République Algérienne Démocratique et Populaire

Ministère de L'Enseignement Supérieur et de la Recherche Scientifique

Université de Ghardaïa

Faculté des Sciences de la Nature et de la Vie et Sciences de la Terre

Département de Biologie

Cours

Biostatistique

Destiné aux étudiants de Master I Biochimie appliquée

Dr. KRAIMAT Mohamed

Année universitaire 2025/2026

Préambule

La biostatistique est la langue commune de la recherche en sciences biologiques. Elle relie les observations du terrain et du laboratoire à des conclusions fiables, reproductibles et utiles pour la décision scientifique. Dans ce cours, Les étudiants inscrits en Master en biochimie appliquée sont censés d'apprendre à formuler des questions quantitatives claires, à concevoir des expériences robustes, à analyser des données avec des méthodes appropriées, puis à communiquer des résultats de manière transparente et critique.

Objectifs du cours

- Situer la biostatistique dans la démarche scientifique : hypothèses, plan d'échantillonnage, puissance et erreurs.
- Choisir et appliquer des méthodes d'analyse adaptées (descriptives, inférentielles et comparatives) aux données biologiques.
- Interpréter les résultats au regard du contexte biologique, des hypothèses et des limites des méthodes.

2Compétences visées

- Concevoir un protocole expérimental (plans complètement aléatoires, en blocs, facteurs croisés, mesures répétées).
- Maîtriser les tests usuels (t, χ^2, corrélations, ANOVA et non paramétriques) et leurs conditions d'application.
- Utiliser un environnement d'analyse (R/RStudio ou équivalent) pour importer, nettoyer, visualiser et modéliser des données.
- Rédiger une section "Matériels & Méthodes / Résultats" claire, avec tableaux, figures et indicateurs d'incertitude.

Prérequis & ressources

- Bases de probabilité, algèbre et notions d'échantillonnage.
- Connaissances fondamentales en biologie (écologie, physiologie, génétique expérimentale).
- Outils recommandés : R ≥ 4.x, RStudio,

Plan du cours –Biostatistique

1. Généralités sur le protocole expérimental et le principe d'expérimentation

1.1. Principe d'expérimentation

1.1.1. Démarche scientifique expérimentale

1.1.2. Source de variation en sciences expérimentales

1.1.3. Notion d'un protocole expérimental

1.2. Nature des variables statistiques

1.2.1. Variables qualitatives

1.2.2. Variables quantitatives

2. Statistiques descriptives

2.1. Notion d'une série de distribution

2.2. Variable aléatoire

2.3. Paramètres caractéristique d'une variable statistique

2.3.1. Paramètres de position

2.3.2. Paramètres de dispersion (variation)

2.3.3. Paramètres de forme

2.4. Notion d'échantillon et de population

3. Statistiques inférentielles

3.1. Lois de probabilités

3.1.1. Loi normale (loi continue)

3.1.1.1. Calcul d'une probabilité dans le cas d'une loi normale

3.1.1.2. Loi normale centrée et réduite

3.1.1.3. Evaluation d'une aire à gauche d'une valeur ou $P(Z \leq z)$

3.1.1.4. Evaluation d'une aire à droite d'une valeur ou $P(Z > z)$

3.2. Inférence statistique

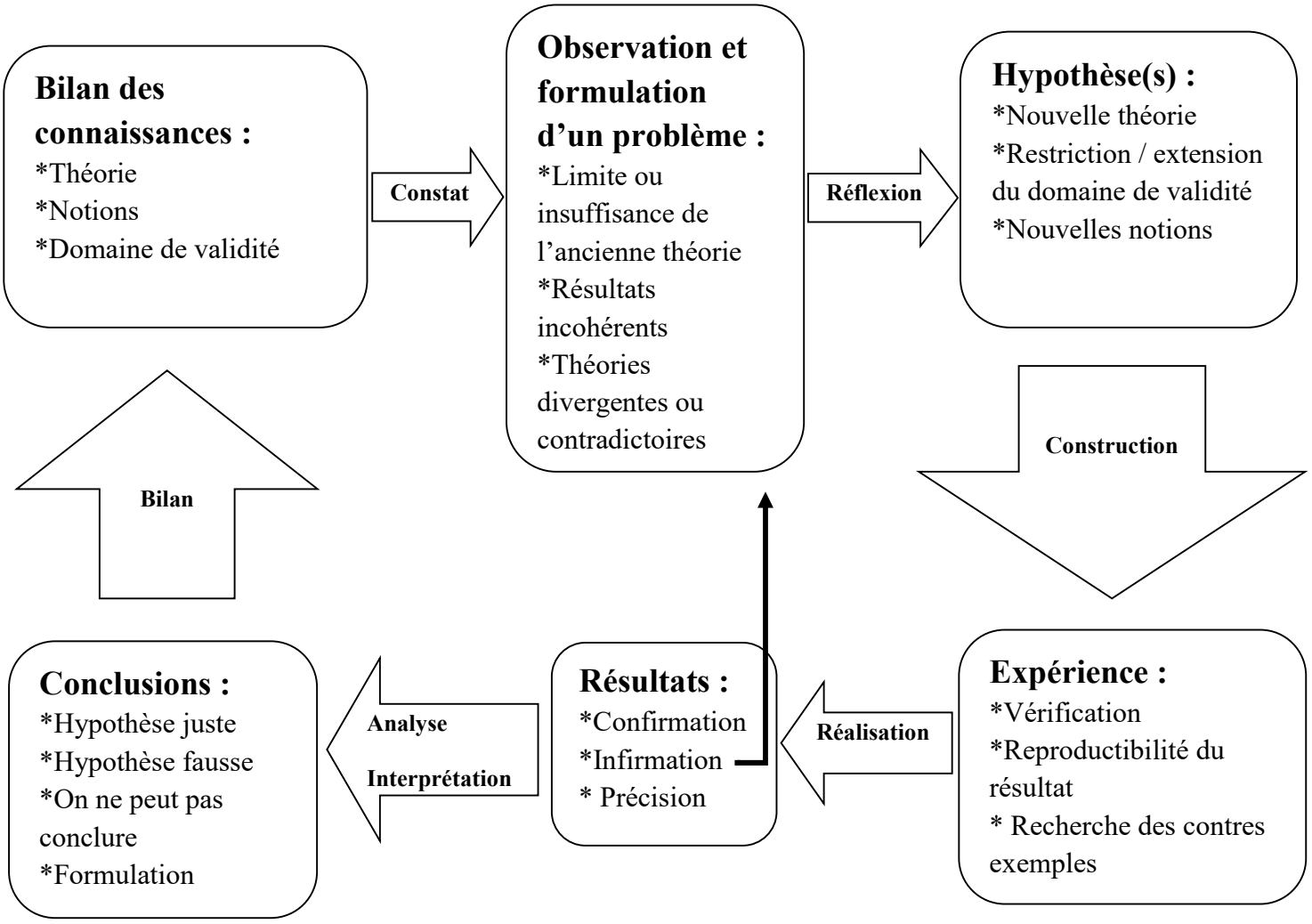
3.2.1. Tests d'hypothèses

1. Généralités sur le protocole expérimental et le principe d'expérimentation

Le mot statistique est dérivé du mot latin *status* qui signifie « état ». En effet, la statistique est un outil pour commencer l'interprétation des phénomènes scientifiques. La biostatistique est un champ scientifique constitué par l'application de la science statistique à la biologie. Il peut s'agir de la conception du volet statistique des études biologiques ou du recueil, de l'analyse et du traitement des données recueillis lors des études biologiques. Contrairement aux sciences exactes, en sciences expérimentales il y a toujours une variation. Mais avant d'aborder les sources de variations et d'incertitude des résultats en sciences expérimentales il faut définir d'abord qu'est-ce qu'une démarche scientifique expérimentale.

1.1. Démarche scientifique expérimentale :

On peut schématiser le processus de la démarche scientifique expérimentale comme suit :



A partir donc des observations préliminaires, on peut établir des hypothèses. La vérification des hypothèses doit passer par une expérimentation. Cependant, la gestion des expériences scientifiques doit reprendre à des caractères déterminant le but et les objectifs recherchés qui sont en lien direct avec les hypothèses et les paramètres à étudier.

1.2. La source de variation en sciences expérimentales

Exemple : une parcelle homogène de 100 m^2, plantée par une culture du blé et produisant un rendement de 50 kg/ 100 m^2.

	5 m	5 m
5 m	P1	P2
5 m	P3	P4

Théoriquement (mathématiquement), les rendements de quatre (04) parcelles sont égaux c'est-à-dire :

P1 = P2 = P3 = P4

Pratiquement, ce n'est pas le cas :

P1 # P2 # P3 # P4

Donc, il y toujours une variation (α) entre les résultats pratiques et les résultats attendus (théoriques), ce qu'on appelle l'erreur α (source de variation).

Tel que :

Variation totale = variation systématique + variation aléatoire

Variation aléatoire : incontrôlable / non maîtrisable.

Variation systématique : qu'on peut la maîtriser surtout par la minimisation des risques d'erreur.

Pour contrôler la source de variation systématique, il nous faut une stratégie ou un programme expérimental performant, appelée aussi « le protocole expérimental ».

1.3. Notion d'un protocole expérimental

a) Définir l'objet de l'expérimentation (objectifs recherchés + la problématique) ;

- Aspect scientifique.

- Retombées socio-économiques.

b) Définir le ou les facteur (s) à étudier ;

En fonction de ou des objectif (s) recherchés :

-Objectif unique : expérience monofactorielle

- Objectif multiple : expérience plurifactorielle.

) Notion d'un facteur

Ensemble d'éléments de même nature. En statistique, le terme variable désigne toujours un facteur étudié.

« Si la caractéristique mesurée peut prendre différentes valeurs, on dit alors que cette caractéristique est une variable ».

Notons que les essais peuvent être simples ou multiples (factorielles), en fonction de la présence ou de l'absence d'interactivité (partielle ou totale) entre les facteurs étudiés (variables).

On distingue alors :

L'expérience factorielle : dans laquelle, il existe une interactivité partielle ou totale.

L'expérience simple : absence d'interactivité.

En termes de facteurs on distingue :

- Le facteur contrôlé : lorsqu'on fait référence aux paramètres utilisés dans la maitrise de l'hétérogénéité par rapport au dispositif expérimental adopté. Le facteur contrôlé dépend donc du dispositif expérimental.
- Le facteur étudié : c'est le facteur objet d'étude contrôlé expérimentalement afin d'évaluer de la variation aléatoire.

En général, les objets de l'étude (de l'expérience), se compose de différents traitements (contrôlés ou étudiés) servant à expliquer les variations obtenues.

c) Définir les unités expérimentales :

C'est l'unité de base faisant l'objet d'au moins d'une observation. L'unité expérimentale est l'objet d'un traitement individuel, indépendamment aux autres unités expérimentales. De même pour son analyse statistique qui devrait se réaliser individuellement.

Exemples :

- Production du miel dans une ruchée : l'unité expérimentale est la ruche.

- Production agricole : l'unité expérimentale est la parcelle

- Etude épidémiologique (une maladie infectieuse des abeilles) : l'unité expérimentale est l'abeille.

- Arboriculture : l'unité expérimentale est souvent exprimée par arbre, un rameau…etc.

D'une manière générale, le dispositif expérimentale (plan d'expérience) est la distribution ou la répartition des objets à étudier dans les unités expérimentales.

) Critères du choix d'un dispositif expérimental

- Répétition / pour chaque dispositif expérimental, il y a un nombre de répétitions à respecter.
- Environnement / Choisir le dispositif expérimental qui s'adapte mieux à l'homogénéité ou l'hétérogénéité du milieu.

Le concept d'un dispositif expérimental est souvent dicté par le gradient d'hétérogénéité, pouvant exister dans l'espace expérimental. On distingue plusieurs dispositifs en fonction du degré d'hétérogénéité.

- Dans des conditions de faiblesse d'hétérogénéité, il est souvent conseillé d'opter pour les dispositifs les plus simples (randomisation totale).

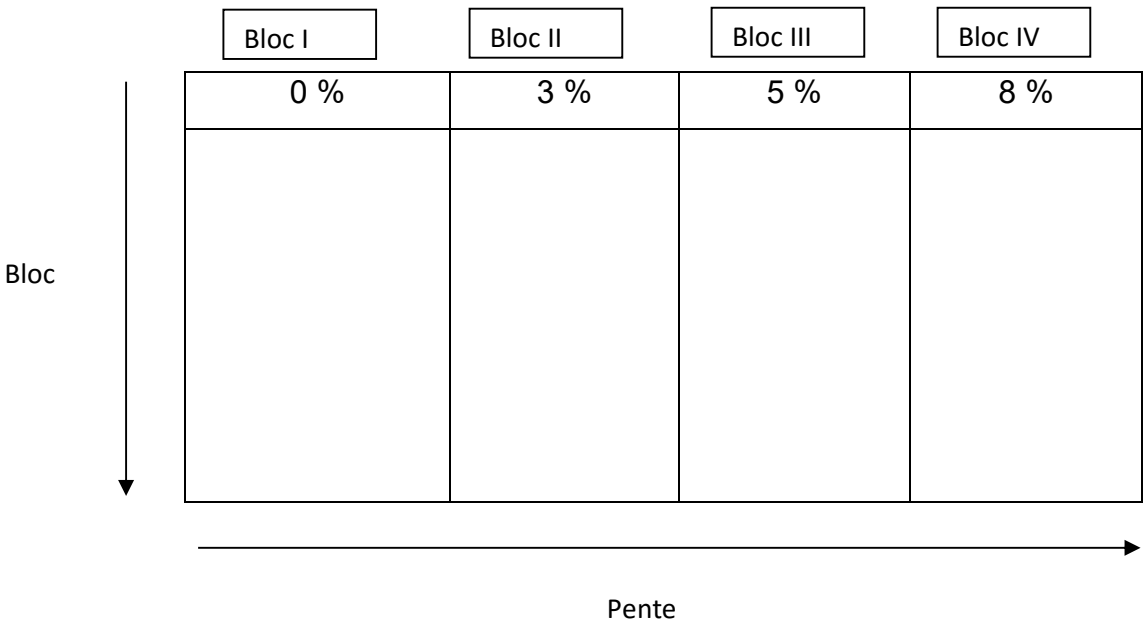
- Dans le cas de la présence d'une seule source d'hétérogénéité, on utilise le dispositif en blocs.

- Dans le cas de la présence de deux gradients d'hétérogénéité différents, on utilise souvent le dispositif du Carré-latin.

Exemples :

*) Dispositif en blocs : dans ce cas le bloc est orienté au sens vertical par rapport au gradient d'hétérogénéité.

Dans une étude de l'influence de la fertilisation phosphatée sur le comportement d'une culture du haricot, implantée sur un terrain hétérogène (il existe plusieurs niveaux de pentes). On utilise le dispositif en blocs, en considérant chaque bloc comme une répétition et en les orientant au sens vertical par rapport à la disposition des pentes.



Facteur contrôlé : bloc ;

Facteur étudié : fertilisation phosphatée.

*) Dispositif en carré-latin

Colonne

Ligne

Dans une étude de l'influence de la fertilisation phosphatée sur le comportement de la culture du haricot cultivée sur une parcelle hétérogène (plusieurs niveaux de pente) et avec plusieurs buses d'irrigation.

Colonne : gradient d'irrigation.

Ligne : gradient de pente.

- Facteur contrôlé : carré-latin ;

- Facteur étudié : fertilisation phosphatée.

1.2. Nature des variables statistiques

1.2.1. Définition

Une variable est une caractéristique étudiée pour une population donnée. Le sexe, la couleur préférée, le nombre de téléviseurs de votre foyer ou encore l'âge sont des variables. Des milliers de variables peuvent être sujet aux études.

Les variables sont en général désignées par les observations ou les caractères observables, mesurables ou encore quantifiables.

Il existe deux types de variables :

1.2.2. Les variables qualitatives : sont des variables représentées par des qualités, telles que le sexe, le programme d'études ou encore l'état civil. Les variables qualitatives s'expriment en modalités. Les modalités sont comme des choix de réponses aux variables étudiées.

1.2.2.1. Les variables qualitatives nominales : sont des variables qui correspondent à des noms, il n'y a aucun ordre précis. Par exemple, le sexe à 2 modalités possibles : féminin ou masculin.

1.2.2.2. Les variables qualitatives ordinales : sont des variables qui contiennent un ordre. Par exemple, le degré de satisfaction par rapport à votre fournisseur cellulaire. Les différentes modalités seraient : très satisfait, satisfait, insatisfait, très insatisfait. Les variables qualitatives ordinales sont très souvent des degrés de satisfaction, d'approbation, …etc.

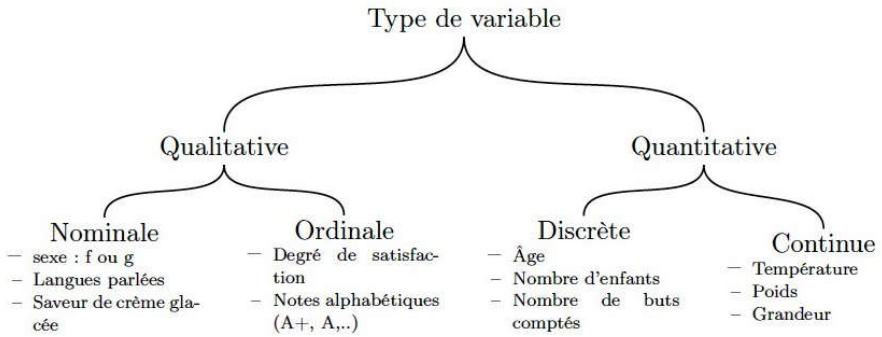
1.2.3. Les variables quantitatives : sont quant à elles des variables représentées par des quantités telles que l'âge, le poids et la taille. Elles s'expriment en valeurs. Les valeurs représentent les choix de réponses aux variables quantitatives.

Pour les variables quantitatives, il y a encore 2 types de variables différentes :

1.2.3.1. Les variables quantitatives discrètes : sont des valeurs que l'on peut énumérer, il est inutile d'utiliser des classes pour les exprimer. Par exemple, le nombre de personnes dans le ménage, …etc.

1.2.3.2. Les variables quantitatives continues : sont des valeurs très nombreuses dont l'énumération serait fastidieuse. Il est donc préférable de les exprimer en classe de largeur égale. Par exemple, le poids des échantillons végétaux,…etc.

On peut résumer la nature des variables statistiques dans le diagramme suivant :



Chaque facteur d'une expérimentation est étudié (ou contrôlé) à plusieurs niveaux, variantes ou modalités.

On parle de variantes quand il s'agit de facteur qualitatif (variétés par exemple), et de niveaux quand il s'agit de facteur quantitatif (dose d'un produit par exemple) et de modalités dans les deux cas.

Les combinaisons de différents niveaux des facteurs étudiés définissent des traitements.

Les résultats d'une expérience sont appréciés par des variables. Nous distinguons :

*) Les variables mesurées ou saisies

*) Les variables élaborées ou calculées.

2. Statistiques descriptives

2.1. Notion d'une série statistique (série de distribution)

Exemple : soit l'étude de l'état civil de 40 employés d'une entreprise donnée :

N = 40

X : représente la variable statistique (état civil).

On admet pour X = {marié, divorcé, célibataire, veuf,….}

$$X = \{x_1, x_2, x_3, \ldots\ldots\}.$$

Pour chaque modalité x_i on compte le nombre d'individus ayant cette modalité à partir des données brutes.

On note ce nombre par n_i et on l'appelle la fréquence absolue (n_i)

$$fi = \frac{ni}{N}$$

Tel que :

f_i : la fréquence relative ;

n_i : la fréquence absolue ;

N : l'effectif total.

Fréquence relative	**Modalité (x_i)**	**Fréquence absolue (n_i)**
0,5	Marié (x_1)	20
0,27	Célibataire (x_2)	11
0,15	Divorcé (x_3)	6
0,07	Veuf (x_4)	3

A chaque modalité x_i on associe une fréquence n_i, l'ensemble des couples (x_i, n_i) est une fonction qu'on appelle distribution des fréquences ou série statistique.

2.1.1 Série statistique pour d'un caractère quantitatif discontinu (discret) :

En considérant un échantillon de taille n (composé de n individu) et appelant X la valeur d'un caractère donné avec des modalités x_1, x_2, x_3, ……..x_n , on a :

- L'effectif total de la série est le nombre n éléments de l'échantillon étudié.

- La fréquence absolue (n_i) : est la répartition de la modalité x_i, n_i fois dans la série statistique.

- La fréquence relative :

$fi = \frac{ni}{N}$ avec $\sum_{i=1}^{n} fi = 1$

Pourcentage = $fi \times 100$

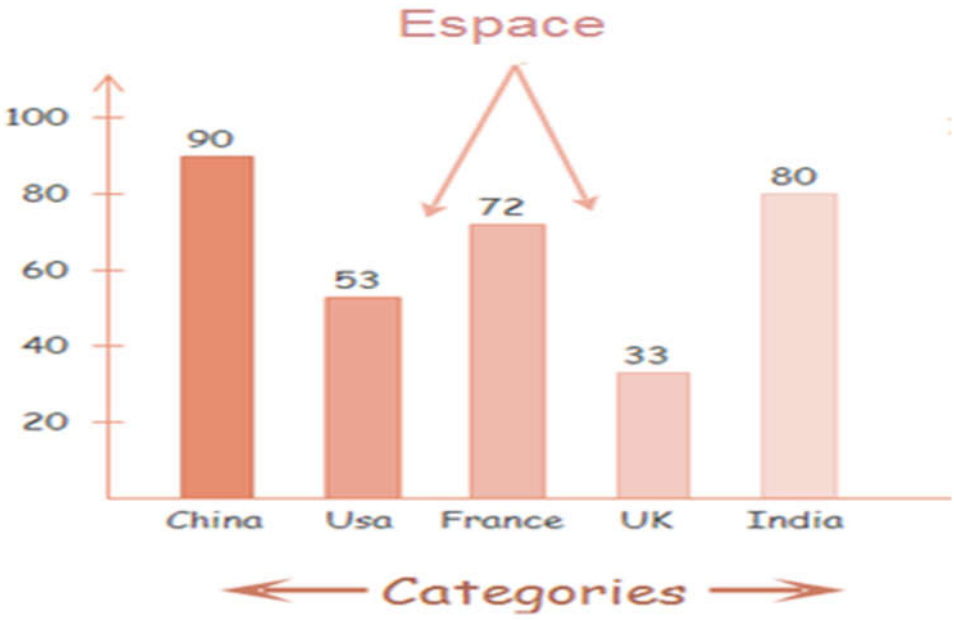
Exemple : on a réparti 150 vaches suivant le nombre de parasités qu'elles hébergent.

Nombre de parasites/vache	**Nombre de vaches correspondant (n_i)**	**Fréquence relative (f_i)**
0	11	0,07
1	22	0,14
2	45	0,30
3	40	0,26
4	19	0,12
5	11	0,07
6	2	0,01
	150	1

Représentation graphique :

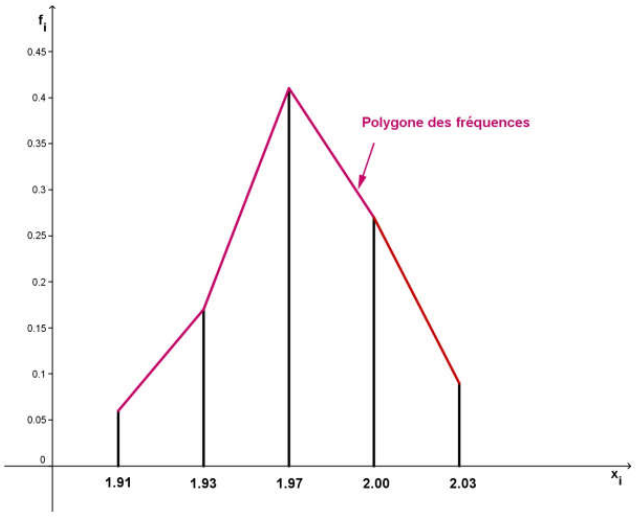
a) Diagramme en Bâtons

C'est un ensemble de bâtons ayant pour abscisse les valeurs : x_1, x_2, x_3, ……..x_n , du caractère et en chacun des points d'abscisse xi correspond une ordonnée proportionnelle à l'effectif n_i de x_i .



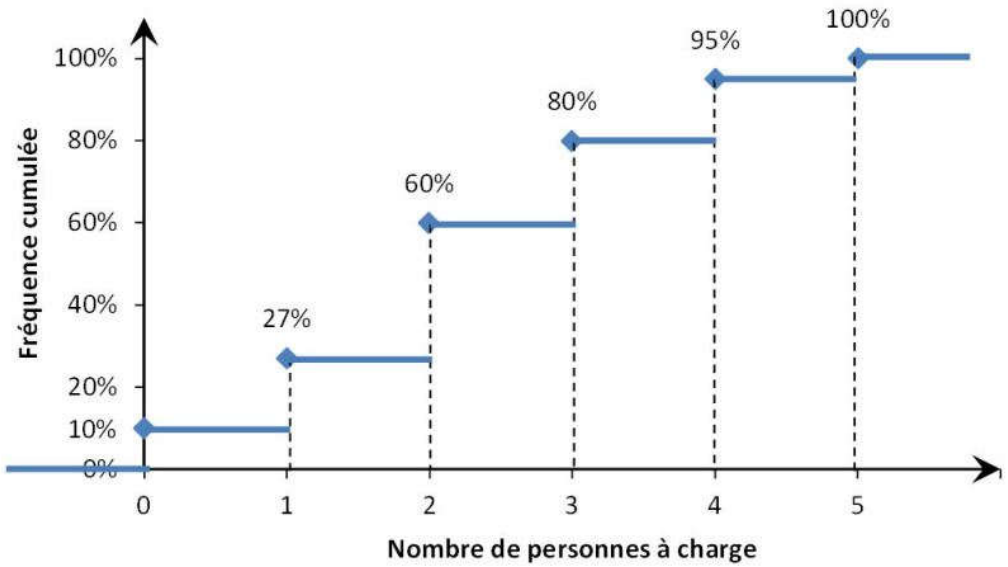
b) Polygone de fréquence

On obtient le polygone des fréquences en joignant par des segments droits les extrémités des bâtons, c'est un graphique linéaire passant par des points ayant pour abscisse x_i et pour ordonnées n_i.



c) Diagramme cumulatif

On appelle l'effectif cumulatif jusqu'à la $i^{ième}$ valeur x_i du caractère la somme $n_1+n_2+\ldots\ldots\ldots\ldots+n_i$ des effectifs obtenus pour les $i^{ième}$ valeurs du caractère de même, la fréquence relative cumulative. Dans ce cas les bâtons ont longueurs proportionnelles aux effectifs cumulés (ou fréquences cumulés), le diagramme prend donc la forme d'un escalier.



2.1.2. Série statistique dans le cas d'un caractère quantitatif continu :

Afin de permettre une étude exacte et d'éviter une répartition de fréquence trop disposée, on constitue des classes, en divisant l'étendue de la série statistique en un certain nombre d'intervalles partiels due de la série inégale.

Chaque classe contiendra toutes les valeurs égales ou supérieures à sa limite inférieure mais strictement inférieure à sa limite supérieure de façons que les classes ne doivent jamais se chevaucher.

Exemple : poids de nouveaux nés (2,240 kg - 4,490 kg)

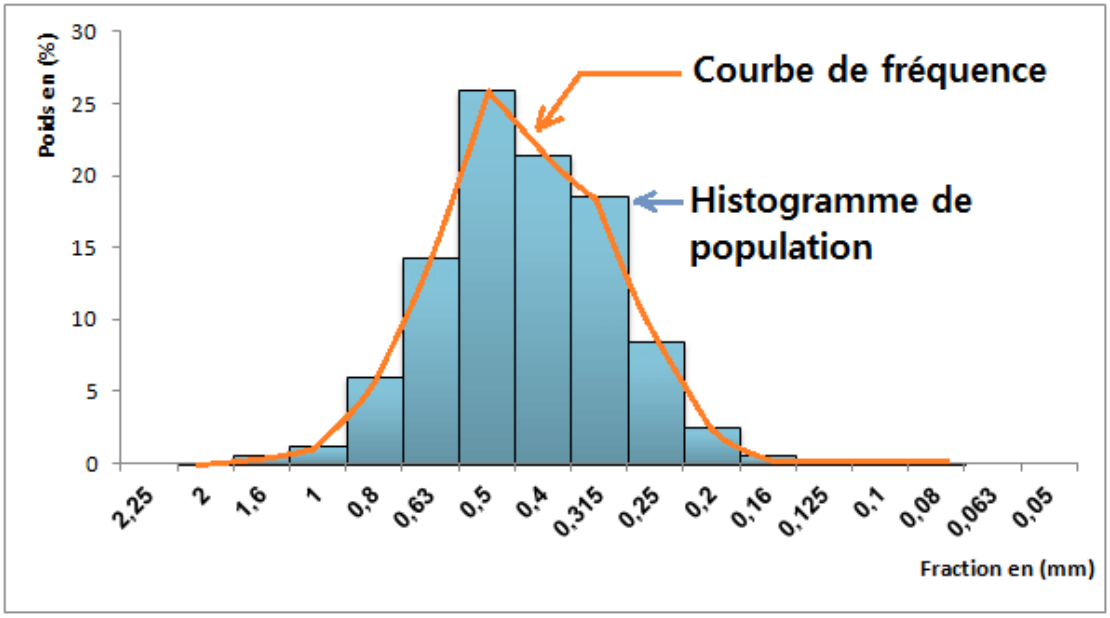
Classe	**Limites de classe (kg)**	**Centre de classe (kg)**	**Effectif/ fréquence absolue (n_i)**	**Fréquence relative (fi)**	**%**
1	[2,2 -2,5[	2,350	5	0,031	3,1
2	[2,5 – 2,8[	2,650	11	0,068	6,8
3	[2,8 – 3,1[	2,950	24	0,148	14,8
4	[3,1 – 3,4[	3,250	40	0,248	24,8
5	[3,4 – 3,7[	3,550	42	0,259	25,9
6	[3,7 – 4,0[	3,850	20	0,124	12,4
7	[4,0 – 4,3[	4,150	13	0,080	8
8	[4,3 – 4,6[	4,450	6	0,037	3,7
			161	1	100

Représentation graphique

a) Histogramme et polygone des fréquences

- L'histogramme est un ensemble de rectangles ayant pour largeur l'amplitude (étendu) de la classe et pour hauteur l'effectif de la classe.

- Les lignes joignant les milieux des bases supérieures des différents rectangles adjacents, la forme ce qu'on appelle polygone des fréquences (absolues ou relatives).



On parle de données discrètes lorsque le nombre possible de valeurs est soit fini soit dénombrable (c'est- à dire le nombre de valeurs est 0 ou 1 ou 2 et ainsi de suite).

On parle de données continues lorsqu'on a un nombre infini de valeurs possibles qui correspondent à une échelle continue de valeurs sans « trou », « interruption » ou « saut ».

Exemples : 1. Données discrètes : le nombre d'œufs pondus par des poules est direct parce qu'il correspond à un comptage.

2. Données continues : les volumes de lait produits par les vaches sont continus parce que ce sont des mesures qui peuvent prendre n'importe quelle valeur dans un intervalle continu.

Une autre façon courante de classer les données est d'utiliser 4 niveaux de mesure : nominal, ordinal, intervalle et rapport.

***) Le niveau nominal de mesure :** est caractérisé par des données qui consistent en noms, labels ou catégories seulement. Les données ne peuvent pas être arrangées suivant un ordre (comme du plus grand au plus petit).

***) Le niveau ordinal de mesure :** dans lequel, on peut arranger les données selon un certain ordre, sous réserve que les différences entre les valeurs soient non déterminées ou qu'elles soient sans signification.

***) Le niveau intervalle de mesure :** est semblable au niveau ordinal avec la propriété supplémentaire que la différence entre deux valeurs a un sens. Cependant, à ce niveau, les données n'ont pas de zéro naturel de référence (pour le quel aucune quantité n'est présente.

***) Le niveau rapport de mesure :** est semblable au niveau intervalle avec la propriété supplémentaire qu'il y a un zéro naturel de référence pour lequel aucune quantité n'est présente. Pour les valeurs à ce niveau, les différences et les rapports ont un sens.

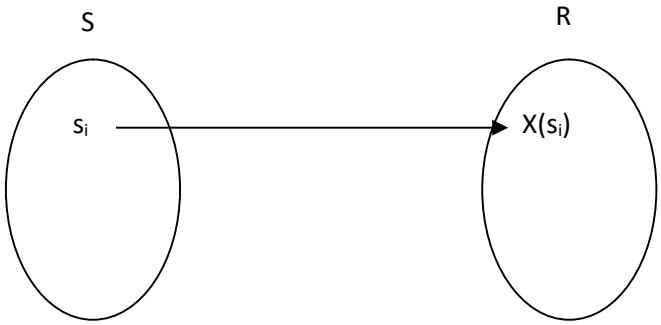
Niveau	Résumé	Exemple	Remarque
Nominal	Catégories seulement. Les données ne peuvent pas être ordonnées.	États où on a rencontré des ours : 5 New York 20 Idaho 40 Wyoming	Catégories ou noms seulement.
Ordinal	Les catégories sont ordonnées mais les différences n'ont pas de sens.	Les ours selon leur agressivité : 5 non agressifs 20 un peu agressifs 40 fortement agressifs	Un ordre est déterminé par « non », « un peu », « fortement ».
Intervalle	Les différences ont un sens mais il n'y a pas de zéro naturel de référence et les rapports n'ont pas de sens.	La température de la tanière des ours : −15 °C −7 °C 4 °C	0 °C ne signifie pas « pas de chaleur ». 40 °C n'est pas deux fois plus chaud que 20 °C.
Rapport	Il y un zéro naturel de référence et les rapports ont un sens.	La distance de migration des ours : 8 km 32 km 64 km	60 km est deux fois plus long que 30 km.

2.2. Les variables aléatoires

2.2.1. Définition

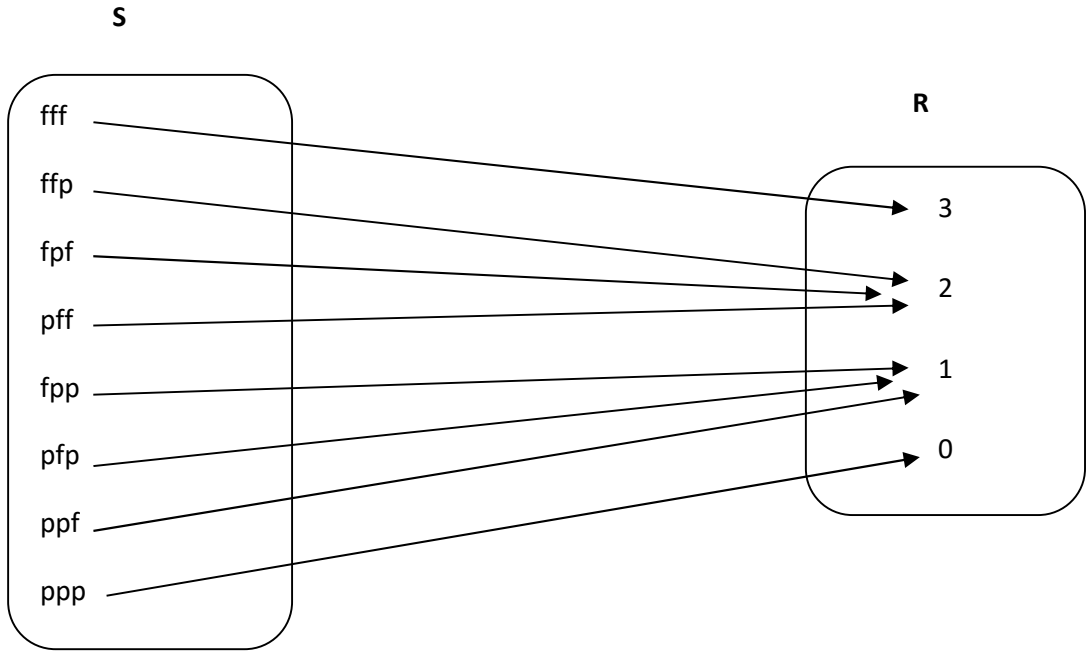
Une variable aléatoire VA est une fonction d'un ensemble S dans R :

$X : s \longrightarrow R \quad : S_i \longrightarrow x(s_i)$



Autrement dit, c'est un processus par lequel on associe à tout élément de S un nombre réel.

Exemple : on lance une pièce de monnaie 3 fois d'affilés, alors la variable aléatoire (VA) X représentant le nombre de coté **face** obtenu est la fonction qui établit les correspondances suivantes :



Exemples :

Ensemble de référence S	**Caractère**	**VA associée**
Elève de classe	Taille	Taille en cm

Etudiants	Sexe	Sexe codé 0 et 1
Participation au marathon...etc.	Rythme cardiaque	Nbre de battement par minute à l'arrivée

2.3. Paramètres caractéristiques

2.3.1. Paramètres de position

Les paramètres de position (mode, médiane, moyenne) permettent de savoir autour de quelles valeurs se situent les valeurs d'une variable statistique.

2.3.1.1. Le mode

Le mode, noté Mo, est la modalité qui admet la plus grande fréquence :

f (Mo) = Max (f*i*) ; i ∈ [1, p]

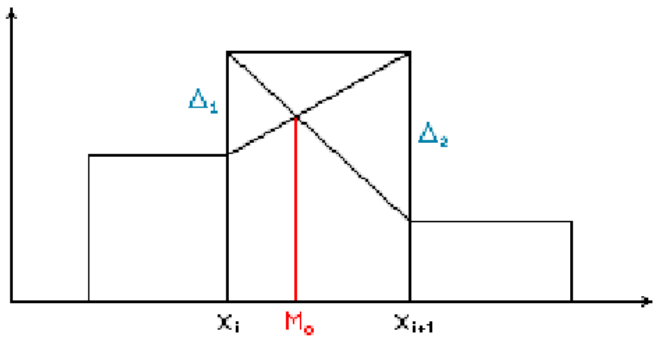
Il est parfaitement défini pour une variable qualitative ou une variable quantitative discrète.

Pour une variable quantitative continue nous parlons de classe modale : c'est la classe dont la densité de fréquence est maximum.

Si les classes ont même amplitude la densité est remplacée par l'effectif ou la fréquence et nous retrouvons la définition précédente.

Nous définissons le mode, pour une variable quantitative continue, en tenant compte des densités de fréquence des 2 classes adjacentes par la méthode suivante.

- **Données rangées** : valeur la plus fréquente.
- **Données condensées** : valeur ayant la plus grande fréquence.
- **Données en classes** : La classe modale [xi, xi + 1[étant déterminée, le mode Mo vérifie :



Dans une proportion, on ne change pas la valeur du rapport en additionnant les numérateurs et en additionnant les dénominateurs :

$$Mo = L + \left(\frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \right) \times h$$

où :

- L = borne inférieure de la classe modale,
- f_1= fréquence de la classe modale,
- f_0= fréquence de la classe précédente,
- f_2= fréquence de la classe suivante.

2.3.1.2. La médiane

La médiane M_e est telle que l'effectif des observations dont les modalités sont inférieures à M_e est égal à l'effectif des observations dont les modalités sont supérieures à M_e.

- **Données rangées :**
  - Si n impair → valeur centrale.
  - Si n pair → moyenne des deux valeurs centrales.
- **Données condensées :**
  - On utilise l'effectif cumulé pour repérer la médiane.
- **Données en classes :**

$$Med = L + \left(\frac{\frac{n}{2} - C}{f_m} \right) \times h$$

où :

- L = borne inférieure de la classe médiane,
- C = effectif cumulé avant la classe médiane,
- fm = fréquence de la classe médiane,

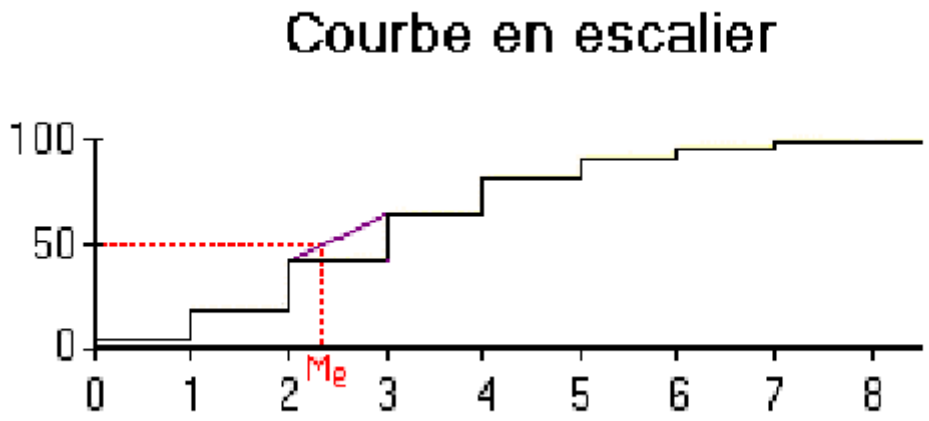
- h = amplitude de la classe.

***) Cas d'une variable discrète**

Exemple : En vue d'établir rationnellement le nombre de postes de travail nécessaires pour assurer à sa clientèle un service satisfaisant, une agence de voyage a fait relever, minute par minute, le nombre d'appels téléphoniques reçus au cours d'une période de 30 jours. Cette opération a fourni, pour la tranche horaire de pointe qui se situe entre onze heures et midi, les résultats suivants :

Nombre d'appels téléphoniques par minute	Nombre de minutes (effectifs)	Fréquence (%)	Fréquence cumulée (%)
0	93	5,2	5,2
1	261	14,5	19,7
2	416	23,1	42,8
3	393	21,8	64,6
4	308	17,1	81,7
5	174	9,7	91,4
6	93	5,2	96,6
7	42	2,3	98,9
8 et plus	20	1,1	100,0
TOTAL	1 800	100,0	

La fréquence cumulée est 42,8 % pour x = 2, et 64,6 % pour x = 3. L'intervalle [2, 3 [est appelé **intervalle médian.** Dans l'intervalle médian, la médiane est calculée par **interpolation linéaire**.



***) Cas d'une variable continue**

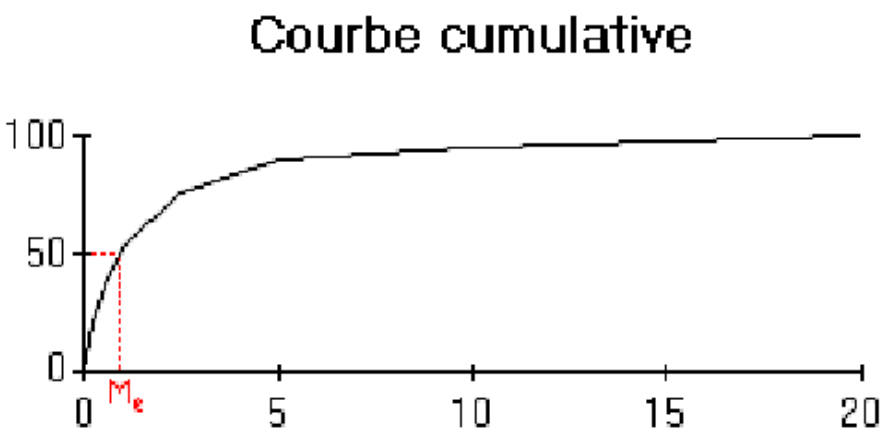
Exemple :

La Fédération nationale de la réparation et du commerce de l'automobile a effectué une enquête auprès de ses adhérents visant à mieux connaître la structure de ce secteur. Cette opération a fourni la répartition suivante des entreprises de la réparation de du commerce de l'automobile selon leur chiffre d'affaires annuel. La masse de chiffres d'affaires correspondant aux entreprises de la première et de la dernière classe s'élève respectivement à 1 714 et 110 145 millions de francs.

Chiffre d'affaires (millions de F)	Nombre d'entreprises (effectif)	Fréquence (%)	Amplitude de classe	Densité de fréquence	Fréquence cumulée
Moins de 0,25	13 712	20,3	0,25	81,2	20,3
0,25 à moins de 0,50	10 674	15,8	0,25	63,2	36,1
0,50 à moins de 1,00	11 221	16,6	0,50	33,2	52,7
1,00 à moins de 2,50	15 496	22,9	1,50	15,3	75,6
2,50 à moins de 5,00	10 043	14,8	2,50	5,9	90,4
5,00 à moins de 10,00	3 347	4,9	5,00	0,98	95,3
10,00 et plus	3 147	4,7	50,0	0,094	100,0
TOTAL	67 640				

La fréquence cumulée est 36,1 % pour x = 0,50, et 52,7 % pour x = 1,00.

L'intervalle [0,50, 1,00 [est **l'intervalle médian.** Dans l'intervalle médian, la médiane est calculée par **interpolation linéaire**.



2.3.1.3. La moyenne

La moyenne ne se définit que pour une variable statistique quantitative.

Pour une variable statistique discrète {(xi, ni)}1 ≤ i ≤ p à valeurs dans IR , la moyenne est la moyenne arithmétique des modalités pondérées par les effectifs.

- **Données rangées** :

$$\bar{X} = \frac{\sum x_i}{n}$$

Exemple : X={5,6,7,8,9} → X¯=355=7\bar{X} = \frac{35}{5} = 7X¯=535=7.

- **Données condensées (fréquences)** :

$$\bar{X} = \frac{\sum f_i x_i}{\sum f_i}$$

Exemple : L'étude de 21 familles a conduit à la distribution suivante le nombre d'enfants dans la famille :

Nombre d'enfants x_i	0	1	2	3	4	5
Nombre de familles n_i	5	3	6	1	3	3

Le nombre moyen d'enfants par famille est

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{i=p} n_i \, x_i$$

$$= \frac{1}{21} (0 \times 5 + 1 \times 3 + 2 \times 6 + 3 \times 1 + 4 \times 3 + 5 \times 3 + 5 \times 3) = \frac{45}{21} = \frac{15}{7}$$

2.3.1.4. Quantiles

Les **quantiles** sont des valeurs qui partagent une distribution de données en intervalles égaux. Ils permettent d'analyser la répartition des données, d'identifier la dispersion et de mieux comprendre la distribution d'une variable biologique (par exemple : taux de glucose, taille, poids, concentration d'un biomarqueur, etc.).

Les quantiles les plus utilisés sont :

- **Médiane (Q_2)** : sépare les données en deux parties égales.
- **Quartiles (Q_1, Q_2, Q_3)** : divisent les données en quatre parties.
- **Déciles (D_1, …, D_9)** : divisent en dix parties.
- **Centiles ou percentiles (P_1, …, P_{99})** : divisent en cent parties.

a) Données non groupées (brutes)

Supposons un échantillon de taille **n** avec les valeurs classées en ordre croissant :

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

Le quantile d'ordre **p** (où 0<p<10 < p < 10<p<1) correspond à la position :

$$k = p \times (n+1)$$

- Si **k** est un entier, le quantile est la valeur x(k).
- Si **k** n'est pas entier, on interpole entre les valeurs voisines.

Exemple :

Données (taux de cholestérol en g/L chez 10 patients) :

1.4 ; 1.6 ; 1.7 ; 1.9 ; 2.0 ; 2.1 ; 2.3 ; 2.5 ; 2.7 ; 3.0

- Pour la **médiane (Q2, p=0.5)** :
  $k = 0.5 \times (10+1) = 5.5$
  On prend la moyenne de $x_{(5)} = 2.0$ et $x_{(6)} = 2.1$.
  $$Q2 = \frac{2.0 + 2.1}{2} = 2.05$$
- Pour le **1er quartile (Q1, p=0.25)** :
  $k = 0.25 \times (11) = 2.75$
  On interpole entre $x_{(2)} = 1.6$ et $x_{(3)} = 1.7$.
  $$Q1 = 1.6 + 0.75 \times (1.7 - 1.6) = 1.675$$

b). Données condensées (valeurs distinctes avec effectifs)

Méthode

Quand plusieurs valeurs se répètent, on utilise la **distribution de fréquences cumulées**.

La position du quantile est donnée par :

$$k = p \times N$$

Avec *N*= effectif total.

Exemple

Taux d'hémoglobine (g/dL) mesurés sur 20 patients :

Valeur	Effectif	Effectif cumulé
10	3	3
11	5	8

Valeur	Effectif	Effectif cumulé
12	6	14
13	4	18
14	2	20

- **Médiane (Q2, p=0.5)**

  $k = 0.5 \times 20 = 10.$

  Le 10e individu est dans la valeur 12 (car cumul = 14 ≥ 10).

  $$Q2 = 12$$

- **Premier quartile (Q1, p=0.25)**

  $k = 0.25 \times 20 = 5.$

  Le 5e individu est dans la valeur 11 (cumul = 8 ≥ 5).

  $$Q1 = 11$$

3.2 Données groupées (par classes)

Lorsque les données sont condensées en classes (tableau de fréquences), on utilise la formule :

$$Q_p = L + \left(\frac{p \times N - F}{f} \right) \times h$$

Avec :

- L = borne inférieure de la classe contenant le quantile
- N = effectif total
- F = effectif cumulé avant la classe du quantile
- f = effectif de la classe du quantile
- h = amplitude de la classe

Exemple :

Concentration en protéine (g/L) répartie chez 40 patients :

Classe (g/L)	Effectif
10–20	5
20–30	8
30–40	12
40–50	10

Classe (g/L)	Effectif
50–60	5

- Médiane (Q2 → position = $N/2 = 20$)

  Le 20e individu se trouve dans la classe [30–40] (car cumul = 25).

Formule :

$$Q2 = L + \left(\frac{20 - 13}{12}\right) \times 10$$

$$Q2 = 30 + \left(\frac{7}{12}\right) \times 10 = 35.83$$

2.3.2. Paramètres de dispersion (ou de variation)

2.3.2.1. Étendue

$$E = X_{\max} - X_{\min}$$

2.3.2.2. Variance et Écart-type

- **Données rangées** :
- **Données condensées :**

$$\sigma^2 = \frac{1}{n}\sum (x_i - \bar{X})^2 \quad ; \quad \sigma = \sqrt{\sigma^2}$$

$$\sigma^2 = \frac{\sum f_i (x_i - \bar{X})^2}{\sum f_i}$$

- **Données en classes :**

$$\sigma^2 = \frac{\sum f_i (m_i - \bar{X})^2}{\sum f_i}$$

où mi est le centre de classe.

2.3.2.3. Coefficient de Variation

$$CV = \frac{\sigma}{\bar{X}} \times 100$$

Permet de comparer deux distributions de natures différentes.

2.3.2.4. Ecart semi-interquartile

La dispersion d'une variable est souvent décrite par des mesures robustes (moins sensibles aux valeurs extrêmes). L'**écart semi-interquartile (ESI)** est l'une d'elles.

Il est défini comme la moitié de l'**étendue interquartile (IQR)**, c'est-à-dire la moitié de la différence entre le troisième quartile (Q3) et le premier quartile (Q1) :

$$ESI = \frac{Q3 - Q1}{2}$$

a). Cas des données brutes (non groupées)

Étapes de calcul

1. Trier les données en ordre croissant.
2. Déterminer les positions de Q1 et Q3 :

- $k_1 = 0.25 \times (n+1)$
- $k_3 = 0.75 \times (n+1)$

Exemple

Données (poids de 8 souris en g) :

20, 22, 23, 25, 27, 30, 31, 35

- Taille de l'échantillon : $n = 8$.
- Position de Q1 : $0.25 \times (8+1) = 2.25$ → entre 2e et 3e valeurs (22 et 23).

$$Q1 = 22 + 0.25 \times (23 - 22) = 22.25$$

- Position de Q3 : $0.75 \times (9) = 6.75$ → entre 6e et 7e valeurs (30 et 31).

$$Q3 = 30 + 0.75 \times (31 - 30) = 30.75$$

- Écart semi-interquartile :

$$ESI = \frac{30.75 - 22.25}{2} = \frac{8.5}{2} = 4.25$$

b). Cas des données condensées (fréquences simples)

Étapes de calcul

Construire un tableau de fréquences et de fréquences cumulées.

Déterminer la position de Q1 et Q3 :

- $k_1 = 0.25 \times N$
- $k_3 = 0.75 \times N$

Exemple

Concentrations de protéine (g/L) chez 16 patients :

Valeur	Effectif	Cumul
12	3	3
13	5	8
14	4	12
15	3	15
16	1	16

- $N = 16.$
- Position Q1 : $0.25 \times 16 = 4.$ → Q1 = 13 (car le 4ᵉ individu est dans la valeur 13).
- Position Q3 : $0.75 \times 16 = 12.$ → Q3 = 14 (car le 12ᵉ individu est dans la valeur 14).
- $ESI = (14 - 13)/2 = 0.5.$

c). Cas des données groupées en classes

Exemple

Glycémie (mg/dL) de 40 patients :

Classe	Effectif	Cumul
60–80	6	6
80–100	10	16
100–120	12	28
120–140	8	36
140–160	4	40

- $N = 40$.
- Position Q1 : $0.25 \times 40 = 10$. → Q1 dans [80–100].

$$Q1 = 80 + \left(\frac{10 - 6}{10}\right) \times 20 = 80 + \frac{4}{10} \times 20 = 88$$

- Position Q3 : $0.75 \times 40 = 30$. → Q3 dans [120–140].

$$Q3 = 120 + \left(\frac{30 - 28}{8}\right) \times 20 = 120 + \frac{2}{8} \times 20 = 125$$

- Écart semi-interquartile :

$$ESI = \frac{125 - 88}{2} = \frac{37}{2} = 18.5$$

3. Paramètres de Forme

a) Asymétrie (Skewness)

$$Sk = \frac{\frac{1}{n}\sum(x_i - \bar{X})^3}{\sigma^3}$$

- Sk = 0 → distribution symétrique.
- Sk > 0 → asymétrie à droite.
- Sk < 0 → asymétrie à gauche.

b) Aplatissement (Kurtosis)

$$K = \frac{\frac{1}{n}\sum(x_i - \bar{X})^4}{\sigma^4}$$

- K = 3 → mésokurtique (comme la normale).
- K > 3 → leptokurtique (pointue).
- K < 3 → platykurtique (aplatie).

2.4. Notions fondamentales : population et échantillon

2.4.1. La 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.

Exemples : l'ensemble des agriculteurs, l'ensemble des champs d'une région ou encore l'ensemble des plantes d'une parcelle.

Une population est finie ou infinie selon que le nombre de ses éléments est fini ou non.

2.4.2. L'échantillon

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

L'objectif est, à partir de la connaissance acquise sur l'échantillon, de faire des déductions correctes sur la population origine.

D'après Snedecor « c'est l'échantillon qu'on observe, mais c'est la population que l'on veut étudier ».

L'échantillon doit être une réduction de la population. Il doit être représentatif. Lorsqu'il provient d'un tirage représentatif.

Un tirage est représentatif lorsque tout élément de la population à représenter peut figurer dans l'échantillon et ce avec une probabilité connue.

3. Statistiques inférentielles

3.1. Lois de probabilités

3.1.1. Loi normale (loi continue)

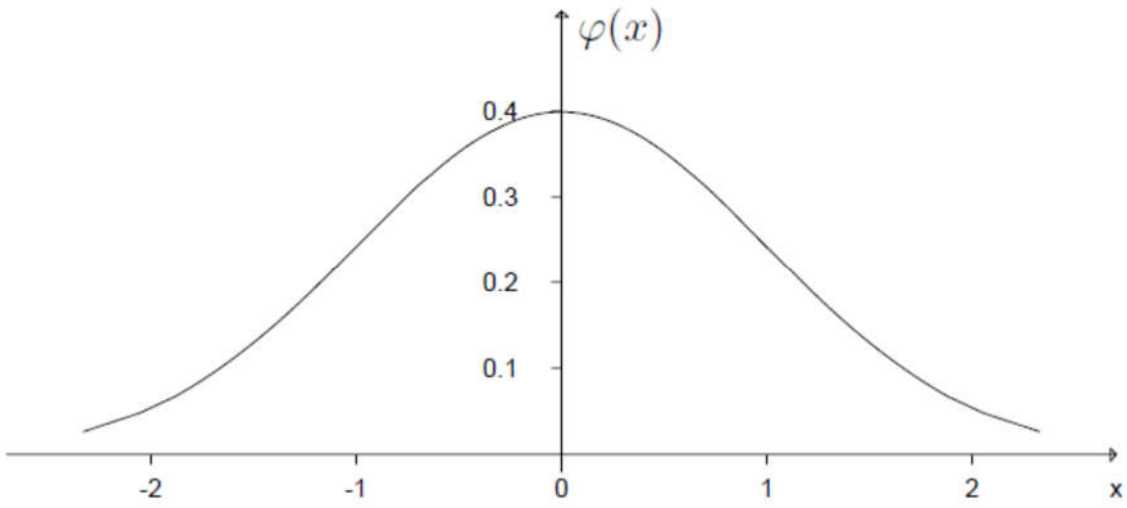
La loi normale ou la loi de Laplace-Gauss s'applique à une variable aléatoire continue qui est la résultante d'un grand nombre de causes indépendantes dont les effets s'additionnent, aucun n'étant prépondérant. De telles conditions sont très souvent rencontrées, la loi normale est le modèle mathématique de très nombreux phénomènes.

Si un phénomène suit une loi normale, on dit que les valeurs sont normalement distribuées.

La loi normale se définit par l'équation :

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

La loi normale est donc caractérisée par ces deux paramètres, la moyenne (μ) et l'écart-type (σ) et elle est notée N (μ, σ^2).



A partir de ce graphique, on peut déduire que :

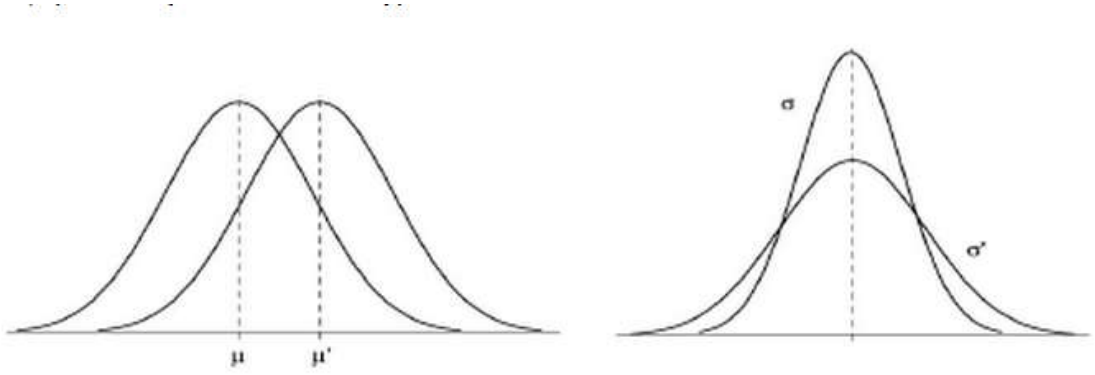
*) La courbe a une forme de cloche (nommée cloche de Gauss) ;

*) Quand X= μ, la courbe atteint son maximum ;

*) Le mode, la médiane et la moyenne sont les même ;

*) La courbe a deux points d'inflexion qui sont μ-σ et μ+σ ;

*) La forme de la courbe dépend de μ et de σ.



En effectuant le changement de variable Z=(X-μ)/σ, la variable Z, variable normale de moyenne nulle et d'écart-type 1 suit la loi normale centrée réduite et on note Z=N(0,1).

Par le changement de variable précédente, toutes les distributions normales se ramènent à une seule : la distribution normale centrée réduite.

3.1.1.1. Calcul d'une probabilité dans le cas d'une loi normale

La probabilité de la variable aléatoire X prend une valeur entre deux nombres a et b est définie par :

$$p(a < X < b) = \int_a^b f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

Puisque les valeurs de probabilité changent d'une courbe normale à une autre selon μ et σ. On a pensé à la loi normale centrée réduite caractérisée par μ=0 et σ=1.

3.1.1.2. Loi normale centrée réduite

Si X suit la loi normale, alors sa transformation en Z suit la loi normale centrée réduite (CR).

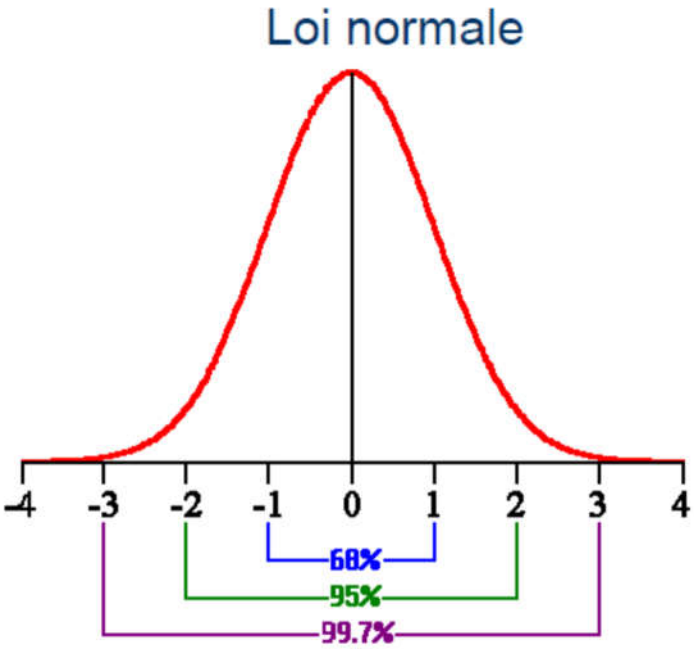
X $\longrightarrow$ $Z = \frac{X-\mu}{\sigma}$, μ=0 et σ=1

N(μ,σ) $\longrightarrow$ N(0,1)

On ramène donc le calcul de la probabilité sur X à un calcul de probabilité sur Z :

$$P(a < X < b) = p\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) = p\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

On remarque que la très grande partie de l'aire comprise entre cette courbe et l'axe horizontale s'étale entre les points d'abscisse -3 et +3, on peut donc négliger toute surface correspondant à des valeurs Z à l'extérieur de cet intervalle.



Pour évaluer une probabilité en termes de surface sous la courbe de la loi normale centrée réduite, on se base sur la table de la variable normale centrée réduite qui donne exclusivement la surface sous la courbe à gauche de la valeur Z non négative.

1.1.3.3. Evaluation d'une aire à gauche d'une valeur ou P(Z ≤ z) :

- Si z ≥ 0 alors P(Z ≤ z) = F_z(Z)……..elle se lit directement sur la table VNCR.

Exemple : P(Z ≤ 1,35) = F_z (1,35) = 0,9115.

- Si z < 0 P(Z ≤ z) = F_z (-Z) = 1-F(Z).

Exemple : Z = -2,13

P(Z ≤ -2,13) = F_z (-Z) = 1- F_z (2,13) = 1- 0,9834 = 0,0166.

1.1.3.4. Evaluation d'une aire à droite d'une valeur Zou P (Z > z) :

- Si z ≥ 0 P(Z > z) = 1 – p(Z ≤ z).

Exemple : z = 0,56

$P(Z > 0,56) = 1 – P(Z \leq 0,56) = 1 - F_z(0,56) = 1 – 0,7213 = 0,2877.$

- Si $z < 0$, $P(Z > -z) = P(Z \leq z)$.

Exemple : $P(Z > -1,75) = P(Z \leq 1,75) = F_z(1,75) = 0,9599$.

3.2. Inférence statistique

3.2.1. Tests d'hypothèses

Il s'agit d'une étape très importante qui sert à éclaircir les décisions qui peuvent être prise dans les différents domaines, et ceux avec le plu de la décision possible. Pour cela, la confirmation ou l'information d'une hypothèse est toujours fait avec une certaine probabilité que l'on voudra aussi forte que possible.

3.2.2. Définition de l'hypothèse

En pratique, on a 2 types d'hypothèses exclusives H_0 et H_1 :

H_0 : hypothèse nulle qui sera rejetée uniquement et qui n'amène pas de changement et d'action à entreprendre dans le cas contraire (si on l'accepte). C'est l'hypothèse à vérifier.

H_1 : hypothèse alternative ou contre-hypothèse qui sera acceptée lorsque H_0 est rejetée. C'est une hypothèse qui amène un changement et qui implique une action à entreprendre.

3.3. Comparaison d'une répartition observée à une répartition théorique (test du X^2)

On veut savoir si une répartition expérimentale est bien conforme à une répartition théorique par le biais du test du X^2. Si on suppose que la répartition de la population suit une loi théorique donnée, on va observer un écart entre l'effectif observé d'une classe et l'effectif théorique de cette même classe. Dans ce cas on est amené à utiliser la somme des écarts quadratique entre l'effectif observé et théorique qui n'est autre que le X^2 observé.

$$X^2 = \frac{\sum(Oi - Ci)^2}{Ci}$$

O_i : effectif observé ;

C_i : effectif théorique.

Le test du X^2 se fait selon les étapes suivantes :

- On pose l'hypothèse nulle H_0 ;

H_0 : il y a conformité entre la répartition théorique et observée.

- Il faut fixer α à l'avance.

- On calcule le X^2 observé.

- Au seuil α et à un degré de liberté ddl correspondant, on lit sur la table du X^2, le X^2 théorique.

- La conclusion sera ainsi :

a) $X^2_{\text{observé}} \geq X^2_{\alpha \text{ théorique}} \longrightarrow$ H_0 est rejetée

b) $X^2_{\text{observé}} < X^2_{\alpha \text{ théorique}} \longrightarrow$ H_0 est acceptée.

Remarque :

Pour appliquer le test X^2, l'effectif théorique par classe doit au moins égal à 5 ; $C_i \geq 5$.

Exemple : On a croisé 2 variétés de plantes différentes ayant comme caractère A et B.

La 1ère génération est homogène. La 2ème génération fait apparaitre 4 phénotypes : AB, Ab, aB et ab.

Si les caractères se transmettent selon les lois Mendel, les proportions théoriques de 4 phénotypes sont : 9/16, 3/16, 3/16 et 1/16.

L'expérience sur un échantillon de 160 plantes a donné : AB : 100, Ab : 18, aB : 24 et ab : 18.

Cette répartition est-elle conforme aux lois de Mendel à un seuil de signification de 5% ?

Solution :

H_o : la répartition observée est conforme aux lois de Mendel avec α = 0,05.

Phénotype	**AB**	**Ab**	**aB**	**ab**	**Total**
Proportion théorique	9/16	3/16	3/16	1/16	1
Effectif théorique	9/16*160 = 90	3/16*160 = 30	3/16*160 = 30	1/16*160 = 10	160

C_i					
Effectif observé O_i	100	18	24	18	160

$$X^2obs = \frac{\sum(Oi - Ci)^2}{Ci} = \frac{(100-90)^2}{90} + \frac{(18-30)^2}{30} + \frac{(24-30)^2}{30} + \frac{(18-10)^2}{10}$$

$X^2obs = 12{,}51$

ddl= k-1= 4-1= 3

α = 0,05

$X^2_{0,05\ ;3} = 7{,}815$ (théorique, lu sur la table de X^2).

$X^2_{\text{observé}} > X^2_{\alpha\ \text{théorique}} \longrightarrow H_0$ est rejetée au seuil de signification α = 5% ou bien H_0 est rejetée au seuil de sécurité de 95%.

3.4. Comparaison d'un pourcentage observé à un pourcentage théorique

La comparaison entre un pourcentage (ou proportion) observé **p** sur un échantillon expérimental et le pourcentage théorique $\mathbf{p_0}$ de la population de l'échantillon est basée sur l'écart réduit **ε**. A savoir :

$$\varepsilon = \frac{(\mathrm{p} - \mathrm{po})}{\sqrt{\frac{\mathrm{poqo}}{\mathrm{n}}}}$$

Au seuil de signification de 5% et avec $q_o = 1 - p_o$

Au seuil de 5% :

Si ε < 1,96 (≈ 2) la différence n'est pas significative ;

Si ε ≥ 1,96 la différence est significative.

Au seuil de 1% :

Si ε < 2,576 ($\approx 2{,}6$) la différence n'est pas significative ;

Si ε ≥ 2,576 la différence est significative.

Exemple : une race de souris présente des tumeurs spontanées avec un taux parfaitement connu soit p_o = 20%. Dans une expérience portant sur 100 souris, on observe 34 atteintes, soit p = 34%. On demande si la différence entre p_o et p est significative ou non.

Solution :

H_o : pas de différence significative entre p et p_o

$$\varepsilon = \frac{(0{,}34 - 0{,}20)}{\sqrt{\frac{0{,}2 \times 0{,}8}{100}}} = 3{,}50$$

ε = 3,5 > 1,96 ⟶ on rejette H_o, donc la différence est significative entre p et p_o au seuil de 5%.

Appliquons le même exemple en employant le X^2.

Solution :

	Tumeur	Pas de tumeur	total
Effectif théorique C_i= np	20	80	100
Effectif observé O_i	34	66	100
% théorique « p_o »	20	80	100%

$$X^2 obs = \frac{\sum(Oi - Ci)^2}{Ci} = \frac{(34-20)^2}{20} + \frac{(66-80)^2}{80}$$

$X^2_{0,05\ ;1} = 3{,}841$ (ddl=2-1=1)

Remarque :

On remarque que le $X^2_{obs} = \varepsilon^2$ ($12{,}25 = (3{,}50)^2$) et que le $X^2_{lu} = t^2$ ($3{,}841 = (1.96)^2$).

En effet, la méthode de comparaison par l'écart réduit et le test du X^2 sont absolument superposables.

3.5. Comparaison d'une moyenne observée à une moyenne théorique

Soit à comparer un échantillon expérimental de moyenne X à une population dont la moyenne μ et l'écart-type σ sont connus.

Prenons le cas des grands échantillons où n ≥ 30, la moyenne X suit donc une loi normale

$X \longrightarrow N(m, \frac{\sigma\ pop}{\sqrt{n}})$

La transformation t qui correspond à la valeur critique de Student suit une loi normale centrée réduite

t ⟶ N(0,1)

$t = \frac{(X-Xo)}{\frac{\sigma}{\sqrt{n}}} = \frac{(X-Xo)}{\sqrt{\frac{SCE}{n(n-1)}}}$ (Variable de Student)

SCE : la somme carrée des écarts

H_o : m = m_o (α = 5%)

Si $t_{1-\alpha} < t_{obs}$ ⟶ la différence n'est pas significative, H_o est acceptée.

Si $t_{1-\alpha} \geq t_{obs}$ ⟶ la différence est significative, H_o est rejetée.

Exemple :

On estime les hauteurs des arbres d'un échantillon de 12 arbres d'une pinède naturelle.

20,4 ; 25,4 ; 25,6; 25,6 ; 26,6 ; 28,6 ; 28,7 ; 29,0 ; 29,8 ; 30,5 ; 30,9 et 31,1 m.

On demande si la moyenne observée est compatible avec l'hypothèse que la moyenne des hauteurs soit égale à 29 m (α = 5%).

Solution :

SCE = 106,16 ; n(n-1) = 132 ; X = 27,68.

H_0 : X_o = 29

$t_{obs} = \frac{(27.68-29)}{\sqrt{\frac{106,16}{132}}} = -1,47$

$t_{1-\alpha}$ = 1,80 (au ddl=12-1=11) ⟶ La différence n'est pas significative, H_o est acceptée. La moyenne observée est donc proche de la moyenne théorique au seuil α = 5%.

3.6. T-test pour échantillons indépendants (Student/Welch)

Objectif

Comparer la moyenne d'une variable quantitative entre **deux groupes indépendants** (p. ex. témoin vs traitement).

Hypothèses

- Observations **indépendantes** entre groupes.
- Variable **continue** (ou assimilée).
- **Student** (variances égales) : normalité approximative et $\sigma_1^2 = \sigma_2^2$.
- **Welch** (variances inégales) : normalité approximative, $\sigma_1^2 \neq \sigma_2^2$ (plus robuste, recommandé par défaut).

Formules

Soient deux échantillons indépendants :

tailles n_1, n_2, moyennes $\bar{x}_1, \bar{x}_2$, écarts-types s_1, s_2.

3.6.1. Test de Student (variances supposées égales)

Pondération/variance « poolée » :

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \quad s_p = \sqrt{s_p^2}.$$

Statistique de test (bilatérale par défaut) :

$$t = \frac{\bar{x}_2 - \bar{x}_1}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \qquad \text{ddl} = n_1 + n_2 - 2.$$

IC à 95 % pour $\mu_2 - \mu_1$:

$$(\bar{x}_2 - \bar{x}_1) \pm t_{0.975,\,\text{ddl}} \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

3.6.2. Test de Welch (variances inégales)

Erreur-type :

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Statistique :

$$t = \frac{\bar{x}_2 - \bar{x}_1}{SE}.$$

Degrés de liberté (Welch–Satterthwaite) :

$$\text{ddl} \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}.$$

IC à 95 % :

$$(\bar{x}_2 - \bar{x}_1) \pm t_{0.975,\,\text{ddl}} \cdot SE.$$

Exemple

Question. Un traitement nutritionnel augmente-t-il l'**activité enzymatique** des plantes ?
Groupes indépendants : **Témoin** (n_1=8) vs **Traitement** (n_2=7).

Données (unités arbitraires) :

- Témoin : 12.1, 11.8, 13.0, 12.4, 11.6, 12.9, 12.2, 11.9
- Traitement : 13.5, 13.2, 13.1, 13.8, 12.7, 13.4, 13.0

Statistiques descriptives :

$$\bar{x}_1 = 12.2375,\, s_1 = 0.5041,\, n_1 = 8$$
$$\bar{x}_2 = 13.2429,\, s_2 = 0.3599,\, n_2 = 7$$

Student (si variances ≈ égales)

$$s_p^2 = \frac{(8-1)\,0.5041^2 + (7-1)\,0.3599^2}{8+7-2} \approx 0.1966, \quad s_p \approx 0.4434.$$
$$t = \frac{13.2429 - 12.2375}{0.4434\sqrt{1/8 + 1/7}} \approx 4.381, \qquad \text{ddl} = 13.$$

Décision (bilatéral) : $|t| = 4.381$ est très supérieur au seuil usuel → **différence significative.**

Welch:

$$SE = \sqrt{\tfrac{0.5041^2}{8} + \tfrac{0.3599^2}{7}} \approx 0.2242, \quad t = \frac{1.0054}{0.2242} \approx 4.484.$$
$$\text{ddl} \approx 12.56.$$

Conclusion identique : **le traitement augmente significativement l'activité enzymatique.**

3.7. T-test pour échantillons appariés

Objectif

Comparer la moyenne d'une variable **mesurée deux fois sur les mêmes sujets** (pré/post, avant/après traitement).

Hypothèses

- Paires correctement **appariées** (mêmes individus).
- **Normalité des différences** $d_i = post_i - pré_i$.
- Mesures à l'échelle **continue** (ou approximativement).

Formules

Statistique de test (bilatérale par défaut) :

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \qquad \mathrm{ddl} = n - 1.$$

IC à 95 % pour μ_d :

$$\bar{d} \pm t_{0.975,\, n-1} \frac{s_d}{\sqrt{n}}.$$

Taille d'effet (Cohen d_z) :

$$d_z = \frac{\bar{d}}{s_d}.$$

Exemple

Question. Un régime réduit-il la **masse** de rongeurs ?
10 rongeurs mesurés **avant** et **après** (g) :

Avant : 220, 215, 210, 225, 230, 218, 222, 219, 224, 216
Après : 214, 210, 206, 219, 222, 212, 216, 213, 219, 211

Différences di=Après−Avant :

[−6, −5, −4, −6, −8, −6, −6, −6, −5, −5]

Calculs :

- $n = 10$
- $\bar{d} = -5{,}7\ \text{g}$
- $s_d = 1{,}0593\ \text{g}$
- $SE = s_d/\sqrt{n} = 0{,}3350\ \text{g}$
- $t = \bar{d}/SE = -17{,}02\,;\ \text{ddl} = 9$

Décision (bilatérale) : $|t| = 17{,}02 \Rightarrow p \ll 0{,}001$ → **diminution significative** après le régime.

IC95 % :

$t_{0.975,9} \approx 2{,}262 \Rightarrow$

$\bar{d} \pm t \cdot SE = -5{,}7 \pm 2{,}262 \times 0{,}3350 \Rightarrow [-6{,}458\ ;\ -4{,}942]\ \text{g}.$

Taille d'effet :

$d_z = \bar{d}/s_d = -5{,}38$ (effet **très grand** en valeur absolue).

Interprétation. La masse moyenne diminue d'environ **5,7 g** (IC95 % **–6,46 à –4,94 g**). L'effet est très marqué.

3.8. Tests non paramétriques (Man-Whitney et Wilcoxon)

3.8.1. Test de Mann–Whitney (U)

But : Comparer deux groupes indépendants (échelle au moins ordinale) sans supposer la normalité.

Hypothèses : indépendance ; formes de distributions similaires (test de tendance centrale).

Procédure : combiner les deux échantillons, classer (rangs) et calculer :

Formules :

$U_1 = n_1\, n_2 + n_1(n_1 + 1)/2 - R_1$; $U_2 = n_1\, n_2 + n_2(n_2 + 1)/2 - R_2$; $U = \min(U_1, U_2)$

Pour grands échantillons : $Z = (U - \mu_U) / \sigma_U$ avec $\mu_U = n_1\, n_2 / 2$ et $\sigma_U = \sqrt{(n_1\, n_2\, (n_1 + n_2 + 1) / 12)}$.

Effet de taille : $r = Z / \sqrt{N}$ ($N = n_1 + n_2$) ; estimateur de localisation : Δ de Hodges–Lehmann (médiane des différences).

Exemple: Activité enzymatique (unités/arbitraires) chez des plantes traitées vs témoins.

Données simulées :

Témoin (n=8)	Traitement (n=7)
12.1	13.5
11.8	13.2

13.0	13.1
12.4	13.8
11.6	12.7
12.9	13.4
12.2	13.0
11.9	

Interprétation (exemple) : U = 10, p = 0,004 (indicatif). On conclut à une activité enzymatique plus élevée sous traitement. Rapporter en plus un effet de taille ($r = Z/\sqrt{N}$).

3.8.2. Test de Wilcoxon pour rangs signés

But : Comparer deux mesures appariées (pré/post) sans supposer la normalité des différences.

Hypothèses : appariement valide ; distribution symétrique des différences autour de la médiane.

Procédure : $d_i = x_i(\text{post}) - x_i(\text{pré})$; ignorer $d_i = 0$; classer $|d_i|$; sommer les rangs signés.

Formules :

W^+ = somme des rangs pour $d_i > 0$; W^- = somme des rangs pour $d_i < 0$; $T = \min(W^+, W^-)$.

Approximation normale pour $n > \sim 10$: $Z = (W - \mu_W) / \sigma_W$ avec $\mu_W = n(n+1)/4$ et $\sigma_W = \sqrt{(n(n+1)(2n+1)/24)}$.

Effet de taille : $r = Z / \sqrt{n}$; estimateur Hodges–Lehmann : médiane des (post−pré).

Exemple : Masse corporelle de 10 rongeurs avant et après un régime protéiné.

ID	Avant	Après
1	220	214
2	215	210
3	210	206
4	225	219
5	230	222

6	218	212
7	222	216
8	219	213
9	224	219
10	216	211

Interprétation (exemple) : W = 4, p = 0,002 (indicatif), suggérant une diminution de masse après le régime. Effet de taille : r ou médiane des différences.

3.9. Analyse de la variance aléatoire (ANOVA)

En sciences biologiques, les expériences impliquent souvent la comparaison de plusieurs traitements ou conditions expérimentales pour évaluer leurs effets sur une variable biologique d'intérêt (croissance, rendement, poids, taux enzymatique, etc.).

L'analyse de la variance (ANOVA) est une méthode statistique utilisée pour tester s'il existe une différence significative entre les moyennes de plusieurs groupes.

Lorsqu'on confronte une variable quantitative à une variable qualitative (nominale ou ordinale), on recourt très généralement à la comparaison de moyennes ou à l'analyse de variance.

Pour comparer plusieurs moyennes, on utilise la variance dont on mesure la différence entre ces moyennes. Si la différence est plus importante, les moyennes des échantillons sont plus dispersées. Cette dispersion est mesurée éventuellement à l'aide d'une variance. Pour juger si la variance est importante ou pas, il faudra disposer d'une référence : la variance dans l'hypothèse d'égalité des moyennes. Le test appliqué considéra donc en une comparaison de variance.

3.9.1. Principe d'analyse de la variance (ANOVA)

En fonction de chaque dispositif, on procède à l'établissement de l'ANOVA par le calccul de :

1) La moyenne générale

2) La somme des carrés des écarts des facteurs SCE (SCE_a, SCE_b, …….)

3) La moyenne des blocs s'il y en a

4) La somme des carrés des écarts de l'interaction SCE_{ab}

5) La somme des carrés des écarts résiduels SCE_r

6) Les carrés moyens factoriels (CM_a, CM_b, ….)

7) Les carrés moyens inter-factoriels CM_{ab}

8) Les carrés moyens résiduels CM_r

9) $F_{observé}$, $F_a = \frac{CMa}{CMr}$; $F_b = \frac{CMb}{CMr}$; $F_{ab} = \frac{CMab}{CMr}$

10) On résume tout dans un tableau appelé « Tableau de l'analyse la variance »

Source de variabilité	**ddl**	**SCE**	**CM**	**F_{obs}**	**P**
Variabilité factorielle Variabilité de l'interaction Variabilité du bloc Variabilité résiduelle		SCE_a ;SCE_b SCE_{ab} SCE_B SCE_r	CM_a ; CM_b CM_{ab} CM_B CM_r	F_a ; F_b F_{ab} F_B	
Variation totale		SCE_t	CM_t		

Finalement, on doit comparer le $F_{observé}$ au $F_{théorique}$ qui doit être lu à partir de la table de **Fischer-Snedecor**.

Si : $F_{obs} \geq F_t$ ⟶ Ho est rejetée. (Au seuil α = 5%)

Si : $F_{obs} < F_t$ ⟶ Ho est acceptée.

3.9.2. Conditions d'application de l'ANOVA :

Théoriquement, l'ANOVA ne peut s'appliquer que sur des données :

*) Qui sont normales, c'est-à- dire obéissent à une loi normale.

*) Qui sont indépendantes et qui n'ont aucun lien de dépendance ou de corrélation.

*) Qui ont la même variance (même dispersion) dans tous les traitements et dans tous les blocs. Cela veut dire que les erreurs doivent être de même ordre de grandeur ou presque, quel que soit le bloc ou le traitement.

3.9.3. ANOVA à un seul facteur

3.9.3.1. Dispositif expérimental complètement aléatoire (randomisation totale)

Lorsqu'un seul facteur expérimental est étudié, et que les traitements sont répartis de façon aléatoire sur les unités expérimentales, on parle d'un dispositif complètement aléatoire (DCA).

En d'autres termes, on cherche à tester l'hypothèse nulle :

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

Contre l'hypothèse alternative :

$$H_1 : \text{au moins une moyenne diffère.}$$

	A_1	A_2		A_j		A_k	
E_1	Y_{11}	Y_{21}		Y_{1j}		Y_{1k}	
E_2	Y_{21}	Y_{22}		Y_{2j}		Y_{2k}	
E_i	Y_{i1}	Y_{2i}		Y_{ij}		Y_{ik}	
E_n	Y_{n1}	Y_{n2}		Y_{nj}		Y_{nk}	
Moyenne	$\bar{Y}_{i1}$	$\bar{Y}_{i2}$		$\bar{Y}_{ij}$		$\bar{Y}_{ik}$	$\bar{Y}_{ij}$
Variance	S^2_{i1}	S^2_{i2}		S^2_{ij}		S^2_{ik}	S^2_{ij}

$\bar{Y}_{i1} = \frac{\sum_{i=1}^{n} Yi1}{ni1}$: Moyenne des observations de l'échantillon A_1

$\bar{Y} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} Yij}{ni}$: Moyenne générale

$S^2_j = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}(Yij-\bar{Y}c)^2}{ni-1}$: Variance de l'échantillon A_j

$S^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}(Yij-\bar{Y})^2}{n-1}$: Variance totale.

***) Estimation du modèle**

Le modèle de l'analyse de la variance à un facteur pour un dispositif complètement aléatoire se résume comme suit :

Variation totale (SCE_t) = Variation factorielle (SCE_a) + Variation résiduelle (SCE_r)

$CM_t = \frac{SCEt}{ddl}$; avec ddl = n-1

$CM_a = \frac{SCEa}{ddl}$; avec ddl = k-1

$CM_r = \frac{SCEr}{ddl}$; avec ddl = n-k

A partir de cela, on calcule $F_{observé}$

$F_{obs}= \frac{CMa}{CMr}$; comparé à la valeur théorique (F_t) de Fischer-Snedecor

Source de variabilité	**ddl**	**SCE**	**CM**	**F_{obs}**	**P**
Variabilité factorielle Variabilité résiduelle	k – 1 n – k	SCE_a SCE_r	CM_a CM_r	$F_a = \frac{CMa}{CMr}$	P_a
Variation totale	n – 1	SCE_t	CM_t		

Exemple Expérimental : Effet d'un engrais sur la croissance des plantes

Description de l'expérience

Un biologiste souhaite étudier l'effet de trois types d'engrais (A, B, C) sur la croissance d'une espèce végétale.

- Les engrais représentent le **facteur de traitement**.
- Chaque engrais est appliqué sur **4 plantes**, choisies **au hasard** parmi 12 plantes homogènes.
- La **hauteur moyenne (en cm)** après 30 jours est mesurée.

Traitement	**Répétition 1**	**Répétition 2**	**Répétition 3**	**Répétition 4**
Engrais A	15	18	16	17
Engrais B	20	19	21	22
Engrais C	13	14	12	15

Calcul de l'ANOVA

Étape 1 : Calcul des moyennes

$$\bar{X}_A = 16.5, \quad \bar{X}_B = 20.5, \quad \bar{X}_C = 13.5$$

$$\bar{X}_G = 16.83 \text{ (moyenne générale)}$$

Étape 2 : Calcul des sommes de carrés

Somme totale des carrés (SCT) :

$$SCT = \sum (X_{ij} - \bar{X}_G)^2 = 92.67$$

Somme des carrés due au traitement (SCTR) :

$$SCTR = n\sum(\bar{X}_i - \bar{X}_G)^2 = 4[(16.5 - 16.83)^2 + (20.5 - 16.83)^2 + (13.5 - 16.83)^2] = 87.34$$

Somme des carrés résiduelles (SCE) :

$$SCE = SCT - SCTR = 92.67 - 87.34 = 5.33$$

Étape 3 : Calcul des moyennes des carrés

$$CM_{traitement} = \frac{SCTR}{k-1} = \frac{87.34}{2} = 43.67$$

$$CM_{erreur} = \frac{SCE}{k(n-1)} = \frac{5.33}{9} = 0.59$$

Étape 4 : Calcul du rapport de Fisher (F)

$$F = \frac{CM_{traitement}}{CM_{erreur}} = \frac{43.67}{0.59} = 73.99$$

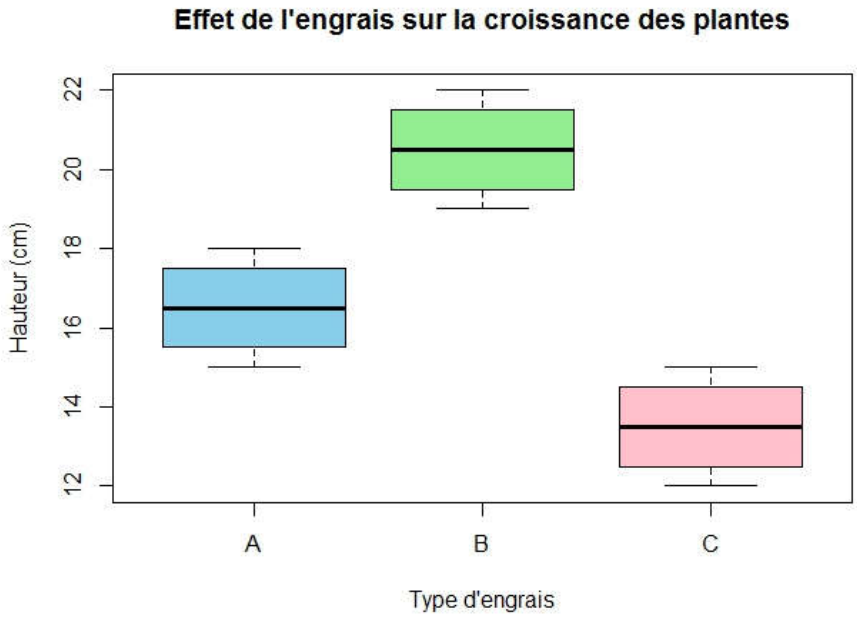
Valeur critique pour $F_{2,9;0.05} \approx 4.26$

Décision : Comme $F_{calculé} > F_{théorique}$, on rejette H_0.

Interprétation

Les différences observées entre les moyennes des engrais sont hautement significatives.

Ainsi, le type d'engrais a un effet réel sur la croissance des plantes.



3.9.3.2. Analyse de la Variance (ANOVA) à un seul facteur en dispositif en blocs aléatoires (RCBD)

1) Contexte et objectif

En sciences biologiques, il est courant que des sources de variabilité non contrôlables (ex. microclimat, gradient de fertilité, lots d'animaux, plateaux de culture) masquent l'effet d'un traitement.
Le dispositif en blocs aléatoires (RCBD) regroupe les unités expérimentales en blocs homogènes, chaque bloc recevant tous les traitements.
Objectif : tester si les moyennes des traitements diffèrent après avoir retiré la variabilité entre blocs.

Modèle statistique

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

- Y_{ij} : observation du **traitement** $i = 1, \ldots, t$ dans le **bloc** $j = 1, \ldots, r$
- μ : moyenne générale
- τ_i : effet du traitement i (contrainte $\sum \tau_i = 0$)
- β_j : effet du bloc j (contrainte $\sum \beta_j = 0$)
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ indépendantes

Hypothèses à tester :

$$H_0 : \tau_1 = \cdots = \tau_t = 0 \quad \text{vs} \quad H_1 : \exists i \neq k \ \tau_i \neq \tau_k$$

	T_1	T_2		T_j		T_k	Moyenne par bloc
$Bloc_1$	Y_{11}	Y_{12}		Y_{1j}		Y_{1k}	$\bar{Y}_{.1}$
$Bloc_2$	Y_{21}	Y_{22}		Y_{2j}		Y_{2k}	$\bar{Y}_{.2}$

$Bloc_i$	Y_{i1}	Y_{i2}		Y_{ij}		Y_{ik}	$\bar{Y}_{.j}$
$Bloc_{no}$	Y_{n1}	Y_{2n}		Y_{nj}		Y_{nk}	$Y_{.no}$
Moyenne par traitement	$\bar{Y}_1$	$\bar{Y}_2$		$\bar{Y}_j$		$\bar{Y}_k$	$\bar{Y}$

n_o : nombre de blocs

k : nombre de traitements

n = no × k ; effectif total

$\bar{Y} = \frac{1}{n} \sum \sum yij$; moyenne générale

$\bar{Y}_{.i} = \frac{1}{k} \sum \sum yi.$; moyenne par bloc i.

$\bar{Y}j. = \frac{1}{no} \sum \sum y.j$; moyenne par traitement j.

***) Estimation du modèle**

En plus du modèle précédent. On doit ajouter l'effet bloc qui correspond à la variabilité due au bloc (SCE_B) :

$$SCE_t = SCE_a + SCE_B + SCE_r$$

$CM_t = \frac{SCEt}{ddl}$; avec ddl = n-1 ; carré moyen total

$CM_a = \frac{SCEa}{ddl}$; avec ddl = k-1 ; carré moyen factoriel

$CM_B = \frac{SCEB}{ddl}$; avec ddl = n_o-1 ; carré moyen bloc

$CM_r = \frac{SCEr}{ddl}$; avec ddl = (n_o-1)(k-1) ; carré moyen résiduel.

Avec n_o : bloc et k : traitement et finalement on peut calculer les valeurs de F observées:

$F_a = \frac{CMa}{CMr}$; $F_B = \frac{CMB}{CMr}$

Source de variabilité	**ddl**	**SCE**	**CM**	**F_{obs}**	**P**
Variabilité factorielle	k – 1	SCE_a	CM_a	$F_a = \frac{CMa}{CMr}$	P_a
Variabilité du bloc	n_o – 1	SCE_B	CM_B	$F_B = \frac{CMB}{CMr}$	P_B
Variabilité résiduelle	(k – 1)(n_o – 1)	SCE_r	CM_r		

Variation totale		SCE_t	CM_t		

2) Exemple expérimental (biologie végétale)

Question : quatre traitements d'engrais (T1–T4) influencent-ils la hauteur (cm) de plantules après 30 jours ?

La serre présente un gradient de lumière ; on le contrôle en formant 5 blocs (B1–B5), chacun recevant les 4 engrais, alloués aléatoirement au sein de chaque bloc.

Bloc \ Trait.	**T1**	**T2**	**T3**	**T4**
B1	18	22	22	25
B2	21	22	25	27
B3	21	24	27	27
B4	21	25	27	30
B5	19	21	24	27

- Totaux traitements : $T_1 = 100,\ T_2 = 114,\ T_3 = 125,\ T_4 = 136$
- Totaux blocs : $B_1 = 87,\ B_2 = 95,\ B_3 = 99,\ B_4 = 103,\ B_5 = 91$
- Total général $T = 475$, nombre d'observations $N = rt = 5 \times 4 = 20$
- Moyenne générale $\bar{X}_G = T/N = 23{,}75$

Moyennes par traitement :

$\bar{X}_{T1} = 20{,}0,\ \bar{X}_{T2} = 22{,}8,\ \bar{X}_{T3} = 25{,}0,\ \bar{X}_{T4} = 27{,}2$

3) Calculs ANOVA

Les sommes de carrés et carrés moyens sont calculés comme suit :

On note le **terme de correction** $C = T^2/N = 475^2/20 = 11281{,}25$.

- **Somme des carrés totale**

$$\mathrm{SCT} = \sum Y_{ij}^2 - C = 191{,}75$$

- **Somme des carrés traitements**

$$\mathrm{SCTR} = \frac{\sum T_i^2}{r} - C = \frac{100^2 + 114^2 + 125^2 + 136^2}{5} - 11281{,}25 = 142{,}15$$

- **Somme des carrés blocs**

$$\mathrm{SCB} = \frac{\sum B_j^2}{t} - C = \frac{87^2 + 95^2 + 99^2 + 103^2 + 91^2}{4} - 11281{,}25 = 40{,}00$$

- **Somme des carrés erreur (résiduelle)**

$$\mathrm{SCE} = \mathrm{SCT} - \mathrm{SCTR} - \mathrm{SCB} = 191{,}75 - 142{,}15 - 40{,}00 = 9{,}60$$

Degrés de liberté (DL)

- Traitements : $t - 1 = 3$
- Blocs : $r - 1 = 4$
- Erreur : $(t - 1)(r - 1) = 3 \times 4 = 12$
- Total : $N - 1 = 19$

Carrés moyens (CM) et statistiques F

$$CM_T = \frac{142{,}15}{3} = 47{,}383 \quad ; \quad CM_B = \frac{40{,}00}{4} = 10{,}000 \quad ; \quad CM_E = \frac{9{,}60}{12} = 0{,}800$$

$$F_{\text{trait}} = \frac{CM_T}{CM_E} = \frac{47{,}383}{0{,}800} = 59{,}23 \quad ; \quad F_{\text{blocs}} = \frac{CM_B}{CM_E} = \frac{10}{0{,}800} = 12{,}50$$

Tableau ANOVA (RCBD)

Source	SC	DL	CM	F
Traitements	142,15	3	47,383	**59,23**
Blocs	40,00	4	10,000	**12,50**
Erreur	9,60	12	0,800	—
Total	191,75	19	—	—

- Aux seuils usuels (ex. $\alpha = 0{,}05$), F_{trait} et F_{blocs} dépassent largement F_α → **effet traitement significatif** et **effet bloc significatif** (les blocs capturent bien un gradient réel).

Interprétation biologique

- L'effet « engrais » est **hautement significatif** : toutes choses égales par ailleurs (variabilité inter-blocs enlevée), les hauteurs diffèrent entre traitements.
- **Classement des moyennes** : T4(27,2)>T3(25,0)>T2(22,8)>T1(20,0)
- L'effet bloc significatif confirme l'intérêt du RCBD : on a **réduit l'erreur résiduelle** (CM_E=0,80) par rapport à un plan complètement aléatoire, augmentant ainsi la **puissance du test**.

Pour identifier **quelles paires** de traitements diffèrent, on réalise un **post-hoc** (ex. Tukey HSD) **sur le facteur Traitement**.

3.9.3.3. Dispositif suivant un carré latin

Avec **r** blocs horizontaux et **r** blocs verticaux et **r** traitements :

BV / BH	1	2	…….	j	……	**r**
1	$Y_{11(1)}$	$Y_{12(2)}$		$Y_{ij(j)}$		$Y_{ir(r)}$

⋮			……..		……	
i	$Y_{i1(1)}$	$Y_{i2(1+2)}$		$Y_{ij(1+j)}$		$Y_{ir(1)}$
⋮			……..		……	
r	$Y_{r1(r)}$	$Y_{n2(r)}$	……	$Y_{nj(r)}$	……	$Y_{nr(r-1)}$

Tel que $Y_{11(1)}$: cellule (1,1) qui a utilisé le traitement 1 ou traitement 1 figurant dans le BH1 et BV_1 :

La moyenne des BH : $\bar{Y}_{i..} = \frac{1}{r}\sum_{jk} Yijk$

La moyenne des BV : $Y_{.j.} = \frac{1}{r}\sum_{jk} Yijk$

La moyenne des traitements : $Y_{..k} = \frac{1}{r}\sum_{jk} Yijk$

La moyenne générale : $Y_{i..} = \frac{1}{r^2}\sum_{jk} Yijk$

***) Estimation du modèle**

La variation totale est constituée de la variation due au facteur étudié plus la variation due aux blocs horizontaux et verticaux, plus la variation résiduelle.

$$SCE_t = SCEa + SCEBH + SCE_{BV} + SCE_r$$

$CM_t = \frac{SCEt}{ddl}$; avec ddl = r^2-1 ; carré moyen total

$CM_a = \frac{SCEa}{ddl}$; avec ddl = r-1 ; carré moyen factoriel

$CM_{BH} = \frac{SCEBH}{ddl}$; avec ddl = r-1 ; carré moyen bloc horizontal

$CM_{BV} = \frac{SCEBV}{ddl}$; avec ddl = r-1 ; carré moyen bloc vertical

$CM_r = \frac{SCEr}{ddl}$; avec ddl = (r-1)(r-2) ; carré moyen résiduel.

Source de variabilité	**ddl**	**SCE**	**CM**	**F_{obs}**	**P**
Variabilité inter-	r – 1	SCE_a	CM_a	$F_a = \frac{CMa}{CMr}$	P_a

traitement Variabilité inter blocs horizontaux Variabilité inter blocs verticaux Variabilité résiduelle	r – 1 r – 1 (r – 1)(r – 2)	SCE_{BH} SCE_{BV} SCE_r	CM_{BH} CM_{BV} CM_r	$F_{BH} = \frac{CMBH}{CMr}$ $F_{BV} = \frac{CMBV}{r}$	P_{BH} P_{BV}
Variation totale		SCE_t	CM_t		

3.9.3.4. ANOVA à deux facteurs selon un dispositif complètement aléatoire (DCA)

1) Contexte et objectif

En sciences expérimentales, on souhaite souvent étudier simultanément deux facteurs (ex. type d'engrais et régime d'irrigation). Dans un dispositif complètement aléatoire (DCA), toutes les unités expérimentales sont homogènes et la randomisation est globale. L'ANOVA à deux facteurs permet de tester les effets principaux et leur interaction.

Modèle (avec interaction)

$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$, avec $\varepsilon_{ijk} \sim N(0, \sigma^2)$ indépendants
- α_i : effet du facteur A (i=1..a), β_j : effet du facteur B (j=1..b)
- $(\alpha\beta)_{ij}$: interaction A×B, k=1..r réplicats par cellule

Hypothèses : H_0(A): $\alpha_1=...=\alpha_a=0$; H_0(B): $\beta_1=...=\beta_b=0$; H_0(AB): $(\alpha\beta)_{ij}=0$ $\forall i,j$.

Tableau d'ANOVA à deux facteurs suivant une randomisation totale

Source de variabilité	**ddl**	**SCE**	**CM**	**F_{obs}**	**P**
Variabilité facteur A Variabilité facteur B Interaction AB Variabilité résiduelle	I – 1 J – 1 (I -1) (J – 1) IJ (k-1)	SCE_a SCE_b SCE_{ab} SCE_r	CM_a CM_b CM_{ab} CM_r	F_a F_b F_{ab}	P_a P_b P_{ab}
Variation totale	n-1	SCE_t	CM_t		

2) Exemple expérimental (biologie végétale)

Question : l'augmentation de l'irrigation modifie-t-elle l'effet de trois engrais sur la hauteur (cm) des plantules après 30 jours ? Plan équilibré : a=3 engrais (A1–A3), b=2 régimes d'irrigation (B1 faible, B2 élevée), r=3 réplicats par cellule (N=18).

A\B	B1	B2
A1	14, 15, 13 (m=14.00)	17, 16, 18 (m=17.00)
A2	16, 17, 15 (m=16.00)	20, 19, 21 (m=20.00)
A3	15, 14, 16 (m=15.00)	22, 23, 21 (m=22.00)

Moyenne générale = 17.33
Moyennes A : A1=15.50, A2=18.00, A3=18.50
Moyennes B : B1=15.00, B2=19.67

3) Calculs ANOVA (formules « raccourcis »)

Terme de correction $C = T^2/N = 5408.000$
SCT = 154.000
SSA = 31.000, SSB = 98.000, SS(AB) = 13.000, SCE = 12.000

Tableau ANOVA

Source	SC	DL	CM	F
A (Engrais)	31.000	2	15.500	15.50
B (Irrigation)	98.000	1	98.000	98.00
A×B	13.000	2	6.500	6.50
Erreur	12.000	12	1.000	
Total	154.000	17		

4) Interprétation

On teste d'abord l'interaction A×B. Si elle est significative, on interprète les effets simples (effet de A à chaque niveau de B, et inversement). Ici, les moyennes par cellule montrent un effet croissant de l'irrigation et un avantage accru de l'engrais A3 sous B2.

5) Conditions de validité

- Normalité des résidus (QQ-plot, Shapiro-Wilk)
- Homogénéité des variances (Levene/Brown–Forsythe)
- Indépendance des observations
- Modèle additif si l'interaction est nulle

3.9.3.5. Tests de Kruskal–Wallis et de Friedman

Les tests paramétriques comme l'ANOVA nécessitent des hypothèses fortes : normalité des données et égalité des variances.
Lorsque ces conditions sont violées (ex. données ordinales, distribution asymétrique, petits échantillons), on utilise des **tests non paramétriques** basés sur les **rangs** plutôt que sur les valeurs brutes.

Les deux tests suivants permettent d'évaluer les différences entre groupes sans supposer la normalité :

- Le **test de Kruskal–Wallis** : alternative non paramétrique à l'**ANOVA à un facteur** (données indépendantes).
- Le **test de Friedman** : alternative non paramétrique à l'**ANOVA à mesures répétées** ou à **un plan en blocs** (données appariées).

a) Test de Kruskal–Wallis

Objectif

Ce test compare **plus de deux échantillons indépendants** pour déterminer s'ils proviennent de la même population.

Hypothèses :

- H_0 : les distributions des groupes sont identiques.
- H_1 : au moins une distribution diffère.

Principe

Les observations de tous les groupes sont **classées (rangées)** ensemble du plus petit au plus grand.
Le test repose sur la **somme des rangs moyens** par groupe.

Statistique de Kruskal–Wallis :

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

avec :

- R_i = somme des rangs du groupe i
- n_i = taille du groupe i
- N = nombre total d'observations

Sous H_0, H suit approximativement une loi du **χ^2** à k−1 degrés de liberté.

Exemple

Un biologiste teste trois régimes alimentaires sur la **prise de poids (g)** de poissons d'aquaculture.

Régime Données (gain en g)

A 20, 18, 16, 22

B 25, 28, 26, 30

C 19, 17, 20, 18

Hypothèse : les trois régimes n'ont pas le même effet sur le gain pondéral.

Étapes :

1. Rassembler les 12 valeurs et les ranger.
2. Calculer la somme des rangs R_A, R_B, R_C.
3. Appliquer la formule du test.

Résultat (exemple calculé) :

$$H = 8.72, \quad \chi^2_{0.05,2} = 5.99 \Rightarrow H > \chi^2 \Rightarrow H_0 \text{ rejetée.}$$

Conclusion : les régimes influencent significativement le gain de poids. Un test post-hoc de **Dunn** ou **Conover** peut identifier les paires significativement différentes.

b) Test de Friedman

Objectif

Le test de Friedman compare **plus de deux traitements appliqués sur les mêmes sujets** ou sur **des blocs homogènes** (expériences appariées). C'est l'alternative non paramétrique à l'**ANOVA à un facteur en blocs aléatoires**.

Hypothèses :

- H_0 : les traitements ont des effets identiques.
- H_1: au moins un traitement diffère.

Principe

Chaque bloc (ou sujet) fournit une **rangée de mesures** classées selon leur rang à l'intérieur du bloc.

Statistique :

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

avec :

- n = nombre de blocs,
- k = nombre de traitements,
- R_j = somme des rangs du traitement j.

Sous H_0, Q suit approximativement une loi du **χ^2** à k−1 degrés de liberté.

Exemple

Un chercheur évalue **trois milieux de culture (A, B, C)** sur la **croissance bactérienne** (densité optique) mesurée sur **5 souches différentes** (blocs).

Souche	A	B	C
1	0.45	0.51	0.55
2	0.38	0.42	0.41
3	0.60	0.66	0.71
4	0.49	0.50	0.52
5	0.33	0.38	0.36

Résultat du test :

$$Q = 9.60, \quad \chi^2_{0.05,2} = 5.99 \Rightarrow Q > \chi^2 \Rightarrow H_0 \text{ rejetée.}$$

Conclusion : il existe une différence significative entre les milieux de culture, le milieu **C** donnant la croissance la plus forte.

Conditions et remarques

Condition	Kruskal–Wallis	Friedman
Données indépendantes	✓	✗
Données appariées (blocs/sujets)	✗	✓

Condition	Kruskal–Wallis	Friedman
Type de mesure	Ordinale ou quantitative non normale	Ordinale ou quantitative non normale
Test post-hoc	Dunn, Conover	Nemenyi, Wilcoxon apparié

Interprétation biologique

Ces tests sont très utilisés en **écologie, agronomie et microbiologie** lorsque :

- les effectifs sont faibles ou hétérogènes ;
- les distributions sont asymétriques ;
- les mesures sont des scores, classements ou intensités.

Ils permettent de **détecter des différences significatives** entre traitements ou conditions expérimentales sans recourir à des hypothèses paramétriques strictes.

Conclusion

Les tests de **Kruskal–Wallis** et de **Friedman** sont des outils essentiels pour les biologistes lorsqu'il s'agit de comparer plusieurs groupes dans des situations réelles où les hypothèses de l'ANOVA ne sont pas respectées. Leur simplicité et leur robustesse en font des méthodes privilégiées pour l'analyse de données expérimentales non normales ou ordinales.

3.10. Corrélation de Pearson et régression linéaire

3.10.1. Régression linéaire simple

La corrélation binaire simple (Corrélation de Pearson) tente de donner une synthèse de la régularité que l'on devine dans le graphique en supposant qu'une droite est capable de « rassembler » au mieux (le plus prés possible de la droite), les divers points du graphique.

Cette corrélation est étudiée au moyen d'un diagramme de dispersion (graphique) et du coefficient de corrélation linéaire (une mesure de la direction et de l'intensité de l'association linéaire entre deux variables).

Le coefficient de corrélation de Pearson est une mesure d'association qui permet d'établir si deux variables mesurées sur le même ensemble d'observations varient de façon analogue ou non.

***) Modèle linéaire de la régression simple**

Soit y et x, deux variables que l'on suppose corrélées linéairement suivant la relation :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Tel que :

β_0 et β_1 : sont des constantes inconnues, considérées comme les paramètres du modèle ;

y : est une variable aléatoire dépendante ;

x : est une variable aléatoire indépendante.

L'expression qui définit le coefficient de corrélation linéaire est la suivante :

$$r = \frac{\sum_{i=1}^{n}(Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{\sum(Xi - \bar{X})^2}\sqrt{\sum(Yi - \bar{Y})^2}}$$

Le coefficient de corrélation linéaire simple prend ses valeurs à l'intérieur de l'intervalle -1 et +1 ($-1 \leq r \leq +1$).

Le signe positif ou négatif du coefficient de la corrélation (r) correspond à l'orientation de la pente de la droite autour de la quelle se regroupent les divers points du nuage de points.

Le coefficient de corrélation (r) est aussi égal à la covariance divisée par le produit des écarts types de x et de y :

$$r = \frac{Cov\ XY}{SxSy}$$

Avec :

$$Cov\ xy = \frac{\sum_{i=1}^{n}(Xi - \bar{X})(Yi - \bar{Y})}{n}$$

3.11. Corrélation de Spearman

- Pour mesurer l'**association monotone** entre deux variables quantitatives (ou ordinales), sans supposer la normalité.

- Plus robuste que Pearson aux valeurs extrêmes et aux non-linéarités monotones (croissant/décroissant).

$$\rho_s = \frac{\sum_{i=1}^{n}(R_{x,i} - \bar{R}_x)(R_{y,i} - \bar{R}_y)}{\sqrt{\sum_{i=1}^{n}(R_{x,i} - \bar{R}_x)^2}\sqrt{\sum_{i=1}^{n}(R_{y,i} - \bar{R}_y)^2}}, \quad \bar{R}_x = \bar{R}_y = \frac{n+1}{2}.$$

3.12. Test d'indépendance et tableau de contingence

Principe général

Le **test du khi-deux (χ^2) d'indépendance** est utilisé pour déterminer s'il existe une **relation statistique entre deux variables qualitatives** (catégorielles).

- **Hypothèses :**
  - H_0 : les deux variables sont **indépendantes** (aucune relation).
  - H_1 : les deux variables sont **dépendantes** (une relation existe).

Tableau de contingence

Un tableau de contingence résume la distribution conjointe des modalités de deux variables qualitatives:

	Modalité B_1	**Modalité B_2**	**Total**
Modalité A_1	a_{11}	a_{12}	$n_1\bullet$
Modalité A_2	a_{21}	a_{22}	$n_2\bullet$
Total	$n\bullet_1$	$n\bullet_2$	N

Chaque case contient un **effectif observé (O_{ij})**. L'hypothèse d'indépendance permet de calculer un **effectif théorique attendu (E_{ij})** :

$$E_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{N}$$

Statistique du test

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

où :

- r: nombre de lignes,
- c: nombre de colonnes.

Le test suit approximativement une **loi du χ^2** à $(r-1)(c-1)$ degrés de liberté.

Exemple:

Une étude sur la **prévalence d'une infection bactérienne** selon le **sexe des animaux** :

Sexe / Infection	Infecté	Non infecté	Total
Mâles	18	22	40
Femelles	32	28	60
Total	50	50	100

Hypothèse H_0 : le sexe et l'infection sont indépendants.
On calcule les effectifs attendus et la statistique χ^2.
Si la p-valeur < 0.05, on rejette H_0 : le sexe influence la probabilité d'infection.

3.13. Mesures d'association : Odds ratio (OR) et risque relatif (RR)

Tableau de contingence 2×2

	Malade (Cas)	Non malade (Témoin)	Total
Exposé	a	b	a+b
Non exposé	c	d	c+d
Total	a+c	b+d	N

Ces deux mesures évaluent **l'association entre une exposition et un événement de santé**.

<u>Risque relatif (RR)</u>

Le **risque** est la probabilité d'être malade :

$$Risque_{exposé} = \frac{a}{a+b}, \quad Risque_{non\ exposé} = \frac{c}{c+d}$$

Le **risque relatif (RR)** est :

$$RR = \frac{Risque_{exposé}}{Risque_{non\ exposé}}$$

- RR = 1 → Pas d'association
- RR > 1 → L'exposition augmente le risque
- RR < 1 → L'exposition protège

Odds ratio (OR)

L'**odds (rapport de chances)** est le rapport entre la probabilité d'un événement et sa non-réalisation.

$$Odds_{exposé} = \frac{a}{b}, \quad Odds_{non\ exposé} = \frac{c}{d}$$

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

- OR = 1 → pas d'association
- OR > 1 → exposition associée à un risque plus élevé
- OR < 1 → exposition protectrice

Remarque

- OR ≈ RR quand la maladie est rare (<10%)
- OR est souvent utilisé dans les **études cas-témoins**
- RR dans les **études de cohorte ou essais cliniques**

Exemple

Étude de l'effet d'un pesticide sur l'apparition d'une tumeur hépatique chez des rats :

	Tumeur	Pas de tumeur	Total
Exposés au pesticide	20	80	100
Non exposés	10	90	100

- $Risque_{exposé} = 20/100 = 0.20$
- $Risque_{non\ exposé} = 10/100 = 0.10$

  → RR = 0.20 / 0.10 = 2

L'exposition **double le risque** de tumeur.

Pour l'odds ratio :

$$OR = \frac{20 \times 90}{80 \times 10} = \frac{1800}{800} = 2.25$$

→ Les rats exposés ont **2.25 fois plus de chances** de développer une tumeur.

Interprétation biologique

- Si **RR ou OR > 1**, l'exposition augmente la probabilité d'un effet biologique néfaste.
- Si **RR ou OR < 1**, l'exposition a un effet protecteur.
- L'importance de l'effet doit toujours être accompagnée d'un **intervalle de confiance à 95 %** pour juger la précision.

3.14. Analyse multidimensionnelle te réduction des données

3.14.1. Analyse en composantes principales (ACP)

But : Réduire la dimensionnalité de variables quantitatives corrélées en un petit nombre de composantes orthogonales expliquant la variance, pour visualiser des patrons (individus, variables) et détecter des gradients.

Prétraitements : centrer et réduire (z-scores) si les échelles diffèrent ; gérer les valeurs manquantes.

Données (centrées-réduites si échelles différentes) : matrice X (n×p).

Matrice de covariance/corrélation : $S = (1/(n-1))\, X^t X$ (si X déjà centrée).

Décomposition spectrale/SVD : $S = P \Lambda P^t$ (Λ : valeurs propres ; P : vecteurs propres).

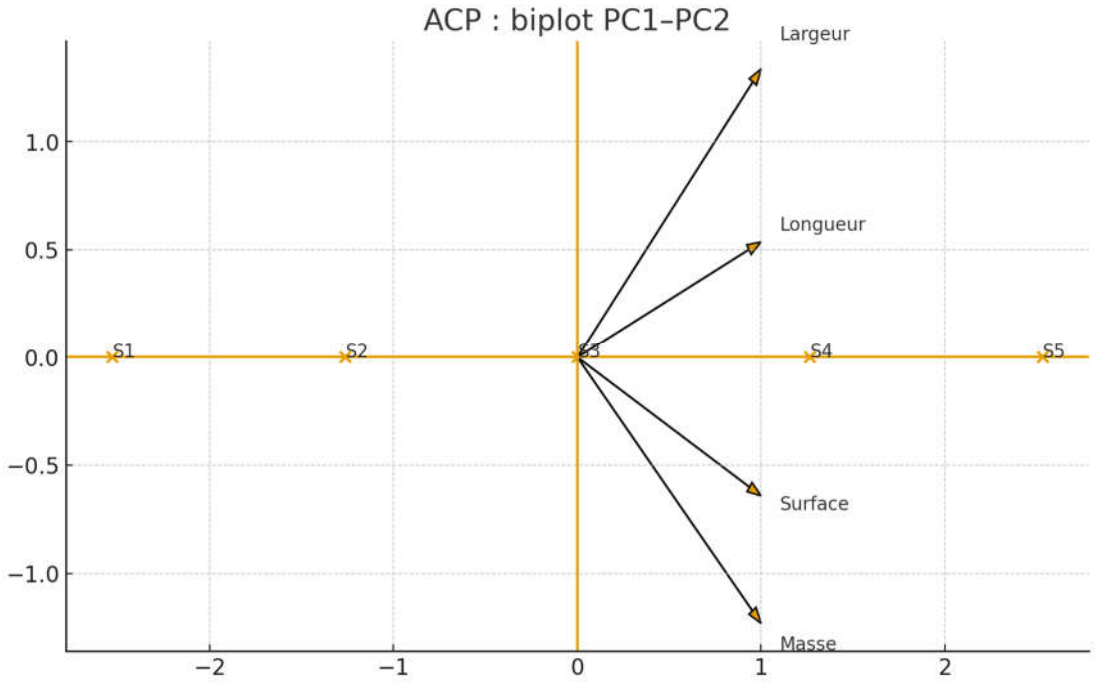
Scores individus : $T = X P$; contributions/qualités (cos^2) pour interpréter les axes.

Variance expliquée : $\lambda_j / \Sigma_k \lambda_k$; choisir q tel que $\Sigma_{j=1..q} \lambda_j \geq 70–85\%$ (indicatif).

Exemple : Morphométrie de 30 spécimens (longueur, largeur, masse, surface foliaire).

Spécimen	Longueur (mm)	Largeur (mm)	Masse (g)	Surface foliaire (cm^2)
1	32	11.1	1.7	13.8
2	34	12.2	1.9	15.6
3	36	13.3	2.1	17.4
4	38	14.4	2.3	19.2
5	40	15.5	2.5	21.0

Résultats typiques : PC1 (taille générale) explique ~70% de la variance ; PC2 (forme) ~15%.



3.14.2. Analyse factorielle des correspondances (AFC)

But : Explorer une table de contingence (fréquences) espèce × site (ou gènes × conditions) pour visualiser les associations entre lignes et colonnes dans un espace de faible dimension.

Prétraitements : éventuellement transformation de profils ; lignes/colonnes rares peuvent déformer l'analyse (poids faibles).

Soit N la table de fréquences, n = somme(N), P = N / n (fréquences relatives).

Masses : r = P 1_c ; c = 1_r^t P (sommes de lignes/colonnes).

Matrice centrée-normalisée : S = D_r^{-1/2} (P − r c^t) D_c^{-1/2}.

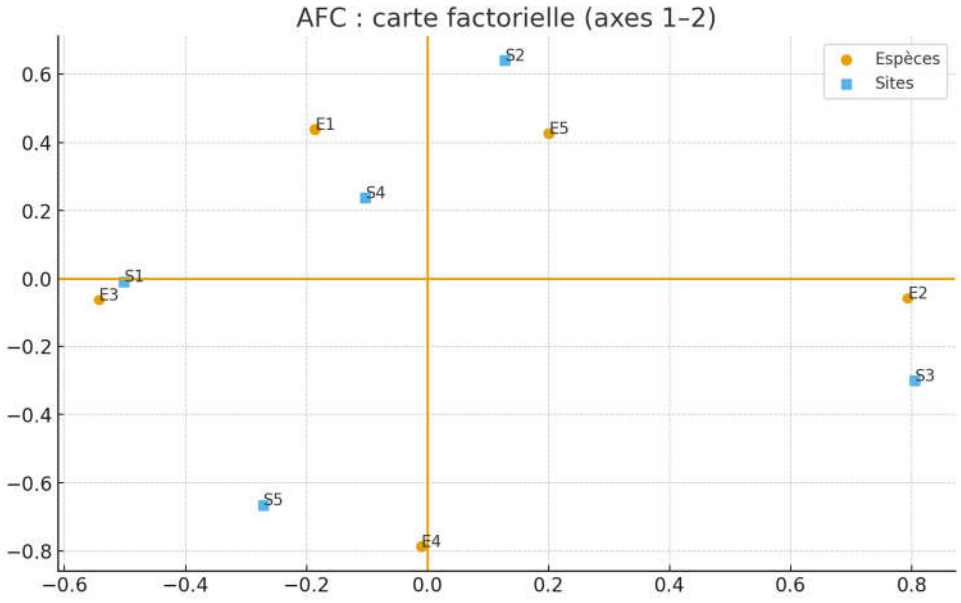
SVD : S = U Σ V^t ; coordonnées factorielles lignes/colonnes à partir de U, V et Σ ; inerties = Σ².

Exemple : Abondances de 5 espèces (E1–E5) sur 6 sites (S1–S6).

	S1	S2	S3	S4	S5
E1	3	7	0	1	2
E2	0	2	5	1	0

E3	6	1	0	4	3
E4	1	0	2	0	5
E5	0	3	1	2	0

Interprétation (typique) : Les axes 1–2 capturent les associations espèces–sites (ex. E3 associée à S1 et S4). Graphiques : nuage des points lignes/colonnes, contributions et qualités de représentation (cos²).



3.14.3. Analyse des correspondances multiples (ACM)

But : Extension de l'AFC à plusieurs variables qualitatives (table disjonctive complète Z).

Burt : $B = Z^t Z$; normalisation analogue à l'AFC ; total d'inertie lié au nombre de modalités (Q−1).

Correction de Benzécri (souvent utilisée) pour interprétation des inerties ; lecture via cartes individus/modèles.

Exemple 50 plantes caractérisées par Couleur_fleur (blanche/jaune/violette), Port (érigé/étalé), Dispersion (anémophile/entomophile). L'ACM révèle des groupements de modalités et d'individus.

Mesures clés : inertie, contributions, cos² ; graphiques des individus et modalités.

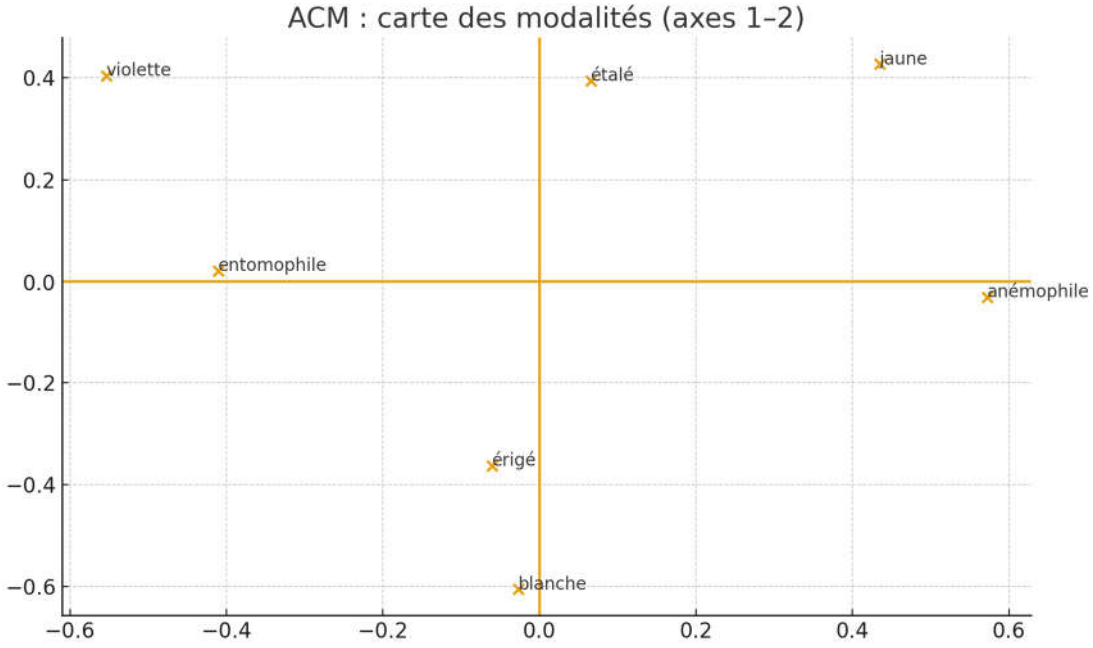
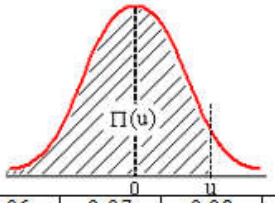

ACM : carte des modalités (axes 1–2)
violette
étalé
jaune
entomophile
anémophile
érigé
blanche
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

Table de Loi Normale
P(x<u)



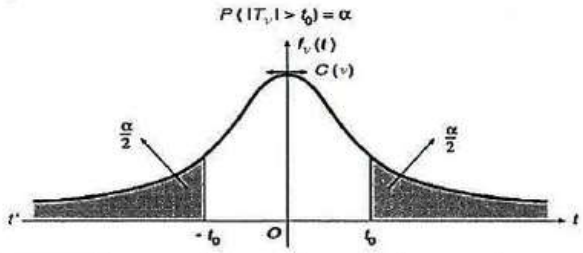
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	.0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8254	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	.0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0.9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	6,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	.0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Loi du khi-deux avec k degrés de liberté
Quantiles d'ordre $1-\gamma$

	γ										
k	0.995	0.990	0.975	0.950	0.900	0.500	0.100	0.050	0.025	0.010	0.005
1	0.00	0.00	0.00	0.00	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.94	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.81	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

Si k est entre 30 et 100 mais n'est pas un multiple de 10, on utilise la table ci-haut et on fait une interpolation linéaire. Si $k > 100$ on peut, grâce au théorème limite central, approximer la loi $\chi^2(k)$ par la loi $N(k, 2k)$.

TABLE DE LA LOI t DE STUDENT



ν \ α	0,50	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001
1	0,000	1,000	3,087	6,314	12,706	31,821	63,657	127,321	636,619
2	0,000	0,816	1,886	2,920	4,303	6,965	9,925	14,089	31,599
3	0,000	0,765	1,638	2,353	3,182	4,541	5,841	7,453	12,924
4	0,000	0,741	1,533	2,132	2,776	3,747	4,604	5,598	8,610
5	0,000	0,727	1,476	2,015	2,571	3,385	4,032	4,773	6,869
6	0,000	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,959
7	0,000	0,711	1,415	1,895	2,365	2,998	3,499	4,029	5,408
8	0,000	0,706	1,397	1,860	2,306	2,896	3,355	3,833	5,041
9	0,000	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,781
10	0,000	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,587
11	0,000	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,437
12	0,000	0,695	1,356	1,782	2,179	2,681	3,055	3,428	4,318
13	0,000	0,694	1,350	1,771	2,160	2,650	3,012	3,372	4,221
14	0,000	0,692	1,345	1,761	2,145	2,624	2,977	3,326	4,140
15	0,000	0,691	1,341	1,753	2,131	2,602	2,947	3,286	4,073
16	0,000	0,690	1,337	1,746	2,120	2,583	2,921	3,252	4,015
17	0,000	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,965
18	0,000	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,922
19	0,000	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,883
20	0,000	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,850
21	0,000	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,819
22	0,000	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,792
23	0,000	0,685	1,319	1,714	2,069	2,500	2,807	3,104	3,768
24	0,000	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,745
25	0,000	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,725
26	0,000	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,707
27	0,000	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,690
28	0,000	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,674
29	0,000	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,659
30	0,000	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,646
40	0,000	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,551
60	0,000	0,679	1,296	1,671	2,000	2,390	2,660	2,915	3,460
80	0,000	0,678	1,292	1,664	1,990	2,374	2,639	2,887	3,416
120	0,000	0,677	1,289	1,658	1,980	2,358	2,617	2,860	3,373
∞	0,000	0,674	1,282	1,645	1,960	2,326	2,576	2,807	3,291

Table de Fisher-Snedecor, $\alpha = 5\ \%$ (95^e centile)

ν_2 (dén.)	ν_1 (numérateur)																			
	1	2	3	4	5	6	7	8	9	10	20	30	40	50	60	80	100	200	500	1 000
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	248.02	250.10	251.14	251.77	252.20	252.72	253.04	253.68	254.06	254.19
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.45	19.46	19.47	19.48	19.48	19.48	19.49	19.49	19.49	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.66	8.62	8.59	8.58	8.57	8.56	8.55	8.54	8.53	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.80	5.75	5.72	5.70	5.69	5.67	5.66	5.65	5.64	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.56	4.50	4.46	4.44	4.43	4.41	4.41	4.39	4.37	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.87	3.81	3.77	3.75	3.74	3.72	3.71	3.69	3.68	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.44	3.38	3.34	3.32	3.30	3.29	3.27	3.25	3.24	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.15	3.08	3.04	3.02	3.01	2.99	2.97	2.95	2.94	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	2.94	2.86	2.83	2.80	2.79	2.77	2.76	2.73	2.72	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.77	2.70	2.66	2.64	2.62	2.60	2.59	2.56	2.55	2.54
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.12	2.04	1.99	1.97	1.95	1.92	1.91	1.88	1.86	1.85
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	1.93	1.84	1.79	1.76	1.74	1.71	1.70	1.66	1.64	1.63
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.84	1.74	1.69	1.66	1.64	1.61	1.59	1.55	1.53	1.52
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.78	1.69	1.63	1.60	1.58	1.54	1.52	1.48	1.46	1.45
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.75	1.65	1.59	1.56	1.53	1.50	1.48	1.44	1.41	1.40
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.72	1.62	1.57	1.53	1.50	1.47	1.45	1.40	1.37	1.36
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.70	1.60	1.54	1.51	1.48	1.45	1.43	1.38	1.35	1.34
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.69	1.59	1.53	1.49	1.46	1.43	1.41	1.36	1.33	1.31
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.68	1.57	1.52	1.48	1.45	1.41	1.39	1.34	1.31	1.30
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.62	1.52	1.46	1.41	1.39	1.35	1.32	1.26	1.22	1.21
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.61	1.50	1.43	1.39	1.36	1.32	1.30	1.23	1.19	1.17
600	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.86	1.59	1.48	1.42	1.38	1.35	1.30	1.28	1.21	1.16	1.14
1 000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.58	1.47	1.41	1.36	1.33	1.29	1.26	1.19	1.13	1.11
2 000	3.85	3.00	2.61	2.38	2.22	2.10	2.01	1.94	1.88	1.84	1.58	1.46	1.40	1.36	1.32	1.28	1.25	1.18	1.12	1.09

Références

Conover, W.J. (1999). Practical Nonparametric Statistics. 3rd ed. Wiley.

Devore, J. L. (2015). Probability and Statistics for Engineering and the Sciences. Cengage Learning.

Dodge, Y. (2003). The Oxford Dictionary of Statistical Terms. Oxford University Press.

Field, A., Miles, J., & Field, Z. (2012). Discovering Statistics Using R. Sage.

Gibbons, J.D., & Chakraborti, S. (2011). Nonparametric Statistical Inference. 5th ed. Chapman & Hall/CRC.

Greenacre, M. (2007). Correspondence Analysis in Practice. 2nd ed. Chapman & Hall/CRC.

Greenacre, M. (2010). Biplots in Practice. Foundation for Open Access Statistics.

Jolliffe, I.T. (2002). Principal Component Analysis. 2nd ed. Springer.

Kirkwood, B. & Sterne, J. (2003). Essential Medical Statistics. Blackwell Science.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1).

Legendre, P., & Legendre, L. (2012). Numerical Ecology. 3rd English ed. Elsevier.

Maxwell, S.E., & Delaney, H.D. (2004). Designing Experiments and Analyzing Data. 2nd ed. Lawrence Erlbaum.

Montgomery, D. C. (2019). Design and Analysis of Experiments. Wiley.

Montgomery, D. C., & Runger, G. C. (2014). Applied Statistics and Probability for Engineers. Wiley.

Pagano, M. & Gauvreau, K. (2018). Principles of Biostatistics.

Saporta, G. (2011). Probabilités, analyse des données et statistique. Technip.

Snedecor, G. W., & Cochran, W. G. (1989). Statistical Methods. Iowa State University Press.

Sokal, R. R. & Rohlf, F. J. (2012). Biometry: The Principles and Practice of Statistics in Biological Research.

Zar, J.H. (2010). Biostatistical Analysis. 5th ed. Pearson.