MINISTÈRE DE l'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ MOULOUD MAMMERI, TIZI-OUZOU

FACULTÉ DES SCIENCES, DÉPARTEMENT DE MATHÉMATIQUES

MEMOIRE DE FIN D'ETUDE

Spécialité : Mathématiques

Option :Mathématiques Appliquées aux Sciences de Gestion

Présenté par :

AIT DAHMANE Dihia
et
HATTAL Thafath

Thème

Exploitation des données d'entreprise en utilisant le logiciel SPSS Cas : Electro-Industrie

Encadré par :
Pr Oukacha Brahim

Devant le jury d'examen composé de :

Pr. AIDENE Mohamed	Professeur UMMTO	**Président**
Pr. OUKACHA Brahim	Professeur UMMTO	**Encadrant**
Dr. KOURAT Hocine	Professeur UMMTO	**Examinateur**
Mme. SADOUNE Hayat	Electro-Industrie	**Maitresse de stage**

Soutenue : le 03/07/2025

Remerciements

Nous remercions tout d'abord Dieu, source de toute sagesse et de force, qui nous a accompagnés et soutenus tout au long de ce parcours. C'est par Sa grâce que nous avons pu mener à bien ce travail.

Au terme de notre formation en mathématiques à la Faculté des Sciences de l'Université Mouloud Mammeri de Tizi-Ouzou, nous tenons à exprimer notre profonde gratitude à toutes les personnes qui ont contribué, de près ou de loin, à notre réussite. Nous adressons tout particulièrement nos remerciements à nos enseignants, et plus spécialement au professeur OUKACHA Brahim, dont l'encadrement et les conseils avisés ont été essentiels dans notre cursus.

Nous remercions également l'ensemble du personnel de l'entreprise Electro Industrie, et particulièrement Madame SADOUNE Hayat, qui nous a accueillies au sein de l'ENEL et nous a fourni tout ce qui est nécessaires à la réalisation de notre travail.

Dédicaces

Je dédie ce mémoire à mes parents, Akli et Ouiza, qui ont façonné ma vie bien plus par leurs actes que par leurs paroles. Leur dévouement, leur droiture et leur force silencieuse m'ont appris à persévérer, à respecter et à rêver sans renier mes racines.

À mes deux frères, à mes quatre sœurs, ainsi qu'aux époux de mes soeurs et à l'épouse de mon frère , pour leur soutien constant, leur solidarité sincère et l'harmonie familiale qu'ils savent préserver avec cœur.

À mes chers neveux et nièces, ma source de joie et de motivation quotidienne, je vous dédie ce mémoire avec tout mon amour, en vous souhaitant un avenir rempli de bonheur et de réussite.

À mes amis, pour leur sérieux, leur loyauté et leur présence précieuse tout au long de ce parcours exigeant.

À ma chère binôme Thafath, pour sa collaboration, son engagement et son soutien indéfectible durant ce travail commun.

Et enfin, à vous qui lisez ces pages, Que ce travail soit pour vous une ouverture vers la réflexion, comme il fut pour moi une expérience de construction et d'engagement.

Merci pour votre lecture et votre attention.

Dédicaces

Je dédie ce modeste travail à mon père, dont le soutien et les sacrifices ont été présents jusqu'à ses derniers jours. Sa confiance inébranlable en moi et tout ce qu'il a fait pour m'accompagner restent une source d'inspiration et de force.

À ma mère et à mes deux grandes sœurs, pour leur courage, leur amour et leur soutien indéfectible, qui m'ont toujours porté et donné la force de persévérer.

À mon petit frère, à qui je souhaite une grande réussite dans tous ses projets.

À toute ma famille, pour leur présence, leur compréhension et leur appui moral, qui m'ont permis de garder confiance et motivation.

À tous mes amis et à ma promotion MASG, pour les moments d'entraide, de partage et de persévérance qui ont jalonné notre parcours.

À vous qui lisez ce mémoire, j'espère que ce travail saura vous apporter autant qu'il m'a apporté.
Merci à vous

Table des matières

Introduction générale

Dans un contexte économique marqué par une concurrence accrue et une évolution rapide des marchés, les entreprises industrielles doivent s'appuyer sur une gestion rigoureuse et une exploitation optimale de leurs données pour assurer leur pérennité et leur développement. La capacité à anticiper les besoins futurs, notamment en matière de production et de gestion des stocks, est devenue un enjeu stratégique majeur. La prévision, en tant qu'outil d'aide à la décision, permet de mieux planifier les activités, d'optimiser les ressources et d'améliorer la réactivité face aux fluctuations de la demande.

L'entreprise Electro Industrie (ENEL), spécialisée dans la fabrication de transformateurs, moteurs électriques et groupes électrogènes, évolue dans un secteur où la précision et la fiabilité des prévisions sont essentielles pour garantir la qualité de la production et la satisfaction des clients. Dans ce cadre, la maîtrise des données issues des différents processus industriels et commerciaux représente un atout fondamental. Toutefois, la complexité et le volume croissant de ces données nécessitent l'utilisation d'outils performants d'analyse statistique et de modélisation.

C'est dans cette optique que ce mémoire s'intéresse à l'exploitation des données de l'entreprise ENEL à travers l'utilisation du logiciel IBM SPSS. Ce logiciel, reconnu mondialement pour ses capacités avancées en traitement statistique, offre une interface conviviale et une large gamme de fonctionnalités adaptées aux besoins des entreprises industrielles. SPSS permet non seulement de structurer et d'analyser les données, mais aussi de mettre en œuvre diverses méthodes de prévision, facilitant ainsi la prise de décision éclairée.

L'objectif principal de ce travail est d'illustrer comment SPSS peut être utilisé efficacement pour répondre à la problématique de prévision au sein d'ENEL. Pour cela, nous avons choisi de nous concentrer sur un modèle spécifique de transformateur fabriqué par l'entreprise, afin d'appliquer différentes méthodes de prévision et d'évaluer leur pertinence. Parmi les techniques exploitées, on retrouve la régression linéaire, qui permet d'étudier les relations entre variables explicatives et la variable à prévoir, ainsi que des modèles de séries temporelles tels que ARMA et ARIMA, qui sont particulièrement adaptés à l'analyse des données chronologiques. D'autres méthodes simples, comme les moyennes mobiles, ont également été utilisées pour comparer les résultats et affiner les prévisions.

Dans ce contexte, la problématique centrale qui guide ce mémoire est la suivante : comment l'entreprise Electro Industrie (ENEL) peut-elle exploiter au mieux ses données industrielles et commerciales à l'aide du logiciel IBM SPSS pour améliorer la fiabilité de ses prévisions de production, notamment pour un modèle spécifique de transformateur ? Plus précisément, quelles méthodes statistiques et modèles de prévision (régression linéaire, modèles ARMA/ARIMA, moyennes mobiles) sont les plus pertinents pour anticiper les besoins futurs et optimiser la gestion des stocks dans un environnement marqué par une forte variabilité de la demande ?

L'utilisation de SPSS dans ce contexte présente plusieurs avantages. D'une part, il facilite le traitement des données brutes en automatisant les calculs complexes et en fournissant des résultats sous forme de graphiques et de tableaux facilement interprétables. D'autre part, il offre une grande souplesse dans le choix des modèles statistiques, permettant d'adapter les analyses aux spécificités des données et aux objectifs de l'entreprise. Cette polyvalence fait de SPSS un outil précieux pour les ingénieurs, les analystes et les décideurs d'ENEL, qui peuvent ainsi mieux anticiper les évolutions du marché et optimiser leurs processus industriels.

En résumé, ce mémoire met en lumière l'importance de l'exploitation des données industrielles à travers des outils statistiques performants comme SPSS, en se focalisant sur la problématique cruciale de la prévision. À travers l'étude d'un cas concret au sein d'ENEL, il démontre comment l'analyse statistique peut contribuer à renforcer la compétitivité de l'entreprise en améliorant la qualité de ses prévisions et en soutenant une gestion proactive et efficace.

Chapitre 1

Rappels statistiques

1.1 introduction

Du fait de la variabilité, on est dans le domaine de l'incertain. Cette science de l'incertain, c'est le défi qu'a relevé la statistique en s'appuyant sur le concept de probabilité.

Dans ce chapitre nous présentons les notions essentielles de la statistique descriptive et inférentielle, apprend comment décrire de façon claire et concise l'information apportée par des observations nombreuses et variées sur un phénomène donné.

Il s'agit de trier ces données, les décrire, les résumer sous forme de tableaux, de graphiques, et sous forme d'un petit nombre de paramètres-clés (moyenne, médiane par exemple).

1.2 Généralités

1.2.1 Statistique descriptive et statistique inférentielle

De manière approximative, il est possible de classer les méthodes statistiques en deux groupes : celui des méthodes descriptives et celui des méthodes inférentielles.

La statistique descriptive. On regroupe sous ce terme les méthodes dont l'objectif est la description des données étudiées ; cette description des données se fait à travers leur présentation (la plus synthétique possible), leur représentation graphique, et le calcul de résumés numériques. Dans cette optique, il n'est pas fait appel à des modèles probabilistes. On notera que les termes de statistique descriptive, statistique exploratoire et analyse des données sont quasiment synonymes.

La statistique inférentielle. Ce terme regroupe les méthodes dont l'objectif principal est de préciser un phénomène sur une population globale, à partir de son observation

sur une partie restreinte de cette population ; d'une certaine manière, il s'agit donc d'induire (ou encore d'inférer) du particulier au général. Le plus souvent, ce passage ne pourra se faire que moyennant des hypothèses de type probabiliste. Les termes de statistique inférentielle, statistique mathématique, et statistique inductive sont eux aussi synonymes.

D'un point de vue méthodologique, la statistique descriptive précède en général la statistique inférentielle dans une démarche de traitement de données : les deux aspects se complètent bien plus qu'ils ne s'opposent

1.3 statistiques descriptives

1.3.1 Définitions :

On appellera :

1 /Individu : l'unité d'observation (exemples : entreprise, chaîne de production) ;

2/ Population : l'ensemble des individus concernés par l'étude (exemples : ensemble des entreprises algériennes, ensemble des pièces sortant de la chaîne) ;

3/ Échantillon : un sous-ensemble de la population dont les individus feront l'objet de l'étude. Le choix de l'échantillon se fait en respectant certaines règles ;

4/ Variable ou caractère statistique : l'aspect de l'unité statistique que l'on va étudier (exemples : situation géographique de l'entreprise, diamètre de la pièce...). On dira que cette variable prend des valeurs (ou modalités).[10]

1.3.2 Types de variables statistiques

On peut définir quatre classes (ou types) dans lesquelles se répartissent les variables statistiques selon la nature de leurs valeurs.

Les variables qualitatives comprennent :

- **Les variables catégorielles (ou nominales)**, qui n'ont pas de structure particulière.
  Exemples : sexe, nationalité, catégorie socioprofessionnelle, contrôle qualitatif d'une pièce, situation de famille.
- **Les variables ordinales**, qui sont ordonnées.
  Exemples : tout jugement qualitatif, mention à un examen.

Les variables quantitatives comprennent :

— **Les variables discrètes**, qui prennent des valeurs entières.
Exemples : nombre d'enfants, nombre de diplômes, poids.
— **Les variables continues**, qui prennent des valeurs réelles.
Exemples : température, fréquence d'un signal, amplitude d'un bruit thermique, valeur boursière.

1.3.3 Distributions statistiques. Effectifs, fréquences

Lorsque le recueil des données a été effectué, on dispose, pour chacun des individus de l'échantillon (ou de la population), de la valeur de la variable étudiée. Le premier traitement consiste alors à relever cette valeur pour chaque individu et ensuite à compter le nombre d'individus pour lesquels la variable prend une valeur donnée.

On associe, à chaque valeur prise par le caractère statistique étudié, son effectif.

Notation : les variables seront notées par des lettres majuscules X, Y, Z. . . ; on note leurs modalités (valeurs) par des lettres minuscules x_i , y_j, z_l et les effectifs associés par n_i , n_j, n_l

Exemple 1.1. *: X = sexe,* x_1*= féminin,* x_2 *= masculin,* n_1 *= nombre de femmes,* n_2 *= nombre d'hommes*

Ce traitement n'est bien sûr directement possible que pour les variables qualitatives ou discrètes, qui n'ont qu'un nombre limité de valeurs possibles, discernables entre elles. Pour les variables continues, on commence par ranger les observations en classes, celles-ci étant des intervalles de la forme $[a_{i-1}, a_i[$. Ensuite, pour chaque classe, on compte le nombre d'individus dont le caractère appartient à la classe : ce nombre est l'effectif de la classe. On note k le nombre de modalités.

Définition 1.1. on appellera distribution statistique des effectifs de la variable X :

L'ensemble des données (x_i, n_i), i = 1,..., k , si X est une variable qualitative ou discrète, L'ensemble des données $([a_{i-1}, a_i[, n_i)$, i = 1,...,k , si X est une variable continuen. Les résultats sont généralement présentés dans un tableau du type du tableau 1.2.[10]

Présentation des variables statistiques			
X est catégorielle, ordinale ou discrete		X est continue	
Modalités	Effectifs	Classes	Effectifs
x_1	n_1	$[a_0, a_1[$	n_1
x_2	n_2	$[a_1, a_2[$	n_2
.	.	.	.
.	.	.	.
x_k	n_k	$[a_{k-1}, a_k[$	n_k
Total (1)	$N = n_1+n_2+...+n_k$	*Total*	$N = n_1+n_2+...+n_k$
(1) N est l'effectif total de l'échantillon			

FIGURE 1.1 – présentation des variables statistiques

Définition 1.2. La fréquence (ou proportion) associée à la valeur du caractère (resp. à la Classe $[a_{i-1},\ a_i[$) est la valeur f_i définie par : $f_i = n_i/N$

La fréquence f_i représente donc la part de l'échantillon pour laquelle la valeur de la variable est x_i (ou appartient à ($[a_{i-1},\ a_i[$). On peut par exemple l'exprimer sous forme de pourcentage (le pourcentage sera alors100 x f_i). [10]

1.3.4 Représentations graphiques des distributions statistiques

Très souvent, on préfère des représentations graphiques à des tableaux .

Ces représentations sont adaptées au type de variable étudiée : nominale, ordinale, discrète ou continue.

a.variables nominales

On dispose pour ces variables de diagrammes en bâtons, ainsi que de diagrammes circulaires (ou en secteurs, ou en « camembert »).

— **Diagramme en bâtons (figure 1.3)**

À chaque modalité x_i, on associe un « bâton » de longueur h_i proportionnelle à la fréquence f_i (ou, si l'on veut, à l'effectif n_i). On a donc $h_i = C \times f_i$ (C est une constante).

Pour une variable nominale, seules les hauteurs sont significatives ; l'ordre et l'écart des x_i ne sont pas significatifs.

— **Diagramme circulaire (figure1.3)**

L'angle de chaque secteur α_i est proportionnel à la fréquence f_i. En degrés, on a

$$a_i = 360 \times f_i$$

C'est la représentation la plus utilisée pour les variables nominales. De surcroît, elle est plus fidèle que la précédente

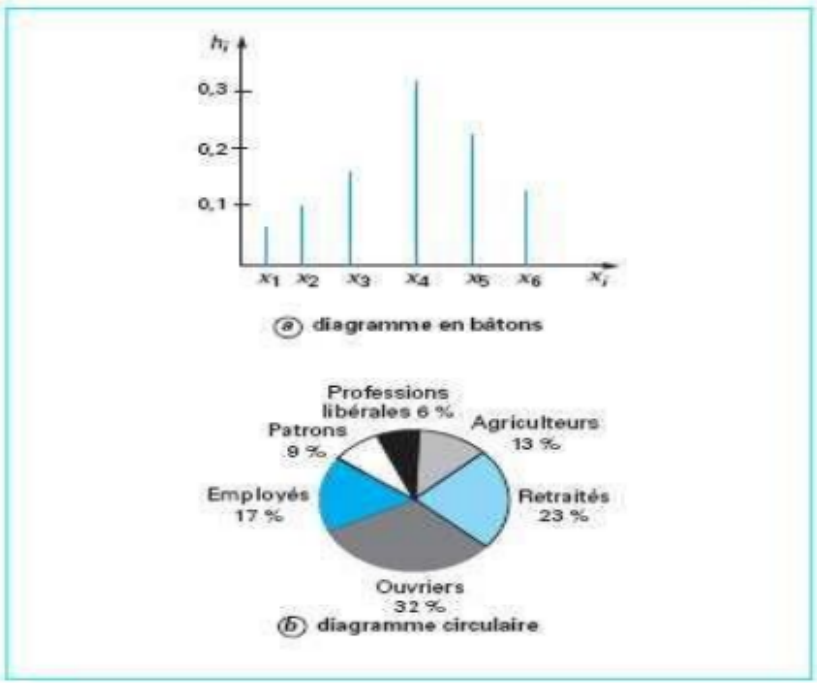


FIGURE 1.2 – Représentations graphiques des variables nominales

b. Variables ordinales et variables discrètes

— **Variables ordinales** On utilise les mêmes représentations que pour les variables nominales. Toutefois, il convient de noter que, pour le diagramme en bâtons, l'ordre des modalités à un sens concret, car il doit correspondre à l'ordre existant entre les valeurs.

— **Variables discrètes** Pour ce type de variables, on préfère le diagramme en bâtons car, dans ce cas, l'ordre et l'écart entre les bâtons sont significatifs

c. Variables continues : histogramme, polygone des fréquences

On considère une variable statistique continue dont les valeurs ont été rangées en classes ($[a_{i-1},\ a_i[$. L'amplitude de la classe ($[a_{i-1},\ a_i[$ est $A_i = a_{i-1}$ - a_i

Pour représenter graphiquement la distribution statistique d'une telle variable, on a recours à un histogramme. Le principe est le suivant : à chaque classe, on fait correspondre un rectangle de base l'intervalle $[a_{i-1}, a_i[$ (pour la classe i) et de hauteur h_i , de sorte que la surface du rectangle soit proportionnelle à l'effectif. Ainsi, on calcule la hauteur h_i du rectangle au moyen de la formule suivante :

$$h_i = \frac{n_i}{a_i - a_{i-1}}$$

D'un point de vue pratique, on constituera un tableau du type du tableau 1.4.

Variables continues: amplitudes et fréquences					
I	**Classes**	**Effectifs** n_I	**Fréquences** f_I	**Amplitudes** A_i	**Hauteurs** h_i
1	$[a_0, a_1[$	n_1	f_1	$a_1 - a_0$	$n_1/(a_1 - a_0)$
2	$[a_1, a_2[$	n_2	f_2	$a_2 - a_1$	$n_2/(a_2 - a_1)$
.	.	.	.	.	.
K	$[a_{k-1}, a_k[$	n_k	f_k	$a_k - a_{k-1}$	$n_k/(a_k - a_{k-1})$

FIGURE 1.3 – Variables continues : amplitudes et fréquence

On obtient ainsi le graphique de la figure 1.5 :

En abscisse, on porte l'ensemble des valeurs prises par la variable, découpé en classes ;

En ordonnée, on porte les hauteurs :

$$h_i = \frac{n_i}{a_i - a_{i-1}}$$

On trace enfin des rectangles

Remarque 1.1. Si les amplitudes sont toutes égales, on porte les effectifs en ordonnée.

La construction de l'histogramme s'opère de la façon suivante :

On calcule la différence de la distribution, différence entre la valeur la plus élevée et la valeur la plus faible.

On partage l'étendue de la distribution, en k classes d'amplitudes égales.

On compte le nombre de valeurs comprises dans chacune des classes.

Puis on reporte ces nombres n_i sur un graphique où l'on porte en abscisse les valeurs du paramètre étudié et en ordonnée les effectifs de chaque classe.

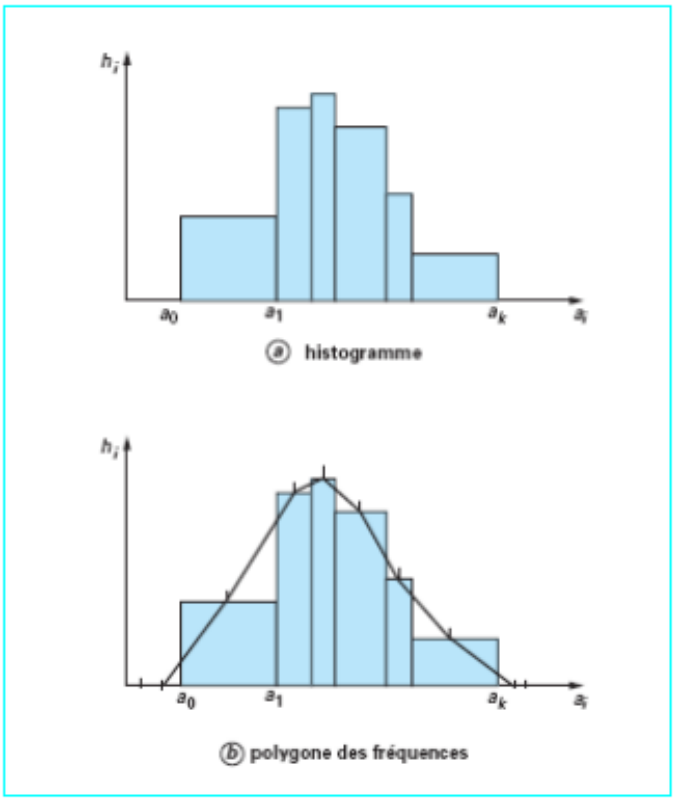


FIGURE 1.4 – Représentations graphiques des variables continues

À partir de l'histogramme d'une variable statistique continue, on peut tracer le

Polygone des fréquences associé (figure 1.5) en procédant de la manière suivante :

- on joint par des morceaux de droites les milieux des segments horizontaux supérieurs des rectangles de l'histogramme ;

- on ajoute à droite et à gauche de l'histogramme des classes fictives, toutes deux de même amplitude et d'effectif nul, ce qui donne alors lieu à deux nouveaux segments.

Remarque 1.2. On ne doit pas « lisser » la courbe.

1.3.5 Fréquences cumulées et fonction de répartition

a. Fréquences cumulées

Pour les variables qualitatives ordinales et pour les variables quantitatives, on peut exploiter la relation d'ordre existant entre les valeurs possibles de la variable. On définit ainsi les distributions cumulées (tableau 1.6).

I	Valeurs	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
1	x_1	n_1	f_1	n_1	f_1
2	x_2	n_2	f_2	n_1+n_2	f_1+f_2
.	.	.	.	.	.
.	.	.	.	.	.
k-1	x_{k-1}	n_{k-1}	f_{k-1}	$n_1+n_2+\ldots+n_{k-1}$	$f_1+f_2+\ldots+f_{k-1}$
K	x_k	n_k	f_k	$n_1+n_2+\ldots+n_k=N$	$f_1+f_2+\ldots+f_k=1$

FIGURE 1.5 – Distributions cumulées

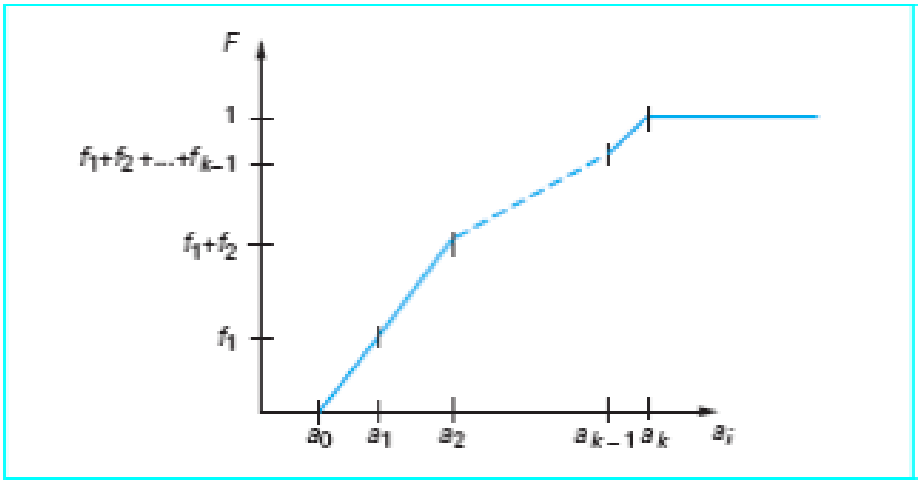


FIGURE 1.6 – Fonction de répartition

b. Fonction de repartition

Cette notion ne concerne que les variables quantitatives.

Définition. La fonction de répartition du caractère X est la fonction F, allant de l'ensemble des réels vers $[0, 1]$, définie par :

$$F(x) = \text{proportion d'individus de l'échantillon dont la valeur de } X \text{ est } < x$$

Soit X une variable continue, dont les valeurs sont rangées en classes

$$[a_0,\, a_1[, \ldots, [a_{k-1},\, a_k[$$

avec des fréquences $f_1, \ldots, f_k$.

— On commence par calculer les valeurs de F aux points du découpage :

$$\begin{aligned} F(a_0) &= 0, \\ F(a_1) &= f_1, \\ F(a_2) &= f_1 + f_2, \\ &\vdots \\ F(a_{k-1}) &= f_1 + f_2 + \ldots + f_{k-1}, \\ F(a_k) &= f_1 + f_2 + \ldots + f_k \end{aligned}$$

— Ensuite, dans chaque classe $[a_{i-1},\, a_i[$, on fait une interpolation linéaire (on relie les points extrêmes par un segment de droite).

— Puis on prolonge la courbe par 0 à gauche de a_0 et par 1 à droite de a_k (figure 1.7).[10]

1.3.6 Caracteristiques d'une distribution tendance centrale et dispersion

Généralités

Jusqu'à présent, nous nous sommes intéressés uniquement à la représentation des données statistiques. Cependant, s'il est vrai que les divers tableaux et graphes définis plus haut « résument » la distribution, ils ne permettent aucune quantification. Le but de ce paragraphe est donc de définir, pour chaque type de distribution statistique, un certain nombre de caractéristiques (ou indicateurs), c'est-à-dire quelques nombres permettant de résumer de manière quantitative (et non plus qualitative) chaque distribution. Bien entendu, n'importe quelle quantité ne peut pas être un indicateur.

Nous nous limiterons ici à 2 types de caractéristiques statistiques : celles dites de tendance centrale, qui donnent un « ordre de grandeur » de la variable étudiée en dégageant la modalité de la variable la plus représentative ;

celles dites de dispersion qui, elles, fournissent des informations sur la façon dont les individus se répartissent (se « dispersent ») autour de la tendance centrale.

Le tableau 1.8 donne les caractéristiques étudiées pour chaque type de variable.

CARACTERISTIQUES D'UNE DISTRIBUTION		
Type de variable	***Tendance centrale***	***Dispersion***
Nominale	Mode	
Ordinale	Mode, médiane, quantiles	Ecart interquartile
Quantitative	Mode, médiane, quantiles, moyenne	Ecart-type, écart interquartile

FIGURE 1.7 – Caractéristiques d'une distribution

Caractéristiques de tendance centrale

— **Mode** Il est défini pour tous les types de variables. On le définit comme suit.

Si X est une variable statistique nominale, ordinale ou discrète, le mode de la distribution associée est la modalité de X la plus représentée, c'est-à-dire celle pour laquelle l'effectif est le plus grand ;

Si X est une variable continue, le mode (ou classe modale) de la distribution associée est la classe dont la hauteur dans l'histogramme est la plus élevée.

— **Médiane et quantiles** Ces indicateurs sont définis pour toutes les variables sauf les variables nominales.

La médiane est la valeur de la variable telle que le nombre d'observations supérieures ou égales à cette valeur est égal au nombre d'observations strictement inférieures à cette valeur. On voit que, par exemple, pour les variables continues, cela revient à chercher un x tel que F(x) = 0,5. En règle générale, cette valeur de x n'existe pas dans le tableau de données dont on dispose.

C'est pourquoi on adopte la définition suivante : la médiane de la distribution de X est donnée par :

pour les variables ordinales ou discrètes :

Si la fréquence cumulée en x_{i-1} est $< 0{,}5$ et celle en x_i est $> 0{,}5$, alors la médiane vaut x_i,

Si la fréquence cumulée en x_{i-1} est égale à 0,5, alors la médiane vaut x_i ;

Pour les variables continues, réparties en classes $[a_{i-1}, a_i[$:

Si $F(a_{i-1}) < 0{,}5$ et $F(a_i) > 0{,}5$, **la classe médiane** est $[a_{i-1}, a_i[$, [et on calcule la médiane par interpolation linéaire sur l'intervalle $[a_{i-1}, a_i[$:]

$$M_{ed} = a_{i-1} + (a_i - a_{i-1})\frac{0{,}5 - F(a_{i-1})}{F(a_i) - F(a_{i-1})} \tag{1.1}$$

Avec F la répartition de X (figure 1.8)

Si $F(a_{i-1}) = 0.5$, la médiane vaut a_{i-1}

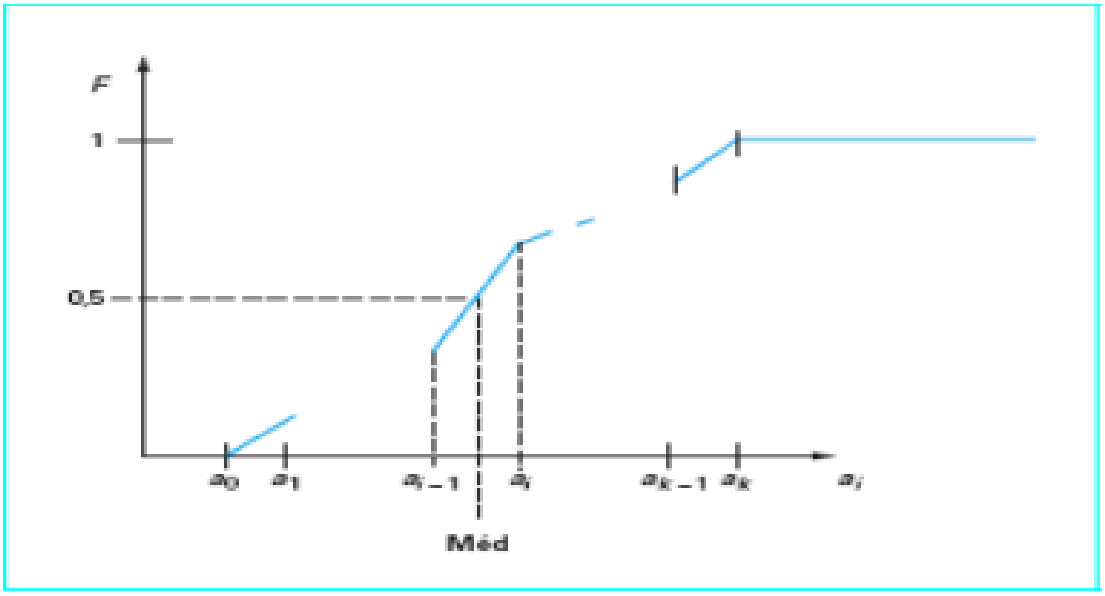


FIGURE 1.8 – Classe médiane

— **Moyenne arithmétique** Elle n'est définie que pour les variables quantitatives et, pour celles- ci, c'est la caractéristique de tendance centrale la plus « naturelle » et la plus utilisée.

La moyenne (arithmétique) d'une variable X sera notée et $N = n_1+n_2+...+n_k$

On définit la moyenne arithmétique de la manière suivante.

Si X est une variable quantitative discrète, donnée par sa distribution d'effectifs $(x_i, n_i), i = 1, ..., k$, alors la moyenne de X est donnée par

$$\bar{X} = \frac{1}{N}(n_1x_1 + n_2x_2 + \ldots + n_kx_k)$$

Si X est une variable continue rangée en classes $[a_{i-1}, a_i[$, la moyenne de X est

$$\bar{X} = \frac{1}{N}(n_1c_1 + n_2c_2 + \ldots + n_kc_k)$$

Où, pour tout i, ci est le centre de la classe $[a_{i-1}, a_i[$, soit

$$Ci = \frac{a_i - a_{i-1}}{2}$$

On dira qu'une variable est centrée si sa moyenne est nulle. Il faut noter les remarques suivantes :

La moyenne peut être définie à l'aide des fréquences $\bar{x} = f_1x_1+f_2x_2+...+f_kx_k$: pour les variables discrètes et $\bar{x} = f_1c_1 + f_2c_2 + ... + f_kc_k$, pour les variables continues ;

Il existe d'autres sortes de moyennes (géométrique, harmonique...).

La moyenne, prenant en compte toutes les valeurs observées, est très sensible aux observations aberrantes ;

Chaque fois que la répartition est assez symétrique (ce qui se traduit par un histogramme proche d'une courbe « en cloche »), la moyenne, la médiane et le mode sont proches. La moyenne est plus élevée que le mode ou la médiane si la répartition est dissymétrique, avec un accent vers les valeurs élevées ; si l'accent est, par contre, sur les valeurs faibles, la moyenne est plus petite que le mode ou la médiane.

Caractéristiques de dispersion

Les caractéristiques de tendance centrale donnent un ordre de grandeur du caractère statistique observé. Il est intéressant d'obtenir des informations sur la variabilité des observations et de leur dispersion autour de la tendance centrale. Intuitivement, une « bonne» caractéristique de dispersion doit être telle que, plus la variabilité est grande autour de la tendance centrale correspondante, plus cette caractéristique doit être grande, et inversement lorsqu'il y a peu de dispersion, la caractéristique doit être voisine de 0. De plus, une caractéristique de dispersion doit toujours être positive.

— **Ecart interquartile** Il est défini pour toutes les variables, excepté les variables nominales.

Définition 1.3. [10]

L'écart interquartile est la distance entre le 1er et le 3eme quartile. Il vaut donc $Q_{0.75} - Q_{0.25}$.

Il représente les valeurs extrêmes d'une dispersion de 50% des effectifs autour de la médiane.

— **Ecart type et variance** Ils ne sont définis que pour les variables quantitatives.

Définition 1.4. [10]

La variance est la moyenne des carrés des écarts à la moyenne, c'est-à-dire

pour une variable discrète :

$$V(\chi) = \frac{1}{N}\left(\sum_{i=1}^{i=k} n_i (x_i - \bar{x})^2\right) = \left(\frac{1}{N}\sum_{i=1}^{i=k} n_i x_i^2\right) - \bar{x}^2$$

pour une variable continue rangée en classes $[a_{i-1}, a_i[$, de centres c_i

$$V(\chi) = \frac{1}{N}\left(\sum_{i=1}^{i=k} n_i (c_I - \bar{x})^2\right) = \left(\frac{1}{N}\sum_{i=1}^{i=k} n_i c_i^2\right) - \bar{x}^2$$

Dans chaque cas, c'est la seconde expression qui sera le plus souvent utilisée pour effectuer les calculs.

L'écart-type est alors la racine carrée de la variance :

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}$$

— **Coefficient de variance** $C_v(x)$

Si l'écart type mesure l'erreur absolue dans l'estimation de la moyenne x , alors le coefficient de variation, noté $C_v(X)$ est :

$$C_v(X) = \frac{\sigma(X)}{\overline{x}}$$

$C_v(X)$ est un facteur adimensionnel utile. Il caractérise la dispersion intrinsèque de la variable.

1.4 Statistiques inférentielles

1.4.1 Rappels de probabilité

Variable aleatoire

Soit Ω l'espace des évènements associé à une expérience aléatoire, A une tribu (ou σ-algèbre) de parties de Ω et P une probabilité sur l'espace (Ω, A).

Définition 1.5. [8] On appelle variable aléatoire (v.a.) définie sur (Ω, A, P) une application

$$X : \Omega \to \mathbb{R} \quad \text{telle que} \quad \forall x \in \mathbb{R}, \quad X^{-1}(]-\infty, x[) \in A.$$

Remarque 1.3. Si $X(\Omega)$ est dénombrable, i.e. si $X(\Omega) = \{x_i\}_{i\in I}$, $I \subset \mathbb{N}$, la v.a. X est dite discrète et sa loi de probabilité est définie par l'ensemble des couples $(x_i, p_i)_{i\in I}$, où $p_i = P(X = x_i)$.

Fonction de répartition : On appelle fonction de répartition (f.r.) d'une v.a. X la fonction F définie par

$$\forall x \in \mathbb{R}, F(x) = P(X < x)$$

où $\{X < x\}$ désigne l'évènement $\{\omega \in \Omega : X(\omega) < x\} = X^{-1}(]-\infty, x[)$.

Densité : Une v.a. X est dite (absolument) continue s'il existe une fonction intégrable f telle que

$$\forall x \in \mathbb{R}, F(x) = \int_{-\infty}^{x} f(t)dt.$$

où F est la f.r. de X.

La fonction f est appelée densité (de probabilité) de la v.a. X et vérifie :

$$\forall x \in \mathbb{R}, f(x) \geq 0 \quad \text{et} \quad \int_{-\infty}^{+\infty} f(x)dx = 1.$$

Indépendance : Les variables aleatoire $X_1...X_k$ de f.r ;respectives $F_1...F_K$, sont dites independantes si :

$$P(X_1 < x_1..., X_k < x_k) = [f(x_1, ..., x_k) = \prod_{i=1}^{k} f_i(x_i)$$

Moment d'une V.a

Espérance Mathématique

On appelle espérance mathématique de la v.a. X le nombre $E(X)$ défini par :

$$E(X) = \sum_{i\in I} x_i p_i$$

si X est une v.a. discrète,

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

si X est v.a. de densité f.

Sous réserve que la série et l'intégrale ci-dessus soient convergentes.

Variance : On appelle variance de la v.a. X le nombre $V(X)$ défini par :

$$V(X) = \sum_{i \in I} (x_i - E(X))^2 p_i$$

si X est une v.a. discrète,

$$V(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x) dx$$

si X est v.a. de densité f.

Sous réserve que les quantités ci-dessus existent.

Écart-type : On appelle écart-type de la v.a. X la quantité

$$\sigma(X) = \sqrt{V(X)}.$$

Moment d'ordre k : On appelle moment non centré d'ordre k (resp. centré d'ordre k) ($k \in \mathbb{N}^*$) de la v.a. X la quantité (lorsqu'elle existe)

$$m_k = E(X^k), \quad \text{resp. } \mu_k = E[(X - E(X))^k].$$

LOIS USUELLES DISCRÈTES :

Loi de Bernoulli :

$$X(\Omega) = \{0, 1\}, \quad X = \begin{cases} 1 & \text{avec probabilité } p \\ 0 & \text{avec probabilité } 1-p \end{cases}, \quad E(X) = p, \quad V(X) = p(1-p).$$

Loi binomiale : Soit $X \sim \mathcal{B}(n, p)$, c'est-à-dire que X suit une loi binomiale de paramètres $n \in \mathbb{N}^*$ et $p \in [0, 1]$.

Alors :

— L'ensemble des valeurs possibles est :

$$X(\Omega) = \{0, 1, \ldots, n\}$$

— La loi de probabilité est donnée par :

$$P(X = x) = C_n^x p^x (1-p)^{n-x}, \quad \text{pour } x \in \{0, 1, \ldots, n\}$$

— L'espérance et la variance sont :

$$\mathbb{E}(X) = np, \quad \text{Var}(X) = np(1-p)$$

Loi de poisson :

$$X \sim > p(\lambda)$$

si $X(\Omega) = NavecP\{X = x\} = e^{-\lambda} \frac{\lambda^x}{x!}$, $\lambda \in \mathbb{R}_+^*$;E(X) = V(X) = λ.

LOIS USUELLES CONTINUES

Loi uniforme :

$$X \sim> \mathcal{U}[a,b]$$

$$\text{si } X(\Omega)=[a,b] \text{ et } f(x)=\begin{cases}\frac{1}{b-a} & \text{si } x\in[a,b]\\ 0 & \text{sinn}\end{cases}$$

$$E(X)=\frac{a+b}{2}, V(X)=\frac{(b-a)^2}{12}.$$

Ioi exponentielle :

la densité d'une v.a. X de loi exponentielle de paramètre θ

$$(\theta>0) \quad est \quad f(x)=\begin{cases}\theta e^{-\theta x} & si\, x>0\\ \\ 0 & sinon\end{cases}$$

$$E(x)=1/\theta, V(x)=1/\theta^2.$$

Loi normale :

$$X \sim> N(m,\sigma) \quad si \quad X(\Omega)=\mathcal{R} \quad et \quad f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\};$$

$$E(X)=m,\ V(X)=\sigma^2.$$

Loi de khi-deux :

Soient $X_1,\ldots,X_n$ n variables aléatoires indépendantes, suivant chacune une loi normale standard (c'est-à-dire centrée réduite, $m=0$, $\sigma=1$).

Alors, la variable aléatoire

$$\sum_{i=1}^{n} X_i^2$$

suit une loi du khi-deux à n degrés de liberté, notée χ_n^2.

Loi de Student :

Soit $U \sim> N(0,1)$ et X_n^2 deux v.a. indépendantes. Le rapport $\frac{U}{\sqrt{X_n^2/n}}$ suit une loi de Student à n degrés de liberté, notée T_n.

Loi de Fisher - Snédécor :

Soit X_n^2 et X_m^2 deux v.a. indépendantes. Le rapport $\frac{X_n^2/n}{X_m^2/m}$ suit une loi de Fisher-Snédécor à n et m degrés de liberté, notée F (n, m).

1.4.2 Estimation ponctuelle

Propriétés d'un estimateur :

Soit X une v.a. dont la loi dépend d'un paramètre inconnu θ, élément d'un sous-ensemble donné Θ de $\mathcal{R}$ appelé espace des paramètres.

On cherche à estimer θ à partir d'un échantillon $(X_1, ..., X_n)$ de v.a. indépendantes Ayant la même loi que X ; on notera $(x_1, ..., x_n)$ l'échantillon observé. Un estimateur T_n de θ sera une v.a.

$T_n = T_n(X_1, ..., X_n)$; la valeur $T_n(x_1, ..., x_n)$ est l'estimation de θ.

Estimateur sans biais :

Un estimateur T_n de θ est dit sans biais si

$$E(T_n) = \theta, \quad \forall \theta \in \Theta$$

Estimateur asymptotiquement sans biais :

Un estimateur T_n de θ est dit asymptotiquement sans biais si

$$\forall \theta \in \Theta, E(T_n) \to \theta \quad \text{quand } n \to \infty.$$

Estimateur convergent :

Un estimateur T_n de θ est dit convergent s'il converge en probabilité vers $\theta : T_n \to \theta$ quand $n \to \infty$.

Théorème 1.1. *[8] Un estimateur sans biais ou asymptotiquement sans biais dont la variance tend vers zéro lorsque n tend vers l'infini est convergent.[8]*

Comparaison des estimateurs :

Définition 1.6. [8] Soient T_n et $T_{n'}$ deux estimateurs sans biais de θ.

T_n est dite plus efficace que $T_{n'}$ si $\forall \theta \in \Theta, V(T_n) \leq V(T_{n'})$.

Définition 1.7. [8] On appelle vraisemblance de l'échantillon $(X_1, ..., X_n)$ L'application $L : \Theta \to \mathbb{R}_+$ définie pour une réalisation particulière $(x_1, ..., x_n)$ par :

$$\forall \theta \in \Theta \quad L(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta),$$

ou $f(x; \theta)$ désigne la densité de la v.a. X.

Définition 1.8. [8] On appelle information de Fisher apportée par la réalisation $(x_1, ..., x_n)$ sur le paramètre θ, la quantité (lorsqu'elle existe)

$$I_n(\theta) = E\left(\frac{\partial L_n L}{\partial \theta}\right)^2$$

Inégalité de FRÉCHET-DARMOIS-CRAMER-RAO (FDCR).
Si X prend ses valeurs dans un ensemble qui ne dépend pas de θ , si la densité $f(x;\theta)$ est deux fois continûment dérivable par rapport à θ , et sous certaines conditions de régularité, tout estimateur T_n sans biais de θ dont la variance existe vérifie l'inégalité FDCR :

$$\forall \theta \in \Theta \quad , V(T_n) \geq \frac{1}{I_n(\theta)}$$

où la quantité d'information de Fisher peut s'écrire

$$I_n(\theta) = \mathbb{E}\left(-\frac{\partial^2 L_n L}{\partial \theta^2}\right)$$

Définition 1.9. [8] Un estimateur T_n sans biais de θ est dit efficace si sa variance est égale à la borne de FDCR :

$$V(T_n) = \frac{1}{I_n(\theta)}$$

Méthode de maximum de vraisemblance

Elle consiste à prendre comme estimation $T(x_1, \ldots, x_n)$ du paramètre θ une valeur qui maximise la vraisemblance :

$$\forall \theta \in \Theta, \quad L(x_1, \ldots, x_n; T_n) \geq L(x_1, \ldots, x_n; \theta)$$

Remarquons que ni l'existence ni l'unicité de T_n ne sont assurées. Dans le cas particulier où X prend ses valeurs dans un ensemble qui ne dépend pas de θ et ou L est deux fois continument dérivable par rapport à θ , T_n est solution du système :

$$\frac{\partial L}{\partial \theta} = 0, \quad \frac{\partial^2 L}{\partial \theta^2} < 0$$

que l'on remplace, si $f(x;\theta) > 0$, par le système en général plus simple :

$$\frac{\partial\ L_n L}{\partial \theta} = 0, \quad \frac{\partial^2\ L_n L}{\partial \theta^2} < 0.$$

Si T_n est l'estimateur du maximum de vraisemblance (EMV) de θ, $g(T_n)$ est l'EMV de $g(\theta)$, pour toute fonction g.

1.4.3 Estimation par intervalle de confiance

Généralités :

Soit X une v.a. dont la loi dépend d'un paramètre réel inconnu θ et α [0,1] un nombre donné.

Définition 1.10. [8] On appelle intervalle de confiance pour le paramètre θ , de niveau de confiance $1-\alpha$, un intervalle qui a la probabilité $1-\alpha$ de contenir la vraie valeur du paramètre .

Construction pratique : soit $(X_1 \ , \ldots, X_n)$ un échantillon de la loi de X et T_n un estimateur de θ . S'il est possible de déterminer $t_1 = t_1(\theta)$ et $t_2 = t_2(\theta)$ tels que

$$P\left\{t_1(\theta) < T_n < t_2(\theta)\right\} = 1-\alpha$$

On cherche à inverser cet intervalle i.e. à déterminer les valeurs $a = a(T_n)$ et $b = b(T_n)$ telles que

$$P\left(\text{a } (T_n) < \theta < b \ (T_n)\right) = 1-\alpha$$

Si, par exemple, t_1 et t_2 sont deux fonctions croissantes :

$$T_n < t_2\left(\theta\right) \quad \Leftrightarrow \quad \theta > t_2^{-1}(T_n) \quad et \quad T_n > t_1(\theta) \quad \Leftrightarrow \quad \theta < t_1^{-1}(T_n) \ ;$$

dans ce cas :

$$\text{a} = t_2^{-1}(T_n) \quad et \quad b = t_1^{-1}(T_n) \ .$$

En fait le choix de t_1 et t_2 reste arbitraire puisqu'une seule équation permet de les déterminer, soit :

$$P\left\{T_n < t_1\right\} + P\left\{T_n > t_2\right\} = \alpha$$

Posons

$$\alpha_1 \ = P\left\{b(T_n) < \theta\right\} \quad \text{et} \quad \alpha_2 = P\left\{\theta < a(T_n)\right\};$$

Si α_1 et α_2 sont non nuls, on dit que l'intérvalle est bilatéral. En raison de la signification concrète du parametre θ, on peut etre amené à construire un intervalle unilatéral de la forme :

$$\mathrm{a}(T_n) < \theta \qquad (\ \alpha_1 = 0,\ \alpha_2 = \alpha\ \ )$$

ou

$$\theta > \ b(T_n) \qquad (\ \ \alpha_1 = \alpha,\ \alpha_2 = 0).$$

Dans le cas d'une loi symetrique, on construira un intervalle bilateral symetrique ($\alpha_1 = \alpha_2 = \alpha/_2$).

Intervalles classiques

Intervalle de l'espérance m d'une loi normale N (m, σ).

Cas où σ est connu.

à partir de l'estimateur $\bar{X_n}$ de m, de loi N $(m, \frac{\sigma}{\sqrt{n}})$, on détermine la valeur u du fractile d'ordre 1- $\frac{\alpha}{2}$ de N (0,1) telle que :

$$P\{-u < \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} < u\} = 1 - \alpha;$$

ce qui conduit à l'intervalle bilatéral symétrique centré en :

$$\overline{X}_n - \frac{\sigma}{\sqrt{n}}u < m < \overline{X}_n + \frac{\sigma}{\sqrt{n}}u.$$

Cas ou σ est inconnu.

la loi de $\bar{X_n}$ dépendant de σ , on utilise comme estimateur de σ^2

la variance empirique $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. La statistique $\sqrt{n}\frac{\overline{X}_n - m}{S_n}$ suit une loi de Student à $n-1$ degrés de liberté permettant de déterminer la valeur t telle que :

$$P\left\{t < \sqrt{n}\frac{\overline{X}_n - m}{S_n} < t\right\} = 1 - \alpha$$

ce qui conduit à l'intervalle bilatéral symétrique :

$$\overline{X}_n - \frac{S_n}{\sqrt{n}}t < m < \overline{X}_n + \frac{S_n}{\sqrt{n}}t.$$

Intervalle de la variance σ^2 d'une loi normale $N(m, \sigma)$

- **Cas ou m est connu**

$$S_n'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$$

est un estimateur sans biais de σ^2 et nS'^2_n/σ^2 suit une loi du khi-deux à n degrés de liberté ; d'où l'intervalle :

$$n\frac{S'^2_n}{c_2} < \sigma^2 < n\frac{S'^2_n}{c_1},$$

avec $\alpha_1 = P\left\{n\frac{S'^2_n}{\sigma^2} < c_1\right\}$ et $1-\alpha_2 = P\left\{n\frac{S'^2_n}{\sigma^2} < c_2\right\}$.

– **Cas ou m est inconnu**

$$(n-1)\frac{S^2_n}{c_2} < \sigma^2 < (n-1)\frac{S^2_n}{c_1},$$

c_1 et c_2 étant les fractions d'ordre respectif α_1 et $1-\alpha_2$ de la loi du Khideux à $(n-1)$ degrés de liberté.

1.4.4 Théorie des tests

Généralités :

Soit X une v.a. dont la loi dépend d'un paramètre réel inconnu θ, élément d'un sous-ensemble Θ de IR. On suppose que Θ est partitionné en deux ensembles Θ_0 etΘ_1 auxquels on associe les hypothèses suivantes :

H_0 :$\theta \in \Theta_0$

H_1 :$\theta \in \Theta_1$

Résoudre un problème de test consiste à prendre, au vu d'un échantillon $(X_1, \ldots, X_n)$ de la loi de X, l'une des deux décisions possibles :

D_0 : accepter H_0

D_1 : refuser H_0 (i.e. accepter H_1).

Cela revient à partitionner IR^n , ensemble des réalisations possibles $(x_1, \ldots, x_n)$ de $(X_1, \ldots, X_n)$, en deux sous-ensembles :

W : ensemble des réalisations $(x_1, \ldots, x_n)$ pour lesquelles on refuse H_0,

$\bar{W}$: ensemble des réalisations $(x_1, \ldots, x_n)$ pour lesquelles on accepte H_0.

Définition 1.11. [8]
La région W de refus de H_0 s'appelle la région critique du test. La région $\bar{W}$ est appelée région d'acceptation.

Résoudre un problème de test consistera donc à déterminer sa région critique.

Chaque décision peut entrainer une erreur :

L'erreur de première espèce consiste à prendre la décision D_1 (refuser H_0) alors que c'est l'hypothèse H_0 qui est vraie.

L'erreur de seconde espèce consiste à prendre la décision D_0 (accepter H_0) alors que c'est l'hypothèse H_1 qui est vraie.

La méthode pratique de construction du test dépendra des conséquences attribuées à chacune de ses deux erreurs possibles.

Méthodes de BAYES

On affecte des probabilités a priori p_0 et $p_1 = 1 - p_0$ aux deux hypothèses H_0 et H_1 et on associe un cout à chaque décision, ce qui est schématisé dans le tableau ci-après

		Décision		Probabilités a priori
		D_0	D_1	
Hypothèse vraie	H_0	C_{00}	C_{01}	P_0
	H_1	C_{10}	C_{11}	P_1

La fonction de vraisemblance $\theta \to L\ (x_1, ..., x_n; \theta)$ est noté L_0 si $\theta \in \Theta_0$ et L_1 si $\theta \in \Theta_1$. Au vu d'une réalisation $(x_1, ..., x_n)$, le théorème de Bayes permet de calculer les probabilites a posteriori π_0 et $\pi_1 = 1 - \pi_\theta$ des hypotheses H_0 et H_1 :

$$\pi_0 \quad = \frac{p_0 L_0}{p_0 L_0 + p_1 L_1} \quad \text{et} \quad \pi_1 \quad = \frac{p_1 L_1}{p_0 L_0 + p_1 L_1}$$

Ceci permet de calculer les espérances du cout de chaque décision :

$$E\left[c(D_0)\right] \, = \, c_{00}\pi_0 \, + \, c_{10}\pi_1 \qquad \text{et} \qquad E\left[c(D_1)\right] \, = \, c_{01}\pi_0 \, + \, c_{11}\pi_1 \, .$$

La règle de décision de Bayes est celle qui associe à la réalisation $(x_1, ..., x_n)$ la décision dont l'espérance du cout est la plus faible.

Méthode de NEYMAN et PEARSON

Les deux hypothèses ne jouent pas un rôle symétrique. On appelle H_0 l'hypothèse nulle du test et H_1 l'hypothèse alternative.

L'hypothèse nulle est celle qui est privilégiée et que l'on considère en général comme la plus vraisemblable.

Définition 1.12. [8] On appelle risque de premier espèce la probabilité de rejeter à tort l'hypothèse nulle :

$$\alpha = P(D_1|H_0) = P(W|\theta \in \Theta_0).$$

On appelle risque de seconde espèce la probabilité d'accepter à tort l'hypothèse nulle :

$$\beta = P(D_0|H_1) = P(\bar{W}|\theta \in \Theta_1).$$

Dans l'optique de Neyman et Pearson, on accroit la dissymétrie du Problème de test en considérant que l'erreur la plus grave consiste à rejeter à tort l'hypothèse nulle. On fixe donc un seuil maximum α_0 au risque de première espèce et on cherche un test qui minimise le risque de seconde espèce.

Définition 1.13. [8] On appelle puissance d'un test la probabilité de rejeter l'hypothèse nulle avec raison :

$$\eta = P\left(D_1 \mid H_1\right) = P\left(W \mid \theta \in \Theta_1\right) = 1 - \beta.$$

La règle de décision de Neyman et Pearson consiste à déterminer la région critique W pour laquelle la puissance n est maximum, sous la contrainte $\alpha \leq \alpha_0$.

Hypothèses simples.

Définition 1.14. [8] Une hypothèse est dite simple si la loi de la v.a. X est parfaitement déterminée dans cette hypothèse. Dans le cas contraire, elle est dite multiple.

Considérons le problème de test suivant entre deux hypothèses simples :

$H_0 \ : \theta = \theta_0$

$H_1 \ : \theta = \theta_1$

Théorème de Neyman et Pearson [8] : Pour tout a fixé dans [0,1], le test de puissance maximum tel que

$P(W|\theta = \theta_0) = \alpha$ est défini par la région critique :

$$W = \{(x_1, ..., x_n) : \frac{L_0}{L_1} \leq k\},$$

où k est une constante déterminée en fonction de α.

Définition 1.15. [8] On appelle niveau d'un test la borne supérieure du risque de première espèce :

$$\sup_{\theta\in\Theta_0} P(W|\theta \in \Theta_0).$$

Définition 1.16. [8] Un test de région critique W^* dit UPP (uniformément le plus puissant) de niveau α si :

$$\forall W, \forall\theta \in \Theta_1, P(W^*|\theta \in \Theta_1) \geq P(W|\theta \in \Theta_1)$$

Test convergent :

Soit W_n la région critique d'un test basé sur un échantillon de taille n et posons $\alpha_n = P(W_n|\theta \in \Theta_0)$ et $\eta_n = P(W_n|\theta \in \Theta_1)$. Le test est dit convergent si

$$\forall\theta \in \Theta_0,\ \alpha_{n+1} \leq \alpha_n$$

$$\forall\theta \in \Theta_1,\ \eta_n \to 1 \quad \text{quand } n \to \infty$$

TEST DU KHI-DEUX

Test d'Indépendance :

Soit X et Y deux caractères (qualitatifs ou quantitatifs) à respectivement r et s modalités. Soit n_{ij} le nombre d'individus d'une population de taille $n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$ qui possèdent simultanément la modalité i $(1 \leq i \leq r)$ du caractère X et la modalité j $(1 \leq j \leq s)$ du caractère Y et p_{ij} la probabilité correspondante.

Au vu de ces observations on désire tester l'indépendance de ces deux caractères i.e. résoudre le problème de test :

$$H_0 : p_{ij} = p_{i.} \times p_{.j} \quad \text{où} \quad p_{i.} = \sum_{j=1}^{s} p_{ij} \quad \text{et} \quad p_{.j} = \sum_{i=1}^{r} p_{ij}$$

$$H_1 : p_{ij} \neq p_{i.} \times p_{.j}.$$

On utilise pour cela la statistique

$$D = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{n_{i.}n_{.j}} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right)$$

La région critique du test de seuil α est $D \geq k$; sous H_0 , la loi approchée de D est $\chi^2_{(r-1)(s-1)}$ et k est donc la valeur lue dans la table telle que $\alpha = P(D \geq k|H_0)$.

Test d'adéquation.

Soit $(x_1, \ldots, x_n)$ un n-échantillon d'une v.a. X et F une fr. donnée. On désire tester :

$$H_0 : X \text{ a pour fr. } F$$

$$H_I : X \text{ n'a pas pour fr. } F.$$

Pour cela on répartit les n observations en k classes $[a_{i-l}, a_i[$ d'effectif n_i $(1 \leq i \leq k)$ et on calcule

$$p_i = F(a_i) - F(a_{i-1}) \text{ puis}$$
$$D = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n.$$

La région critique du test de seuil α est

$$D \geq k,$$

la valeur de k étant lue dans la table, avec $\alpha = P(D \geq k \mid H_0)$ et D de loi approchée χ^2_{k-1} sous H_0.

L'effectif np_i de la classe $[a_{i-1}a_i[$ doit être supérieur ou égal à 5 ; sinon, on regroupe deux (ou plus) classes consécutives.

TEST DE NORMALITE

Test de Kolmogorov-Smirnov

Une approche élégante des tests de normalité est le test de conformité de Kolmogorov-Smirnov. Ce test non paramétrique consiste à comparer la distribution de fréquences relatives cumulées d'une variable observée avec la distribution théorique que cette variable aurait si elle était distribuée normalement. On superpose les deux distributions, on cherche la classe ou l'écart entre la distribution théorique et la distribution observée, et on vérifie dans une table conçue à cet effet ou en calculant directement la valeur critique D, si cet écart est significativement grand, i.e si l'hypothèse de normalité H_0 : distribution normale peut être rejetée au seuil considéré. L'idée est que, dans une

distribution relative cumulée observée, chaque classe peut diverger un peu (en plus ou en moins) par rapport au niveau qui serait le sien sous une distribution normale, mais si une classe est particulièrement éloignée de sa position théorique, cela signifie qu'une ou plusieurs autres le sont aussi (dans l'autre sens), ce qui veut dire que c'est l'ensemble de la distribution qui n'est pas conforme à la loi normale.

Remarque 1.4. Les premières tables de Kolmogorov-Smirnov se basaient sur le fait qu'on connaissait les vrais paramètres de la distribution théorique (moyenne et écart-type). Ce n'est pratiquement jamais le cas, et en cas de calcul fondé sur des paramètres estimés à partir des données, les tables originales sont trop conservatrices (on accepte trop souvent l'hypothèse nulle de normalité).

Définition 1.17. [9] Soit un échantillon X_i ,$i = 1..n$ qui présente la distribution empirique, et X la variable aléatoire de cet échantillon.

Déclaration de l'hypothèse H_0 : La distribution de la variable X suit une loi normale.

$$F_{rel}(X_i) \neq F_{relth}(X_i)$$

telle que $F_{rel}(X_i)$ la fonction cumulative empirique et $F_{relth}(X_i)$ la fonction cumulative théorique.

H_I : La distribution de la variable X ne suit pas une loi normale.

$$F_{rel}(X_i) = F_{relth}(X_i),$$

tel que rel th c'est réel théorique.

On utilise un test de Kolmogorov-Smirnov ou :

$$D_{obs} = \max(|F_{rel}(X_i) - F_{relth}(X_i)|).$$

La variable x continue est la condition pour utiliser le test de Kolmogorov.

Si H_0 est vraie, la variable D_{obs} suivra une distribution selon la fonction de Kolmogorov-Smirnov.

On rejette H_0 au seuil $\alpha = 0.05$, si $D_{obs} > D_\alpha$.

Pour $\alpha = 0.05 : D_\alpha = \frac{0.895}{S}$.

pour $\alpha = 0.01 \ : D_\alpha = \frac{1.035}{S}$.

ou $S = \sqrt{n} - 0.01 + \frac{0.85}{n}$ n étant le nombre d'individus (et non de classe)

Procédure du test de Kolmogorov-Smirnov :

La procédure s'agence comme suit

►**Première étape** :

Tri des données brutes en ordre croissant.

►**Deuxième étape :**

Centrage et réduction des valeurs de X.

►**Troisième étape :**

Trouver les valeurs de $\hat{F}$ correspondantes tel que $\hat{F}$ est la fonction cumulative empirique.

►**Quatrième étape :**

Calculer les fréquences observées cumulées F_cum qui est la fonction cumulative.

►**Cinquième étape :**

Trouver la valeur maximale de d_i^+ et d_i^- , et le comparer à notre valeur critique D

Test de Shapiro-Wilk

Description :

Ce test est basé sur la statistique W. En comparaison des autres tests, il est particulièrement puissant pour les petits effectifs ($n > 50$). La statistique du test s'écrit

$$W = \frac{\sum_{i=1}^{n/2} a_i (x_{(n-i+1)} - x_i)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

où $x_{(i)}$ correspond à la série des données triées.

$\left[\frac{n}{2}\right]$ est la partie entière du rapport $\frac{n}{2}$.

a_i sont des constantes générées à partir de la moyenne et de la matrice de variance-covariance des quantiles d'un échantillon de taille n suivant la loi normale. Ces constantes sont fournies dans des tables spécifiques.

La statistique W peut donc être interprétée comme le coefficient de détermination (le carré du coefficient de corrélation) entre la série des quantiles générées à partir de la loi normale et les quantiles empiriques obtenus à partir des données. Plus W est élevé, plus la compatibilité avec la loi normale est crédible. La région critique, rejet de la normalité, s'écrit :

$$R.C : W < W_{crit}.$$

Procédure du test de Shapiro-Wilk :

La procédure s'agence de la manière suivante :

▶ **Première étape :**

Trier les données x_i, nous obtenons la série $x_{(i)}$.

▶ **Deuxième étape :**

Calculer les écarts $(x_{(n-i+1)} - x_{(i)})$.

▶ **Troisième étape :**

Lire dans la table pour n donnée, les valeurs des coefficients a_i.

▶ **Quatrième étape :**

Former le numérateur de W, nW.

▶ **Cinquième étape :**

Former le dénominateur de W, dW.

▶ **Sixième étape :**

Déduction de W.

1.4.5 Le modèle linéaire

Généralités

On cherche à approximer une liaison inconnue $Y = f(X_1, ..., X_p)$ par une relation linéaire :

$$Y = b_0 + b_1 X_1 + ... + b_p X_p + \epsilon,$$

ϵ tant une v.a.. La variable Y est la variable à expliquer, ou endogène, ou dépendante, les variables $X_1, ..., X_p$ sont explicatives, ou exogènes, ou indépendantes. Pour un échantillon de taille n, le modèle s'écrit :

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ji} + \epsilon_i \quad (i = 1 \text{ à } n),$$

avec les hypothèses : $E(\epsilon_i) = 0$, $v(\epsilon_i) = \sigma^2$, $\text{Cov}(\epsilon_i, \epsilon_k) = 0\,(i \neq k)$.

LA RÉGRESSION SIMPLE [8]

Pour $p = 1$ le modèle devient : $y_i = a + bx_i + \epsilon_i \, (i = 1 \text{ à } n)$.

Estimation des paramètres a et b.

La méthode d'estimation est celle des moindres carrés ordinaires (MCO), et consiste à trouver les estimateurs $\hat{a}$ et $\hat{b}$ de a et b qui minimisent la somme des carrés des erreurs

$$Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

La résolution de $\min_{a,b} \sum_{i=1}^{n} \epsilon_i^2$ conduit à :

$$\hat{a} = \bar{y} - \hat{b}\bar{x}; \quad \hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Remarque : en désignant par S_x^2 (resp. S_y^2) la variance empirique $\frac{1}{n}\sum(x_i - \bar{x})^2$ de $(x_1, ..., x_n)$ (resp. $(y_1, ..., y_n)$) et par ρ le coefficient de corrélation linéaire empirique $\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{nS_xS_y}$ entre x et y, $\hat{b}$ s'écrit $\rho\frac{S_y}{S_x}$.

Qualité de l'ajustement linéaire.

On appelle : $\hat{y}_i = \hat{a} + \hat{b}x_i$ la valeur ajustée de y_i ;

$y = \hat{a} + \hat{b}x$:la droite des moindres carrés, ou droite de régression de y en x ;

$\hat{\epsilon}_i = y_i - \hat{y}_i$: le résidu en i ;

$\nu_E = \frac{1}{n}\sum_{l=1}^{n}(\hat{y}_i - \overline{\mathbf{y}})^2$: la variance expliquée (par le modèle) ;

$\nu_R = \frac{1}{n}\sum_{l=1}^{n}\hat{\epsilon}_l^2$ la variance résiduelle

On montre (équation d'analyse de la variance) que :

$$v_T = \frac{1}{n}\sum_{l=1}^{n}(y_l - \bar{y})^2 = v_E + v_R \, .$$

La qualite du modele est jugee le coefficient de determination de la regression R^2 :

$$R^2 = \frac{v_E}{v_T} = \frac{\sum_{l=1}^{n}(y_l - \bar{y})^2}{\sum_{l=1}^{n}(y_l - \bar{y})^2} \, .$$

Autres formes :

$$\mathrm{R}^2 = \frac{\tilde{b}^2 \sum_{l=1}^{n}(x_i - \bar{x})^2}{\sum_{l=1}^{n}(y_l - \bar{y})^2} = 1 - \frac{\sum_{l=1}^{n}\hat{\varepsilon}_l^2}{\sum_{l=1}^{n}(y_l - \bar{y})^2} = \rho^2.$$

Proprietes statistiques de $\hat{a}$ et $\hat{b}$.

Les estimateurs $\hat{a}$ et $\hat{b}$ sont sans biais et convergents pour a et b.

$$E(\hat{a}) = a, E\left(\hat{b}\right) = b$$

$$V(\ \hat{a}\ ) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_l - \bar{x})^2}\right], V\left(\hat{b}\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$COV\left(\widehat{a},\ \bar{\hat{b}}\right) = -\frac{\sigma^2\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,.$$

$S^2 = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_l^2}{n-2}$ est un estimateur sans biais de σ^2 ; $V(\hat{a})$, $V\left(\hat{b}\right)$, $COV\left(\widehat{a},\ \bar{\hat{b}}\right)$ sont estimes en remplaçant σ^2 par S^2.

Remarque : $R^2 = 1 - \frac{(n-2)S^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$, soit $\quad S^2 = \frac{1-R^2}{n-2}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Hypothèse de normalité des ε_i.

Soit $\varepsilon_i \sim N(0, \sigma)$, $i = 1$ à n.

Alors $Y_i \sim N(a + bx_i, \sigma)$, $(n-2)\frac{S^2}{\sigma^2} \sim \chi^2_{n-2}$.

Loi de $\hat{b}$: $\hat{b} \sim N\left(b, \frac{\sigma}{\sqrt{\sum(x_i-\bar{x})^2}}\right)$.

Loi de $\hat{a}$: $\hat{a} \sim N\left(a, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}}\right)$.

Remarque 1.5. la méthode des moindres carrés est équivalente à celle du maximum de vraisemblance dans le cas d'aléas normaux.

Intervalle de confiance pour b.

$$p\left(\hat{b} - \frac{tS}{\sum(x_i - \bar{x})^2} < b < \hat{b} + \frac{S}{\sum(x_i - \bar{x})^2}\right) = 1 - \alpha$$

t fractile d'ordre $1 - \alpha/2$ de T_{n-2}.

Test $b = b_0$ contre $b \neq b_0$ ("Test de Student"). Au risque α, on acceptera $b = b_0$ si :

$$\frac{|\hat{b} - b_0|}{S/\sqrt{\sum(x_i - \bar{x})^2}} < t,$$

t fractile d'ordre $1 - \alpha/2$ de T_{n-2}.

Prévision. A partir de x_{n+1}, réalisation supposée connue de X en $n+1$, on prévoit y_{n+1} par

$$\hat{y}_{n+1} = \hat{a} + \hat{b}x_{n+1}.$$

Sous l'hypothèse $\epsilon_i \sim N(0, \sigma)$ pour tout i, l'intervalle de confiance de niveau $1 - \alpha$ pour $E(y_{n+1})$ appelé ici intervalle de prévision, est

$$\left[\hat{y}_{n+1} - tS\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}\right)^{1/2} ; \quad \hat{y}_{n+1} + tS\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x}_n)^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}\right)^{1/2}\right]$$

t fractile d'ordre 1 - $\alpha/2$ de $T_n - 2$

1.5 Conclusion

Ce premier chapitre a permis d'établir un cadre théorique solide en rappelant les notions fondamentales de la statistique, tant descriptive qu'inférentielle, qui constituent les bases indispensables à toute analyse de données. Ces concepts offrent un socle méthodologique essentiel pour comprendre et exploiter efficacement les données. Ce cadre théorique prépare ainsi la suite du mémoire, qui se concentrera sur la problématique de la prévision à travers l'utilisation du logiciel SPSS, un outil puissant et adapté pour modéliser et anticiper les évolutions futures des données. Cette transition vers les méthodes de prévision représente une étape clé pour optimiser l'exploitation des données et soutenir la prise de décision.

Chapitre 2

Calculs de prévision

2.1 Introduction

Ce deuxième chapitre aborde les calculs de prévision, une étape clé pour anticiper les évolutions futures à partir des données disponibles. Il présente l'utilisation du logiciel SPSS, qui facilite l'analyse statistique et la mise en œuvre des différentes méthodes de prévision. Ce chapitre pose ainsi les bases pratiques nécessaires pour appliquer ces techniques et exploiter au mieux les données dans un contexte décisionnel.

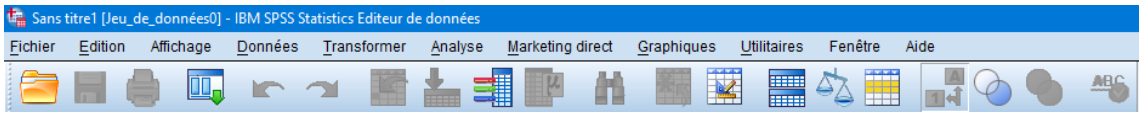
2.2 SPSS

Définition 2.1. [1] [2] [3] :

SPSS (Statistical Package for the Social Sciences) est un logiciel utilisé principalement pour l'analyse statistique et la gestion de données dans divers domaines comme l'économie, la santé, le marketing, et les sciences sociales. Il permet de traiter, analyser et documenter des données à travers une interface conviviale avec des menus déroulants ou via un langage de commandes pour des analyses plus complexes. SPSS facilite la manipulation des données sous forme de tableaux où chaque ligne représente un cas (individu, foyer, etc.) et chaque colonne une variable (âge, sexe, revenu, etc.). Il est accessible à tous les niveaux d'utilisateurs et offre des fonctionnalités avancées comme l'apprentissage automatique et l'intégration avec des langages comme Python et R

L'éditeur de données

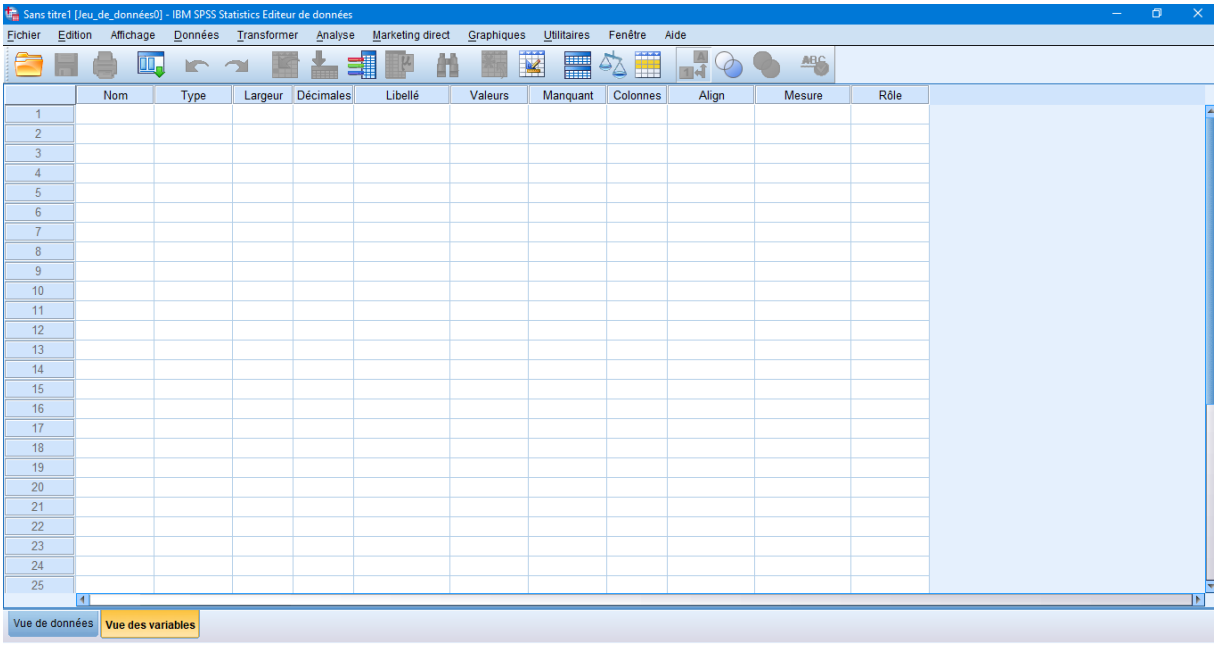
Une fois le logiciel ouvert, nous voyons l'éditeur de donnée. L'éditeur de données contient la grille (matrice) de données (« affichage des données » ou Data View ») et les descriptions des variables (« affichage des variables » = « Variable View »). Dans la partie supérieure de l'éditeur de données nous avons, comme dans Word et Excel, des menus déroulants :



Menu Fichier ou File : créer un nouveau fichier SPSS, ouvrir un fichier compatible existant (SPSS, Excel, ACCES, etc.), enregistrer le fichier, etc. Pour ouvrir un fichier de données SPSS (extension .sav) par exemple, nous allons utiliser les menus déroulants. Cliquer sur « Fichier ou File », puis sur « Ouvrir ou Open » et finalement sur « Données ou Data... ». Là, vous pouvez chercher votre fichier dans le répertoire où vous l'avez enregistré. Une fois nos données ouvertes, nous pouvons explorer les deux affichages mentionnés en dessous : affichage des données et affichage des variables.

Affichage des variables :

Chaque ligne représente une variable. Les colonnes décrivent les caractéristiques des variables.



Ces caractéristiques sont entre-autres :

— **Nom** : nom de la variable qui doit être unique

— **Type** : nature de la variable (numérique, date, dollar, etc.)

— **Largeur** : nombre de chiffres accordés à la donnée (décimaux inclus)

— **Décimales** : nombre de décimaux

— **Libellé** : étiquette ou description de la variable

— **Valeurs** : valeurs définies et leur description (p.ex. 1 = Femme, 2 = Homme)

— **Manquant** : attribution de certaines valeurs comme codes pour valeurs manquantes

— **Colonnes** : largeur des colonnes dans la vue de données

— **Align** : Alignement des valeurs des variables dans les cellules de la grille de données (à droite, à gauche, centrées)

— **Mesure** : Description de l'échelle de mesure (continue, ordinale ou nominale) cellules de la grille de données (à droite, à gauche, centrées)

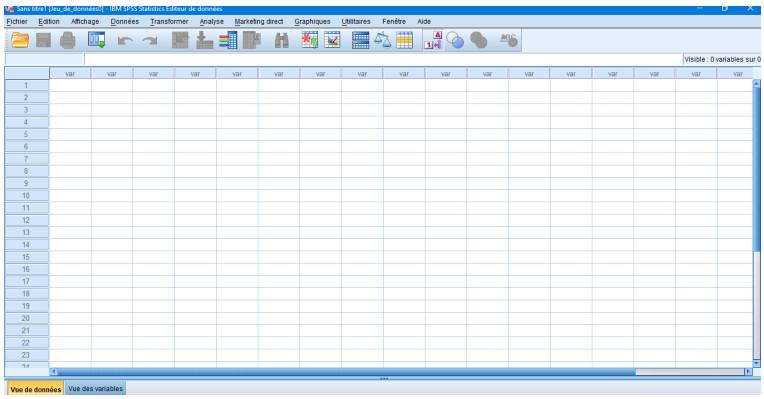
— **Rôle** :indique la fonction de la variable dans l'analyse (par exemple, variable indépendante ou dépendante).

Affichage des données :

Chaque ligne représente un cas, par exemple un sujet (case)

Chaque colonne représente une variable (variable)

Chaque cellule contient une valeur d'un cas sur une variable



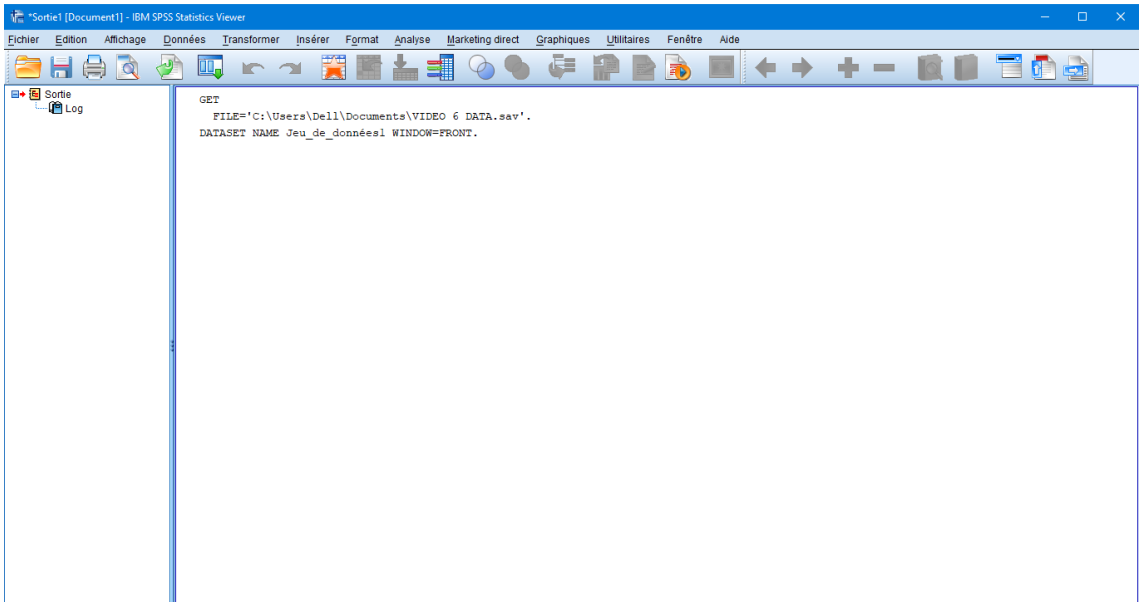
2.2.1 Les différentes fenêtres de SPSS (données, résultats et syntaxe)

Fenêtre de données (Data Editor)

Cette fenêtre permet d'entrer des données, de les modifier ou de les effacer. Il est rare que l'on tape les données manuellement dans SPSS car il y a trop d'erreurs de saisie possibles. On va plutôt ouvrir un fichier déjà existant ou on procède par copier-coller.

Fenêtre des résultats (Output Editor)

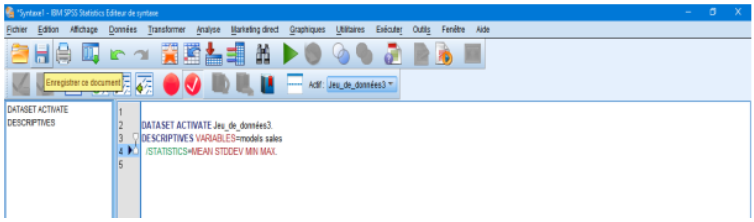
Cette fenêtre apparaît après qu'une commande d'analyse a été effectuée, et contient les résultats de cette analyse. Les résultats apparaissent à droite dans la fenêtre. A gauche, figure une table des matières des résultats générés par SPSS. Les résultats peuvent être imprimés tels quels (mais avec le risque de voir un tableau s'imprimer sur plusieurs pages. Il est également possible de copier les tableaux qui nous intéressent pour les coller ensuite dans Word, Excel ou dans un autre logiciel. Il est possible de copier un tableau de deux manières. En cliquant sur le tableau en appuyant sur le bouton de droite de la souris, SPSS vous propose de copier (copy) ou de copier l'objet (copy object). Copier correspond à copier les valeurs, mais lorsqu'il est collé il peut perdre son format (utile pour copier les résultats dans une feuille Excel par exemple). Copier les objets correspond à copier les valeurs et le format du tableau : une fois collé, impossible de modifier les cellules du tableau (utile pour copier les résultats dans Word).



Fenêtre de syntaxe (Syntax Editor)

Jusqu'à maintenant, nous avons vu comment travailler avec les menus déroulants. Il existe une autre manière de lancer des analyses : passer par la fenêtre de syntaxe.

Cette fenêtre permet d'écrire les commandes d'analyses statistiques. Elle fonctionne comme un traitement de texte simple.



Entrer les données

Il y a plusieurs manières d'entrer les données :

* directement dans SPSS

*dans Excel ; puis nous importons les données dans SPSS (par menu Fichier ou par copier-coller)

* dans un éditeur de texte, puis nous importons les données dans SPSS (pas recommandé, sujet à erreur !)

* scannage des données : pour cela il est nécessaire d'avoir un hardware avec logiciel et mise en page spécialisés

Encodage

Il est recommandé de résumer les informations les plus importantes sur les variables rassemblées dans un « tableau de codage ». Ce tableau de codage à deux utilités à deux moments bien précis :

* Pendant l'entrée des données : comme règle de codage des valeurs des variables

* Après l'entrée des données : comme description compacte du fichier des données

2.3 La Prévison :

Définition 2.2. [11]

La prévision est une activité scientifique et technique qui transforme des données passées en connaissances prospectives, permettant ainsi d'agir de manière proactive

sur le futur plutôt que de le subir ; elle peut être aussi définie comme étant « une appréciation sur les valeurs futures d'une variable quantitative »

Nous partons d'une série d'observations à travers le temps portant sur une variable y quelconque, de l'instant 1 jusqu'à l'instant T ; il s'agit d'une série chronologique ou encore d'une série temporelle. Nous cherchons à prévoir la valeur qui sera atteinte par y à un instant futur T+h, ou encore à l'horizon h

Plusieurs méthodes de prévision existent, elles peuvent être regroupées en deux grandes classes :

1. Méthodes extrapolatives (courbes de croissance, lissage par les moyennes mobiles, modélisation ARMA...) : Ces méthodes utilisent le passé de la variable elle-même. Seul le passé de la variable est utilisé en vue de la prévoir sans apport d'information extérieure.[12]

2. Méthodes explicatives (régression linéaire, courbes de croissance...) :

Celles ci utilisent les valeurs passées et présentes d'une ou de plusieurs variables pour prévoir y. L'ensemble d'information utilisé comporte des facteurs extérieurs qui peuvent influencer le futur de y en plus du passé de la variable y elle-même.[12]

Nous présentons dans ce qui suit quelques méthodes de prévision extrapolatives et quelques méthodes de prévision explicatives.

Les modèles de prévision traditionnels sont basés sur l'extrapolation des tendances passées.

La prévision classique est souvent faite sous l'hypothèse de stabilité du système en vue de dégager un scénario tendanciel. Autrement dit : Les modèles de prévision traditionnels reposent sur l'idée que le futur va suivre les mêmes tendances que le passé. Ils utilisent donc les données historiques pour prolonger ces tendances dans le temps, en supposant que le système étudié reste stable et ne subit pas de changements importants. Ainsi, la prévision classique cherche à dégager un scénario basé sur la continuité des comportements passés, sans prendre en compte d'éventuelles évolutions ou ruptures dans le fonctionnement du système.

2.4 Les méthodes de prévision : Les méthodes extrapolatives :

2.4.1 Méthodes des courbes de croissance

Il s'agit de méthodes extrapolatives de la tendance passée de la variable considérée. A titre d'exemple, supposons que l'on identifie une tendance de type linéaire suivie par une variable y sur le passé :

$$\text{Tendance linéaire } : y_t = \alpha + \beta t$$

t étant la variable temps ; α et β sont des paramètres à estimer. La prévision de la variable y, à l'instant (T+1) par exemple, est basée sur l'extrapolation de cette tendance, sous l'hypothèse d'invariance des paramètres α et β :

$$\hat{y}_{t+1} = \alpha + \beta(t+1)$$

La tendance identifiée peut être d'une autre nature que linéaire, elle peut être de type exponentielle, quadratique ou autre :

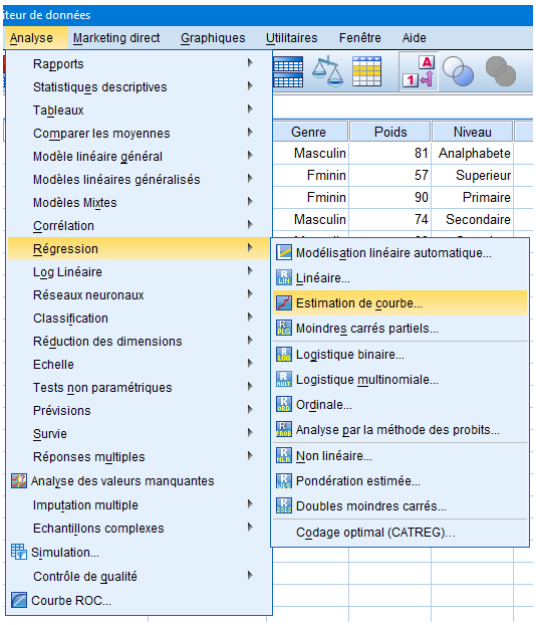
$$\text{Exponentielle : } \quad y = \alpha\beta^t$$

$$\text{Quadratique : } \quad y = \alpha + \beta t + \delta t^2$$

La prévision est toujours basée sur l'extrapolation de la tendance en supposant la stabilité des coefficients. Ce type de modèles permet certes d'approcher le fonctionnement des variables mais, en simplifiant la réalité, peut finir par transformer, par déformer et par s'éloigner de la réalité. [12]

Procedure

▶ Allez dans le menu Analyse > Régression >Estimation des courbe ...(ou Analyse > Régression > Courbe...selon la version).



►Sélectionnez la variable dépendante (la série à prévoir) et la variable indépendante (le temps ou l'indice chronologique).

►Choisissez le type de courbe à ajuster : linéaire, exponentielle, logarithmique, quadratique, etc.

►Cliquez sur OK pour obtenir les paramètres du modèle, les statistiques d'ajustement et les valeurs ajustées.

Pour obtenir les prévisions, utilisez les paramètres estimés pour extrapoler les valeurs futures (SPSS affiche les valeurs ajustées et permet de calculer les prévisions pour de nouvelles valeurs de la variable temps).

2.4.2 Méthodes de prévision par moyenne mobile

Les moyennes mobiles sont à la base d'une méthode de prévision qui consiste à utiliser la moyenne des ν dernières observations disponibles comme prévision pour la date suivante. On parle alors de méthode de prévision par moyenne mobile d'ordre ν .

La méthode présentée ci-dessous s'inscrit dans le cadre d'un**modèle de décomposition de type additif**. On suppose que le mouvement extra-saisonnier est une fonction quelconque du temps, que le mouvement saisonnier est rigoureusement périodique et que le mouvement accidentel est de faible amplitude et de moyenne nulle.

L'observation x_k, relative à la k-ième période, se décompose alors sous la forme :

$$x_k = T_k + S_j + A_k$$

où T_k désigne la tendance (trend), S_k le facteur saisonnier, et A_k le facteur accidentel.[12]

Détermination du trend [7]

La détermination du trend est différente selon que le mouvement saisonnier comprend un nombre de périodes d'observations ν impair ($\nu = 2p + 1$, p entier) ou pair ($\nu = 2p$). (Par exemple dans le cas de fluctuations trimestrielles, $\nu = 4 = 2 \times p$ avec $p = 2$).

À chaque chronique $x(t)$, on peut associer sa moyenne mobile d'ordre ν :

— Lorsque le mouvement saisonnier comprend un nombre impair $\nu = (2p + 1)$ de saisons, on attribue pour trend :
  — à la $(p + 1)$-ème période, la valeur $T_{p+1} = \frac{x_1+x_2+\cdots+x_\nu}{\nu}$,
  — à la $(p + 2)$-ème période, la valeur $T_{p+2} = \frac{x_2+x_3+\cdots+x_{\nu+1}}{\nu}$,
  — à la $(p + 3)$-ème période, la valeur $T_{p+3} = \frac{x_3+x_4+\cdots+x_{\nu+2}}{\nu}$.

— Lorsque le mouvement saisonnier a un nombre pair $\nu = 2p$ de saisons dans l'année on attribue pour trend

$$\text{à la période } (p+1),\ \text{la valeur } T_{p+1} = \frac{0.5x_1 + x_2 + \cdots + x_\nu + 0.5x_{\nu+1}}{\nu},$$
$$\text{à la période } (p+2),\ \text{la valeur } T_{p+2} = \frac{0.5x_2 + x_3 + \cdots + x_{\nu+1} + 0.5x_{\nu+2}}{\nu},$$
$$\text{à la période } (p+3),\ \text{la valeur } T_{p+3} = \frac{0.5x_3 + x_4 + \cdots + x_{\nu+2} + 0.5x_{\nu+3}}{\nu}, \ldots$$

Détermination de la composante saisonnière S_j [7]

Il convient au préalable de déterminer l'écart entre la valeur d'observation x_k et la valeur de son trend T_k précédemment calculé, soit $S'_k = x_k - T_k$ pour $k = p + 1$, $p + 2, \ldots, n - p$;

de calculer pour chacune des ν saisons (et donc pour chaque $j = 1, 2, \ldots, \nu$) la valeur moyenne $\bar{S}'_j$ des écarts constatés S'_k relatifs à cette saison :

Afin de respecter l'hypothèse selon laquelle le mouvement saisonnier est rigoureusement périodique, on soustrait à chaque $\bar{S}'_j$ la moyenne générale $\bar{S}' = \frac{\bar{S}'_1+\cdots+\bar{S}'_{p+1}+\cdots+\bar{S}'_\nu}{\nu}$ de ces ν moyennes. Le coefficient saisonnier $S_j = \bar{S}'_j - \bar{S}'$.

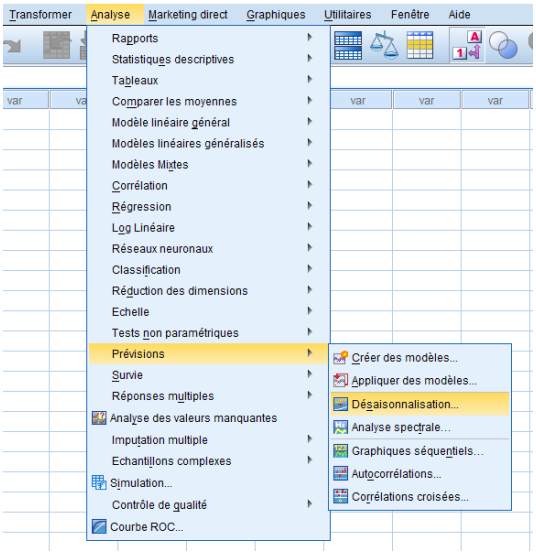
Désaisonnalisation [7]

Pour désaisonnaliser la série, il suffit alors de retrancher à chaque relevé x_k sa composante saisonnière afin de lui substituer soit $x'_k = x_k - S_j$, que l'on décompose sous la forme $x'_k = T_k + A_k$. Donc l'aléas $A_k = x_k - T_k - S_j$.

Procedure

Pour appliquer la méthode de la moyenne mobile dans SPSS :

►Allez dans le menu Analyse > Prévisions > Désaisonnalisation ou Moyennes mobiles



►Sélectionnez la série temporelle à analyser et validez

Il est aussi possible de procéder manuellement comme suit :

►Créez une nouvelle variable pour certains versions par la methode la moyenne mobile en utilisant la fonction :

$$MOVAVE(nom\_variable, k)$$

où nom-variable est la série à lisser et k l'ordre de la moyenne mobile (par exemple, 3 pour une moyenne mobile sur 3 périodes). et pour d'autre versions on écrit :

$$(variable + LAG(variable, 1) + LAG(variable, 2))/3$$

►Validez pour créer la nouvelle variable contenant les valeurs lissées.

Pour visualiser la série lissée et la série originale , allez dans Analyse > Previsons > Graphiques sequentiels puis indiquez Les variables et Le libellé axe des temps et on clique sur OK

2.4.3 Modèles ARMA

Il s'agit d'une méthode de prévision qui utilise l'information contenue dans la série elle même en vue de faire des prévisions.

Pour cela, il faut commencer par modéliser la série chronologique selon un processus ARMA (AutoRegressif Moving Average) d'ordre p et q. Les processus ARMA serviront d'abord de modèle pour décrire l'évolution des séries chronologiques et ensuite pour les prévoir.

Un modèle ARMA(p,q) : processus autorégressif d'ordre p et moyenne mobile d'ordre q pour une variable y s'écrit comme suit :

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_p y_{t-p} + \varepsilon_t - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2} - \cdots - \beta_q \varepsilon_{t-q}$$

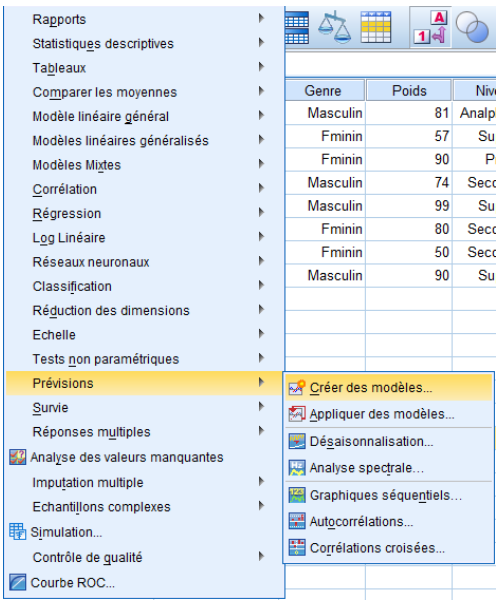
Les α et les β sont des coefficients à estimer. p et q sont des paramètres à identifier

y est alors une expression de son propre passé mais aussi du passé des erreurs ϵ . Les erreurs (de prévision) ϵ peuvent contenir des informations pertinentes pour la prévision. Ce sont des innovations, dans le sens où elles apportent du neuf par rapport au passé de la variable y.[12]

Procédure

Pour appliquer un modèle ARMA dans SPSS :

►Allez dans le menu Analyse > Prévisions > Créer des modèles



►Sélectionnez la variable dépendante à modéliser

►Dans la liste des méthodes, choisissez Modélisateur expert pour que SPSS sélectionne automatiquement le meilleur modèle, ou sélectionnez ARIMA pour spécifier manuellement un modèle ARMA

Pour un modèle ARMA, fixez le paramètre de différenciation d à 0 et indiquez les ordres p (autorégressif) et q (moyenne mobile) selon l'analyse des autocorrélations

►Cliquez sur l'onglet Options pour définir la période de prévision, enregistrer les prévisions, les intervalles de confiance et les résidus

►Lancez la modélisation. SPSS affiche les coefficients estimés, les diagnostics du modèle et les graphiques des valeurs observées et prévues

Les prévisions pour les périodes futures sont générées automatiquement.

Exemple 2.1. *[12]* ***Exemple de calcul des prévisions à partir d'un modèle ARMA(1,1) :***

$$y_{T+1} = \alpha_1 y_T + \varepsilon_{T+1} - \beta_1 \varepsilon_T$$

$$y_{T+2} = \alpha_1 y_{T+1} + \varepsilon_{T+2} - \beta_1 \varepsilon_{T+1}$$

$$Etc$$

Le terme moyenne mobile ici n'a rien à voir avec la méthode de lissage et de prévision par la moyenne mobile.

Pour obtenir des prévisions, on remplace les innovations inconnues par leur moyenne qui est égale à 0 et celles du passé par leurs estimations. Les valeurs futures de y sont remplacées par leurs prévisions :

$$\hat{y}_{T+1} = \alpha_1 y_T + 0 - \beta_1 \hat{\varepsilon}_T$$

$$\hat{y}_{T+2} = \alpha_1 \hat{y}_{T+1} + 0 + 0$$

$$Etc$$

2.4.4 Modèle ARIMA (AutoRegressive Integrated Moving Average)

Alors que le modèle ARMA suppose que la série temporelle est stationnaire, ce qui limite son application aux séries dont les propriétés statistiques sont constantes dans le temps, le modèle ARIMA introduit la différenciation pour traiter les séries non

stationnaires, très fréquentes en pratique. Cette différenciation permet de stabiliser la moyenne de la série et d'améliorer la qualité des prévisions.

Le modèle ARIMA est une extension du modèle ARMA qui permet de modéliser des séries chronologiques non stationnaires, c'est-à-dire des séries dont la moyenne ou la variance évoluent dans le temps. Alors que le modèle ARMA combine un processus autorégressif (AR) d'ordre p et un processus de moyenne mobile (MA) d'ordre q. Le modèle ARIMA ajoute une étape de différentiation d'ordre d qui vise a rendre la série stationnaire avant d appliquer le modèle ARMA

Un modèle ARIMA (p,d,q) s'écrit donc en trois étapes :

Intégration (I) : la série initiale est différenciée fois pour éliminer la tendance ou la non –stationnarité. La série différenciée est notée.

Auto régression (AR) : la valeur courante de la série différenciée est exprimée comme une combinaison linéaire de ses p valeurs passées.

Moyenne mobile (MA) :la valeur courante est aussi influencée par une combinaison linéaire des erreurs de prévision passées jusqu'à l'ordre q.

Mathématiquement, le modèle ARIMA (p, d ,q) s'écrit :

$$\Delta^d y_t = \alpha_1 \Delta^d y_{t-1} + \alpha_2 \Delta^d y_{t-2} + \cdots + \alpha_p \Delta^d y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \cdots + \beta_q \varepsilon_{t-q}$$

Ou $\Delta^d y_t$ est la série différenciée d'ordre d , α_i et β_j sont les coefficients à estimer et est un bruit blanc

Dans SPSS, la procédure ARIMA permet de spécifier les ordres p ,d ,q et d'estimer automatiquement les paramètres du modèle.[2]

Procédure

La procédure de modélisation de série chronologique évalue le lissage exponentiel, le processus autorégressif moyenne mobile intégré (ARIMA – Auto Regressive Integrated Moving Average) univarié et les modèles ARIMA multivariés (ou modèles des fonctions de transfert) pour les séries chronologiques, et produit des prévisions.

►Allez dans Analyse > Prévisions > Créer des modèles traditionnels.

►Sélectionnez la variable dépendante.

►Sous Méthode de modélisation, choisissez ARIMA.

►Spécifiez les ordres du modèle :

p : ordre autorégressif ; d : degré de différenciation (nombre de fois que la série est différenciée pour stationnariser) ; q : ordre moyenne mobile.

On peut utiliser l'option Expert Modeler pour que SPSS sélectionne automatiquement le meilleur modèle ARIMA.

►Cliquez sur OK pour exécuter.

Consultez les résultats, diagnostics et prévisions.

Remarques sur les données du modélisateur de séries chronologiques

Données : La variable dépendante et les variables indépendantes doivent être numériques.

Hypothèses : La variable dépendante et toute variable indépendante sont traitées en tant que séries chronologiques, ce qui veut dire que chaque observation représente un point dans le temps, avec des observations successives séparées par un intervalle de temps constant.

Stationnarité : Pour les modèles ARIMA personnalisés, les séries chronologiques à modéliser doivent être stationnaires. La méthode la plus efficace pour transformer une série non stationnaire en une série stationnaire nécessite l'utilisation d'une différence, disponible dans la boîte de dialogue Créer la série chronologique.

Prévisions : Pour produire des prévisions à l'aide de modèles avec des variables indépendantes (explicatives), l'ensemble de données actif doit contenir des valeurs de ces variables pour toutes les observations de la période de prévision. En outre, les variables indépendantes ne doivent pas contenir de valeurs manquantes dans la période d'estimation.

Définition des dates

Bien que ce ne soit pas une obligation, il est recommandé d'utiliser la boîte de dialogue Définir des dates pour indiquer la date correspondant à la première observation et l'intervalle de temps entre les observations suivantes. Ceci est effectué avant d'utiliser le modélisateur de séries chronologiques et résulte en un ensemble de variables qui indiquent la date associée à chaque observation. Ceci définit également une périodicité supposée des données, par exemple une périodicité de 12 si l'intervalle de temps entre les observations suivantes est d'un mois. Cette périodicité est requise si vous voulez créer des modèles saisonniers. Si vous ne voulez pas créer de modèles saisonniers et ne voulez

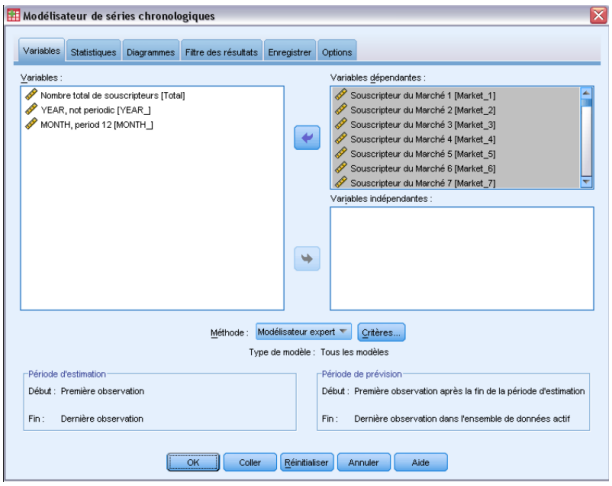
pas d'étiquette de date dans vos résultats, vous pouvez ignorer la boîte de dialogue Définir des dates. L'étiquette associée à chaque observation est alors simplement le numéro de l'observation.[2]

Pour utiliser le modélisateur de séries chronologiques

►A partir des menus, sélectionnez :

►Analyse > Prévisions > Créer des modèles...

Dans l'onglet Variables, sélectionnez une ou plusieurs variables dépendantes à modéliser



►Depuis la liste déroulante des méthodes, sélectionnez une méthode de modélisation. Pour la modélisation automatique, conservez la méthode du modélisateur expert par défaut. Vous invoquerez ainsi le modélisateur expert pour déterminer le modèle le plus approprié pour chacune des variables dépendantes.

Pour produire des prévisions :

►Cliquez sur l'onglet Options.

►Indiquez la période de prévision. Un diagramme sera produit, incluant les prévisions et les valeurs observées.

Sinon, vous pouvez :

►Sélectionnez une ou plusieurs variables indépendantes. Les variables indépendantes sont traitées comme des variables explicatives dans les analyses de régression, mais sont facultatives. Elles peuvent être incluses dans les modèles ARIMA mais pas dans les modèles de lissage exponentiel. Si vous spécifiez le modélisateur expert comme méthode

de modélisation et incluez des variables indépendantes, seuls les modèles ARIMA seront pris en considération.

Time Series Modeler

►Cliquez sur Critères pour indiquer les détails de modélisation.

►Enregistrer les prévisions, les intervalles de confiance et les résidus du bruit.

►Enregistrez les modèles estimés au format XML. Les modèles enregistrés peuvent être appliqués aux données nouvelles ou revus pour obtenir des prévisions mises à jour sans besoin de reconstruire les modèles. Ceci est effectué via la procédure Apply Time Series Models.

►Obtenir des statistiques récapitulatives dans tous les modèles estimés.

►Spécifier les fonctions de transfert pour les variables indépendantes dans les modèles ARIMA personnalisés.

►Activer la détection automatique de valeurs éloignées.

Points dans le temps spécifiques à un modèle comme valeurs éloignées pour les modèles ARIMA personnalisés.

2.5 Méthodes explicatives

2.5.1 Prévision par la régression linéaire

Celle ci est intéressante dans la mesure où elle permet d'introduire les facteurs extérieurs qui influencent le futur de la variable analysée (y).

Le point de départ de ce type de méthodes est un modèle économique. Celui-ci consiste en une présentation formalisée d'un phénomène sous forme d'équations, une présentation schématique et partielle d'une réalité plus complexe.[12]

Exemple 2.2. *Exemple : A un niveau microéconomique, au niveau de la forme :*

$$\textit{Demande de travail} = f\textit{(production, salaire)}$$

Cette relation découle d'un problème de maximisation de profit sous contrainte technologique

$$\textit{Fonction de production} : L = f(y, w)$$

Partant du modèle économique nous définissons un modèle économétrique : il s'agit d'un modèle économique faisant intervenir l'aléatoire. Un modèle économétrique comporte une variable aléatoire dite variable erreur. En effet, les phénomènes économiques sont d'une telle complexité qu'il n'est pas raisonnable de penser pouvoir expliquer une variable par, une, deux ou plus de variables explicatives. Un terme d'erreur est ajouté au modèle économique pour tenir compte des variables omises, des erreurs de mesure et autres.

Le modèle économétrique associé au modèle économique de l'exemple précédent (demande de travail) peut s'écrire simplement comme suit :

$$L = a_0 + a_1 y + a_2 w + \varepsilon$$

Dans le cas général, la variable y (endogène) est expliquée par k variables explicatives :

$$y_t = a_0 + a_1 x_t^1 + a_2 x_t^2 + \cdots + a_k x_t^k + \varepsilon_t \qquad (t = 1, \ldots, T)$$

• Si les variables sont en niveau différentes variables explicatives.

$$a_j = \frac{\partial y}{\partial x^j} \approx \frac{\Delta y}{\Delta x^j}$$

• Si les variables sont en log, les coefficients sont des élasticités

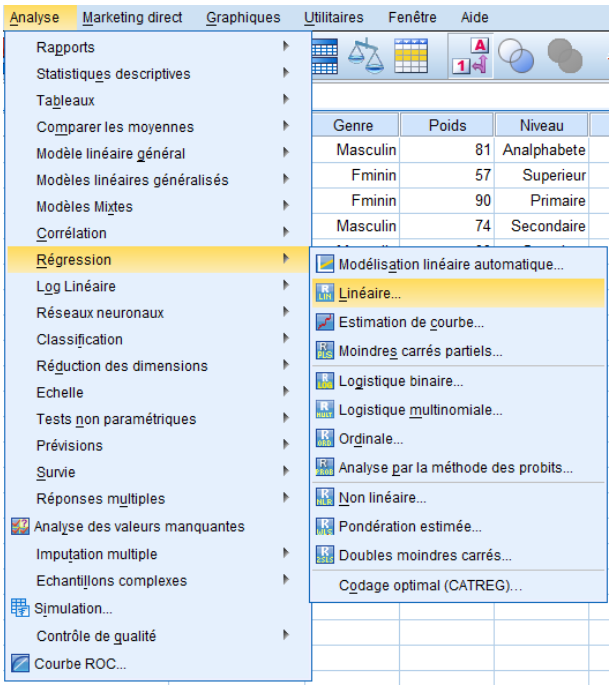
$$a_j = \frac{\partial \ln y}{\partial \ln x^j} \approx \frac{\Delta y / y}{\Delta x^j / x^j}$$

Plusieurs méthodes d'estimation des paramètres (les a_j) sont disponibles (la méthode des moindres carrés ordinaire MCO, la méthode des moindres carrés généralisée MCG et autres). La Prévision de y nécessite la connaissance (ou une approximation) des x à l'instant (T+1)

$$\hat{y}_{T+1} = \hat{a}_0 + \hat{a}_1\, x_{T+1}^1 + \cdots + \hat{a}_k\, x_{T+1}^k + 0$$

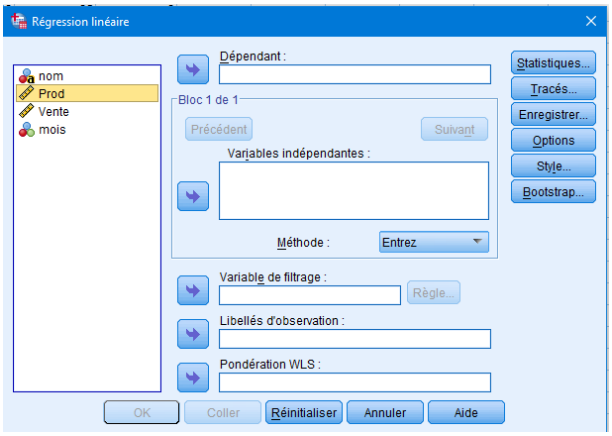
Les sont les estimations des coefficients obtenues par une méthode économétrique donnant les meilleures estimations.[12] **Procédure**

▶Aller dans Analyse > Régression > Linéaire...



►Sélectionner la variable dépendante (à prévoir) dans la case « Variable dépendante ».

►Sélectionner une ou plusieurs variables explicatives dans la case « Variables indépendantes ».



►Cliquer sur OK pour obtenir les résultats : coefficients estimés, statistiques d'ajustement, analyse des résidus.

Pour obtenir une prévision, utiliser la boîte de dialogue Enregistrer (avant de cliquer sur OK) et cocher «non standardisés» pour que SPSS crée une nouvelle variable contenant les prévisions pour chaque observation. Pour prévoir une valeur future, il suffit d'entrer les valeurs des variables explicatives dans une nouvelle ligne de données et de relancer la commande pour obtenir la prévision correspondante.

2.5.2 Modèle de fonction de transfert

Il s'agit d'une extension de la modélisation ARMA au cas où la variable y à prévoir peut être reliée à son passé mais aussi au présent et au passé d'autres variables x :

$$y_t = \delta_1 y_{t-1} + \ldots + \delta_r y_{t-r} + \lambda_0 x_t + \lambda_1 x_{t-1} + \ldots + \lambda_s x_{t-s} + \varepsilon_t$$

Des méthodes existent pour identifier r et s ; il s'agit d'utiliser la fonction d'autocorrélation (pour r) et les corrélations croisées (entre x et y) pour déterminer s.

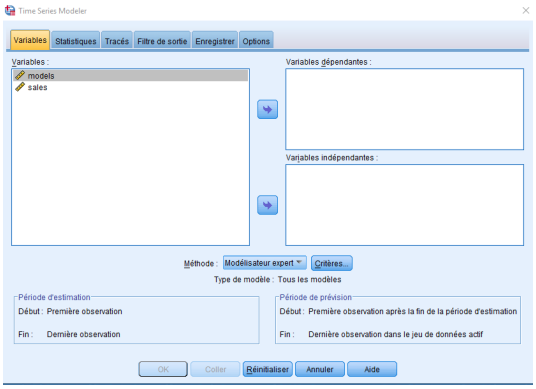
$$y_{T+1} = \delta_1 y_T + \cdots + \delta_r y_{T-r} + \lambda_0 x_{T+1} + \lambda_1 x_T + \cdots + \lambda_s x_{T-s}$$

L'estimation de cette prévision nécessite la connaissance (ou une approximation) de la valeur de x à l'instant (T+1) [12]

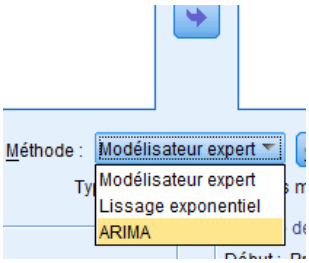
Procedure

▶Aller dans Analyse > Prévisions > Créer des modèles... (ou Analyse > Séries chronologiques > Modèle ARIMA... selon la version).

▶Sélectionner la variable dépendante à modéliser.



Dans la liste des méthodes, choisir ARIMA.



►Cliquer sur Variables indépendantes et sélectionner les variables explicatives à inclure dans le modèle (ces variables seront traitées comme des fonctions de transfert).

►Spécifier les ordres du modèle ARIMA pour la variable dépendante (p, d, q) Ensuite, dans la partie Fonction de transfert (ou Transfer Function), pour chaque variable indépendante, spécifiez :

— **Ordre du numérateur** (nombre de retards positifs)

— **Ordre du dénominateur** (nombre de retards négatifs)

— **Délai** (retard entre la variable indépendante et la variable dépendante)

Ces paramètres modélisent l'impact retardé des variables indépendantes sur la variable dépendante.

►Cliquer sur l'onglet Options pour définir la période de prévision, enregistrer les prévisions, les intervalles de confiance et les résidus.

►Lancer la modélisation. SPSS affiche les coefficients estimés, les diagnostics du modèle, et les graphiques des valeurs observées et prévues.

Les prévisions pour les périodes futures sont générées automatiquement, à condition de fournir les valeurs futures (ou prévues) des variables explicatives.

Remarque 2.1. Pour toutes les méthodes, il est recommandé de vérifier la qualité de l'ajustement (statistiques d'ajustement, analyse des résidus) avant d'utiliser les prévisions.

2.6 Critères souvent utilisés pour juger de la validité de la méthode de prévision

Introduction

Selon Murphy (cité dans Jolliffe et Stephenson, 2003), une " bonne prévision " peut l'être selon deux aspects : i) la qualité, qui examine la correspondance entre observations et prévisions, et ii) la valeur (ou utilité), qui concerne la valeur économique de la

prévision pour un utilisateur décideur.
La qualité d'une prévision est en général jugée par rapport à l'observation. Dans le cadre simple de l'évaluation d'une prévision à scénario unique, on peut directement dire que la prévision est "correcte" ou "fausse" une fois l'événement observé. En revanche, dans le cadre des prévisions probabilistes ou des prévisions d'ensemble certaines particularités sont à prendre en compte dans l'évaluation de la qualité des prévisions. Puisque ces prévisions attribuent une probabilité à l'occurrence et à la magnitude d'un événement, les prévisionnistes doivent également évaluer cette information sur l'incertitude de la prévision émise. Il s'agit de déterminer comment la probabilité des événements prévus dans l'ensemble correspond à la fréquence à laquelle les événement s sont observés (évaluation de la cohérence statistique des prévisions). Pour cela, il est nécessaire de comparer une longue série de prévisions avec la série correspondante des observations. Selon Jolliffe et Stephenson (2003), cette évaluation de la qualité d'une prévision d'ensemble consiste à trouver où se situe l'observation par rapport à la gamme de valeurs prévues par les différents membres de la prévision d'ensemble.

La prévision des valeurs futures d'une variable y peut se faire en utilisant différentes méthodes. Les prévisions obtenues peuvent être comparées et appréciées selon plusieurs critères parmi lesquels :

2.6.1 L'erreur moyenne (Mean Error, ME) :

Définition

L'erreur moyenne, notée $\bar{e}$, est une mesure simple qui évalue le biais global d'une méthode de prévision. Elle correspond à la moyenne arithmétique des erreurs de prévision sur un ensemble de données.

Pour une série de prévisions sur n périodes, l'erreur à l'instant t est :

L'erreur moyenne : (mean error) :

$$\bar{e} = \frac{1}{n} \sum e_t$$

$$\text{Tq} : e_t = y_t - \hat{y}_t$$

Ou :

— y_t est la valeur observée à l'instant t

— $\hat{y}_t$ est la valeur prévue à l'instant t

Donc l'erreur moyenne calculée par :

$$\bar{e} = \frac{1}{n}\sum_{t=1}^{n} e_t = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)$$

Dans SPSS, l'erreur moyenne peut être calculée en créant une variable d'erreur (différence entre observé et prévu) puis en calculant la moyenne de cette variable via les fonctions descriptives.[12]

2.6.2 Le carré moyen des erreurs (Mean Square Error, MSE)

Le carré moyen des erreurs, ou Mean Square Error (MSE), est l'un des critères statistiques les plus utilisés pour évaluer la performance d'une méthode de prévision. Le MSE mesure la moyenne des carrés des écarts entre les valeurs observées et les valeurs prédites par un modèle. Formellement, si y_t désigne la valeur observée à l'instant t et $\hat{y}_t$ la valeur prévue, l'erreur de prévision à l'instant t est $e_t = \ y_t - \hat{y}_t$. Le MSE se calcule alors selon la formule suivante :

$$\text{MSE} = \frac{1}{n}\sum e_t^2 = \frac{1}{n}\sum (y_t - \hat{y}_t)^2$$

Le MSE présente plusieurs avantages majeurs. Tout d'abord, en élevant les erreurs au carré, il pénalise fortement les grandes erreurs, ce qui le rend particulièrement sensible aux prévisions très éloignées des valeurs réelles. Cela permet d'identifier rapidement les modèles qui commettent des erreurs importantes, même si ces erreurs sont rares. De plus, le MSE est une mesure continue et différentiable, ce qui facilite son utilisation dans l'optimisation des modèles statistiques et des algorithmes d'apprentissage automatique.

Le MSE est largement utilisé dans la comparaison de plusieurs modèles de prévision : le modèle qui présente le MSE le plus faible est généralement considéré comme le plus performant sur l'échantillon de test. Il est aussi utilisé pour ajuster les paramètres des modèles lors de la phase d'apprentissage, notamment dans les méthodes de régression, les réseaux de neurones, ou encore les modèles de séries temporelles comme ARMA ou ARIMA.[12]

2.6.3 La racine carrée de l'erreur quadratique moyenne (Root Mean Square Error, RMSE)

La racine carrée de l'erreur quadratique moyenne, ou Root Mean Square Error (RMSE), représente l'un des critères les plus importants et les plus utilisés pour évaluer

la précision d'un modèle de prévision. Le RMSE est défini comme la racine carrée du carré moyen des erreurs (MSE) et se calcule selon la formule suivante :

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum e_t^2} = \sqrt{\frac{1}{n}\sum (y_t - \hat{y}_t)^2}$$

(exprimé en la même unité que la variable).

et étant l'erreur de prévision pour un instant t (la différence entre la valeur observée et la valeur prévue par une méthode quelconque).

Le RMSE présente plusieurs avantages majeurs par rapport aux autres critères d'évaluation. Tout d'abord, contrairement au MSE qui s'exprime dans le carré de l'unité de la variable étudiée, le RMSE est exprimé dans la même unité que la variable d'origine, ce qui facilite considérablement son interprétation. Par exemple, si la variable étudiée est mesurée en euros, le RMSE sera également exprimé en euros, permettant ainsi une compréhension directe de l'ampleur moyenne des erreurs de prévision.

En outre, le RMSE conserve la propriété du MSE de pénaliser davantage les grandes erreurs que les petites, car il prend en compte le carré des écarts avant d'en extraire la racine carrée. Cette caractéristique est particulièrement utile dans les contextes où les grandes erreurs de prévision sont plus préjudiciables que les petites, le RMSE offre ainsi une mesure de l'écart-type des erreurs de prévision, indiquant la dispersion des valeurs prédites autour des valeurs réelles.

La meilleure méthode est celle qui fournie les valeurs les plus faibles pour ces critères.[12]

2.7 conclusion

En conclusion, ce chapitre a permis de mettre en lumière l'importance des calculs de prévision dans l'analyse des données, en insistant sur le rôle central du logiciel SPSS dans cette démarche. Grâce à ses fonctionnalités avancées, SPSS offre un environnement adapté pour appliquer diverses méthodes de prévision, L'utilisation de SPSS facilite ainsi la manipulation des données, l'automatisation des calculs et la visualisation des résultats, ce qui constitue un atout majeur pour les analystes et décideurs. En combinant différentes méthodes, il est possible d'adapter les prévisions aux spécificités des données et aux objectifs de l'entreprise, améliorant ainsi la précision et la pertinence des décisions stratégiques.

Ce chapitre pose donc les fondations pratiques nécessaires pour exploiter pleinement les données à travers des techniques de prévision robustes, préparant ainsi le terrain pour leur application concrète dans le contexte industriel étudié. La maîtrise de ces outils est essentielle pour anticiper les évolutions futures et optimiser la gestion des ressources, contribuant ainsi à la performance globale de l'entreprise.

Chapitre 3

Analyse prévisionnelle des données de l'entreprise ENEL – Application SPSS

3.1 Introduction

Ce troisième chapitre est consacré à l'analyse prévisionnelle des données, étape essentielle pour transformer les données collectées en informations exploitables. Après avoir présenté les concepts théoriques et les méthodes de prévision, ce chapitre s'attache à leur application concrète à travers le logiciel SPSS. Il débutera par une brève présentation de l'entreprise afin de mieux situer le contexte de l'analyse. Ensuite, nous mettrons en œuvre les différentes techniques de prévision précédemment étudiées, en utilisant SPSS pour traiter et modéliser les données. Enfin, les résultats obtenus seront présentés et interprétés

3.2 Representation de l'entreprise

Dans ce point nous allons présenter l'entreprise à travers plusieurs critères.

3.2.1 Historique de l'entreprise Electro-Industries

Electro-industrie est l'une des unités de production de SONELEC, qui a été l'une des plus importantes entreprises du pays. Cette entreprise, possède plusieurs unités de production réparties à travers le territoire, est créée en 1969. Celle-ci a existé jusqu'à la restructuration des secteurs industriels en plusieurs entreprises juridiquement indépendantes composées des unités commerciales et de production en 1983.

L'ENEL est l'une de ces entreprises qui a occupé une place dans le secteur industriel. Créée en 1985 par une convention qui est signée entre SONELEC et les patrimoines Allemands en l'occurrence :

- SIMENS pour les produits alternateurs, générateurs et les groupes électrogènes ;
- TRAFO-UNION pour le produit transformateur :
- FRITZ-WERNER pour la partie engineering du projet ;
- La construction et l'infrastructure sont réalisées par les entreprises algériennes telles qu'ECOTEC, COSIDER et BATIMETAL

L'ENEL a deux secteurs de productions essentiels. Le premier est le secteur des transformateurs, qui a commencé la production à la même année de création 1985. Le deuxième est le secteur des moteurs/alternateurs qui a commencé la production en 1986. Ces produits sont fabriqués sous la licence SIEMENS jusqu'en 1992.

En 1991 une extension de ses capacités de production de transformateur de 1500 à 5000 unités/an, développement de la gamme de moteurs monophasés, développement de l'activité des groupes électrogènes, développement de moteurs destinés à la climatisation, extension verticale de la gamme de transformateurs (2000 KVA) et l'extension horizontale de la gamme du moteur en types et variantes.

L'ENEL, a connu une autre restructuration en 1999. Elle a changé de statut pour devenir une entreprise autonome Electro-Industrie. Cette dernière est spécialisée dans la fabrication et la commercialisation des transformateurs, moteurs électriques et la commercialisation des groupes électrogènes (activité insignifiante).

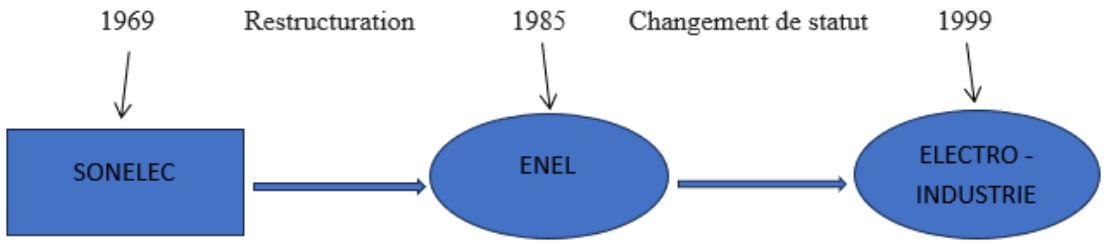


FIGURE 3.1 – Evolution de l'Electro-Industrie.

Source : réalisé par nos soins à partir des documents internes de l'entreprise.

En effet, les produits fabriqués par l'entreprise sont réalisés et contrôlés en suivant les normes DIN/VDE (institut allemand des normes/ groupe allemand d'électricité). Ses produits sont destinés essentiellement au marché algérien, avec une part de marché qui dépasse les 70%. [13]

3.2.2 La situation géographique et la superficie de l'entreprise

L'Electro-Industries est située à mi-chemin entre les deux localités de Fréha et d'Azazga à 30 KM du chef-lieu de la wilaya de Tizi-Ouzou, et à 08KM du chef-lieu de la daïra d'Azazga. Elle occupe une superficie de 35 hectares dont 7 hectares sont bâtis, et sur lesquels se trouvent les unités de production ainsi que la direction générale.[13]

3.2.3 Le statut juridique et le capital social

Conformément à la loi de 88/01 du 13/01/1988 qui adopte plusieurs règles pour la création des EPE, Electro-Industrie est une entreprise publique économique, Société Par Action (EPE-SPA) avec un capital social de quatre milliards sept cent cinquante-trois millions de dinars (4 753 000 000 DA) détenus totalement par le Groupe ELEC EL DJAZAIR pour le compte de l'Etat. Cette entreprise a été créée dans le cadre du « projet de l'industrie industrialisant » dont l'objectif est de réduire la dépendance extérieure.[13]

3.2.4 Le domaine d'activité de l'entreprise

Electro-Industries est leader national et continentale dans le domaine de l'industrie électrotéchnique. Son activité principale étant la conception, la fabrication et la commercialisation des transformateurs de distribution, moteurs électriques, alternateurs et des groupes électrogènes. On peut les scinder en deux catégories comme le montre le tableau suivant :[13]

Activité principale	**Activité secondaire**
Conception, fabrication et commercialisation des : - Transformateurs de distribution - Moteurs électriques	- Fabrication des groupes électrogènes - Maintenance des équipements de production. - Rénovation et réparation des moyens de fabrication - Contrôle et vérification des matières - Activités de sous-traitance : laboratoire physique et chimie, moulage sous-pression des pièces en aluminium, découpage des pièces en tôles d'acier, fabrication des pièces d'usinage spécifiques, métrologie.

FIGURE 3.2 – Activités principales et secondaires d'Electro-Industries.

Depuis 2016 Electro-Industries est composée de trois unités, toutes situées sur un même site :

Unité de fabrication de transformateurs de distribution (UTR)

Elle prend en charge la production des transformateurs de distribution (moyenne tension/basse tension), c'est l'unité qui génère plus de rentabilité pour l'entreprise.

L'unité transformateur a une gamme de production très diversifiée qui ont une puissance allant de 50 à 2000 Kilo Volt Ampère (KVA), avec une tension usuelle en moyenne tension de 5.5, 10 et 30 Kilo Volt (KV), et une tension usuelle en basse tension de 0.4 KV. Les transformateurs fabriqués par Electro-Industrie sont des transformateurs abaisseurs et non élévateurs de tension.

La capacité théorique de production des transformateurs est de 5000 unités /an, elle réalise 90% du chiffre d'affaires et répond à 70% de la demande du marché. Alors que la capacité réelle de production est de 3 455 unités en 2017, soit un taux réel de 69%.

Unité de fabrication des moteurs électriques, alternateurs, groupes électrogènes (UMAGE)

Cette unité se positionne en seconde place dans l'activité de l'entreprise, et occupe la plus grande surface dans cette dernière.

Elle se spécialise dans la fabrication des moteurs électriques destinés à être montés dans différentes machines telles que les pompes à eau, les broyeuses, les cintreuses ; Elle fabrique aussi des groupes électrogènes et des alternateurs qui sont nécessaires à leur montage. La puissance de ces produits est de :

— Puissance des moteurs électriques : de 0.25 à 400 KVA ;
— Puissance des alternateurs : de 17.5 à 200 KV ;
— Puissance des groupes électrogènes : de 100 à 200 KVA.

Unité de prestations techniques (UPT)

Pour maintenir ses équipements et installations, l'entreprise dispose de sa propre unité de prestation technique pour les deux unités de production (UTR et UMAGE).

3.2.5 Les structures organisationnelles d'Electro-Industrie

Les structures organisationnelles de l'El sont représentées par l'organigramme suivant . Cet organigramme est de type fonctionnel avec domination des liens hiérarchiques verticaux.

Nous retrouvons trois niveaux de structure [13] :

Les structures gérées par des assistants

Elles sont au nombre de six et se composent du : Contrôle de gestion, secrétariat, l'audit interne, la sécurité interne, la qualité hygiène sécurité environnement et la communication.

Les structures gérées par des directeurs

Il existe cinq directions qui sont en relation avec : Les ressources humaines et organisation, la fonction commerciale et marketing, les achats et approvisionnement, la finance et la Comptabilité, le développement industriel et partenariat.

Les structures gérées par les assistants ainsi que les directions sont liées aux processus de management et support, indispensables au fonctionnement des activités de production.

Les structures représentant les unités opérationnelles

On trouve trois unités au sein de l'El qui sont : L'unité transformatrice, l'unité moteurs électriques et l'unité prestations techniques au sein desquelles se réalisent les activités de développement, de production, de contrôle, des produits finis et des outillages. Ces structures sont toutes rattachées à la direction générale qui a pour mission d'assurer la coordination de l'ensemble des activités.[13]

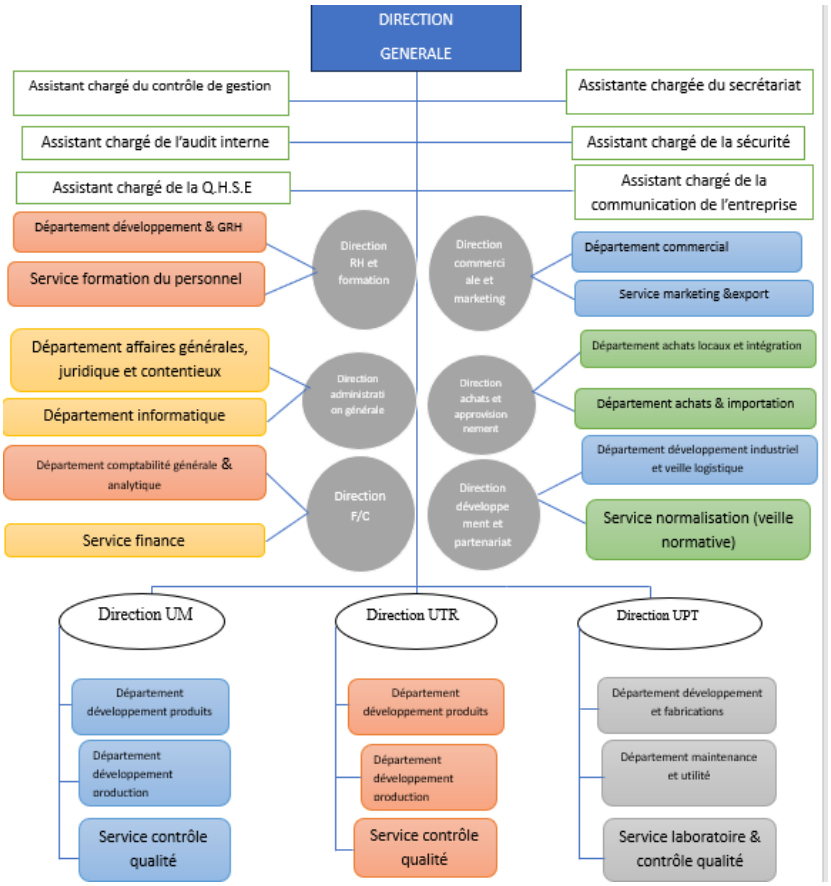


FIGURE 3.3 – Organigramme d'Electro-Industries.

3.2.6 L'environnement d'Electro-Industrie

L'environnement est un ensemble de facteurs externes qui ont une incidence sur les décisions de l'entreprise et sur l'évolution de son activité, parmi lesquels, nous pouvons citer [13] :

Les clients potentiels de l'entreprise

Les principaux clients de l'entreprise sont SONELGAZ, KAHRIF et leurs filiales pour le produit transformateur, et pour le moteur électrique on trouve PONAL, ERID, ENMTP et divers opérateurs publics et privés ainsi que les particuliers.

Les distributeurs et les intermédiaires sont des entreprises privées ou des agents agréés revendeurs principalement pour les entreprises nationales.

Les fournisseurs de l'entreprise

Les principaux fournisseurs locaux sont SIDER, NAFTAL et les divers opérateurs publics et privés. L'entreprise Electro-Industrie travaille aussi avec des fournisseurs étrangers qui sont principalement des entreprises françaises SOOFILIS, PROCELIS et MATELEC ainsi que l'entreprise portugaise ASEMETAL

Les concurrents de l'entreprise

Les principaux concurrents nationaux d'Electro-Industrie sont SWEIDY sis à Ain-Defla et l'entreprise NUCON sis à Sor-El-Ghozlane. L'entreprise est concurrencée par des sociétés étrangères principalement l'entreprise allemande SIEMENS et deux entreprises Portugaises LEPOYSOMMER et EFACEC.

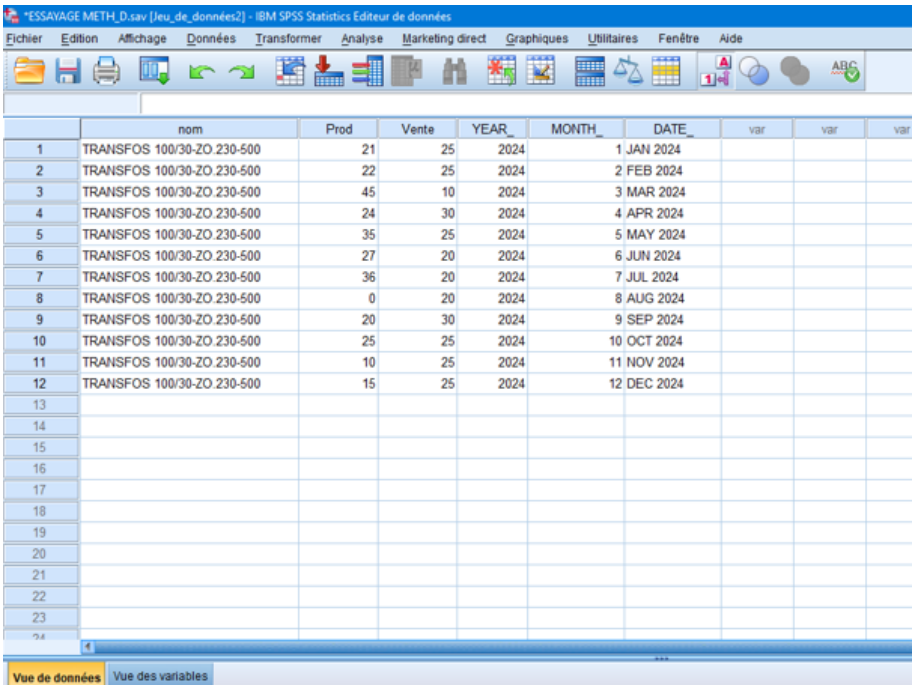
Les services de l'Etat

Electro-Industrie est en relation avec les divers services de l'Etat tels que :

— Les banques : la BEA et la BDL, qui contribuent au financement de ses projets ;
— Les assurances : la CAAT et la SAA pour une assurance tout risque humain et matériel ;
— Les transports : AIR ALGERIE, la CNAN pour assurer ses importations et ses exportations
— La douane : qui assure le dédouanement de la marchandise et de la matière première dans le cadre de l'importation et de l'exportation ;
— Les services des impôts : pour assurer le règlement de ses différents impôts et taxes dus à payer à la recette des impôts tels que la TVA, la TAP, I'IBS et l'IRG/salaire ;
— Les services sociaux : tels que la CNAS et la CASNOS.

3.3 Application des méthodes de prévision : étude et résultats

On a la base de données suivante pour le transformateurTRANSFOS 100/30-ZO.230-500



	nom	Prod	Vente	YEAR_	MONTH_	DATE_	var	var	var
1	TRANSFOS 100/30-ZO 230-500	21	25	2024	1	JAN 2024			
2	TRANSFOS 100/30-ZO 230-500	22	25	2024	2	FEB 2024			
3	TRANSFOS 100/30-ZO 230-500	45	10	2024	3	MAR 2024			
4	TRANSFOS 100/30-ZO 230-500	24	30	2024	4	APR 2024			
5	TRANSFOS 100/30-ZO 230-500	35	25	2024	5	MAY 2024			
6	TRANSFOS 100/30-ZO 230-500	27	20	2024	6	JUN 2024			
7	TRANSFOS 100/30-ZO 230-500	36	20	2024	7	JUL 2024			
8	TRANSFOS 100/30-ZO 230-500	0	20	2024	8	AUG 2024			
9	TRANSFOS 100/30-ZO 230-500	20	30	2024	9	SEP 2024			
10	TRANSFOS 100/30-ZO 230-500	25	25	2024	10	OCT 2024			
11	TRANSFOS 100/30-ZO 230-500	10	25	2024	11	NOV 2024			
12	TRANSFOS 100/30-ZO 230-500	15	25	2024	12	DEC 2024			
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									
23									

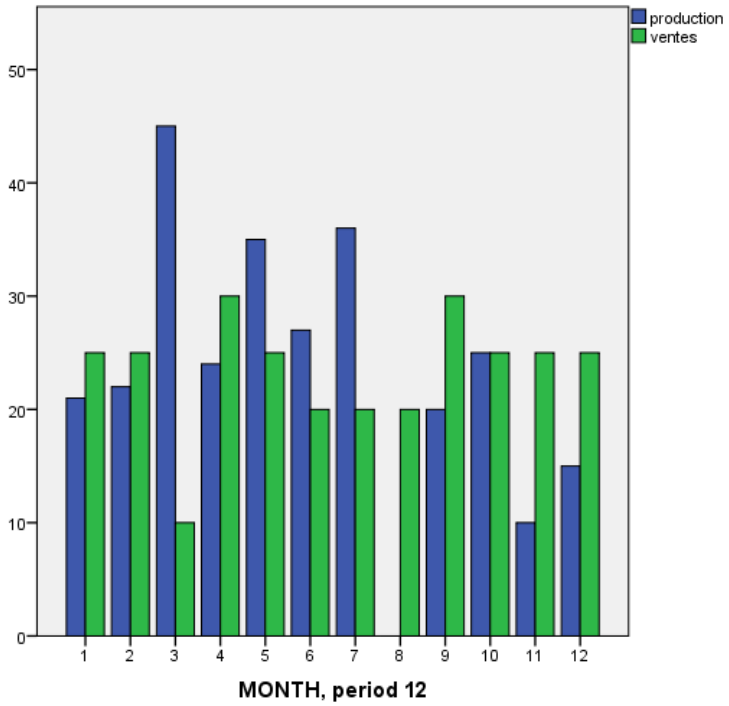


FIGURE 3.4 – Representation graphique

Statistiques descriptives

Statistiques

		ventes	production
N	Valide	12	12
	Manquant	12	12
Moyenne		23,00	23,00
Médiane		25,00	23,00
Mode		25	0[a]
Ecart type		5,000	12,033
Variance		28,000	144,000
Plage		20	45
Minimum		10	0
Maximum		30	45
Somme		280	280
Percentiles	25	20,00	16,00
	50	25,00	23,00
	75	25,00	33,00

a. Présence de plusieurs modes. La plus petite valeur est affichée.

tables de fréquences

production

		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	0	1	8,0	8,0	8,0
	10	1	8,0	8,0	16,0
	15	1	8,0	8,0	25,0
	20	1	8,0	8,0	33,0
	21	1	8,0	8,0	41,0
	22	1	8,0	8,0	50,0
	24	1	8,0	8,0	58,0
	25	1	8,0	8,0	66,0
	27	1	8,0	8,0	75,0
	35	1	8,0	8,0	83,0
	36	1	8,0	8,0	91,0
	45	1	8,0	8,0	100,0
	Total	12	100,0	100,0	

ventes

		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	10	1	8,0	8,0	8,0
	20	3	25,0	25,0	33,0
	25	6	50,0	50,0	83,0
	30	2	16,0	16,0	100,0
	Total	12	100,0	100,0	

— **Variable « ventes »**

— Moyenne : 23

En moyenne, il s'est vendu 23 unités par mois sur la période étudiée.

— Médiane : 25

La moitié des observations sont inférieures ou égales à 25, ce qui montre une légère asymétrie vers le bas par rapport à la moyenne.

- Mode : 25
  La valeur la plus fréquente est 25, ce qui confirme la tendance centrale autour de ce chiffre.
- Écart-type : 5,000
  Les ventes mensuelles varient en moyenne de 5 unités autour de la moyenne, ce qui indique une variabilité modérée.
- Variance : 28 000
  Cette valeur traduit la dispersion globale des ventes.
- Plage : 20 (min = 10, max = 30)
  Les ventes mensuelles varient entre 10 et 30 unités, ce qui donne l'étendue des fluctuations possibles.
- Percentiles (25, 50, 75) : 20, 25, 25
  25 % des valeurs sont inférieures à 20, 50 % à 25, et 75 % à 25, ce qui montre une concentration des valeurs élevées autour de 25.

Interprétation : La variable « ventes » présente une distribution relativement concentrée autour de 25, avec une variabilité modérée. Cela suggère que des méthodes de prévision simples comme la moyenne mobile ou la régression linéaire pourraient bien fonctionner,

- **Variable « production »**
  - Moyenne : 23
    La production mensuelle moyenne est identique à celle des ventes, soit 23 unités.
  - Médiane : 23
    La moitié des observations sont inférieures ou égales à 23, ce qui indique une distribution centrée.
  - Mode : 0
    La valeur la plus fréquente est 0, ce qui révèle la présence de mois sans production.
  - Écart-type : 12,033
    La production varie fortement d'un mois à l'autre, avec un écart-type nettement supérieur à celui des ventes.
  - Variance : 144,000
    Cette forte variance confirme la grande dispersion des valeurs.

— Plage : 45 (min = 0, max = 45)
La production mensuelle varie de 0 à 45 unités, ce qui indique des fluctuations très importantes.
— Percentiles (25, 50, 75) : 16, 23, 33
25 % des valeurs sont inférieures à 16, 50 % à 23, et 75 % à 33, ce qui montre une distribution plus étalée que pour les ventes.

Interprétation : La variable « production » est beaucoup plus dispersée que les ventes, avec des périodes de production nulle et des pics importants. Cette forte variabilité nécessite l'utilisation de méthodes de prévision capables de gérer les fluctuations soudaines, comme les modèles ARIMA ou les moyennes mobiles pondérées.

3.4 Prévision

3.4.1 méthode des moyennes mobiles

Production

La méthode des moyennes mobiles pour la production n'est pas applicable vu la saisonnalité aout

Ventes :

On n'a pas de valeurs nulles (saisonnalité) donc appliquant la méthode des moyennes mobiles on aura ce résultat :

nom	production	ventes	YEAR_	MONTH_	DATE_	moymobvent
TRANSFOS 100/30-ZO 230-500	21	25	2024	1	JAN 2024	.
TRANSFOS 100/30-ZO 230-500	22	25	2024	2	FEB 2024	.
TRANSFOS 100/30-ZO 230-500	45	10	2024	3	MAR 2024	20
TRANSFOS 100/30-ZO 230-500	24	30	2024	4	APR 2024	22
TRANSFOS 100/30-ZO 230-500	35	25	2024	5	MAY 2024	22
TRANSFOS 100/30-ZO 230-500	27	20	2024	6	JUN 2024	25
TRANSFOS 100/30-ZO 230-500	36	20	2024	7	JUL 2024	22
TRANSFOS 100/30-ZO 230-500	0	20	2024	8	AUG 2024	20
TRANSFOS 100/30-ZO 230-500	20	30	2024	9	SEP 2024	23
TRANSFOS 100/30-ZO 230-500	25	25	2024	10	OCT 2024	25
TRANSFOS 100/30-ZO 230-500	10	25	2024	11	NOV 2024	27
TRANSFOS 100/30-ZO 230-500	15	25	2024	12	DEC 2024	25

La prévision pour chaque mois de 2025 :
Pour prévoir janvier 2025, On Prend la moyenne des 3 (ou N) derniers mois de 2024. Pour février 2025, prenez la moyenne des 3 derniers mois connus (par exemple, novembre, décembre 2024 et janvier 2025 prévisionnel), etc. Dans SPSS on calcul la variable moymob = (moymobvent + LAG (moymobvent,1) + LAG (moymobvent,2))/3

et on ajoute à chaque fois dans le mois prochain de (moymobvent) la dernière valeur obtenue dans (moymob)

Résultats :

nom	production	ventes	YEAR_	MONTH_	DATE_	moymobvent
TRANSFOS 100/30-ZO 230-500	.	.	2025	1	JAN 2025	25
TRANSFOS 100/30-ZO 230-500	.	.	2025	2	FEB 2025	26
TRANSFOS 100/30-ZO 230-500	.	.	2025	3	MAR 2025	25
TRANSFOS 100/30-ZO 230-500	.	.	2025	4	APR 2025	25
TRANSFOS 100/30-ZO 230-500	.	.	2025	5	MAY 2025	25
TRANSFOS 100/30-ZO 230-500	.	.	2025	6	JUN 2025	25
TRANSFOS 100/30-ZO 230-500	.	.	2025	7	JUL 2025	25
TRANSFOS 100/30-ZO 230-500	.	.	2025	8	AUG 2025	25
TRANSFOS 100/30-ZO 230-500	.	.	2025	9	SEP 2025	25
TRANSFOS 100/30-ZO 230-500	.	.	2025	10	OCT 2025	25
TRANSFOS 100/30-ZO 230-500	.	.	2025	11	NOV 2025	25
TRANSFOS 100/30-ZO 230-500	.	.	2025	12	DEC 2025	25

Intérprétation :

MAPE (Mean Absolute Percentage Error)

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \times 100 \approx 21\%$$

RMSE (Root Mean Squared Error)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2} \approx 5.2$$

Les prévisions montrent une demande mensuelle stable autour de 22 à 25 unités, avec une erreur moyenne d'environ 21%, ce qui indique que le modèle est globalement fiable. Cependant, la forte erreur en mars montre que le modèle peut surestimer la demande lors de variations soudaines. Ce modèle reste un bon outil de planification pour les mois réguliers, à condition de rester vigilant face aux fluctuations ponctuelles.

3.4.2 Méthode des courbes de croissance

Définition 3.1. :

R-deux (R^2) : Mesure la proportion de la variance expliquée par le modèle (0 = aucune explication, 1 = explication parfaite).

F : Teste la significativité globale du modèle (plus F est élevé, plus le modèle est significatif).

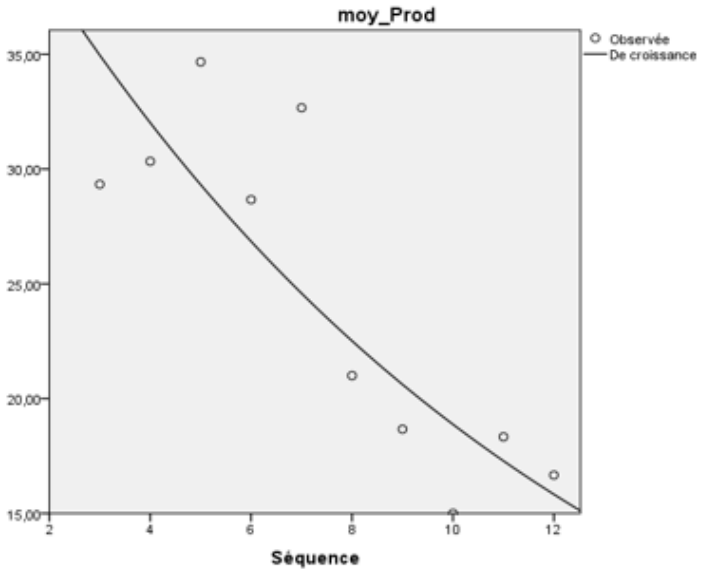
ddI1 & ddI2 : Degrés de liberté (ddI1 = nombre de paramètres estimés, ddI2 = nombre d'observations moins le nombre de paramètres).

Sig. (p-value) : Indique si le modèle est statistiquement significatif (valeur $< 0{,}05$ = significatif).

Les Constante : b1, b2 : Coefficients du modèle.

Production :

DATE_	courbe_de_croissance
JAN 2025	14
FEB 2025	13
MAR 2025	12
APR 2025	11
MAY 2025	10
JUN 2025	9
JUL 2025	9
AUG 2025	8
SEP 2025	7
OCT 2025	7
NOV 2025	6
DEC 2025	6



Récapitulatif du modèle et estimations de paramètres

Variable dépendante: moy_Prod

Equation	Récapitulatif des modèles					Estimations des paramètres		
	R-deux	F	ddl1	ddl2	Sig.	Constante	b1	b2
Quadratique	,737	9,825	2	7	,009	33,897	-,173	-,125
S	,502	8,076	1	8	,022	2,720	2,731	
De croissance	,744	23,202	1	8	,001	3,818	-,088	

Remarque : on n'a pas appliqué la méthode directement sur les valeurs réelles mais plutôt sur les valeurs lissées qu'on a trouvé par la méthode des moyennes mobiles et sa est du à l'effet de saisonnalité

Interprétation :

Le modèle de croissance appliqué aux données de production présente un bon ajustement global, avec un coefficient de détermination $R^2 = 0{,}744$, indiquant qu'il explique 74,4 % de la variance de la variable dépendante. La valeur de $F = 23{,}202$ et la significativité (Sig. = 0,001) confirment que le modèle est statistiquement significatif. Les paramètres estimés montrent une constante de 3,818 et un coefficient de tendance $b_1 = -0{,}088$, traduisant une dynamique décroissante de la production au fil du temps.

Sur le plan de la précision des prévisions, le modèle affiche un MAPE d'environ 21 % et un RMSE de 5,2, ce qui traduit une erreur modérée mais acceptable pour un usage de planification. La courbe de croissance permet de lisser les variations et de visualiser clairement la tendance, avec des valeurs prévues allant de 42 à 16 unités. Toutefois, elle ne capture pas fidèlement les fluctuations extrêmes, comme les chutes brutales, et tend à sous-estimer les pics de production.

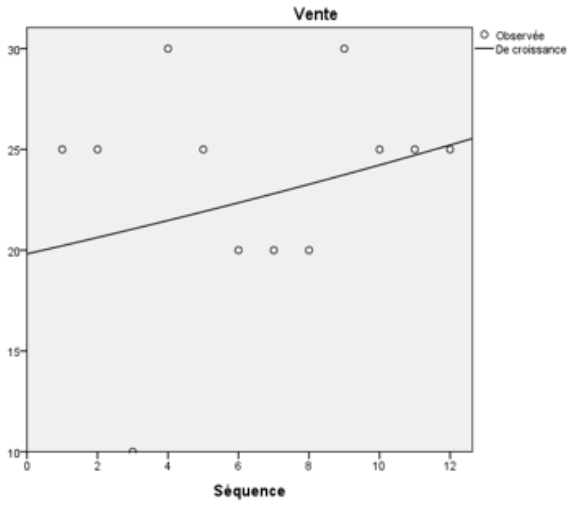
Ventes :

Récapitulatif du modèle et estimations de paramètres

Variable dépendante: Vente

Equation	Récapitulatif des modèles					Estimations des paramètres		
	R-deux	F	ddl1	ddl2	Sig.	Constante	b1	b2
Quadratique	,070	,339	2	9	,721	23,636	-,692	,077
S	,001	,006	1	10	,939	3,124	-,027	
De croissance	,062	,660	1	10	,435	2,987	,020	

DATE_	courbe_de_croissance_vente
JAN 2025	26
FEB 2025	26
MAR 2025	27
APR 2025	27
MAY 2025	28
JUN 2025	28
JUL 2025	29
AUG 2025	30
SEP 2025	30
OCT 2025	31
NOV 2025	31
DEC 2025	32



Interprétation :

Le modèle de croissance appliqué aux données de production présente un ajustement très faible, avec un coefficient de détermination $R^2 = 0{,}062$, indiquant qu'il n'explique

que 6,2 % de la variance de la variable dépendante. La valeur de $F = 0{,}660$ et la significativité (Sig. = 0,435) montrent que le modèle n'est pas statistiquement significatif. Les paramètres estimés indiquent une constante de 2,987 et un coefficient de tendance $b_1 = 0{,}020$, traduisant une très légère dynamique haussière, mais statistiquement non significative.

Sur le plan de la précision des prévisions, le modèle affiche un MAPE d'environ 18 % et un RMSE de 3,6, ce qui reflète une erreur modérée, acceptable uniquement pour des tendances générales. Les valeurs prévues s'échelonnent de 20 à 25 unités, alors que les données réelles varient de 10 à 30 unités, révélant que le modèle sous-estime les pics et surestime certains creux, notamment en mars et septembre.

Ainsi, bien que la méthode des courbes de croissance permette une lecture lissée et simplifiée des données, elle ne reflète pas fidèlement les fluctuations réelles de la production. Ce modèle est donc inadapté pour des prévisions précises ou une planification stratégique, mais peut servir de support pour dégager une tendance globale, à condition de rester attentif aux variations ponctuelles importantes.

3.4.3 Modèle ARMA

L'application de ce modèle donne les resultat suivantes :

Qualité de l'ajustement

Statistiques d'ajustement	Moyenne	Erreur standard	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
R-deux stationnaire	,295	.	,295	,295	,295	,295	,295	,295	,295	,295	,295
R-deux	,295	.	,295	,295	,295	,295	,295	,295	,295	,295	,295
RMSE	10,032	.	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032	10,032
MAPE	31,239	.	31,239	31,239	31,239	31,239	31,239	31,239	31,239	31,239	31,239
MaxAPE	81,052	.	81,052	81,052	81,052	81,052	81,052	81,052	81,052	81,052	81,052
MAE	6,927	.	6,927	6,927	6,927	6,927	6,927	6,927	6,927	6,927	6,927
MaxAE	12,077	.	12,077	12,077	12,077	12,077	12,077	12,077	12,077	12,077	12,077
BIC normalisé	5,484	.	5,484	5,484	5,484	5,484	5,484	5,484	5,484	5,484	5,484

Statistiques du modèle

Modèle	Nombre de prédicteurs	Statistiques de la qualité de l'ajustement	Ljung-Box Q(18)			Nombre de valeurs extrêmes
		R-deux stationnaire	Statistiques	DL	Sig.	
Prod-Modèle_1	1	,295	.	0	.	0

Interprétation :

— R^2 (Stationnaire et Standard) = 0.295(29,5%)

- Le modèle explique environ 29,5 % de la variance de la série corrigée.

• Cela reste une performance modérée, indiquant que près de 70 % de la variance reste inexpliquée.

• Le traitement des valeurs manquantes (zéro d'août) a amélioré la cohérence des données, mais la saisonnalité limite encore la qualité de l'ajustement

— RMSE (Root Mean Square Error) : L' Erreur quadratique moyenne(RMSE) est de 10,032 indique une erreur moyenne assez importante, ce qui confirme la difficulté du modèle à bien ajuster la série.

— MAPE (Mean Absolute Percentage Error) : 31,239 L' Erreur absolue moyenne en pourcentage de 31,2 %. En moyenne, le modèle se trompe de plus de 30 % dans ses prévisions, ce qui est une erreur importante pour un modèle de prévision.

— MaxAPE (Maximum Absolute Percentage Error) : 81,052 Cette statistique de l' Erreur absolue maximale en pourcentage atteind 81 %, ce qui souligne la présence de valeurs aberrantes ou de périodes où le modèle ajuste très mal les données (probablement autour du mois AOUT).

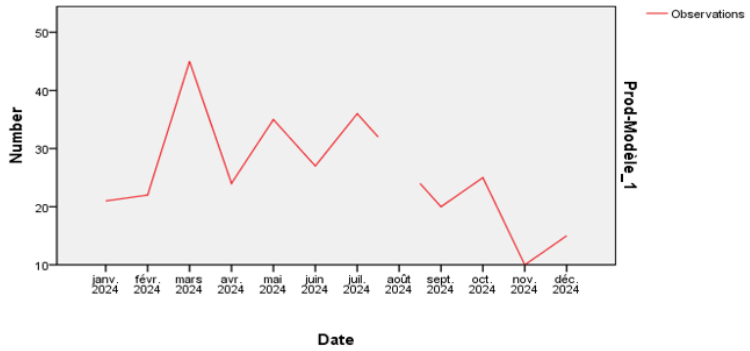
— MAE (Mean Absolute Error) : 6,927 L'erreur absolue moyenne est de 6,927, En moyenne, les prévisions s'écartent des observations de presque 7 unités ce qui confirme un ajustement modéré du modèle.

— BIC (Bayesian Information Criterion) : 5,484 Le BIC (Bayesian Information Criterion) normalisé est utilisé pour comparer différents modèles. Ici, sa valeur seule,(sans autre modèle pour comparaison)ne permet pas d'interprétation directe, mais une valeur plus faible indique généralement un meilleur compromis entre qualité d'ajustement et complexité du modèle. Et donc cette valeur ne suffit pas à juger seule de la qualité Dans le 2eme tableau intitulé « statistiquues du modèles » ona :

• Test de Ljung-Box : Avec 0 degré de liberté (DL), ce test n'a pas pu être calculé correctement, ce qui est inhabituel et pourrait indiquer un problème dans la spécification du modèle

• Nombre de valeurs extrêmes : 0 - Aucune valeur aberrante n'a été détectée par le modèle

Analyse du graphique de la série temporelle

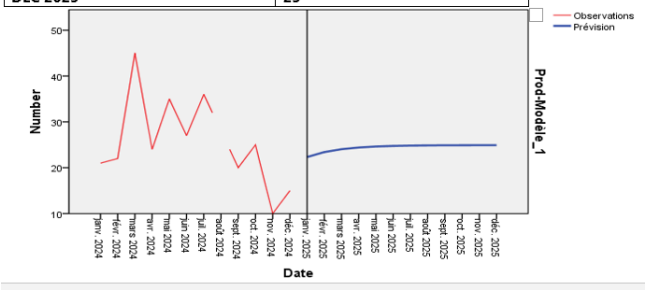


Interpretation : Le graphique montre une tendance à la baisse sur l'ensemble de l'année, avec des valeurs plus élevées au premier semestre et une diminution progressive au second semestre. Autrement dit : La série présente :

- Un pic important en mars (environ 45 unités)
- Des pics secondaires en mai et juillet (environ 35 unités)
- L absence d observation en août, qui correspond au mois traité comme valeur manquante (mois de congé)
- Une reprise en octobre suivie d'une nouvelle baisse en fin d'année Traitement de la valeur d'août : On peut voir que la valeur d'août n'est pas à zéro dans le graphique, ce qui confirme que votre traitement de cette valeur comme "manquante" a bien fonctionné. Le modèle a probablement interpolé cette valeur.

Prévision de production :

Périodes :	Valeurs
JAN 2025	22
FEB 2025	23
MAR 2025	24
APR 2025	24
MAY 2025	25
JUN 2025	25
JUL 2025	25
AUG 2025	25
SEP 2025	25
OCT 2025	25
NOV 2025	25
DEC 2025	25



Conclusion : Même après avoir déclaré le zéro d'août comme valeur manquante, le modèle ARMA montre une capacité limitée à modéliser la série de production. Les erreurs restent élevées, surtout en pourcentage, c'est à dire ; Le traitement des données manquantes a aidé, mais il ne suffit pas à lui seul pour obtenir un bon ajustement

Ventes :

Qualité de l'ajustement

Statistiques d'ajustement	Moyenne	Erreur standard	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
R-deux stationnaire	,409	.	,409	,409	,409	,409	,409	,409	,409	,409	,409
R-deux	,409	.	,409	,409	,409	,409	,409	,409	,409	,409	,409
RMSE	4,837	.	4,837	4,837	4,837	4,837	4,837	4,837	4,837	4,837	4,837
MAPE	17,589	.	17,589	17,589	17,589	17,589	17,589	17,589	17,589	17,589	17,589
MaxAPE	87,249	.	87,249	87,249	87,249	87,249	87,249	87,249	87,249	87,249	87,249
MAE	3,131	.	3,131	3,131	3,131	3,131	3,131	3,131	3,131	3,131	3,131
MaxAE	8,725	.	8,725	8,725	8,725	8,725	8,725	8,725	8,725	8,725	8,725
BIC normalisé	3,981	.	3,981	3,981	3,981	3,981	3,981	3,981	3,981	3,981	3,981

Statistiques du modèle

Modèle	Nombre de prédicteurs	Statistiques de la qualité de l'ajustement	Ljung-Box Q(18)			Nombre de valeurs extrêmes
		R-deux stationnaire	Statistiques	DL	Sig.	
Vente-Modèle_1	1	,409	.	0	.	0

Interprétation :

— R^2 (R-deux) stationnaire : 0.409 Cela signifie que 40.9% de la variabilité des ventes est expliquée par le modèle ARMA. Bien que ce ne soit pas très élevé, cela indique que le modèle capture une partie significative de la dynamique des données.

— R^2 standard : 0.409 (identique au R^2 stationnaire), ce qui confirme la cohérence du modèle.

• Erreurs de Prédiction

— RMSE (Root Mean Square Error) : 4.837

L'erreur quadratique moyenne est relativement élevée, ce qui suggère que les prédictions peuvent s'écarter significativement des valeurs réelles.

— MAPE (Mean Absolute Percentage Error) : 17.589%

En moyenne, l'erreur de prédiction est de 17.6%, ce qui est modéré mais pourrait être amélioré.

— MaxAPE : 87.249% L'erreur maximale atteint 87.2%, indiquant des points où le modèle performe très mal (possiblement des valeurs aberrantes ou des variations soudaines).

— MAE (Mean Absolute Error) : 3.131

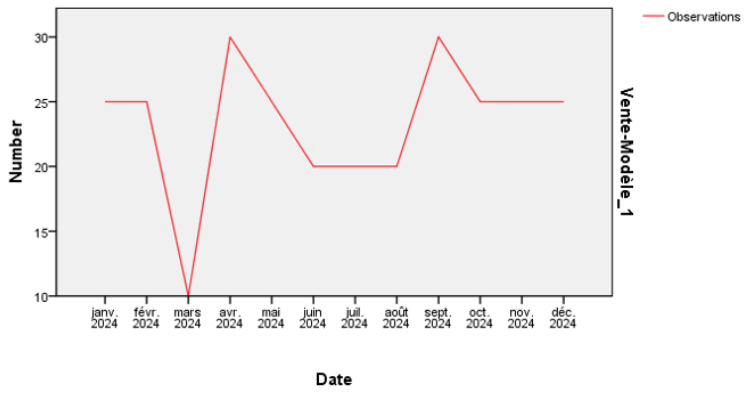
L'erreur absolue moyenne est de 3.131 unités, ce qui donne une idée de la magnitude des erreurs.

— BIC normalisé : 3.981 Un BIC plus faible serait préférable, mais cette valeur aide à comparer ce modèle avec d'autres modèles alternatifs.

— Test de Ljung-Box (Q(18))

• Résultat : Non significatif (Sig. = .0)

Cela suggère que les résidus du modèle ne présentent pas d'autocorrélation significative, ce qui est une bonne indication que le modèle ARMA a bien capturé la structure des données.



Interprétation du graphique « Vente » :

Variabilité importante :

On observe une forte variation des ventes au cours de l'année.

Les ventes commencent autour de 25 unités en janvier et février.

Il y a une chute brutale en mars, atteignant un minimum proche de 10 unités.

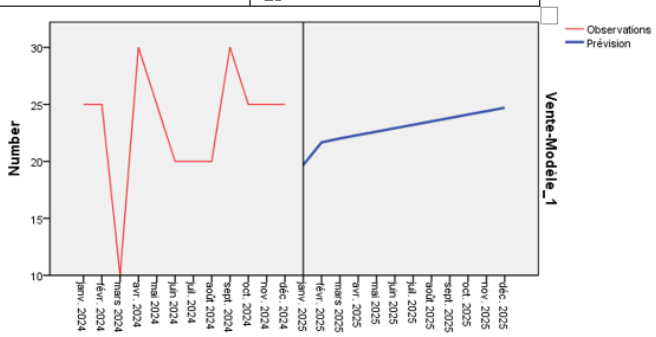
Dès avril, les ventes remontent fortement pour atteindre un pic à plus de 30 unités.

Après ce pic, les ventes redescendent à 20 unités et restent stables de mai à août.

En septembre, un nouveau pic est observé (plus de 30 unités), suivi d'une stabilisation autour de 25 unités jusqu'à la fin de l'année.

Prévision de la ventes :

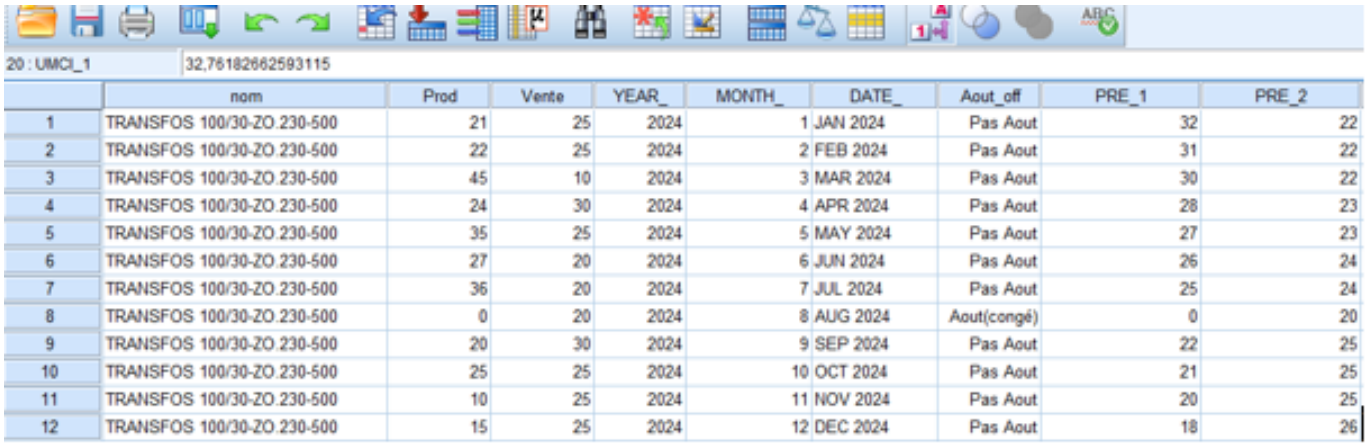
Périodes	valeurs
JAN 2025	20
FEB 2025	22
MAR 2025	22
APR 2025	22
MAY 2025	23
JUN 2025	23
JUL 2025	23
AUG 2025	24
SEP 2025	24
OCT 2025	24
NOV 2025	24
DEC 2025	25



3.4.4 Prévision par la régression linéaire

Production

Pour obtenir les prévisions pour chaque observation en tenant compte que le mois d'août est un congé avec une production nulle, on crée une nouvelle variable appelée AOUT-OFF. Dans la vue des variables, on attribue à cette variable des valeurs codées ainsi : 0 pour « Pas août » et 1 pour « Août (congé) », en utilisant la fonction d'étiquetage des valeurs dans SPSS. Ensuite, on remplit cette colonne dans la vue des données en mettant 1 pour août et 0 pour les autres mois. Lors de la procédure de régression linéaire, on place Prod en variable dépendante, et à la fois MONTH et AOUT-OFF comme variables indépendantes. Cette méthode permet au modèle de prévoir une production proche de zéro pour le mois d'août, en prenant en compte ce congé dans les prévisions. On trouve donc les résultats suivants : PRE-1 est la prévision de chaque observation de production (Prod), tandis que PRE-2 est la prévision correspondante pour les ventes (Vente).



20 : UMCI_1 | 32,76182662593115

	nom	Prod	Vente	YEAR_	MONTH_	DATE_	Aout_off	PRE_1	PRE_2
1	TRANSFOS 100/30-ZO.230-500	21	25	2024	1	JAN 2024	Pas Aout	32	22
2	TRANSFOS 100/30-ZO.230-500	22	25	2024	2	FEB 2024	Pas Aout	31	22
3	TRANSFOS 100/30-ZO.230-500	45	10	2024	3	MAR 2024	Pas Aout	30	22
4	TRANSFOS 100/30-ZO.230-500	24	30	2024	4	APR 2024	Pas Aout	28	23
5	TRANSFOS 100/30-ZO.230-500	35	25	2024	5	MAY 2024	Pas Aout	27	23
6	TRANSFOS 100/30-ZO.230-500	27	20	2024	6	JUN 2024	Pas Aout	26	24
7	TRANSFOS 100/30-ZO.230-500	36	20	2024	7	JUL 2024	Pas Aout	25	24
8	TRANSFOS 100/30-ZO.230-500	0	20	2024	8	AUG 2024	Aout(congé)	0	20
9	TRANSFOS 100/30-ZO.230-500	20	30	2024	9	SEP 2024	Pas Aout	22	25
10	TRANSFOS 100/30-ZO.230-500	25	25	2024	10	OCT 2024	Pas Aout	21	25
11	TRANSFOS 100/30-ZO.230-500	10	25	2024	11	NOV 2024	Pas Aout	20	25
12	TRANSFOS 100/30-ZO.230-500	15	25	2024	12	DEC 2024	Pas Aout	18	26

Normalité des résidus :

Productions

Descriptives

Statistiques descriptives

	N	Skewness		Kurtosis	
	Statistiques	Statistiques	Erreur std.	Statistiques	Erreur std.
residu	12	,405	,637	-,676	1,232
N valide (liste)	12				

Tests de normalité

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistiques	ddl	Sig.	Statistiques	ddl	Sig.
residu	,119	12	,200*	,955	12	,712

*. Il s'agit de la borne inférieure de la vraie signification.

a. Correction de signification de Lilliefors

Skewness (Asymétrie) : 0,405 Cela indique une légère asymétrie à droite de la distribution des résidus. Une valeur proche de 0 indique une distribution symétrique. Ici, la valeur est faible, donc la distribution est presque symétrique.

Erreur standard (Skewness) : 0,637 C'est l'incertitude associée à la mesure de l'asymétrie.

Kurtosis (Aplatissement) : -0,676 Cela indique que la distribution des résidus est légèrement plus aplatie que la distribution normale (kurtosis = 0 pour une normale). Une valeur négative indique une distribution plus plate.

Erreur standard (Kurtosis) : 1,232 C'est l'incertitude associée à la mesure du kurtosis.

Kolmogorov-Smirnov (K-S)

Statistique : 0,119

ddl : 12 (degrés de liberté = nombre de résidus)

Sig. (p-value) : 0,200 La p-value (> 0,05) indique que l'on ne rejette pas l'hypothèse de normalité : les résidus suivent une distribution normale.

Shapiro-Wilk

Statistique : 0,955

ddl : 12

Sig. (p-value) : 0,712 La p-value est largement supérieure à 0,05, donc on ne rejette pas non plus l'hypothèse de normalité pour ce test.

Ventes

Descriptives

Statistiques descriptives

	N	Skewness		Kurtosis	
	Statistiques	Statistiques	Erreur std.	Statistiques	Erreur std.
residu_vente	12	-1,111	,637	2,116	1,232
N valide (liste)	12				

Tests de normalité

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistiques	ddl	Sig.	Statistiques	ddl	Sig.
residu_vente	,177	12	,200*	,921	12	,297

Skewness (Asymétrie) : -1,111

Cette valeur négative indique que la distribution des résidus est asymétrique à gauche, c'est-à-dire qu'elle présente une queue plus longue du côté des valeurs faibles. Plus la valeur absolue s'éloigne de zéro, plus l'asymétrie est marquée. Ici, l'asymétrie est modérée.

Erreur standard de skewness : 0,637

Cette valeur permet d'évaluer la significativité de l'asymétrie observée.

Kurtosis (Aplatissement) : 2,116

Cette valeur positive indique que la distribution des résidus est plus

«pointue » que la distribution normale (le kurtosis d'une loi normale est 0). Cela signifie que les résidus présentent plus de valeurs extrêmes (outliers) que ce qui serait attendu sous une distribution normale.

Erreur standard de kurtosis : 1,232

Elle permet d'évaluer la significativité de l'aplatissement observé.

Deux tests de normalité ont été réalisés pour vérifier si la distribution des résidus s'écarte significativement de la normale :

Test de Kolmogorov-Smirnov :

Statistique : 0,177

Degrés de liberté (ddl) : 12

Signification (Sig.) : 0,200 La p-value (0,200) est largement supérieure au seuil de 0,05. On ne rejette donc pas l'hypothèse de normalité : la distribution des résidus ne diffère pas significativement d'une loi normale.

Test de Shapiro-Wilk :

Statistique : 0,921

Degrés de liberté (ddl) : 12

Signification (Sig.) : 0,297 Là encore, la p-value (0,297) est supérieure à 0,05. Le test confirme que la distribution des résidus ne s'écarte pas significativement de la normale.

Pour des Previsions des mois prochains de l'année 2025 (ce qui est le but) on remplie jusqu'à la ligne 24 les differents informations sauf la case production . Cette derniere on la laisse vide du janvier 2025 jusqu'au decembre 2025 puis on relance la prevision et on trouve :

13	TRANSFOS 100/30-ZO.230-500	.	.	2025	1	JAN 2025	Pas Aout	32
14	TRANSFOS 100/30-ZO.230-500	.	.	2025	2	FEB 2025	Pas Aout	31
15	TRANSFOS 100/30-ZO.230-500	.	.	2025	3	MAR 2025	Pas Aout	30
16	TRANSFOS 100/30-ZO.230-500	.	.	2025	4	APR 2025	Pas Aout	28
17	TRANSFOS 100/30-ZO.230-500	.	.	2025	5	MAY 2025	Pas Aout	27
18	TRANSFOS 100/30-ZO.230-500	.	.	2025	6	JUN 2025	Pas Aout	26
19	TRANSFOS 100/30-ZO.230-500	.	.	2025	7	JUL 2025	Pas Aout	25
20	TRANSFOS 100/30-ZO.230-500	.	.	2025	8	AUG 2025	Aout(congé)	0
21	TRANSFOS 100/30-ZO.230-500	.	.	2025	9	SEP 2025	Pas Aout	22
22	TRANSFOS 100/30-ZO.230-500	.	.	2025	10	OCT 2025	Pas Aout	21
23	TRANSFOS 100/30-ZO.230-500	.	.	2025	11	NOV 2025	Pas Aout	20
24	TRANSFOS 100/30-ZO.230-500	.	.	2025	12	DEC 2025	Pas Aout	18

Interprétation des résultats :

MAPE (Mean Absolute Percentage Error)

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \times 100 \approx 23\%$$

RMSE (Root Mean Squared Error)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2} \approx 7{,}1$$

Les prévisions montrent une demande mensuelle estimée décroissante, allant de 32 à 18 unités, ce qui reflète une tendance linéaire descendante selon le modèle. L'erreur moyenne d'environ 23 % indique que la précision du modèle est modérée, avec une erreur absolue significative dans certains mois. Le RMSE de 7,1 confirme cette dispersion importante entre les valeurs prévues et les valeurs réelles.

La méthode de régression linéaire permet de dégager une tendance générale, mais elle surestime fortement la production en mars (valeur réelle = 0, prédite = 25), ce qui diminue sa fiabilité dans le cas de variations brutales. Le modèle reste néanmoins un outil valable pour des prévisions de tendance sur des périodes relativement stables, mais il nécessite une surveillance rapprochée en cas de perturbations inhabituelles de la production.

Vente :

On procède de la meme facon et on retrouve les résultat suivant :

13	TRANSFOS 100/30-ZO.230-500	.	.	2025	1	JAN 2025	Pas Aout	22
14	TRANSFOS 100/30-ZO.230-500	.	.	2025	2	FEB 2025	Pas Aout	22
15	TRANSFOS 100/30-ZO.230-500	.	.	2025	3	MAR 2025	Pas Aout	22
16	TRANSFOS 100/30-ZO.230-500	.	.	2025	4	APR 2025	Pas Aout	23
17	TRANSFOS 100/30-ZO.230-500	.	.	2025	5	MAY 2025	Pas Aout	23
18	TRANSFOS 100/30-ZO.230-500	.	.	2025	6	JUN 2025	Pas Aout	24
19	TRANSFOS 100/30-ZO.230-500	.	.	2025	7	JUL 2025	Pas Aout	24
20	TRANSFOS 100/30-ZO.230-500	.	.	2025	8	AUG 2025	Aout(congé)	20
21	TRANSFOS 100/30-ZO.230-500	.	.	2025	9	SEP 2025	Pas Aout	25
22	TRANSFOS 100/30-ZO.230-500	.	.	2025	10	OCT 2025	Pas Aout	25
23	TRANSFOS 100/30-ZO.230-500	.	.	2025	11	NOV 2025	Pas Aout	25
24	TRANSFOS 100/30-ZO.230-500	.	.	2025	12	DEC 2025	Pas Aout	26

Interprétation des résultats :

MAPE (Mean Absolute Percentage Error)

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \times 100 \approx 11\%$$

RMSE (Root Mean Squared Error)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2} \approx 2{,}7$$

Les prévisions indiquent une estimation relativement constante, allant de 22 à 26 unités, traduisant une tendance légèrement croissante. L'erreur moyenne d'environ 11 % montre que le modèle est précis et fiable, notamment pour un usage opérationnel courant. Le RMSE de 2,7 indique une faible dispersion entre les valeurs prévues et observées, ce qui confirme la qualité de l'ajustement.

Cependant, on note une légère sous-estimation lors des pics (par exemple en mars, avec une valeur réelle de 30 contre une prédiction de 23). Malgré cela, le modèle capte correctement la dynamique générale des ventes, sans être affecté par des fluctuations majeures.

3.5 Conclusion

L'étude comparative des différentes méthodes de prévision appliquées aux séries de production et de ventes a permis d'évaluer la performance de chaque modèle à l'aide d'indicateurs d'erreur tels que le **MAPE** et le **RMSE**.

Pour la variable des **ventes**, les résultats montrent que les méthodes les plus fiables sont les *moyennes mobiles* et la *courbe de croissance*, avec un MAPE respectif de **11 %** et **18 %**, et un RMSE de **2,7** et **3,6**, ce qui traduit une bonne précision dans un contexte de tendance relativement stable. À l'inverse, la *régression linéaire* appliquée aux ventes reste pertinente pour dégager une tendance globale (MAPE de **11 %**, RMSE de **2,7**), mais elle tend à sous-estimer les pics. Le modèle *ARMA* obtient également des résultats modérés (MAPE $\approx$ **17,6 %**, RMSE $\approx$ **4,8**), avec un bon ajustement global ($R^2 \approx$ **41 %**), bien qu'il présente une erreur maximale élevée sur certains mois.

Pour la variable **production**, les prévisions sont globalement moins précises en raison de variations plus marquées et d'effets saisonniers (notamment l'arrêt d'août). La *régression linéaire* obtient un MAPE de **23 %** et un RMSE de **7,1**, ce qui reste acceptable pour des prévisions de tendance, mais insuffisant pour capter des changements soudains. La *courbe de croissance* offre une performance légèrement meilleure (MAPE $\approx$ **21 %**, RMSE $\approx$ **5,2**) et un bon ajustement ($R^2 \approx$ **74,4 %**), ce qui en fait un modèle plus adapté pour une planification stratégique de moyen terme. En revanche, le modèle *ARMA* pour la production affiche des résultats plus faibles (MAPE $\approx$ **31 %**, RMSE $\approx$ **10**), avec une variabilité importante non captée ($R^2 \approx$ **29,5 %**), ce qui limite sa fiabilité opérationnelle.

Conclusion générale et perspectives

Au terme de cette étude, il apparaît clairement que le choix de la méthode de prévision ne peut se faire indépendamment de la nature des données à modéliser. En analysant les séries chronologiques de ventes et de production de l'entreprise ELECTRO-INDUSTRIE, nous avons observé deux comportements statistiques distincts. La variable « ventes » présente une distribution relativement stable et centrée autour de 23 à 25 unités, avec un écart-type modéré de 5. Cette régularité traduit une faible variabilité mensuelle, ce qui rend les ventes particulièrement adaptées à des méthodes simples mais efficaces. Les résultats obtenus confirment cette intuition : les **moyennes mobiles** ont affiché une précision remarquable, avec un MAPE de 11 % et un RMSE de 2,7, ce qui en fait la méthode la plus fiable pour anticiper les ventes à court terme. La **régression linéaire**, bien que légèrement moins flexible face aux pics de vente, a également montré une performance acceptable, traduisant une bonne capacité à identifier une tendance globale.

À l'inverse, les données de production sont marquées par une grande dispersion, avec une moyenne similaire à celle des ventes (23), mais un écart-type beaucoup plus élevé (12,03), un mode nul (révélant des mois sans production), et une plage de variation allant de 0 à 45 unités. Cette forte variabilité complique la tâche des méthodes simples, qui peinent à s'adapter aux ruptures de tendance et aux pics soudains. Dans ce contexte, des méthodes plus robustes sont nécessaires. Parmi celles testées, la **courbe de croissance** a fourni les meilleurs résultats globaux pour la production, avec un MAPE de 21 %, un RMSE de 5,2, et un coefficient de détermination R^2 de 74,4 %, indiquant une très bonne capacité d'ajustement aux données historiques. En revanche, les **modèles ARMA**, pourtant théoriquement puissants, se sont montrés plus sensibles à l'irrégularité des données, avec des erreurs plus importantes (MAPE de 31 %, RMSE de 10) et une capacité explicative limitée ($R^2 \approx 29{,}5\,\%$).

Ainsi, la réponse à notre problématique est double : les **moyennes mobiles** s'imposent comme la méthode de référence pour les prévisions de ventes, grâce à leur simplicité, leur précision et leur capacité d'adaptation à une série stable, tandis que la **courbe de croissance** est la plus adaptée pour la prévision de la production, notamment dans une optique de planification stratégique à moyen terme. Toutefois, aucune méthode ne peut prétendre être universelle. Il est donc conseillé à ELECTRO-INDUSTRIE

d'adopter une approche **flexible et combinée**, utilisant plusieurs modèles en parallèle pour valider ou affiner les prévisions, en particulier lors des périodes de forte instabilité.

Enfin, cette démarche d'analyse et de sélection méthodique des outils de prévision offre à l'entreprise ELECTRO-INDUSTRIE des bénéfices concrets et durables. En anticipant plus précisément les volumes de ventes et de production, l'entreprise pourra **optimiser la gestion des stocks**, **mieux ajuster sa capacité de production**, **réduire les coûts liés aux surstocks ou aux ruptures**, et **améliorer la réactivité de sa chaîne d'approvisionnement**. À plus long terme, une politique de prévision fiable permet aussi de soutenir des décisions stratégiques, telles que les investissements, les recrutements ou les négociations commerciales. En intégrant ces outils statistiques dans ses processus internes, ELECTRO-INDUSTRIE renforce ainsi sa capacité à piloter son activité de façon plus rigoureuse, agile et compétitive, dans un environnement de plus en plus incertain.

Dans la continuité de ce travail, il serait pertinent de collecter des données comportant un plus grand nombre de variables explicatives. Cela permettrait non seulement d'enrichir l'analyse, mais aussi d'explorer davantage l'utilisation de méthodes telles que la régression linéaire multiple, les modèles ARIMA et les fonctions de transfert. Ces démarches offriraient une meilleure compréhension des relations entre les variables, contribueraient à améliorer la qualité des prévisions réalisées et faciliteraient une prise de décision plus éclairée. Par ailleurs, des analyses futures pourraient intégrer l'étude des effets externes afin de mieux appréhender l'influence de facteurs extérieurs sur les résultats, ce qui renforcerait la pertinence et la portée des conclusions.

Bibliographie

[1] Pearson Education France. (2010, 26 novembre). *Analyse de données avec SPSS*. IBM.

[2] IBM. (s.d.). *IBM SPSS Forecasting Documentation* [Documentation PDF]. IBM.

[3] AFRISTAT. (s.d.). *Manuel d'initiation au traitement de données sous SPSS* [Manuel PDF]. AFRISTAT.

[4] GRAIE – Groupe de Travail Autosurveillance, Sous-groupe Modélisation. (mars 2018). *Critères & indicateurs d'auto-évaluation des modèles dans le cadre de l'autosurveillance réglementaire des systèmes d'assainissement* (Version 1, Document de travail). Disponible en ligne : `http://www.graie.org/graie/graiedoc/reseaux/autosurv/GRAIE-Criteres-INDICATEURS-AUTOEVALUTIONdesMODELES-AUTOSURVEILLANCE-WEB18-v1.pdf`

[5] Université de Montpellier. (2020). *Qualité des prévisions et critères d'évaluation des modèles statistiques* [Cours de statistiques appliquées]. Disponible en ligne : `https://www.mcours.net/cours/pdf/yass3/yass3cli1949.pdf`

[6] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting : Principles and Practice* (3rd ed.). OTexts. Disponible en ligne : `https://otexts.com/fpp3/accuracy.html`

[7] Pupion, P.-C. (2018). *Statistiques pour la gestion : Applications avec Excel et SPSS* (2e éd.). Dunod.

[8] Lecoutre, J.-P., Legait, S., & Tassi, P. (1998). *Statistiques : Exercices corrigés avec rappels de cours.*

[9] Belharet, N., & Haouame, A. (2013). *Les tests de normalité* [Mémoire de master de Probabilités et Statistiques, Université Mouloud Mammeri de Tizi-Ouzou (UMMTO)]. Promotion 2012/2013.

[10] Rahmani, N. (2011). *Méthodes stochastiques de calcul de stabilité des pentes* [Mémoire de Master en Génie Civil (Option : Géotechnique et environnement), Université Mouloud Mammeri de Tizi-Ouzou (UMMTO)].

[11] Hyndman, R. J., Athanasopoulos, G. (2018). Forecasting : Principles and Practice (2nd ed.). OTexts

[12] Bouzaïane, L., Mouelhi, R. (2006-2007). Méthodes de prévision (Projet de M2PA). Université Virtuelle de Tunis

[13] Electro-Industrie(2000). [Document interne]. Electro-Industrie, Azazga, Algérie