Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

ECOLE NATIONALE SUPERIEURE DE MANAGEMENT "ENSM"

Pôle universitaire de Koléâ, Tipaza

Polycopie de cours

STATISTIQUE ET ANALYSE DE DONNEES

Réalisée par : Dr. **Messaoud ZEROUTI**

Année Universitaire : 2018/2019

SOMMAIRE

INTRODUCTION GENERALE

Ce support de cours "Statistique et Analyse de données" constitue une introduction à l'analyse de données statistique en sciences sociales et il est destiné aux étudiants de première année master de l'ENSM ; Spécialité "Management Marketing".

La statistique concerne pratiquement tous les domaines d'application, aucun n'en est exclu ; elle permet de faire une exploration et une analyse d'un vaste ensemble de données.

Cette omniprésence s'accompagne bien souvent de l'absence de regard critique tant sur l'origine des données que sur la manière de les traiter. La facilité d'utilisation des logiciels de traitement statistique permet de fournir quasi instantanément des graphiques et des résultats numériques, d'où l'importance de s'intéresser aux fondements des méthodes d'analyse de données ainsi que leurs périmètres d'utilisation.

Le polycopie est structuré en huit chapitres où on présente les méthodes essentielles de la statistique et de l'Analyse de données (univariées, bivariées et multivariées) et explique comment les appliquer à des problèmes concrets des sciences sociales. Il expose de façon claire et pédagogique toutes les notions importantes relatives à chaque méthode. En plus, de nombreux exemples issus de champs d'application variés sont traités ainsi que des exercices d'application pour entrainement sont proposés à la fin de chaque chapitre.

Par ailleurs, le choix de commencer par un chapitre portant sur les techniques d'échantillonnage est justifié par l'hétérogénéité des étudiants que nous recevant l'ENSM[1], c'est-à-dire des étudiants venant de spécialités différentes (gestion, économie, informatique, droit, langue…). Ensuite, la première partie de ce document a pour but d'initier les étudiants aux méthodes d'analyse de données univariées (représentation tabulaire et graphique d'une série statistique). La deuxième partie, vise à développer les méthodes d'analyse de données bivariées (commençant par les tableaux croisés jusqu'à la régression simple en passant par les tests statistique). Enfin, dans la troisième partie nous allons focaliser nos propos sur l'une des méthodes de base de l'analyse de données multivariées qui est la méthode d'Analyse en Composante Principale « ACP ».

[1] L'acronyme ENSM désigné : Ecole Nationale Supérieure de Management et son slogan est l'école de la deuxième compétence.

CHAPITRE 1 : APERÇU SUR LES TECHNIQUES D'ECHANTILLONNAGE.

1-1- Quelques concepts de bases

1-2- Les caractéristiques d'un échantillon

1-3- Les méthodes d'échantillonnage

1-4- Choix d'une méthode d'échantillon.

1-5- La taille de l'échantillon.

1-1- Introduction :

L'échantillonnage permet aux statisticiens ou aux chargés d'étude de tirer des conclusions au sujet d'un tout, en n'en examinant qu'une partie. Les chercheurs ne s'intéressent pas à l'échantillon lui-même, mais à ce qu'il est possible d'apprendre à partir de l'enquête et à la façon dont on peut appliquer cette information à l'ensemble de la population.

1-2- Quelques concepts de bases :

- ✓ On appelle **enquête** l'ensemble des opérations qui ont pour but de collecter de façon organisée des informations relatives à un groupe d'individus, d'éléments ou d'unités statistiques.
- ✓ Les **individus** ou les éléments en question, appelées aussi **unité statistique** peuvent être aussi bien des personnes humaines que des animaux, des familles, machines, entreprises, etc…
- ✓ L'ensemble des individus auxquelles on s'intéresse est appelé **population** ou **univers**.
- ✓ Lorsque toutes les unités de la population sont observées individuellement, l'enquête est appelée **recensement**.
- ✓ La partie de la population qui est réellement observée constitue **l'échantillon,** et l'opération de choix de cette population est précisément l'opération **d'échantillonnage**.
- ✓ Une **base d'échantillonnage** ou une **Base de sondage** est une liste des unités d'échantillonnage possibles.

1-3- Les caractéristiques d'un échantillon :

L'échantillon ou la population (cible) peut être différente selon le problème étudié. Elle est caractérisée par les critères suivants :

a- L'objectif de l'enquête : la population peut être définie dans l'objet même de l'enquête.

Exemple : Enquête auprès des consommateurs sur le prix du Yaourt.

b- Définir la population cible : C'est la population totale pour laquelle on a besoin de l'information. Il faut définir les unités qui composent la population sous forme de caractéristiques l'identifiant (nature des données, emplacement géographique, dates ou encore critères sociodémographiques…). Cependant, la *population observée* est différente de la

population cible. En effet, la population cible est la population que nous **voulons observer**, tandis que la population observée est la population que nous **pouvons observer.** Et les conclusions ne s'appliqueront qu'à la population réellement observée (cible). L'utilisateur des résultats doit en être informé.

c- Le type d'échantillonnage adopté : la nature des documents disponibles pour construire l'échantillon peut amener à restreindre la population à interroger.
Exemple : si on analyse les listes électorales ; l'échantillon est limité aux personnes dont l'âge dépasse 18 ans (âge de voter).

d- Les contraintes matérielles imposées : Pour des raisons budgétaires ou de délais d'exécution, la population peut être restreinte à une portion plus accessible.

Exemple : Dans une enquête sur l'opinion des jeunes à l'égard du code de la route, on limitera la population aux jeunes qui possèdent un permis de conduire.

1-4- Les méthodes d'échantillonnage :

A l'exception des recensements qui se réalisent auprès de l'ensemble des individus composant la population, toutes les autres enquêtes se limitent à interroger un échantillon (une partie) de la population.
L'échantillon doit être une partie représentative de la population et refléter, dans sa composition, la diversité réelle de la population.

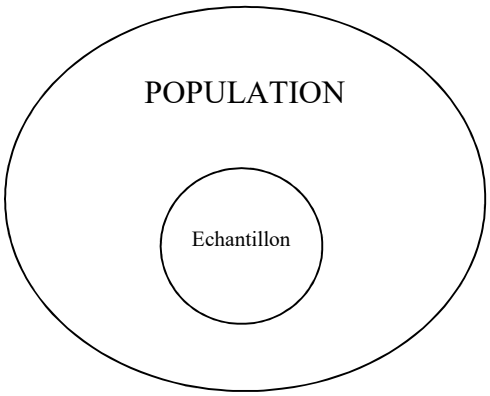


Figure : Echantillon-Population

Toute la question est de savoir si cet échantillon est représentatif ou pas de cette population. Si ce n'est pas le cas, l'enquête réalisée ne sera pas représentative et ne présentera aucune validité externe.

Pour cela ; on fait appel aux méthodes d'échantillonnage pour sélectionner les individus représentatifs d'une population. Il existe principalement les *méthodes aléatoires ou probabilistes* et les *méthodes empiriques.* La différence entre les deux tient au fait que dans le cas de l'échantillonnage probabiliste chaque unité a une « chance » d'être sélectionnée et que cette chance peut être quantifiée, ce qui n'est pas vrai pour l'échantillonnage non probabiliste ; dans ce cas, chaque unité incluse à l'intérieur d'une population n'a pas une chance égale d'être sélectionnée.

1-4-1 Méthodes aléatoires ou probabilistes :

Tous les membres de l'univers (population) ont une chance égale de faire partie de l'échantillon. Plusieurs méthodes existent :

a. Echantillon au hasard simple ou "Sondage aléatoire Simple" :

Dans un échantillon aléatoire simple, chaque élément de la population a la même chance d'être choisi. En outre, chaque échantillon possible d'une traille « n » a la même probabilité d'être sélectionné. Cela implique que chaque élément est choisi indépendamment de tout autre élément. Les unités qui constituent l'échantillon sont désignées par tirage au hasard.

Le chargé d'étude doit dresser une liste de toutes les unités incluses dans la population observée (base de sondage ou d'échantillonnage) pour sélectionner un échantillon aléatoire simple. Un échantillonnage aléatoire simple peut s'effectuer avec remise (tirage non exhaustif) ou sans remise (tirage exhaustif).

Avantages de cette méthode :

- On peut espérer à un échantillon « représentatif » puisque la méthode donne à chaque individu de la population une chance égale d'être sélectionné ;
- Facile à mettre en œuvre.

Cependant, cette méthode souffre de certains inconvénients, en particulier :

- La difficulté dans la construction de la base de sondage apte à constituer un échantillon aléatoire simple ;

- Augmentation du temps et le coût de collecte des données, si la population est très dispersée géographiquement, de ce fait les frais de déplacement des enquêteurs seront élevés.
- Elle ne permet pas toujours d'obtenir un échantillon représentatif. Bien qu'en moyenne, les échantillons constitués représentent bien la population, il arrive qu'un échantillon aléatoire simple donné présente une image grossièrement déformée de la population ciblée.

b. Echantillonnage systématique :

L'échantillonnage systématique est une méthode qui exige aussi l'existence d'une liste de la population où chaque individu est numéroté de 1 jusqu'à N.

Notons n, le nombre d'individus que doit comporter l'échantillon (la taille de l'échantillon). L'entier voisin de N/n sera noté « r » et appelé « raison de sondage » ou « pas de sondage ».

Choisir au hasard un entier naturel d entre 1 et r (cet entier sera le point de départ).

L'individu dont le numéro correspond à d est le premier individu, pour sélectionner les autres, il suffit d'ajouter à d la raison de sondage : les individus choisis seront alors ceux dont les numéros correspondent à : d + r ; d + 2r ; d + 3r etc.

Avantages : facile à sélectionner parce qu'un seul individu est choisi au hasard. On peut obtenir une bonne précision parce que la méthode permet de répartir l'échantillon dans l'ensemble de la liste.

Inconvénients : dans le cas où le pas de sondage égale à la périodicité des données. Les données peuvent être biaisées à cause de la périodicité.

Exemple : Étude des déplacements en autobus sur 365 jours en prenant un échantillon de taille 60.
(N=365 jours et n=60).
Dans ce cas : d= 365/60= 7.
Donc, le jour sélectionné au départ sera le même pour tout l'échantillon.

Exemple 2 : On a une population de 400 individus, on veut un échantillon de 100 individus
R = 4

On a donc 4 échantillons possibles :

1, 5, 9, …. 397

2, 6, 10, … 398

3, 7, 11, ….399

4, 8, 12, … 400

c. Echantillonnage stratifié :

Une strate est un groupement homogène d'unités statistiques ; triées de la population totale et reliées entre elle par un caractère lié à l'objet de l'enquête.

L'univers statistique est découpé en sous-ensembles homogènes. La stratification se fait selon les problèmes étudiés, par exemple :

- Classement des entreprises suivant le personnel.
- Classement des ménages suivant leurs revenus ou le nombre de personne qui les composent.
- Classement de la population suivant les communes.

Soit N : la taille de la population

Ni : la taille de la strate «i», (i allant de 1 à k).

K : le nombre de strates

ni : l'échantillon lié à la strate :

ni %= $\frac{Ni}{N}$ 100

Exemple : Une entreprise possède 3000 clients. Ils sont repartis en 2 strates : 2100 femmes et 900 hommes. On vérifie que l'échantillon total est reparti proportionnellement à l'importance de chaque strate dans la population.

Tableau : Répartition des clients d'une entreprise selon le genre

	Population Ni	taille de l'Echantillon (ni %)
Strate 1 (femmes)	2100	70
Strate 2 (hommes)	900	30
Total	3000	100

Pour un échantillon de taille 300, il doit être composé de 140 femmes et de 60 hommes.

Avantages

Il est peu probable de choisir un échantillon absurde puisqu'on s'assure de la présence proportionnelle de tous les divers sous-groupes composant la population.

Inconvénients

La méthode suppose l'existence d'une liste de la population (base de sondage). Il faut aussi connaître comment cette population se répartit selon certaines strates, c'est-à-dire les critères à adopter pour constituer les strates.

c. Echantillon en grappes :

L'échantillonnage en grappe limite la portée de la construction de la base de l'échantillon et des travaux de terrain connexes à un ensemble ou un échantillon, surtout quand la population est dispersée géographiquement. Cette méthode, simple et économique, permet de choisir un échantillon en plusieurs étapes.

Tout d'abord, on divise la population en sous-groupes (grappes) mutuellement exclusifs (chaque individu n'appartient qu'à une seule grappe) et collectivement exhaustif (la somme des éléments de toutes les grappes donnera la population).

Ensuite, choisir au hasard un échantillon en grappes en s'appuyant sur une des techniques d'échantillonnage probabiliste comme l'échantillonnage aléatoire simple. Pour chaque grappe sélectionnée, soit on inclut tous les éléments dans l'échantillon, il s'agit dans ce cas d'un échantillonnage en grappes *simple*, soit on extrait un échantillon d'éléments de façon probabiliste, la procédure est dite *d'échantillonnage à deux degrés.*

Avantages :

Réduction des coûts de collecte de données.

Inconvénients :

Effet de grappe (variance intra-grappes qui est faible) dû à l'existence de similarité entre individus d'une même grappe.

1-4-2- Méthodes empiriques ou non probabilistes :

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.

Dans le cas de l'échantillonnage probabiliste, chaque unité a une chance d'être sélectionnée. Ce n'est plus le cas dans l'échantillonnage non probabiliste. On se fixe alors

comme règle que l'échantillon retenu doit avoir la même composition que la population mère par rapport à une ou plusieurs caractéristiques.

Elles ne font plus appel au hasard, et elles sont plutôt empiriques. Il existe, en effet plusieurs méthodes :

a. Echantillon par choix raisonné : Cette méthode consiste à procéder par un choix raisonné, c'est à dire, le chargé d'étude choisit les unités les plus représentatives de la population. Plusieurs méthodes sont utilisées :

- **Méthode des quotas** : En pratique, c'est la méthode la plus utilisée. Elle consiste à choisir un ensemble de critères appelés « caractères de contrôle » et construire un modèle réduit de la population totale. Les caractères de contrôle sont, par exemple : revenu, age, sexe, catégorie socioprofessionnelle, Etc…

Les avantages :

- ✓ Elle n'exige pas une énumération complète des unités de l'univers.
- ✓ Les unités à relever étant choisis par les enquêteurs.

Les inconvénients :

- ✓ La méthode ne permet pas le calcul des erreurs, seul les échantillons aléatoires le permettent.
- ✓ Les résultats peuvent être entachés d'erreurs systématiques introduite suite à une faute de jugement des enquêteurs (biais de sélection).
- ✓ La qualité des enquêtes repose sur la qualité du travail des enquêteurs.

- **Méthode de convenance** : Pour faciliter la recherche des personnes à interroger, on peut décider de les interroger la où il y a le plus de chance de trouver les unités directement concernées par le sujet de l'enquête. Par exemple : des personnes interceptées dans la rue, dans des magasins ou dans des centres commerciaux…

Exemple 1 : un journaliste qui réalise un micro-trottoir à l'occasion d'une manifestation.

Exemple 2 : Si un enquêteur doit interroger des acheteurs de matelas, il pourra se placer à la sortie des magasins de literie.

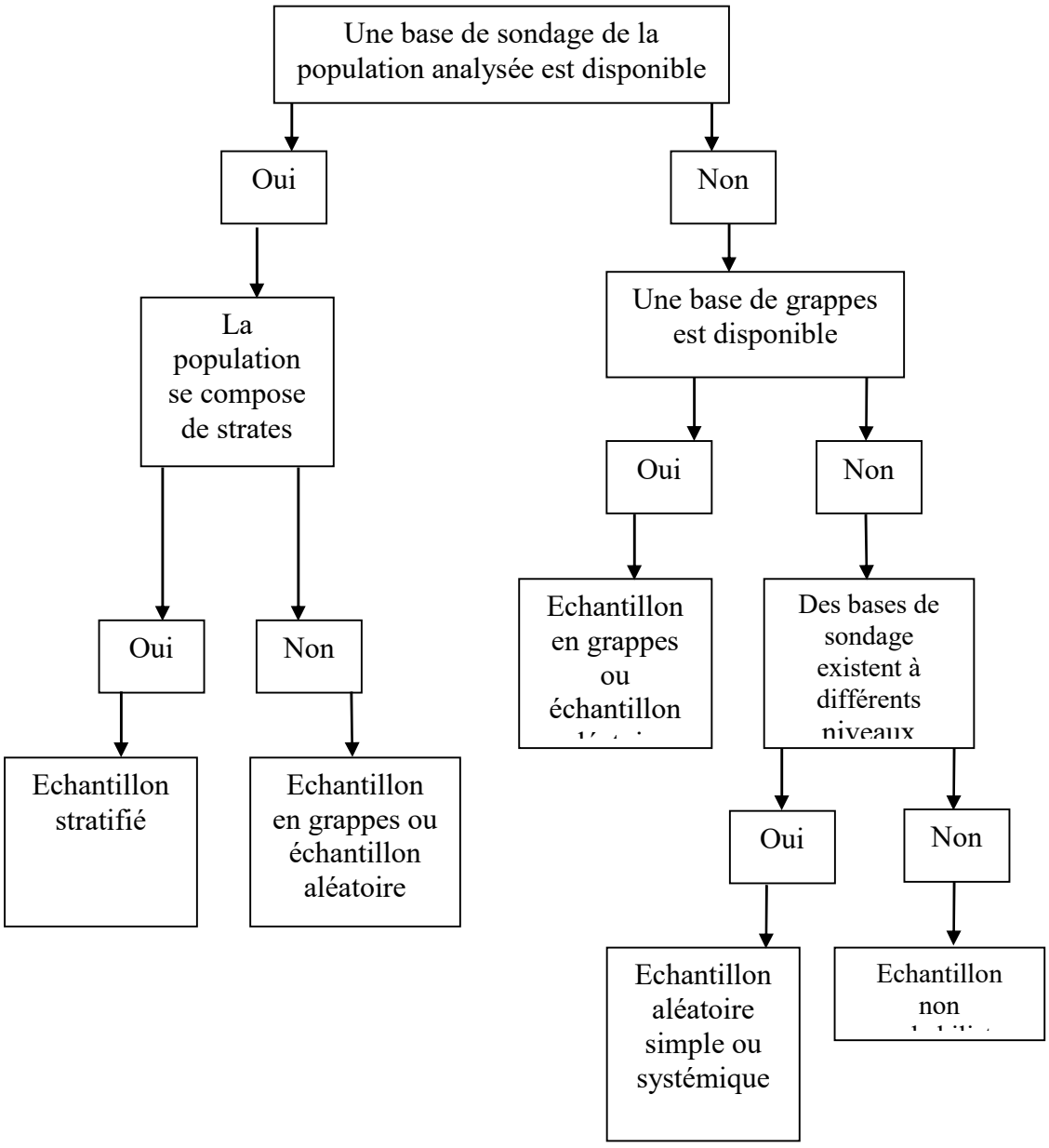
En effet, cette manière de faire n'interroge pas nécessairement la population cible mais aussi des personnes qui ne soit pas concernées par l'étude.

- **Echantillon de jugement** : Ce type d'échantillon se veut représentatif dans la mesure où le chercheur va interroger les individus les plus susceptibles d'éclairer et d'apporter une information pertinente sur le problème à résoudre.

b. Echantillon pseudo-aléatoire : Il s'agit de méthodes qui sans être des méthodes probabilistes s'en rapprochent le plus. Elle consiste à imposer à l'enquêteur un itinéraire en lui indiquant exactement les points du circuit. L'identification de ces points d'enquête résulte de la combinaison de tirage au hasard.

1-5- PROCEDURE DE Choix d'une méthode d'échantillon :

Il serait faux de dire que toutes les méthodes d'échantillonnage se valent. Toutefois, nous sommes souvent restreints à choisir une méthode. Ce choix peut être résumé par la figue suivante :



1-6- La taille de l'échantillon :

Combien de personnes doit on interroger pour que l'échantillon soit représentatif de la population étudiée ? Autrement dit :

1-Est-ce qu'un échantillon de taille 500 suffit pour une population de taille 10 000 ?

2. Quelle est la taille de l'échantillon qui assure tel degré précision ?

Il est impossible de répondre par oui ou par non à ces questions. Un échantillon doit fournir une estimation aussi précise que possible d'une variable.

La taille de l'échantillon dépend généralement des critères suivants :

- Taille de la population d'étude
- Critères statistiques
- Contraintes de terrain observées et des questions auxquelles on désire répondre par le moyen de l'enquête

1-6-1- La taille de la population :

Pour déterminer la taille de l'échantillon il est important de connaitre la taille de population. En effet, plus la taille de la population est grande, plus la taille de l'échantillon sera grande afin de tenir compte des caractéristiques de la population.

1-6-2- Critères statistiques

a- Niveau de précision = erreur d'échantillonnage : Niveau de précision estime l'intervalle de confiance dans lequel on va situer la valeur réelle de la population et il est exprimé en points de pourcentage (ex : +/- 5%).

Si la valeur estimée est un pourcentage alors la valeur réelle est comprise entre « la valeur estimée moins le niveau de précision » et « la valeur estimée plus le niveau de précision ».

Si la valeur estimée est un nombre, la largeur de l'intervalle se calcule en multipliant la valeur estimée par le niveau de précision adopté ; la valeur réelle de la population est « valeur estimée moins largeur de l'intervalle » ; « valeur estimée plus la largeur de l'intervalle ».

Plus le degré de précision est élevé, plus l'intervalle sera étendu et inversement.

b- Niveau de confiance : Il y a toujours un risque que l'échantillon sélectionné ne représente pas la population étudiée. Le niveau de confiance (ou marge d'erreur) permet d'indiquer le pourcentage de chances que l'échantillon sélectionné soit représentatif de la population étudiée.

Exemple : moyenne du revenu des foyers. Les valeurs sont :

« 84;85;84;86;…;87;83;84;84;… ;83;81;87;… ;86;82;79 ;…;85;85;86;85;89;84 »

Moyenne = 85UM

Les valeurs obtenues pour ces échantillons suivent une distribution normale autour de la moyenne réelle. Certaines sont proches de la valeur réelle (ex :84;85;86), d'autres sont plus éloignées (ex :79;81;89…).

On dit alors, 95% des valeurs obtenues gravitent autour de la valeur réelle de la population avec une différence de moins de deux écart-types. Autrement dit, un niveau de confiance de 95% assure que, parmi 100 échantillons tirés aléatoirement, 95 donnent une valeur estimée égale à la valeur réelle de la population totale.

En conclusion, plus le niveau de confiance retenu est fort, moins le risque de tirer un échantillon éloigné de la population étudiée est élevé.

c- Degré de variabilité : Ce critère détermine la ressemblance (degré d'homogénéité) des individus de la population selon leurs caractéristiques communes. En effet, moins les individus d'une population se ressemblent, plus l'échantillon doit être grand pour atteindre un même degré de précision.

Par exemple : Une proportion de 50% indique une plus forte variabilité que 20% ou 80%.

Cette proportion est suspectée, mais rarement quantifiable d'avance, il est donc d'usage d'utiliser la variabilité maximale (P=0.5) pour éviter les risques d'erreurs.

Ainsi lorsque l'on sait que les mesures recueillies seront très proches les unes des autres on veillera à retenir un degré de précision plus fin (ex :+/-3%), notamment lorsque notre étude vise à comparer deux types de populations très homogènes.

En revanche, lorsque l'objectif de l'étude est simplement de décrire les comportements de populations que l'on sait à priori différenciées, le degré de précision sera moins fin (e=+/-14%).

5-3- Calcul de la taille de l'échantillon :

Lorsqu'on se limite aux critères cités précdemment, le calcul de la taille de l'échantillon sera déterminé comme suit :

a- Petite population/Etudes similaires : Lorsque la population est petite (200 individus ou moins), il est préférable de l'enquêter dans sa totalité, car les coûts associés au déploiement de

l'enquête seront les mêmes si l'on enquête 50 ou 200 individus, et enquêter toute la population évite les erreurs d'échantillonnage, plus conséquentes lorsque la population totale est petite.

Lorsque des enquêtes similaires ont été menées sur la même population, il est préférable d'utiliser la même taille d'échantillon qui permettra des comparaisons intéressantes. Cette approche doit être retenue dans le cas où le plan d'échantillonnage est valide et répond aux attentes et conditions de l'étude en cours (même population ciblée, mêmes objectifs que l'enquête précédente, mêmes zones accessibles, pas de déplacement de population de masse…).

b- Tables statistiques : Lorsque les critères de choix sont standards et prédéfinis, on peut se référer aux tables statistiques existantes qui présentent les différentes tailles d'un échantillon aléatoire simple selon la taille de la population et le niveau de précision désiré (pour des niveaux de confiance (95%) et d'hétérogénéité (P=0.5)). A noter que :

- Ces tailles représentent le nombre d'individus effectivement enquêtés. Il est donc important de prévoir un échantillon complémentaire pour pallier aux phénomènes de réponses erronées et de non-réponse.
- La table concernant les populations de petite taille part du postulat que la population suit une distribution normale.

Tableau 1-1 : Différentes tailles d'un échantillon aléatoire simple (pour des niveaux de confiance (95%) et d'hétérogénéité (P=0.5))

Taille de la population	Taille de l'échantillon selon la précision :			
	+/-3%	+/-5%	+/-7%	+/-10%
100	-	81	67	51
125	-	96	78	56
150	-	110	86	61
175	-	122	94	64
200	-	134	101	67
225	-	144	107	70
250	-	154	112	72
275	-	163	117	74
300	-	172	121	76
325	-	180	125	77
350	-	187	129	78
375	-	194	132	80
400	-	201	135	81
425	-	207	138	82
450	-	212	140	82

Taille de la population	Taille de l'échantillon selon la précision :			
	+/-3%	+/-5%	+/-7%	+/-10%
500	all	222	145	83
600	all	240	152	86
700	all	255	158	88
800	all	267	163	89
900	all	277	166	90
1 000	all	286	169	91
2 000	714	333	185	95
3 000	811	353	191	97
4 000	870	364	194	98
5 000	909	370	196	98
6 000	938	375	197	98
7 000	959	378	198	99
8 000	976	381	199	99
9 000	989	383	200	99
10 000	1000	385	200	99
15 000	1034	390	201	99
20 000	1053	392	204	100
25 000	1064	394	204	100
50 000	1087	397	204	100
100 000	1099	398	204	100
<100 000	1111	400	204	100

Par exemple, pour une population d'étude qui compte 4000 unités et les critères retenus sont standards, la population suivant à priori une distribution normale et présentant un degré de variation acceptable (niveau de confiance : 95%, degré de variabilité par défaut (P =0.5), niveau de précision :+/-5%).

Dans le tableau d'échantillonnage, au croisement des valeurs 4000 (ligne) et « +/-5% » (colonne), on obtient la taille requise pour disposer d'un échantillon représentatif de la population (364 unités). Nous devons ajouter 10% en plus pour pallier aux phénomènes de non réponse et réponses erronées, ce qui donne un échantillon de 364+364x10%=400.

c- Formules mathématiques

Soit la formule simplifiée suivante : $n = \frac{N}{1+N\ x\ e^2}$

Avec : N = taille de la population, e = niveau de précision

Application numérique sur l'exemple précédent : n= $\frac{4000}{(1+4000\ \text{x}\ (0.05\text{x}0.05))} = \frac{4000}{11} = 364$

Formule pour les proportions : Lorsque la prévalence estimative de la caractéristique étudiée (degré de variabilité) est connue, on peut calculer la taille de l'échantillon requise en utilisant la formule suivante : $\frac{t^2+px\ q}{e^2} = \frac{t^2+px\ (1-p)}{e^2}$

Avec : e = niveau de précision ; p = degré de variabilité (taux de prévalence estimative) ;

t = valeur type associée au niveau de confiance requis (95% -> 1,96).
Cette formule renvoie une taille d'échantillon plus grande que précédemment, mais elle est utile lorsque nous ne connaissons pas l'effectif de la population totale.

- Formule pour les proportions, populations finies

Si l'on dispose du taux de variabilité, et de la taille de la population, on peut réduire substantiellement la taille minimum de l'échantillon obtenue par la formule précédente :

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$

Avec : N = taille de la population ; no = taille de l'échantillon obtenue par la « formule pour les proportions ».

d- Contraintes de terrain :

Parfois certaines contraintes empêchent de tirer le nombre d'individus optimum estimé par les formules précédentes pour faire partie de l'échantillon.

Les principales contraintes rencontrées lors d'enquêtes de terrain sont :

- Accessibilité de la population à enquêter ;
- Le temps nécessaire pour communiquer les résultats ;
- Le nombre de personnes mobilisés pour l'enquête ;
- Le nombre de questionnaires à remplir par les personnes à interroger…

PARTIE 1 :

STATISTIQUE ET ANALYSE DES DONNEES UNIVARIEES

INTRODUCTION DE LA PARTIE 1 :

La statistique est un ensemble de méthodes et de techniques mathématique permettant de présenter, de décrire et de résumer un vaste ensemble de données relatives à un certains nombres d'objets. Ces derniers peuvent prendre plusieurs formes :

1- Des objets ordinaires : Par exemples : les entreprises cotées en bourse en Algérie ; les établissements d'enseignement supérieurs d'Alger.
2- Des êtres vivants : tels que : les vaches dans la zone des hauts plateaux en 2017 ; les visiteurs du Show-room de marque Toyota durant le salon de l'automobile d'Alger de 2015.
3- Des faits : les naissances enregistrées à l'état civil à la commune d'Alger centre durant la décennie 2005-2015 ; les accidents de la route durant le mois de ramadan de 2017.

Par ailleurs, toute étude statistique se décompose au moins en deux phases :

- Phase de rassemblement ou de collecte des données qui se fait soit par simple observation, ou par expérimentation.
- Phase d'analyse des données et d'interprétation des résultats : elle comporte deux étapes :
  - ✓ Etape descriptive ou exploratoire.
  - ✓ Etape inductive ou inférentielle.

L'objet de cette première partie est présenté les techniques les plus utilisés en sciences sociales pour analyser une seule série statistique. Pour ce faire, nous allons procéder en deux temps. Dans un premier temps, nous allons présenter les concepts de bases ainsi que les présentation (tabulaires ensuite graphique) des séries statistiques. Dans un second temps, nous focalisons nos propos sur les indicateurs de synthèse numérique d'une série statistique (indicateurs de position et de dispersion).

CHAPITRE 2 :
CONCEPTS DE BASES ET PRESENTATION DES SERIES STATISTIQUES

L'objectif de ce chapitre est de présenter les concepts de bases de l'analyse statistique ainsi que la procédure à suivre pour représenter sur tableau ensuite graphiquement une seule série statistiques, et ce afin de faciliter aux chargé d'études et ou manager la lecture et l'analyse des données statistiques.

2-1- LE VOCABULAIRE DE LA STATISTIQUE :

2-1-1- Quelques concepts de la statistique :

a- Population : Il s'agit d'un groupe d'individus ou d'unités statistiques, qui fait l'objet de l'étude statistique.

Exemples :

- Tous les étudiants d'une université ;
- Tous les travailleurs d'une entreprise ;
- L'ensemble des consommateurs des boissons gazeux de marque Hammoud sur le territoire national ;
- Tous les produits fabriqués par une entreprise.

b- Individus : Un individu ou une unité statistique est l'élément de base de la population. La totalité des individus correspond à la population.

Exemples :

- Etudiant "x" d'une université ;
- Employé "Y" d'un organisme ;
- Le produit "P" fabriqué par une entreprise.

c- Caractère : Pour étudier une population, on la divise en sous-ensembles qui seront déterminées par rapport à un ou plusieurs critères : ce sont les caractères statistiques ou variables statistiques.

Par exemple, les travailleurs d'une entreprise peuvent être classés selon les caractères suivants : sexe, âge, résidence, salaires, CSP etc…

d- Modalités : On appelle *modalité* d'un caractère les différentes situations possible (numérique ou pas) que peut prendre un caractère. Les modalités d'un même caractère doivent être incompatibles et exhaustives, de sorte qu'un individu n'appartient qu'à une seule modalité.

Exemples :

- Le sexe (masculin ou féminin) ;
- L'âge des employés d'une entreprise (de 20 à 30 ans ; de 30 à 40 ans ….).

2-1-2- Les différents types de caractères : La littérature distingue deux types de caractères : Qualitatif et Quantitatif.

a- Caractère Qualitatif : Un caractère est qualitatif si ses modalités échappent à la mesure ou au comptage.

Exemples : le sexe, la profession, la résidence, etc….

On distingue deux types de caractères qualitatifs : Nominale et ordinale[1].

- Caractère qualitative nominale : est un caractère dont les modalités sont des noms, par conséquent ces modalités ne peuvent pas être ordonnées ni additionnées.
  Tel est le cas de : la Catégorie socioprofessionnelle, le sexe, le lieu de résidence, la nationalité.
- Caractère qualitative ordinale[2] : Est un caractère qui permet d'apprécier le degré d'appartenance ou non à une catégorie donnée, de plus haut jusqu'au plus bas, de plus important au plus faible.
  On peut donner comme exemples : niveau de satisfaction, niveau d'accord ou de désaccord… Les modalités de ce caractère peuvent être codifier par des échelles allant par exemple de : 1 à 3, 1 à 5, 1 à 7.

b- Caractère quantitatif : Ses modalités sont mesurables. Les nombres correspondant à la mesure d'un individu sont appelés modalités de la variable quantitative.

On distingue deux types de variables quantitatives : discrète et continue.

- **Variable quantitative discrète :** Elle prend des valeurs isolées dans son domaine de variations.

Exemples :

- Le nombre de voitures vendues en 2004 de marque Mercedes en Algérie.
- Le nombre de vols de la compagnie Air Algérie enregistré durant le mois de Mai 2012.

[1] Une variable qualitative est dite dichotomique si elle ne peut prendre que deux modalités.

[2] Un caractère de type quantitatif peut être converti en un caractère qualitatif ordinale en classant les individus par rapport à ce caractère quantitatif du plus bas jusqu'au plus élevé ou l'inverse.

- Le nombre d'articles vendus dans un magasin en fin de journées.

- **Variable quantitative continue :** Le nombre de valeurs possibles au sein de chaque classe est infini ; il est donc nécessaire de définir les modalités en groupant en classes ces valeurs.

Exemples :

- La taille, le poids, ou l'âge d'un individu.
- La distance entre deux villes.
- Les salaires des travailleurs ou le nombre d'années d'expérience.

La longueur de chaque classe est appelée "**amplitude**". Ces dernières peuvent être constantes ou variables.

Exemples :

- *Amplitude constante :* Pour la variable Age des employés d'une entreprise, nous allons proposer : 15 ans à 20 ans ; 20 ans à 25 ans ; 25 à 30 ; 30 à 35 ans, ….
- *Amplitude variable :* Durée du chômage : moins de 3 mois, entre 3 mois et 1 an, et plus d'une année[3].

On évite, en général, de constituer plus d'une dizaine de classes afin de faciliter les analyses et les interprétations des résultats.

Le gestionnaire est souvent confronté à une masse de données qu'il peut être utile de chercher à interpréter. On recourt alors à un certain nombre méthodes pour classer au mieux ces informations. On peut en faire une synthèse tabulaire, graphique[4].

2-2- PRESENTATIONS TABULAIRES DES STATISTIQUES :

2-2-1- La structure et les étapes de construction d'un tableau statistique :

a- La structure d'un tableau statistique :

Une population de « N » individus étudiée par rapport à un caractère déterminé donne lieu à un tableau statistique.

Un seul caractère donne un tableau à une dimension (colonne) où chaque case est une modalité du caractère. Si on a deux caractères, on aura un tableau à deux dimensions (deux colonnes), et ainsi de suite.

[3] L'utilisation des amplitudes variables est justifiée par le faite que la dispersion du caractère étudié est très élevée. Tel est le cas de l'exemple cité plus haut où plus la durée de chômage augmente ; le nombre de personnes concernées sera de plus en plus faible.

[4] Nous aborderons dans le prochain chapitre les synthèses numériques pour une variables quantitative.

Exemple : avec les caractères sexe et état matrimonial, on peut obtenir les deux tableaux suivants (à deux dimensions chacun) :

Tableau2-1 :

Représentation tabulaire de la distribution d'une population par rapport à un caractère

Individus	Sexe
1	Masculin
2	Féminin
3	Féminin
4	Masculin
5	Masculin
. . . .	. . . .
N	Féminin

Tableau 2-2 :

Représentation tabulaire de la distribution d'une population par rapport à deux caractères[5].

Individus	**Sexe**	**Etat matrimonial**
1	Masculin	Marié
2	Féminin	Célibataire
3	Féminin	Divorce
4	Masculin	Veuf
5	Masculin	Célibataire
. . . .	. . . .	. . . .
N	Féminin	Mariée

Si on a plusieurs caractères, le nombre de sous-ensembles incompatible et exhaustifs, c'est-à-dire, le nombre de cases est égal au produit des modalités des différents caractères.

b- Les étapes de construction d'un tableau statistique :

Pour construire un tableau, on doit déterminer les informations suivantes :

- **Modalité :** Il s'agit de l'identification de l'ensemble des modalités que peuvent prendre l'ensemble des individus par rapport au caractère étudié.

[5] Nous reviendrons dans les prochains chapitres à l'étude du croisement de deux variables.

- **Effectif** : c'est le nombre d'individus associé à chaque modalité, et on le note "n_i"

On a donc par définition :

$n_1 + n_2 + \ldots\ldots + + n_p$ = n, ou encore : $\sum_{i=1}^{n} n_i$ = n.

- **Fréquence :** c'est le rapport de l'effectif « n_i » sur l'effectif total « n », il est souvent exprimé en pourcentage :

$f_i = \frac{n_i}{n} * 100$ D'où : $f_1 + f_2 + \ldots. + f_p = \frac{n_1}{n} * 100 + \frac{n_2}{n} * 100 + \ldots. + \frac{n_p}{n} * 100 = 100\%$.

- **Fréquences cumulées** : on appelle fréquence cumulée croissante (ou décroissante) associée à la valeur "f_i" La somme des fréquences associées aux valeurs inférieures ou égales à "f_i" :

$F_1 = f_1$

$F_2 = f_1 + f_2$

$F_3 = f_1 + f_2 + f_3$

……

……..

$F_p = f_1 + f_2 + \ldots. + f_p = 100\%$.

- **Effectifs cumulés :** on appelle effectif cumulé croissant associé à la valeur n_p la somme des effectifs associés aux valeurs inférieures ou égales à n_p :

$N_1 = n_1$

$N_2 = n_1 + n_2$

$N_3 = n_1 + n_2 + n_3$

……

……

$N_p = n_1 + n_2 + \ldots\ldots + + n_p = N$.

2-2-2- Tableau associé à un caractère qualitatif :

La présentation tabulaire d'un caractère qualitatif nous amène à suivre les étapes citées précédemment. La forme générale prend la forme suivante :

Tableau 2-3 : Exemple d'un tableau associé à un caractère qualitatif **:**

Caractère	Effectifs	Fréquence	Effectifs cumulés	Fréquences cumulées
Modalité 1	n_1	f_1	N_1	F_1
Modalité 2	n_2	f_2	N_2	F_2
Modalité 3	n_3	f_3	N_3	F_3
. .	. .	. .	. .	. .
Modalité i	n_i	f_i	N_i	F_i

Exemple **:** Le tableau suivant représente la répartition des travailleurs d'une entreprise selon la Catégorie Socio-Professionnelle "CSP".

Tableau 2-4 : Répartition de l'effectif selon la CSP

Catégorie	Effectifs	Fréquence	Effectifs cumulés	Fréquences cumulées
Ouvriers	72	0.60	72	0.60
Techniciens	24	0.20	96.	0.80
Ingénieurs	18	0.15	114	0.95
Cadres dirigeants	6	0.05	120	1.00
Total	120	1.00		

2-2-3- Tableau associé à un caractère quantitatif :

a- Tableau associé à un caractère quantitatif discret : La représentation sous forme d'un tableau d'un caractère quantitatif prend la structure suivante :

Tableau 2-5 : Exemple d'un tableau associé à un caractère quantitatif discret :

Valeurs observées	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
X_1	n_1	f_1	N_1	F_1
X_2	n_2	f_2	N_2	F_2
. . .	. . .	. . .	. . .	. . .
Xi	n_i	f_i	N_i	Fi

Exemple : Soit la distribution du nombre d'enfants par famille pour une population de 150 ménages.

Tableau 2-6 : répartition du nombre d'enfants par famille

Nombre d'enfants x_i	Effectifs (ni)	Fréquences (fi)	Effectifs cumulés (Ni)	Fréquences cumulées (Fi)
1 2 3 4	30 45 15 60	0.2 0.3 0.1 0.4		0.2 0.5 0.6 1
Total	150	1		

b- Tableau associé à un caractère quantitatif continu : La représentation sous forme d'un tableau d'un caractère quantitatif continu prend la structure suivante :

Tableau 2-7 : Représentation d'un tableau associé à un caractère qualitatif continu

Classes	Centre Xi	Effectif n_1	Fréquence (%)
$[e_1\ e_2[$ $[e_2\ e_3[$ . . . $[e_k\ e_{k+1}[$	x_1 x_2 . . x_k	n_1 n_2 . . . n_k	f_1 f_2 . . . f_k
Total		N	100%

Avec : $x_k = (e_k + e_{k+1})/2$: est le centre de la classe i ;

On appelle amplitude de la classe i : $[e_k ; e_{k+1}[$, et on note a_i , le nombre défini par :

$a_i = e_k - e_{k-1}$

Exemple : Nombre d'années d'expérience dans une entreprise pour une population de 80 salariés.

Tableau 2-8 : Nombre d'années d'expérience dans une entreprise.

Xi	C_i	n_i	f_i (%)	N_i	F_i
[0 5[[5 10[[10 15[[15 20[	2.5 7.5 12.5 17.5	15 20 30 15	18.7 25 37.5 18.75	15 35 65 80	18,7 43,7 81,2 100

2-3- REPRESENTATIONS GRAPHIQUES :

Les représentations graphiques ont l'avantage de proposer une image plus élaborée et plus synthétique de l'ensemble des observations.

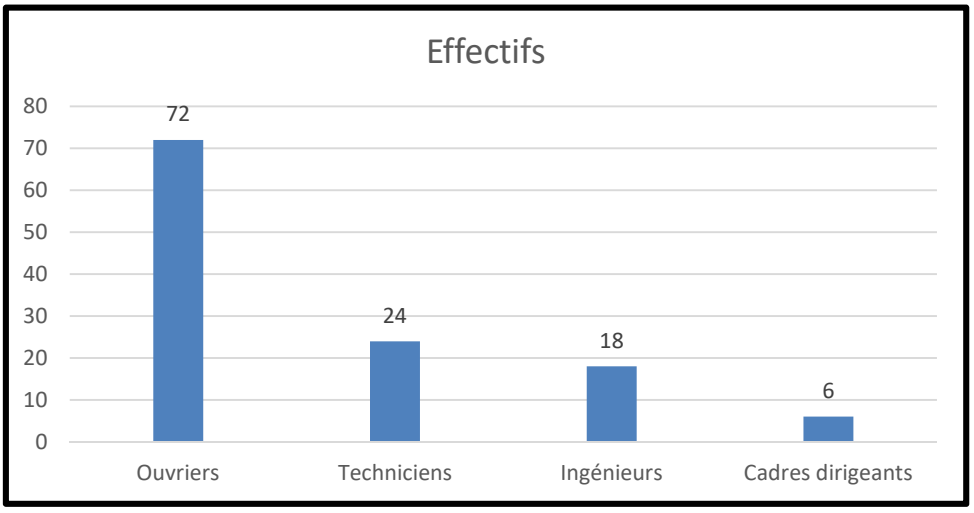
2-3-1- Caractère qualitatif : La représentation graphique d'un caractère qualitatif se fait de deux manières : Diagramme à bande ou diagramme à secteur en camembert.

a- Diagramme à bande : On utilise des rectangles où la hauteur étant proportionnelle à l'effectif ou à la fréquence. Des rectangles sont appelés **tuyaux d'orges**.

Reprenons l'exemple 1, qui présente la répartition du personnel d'une entreprise selon les quatre catégories socioprofessionnelles.

72 ouvriers ; 24 techniciens ; 18 ingénieurs ; et 6 cadres dirigeants.

Figure 2-1 : Diagramme à bande de la distribution des employés selon la catégorie socioprofessionnelle.

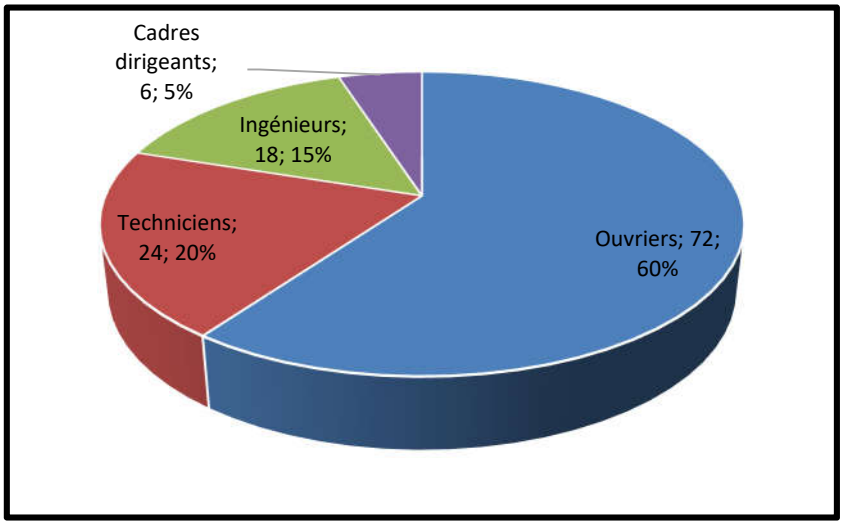


b- Diagramme à secteurs circulaires « en camembert » : La surface du cercle représente la population totale, les effectifs (ou les fréquences) sont représentés par des secteurs dont la surface est proportionnelle à ces effectifs (ou à ces fréquences).

L'angle intérieur du cercle caractérisent chaque modalité est :

$Oi = 360*fi = (n_i / n) * 360.$

Figure 2-2 : Diagramme en camembert de la distribution des employés selon la catégorie socioprofessionnelle.



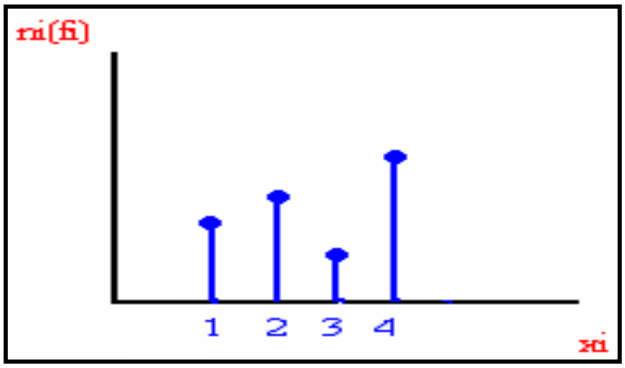
2-3-2- Caractère quantitatif :

a- Caractère quantitatif discret : Elle s'effectue de deux manières :

- **Diagrammes en bâtons :**

La représentation graphique de ce caractère se fera de la manière suivante : On positionne les modalités sur l'axe des abscisses et les effectifs ni (fréquences fi) sur l'axe des ordonnées. La hauteur des bâtons correspond à l'effectif « ni » (ou à la fréquence « fi ») associé à chaque modalité.

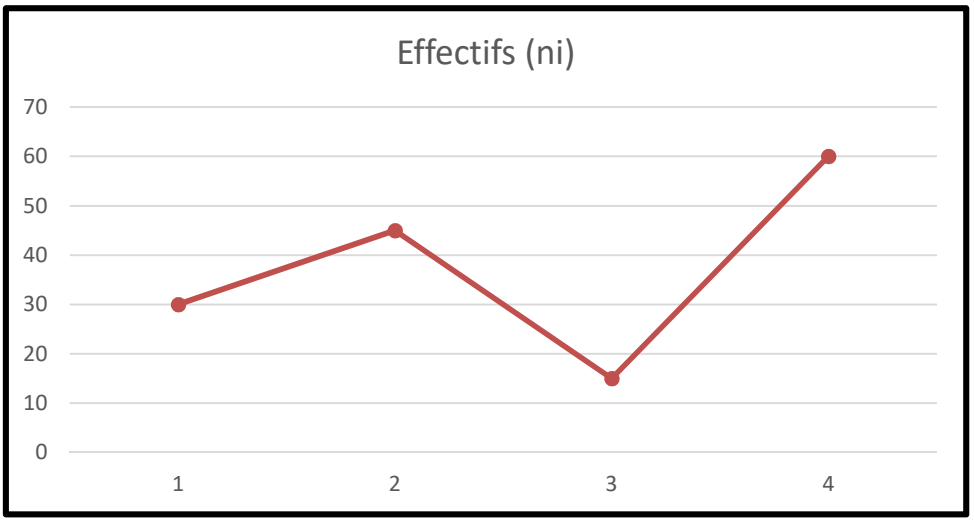
Figure 2-3 : Diagramme de la répartition du nombre de personnes par foyer.



- **Courbe :**

La courbe est une autre forme de représentation d'une série statistique graphiquement. Il s'agit de la courbe reliant les sommets du diagramme en bâtons.

Figure 2-4 : Courbe de la répartition du nombre de personnes par foyer.

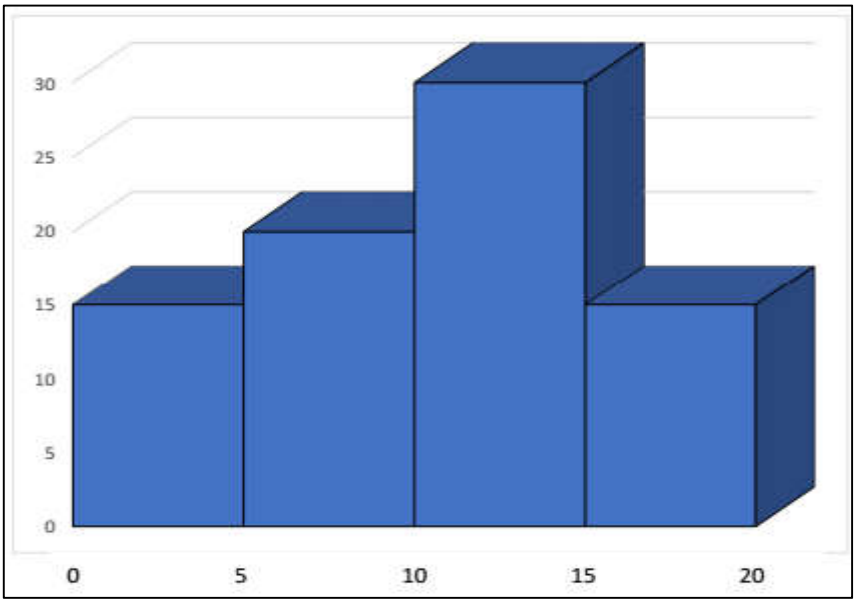


b- Caractère quantitatif continue :

C'est un diagramme composé d'un ensemble de rectangles d'aire proportionnelle aux effectifs (fréquences) et de bases déterminées par les extrémités de classe.

Traçons l'histogramme de la répartition du nombre d'années d'expérience dans une entreprise.

Figure 2-5 : Répartition du nombre d'années d'expérience dans une entreprise.



Comme on peut utiliser le polygone des effectifs ou des fréquences qui est la ligne polygonale joignant les points des centres de classe de la série d'abscisses x_i et l'ordonnée ni ou f_i correspondante.

EXERCICES D'APPLICATION DU CHAPITRE 2 :

Exercice 1 : Indiquer parmi les caractères suivants lesquels sont qualitatifs, quantitatifs discrets ou quantitatifs continus,

La profession ; La taille ; Le sexe ; Le nombre d'enfants ; Le lieu de résidence ; La nationalité ; La langue maternelle ; La pointure des chaussures ; L'état matrimonial ; Le poids ; Le nombre d'animaux domestiques possédés ; Le nombre de véhicules par famille ; Le nombre de pièces par maison ; Le revenu annuel.

Exercice 2 : Dans les quatre cas suivants, définir la population, le caractère étudié et les valeurs possibles du caractère (modalités).

- Le diamètre des boulons produits par une machine.
- Le salaire moyen annuel des ouvriers du bâtiment en Algérie de 1962 à ce jour.
- Les couleurs des voitures commercialisées de marque Toyota durant l'année 2013.

Exercice 3 : On trouvera ci-dessous l'année de naissance de chacun des trente-deux membres d'un club Sportif :

1928	1926	1926	1925	1933	1933	1931	1930
1930	1936	.1937	1929	1920	1929	1928	1940
1924	1933	1921	1936	1942	1936	1943	1934
1938	1925	1929	1935	1942	1937	1929	1931

1. a) Définir la population observée et le caractère statistique étudié.
   b) Le caractère statistique est-il qualitatif ou quantitatif, discret ou continu?
2. *a)* Ordonner les informations ci-dessus en faisant un tableau statistique donnant les différentes modalités prises par le caractère et les effectifs correspondants.
   b) Quel est le nombre de modalités prises par le caractère ?
   c) Représenter graphiquement cette série statistique ?
   d) Quels sont les avantages et les inconvénients de cette mise en ordre des données ?

3.a) Compléter le tableau ci-dessus par les colonnes des effectifs cumulés croissants ?

d) Faire le graphique des effectifs cumulés croissants ?

4. a) Répartir les observations par classe de quatre années en prenant 1920 comme limite inférieure de la première classe et compléter le tableau statistique obtenu pour pouvoir tracer les histogrammes suivants :

- Histogramme des effectifs ;
- Polygone des effectifs ;
- Courbes des fréquences cumulées ascendantes et descendantes

b) Tracer ces diagrammes

c) Quels sont les avantages et les inconvénients de cette nouvelle mise en ordre ?

Exercice 4 : Dans un groupe de 400 ménages possédant une automobile, on trouve la répartition suivante selon la marque du constructeur :

- Renault 110 Volkswagen 35
- Peugeot 80 Fiat 32
- Citroen 50 Ford 30
- Kia 40 Opel 23

1- Quel est le caractère statistique étudié, et quelle est sa nature ?

2- Quelles représentations graphiques proposez-vous pour ces distributions ?

Exercice 5 : On suppose que le nombre de maladies hospitalisés dans un centre hospitalo-universitaire de la ville d'Alger durant 24 heures le 05/01/96 a été comme suit :

Horaires	0-3	03-06	06-09	09-12	12-15	15-18	18-21	21-24
Nombre de malades	5	10	15	30	40	11	6	3

1. Déterminer l'unité statistique et la variable statistique ?
2. Calculer les fréquences relatives puis cumulées en (%), Que constatez-vous ?
3. Tracer la courbe cumulative ascendante. Déterminer le nombre de malades hospitalisés dans le centre entre 6h et 12h graphiquement et par le calcul ?

CHAPITRE 3 :

CARACTERISTIQUES DES DISTRIBUTIONS STATISTIQUES UNIVARIEES

L'étude d'une variable statistique quantitative fait souvent l'objet de l'utilisation des indicateurs de synthèses numériques. Cela nous amène à s'intéresser dans le présent chapitre à la présentation des indicateurs de tendance centrale (ou de position) et de dispersion.

3-1- Caractéristiques de tendance centrale

Les trois caractéristiques de tendance centrale les plus utilisées sont : Le mode, la médiane et la moyenne arithmétique. On peut leur ajouter les quartiles et les centiles ainsi que les moyennes géométrique et harmonique dont l'usage s'impose dans certains cas particuliers.

3-1-1- Le mode : C'est la valeur dominante ou la réponse la plus souvent rencontrée dans une distribution statistique d'une variable quantitative.

Exemple :

Une entreprise, dans le domaine pharmaceutique, veut lancer la fabrication d'un médicament dans un marché qu'elle ne maitrise pas. Il sera souhaitable pour elle de se lancer, dans un premier temps, dans la fabrication d'un produit de large consommation, ensuite, elle s'orientera vers d'autre produits en fonction de la population ciblée.

a- Le mode pour le cas discret :

Le mode est la modalité (x_i) telle que la fréquence (f_i) correspondante est la plus élevée. On le note Mo.

Exemple : Dans une PME, on a comptabilisé pendant un an le nombre de jours d'absence pour l'arrêt-maladie de chacun des 12 employés. On s'intéresse au nombre de jours d'absence le plus demandé.

Tableau 3-1 : La distribution du nombre de jours d'absence pour l'arrêt-maladie

Nb jours d'absence	0	3	5	7	8	11	13
Effectif ni	1	4	2	1	1	2	1

Le mode est de : MO = 3 jours

Remarque : il se peut que dans certains cas l'effectif maximal se répète deux ou plusieurs fois pour différente valeur de la variable. On parle alors de distribution à deux modes (bimodale) ou à plusieurs modes (plurimodales)

b- mode pour le cas continu :

Dans le cas continue, on parle de la classe modale $[a_i, a_{i+1}]$, c'est la classe correspondante au plus grand effectif. Pour déterminer la valeur du Mode qui appartient à cette classe on utilise la formule suivante :

$Mo=L_1 +A\ [\ \Delta_1/(\Delta_1+\Delta_2)\ ]$

Où :

- L_1: borne inférieure de la classe modale.
- $\Delta_1=n_i - n_{i-1}$: excédent d'effectif de la classe modale à l'effectif de la classe précédente.
- $\Delta_2=n_i - n_{i+1}$: excèdent d'effectif de la classe modale à l'effectif de la classe suivante.
- $A = a_{i+1} - a_i$: amplitude de la classe modale.

Exemple :

Déterminons le mode pour des données portant sur une variable quantitatives continues qui la répartition du salaire mensuel moyen de 33 employés d'une entreprise de l'année 2000.

Tableau 3-2 : Distribution du salaire mensuel moyen en milliers de DA

$[a_{i-1}\ a_i[$	ni
[10 20[	5
[20 30[	10
[30 40[	15
[40 50[	3
Total	n=35

La classe modale est : [30 - 40[, c'est la classe l'effectif le plus élevé.

Le mode Mo = 30 + 10 [5/(5+12)] = 30+ 2,94= 33

3-1-2- La médiane :

On appelle médiane d'une distribution, et on note **"Me"** la valeur de la variable partageant les observations classées par ordre croissant en deux groupes de même effectifs (50%).

a- La médiane pour le cas discret :

- **Données non groupées** : Pour déterminer la médiane il faut, tout d'abord, ordonner les effectifs (ou les fréquences) par ordre croissant ou décroissant. Ensuite, la médiane est la valeur de la variable située au milieu.

Remarques :

- Si le nombre d'observateur "n" est impair, la médiane est la valeur située à la position (n+1)/2 ;
- Si le nombre d'observateur n est pair, la médiane se trouve à l'intérieur de l'intervalle médian compris entre les deux valeurs centrales situées aux positions n/2 et (n/2) +1.

Exemple :

Le tableau ci-dessous donne la distribution d'un groupe de neuf ménages selon le nombre de personne par ménage. La variable statistique est le "***nombre de personne par ménage***".

Tableau 3-3 : Distribution du nombre de personne par ménage

N° du ménage	1	2	3	4	5	6	7	8	9
Nombre de personne	3	1	4	6	2	4	3	5	7

En ordonnant la série suivant les valeurs croissantes, on obtient :

Tableau 3-4 : Classement du nombre de personne par ménage par ordre croissant.

N° du ménage	2	5	1	7	6	3	8	4	9
Nb personnes	1	2	3	3	4	4	5	6	7
	4 valeurs				Me	4valeurs			

La médiane Me = 4

Interprétation[6] : Il y a autant de ménages composés de moins de 4 personnes que de ménages composés de plus de 4 personnes.

Exemple : Examinons le cas où la taille de l'échantillon est paire (12 ménages).

Tableau 3-5 : Distribution statistique de 12 ménages par rapport au nombre de personnes

N° du ménage	1	2	3	4	5	6	7	8	9	10	11	12
Nb de personnes	5	3	2	3	6	3	5	4	7	2	1	4

On ordonne la série la série par ordre croissant selon le nombre de personnes par ménage.

Tableau 2-6 : Classement du nombre de personne par ménage par ordre croissant.

N° du ménage	11	3	10	2	4	6	8	12	1	7	5	9
nb de personnes	1	2	2	3	3	3	4	4	5	5	6	7
	5 valeurs					Intervalle Médian		5 valeurs				

[6] La signification de la médiane est de plus en plus intéressante quand on travaille sur des échantillons de taille importante.

La médiane se situe entre la sixième et la septième position.

La moyenne de ces deux valeurs :

Me=(3+4)/2=3,5= 4

- **Données groupées :** Elle se détermine à partir des fréquences cumulées ou des effectifs cumulés.

Exemple :

Soit la distribution du nombre d'article vendu dans un magasin microordinateurs par jours pendant 3534 jours.

Tableau 2-7 : Distribution des ventes des microordinateurs par jours

Valeur x_i	Effectif	Effectifs cumulés	Fréquences cumulées
1	191	191	5.4
2	625	816	23.1
3	1293	2109	59.7
4	1084	3193	90.4
5	341	3534	100.0
Total	3534		

Le nombre de jours est pair, donc la moitié de l'effectif total est : 3534/2 = 1767.

Me= 3, cette modalité correspond à la fréquence cumulée qui associée à 50 % ou à celle qui vient après 50%.

b- La médiane pour le cas continu :

Elle se détermine à partir des fréquences cumulées, ou d'effectifs cumulés.

On procède en premier lieu au calcul des effectif cumulé, ensuite on détermine la classe médiane ; c'est la classe correspond au 50% des effectif ou celle qui vienne après.

Pour calculer la valeur approximative de la médiane, on utilise la formule suivante :

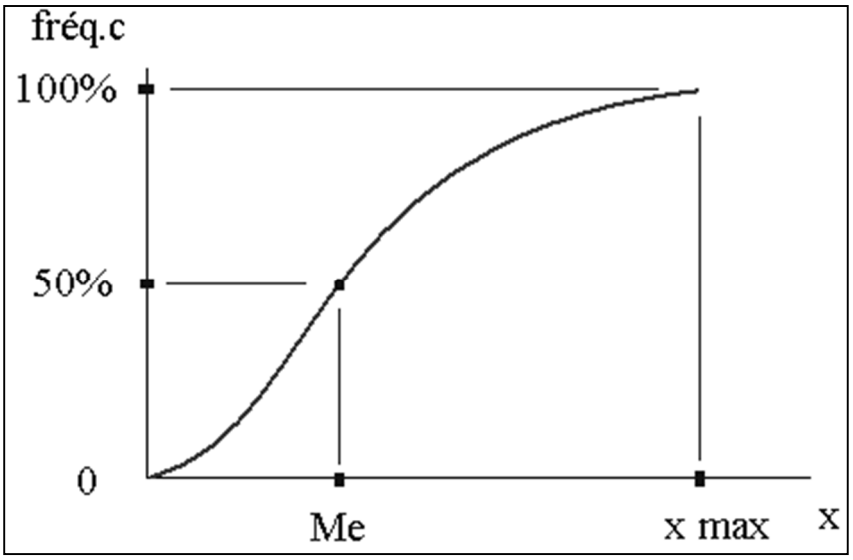
$$\frac{(Me - Xi)}{(Xj - Xi)} = \frac{(n/2 - Ni)}{(Nj - Ni)}$$

Tel que : X_j et X_i se sont respectivement la borne supérieure et la borne inférieure de la classe médiane. N_j est l'effectif cumulé de la classe médiane, et N_i est l'effectif cumulé de la classe qui précède la classe médiane.

Si n est impaire on calcule (n+1)/2 au lieu n/2.

Remarque : La médiane peut être déterminer graphiquement, et ce à partir de la courbe des effectifs ou des fréquences cumulées.

Figure 3-1 : Détermination de la médiane graphiquement



La médiane est la valeur de la variable associée à la fréquence cumulée 50%.

3-1-3- La moyenne arithmétique :

C'est l'indicateur qui permet de situer (repérer) la valeur moyenne d'une série statistique. Elle est égale à la somme des valeurs prises par cette variable divisée par le nombre d'observation : $\bar{X} = \frac{1}{n} \sum X_i$

Soit une variable statistique Xi qui peut prendre les valeurs $x_1,\ldots..,x_k$ auxquelles correspondent respectivement les effectifs $n_1,\ldots..,n_k$; la moyenne arithmétique pondérée est :

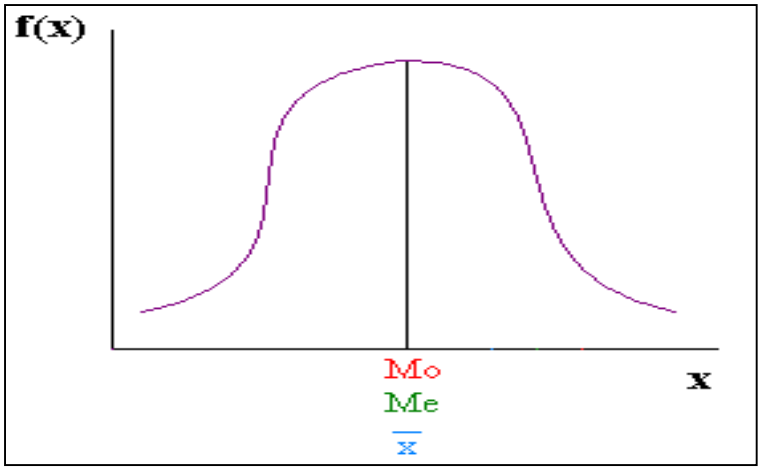
$\bar{X} = \frac{1}{n} \sum n_i X_i = \frac{1}{n} [(n_1x_1+n_2x_2+\ldots+n_kx_k)] = f_1x_1+f_2x_2+\ldots+f_kx_k$

Ou $f_i = \frac{ni}{n}$; $\sum f_i = 1$.

3-1-4- Positions graphique du mode, de la médiane et de la moyenne :

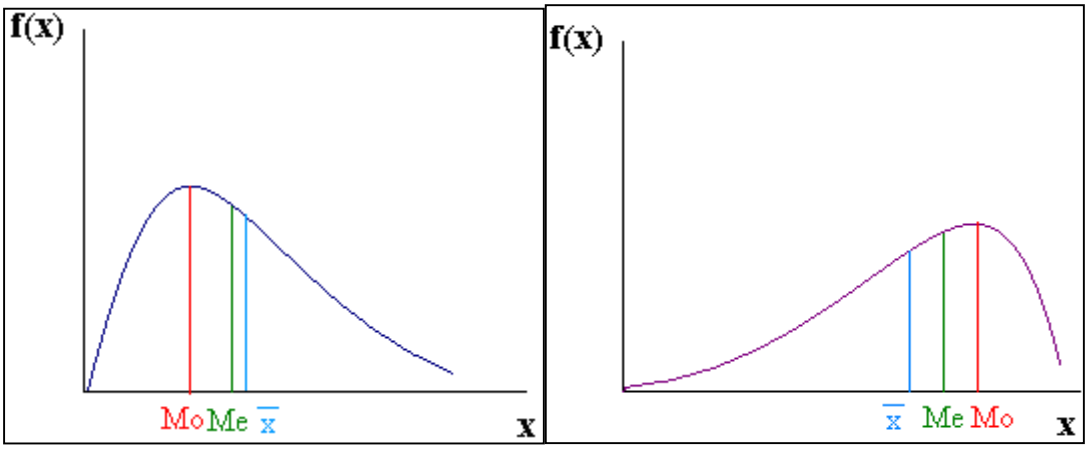
a- Distribution symétrique : Lorsque la distribution est symétrique, les 3 caractéristiques de tendance centrale sont confondues :

Figure 3-2 : Position des indicateurs de position pour une distribution symétrique



b- Distribution asymétrique : Lorsque la distribution est asymétrique, la médiane est généralement comprise entre le mode et la moyenne, et plus proche de cette dernière.

Figure 3-3 : Position des indicateurs de position pour une distribution non symétrique



3-1-4- Généralisation : Moyenne géométrique et harmonique :

Il existe d'autres types de moyennes dont l'utilisation est recommandée dans certains cas.

a- La moyenne géométrique :

On appelle moyenne géométrique d'une distribution statistique, et on la note G, la racine $n^{ième}$ du produit des n valeurs observées, soit :

$\bar{G} = (x_1 * x_2 * \ldots * x_n)^{1/n}$.

Remarque :

- Une moyenne géométrique est nulle si une seule valeur est nulle.
- Elle est généralement utilisée pour calculer une moyenne de ratios ou d'indices.

b- Moyenne Harmonique

On appelle moyenne Harmonique d'une distribution et on note H, la moyenne arithmétique des inverses de n valeurs observées :

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

Remarque :

- Une moyenne Harmonique ne peut être calculée que si toutes les valeurs observées sont non nulles.
- En pratique, elle est utilisée lorsqu'une même somme est investie dans des biens avec des prix différents et qu'on souhaite déterminer le prix moyen des biens.

3-2- CARACTERISTIQUES DE DISPERSION :

Les caractéristiques de dispersion les plus fréquentes utilisées sont : l'étendue, l'écart absolu moyen, la variance, l'écart-type et le coefficient de variation.

3-2-1- L'étendue :

L'étendue est la différence entre la plus petite valeur et la plus grande valeur d'une distribution statistique.

$$W = \max_i(x_i) - \min_i(x_i)$$

Exemple : Considérons les deux séries suivantes :

Xi=8, 9,9, 10, 10, 10, 11, 11,12

Yi=1, 3, 3, 10, 10, 10, 17, 17,19

On peut imaginer, par exemple, que Xi et Yi sont les notes d'espagnol des groupes A et B respectivement. Les deux séries ont même moyenne arithmétique (10), même médiane (10) et même mode (10). Cependant, on constate que ces séries sont très différentes : les valeurs de la première sont fortement concentrées autour de 10, alors que la deuxième présente une forte variabilité. Cette dernière, appelée dispersion, il est donc le complément indispensable à la moyenne pour un résume numérique d'une distribution statistique.

Pour les deux séries précédentes, on a :

Wx =12-8=4 et Wy = 19-1=18

On considère la distribution de la série Zi dont les valeurs sont :

Zi= 1,10,10,10,10,10,10,10,10,10,10,10,10,10,19

L'étendue de cette série est également 18, mais la variabilité est très différente on voit ici la limite de cette notion, c'est à dire qu'elle ne prend en compte que les deux valeurs extrêmes et ignore totalement les autres valeurs.

3-2-2- Ecart absolu moyen :

La somme des écarts à la moyenne arithmétique ne peut servir à mesurer la dispersion, car elle est toujours nulle : $\sum(xi - \bar{x}) = 0$.

Par contre, en prenant la valeur absolue| xi - $\overline{x}$ | on a :

On donne les séries {x1 ,……. xp} avec les effectifs respectifs {n1,…….np } vérifiant n1+…+ n_p = n . L'écart absolu moyen de cette série est la moyenne arithmétique des valeurs absolues des écarts à la moyenne arithmétique :

$$EAM = \frac{1}{n}\sum ni\left|x_i - \bar{x}\right|$$

Plus la valeur de l'EAM est importante par rapport à la moyenne, plus la distribution statistique de la variable étudiée est dispersée.

3-2-3- Variance et écart-type

Considérant la série Xi = {x1 …. xp} avec les effectifs respectifs {n1….np } vérifiant n1 +…+ np = n.

La variance de cette série est la moyenne arithmétique des carrées des écarts à la moyenne arithmétique.

$$V(x) = \frac{1}{n} ni \sum_{i=1}^{p} (x_i - \bar{x})^2$$

Quat à son interprétation, il faut calculer l'écart type.

L'écart-type σ_x est égal à la racine carrée de la variance : $\sigma_x = \sqrt{var(x)}$

Exemple : Soient les trois séries suivantes X, Y et Z :

X : 6,8,10,12,14

Y : 8,9,10,11,12

Z : 0,5,10,15,20

Il est aisé de voir que la moyenne arithmétique est égale à 10 pour chacune de ces trois séries, mais l'écart-type est respectivement :

V(x) = 8, donc l'écart-type sera 2,828.

V(Y) = 2, donc l'écart-type sera 1,414.

V(Z) = 50, donc l'écart-type sera 7,071.

EXERCICES D'APPLICATION DU CHAPITRE 3 :

Exercice 1 : Un bureau d'étude a recensé le nombre d'enfants âgés de 0 à 24 ans par famille à Alger en 2000. Les résultats figurent dans le tableau ci-dessous.

Nombre d'enfants X_i	Nombre de familles (en milliers) n_i	Fréquences (%)	Effectifs cumulés croissants	Effectifs cumulés décroissants	Fréquences cumulées croissantes	Fréquences cumulées décroissantes
0	6 064					
1	3 664					
2	3 343					
3	1 349					
4	348					
5	116					
6 et plus	81					
Total						

1- Calculer l'effectif total et remplir la colonne des fréquences.

2- Remplir les colonnes des effectifs et fréquences cumulés croissants et décroissants.

3- Quelle est la proportion de familles ayant 4 enfants ou plus ?

4- Combien de familles qui ont 3 enfants ou moins ?

5- Dessiner le diagramme des fréquences et celui des fréquences cumulées croissantes ?

6- Calculer et commenter les mesures de tendance centrale et de dispersion ?

Exercice 2 : On étudie la distribution des 1000 premières entreprises bénéficiaires d'une subvention gouvernementale d'aide à l'embouche des jeunes selon leur chiffre d'affaire. Les données sont les suivantes :

Chiffre d'affaire (10^6)	**Nombre d'entreprises**	Amplitude A	N_i corrigés= N_i/A
[0 ; 2[	12		
[2 ; 5[	54		
[5 ; 10[	314		
[10 ; 50[	407		
[50 ; 100[	156		
[100 ; 500[	52		
[500 ; 1000[	01		
$\sum$	**1000**		

- Calculer les valeurs de tendance centrale ?

Remarque : Il faut tenir compte des amplitude variables dans le calcul des différents indicateurs.

Exercice 3 : Une enquête portant sur les dépenses qu'un ménage a effectué, en moyen durant un mois, pour acheter une voiture d'occasion. Les résultats sont résumés dans le tableau suivant :

Prix d'achat (*1000 DA)	Nombre de famille
[3000 ; 4000[	5
[4000 ; 5000[	60
[5000 ; 6000[	15
[6000 ; 7000[	95
[7000 ; 8000[	30
[8000 ; 9000[	5

1- Calculer les indicateurs de position et de dispersion ?

2- Calculer le troisième et le premier quartile et commenter ? Déduire l'intervalle interquartile ?

Exercice 4 : Afin d'étudier les caractéristiques des employés d'une entreprise spécialisée dans le secteur de l'énergie, nous avons choisi les variables suivantes : Genre, Catégorie socioprofessionnelle, Lieu de résidence, Salaire, Nombre de personne par foyer, Nombre de pièces par maison, distance entre lieu de travail et lieu de résidence.

1- Proposez au moins deux modalités pour chacune des variables ci-dessus ?

2- Supposant qu'on a introduit dans l'étude la variable « *type de logement* » dont la codification et la suivante : "1" pour studio ; "2" pour un logement de type F2 ; "3" pour un logement de type F3 et "4" pour un logement de type F4 et plus. Après une enquête auprès d'un échantillon des employés on a obtenu les résultats suivants : 2, 4, 2, 1, 3, 2, 4, 1, 4, 1, 3, 2, 2, 1, 3, 3, 3, 4, 2, 3, 1, 2, 2, 3, 3, 2, 3, 3, 2, 1, 2, 2, 3, 2, 2, 4, 4, 4, 3.
   a- Quelle est la taille de la population étudiée ?
   b- Construire le tableau donnant les effectifs, les fréquences en pourcentages, les effectifs cumulés croissants et les fréquences cumulées décroissantes ?
   c- Représentez graphiquement le diagramme des effectifs (simples) ?
   d- Calculez et commentez les indicateurs de position et de dispersion ?

Exercice 5 : On considère la variable “Temps vécu dans un logement en année” pour laquelle on a obtenu le tableau d’effectifs suivants :

xi	[0 ;1 [	[1 ;2[	[2 ;3[	[3 ;5[	[5 ;11[	[11 ;16[	[16 ;21[	[21 ;26]
ni	35	36	32	25	20	18	16	7

1. Quel est le type de cette variable ?
2. Déterminer les valeurs de : Mode, Moyenne, médiane ? Commentez ?
3. Calculer l’écart-type et Commentez le résultat ?

A cause d’une erreur de saisie, la borne supérieure 26 a été remplacée par 66 ; cela a-t-il un impact sur les valeurs des indicateurs précédents ?

Conclusion de la partie 1 :

L'analyse de données univariée est une étape importante dans toute étude statistique. En effet, le chargé d'étude détermine les besoins en information en fonction du problème qu'il cherche à résoudre et ensuite il collecte ces informations. Cependant, disposé de l'information sur un sujet n'est pas une fin en soi ; il faut savoir lire et interpréter ces chiffres. Cela nous amène à travailler sur présentation des tableaux synthétisant toute l'information, ensuite transformer ces tableaux en graphique adapter à la nature de la variable et enfin faire des synthèses numériques à l'aide des indicateurs de position et de dispersion.

Cependant, les analyses de données univariées ne sont pas suffisantes pour prendre de bonnes décisions ou analyser en profondeur l'ensemble des variables qui font l'objet de l'étude. A cet effet, il est souvent nécessaire de recourir à des méthodes d'analyse de donnés bivariées et multivariées. Cela nous amène à aborder dans la deuxième partie les analyses de données bivariées, c'est-à-dire analyser la relation entre deux variables.

PARTIE 2 :

STATISTIQUE ET ANALYSE DES DONNEES BIVARIEES

INTRODUCTION DE LA PARTIE 2 :

Les méthodes d'analyse de données univariées ont montrées leurs limites en traitant chacune des variables séparément des autres variables. Compte tenu de la complexité du monde réel, le recours à des méthodes d'analyse de données bivariées est nécessaire afin d'apporter plus de précisions et de clartés à nos résultats.

L'objet de cette partie est de présenter les méthodes d'analyse de données bivariées. Dans un premier temps nous exposons les distributions bivariées, ensuite nous focalisons nos propos sur le croisement de deux variables (test de khi-deux, analyse de la corrélation). Enfin, cette partie s'achève par une analyse de la régression qui une introduction à la modélisation économétrique.

CHAPITRE 4 :
CARACTERISTIQUES DES DISTRIBUTIONS STATISTIQUES BIVARIEES

4-1- Introduction :

Les distributions statistiques à deux caractères sont présentées sous forme de tableaux à deux dimensions dont les distributions marginales sont les distributions de chacun des deux caractères, étudiés séparément, sans conditions quant à la modalité prise par l'autre caractère.

Exemple : Soit la distribution de 29 salariés d'une entreprise selon l'âge et le revenu mensuel.

Tableau 4-1 : Distribution de l'âge et du revenu

X/Y	<20	20-30	30-40	40-50	50-60	∑
<20	1	0	0	0	0	1
20-30	1	3	1	1	0	6
30-40	1	1	0	2	0	4
40-50	1	0	4	2	4	11
50-60	0	1	0	3	2	6
>60	0	0	0	1	0	1
∑	4	5	5	9	6	29

Ce tableau permet d'obtenir les deux distributions marginales suivantes :

Tableau 4-2 : Distributions marginales de l'âge et du revenu.

Pour x	
X	∑
<20	**1**
20-30	**6**
30-40	**4**
40-50	**11**
50-60	**6**
>60	**1**
∑	**29**

Pour y	
Y	∑
<20	**4**
20-30	**5**
30-40	**5**
40-55	**9**
50-60	**6**
∑	**29**

4-2- Tableau de contingence : Considérons N individus décrits simultanément selon deux caractères X et Y.

X possède k modalités : x1, x2, x3, ……., xi, …….x_k

Y possède p modalités : y1, y2, y3,…….., yj,……..y_p

Tableau 4-3 : Structure d'un tableau de contingence

Y / X	y_1	……	Y_j	……..	Y_p	Total
X_1	n_{11}					$n_{1.}$
_ _ _		.. .. ..				
x_i			n_{ij}			$n_{i..}$
_ —						
x_k						$n_{k.}$
Total	$n_{.1}$		$n_{.j}$		$n_{.p}$	**N**

Les paramètres utilisés pour caractériser les distributions à deux variables sont de deux types :

- Les paramètres qui concernent une seule variable, ils servent à caractériser les diverses distributions marginales.

- Les paramètres qui servent à décrire les relations qui existent entre les deux séries d'observations considérées simultanément.

4-3- Moyennes et variances marginales :

Le calcul de la moyenne et de l'écart type pour les distributions marginales se fait Comme toute distribution uni variée.

Dans la notation complète des modalités des variables et des effectifs à deux dimensions, on a pour x :

$\bar{X}= 1/n_{..} \sum n_{i.} * x_i$ $V(x)=1/n_{..} \sum n_{i.} * (x_i-X)^2$; Avec : $n_{..}=\sum n_{i.}$

Pour Y :

$\bar{Y}= 1/n \sum n_j * y_j$ $V(y)=1/n \sum n_j * (y_j-Y)^2$; Avec : $n_{..} \sum n_{.j}$

Considérant l'exemple précédent, on a :

$\bar{X}= 1/n \sum n_i * x_i = 41,20$ $V(x)=1/n \sum n_i * (x_i-X)^2 =147,68.$

$\bar{Y} = 1/n \sum n_j * y_j = 38,31$ $V(y)=1/n \sum n_j * (y_j -Y)^2 = 1724,7.$

4-4- Distributions conditionnelles :

Les distributions statistiques à deux caractères sont présentées sous forme de tableaux à deux dimensions dont chaque distribution conditionnelle est la distribution d'un caractère si l'autre caractère prend l'une de ses modalités.

Si on reprend les données de l'exemple précédent :

Si le salaire est compris entre 30 et 40 mille dinars, l'âge (Xi) des salaries va prendre les valeurs suivantes :

Tableau 4-4 : Distribution conditionnelle de l'âge par rapport au salaire

X	<20	20-30	30-40	40-50	50-60	>60	∑
30-40	0	1	0	4	0	0	5

Si l'âge est compris entre 50 et 60 ans, le salaire (Yi) va prendre les valeurs suivantes :

Tableau 4-5 : *Distribution conditionnelle du salaire par rapport à l'âge.*

Y	**<20**	**20-30**	**30-40**	**40-55**	**50-60**	∑
50-60	0	1	0	3	2	**6**

4-5- Moyennes et variances conditionnelles :

Pour toutes distributions conditionnelles, on peut calculer la moyenne et la variance conditionnelle comme pour toutes distributions statistique uni variée (un seul caractère).

On parle alors, de moyenne et variance de x si $y=y_j$

Ou moyenne et variance de y si $x=x_i$

Reprenant l'exemple précédent et on calcule la moyenne et la variance conditionnelle de la variable revenu (Y) si l'âge (X) est compris entre 20 et 30ans ?

$\bar{Y} = 1/n\sum n_j * y_j = 41$ $V(y)=1/n \sum n_j * (y_j - Y)^2 = 1745$

EXERCICE D'APPLICATION DU CHAPITRE 4 :

Exercice 1 : Dans une entreprise activant dans le secteur des travaux publics, on relève le tableau des jours d'absences des employés "Y", durant le mois de mai 2015, en fonction de leurs âge "X" :

Y X	0	1	2	3	Distributions Marginales Xi
[20-30[	2	10	6	2	
[30-40[	15	10	5	0	
[40-50[	2	18	15	5	
[50-60[	0	2	4	4	
Distributions Marginales de Yi					

1- Calculer la moyenne et l'écart-type des deux variables X et Y ?

2- Calculer et commenter le coefficient de variation des deux variables X et Y ?

3- Quelle est la distribution du nombre de jours d'absence quand l'âge est fixé dans la classe [30-40[? Déduire la moyenne et l'écart-type pour cette distribution et commenter ?

4- Quelle est la distribution de l'âge quand le nombre de jours d'absence est fixé à deux jours ? Déduire la moyenne et l'écart-type pour cette distribution ?

Exercice 2 : Soit la relation entre la catégorie socioprofessionnelle et le revenu mensuel (en milliers de dinars) de 120 employés d'une entreprise spécialisée dans la production des doublets en plastiques.

	Ouvriers	techniciens	ingénieurs	cadre	Total
20- 30	25	3	0	0	
30-40	10	15	0	0	
40-50	4	12	10	0	
50-60	0	5	12	0	
60-70	0	0	5	0	
70-80	0	0	3	6	
80-90	0	0	4	5	
90-100	0	0	0	1	
Total					

1- Calculer les distributions marginales des deux variables X et Y ?

2- Calculer la moyenne et l'écart-type de la variable revenu mensuel ?

3- Représenter graphiquement les distributions marginales des deux variables ?

CHAPITRE 5 : TEST D'INDEPENDANCE DE KHI DEUX.

5- 1- Démarche générale d'un test d'hypothèse :

L'analyse de base des données implique obligatoirement les tests d'hypothèses. Les exemples en marketing générés par les tests d'hypothèses sont nombreux, par exemple :

- Le magasin « X » est fréquenté par plus de 20% des foyers de la région ;
- Le fait de connaitre un restaurant se traduit par une préférence plus marquée en sa faveur ;
- Un hôtel possède une image plus haute de gamme que son concurrent direct.

D'autres questions peuvent être posées en marketing, mais afin de donner des arguments solides à nos réponses il est nécessaire de faire appel à des test d'hypothèses. En effet, la démarche scientifique nous oblige de passer en revue les concepts de distribution d'échantillonnage, de marge d'erreur et d'intervalle de confiance concernent le test d'hypothèses.

a- Formulation de l'hypothèse nulle H_0 et l'hypothèse alternative H_1 :

L'hypothèse nulle H_o exprime l'absence de différence ou d'effet. Si l'hypothèse nulle n'est pas rejetée, aucun changement ne se produit. L'hypothèse alternative H_1 exprime l'attente d'une différence ou d'un effet quelconque.

Par exemple dans le cadre d'étude du comportement du consommateur, l'adoption de l'hypothèse alternative engendre des changements en termes d'opinions ou de comportements. L'hypothèse alternative se définit donc comme l'opposée de l'hypothèse nulle.

Dans la pratique, la vérification porte toujours sur l'hypothèse nulle, laquelle fait référence à une valeur spécifique d'un paramètre de la population (Moyenne, Proportion, Variance…), et non à un indicateur mesuré sur l'échantillon.

Dans une étude de cas, l'hypothèse alternative se trouve formulée de telle sorte que son rejet aboutisse à l'adoption de la conclusion souhaitée. L'hypothèse alternative présente la conclusion que l'on cherche à motivée.

Dans la pratique, on distingue deux types de test d'hypothèse : test unilatéral et bilatéral.

b- Choisir un test approprié :

Pour tester une hypothèse nulle, il est nécessaire de choisir une technique statistique approprié. Le chargé d'étude doit tenir compte du mode de calcul de la statistique du test et de la distribution suivie par la statistique de référence (la moyenne, par exemple).

La statistique du test mesure la proximité de l'échantillon vis-à-vis de l'hypothèse nulle. Elle s'aligne généralement sur une distribution classique (normale, Student ou encore Khi-deux).

c- Choisir le niveau de signification α :

Dès que l'on cherche à dégager des inférences (inductions analogies généralisations…) par rapport à une population, on prend le risque d'aboutir à une conclusion erronée. Deux types d'erreurs peuvent survenir :

- *Rejeter H_0 alors que H_o est vraie :* cette erreur s'appelle erreur de première espèce ;
- Accepter *H_0 alors que H_1 est vraie :* On appelle cette erreur, erreur de seconde espèce ;

Il est clair que lors d'une observation d'un échantillon une décision, seulement, peut être prise, soit rejeter H_0 soit on l'accepte.

Par conséquent, l'espace Ω de tous les échantillons possibles sera divisé en deux parties qu'on notera Ω_a et Ω_r.

Ω_a : Ensemble d'acceptation et Ω_r : Ensemble de rejet.

On appelle seuil du test, que l'on note α, la probabilité d'erreur de première espèce, définie par : $\alpha = P(\Omega_r/H_0)$= Probabilité de rejeter H_0 à tort.

On appelle puissance du test, que l'on note 1-β, le complémentaire à 1 de la probabilité d'erreur de seconde espèce : $\beta = P(\Omega_a/H_1)$= Probabilité d'accepter H_0 à tort.

d- Collecter les données et calculer la statistique du test :

On détermine la taille de l'échantillon en fonction des critères cités dans le chapitre 1, tels que : erreur d'échantillonnage α souhaitées, degré de précision et des considérations

d'ordre matériel et humaines. Ensuite, nous procédons à la collecte d'informations nécessaire pour la résolution du problème posé. Enfin, nous calculons la statistique avec la formule adaptée aux tests.

e- Déterminer la probabilité (Valeur critique) :

Il est à noter que lors de la détermination de la valeur critique de la statistique du test, la surface située à droite de la valeur critique est égale à « α » pour un test unilatéral, et « α/2 » pour un test bilatéral.

f- Comparer la probabilité (valeur critique) et prendre une décision :

La décision à prendre est la suivante :

- Si la probabilité de la statistique du test calculée est inférieure au niveau de signification (α), alors rejeter H_0 ;
- Si la statistique du test calculée est supérieure à la statistique du test critique, alors rejeter H_0.

Enfin, la conclusion tirée du test d'hypothèses doit s'exprimer en des termes adaptés au problème de l'étude.

Exemple : Supposant un fabriquant des moteurs des appareils d'électroménagers d'une durée de vie moyenne de 3000 heures, et un écart type de $\rho = 150\ heures$.

Suite à la modification d'un certain nombre de composante de ces moteurs, le fabriquant c'est rendu compte que les nouveaux moteurs ont une durée de vie moyenne supérieure à celle des anciens. Afin de vérifier s'il y a une amélioration, on a tiré un échantillon de 50 nouveaux moteurs, on a trouvé que la durée de vie moyenne est de 3250 heures et un écart type de 150 heures. Quelles conclusions peut-on porter au risque d'erreur de α= 5% ?

$Z = \frac{\mu - \mu_0}{\delta/\sqrt{n-1}} = \frac{3250-3000}{150/7} = 11{,}66$

La Valeur critique du test se détermine à partir de la table de la loi de Student au risque d'erreur 5/2 % = 2,5% et un degré de liberté (n-1= 49).

$T_{(\alpha/2)}$= 1,96.

Comparaison : Z=11,66 > $t_{\alpha/2}$= 1,96. On rejette H_0 et on accepte H_1.

Pour conclure, on dit que les nouveaux moteurs ont durée de vie moyenne supérieure à celle des anciens.

5-2- Méthodologie de calcul du test de Khi-deux :

Le test de khi-deux s'appliqué à des données classées selon un tableau de contingence où on note dans chaque case l'effectif observé correspondant à une modalité X_i d'une variable x et une modalité y_j d'une autre variable Y.

Le principe du test d'indépendance consiste à comparer ces effectifs observés o_{ij} aux effectifs attendus c_{ij} si X et Y sont indépendantes.

Pour ce faire, **o**n formule les deux hypothèses de travail :

H_0 : X et Y sont indépendantes.

H_1 : X et Y sont liées.

Tableau 5-1 : Tableau des effectifs observé (O_{ij}*)*

	y_1	……	Y_j	……..	Y_k	Total
X_1						n_1.
– – –						
x_i			O_{ij}			$n_{i.}$
– —						
x_I						n_I
Total	$n._1$		$n._j$		$n._k$	N

- On calcule les effectifs attendus c_{ij} sous H_0.

- On Calcule la statistique : $\chi_0^2 = \frac{\sum (O_{ij} - C_{ij})^2}{C_{ij}}$ suit une loi du χ^2 à (I-1) (k-1) ddl.

"I" étant le nombre de lignes et "k" le nombre de colonnes.

Les conditions de validité du test sont : $c_{ij} \geq 5$ pour tout (i,j).

Si ces conditions ne sont pas vérifiées, on fait des regroupements de classes.

Comment déterminer c_{ij} ?

C_{ij} est l'effectif attendu si X et Y sont indépendants

$C_{ij} = Np_{ij}$ avec p_{ij} = probabilité ($X=x_i$ et $Y=y_j$)

X et Y étant indépendantes, $p_{ij} = p(X=x_i)\ p(Y=y_j)$

Donc $p_{ij} = \frac{n_{i.}}{N}\frac{n_{.j}}{N}$ et $c_{ij} = Np_{ij} = \frac{n_{j.}n_{i.}}{N}$

L'effet attendu c_{ij} s'obtient en faisant le rapport :

$$C_{ij} = \frac{(\text{Total de la i ième ligne})(\text{Total de la j ième colonne})}{Total}$$

Tableau 5-2 : Tableau des effectifs observé (C_{ij})

	y_1	……	Y_j	……..	Y_k
X_1					
_ _ _					
x_i			C_{ij}		
_ __					
x_I					

On calcule ensuite les $\chi_0^2 = \sum \frac{(O_{ij} - C_{ij})^2}{C_{ij}}$ et cette statistique suit une loi du χ^2 à (I-1) (k-1) ddl dans le cas des grands échantillons (avec $c_{ij} \geq 5$).

Le nombre de ddl (I-1) (k-1) est égal au nombre d'effectifs calculés indépendants. En effet la connaissance de (I-1) (k-1) c_{ij} permet de déduire les autres, puisque $\sum c_{ij} = n_i . \sum c_i = n_{.j}$

La règle de décision : On compare χ^2_0 calculé avec le $\chi^2_{(I-1)\,(k-1)}$ lu sur la table de khi deux

* Si : $\chi^2_0 > \chi^2_{(I-1)(k-1)}$ on rejette H_0 , c'est-à-dire, il y a un lien entre X et Y.

5-3- Exemple d'application du test de Khi-deux :

Application de test d'homogénéité sur la production nationale des céréales par région :

On dispose dans le cadre de cette étude de données sur la production des céréales durant l'année 2000 (année sèche), présenté sous forme d'un tableau de contingence, ces dernières sont réparties par wilayas qui représentent à leurs tour les différentes régions du pays (Est, Ouest, Centre, Sud).

On veut savoir si ces régions sont homogènes ou si au contraire, certaines encouragent plus la production d'un produit que d'autres.

Autrement dit, y a-t-il un lien entre le type de produit céréalier et la région ?

Tableau 5-3 : Les effectifs observés (O_{ij})

O_{ij}	Blé dur	Blé tendre	Orge	Somme (∑)
Adrar	76890	106650	32590	216130
Bouira	44990	27540	24150	96680
Tlemcen	28000	20000	13000	61000
S.B.Abbes	94000	118000	56200	268200
Annaba	276280	100230	36650	413160
Boumerdes	119800	46000	17500	183300
Souk-Ahras	225000	92400	38500	355900
Somme (∑)	864960	510820	218590	1594370

On forme les hypothèses de travail :

* H_0 : La production des céréales est hétérogène dans les différentes régions

* H_1 : La production des céréales est homogène entre les différentes régions

Tableau 5-4 : Les effectifs attendus (théoriques) C_{ij} sous H_0.

C_{ij}	Blé dur	Blé tendre	Orge
Adrar	117252.46	69245.863	29631.6769
Bouira	30975.2928	30975.2928	13254.9416
Tlemcen	33093.0462	19543.7822	8363.17166
S.B.Abbes	145500.901	85928.5636	36770.5351
Annaba	224142.999	132372.279	56644.7214
Boumerdes	99441.8912	58727.4635	25130.6453
Souk-Ahras	193078.937	114026.755	48794.3081

Les conditions de validité du test (C_{ij} >5) sont vérifiées.

* On calcule ensuite les $\chi^2_{ij} = \frac{(O_{ij}-C_{ij})^2}{C_{ij}}$

Tableau 5-5 : Les valeur de χ^2_{ij} sous H_0.

χ^2_{ij}	Blé dur	Blé tendre	Orge
Adrar	13894,1919	20204,3762	295,348643
Bouira	1060,97906	380,988705	8955,32413
Tlemcen	783,823859	10,6496631	2570,81618
S.B.Abbes	18229,0475	11970,1411	10266,4839
Annaba	12127,3779	7804,70151	7057,83121
Boumerdes	4167,78674	2758,30621	2316,96192
Souk-Ahras	5277,39749	4101,81406	2171,82666

$\chi^2_0 = \sum \frac{(O_{ij}-C_{ij})^2}{C_{ij}} = 136406{,}17.$

On a I= 7, et K= 3 donc ddl = (I-1) (K-1) = 12, au seuil de signification de 95%,

χ^2_{12}= 21,026 (lu sur la table de khi deux).

La décision est :

On a $\chi^2_0 > \chi^2_{12}$ donc H_0 est rejeté, c'est-à-dire la production des céréales est homogène dans les régions durant cette année. Cela signifié que les pouvoirs publics ne favorisent pas une région par rapport une autre pour la production des trois types de produits céréaliers.

5-4- Test de Cramer (V) :

L'une des limites du test de khi-deux c'est qu'il ne nous renseigne pas sur l'intensité de relation entre deux variables. Dans le cadre de l'analyse des résultats 'une enquête, il est très long de voir tous les croisements possibles entre les variables deux à deux, donc à l'aide du test de cramer, on peut cibler les relations de dépendance les plus fortes à analyser seulement. C'est à dire il permet donc d'indiquer si la relation de dépendance entre deux variables est forte, moyenne ou faible.

Le test de Cramer ou le « V » de Cramer se calcule comme suit : $\mathbf{V} = \sqrt{\frac{\chi 2}{\mathbf{DDL}+N}}$

Avec : v : la valeur du test de Cramer ;
$\chi 2$: est la valeur du test de Khi-deux ;
N : la taille de l'échantillon.

La règle de décision :
Quatre situations à distinguer quant à l'interprétation du test de Cramer :

1- Si : V < 0,1 : Relation nulle ou très faible ;
2- Si : 0,1 =< V < 0,2 : Relation faible ;
3- Si : 0,2 =< V < 0,3 : Relation moyenne ;
4- Si : 0,3 =< V < 1 : Relation Forte.

EXERCICES D'APPLICATION DU CHAPITRE 5 :

Exercice1 :

Afin d'analyser la relation entre le choix d'une spécialité à l'ENSM et le sexe des étudiants, nous disposons des données un échantillon de taille 50 :

Spécialité Sexe	Management Marketing	Management des ressources humaines	Management stratégique et système d'information	Management par la qualité	Management des organisations
Masculin	6	4	8	5	4
Féminin	4	8	5	2	4

Question :

Analyser la relation en la spécialité et sexe en utilisant le test de khi-deux ?

Calculer le V de Cramer et déduire l'intensité de relation entre le sexe et la spécialité choisies ?

Exercice 2 :

Soit la relation entre le lieu d'achat d'un téléphone portable et la catégorie socioprofessionnelle (CSP).

Lieu d'achat CSP	Point de vente agrée par la marque	Ailleurs[1]
Cadre	25	80
Non cadre	75	50

Question :

1- Analyser la relation en la spécialité et sexe en utilisant le test de khi-deux ?

2- Calculer le V de Cramer et déduire l'intensité de relation entre le lieu d'achat et la CSP ?

[1] L'ensemble des magasins qui vendent des téléphones portables multimarque.

Exercice 3 :

Afin de montrer est ce qu'il y a une relation entre le type de logement (X_i) et la catégorie socioprofessionnelle (Y_i) on a interrogé une population de 238 individus ensuite on a construit le tableau croisé suivant :

Y_i X_i	F1	F2	F3	F4
Ouvrier	12	20	10	13
Technicien	22	18	9	13
Ingénieur	8	10	17	24
Cadre Supérieur	5	11	21	25

Questions :

1- Calculez les distributions marginales ?

2- En utilisant le test de khi-deux, démontrez la nature de la relation entre le type de logement (X_i) et la catégorie socioprofessionnelle (Y_i) ? Etudier l'intensité de la relation entre ces deux variables ?

CHAPITRE 6 :

LIAISON ENTRE DEUX VARIABLES QUANTITATIVES : LA CORRELATION.

6-1- Introduction :

En présence de deux variables quantitatives, on peut se demander s'il existe un lien linéaire entre ces deux variables ou si, ou contraire, l'une évolue indépendamment de l'autre. Lorsqu'une liaison existe entre deux variables, on peut s'interroger sur son sens (proportionnel ou opposé) et sur son importance (forte ou faible).

Pour répondre à ces questionnements, on peut faire une représentation graphique où positionne l'une des variables en abscisse et l'autre en ordonnées "Nuage des points". Comme on peut mesurer cette relation à l'aide d'un indicateur statistique : La covariance ou encore le coefficient de corrélation.

6-2- Etude graphique "Nuage des points" :

Selon la nature de la relation qui existe entre deux variables statistiques quantitatives, ce lien peut être plus ou moins fort, et on distinguera trois cas : Absence de relation ; dépendance totale et dépendance partielle.

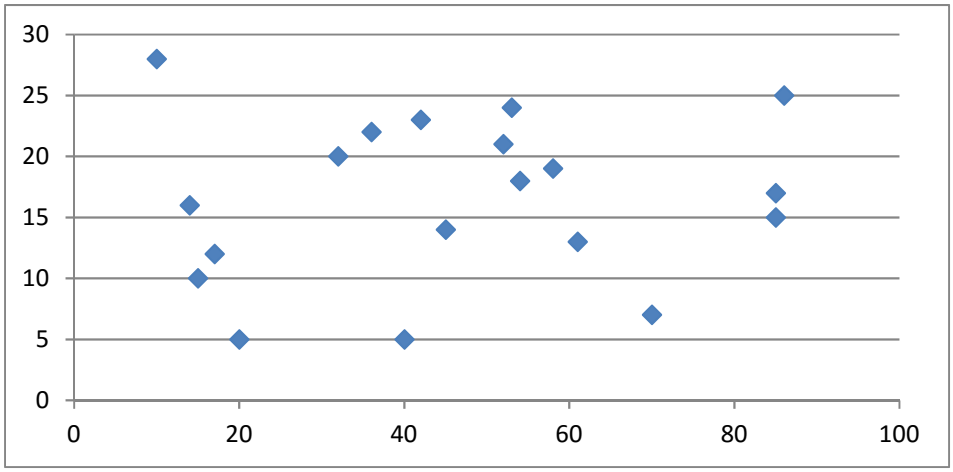
a- Absence de relation linéaire :

La liaison entre deux variables est absente ou nulle s'il n'existe aucune relation entre elles. Autrement dit, la connaissance de la valeur prise par l'une des variables n'apporte rien dans la connaissance de la valeur prise par l'autre variable.

Le nuage des points en cas d'indépendance entre les variables ne prend pas de forme particulière[2] et elle est en générale sous la forme suivante :

[2] Les points du nuage sont répartis d'une façon aléatoire.

Figure 6-1 : Nuage des points en cas d'absence de corrélation



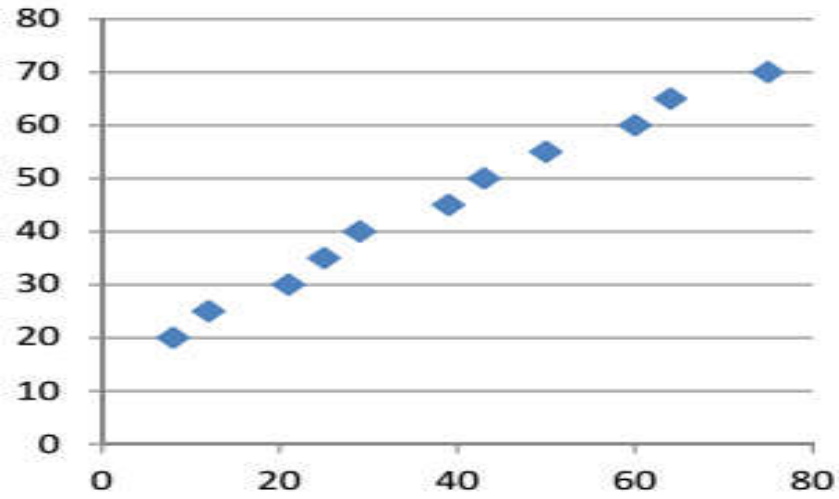
b- Dépendance entre les deux variables :

La relation entre deux variables est dite proportionnelle ou inverse, s'il y a dépendance (totale ou partielle) entre les deux variables. Dans ce cas, la connaissance de la valeur prise par l'une des variables permet de connaitre approximativement[3] la valeur prise par l'autre variable.

E*xemple :* le lien entre : le Taux d'intérêt et l'investissement, le Prix et la demande d'un bien sur le marché (fonction de demande ou d'offre).

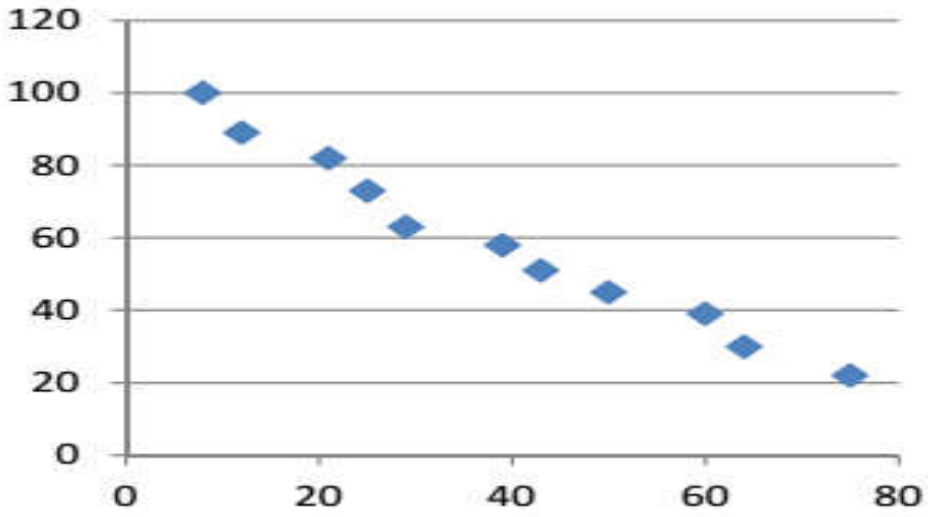
Le nuage de points en présence d'une dépendance entre les variables prend, selon la nature de la relation, une des deux formes suivantes :

Figure 6-2 : Nuage des points en cas de corrélation positive.



[3] Cette approximation et relative au niveau de dépendance entre les deux variables.

Figure 6-3 : Nuage des points en cas de corrélation négative



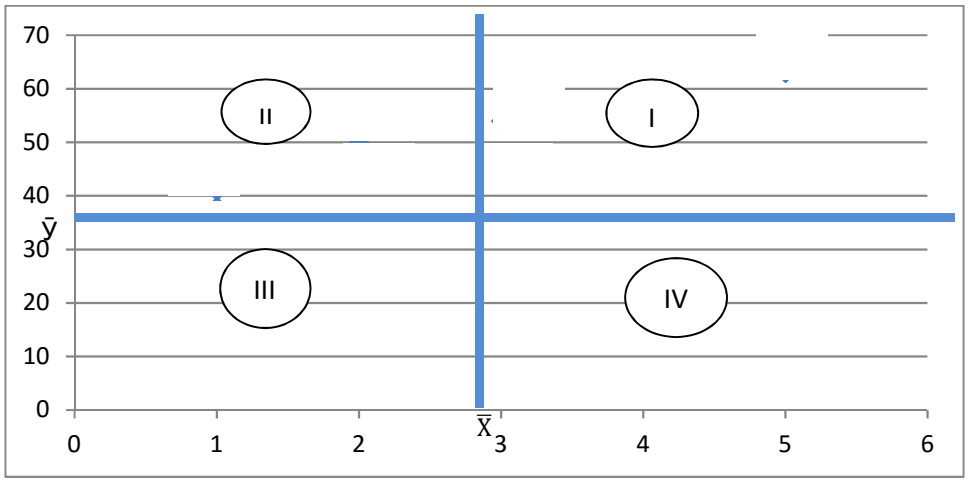
6-3- La covariance :

C'est un indicateur statistique qui permet de mesurer le lien linéaire entre deux variables quantitatives. Sa formule de calcul est la suivante :

Cov (x, y)= 1/n $\sum(X_i - \bar{X})(Y_i - \bar{y})$

Quant à l'interprétation de ce coefficient on utilise le graphique suivant, où on distingue quatre cadrans qui sont délimités par les moyennes des deux variables :

Figure 6-4 : Les quatre cadrans d'un nuage des points.



- Cadran I : $X_i > \bar{X}$ et $Y_i > \bar{y}$;
- Cadran II : $X_i < \bar{X}$ et $Y_i > \bar{y}$;
- Cadran III: $X_i < \bar{X}$ et $Y_i < \bar{y}$;
- Cadran IV : $X_i > \bar{X}$ et $Y_i < \bar{y}$.

Quant à l'interprétation de la covariance, on a les trois cas suivants :

- Si la Cov (X, Y) >0 : les points qui ont la plus grande influence sur la covariance se trouvent dans les cadrans I et III. Et une valeur positive de la covariance relève une relation linéaire positive entre X et Y ;
- Si Cov (X, Y)<0 : cela signifie que les points situés dans les cadrans II et IV qui ont la plus grande influence sur la covariance. Et une valeur négative de la covariance relève une relation linéaire négative entre X et Y ;
- Si les points du nuage sont répartis de façon uniforme entre les quatre cadrans, la valeur de la covariance sera proche de « 0 », impliquant l'absence d'une relation linéaire entre les deux variables X et Y.

En résumé :

- ✓ Une valeur positive importante[4] de la covariance implique une forte relation linéaire positive, et vice versa.
- ✓ Une valeur proche de zéro implique l'absence d'une relation linéaire entre les deux variables.

Cependant, si la valeur de la covariance est élevée par rapport aux valeurs prises par une des variables et faible par rapport aux valeurs prises par la deuxième variable, on pourra par jugé de la nature de la relation entre les deux variables. Et pour remédier à ce problème, on fait appel à un autre indicateur qui est : le *coefficient de corrélation.*

6-4- Le coefficient de corrélation :

C'est un indicateur statistique qui permet de mesurer l'intensité de la relation entre deux variables quantitatives X et Y.

Ce coefficient noté "r" et il se calcule à l'aide la formule suivante : $r=\frac{\text{cov}\,(X,Y)}{\delta x \delta y}$.

Le coefficient de corrélation a les caractéristiques suivantes :

- r : est sans unité de mesure.
- r est toujours compris entre -1 et +1.
- Si r est proche de 1 ; la corrélation est positive

[4] L'importance ici c'est par rapport aux valeurs prises par les deux variables (les moyennes par exemple).

- Si r est proche de -1 ; la corrélation est négative.
- Si r est proche de -0,5 ou 0,5 ; la corrélation est moyenne (négative ou positive).
- Si r est proche de 0 ; la relation linéaire est absente.

Dans le cas non pondéré, les données se présentent sous forme d'une série de n couples (X_i, Y_i), alors :

$$r = \frac{\sum_{i=1}^{i=n}[(xi - \overline{X})\,(Yi - \overline{Y})]}{\sqrt{[\Sigma(xi - \bar{x})^2][\Sigma(yi - \bar{y})2}}$$

Dans le cas pondéré, on a :

$$r = \frac{\sum_{i=1}^{i=n} ni[(xi - \overline{X})\,(Yi - \overline{Y})]}{\sqrt{[\Sigma ni(xi - \bar{x})^2][\Sigma ni(yi - \bar{y})2}}$$

6-5- Exemple d'application de la corrélation :

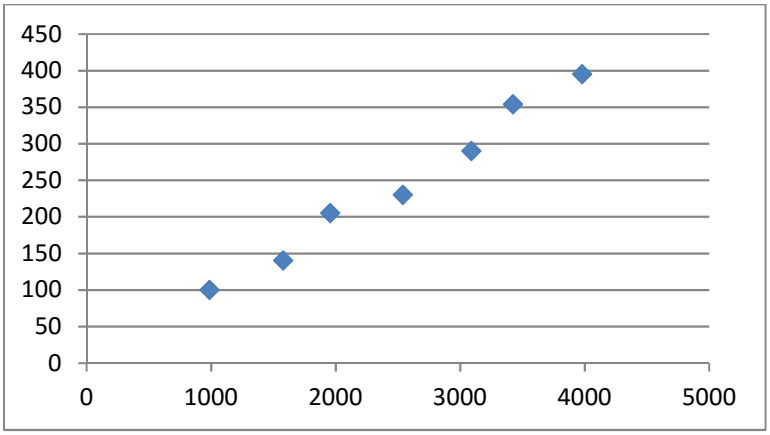
Une entreprise souhaite savoir s'il existe une corrélation entre les dépenses publicitaires qu'elle a engagées sur un produit et le nombre d'unités vendues de ce produit. Pour ce faire on dispose les données annuelles suivantes :

Tableau 6-1 : Relation entre dépenses publicitaires et ventes

Années	1	2	3	4	5	6	7	∑	Moyenne
Nombres d'unités vendues (X_i)	987	1578	1956	2540	3091	3426	3980	17558	2508,28
Dépenses publicitaires (Y_i)	100	140	205	230	290	354	395	1714	244,85

- On procède en premier lieu à l'étude graphique de la corrélation entre les dépenses publicitaires et le nombre d'unités annuellement.

Figure 6-5 : Nuage de points dépenses publicitaire et ventes



On remarque sur le graphe que les points du nuage des points sont alignés d'une façon croissante, on s'attend donc à une valeur importante positive de la covariance entre les dépenses publicitaires et le nombre d'unités vendues de ce produit.

- On passe au calcul de la covariance pour confirmer la première réponse.

Cov (X_i, Y_i) = **1/n** $\sum$**$(X_i - \bar{X})(Y_i - \bar{y})$**

$\bar{X} = \frac{1}{n}\sum Xi = 2508{,}28571$, $\bar{y}$= 1/n $\sum Y_i$= 244,857143

Tableau 6-2 : Calculs de la covariance entre dépenses publicitaires et ventes

Années	1	2	3	4	5	6	7	$\sum$
$X_i - \bar{X}$	-1521,28	-930,28	-552,28	31,71	582,71	917,71	1471,71	0
$(X_i - \bar{X})^2$	2314310,22	865431,51	305019	1005	339555,93	842199,51	2165942,94	6833465,4
$Y_i - \bar{y}$	-144,85	-104,85	-39,85	-14,85	45,14	109,14	150,14	
$(Y_i - \bar{y})^2$	20983,59	10995,02	1588,59	220,73	2037,87	11912,16	22542,87	70280,85
$(X_i - \bar{X})$ $(Y_i - \bar{y})$	220369,10	97547,10	22012,53	-471,18	26305,38	100161,95	220967,38	686892,28

Cov (X_i, Y_i) = **1/n** $\sum$**$(X_i - \bar{X})(Y_i - \bar{y})$=686892,28/7=** 98127,4694

La covariance prend une valeur positive importante, donc il existe une forte corrélation positive entre les dépenses publicitaires et le nombre d'unités vendues de ce produit. Et il confirme aussi l'allure du graphique ci-dessus.

Cependant, il arrive des fois que l'interprétation de la covariance est ambigüe. Supposant qu'on a trouvé une valeur comprise entre 400 et 950. Dans ce cas la covariance est élevée par rapport aux dépenses publicitaires et basse par rapport au nombre d'unités vendues. Donc, on peut juger de la nature du lien qu'il existe entre les deux variables, et pour pallier à ce problème on doit calculer le coefficient de corrélation « r ».

$$r = \frac{\sum_{i=1}^{i=n}[(xi - \overline{X})\,(Yi - \overline{Y})]}{\sqrt{[\Sigma(xi - \overline{x})^2][\Sigma(yi - \overline{y})2}$$

$$r = \frac{686892,28}{(6833465,4*70280,85)1/2} = 0,99.$$

Puisque le coefficient de corrélation est proche de 1, on dit qu'il existe une forte corrélation positive entre les dépenses publicitaires et le nombre d'unités vendues de ce produit.

6-6- Corrélation et causalité :

Une valeur élevée du coefficient de corrélation linéaire (en valeur absolue) ne signifie pas forcément qu'il existe un lien de dépendance réel, de causalité, entre les deux phénomènes décrits. On peut observer une liaison statistique importante entre deux variables même si aucun lien de causalité direct n'existe entre ces deux phénomènes. Il ne faut pas donc confondre liens de causalité et liaisons statistiques.

EXERCICE D'APPLICATION DU CHAPITRE 6 :

Exercice 1 : Les observations portant sur la quantité vendue d'un bien (Q) et son prix (P) se résument comme suit :

$\sum_{i=1}^{100} Q_i = 7930$; $\sum_{i=1}^{100} P_i = 2470$; $\sum_{i=1}^{100} Q^2 = 650950$; $\sum_{i=1}^{100} P^2 = 61570$;

$\sum_{i=1}^{100} Q_i P_i = 192840$

1. Calculer les moyennes et les variances empiriques des deux variables Q et P ?

2. Calculer la covariance entre Q et P. En déduire le coefficient de corrélation linéaire entre ces deux variables.

3. Selon le résultat de la question 2, préciser la nature de la relation entre Q et P (s'agit-il d'une fonction d'offre ou de demande ?).

Exercice 2 : Reprenant l'énoncé du premier exercice du chapitre3.

Y **X**	**0**	**1**	**2**	**3**	**Distributions Marginales Xi**
[**20-30**[	2	10	6	2	
[**30-40**[	15	10	5	0	
[**40-50**[	2	18	15	5	
[**50-60**[	0	2	4	4	
Distributions Marginales de Yi					

Question : Analyser l'existence ou l'inexistence de relation entre le nombre des jours d'absences des employés "Y" et leurs âge "X" ?

CHAPITRE 7 : MODELE DE REGRESSION LINEAIRE SIMPLE

7-1- Introduction :

Dans le cadre d'une analyse économétrique, nous pouvons considérer qu'un modèle consiste en une *présentation formalisée d'un phénomène* sous forme d'équation dont les variables sont des grandeurs économiques. En effet, l'objectif du modèle (économétrique) est de représenter *les traits les plus marquants* d'une réalité qu'il cherche à styliser (clarifier), d'où l'hypothèse "*toutes choses égales par ailleurs"* ou selon l'expression latine "*Cétirus Paribus".* Pour ce faire, le modélisateur émet des hypothèses et explicite des relations.

Rôle du modèle change en fonction de l'objet de l'étude ; soit pour valider une théorie économique dans un contexte particulier, soit pour faire des investigations (mise en évidence des relation qui n'étais pas auparavant pressenties, simulation ou prévisions à court terme).

7-2- Présentation du Modèle de Régression Linéaire Simple "MRLS" :

7-2-1- Généralité sur le MRLS : le modèle de régression linéaire simple permet d'expliquer une variable endogène (expliquée, dépendante) "Y_i" en fonction d'une autre variable explicative (exogène, indépendante) "X_i", la forme générale du MRLS est suivante :

$Y_i = a + b\,X_i + \xi_i$. i=1…n.

Avec:

Y_i : est une variable endogène, inconnue et aléatoire.

X_i : Est une variable exogène, connue (mesurer sans erreur) et non aléatoire.

ξ_i : est le terme d'erreur ou l'aléa, inconnue et non observé qui permet de prendre en compte le fait que la variable Y_i est affectée par d'autre variables que la variable X_i. Autrement dit, le fait que la variable X_i n'explique pas pleinement la variable Y_i et d'où l'hypothèse *Toutes choses égales par ailleurs.*

n : représente le nombre d'observation ou la taille de l'échantillon.

7-2-2- Exemple introductif du MRLS : Considérons l'analyse, pour l'année 2010, des dépenses de consommation des travailleurs d'une entreprise en communications téléphoniques en fonction de leurs revenus mensuels.

$C_i = a + b\ R_i + \xi_i$.

Où : C_i : représente les dépenses en communications téléphoniques ;

a : les dépenses minimales en communications téléphoniques quand le revenu s'annule ;

b : Le degré de sensibilité des dépenses en communications téléphoniques suite à la variation du revenu (l'élasticité) ;

R_i : Revenu mensuel ;

ξ_i : le terme d'erreur : on note que les travailleurs dont le revenu est le même ne dépensent le même montant en communications téléphoniques, ils existent donc d'autres variables (facteurs), non prise en charge par le modèle, qui peuvent expliquer le comportement des travailleur ; notamment étendu du réseau d'amis, le goût pour le bavardage… Dans le modèle économétrique, ces facteurs sont pris en compte par l'aléa ξ_i.

7-3- Estimation du modèle avec la méthode des Moindres Carrés Ordinaires "MCO".

C'est la méthode d'estimation de base des paramètres a et b du modèle de régression linéaire simple : $Y_i = a + b\ X_i + \xi_i$.

a- Hypothèses de MCO sur le MRLS :

H_1 : Espérance des aléas est nulle. $E(\xi_i) = 0$. i= 1…n.

H_2 : Les aléas sont homoscédastiques (leur variance est constante) et non autocorrélés (leur covariance est nulle).

- $Var(\xi_i) = \sigma^2$.
- $Cov(\xi_i, \xi_j) = 0$, i, j=1…n, $i \neq j$.

Si H_1 et H_2 sont vérifiées, l'aléa ξ_i est un Bruit Blanc (White Nouse).

H_3 : La covariance entre les l'aléa et la variable explicative X_i est nulle. Cov (X_i, ξ_i)=0.

H_4 : La variable X_i est mesurée sans erreur, et sa variance non nulle.

H_5 : l'aléa suit une loi normale de moyenne zéro et d'écart type σ. $\xi_i \rightarrow$ N (0, σ).

b- Estimation des paramètres a et b par la méthode MCO :

Le principe de MCO consiste à minimiser la somme des résidus aux carrés.

On a : $e_i = Y_i - \hat{Y}$, $\sum e_i^2 = \sum (Y_i - \hat{Y})^2 = \sum (Y_i - \hat{a} - \hat{b}X_i)^2$. On cherche à minimiser : $S = \sum e_i^2$.

D'après les conditions du premier ordre :

$$\frac{dS}{da} = 0 \; et \; \frac{dS}{db} = 0$$

La solution est donnée par les deux équations :

$\frac{dS}{da} = -2 \sum(Yi - \hat{a} - \hat{b}X_i) = 0$……1

$\frac{dS}{db} = -2 \sum(Yi - \hat{a} - \hat{b}X_i) X_i = 0$ ……2

De (1) et (2) on a déduit :

$\hat{b} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}$ **et** $\hat{a} = \overline{Y} - \hat{b}\overline{X}$

La droite de régression de l'échantillon est alors donnée par :

$\hat{Y} = \hat{a} + \hat{b}X_i$, où : i=1…n.

Les caractéristiques de la droite sont les suivantes :

- La droite de régression passe le point moyen $(\overline{Y}, \overline{X})$, par conséquent : $\overline{Y} = \overline{\hat{Y}}$.
- La somme des résidus est nulle : $\sum e_i = 0$. Les valeurs positives éliminent les valeurs négatives.
- Le vecteur des résidus et de la variable explicative sont orthogonales. $\sum e_i X_i = 0$.

7-4- Tests de validation du modèle de régression linéaire simple :

a- Tests de significativités des paramètres a et b :

On a[5] : $v(\hat{a})=\frac{\sigma^2\sum {X_i}^2}{n\ \sum(X_i-\bar{X})^2}$ $v(\hat{b})=\frac{\sigma^2}{\sum(X_i-\bar{X})^2}$, $\hat{\sigma}^2=\frac{1}{n-2}\sum {e_i}^2$, $E(\hat{\sigma}^2)=\sigma^2$

$\text{Cov}(\hat{a},\hat{b})=\frac{-\bar{X}\sigma^2}{\sum(X_i-\bar{X})^2}$

**** Test de signification du paramètre $\hat{a}$:***

$$\begin{cases} H_0: \hat{a}=0 \\ H_1: \hat{a}\neq 0 \end{cases}$$

La statistique du test : $T_{cal}=\frac{\hat{a}}{\hat{\sigma}_{\hat{a}}} \rightsquigarrow T_{(n-2)}$.

Règle de décision :

Avec « α » un risque d'erreur fixé généralement à : 5%.

Si : $|T_{cal}| > T_{(\alpha/2;n-2)}$, alors l'hypothèse H_o est rejetée au seuil de α% et le paramètre $\hat{a}$ peut être considéré significativement différent de zéro.

**** Test de signification du paramètre $\hat{b}$:***

$$\begin{cases} H_0: \hat{b}=0 \\ H_1: \hat{b}\neq 0 \end{cases}$$

La statistique du test : $T_{cal}=\frac{\hat{b}}{\hat{\sigma}_{\hat{b}}} \rightsquigarrow T_{(n-2)}$.

Règle de décision :

Avec « α » un risque d'erreur fixé généralement à : 5%.

Si : $|T_{cal}| > T_{(\alpha/2;n-2)}$, alors l'hypothèse H_o est rejetée au seuil de α% et le paramètre $\hat{b}$ peut être considéré significativement différent de zéro.

[5] Pour une démonstration mathématique de ces formules, le lecteur pourra se référer à l'ouvrages de : Bourbounais (2005) et Cadoret (2009).

b- Tests sur la qualité globale d'ajustement du modèle :

*** Analyse de la variance "ANOVA" :**

Le test de l'ANOVA a pour but de tester la qualité globale d'ajustement du modèle.

En effet, la qualité de l'estimation est traduite par l'équation suivante :

$$\sum(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})^{\boldsymbol{2}} = \sum\left(\widehat{\boldsymbol{Y}}_{\boldsymbol{i}} - \overline{\boldsymbol{Y}}\right)^{\boldsymbol{2}} + \sum\left(\boldsymbol{Y_i} - \widehat{\boldsymbol{Y}}\right)^{\boldsymbol{2}}$$

La somme des carrés totale (SCT)= La somme des carrés estimés (expliqués par le modèle) "SCE" + La somme des carrés résiduelles (non expliqués par le modèle) "SCR".

L'équation : SCT= SCE+ SCR, s'appelle *Équation d'analyse de la variance.*

Tableau 7-1 : Analyse de la variance.

Source de variation	Somme des carrés	Degrés de liberté	Somme des carrés moyens
Modèle	SCE	P	SCE/p
Résidus	SCR	n-p-1	SCR/(n-p-1)
Total	SCT	n-1	SCT/(n-1)

Où : n représente la taille de l'échantillon, et p : le nombre de paramètres du modèle (constante exclue).

*** Coefficient de détermination R^2 :**

C'est un indicateur synthétique issu du tableau d'analyse de la variance, il permet d'évaluer la qualité globale du modèle construit.

$R^2 = \frac{SCE}{SCT}$ = 1- $\frac{SCR}{SCT}$, et il est compris entre 0 et 1.

Le coefficient de détermination R^2 est égal au carré du coefficient de corrélation "r".

*** Test de signification de R^2 :**

Afin de s'assurer si le modèle est intéressant, on procède au test d'hypothèse suivant :

- $H0 : R^2 = 0$;
- $H1 : R^2 \# 0$.

La statistique du test est : $F_{cal} = \frac{\frac{R^2}{p}}{\frac{(1-R^2)}{(n-p-1)}} = \frac{\frac{R^2}{1}}{\frac{(1-R^2)}{(n-2)}} \rightsquigarrow F_{(1\ ;\ n-2)}$.

Règle de décision : Si $F_{cal} > F_{(\alpha;\ 1;\ n-2)}$ alors l'hypothèse H_0 est rejetée au seuil α% (5%, 10%...)

7-5- Prévision à court terme de la variable expliquée Y_i :

Dans le cadre du modèle linéaire simple : $Y_i = a + b\ X_i + \xi_i$, on se pose la question suivante : Pour une valeur future X_{i+h} de X_i, quelle valeur peut-on prédire pour la variable expliquée ?

Soit : $\hat{Y} = \hat{a} + \hat{b}X_i$, le modèle de régression linéaire estimé par la méthode MCO. Soulignons que les paramètres â et $\hat{b}$ sont estimés sur un échantillon comprenant les observations i=1,2….n, sans inclure la réalisation (i+h), et puisque : $Y_{i+h} = a + b\ X_{i+h} + \xi_{i+h}$, il est naturel de proposer la prédiction : $\hat{Y}_{i+h} = \hat{a} + \hat{b}X_{i+h}$.

EXERCICES D'APPLICATIONS DU CHAPITRE 7 :

Exercice n°1 : Reprenant l'énoncé de l'exercice 1 du chapitre 5, et on considère à présent le modèle linéaire suivant :

$Q_i = \alpha + \beta P_i + \xi_i$, i =1, ⋯ , 100.

Où ξ_i, i =1, ⋯, 100. sont des termes aléatoires identiquement et indépendamment distribués d'espérance mathématique zéro et de variance σ^2. , α et β sont des paramètres à estimer.

(a) Donner les expressions des estimateurs de α et β obtenus par la Méthodes des moindres carrés ordinaires.

(b) Calculer les valeurs numériques des estimateurs $\widehat{\alpha}$ et $\widehat{\beta}$.

(c) Décomposer la variance de Q en variance expliquée et variance résiduelle.

(d) Déterminer la valeur numérique de l'estimation sans biais de la variance des erreurs, notée$\hat{\sigma}^2$.

(e) Calculer la valeur de l'estimation de la variance de $\widehat{\beta}$.

(f) En admettant la normalité de E_i :

i. Construire un intervalle de confiance au niveau 95% pour le paramètre β.

ii. Tester l'hypothèse H_0 : β = −5 contre l'hypothèse alternative H_1 : β ≠ −5 pour un niveau de confiance de 95%.

Exercice n°2 : Sur une période de dix ans, on dispose des informations sur le chiffre d'affaire (Y_t), mesurée en millions de dinars et le taux d'intérêt (R_t) mesuré en pourcentage.

Ces informations sont résumées comme suit :

$\sum_{t=1}^{10} E_t = 435,\ \sum_{t=1}^{10} R_t = 43{,}5,\ \sum_{t=1}^{10} E_t^2 = 20775,$

$\sum_{t=1}^{10} R_t^2 = 215{,}25,\ \sum_{t=1}^{10} E_t R_t = 2106{,}5$

On se propose d'estimer une relation linéaire dénié par :

$E_t = \alpha + \beta R_t + E_t$, t =1, ⋯, 10 …….…. (1)

Avec α et β des paramètres à estimer ; E_t, t =1, ⋯, 10 sont des termes aléatoires identiquement et indépendamment distribués d'espérance mathématique zéro et de variance σ^2. On suppose que ces termes suivent la loi normale.

1. (a) Donner l'interprétation économique des deux paramètres α et β. Quels sont les signes attendus de ces paramètres ?

(b) Quelle interprétation économique peut-on donner au terme d'erreur E_t ?

2. Estimer les paramètres α et β par la méthode des moindres carrés ordinaires.

3. Calculer la valeur numérique de l'estimateur sans biais de la variance des termes d'erreurs.

4. (a) Calculer le coefficient de corrélation linéaire simple entre les deux variables E_t et R_t ?

(b) En déduire le coefficient de détermination.

(c) Testez l'hypothèse H0 : $R^2=0$ contre H1 : $R^2=1$, au seuil de 5% ?

5. Tester la significativité de l'effet de la variable R_t sur E_t, au seuil de 5% ?

6. Pour l'année 11, le taux d'intérêt observé est de 4. Construire un intervalle de prévision pour l'épargne de l'année correspondante, au niveau de 95%.

Exercice n°3 : Les données suivantes correspondent aux dépenses publicitaires (en million de DA) et les ventes (en million de DA) de sept grandes marques de boissons non alcoolisés.

Marque	Coca-cola	Pepsi	Coca-Light	Sprit	Dr-Pepper	Red Bool	Hemmoud
Dépenses pub (D_t)	131,3	92,4	60,4	55,7	40,2	29,0	11,6
Ventes (V_t)	1929,2	1384,6	811,4	541,5	536,9	625,6	219,5

1- Représentez le nuage de points associé à ces données ?

2- Quelle relation entre les deux variables, le nuage de points indique-t-il ?

3- Supposant que la relation entre les deux variables à la forme suivante :

$V_t = a + b\, D_t + E_t$.

a- Rappeler les hypothèses classiques du terme d'erreur permettant de calculer les paramètres avec la méthode des Moindres Carrés Ordinaires "MCO".

b- Estimez les paramètres a et b par la méthode MCO ?

c- Tracez la droite de régression et Interprétez sa pente ?

d- Testez la significativité des deux paramètres a et b au seuil de 5%?

CONCLUSION DE LA PARTIE 2 :

Les méthodes d'analyse de données bivariées permettant de traiter le croisement entre de deux variables, qualitative ou quantitative selon le cas, ont été présentées dans cette parties. En effet, l'analyse de deux variables qualitatives a pour but de s'interroger sur la dépendance ou l'indépendance entre ces deux variables (test de khi-deux) ainsi que l'intensité de cette relation (test de Cramer).

Par ailleurs, l'étude du système de relation entre deux variables quantitatives permet de vérifier l'existence ou l'inexistence du lien linéaire entre ces deux variables (nuage de pointe ou covariance) et de mesure l'importance de cette relation (coefficient de corrélation). Enfin, le modèle de régression linéaire vise à analyser l'impact, s'il existe, d'une variable quantitative sur une autre variable quantitative.

PARTIE 3 :
INTRODUCTION A L'ANALYSE DE DONNEES MULTIVARIEES

INTRODUCTION DE LA PARTIE 3

L'expression Analyse De Données "ADD" possède aujourd'hui un sens très précis en mathématiques. Elle désigne l'ensemble des méthodes ou de techniques qui permettant de traiter des situations qui impliquent un grand nombre de *caractères* et *d'individus*. Ces méthodes nécessitent beaucoup de calculs et elles ont fleuri depuis que l'informatique leur permet d'être mises en œuvre.

Souvent très élaborées, elles sont néanmoins très « *descriptives* ». Elles reposent pour beaucoup sur une analyse *géométrique* de la représentation des données dans un espace (*abstrait*) de petite dimension. Elles s'intéressent à la fois aux caractères (détermination des liaisons) et aux individus (par exemple sous-structures du nuage de points des individus).

La méthode qui se trouve dans le prolongement direct de la corrélation et de la régression linéaire s'appelle « l'*analyse en composantes principales* ». Les autres méthodes portent les noms d'analyse (factorielle) discriminante, d'analyse des correspondances, de classification hiérarchique, etc.

Toutes les méthodes d'analyse des données débouchent (mais non exclusivement) sur des représentations graphiques, et certaines sont fréquemment utilisées dans les études économiques et sociales.

L'objet de cette partie est de faire une introduction à l'analyse de données multi variées à travers la présentation la méthode d'analyse en composante principale "ACP".

CHAPITRE 8 : ANALYSE EN COMPOSANTE PRINCIPALE

8-1- Les objectifs de l'ACP :

L'Analyse en Composantes principales (ACP) fait partie des méthodes d'analyses descriptives multivariées. Le but de cette analyse est de résumer le maximum d'informations possibles pour :

A. Faciliter l'interprétation d'un plus grand nombre de données initiales.
B. Donner plus de sens aux données réduites.

L'ACP permet donc de réduire des tableaux de grandes tailles en un petit nombre de variables (2 ou 3 généralement) tout en conservant le maximum d'information.

On utilise l'ACP pour faire apparaître :

- Les liaisons entre variables : les systèmes de relation qui existent entre elles. C'est-à-dire : *leurs associations ou leurs oppositions.*
- La répartition des individus les uns par rapport aux autres, en relation avec les variables traitées. Autrement dit, les individus qui représentent des caractéristiques communes ou antagonistes.

L'Analyse en Composantes Principales (ACP) consiste à transformer des variables liées entre elles dites "corrélées" en nouvelles variables décorrélées les unes avec autres.

Ces nouvelles variables sont nommées *"COMPOSANTES PRINCIPALES"*.

Elle permet aux praticiens de réduire l'information en un nombre de composantes plus limité que le nombre initial de variables.

8-2- Types de tableaux pouvant être traités par une ACP :

L'analyse en composantes principales peut être appliquée sur :

- Les tableaux de mesures : variables quantitatives ;
- Les tableaux de notes : Variables qualitatives ordinales ;
- Les tableaux de rangs : classement des individus de 1 à n, du meilleur au mauvais, du plus rapide au plus lent, etc.

8-3- Le tableau de donnée initial :

On suppose que l'on dispose des observations de "**P**" variables quantitatives sur **"n"** individus. Les valeurs sont "rangées" dans un tableau à **n** lignes et **p** colonnes ; on note X la matrice associée à ce tableau :

$$X = \begin{bmatrix} \ldots\ldots & \ldots\ldots & \ldots\ldots\ldots \\ \ldots\ldots & X_i^j & \ldots\ldots\ldots \\ \ldots\ldots & \ldots\ldots & \ldots\ldots\ldots \end{bmatrix}$$

Où x^i_j est la valeur prise par la variable j sur l'individu i.

Une variable j sera alors identifiée au vecteur $x^j = [x^j_1 \ldots\ldots\ldots x^j_n]^t$ et un individu i sera identifié au vecteur $e_i = [x^1_i \ldots\ldots\ldots x^p_i]$.

On peut interpréter géométriquement les lignes et les colonnes du tableau ***X*** par des points dans deux espaces différents : l'espace des variables et l'espace des individus.

- L'espace des individus : Les *n* lignes peuvent être considérés comme *n* points de l'espace des individus à *p* dimensions. Deux points sont très proches si les ***p*** coordonnées de ces deux points sont très proches (mêmes valeurs pour les différentes variables).
- L'espace des variables : Les *p* colonnes peuvent être considérés comme *p* points dans un espace à *n* dimensions. Cet espace est appelé l'espace des variables. Si les valeurs prises par deux variables sont très voisines pour l'ensemble des individus, ces variables seront très proches (ce qui peut signifier que les variables mesurent la même chose ou encore qu'elles soient liées par une relation particulière).

  Les données sont les mesures effectuées sur n unités $\{u_1, u_2, \ldots, u_i, \ldots u_n\}$. Les p variables qui représentent ces mesures sont $\{v_1, v_2, \ldots, v_j, \ldots v_p\}$.

8-4- Individus et variables supplémentaires

a- Individus supplémentaires :

Afin de faciliter l'interprétation des résultats, on peut introduire dans le tableau de données de départ des données que l'on appelle individus supplémentaires.

Les unités statistiques supplémentaires sont des unités statistiques sur lesquelles on dispose des observations sur l'ensemble des variables mais dont on ne tient pas compte dans le calcul des paramètres statistiques.

L'intérêt des données supplémentaires est de caractériser les graphiques des groupes d'unités statistiques supplémentaires.

b- Variables supplémentaires :

Ce sont des variables n'ayant pas de rapport direct avec l'analyse mais que l'on souhaite voir représentées dans les graphiques.

Certains auteurs utilisent les termes de variables actives pour les variables de départ et passives pour les variables supplémentaires.

8-5- Transformation des données initiales :

a- Données centrées :

On appelle données centrées le tableau de données :

$\tilde{X} = (X_i^j - \bar{X}^j)$

Dans ce tableau, la somme des valeurs d'une même colonne est nulle.

b- Données centrées réduites :

On appelle données centrées réduites le tableau de données :

$\hat{X} = \frac{(X_i^j - \bar{X}j)}{\delta j}$

Dans ce tableau, la somme des valeurs d'une même colonne est nulle et la somme des carrés des valeurs d'une même colonne vaut 1.

c- Matrice de variances/covariances :

On appelle matrice de variances/covariances de X la matrice centrée au carrée.

La diagonale de cette matrice contient les variances de chaque variable étudiée et les autres valeurs sont des covariances.

d- Matrice de corrélations :

On appelle matrice de corrélations R de X la matrice centrée réduite au carrée.

La diagonale de cette matrice contient des 1 et les autres valeurs représentent les coefficients de corrélation entre les variables deux à deux. Ainsi, R résume la structure des dépendances linéaires entre les p variables.

***e-* La diagonalisation de la matrice de corrélation :**

On diagonalise la matrice de corrélation pour obtenir les valeurs propres en calculant le déterminant de la matrice suivante : $|\mathrm{cor}(X) - \lambda I)| = 0$

La résolution de polynôme caractéristique issu de calcul de déterminant donne les valeurs propres λi.

Chaque valeur propre nous renseigne sur la quantité d'information située sur chaque axe factoriel.

8-6- La construction des espaces factoriels :

Définir l'espace factoriel revient à :

- Définir *q* nouvelles variables comme axes du repère du nuage de points-individus : les composantes principales.
- Définir q nouveaux individus comme axes du repère du nuage de points-variables.

a- L'espace des individus :

L'analyse du nuage de point utilise la notion fondamentale de distance. On munit l'espace des individus de la distance euclidienne classique.

Le principe de la méthode est d'obtenir une représentation approchée du nuage des n individus dans un sous-espace de dimension faible.

Le choix de l'espace de projection s'effectue selon le critère suivant qui revient à déformer le moins possible les distances en projection. Le sous-espace de dimension k recherché est tel que la moyenne des carrés des distances entres projections soit le plus grand possible, en d'autres termes il faut que l'inertie du nuage projeté sur le sous-espace H_k soit maximale. Autrement dit, *L'inertie* est la quantité réelle qui mesure la dispersion des individus dans le nouvel espace à p dimensions.

b- L'espace des variables :

Changement d'origine : g = 0 (centrage des variables). La recherche des sous-espaces H_k se fait de proche en proche pour k=1 à p.

K : les valeurs propres et p : le nombre de variables initiales.

La détermination de H1 revient à chercher une droite passant par l'origine qui s'ajuste le mieux au nuage de points-individus (maximisant l'inertie expliquée). La détermination de cette droite, revient à déterminer un vecteur unitaire u_1 porté par cette droite avec $d(0,u_1)=1$.

Une fois u_1 déterminé, on peut démontrer que le sous-espace H1 s'ajustant au mieux au nuage de points contenant nécessairement u_1. Pour déterminer le sous-espace H2, on recherche u_2 tel que u_2 **perpendiculaire** à $u_1$et tel que la droite portée par u2, passant par0, et ait une inertie maximale. On peut démontrer que le sous-espace H3 contient nécessairement u1 et u2…

Les vecteurs $u_1, u_{2,\ldots,} u_p$ peuvent s'obtenir à partir de la matrice d'inertie C (covariance ou corrélation) entre les variables du tableau.

Cette matrice est telle qu'il existe p vecteurs et λ constantes qui vérifient l'équation matricielle suivante *: $C.v = \lambda$.*

Les p vecteurs sont les vecteurs propres et les constantes λ associées sont les valeurs propres. Ces vecteurs sont orthogonaux deux à deux et unitaires (de longueur égale à 1). Ils peuvent être rangés par ordre décroissant des valeurs propres associées : le premier vecteur propre v_1 est associé à la valeur propre la plus élevé λ1.

Ces vecteurs sont les vecteurs u_1 à u_p recherchés.

Les droites engendrées par ces vecteurs propres sont appelées respectivement le 1er, 2ème, et pième axe principal d'inertie du nuage.

8-7- Etude de cas : Consommation du gaz naturel dans le secteur résidentiel

L'objet de cette étude est de faire une analyse descriptive des principaux déterminants de la consommation du gaz naturel dans le secteur résidentiel, et ce durant la période allant de 1990 jusqu'à 2012.

a- Statistiques descriptives (analyse univariée) :

Tableau 8-1 : Statistiques descriptives

Variable	Min	Max	Moyenne	Ecart-type	CV
REV	104029,27	1258164,11	542778,83	315852,18	58,19
PE	3,46	31,84	21,75	10,45	48,08
PG	0,07	0,74	0,52	0,25	47,10
PGO	11041,67	16309,52	13545,87	2207,27	16,29
COEL	314520,07	1412034,00	679604,16	290660,20	42,77
COGN	797054,71	5350950,00	2065608,83	1203476,62	58,26
COGO	1349,03	125205,08	69002,89	48673,14	70,54

Avec : **Min :** Minimum ; **Max :** Maximum ; **Moy :** Moyenne ; **CV :** Coefficient de variation ; REV : Revenu ; PE : Prix d'électricité ; PG : Prix du gaz naturel ; PGO : prix du gazoeile ; COEL : consommation d'électricité ; COGN : Consommation du gaz naturel et COGO : Consommation du gazole.

Les résultats de l'analyse univariée permettent d'apprécier l'importance voire le poids de chaque variable dans notre analyse.

On en déduit alors à partir de ces résultats de statistiques descriptive que la variable consommation du gaz naturel a le poids le plus important dans notre base de données, parce qu'elle détienne la moyenne la plus élevée et on trouve en deuxième et troisième position les variables consommation d'électricité et revenu des ménages.

Quant à la représentativité des moyennes des variables, on déduit que les variables : PE, REV, COEL et COGN sont *moyennement représentés* du fait de leur coefficient de variation[1] qui est moyen, c'est à dire tournent autour de 50%.

Cependant, nous avons la variable COGO dont la moyenne est non représentative, avec un coefficient de variation de 70,54%. Enfin la variable ayant la moyenne la plus représentative dans notre base de données est le prix du gazole (PGO), elle se caractérise avec une moyenne élevée et un coefficient de variation faible (16,29%).

b- ***Matrice de corrélation (analyse bivariée) :***

Elle nous indique le niveau des relations entre les variables prises deux à deux.

Tableau 8-2 ***:*** *Matrice de corrélation (résultat ACP).*

Variables	**REV**	**PE**	**PG**	**PGO**	**COEL**	**COGN**	**COGO**
REV	**1**	0,836	0,815	0,841	0,980	0,962	-0,929
PE	0,836	**1**	0,998	0,753	0,843	0,682	-0,862
PG	0,815	0,998	**1**	0,725	0,824	0,654	-0,839
PGO	0,841	0,753	0,725	**1**	0,846	0,826	-0,933
COEL	0,980	0,843	0,824	0,846	**1**	0,931	-0,921
COGN	0,962	0,682	0,654	0,826	0,931	**1**	-0,875
COGO	-0,929	-0,862	-0,839	-0,933	-0,921	-0,875	**1**

[1] Cet indicateur permet de mesurer la représentativité de l'indicateur moyenne arithmétique.

Le résultat de la matrice de corrélation indique qu'il existe de fortes corrélations positives entre toutes les variables prise deux à deux, exception faite pour la variable COGO qui a une corrélation négative avec l'ensemble des variables.

Plus précisément, la forte corrélation positive (corrélation supérieur 0.9) se manifeste pour les couples de variables suivants : (PG, PE), (REV, COEL), (COGN, COEL). Quant à la corrélation négative elle se manifeste pour les couples des variables suivants : (COGO, REV), (COGO, PE), (COGO, PGO).

c- Applicabilité de l'ACP (Test de Bartlett) :

Le test d Bartlett sert à comparer la matrice de corrélation avec la matrice identité, pour voir si les deux matrices se ressemblent.

Tableau 8-3 : Test de Bartlett (Résultat ACP) :

Khi² (Valeur observée)	363,015
Khi² (Valeur critique)	32,671
DDL	21
p-value	< 0,0001
Alpha	0,05

Interprétation du test :

H_0: Il n'y a pas de corrélation significativement différente de 0 entre les variables.
H_1: Au moins l'une des corrélations entre les variables est significativement différente de zéro.

Etant donné que la p-value calculée est inférieure au risque d'erreur alpha=0,05, on doit rejeter l'hypothèse nulle H_0, et retenir l'hypothèse alternative H_1.

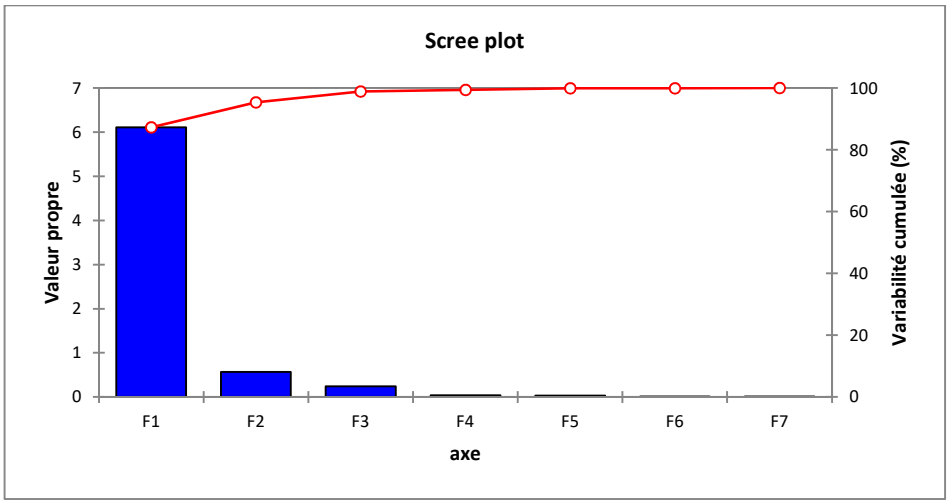
d- Valeurs propres :

Les valeurs représentent la masse d'information expliquée par chaque axe (nouvelle variable extraite à partir de la base de données de départ).

Tableau 8-4 : Valeurs propres (Résultat ACP).

	F1	F2	F3	F4	F5	F6	F7
Valeur propre	6,114	0,564	0,244	0,041	0,031	0,006	0,001
Variabilité (%)	87,348	8,052	3,479	0,580	0,449	0,083	0,010
% cumulé	87,348	95,400	98,879	99,459	99,907	99,990	100,000

Figure 8-1 : Valeurs propres.

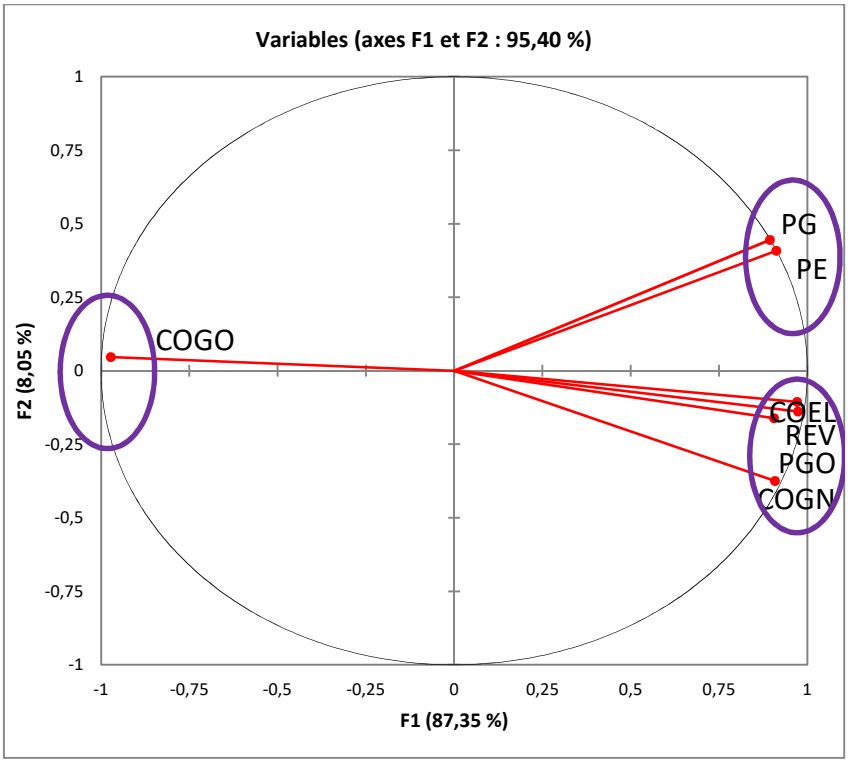


A partir du graphique, on peut déduire le nombre d'axes à prendre en considération pour l'analyse. Le nombre d'axes représente le nombre de points jusqu'au point d'inflexion.

Dans notre cas, le nombre de points est arrêté à « 02 », soit l'équivalent de deux axes, ce résultat est confirmé par le tableau des valeurs propres qui correspond à la colonne F2 et du cumule de la variabilité, soit un peu plus de 95% de l'information qui est prise en charge par les deux premiers axes.

e- L'analyse multivariée des variables :

Figure 8-2 : Analyse multivariée des variables.



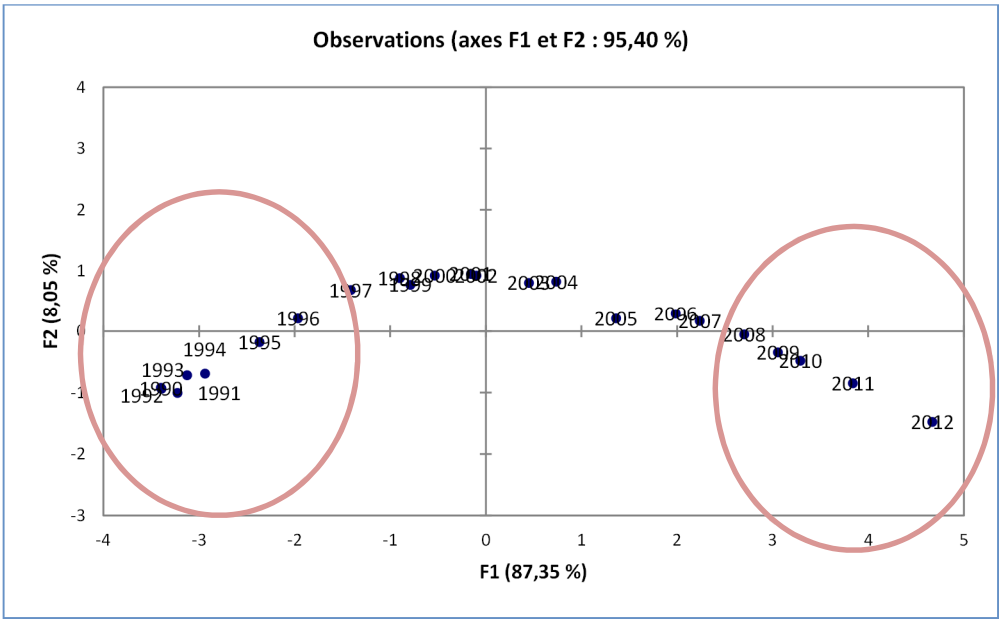
Comme le montre le graphique ci-dessus, on distingue trois groupes de variables fortement corrélées entre elles et opposent en même temps les autres groupes de variables.

- Groupe (1) : composé des variables suivantes : COEL, PGO, REV et COGN.
- Groupe (2) : contient les variables suivantes : PG et PE.
- Groupe (3) : représenté uniquement par la variable COGO.

Interprétation : l'interprétation se fait par rapport aux deux axes (F1 et F2).
Par rapport à l'axe 1 (F1) : les variables PGO, REV, COEL, PG, PE et COGN sont corrélés positivement entre elles et négativement avec la variable COGO.
Par rapport à l'axe 2 (F2) : les variables COEL, PGO, REV et COGN sont corrélés négativement avec les variables PG et PE.

f- Analyse multivariée des individus :

Figure 8-3 : analyse multivariée des individus.



L'analyse multivariée des individus, consiste à calquer le graphique des valeurs propres sur celui des individus afin d'en tirer une interprétation optimale.

De ce fait, on peut procéder à l'interprétation suivante :

L'analyse multivariée des individus dégage une courbe, celle-ci est caractérisée par une forte

La consommation du gasoil (COGO) durant la période (1991-1997) et une faible consommation des autres produits énergétiques (COGN, COEL), ainsi que de faibles valeurs pour le revenu/ménage et les prix des produits énergétiques étaient relativement faible dans cette période.

Contrairement, durant la période 2008-2012 où on assiste clairement à un changement du comportement du consommateur. En effet, on observe un effet de substitution du gasoil par l'électricité et le gaz. Egalement, cette période se caractérise par l'augmentation du revenu/ménage ainsi que l'augmentation des prix des produits énergétiques.

Tous ces changements peuvent être liés à certains facteurs dont les principales sont les suivants :

1. Augmentation des salaires, ce qui entraîne une augmentation du revenu des ménages.

2. Compte tenu des subventions sur les prix des produits énergétiques, la consommation sur ces derniers continuent à accroitre.
3. Pour des motifs d'ordre environnementaux, la consommation du gasoil a été remplacée par la consommation de l'électricité et du gaz.
4. Les ménages utilisent des équipements plus adaptés au développement de la société, ce qui favorise l'utilisation de certains équipements à usage électrique.

EXERCICE D'APPLICATION DU CHAPITRE 8 :

Exercice 1 :

Une étude portant sur les avantages fondamentaux recherchés par les consommateurs lors de l'achat d'un dentifrice.

Tableau de données de l'enquête

N°	V1	V2	V3	V4	V5	V6
1	7	3	6	4	2	4
2	1	3	2	4	5	4
3	6	2	7	4	1	3
4	4	5	4	6	2	5
5	1	2	2	3	6	2
6	6	3	6	4	2	4
7	5	3	6	3	4	3
8	6	4	7	4	1	4
9	3	4	2	3	6	3
10	2	6	2	6	7	6
11	6	4	7	3	2	3
12	2	3	1	4	5	4
13	7	2	6	4	1	3
14	4	6	4	5	3	6
15	1	3	2	2	6	4
16	6	4	6	3	3	4
17	5	3	6	3	3	4
18	7	3	7	4	1	4
19	2	4	3	3	6	3
20	3	5	3	6	4	6
21	1	3	2	3	5	3
22	5	4	5	4	2	4
23	2	2	1	5	4	4
24	4	6	4	6	4	7
25	6	5	4	2	1	4
26	3	5	4	6	4	7
27	4	4	7	2	2	5
28	3	7	2	6	4	3
29	4	6	3	7	2	7
30	2	3	2	4	7	2

Le sondage est réalisé dans un centre commercial, avec un échantillon de 30personnes, qui ont donné leur avis sur les affirmations suivantes, sur une échelle de 1 à 7 (1 : en total désaccord ; 7 : entièrement d'accord).

V_1 : Il est important d'utiliser un dentifrice qui prévient la formation des caries
V_2 : Un dentifrice doit rendre les dents brillantes ;
V_3 : Un dentifrice doit renforcer les gencives ;
V_4 : Un dentifrice doit rafraichir l'haleine ;
V_5 : La prévention contre les caries n'est pas un avantage important du dentifrice ;
V_6 : Un dentifrice doit, avant tout, donner de belles dents.

Question :

Appliquez la méthode d'Analyse en Composante Principe "ACP" sur le tableau ci-dessus, et interprétez les résultats obtenus.

Exercice 2 :

Lors d'une étude sur la relation entre le comportement au sein du foyer et comportement lié aux achats, on a obtenu des données sur les rapports de styles de vie suivants sur une échelle en sept points (1= n'est pas d'accord, 7= est d'accord) :

V_1 : Je préfère passer une soirée tranquille à la maison plutôt que d'aller à la fête ;

V_2 : Je préfère toujours les prix, même sur les petits articles ;

V_3 : Les magazines sont plus intéressants que les films ;

V_4 : Je n'achète pas les produits affichés sur les panneaux publicitaires ;

V_5 : Je suis casanier (qui aime rester à la maison) ;

V_6 : Je garde et utilise les coupons de réduction ;

V_7 : Les entreprises gaspillent beaucoup d'argent en publicité.

Les données obtenues à partir d'un échantillon de 25 individus sont regroupées ci-après :

N°	V1	V2	V3	V4	V5	V6	V7
1	6	2	7	6	5	3	5
2	5	7	5	6	6	6	4
3	5	3	4	5	6	6	7
4	3	2	2	5	1	3	2
5	4	2	3	2	2	1	3
6	2	6	2	4	3	7	5
7	1	3	3	6	2	5	7
8	3	5	1	4	2	5	6
9	7	3	6	3	5	2	4
10	6	3	3	4	4	6	5
11	6	6	2	6	4	4	7
12	3	2	2	7	6	1	6
13	5	7	6	2	2	6	1
14	6	3	5	5	7	2	3
15	3	2	4	3	2	6	5
16	2	7	5	1	4	5	2
17	3	2	2	7	2	4	6
18	6	4	5	4	7	3	3
19	7	2	6	2	5	2	1
20	5	6	6	3	4	5	3
21	2	3	3	2	1	2	6
22	3	4	2	1	4	3	6
23	2	6	3	2	1	5	3
24	6	5	7	4	5	7	2
25	7	6	5	4	6	5	3

Faites une Analyse en Composante Principale "ACP" sur ces données ?

Exercice 3 :

Afin d'étudier les caractéristiques de 24 marques de voiture, nous disposons de la base de données suivantes.

Numéro	*Modèle*	*Cylindrée*	*Puissance*	*Vitesse*	*Poids*	*Longueur*	*Largeur*
1	HONDA Civic	1396	90	174	850	369	166
2	RENAULT 19	1721	92	180	965	415	169
3	FIAT Tipo	1580	83	170	970	395	170
4	PEUGEOT 405	1769	90	180	1080	440	169
5	RENAULT 21	2068	88	180	1135	446	170
6	CITROEN BX	1769	90	182	1060	424	168
7	BMW 530i	2986	188	226	1510	472	175
8	ROVER 827i	2675	177	222	1365	469	175
9	RENAULT 25	2548	182	226	1350	471	180
10	OPEL Omega	1998	122	190	1255	473	177
11	PEUGEOT 405	1905	125	194	1120	439	171
12	FORD Sierra	1993	115	185	1190	451	172
13	BMW 325iX	2494	171	208	1300	432	164
14	AUDI 90 Quattro	1994	160	214	1220	439	169
15	FORD Scorpio	2933	150	200	1345	466	176
16	RENAULT Espace	1995	120	177	1265	436	177
17	NISSAN Vanette	1952	87	144	1430	436	169
18	VW Caravelle	2109	112	149	1320	457	184
19	FORD Fiesta	1117	50	135	810	371	162
20	FIAT Uno	1116	58	145	780	364	155
21	PEUGEOT 205	1580	80	159	880	370	156
22	PEUGEOT 205	1294	103	189	805	370	157
23	SEAT Ibiza SX I	1461	100	181	925	363	161
24	CITROEN AX Sport	1294	95	184	730	350	160

Moyenne et écart-type des colonnes

	Moyenne	Ecart-type
Cylindrée	1906,125	516,794
Puissance	113,667	37,968
Vitesse	183,083	24,685
Poids	1110,833	225,442
Longueur	421,583	40,470
Largeur	168,833	7,493

Matrice de corrélation :

	Cylindrée	Puissance	Vitesse	Poids	Longueur	Largeur
Cylindrée	1	**0,861**	**0,693**	**0,905**	**0,864**	**0,709**
Puissance	**0,861**	1	**0,894**	**0,746**	**0,689**	**0,552**
Vitesse	**0,693**	**0,894**	1	**0,491**	**0,532**	0,363
Poids	**0,905**	**0,746**	**0,491**	1	**0,917**	**0,791**
Longueur	**0,864**	**0,689**	**0,532**	**0,917**	1	**0,864**

Largeur	**0,709**	**0,552**	0,363	**0,791**	**0,864**	1

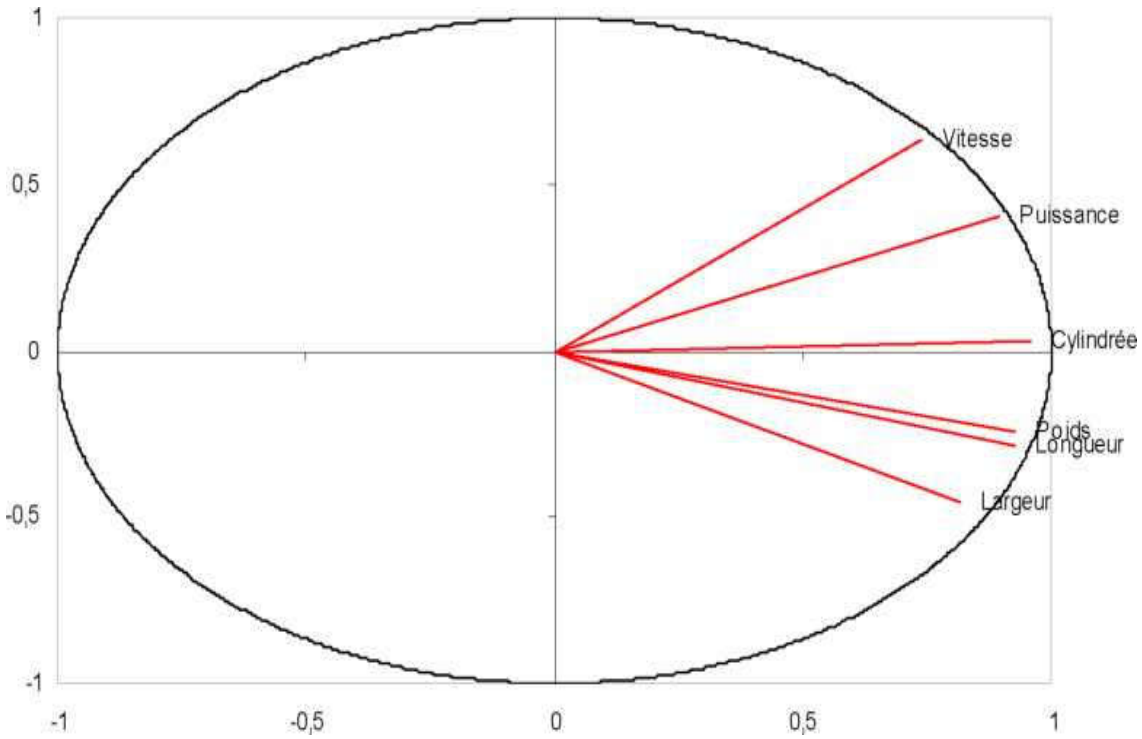
Test de sphéricité de Bartlett :

Khi^2 (valeur observée)	178,583
Khi^2 (valeur critique)	24,996
ddl	15
p-value unilatérale	< 0.0001
Alpha	0,05

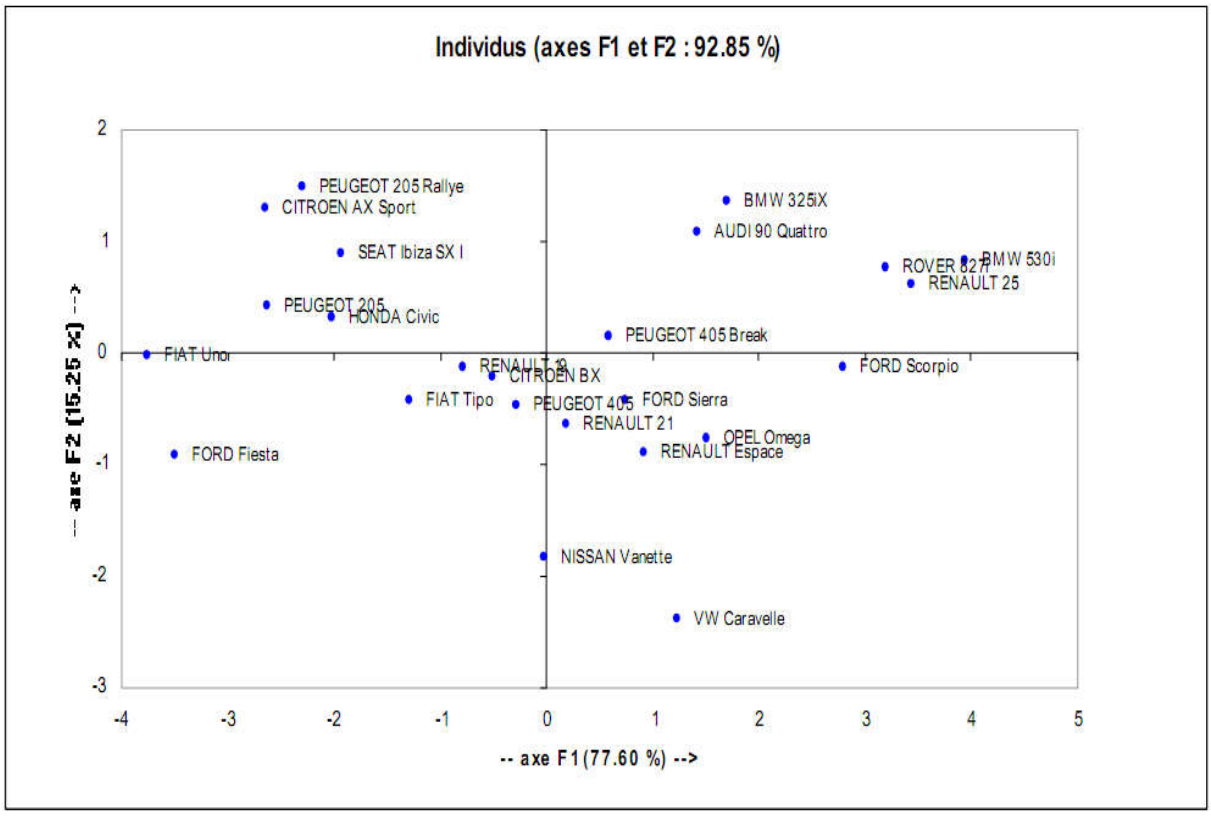
Valeurs propres

	F1	F2	F3	F4	F5	F6
Valeur propre	4,656	0,915	0,240	0,103	0,065	0,021
% variance	77,600	15,254	4,007	1,712	1,078	0,349
% cumulé	77,600	92,854	96,861	98,573	99,651	100,00

Représentation des variables (axes F1 et F2 : 92.85 %)



Position des individus :



Questions :

1. Commenter les tableaux ci-dessus ?
2. L'ACP utilisée est-elle normée ?
3. Quel est le pourcentage d'inertie du 1er plan factoriel ?
4. Quelle est la qualité de la représentation du 1erplan factoriel ? Avons-nous besoin de l'axe 3 ?
5. Quels sont les individus et les variables qui contribuent le plus à la construction du 1er et du 2ème axe factoriel ?
6. Renault 25 et Fiat Tipo Sont-t-ils bien représenté sur l'Axe1 ? et Sur l'axe 2 ?

CONCLUSION DE LA PARTIE 3 :

Les méthodes d'analyses de données multivariées sont nombreuses et permettent de de synthétiser l'information portant sur des bases de données de grandes tailles en quelques résultats très facile à lire et à expliquer. Elles sont très élaborées et elles font appelle souvent à des représentations géométriques sur des plans (espaces vectoriels de deux dimensions).

L'analyse en composante principale « ACP » est la méthode de base dans la mesure où les autres méthodes sont des prolongements de cette dernière.

Par ailleurs, l'ACP présente certaines limites, notamment son application sur des données quantitatives ou qualitatives ordinaires. Afin de pallier à cette limite, d'autre méthodes ont été développés, tels que l'analyse factorielle des correspondances, l'analyse discriminante, l'analyse des correspondances multiples…

CONCLUSION GENERALE

Le présent support de cours est le fruit de presque une dizaine d'années d'enseignement aux profils des étudiants universitaires (à l'Université de Mouloud Mammeri de Tizi Ouzou, à l'EHEC (Ex-INC), à l'ENSM) ainsi que pour le compte des personnels des entreprises économique (publiques et privés) et d'administration publiques.

Par ailleurs, le recours aux méthodes d'analyse de données est de nos jours très indispensable pour donner un sens à l'action prise. En effet, les méthodes d'analyse de données univariées permettent de décrire une seule variable statistique (présentation tabulaires, graphiques et synthèses numériques) et les méthodes d'analyse de données bivariées, quant à elles, permettent de traiter le système de relation qui existe entre deux variables, deux à deux mutuellement.

La méthode d'Analyse en Composante Principale (ACP) est une méthode d'analyse de données multivariée qui a pour but de faire une analyse descriptive d'une base de données comportant plusieurs variables (quantitatives ou qualitatives ordinales) et qui sont mesurées sur plusieurs individus. Cependant, la présence de bases de données avec des variables quantitatives et qualitatives et l'essor de l'outil informatique (logiciels spécialisés dans le traitement de données statistiques tels que : SPSS, Xlstat et Stata) nous amènent à traiter dans la prochaine édition de ce cours les autres méthodes d'analyse de données multivariées, tels que : l'AFC, l'ACM, la CAH, la régression multiple, la régression logistique...).

ANNEXES

ANNEXES 1 : TABLES STATISTIQUES

ANNEXE 2 : APERÇU THEORIQUE SUR LES TESTS DE CONFORMITE « Z »

ANNEXE 3 : APERÇU THEORIQUE SUR LES TESTS D'HOMOGENEITE «TEST t»

ANNEXES 1 : TABLES STATISTIQUES

Annexe 1-1 : Table de Khi-deux.

$$P(\chi_\nu^2 \geq \chi_{\nu,\alpha}^2) = \alpha$$

1 – α	0,001	0,005	0,01	0,025	0,05	0,1	0,5	0,9	0,95	0,975	0,99	0,995	0,999
α	0,999	0,995	0,99	0,975	0,95	0,9	0,5	0,1	0,05	0,025	0,01	0,005	0,001
ν = ddl													
1	0,00	0,00	0,00	0,00	0,00	0,02	0,45	2,71	3,84	5,02	6,63	7,88	10,83
2	0,00	0,01	0,02	0,05	0,10	0,21	1,39	4,61	5,99	7,38	9,21	10,60	13,82
3	0,02	0,07	0,11	0,22	0,35	0,58	2,37	6,25	7,81	9,35	11,34	12,84	16,27
4	0,09	0,21	0,30	0,48	0,71	1,06	3,36	7,78	9,49	11,14	13,28	14,86	18,47
5	0,21	0,41	0,55	0,83	1,15	1,61	4,35	9,24	11,07	12,83	15,09	16,75	20,51
6	0,38	0,68	0,87	1,24	1,64	2,20	5,35	10,64	12,59	14,45	16,81	18,55	22,46
7	0,60	0,99	1,24	1,69	2,17	2,83	6,35	12,02	14,07	16,01	18,48	20,28	24,32
8	0,86	1,34	1,65	2,18	2,73	3,49	7,34	13,36	15,51	17,53	20,09	21,95	26,12
9	1,15	1,73	2,09	2,70	3,33	4,17	8,34	14,68	16,92	19,02	21,67	23,59	27,88
10	1,48	2,16	2,56	3,25	3,94	4,87	9,34	15,99	18,31	20,48	23,21	25,19	29,59
11	1,83	2,60	3,05	3,82	4,57	5,58	10,34	17,28	19,68	21,92	24,73	26,76	31,26
12	2,21	3,07	3,57	4,40	5,23	6,30	11,34	18,55	21,03	23,34	26,22	28,30	32,91
13	2,62	3,57	4,11	5,01	5,89	7,04	12,34	19,81	22,36	24,74	27,69	29,82	34,53
14	3,04	4,07	4,66	5,63	6,57	7,79	13,34	21,06	23,68	26,12	29,14	31,32	36,12
15	3,48	4,60	5,23	6,26	7,26	8,55	14,34	22,31	25,00	27,49	30,58	32,80	37,70
16	3,94	5,14	5,81	6,91	7,96	9,31	15,34	23,54	26,30	28,85	32,00	34,27	39,25
17	4,42	5,70	6,41	7,56	8,67	10,09	16,34	24,77	27,59	30,19	33,41	35,72	40,79
18	4,90	6,26	7,01	8,23	9,39	10,86	17,34	25,99	28,87	31,53	34,81	37,16	42,31
19	5,41	6,84	7,63	8,91	10,12	11,65	18,34	27,20	30,14	32,85	36,19	38,58	43,82
20	5,92	7,43	8,26	9,59	10,85	12,44	19,34	28,41	31,41	34,17	37,57	40,00	45,31
21	6,45	8,03	8,90	10,28	11,59	13,24	20,34	29,62	32,67	35,48	38,93	41,40	46,80
22	6,98	8,64	9,54	10,98	12,34	14,04	21,34	30,81	33,92	36,78	40,29	42,80	48,27
23	7,53	9,26	10,20	11,69	13,09	14,85	22,34	32,01	35,17	38,08	41,64	44,18	49,73
24	8,08	9,89	10,86	12,40	13,85	15,66	23,34	33,20	36,42	39,36	42,98	45,56	51,18
25	8,65	10,52	11,52	13,12	14,61	16,47	24,34	34,38	37,65	40,65	44,31	46,93	52,62
26	9,22	11,16	12,20	13,84	15,38	17,29	25,34	35,56	38,89	41,92	45,64	48,29	54,05
27	9,80	11,81	12,88	14,57	16,15	18,11	26,34	36,74	40,11	43,19	46,96	49,65	55,48
28	10,39	12,46	13,56	15,31	16,93	18,94	27,34	37,92	41,34	44,46	48,28	50,99	56,89
29	10,99	13,12	14,26	16,05	17,71	19,77	28,34	39,09	42,56	45,72	49,59	52,34	58,30
30	11,59	13,79	14,95	16,79	18,49	20,60	29,34	40,26	43,77	46,98	50,89	53,67	59,70

Pour $\nu > 30$, La loi du χ^2 peut –être approximée par la loi normale $N(\nu, \sqrt{\nu})$

Annexe 1-2 : Table de Student

(Valeurs de T ayant la probabilité P d'être dépassée en valeur absolue)

DL / P	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,929
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,263	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0 537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0 688	0,961	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,137	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0 530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,649
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,656
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
80	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
Infini	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Annexe 1-3 : Table de Fisher.

$$P(F_{v_1,v_2} < \mathsf{f}_{v_1,v_2,\alpha}) = \alpha$$

$\alpha = 0{,}95$

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	100	200	500	•
1	161	200	216	225	230	234	237	239	241	242	246	248	250	252	253	254	254	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,70	8,66	8,62	8,58	8,55	8,54	8,53	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,86	5,80	5,75	5,70	5,66	5,65	5,64	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,50	4,44	4,41	4,39	4,37	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,81	3,75	3,71	3,69	3,68	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,44	3,38	3,32	3,27	3,25	3,24	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,08	3,02	2,97	2,95	2,94	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,86	2,80	2,76	2,73	2,72	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,70	2,64	2,59	2,56	2,55	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,72	2,65	2,57	2,51	2,46	2,43	2,42	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,62	2,54	2,47	2,40	2,35	2,32	2,31	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,53	2,46	2,38	2,31	2,26	2,23	2,22	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,46	2,39	2,31	2,24	2,19	2,16	2,14	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,25	2,18	2,12	2,10	2,08	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,35	2,28	2,19	2,12	2,07	2,04	2,02	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,31	2,23	2,15	2,08	2,02	1,99	1,97	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,27	2,19	2,11	2,04	1,98	1,95	1,93	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,23	2,16	2,07	2,00	1,94	1,91	1,89	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	2,04	1,97	1,91	1,88	1,86	1,84
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,15	2,07	1,98	1,91	1,85	1,82	1,80	1,78
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,11	2,03	1,94	1,86	1,80	1,77	1,75	1,73
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,07	1,99	1,90	1,82	1,76	1,73	1,71	1,69
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,04	1,96	1,87	1,79	1,73	1,69	1,67	1,65
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,01	1,93	1,84	1,76	1,70	1,66	1,64	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,92	1,84	1,74	1,66	1,59	1,55	1,53	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,87	1,78	1,69	1,60	1,52	1,48	1,46	1,44
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,84	1,75	1,65	1,56	1,48	1,44	1,41	1,39
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,79	1,70	1,60	1,51	1,43	1,38	1,35	1,32
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,77	1,68	1,57	1,48	1,39	1,34	1,31	1,28
200	3,89	3,04	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88	1,72	1,62	1,52	1,41	1,32	1,26	1,22	1,19
500	3,86	3,01	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,69	1,59	1,48	1,38	1,28	1,21	1,16	1,11
•	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,67	1,57	1,46	1,35	1,24	1,17	1,11	1,00

ANNEXE 2 :

APERÇU THEORIQUE SUR LES TESTS DE CONFORMITE « Z »

Annexe 2-1- Test de conformité d'une moyenne par rapport à une norme :

1-1- La variance de la PM est connue :

Hypothèses	Statistique du test	Décision
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$Z= \frac{\mu - \mu 0}{\rho/\sqrt{n}}$	RH_0 si $Z \geq \mu_{\alpha/2}$ $\bar{R}H_0$ si $Z < \mu_{\alpha/2}$
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	$Z= \frac{\mu - \mu 0}{\rho/\sqrt{n}}$	RH_0 si $Z \geq \mu_{\alpha}$ $\bar{R}H_0$ si $Z < \mu_{\alpha}$
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$Z= \frac{\mu - \mu 0}{\rho/\sqrt{n}}$	RH_0 si $Z < \mu_{\alpha}$ $\bar{R}H_0$ si $Z \geq \mu_{\alpha}$

μ_α : Valeur critique lue sur la table de la loi normale centrée réduite au risque d'erreur α ou (α/2)%.

1-2- La variance de la PM est inconnue :

Hypothèses	Statistique du test	Décision
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$Z= \frac{\mu - \mu 0}{\delta/\sqrt{n-1}}$	RH_0 si $Z \geq t_{\alpha/2\ ;\ n-1}$ $\bar{R}H_0$ si $Z < t_{\alpha/2}$
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	$Z= \frac{\mu - \mu 0}{\delta/\sqrt{n-1}}$	RH_0 si $Z \geq t_{\alpha/2}$ $\bar{R}H_0$ si $Z < t_{\alpha/2}$
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$Z = \frac{\mu - \mu 0}{\delta/\sqrt{n-1}}$	RH_0 si $Z < t_{\alpha/2}$ $\bar{R}H_0$ si $Z \geq t_{\alpha/2}$

δ : représente l'écart type de l'échantillon.

$t_{\alpha/2}$: valeur critique lu sur la table de Student au risque d'erreur α/2 et un degré de liberté (n-1);

Remarque : *On met toujours les résultats de la statistique calculée « Z » en valeur absolue.*

Annexe 2-2- Test de conformité d'une proportion :

Hypothèses	**Statistique du test**	**Décision**
$H_0 : P = f$ $H_1 : P \neq f$	$Z= \frac{P-f}{\sqrt{f(1-f)/n}}$	RH_0 si $Z \geq U_{\alpha/2}$ $\bar{R}H_0$ si $Z < U_{\alpha/2}$
$H_0 : p= f$ $H_1 : P > f$	$Z= \frac{P-f}{\sqrt{f(1-f)/n}}$	RH_0 si $Z \geq U_{\alpha}$ $\bar{R}H_0$ si $Z < U_{\alpha}$
$H_0 : p= f$ $H_1 : P < f$	$Z= \frac{P-f}{\sqrt{f(1-f)/n}}$	RH_0 si $Z < U_{\alpha}$ $\bar{R}H_0$ si $Z \geq U_{\alpha}$

Annexe 2-3- Test de conformité d'une variance :

Hypothèses	**Statistique du test**	**Décision**
$H_0 : \sigma_e^2 = \sigma_0^2$ $H_1 : \sigma_e^2 \neq \sigma_0^2$	$\chi 2 = \frac{(n-1)\sigma e2}{\sigma 2}$	RH_0 si $\chi^2 \geq U_{\alpha/2}$ $\bar{R}H_0$ si $\chi^2 < U_{\alpha/2}$
$H_0 : \sigma_e^2 = \sigma_0^2$ $H_1 : \sigma_e^2 > \sigma_0^2$	$\chi 2 = \frac{(n-1)\sigma e2}{\sigma 2}$	RH_0 si $\chi^2 \geq U_{\alpha}$ $\bar{R}H_0$ si $\chi^2 < U_{\alpha}$
$H_0 : \sigma_e^2 = \sigma_0^2$ $H_1 : \sigma_e^2 < \sigma_0^2$	$\chi 2 = \frac{(n-1)\sigma e2}{\sigma 2}$	RH_0 si $\chi^2 < U_{\alpha}$ $\bar{R}H_0$ si $\chi^2 \geq U_{\alpha}$

ANNEXE 3 :

APERÇU THEORIQUE SUR LES TESTS D'HOMOGENEITE « TEST "T" »

Annexe 3-1 : Test d'homogénéité de deux populations (Test de Student 't'):

Soient deux échantillons indépendants de tailles n_1 et n_2, et chacun des deux suit une loi normale : N (μ_1, σ_1), N (μ_2, $\sigma 2$).

H_0 : $\mu_1=\mu_2$

H1 : $\mu_1\neq\mu_2$

Statistique de test : $T = \left|\frac{\mu 1-\mu 2}{\frac{\sigma 1}{\sqrt{n1}}+\frac{\sigma 2}{\sqrt{n2}}}\right|$

Règle de décision : Si n_1 et n_2 sont inférieurs à 30, on compare la statistique t avec la statistique lue sur la table de Student. Et si sont supérieurs à 30, on compare la statistique t avec la statistique lue sur la table de la loi normale centrée réduite.

Si le caractère étudié est un caractère qualitatif, on utilise la statistique Z suivante :

$$t=\left|\frac{P1-P2}{\sqrt{Pq(\frac{1}{n1}+\frac{1}{n2})}}\right| \quad \text{Où : } P=\frac{n1*P1+n2*P2}{n1+n2}$$

REFERENCES BIBLIOGRAPHIQUES :

- ✓ Bourbonnés, R (2005), Manuel d'Econométrie, Edition Dunod.
- ✓ Cadoret, I ; Benjamin, C ; Martin, F ; Herrard, N et Tanguy, S (2009), Econométrie appliquée : Méthodes, Application et corrigés, Edition De Boeck.
- ✓ David, R. Anderson, (2015), Statistiques pour l'économie et la gestion, Ed De Boeck ;
- ✓ Escofier, P (1998), Analyses Factorielles Simples et Multiples, Objectifs, Méthodes et Interprétation, Dunod.
- ✓ Fenneteau, H et Bialès, C (1993), Analyse statistique des données : Applications et cas pour le Merketing, Ed Ellipses, Paris.
- ✓ Godelieve, M-S et Rafael, C (2013), Analyser les données en sciences sociales ; Ed P.I.E Peter Lang, Bruxelles.
- ✓ Goldfard, B et Pardoux, C (2011), Introduction à la méthode statistique; ed Dunod, 6eme édition, Paris.
- ✓ Hahn, C et Macé, S ; Méthodes statistiques appliquées au management, 2e Ed PEARSON.
- ✓ Hamrouni, A (2017), Analyse des données : Traitement statistiques des enquetes avec SPSS; Editions universitaires européennes; France.
- ✓ HUBLIER, J et RAIMBOURG, P (1996), Statistiques pour l'économie, Ed Bréal.
- ✓ Morineau, L-P (1995) : Statistique Exploratoire Multidimensionnelle, Dunod.
- ✓ Malhotra, N, Décaudin, J-M et Bouguerra, A. (2004), Études marketing avec SPSS. 4^{e} édition, Pearson Éducation.
- ✓ Py, B (2007) ; Statistique descriptive : Nouvelle méthode pour bien comprendre et réussir ; 5e éd Economica.
- ✓ Saporta, G (1990), Probabilités, Analyse des Données et Statistiques, Technip.
- ✓ Tenenhaus, M (1995) : Méthodes Scientifiques de Gestion, Dunod.
- ✓ Tuffery, S (2005), Data mining et statistique décisionnnelle, Editions Technip.