République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université de Tissemsilet
Faculté des Sciences et Technologie

Cours : analyse statistique des données

Support pédagogique

Mouaissia Wahiba « MCB »

Destiné aux étudiants de Master 1

PROTECTION DES ECOSYSTEMES

2023-2024

Table Des Matières

Table Des Matières

Table Des Matières

Introduction

Tout résultat de recherche biologique résulte d'une expérimentation qui s'appuie sur une méthodologie statistique rigoureuse, et dont les résultats sont analysés en termes statistiques. Leur objectif consiste à caractériser une population à partir d'une image obtenue par cette observation constituée à l'aide d'un échantillon issu de cette population. On peut alors chercher à extrapoler l'information obtenue à partir de cette échantillon.

Ce polycopié présente le fondement des méthodes statistiques et donne les outils de base et leur mise en application en Biostatistique. Il permet aux étudiants d'identifier les outils statistiques nécessaires à la résolution des problèmes occurrents dans le domaine de la biologie et la science de la nature et la vie.

Il reprend les éléments de bases des statistiques descriptives et les méthodes statistiques d'estimation (inférence statistique) ainsi que les principaux facteurs de choix des tests d'hypothèse, de manière à satisfaire les conditions d'application des méthodes de l'inférence paramétriques et son alternative non paramétriques.

Il fournit également, des outils statistiques qui permettent de généraliser, dans certaines conditions, les conclusions obtenues par la statistique descriptive à partir de la fraction des individus (échantillon) que l'on a observé ou étudié expérimentalement, à l'ensemble des individus constituant la population.

La compétence visée par ce cours, dans son ensemble, est « ***d'être capable de comprendre l'intérêt de l'analyse statistique, et mieux saisir l'importance de l'information numérique présentée sous diverses formes, et comment les rendre utile pour l'explication des phénomènes biologiques diverses et d'interpréter correctement les résultats de nouvelles recherches, et d'adopter un mode de raisonnement qui soit à même d'aider à la décision dans l'expérience biologique via la signification des tests d'hypothèses utilisés*** ».

Il apporte une formation la plus exhaustive possible en statistique descriptive et inférentielle directement exploitable dans de nombreux Masters académique. Il donne des compétences indispensables pour la collecte et le traitement de données expérimentales.

1. Prérequis et objectifs

L'objectif de ce chapitre est d'étudier comment à partir de tableaux et de graphiques il est possible de résumer les principales caractéristiques de leur distribution, comparer la composition, la position, la variabilité, etc. de plusieurs groupes, mettre en évidence des relations entre variables et détecter les lois qui régentent le phénomène étudié.

La statistique descriptive a pour but de présenter les données sur une forme telle qu'on puisse prendre des connaissances et les exploiter facilement. Elle a donc pour but de d´écrire et non d'expliquer.

2. Définitions et vocabulaire statistique

2.1 Population

Une population désigne l'ensemble des éléments, individus ou unités auquel on s'intéresse à l'un ou plusieurs de leurs caractères quantitatifs ou qualitatifs.

Population = Ensemble d'unités statistiques de même nature sur lequel on recherche des informations quantifiables.

Exemples

- Ensemble des arbres d'une forêt (unité statistique = arbre)
- Ensemble des plantes d'une parcelle……etc.

Dans la plupart des cas, il est difficile d'obtenir l'information à partir de la population dans son ensemble. On utilise alors un échantillon pour tirer des conclusions sur la population.

2.2 Echantillons

Un échantillon est une fraction de la population sur laquelle porte l'observation d'un ou des caractères étudiés.

Exemples

- Si nous intéressons à la pollution de l'eau d'un barrage en période hivernale, nous constituerons nos échantillons d'eau au hasard durant les trois mois d'hiver et dans plusieurs sites pour obtenir un échantillon suffisamment représentatifs de la qualité de cette eau.

2.3 Échantillonnage

L'ensemble des opérations destinées à former l'échantillon, il consiste à :

- Sélection d'une partie de la population
- Étude de certaines caractéristiques de l'échantillon
- Dégager les inférences relatives à la population

2.4 Individu (ou unité statistique)

Les individus sont les éléments de la population statistique étudiée.

Chaque individu peut avoir un ou plusieurs caractères. L'ensemble des données numériques relatives à ces caractères constitue une série statistique ou distribution statistique.

2.5 Taille

Elle représente le nombre d'individus d'un échantillon ou d'une population.

Elle est symbolisée par :

- « $\boldsymbol{n}$ » dans le cas d'un échantillon;
- « $\boldsymbol{N}$ » dans le cas d'une population.

2.6 Modalités

Ce sont les différentes manières que peut présenter un caractère.

- Etat de santé : est un variable qui présente deux modalités *malade et sain*
- Sexe : est un variable qui comprends deux modalités * féminin ou masculin*
- Couleur des yeux : est un variable qui présente quatre modalités *noire, bleu, vert et noisette*

2.7 Caractère (ou variable statistique)

C'est l'aspect particulier que l'on désire étudier, c'est le variable observés ou mesuré sur les individus d'une population statistique.

Exemple

Concernant un groupe de lapins, on peut s'intéresser à leur âge, leur sexe, leur taille…etc.

2.8 Variable quantitative

Une variable statistique est quantitative si **ses valeurs sont des nombres exprimant une quantité**, sur lesquels les opérations arithmétiques (somme, etc...) ont un sens.

➢ Variable quantitative continue

C'est une variable ne prenant que **des valeurs infini** qu'elle peut prendre n'importe quelles valeurs dans un intervalle donnée (poids, revenu, longueur, âge, dosage biologique, taille, température… etc.)

➢ Variable quantitative discrète

C'est une variable qui ne prenant que **des valeurs fini** (plus rarement décimales), ces valeurs sont distincts et séparés (nombre d'enfants dans une famille, classe d'âges...)

2.9 Variable qualitative

Une variable statistique est qualitative si ses valeurs, ou modalités (grandeurs non quantifiables comme couleurs, sexe, mention etc.), s'expriment de façon littérale ou par un codage sur lequel les opérations arithmétiques telles que moyenne, somme, ..., n'ont pas de sens.

➢ Variable qualitative nominale

C'est une variable qualitative dont les modalités ne sont pas ordonnées (Sexe, groupe sanguin, Couleur des yeux, Espèces, …etc).

➢ Variable qualitative ordinale

C'est une variable qualitative dont les modalités sont naturellement ordonnées (Mention au bac, Niveau d'études, Seuil de gravité d'une maladie,…etc).

3. Définition de la statistique descriptive

La statistique descriptive est un ensemble de méthodes permettant de d´écrire et d'analyser des phénomènes susceptibles d'être dénombrés et classés.

- STATISTIQUE DESCRIPTIVE A UNE DIMENSION : Elle concerne **une seule variable** ou **une seule caractéristique** d'une variable à la fois ;
- STATISTIQUE DESCRIPTIVE A DEUX (OU PLUSIEURS) DIMENSIONS : Elle peut aussi s'attacher à **deux (ou plusieurs) variables**,

4. Statistique descriptive à une dimension ou statistique univariée

Le calcul de certains paramètres permet de caractériser de façon simple les séries statistiques observées.

- Paramètres de position
- Paramètres de dispersion
- Paramètres de dissymétrie et d'aplatissement

4.1. Paramètres de position

Ils sont des indicateurs statistiques de tendance centrale (dits aussi de position) permettent de savoir autour de quelles valeurs se situent les valeurs d'une variable statistique. Ces paramètres servent à caractériser (le milieu) de la distribution. Ce sont simplement :

- La moyenne arithmétique (moyenne),
- La médiane.
- Le mode.
- Les deux quartile ($\boldsymbol{Q1}$ et $\boldsymbol{Q3}$).

a) Moyenne arithmétique $\bar{x}$

La moyenne arithmétique est la somme des valeurs observées divisée par le nombre d'observations.

Calcule

Série brute $\bar{x}$	Series groupées	Série groupée en classes
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ x une variable statistique discrète $x1, x2, \ldots, xn$ ses valeurs, Avec $n = \sum n_i = 1n_i$ n est l'effectif total.	$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i \cdot x_i$ $= n_1.x_1 + n_2.x_2 + \cdots n_k.x_k$ $= \frac{n_1}{n}.x_1 + \frac{n_2}{n}.x_2 + \cdots \frac{n_k}{n}.x_k$ $= f_1.x_1 + f_2.x_2 + \cdots f_k.x_k$ $= \sum_{i=1}^{k} f_i \cdot x_i$ x une variable statistique discrète $x1, x2, \ldots, xk$ ses valeurs, Les effectifs $n1, n2, \ldots\ldots, nn$; $n = \sum k_i = 1n_i$ n l'effectif total.	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i \cdot c_i$ Cas d'une variable statistique continue, Les observations sont groupées dans des classes ; La même formule que le cas discret, sauf remplace les xi par les centres des classes ci :

Tableau 2 Calcule des fréquences pour une série groupée

Valeurs de la variable	Effectifs	Fréquences
x_1	n_1	$f_1 = n_1/n$
…	…	…
x_i	n_i	$f_i = n_i/n$
…	…	…
x_k	n_k	$f_k = n_k/n$

Tableau 3 Calcule des fréquences pour une série groupée en classes

Classes	Effectifs	Fréquences	Centres de classes
$[e_1 - e_2[$	n_1	$f_1 = n_1/n$	$c_1 = (e_1 + e_2)/2$
$[e_2 - e_3[$	n_1		$c_2 = (e_2 + e_3)/2$
…	…	…	…
$[ek - ek+1[$	n_k	$f_i = n_i/n$	$c_k = (e_k + e_{k+1})/2$

b) Médiane

La médiane correspond à la valeur du milieu de la distribution. On l'appelle également parfois « percentile 50 » : c'est la valeur centrale par excellence.

- Si $\boldsymbol{n}$ est impair : $\boldsymbol{M = x_{(n+1)/2}}$
- Si $\boldsymbol{n}$ est pair : $\boldsymbol{M = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}}$
- Cas de la série groupés en classes (variable quantitatif continue) :

$$\hat{x} = L_i + \left(\frac{\frac{n}{2} - \sum_{i=1}^{<\hat{x}} n_i}{n_{\hat{x}}} \right) . a$$

L_i: borne inferieure de la classe médiane (classe qui divise l'effectif en deux) ;
n : effectif total ;
$\sum_{i=1}^{<\hat{x}} n_i$: Somme des effectifs correspondant à toutes les classes inferieures à la classe médiane ;
$n_{\hat{x}}$: Effectif de la classe médiane ;
a : amplitude de la classe médiane.

Remarque :

Il est Nécessaire de classer les observations par ordre croissant.

La médiane a comme propriété d'être peu sensible aux valeurs extrêmes.

c) Mode

C'est la valeur distincte correspondant à l'effectif le plus élevé (la valeur la plus fréquente dans l'échantillon) ; il peut être calculé pour les types de variable quantitative et qualitative ;

Quand une variable continue est groupée en classes, on peut définir une classe modale (classe correspondant à l'effectif le plus élevé) ; le mode se calcule par la formule :

$$M0 = L_i + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) . a$$

: borne inferieure de la classe modale (classe correspondant au plus grand effectif) ;
Δ_1: Excédent de l'effectif de la classe modale par rapport à l'effectif de la classe précédente ;
Δ_2: Excédent de l'effectif de la classe modale par rapport à l'effectif de la classe suivante ;
: Amplitude de la classe modale.

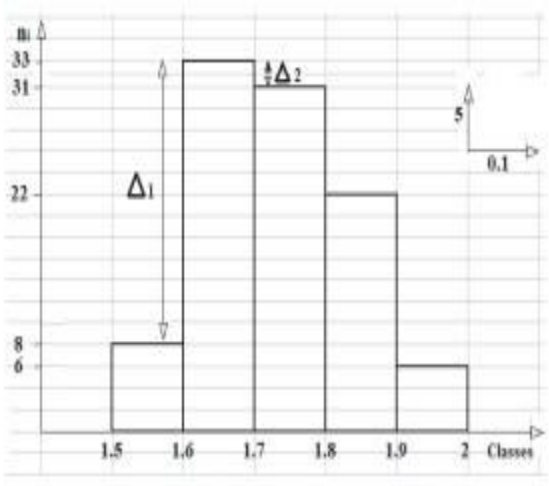


Figure 1 : Histogramme des effectifs avec une illustration de Δ_1 et Δ_2

Remarque2:

Une distribution est unimodale si elle présente un maximum marqué, et pas d'autres maxima.

La lecture s'effectue sur le diagramme en bâtons ou l'histogramme.

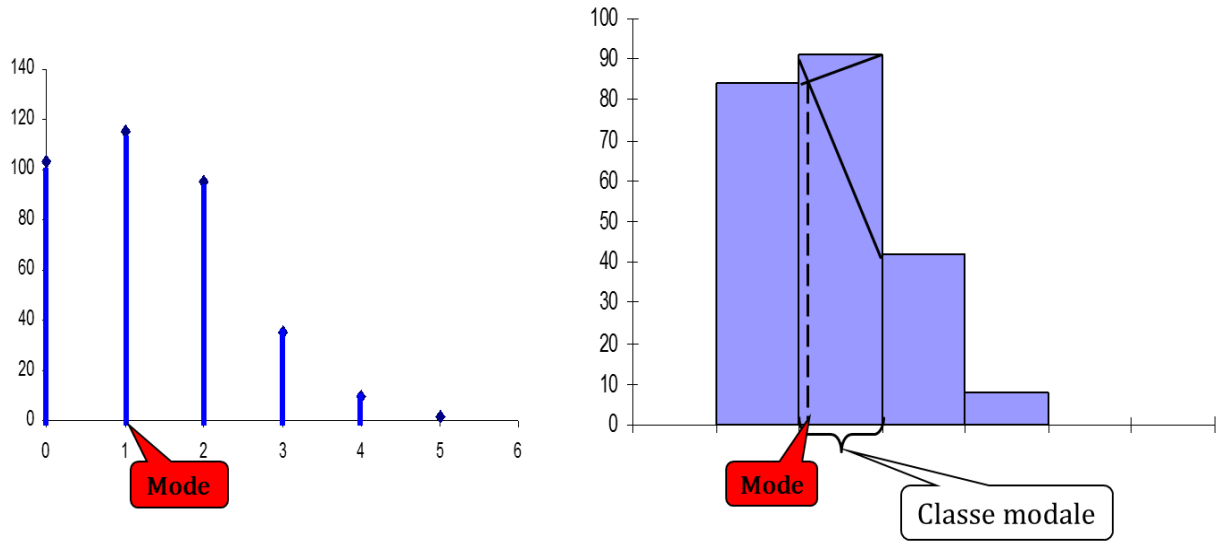


Figure 2 : Diagramme en bâtons et histogramme avec une illustration de mode Le mode correspond à l'abscisse du maximum, c.à.d. la valeur la plus fréquente.

Si la distribution présente 2 ou plus maxima relatifs, on dit qu'elle est bimodale ou plurimodale.

La population est composée de plusieurs sous-populations ayant des caractéristiques de tendance centrale différentes.

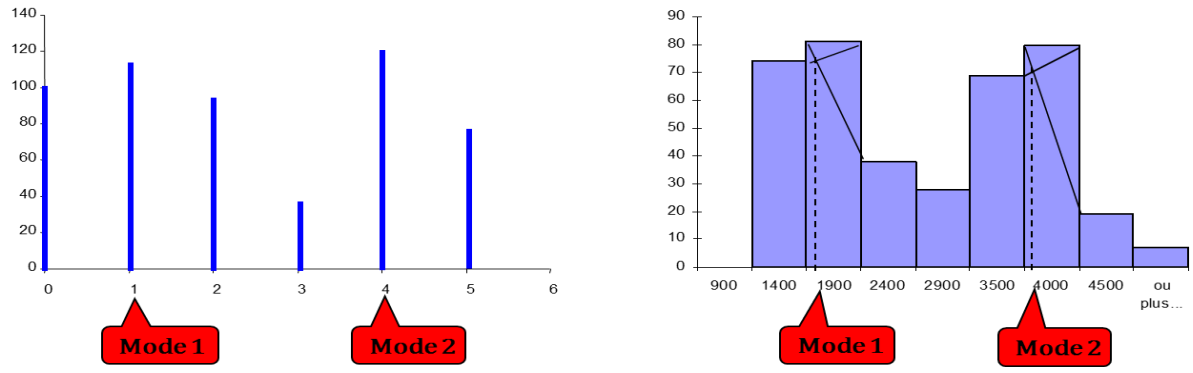


Figure 3 : Diagramme en bâtons et histogramme avec une illustration d'une distribution bimodale

d) Fractiles ou Quantiles

▶ **Série brute**

On appelle fractiles ou quantiles d'ordre $\boldsymbol{k}$ les $(\boldsymbol{k}-\boldsymbol{1})$ valeurs qui divisent les observations en $\boldsymbol{k}$ parties d'effectifs égaux.

1 médiane $\boldsymbol{M}$ qui divise les observations en 2 parties égales ;
3 quartiles $\boldsymbol{Q_1}$, $\boldsymbol{Q_2}$, $\boldsymbol{Q_3}$ qui divisent les observations en 4 parties égales ;
9 déciles $\boldsymbol{D_1}$, $\boldsymbol{D_2}$, … , $\boldsymbol{D_9}$ qui divisent les observations en 10 parties égales ;
99 centiles $\boldsymbol{C_1}$, $\boldsymbol{C_2}$, … , $\boldsymbol{C_{99}}$ qui divisent les observations en 100 parties égales.

▶ **Série groupée**

Les trois quartiles les plus couramment utilisés en biostatistique sont ceux qui divisent l'ensemble des observations en quatre sous-ensembles de même effectif,

$\boldsymbol{Q_1}$: $\boldsymbol{x_i}$ tel que $\boldsymbol{F_i = 0,25 => 1/4}$ des valeurs lui sont inférieures, $\boldsymbol{3/4}$ lui sont supérieures.
$\boldsymbol{Q_2}$ = Médiane; $\boldsymbol{x_i}$ tel que $\boldsymbol{F_i = 0,5 => 1/2}$
$\boldsymbol{Q_3}$: $\boldsymbol{x_i}$ tel que $\boldsymbol{F_i = 0,75 => 3/4}$ des valeurs lui sont inférieures, $\boldsymbol{1/4}$ lui sont supérieures.

▶ **Série groupée en classes**

Dans ce cas (variable quantitative continue), les quartiles (1 et 3) sont donnés par la formule suivante :

Quartile 1

$$\boldsymbol{Q_1 = L_i + \left(\frac{\frac{n}{4} - \sum_{i=1}^{<Q_1} n_i}{n_{Q_1}} \right) . a}$$

$\boldsymbol{L_i}$: borne inferieure de la classe Q_1 ;
$\boldsymbol{n}$: Effectif total ;
$\sum_{i=1}^{<Q_1} \boldsymbol{n_i}$: Somme des effectifs correspondant à toutes les classes inferieures à la classe Q_1;
$\boldsymbol{n_{Q_1}}$: Effectif de la classe Q_1;
$\boldsymbol{a}$: amplitude de la classe Q_1.

Quartile 3

$$Q_3 = L_i + \left(\frac{\frac{3n}{4} - \sum_{i=1}^{<Q_3} n_i}{n_{Q_3}} \right) . a$$

L_i: borne inferieure de la classe Q_3 ;
n: Effectif total ;
n_{Q_3}: Effectif de la classe Q_3;
a: Amplitude de la classe Q_3.
$\sum_{i=1}^{<Q_3} n_i$: Somme des effectifs correspondant à toutes les classes inferieures à la classe Q_3.

4.2. Paramètres de dispersion

La moyenne ne donne qu'une information partielle. En effet, il est aussi important de pouvoir mesurer combien ces données sont dispersées autour de la moyenne, les données des deux variables ont la même moyenne, mais vous sentez bien qu'elles sont de nature différente.

- Les paramètres de dispersion évaluent le niveau d'étalement de la série autour de la valeur centrale.
- Ils Permettent de chiffrer la variabilité des valeurs observées autour d'un des paramètres de position.
- Ils complètent les paramètres de position en permettant de comparer des séries dont les paramètres de position sont proches, mais où la forme de la dispersion est très différente.

Les indicateurs statistiques de dispersion usuels sont :

- La variance,
- L'écart-type,
- L'étendue,
- Le coefficient de variation et
- L'écart-interquartile.

a) Variance

La variance est la somme des carrées des écarts (SCE) à la moyenne divisée par le nombre d'observations (n). C'est une autre manière de procéder qui tient compte de toutes les données, et non pas seulement des valeurs extrêmes.

On considère les données (j) de la $j^{ème}$ variable, l'idée est de calculer la somme, pour chacune des données de cette variable, des distance à la moyenne, et de diviser par le nombre de données.

▶ **Série brute**

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

▶ **Série groupée**

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} n_i {x_i}^2 - \bar{x}^2$$

▶ **Série groupée en classes**

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} n_i {c_i}^2 - \bar{x}^2$$

b) Ecart type

L'écart type est égal à la racine carrée de la variance

$$S = \sqrt{V}$$

c) Coefficient de variation

Il est donné par le rapport de l'écart type à la moyenne multiplié par 100

$$CV = \left(\frac{S}{\bar{x}}\right).100$$

d) Etendue

L'étendue est la différence entre la plus grande et la plus petite valeur observée.

$$E = x_{max} - x_{min}$$

Remarque 3

- L'étendue est un indicateur instable étant donné qu'il ne dépend que des valeurs extrêmes.
- Vous pouvez avoir un grand nombre de données qui sont similaires, mais qui ont une plus grande et plus petite valeur qui sont très différentes, elles auront alors une étendue très différente, mais cela ne représente pas bien la réalité des données.

e) Ecart interquartile (intervalle)

Q_1 et Q_3 contient 50% de la population laissant à droite 25% et à gauche 25%.
Cet intervalle est donné par : Q_1–Q_3.
L'écart interquartile est la différence entre le troisième et le premier quartile.

$$IQ = Q_3 - Q_1$$

4.3. Paramètres de forme

L'habitude de lire des histogrammes ou plus généralement des courbes de distribution de fréquence (représentation graphique de la densité de fréquence) rend sensible à la « forme » de la courbe.

Les statisticiens ont expliqué deux critères pour décrire la forme d'une telle courbe : sa symétrie et son aplatissement.

a) Coefficient d'asymétrie (skewness)

Il mesure l'asymétrie d'une distribution, c'est-à-dire si elle "penche" d'un côté ou de l'autre.

Une série a une distribution symétrique, si ses valeurs sont réparties dans les mêmes proportions autour de la valeur centrale (également dispersées de part et d'autre du mode, la moyenne et la médiane).

Remarque 4 :

Une distribution parfaitement symétrique : Moyenne = Médiane = Mode

Calcule

$$\alpha_3 = \frac{m_3}{S^3}$$

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

S^3 *est le cube de l'écart type de la distribution*

Lecture :

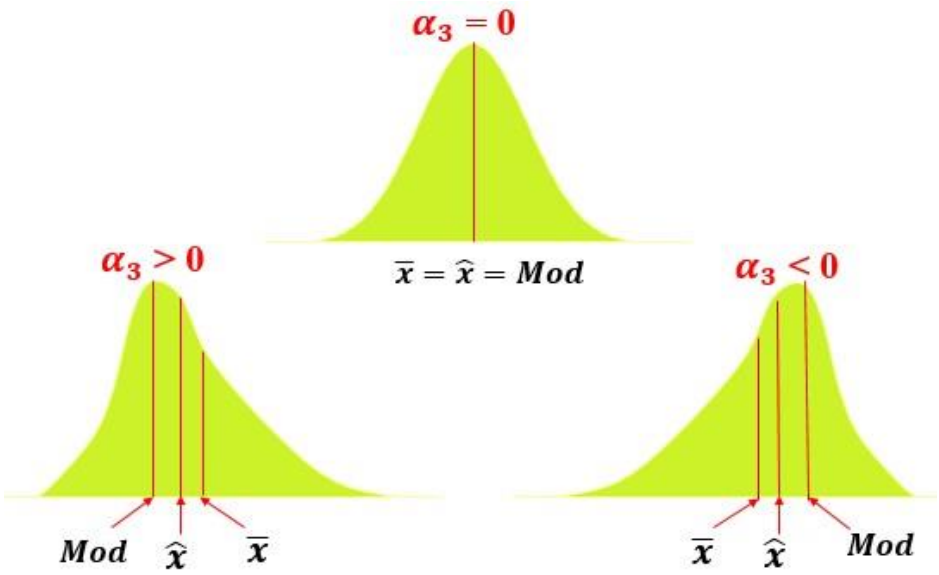


Figure 4 : Représentation schématique de l'asymétrie d'une distribution (skewness)

b) Coefficient d'aplatissement (kurtosis)

Il mesure l'aplatissement d'une distribution. Une distribution est plus ou moins aplatie selon que les fréquences des valeurs voisines des valeurs centrales diffèrent peu ou beaucoup les unes par rapport aux autres.

Calcule :

$$\alpha_4 = \frac{K_4}{S^4}$$

$$K_4 = \frac{n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3(n-1)[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2}{(n-1)(n-2)(n-3)}$$

S^4 *est la quaterieme puissance de l'écart type de la distribution*

Lecture

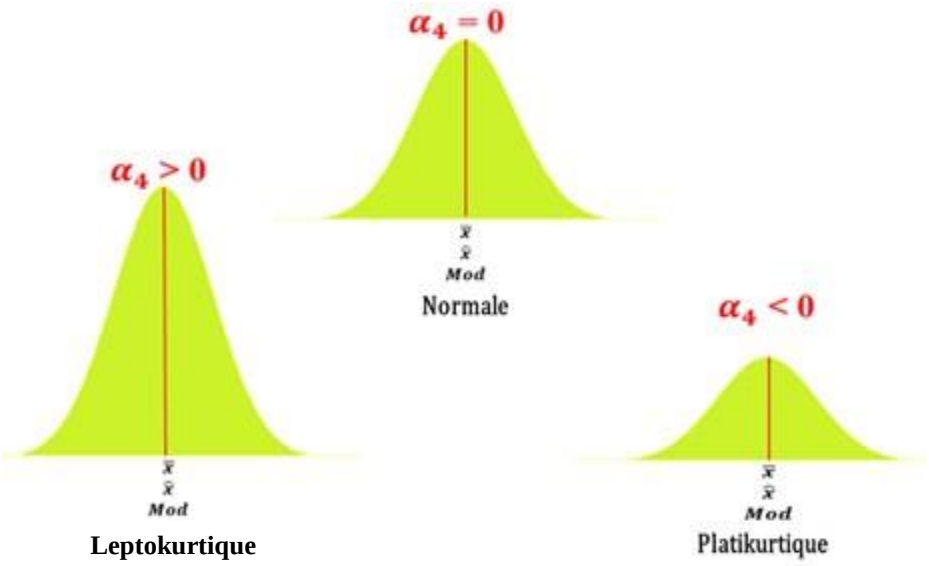


Figure 5 : Représentation schématique de l'aplatissement d'une distribution (kurtosis).

Chapitre 02: Régressions simples et multiples (Statistique descriptive à deux dimensions)

1. Régression linéaire simple

1.1. Objectif

La méthode de la régression a pour but de décrire la relation entre

- Un variable ***dépendant*** (variable *expliquée*) ($\boldsymbol{y}$)
- Un variable ***indépendant*** ou prédictive ($\boldsymbol{x}$).

La corrélation permet d'étudier la liaison rencontrée fréquemment entre deux variables quantitatives.

1.2. Représentation graphique

Le but de la régression simple est de chercher une fonction f telle que $y_i = f(x_i)$. L'étude de cette dernière débute toujours par un tracé des observations sous forme de diagramme de dispersion ou nuage de points de n couples ($\boldsymbol{x_i}$, $\boldsymbol{y_i}$), numérotés de $\boldsymbol{i} = \mathbf{1}$ à $\boldsymbol{i} = \boldsymbol{n}$.

On note $\bar{\boldsymbol{x}}$, $\bar{\boldsymbol{y}}$, $\boldsymbol{s^2_x}$, $\boldsymbol{s^2_y}$, les moyennes et les variances des séries ($\boldsymbol{x_i}$) et ($\boldsymbol{y_i}$).

Cette première représentation permet de savoir si le modèle linéaire est pertinent.

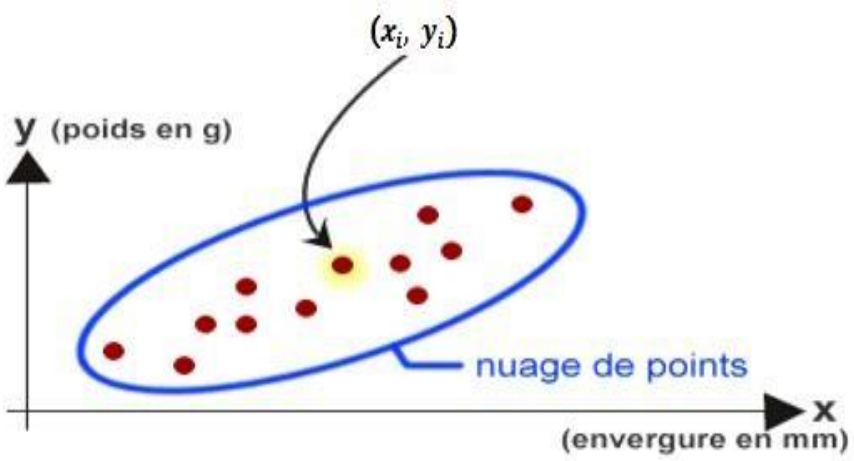


Figure 6 : Diagramme de dispersion des couples d'observations ($\boldsymbol{x_i}$, $\boldsymbol{y_i}$)

1.3. Réduction des données

Les paramètres utilisés pour caractériser les séries statistiques doubles sont deux types :

- Les uns ne concernent qu'une variable à la fois, ils servent à caractériser les distributions marginales ou conditionnelles (la covariance).
- Les autres servent à décrire les relations existant entre les deux séries d'observation (coefficient de corrélation et le coefficient de détermination)

Chapitre 02: Régressions simples et multiples (Statistique descriptive à deux dimensions)

a) Corrélation

La corrélation est une statistique qui caractérise l'existence ou l'absence d'une relation entre **deux échantillons** de valeurs prise sur **un même groupe de sujets**.

Elle permet de quantifier cette relation par le signe de la corrélation (positive et négative), et par la force de cette corrélation.

▶ Coefficient de corrélation « r »

Pour calculer le coefficient de corrélation,

Calculer la covariance entre deux échantillons. Il se calcule comme suit :

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Une façon d'atteindre cette mesure est d'utiliser le produit des différences, comme suit :

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

Notes

La covariance est une mesure de la variance présente dans deux échantillons simultanément.

- Si les deux échantillons co-varient, la covariance devrait être **grande**,
- S'ils ne co-varient pas, la covariance devrait être **modérément faible**.
- Si $\boldsymbol{cov}\ (\boldsymbol{x}, \boldsymbol{y}) > \mathbf{0}$: cela suggère que, les grandes valeurs de x sont généralement associées aux grandes valeurs de y et les petites valeurs de x aux petites valeurs de y (le nuage de points à une orientation ascendante).
- Si $\boldsymbol{cov}\ (\boldsymbol{x}, \boldsymbol{y}) < \mathbf{0}$: cela suggère que, les grandes valeurs de x sont généralement associées aux petites valeurs de y et les petites valeurs de x aux grandes valeurs de y (le nuage à une orientation descendante).

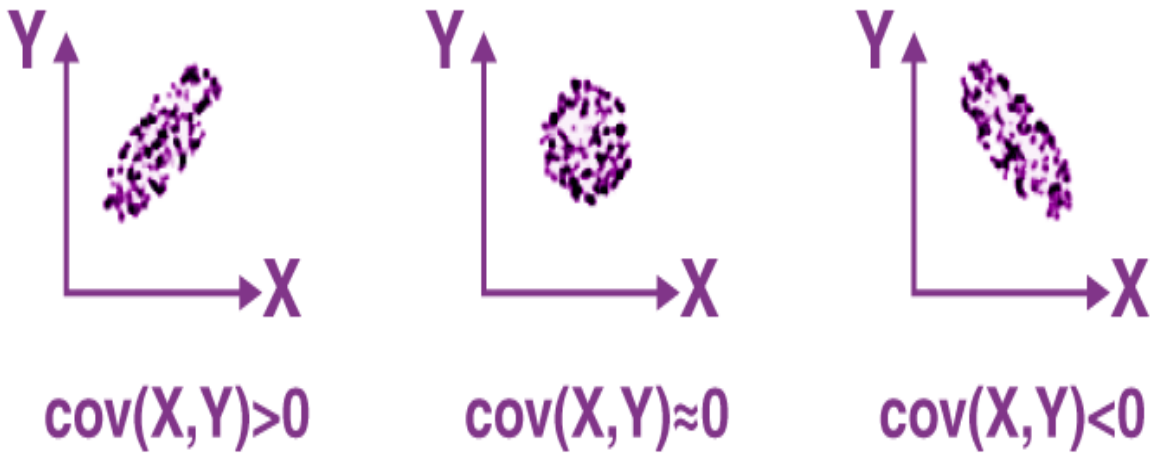


Figure 7 : Variation de la covariance

Le coefficient de corrélation linéaire de Bravais-Pearson, est une mesure de la liaison entre les variables.

$$r = \frac{Cov(x,y)}{\sqrt{v(x)}\sqrt{v(y)}}$$

Les illustrations « scatterplot » suivantes donnent quelques valeurs possibles pour le coefficient de corrélation.

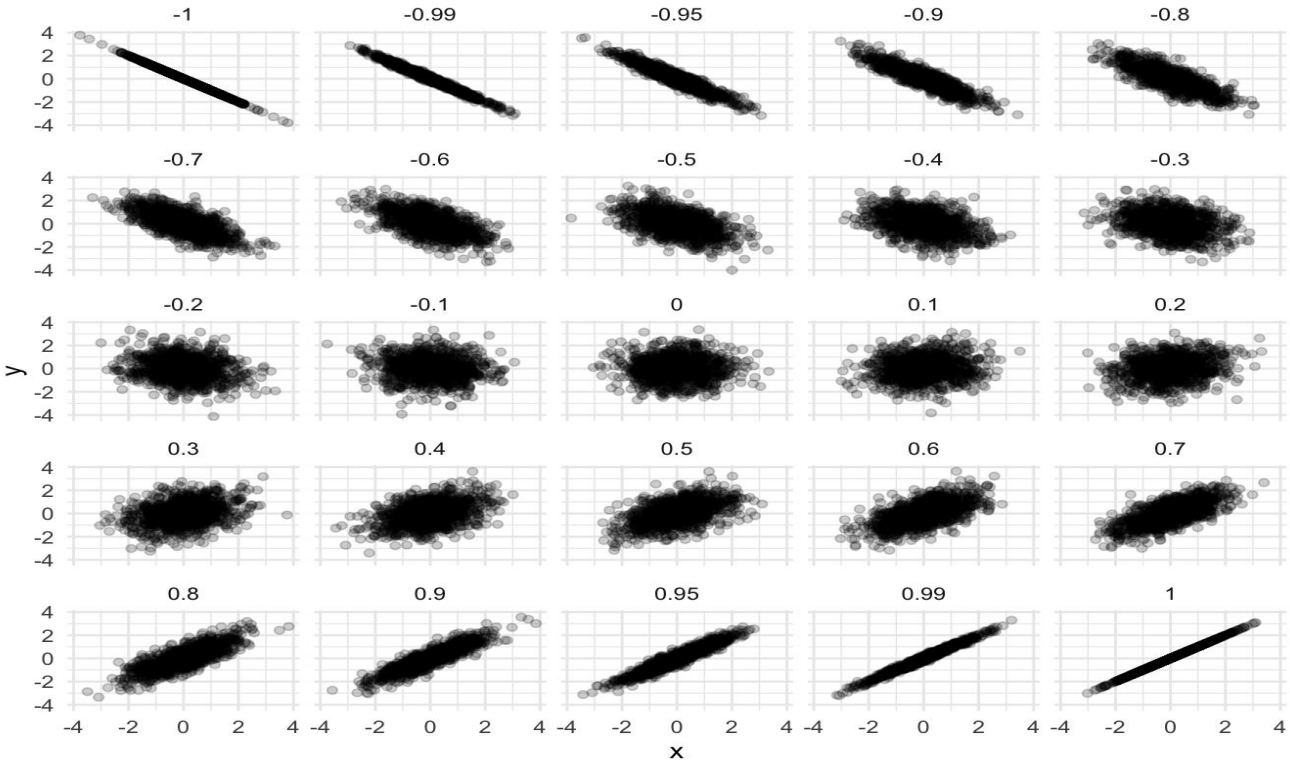


Figure 8 : diagrammes de dispersion correspondant à différentes valeurs de la corrélation.

Note :

Le coefficient de corrélation varie dans l'intervalle suivant : $-1 \leq r \leq 1$

- Si les deux variables varient indépendamment l'une de l'autre, sa valeur est de 0 (absence de corrélation linéaire entre les deux mesures).
- Si les deux variables évoluent parallèlement (y augmente lorsque x augmente), sa valeur sera positive, avec un maximum de 1 (lorsque l'évolution de y est directement proportionnelle à celle de x) ($corrélation\ linéaire\ possible$).
- Si les deux variables évoluent à l'inverse l'une de l'autre, sa valeur sera négative, avec un minimum de -1($corrélation\ linéaire\ possible$).

▶ **Coefficient de détermination « R^2 »**

C'est la variance expliquée par le modèle de régression

$$(R)^2 = \frac{Cov(x,y)^2}{V(x)V(y)}$$

Tableau 4 : variation du coefficient de détermination R^2

R^2 est nul	R^2 vaut 1	R^2=1
L'équation de la droite de régression détermine 0% de la distribution des points.	L'équation de la droite de régression est capable de déterminer 100% de la distribution des points.	Les points sont exactement alignés sur la droite de régression

Notes

Plus **le coefficient de détermination** se rapproche **de 0,** plus le nuage de points est **diffus autour de la droite de régression**.

Au contraire, plus le R^2 **tend vers 1**, plus le nuage de points **se rapproche de la droite de régression**.

b) Droite de régression par la méthode des moindres carrés

Les données $\{(x_i, y_i), i = 1, \ldots, n\}$ peuvent être représentées par un nuage de n points dans le plan (x, y), le diagramme de dispersion.

Le centre de gravité de ce nuage peut se calculer facilement : il s'agit du point de coordonnées

$$(\bar{x}, \bar{y}) = (\frac{1}{n}\sum_{i=1}^{n} x_i ; \frac{1}{n}\sum_{i=1}^{n} y_i)$$

Rechercher une relation affine entre les variables X et Y revient à rechercher une droite qui s'ajuste le mieux possible à ce nuage de points.

Parmi toutes les droites possibles, on retient celle qui jouit d'une propriété remarquable : c'est celle qui rend minimale la somme des carrés des écarts des valeurs observées yi à la droite $\boldsymbol{\hat{y}_i = ax_i + b}$.

Si ε_i représente cet écart, appelé aussi résidu, le principe des moindres carrés ordinaire (MCO) consiste à choisir les valeurs de a et de b qui minimisent (Figure 9).

$$E = \sum_{i=0}^{n} \varepsilon_i^2 = \sum_{i=0}^{n} (y_i - (ax_i + b))^2$$

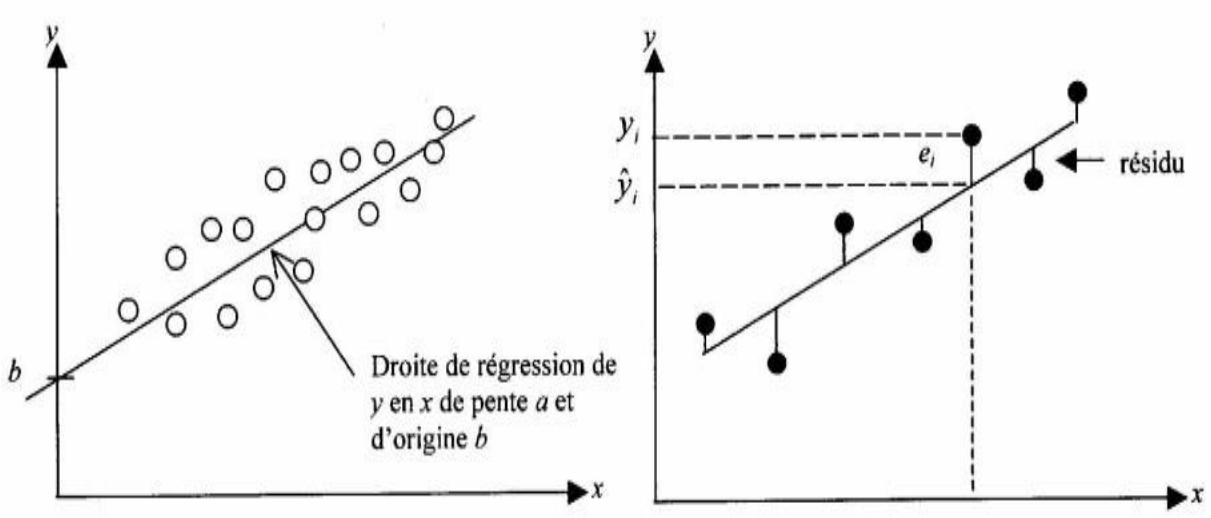


Figure 9 : La régression linéaire selon la méthode des moindres carrés

Un calcul montre que ces valeurs, notées $\hat{a}$ et $\hat{b}$, sont égales à :

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

et

$$\boldsymbol{\hat{b} = \bar{y} - \hat{a}\bar{x}}$$

On exprime souvent $\hat{a}$ au moyen de la variance de X, s_x^2, et de la covariance des variables aléatoires X et , cov_{xy}:

$$\boldsymbol{\hat{a} = \frac{Cov(x,y)}{V(x)}}, \text{avec } \boldsymbol{S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \text{et } \boldsymbol{Cov_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}$$

2. Régression linéaire multiple

2.1 Objectif

Il s'agit d'étudier les variations d'une variable quantitative Y (variable dite ou à expliquer dépendante, supposée aléatoire) en fonction de p (p > 1) variables explicatives X_1 ,X_2 ,X_3 ,...X_P explicative (variables aussi dites indépendantes).

Les variables explicatives peuvent être uniquement quantitatives, uniquement qualitatives (auquel cas on retombe sur l'ANOVA présentée au chapitre 7), ou un mélange de variables quantitatives et qualitatives. Dans ce dernier cas, le modèle de régression linéaire multiple est aussi appelé ANCOVA

2.2 Modèle

Une variable quantitative Y dite à expliquer est mise en relation avec **p variables quantitatives $X^1, \ldots, X^p$** dites explicatives.

Les données sont supposées provenir de l'observation d'un échantillon statistique de taille **n (n > p + 1) de $R^{(p+1)}$** :

$$n(n > p+1) de\ \mathbb{R}^{(p+1)}:$$
$$\left(x_i^1, \ldots, x_i^j, \ldots, x_i^p, y_i\right) \quad i = 1, \ldots, n.$$

L'écriture du modèle linéaire dans cette situation conduit à supposer que l'espérance de Y appartient au sous-espace de R^n engendré par **{1, $X^1, \ldots, X^p$}** où 1 désigne le vecteur de R^n constitué de "1" .

C'est-à-dire que les **(p + 1)** variables aléatoires vérifient :

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

Avec les hypothèses suivantes :

1. Les ε_i sont des termes d'erreur, non observés, indépendants et identiquement distribués ;

$$\mathbf{E}(\varepsilon_i) = 0, \quad \mathbf{Var}(\varepsilon) = \sigma^2 \mathbf{I}.$$

2. Les termes x^j sont supposés déterministes (facteurs contrôlés) ou bien l'erreur ε est indépendante de la distribution conjointe de $X^1, \ldots, X^p$.

On écrit dans ce dernier cas que :

$$E\left(\frac{Y}{X^1}, \ldots, X^P\right) = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \cdots + \beta_p X^p \quad et\ Var\left(\frac{Y}{X^1}, \ldots, X^P\right) = \sigma^2$$

3. Les paramètres inconnus $\beta_0, \ldots, \beta_p$ sont supposés constants.

4. En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur $\varepsilon(N(0, \sigma^2 I))$.Les ε_i sont alors i.i.d. de loi **N (0, σ^2).**

Les données sont rangées dans une matrice **X (n × (p + 1))** de terme général x_j^i, dont la première colonne contient le vecteur 1 (x^i_0 = 1), et dans un vecteur Y de terme général y_i. En notant les vecteurs ε = [ε1 · · εp] ′ et **β = [$\beta_0 \beta_1 \cdots \beta_p$]′**, le modèle s'écrit matriciellement :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

*** Estimation**

Conditionnellement à la connaissance des valeurs des X^j, les paramètres inconnus du modèle : le vecteur β et σ^2 (paramètre de nuisance), sont estimés par minimisation du critère des **moindres carrés (M.C.)** ou encore, en supposant (iv), par **maximisation de la vraisemblance (M.V.).** Les estimateurs ont alors les mêmes expressions, l'hypothèse de normalité et l'utilisation de la vraisemblance conférant à ces derniers des propriétés complémentaires.

Attention, de façon abusive mais pour simplifier les notations, estimateurs et estimations des paramètres β, c'est-à-dire la réalisation de ces estimateurs sur l'échantillon, sont notés de la même façon b.

2.3 Estimation par moindres carrés.

L'expression à minimiser sur β ∈ Rp+1 s'écrit :

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i^1 - \beta_2 x_i^2 - \cdots \beta_p x_i^p)^2 = \|y - x\beta\|^2$$
$$= (y - x\beta)'(y - x\beta)$$
$$= y'y - 2\beta' x'y + \beta' x' x\beta$$

Par dérivation matricielle de la dernière équation on obtient les "équations normales" :

$$x'y - x'x\beta = 0$$

Dont la solution correspond bien à un minimum car la matrice 2X′X est semi définie-positive.

Nous faisons l'hypothèse supplémentaire que la matrice X′X est inversible, c'est-à-dire que la matrice X est de rang (p + 1) et donc qu'il n'existe pas de colinéarité entre ses colonnes.

En pratique, si cette hypothèse n'est pas vérifiée, il suffit de supprimer des colonnes de X et donc des variables du modèle.

Des diagnostics de colinéarité et des aides au choix des variables sont explicités dans une présentation détaillée du modèle linéaire. Alors, l'estimation des paramètres βj est donnée par :

$$\mathbf{b = (X'X)\ -1X'y}$$

et les valeurs ajustées (ou estimées, prédites) de y ont pour expression :

$$\mathbf{y = Xb = X(X'X)^{-1}X'y = Hy}$$

Où H = X(X′X) $^{-1}$X′ est appelée **"*hat matrix*"** ; elle met un chapeau à y.

On note

$$\mathbf{e = y - \hat{}y = y - Xb = (I - H)\ y}$$

Le vecteur des résidus ; c'est la projection de y sur le sous-espace orthogonal de Vect(X) dans R^n.

2.4 Analyse d'une régression multiple

a) Présentation des données

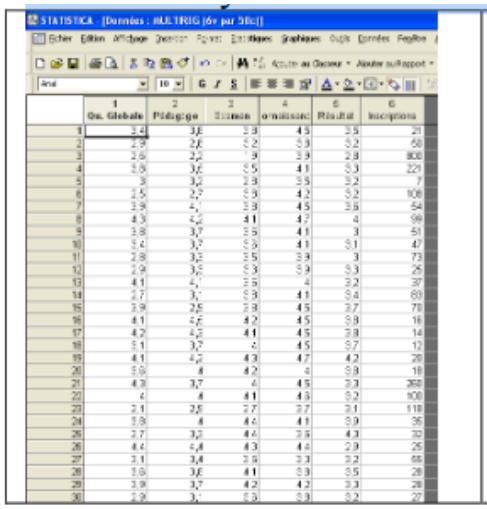
Une association a réalisé une évaluation des cours d'une université durant un semestre. Notre exemple agit d'une base de données de 6 colonnes et de 50 observations.

- La première est la qualité globale des exposés,

- Ensuite les aptitudes pédagogiques du professeur,
- La qualité des examens,
- La connaissance dont témoigne l'enseignant sur sa matière selon le point de vue des étudiants,
- les résultats auxquels s'attendent les étudiants pour ce cours (très bon à insuffisant) le nombre d'inscriptions à son cours.

Le questionnaire contenait des échelles en 5 points (de très mauvais à excellent) :

La qualité globale perçue du cours est la VD. Les 5 autres variables sont les prédicteurs. Les 50 observations correspondent à différents cours par exemple la 4ème ligne est le 4ème cours. Il s'agit ici des données moyennes obtenues pour chaque cours sur chaque critère.

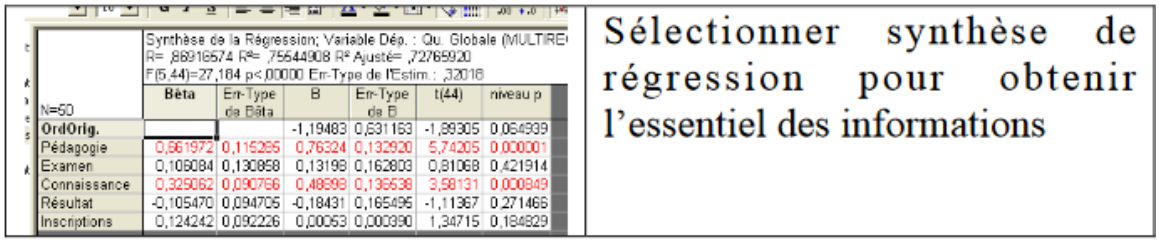


1 - étude de la distribution des valeurs pour les différentes variables (faire étude descriptive des données / distributions, moyennes et variances etc.) :
Elles sont à peu près distribuées normalement, la variabilité est raisonnable, les notes sont sensiblement supérieures à 3. Il y a un biais positif. Les enseignants sont jugés médiocres pédagogues, mais ils semblent bien maîtriser leurs cours (deux scores très bas pour les qualités pédagogiques) et on note deux valeurs extrêmes pour la fréquentation des cours (220 et 800). Le cours 3 est fort en nombre, de faible qualité pédagogique et l'examen est jugé peu adapté.

b) R multiple, R^2 et erreur standard d'estimation

Le R multiple (0.869) est significatif, F(5, 44) = 27.18, p < 0. 0001.

L'ensemble des variables (les 5 variables impliquées dans le modèle : pédagogie, examen, connaissance, résultat, inscriptions) explique près de 75% de la variance (R^2 = .755), et l'erreur standard d'estimation est peu importante (0,32).



Synthèse de la Régression; Variable Dép. : Qu. Globale (MULTIRE
R= ,86916574 R²= ,75544908 R² Ajusté= ,72765920
F(5,44)=27,184 p<,00000 Err-Type de l'Estim : ,32016

N=50	Bêta	Err-Type de Bêta	B	Err-Type de B	t(44)	niveau p
OrdOrig.			-1,19483	0,631163	-1,89305	0,064939
Pédagogie	0,861972	0,115295	0,76324	0,132920	5,74205	0,000001
Examen	0,106084	0,130858	0,13198	0,162803	0,81068	0,421914
Connaissance	0,329062	0,090766	0,48998	0,135138	3,58131	0,000849
Résultat	-0,105470	0,094705	-0,18431	0,165496	-1,11367	0,271466
Inscriptions	0,124242	0,092226	0,00053	0,000390	1,34715	0,184829

Sélectionner synthèse de régression pour obtenir l'essentiel des informations

c) Les coefficients de régression (B) et les coefficients de régression standardisés (b)

- **Coefficients standardisés b (Beta)**

La lecture du tableau nous permet de repérer les coefficients de régression standardisés b (Beta) au niveau de la première colonne (colonne Beta).

Note : Un coefficient exprime pour une variable indépendante le poids explicatif qu'elle exerce sur la variable dépendante. Plus ce coefficient est important (oscille entre + 1 et -1) plus le poids de la variable est important.

Dans le cas présent les variables b Pédagogie = + 0.66 /p < 0.000001/ et b Connaissance = + 0.32 / p <0.0008/ ont un rôle prédicteur signifiant sur l'évaluation de la qualité de l'enseignement. Cela veut dire que dans le cas de la variable pédagogie, que l'évaluation de la pédagogie de l'enseignant est dépendante de la qualité de son enseignement. Il en est de même pour la variable Connaissance. Cela se lit comme une corrélation mais c'est davantage qu'une corrélation.

Le b signifie également que la variable (connaissance) exerce une influence directe sur la qualité globale perçue de l'enseignement indépendamment de l'effet potentiel de toutes les autres variables qui ont été introduites dans le modèle.

Le b exprime l'effet « net », ou un effet principal de la VI sur la VD, sachant que les effets des autres VI sur la VD ont été contrôlés ou maintenus **constants (c'est-à-dire que les valeurs des VI ne changent pas).** Autrement dit la variation de la VI connaissance entraîne une variation positive de la VD quand les autres variables restent fixes.

Finalement, les b permettent de faire le graphe de régression à partir de la constitution de la droite de régression.

Les trois autres valeurs b ne sont pas significatives. La variable examen ne contribue pas à la prédiction de la VD. Cette variable était pourtant corrélée avec la VD à 0.596, et le r était significatif. Ceci montre qu'une corrélation entre deux mesures ne dit pas tout et notamment sur la contribution et la prédiction de la VD par une autre mesure. Un test t est réalisé pour rendre compte de la significativité du b.

Finalement, les b permettent de faire le graphe de régression à partir de la constitution de la droite de régression.

- **Coefficients non standardisés**

A quoi servent les coefficients de régression non standardisés ?
Elle permet de prédire théoriquement la VD (Qualité globale) pour chaque sujet (ici il s'agit d'un cours) et elle permet de calculer l'ajustement du modèle aux données observées (l'erreur).

La colonne 3 nommée B donne les coordonnées de l'équation ou les coefficients de régressions non standardisés qui permettent de rendre compte de la justesse du modèle.

Qualité globale = -1.195 + 0.001 Inscriptions + 0.132 Examen - 0.184 Résultat + 0.489 Connaissance + 0.763 Pédagogie

Prenons les valeurs observées du cours 1 (première ligne de la base de données) pour les 5 variables indépendantes et pour la VD:

- VD (Qualité globale observée) = 3.4,
- Inscription observé = 21,
- Examen observé = 3.8,
- Résultat observé = 3.5,

- Connaissance observé = 4.5
- Pédagogie observé = 3.8

Prenons l'équation de la régression multiple :

Qualité globale prédite = -1.195 + 0.001 x (Inscription observé) + 0.132 x (Examen observé) - 0.184 x (Résultat observé) + 0.489 x (Connaissance observé) + 0.763 x (Pédagogie observé)

Le tableau suivant donne les erreurs pour chaque cours entre la valeur prédite et la valeur observée.

Valeurs Prévues & Résidus (MULTIREG)
Var. dépendante : Qu. Globale

N° d'Obs.	Valeur Observée	Valeur Prévue	Résidus	Standard Val.Prév	Standard Résidus	E
1	3,400000	3,773387	-0,373387	0,41690	-1,16617	0
2	2,900000	2,659205	0,240796	-1,67045	0,75206	0
3	2,600000	2,546428	0,053572	-1,88194	0,16732	0
4	3,800000	3,451187	0,348812	-0,18530	1,08942	0
5	3,000000	2,742416	0,257584	-1,51441	0,80449	0
6	2,500001	2,888142	-0,388141	-1,24114	-1,21225	0
7	3,900000	4,001268	-0,101268	0,84624	-0,31628	0
8	4,300000	4,164907	0,135093	1,15310	0,42192	0
9	3,800000	3,582992	0,217008	0,06187	0,67776	0
10	3,400000	3,562460	-0,162459	0,02336	-0,50740	0
11	2,800000	3,178263	-0,378263	-0,69709	-1,18140	0
12	2,900000	3,071351	-0,171351	-0,89758	-0,53517	0
13	4,100000	3,795170	0,304830	0,45975	0,95205	0
14	2,700000	3,094539	-0,394539	-0,85410	-1,23223	0
15	3,900000	3,075361	0,824639	-0,89006	2,57553	0
16	4,100000	4,302525	-0,202525	1,41116	-0,63253	0
17	4,200000	4,135629	0,064371	1,09619	0,20105	0
18	3,100000	3,681869	-0,581869	0,24728	-1,81730	0
19	4,100000	4,112926	-0,012926	1,05662	-0,04036	0
20	3,600000	3,677466	-0,077466	0,23903	-0,24194	0
21	4,300000	3,886913	0,414087	0,62992	1,29328	0
22	4,000000	4,111333	-0,111333	1,05263	-0,34772	0
23	2,100000	2,674804	-0,574804	-1,64120	-1,79524	0
24	3,800000	3,743263	0,056737	0,36241	0,17720	0
25	2,700000	2,889206	-0,189206	-1,23915	-0,59093	0
26	4,400000	4,361107	0,038893	1,52102	0,12147	0
27	3,100000	2,928074	0,171926	-1,16626	0,53696	0
28	3,600000	3,474370	0,125629	-0,14182	0,39237	0

L'équation prédit que :

1/ Qualité globale prédite = -1.195 + 0.001 x 21 (d'Inscription) + 0.132 x 3.8 (d'Examen) - .184 x 3.5 (de Résultat) + .489 x 4.5 (de Connaissance) + .763 x 3.8 (de Pédagogie)

2/ Qualité globale prédite = -1.195 + (0.001 x 21) + (0.132 x 3.8) – (.184 x 3.5) + (.489 x 4.5) + (.763 x 3.8) = **3.773**

3/ On compare la valeur observée de Qualité globale (**3.4**) à la valeur prédite de Qualité globale (**3.733**):
résidu sujet 1 = 3.4 – **3.773** = - 0.373. Il y a une surestimation (erreur ou résidu) de .3733 du modèle pour le cours numéro 1.

Vous retrouvez ce tableau dans analyse des résidus.

Que signifient les coefficients de régression non standardisés ?

L'équation (**Qualité globale = -1.195 + 0.001 Inscriptions + 0.132 Examen - 0.184 Résultat + 0.489 Connaissance + 0.763 Pédagogie)** révèle qu'à chaque fois qu'un étudiant de plus viendrait suivre le cours de l'enseignant, l'évaluation globale de l'enseignant par les étudiants s'améliorerait de 0.001 point. Donc un cours noté 3,1 quand il y a 10 étudiants qui composent la classe passerait à 3,101 quand un 11ème étudiant viendrait suivre ce cours).

Attention :

Cette lecture n'est vraie que si, dans l'absolu, aucune autre variable que la taille des effectifs n'est modifiée. Ceci n'est pas possible dans la réalité, car si vous augmentez les effectifs, vous allez modifier certainement d'autres paramètres, comme la pédagogie et en retour, la pédagogie peut modifier les valeurs des autres scores.

Par exemple, pour la variable connaissance, si la qualité globale de l'enseignant est notée 3,1 par ses étudiants, et que la pertinence des connaissances enseignées augmente d'une unité (par exemple note moyenne qui passe de 3 à 4), la qualité globale de l'enseignant sera de 3.589 (soit + 0.489).

Attention, ces scores de régression n'indiquent pas l'importance relative des différents prédicteurs (ce sont les Beta qui apportent cette information). Ce n'est pas parce que la valeur de l'aptitude pédagogique 0,763 est plus grande que la valeur taille des effectifs 0,001 qu'elle explique davantage la VD.

La lecture des corrélations partielles à partir des données dans ce tableau, on voit que, pour la variable pédagogie, la corrélation partielle est 0.6544. Cela veut dire que l'effet de la pédagogie (VI1) sur la qualité globale est pur des effets des autres VI sur elle-même et sur la VD. Lorsque cette valeur est élevée au carré (.6544 * .6544), nous obtenons un R2 = 0.43. Cela veut dire que 42,84 % de la variation touchant la qualité globale qui ne peut être expliquée par les autres prédicteurs peut être expliquée par les aptitudes pédagogiques.

d) Problèmes de la colinéarité inter-corrélations

Lorsque plusieurs variables explicatives apportent le même type d'information, plusieurs phénomènes peuvent apparaître :

- qualité des estimations perturbée (variance très grande) ;
- valeurs des coefficients contradictoires (signes opposes) ;
- coefficients devenant non significatifs.

C'est ce que l'on nomme le ***problème de colinéarité*.**

Le critère utilisé pour juger de la colinéarité entre les variables explicatives est le facteur d'inflation de la variance VIF (variance inflation factor) ou r^2 ne désigne rien d'autre que le coefficient de corrélation multiple au carré (coefficient de J détermination) lorsque l'on régresse la j-^{ième} variable explicative Xj sur l'ensemble des autres régresseurs.

Introduction aux statistiques multivariées

1. Objectifs de l'analyse statistique multivariée

L'analyse statistique multivariée constitue un ensemble de méthodes statistiques qui ont pour but de résumer les données issues de plusieurs variables en minimisant la déperdition de l'information.

2. Critères de choix des méthodes statistiques

Le choix d'une méthode dépend de :

- **L'objectif initial,**
- **Les types de variables manipulées,**
- **La forme des résultats à obtenir.**

3. Types de méthodes multivariées

Il existe différents groupes de méthodes multivariées

3.1. METHODES DESCRIPTIVES :

- L'analyse en composantes principales (**ACP**) cherche à représenter dans un espace de dimension faible (<< *p*) un nuage de points représentant *n* individus, ou objets, décrits par *p variables quantitatives* (donc de dimension *p*) en utilisant les corrélations existant entre ces variables.

- L'analyse des correspondances (**AFC** ou **ACM**) étudie les proximités entre individus décrits par deux ou plusieurs *variables qualitatives* ainsi que les proximités entre les modalités de ces variables.

- Les méthodes de classification (***clustering***) ou de typologie procèdent par regroupement des individus en classes homogènes (classifications hiérarchiques, arbres phylogénétiques, moyennes mobiles (*K-means*).

3.2. METHODES EXPLICATIVES ET PREDICTIVES:

- L'analyse discriminante (**AFD**) étudie la prévision d'une variable qualitative par des variables numériques. C'est une méthode géométrique en espace réduit.

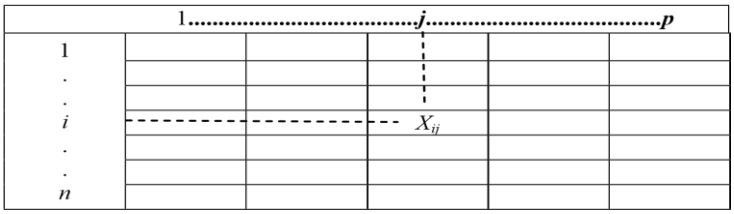
- Les **arbres de décision** et régressions (**GLM**) étudient la prévision d'une variable qualitative ou quantitative dépendante par une combinaison linéaire de variables explicatives (modèles de régression)

4. Analyse en composantes principales (ACP)

4.1.L'ACP : de quoi s'agit-il ?

L'analyse en composantes principales (ACP) cherche à représenter dans un espace de dimension faible (<< $\boldsymbol{p}$) un nuage de points représentant $\boldsymbol{n}$ individus, ou objets, décrits par $\boldsymbol{p}$ variables quantitatives (donc de dimension $\boldsymbol{p}$) en utilisant les corrélations existant entre ces variables.

Tableau 5. Présentation des données



X_{ij} = valeur de la variable *j* prise par l'individu *i*

4.2.Types de tableau de l'ACP

❖ Tableaux de mesures

Dosages, densités optiques (absorbance), comptages (nombres d'individus, d'événements, etc…). Les variables (descripteurs) sont continuées ou entières, quantitatives.

❖ Tableaux de notes

Evaluation d'intensité de maladie, de qualité, etc. L'œil (l'odorat, le toucher…) remplace l'instrument de mesure. L'intervalle de notation doit être suffisamment grand (8 à 10 classes de notes au moins). Il s'agit de variables qualitatives ordinales.

❖ Tableaux de rangs

Les variables sont des rangs, les n observations sont classées de 1 à n (du plus faible au plus fort, du plus rapide au plus lent, etc.)

Remarque

- ❖ L'ACP est fortement influencé par l'ordre de grandeur des variables.
- ❖ Les variables ayant les plus grandes variances engendrent les premières composantes.
- ❖ En cas de fortes hétérogénéité des dimensions (mesures dans des unités différentes) on recommande d'effectuer l'ACP sur des données centrées-réduites (matrice des corrélations). Sinon, effectuer l'analyse sur les données centrées seulement (matrice de dispersion).
- ❖ Cette remarque est sans objet pour les tableaux de rangs (variances identiques).

Attention : une variable à faible variance (donc de peu d'intérêt et devant être éliminée avant analyse) se retrouvera avec un poids équivalent aux autres variables après normalisation par centrage et réduction, ce qui n'est pas souhaitable.

4.3.L'ACP permet :

- Représenter les variables en fonction de leurs corrélations :
  - Quelles sont les variables *corrélées*, *anti corrélées* et *non corrélées* entre elles.

- Représenter les individus en fonction de leurs proximités :
  - Quels sont les individus *ressemblants* et les individus *dissemblants* (éloignés),
  - Comment les individus se situent sur les composantes qui sont des *axes synthétiques ayant une signification biologique*, écologique ou autre, dépendant de la nature des variables mesurées et de la nature des individus.

4.4.Principe de l'ACP

On substitue aux variables initiales (X) des "indices synthétiques" (F) qui sont des combinaisons linéaires de ces variables. Ces indices sont appelés axes ou composantes principales.

- Le premier axe (F1) sera tel que la variance des coordonnées des individus projetés sur cet axe sera maximale. Il explique donc un % de la variance totale du tableau.

- Le second (F2) sera orthogonal au premier (corrélation nulle avec F1) et aura la variance la plus élevée possible (mais inférieure à la première)

- Et ainsi de suite pour tous les autres (il y a exactement p composantes).

Facteurs centrés-réduits résumant les données $F_h = \sum_{j=1}^{p} u_{hj} X_j$ (non corrélés entre eux)

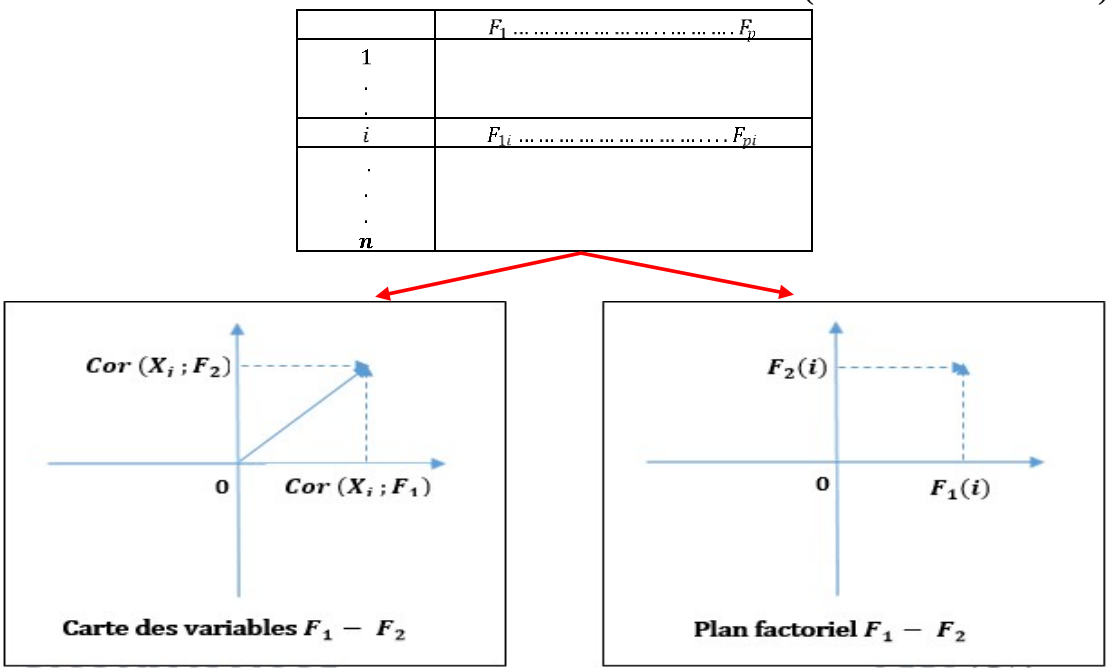


Figure 10 Exemple d'un tableau des données ; cercle de corrélation et carte factorielle L'Analyse en Composante Principale (ACP) fait partie des analyses descriptives multivariées.

4.5.Examiner les données

- ❖ Réaliser des histogrammes pour chaque variable. L'utilisation du plot pairs(données), permet de prendre connaissance des données et détecter des erreurs de mesure et autres.
- ❖ Examiner les variances et écart-types de chaque variable.
- ❖ Notez que si des variables sont très fortement corrélées, elles sont redondantes. Deux variables fortement corrélées disent la même chose ! Il convient peut-être d'en éliminer une.
- ❖ Décider des variables et observations supplémentaires.

4.6.Analyse des résultats

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

Analyser les résultats d'une ACP, c'est répondre à trois questions :

- ❖ **Les données sont-elles factorisables ?**
- ❖ **Combien de facteurs retenir?**
- ❖ **Comment interpréter les résultats?**

a) Données factorisables

Il convient donc d'observer la matrice des corrélations « Correlation Matrix ».

- Si plusieurs variables sont corrélées (> 0,5), la factorisation est possible.
- Si non, la factorisation n'a pas de sens et n'est donc pas conseillée.

Tableau 6 Exemple de la matrice de corrélation entre les variables biométriques des trois espèces d'Iris de FISHER (*I. setosa, I. verginica, I. versicolor*).

Variables	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	1,00			
Sepal Width	-0,12	1,00		
Petal Length	0,87	-0,43	1,00	
Petal Width	0,82	-0,36	0,96	1,00

b) Nombre de facteurs

Trois règles sont applicables

- ▶ **1ére règle (la règle de Kaiser)** : retenir que les facteurs aux valeurs propres supérieures à 1.

▶ **2éme règle** : choisir le nombre d'axe en fonction de la restitution minimale d'information souhaiter. Sélectionner le modèle restitue au moins 80% de l'information.

Pour ces deux premières règles, on examine le tableau « Total Variance Explained ».

Tableau 7 nombre des valeurs propres obtenues par les variables biométriques des trois espèces d'Iris de FISHER (*I. setosa, I. verginica, I. versicolor*).

Value number	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	2,92	72,96	2,92	72,96
2	0,91	22,85	3,83	95,81
3	0,15	3,67	3,98	99,48
4	0,02	0,52	4,00	100,00

▶ **3éme méthode : le « Scree-test » ou test du coude.**

Observer le graphique des valeurs propres et ne retenir que les valeurs qui se trouvent à gauche du point d'inflexion.

Partir des composants qui apportent le moins d'information (qui se trouvent à droite), puis relire par une droite les points presque alignés et on ne retient que les axes qui sont au-dessus de cette ligne.

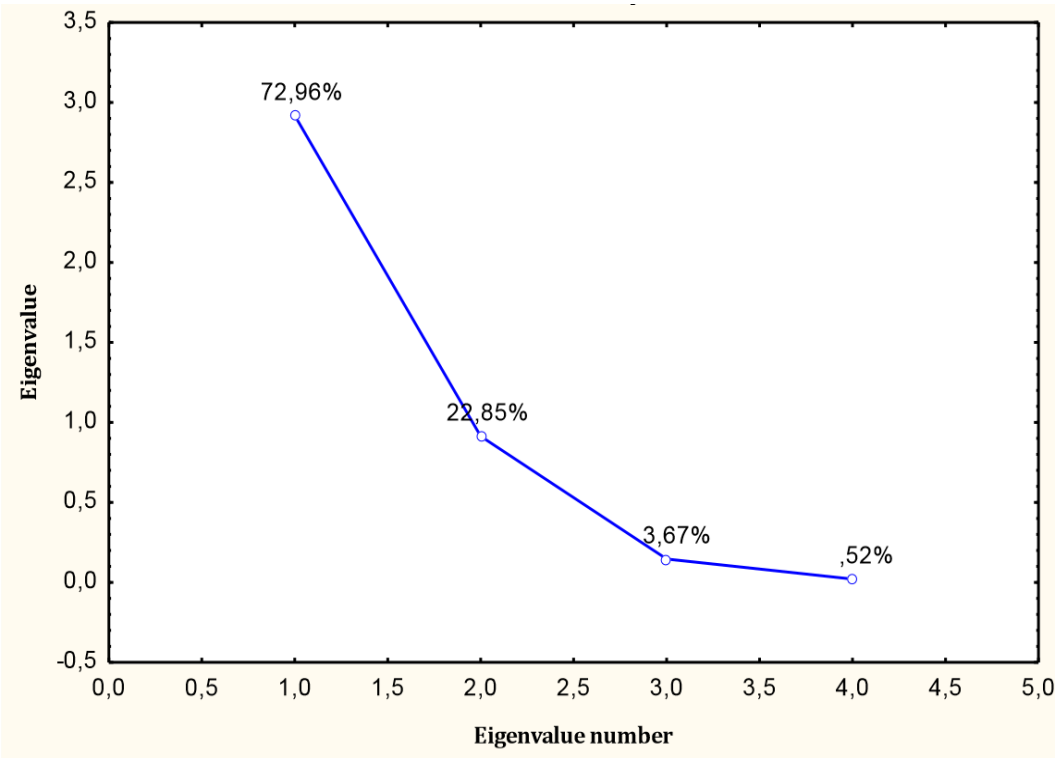


Figure 11 Graphe de coude obtenu par les variables biométriques des trois espèces d'Iris de FISHER (*I. setosa, I. verginica, I. versicolor*).

c) Interprétation des résultats

En vue de passer du nombre de variables initiales à un nombre réduit de variables obtenues par combinaison linéaire des premières : les composantes principales ; L'algorithme de l'ACP effectue sur la matrice individus/variables différentes opérations.

- Examiner les statistiques élémentaires (Les moyennes et les écarts-type sont à comparer aux références que l'on peut connaître).
- Il ne faut pas oublier que, sur le plan des variables, l'ACP est avant tout l'étude de la matrice des corrélations (ou des variances- covariances).
- Le calcul des composantes principales est fait par une technique mathématique appelée la diagonalisation. La diagonalisation fournit une valeur propre par axe principal et un vecteur propre associé.
- Les valeurs propres sont les variances des coordonnées des individus sur les axes principaux. Chaque valeur propre représente la part de l'inertie du nuage portée par l'axe correspondant.
- Chaque vecteur propre est constitué des coefficients à affecter aux variables initiales pour calculer l'axe principal correspondant.
- L'étude des variables est faite en dressant les cercles de corrélation. Ils servent à analyser les corrélations entre variables de départ et à interpréter les axes principaux.
- On attribue à chaque variable des coordonnées égales à ses corrélations avec les axes principaux. Les variables se placent sur une sphère de dimension $\boldsymbol{p}$ et de rayon $\mathbf{1}$. Les cercles de corrélation sont donc les projections des variables sur les plans considérés.

Tableau 8 Exemple des coordonnées factorielles des variables, basées sur les corrélations obtenues par les variables biométriques des trois espèces d'Iris de FISHER **(*I. setosa, I. verginica, I. versicolor*).**

Variable	Factor 1	Factor 2
Sepal Length	**-0,89**	0,36
Sepal Width	0,46	**0,88**
Petal Length	**-0,99**	0,02
Petal Width	**-0,96**	0,06

On peut retenir les règles d'interprétation suivantes :

- ✓ On ne peut interpréter que les variables proches du cercle, donc bien représentées sur le plan considéré,
- ✓ Ce sont les directions des variables par rapport aux axes qui sont à interpréter,
- ✓ Deux variables dont les directions forment un angle faible sont corrélées positivement,
- ✓ Deux variables qui sont diamétralement opposées sont corrélées négativement,
- ✓ Deux variables dont les directions sont perpendiculaires sont non corrélées linéairement.

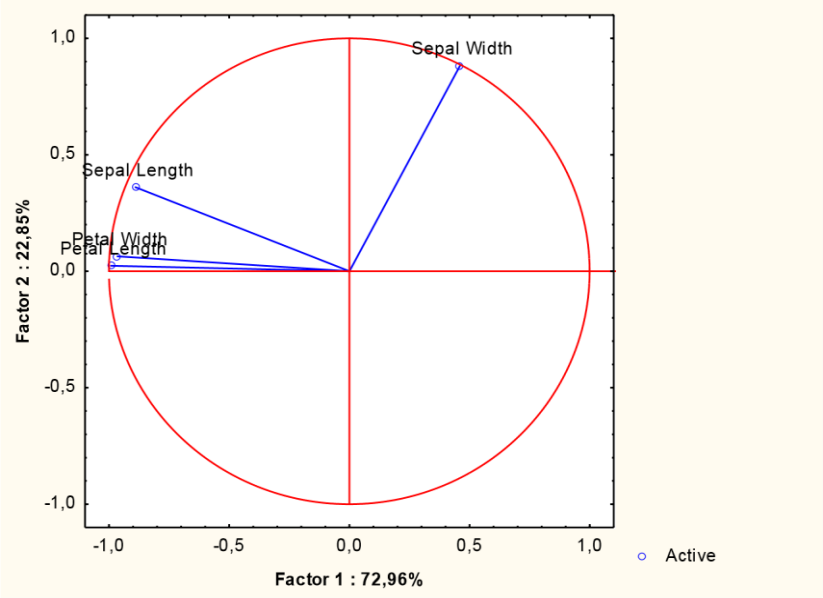


Figure 12 Cercle de corrélation des variables biométrique (Sepal Length, Petal Width, Sepal Length et Sepal Width) avec les deux premiers axes (1 et 2).

d) Examen des individus

- L'étude des individus est faite par les représentations graphiques dans les plans principaux choisis.
- On analyse la répartition des individus, en se demandant quels sont ceux qui se ressemblent, et quels sont ceux qui diffèrent.
- On se fait une idée de la qualité de la représentation d'un individu en analysant les cosinus carrés des angles que font les axes principaux avec la direction de l'individu par rapport à l'origine :
- Il faut également tenir compte du positionnement de chaque variable sur chaque axe, les variables à éliminer sont les variables qui sont :

  - **Soit proches du centre sur l'ensemble des axes retenus.**
  - **Soit au milieu d'un quart de cercle sur les axes retenus.**
  - **Soit les variables qui forment un axe à elles toute seule.**

L'ACP permet d'introduire également des variables et des individus supplémentaires qui ne participent pas à l'analyse, mais qui sont représentés en projection sur les axes formés par les variables et les individus actifs.

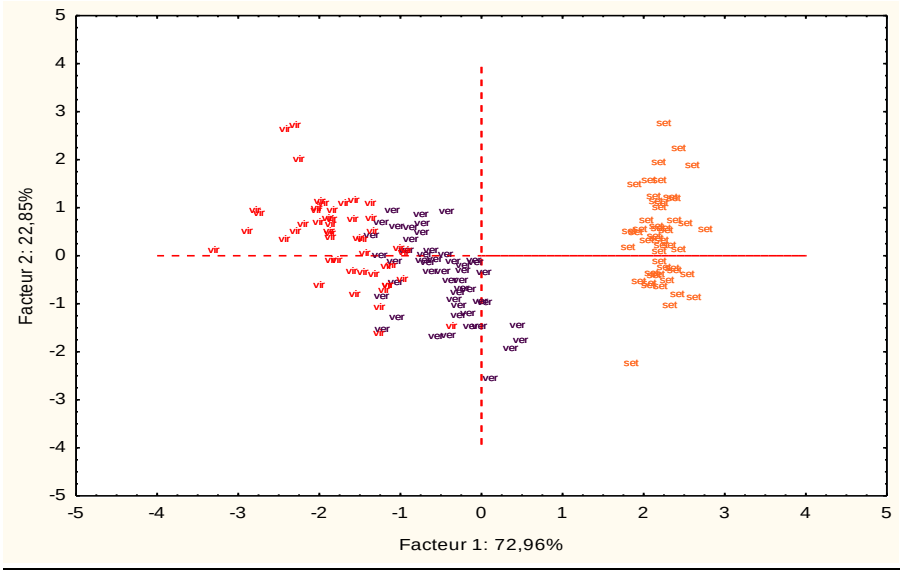


Figure 13 Projection des individus sur le plan factoriel (1 ,2)

4.7.Difference entre une ACP normée et une ACP non normée.

Le tableau suivant montre la différence entre une ACP normée et une ACP non normée.

Tableau 9 : comparaison entre une ACP normée et une ACP non normée.

ACP normée 'données centrées et réduites'	**ACP non normée** 'données simplement centrées'
Diagonaliser la matrice des corrélations	Diagonaliser la matrice des variances-covariances.
la somme des valeurs propres est égale à p (le nombre de variables)	La somme des valeurs propres est égale à la somme des variances des ***p*** variables.
Les variables initiales sont les variables centrées et réduites	Les variables initiales les variables centrées uniquement

1. Objectif

L'analyse discriminante est une technique d'analyse des données qui vise à décrire, expliquer et prédire l'appartenance d'un individu à des groupes prédéfinis. À l'origine, cette méthode a été étudiée par Ronald Fisher dès 1936, dans le but de reconnaitre le type d'iris (*setosa, virgina*, et *versicolor*) à l'aide de la longueur de ses pétales et sépales.

Précisons aussi que, la technique d'analyse discriminante donne lieu à deux principales approches.

- D'une part, **l'analyse factorielle discriminante** (ou **analyse discriminante descriptive**), qui est une méthode factorielle ou descriptive, qui comme l'ACP et l'AFC, qui a pour but de proposer un nouveau système de représentation, des variables latentes formées à partir de combinaisons linéaires des variables prédictives, qui permettent de discerner le plus possible les groupes d'individus.
- D'autre part, **l'analyse discriminante linéaire**, qui est une méthode prédictive consistant à construire une fonction de classement (règle d'affectation, …) permettant de prédire la classe dans lequel appartient un individu à partir des valeurs prises par les variables prédictives. Dans ce sens, il appartient à la seconde famille des méthodes de classification comme le stipulent.

2. Problématique

Voici la situation-type traitée par l'analyse discriminante : on a un ensemble d'individus appartenant chacun à un groupe, le nombre de groupes étant fini et >1. Deux problèmes se posent à nous :

● Trouver une représentation des individus qui sépare le mieux les groupes (analyse discriminante descriptive) ou trouver des règles d'affectation des individus à leur groupe (analyse discriminante prédictive).

● Une autre formulation est la suivante: on a un ensemble d'individus caractérisés par une variable *à expliquer Y qualitative* et des variables *explicatives* X_i *quantitatives*.

3. Principe de l'AFD

On a une variable cible (*à expliquer) Y qualitative à k modalités*, correspondant à *k* groupes G_i dont on ne note ni les effectifs. L'effectif total est n. On a d'autre part *p variables explicatives* X_j *continues*.

L'analyse factorielle discriminante consiste à remplacer les X_j par des axes discriminants, c'est-à-dire des combinaisons linéaires des X_j prenant les valeurs les plus différentes possible pour des individus différant sur la variable cible.

On reconnaîtra dans ce mécanisme une analyse en composantes principales du nuage des *k centres de gravité* des classes (pondérés par *n/n*). Les axes sont au nombre de *k-1* ou *p*, le plus petit des deux.

La figure suivante illustre simplement l'*approche géométrique descriptive* :

Dans cet exemple, on voit que :

- L'axe « x » sépare bien les groupes « B » et « C » mais non les groupes « A » et « B »,
- L'axe « y » sépare bien les groupes « A » et « B » mais non les groupes «B» et « C »,
- et l'axe « Z », combinaison linéaire de « x » et « y », sépare bien les trois groupes.

La droite d'équation z = 1 sépare les « B » et les « C », tandis que la droite d'équation z = - 1 sépare les « A » et les « B » : donc « z » est une fonction de score.

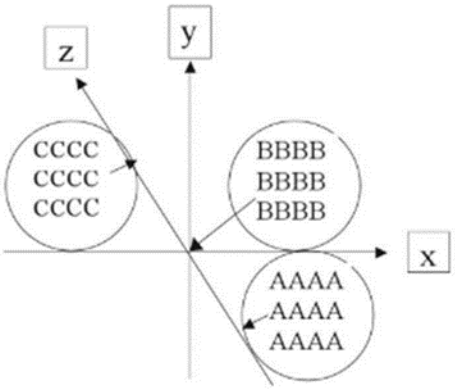


Figure 14: Approche géométrique descriptive de l'AFD

Mathématiquement, les *n* individus forment un nuage de n points dans R_P, forme des *k sous-nuages* G_i à différencier **la variance interclasse** (« ***between*** ») est par définition la variance des *barycentres* $\boldsymbol{g_i}$ ***centroïdes*** ») des classes G_i, et la matrice des covariances « *between* » est $B = 1/n \sum ni(g_i - g)(g_i - g)'$.

La variance intraclasse (« ***within*** ») est par définition la moyenne pondérée des variances des classes Gi, et la matrice des covariances « ***within*** » est $W = 1/n \sum ni\ Vi$ calculée à partir de la matrice des covariances V_i de chaque classe G_i.

D'après le *théorème de Huygens : B + W = V* la matrice des covariances totale.

II est généralement impossible de trouver un axe *u* qui, pour satisfaire l'objectif de l'analyse discriminante, simultanément :

- Maximise la variance *interclasse* sur *u: max u'Bu*;
- Minimise la variance *intraclasse* sur *u: min u'Wu*.

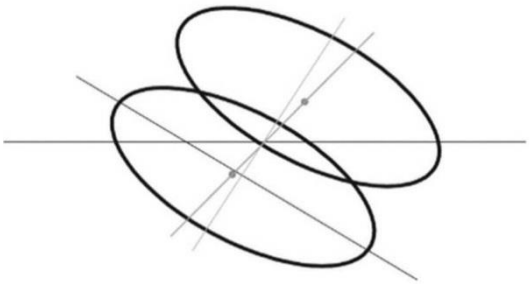


Figure 15: Double objectifs de l'AFD

On le voit bien sur la figure ci-dessus. Rechercher le max de dispersion *interclasse* nous fait choisir un axe *u* parallèles au segment joignant les *centroïdes*, tandis que rechercher le min de dispersion *intraclasse* nous fait choisir un axe *u* perpendiculaire à l'axe principal des *ellipses*. On suppose l' *homoscédasticité*, c'est-à-dire l'égalité de toutes les matrices des covariances V_i : c'est l'hypothèse de base de l'analyse factorielle discriminante.

A quoi correspondent *géometriquement* l'axe *u* et la métrique W^{-1} ? L'axe u est celui de l'ACP à laquelle nous faisions plus haut allusion, l'ACP sur le nuage des centroïdes g_i, mais c'est un axe sur lequel les points sont projetés obliquement et non orthogonalement

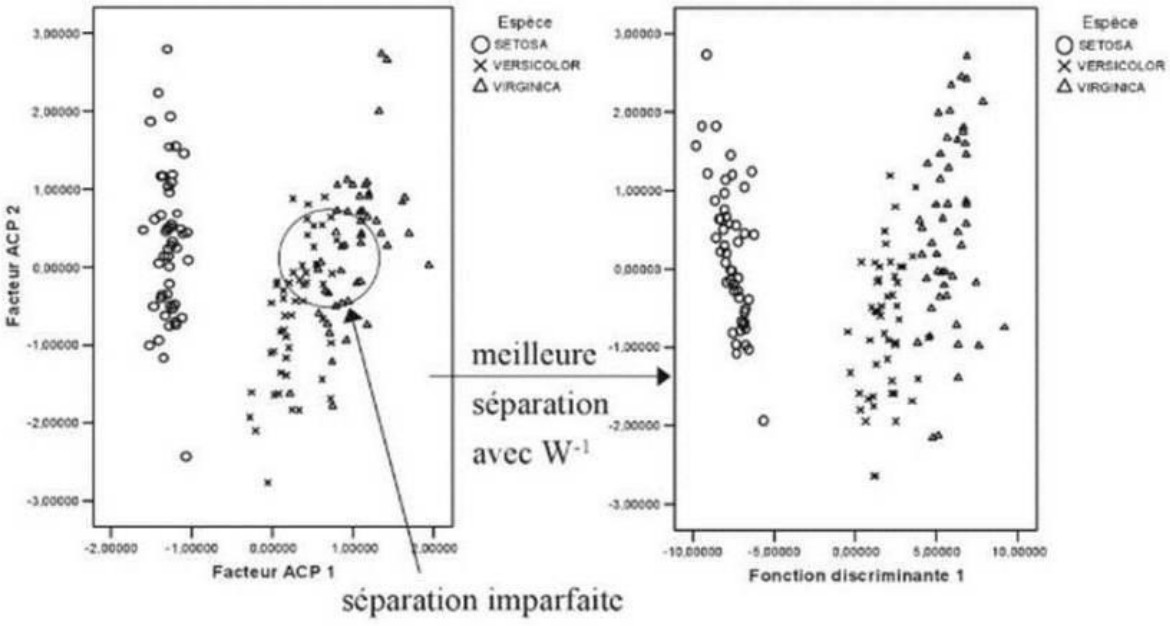


Figure 16: Comparaisons entre ACP et AFD appliqués sur les iris de Fisher

Sans cette oblicité, qui correspond aux métriques équivalentes V^{-1} et W^{-1}, il s'agirait d'une simple ACP, dans laquelle les groupes seraient moins bien séparés. Dans cette métrique, l'éloignement de deux points ne dépend pas seulement d'une mesure *euclidienne*, mais aussi de la variance et de la corrélation des variables.

4. Interprétation de l'analyse discriminante

Le fameux fichier IRIS permet d'illustrer la méthode. Il a été proposé et utilisé pour illustrer l'analyse discriminante. Il comporte 150 fleurs décrites par 4 variables (longueur et largeur des pétales et sépales) et regroupées en 3 catégories (*Setosa, Versicolor et Virginica*).

L'exemple historique d'analyse discriminante est celui des *iris de Fisher*

Figure 17: Les iris de Fisher

L'objectif est de produire un plan factoriel (3 catégories ⇒ 2 axes) permettant de distinguer au mieux ces catégories, puis d'expliquer leurs positionnements respectifs.

4.1.Axes factoriels

Les deux axes sont globalement significatifs. En effet, le lambda de Wilks de nullité des deux axes est égal à 0.023525. Le KHI-2 de Bartlett est égal à 545.57, avec un degré de liberté égal à (2x (4-3+2+1)) = 8, il est très hautement significatif (p-value très petite).

Nous constatons néanmoins que le premier axe traduit 99,1% de la variance expliquée. Nous pouvons légitimement nous demander si le second axe est pertinent pour la discrimination des groupes.

Il suffit pour cela de tester la nullité du dernier axe. Le lambda est plus élevé (0.78), ce qui se traduit par un KHI-2 plus faible (35.64) à (1 x (4-3+1+1)) = 3 degrés de liberté, il reste néanmoins significatif si l'on se fixe un niveau de confiance à 5%. Partant de ce résultat, nous serions amenés à conserver les deux axes. Nous verrons plus bas que ce résultat est à relativiser.

Tableau 10 : résultats des axes factoriels.

Axe	Val. propre	Proportion	Canonical R	Wilks	KHI-2	D.D.L.	p-value
1	32.272	0.991	0.985	0.024	545.58	8	0.0
2	0.277	1.0	0.466	0.783	35.6	3	0.0

4.2.Représentation graphique

En projetant les points dans le plan factoriel, nous obtenons le positionnement

Nous distinguons bien les trois catégories de fleurs. Nous constatons également que le premier axe permet déjà de les isoler convenablement. Sur le second axe, même si les centres de gravité des groupes semblent distincts, la différenciation n'est pas aussi tranchée.

Nous retrouvons bien dans ce graphique ce que l'on pressentait avec la proportion de variance expliquée. Le premier axe suffit largement pour discriminer les groupes. Le second axe, même s'il est statistiquement significatif, n'apporte pas un réel complément d'informations.

Très souvent, les techniques visuelles emmènent un contrepoint très pertinent aux résultats numériques bruts.

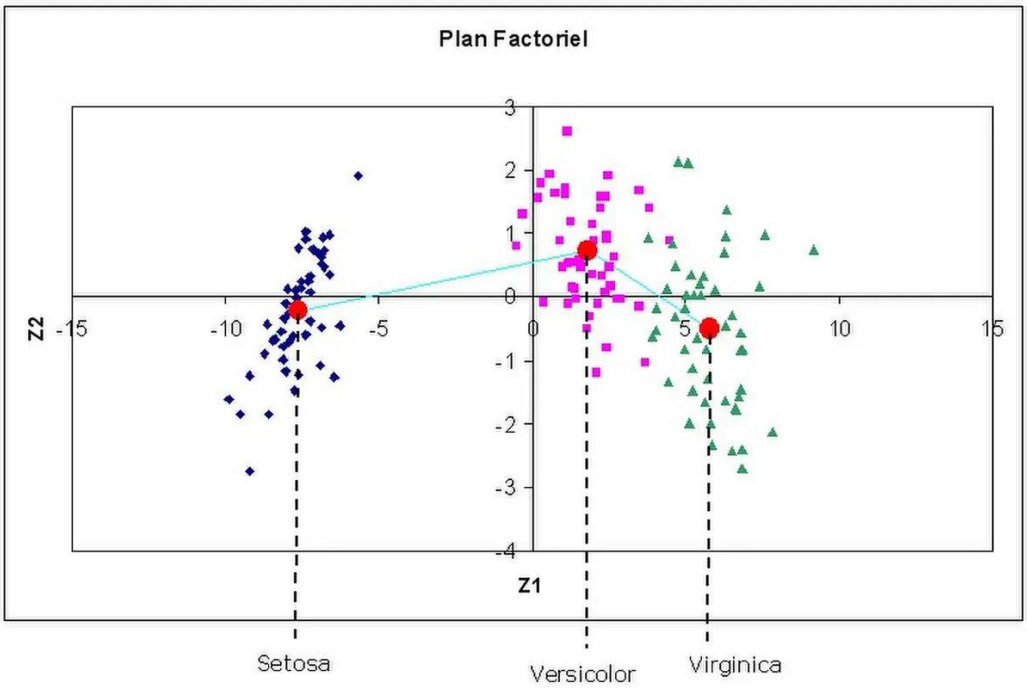


Figure 18 : premier plan factoriel de l'AFD des iris de Fisher.

4.3.Projection des individus supplémentaires

Pour projeter des observations supplémentaires dans le plan factoriel, les logiciels fournissent les équations des fonctions discriminantes. Il suffit de les appliquer sur la description de l'individu à classer pour obtenir ses coordonnées dans le nouveau repère.

Tableau 11 : corrélation des variables avec les axes factorielles

Variables	Axe 1	Axe 2
Sepal Length	-0.819	-0.033
Sepal Width	-1.548	-2.155
Petal Length	2.185	0.930
Petal Width	2.854	-2.806
Constante	-2.119	6.640

4.4.Interprétation des axes

Dernier point, et non des moindres, il nous faut comprendre le positionnement relatif des groupes, c.-à-d. expliquer à l'aide de variables initiales l'appartenance aux catégories.

Pour cela, à l'instar des techniques factorielles telles que l'analyse en composantes principales (ACP) -- *l'analyse factorielle discriminante peut être vue comme un cas particulier de l'ACP d'ailleurs* -- les logiciels fournissent la matrice de corrélation. À la différence de l'ACP, trois types de corrélations peuvent être produits : la corrélation globale entre les axes et les variables initiales ; la corrélation intra-classes, calculée à l'intérieur des groupes ; la corrélation inter-classes calculée à partir des centres de gravité des groupes pondérés par leurs fréquences.

Dans l'exemple IRIS, si nous nous en tenons au premier axe, nous obtenons les corrélations suivantes.

Tableau 12: corrélation inter-classes et corrélation intra-classes.

Variables	Total	Intra-groupes	Inter-groupes
Sep Length	0.792	0.222	0.992
Sep Width	-0.523	-0.116	-0.822
Pet Length	0.985	0.705	1.000
Pet Width	0.973	0.632	0.994

La corrélation inter-classes qui traduit le positionnement des groupes sur les axes indique ici que les *Virginica* ont plutôt des longueurs de sépales, des longueurs et des largeurs de pétales importantes. Les *Setosa* possèdent à l'inverse des longueurs de sépales, des longueurs et des largeurs de pétales réduites. Les *Versicolor* occupent une position intermédiaire. La lecture est inversée concernant la largeur des sépales.

1. Préambule descriptif de l'AFC

L'AFC (ou analyse des correspondances) est une technique descriptive et exploratoire destinée à analyser des tables à double entrée ou plus contenant des données de type fréquences ou effectifs tirées de variables de nature catégorielles.

2. Jeu de données pour réaliser une Analyse Factorielle des Correspondances

Les données correspondent à une enquête dans laquelle les personnes interrogées donnent leurs opinions sur un film qu'elles viennent de voir. On leur demande également leur tranche d'âge.

3. Objectif

Les objectifs de cette méthode sont d'étudier l'association entre deux variables (lignes et colonnes d'un tableau de contingence) et les similitudes entre les catégories de chaque variable respectivement (lignes et colonnes respectivement). L'OBJECTIF est de connaitre l'organisation des données ; et d'en connaitre la configuration après analyse.

Lorsque plus de deux variables sont utilisés dans une enquête, la meilleure technique à utiliser est l'**Analyse des Correspondances Multiples (MCA)**.

4. L'AFC : de quoi s'agit -il

Un tableau est constitué de données provenant des mesures faites sur deux ensembles de caractères disposés pour l'un en ligne et pour l'autre en colonne. C'est une technique très efficace pour analyser les tableaux de contingence.

On note souvent I l'ensemble des modalités de la variable X et J celui des modalités de Y. Le cardinal de I est noté n et celui de J est noté m.

Pour chercher les liaisons entre X et Y nous allons croiser les deux partitions pour obtenir un tableau de contingence indexé par I × J (on définit un ordre sur I et J, qui peut être éventuellement arbitraire, afin de pouvoir construire ce tableau).

Dans la case associée à la i-^{ème} ligne et à la j-^{ème} colonne on écrit l'effectif des individus ayant la i-^{ème} modalité pour la variable X et la j-^{ème} modalité pour la variable Y, celui-ci est noté k_{ij}.

Tableau 13: type de tableau de contingence complété par ses marges.

Y / X	...	j-ème colonne	...	marge
...	...	...	...	...
i-ème ligne	...	k_{ij}	...	$k_{i.}$
...	...	...	...	...
marge	...	$k_{.j}$	...	N

On pose :

$$k.j = \sum_{i-1}^{n} k_i\, j \; et \; k_i = \sum_{j-1}^{m} k_{ij}$$

$$\text{k.j} = \sum_{i-1}^{n} kij \text{ et ki} = \sum_{j-1}^{m} kij$$

k.j : l'effectif marginal de la j-ème modalité de Y,
ki. : l'effectif marginal de la i-ème modalité de X.

Les éléments du tableau de contingence divisés par l'effectif total N constituent le tableau des fréquences où l'on note f_j^i l'élément générique.

Note

Le terme de correspondance provient du fait que l'on cherche à mettre ces deux ensembles de caractères (lignes : ensemble I et colonnes : Ensemble J) en correspondance afin d'en connaitre la structure et l'organisation sous-jacentes.

L'AFC est une technique très proche de l'ACP. On pourrait dire que l'AFC est un cas particulier de l'ACP

Exemple 01

On étudie l'efficacité de différents produits de beauté sur l'amélioration de la souplesse de la peau. L'efficacité des produits sur les peaux rencontrées a été notée. Une AFC pourra aider à décrire l'efficacité des produits peaux des types de peaux différents

Exemple 02

La présence (1) ou l'absence (0) de symptômes pathologiques pour des catégories de métiers a été enregistrée.

L'ensemble I est celui des symptômes, l'ensemble J celui des catégories de métiers et le tableau est composé de 0(absence) 1(présence). Une AFC permettra de décrire la représentation des symptômes suivant les types de métiers.

5. Principe de la démarche de l'analyse des correspondances et présentation des concepts

Le principe de l'AFC est de réaliser une synthèse des données en présence (présentées dans le tableau de données). Il s'agit de faire **une réduction de l'espace de représentation de ces données** sur un nombre **minimum de dimensions** qui sont censées bien représenter toutes ces données ; de la façon la plus juste, sans trop de perte d'information après réduction. On **représente par des axes de projection** cette configuration des données autour d'un nombre réduit de dimension.

L'AFC recherche ces **composantes principales** ou axes pour les lignes et pour les colonnes. Elle les représentera sous forme de **graphe**.

Pour les lignes et les colonnes, l'AFC calcule **un premier axe** qui donne une indication sur **la proportion maximale** de la variation totale des individus. Ensuite l'AFC calcule recherche séparément pour les lignes et pour les colonnes, **une seconde composante** qui représente **une part de la variation** qui n'est pas expliquée

par le premier axe. Cette seconde composante n'est pas corrélée avec la première et représentera moins bien les données que la première composante.

L'AFC calculera ainsi de suite, séparément pour les lignes et pour les colonnes, tous les axes ou composantes nécessaires à la présentation des données. L'AFC donnera alors pour les lignes et pour les colonnes, une représentation de la totalité de la variation du nuage après l'extraction de toutes les composantes.

L'AFC représentera sous formes de **représentations graphiques** des **projections du nuage des individus** en lignes et des individus en colonnes sur les axes principaux.

6. Interprétation les résultats de l'Analyse Factorielle des Correspondances

6.1 Les profils en ligne et les profils en colonne

Avant de commencer l'interprétation, il est utile d'introduire le concept de **profil**. Les calculs des composantes ou axes se font à partir **des profils en ligne et des profils en colonnes**.

Un profil est l'ensemble des fréquences divisées par leur total, c'est à dire les fréquences relatives. En d'autres termes, un profil reflète la façon dont la catégorie d'une variable varie selon les catégories de l'autre variable.

Le premier résultat affiché est le **test d'indépendance entre les lignes et les colonnes**, basé sur une statistique **du Khi².**

- Si la valeur du **Khi² observée** est supérieure à **la valeur critique**
- Si la **p-value** est **inférieure** au niveau **alpha choisi**,

Alors on peut conclure que les lignes et les colonnes du tableau de contingence sont liées de manière significative.

Exemple : Les données correspondent à une enquête dans laquelle les personnes interrogées donnent leurs opinions sur une pommade qu'elles viennent d'appliqué. On leur demande également leur tranche d'âge.

Dans notre exemple, il est fortement probable que des différences réelles existent entre les profils d'appréciation de la pommade allergique et les différents groupes d'âge.

Tableau 14 : Test d'indépendance entre les lignes et les colonnes.

Khi² (Valeur observée)	148,268
Khi² (Valeur critique)	28,869
DDL	18
p-value	<0,0001
alpha	0,050

6.2 Interprétation du test :

H_0 : les lignes et les colonnes du tableau sont indépendantes
Ha : il existe un lien entre les lignes et les colonnes du tableau.

Etant donné que la p-value calculée est inférieure au niveau de signification α=0.05 ; on doit retenir Ha

6.3 Les valeurs propres et les vecteur propres

Les **valeurs propres quantifient la part de l'information expliquée par les différents axes**. C'est à partir des valeurs propres que l'on peut décider du nombre d'axes que l'on conserver et sur lesquels on va projeter le nuage de lignes et le nuage de colonnes.

Notes : Le nombre de valeurs propres possibles correspond à la valeur inf (n ; p) -1

- Les valeurs propres sont comprises entre 0 et 1 et expriment la part de variation du nuage des points expliqués par l'axe correspondant.
- Plus la valeur propre est **proche de 1**, plus les profils des points représentés dans l'axe sont différents et plus la part de l'information expliquée par l'axe est importante.
- Plus la valeur propre est proche de 0, plus les profils des points sont semblables et moins l'information expliquée par l'axe est importante

Dans notre exemple, la somme des deux premières valeurs propres représente 97% de l'inertie totale, l'analyse est donc de bonne qualité.

Tableau 15 : valeurs propres et pourcentage d'inertie

	F1	F2	F3
Valeur propre	0,095	0,012	0,003
Inertie %	86,640	10,674	2,685
% cumulé	86,640	97,315	100,000

6.4 Combien d'axe à retenir

Une règle d'application utile : une inertie qui explique moins de **100/p(colonne)** ou **100/n (ligne)** n'est pas intéressante.

On peut regarder la quantité d'information que l'on ne perd pas certains axes. Ensuite on considère la simplification que peut nous apporter le fait de ne pas prendre tous les axes.

6.5 Le pourcentage d'inertie

Le pourcentage d'inertie est la part de l'information représentée par chaque axe. Autrement dit, elle donne le pourcentage de la variance expliquée par l'axe ou la composante que l'AFC a calculée. Plus cette valeur est importante, plus elle rend compte du pouvoir explicatif des données par l'axe.

Cette valeur est lue à la colonne **« pourcentage d'inertie ».** Elle est obtenue en faisant le rapport de chaque valeur propre singulière à la somme de toutes les valeurs propres singulières.

Pour notre exemple Une série de tableaux est ensuite affichée pour les lignes (et les colonnes respectivement).

Un premier tableau contient les **poids**, les **distances** et **distances quadratiques à l'origine**, les **inerties** et **inerties relatives** des lignes (et respectivement des colonnes). Les poids sont des proportions marginales utilisées pour pondérer les profils des points lors du calcul des distances.

Plus la distance à l'origine est grande, plus le profil de la catégorie est différent du profil moyen (plus la catégorie participe à la dépendance entre les deux variables).

Les groupes d'âge 25-34, 35-44 et 45-54 ont la distance la plus courte à l'origine, ce qui indique que les profils de ces groupes sont proches du profil moyen.

Tableau 16: poids, distances quadratiques à l'origine, inerties et inerties relatives (lignes)

	Poids (relatif)	Distance	Distance²	Inertie	Inertie relative
16-24	0,153	0,718	0,516	0,079	0,720
25-34	0,169	0,117	0,014	0,002	0,021
35-44	0,203	0,152	0,023	0,005	0,043
45-54	0,189	0,124	0,015	0,003	0,027
55-64	0,127	0,235	0,055	0,007	0,064
65-74	0,114	0,239	0,057	0,007	0,060
75+	0,046	0,396	0,157	0,007	0,066

L'analyse des tableaux **profils lignes** (respectivement colonnes) ainsi que le **profil moyen** montre que les profils des groupes d'âge 25-34, 35-44 et 45-54 sont proches les uns des autres et du profil moyen. Ce dernier résultat confirme l'observation faite en analysant les distances à l'origine.

Tableau 17 : **profils lignes** (respectivement colonnes) et le **profil moyen.**

	MAUVAIS	MOYEN	BON	TRÈS BON	Somme
16-24	0,333	0,237	0,232	0,198	1,000
25-34	0,646	0,197	0,061	0,096	1,000
35-44	0,616	0,236	0,043	0,105	1,000
45-54	0,621	0,223	0,047	0,109	1,000
55-64	0,709	0,151	0,035	0,105	1,000
65-74	0,684	0,135	0,032	0,148	1,000
75+	0,645	0,113	0,016	0,226	1,000
Moyenne	0,608	0,184	0,067	0,141	1,000

Les **distances entre les lignes** (respectivement colonnes) fournissent des informations sur la similitude entre les catégories. Encore une fois, les groupes d'âge 25-34, 35-44 et 45-54 semblent être similaires avec des distances inférieures à 0,2.

Tableau 18 : Distance du Khi^2 (lignes)

	16-24	25-34	35-44	45-54	55-64	65-74	75+
16-24	0	0,810	0,832	0,821	0,937	0,910	0,943
25-34	0,810	0	0,119	0,093	0,165	0,232	0,440
35-44	0,832	0,119	0	0,034	0,227	0,273	0,448
45-54	0,821	0,093	0,034	0	0,202	0,244	0,424
55-64	0,937	0,165	0,227	0,202	0	0,131	0,365
65-74	0,910	0,232	0,273	0,244	0,131	0	0,235
75+	0,943	0,440	0,448	0,424	0,365	0,235	0

Les **coordonnées principales et coordonnées standard des lignes** (respectivement colonnes) sont ensuite affichées. Les coordonnées standard sont le résultat de la division des coordonnées principales par la racine carrée de la valeur propre du facteur correspondant. La somme des carrés pondérée des coordonnées standard est égale à 1 pour chaque facteur.

Les **contributions des lignes** (respectivement colonnes) sont ensuite affichées. Les contributions correspondent à l'importance de chaque catégorie pour chaque facteur (dimension). La somme des contributions est égale à 1 pour chaque facteur. En général, si la contribution est supérieure à 1/I avec I le nombre de lignes (respectivement 1/J avec J le nombre de colonnes), la catégorie est importante pour le facteur donné.

Dans notre exemple, le groupe des 16-24 ans est important pour le facteur F1, les groupes des 65-74 ans et 75 ans et plus sont importants pour le facteur F2.

Tableau 19: résultats de contributions (lignes)

	Poids (relatif)	F1	F2	F3
16-24	0,153	0,830	0,008	0,003
25-34	0,169	0,013	0,045	0,178
35-44	0,203	0,023	0,165	0,206
45-54	0,189	0,019	0,077	0,084
55-64	0,127	0,062	0,020	0,298
65-74	0,114	0,040	0,227	0,015
75+	0,046	0,013	0,457	0,216

Le tableau suivant contient les **cosinus carrés des lignes** (respectivement colonnes). Les cosinus carrés représentent l'importance de chaque facteur pour chaque catégorie. La somme des cosinus carrés est égale à 1 pour chaque catégorie.

Dans notre exemple, la quasi-totalité de la variance du groupe des 16-24 ans est attribuée au facteur F1.

Tableau 20: résultats de cosinus carrés des lignes

	F1	F2	F3
16-24	0,999	0,001	0,000
25-34	0,549	0,226	0,225
35-44	0,458	0,412	0,129
45-54	0,607	0,309	0,084
55-64	0,843	0,032	0,125
65-74	0,587	0,407	0,007
75+	0,167	0,744	0,089

Le **graphique symétrique des lignes et colonnes** est le plus couramment utilisé. Les profils des lignes et des colonnes sont superposés dans un même espace (en coordonnées principales). Les points correspondants aux lignes et aux colonnes étant également espacés, ce graphique est très pratique. Les distances entre les points-lignes (respectivement points-colonnes) correspondent aux distances du Khi² entre les lignes (respectivement entre les colonnes).

Les groupes d'âge 25-34, 35-44 et 45-54 sont presque superposés, indiquant des profils très similaires. La **proximité entre les points-lignes et les points-colonnes** ne peut pas être interprétée directement sur ce graphique.

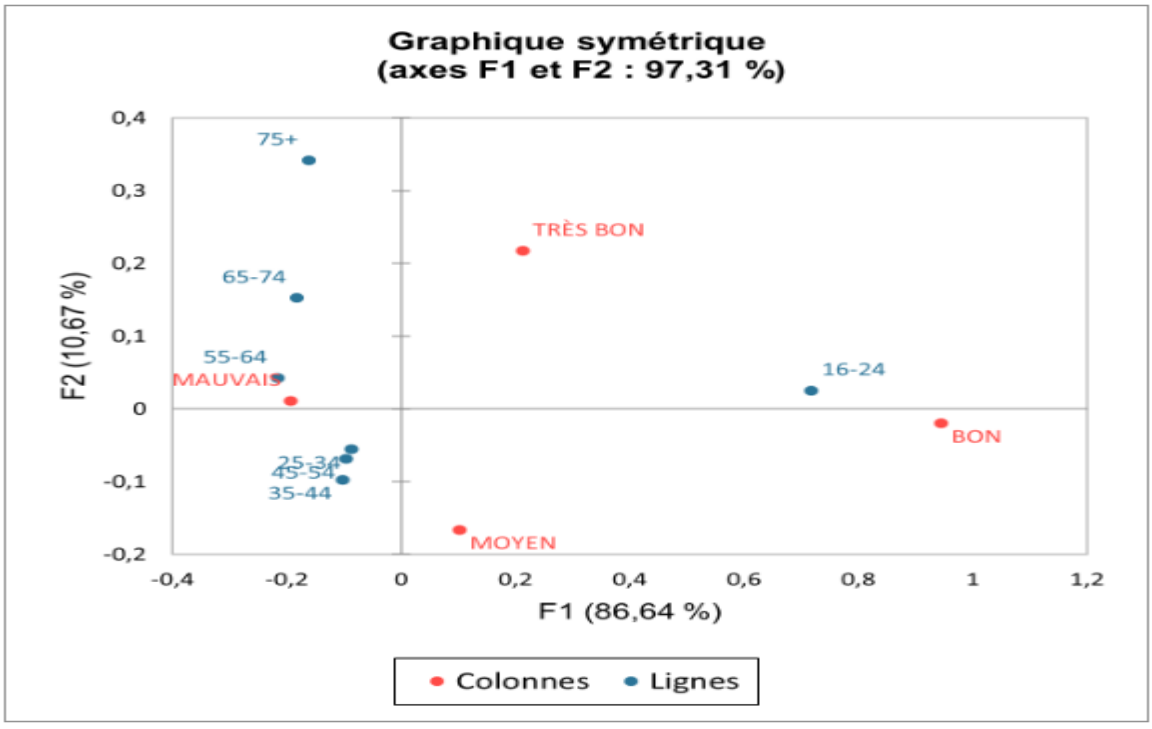


Figure 19 : graphique symétrique des lignes et colonnes.

Des **ellipses de confianc**e peuvent être ajoutées sur les graphiques symétriques des lignes ou des colonnes, comme illustré sur le graphique symétrique des lignes ci-dessous. Si l'origine se trouve dans l'ellipse d'une catégorie donnée, cette catégorie ne contribue pas à la dépendance entre les variables.

Dans notre exemple, les ellipses confirment que les groupes d'âge 25-34, 34-45 et 45-54 ne contribuent pas à la dépendance entre les variables. Le groupe des 16-24 ans contribuent à la dépendance entre les variables.

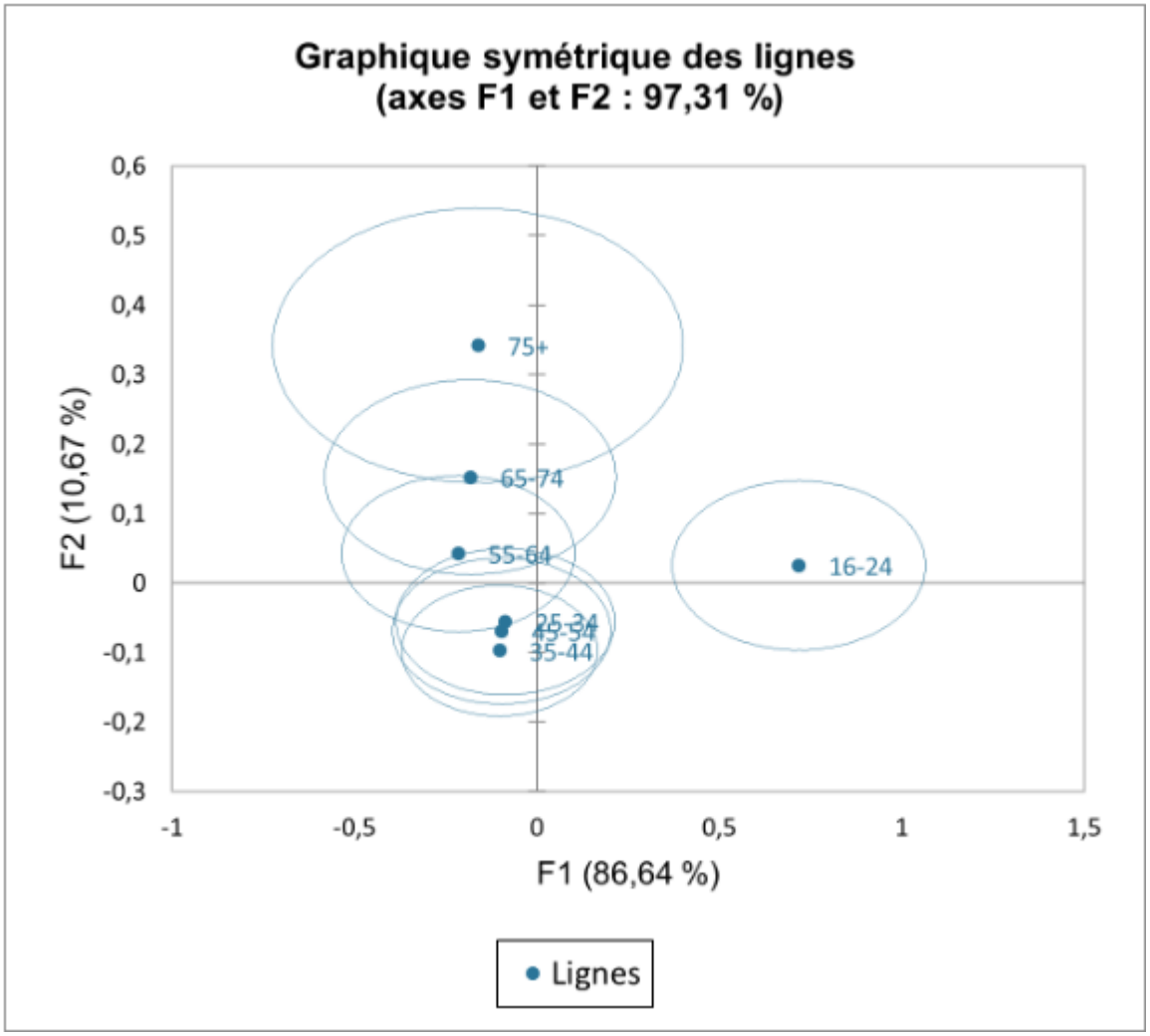


Figure 20 : graphique symétrique des lignes.

Sur le **graphique asymétrique des lignes**, les colonnes sont représentées dans l'espace des lignes (coordonnées standard pour les colonnes et coordonnées principales pour les lignes).

Inversement, le graphique asymétrique des colonnes correspond aux lignes représentées dans l'espace des colonnes. Les distances entre lignes et colonnes peuvent être interprétées en projetant les points-lignes sur les vecteurs-colonnes.

Le choix de la représentation dans l'espace des lignes ou l'espace des colonnes dépend de l'interprétation appropriée.

Dans notre exemple, nous choisissons d'interpréter les groupes d'âge dans l'espace des niveaux d'appréciation. La première dimension oppose « BON » à « MAUVAIS ». Le groupe des 16-24 ans comprend une proportion plus grande de « BON » par rapport aux proportions de « BON » dans les autres tranches d'âge. Cependant, cela ne signifie pas que la qualification « BON » a la plus grande proportion parmi les autres proportions au sein du groupe des 16-24 ans. Les profils lignes ne sont pas très différents du profil moyen (points proches de l'origine).

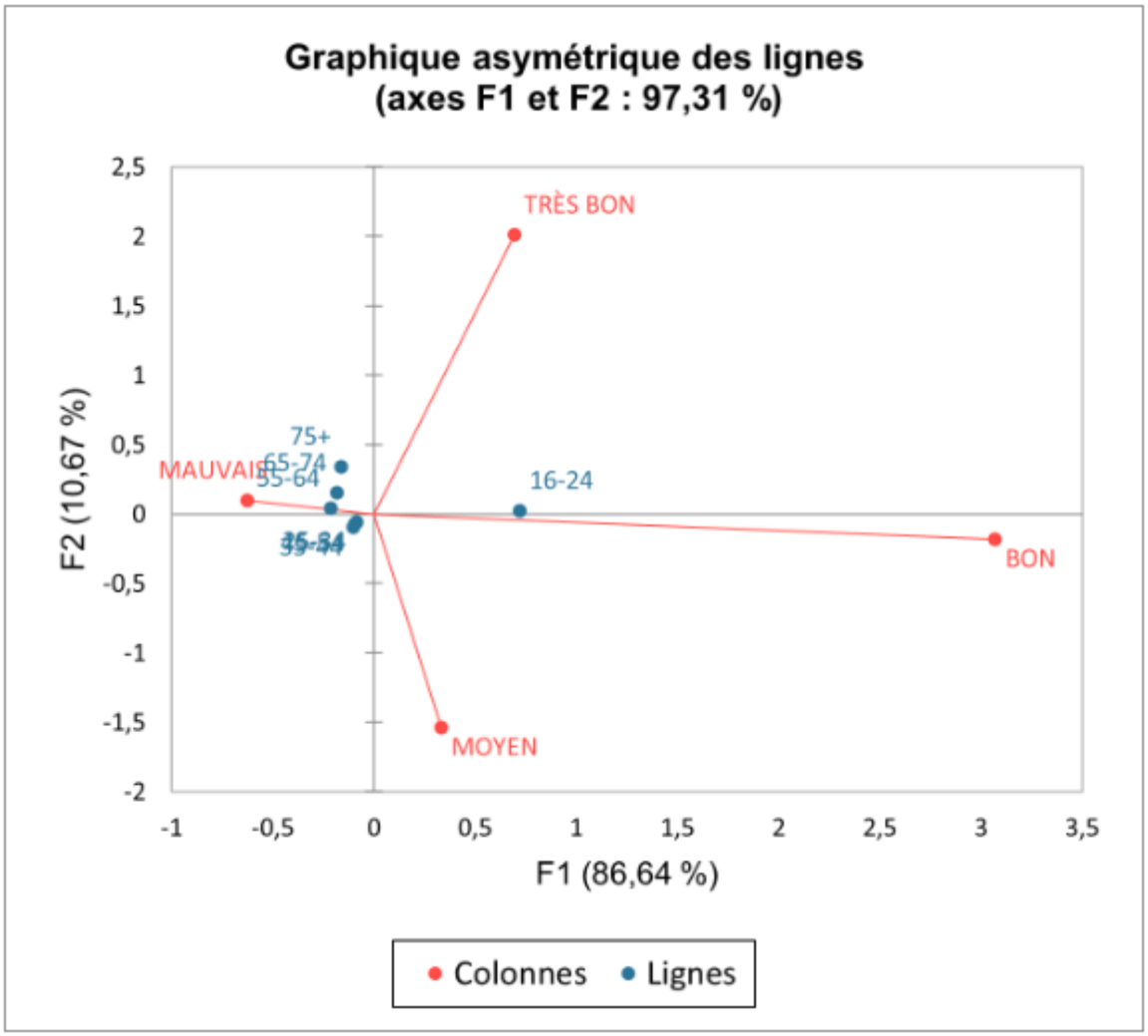


Figure 21 : graphique asymétrique des lignes.

Les **coordonnées de contribution des lignes et des colonnes** sont ensuite affichées. Les coordonnées de contribution sont obtenues en divisant les coordonnées standard par la racine carrée de la masse de la catégorie donnée.

Sur le **biplot de contribution des lignes**, les lignes sont en coordonnées de contribution et les colonnes sont en coordonnées principales, et inversement pour le biplot de contribution des colonnes. Sur le biplot des contributions des lignes (respectivement des colonnes), les distances des points lignes (respectivement colonnes) à l'origine sont liées à leur contribution au graphique.

Dans notre exemple, sur le biplot de contribution des lignes, les positions des points des lignes sont inchangées par rapport au graphique asymétrique. Les points colonnes sont plus proches de l'origine (voir les échelles des deux représentations).

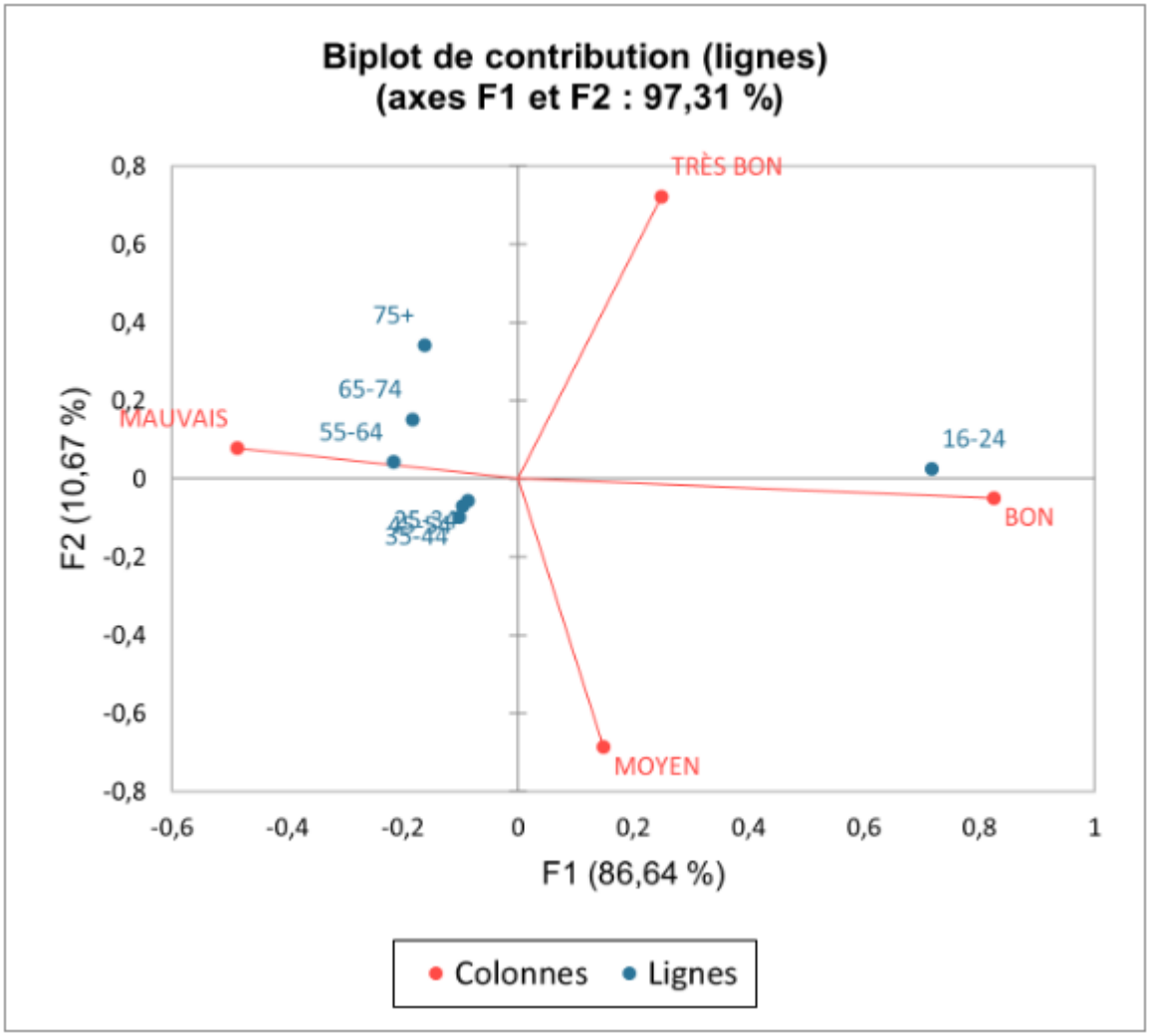


Figure 22: biplot de contribution des lignes.

CHAPITRE 06 : Classifications hiérarchique ascendante et nuées dynamiques

1. Définitions

Classifier, c'est regrouper entre eux des objets similaires selon tel ou tel critère. Les diverses techniques de classification (ou d'"analyse typologique", de "taxonomie", ou "taxinomie" ou encore "analyse en clusters" (amas)) visent toutes à répartir n individus, caractérisés par p variables X1, X2, ..., Xp en un certain nombre m de sous-groupes aussi homogènes que possible.

Les classifications hiérarchiques : pour un niveau de précision donné, deux individus peuvent être confondus dans un même groupe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront à deux sous-groupes différents. Le résultat d'une classification hiérarchique n'est pas une partition de l'ensemble des individus. C'est une hiérarchie de classes telles que : - toute classe est non vide - tout individu appartient à une (et même plusieurs) classes - deux classes distinctes sont disjointes, ou vérifient une relation d'inclusion (l'une d'elles est incluse dans l'autre) - toute classe est la réunion des classes qui sont incluses dans elle.

2. Principes De La Classification Ascendante Hiérarchique

La classification ascendante hiérarchique (CAH) est une méthode de classification itérative dont le principe est simple.

Les *matrices d'association* sont nécessaires pour utiliser les méthodes de groupement. Le *groupement* n'est pas une méthode statistique en tant que telle puisqu'elle ne teste pas d'hypothèse, mais permet de déceler des *structures* dans les données en partitionnant soit les objets, soit les descripteurs.

1. On commence par calculer la dissimilarité entre les N objets.
2. Puis on regroupe les deux objets dont le regroupement minimise un critère d'agrégation donné, créant ainsi une classe comprenant ces deux objets.
3. On calcule ensuite la dissimilarité entre cette classe et les N-2 autres objets en utilisant le critère d'agrégation. Puis on regroupe les deux objets ou classes d'objets dont le regroupement minimise le critère d'agrégation.

On continue ainsi jusqu'à ce que tous les objets soient regroupés.

Ces regroupements successifs produisent un arbre binaire de classification (dendrogramme), dont la racine correspond à la classe regroupant l'ensemble des individus. Ce dendrogramme représente une hiérarchie de partitions. On peut alors choisir une partition en tronquant l'arbre à un niveau donné, le niveau dépendant soit des contraintes de l'utilisateur (l'utilisateur sait combien de classes il veut obtenir), soit de critères plus objectifs.

3. Méthodes De Groupement

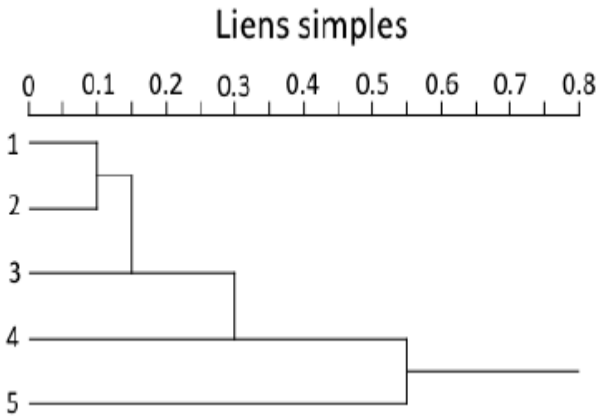
Il y a plusieurs familles de méthodes de groupements, mais nous présenterons uniquement un aperçu de trois méthodes :

- le groupement agglomératif, et hiérarchique à liens simples
- le groupement agglomératif hiérarchique à liens complets

- méthode.de Ward

Dans les méthodes hiérarchique, les éléments des petits ensembles se regroupent en groupes plus vastes de rang supérieur, et ainsi de suite (par exemple : espèces, genres, familles, ordre). Avant de faire le groupement, il faut créer une matrice d'association entre les objets. La matrice d'association est premièrement classée en ordre croissant de distances. Ensuite, les groupes sont formés de manière hiérarchique selon les critères spécifiques à chaque méthode.

Prenons un exemple tout simple d'une matrice de distances euclidiennes entre 5 objets dont on a ordonné les distances en ordre croissant Z



	2	3	4	5
1	0.10	0.15	0.40	0.80
2		0.60	0.35	0.70
3			0.30	0.65
4				0.75

Dist	Paire
0.10	1-2
0.15	1-3
0.30	3-4
0.35	2-4
0.40	1-4
0.65	3-5
0.70	2-5
0.75	4-5
0.80	1-5

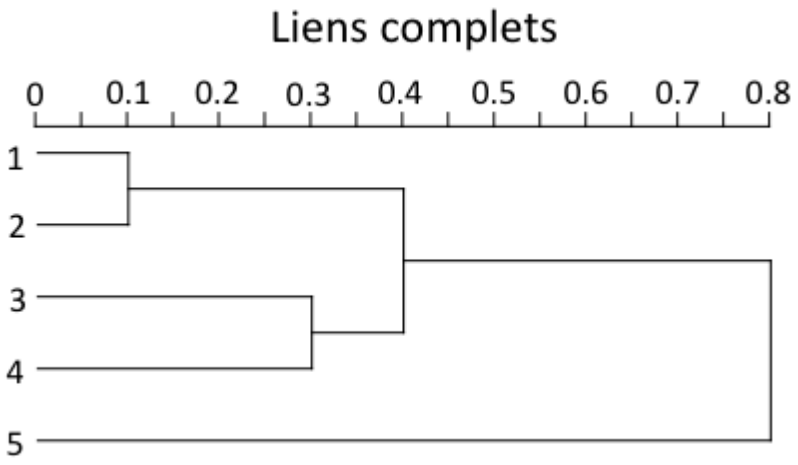


Figure 23 : Dendrogramme de Classification

3.1 Groupement agglomératif, et hiérarchique à liens simples

Pour le *groupement agglomératif à liens simples*, les deux objets les plus proches se regroupent en premier. Ensuite, un deuxième groupe est formé à partir des deux objets les plus proches suivant (il se peut que ce soit deux objets différents, ou bien un objet et le groupe formé précédemment), et ainsi de suite. Cette méthode forme généralement de longues chaînes de groupes (dans l'exemple ci-haut, les objets 1 à 5 se regroupent successivement).

3.2 Groupement agglomératif, et hiérarchique à liens complexe

À l'inverse, pour le *groupement agglomératif à liens* complexe, un objet se regroupe à un autre objet/groupe seulement lorsqu'il est aussi *complet* lié à l'élément le plus éloigné de ce groupe. Ainsi, quand deux groupes fusionnent, tous les éléments des deux groupes sont liés à la distance considérée (ci-haut, le groupe 3-4 ne se lie au groupe 1-2 qu'à la distance à laquelle tous les autres éléments sont déjà liés). C'est pour cette raison que le groupement à liens complets forme généralement plusieurs petits groupes séparés et qu'elle peut être plus appropriée pour relever des contrastes ou des discontinuités dans les données.

Comparons ces deux méthodes en utilisant les données d'abondances de poissons de la rivière Doubs. Les données d'abondances ont été au préalable transformées par la méthode *Hellinger*. Puisque les fonctions de groupement requièrent une matrice de distances, la première étape sera de générer une *matrice de distances Hellinger*. *Est-ce que les deux dendrogrammes sont très différents?*

On remarque que pour le groupement à liens simple, plusieurs objets s'enchaînent (par exemple les sites 19, 29, 30, 20, 26, etc.) alors que des groupes plus distincts peuvent être observés dans le groupement à liens complets.

3.3 Méthode.de Ward

La méthode de Ward diffère légèrement des deux méthodes précédentes. Le critère utilisé est la méthode des moindres carrés (comme dans les modèles linéaires). Ainsi, des objets/groupes fusionnent de façon à minimise la variance intragroupes. Pour débuter, chaque objet est considéré comme un groupe. À chaque étape, la paire de groupes à fusionner est celle qui résulte à la plus petite augmentation de la somme des carrés des écarts intra-groupes.

Le dendrogramme produit par défaut montre les distances au carré. Afin de comparer ce dendrogramme à celui du groupement à liens simples et à liens complets, il faut calculer la racine carrée des distances.

Les groupements générés par la méthode de *Ward* ont tendance à être plus sphériques et à contenir des quantités plus similaires d'objets.

4. Quelle méthode choisir?

Le choix de la bonne mesure d'association et de la bonne méthode de groupement dépend de l'objectif. Qu'est-ce qu'il est plus intéressant de démontrer : des gradients? Des contrastes? Il est également important de tenir en compte les propriétés de la méthode utilisée dans l'interprétation des résultats.

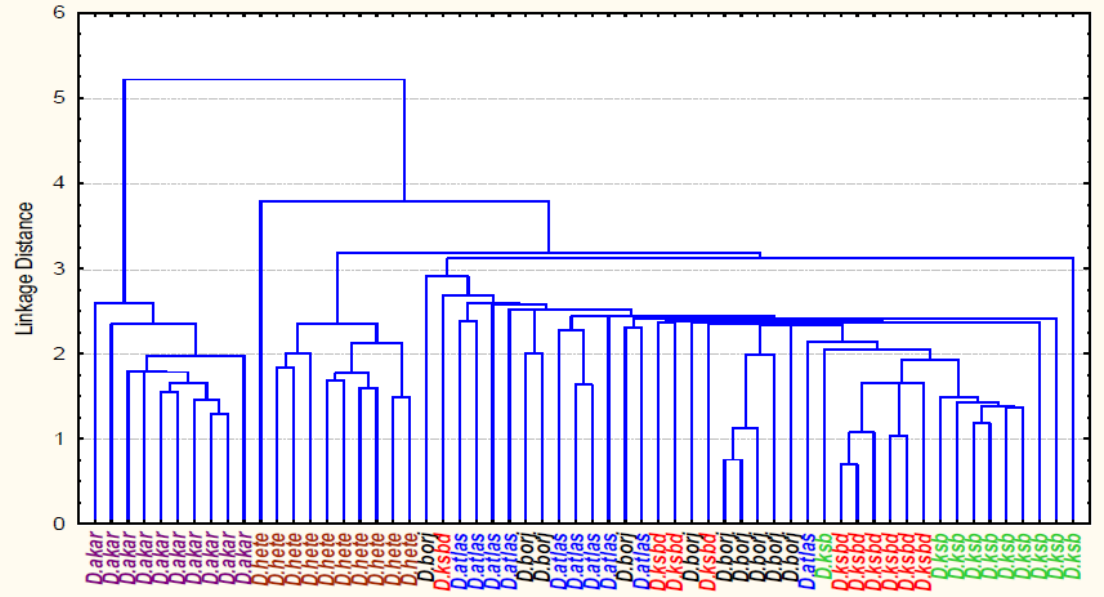
Si plus d'une méthode semble adéquate pour répondre à une question biologique, comparer les dendrogrammes serait une bonne option. Encore une fois, le groupement n'est pas une analyse statistique, mais il est possible de tester les résultats et d'identifier des partitions ayant un sens biologique. Il est également possible de déterminer le nombre de groupes optimal et de performer des tests statistiques sur les résultats. Les méthodes de groupement peuvent aussi être combinées à une ordination pour distinguer des groupes de sites.

Analyse hiérarchique

La matrice de cette analyse est représentée par un tableau de 60 lignes correspond à 06 espèces (*D. ksibii*, *D. ksibiodes*, *D. borjensus*, *D. akaraicus* et *D. atlasensis*) avec 10 individus pour chaque espèces et 16 colonnes (Variables morpho métriques: longueur du corps, largeur du corps et du hapteur, longueur et la largeur de la barre médiane, longueur et la largeur de la barre ventrale, longueur de la garde, longueur de la manche, longueur du crochetons, longueur de la manche, distance entre la manche et lame, entre la manche et la pointe, entre la garde et la pointe, entre la garde et la lame, entre la lame et la pointe, entre la garde et la manche).

L'analyse canonique (chapitre 06) permet de regrouper les six espèces de Monogènes en trois groupes d'espèces. Cependant, la classification à posteriori enregistre 93,33% comme un taux global de bon classement et elle propose un reclassement de certains spécimens des espèces analysées.

Cette observation initiale s'avère être mauvaise, ce qui conduit à une classification hiérarchique Nous pouvons cependant remarquer que les classes des *D. akaraicus* et *D. heteromorphus* ont été correctement classifiés, la classe *D. ksibii* est presque intégralement correcte. En effet, la troisième classe formée par une confusion entre les espèces de *D. borjensis*, *D. atlasensis* et *D. ksibiodes* (Fig.35).



(***D.akar***: *D.akaraicuss*; ***D.hete***: *D.heteromorphus*; ***D.ksb***: *D.ksibii*; ***D.ksbd***: *D.ksibiodes*; ***D.borj***: *D.borjensis*)..

Figure 24: Classification hiérarchique des six espèces de *Dactylogyrus* (hapteur de type carpathicus).

1. Introduction

Pour étudier un phénomène biologique, chaque fois que cela est possible, il faut se mettre en situation d'expérimentation où l'expérimentateur a le contrôle sur l'administration d'un facteur.

Le plus souvent, on constitue un groupe avec le facteur étudié (par exemple un traitement médicamenteux), et un autre groupe sans le facteur. S'il y a une différence entre les deux groupes, et seulement dans ce cas d'étude, on peut parler de causalité car seul le facteur étudié varie entre les deux populations. Quelquefois, la situation de type expérimental n'est pas possible, et le facteur étudié est aléatoire : il est distribué au hasard, et l'expérimentateur ne peut pas le modifier. On est alors dans une situation d'observation. Par exemple : tabac et grossesse, transferts des nouveau-nés et facteurs et autres maladies.

On peut parler de facteurs de risque soumis à 3 conditions :
(1) le facteur doit être statistiquement associé à la maladie,
(2) sa présence doit précéder l'apparition de la maladie,
(3) l'association ne doit pas être due à une source d'erreur, ou au hasard.

De plus, une relation de causalité est toujours difficile à mettre en évidence, en tout cas jamais de manière simple.

Un facteur, pour être en relation de CAUSALITE, doit être soumis à plusieurs conditions :
(1) Il doit être présent avant ou au début de la maladie,
(2) l'association statistique doit être forte,
(3) l'association doit être constante dans toutes les études,
(4) l'association doit être spécifique (c'est-à-dire qu'il n'y a pas d'autres facteurs de causalité),
(5) l'association augmente avec l'exposition au facteur (dose-effet),
(6) les données scientifiques sont cohérentes avec l'association (effet biologique connu expérimentalement par exemple).

2. Les Tests Statistiques

On traduit la question biologique en hypothèses statistiques. Pour cela, il est nécessaire d'observer le problème avec une approche statistique et de décider des paramètres des échantillons à comparer (moyenne, variance…). On définit deux hypothèses :

Confronté à des phénomènes complexes et aléatoires, la prise de décision est difficile et les outils adaptés de la théorie des tests ont pour objet de guider les choix entre différentes alternatives. De façon générale, il s'agira de décider si des différences observées entre un modèle posé a priori et des observations sont significatives ou peuvent être considérées comme étant dues au simple effet du hasard consécutif aux aléas du tirage d'un échantillon.

Réaliser un test statistique consiste à mettre en œuvre une procédure permettant :

- De confronter une hypothèse avec la réalité, ou plus exactement, avec ce que l'on perçoit de la réalité à travers les observations à disposition ;
- De prendre une décision à la suite de cette confrontation. Si les problèmes traités par l'estimation (ponctuelle ou par intervalle de confiance) sont de type quantitatif, i.e. conduisent à un résultat numérique, ceux traités par les tests d'hypothèses sont d'ordre qualitatif, i.e. conduisent à une réponse du type rejet/acceptation de l'hypothèse statistique considérée.

Le test statistique est formulé dans **le but** de répondre à une problématique expérimentale ou clinique, comme par exemple :
- Est-ce que le médicament testé est-il efficace ?
- Les pièces sortant d'une machine sont-elles conformes ?

- Un nouveau mode de culture bactérienne est-il plus efficace ?
- Quels sont les gènes significativement différentiellement exprimés dans un tissu pathologique ? ...

Sont autant de questions auxquels des tests statistiques peuvent apporter des réponses sous **4 conditions :**

1. La question est posée de sorte qu'il n'y ait que 2 réponses possibles : oui/non,
2. Une expérimentation planifiée fournit des données relatives à cette question,
3. Les données sont considérées comme la réalisation de variables aléatoires décrites par un modèle statistique
4. La réponse à la question se traduit par l'acceptation ou le rejet d'une hypothèse (notée H_0) caractéristique du modèle précédent.

Notes

Dans ces conditions et avec une marge d'erreur plus ou moins bien contrôlée,
Accepter l'hypothèse, fait répondre **Non** à la question en considérant que les différences observées entre le modèle et la réalité des observations sont imputables au seul hasard.

Rejeter l'hypothèse fait **Oui** à la question : les différences sont jugées significatives car trop improbables ou invraisemblables.

3. Principe D'un Test D'hypothèses

Les tests d'hypothèse constituent un autre aspect important de l'inférence statistique. Le principe général d'un test d'hypothèse peut s'énoncer comme suit :

• On étudie une population dont les éléments possèdent un caractère (mesurable ou qualitatif) et dont la valeur du paramètre relative au caractère étudié est inconnue.
Une hypothèse est formulée sur la valeur du paramètre : cette formulation résulte de considérations théoriques, pratiques ou encore elle est simplement basée sur un pressentiment.

• On veut porter un jugement sur la base des résultats d'un échantillon prélevé de cette population.
Pour décider si l'hypothèse formulée est supportée ou non par les observations, il faut une méthode qui permettra de conclure si l'écart observé entre la valeur de la statistique obtenue dans l'échantillon et celle du paramètre spécifiée dans l'hypothèse est trop important pour être uniquement imputable au hasard de l'échantillonnage.

La construction d'un test d'hypothèse consiste en fait à déterminer entre quelles valeurs peut varier la variable aléatoire, en supposant l'hypothèse vraie, sur la seule considération du hasard de l'échantillonnage.

4. Définition des concepts utiles à l'élaboration des Tests d'hypothèse

➢ Hypothèse statistique

Une hypothèse statistique est un énoncé (une affirmation) concernant les caractéristiques (valeurs des paramètres, forme de la distribution des observations) d'une population.

➢ Test d'hypothèse

Un test d'hypothèse (ou test statistique) est une démarche qui a pour but de fournir une règle de décision permettant, sur la base de résultats d'échantillon, de faire un choix entre deux hypothèses statistiques.

➢ Hypothèse nulle (H0) et hypothèse alternative (H1)

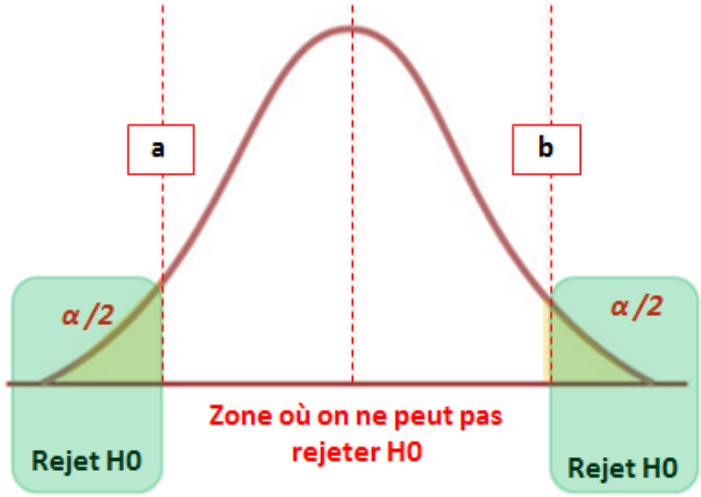
L'hypothèse selon laquelle on fixe à priori un paramètre de la population à une valeur particulière s'appelle l'hypothèse nulle et est notée H_0. N'importe quelle autre hypothèse qui diffère de l'hypothèse H_0 s'appelle l'hypothèse alternative (ou contre-hypothèse) et est notée H_1.

5. Seuil de signification du test

Le risque, consenti à l'avance et que nous notons α de rejeter à tort l'hypothèse nulle H_0 alors qu'elle est vraie, s'appelle le seuil de signification du test et s'énonce en probabilité ainsi :
$\alpha = P$ (rejeter H_0/ H_0 vraie=).

A ce seuil de signification, on fait correspondre sur la distribution d'échantillonnage de la statistique une région de rejet de l'hypothèse nulle (appelée également région critique). L'aire de cette région correspond à la probabilité **α.**

Si par exemple, on choisit α = 0.05, cela signifie que l'on admet d'avance que la variable d'échantillonnage peut prendre, dans 5% des cas, une valeur se situant dans la zone de rejet de H_0, bien que H_0 soit vraie et ceci uniquement d'après le hasard de l'échantillonnage. Sur la distribution d'échantillonnage correspondra aussi une région complémentaire, dite région d'acceptation de H_0 (ou région de non-rejet) de probabilité 1−α.



Soit V la valeur critique du test statistique

Si V ∈]a ; b [alors on ne peut pas rejeter l'hypothèse H0

Si V ∉] a ; b [alors on peut rejeter l'hypothèse H0

Où a et b sont des valeurs seuil à lire dans les tables des différents tests d'hypothèses

Figure 25 : Zones d' acceptations et de rejet des hypothèses

6. Les critères de décision des tests

➢ **Règle de décision 1 :**
Sous l'hypothèse « H_0 est vraie » et pour un seuil de signification α fixé (5% par défaut) :

- Si la valeur statistique S_{obs} calculée appartient à la région critique alors l'hypothèse H_0 est rejetée au risque d'erreur α et l'hypothèse H1 est acceptée
- Si la valeur statistique S_{obs} calculée n'appartient à la région critique alors l'hypothèse H_0 ne peut pas être rejetée.

Remarque : Le choix du niveau de signification ou risque α est lié aux conséquences pratiques de la décision, en général on choisira **α = 0,05 ; 0,01 ou 0,001**

- **Règle de décision 2 :**
  La probabilité critique α telle que $\mathbf{P\ (S \geq S_{obs}\ ) = \alpha_{obs}}$ est évaluée
- Si $\boldsymbol{\alpha_{obs} \geq \alpha}$, l'hypothèse H_0 est acceptée car le risque d'erreur de rejeter H_0 alors qu'elle est vraie est trop important
- Si $\boldsymbol{\alpha_{obs} < \alpha}$, l'hypothèse H_0 est rejetée car le risque d'erreur de rejeter H_0 alors qu'elle est vraie est très faible

7. Différents type de tests d'hypothèse

Avant d'appliquer tout test statistique, il s'agit de bien définir le problème posé. En effet, selon les hypothèses formulées, vous appliquerez soit un ***test bilatéral***, soit un ***test unilatéral.***

7.1 Test Bilatéral

Le test bilatéral s'applique quand vous cherchez une différence entre deux paramètres, ou entre un paramètre et une valeur donnée sans se préoccuper du signe ou du sens de la différence.
Dans ce cas, la zone de rejet de l'hypothèse principale se fait de part et d'autre de la distribution de référence.

7.2 Test Unilatéral

Un test unilatéral s'applique quand vous cherchez à savoir si un paramètre est supérieur (ou inférieur) à un autre ou à une valeur donnée.
La zone de rejet de l'hypothèse principale est située d'un seul côté de la distribution de probabilité de référence.

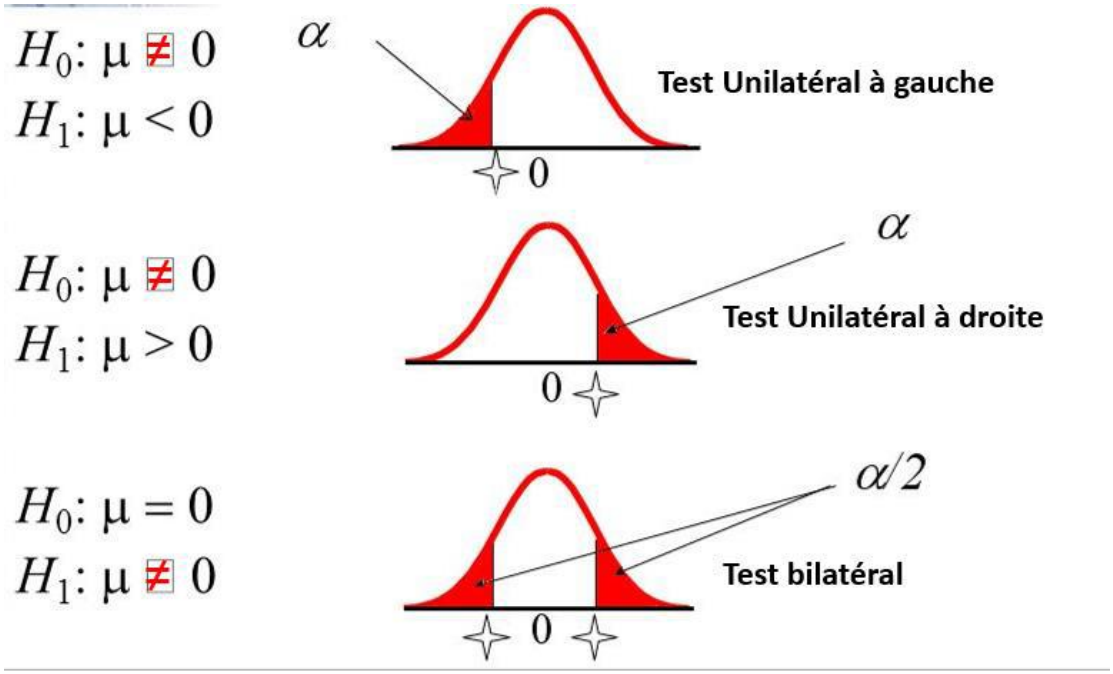


Figure 26 : Différents type de test d'hypothèse

8. Tests d'égalité de deux variances

8.1 Test F de Fisher (Echantillons indépendants)

Le test d'égalité de deux variances est relatif à l'hypothèse nulle :

$$\boldsymbol{H_0: \sigma_1^2 = \sigma_2^2}$$

Calcule :

$$F_{obs} = \frac{\hat{\sigma}^2_{max}}{\hat{\sigma}^2_{min}}$$

Lorsque les effectifs sont égaux $n_1 = n_2 = n$, nous avons la formule suivante :

$$F_{obs} = \frac{\hat{\sigma}^2_{max}}{\hat{\sigma}^2_{min}} = \frac{SCE_{max}}{SCE_{min}}$$

Avec :
F: variable de Fisher-Snedecor ;
$\hat{\sigma}$: La plus grande de ces deux variances
$\hat{\sigma}^2_{min}$: La plus petite de ces deux variances
Le nombre de degré de liberté (ddl)
$k_1 = (n_1 - 1)$ ddl $nombre\ de\ degré\ de\ liberté\ de\ la\ variance\ la\ plus\ grande\ \hat{\sigma}^2_{max}$
$k_2 = (n_2 - 1)$ $ddl\ le\ nombre\ de\ degré\ de\ liberté\ de\ la\ variance\ la\ plus\ petite\ \hat{\sigma}^2_{min}$
$\alpha = 0,05$

Règle de discision et interprétation

On rejette H_0 (hypothèse nulle) lorsque $F_{obs} \geq F_{1-\alpha/2} \Rightarrow RH_0$ donc il existe de différence significative entre les variances des deux populations, c.-à-d : $\hat{\sigma}_1 \neq \hat{\sigma}_2$

On accepte H_0 (hypothèse nulle) lorsque $F_{obs} < F_{1-\alpha/2} \Rightarrow AH_0$ donc il n'existe pas de différence significative entre les variances des deux populations, c.-à-d : $\hat{\sigma}_1 = \hat{\sigma}_2$

8.2 Echantillons non indépendants

Pour des échantillons :

- non indépendants,
- associés par paires,

L'hypothèse nulle (H_0) d'égalité des variances est la suivante :

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Calcul:

$$t_{obs} = \frac{|SCE_1 - SCE_2|\sqrt{n-2}}{2\sqrt{SCE_1.SCE_2 - SPE^2}}$$

SCE_1, SCE_2 et SPE^2 sont relative aux deux séries d'observation, nécessairement de même effectif

Avec: $\{\alpha = 0,05\ et\ (n-2)\}$

Règle de discision et interprétation

On **rejette H_0** (hypothèse nulle) lorsque $t_{obs} \geq t_{1-\alpha/2} \Rightarrow RH_0$ donc il existe de différence significative entre les variances des deux populations.

On **accepte** $\boldsymbol{H_0}$ (hypothèse nulle) lorsque $\boldsymbol{t_{obs} < t_{1-\alpha/2}}$ $\Rightarrow \boldsymbol{AH_0}$ donc il n'existe pas de différence significative entre les variances des deux populations.

9. Tests d'égalité de plusieurs variances

Deux tests sont couramment utilisés en vue de contrôler **l'égalité des variances de plusieurs populations**, à partir **d'échantillons indépendants** :

9.1 Test de HARTLEY

Propriétés du test

- Le test de HARTLEY permet de vérifier rapidement le test d'égalité des variances : $\boldsymbol{H}$
- Il est basé sur la comparaison du rapport des deux variances estimées extrêmes
- C'est un test très rapide qui s'applique en principe qu'a des échantillons de même effectif.
- Lorsque les effectifs des échantillons sont constants et égale à $\boldsymbol{n}$: $\boldsymbol{n_1 = n_2 = n_p = \cdots n}$

$$\boldsymbol{H_{obs} = \frac{\hat{\sigma}^2_{max}}{\hat{\sigma}^2_{min}} = \frac{SCE_{max}}{SCE_{min}}}$$

Avec :

$\boldsymbol{\alpha}$ = 0, 05;
$\boldsymbol{p}$: $Nbr\ de\ population$;
$\boldsymbol{k}$ = $(n-1)$ $\boldsymbol{ddl}$

Règle de discision et interprétation

Acceptation d'H_0 (hypothèse nulle) lorsque $H_{obs} < H_{1-\alpha}$ $\Rightarrow AH_0$ donc il n'existe pas de différence significative entre les variances des $\boldsymbol{p}$ populations.

Rejet d'H_0 (hypothèse nulle) lorsque $H_{obs} \geq H_{1-\alpha}$ $\Rightarrow RH_0$ donc il existe de différence significative entre les variances des $\boldsymbol{p}$ populations.

9.2 Test de BARTLETT

Propriétés du test :

- **Ce test nécessite des calculs relativement longs**, mais s'applique indifférâmes à **des échantillons d'effectifs égaux** ou **inégaux.**
- Pour $\boldsymbol{p}$ population

$$\boldsymbol{H_0: \sigma_1^2 = \sigma_2^2 = \cdots \sigma_P^2}$$

1/ Calculer pour chaque échantillon la somme des carrés des écarts ($\boldsymbol{SCE}$) et la variance estimé $\boldsymbol{\hat{\sigma}^2}$

$$\boldsymbol{SCE = \sum_{j=1}^{n_i} x_{ij}^2 - \frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}^2\right)^2}$$

$$\boldsymbol{\hat{\sigma}_i^2 = \frac{SCE_i}{n_i - 1}}$$

2/ Calculer aussi la SCE_g et $\hat{\sigma_g}^2$ relative à l'ensemble des observations globale :

$$SCE_g = \sum_{i=1}^{p} SCE_i$$

$$\hat{\sigma}_g^2 = \frac{SCE_g}{n. - p}$$

$$n. = \sum_{i=1}^{p} n_i$$

p: *Nbre des échantillons*
$n.$: *la somme des effectifs de tous les échantillons*

L'expression relative à ce test est alors :

$$x^2 = \frac{[n. - P]\log_e \hat{\sigma}_g^2 - \sum[(n_i - 1)\log_e \hat{\sigma}_i^2]}{1 + \frac{1}{3(P-1)}(\sum \frac{1}{n_i - 1} - \frac{1}{n. - P})}$$

Avec : $\{\alpha = 0{,}05 \; et \; (p-1)ddl\}$

Règle de discision et interprétation

On rejette H_0 (hypothèse nulle) lorsque $\chi_{obs2} \geq \chi^2_{1-\alpha}$ $\Rightarrow$ RH_0 donc il existe de différence significative entre les variances des p populations.

On accepte H_0 (hypothèse nulle) lorsque $\chi^2_{obs} < \chi_1{}^2_{-\alpha}$ $\Rightarrow$ AH_0 donc il n'existe pas de différence significative entre les variances des p populations.

Remarques

- Le test de BARTLETT est très sensible à la non normalité des populations parent quel que soit les effectifs des échantillons.

- Ce test n'est pas très robuste par rapport à la non normalité ; mieux vaut utiliser le test de LEVENE ou de BROWN-FORSYTHE.

- L'approximation χ^2 n'est satisfaisante que si les effectifs n_i sont au moins égaux à 4, et si le nombre de population p n'est pas trop élevé par rapport aux effectifs. Ce test ne permet donc pas de comparer les variances d'un grand nombre de petit échantillon.

- Pour deux échantillons le test de HARTLEY est strictement équivalent de test F de Fisher, sauf dans ce cas ce test est moins sensible que le test de BARTLETT car il le fait intervenir explicitement (d'une façon claire) que des valeurs observées de deux échantillons d'autre part comme le test de BARTLETT, le test de HARTLEY est très sensible à la non normalité des populations.

10. Tests paramétriques (Méthodes statistiques relative à une ou à deux moyennes)

Ces méthodes sont des méthodes d'inférence statistique relatives aux **moyennes** pour **une** ou **deux populations.**

Ces méthodes sont semis à 3 conditions qui sont les suivantes :

- **Population normale ;**
- **Des échantillons aléatoires et simples ;**
- **L'égalité des variances.**

10.1 Test de Student « Test de comparaison de deux moyennes »

Ce test permet de comparer :

- Les moyennes de deux échantillons indépendants
- Une moyenne d'un échantillon a une valeur donnée
- Les moyennes de deux échantillons appariées

a) Comparaison des moyennes de deux échantillons indépendants

▶ **CAS des populations de même variance**

- Ce test est appeler t de STUDENT ou test de STUDENT-FISHER
- On supposant satisfaite les conditions présente et on admettant que les échantillons sont indépendant et que les populations sont de même variance en peut donner le test d'hypothèse suivants :

$\boldsymbol{H_0}$: $m_1 = m_2$

Calcule:

- **Quant $n_1 \neq n_2$**

$$t_{obs} = \frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{SCE_1 + SCE_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Avec :

$\alpha = 0{,}05$ **et** $k = (n_1 + n_2 - 2)$ **ddl ou** $2(n-1)$ **ddl**

- **Quant $n_1 = n_2$**

$$t_{obs} = \frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{SCE_T + SCE_{Sp}}{n(n-1)}}}$$

Avec :

$\alpha = 0{,}05$ **et** $k = n(n-1)$**ddl**

Règle de discision et interprétation

On rejette $\boldsymbol{H_0}$ (hypothèse nulle) lorsque $t_{obs} \geq t_{1-\alpha/2}$ $\Rightarrow m_1 \neq m_2$ donc il existe de différence significative entre les moyennes des deux populations (1 et 2).

On accepte $\boldsymbol{H_0}$ (hypothèse nulle) lorsque $t_{obs} < t_{1-\alpha/2}$ $\Rightarrow m_1 = m_2$ donc il n'existe pas de différence significative entre les moyennes des deux populations (1 et 2).

Intervalle de confiance

Quand on rejette l'hypothèse nulle, on calcule aussi l'intervalle de confiance de la différence des moyennes des deux populations.

- **Quant $n_1 \neq n_2$**

$$(\overline{x_1} - \overline{x_2}) \pm t_{1-\alpha/2} \sqrt{\frac{SCE_1 + SCE_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Avec:
$\alpha = 0{,}05$ ***et*** $k = (n_1 + n_2 - 2)$ ***ddl ou*** $2(n-1)$ ***ddl***

- **Quant $n_1 = n_2$**

$$(\overline{x_1} - \overline{x_2}) \pm t_{1-\alpha/2} \sqrt{\frac{SCE_1 + SCE_2}{n(n-1)}}$$

Avec:
$\alpha = 0,05$ ***et*** $k = (n-1)$

▶ **Cas des populations de variances hétérogènes**

Plusieurs auteurs ont montré que l'hypothèse de normalité est secondaire dans le test d'égalité de 2 moyennes

De même l'hypothèse d'égalité des variances n'est pas fondamentale lorsque les effectifs sont égaux ($n_1 = n_2$).

Quand le test n'est pas sensible à la non normalité et les inégalités des variances on dit alors que le test est robuste par contre lorsque ($\neq n_2$) il est absolument indispensable de s'assurer de l'égalité des variances si cette hypothèse n'est pas vérifier il faut alors procéder à une transformation de variable (à titre d'exemple : transformation logarithmique ; transformation racine carrée…etc) destiné a stabilisé les variances et utilisé ensuite le test *t* de STUDENT décrire si dessus.

b) Échantillon non indépendant (associés par paires ou par couples)

En ce qui concerne les échantillons non indépendants, nous envisagerons successivement le test t par paires et la détermination des limites de confiance de la différence des moyennes.

Les tests relatifs aux échantillons non indépendants, ou associés par couples sont basés sur le calcul des différences entre les couples d'observations.

Dans la mesure où tester l'égalité des moyennes des deux populations est alors strictement équivalent à tester la nullité de la moyenne des différences ces tests sont évidemment étroitement associés aux tests de conformité.

Exemple

Les mêmes individus sont soumis à 2 méthodes différentes, à 2 techniques différentes mais le but est de comparer ces 2 méthodes.

Les conditions d'applications des tests sont alors :

- **Le caractère aléatoire et simple des échantillons.**

- **La normalité de la population des différences.**

On pose l'hypothèse nulle suivante :

$\boldsymbol{H_0}$: $m_1 = m_2$

Calculer la différence $\boldsymbol{d_i} = (\boldsymbol{x_i} - \boldsymbol{x_j})$

Calcule:

Ce test est appeler t de STUDENT par pairs ou par couples

$$\boldsymbol{t_{obs}} = \frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{SCE_d}{n(n-1)}}} = \frac{|\overline{d}|}{\sqrt{\frac{SCE_d}{n(n-1)}}}$$

Avec :
$\boldsymbol{\alpha}$ = 0,05 $\boldsymbol{et}$ $\boldsymbol{k}$ = ($\boldsymbol{n}$ – **1**) $\boldsymbol{ddl}$

Règle de discision et interprétation

On rejette $\boldsymbol{H_0}$ (hypothèse nulle) lorsque $\boldsymbol{t_{obs}} \geq \boldsymbol{t_{1-\alpha/2}}$ $\Rightarrow$ $\boldsymbol{m_1} \neq \boldsymbol{m_2}$ donc il existe de différence significative entre les moyennes des deux couples (1 et 2).

On accepte $\boldsymbol{H_0}$ (hypothèse nulle) lorsque $\boldsymbol{t_{obs}} < \boldsymbol{t_{1-\alpha/2}}$ $\Rightarrow$ $\boldsymbol{m_1} = \boldsymbol{m_2}$ donc il n'existe pas de différence significative entre les moyennes des deux couples (1 et 2).

Intervalle de confiance

Quand on rejette l'hypothèse nulle, en calcule l'intervalle de confiance de la différence des moyennes des deux paires d'observation.

$$\overline{\boldsymbol{d}} \pm \boldsymbol{t_{1-\alpha/2}} \sqrt{\frac{SCE_d}{n(n-1)}}$$

Avec :

$\boldsymbol{\alpha}$ = 0,05 $\boldsymbol{et}$ $\boldsymbol{k}$ = (n – 1) $\boldsymbol{ddl}$

c) Test de conformité d'une moyenne

Les tests de conformité sont destinés à vérifier si un échantillon peut être considéré comme extrait d'une population donnée ou représentatif de cette population, vis-à-vis d'un paramètre comme la moyenne, la variance ou la fréquence observée.

Ce test a pour but de vérifier si la moyenne $\boldsymbol{m}$ d'une population est ou n'est pas égale à une valeur théorique $\boldsymbol{m_0}$.

Exemple :

Une usine veut vérifier le bon fonctionnement de ces machines car l'usure des machines peut impliquer une déviation aux normes imposées. Nous tirons aléatoirement un certain nombre d'éléments de la production, nous calculons la moyenne et nous comparons celle-ci avec la norme imposée. Les hypothèses à tester sont :

- hypothèse nulle : H0: $\mu = \mu_0$

➢ hypothèse alternative :
o H1: μ > μ0 (test unilatéral à droite)
o H1: μ < μ0 (test unilatéral à gauche)
o H1: μ □μ0 (test bilatéral symétrique)

Calcule
La formulation de l'hypothèse nulle est comme suite :

$$\boldsymbol{H_0}: m = m_0 \quad \Rightarrow \quad m - m_0 = \mathbf{0}$$

Il est également basé sur les distributions t de STUDENT. Pour réaliser ce test, on doit calculer la quantité suivante :

$$\boldsymbol{t_{obs}} = \frac{|\bar{x} - \boldsymbol{m_0}|}{\hat{\sigma}/\sqrt{n}} = \frac{|\bar{x} - \boldsymbol{m_0}|}{\sqrt{\frac{SCE}{n(n-1)}}}$$

Avec : $= 0{,}05$ $\boldsymbol{et}$ $(n-1)$

Règle de discision et interprétation

On rejette $\boldsymbol{H_0}$ (hypothèse nulle) lorsque $\boldsymbol{t_{obs} \geq t_{1-\alpha/2}} \Rightarrow \boldsymbol{m \neq m_0}$ donc il existe de différence significative entre la moyenne de la population m et la moyenne de référence m_0.

On accepte $\boldsymbol{H_0}$ (hypothèse nulle) lorsque $\boldsymbol{t_{obs} < t_{1-\alpha/2}} \Rightarrow \boldsymbol{m = m_0}$ donc il n'existe pas de différence significative entre la moyenne de la population m et la moyenne de référence m_0.

11. Tests non paramétriques

Si l'on ne peut pas considérer que la loi de $\boldsymbol{X}$ est normale ou bien si les données ne peuvent faire l'objet d'opérations arithmétiques (données qualitatives ou comptages) on adoptera des tests non paramétriques (le cas où la variable est une mesure mais de distribution non normale est particulier (test de randomisation).

Les tests non paramétriques sont des tests de comparaison des médianes

Les tests non paramétriques nécessitent seulement les rangs dans la liste ordonnée de toutes les valeurs.

11.1 Test de MANN-WHITNEY (Deux échantillons indépendants)

Propriétés

- Ce test est utilisé le plus souvent pour comparer deux populations à partir d'un échantillon indépendant,
- Cas des données continues,
- Un test des rangs ou la sommes des rangs

Principe

- Son principe est de classer l'ensemble des observations des deux échantillons par ordre croissant,
- Déterminer les rangs de chacune d'entre elles dans cet ensemble,
- Calculer la somme des rangs relative, par exemple, au premier échantillon.

- Nous désignerons cette somme par $\boldsymbol{X_1}$.

Statistique du test

On peut démontrer que l'hypothèse nulle et la quantité sont les suivantes :

$$\boldsymbol{H_0}: \hat{\boldsymbol{x}}_1 = \hat{\boldsymbol{x}}_2$$

$$\boldsymbol{U_{obs}} = \frac{|\boldsymbol{X_1} - \boldsymbol{n_1}(\boldsymbol{n_1} + \boldsymbol{n_2} + \boldsymbol{1})/\boldsymbol{2}|}{\sqrt{\boldsymbol{n_1}.\boldsymbol{n_2}(\boldsymbol{n_1} + \boldsymbol{n_2} + \boldsymbol{1})/\boldsymbol{12}}}$$

Règle de décision et interprétation

On rejette H_0 (hypothèse nulle) lorsque $U_{obs} \geq U_{1-\alpha/2}$ $\Rightarrow \hat{x}_1 \neq \hat{x}_2$ donc il existe de différence significative entre les médianes des deux populations.

On accepte H_0 (hypothèse nulle) lorsque $\boldsymbol{U_{obs}} < \boldsymbol{U_{1-\alpha/2}}$ $\Rightarrow \boldsymbol{\hat{x}_1} = \boldsymbol{\hat{x}_2}$ donc il n'existe pas de différence entre les médianes des deux populations.

11.2 Test de WILCOXON (Deux échantillons appariés ou non-indépendants)

Propriétés

- Le test de **WILCOXON** est un test non paramétrique le plus courant,
- C'est un test de comparaison de deux échantillons non indépendants, Cas des données continues.
- Un test des rangs
- Parfois appelé test des rangs par paires ou test des signes.
- La réalisation de ce test nécessite le calcul des différences entre couples d'observations, la détermination des rangs relatifs aux différences négatives ou aux différences positives.

Principe

Il est généralement suggéré d'éliminer de l'analyse les éventuelles égales à zéro et de réduire en conséquence la valeur de n, le test ne tenant compte que des différences strictement négatives ou positives.

Statistique du test

$$\boldsymbol{W_{obs}} = \frac{|\boldsymbol{X_-} - \boldsymbol{n}(\boldsymbol{n} + \boldsymbol{1})/\boldsymbol{4}|}{\sqrt{\boldsymbol{n}(\boldsymbol{n} + \boldsymbol{1})(\boldsymbol{2n} + \boldsymbol{1})/\boldsymbol{24}}}$$

Avec : $\boldsymbol{X_-}$ la somme des rangs correspondant aux différences négatives,

Règle de décision et interprétation

On accepte H_0 si $\boldsymbol{W_{obs}} < \boldsymbol{W_{1-\alpha/2}}$ $\Rightarrow \boldsymbol{\hat{x}_1} = \boldsymbol{\hat{x}_2}$ donc il n'existe pas de différence entre les médianes des deux populations.

On rejette H_0 si $\boldsymbol{W_{obs}} \geq \boldsymbol{W_{1-\alpha/2}}$ $\Rightarrow \boldsymbol{\hat{x}_1} \neq \boldsymbol{\hat{x}_2}$ donc il existe de différence significative entre les médianes des deux populations.

11.3 Test de KRUSKALL-WALLIS (Plus de deux échantillons indépendants)

Propriétés

- Ce test non paramétrique utilisé le plus couramment pour comparer $\boldsymbol{p}$ populations

Principe

- Le test de KRUSKAL-WALLIS nécessite le classement de l'ensemble des $\boldsymbol{n}$.
- Observations : le calcul des sommes des rangs $\boldsymbol{X_i.}$, relatives aux $\boldsymbol{p}$ échantillons.

Statistique du test

Ce test nécessite la détermination de la quantité :

$$x^2_{obs} = \frac{12}{n.(n.+1)} \sum \left(\frac{X_{i.}^2}{n_i}\right) - 3(n.+1)$$

Règle de décision et interprétation

On accepte H_0 (l'hypothèse nulle) si $\chi^2_{obs} < \chi^2_{1-\alpha} \Rightarrow \quad \hat{x_1} = \hat{x}_2 = \cdots \hat{x_p}$ donc il n'existe pas de différence significative entre les médianes des $\boldsymbol{p}$ populations.

On rejette H_0 (l'hypothèse nulle) si $\chi^2_{obs} \geq \chi^2_{1-\alpha} \Rightarrow \quad \hat{x_1} \neq \hat{x}_2 \neq \cdots \hat{x_p}$ donc il existe de différence significative entre les médianes des $\boldsymbol{p}$ populations.

Avec: $\boldsymbol{k} = (p-1)\ \boldsymbol{ddl}$

Remarque

Face à de petits échantillons et une variable aléatoire de loi inconnue (ou connue pour être éloignée de la loi normale) on peut quand même deux moyennes observées en utilisant le test U de MANN-WHITNEY (test W de WILCOXON), équivalent non paramétrique du test $\boldsymbol{t}$ de STUDENT.

Dans le cas de plusieurs moyennes à comparer simultanément, on peut utiliser le test H de KRUSKALL-WALLIS, équivalent non paramétrique de l'ANOVA.

L'utilisation d'un test non paramétrique s'accompagnant d'une légère perte de puissance, ces tests ne sont utilisés que dans le cas où on ne peut utiliser les tests paramétriques.

12. Test de Khi deux

On utilise un test de khi2 si notre recherche comporte deux groupes (ou deux mesures) et que votre variable dépendante est qualitative et cela dans le but de :

- De comparer les fréquences de ces deux groupes afin d'inférer une relation entre les deux groupes
- De rejeter ou non l'hypothèse nulle, donc de prendre une décision.

L'expression test du khi-carré recouvre plusieurs tests statistiques

- le test d'ajustement ou d'adéquation, qui compare globalement la distribution observée dans un échantillon statistique à une distribution théorique, celle du khi-carré. , trois tests principalement :
- Le test d'indépendance du khi-carré qui permet de contrôler l'indépendance de deux caractères dans une population donnée.
- le test d'homogénéité du khi-carré qui teste si des échantillons sont issus d'une même population.

12.1. Test d'ajustement du Khi-deux

Le but de ce test est de comparer une distribution théorique d'un caractère à une distribution observée. Pour cela, le caractère doit prendre un nombre fini de valeurs, ou bien ces valeurs doivent être rangées en un nombre fini de classes.

➢ Données :

o un caractère A dont les valeurs possibles sont réparties en k classes $A_1,...,A_k$.

➢ La probabilité théorique dans chacune des classes est notée $p_1,...,p_k$.

o n observations, qui donnent un effectif n1 pour la classe $A_1,...,n_k$ pour la classe A_k.

Bien sûr, on doit avoir $n_1+...+n_k=n$.

Hypothèse testée : "La distribution observée est conforme à la distribution théorique" avec un risque d'erreur a.

➢ Déroulement du test :

1. On calcule les effectifs théoriques npj.
2. On calcule la valeur observée de la variable de test :

$$x^2 = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i}$$

➢ On cherche la valeur critique dans la table de la loi du Khi-2 à k-1 degrés de liberté.

➢ On cherche la valeur critique x_a^2 dans la table de la loi du Khi-2 à k-1 degrés de liberté.

➢ Si $x^2 < x_a^2$, on accepte l'hypothèse H_0, sinon on la rejette.

12.2 Calcule et structuration du test khi 2

Soient deux distributions A (Observée) et B (théorique) rangée de la même façon suivant les différentes valeurs (1, 2, 3,…. K) et que peut prendre un caractère étudié.

Pour pouvoir calculer la probabilité du test khi 2 on doit calculer les effectifs théoriques (**T_i**) à partir des effectifs observés (**O_i**)

Caractère	Répartition A (effectifs observés) « Oi »		Répartition B (effectifs théoriques) « Ti »
1	n_1	Il faut calculer les effectifs théoriques à partir des effectifs observés	N^T_1
2	n_2		N^T_2
3	n_3		N^T_3
.	.		.
.	.		.
K	n_k		N^T_K
	$\sum n = N$		$\sum n^T = N$

Pour obtenir les effectifs théoriques

Données observées

n_{11}	...	n_{1j}	$n_{1.}$
...	...	...	...
n_{i1}	...	n_{ij}	$n_{i.}$
$n_{.1}$	...	$n_{.j}$	$n_{..}$

Données théoriques

$n^*_{11}=(n_{.1}*n_{1.})/n_{..}$	...	$(n_{.j}*n_{1.})/n_{..}$	$n_{1.}$
...	...	...	...
$(n_{.1}*n_{i.})/n_{..}$	...	$(n_{.j}*n_{i.})/n_{..}$	$n_{i.}$
$n_{.1}$	...	$n_{.j}$	$n_{..}$

Statistiques de test

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

Où n^* est l'effectif théorique c'est-à-dire l'effectif que l'on aurait eu si les variables étaient indépendantes : $n^*_{ij} = \frac{n_{i.} n_{.j}}{n}$

Effectifs réels

Sexe / Couleur	Bleu	Rouge	Total
Homme	46	44	90
Femme	84	26	110
Total	130	70	200

Effectifs théoriques

Sexe / Couleur	Bleu	Rouge	Total
Homme	58,5	31,5	90
Femme	71,5	38,5	110
Total	130	70	200

La problématique du test ou le problème posé est : les répartitions A et B sont-elles conformes ou différentes ?

Intuitivement, on voit que si $n_i = n_{ti}$, on peut conclure que les deux répartitions A et B sont identiques

Si $n_i \neq n_t$

$_i$, il faut alors étudier l'importance statistique des différences $n_i - n_t$

$_i$. Donc pour pouvoir répondre à la question du test, il faut comparer le résultat des deux tests khi2 celui observé à celui théorique.

La probabilité du test khi 2 observé est donnée par la formule mathématique suivante :

$$x^2_{obs} = \frac{(n_1 - n'_1)^2}{n'_1} + \frac{(n_2 - n'_2)^2}{n'_2} + \cdots + \frac{(n_k - n'_k)^2}{n'_k}$$

Qui peut aussi s'écrire :

$$x^2_{obs} = k\frac{(n_i - n'_i)^2}{n'_i} = \sum \frac{(O_i - T_i)^2}{T_i}$$

Donc les critères de décision (α = 5%)

- **P (x^2_α)** se calcule en fonction du degré de liberté (ddl) et α
- **La formule du « ddl » = (nbr lignes -1) x (nbr colonnes -1)**
- On doit calculer la probabilité de (**Khi2 observé**) et la comparer à la probabilité de (**khi2 théorique ou khi2 alpha**) qui se calcule du tableau :

Extrait de la table du χ^2

α / ν	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,001
1	0,0002	0,001	0,004	0,016	2,71	**3,84**	5,02	**6,63**	10,83
2	0,02	0,05	0,10	0,21	4,61	**5,99**	7,38	**9,21**	13,82
3	0,11	0,22	0,35	0,58	6,25	**7,81**	9,35	**11,34**	16,27
4	0,30	0,48	0,71	1,06	7,78	**9,49**	11,14	**13,28**	18,47
5	0,55	0,83	1,15	1,61	9,24	**11,07**	12,83	**15,09**	20,51

Exemple : pour un ddl = 3, α = 5% (0,05), la probabilité de x^2_α**=7,81**

Rejet de $H_0 \Leftrightarrow \mathbb{P}_{H_0}\left(\sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} > \sum_{i=1}^{k} \frac{(N_{i,obs} - np_i)^2}{np_i}\right) < \alpha$

Remarque :
Le test de khi2 est toujours unilatéral
Si Khi2 *obs* < *khi2* α : on accepte H_0 ⟹ **Si khi 2 obs > Khi 2 theo = H_0 rejeté**
Si Khi2 *obs* > khi2 α : on rejette H_0 ⟹ **Si Khi 2 obs < khi 2 théo = H_0 accepté**

Seuils de signification du test statistique
Si P ($\chi 2 < 0,05$) => *P (différence significative)
P ($\chi 2 < 0,01$) => ** P (différence hautement significative)
P ($\chi 2 < 0,001$) => *** P (différence très hautement significative)

Exemple :

Sur 100 lancés, on a les résultats suivants :

Face	1	2	3	4	5	6
$N_{i,obs}$: Effectifs	7	18	26	15	18	16
np_i	16.67	16.67	16.67	16.67	16.67	16.67
$\frac{(N_{i,obs}-np_i)^2}{np_i}$	5.61	0.11	5.23	0.17	0.11	0.03

Les effectifs théoriques sont par ordre, n_{pi} = 100 x 1/6 = 16,67

$$x_{obs}^2 = \sum_{i=1}^{k} \frac{(N_{i,obs} - np_i)^2}{np_i} = 11,24$$

On doit calculer la probabilité la probabilité de χ^2 **théorique** en fonction du **ddl** et **α**
ddl = (nbr colonnes – 1) x (nbr lignes -1) = (6-1) x (2-1) = 5
α = 5% (0,05)
A l'aide des tables, on trouve la valeur de la probabilité du **khi2 α : 11.07**
- Khi2 *obs* > khi 2 α : On rejette H_0

L'aide des tables, on trouve la valeur de la probabilité du **khi2 α : 11.07**
- Khi2 *obs* > khi 2 α : On rejette H_0

13. Analyse de la variance

L'analyse de la variance à 1 critère de classification ou à un facteur a pour but de comparer les moyennes de plusieurs populations supposant normales, de même variance (homoscédasiticité) et à partir d'échantillon aléatoire simple et indépendant les uns des autres.
Il permet d'identifier les sources de variation qui permet d'expliquer les différences existant entre plusieurs séries d'observations selon un facteur étudié.

13.1 Principe d'analyse de la variance

L'hypothèse nulle s'écrit:

: $m_1 = m_2 = m_3 = \ldots\ldots\ldots\ldots\ldots m_p$ $(p > 2)$

Pour tester l'égalité des moyennes pour « $\boldsymbol{p}$ » populations on prélève un échantillon aléatoire simple pour chaque population.

Les moyenne de ce $\boldsymbol{p}$ échantillon est la moyenne de l'ensemble des données, ou d'observations permettent de définir 2 types de variation.

- Ecart existant entre les échantillons => donne la variation factorielle.
- Ecart existant dans les échantillons => donne la variation résiduelle.

Note :
L'écart total = L'écart dans les échantillons + L'écart entre les échantillons

13.2 Décomposition de la variation totale

Nous supposerons qu'on dispose au départ de $\boldsymbol{p}$ échantillons ou séries d'observations, d'effectifs $\boldsymbol{n_i}$ ($\boldsymbol{i}$ =, . . . , $\boldsymbol{p}$), et nous désignerons l'effectif total par $\boldsymbol{n_.}$:

$$\boldsymbol{n_. = \sum_{i=1}^{p} n_i}$$

Nous désignerons aussi les différentes observations par le symbole $\boldsymbol{x_{ik}}$ ($\boldsymbol{i}$ = , . . . , $\boldsymbol{p}$ $\boldsymbol{et}$ $\boldsymbol{k}$ = $\boldsymbol{1}$, . . . , $\boldsymbol{n_i}$), la valeur $\boldsymbol{x_{ik}}$ étant donc la $\boldsymbol{k^{éme}}$ observation du $\boldsymbol{i^{éme}}$ échantillon.

On peut déduire $\boldsymbol{p}$ somme d'observation $\boldsymbol{X_{..}}$, et $\boldsymbol{p}$ somme des carrés $\boldsymbol{T}$ relatifs aux $\boldsymbol{p}$ échantillons, et une somme des carrés des écarts résiduels $\boldsymbol{SCE_r}$.

$$\boldsymbol{X_{i.} = \sum_{k=1}^{n_i} x_{ik}}$$

$$\boldsymbol{X_{..} = \sum_{i=1}^{p} x_{i.} = \sum_{i=1}^{p} \sum_{k=1}^{n_i} x_{ik}}$$

$$\boldsymbol{C = \frac{X_{..}^2}{n_.} = \frac{\left(\sum_{i=1}^{p} \sum_{k=1}^{n_i} x_{ik}\right)^2}{n}}$$

$$\boldsymbol{T = \sum_{i=1}^{p} \sum_{k=1}^{n_i} x_{ik}^2}$$

$$\boldsymbol{SCE_r = \sum_{i=1}^{p} \sum_{k=1}^{n_i} x_{ik}^2 - \frac{X_{i.}^2}{n_i}}$$

$$\boldsymbol{SCE_t = T - C}$$

Tableau 22 Tableau explicatif des données brutes de l'analyse de la variance

i / k	1	2	.	.	p	Totaux
1	n_{11}	n_{21}	.	.	n_{p1}	
2	n_{12}	n_{22}	.	.	n_{p2}	
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
n_i	n_1	n_2	.	.	n_p	$n_.$
$X_{i.} = \sum_{k=1}^{n_i} x_{ik}$	$X_{1.}$	$X_{2.}$	.	.	$X_{p.}$	$X_{..}$
$\sum_{k=1}^{n_i} x_{ik}^2$	$\sum_{k=1}^{n_i} x_{1k}^2$	$\sum_{k=1}^{n_i} x_{2k}^2$	.	.	$\sum_{k=1}^{n_i} x_{pk}^2$	T
$SCE_i = \sum_{k=1}^{n_i} x_{ik}^2 - \frac{X_{i.}^2}{n_i}$	SCE_1	SCE_2	.	.	SCE_p	SCE_r

On constate ainsi que **la somme des carrés des écarts** par rapport à la moyenne générale, également appelée **somme des carrés des écarts totale**, peut elle aussi être divisée en deux composantes additives : ***une somme des carrés des écarts factorielle*** ou entre échantillons, et ***une somme des carrés des écarts résiduelle*** ou dans les échantillons.

En désignant la somme totale par SCE_t et ses deux composantes respectivement par SCE_f et SCE_r, on peut résumer l'équation d'analyse de la variance sous la forme condensée :

$$SCE_t = SCE_f + SCE_r$$

Les nombres de degrés de liberté peuvent être associés aux différentes sommes des carrés des écarts. Ces nombres de degrés de liberté sont aussi additifs et se présentent de la manière suivante :

$$n. - 1 = (p - 1) + (n. - p)$$

Enfin, en divisant les sommes des carrés des écarts par leurs nombres de degrés de liberté respectifs, on définit des quantités appelées carrés moyens, à savoir un carré moyen total, un carré moyen factoriel ou entre échantillons, et un carré moyen résiduel ou dans les échantillons :

- Variation factorielle => mesure par carré moyen factorielle (CM_f)
- Variation résiduelle => mesure par carré moyen résiduelle (CM_r)

$$CM_t = \frac{SCE_t}{n. - 1} \quad ; \quad CM_f = \frac{SCE_f}{p - 1} \quad ; \quad CM_r = \frac{SCE_r}{n. - p}$$

L'égalité des moyennes doit donc correspond à la valeur élevée du rapport des carrés moyens

$$F_{obs} = \frac{CM_f}{CM_r}$$

Pour réaliser le test d'égalité des moyennes en doit comparer le carré moyen factorielle CM_f ou carré moyen résiduelle pour cela on calcule le rapport $F_{obs} = \frac{CM_f}{CM_r}$ et on le compare, par la suite, avec la valeur $F_{1-\alpha}$.

Avec :

$\alpha = 0,05 ; 0,01 ; 0,001$

$\boldsymbol{k_1} = (\boldsymbol{p} - \mathbf{1})$ $\boldsymbol{dl}$ $Correspond$ à $\boldsymbol{CM_f}$
$\boldsymbol{k_2} = (\boldsymbol{n}. - \boldsymbol{p})$ $\boldsymbol{dl}$ $\boldsymbol{C}orrespond$ à $\boldsymbol{CM_r}$

Règle de décision et interprétation

Si au niveau $\boldsymbol{\alpha} = \mathbf{0},,$ $F_{obs} < F_{1-\alpha}$ ⇒ On accepte l'hypothèse nulle donc il n'existe pas des différences significatives entre les moyennes ⇒ $\boldsymbol{NS}$

Si au niveau $\boldsymbol{\alpha} = \mathbf{0},,$ $F_{obs} \geq F_{1-\alpha}$⇒On rejette l'hypothèse nulle donc il existe des différences significatives entre les moyennes ⇒ (*).

Si au niveau $\boldsymbol{\alpha} = \mathbf{0},,$ $F_{obs} \geq F_{1-\alpha}$ ⇒On rejette l'hypothèse nulle donc il existe des différences hautement significatives entre les moyennes ⇒ (**).

Si au niveau $\boldsymbol{\alpha} = \mathbf{0},\mathbf{1}$, $F_{obs} \geq F_{1-\alpha}$ ⇒ On rejette l'hypothèse nulle donc il existe des différences très hautement significatives entre les moyennes ⇒ (***).

L'ensemble des résultats peut être présenté sous la forme d'un tableau d'analyse de la variance.

Tableau 23 : Tableau d'analyse de la variance à un critère de classification.

Source de variation	$\boldsymbol{ddl}$	$\boldsymbol{Somme\ des\ carrés\ des\ ecarts}$	$\boldsymbol{Carré\ Moyens}$	$\boldsymbol{Fobs}$
$\boldsymbol{Variation\ inter - espéces}$	$\boldsymbol{P\_1}$	SCE_f	CM_r	...***
$\boldsymbol{Variation\ intra - espéces}$	$\boldsymbol{n.-P}$	SCE_r	CM_r	
$\boldsymbol{Totale}$	$\boldsymbol{n.-1}$	SCE_t		

Remarque

Travailler sur des plans équilibrés (même effectif dans chaque échantillon) atténue (rend moins important) l'effet néfaste de l'hétérogénéité des variances.

1. Introduction

La théorie des séries temporelles (chronologiques) est appliquée de nos jours dans des domaines aussi variés que la biologie, la médecine ou la démographie, pour n'en citer qu'une petite partie. Leur particularité vient de l'introduction du temps dans l'analyse de ces données : on étudie une suite de couples de la forme (t; Yt), où Yt est l'observation de la variable à l'instant t. L'étude des séries chronologiques (ou séries temporelles ou chroniques - time séries en anglais) permet de d'écrire, expliquer, contrôler, prévoir des phénomènes Evoluant au cours du temps

2. Objectifs

Les principaux objectifs de la modélisation des séries temporelles sont les suivants

- Comparer deux séries temporelles. Par exemple, l'évolution démographique de deux régions ou deux séquences d'ADN.
- Dans le séquençage du génome, détecter les parties de l'ADN qui contiennent de l''information.
- Décrire l'évolution
- Permettre l'explication des fluctuations.
- Faciliter la prévision (le passé peut expliquer le futur)
- Prédire l'évolution future de la série temporelle 'partir des valeurs qui ont été observées'.

3. Qu'est-ce qu'une série temporelle ?

3.1. Définitions

On appelle série chronologique la succession de valeurs que prend une variable (ou caractéristique) au cours du temps pour N périodes successives pour un même individu (ou cas). Donc, une série chronologique est une suite d'observations d'une variable statistique au cours du temps. Une série chronologique peut être identifiée 'une série statistique' deux variables t et Yt, où t représente le temps. On la note (t ; Yt)

Exemple : Nombre annuel de naissance en Algérie

3.2. Domaines d'application.

On trouve des exemples de séries chronologiques univariées dans de très nombreux domaines :

- En Médecine / biologie : suivi des Evolutions des pathologies, analyse d'électro-encéphalogrammes et d'électrocardiogrammes.
- En Sciences de la Terre et de l'espace : indices de marées, variations des Phénomènes physiques (Météorologie), Evolution des taches solaires, phénomènes d'avalanches.
- En Démographie : évolution de la population.

Remarque :
Les dates d'observations sont généralement ordonnées de manière régulière dans le temps ; on manipule

- Des séries journalières (cours d'une action en bourse)
- Des séries mensuelles (température mensuelle)
- Des séries trimestrielles (précipitation trimestrielle)
- Des séries annuelles (chiffre annuel de la production agricole).

4. Principe

Pour décomposer une série chronologique on doit commencer par :

- Tracer son graphique
- Choisir un modèle de composition (additif ou multiplicatif)
- Estimer la tendance Ct
- Estimer les variations saisonnières.

5. Présentations d'une série chronologique

Les données d'une série chronologique **trimestrielle, mensuelle, journalière** sont présentées :

• soit repérées à l'aide d'un seul indice t qui représente le nombre de mois entre la première observation et l'observation de la donnée en question + 1, (la 1re observation étant Y_1). Donc, sous la forme d'un tableau à deux colonnes, contenant *np* lignes.

t	1	2	.	.	.	n_p
y_t	y_1	y_2	.	.	.	y_{np}

• soit repérées à l'aide de deux indices i et j, où i représente l'année de l'observation, j son « mois ». Le « mois » peut être un trimestre, un mois, un jour. On notera indifféremment la série (Y_t) et $(Y_{i,j})$.

Donc, la présentation des résultats se fait sous la forme d'un tableau contenant p colonnes et n lignes : une colonne par mois et une ligne par année

	mois					
	1	2	…	*j*	…	*p*
Année 1	y_{11}	y_{12}	…	y_{1j}	…	y_{1p}
Année 2	y_{21}	y_{22}	…	y_{2j}	…	y_{2p}
.	.	.	…	.		
.	.	.	…	.		
.	.	.				
i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{ip}
.	.					
n	y_{n1}	y_{n2}	…	y_{nj}	…	y_{np}

Exemple : Le lien qui existe entre Yt et Yi,j :

- Soit une série trimestrielle. $Y_{4,2} = 4370$ donc $4370 = Yt$ avec $t = (4 - 1) \times 4 + 2 = 14$.
- Soit une série mensuelle. $Y_{26} = 542$ donc $542 = Y_{i,j}$ or $26 = 2*12 + 2$ donc i= 3, j =2.

Ces 2 modes de repérages donnent deux types de présentation des données et deux graphiques différents.

6. Représentation graphiques d'une série chronologique

Essayer de repérer les caractéristiques des séries chronologiques, comme

- Une tendance
- Un phénomène périodique
- Un cycle
- Des variations accidentelles
- Des fluctuations (variations) irrégulières

a) Graphe de la série chronologique.

Pour représenter graphiquement la série chronologique $\{y_t\}_{t\in T}$

1- Dessiner le nuage formé par les points $(t_j\ ;\ y_j)$ $1 \leq j \leq n$
2- Relier les points entre eux par des segments de droite, pour indiquer la chronologie

On représente l'évolution de la grandeur considère sur l'ensemble de la période observée.

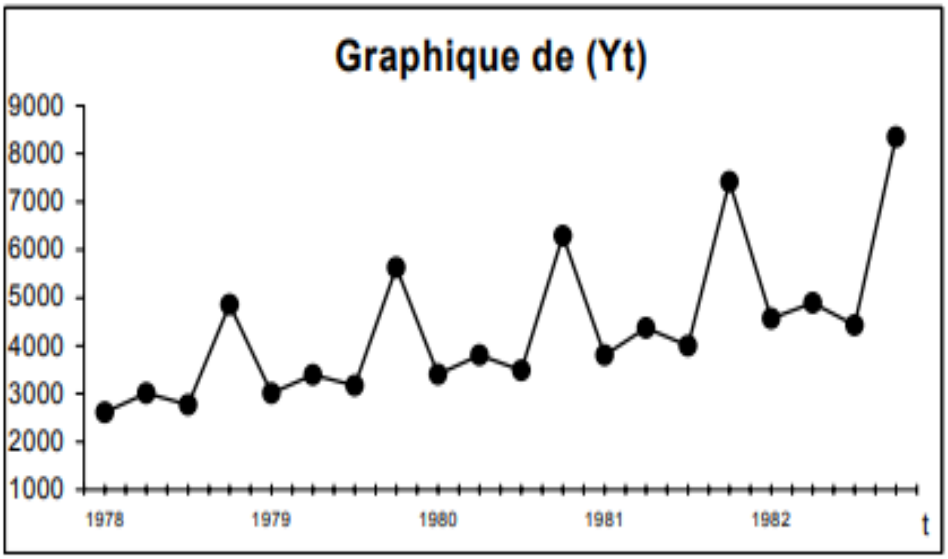


Figure 27 : graphique de Yt .

b) Graphiques des courbes superposées.

Pour le deuxième cas :

1- Représenter les points $(j\ ;\ Y_{i,j})$ que l'on relie par des segments de droites,
2- Relier les points entre eux par des segments de droite, ceci pour chacune des années i.

On représente ainsi l'évolution annuelle de la grandeur au cours des mois (pour chacune des années). On peut ainsi comparer le même mois j des différentes années, mais on ne voit pas l'évolution globale.

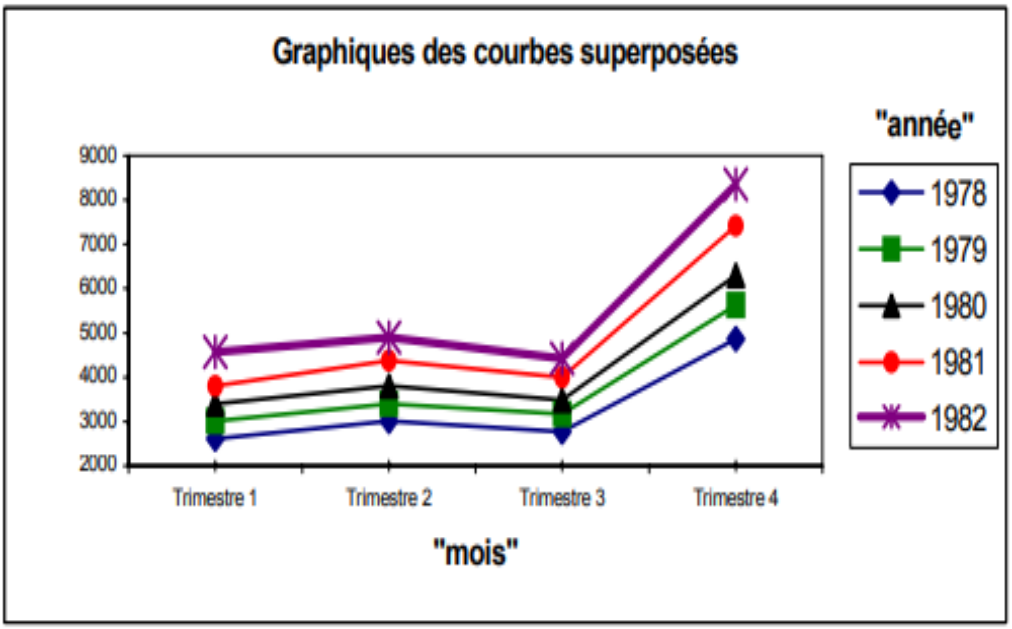


Figure 28 : Graphiques des courbes superposées

7. Composantes fondamentales d'une série chronologique

Le but de la décomposition d'une série chronologique est de distinguer dans l'évolution de la série :

- Une tendance « générale »,
- Des variations saisonnières qui se répètent chaque année,
- Des variations accidentelles imprévisibles.

L'intérêt de ceci est :

- De mieux comprendre, et de mieux décrire l'évolution de la série.
- D'autre part, de prévoir son évolution (à partir de la tendance et des variations saisonnières).

7.1 La tendance à long terme ou *Trend* : Tt

Elle traduit généralement l'évolution générale du phénomène observé (croissance, stagnation, . .) Figure 29.
C'est une courbe (ici une droite) que l'on note Tt (fonction de *t*).
Elle traduit l'aspect général (moyen) de la série.

Exemples :

- **Tendance linéaire :** $Tt = at + b$
- **Tendance quadratique** : $Tt = at^2 + bT + c$
- **Tendance logarithmique** : $Tt = \log(t)$

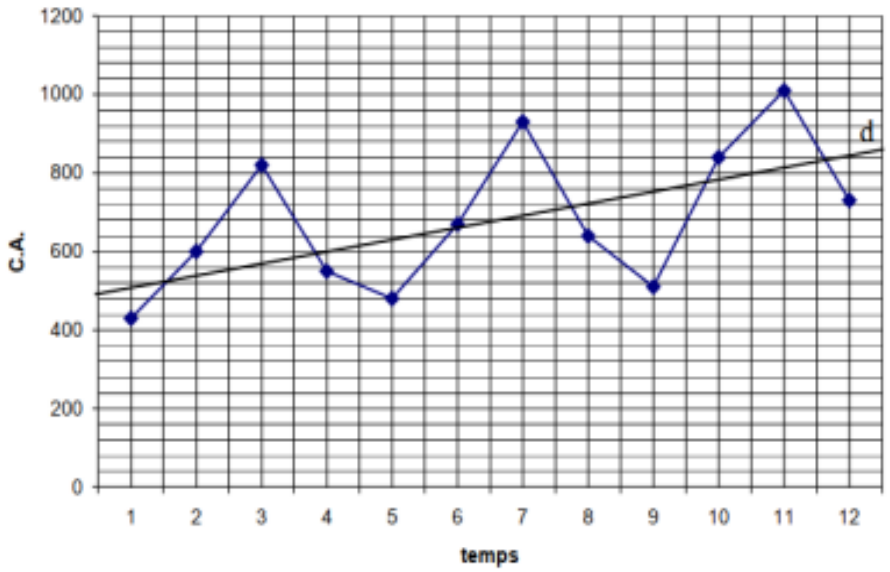


Figure 29 : évolution du C.A par trimestre

7.2 La composante saisonnière : St

Elle correspond à des variations régulières inconnues mais détectables sur des périodes courtes, généralement à l'intérieur d'une année : semaine, mois, trimestre.

La composante saisonnière est donc totalement déterminée par p coefficients saisonniers :

$$Cs_1; : : : ; C_{sj} ; :::; C_{sp}$$

7.3 La composante accidentelle (résiduel) : At

Le mouvement accidentel ou résiduel (composante accidentelle ou résiduelle) correspond à des variations accidentelles de forte amplitude dues à des phénomènes imprévisibles.

Par exemple, la grève, le risque de guerre, forte baisse des températures en été qui entraine une augmentation de l'électricité (chauffage) sur cette période). Cette partie est aléatoire.

7.4 La composante cyclique : *Ct*

Elle correspond à des variations connues régulières, pour des séries très longues.

Par exemple, des fluctuations correspondant à des périodes de prospérité ou de récession. En général, la composante cyclique n'est pas détectable sur la période étudiée et on suppose alors que *Ct* n'existe pas.

Les modèles de décomposition déterministe

8. Les modèles de composition et de déterministes :

On étudiera deux modèles de décomposition déterministes :

a. Le modèle additif

Si Ct ; St et At sont indépendants de Tt, le modèle est défini par

$$Yt = Tt + Ct + St + At ;$$

Dans ce modèle, le nuage de points à une enveloppe d'épaisseur plus ou moins constante

b. Le modèle multiplicatif :

1ere forme du modèle multiplicatif

On emploie le modèle multiplicatif lorsque l'enveloppe du nuage de points s'élargit au fur et à mesure que la tendance générale croit (et est de plus en plus resserrée au fur et à mesure que le trend diminue tout en restant positif).

Le terme Yt est alors vu comme le produit de la tendance générale Tt, de la composante saisonnière St et de la composante aléatoire

$$At : Yt = Tt \times St \times At$$

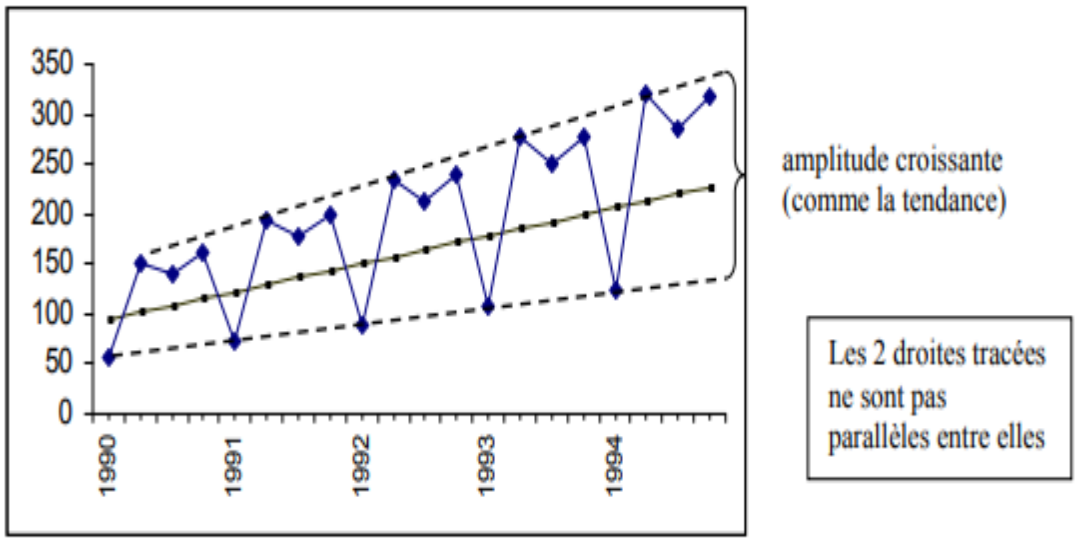


Figure 30 : première forme du modèle additif

2eme forme du modèle multiplicatif

Dans le cas d'une série (Yt) à valeurs positives, ce 2eme modèle multiplicatif se ramène à un modèle additif en considérant la série

$$(\ln(Yt)) : \ln(Yt) = \ln(Ct) + \ln(St) + \ln(\varepsilon t).$$

La seule différence entre les 2 modèles multiplicatifs est dans l'estimation des εt, qui n'a pas une grande importance.

9. Choix du modèle

Le tableau suivant résume les méthodes et ses propriétés pour choisir un modèle.

Tableau 24 : méthodes de choix les modèle

Méthodes	Propriétés	Modèles	
		modèle additif	**modèle multiplicatif**
Méthode du profil	On utilise le graphique des courbes superposées	Si les différentes courbes sont à peu près parallèles	Si les pics et les creux s'accentuent
Méthode de la bande	On utilise le graphe de la série et la droite passant par les minima et celle passant par les maxima.	Si ces 2 droites sont à peu près parallèles	Si ces 2 droites ne sont pas parallèles
Méthode du ableau de Buys et Ballot	On calcule, pour chacune des années, **la moyenne et l'écart type**. On **trace les points d'abscisse la moyenne et d'ordonnée l'écart type de la même année.** On trace **la droite des moindres carrés de ces points**.	Si l'écart type est indépendant de la moyenne La pente (a) de la droite des moindres carrés est très proche de 0	Si l'écart type est fonction de la moyenne, la pente (a) de la droite des moindres carrés n'est pas nulle.

[1]. Dagnelie, P. (1988). Quelques notions d'expérimentation agronomique utilisées en recherche biopharmaceutique. Revue de Statistique Appliquée, 36(2), 23-36.

[2]. Dagnelie, P. (2003). Principes d'expérimentation : planification des expériences et analyse de leurs résultats. Presses agronomiques de Gembloux.

[3]. Dagnelie, P. (2006). Statistique théorique et appliquée. Tome 1, statistique descriptive et bases de l'inférence statistique. De Boeck Université. Paris, Bruxelles.

[4]. Dagnelie, P. (2013). Statistique théorique et appliquée. Tome 2. Inférence statistique à une et à deux dimensions. De Boeck et Larcier sa, Département de Boeck Université, Paris, Bruxelles.

[5]. Golmard, J.L. Mallet, A. Morice. V. (2007). Biostatistique. Faculté médecine Université Pierre et Marie Curie, 181p.

[6]. Gharout, H. (2019). Cours de Biostatistique. Faculté des Sciences de la Nature et de la Vie, Université Abderrahmane Mira de Bejaia, 70p.

[7]. Fisher, R.A. (1925). Statistical methods for research workers. Oliver and Boyd, Edinburgh.

[8]. Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd edition. Lawrence Erlbaum Assoc., Publ., Hillsdale, New Jersey.

[9]. Edgington, E.S. (1995). Randomization tests. 3rd edition. Marcel Dekker Inc., New York.

[10].Scherrer, B. (1984). Biostatistique. Gaëtan Morin Éditeur, Chicoutimi.

[11].Montgomery, D. C., Peck, E. A. & Geoffrey Vining, G. (2001). Introduction to linear regression analysis. John Wiley, New-York, 3 edition.

[12]. Allalga, A. (2020). Cours de Biostatistique. Faculté des Sciences de la Nature et de la Vie, Université Mohamed Cherif Messaadia, Souk-Ahras, 75p.

[13]. Menaa, M. (2020). Cours d'Analyse Multivariée & Modélisation Statistique sous R. Faculté des Sciences de la Nature et de la Vie, Université Mohamed Cherif Messaadia, Souk-Ahras, 35p.

[14]. Florence. N. Cours de généralité sur les series chronologiques. Département de Réseaux et Télécommunications. Université Côte d'Azur. 6p.