2987	1.6759	2.0086	2.4033	2.6778	2.937	3.496
60	0.1262	0.2545	0.3872	0.5272	0.6786	0.8477	1.0455	1.2958	1.6706	2.0003	2.3901	2.6603	2.9146	3.4602
70	0.1261	0.2543	0.3869	0.5268	0.678	0.8468	1.0442	1.2938	1.6669	1.9944	2.3808	2.6479	2.8987	3.435
80	0.1261	0.2542	0.3867	0.5265	0.6776	0.8461	1.0432	1.2922	1.6641	1.9901	2.3739	2.6387	2.887	3.4164
infini (loi normale)	0.1257	0.2533	0.3853	0.5244	0.6744	0.8416	1.0364	1.2816	1.6449	1.96	2.3264	2.5759	2.8072	3.2908

Exemple : $\nu = 10 d.d.l.$ $P = P(|T_{10}| \leq t) = 0.95 \Rightarrow t = \pm 2.2281$
$P = P(T_{10} \leq t) = 0.95 \Rightarrow t = +1.8125$

Le manuel "Statistique descriptive" offre le matériel théorique et pratique, nécessaire à apprendre d'après le programme en "Bases de la statistique – IIe partie» des spécialités Gestion et Economie de la filière de gestion à l'Université de Sofia « Sv. Kliment Ohridski ». On y donne des notions de base, liées à la statistique descriptive. Les thèmes considérés sont : Série statistique uni- et bivariées, Paramètres de position, de dispersion, de forme et de concentration, Indice de Gini, Tableau de contingence, Corrélation linéaire, Ajustement linéaire ; Séries chronologiques, Prévision, Moyennes mobiles ; Echantillonnage, Distribution de la moyenne, de la dispersion et de la fréquence échantillonnalles.

Le manuel a pour but de donner des connaissances théoriques et de développer des compétences pratiques pour le travail avec la terminologie et les méthodes de base de la statistique contemporaine comme un utile de planification, analyse, représentation et interprétation d'études expérimentales et non expérimentales.

eISSN 2367-8666