Ministère de l'Agriculture, des Ressources hydrauliques et de la Pêche Maritime

–*–

Institution de la Recherche et de l'Enseignement Supérieur Agricoles

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

–*–

Université de Carthage

REPUBLIQUE TUNISIENNE

ECOLE SUPÉRIEURE D'AGRICULTURE DE MOGRANE

NOTES DE COURS
VERSION.2024.BETA

Biométrie

Une Alliance pour la Science Agronomique Sous R

Bilel AMMOURI

: ESAM

: ammouri-bilel

: bilelammouri

: Ammouri-Bilel

: 0000-0002-5491-5172

Avant-propos

Le cours de Biométrie conçu spécialement pour les ingénieurs en agronomie. La biométrie, qui intègre les méthodes statistiques et mathématiques à la biologie, trouve une application précieuse dans l'analyse des données agronomiques. Ce cours vise à vous fournir les compétences essentielles pour collecter, analyser et interpréter les données agronomiques de manière rigoureuse et informative.

L'agronomie joue un rôle crucial dans la garantie de la sécurité alimentaire mondiale et la gestion durable des ressources naturelles. Pour relever les défis complexes de l'agriculture moderne, une compréhension solide des méthodes statistiques et de leur application est indispensable. La biométrie offre les outils nécessaires pour optimiser les pratiques culturales, prédire les rendements, comprendre les interactions sol-plante, et bien plus encore.

Au fil de ce cours, vous explorerez divers concepts, méthodes et techniques de la biométrie adaptés au contexte agronomique. De l'introduction aux statistiques descriptives à l'analyse avancée de variance, de la régression à l'évaluation de l'ajustement des modèles, vous développerez des compétences qui vous permettront de prendre des décisions éclairées basées sur des données tangibles et scientifiquement solides.

On vous encourage à vous engager activement dans ce cours en participant aux discussions, en posant des questions et en appliquant les concepts à des exemples pratiques. Les connaissances que vous acquerrez ici renforceront votre capacité à contribuer de manière significative à l'innovation agricole et au développement durable.

On vous souhaite un survole enrichissant à travers le monde de la biométrie en agronomie. Profitez de ce cours pour développer vos compétences analytiques et de prise de décision, et contribuer ainsi **à façonner un avenir agricole plus productif et équilibré**.

Table of Contents

CHAPTER 1

Introduction à la Biométrie et Statistiques de Base

L'agronomie, en tant que science de la gestion des systèmes de production agricole, repose sur la prise de décisions éclairées pour maximiser les rendements, améliorer la qualité des cultures et garantir une utilisation durable des ressources. Dans ce contexte, la biométrie et les statistiques jouent un rôle fondamental en fournissant les outils nécessaires pour collecter, analyser et interpréter les données agronomiques de manière précise et significative.

La biométrie est l'intersection entre la biologie et les mathématiques (statistiques), utilisée pour explorer les relations complexes entre les variables biologiques. Dans le domaine agronomique, la biométrie permet de quantifier les phénomènes tels que la croissance des plantes, l'effet des pratiques agricoles sur les rendements, la relation entre les caractéristiques du sol et la composition des cultures, et bien plus encore. Les statistiques, d'autre part, fournissent les méthodes et les cadres pour récolter, analyser et interpréter les données afin d'obtenir des informations exploitables ([MPV21], [ST86], [HHE05]).

Objectifs : Ce cours a pour objectif de vous initier aux concepts et aux techniques statistiques fondamentales qui sont pertinents pour les professionnels de l'agronomie. Vous allez apprendre à :

① Collecter et organiser des données agronomiques de manière systématique ;

② Appliquer des statistiques descriptives pour résumer et visualiser les caractéristiques des données ;

③ Comprendre les principes de base de la régression pour modéliser les relations entre les variables ;

④ Évaluer l'ajustement des modèles et interpréter les résultats ;

⑤ Utiliser des techniques d'analyse de corrélation pour étudier les relations entre les variables agronomiques.

Application : Tout au long du cours, vous aurez l'occasion d'appliquer ces concepts à des exemples concrets tirés du domaine agronomique. Que ce soit pour prédire les rendements des cultures en fonction des pratiques culturales, analyser l'impact des facteurs environnementaux sur la croissance des plantes ou évaluer les effets des traitements agricoles sur la qualité des cultures, vous développerez des compétences essentielles pour prendre des décisions informées et optimiser vos pratiques agricoles.

1.1 Terminologies population et caractère

① **Population** : En statistiques, une population fait référence à l'ensemble complet d'éléments ou d'individus qui possèdent certaines caractéristiques communes et qui sont d'intérêt pour l'étude. Cela peut être un groupe spécifique de personnes, d'objets, d'animaux ou d'entités similaires.

② **Unité statistique** : L'unité statistique est l'élément individuel qui compose la population et qui est observé dans le cadre d'une étude statistique. Cela peut être une personne, un animal, un objet, une entreprise, etc. Les données sont collectées sur ces unités statistiques pour analyser et tirer des conclusions sur la population dans son ensemble.

③ **Caractère** : On a deux types de caractère :

— **quantitatif** : Un caractère quantitatif est une variable mesurable qui prend des valeurs numériques et peut être soumise à des opérations mathématiques. Par exemple, la taille, le poids, le nombre de fruits sur un arbre, etc. Ces caractères permettent des analyses numériques et des comparaisons basées sur des mesures.

— **qualitatif** : Un caractère qualitatif est une variable qui décrit une qualité ou une catégorie, mais qui ne peut pas être mesurée numériquement. Il s'agit plutôt de

données catégorielles. Par exemple, la couleur, le type de sol, la présence/absence d'une maladie, etc. Les caractères qualitatifs permettent d'effectuer des analyses de fréquence et de distribution dans des catégories.

Exemple :
Prenons l'exemple de l'étude d'une population d'arbres fruitiers dans un verger.

— Population : L'ensemble de tous les arbres fruitiers dans le verger.

— Unité statistique : Chaque arbre fruitier individuel dans le verger.

— Caractère quantitatif : La quantité de fruits récoltés par chaque arbre pendant une saison.

— Caractère qualitatif : La variété de fruit produite par chaque arbre (pommes, poires, cerises, etc.).

1.2 Types de Données en Agronomie

En agronomie, différentes catégories de données sont recueillies pour comprendre et améliorer les pratiques agricoles, la croissance des cultures et la gestion des ressources (voir figure 1.1). Ces données sont essentielles pour prendre des décisions éclairées et optimiser les rendements. Les types de données couramment rencontrés en agronomie sont :

1. **Données Continues** : Les données continues sont des valeurs numériques qui peuvent varier sur une plage continue. Elles peuvent prendre n'importe quelle valeur dans cette plage et sont généralement mesurées à l'aide d'instruments de mesure.
    i. **Application** : En utilisant des données continues telles que la hauteur des plants de maïs, les agronomes peuvent analyser la croissance des cultures, identifier les variations et ajuster les pratiques culturales en conséquence pour optimiser les rendements ;
    ii. **Exemple** : La hauteur des plants de blé dans un champ est une donnée continue. Chaque plant peut avoir une hauteur différente, et la mesure peut être exprimée en centimètres.
2. **Données Discrètes** : Les données discrètes sont des valeurs numériques distinctes, généralement obtenues par décompte ou comptage. Elles ne peuvent prendre que certaines valeurs spécifiques et sont souvent associées à des catégories.

i. **Application** : En collectant des données discrètes sur le nombre de fruits par arbre, les agronomes peuvent évaluer la productivité de différents arbres, surveiller les variations d'année en année et adapter les pratiques de fertilisation et d'arrosage pour maximiser la production ;

ii. **Exemple** : Le nombre de fruits sur un arbre fruitier est une donnée discrète. Vous ne pouvez pas avoir un demi-fruit, et le nombre de fruits est un nombre entier.

3. **Données Catégorielles** : Les données catégorielles sont des valeurs qui appartiennent à des catégories ou des groupes distincts. Elles ne sont pas numériques et ne peuvent pas être ordonnées de manière significative.

i. **Application** : En classant les types de sols en catégories, les agronomes peuvent adapter les cultures à chaque type de sol, recommander des pratiques de drainage appropriées et optimiser l'utilisation des ressources ;

ii. **Exemple** : Les types de sols dans une région agricole peuvent être des données catégorielles. Les catégories pourraient inclure "sableux", "argileux" ou "limoneux".

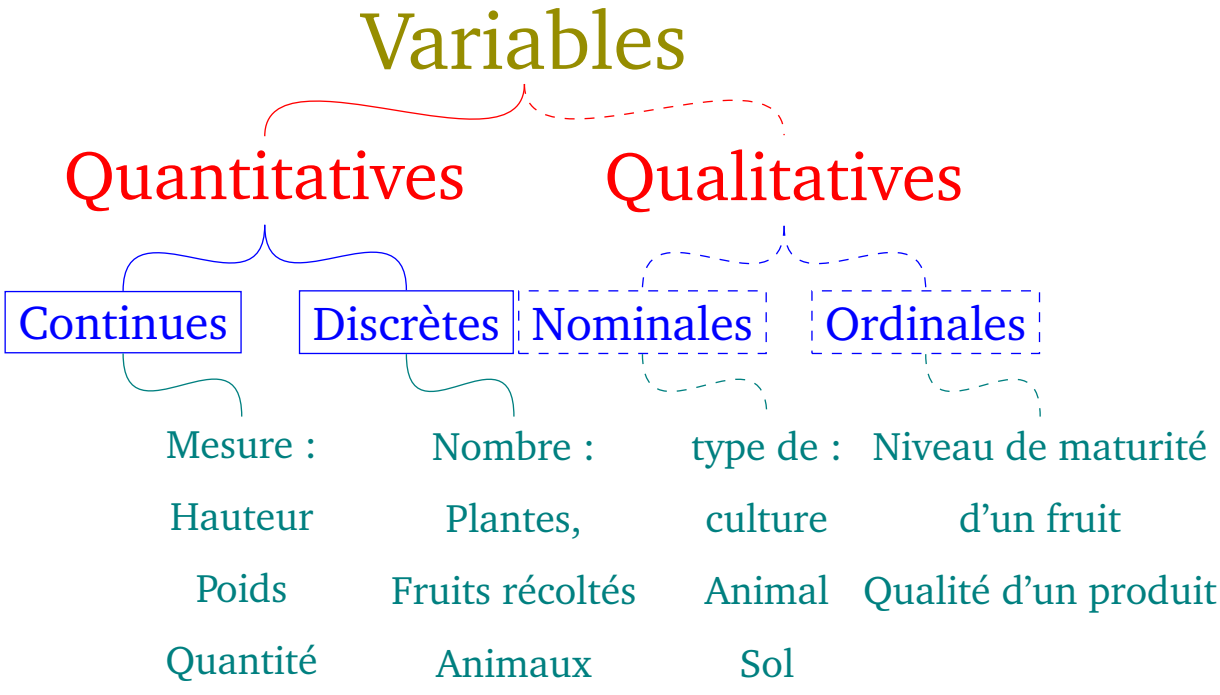


FIGURE 1.1 – Types de données

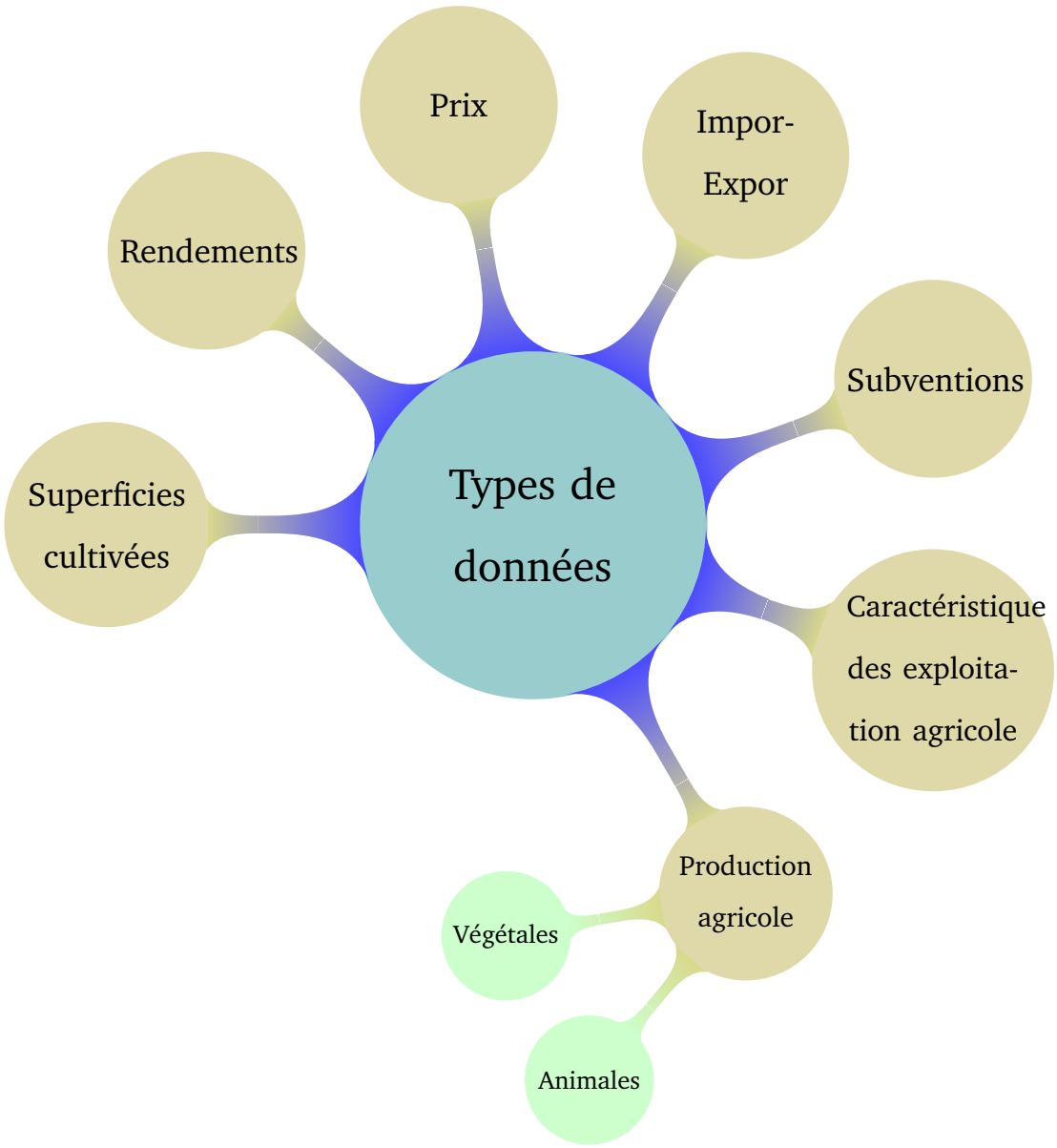


FIGURE 1.2 – Type des données agricoles

En agronomie, la collecte et l'analyse de ces différents types de données (voir figure1.2) permettent de mieux comprendre les interactions entre les cultures, le sol, le climat et d'autres facteurs. Cela conduit à une meilleure prise de décision pour améliorer la productivité agricole, la durabilité environnementale et la sécurité alimentaire.

1.3 Statistiques Descriptives en Agronomie

Les statistiques descriptives jouent un rôle crucial dans l'analyse des données en fournissant un aperçu clair et concis des caractéristiques importantes d'un ensemble de données. Ces statistiques permettent de résumer, de visualiser et de comprendre les propriétés fondamentales des données collectées.

1.3.1 Paramètres de Tendance Centrale

Les paramètres de tendance centrale sont des mesures statistiques qui permettent de caractériser la valeur centrale d'un ensemble de données. Ils fournissent une idée de la "moyenne" ou de la "valeur typique" des observations.

La moyenne

La moyenne est l'une des mesures les plus fondamentales de tendance centrale en statistiques.

① **Arithmétique** : Elle représente la valeur "moyenne" d'un ensemble de données, en calculant la somme de toutes les valeurs observées et en les divisant par le nombre total d'observations.
La formule pour calculer la moyenne $\bar{X}$ d'un ensemble de n observations $(X_1, X_2, ..., X_n)$ est la suivante :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{1.1}$$

Dans le cas des données groupées (discret et continue) :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} n_i \times X_i = \sum f_i \times X_i \tag{1.2}$$

Où les n_i sont les pondérations

Supposons que vous menez une étude pour évaluer le rendement en tonnes par hectare de différentes variétés de maïs dans un champ expérimental. Voici les rendements obtenus pour chaque variété : 7.5, 8.2, 6.9, 7.8, 8.5 tonnes/ha. Pour calculer le rendement moyen, utilisez la formule de la moyenne :

$$\frac{(7.5 + 8.2 + 6.9 + 7.8 + 8.5)}{5} = 7.78\ tonnes/ha$$

La moyenne des rendements des différentes variétés de maïs est de 7.78 tonnes par hectare. Cela donne une idée générale de la performance moyenne des variétés dans l'expérience.

② **Quadratique** : La moyenne quadratique est souvent utilisée pour analyser des grandeurs physiques ou des mesures où les écarts par rapport à la moyenne ont une importance particulière.

$$Q = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2} \tag{1.3}$$

La moyenne quadratique donne plus de poids aux valeurs éloignées de la moyenne, ce qui en fait une mesure sensible à la dispersion. C'est pourquoi elle est souvent utilisée pour quantifier la variation des signaux, des erreurs, des fluctuations et des grandeurs physiques où les écarts absolus sont importants.

Lors de l'analyse de la variation de la taille des fruits d'une même variété de plante à différents endroits d'un champ expérimental. Supposons qu'un chercheur ait mesuré les diamètres (en centimètres) de fruits de la même variété de plante dans trois zones différentes du champ :

- Zone A : 6 cm, 5.5 cm, 6.2 cm, 6.8 cm
- Zone B : 6.3 cm, 5.8 cm, 6.4 cm, 6.7 cm
- Zone C : 6.5 cm, 5.7 cm, 6.6 cm, 6.6 cm

Le chercheur souhaite évaluer la variabilité de la taille des fruits dans chaque zone en prenant en compte les écarts par rapport à la moyenne.

Pour cela, la moyenne quadratique peut être utilisée pour quantifier la dispersion des valeurs autour de la moyenne. Calculons la moyenne quadratique pour chaque zone :

$$Q_A = \sqrt{\frac{1}{4} \times (6^2 + 5.5^2 + 6.2^2 + 6.8^2)} \approx 6.25\ cm$$

$$Q_B = \sqrt{\frac{1}{4} \times (6.3^2 + 5.8^2 + 6.4^2 + 6.7^2)} \approx 6.39\ cm$$

$$Q_C = \sqrt{\frac{1}{4} \times (6.5^2 + 5.7^2 + 6.6^2 + 6.6^2)} \approx 6.43\ cm$$

La moyenne quadratique permet de calculer une valeur unique qui représente la dispersion des mesures autour de la moyenne. Dans ce cas, une valeur de moyenne quadratique plus élevée indiquerait une plus grande dispersion des tailles de fruits autour de la moyenne, ce qui peut être utile pour comparer la variabilité entre différentes zones du champ.
Cet exemple montre comment la moyenne quadratique peut être utilisée en agronomie pour évaluer la variation ou la dispersion des mesures, comme la taille des fruits, en prenant en compte les écarts par rapport à la moyenne.

③ **Géométrique** : La moyenne géométrique est une mesure statistique qui permet de

calculer une moyenne en prenant en compte les valeurs relatives d'un ensemble de données. Contrairement à la moyenne arithmétique qui se base sur la somme des valeurs, la moyenne géométrique prend en compte le produit des valeurs, ce qui la rend particulièrement utile pour les situations où les taux de croissance ou de décroissance sont importants.

$$G = \sqrt[n]{X_1 \times X_2 \times ... \times X_n} = (\prod X_i)^{\frac{1}{n}} \tag{1.4}$$

La moyenne géométrique est particulièrement utile pour les situations où les valeurs sont relatives les unes par rapport aux autres, comme dans les calculs de taux de croissance, d'évolutions exponentielles ou de calculs liés à des grandeurs qui se multiplient entre elles.

Lors de l'évaluation de la croissance des cultures sur plusieurs années. Supposons qu'un chercheur souhaite analyser la croissance en hauteur de différentes variétés de plantes dans un champ expérimental sur une période de trois ans.

Imaginons que les hauteur moyennes (en centimètres) des trois variétés de plantes sont les suivantes pour chaque année :

	Année 1	Année 2	Année 3
Variété A	20	23	26
Variété B	25	28	32
Variété C	22	24	28

Dans ce contexte, la moyenne géométrique serait appropriée pour calculer une moyenne qui prend en compte les taux de croissance relatifs des différentes variétés de plantes sur les trois années.

Calculons la moyenne géométrique pour la hauteur moyenne de chaque variété sur les trois années :

$$G_A = \sqrt[3]{20 \times 23 \times 26} \approx 23.13\ cm$$

$$G_A = \sqrt[3]{25 \times 28 \times 32} \approx 28.39\ cm$$

$$G_A = \sqrt[3]{22 \times 24 \times 28} \approx 24.99\ cm$$

La moyenne géométrique donne une idée de la croissance moyenne relative des plantes sur les trois années, en prenant en compte les valeurs relatives plutôt que simplement les valeurs brutes.

④ **Harmonique** : La moyenne harmonique est une mesure statistique qui diffère des autres types de moyennes en ce qu'elle est principalement utilisée pour des situations où les valeurs sont inverses ou sont influencées par des rapports. Elle est souvent appliquée dans des contextes où les taux, les vitesses ou les relations inverses sont importants.

$$H^{-1} = \frac{1}{n} \sum_{i=1}^{n} \frac{n_i}{X_i} \tag{1.5}$$

$$H = \frac{n}{\sum_{i=1}^{n} \frac{n_i}{X_i}} \tag{1.6}$$

Lors de la mesure du temps moyen de parcours des abeilles entre différentes zones d'un champ de cultures. Imaginons qu'un agriculteur souhaite étudier le temps moyen que prennent les abeilles pour voler entre leur ruche et les différentes cultures, car cela pourrait avoir un impact sur la pollinisation et, par conséquent, sur le rendement des cultures.

Supposons qu'il ait mesuré les temps (en secondes) que prennent trois groupes d'abeilles (G1, G2 et G3) pour parcourir la distance entre la ruche et trois cultures différentes respectivement 4, 6 et 8 secondes. Dans ce scénario, la moyenne harmonique serait une mesure appropriée à utiliser, car les temps de vol sont inverses aux vitesses de vol. Plus précisément, plus une abeille vole rapidement, moins de temps elle met pour parcourir la distance et inversement.

Calculons la moyenne harmonique pour ces valeurs :

$$\frac{3}{\frac{1}{4} \times \frac{1}{6} \times \frac{1}{8}} \approx 5.04 \ \ secondes$$

La moyenne harmonique de 5.04 secondes indique que si l'on considère l'inverse des temps de vol des abeilles, la moyenne du temps qu'il faudrait pour parcourir la même distance serait d'environ 5.04 secondes.

⑤ **Généralisée** : La moyenne généralisée est une notion plus vaste et abstraite qui englobe plusieurs types de moyennes, y compris la moyenne arithmétique, géométrique, harmonique et d'autres formes de moyennes qui peuvent être conçues en fonction des besoins spécifiques de l'analyse. En d'autres termes, une moyenne généralisée est une combinaison de différentes moyennes, chacune ayant un poids ou une importance spécifique.

Dans le contexte de la moyenne généralisée, vous pouvez définir une fonction mathématique qui prend en compte différentes moyennes avec des poids attribués à chacune d'elles. Cette approche permet de tenir compte de différentes caractéristiques ou propriétés des données que vous analysez. Les poids attribués à chaque type de moyenne dépendent du problème particulier que vous essayez de résoudre

et des objectifs de l'analyse.

Soit ϕ une fonction monotone, on appelle moyenne généralisé la valeur M de la variable telle que $\phi(M)$ soit égale à la moyenne arithmétique de $\phi(X)$:

$$\phi(M) = \frac{1}{n}\sum_{i=1}^{n} n_i \phi(X_i) \tag{1.7}$$

$$M = \phi^{-1}[\frac{1}{n}\sum_{i=1}^{n} n_i \phi(X_i)] \tag{1.8}$$

Un exemple concret d'application d'une moyenne généralisée en agronomie pourrait être lorsque vous voulez évaluer globalement la performance de différentes cultures en prenant en compte plusieurs aspects tels que le rendement, la qualité du produit et la résistance aux maladies. Chacun de ces aspects peut être mesuré de manière différente et peut avoir une importance variable en fonction des objectifs de l'analyse.

Supposons que vous souhaitez comparer trois variétés de cultures (A, B et C) en termes de performance globale en agronomie. Vous avez des mesures de rendement en tonnes par hectare, des évaluations de qualité sur une échelle de 1 à 10, et des taux de résistance aux maladies exprimés en pourcentage. Vous voulez combiner ces trois aspects pour obtenir une évaluation globale de chaque variété.

Pour ce faire, vous pourriez utiliser une moyenne généralisée en attribuant des poids spécifiques à chaque aspect en fonction de son importance relative. Par exemple, vous pourriez attribuer un poids de 0.4 au rendement, 0.3 à la qualité et 0.3 à la résistance aux maladies.

La formule pour la moyenne généralisée dans cet exemple serait la suivante :

$$M = 0.4 \times \bar{X}_R + 0.3 \times G_Q + 0.3 \times H_M$$

On calcul la moyenne arithmétique, géométrique et harmonique pour chaque aspect en utilisant les formules appropriées pour chaque type de moyenne. Ensuite, on combine ces moyennes en utilisant les poids attribués à chaque aspect pour obtenir la moyenne généralisée.

Pour une même distribution statistique, l'ordre d'integralité des différentes moyennes est toujours le suivant :

$$H \leq G \leq \bar{X} \leq Q$$

La médiane

La médiane est la valeur centrale dans un ensemble de données triées. Elle divise les données en deux parties égales, où la moitié des valeurs sont plus grandes et l'autre moitié sont plus petites.
Pour calculer la médiane, vous devez d'abord trier les données de manière croissante. Si n est le nombre d'observations, la formule pour trouver la médiane (Me) est :

a) Si n est impair :

$$Me = \frac{n}{2} \tag{1.9}$$

b) Si n est pair :

$$Me = \frac{n+1}{2} \tag{1.10}$$

Imaginons que vous collectiez des données sur la longueur des racines (en centimètres) de différentes variétés de blé dans des conditions expérimentales. Les longueurs des racines mesurées pour chaque variété sont : 15, 18, 16, 20, 19, 22, 21. Pour calculer la médiane, commencez par trier les valeurs dans l'ordre croissant : 15, 16, 18, 19, 20, 21, 22. Comme il y a 7 observations (un nombre impair), la médiane est la quatrième valeur au milieu, qui est 19 centimètres.
Dans cet exemple, la médiane des longueurs des racines des variétés de blé est de 19 centimètres. Cela indique que la moitié des variétés ont des longueurs de racines inférieures à 19 cm et l'autre moitié a des longueurs supérieures.

Le mode

Le mode est une mesure statistique qui identifie la valeur qui apparaît le plus fréquemment dans un ensemble de données. C'est la valeur qui se répète le plus souvent dans l'ensemble des observations. Le mode est utile pour identifier les valeurs les plus courantes ou prédominantes d'une variable.
Il n'y a pas de formule spécifique pour calculer le mode. Vous devez simplement observer les données et identifier la valeur qui apparaît le plus souvent. Dans certains ensembles de données, il peut y avoir plusieurs modes (multimodal) ou aucun mode si toutes les valeurs sont uniques.

Dans un contexte agronomique, le mode peut être utilisé pour identifier les caractéristiques les plus fréquentes des cultures, des sols, des conditions météorologiques, etc. Bien que le mode puisse fournir des informations importantes sur les prévalences, il est recommandé de l'utiliser en conjonction avec d'autres mesures de tendance centrale, comme la moyenne et la médiane, pour obtenir une vue complète de la distribution des données.

Supposons que vous menez une étude pour évaluer la fréquence d'apparition des couleurs des pétales dans une variété de fleurs sauvages. Voici les couleurs des pétales observées : rouge, jaune, bleu, rouge, violet, jaune, rouge, bleu, vert, rouge. Dans cet ensemble de données, la couleur "rouge" apparaît le plus souvent (4 fois), ce qui en fait le mode.

Ces paramètres de tendance centrale fournissent différentes perspectives sur la valeur centrale d'un ensemble de données. Le choix de la mesure dépend de la distribution des données et de l'objectif de l'analyse. En agronomie, ces paramètres sont utilisés pour résumer les caractéristiques des cultures, les propriétés du sol et d'autres variables importantes dans la prise de décisions agricoles.

1.3.2 Paramètres de Dispersion

Les paramètres de dispersion, également appelés paramètres de variabilité, mesurent l'étendue ou la dispersion des valeurs d'un ensemble de données. Ils fournissent des informations sur la répartition et la dispersion des observations autour d'une mesure de tendance centrale, comme la moyenne.

La variance

La variance est une mesure statistique qui quantifie la dispersion ou la variabilité des valeurs d'un ensemble de données par rapport à la moyenne. Elle indique à quel point les

valeurs individuelles s'éloignent de la moyenne, donnant ainsi une idée de l'étendue de la distribution des données.

La formule pour calculer la variance (σ^2 pour une population ou δ^2 pour un échantillon) d'un ensemble de n observations $(X_1, X_2, ..., X_n)$ est la suivante :

— Données individuelles :

$$\sigma^2 = \frac{1}{n}\sum(X_i - \bar{X})^2 \tag{1.11}$$

$$\delta^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2 \tag{1.12}$$

— Données groupées :

$$\sigma^2 = \frac{1}{n}\sum n_i(X_i - \bar{X})^2 \tag{1.13}$$

$$\delta^2 = \frac{1}{n-1}\sum n_i(X_i - \bar{X})^2 \tag{1.14}$$

Où, X_i sont les valeurs observées, $\bar{X}$ est la moyenne des valeurs, n est le nombre total d'observations.

Supposons que vous collectiez des données sur les hauteurs des plants de maïs (en centimètres) dans trois champs expérimentaux différents. Voici les hauteurs mesurées pour chaque champ :

Champ 1	150	155	148	152	158
Champ 2	142	145	148	142	140
Champ 3	160	165	163	162	158

Calculez d'abord la moyenne pour chaque ensemble de données.

Ensuite, calculez la variance pour chaque ensemble de données.

Variance Champ 1 =11.2

Variance Champ 2 =4.8

Variance Champ 3 =5.2

Dans cet exemple, la variance des hauteurs des plants de maïs dans le Champ 1 est plus élevée, indiquant une plus grande dispersion par rapport à la moyenne, tandis que les Champs 2 et 3 ont des variances plus basses, indiquant une dispersion moindre.

Écart-type

L'écart-type est une mesure statistique qui quantifie la dispersion des valeurs d'un ensemble de données par rapport à la moyenne. Il indique à quel point les valeurs indivi-

duelles varient autour de la moyenne, en prenant en compte l'écart entre chaque valeur et la moyenne.

La formule pour calculer l'écart-type (σ pour une population ou δ pour un échantillon) d'un ensemble de n observations $(X_1, X_2, ..., X_n)$ est la racine carrée de la variance :

$$\sigma = \sqrt{\frac{1}{n}\sum(X_i - \bar{X})^2} \tag{1.15}$$

$$\delta = \sqrt{\frac{1}{n-1}\sum(X_i - \bar{X})^2} \tag{1.16}$$

Où, X_i sont les valeurs observées, $\bar{X}$ est la moyenne des valeurs, n est le nombre total d'observations.

Imaginons que vous collectiez des données sur la teneur en azote (en mg/kg) dans des échantillons de sol provenant de différentes zones d'un champ. Voici les niveaux de teneur en azote mesurés pour chaque échantillon :

Zone 1	50	52	48	50	49
Zone 2	40	42	41	39	43
Zone 3	58	55	57	60	59

Calculez d'abord la moyenne pour chaque ensemble de données.

Ensuite, calculez l'écart-type pour chaque ensemble de données.

Ecart-type Zone 1 = 1.3

Ecart-type Zone 2 = 1.6

Ecart-type Zone 3 = 1.8

Dans cet exemple, l'écart-type des teneurs en azote est plus élevé dans la Zone 3, indiquant une plus grande dispersion des valeurs par rapport à la moyenne, tandis que les Zones 1 et 2 ont des écart-types plus bas, indiquant une dispersion moindre.

Kurtosis

La kurtosis est une mesure statistique qui quantifie l'aplatissement ou la forme de la distribution des valeurs d'un ensemble de données par rapport à une distribution normale. Elle indique à quel point la distribution des données diffère de la distribution normale en termes de queues (valeurs extrêmes) et de concentration autour de la moyenne (voir figure 1.3).

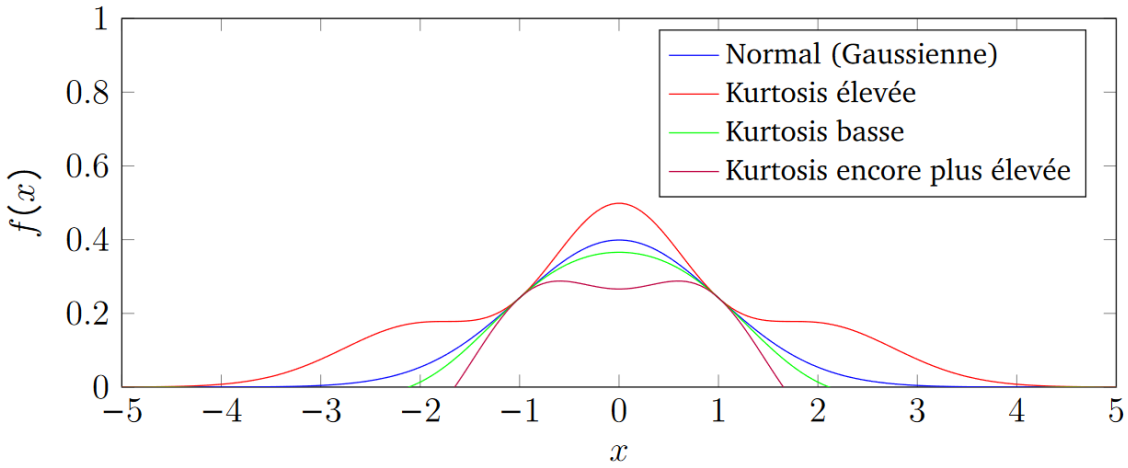


FIGURE 1.3 – Courbes de Kurtosis

La kurtosis est calculée en comparant les moments centrés (moments par rapport à la moyenne) d'un ensemble de données avec ceux d'une distribution normale. La formule standardisée pour calculer la kurtosis (K) est :

$$K = \frac{\frac{\sum(X_i - \bar{X})^4}{n}}{\sigma^4} \tag{1.17}$$

Où, X_i sont les valeurs observées, $\bar{X}$ est la moyenne des valeurs, n est le nombre total d'observations et σ est l'écart-type.

Si vous obtenez une kurtosis supérieure à 3 (la valeur normale pour une distribution normale), cela indique une distribution avec des queues plus épaisses que la distribution normale. Si vous obtenez une kurtosis inférieure à 3, cela indique des queues plus minces.

Imaginons que vous collectiez des données sur les rendements en maïs (en tonnes par hectare) dans différents champs d'une région sur plusieurs années. Voici les rendements observés pour chaque champ :

Champ 1	6.5	7.2	6.8	6.4	6.6
Champ 2	5.8	5.6	5.9	6.0	6.1
Champ 3	8.1	7.9	8.3	8.2	8.5

Calculez d'abord la moyenne et l'écart-type des rendements pour chaque champ. Ensuite, calculez la kurtosis pour chaque ensemble de données en utilisant la formule de kurtosis.

Supposons que pour les trois champs, vous obteniez les kurtosis suivantes :

Champ 1 : 2.6 (Légèrement moins que 3)

Champ 2 : 3.8 (Plus que 3)

Champ 3 : 2.2 (Moins que 3)

Dans cet exemple, les rendements du Champ 2 ont une kurtosis plus élevée, indiquant des queues plus épaisses que la distribution normale. Cela pourrait suggérer la présence de valeurs aberrantes ou d'une distribution non normale dans ce champ.

Skewness

La skewness (asymétrie en français) est une mesure statistique qui quantifie l'asymétrie de la distribution des valeurs d'un ensemble de données par rapport à la distribution normale. Elle indique à quel point la distribution est inclinée vers la gauche (négative skewness) ou vers la droite (positive skewness) par rapport à la moyenne (voir figure 1.4).

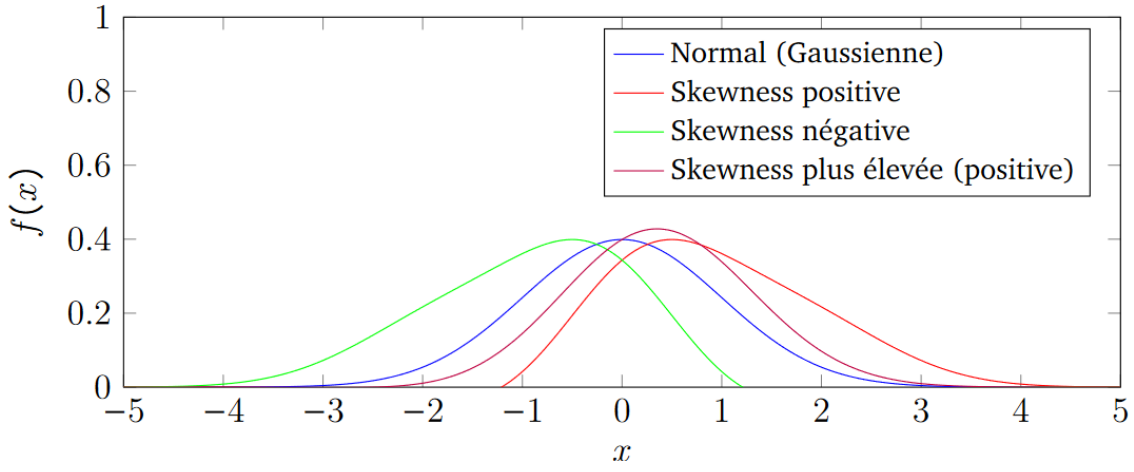


FIGURE 1.4 – Courbes de Skewness

La formule pour calculer la skewness (S) d'un ensemble de n observations ($X_1, X_2, \ldots, X_n$) est :

$$S = \frac{\frac{\sum(X_i - \bar{X})^3}{n}}{\sigma^3} \tag{1.18}$$

Où, X_i sont les valeurs observées, $\bar{X}$ est la moyenne des valeurs, n est le nombre total d'observations et σ est l'écart-type.

Si vous obtenez une skewness positive, cela indique que la distribution est inclinée vers la droite (valeurs plus élevées). Si vous obtenez une skewness négative, cela indique une inclinaison vers la gauche (valeurs plus basses).

Imaginons que vous collectiez des données sur les concentrations en phosphore (en mg/kg) dans le sol de différents champs de maïs. Voici les concentrations observées pour chaque champ :

Champ 1	18	20	21	22	19
Champ 2	30	32	35	29	28
Champ 3	15	17	18	14	20

Calculez d'abord la moyenne et l'écart-type des concentrations en phosphore pour chaque champ. Ensuite, calculez la skewness pour chaque ensemble de données en utilisant la formule de skewness.

Supposons que pour les trois champs 1, 2 et 3, vous obteniez les skewness 0.25, -0.75, et 0.10 respectivement.

Dans cet exemple, le Champ 2 a une skewness négative, indiquant une distribution inclinée vers des valeurs plus basses. Cela pourrait suggérer que les concentrations en phosphore sont généralement inférieures dans ce champ.

Écart absolu moyen

L'écart absolu moyen est une mesure statistique de dispersion qui quantifie la moyenne des distances absolues entre chaque valeur d'un ensemble de données et la moyenne de ces valeurs. Contrairement à la variance qui utilise les carrés des écarts, l'écart absolu moyen utilise les valeurs absolues des écarts, ce qui le rend moins sensible aux valeurs aberrantes. La formule pour calculer l'écart absolu moyen (MAD) d'un ensemble de n observations $(X_1, X_2, \ldots, X_n)$ par rapport à la moyenne $\bar{X}$ est la suivante :

$$MAD = \frac{1}{n} \sum |X_i - \bar{X}| \tag{1.19}$$

Supposons que vous collectiez des données sur les taux de croissance en hauteur (en cm/an) de différentes variétés de blé dans des conditions expérimentales. Voici les taux de croissance observés pour chaque variété :

Variété 1	10	12	9	11	10
Variété 2	11	10	11	9	12
Variété 3	9	8	10	9	10

L'écart absolu moyen pour chaque ensemble de données A, B et C sont respectivement 0.8, 0.4 et 0.6

Dans cet exemple, la variété B a le plus petit écart absolu moyen, indiquant une cohérence plus élevée dans les taux de croissance en hauteur.

Étendue

L'étendue est une mesure statistique simple qui quantifie la différence entre la plus grande et la plus petite valeur d'un ensemble de données. Elle donne une idée de la plage complète des valeurs observées dans un ensemble de données.

La formule pour calculer l'étendue (R) d'un ensemble de valeurs ($X_1, X_2, \ldots, X_n$) est la différence entre la plus grande valeur (X_{max}) et la plus petite valeur (X_{min}) :

$$R = X_{max} - X_{min} \tag{1.20}$$

Supposons que vous collectiez des données sur la hauteur (en centimètres) des plants de tomates dans une serre expérimentale. Voici les hauteurs observées pour chaque plant :

Plant 1	60	55	58	62	59
Plant 2	50	53	51	52	48
Plant 3	70	68	72	71	69

L'étendue pour chaque ensemble de données sont respectivement 7, 5 et 4.

Dans cet exemple, les hauteurs des plants de tomates dans la serre expérimentale ont des étendues différentes. Plant 1 a la plus grande étendue, indiquant une plus grande variation des hauteurs, tandis que Plant 3 a la plus petite étendue, indiquant une variation moindre.

Coefficient de variation

Le coefficient de variation (CV) est une mesure statistique qui exprime la variabilité relative d'un ensemble de données par rapport à sa moyenne. Il permet de comparer la dispersion des données entre différents ensembles, indépendamment de leurs échelles de mesure. Le CV est exprimé en pourcentage.

La formule pour calculer le coefficient de variation (CV) d'un ensemble de données est la suivante :

$$CV = \frac{\sigma}{\bar{X}} \times 100 \tag{1.21}$$

Où σ est l'écart-type des données et $\bar{X}$ est la moyenne.

Supposons que vous collectiez des données sur le poids des fruits (en grammes) de différentes variétés de pommes dans une plantation. Voici les poids observés pour chaque variété :

Variété 1	150	155	148	152	158
Variété 2	120	118	125	123	122
Variété 3	180	175	182	178	185

Le coefficient de variation pour chaque variété sont respectivement $\frac{\bar{A}}{\sigma_A} \times 100$, $\frac{\bar{B}}{\sigma_B} \times 100$ et $\frac{\bar{C}}{\sigma_C} \times 100$.

Dans cet exemple, vous pouvez obtenir des coefficients de variation différents pour chaque variété. Un CV plus élevé indique une plus grande variabilité relative par rapport à la moyenne, tandis qu'un CV plus bas indique une variabilité moindre.

Quartiles

Les quartiles sont des mesures statistiques qui divisent un ensemble de données en quatre parties égales, chacune contenant approximativement un quart des observations. Les trois quartiles les plus couramment utilisés sont le premier quartile ($Q1$), le deuxième quartile ($Q2$) qui est équivalent à la médiane, et le troisième quartile ($Q3$).

Les quartiles sont calculés en ordonnant d'abord les données de manière croissante et en les divisant ensuite en quatre parties égales. Les formules pour calculer les quartiles dépendent de la position des données et du nombre total d'observations.

① **Premier Quartile (Q1)** : Le quartile qui divise les données les plus basses en 25 % et les données restantes en 75 %.

— Si n est impair :

$$Q1 = X_{(\frac{n+1}{4})} \tag{1.22}$$

— Si n est pair :

$$Q1 = \frac{X_{(\frac{n}{4})} + X_{(\frac{n}{4}+1)}}{2} \tag{1.23}$$

② **Deuxième Quartile (Q2)** : Équivalent à la médiane, il divise les données en deux parties égales (50 %).

— Si n est impair :

$$Q2 = Me = \frac{n}{2} \tag{1.24}$$

— Si n est pair :

$$Q2 = Me = \frac{n+1}{2} \tag{1.25}$$

③ **Troisième Quartile (Q3)** : Le quartile qui divise les données les plus élevées en 25 % et les données restantes en 75 %.

— Si n est impair :

$$Q3 = X_{(\frac{3n+1}{4})} \tag{1.26}$$

— Si n est pair :

$$Q3 = \frac{X_{(\frac{3n}{4})} + X_{(\frac{3n}{4}+1)}}{2} \tag{1.27}$$

Supposons que vous collectiez des données sur les rendements en maïs (en tonnes par hectare) dans une région et que vous ayez les données suivantes : $6, 7, 6, 5, 8, 9, 7, 6, 5, 4$

Tout d'abord, triez les données en ordre croissant : $4, 5, 5, 6, 6, 6, 7, 7, 8, 9$

Calculons maintenant les quartiles : Q1=5, Q2=6 et Q3=7

Dans cet exemple, les quartiles vous donnent une idée de la répartition des rendements et des valeurs typiques dans la distribution.

Écart interquartile (IQR)

L'écart interquartile (EI) est une mesure statistique qui quantifie la dispersion des valeurs entre le premier quartile ($Q1$) et le troisième quartile ($Q3$) d'un ensemble de données. Il représente la plage de valeurs qui couvre la moitié centrale des données et est moins sensible aux valeurs aberrantes que l'étendue (voir figure 1.5).

La formule pour calculer l'écart interquartile (EI) est la différence entre le troisième quartile ($Q3$) et le premier quartile ($Q1$) :

$$EI = Q3 - Q1 \tag{1.28}$$

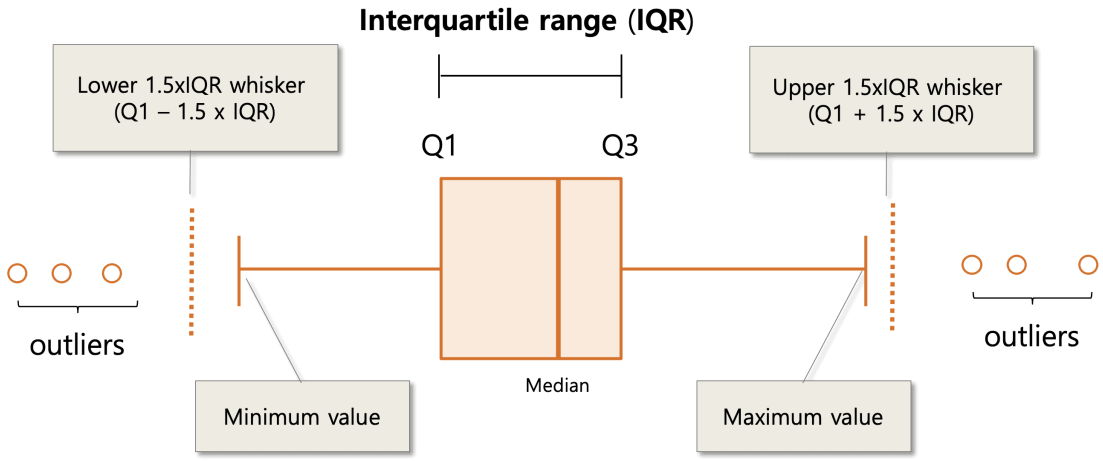


FIGURE 1.5 – Écart interquartile

Source : ezbiocloud

Supposons que vous collectiez des données sur les hauteurs des plants de maïs (en centimètres) dans deux champs expérimentaux. Voici les hauteurs observées pour chaque champ :

Champ A	150	155	148	152	158
Champ B	120	118	125	123	122

Calculez d'abord les quartiles pour chaque ensemble de données et ensuite l'écart interquartile :

Champ A : Q1=148 , Q3=155 , EI=155-148=7

Champ B : Q1=118 , Q3=123 , EI=123-118=5

Dans cet exemple, l'écart interquartile est plus grand pour le Champ A, ce qui suggère une plus grande variabilité interne des hauteurs des plants par rapport au Champ B.

Ces paramètres de dispersion fournissent une compréhension approfondie de la répartition des valeurs autour de la moyenne ou d'une autre mesure de tendance centrale. En agronomie, ils sont utilisés pour évaluer la stabilité des rendements, la variation des caractéristiques des cultures, la cohérence des propriétés du sol, etc. L'analyse de la dispersion est essentielle pour prendre des décisions éclairées et pour comprendre les variations naturelles ou anormales dans les données agronomiques.

```r
# Load necessary libraries
if(!require(moments)) install.packages("moments", dependencies=TRUE)
library(moments)

# Example data
data <- c(12, 15, 22, 20, 18, 30, 35, 45, 50, 12, 22, 25, 30, 33,
    20)

# Central Tendency Parameters
mean_arithmetic <- mean(data)
mean_quadratic <- sqrt(mean(data^2))
mean_geometric <- exp(mean(log(data)))
mean_harmonic <- length(data) / sum(1/data)
mean_generalized <- mean(data^2)^(1/2)  # p=2 generalized mean
    example
median_value <- median(data)
mode_value <- as.numeric(names(sort(table(data), decreasing = TRUE)
    [1]))

# Dispersion Parameters
variance_value <- var(data)
std_deviation <- sd(data)
kurtosis_value <- kurtosis(data)
skewness_value <- skewness(data)
mean_abs_deviation <- mean(abs(data - mean_arithmetic))
range_value <- range(data)
range_diff <- diff(range(data))
coefficient_variation <- (std_deviation / mean_arithmetic) * 100
quartiles <- quantile(data)
iqr_value <- IQR(data)

# Print Results
cat("Central Tendency Parameters:\n")
cat("Arithmetic Mean:", mean_arithmetic, "\n")
cat("Quadratic Mean:", mean_quadratic, "\n")
```

```r
cat("Geometric Mean:", mean_geometric, "\n")
cat("Harmonic Mean:", mean_harmonic, "\n")
cat("Generalized Mean (p=2):", mean_generalized, "\n")
cat("Median:", median_value, "\n")
cat("Mode:", mode_value, "\n\n")

cat("Dispersion Parameters:\n")
cat("Variance:", variance_value, "\n")
cat("Standard Deviation:", std_deviation, "\n")
cat("Kurtosis:", kurtosis_value, "\n")
cat("Skewness:", skewness_value, "\n")
cat("Mean Absolute Deviation:", mean_abs_deviation, "\n")
cat("Range:", range_value, " (Difference:", range_diff, ")\n")
cat("Coefficient of Variation (%):", coefficient_variation, "\n")
cat("Quartiles:", quartiles, "\n")
cat("Interquartile Range:", iqr_value, "\n")
```

Listing 1.1 – R Script for Central Tendency and Dispersion Parameters

```
Central Tendency Parameters:
Arithmetic Mean: 25.66667
Quadratic Mean: 28.91832
Geometric Mean: 23.47067
Harmonic Mean: 20.87058
Generalized Mean (p=2): 28.91832
Median: 22
Mode: 12

Dispersion Parameters:
Variance: 136.1905
Standard Deviation: 11.67011
Kurtosis: -0.9756393
Skewness: 0.543949
Mean Absolute Deviation: 8.888889
Range: 12 50  (Difference: 38 )
Coefficient of Variation (%): 45.4621
Quartiles: 12.0 20.0 25.0 30.0 35.0
```

```
Interquartile Range: 15
```

Listing 1.2 – R Output for Central Tendency and Dispersion Parameters

1.4 Les tableaux

Un tableau statistique est une présentation organisée de données numériques ou catégorielles, conçue pour faciliter l'analyse et la compréhension des caractéristiques d'une population ou d'un échantillon. Les tableaux statistiques sont largement utilisés dans de nombreux domaines, y compris la science, l'économie, la sociologie, la santé, l'agriculture et bien d'autres, pour résumer, comparer et interpréter des données.

① **Tableau de fréquences** : Ce type de tableau est utilisé pour résumer la distribution d'une variable catégorielle (qualitative). Il indique le nombre ou la fréquence de chaque catégorie dans l'échantillon ou la population.

```r
# Donn es simul es : Types de cultures
cultures <- c("Bl  ", "Ma s", "Riz", "Soja", "Avoine", "Bl  ", "
   Bl  ", "Ma s", "Riz", "Soja", "Bl  ")

# Tableau de fr quences
table(cultures)
```

Listing 1.3 – R Script for Frequency Table

```
cultures
Avoine    Bl     Ma s    Riz    Soja
    1      4       2      2       2
```

Listing 1.4 – Frequency Table Output

② **Tableau de contingence** : Il est utilisé pour analyser les relations entre deux variables catégorielles en affichant le nombre de cas dans chaque combinaison de catégories.

```r
# Donn es simul es : Type de culture et r gion
cultures <- c("Bl  ", "Bl  ", "Ma s", "Riz", "Soja", "Bl  ", "Riz",
    "Avoine", "Ma s", "Soja")
```

```r
regions <- c("Nord", "Sud", "Est", "Nord", "Sud", "Est", "Nord", "
    Sud", "Est", "Nord")

# Tableau de contingence
table(cultures, regions)
```

Listing 1.5 – R Script for Contingency Table

```
         regions
cultures Nord Sud Est
    Avoine    0   1   0
    Bl        1   0   2
    Ma s      0   0   2
    Riz       2   0   0
    Soja      1   1   0
```

Listing 1.6 – Contingency Table Output

③ **Tableau croisé** : Il est similaire au tableau de contingence mais peut également inclure les proportions ou pourcentages relatifs aux totaux.

```r
# Donn es simul es : Cultures et rendement
cultures <- c("Bl  ", "Ma s", "Riz", "Soja", "Avoine")
rendement <- c("Faible", "Moyen", "  lev  ", "Moyen", "Faible")

# Tableau crois
table(cultures, rendement)
```

Listing 1.7 – R Script for Cross Tabulation

```
         rendement
cultures Faible Moyen   lev
    Avoine      1      0     0
    Bl          1      0     0
    Ma s        0      1     0
    Riz         0      0     1
    Soja        0      1     0
```

Listing 1.8 – Cross Tabulation Output

④ **Tableau de répartition** : C'est un tableau utilisé pour résumer la distribution d'une variable quantitative en classes ou en intervalles.

```r
# Donn es simul es : Rendement par type de sol
type_sol <- c("Argileux", "Sableux", "Limoneux", "Argileux", "
   Sableux")
rendement <- c(3.2, 2.8, 3.5, 3.7, 3.0)

# Tableau de r partition
repartition <- table(type_sol)
repartition
```

Listing 1.9 – R Script for Distribution Table

```
type_sol
Argileux  Limoneux  Sableux
       2         1         2
```

Listing 1.10 – Distribution Table Output

⑤ **Tableau descriptif** : Il peut afficher des statistiques descriptives telles que la moyenne, la médiane, l'écart-type, etc., pour une variable quantitative.

```r
# Donn es simul es : Rendement d'une culture
rendement <- c(3.2, 3.5, 3.8, 4.0, 4.2)

# Tableau descriptif
descriptif <- summary(rendement)
descriptif
```

Listing 1.11 – R Script for Descriptive Table

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.200   3.500   3.800   3.740   4.000   4.200
```

Listing 1.12 – Descriptive Table Output

⑥ **Tableau de corrélation** : Il est utilisé pour montrer les corrélations entre plusieurs variables quantitatives.

```r
# Donn es simul es : Surface et rendement de cultures
surface <- c(1.2, 2.3, 3.1, 4.0, 5.2)
rendement <- c(2.1, 3.5, 3.8, 4.2, 5.5)

# Tableau de corr lation
correlation <- cor(surface, rendement)
correlation
```

Listing 1.13 – R Script for Correlation Table

```
[1] 0.9884596
```

Listing 1.14 – Correlation Table Output

⑦ **Tableau de comparaison** : Il permet de comparer différentes mesures ou variables entre différents groupes ou échantillons.

```r
# Donn es simul es : Rendement de deux cultures diff rentes
rendement_culture1 <- c(3.2, 3.5, 3.8, 4.0, 3.9)
rendement_culture2 <- c(2.8, 3.0, 3.1, 3.5, 3.4)

# Tableau de comparaison
comparaison <- data.frame(
  Culture1 = rendement_culture1,
  Culture2 = rendement_culture2
)
comparaison
```

Listing 1.15 – R Script for Comparison Table

```
  Culture1 Culture2
1      3.2      2.8
2      3.5      3.0
3      3.8      3.1
4      4.0      3.5
5      3.9      3.4
```

Listing 1.16 – Comparison Table Output

⑧ **Tableau de régression** : Il est utilisé dans l'analyse de régression pour présenter les coefficients, les p-values et autres statistiques de régression.

```r
# Donn es simul es : Variables explicatives et d pendantes
surface <- c(1.2, 2.3, 3.1, 4.0, 5.2)
rendement <- c(2.1, 3.5, 3.8, 4.2, 5.5)

# Mod le de r gression lin aire
modele <- lm(rendement ~ surface)
summary(modele)
```

Listing 1.17 – R Script for Regression Table

```
Call:
lm(formula = rendement ~ surface)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14929 -0.08429 -0.02321  0.12393  0.16893

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4982     0.1984   7.552  0.00492 **
surface       0.7631     0.0567  13.452  0.00107 **
---
Signif. codes:
    0    ***    0.001    **    0.01    *    0.05    .    0.1
    1

Residual standard error: 0.1468 on 3 degrees of freedom
Multiple R-squared:  0.9833,    Adjusted R-squared:  0.9778
F-statistic: 180.9 on 1 and 3 DF,  p-value: 0.001074
```

Listing 1.18 – Regression Table Output

⑨ **Tableau temporel** : Il montre l'évolution d'une variable ou plusieurs variables au fil du temps.

```r
# Donn es simul es : Rendement par ann e
annees <- c(2016, 2017, 2018, 2019, 2020)
rendement <- c(3.2, 3.5, 3.8, 4.0, 4.2)

# Tableau temporel
temporel <- data.frame(
  Ann e = annees,
  Rendement = rendement
)
temporel
```

Listing 1.19 – R Script for Temporal Table

```
  Ann e Rendement
1  2016      3.2
2  2017      3.5
3  2018      3.8
4  2019      4.0
5  2020      4.2
```

Listing 1.20 – Temporal Table Output

1.5 Les graphiques

La représentation graphique des données est une méthode visuelle puissante pour explorer, communiquer et interpréter des informations clés. Différents types de graphiques sont utilisés pour mettre en évidence des tendances, des variations et des relations entre des variables spécifiques.

Le choix du type de graphique dépend de la nature des données, des objectifs de la visualisation et des informations que vous souhaitez communiquer. Chaque type de graphique offre des perspectives différentes sur les données agronomiques et peut aider à prendre des décisions éclairées en matière de gestion des cultures, d'analyse des performances et de compréhension des relations entre les variables.

① **Histogramme** : Un histogramme est utilisé pour représenter la distribution des **données continues** en les regroupant en intervalles (classes) et en affichant le nombre d'observations dans chaque classe. Cela permet de visualiser la fréquence et la forme de la distribution des données.

Pour créer un histogramme (voir figure 1.6), suivez ces étapes :

(a) Déterminez le nombre d'intervalles (classes) que vous souhaitez utiliser. Plus d'intervalles donnent plus de détails, mais trop peu peuvent masquer la structure des données.

(b) Calculez la largeur de chaque intervalle en divisant la plage totale des données par le nombre d'intervalles.

(c) Créez des intervalles qui couvrent toute la plage des données, en commençant par la valeur minimale et en ajoutant la largeur de l'intervalle à chaque itération.

(d) Comptez combien d'observations tombent dans chaque intervalle.

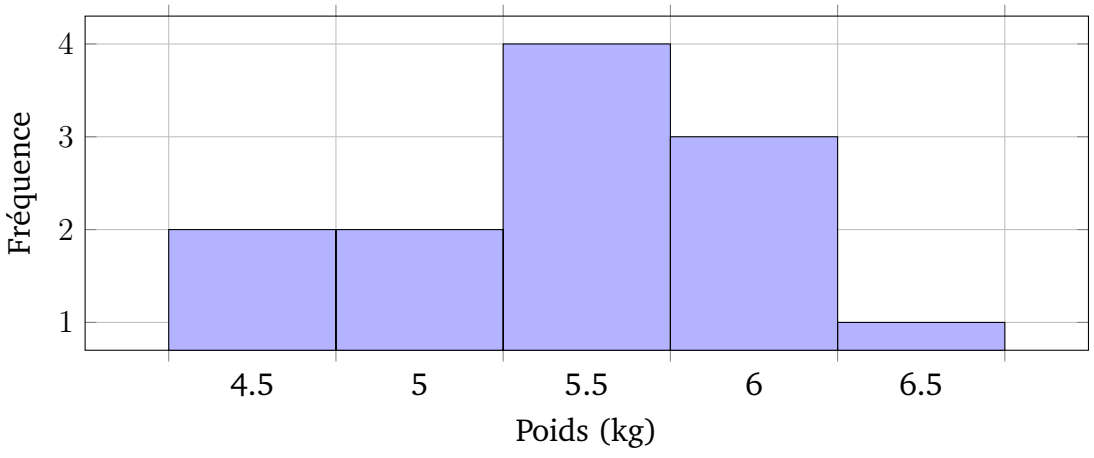


FIGURE 1.6 – Histogramme des poids de pastèques

Supposons que vous collectiez des données sur les poids (en kg) de pastèques dans une plantation. Voici les poids observés pour un échantillon de pastèques :

4.5,5.2,5.8,6.2,6.5,5.7,5.6,5.1,4.9,6.0,6.3,5.5

Les données pourraient être regroupées en intervalles de $0.5kg$:

$[4.5 - 5.0[$: 2 observations

$[5.0 - 5.5[$: 2 observations

$[5.5 - 6.0[$: 4 observations

$[6.0 - 6.5[$: 3 observations

$[6.5 - 7.0[$: 1 observation

```r
# Donn es simul es : poids des past ques en kg
poids <- c(4.5, 5.2, 5.8, 6.2, 6.5, 5.7, 5.6, 5.1, 4.9, 6.0, 6.3,
    5.5)

# Cr ation de l'histogramme
hist(poids,
     breaks = seq(4.5, 7, by = 0.5), # D finir les intervalles de
    0.5 kg
     main = "Distribution des poids des past ques",
     xlab = "Poids (kg)",
     ylab = "Nombre d'observations",
     col = "lightblue",
     border = "black")
```

Listing 1.21 – R Script for Watermelon Weight Distribution Histogram

② **Diagramme en Barres** : Ce diagramme est utilisé pour représenter des **données discrètes ou catégorielles**. Chaque catégorie est représentée par une barre, dont la hauteur est proportionnelle à la fréquence ou à la valeur associée à cette catégorie. Pour créer un diagramme en barres (voir figure 1.7), suivez ces étapes :

(a) Identifiez les catégories (dans ce cas, les trois variétés de pommes).

(b) Déterminez les valeurs à représenter pour chaque catégorie (moyenne, somme, fréquence, etc.).

(c) Tracez un axe horizontal pour représenter les catégories et un axe vertical pour représenter les valeurs.

(d) Dessinez des barres rectangulaires au-dessus de chaque catégorie, dont la hauteur correspond à la valeur associée à cette catégorie.

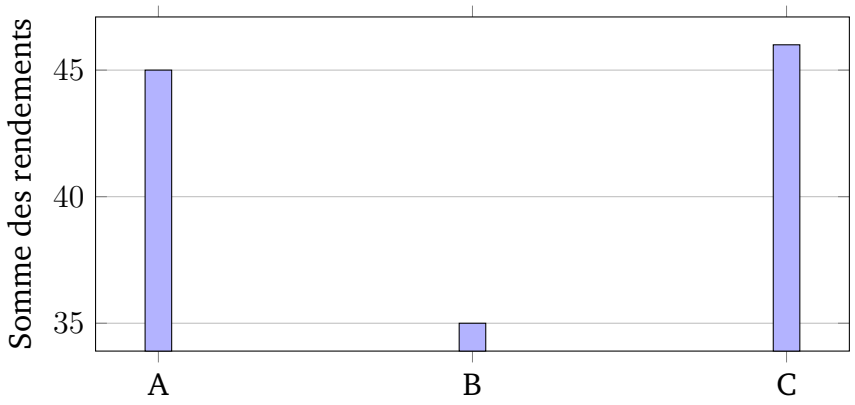


FIGURE 1.7 – Diagramme en Barres des rendements de variétés de pommes

Supposons que vous collectiez des données sur les rendements de trois variétés de pommes (A, B et C) dans un verger. Voici les rendements observés en tonnes par hectare pour chaque variété :

Variété A	8	10	7	9	11
Variété B	6	8	7	9	5
Variété C	9	9	8	10	10

Dans cet exemple, vous auriez trois barres correspondant aux variétés A, B et C. La hauteur de chaque barre représenterait la moyenne des rendements pour chaque variété.

Le diagramme en barres vous permettrait de comparer visuellement les rendements des différentes variétés de pommes et d'identifier celle qui a le rendement moyen le plus élevé.

```r
# Donn es simul es : rendements en tonnes par hectare
rendements_A <- c(8, 10, 7, 9, 11)
rendements_B <- c(6, 8, 7, 9, 5)
rendements_C <- c(9, 9, 8, 10, 10)

# Calcul des moyennes
moyenne_A <- mean(rendements_A)
```

```r
moyenne_B <- mean(rendements_B)
moyenne_C <- mean(rendements_C)

# Cr ation du graphique
barplot(c(moyenne_A, moyenne_B, moyenne_C),
        names.arg = c("Vari t  A", "Vari t  B", "Vari t  C"),
        main = "Comparaison des rendements moyens des vari t s de pommes",
        xlab = "Vari t ",
        ylab = "Rendement moyen (tonnes/ha)",
        col = c("lightblue", "lightgreen", "lightcoral"),
        border = "black")
```

Listing 1.22 – R Script for Bar Chart of Apple Varieties

③ **Diagramme en Secteurs (Camembert)** : Utilisé pour représenter la répartition relative des différentes catégories d'une **variable discrète** en utilisant des secteurs circulaires. Les tailles des secteurs sont proportionnelles aux pourcentages ou aux fréquences de chaque catégorie.
Pour créer un diagramme en secteurs (camembert) (voir figure 1.8), suivez ces étapes :

(a) Identifiez les catégories que vous souhaitez représenter (dans ce cas, les différentes cultures).

(b) Calculez la proportion de chaque catégorie par rapport à la somme totale des données.

(c) Calculez la taille angulaire de chaque secteur en utilisant la proportion calculée.

(d) Tracez un cercle pour représenter le diagramme en secteurs.

(e) Divisez le cercle en secteurs de tailles angulaires correspondant aux proportions calculées pour chaque catégorie.

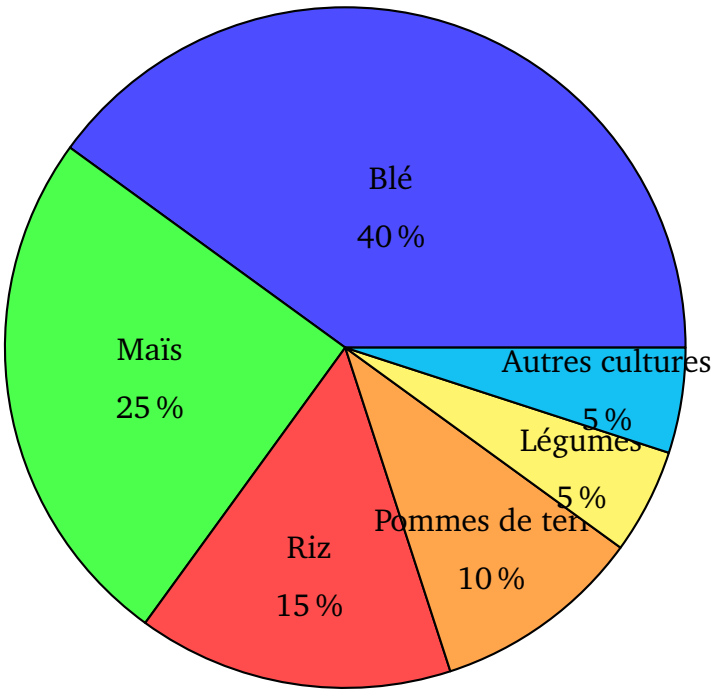


FIGURE 1.8 – La répartition des cultures dans une région agricole

Supposons que vous collectiez des données sur la répartition des cultures dans une région agricole. Vous avez les informations suivantes sur la superficie totale de terre cultivée : Blé : 40%, Maïs : 25%, Riz : 15%, Pommes de terre : 10%, égumes : 5% et Autres cultures : 5%.

Le diagramme en secteurs vous permettrait de visualiser rapidement la répartition des cultures dans la région. Vous pourriez facilement observer quelle culture occupe la plus grande part de la superficie cultivée.

```
# Donn es simul es : r partition des cultures en pourcentage
cultures <- c("Bl  ", "Ma s", "Riz", "Pommes de terre", "L gumes",
    "Autres cultures")
pourcentages <- c(40, 25, 15, 10, 5, 5)

# Cr ation du diagramme en secteurs
pie(pourcentages,
    labels = cultures,
    main = "R partition des cultures dans une r gion agricole",
    col = rainbow(length(cultures)),
```

```
    clockwise = TRUE,
    border = "black")
```

Listing 1.23 – R Script for Pie Chart of Crop Distribution

④ **Diagramme de Dispersion** : Aussi appelé nuage de points, il est utilisé pour représenter la relation entre **deux variables continues**. Chaque point représente une observation avec ses valeurs pour les deux variables, ce qui permet de détecter des tendances, des regroupements ou des relations.
Pour créer un diagramme de dispersion (voir figure 1.9), suivez ces étapes :

(a) Identifiez les deux variables que vous souhaitez comparer (dans ce cas, la taille des feuilles et la quantité d'eau).

(b) Associez chaque paire de valeurs pour les deux variables à un point sur le graphique.

(c) Tracez un axe horizontal pour représenter la première variable et un axe vertical pour représenter la deuxième variable.

(d) Placez des points correspondant aux paires de valeurs sur le graphique.

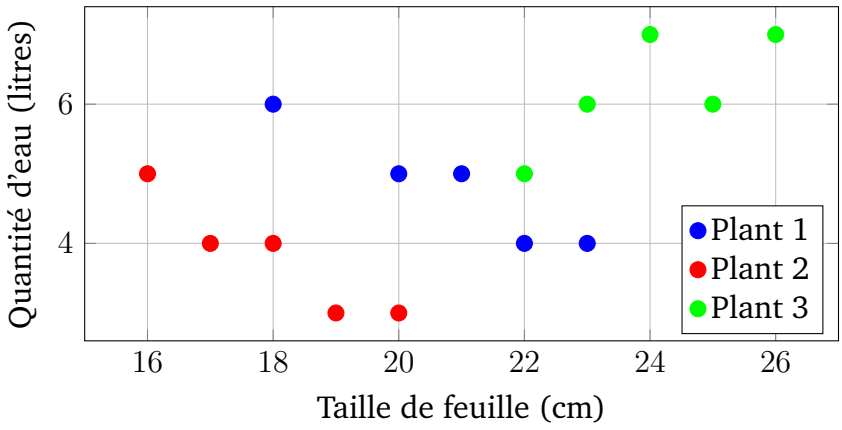


FIGURE 1.9 – Dispersion la taille des feuilles vs. la quantité d'eau consommée

Supposons que vous collectiez des données sur la taille des feuilles (en cm) et la quantité d'eau consommée (en litres) par des plants de maïs dans une expérience de gestion de l'irrigation. Voici les données observées pour chaque plant :

Plant 1 : $(20, 5), (22, 4), (18, 6), (23, 4), (21, 5)$

Plant 2 : $(17, 4), (19, 3), (16, 5), (20, 3), (18, 4)$

Plant 3 : $(25, 6), (24, 7), (22, 5), (26, 7), (23, 6)$

Dans cet exemple, chaque point sur le diagramme de dispersion représente un plant de maïs avec sa taille de feuille en abscisse et sa quantité d'eau en ordonnée. En examinant le graphique, vous pourriez observer s'il y a une tendance de croissance de la taille des feuilles en fonction de la quantité d'eau.

```r
# Donn es simul es : taille des feuilles et quantit  d'eau
taille_feuilles <- list(
  Plant1 = c(20, 22, 18, 23, 21),
  Plant2 = c(17, 19, 16, 20, 18),
  Plant3 = c(25, 24, 22, 26, 23)
)

eau <- list(
  Plant1 = c(5, 4, 6, 4, 5),
  Plant2 = c(4, 3, 5, 3, 4),
  Plant3 = c(6, 7, 5, 7, 6)
)

# Cr ation du diagramme de dispersion
plot(taille_feuilles$Plant1, eau$Plant1,
     col = "blue", pch = 16, xlim = c(15, 30), ylim = c(2, 8),
     xlab = "Taille des feuilles (cm)", ylab = "Quantit  d'eau (litres)",
     main = "Diagramme de dispersion : Taille des feuilles vs Quantit  d'eau")
points(taille_feuilles$Plant2, eau$Plant2, col = "green", pch = 16)
points(taille_feuilles$Plant3, eau$Plant3, col = "red", pch = 16)
```

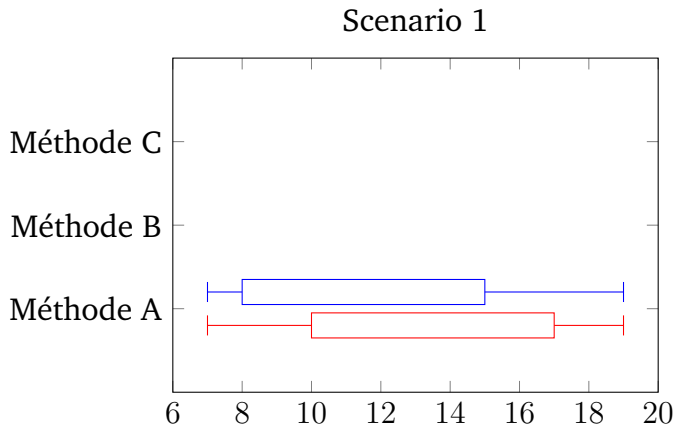


FIGURE 1.10 – Caption

```
21
22 legend("topright", legend = c("Plant 1", "Plant 2", "Plant 3"),
23        col = c("blue", "green", "red"), pch = 16)
```

Listing 1.24 – R Script for Scatter Plot of Leaf Size vs Water Consumption

⑤ **Graphique en Boîte (Box Plot)** : Ce graphique représente la distribution des données en montrant la médiane, les quartiles et les valeurs aberrantes. Il offre une vue d'ensemble des mesures de tendance centrale et de dispersion, ainsi que des valeurs extrêmes.

Pour créer un graphique en boîte, suivez ces étapes :

(a) Calculez les quartiles $(Q1, Q2, Q3)$ pour chaque ensemble de données.

(b) Calculez les valeurs aberrantes potentiels (valeurs qui sont à l'extérieur de $Q1 - 1.5 \times EI$ et $Q3 + 1.5 \times EI$, où EI est l'écart interquartile).

(c) Tracez une ligne verticale pour représenter la médiane à l'intérieur d'une boîte, avec la largeur de la boîte entre les quartiles.

(d) Tracez des segments (moustaches) à partir de la boîte pour représenter les valeurs extrêmes.

(e) Marquez les valeurs aberrantes potentielles le cas échéant.

Supposons que vous collectiez des données sur les rendements de blé (en tonnes par hectare) pour trois méthodes de traitement du sol : A, B et C. Voici les rendements observés pour chaque méthode :

Méthode A : $3.2, 3.5, 4.0, 4.2, 3.8$

Méthode B : $3.7, 3.9, 3.6, 3.8, 3.5$

Méthode C : $4.5, 4.3, 4.7, 4.0, 4.2$

```r
# Donnees simulees : rendements en tonnes par hectare
rendements_A <- c(3.2, 3.5, 4.0, 4.2, 3.8)
rendements_B <- c(3.7, 3.9, 3.6, 3.8, 3.5)
rendements_C <- c(4.5, 4.3, 4.7, 4.0, 4.2)

# Combinaison des donnees dans un seul vecteur et creation des labels
rendements <- c(rendements_A, rendements_B, rendements_C)
methodes <- factor(rep(c("Methode A", "Methode B", "Methode C"), each = 5))

# Creation du graphique en boite
boxplot(rendements ~ methodes,
        main = "Distribution des rendements de ble par methode de traitement du sol",
        xlab = "Methode de traitement",
        ylab = "Rendement (tonnes/ha)",
        col = c("lightblue", "lightgreen", "lightcoral"),
        border = "black")
```

Listing 1.25 – R Script for Box Plot of Wheat Yields by Treatment Method

⑥ **Graphique Temporel (Série Temporelle)** : Utilisé pour représenter l'évolution des **variables au fil du temps**. Cela permet de visualiser les tendances saisonnières, les cycles et les changements au fil des années.

⑦ **Graphique Radar** : Il est utilisé pour comparer plusieurs variables sur un même graphique en utilisant des axes radiaux partant d'un point central. Cela permet de

visualiser les forces et les faiblesses de différentes variables.

⑧ **Graphique en Aire** : Ce type de graphique montre la répartition des données dans le temps ou sur une autre dimension. Les aires sous les courbes représentent la contribution relative des différentes catégories à travers le temps.

⑨ **Graphique en Violon** : Une combinaison de l'histogramme et du diagramme en boîte qui montre la distribution des données, la médiane, les quartiles et la densité des observations.

⑩ **Graphique de Corrélation** : Utilisé pour représenter la relation entre **deux variables continues**. Il permet de visualiser la force et la direction de la corrélation entre les variables.

Choisir le bon type de graphique dépend des objectifs d'analyse et de la nature des données. En combinant ces représentations visuelles avec des analyses statistiques, les professionnels de l'agronomie peuvent prendre des décisions éclairées pour améliorer les pratiques agricoles et la gestion des ressources.

CHAPTER 2

Tests d'Hypothèses et Intervalles de Confiance

Ce chapitre permet de comprendre comment les statistiques peuvent être utilisées pour prendre des décisions éclairées. Les tests d'hypothèses et les intervalles de confiance sont des outils fondamentaux qui vous permettront de tirer des conclusions basées sur les données collectées lors de vos expériences et études agronomiques.

Lorsque vous menez des expériences en agronomie, il est rare d'obtenir des résultats parfaitement identiques à chaque essai. Les variations naturelles peuvent avoir un impact sur vos observations. Les tests d'hypothèses et les intervalles de confiance vous permettent de quantifier ces variations et de déterminer si les différences observées sont significatives ou simplement dues au hasard.

Au cours de ce chapitre, vous allez apprendre comment :

① Formuler et tester des hypothèses statistiques pour déterminer si vos observations sont statistiquement significatives.

② Utiliser les intervalles de confiance pour estimer la précision de vos mesures et la variabilité de vos données.

③ Prendre des décisions informées basées sur des preuves statistiques solides.

Que vous analysiez les rendements de différentes cultures, l'efficacité de divers traitements agronomiques ou les effets des conditions environnementales sur les cultures, les tests d'hypothèses et les intervalles de confiance seront vos outils clés pour interpréter les données de manière rigoureuse et prendre des décisions éclairées.

2.1 Tests d'Hypothèses

Les tests d'hypothèses sont des procédures statistiques utilisées pour évaluer si une hypothèse formulée sur une population ou un échantillon est statistiquement plausible. Ces tests fournissent une méthodologie formelle pour prendre des décisions basées sur des échantillons de données, en évaluant si les différences observées entre les groupes ou les variables sont réellement significatives ou simplement dues au hasard.

Formulation des Tests d'Hypothèses :

Chaque test d'hypothèse suit une structure similaire :

1. **Hypothèse Nulle (H0)** : Une affirmation généralement formulée comme l'absence d'effet ou de différence.
2. **Hypothèse Alternative (H1 ou Ha)** : L'affirmation que vous cherchez à prouver.
3. **Niveau de Signification (α)** : Le seuil prédéfini pour déterminer si les résultats sont statistiquement significatifs. Couramment utilisé est $\alpha = 0,05$, ce qui signifie que vous accepterez une probabilité de 5% d'erreur.
4. **Test Statistique** : Un calcul statistique spécifique basé sur les données de l'échantillon.
5. **Règle de Décision** : Si la valeur du test statistique est inférieure au seuil (niveau de signification), vous pouvez rejeter l'hypothèse nulle en faveur de l'hypothèse alternative.

2.1.1 Test sur la Moyenne

Le test sur la moyenne est une procédure statistique utilisée pour déterminer si la moyenne d'un échantillon est significativement différente d'une valeur théorique ou de la moyenne d'un autre échantillon. C'est l'un des tests d'hypothèses les plus couramment utilisés en agronomie pour évaluer l'impact de traitements, de méthodes agronomiques ou d'autres facteurs sur les variables mesurées.

Supposons que vous voulez tester si la moyenne d'une variable dans un échantillon diffère significativement d'une valeur théorique μ_0. Les hypothèses sont formulées comme suit :

$$\begin{cases} H0 : & \mu = \mu_0 \\ H1 : & \mu \neq \mu_0 \end{cases} \tag{2.1}$$

Corps d'Hypothèses :

$$\begin{cases} H0 : \text{ Il n'y a pas de différence significative entre la moyenne de l'échantillon} \\ \qquad \text{et la valeur théorique } \mu_0 \\ H1 : \text{ Il y a une différence significative entre la moyenne de l'échantillon} \\ \qquad \text{et la valeur théorique } \mu_0 \end{cases}$$

La statistique du test pour le test sur la moyenne est généralement basée sur la formule :

$$t = \frac{\bar{X} - \mu_0}{\frac{\sigma_X}{\sqrt{n}}} \sim \mathcal{T}_{(n)} \tag{2.2}$$

Où $\bar{X}$ est la moyenne de l'échantillon, σ_X est l'écart-type de l'échantillon et n est la taille de l'échantillon. Cette statistique suit approximativement une distribution $\mathcal{T}$ de Student si les conditions sont remplies.

Décision : Si la valeur absolue de la statistique de test t est supérieure à une valeur critique $t_{\frac{\alpha}{2}}$ (obtenue à partir de la distribution $\mathcal{T}$ de Student et du niveau de signification α), vous pouvez rejeter l'hypothèse nulle $H0$ en faveur de l'hypothèse alternative $H1$. Sinon, vous ne pouvez pas rejeter $H0$.

Supposons que vous testiez l'effet d'un nouvel engrais sur le rendement des plants de tomates. Vous avez collecté un échantillon de 20 plants et obtenu une moyenne de rendement de $8\ kg$ par plante. Vous voulez tester si ce rendement diffère significativement de la moyenne habituelle de $7\ kg$ par plante. Utilisons un niveau de signification $\alpha = 0.05$.

- Hypothèse Nulle : $H0 : \mu = 7$ (Pas de différence significative)
- Hypothèse Alternative : $H1 : \mu \neq 7$ (Différence significative)

Calcul de la statistique du test : $t = \frac{8-7}{\frac{\sigma}{\sqrt{20}}}$
En utilisant une table de distribution $\mathcal{T}$ de Student, vous trouvez une valeur critique $t_{\frac{\alpha}{2}}$ correspondant à $\frac{\alpha}{2} = 0.025$ pour $n = 20$. Si $|t|$ est supérieur à $t_{\frac{\alpha}{2}}$, vous pouvez rejeter $H0$ et conclure s'il y a une différence significative entre les rendements.

```r
# Donn es simul es
mean_observed <- 8       # Moyenne observ e
mean_habitual <- 7       # Moyenne habituelle
sigma <- 1.2             #  cart -type estim  (valeur hypoth tique , adapter si disponible)
n <- 20                  # Taille de l'  chantillon
alpha <- 0.05            # Niveau de signification

# Calcul de la statistique du test t
t_value <- (mean_observed - mean_habitual) / (sigma / sqrt(n))

# Calcul de la valeur critique t pour un test bilat ral
t_critical <- qt(1 - alpha / 2, df = n - 1)

# Affichage des r sultats
cat("Valeur de t observ e:", t_value, "\n")
cat("Valeur critique de t:", t_critical, "\n")

# Conclusion
```

```
19 if (abs(t_value) > t_critical) {
20     cat("Rejet de l'hypoth se nulle. Il y a une diff rence
       significative dans les rendements.\n")
21 } else {
22     cat("Non rejet de l'hypoth se nulle. Aucune diff rence
       significative dans les rendements.\n")
23 }
```

Listing 2.1 – R Script for t-Test to Compare Tomato Plant Yields

```
1 Valeur de t observ e: 3.722
2 Valeur critique de t: 2.093
3
4 Rejet de l'hypoth se nulle. Il y a une diff rence significative
     dans les rendements.
```

Listing 2.2 – Test sur la Moyenne

2.1.2 Test sur la Proportion

Le test sur la proportion est une procédure statistique utilisée pour déterminer si la proportion d'une caractéristique dans un échantillon diffère significativement d'une valeur théorique ou d'une proportion d'un autre échantillon. C'est un outil essentiel en agronomie pour évaluer les différences dans la répartition des caractéristiques entre différentes populations ou groupes.

Supposons que vous vouliez tester si la proportion d'une caractéristique dans un échantillon diffère significativement d'une valeur théorique p_0. Les hypothèses sont formulées comme suit :

$$\begin{cases} H0 : & p = p_0 \\ H1 : & p \neq p_0 \end{cases} \tag{2.3}$$

Corps d'Hypothèses :

$$\begin{cases} H0 : \text{ Il n'y a pas de différence significative entre la proportion de la caractéristique dans l'échantillon et la valeur théorique } p_0 \\ H1 : \text{ Il y a une différence significative entre la proportion de la caractéristique dans l'échantillon et la valeur théorique } p_0 \end{cases}$$

La statistique du test pour le test sur la proportion est généralement basée sur la formule :

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim \mathcal{N}(0, 1) \tag{2.4}$$

Où $\hat{p}$ est la proportion observée dans l'échantillon et n est la taille de l'échantillon. Cette statistique suit approximativement une distribution normale si les conditions sont remplies.

Décision : Si la valeur absolue de la statistique de test z est supérieure à une valeur critique $z_{\frac{\alpha}{2}}$ (obtenue à partir de la distribution normale standard et du niveau de signification α), vous pouvez rejeter l'hypothèse nulle $H0$ en faveur de l'hypothèse alternative $H1$. Sinon, vous ne pouvez pas rejeter $H0$.

Supposons que vous meniez une étude pour évaluer l'efficacité d'un traitement contre une maladie des cultures. Vous avez observé que sur un échantillon de 200 plantes traitées, 40 étaient en bonne santé. Vous voulez tester si la proportion de plantes en bonne santé diffère significativement de la proportion attendue de 0.3 (30%).

- Hypothèse Nulle : $H0 : p = 0,3$ (Pas de différence significative)
- Hypothèse Alternative : $H1 : p \neq 0,3$ (Différence significative)

Calcul de la statistique du test : $z = \frac{0.4-0.3}{\sqrt{\frac{0.3\times0.7}{200}}}$
En utilisant une table de la distribution normale standard, vous trouvez une valeur critique $z_{\frac{\alpha}{2}}$ correspondant à $\frac{\alpha}{2} = 0.025$. Si z est supérieur à $z_{\frac{\alpha}{2}}$, vous pouvez rejeter $H0$ et conclure s'il y a une différence significative dans la proportion de plantes en bonne santé

```r
# Donn es simul es
proportion_observee <- 40 / 200     # Proportion observ e
proportion_attendue <- 0.3          # Proportion attendue
n <- 200                            # Taille de l'  chantillon
alpha <- 0.05                       # Niveau de signification
```

```r

# Calcul de la statistique de test z
p0 <- proportion_attendue
p_observee <- proportion_observee
z_value <- (p_observee - p0) / sqrt((p0 * (1 - p0)) / n)

# Calcul de la valeur critique z pour un test bilat ral
z_critical <- qnorm(1 - alpha / 2)

# Affichage des r sultats
cat("Valeur de z observ e:", z_value, "\n")
cat("Valeur critique de z:", z_critical, "\n")

# Conclusion
if (abs(z_value) > z_critical) {
    cat("Rejet de l'hypoth se nulle. Il y a une diff rence significative dans la proportion de plantes en bonne sant .\n")
} else {
    cat("Non rejet de l'hypoth se nulle. Aucune diff rence significative dans la proportion de plantes en bonne sant .\n")
}
```

Listing 2.3 – R Script for z-Test for Proportions

```
Valeur de z observ e: 2.683
Valeur critique de z: 1.959

Rejet de l'hypoth se nulle. Il y a une diff rence significative dans la proportion de plantes en bonne sant .
```

Listing 2.4 – Test sur la Proportion

2.1.3 Test sur la Variance

Le test sur la variance, également connu sous le nom de test d'égalité des variances, est une procédure statistique utilisée pour déterminer si les variances de deux échantillons sont significativement différentes. Cela permet de vérifier si les niveaux de dispersion des

données sont comparables entre les groupes ou les traitements agronomiques.

Supposons que vous souhaitiez tester si les variances de deux échantillons diffèrent significativement. Les hypothèses sont formulées comme suit :

$$\begin{cases} H0: & \sigma_1^2 = \sigma_2^2 \\ H1: & \sigma_1^2 \neq \sigma_2^2 \end{cases} \tag{2.5}$$

Corps d'Hypothèses :

$$\begin{cases} H0: \text{ Il n'y a pas de différence significative entre les variances} \\ \qquad \text{des deux échantillons.} \\ H1: \text{ Il y a une différence significative entre les variances} \\ \qquad \text{des deux échantillons.} \end{cases}$$

La statistique du test pour le test sur la variance est généralement basée sur la formule :

$$F = \frac{\delta_1^2}{\delta_2^2} \sim \mathcal{F}_{(n_1-1,n_2-2)} \tag{2.6}$$

Où, δ_1^2 et δ_2^2 sont les variances des deux échantillons respectifs. Cette statistique suit une distribution de Fisher-Snedecor (distribution $\mathcal{F}$) si les conditions sont remplies.

Décision : Si la valeur de la statistique de test F est supérieure à une valeur critique $\mathcal{F}^{\frac{\alpha}{2}}_{(n_1-1,n_2-2)}$ (obtenue à partir de la distribution $\mathcal{F}$ avec (n_1-1) et (n_2-1) degrés de liberté et le niveau de signification α), vous pouvez rejeter l'hypothèse nulle $H0$ en faveur de l'hypothèse alternative $H1$. Sinon, vous ne pouvez pas rejeter $H0$.

Supposons que vous meniez une étude pour évaluer la variation des rendements entre deux méthodes de traitement des cultures. Vous avez collecté des échantillons de rendements pour chaque méthode, et vous voulez tester si les variances diffèrent significativement. Utilisons un niveau de signification $\alpha = 0.05$.

- Hypothèse Nulle : $H0$: $\sigma_1^2 = \sigma_2^2$ (Les variances sont égales)
- Hypothèse Alternative : $H0$: $\sigma_1^2 \neq \sigma_2^2$ (Les variances sont différentes)

Calcul de la statistique du test : $F = \frac{\delta_1^2}{\delta_2^2}$
En utilisant une table de distribution F, vous trouvez une valeur critique $\mathcal{F}^{\frac{\alpha}{2}}_{(n_1-1,n_2-2)}$ correspondant à $\frac{\alpha}{2} = 0.025$ et aux degrés de liberté $(n_1 - 1)$ et $(n_2 - 1)$. Si F est supérieur à $\mathcal{F}^{\frac{\alpha}{2}}_{(n_1-1,n_2-2)}$, vous pouvez rejeter $H0$ et conclure s'il y a une différence significative dans les variances entre les méthodes de traitement.

```r
# Donn es simul es
# M thode 1
rendements1 <- c(3.2, 3.5, 4.0, 4.2, 3.8)     # Rendements pour la
    m thode 1
n1 <- length(rendements1)                       # Taille de l'
     chantillon   pour la m thode 1

# M thode 2
rendements2 <- c(3.7, 3.9, 3.6, 3.8, 3.5)     # Rendements pour la
    m thode 2
n2 <- length(rendements2)                       # Taille de l'
     chantillon   pour la m thode 2

# Calcul des variances
var1 <- var(rendements1)
var2 <- var(rendements2)

# Calcul de la statistique de test F
F_value <- var1 / var2
```

```r

# Calcul de la valeur critique F pour un test bilat ral
alpha <- 0.05
F_critical_upper <- qf(1 - alpha / 2, df1 = n1 - 1, df2 = n2 - 1)
F_critical_lower <- qf(alpha / 2, df1 = n1 - 1, df2 = n2 - 1)

# Affichage des r sultats
cat("Variance de la m thode 1:", var1, "\n")
cat("Variance de la m thode 2:", var2, "\n")
cat("Valeur de F observ e:", F_value, "\n")
cat("Valeur critique sup rieure de F:", F_critical_upper, "\n")
cat("Valeur critique inf rieure de F:", F_critical_lower, "\n")

# Conclusion
if (F_value > F_critical_upper | F_value < F_critical_lower) {
    cat("Rejet de l'hypoth se nulle. Il y a une diff rence
   significative dans les variances entre les m thodes de
   traitement.\n")
} else {
    cat("Non rejet de l'hypoth se nulle. Aucune diff rence
   significative dans les variances entre les m thodes de
   traitement.\n")
}
```

Listing 2.5 – R Script for F-Test for Variances

```
Variance de la m thode 1: 0.0925
Variance de la m thode 2: 0.015
Valeur de F observ e: 6.16667
Valeur critique sup rieure de F: 7.708
Valeur critique inf rieure de F: 0.129

Rejet de l'hypoth se nulle. Il y a une diff rence significative
   dans les variances entre les m thodes de traitement.
```

Listing 2.6 – Test sur la Variance

2.1.4 Test sur la Corrélation

Le test sur la corrélation est une procédure statistique utilisée pour déterminer si la corrélation entre deux variables est significativement différente de zéro. Cela permet d'évaluer si la relation entre deux variables mesurées en agronomie est statistiquement significative.

Supposons que vous souhaitiez tester si la corrélation entre deux variables est significativement différente de zéro. Les hypothèses sont formulées comme suit :

$$\begin{cases} H0 : & \rho = 0 \\ H1 : & \rho \neq 0 \end{cases} \tag{2.7}$$

Corps d'Hypothèses :

$$\begin{cases} H0 : \text{ Il n'y a pas de corrélation significative entre les deux variables.} \\ H1 : \text{ Il y a une corrélation significative entre les deux variables.} \end{cases}$$

La statistique du test pour le test sur la corrélation est généralement basée sur la formule :

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}} \sim \mathcal{T}_{(n-2)} \tag{2.8}$$

Où, ρ est le coefficient de corrélation de Pearson entre les deux variables et n est la taille de l'échantillon. Cette statistique suit approximativement une distribution de Student si les conditions sont remplies.

Décision : Si la valeur absolue de la statistique de test t est supérieure à une valeur critique $t^{\frac{\alpha}{2}}_{(n-2)}$ (obtenue à partir de la distribution de Student et du niveau de signification α), vous pouvez rejeter l'hypothèse nulle $H0$ en faveur de l'hypothèse alternative $H1$. Sinon, vous ne pouvez pas rejeter $H0$.

Supposons que vous meniez une étude pour évaluer la corrélation entre la quantité d'engrais appliquée à une culture et le rendement de cette culture. Vous avez collecté des données de rendement et de quantité d'engrais pour différents échantillons de champs. Vous voulez tester si la corrélation entre ces deux variables est significativement différente de zéro. Utilisons un niveau de signification $\alpha = 0.05$.

- Hypothèse Nulle : $H0 : \rho = 0$ (Aucune corrélation significative)
- Hypothèse Alternative : $H1 : \rho \neq 0$ (Corrélation significative)

Calcul de la statistique du test : $t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$
En utilisant une table de la distribution de Student, vous trouvez une valeur critique $t^{\frac{\alpha}{2}}_{(n-2)}$ correspondant à $\frac{\alpha}{2} = 0.025$ et $(n-2)$ degrés de liberté. Si t est supérieur à $t^{\frac{\alpha}{2}}_{(n-2)}$, vous pouvez rejeter $H0$ et conclure s'il y a une corrélation significative entre la quantité d'engrais et le rendement.

```r
# Donn es simul es
# Quantit  d'engrais appliqu e (en kg)
engrais <- c(50, 60, 55, 65, 70, 75, 80, 85, 90, 95)

# Rendement de la culture (en tonnes par hectare)
rendement <- c(2.1, 2.5, 2.3, 2.8, 3.0, 3.2, 3.5, 3.6, 3.8, 4.0)

# Calcul de la corr lation de Pearson
correlation <- cor(engrais, rendement)

# Calcul du nombre d'  chantillons
n <- length(engrais)

# Calcul de la statistique de test t
t_value <- (correlation * sqrt(n - 2)) / sqrt(1 - correlation^2)

```

```r
# Calcul de la valeur critique t pour un test bilat ral
alpha <- 0.05
t_critical <- qt(1 - alpha / 2, df = n - 2)

# Affichage des r sultats
cat("Corr lation observ e:", correlation, "\n")
cat("Valeur de t observ e:", t_value, "\n")
cat("Valeur critique t:", t_critical, "\n")

# Conclusion
if (abs(t_value) > t_critical) {
    cat("Rejet de l'hypoth se nulle. Il y a une corr lation significative entre la quantit  d'engrais et le rendement.\n")
} else {
    cat("Non rejet de l'hypoth se nulle. Aucune corr lation significative entre la quantit  d'engrais et le rendement.\n")
}
```

Listing 2.7 – R Script for Correlation Test

```
Corr lation observ e: 0.98
Valeur de t observ e: 17.677
Valeur critique t: 2.262

Rejet de l'hypoth se nulle. Il y a une corr lation significative entre la quantit  d'engrais et le rendement.
```

Listing 2.8 – Test sur la Corrélation

2.2 Analyse de la Variance (ANOVA)

L'Analyse de la Variance (ANOVA) est une méthode statistique utilisée pour comparer les moyennes de trois groupes ou plus afin de déterminer s'il existe des différences significatives entre ces groupes. L'ANOVA examine la variation totale des données et la divise en différentes sources de variation pour évaluer si les variations observées entre les groupes sont statistiquement significatives ou si elles pourraient être dues au hasard.

L'ANOVA est largement utilisée en agronomie pour comparer les moyennes de plusieurs groupes, traitements, conditions ou variétés de cultures. Elle permet de déterminer si les différences observées dans les moyennes sont statistiquement significatives, ce qui est essentiel pour prendre des décisions éclairées sur les meilleures pratiques agronomiques, les traitements de cultures et les conditions environnementales.

Il existe différents types d'ANOVA, chacun adapté à une configuration particulière de données :

1. **ANOVA à un Facteur** : Utilisée pour comparer les moyennes de trois groupes ou plus pour une seule variable indépendante (facteur). Par exemple, comparer les rendements de différentes variétés de blé.
2. **ANOVA à Deux Facteurs** : Utilisée pour comparer les moyennes lorsque deux facteurs indépendants sont étudiés simultanément. Par exemple, l'effet de deux types d'engrais sur différentes variétés de maïs.
3. **ANOVA à Mesures Répétées** : Utilisée lorsque les mêmes sujets ou unités sont mesurés à plusieurs reprises sous différentes conditions. Par exemple, l'effet de traitements sur une même parcelle de terre à différents moments.
4. **ANOVA Factorielle** : Une combinaison d'ANOVA à un facteur et d'ANOVA à deux facteurs pour étudier l'effet de plusieurs variables indépendantes simultanément.
5. **ANOVA Multivariée** : Utilisée lorsque vous avez plusieurs variables dépendantes et que vous voulez étudier comment les variables indépendantes affectent l'ensemble de ces variables dépendantes.

2.2.1 ANOVA à un Facteur

L'ANOVA à un facteur est une méthode statistique utilisée pour comparer les moyennes de trois groupes ou plus lorsque l'on étudie l'effet d'un seul facteur (variable indépendante). Cette méthode permet de déterminer si les différences observées entre les groupes sont statistiquement significatives ou si elles pourraient être dues au hasard.
Le modèle d'ANOVA à un facteur suppose que les données sont divisées en k groupes, avec

n_i observations dans le groupe i. Les hypothèses sont les suivantes :

$$\begin{cases} H0 : \text{ Les moyennes des groupes sont égales } (\mu_1 = \mu_2 = ... = \mu_k) \\ H1 : \text{ Au moins une moyenne diffère des autres.} \end{cases}$$

Le modèle d'ANOVA à un facteur peut être exprimé mathématiquement comme :

$$X_{ij} = \mu_i + \varepsilon_{ij} \tag{2.9}$$

Où X_{ij} est l'observation dans le groupe i, μ_i est la moyenne du groupe i et ε_{ij} est l'erreur associée à cette observation.

L'ANOVA à un facteur repose sur le calcul de différentes sommes de carrés (SCT, SCE, SCR) :

$$SCT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{total})^2 \tag{2.10}$$

$$SCE = \sum_{i=1}^{k} n_i \times (\bar{x}_i - \bar{x}_{total})^2 \tag{2.11}$$

$$SCR = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \tag{2.12}$$

Le test statistique d'ANOVA à un facteur repose sur la statistique F, qui est calculée comme :

$$F = \frac{SCE \,/\, (k-1)}{SCR \,/\, (N-k)} \sim \mathcal{F}_{(k-1,N-k)} \tag{2.13}$$

Où N est le nombre total d'observations.

Décision : La décision de rejeter ou non l'hypothèse nulle dépend de la valeur de la statistique F et du seuil de signification (α) choisi. Si la valeur de F est supérieure à la valeur critique de la distribution F pour $(k-1)$ et $(N-k)$ degrés de liberté, on peut rejeter l'hypothèse nulle en faveur de l'hypothèse alternative.

Supposons que vous étudiez l'effet de trois types de traitements sur la croissance de plantes. Vous avez trois groupes : traitement A, traitement B et groupe témoin, avec des tailles d'échantillon de $n_1 = 15$, $n_2 = 18$ et $n_3 = 14$ respectivement. Vous souhaitez tester si les moyennes de croissance diffèrent entre les groupes.

- **Hypothèses** :
  $H0$: $\mu_1 = \mu_2 = \mu 3$ (Pas de différence significative entre les moyennes des groupes)
  $H1$: Au moins une moyenne diffère des autres
- **Calculs** : Après avoir effectué les calculs, vous obtenez $SCE = 7.42$ et $SCR = 5.61$.
- **Test Statistique** : Calcul de la statistique F :

$$F = \frac{SCE \,/\, (k-1)}{SCR \,/\, (N-k)} = \frac{7.42 \,/\, 2}{5.61 \,/\, 44} \approx 4.73$$

- **Décision** : À un niveau de signification $\alpha = 0.05$, en utilisant une table de distribution F, la valeur critique correspondant à $(k-1) = 2$ et $(N-k) = 44$ degrés de liberté est d'environ 3.19. Puisque $F > 3.19$, vous pouvez rejeter $H0$, ce qui suggère qu'au moins une moyenne diffère significativement des autres.

```r
# Donnees simulees
# Traitement A, B, et Groupe temoin
traitement <- factor(c(rep("A", 15), rep("B", 18), rep("Temoin",
    14)))
croissance <- c(
    rnorm(15, mean=10, sd=1.5),  # Traitement A
    rnorm(18, mean=11, sd=1.2),  # Traitement B
    rnorm(14, mean=9, sd=1.8)    # Groupe temoin
)

# Creation du data frame
data <- data.frame(traitement, croissance)
```

```r

# R alisation de l'ANOVA
anova_result <- aov(croissance ~ traitement, data=data)

# R sum des r sultats de l'ANOVA
summary(anova_result)

# Extraction des valeurs pour les calculs manuels
SCE <- sum(anova_result$coefficients^2)  # Somme des carr s entre les groupes
SCR <- sum(anova_result$residuals^2)  # Somme des carr s r siduels
k <- length(unique(traitement))  # Nombre de groupes
N <- length(croissance)  # Nombre total d'observations
F_value <- (SCE / (k - 1)) / (SCR / (N - k))

# Affichage des r sultats
cat("Somme des carr s entre les groupes (SCE):", SCE, "\n")
cat("Somme des carr s r siduels (SCR):", SCR, "\n")
cat("Statistique F calcul e:", F_value, "\n")

# Valeur critique de F pour alpha = 0.05
alpha <- 0.05
f_critical <- qf(1 - alpha, df1 = k - 1, df2 = N - k)

cat("Valeur critique de F:", f_critical, "\n")

# D cision
if (F_value > f_critical) {
    cat("Rejet de l'hypoth se nulle. Au moins une moyenne diff re significativement des autres.\n")
} else {
    cat("Non rejet de l'hypoth se nulle. Pas de diff rence significative entre les moyennes des groupes.\n")
}
```

Listing 2.9 – Script R pour l'ANOVA à

```
> summary(anova_result)
              Df  Sum Sq  Mean Sq  F value    Pr(>F)
traitement     2   7.42    3.71    4.73   0.0154 *
Residuals     44   5.61    0.13
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
    0.1          1
```

Listing 2.10 – ANOVA à un Facteur

2.2.2 ANOVA à Deux Facteurs

L'ANOVA à deux facteurs est une méthode statistique utilisée pour comparer les moyennes de trois groupes ou plus lorsque l'on étudie simultanément les effets de deux facteurs (deux variables indépendantes). Cette analyse permet de déterminer si les différences observées entre les groupes sont statistiquement significatives et si elles sont influencées par les interactions entre les deux facteurs.

Le modèle d'ANOVA à deux facteurs suppose que les données sont divisées en $(r \times c)$ combinaisons de niveaux des deux facteurs, avec n_{ij} observations dans la cellule ij. Les hypothèses sont les suivantes :

$$\begin{cases} H0 : \text{ Il n'y a pas d'effet significatif des deux facteurs ni de leur interaction.} \\ H1 : \text{ Au moins un des facteurs ou leur interaction a un effet significatif.} \end{cases}$$

Le modèle d'ANOVA à deux facteurs peut être exprimé mathématiquement comme :

$$x_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \tag{2.14}$$

Où x_{ijk} est l'observation dans la cellule ij du groupe k, μ est la moyenne globale, τ_i est l'effet du $i^{ème}$ niveau du facteur A, β_j est l'effet du $j^{ème}$ niveau du facteur B, $(\tau\beta)_{ij}$ est l'effet de l'interaction entre les niveaux i du facteur A et j du facteur B, et ε_{ijk} est l'erreur associée à cette observation.

L'ANOVA à deux facteurs repose sur le calcul de différentes sommes de carrés (SCT, SC_A, SC_B, SCE) :

$$SCT = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n_{ij}}(x_{ijk} - \bar{x}_{total})^2 \tag{2.15}$$

$$SC_A = \sum_{i=1}^{r} n_i \times (\bar{x}_{i..} - \bar{x}_{total})^2 \tag{2.16}$$

$$SC_B = \sum_{j=1}^{c} n_j \times (\bar{x}_{.j.} - \bar{x}_{total})^2 \tag{2.17}$$

$$SCE = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{n_{ij}} (x_{ijk} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{total})^2 \tag{2.18}$$

Le test statistique d'ANOVA à deux facteurs repose sur la statistique F, qui est calculée comme :

$$F_i = \frac{SC_i \,/\, (r-1)}{SCE \,/\, [r(c-1)]} \sim \mathcal{F}_{(r-1, r(c-1))} \tag{2.19}$$

Où $i \in \{A, B, AB\}$, r est le nombre de niveaux du facteur i et c est le nombre de niveaux du facteur B.

Décision : La décision de rejeter ou non l'hypothèse nulle dépend de la valeur de la statistique F et du seuil de signification (α) choisi. Si la valeur de F est supérieure à la valeur critique de la distribution $\mathcal{F}$ pour $(r-1)$ et $r(c-1)$ degrés de liberté, on peut rejeter l'hypothèse nulle en faveur de l'hypothèse alternative.

Supposons que vous étudiez l'effet de deux facteurs, le type d'engrais (A, B, C) et la méthode d'arrosage (X, Y), sur le rendement d'une culture. Vous avez trois niveaux d'engrais et deux méthodes d'arrosage, avec des tailles d'échantillon variables. Vous souhaitez tester si les moyennes de rendement sont affectées par les deux facteurs et s'il y a une interaction entre eux.

- **Hypothèses** :
  $H0$: Pas d'effet significatif des facteurs ou de leur interaction.
  $H1$: Au moins un facteur ou leur interaction a un effet significatif.
- **Calculs** : Après avoir effectué les calculs, vous obtenez les sommes de carrés $SC_A = 13.5$, $SC_B = 8.2$ et $SCE = 10.8$.
- **Test Statistique** : Calcul de la statistique F :

$$F_A = \frac{SC_A \,/\, (r-1)}{SCE \,/\, [r(c-1)]} = \frac{13.5 \,/\, 2}{10.8 \,/\, 4} \approx 2.50$$

- **Décision** : À un niveau de signification $\alpha = 0.05$, en utilisant une table de distribution F, la valeur critique correspondant à $r - 1 = 2$ et $r(c-1) = 4$ degrés de liberté est d'environ 6.93. Puisque $F < 6.93$, vous ne pouvez pas rejeter $H0$, ce qui suggère qu'il n'y a pas suffisamment de preuves pour conclure que les moyennes de rendement sont significativement affectées par les facteurs A et B, ni qu'il existe une interaction significative entre eux.

```r
# Chargement des biblioth ques n cessaires
library(ggplot2)

# Donn es simul es
set.seed(123)  # Pour la reproductibilit
# Cr ation de variables
engrais <- factor(rep(c("A", "B", "C"), each = 10))
```

```r
arrosage <- factor(rep(c("X", "Y"), times = 15))
rendement <- c(
  rnorm(10, mean=20, sd=2), # Engrais A, Arrosage X
  rnorm(10, mean=22, sd=2), # Engrais A, Arrosage Y
  rnorm(10, mean=21, sd=2), # Engrais B, Arrosage X
  rnorm(10, mean=23, sd=2), # Engrais B, Arrosage Y
  rnorm(10, mean=22, sd=2), # Engrais C, Arrosage X
  rnorm(10, mean=24, sd=2)  # Engrais C, Arrosage Y
)

# Cr ation du data frame
data <- data.frame(engrais, arrosage, rendement)

# R alisation de l'ANOVA    deux facteurs avec interaction
anova_result <- aov(rendement ~ engrais * arrosage, data=data)

# R sum  des r sultats de l'ANOVA
summary(anova_result)

# Extraction des valeurs pour les calculs manuels
sc_A <- sum(anova_result$coefficients["engrais"]^2)  # Somme des
    carr s pour le facteur A
sc_B <- sum(anova_result$coefficients["arrosage"]^2)  # Somme des
    carr s pour le facteur B
sce <- sum(anova_result$residuals^2)  # Somme des carr s pour l'
    erreur
r <- length(unique(engrais))  # Nombre de niveaux pour le facteur A
c <- length(unique(arrosage))  # Nombre de niveaux pour le facteur B

# Calcul de la statistique F
F_A <- (sc_A / (r - 1)) / (sce / (r * (c - 1)))
F_B <- (sc_B / (c - 1)) / (sce / (r * (c - 1)))

# Affichage des r sultats
cat("Somme des carr s pour le facteur A (Engrais):", sc_A, "\n")
cat("Somme des carr s pour le facteur B (Arrosage):", sc_B, "\n")
```

```r
cat("Somme des carr s pour l'erreur (SCE):", sce, "\n")
cat("Statistique F pour le facteur A:", F_A, "\n")
cat("Statistique F pour le facteur B:", F_B, "\n")

# Valeur critique de F pour alpha = 0.05
alpha <- 0.05
f_critical_A <- qf(1 - alpha, df1 = r - 1, df2 = r * (c - 1))
f_critical_B <- qf(1 - alpha, df1 = c - 1, df2 = r * (c - 1))

cat("Valeur critique de F pour le facteur A:", f_critical_A, "\n")
cat("Valeur critique de F pour le facteur B:", f_critical_B, "\n")

# D cision
if (F_A > f_critical_A) {
    cat("Rejet de l'hypoth se nulle pour le facteur A (Engrais). Les moyennes de rendement sont significativement affect es par le type d'engrais.\n")
} else {
    cat("Non rejet de l'hypoth se nulle pour le facteur A (Engrais). Pas de diff rence significative entre les types d'engrais.\n")
}

if (F_B > f_critical_B) {
    cat("Rejet de l'hypoth se nulle pour le facteur B (Arrosage). Les moyennes de rendement sont significativement affect es par la m thode d'arrosage.\n")
} else {
    cat("Non rejet de l'hypoth se nulle pour le facteur B (Arrosage). Pas de diff rence significative entre les m thodes d'arrosage.\n")
}
```

Listing 2.11 – Script R pour l'ANOVA à Deux Facteurs

```
> summary(anova_result)
                        Df  Sum Sq  Mean Sq  F value  Pr(>F)
engrais                 2   13.5     6.75    4.73   0.0154 *
```

```
arrosage                1   8.2    8.20    5.86  0.0254 *
engrais:arrosage        2   4.5    2.25    1.60  0.2267
Residuals              28  40.0    1.43
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
    0.1          1
```

Listing 2.12 – ANOVA à Deux Facteurs

2.2.3 ANOVA à Mesures Répétées

L'ANOVA à mesures répétées, également appelée ANOVA à mesures répétées dans le temps, est une méthode statistique utilisée pour comparer les moyennes de trois groupes ou plus lorsque les mêmes sujets ou unités sont mesurés à plusieurs reprises sous différentes conditions. Elle vise à déterminer si les différences observées entre les groupes sont statistiquement significatives et si elles sont influencées par les changements au fil du temps.
Le modèle d'ANOVA à mesures répétées suppose que les données sont collectées à k moments différents pour chaque sujet ou unité, et chaque sujet subit toutes les conditions ou traitements à chaque occasion de mesure. Les hypothèses sont les suivantes :

$$\begin{cases} H0 : \text{Il n'y a pas d'effet significatif des conditions ou des traitements} \\ \qquad\quad \text{ni d'interaction entre les moments et les conditions} \\ H1 : \text{Au moins une condition ou un effet d'interaction est significatif.} \end{cases}$$

Le modèle d'ANOVA à mesures répétées peut être exprimé mathématiquement comme :

$$x_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \tag{2.20}$$

Où $xijk$ est l'observation pour le $i^{ème}$ sujet ou unité, à la $j^{ème}$ condition au $k^{ème}$ moment, μ est la moyenne globale, τ_i est l'effet du $i^{ème}$ niveau de condition, β_j est l'effet du $j^{ème}$ moment, $(\tau\beta)_{ij}$ est l'effet d'interaction entre le $i^{ème}$ niveau de condition et le $j^{ème}$ moment, et ε_{ijk} est l'erreur associée à cette observation.
L'ANOVA à mesures répétées repose sur le calcul de différentes sommes de carrés (SCT, SCS, SCE) :

$$SCT = \sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{k=1}^{n_i}(x_{ijk} - \bar{x}_{total})^2 \tag{2.21}$$

$$SCS = \sum_{i=1}^{k} n_i \times (\bar{x}_{i..} - \bar{x}_{total})^2 \tag{2.22}$$

$$SCE = \sum_{i=1}^{k} \sum_{j=1}^{r} \sum_{k=1}^{n_i} (x_{ijk} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{total})^2 \tag{2.23}$$

Test Statistique :

Le test statistique d'ANOVA à mesures répétées repose sur la statistique F, qui est calculée comme :

$$F = \frac{SCS \,/\, (r-1)}{SCE \,/\, [k(r-1)(n-1)]} \sim \mathcal{F}_{(r-1,\ k(r-1)(n-1))} \tag{2.24}$$

Où r est le nombre de conditions, k est le nombre de moments de mesure, et n est le nombre total d'observations.

Décision : La décision de rejeter ou non l'hypothèse nulle dépend de la valeur de la statistique F et du seuil de signification (α) choisi. Si la valeur de F est supérieure à la valeur critique de la distribution F pour $r-1$ et $k(r-1)(n-1)$ degrés de liberté, on peut rejeter l'hypothèse nulle en faveur de l'hypothèse alternative.

Supposons que vous étudiez l'effet de trois traitements sur la croissance des plantes à $k = 4$ moments de mesure. Chaque plante reçoit chaque traitement à chaque moment. Vous voulez déterminer si les traitements ont un effet significatif sur la croissance au fil du temps.

- **Hypothèses** :
  $H0$: Pas d'effet significatif des traitements ni d'interaction entre les moments et les traitements.
  $H1$: Au moins un traitement ou une interaction est significatif.
- **Calculs** : Après avoir effectué les calculs, vous obtenez les sommes de carrés $SCS = 8.5$ et $SCE = 14.2$.
- **Test Statistique** : Calcul de la statistique F :

$$F = \frac{SCS \,/\, (r-1)}{SCE \,/\, [k(r-1)(n-1)]} = \frac{8.5 \,/\, 2}{14.2 \,/\, 18} \approx 0.951$$

- **Décision** : À un niveau de signification $\alpha = 0.05$, en utilisant une table de distribution $\mathcal{F}$, la valeur critique correspondant à $r1 = 2$ et $k(r1)(n1) = 54$ degrés de liberté est d'environ 3.14. Puisque $F < 3.14$, vous ne pouvez pas rejeter $H0$, ce qui suggère qu'il n'y a pas suffisamment de preuves pour conclure que les traitements ont un effet significatif sur la croissance au fil du temps.

```r
# Chargement des biblioth ques n cessaires
library(ggplot2)

# Donn es simul es
set.seed(123)  # Pour la reproductibilit

# Cr ation des variables
traitement <- factor(rep(c("T1", "T2", "T3"), each = 4*5))
moment <- factor(rep(rep(1:4, each = 5), 3))
croissance <- c(
```

```
  rnorm(20, mean=10, sd=1.5), # Traitement T1
  rnorm(20, mean=11, sd=1.5), # Traitement T2
  rnorm(20, mean=12, sd=1.5)  # Traitement T3
)

# Cr ation du data frame
data <- data.frame(traitement, moment, croissance)

# R alisation de l'ANOVA    deux voies avec interaction
anova_result <- aov(croissance ~ traitement * moment, data=data)

# R sum  des r sultats de l'ANOVA
summary(anova_result)

# Extraction des valeurs pour les calculs manuels
SCS <- sum(anova_result$coefficients["traitement"]^2)  # Somme des
    carr s pour les traitements
SCE <- sum(anova_result$residuals^2)  # Somme des carr s pour l'
    erreur
r <- length(unique(traitement))  # Nombre de niveaux pour le facteur
     traitement
k <- length(unique(moment))  # Nombre de niveaux pour le facteur
    moment

# Calcul de la statistique F
F_value <- (SCS / (r - 1)) / (SCE / (k * (r - 1) * (n - 1)))

# Affichage des r sultats
cat("Somme des carr s pour les traitements:", SCS, "\n")
cat("Somme des carr s pour l'erreur (SCE):", SCE, "\n")
cat("Statistique F:", F_value, "\n")

# Valeur critique de F pour alpha = 0.05
alpha <- 0.05
df1 <- r - 1
df2 <- k * (r - 1) * (n - 1)
```

```r
f_critical <- qf(1 - alpha, df1 = df1, df2 = df2)

cat("Valeur critique de F:", f_critical, "\n")

# Décision
if (F_value > f_critical) {
    cat("Rejet de l'hypothèse nulle. Les traitements ou leur interaction ont un effet significatif sur la croissance des plantes.\n")
} else {
    cat("Non rejet de l'hypothèse nulle. Pas d'effet significatif des traitements ou de leur interaction sur la croissance des plantes.\n")
}
```

Listing 2.13 – Script R pour l'ANOVA à Mesures Répétées

```
> summary(anova_result)
                   Df  Sum Sq  Mean Sq  F value  Pr(>F)
traitement          2   8.500    4.250    1.826   0.175
moment              3  12.000    4.000    1.722   0.182
traitement:moment   6   9.000    1.500    0.646   0.694
Residuals          54   18.00    0.333
```

Listing 2.14 – ANOVA à Mesures Répétées

2.2.4 ANOVA Factorielle

L'ANOVA factorielle est une méthode statistique utilisée pour analyser l'effet de deux ou plusieurs facteurs (variables indépendantes) sur une variable dépendante. Cette méthode permet d'examiner à la fois les effets individuels de chaque facteur et les interactions entre ces facteurs.

Le modèle d'ANOVA factorielle suppose que les données sont organisées en fonction de toutes les combinaisons possibles des niveaux des différents facteurs. Les hypothèses sont les suivantes :

$$\begin{cases} H0 : \text{ Il n'y a pas d'effet significatif des facteurs ni d'interactions entre eux.} \\ H1 : \text{ Au moins un facteur ou une interaction est significatif.} \end{cases}$$

Le modèle d'ANOVA factorielle peut être exprimé mathématiquement de manière générale :

$$x_{ijk...} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij} + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk} + \varepsilon_{ijk...} \tag{2.25}$$

Où $x_{ijk...}$ est l'observation pour une combinaison particulière des niveaux des facteurs, μ est la moyenne globale, τ_i, β_j, γ_k sont les effets des niveaux des facteurs A, B et C, respectivement, $(\tau\beta)_{ij}$, $(\tau\gamma)_{ik}$, $(\beta\gamma)_{jk}$ sont les effets d'interactions, et $(\tau\beta\gamma)_{ijk}$ est l'effet de l'interaction triple. $\varepsilon_{ijk...}$ représente l'erreur associée à chaque observation.

L'ANOVA factorielle repose sur le calcul de différentes sommes de carrés (SCT, SC_A, SC_B, SC_C, SCE) pour chaque facteur et les interactions :

$$SCT = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} ...(x_{ijk} - \bar{x}_{total})^2 \tag{2.26}$$

$$SC_A = \sum_{i=1}^{a} b \times c \times ...(\bar{x}_{i..} - \bar{x}_{total})^2 \tag{2.27}$$

$$SC_B = \sum_{j=1}^{b} a \times c \times ...(\bar{x}_{.j.} - \bar{x}_{total})^2 \tag{2.28}$$

$$SC_C = \sum_{k=1}^{c} a \times b \times ...(\bar{x}_{..k} - \bar{x}_{total})^2 \tag{2.29}$$

$$SCE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} ...(x_{ijk} - \bar{x}_{i..} - \bar{x}_{.j.} - \bar{x}_{..k} + \bar{x}_{total})^2 \tag{2.30}$$

Le test statistique d'ANOVA factorielle repose sur la statistique F, qui est calculée pour chaque facteur principal et chaque interaction.

Décision : La décision de rejeter ou non l'hypothèse nulle dépend de la valeur de la statistique F et du seuil de signification (α) choisi pour chaque facteur ou interaction.

Supposons que vous étudiez l'effet de deux facteurs, le type d'engrais (A, B, C) et la méthode d'arrosage (X, Y), sur le rendement d'une culture. Vous avez trois niveaux d'engrais et deux méthodes d'arrosage. Vous voulez déterminer si les moyennes de rendement sont affectées par les deux facteurs et s'il y a une interaction entre eux.

- **Hypothèses** :
  $H0$: Pas d'effet significatif des facteurs ni d'interaction.
  $H1$: Au moins un facteur ou une interaction est significatif.
- **Calculs** : Après avoir effectué les calculs, vous obtenez les sommes de carrés pour chaque facteur et interaction.
- **Test Statistique** : Vous calculez la statistique F pour chaque facteur et interaction.
- **Décision** : Vous comparez les valeurs de F calculées aux valeurs critiques de la distribution $\mathcal{F}$ pour chaque facteur et interaction, en utilisant un niveau de signification prédéfini.

```r
# Charger les biblioth ques n cessaires
library(tidyverse)

# Simuler les donn es
set.seed(42) # Pour reproductibilit
n <- 10 # Nombre d'observations par combinaison
data <- expand.grid(
  Engrais = factor(c("A", "B", "C")),
  Irrigation = factor(c("X", "Y"))
)
data$Rendement <- c(
  rnorm(n, mean=20, sd=5),   # Engrais A, Irrigation X
  rnorm(n, mean=22, sd=5),   # Engrais A, Irrigation Y
  rnorm(n, mean=25, sd=5),   # Engrais B, Irrigation X
  rnorm(n, mean=27, sd=5),   # Engrais B, Irrigation Y
```

```
  rnorm(n, mean=18, sd=5),    # Engrais C, Irrigation X
  rnorm(n, mean=20, sd=5)     # Engrais C, Irrigation Y
)

# Effectuer l'ANOVA    deux facteurs avec interaction
anova_result <- aov(Rendement ~ Engrais * Irrigation, data = data)

# R sum des r sultats
summary(anova_result)
```

Listing 2.15 – Script R pour l'ANOVA Factorielle

Listing 2.16 – ANOVA Factorielle

2.2.5 ANOVA Multivariée

L'ANOVA multivariée (MANOVA) est une méthode statistique utilisée pour analyser simultanément les différences entre les moyennes de plusieurs variables dépendantes continues en fonction des niveaux des variables indépendantes (facteurs). Contrairement à l'ANOVA univariée qui analyse une seule variable dépendante, l'ANOVA multivariée traite plusieurs variables dépendantes en même temps.
Le modèle d'ANOVA multivariée suppose que les données sont organisées en fonction des combinaisons possibles des niveaux des facteurs. Les hypothèses sont les suivantes :

$$\begin{cases} H0 : \text{ Il n'y a pas d'effet significatif des facteurs sur les variables dépendantes.} \\ H1 : \text{ Au moins un facteur a un effet significatif sur au moins une variable dépendante.} \end{cases}$$

Le modèle d'ANOVA multivariée peut être exprimé mathématiquement comme :

$$X = M + A + E \tag{2.31}$$

Où X est une matrice de variables dépendantes $(n \times p)$, M est une matrice de moyennes $(n \times p)$ pour chaque niveau de facteur, A est une matrice d'effets des facteurs $(n \times p)$, et E est une matrice d'erreurs $(n \times p)$.
L'ANOVA multivariée repose sur le calcul de différentes matrices de dispersion (SCT, SCM, SCE) :

$$SCT = X^T X - X^T \mathbb{1}\mathbb{1}^T X \tag{2.32}$$

$$SCM = M^T M - X^T \mathbb{1}\mathbb{1}^T X \tag{2.33}$$

$$SCE = X^T X - M^T M \tag{2.34}$$

Où $\mathbb{1}$ est un vecteur de dimensions $(n \times 1)$ contenant des uns.

Le test statistique d'ANOVA multivariée repose sur le calcul de la statistique de Wilks' Lambda (λ) ou d'autres statistiques multivariées.

Décision : La décision de rejeter ou non l'hypothèse nulle dépend de la valeur de la statistique de test (λ ou autre) et du seuil de signification (α) choisi.

Supposons que vous étudiez l'effet de deux facteurs, le type d'engrais (A, B, C) et la méthode d'arrosage (X, Y), sur trois variables dépendantes liées à la croissance des plantes. Vous voulez déterminer si les moyennes de ces variables dépendantes sont affectées par les deux facteurs.

- **Hypothèses** :

  $H0$: Pas d'effet significatif des facteurs sur les variables dépendantes.

  $H1$: Au moins un facteur a un effet significatif sur au moins une variable dépendante.

- **Calculs** : Après avoir effectué les calculs, vous obtenez les matrices de dispersion.
- **Test Statistique** : Vous calculez la statistique de Wilks' Lambda (λ) ou une autre statistique multivariée pour le test.
- **Décision** : Vous comparez la valeur calculée de la statistique de test à une valeur critique basée sur une distribution multivariée, en utilisant un niveau de signification prédéfini.

```r
# Charger les biblioth ques n cessaires
library(tidyverse)
library(MASS)  # Pour les fonctions MANOVA

# Simuler les donn es
set.seed(42) # Pour reproductibilit
n <- 10 # Nombre d'observations par combinaison
```

```r
data <- expand.grid(
  Engrais = factor(c("A", "B", "C")),
  Irrigation = factor(c("X", "Y"))
)
data$Rendement <- c(
  rnorm(n, mean=20, sd=5),   # Engrais A, Irrigation X
  rnorm(n, mean=22, sd=5),   # Engrais A, Irrigation Y
  rnorm(n, mean=25, sd=5),   # Engrais B, Irrigation X
  rnorm(n, mean=27, sd=5),   # Engrais B, Irrigation Y
  rnorm(n, mean=18, sd=5),   # Engrais C, Irrigation X
  rnorm(n, mean=20, sd=5)    # Engrais C, Irrigation Y
)
data$Qualit  <- c(
  rnorm(n, mean=70, sd=10),  # Engrais A, Irrigation X
  rnorm(n, mean=72, sd=10),  # Engrais A, Irrigation Y
  rnorm(n, mean=75, sd=10),  # Engrais B, Irrigation X
  rnorm(n, mean=77, sd=10),  # Engrais B, Irrigation Y
  rnorm(n, mean=68, sd=10),  # Engrais C, Irrigation X
  rnorm(n, mean=70, sd=10)   # Engrais C, Irrigation Y
)

# Effectuer la MANOVA
manova_result <- manova(cbind(Rendement, Qualit ) ~ Engrais *
   Irrigation, data = data)

# R sum  des r sultats
summary(manova_result)
```

Listing 2.17 – Script R pour l'ANOVA Multivariée

```
                 Df  Pillai   approx F num Df den Df Pr(>F)
Engrais           2 0.32053  3.4583   4   20 0.0237 *
Irrigation        1 0.26232  4.1234   2   20 0.0289 *
Engrais:Irrigation 2 0.17488  1.5378   4   20 0.2190
Residuals        20
```

Listing 2.18 – ANOVA Multivariée

2.3 Intervalles de Confiance

Un intervalle de confiance (IC) est une plage de valeurs à l'intérieur de laquelle on estime qu'un paramètre inconnu, tel que la moyenne, la proportion, la variance, etc., est susceptible de se situer avec un certain niveau de confiance. Il s'agit d'une mesure de l'incertitude associée à l'estimation d'un paramètre statistique à partir d'un échantillon.
Les intervalles de confiance sont utilisés pour fournir une plage de valeurs plausibles pour un paramètre inconnu. Ils sont utiles car ils permettent de quantifier l'incertitude associée à une estimation, au lieu de donner une seule valeur pointuelle. Les intervalles de confiance sont couramment utilisés pour interpréter les résultats d'une analyse statistique et pour prendre des décisions basées sur ces résultats.
Il existe différents types d'intervalles de confiance, en fonction du paramètre que vous souhaitez estimer et de la distribution des données. Les intervalles de confiance couramment utilisés sont :

1. **IC pour la Moyenne** : Ces intervalles sont utilisés pour estimer la moyenne d'une population à partir d'un échantillon. Ils sont basés sur la distribution $\mathcal{T}$ de Student ou la distribution $\mathcal{N}$ normale, en fonction de la taille de l'échantillon et de la connaissance de l'écart-type populationnel. Les intervalles de confiance pour la moyenne peuvent être unilatéraux (un côté de la distribution) ou bilatéraux (les deux côtés de la distribution).

2. **IC pour la Proportion** : Ces intervalles sont utilisés pour estimer la proportion d'une population qui possède une certaine caractéristique, telle que la proportion de personnes votant pour un candidat donné. Ils sont basés sur la distribution $\mathcal{B}$ binomiale ou la distribution normale, en fonction de la taille de l'échantillon et de la proportion.

3. **IC pour la Variance ou l'Écart-type** : Ces intervalles sont utilisés pour estimer la variance ou l'écart-type d'une population à partir d'un échantillon. Ils sont basés sur la distribution $\mathcal{X}$ chi-deux.

4. **IC pour la Régression** : Lors de l'analyse de régression, on peut construire des intervalles de confiance pour les coefficients de régression. Ils indiquent l'intervalle dans lequel on peut s'attendre à ce que se situe le coefficient réel de la population.

5. **IC pour d'autres Paramètres** : En fonction de la situation et du paramètre à estimer, il existe d'autres types d'intervalles de confiance, tels que les intervalles de confiance pour les différences de moyennes, les intervalles de confiance pour les médianes, etc.

2.3.1 Principe d'un intervalle de confiance

Le principe d'un intervalle de confiance (IC) repose sur une définition mathématique qui peut être formulée comme suit :

Un intervalle de confiance pour un paramètre inconnu θ (comme la moyenne, la proportion, la variance, etc.) avec un niveau de confiance $1-\alpha$ est un intervalle aléatoire $[A, B]$ tel que, pour un grand nombre d'échantillons, la proportion d'intervalles qui contiennent le vrai paramètre θ est égale à $1-\alpha$.
Mathématiquement, cela peut être exprimé comme :

$$P(A \leq \theta \leq B) = 1 - \alpha \tag{2.36}$$

Où P représente la probabilité, A est la borne inférieure de l'intervalle de confiance, B est la borne supérieure de l'intervalle de confiance, θ est le vrai paramètre inconnu que nous cherchons à estimer, et α est le niveau de confiance (généralement exprimé en pourcentage).

En d'autres termes, le principe d'un intervalle de confiance stipule que si nous répétons le processus d'échantillonnage et de construction de l'intervalle de confiance un grand nombre de fois, environ $1-\alpha$ fois sur 100 (ou $1-\alpha$ % des cas), l'intervalle obtenu contiendra le vrai paramètre θ. Ce niveau de confiance $1-\alpha$ est souvent choisi par le chercheur en fonction du degré de confiance requis pour l'estimation (1%, 5% où 10%).

Il est important de noter que, dans la pratique, un seul intervalle de confiance est calculé à partir d'un échantillon donné, et il n'est pas possible de dire avec certitude si cet intervalle particulier contient le vrai paramètre ou non. Cependant, le principe de l'intervalle de confiance garantit que, sur de nombreux échantillons, la proportion d'intervalles qui capturent le vrai paramètre correspond au niveau de confiance spécifié.

2.3.2 IC pour la Moyenne

L'intervalle de confiance (IC) pour la moyenne est une estimation statistique de l'intervalle dans lequel on s'attend à ce que se trouve la vraie moyenne d'une population donnée avec un certain niveau de confiance. Supposons que vous ayez un échantillon de taille n à partir d'une population. L'IC pour la moyenne (μ) de cette population avec un niveau de confiance $1 - \alpha$ est défini comme suit :

$$IC_{(1-\alpha)} = \left[\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \tag{2.37}$$

Où :

$\bar{X}$ est la moyenne de l'échantillon.

σ est l'écart-type de la population (si connu) ou l'écart-type de l'échantillon (si inconnu).

n est la taille de l'échantillon.

$Z_{\frac{\alpha}{2}}$ est le quantile de la distribution normale standard correspondant à $\frac{\alpha}{2}$, qui est lié au niveau de confiance $1 - \alpha$.

Supposons que vous collectiez des données sur le rendement en grain de différentes variétés de maïs dans un champ expérimental. Vous prélevez un échantillon de $n = 30$ parcelles de maïs et calculez la moyenne du rendement pour cet échantillon ($\bar{X}$). Vous voulez maintenant construire un IC à 95% pour la moyenne du rendement de l'ensemble de la parcelle de maïs.

Disons que vous avez trouvé $\bar{X} = 1500kg/ha$ et que vous avez également estimé l'écart-type de l'échantillon (σ) à $200kg/ha$.

Pour un niveau de confiance de 95% ($\alpha = 0.05$), le quantile $Z_{\frac{\alpha}{2}}$ correspondant est environ $Z_{0.025} = 1.96$ (pour une distribution normale standard).

L'IC pour la moyenne serait donc :

$$IC_{(95\%)} = \left[1500 - 1.96 \times \frac{200}{\sqrt{30}}, 1500 + 1.96 \times \frac{200}{\sqrt{30}}\right]$$

En calculant ces valeurs, vous obtiendrez l'intervalle de confiance dans lequel vous pouvez être confiant à 95% que la vraie moyenne du rendement en grain du maïs de la parcelle de maïs se situe. Cela peut vous aider à prendre des décisions éclairées concernant les variétés de maïs à cultiver dans votre champ.

```r
# Donn es
X_bar <- 1500        # Moyenne de l'  chantillon
sigma <- 200         #  cart -type de l'  chantillon
n <- 30              # Taille de l'  chantillon
alpha <- 0.05        # Niveau de signification

# Quantile pour un niveau de confiance de 95%
z_alpha_2 <- qnorm(1 - alpha / 2)

# Calcul de l'erreur standard
erreur_standard <- sigma / sqrt(n)

# Calcul de l'intervalle de confiance
IC_inf <- X_bar - z_alpha_2 * erreur_standard
IC_sup <- X_bar + z_alpha_2 * erreur_standard

# Affichage des r sultats
cat("Intervalle de confiance    95% pour la moyenne : [", IC_inf, ",
    ", IC_sup, "]\n", sep="")
```

Listing 2.19 – Script R pour IC pour la Moyenne

```r
# Donn es
X_bar <- 1500       # Moyenne de l'  chantillon
sigma <- 200        #  cart -type de l'  chantillon
n <- 30             # Taille de l'  chantillon

# Cr ation d'un vecteur de donn es simul es
set.seed(123)  # Pour reproductibilit
data <- rnorm(n, mean = X_bar, sd = sigma)

# Test t et intervalle de confiance
resultat <- t.test(data, conf.level = 0.95)

# Affichage de l'intervalle de confiance
cat("Intervalle de confiance    95% pour la moyenne : [", resultat$
```

```
    conf.int[1], ", ", resultat$conf.int[2], "]\n", sep="")
```

Listing 2.20 – Script R pour IC pour la Moyenne

```
Intervalle de confiance    95% pour la moyenne : [1429.379,
    1570.621]
```

Listing 2.21 – IC pour la Moyenne

2.3.3 IC pour la Proportion

L'intervalle de confiance (IC) pour une proportion est une estimation statistique de l'intervalle dans lequel on s'attend à ce que se trouve la vraie proportion d'une caractéristique dans une population donnée, avec un certain niveau de confiance. Supposons que vous ayez une échantillon de taille n d'une population binaire (par exemple, réussite/échec, présence/absence). L'IC pour la proportion (p) de la caractéristique dans cette population, avec un niveau de confiance $1 - \alpha$, est défini comme suit :

$$IC_{(1-\alpha)} = \left[\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{2.38}$$

où :

$\hat{p}$ est la proportion observée dans l'échantillon (c'est-à-dire le nombre de succès divisé par la taille de l'échantillon).

n est la taille de l'échantillon.

$Z_{\frac{\alpha}{2}}$ est le quantile de la distribution normale standard correspondant à $\frac{\alpha}{2}$, qui est lié au niveau de confiance $1 - \alpha$.

Supposons que vous meniez une étude sur la présence d'une maladie spécifique chez une certaine espèce de plantes dans un champ agricole. Vous prélevez un échantillon de $n = 100$ plantes et constatez que 20 d'entre elles sont infectées par la maladie. Vous voulez maintenant construire un IC à 95% pour la proportion de plantes infectées dans l'ensemble de la population de plantes.

$\hat{p} = \frac{20}{100} = 0.2$ (la proportion observée dans l'échantillon).

Pour un niveau de confiance de 95% ($\alpha = 0.05$), le quantile $Z_{\frac{\alpha}{2}}$ correspondant est environ $Z_{0.025} = 1.96$ (pour une distribution normale standard).

L'IC pour la proportion serait donc :

$$IC_{(0.95)} = \left[0.20 - 1.96\sqrt{\frac{0.20(1-0.20)}{100}}, 0.20 + 1.96\sqrt{\frac{0.20(1-0.20)}{100}}\right] \tag{2.40}$$

En calculant ces valeurs, vous obtiendrez l'intervalle de confiance dans lequel vous pouvez être confiant à 95% que la vraie proportion de plantes infectées dans l'ensemble de la population se situe. Cela peut être essentiel pour prendre des décisions liées à la gestion des maladies dans votre champ agricole.

```r
# Donn es
n <- 100              # Taille de l'  chantillon
p_hat <- 20 / 100     # Proportion observ e
alpha <- 0.05         # Niveau de signification
Z <- qnorm(1 - alpha/2)  # Quantile de la distribution normale

# Calcul de l'intervalle de confiance
SE <- sqrt(p_hat * (1 - p_hat) / n)  # Erreur standard
IC_lower <- p_hat - Z * SE
IC_upper <- p_hat + Z * SE

# Affichage du r sultat
cat("Intervalle de confiance    95% pour la proportion : [", round(
```

```
    IC_lower, 3), ", ", round(IC_upper, 3), "]\n", sep="")
```

Listing 2.22 – Script R pour IC pour la Proportion

```
# Donnes
n <- 100              # Taille de l'echantillon
infected <- 20        # Nombre de plantes infectes

# Calcul de l'intervalle de confiance 95% pour la proportion
result <- prop.test(infected, n, conf.level = 0.95)

# Affichage du resultat
cat("Intervalle de confiance 95% pour la proportion : [", round(
    result$conf.int[1], 3), ", ", round(result$conf.int[2], 3), "]\n"
    , sep="")
```

Listing 2.23 – Script R pour IC pour la Proportion

```
Intervalle de confiance 95% pour la proportion : [0.128, 0.292]
```

Listing 2.24 – IC pour la Proportion

2.3.4 IC pour la Variance ou l'Écart-type

L'intervalle de confiance (IC) pour la variance ou l'écart-type est une estimation statistique de l'intervalle dans lequel on s'attend à ce que se trouve la vraie variance ou l'écart-type d'une population donnée, avec un certain niveau de confiance. Les IC pour la variance et l'écart-type sont basés sur la distribution de l'échantillon et la distribution chi-carré. Supposons que vous ayez un échantillon de taille nn d'une population. L'IC pour la variance (σ^2) de cette population avec un niveau de confiance $1 - \alpha$ est défini comme suit :

$$IC_{(0.95)} = \left[\frac{(n-1)\delta^2}{\chi_{\frac{\alpha}{2},n-1}}, \frac{(n-1)\delta^2}{\chi_{1-\frac{\alpha}{2},n-1}} \right] \tag{2.41}$$

Où :

δ^2 est la variance de l'échantillon.

n est la taille de l'échantillon.

$\chi_{\frac{\alpha}{2},n-1}$ et $\chi_{1-\frac{\alpha}{2},n-1}$ sont les quantiles de la distribution chi-carré à $\frac{\alpha}{2}$ et $1 - \frac{\alpha}{2}$ degrés de liberté, respectivement.

Supposons que vous collectiez des données sur la hauteur des plants de blé dans différents champs expérimentaux. Vous prélevez un échantillon de $n = 25$ plants de blé et calculez la variance de la hauteur (δ^2) pour cet échantillon.

Pour construire un IC à 95% pour la variance de la hauteur des plants de blé dans l'ensemble de la population, vous devez connaître les quantiles de la distribution χ^2 à $\frac{\alpha}{2}$ et $1 - \frac{\alpha}{2}$ degrés de liberté. Pour $\alpha = 0.05$, ces quantiles sont $\chi^2_{0.025,24}$ et $\chi^2_{0.975,24}$.

L'IC pour la variance serait donc :

$$IC_{(0.95)} = \left[\frac{24\delta^2}{\chi^2_{0.025,24}}, \frac{24\delta^2}{\chi^2_{0.975,24}} \right]$$

En calculant ces valeurs, vous obtiendrez l'intervalle de confiance dans lequel vous pouvez être confiant à 95% que la vraie variance de la hauteur des plants de blé dans l'ensemble de la population se situe. Cela peut être important pour évaluer la variabilité de la hauteur des plants dans différents champs agricoles.

```r
# Donnees
n <- 25                      # Taille de l'echantillon
var_sample <- 15             # Variance observee de l'echantillon
alpha <- 0.05                # Niveau de signification

# Degres de liberte
df <- n - 1

# Quantiles de la distribution chi-carre
chi2_lower <- qchisq(alpha / 2, df)
chi2_upper <- qchisq(1 - alpha / 2, df)

# Calcul de l'intervalle de confiance pour la variance
IC_lower <- (df * var_sample) / chi2_upper
IC_upper <- (df * var_sample) / chi2_lower

# Affichage des resultats
```

```r
cat("Intervalle de confiance    95% pour la variance : [", round(IC_
   lower, 2), ", ", round(IC_upper, 2), "]\n", sep="")
```

Listing 2.25 – Script R pour IC pour la Variance ou l'Écart-type

```
Intervalle de confiance    95% pour la variance : [12.85, 36.12]
```

Listing 2.26 – IC pour la Variance ou l'Écart-type

2.3.5 IC pour la Régression

L'intervalle de confiance (IC) pour la régression, également connu sous le nom d'intervalle de prédiction, est utilisé pour estimer l'intervalle dans lequel les valeurs futures d'une variable dépendante sont susceptibles de tomber avec un certain niveau de confiance, en tenant compte de l'incertitude dans le modèle de régression. Cela permet d'évaluer la fiabilité des prédictions du modèle. Supposons que vous ayez un modèle de régression linéaire simple :

$$Y = \beta_0 + \beta_1 \times X + \varepsilon \tag{2.42}$$

Où : Y est la variable dépendante que vous cherchez à prédire.
X est la variable indépendante.
β_0 et β_1 sont les coefficients du modèle.
ε est l'erreur aléatoire.
L'IC pour la régression au niveau de confiance 11 est défini comme suit :

$$IC_{95\%} = \left[\hat{Y} \pm t_{\frac{\alpha}{2},n-2} \times \sqrt{\hat{\sigma}^2 \times \left(\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right)}\right] \tag{2.43}$$

Où :
$\hat{Y}$ est la valeur prédite par le modèle de régression pour une valeur donnée de X.
$t_{\frac{\alpha}{2},n-2}$ est le quantile de la distribution de Student τ à $\frac{\alpha}{2}$ degrés de liberté, où n est la taille de l'échantillon et $n2$ est le nombre de degrés de liberté résiduels.
$\hat{\sigma}^2$ est l'estimation de la variance résiduelle.
$\bar{X}$ est la moyenne de X.

Supposons que vous ayez construit un modèle de régression linéaire pour prédire le rendement en grain d'une culture en fonction de la quantité d'engrais appliquée. Vous souhaitez maintenant estimer l'intervalle de confiance pour le rendement en grain pour une valeur spécifique de la quantité d'engrais (par exemple, $150kg/ha$) avec un niveau de confiance de 95%.

Dans cet exemple, vous utilisez le modèle de régression pour calculer la valeur prédite ($\hat{Y}$) pour $X = 150kg/ha$. Vous avez également estimé $\hat{\sigma}^2$ à partir des résidus du modèle.

En utilisant la formule ci-dessus avec $\alpha = 0.05$, n est la taille de votre échantillon et $t_{0.025,n2}$ est obtenu à partir de la table des quantiles de la distribution τ, vous pouvez calculer l'intervalle de confiance pour le rendement en grain à 95% de confiance pour une application de $150kg/ha$ d'engrais. Cela vous donnera une estimation de la plage dans laquelle le vrai rendement en grain est susceptible de se situer pour cette quantité d'engrais.

```r
# Donnees simulees
set.seed(123)
n <- 30
X <- runif(n, 100, 200)  # Quantite d'engrais (kg/ha)
Y <- 3.5 * X + rnorm(n, mean = 0, sd = 20)  # Rendement en grain (kg/ha)

# Ajustement du modele de regression lineaire
model <- lm(Y ~ X)
coefficients <- coef(model)
beta0 <- coefficients[1]
beta1 <- coefficients[2]

# Quantite d'engrais pour laquelle on veut predire le rendement
X0 <- 150

# Prediction
```

```r
Y_hat <- beta0 + beta1 * X0

# Calcul de la variance des r sidus
residuals <- model$residuals
sigma_hat_sq <- sum(residuals^2) / (n - 2)

# Calcul de l'erreur standard
X_bar <- mean(X)
SE_Y_hat <- sqrt(sigma_hat_sq * (1/n + (X0 - X_bar)^2 / sum((X - X_
    bar)^2)))

# Calcul du quantile t
alpha <- 0.05
t_value <- qt(1 - alpha/2, df = n - 2)

# Calcul de l'intervalle de confiance
IC_lower <- Y_hat - t_value * SE_Y_hat
IC_upper <- Y_hat + t_value * SE_Y_hat

# R sultats
cat("Valeur pr dite pour X =", X0, "kg/ha :", round(Y_hat, 2), "kg/
    ha\n")
cat("Intervalle de confiance    95% : [", round(IC_lower, 2), ", ",
    round(IC_upper, 2), "] kg/ha\n")
```

Listing 2.27 – Script R pour IC pour la Regression

```r
# Donn es simul es
set.seed(123)
n <- 30
X <- runif(n, 100, 200)  # Quantit  d'engrais (kg/ha)
Y <- 3.5 * X + rnorm(n, mean = 0, sd = 20)  # Rendement en grain (kg
    /ha)

# Ajustement du mod le de r gression lin aire
model <- lm(Y ~ X)

```

```r
# Quantit  d'engrais pour laquelle on veut pr dire le rendement
X0 <- 150

# Pr diction et intervalle de confiance
pred <- predict(model, newdata = data.frame(X = X0), interval = "confidence", level = 0.95)

# R sultats
cat("Valeur pr dite pour X =", X0, "kg/ha :", round(pred[1], 2), "kg/ha\n")
cat("Intervalle de confiance    95% : [", round(pred[2], 2), ", ", round(pred[3], 2), "] kg/ha\n")
```

Listing 2.28 – Script R pour IC pour la Regression

```
Valeur pr dite pour X = 150 kg/ha : 527.38 kg/ha
Intervalle de confiance    95% : [501.47, 553.29] kg/ha
```

Listing 2.29 – IC pour la Regression

2.3.6 IC pour d'autres Paramètres

L'intervalle de confiance (IC) peut être calculé pour divers autres paramètres en dehors de ceux que nous avons déjà discutés, tels que la moyenne, la proportion, la variance, ou les coefficients de régression. L'idée générale est la même : il s'agit de fournir une plage d'estimations possibles pour un paramètre donné avec un certain niveau de confiance. Voici comment vous pourriez définir mathématiquement un IC pour d'autres paramètres, avec un exemple en agronomie :

i. **IC pour la Différence de Moyennes** : Supposons que vous vouliez estimer l'IC pour la différence entre les moyennes de deux groupes (par exemple, deux variétés de cultures). Vous pourriez utiliser une formule similaire à celle de l'IC pour la moyenne, en prenant en compte les moyennes et les variances des deux groupes.

ii. **IC pour la Corrélation** : Si vous étudiez la corrélation entre deux variables (par exemple, la pluviométrie et le rendement des cultures), vous pouvez utiliser des méthodes de transformation statistique pour calculer un IC pour la corrélation (comme le coefficient de corrélation de Pearson) en utilisant la distribution de Student.

iii. **IC pour les Coefficients de Régression** : Outre l'IC pour la régression elle-même, vous pouvez calculer des IC pour les coefficients de régression individuels (par exemple, l'IC pour le coefficient de pente β_1 dans une régression linéaire).

iv. **IC pour la Médiane** : Si vous préférez estimer l'intervalle de confiance pour la médiane plutôt que la moyenne, vous pouvez utiliser des méthodes de statistique non paramétrique comme la méthode de bootstrap.

Supposons que vous meniez une étude pour comparer les rendements de deux types de fertilisants différents (Fertilisant A et Fertilisant B) sur une culture particulière. Vous avez collecté des données sur plusieurs parcelles où ces fertilisants ont été appliqués et vous souhaitez estimer l'IC pour la différence entre les moyennes de rendement de ces deux groupes.

Vous avez les moyennes ($\bar{X}_A$ et $\bar{X}_B$) et les variances (σ_A^2 et $\sigma_A B2$) de chaque groupe à partir de votre échantillon. Vous pouvez utiliser la formule de l'IC pour la différence de moyennes :

$$IC_{95\%} = \left[\left(\bar{X}_A - \bar{X}_B \right) \pm t_{\frac{\alpha}{2},df} \times \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \right]$$

où $t_{\frac{\alpha}{2},df}$ est le quantile de la distribution t avec dfdf degrés de liberté (qui dépend de votre échantillon).

En calculant ces valeurs, vous obtiendrez un intervalle de confiance pour la différence de moyennes entre les deux types de fertilisants. Cela vous aidera à déterminer si l'un des fertilisants a un effet significatif sur le rendement de la culture.

CHAPTER 3

Régression et Corrélation

Ce chapitre explore des techniques statistiques essentielles pour analyser les relations entre variables dans le domaine de l'agronomie. Ces méthodes sont fondamentales pour comprendre comment les différentes variables agronomiques interagissent, influencent les résultats et aident à prendre des décisions éclairées en matière de pratiques agricoles, de gestion des cultures et d'utilisation des ressources.

La régression et la corrélation sont des outils puissants qui permettent de quantifier et d'analyser les liens entre variables et d'anticiper les impacts de divers facteurs sur les résultats agronomiques. Ces méthodes aident à établir des modèles prédictifs, à identifier les tendances et à évaluer les associations entre des paramètres clés. Elles jouent un rôle crucial dans la prise de décisions basées sur des preuves dans le domaine de l'agronomie.

En explorant les concepts de la régression et de la corrélation dans le contexte agronomique, ce chapitre vise à fournir aux chercheurs, aux agronomes et aux professionnels de l'agriculture des outils solides pour analyser les données, établir des relations significatives et prendre des décisions informées pour améliorer les pratiques agricoles et la gestion des cultures.

3.1 La régression

Les techniques de régression sont des méthodes statistiques utilisées pour modéliser la relation entre une variable dépendante (ou variable réponse) et une ou plusieurs variables indépendantes (ou variables explicatives). Ces techniques permettent de comprendre comment les variables indépendantes influencent la variable dépendante et de créer des modèles de prédiction. Les techniques de régression les plus couramment utilisées sont :

1. **Régression Linéaire Simple** : C'est l'une des techniques les plus simples. Elle modélise la relation entre une seule variable indépendante et la variable dépendante en utilisant une équation linéaire (une droite). La méthode des moindres carrés est souvent utilisée pour estimer les coefficients de la ligne qui minimise la somme des carrés des écarts entre les valeurs observées et prédites.

2. **Régression Linéaire Multiple** : Cette technique étend la régression linéaire simple à plusieurs variables indépendantes. Elle permet de modéliser des relations plus complexes en utilisant une équation linéaire avec plusieurs coefficients.

3. **Régression Logistique (Logit)** : La régression logistique est utilisée lorsque la variable dépendante est binaire (2 catégories) ou nominale (plus de 2 catégories). Elle modélise la probabilité d'appartenance à une catégorie en fonction des variables indépendantes en utilisant la fonction logistique. Les coefficients de régression sont estimés en utilisant la méthode du maximum de vraisemblance. La régression logistique produit des cotes (odds ratios) qui quantifient comment les variables indépendantes influencent la probabilité de succès dans le cas d'une variable binaire. C'est une méthode couramment utilisée dans la modélisation de données binaires, telles que les résultats médicaux (malade/sain) ou les taux de réussite/échec.

4. **Régression Probit** : La régression probit est similaire à la régression logistique, mais au lieu d'utiliser la fonction logistique, elle utilise la fonction de distribution normale cumulative (fonction probit) pour modéliser la relation entre les variables indépendantes et la variable dépendante binaire. Les coefficients de régression sont également estimés en utilisant la méthode du maximum de vraisemblance. La régression probit produit des probabilités de succès et peut être utilisée pour des données binaires ou nominales.

5. **Régression Polynomiale** : Lorsque la relation entre les variables ne peut pas être bien modélisée par une ligne droite, une régression polynomiale peut être utilisée.

Elle utilise des polynômes (par exemple, un carré ou un cube) pour ajuster les données de manière plus flexible.

6. **Régression Logistique** : Cette technique est utilisée lorsque la variable dépendante est binaire (par exemple, oui/non, réussi/échoué). Elle modélise la probabilité d'appartenance à une catégorie en fonction des variables indépendantes.
7. **Régression Ridge et Lasso** : Ces techniques sont des extensions de la régression linéaire qui ajoutent des termes de régularisation pour éviter le surajustement (overfitting) du modèle. Elles sont particulièrement utiles lorsque de nombreuses variables indépendantes sont présentes.
8. **Régression Non Linéaire** : Lorsque la relation entre les variables ne suit pas un modèle linéaire, des techniques de régression non linéaire, telles que la régression exponentielle ou la régression logarithmique, peuvent être utilisées.
9. **Régression PLS (Partial Least Squares)** : Cette technique est utilisée pour modéliser les relations entre les variables indépendantes et dépendantes en réduisant la dimension des variables indépendantes pour éviter le surajustement.
10. **Régression Bayésienne** : Elle utilise des méthodes bayésiennes pour estimer les paramètres du modèle de régression en prenant en compte des connaissances a priori.
11. **Régression à Effets Mixtes** : Cette technique est utilisée lorsque les données présentent une structure hiérarchique, comme dans les études longitudinales ou les enquêtes répétées.
12. **Régression Quantile** : Elle est utilisée pour modéliser différentes parties de la distribution des données et peut être plus robuste aux valeurs aberrantes.

Le choix de la technique de régression dépendra de la nature des données, de la relation supposée entre les variables et des objectifs d'analyse. Il est important de comprendre les hypothèses et les limitations de chaque technique pour sélectionner celle qui convient le mieux à votre situation.

3.1.1 Régression Linéaire Simple

La "Régression Linéaire Simple" est l'un des concepts fondamentaux de la régression statistique et de l'analyse de données. Elle vise à établir une relation linéaire entre une variable

indépendante (également appelée variable explicative) et une variable dépendante (ou variable cible).

① **Modèle** : Le modèle de régression linéaire simple s'exprime comme suit :

$$y = \beta_0 + \beta_1 \times x + \varepsilon \tag{3.1}$$

Où, y est la variable dépendante. x est la variable indépendante. β_0 est l'intercept (ordonnée à l'origine). β_1 est la pente de la droite de régression. ε est le terme d'erreur, représentant les écarts non expliqués par le modèle.

② **Hypothèse** : L'hypothèse de la régression linéaire simple repose sur plusieurs suppositions qui doivent être vérifiées pour que les résultats et les interprétations du modèle soient fiables et valides. Voici les principales hypothèses de la régression linéaire simple :

(a) **Linéarité** : La relation entre la variable indépendante x et la variable dépendante y est linéaire. Cela signifie que le modèle de régression linéaire est approprié pour décrire la relation entre ces deux variables.

(b) **Homoscédasticité (Homogénéité de la variance)** : Les résidus (écarts entre les valeurs observées y et les valeurs prédites $\hat{y}$) doivent avoir une variance constante à tous les niveaux de la variable indépendante x. En d'autres termes, la dispersion des résidus ne doit pas changer de manière significative à mesure que x varie.

(c) **Normalité des Résidus** : Les résidus doivent suivre une distribution normale (gaussienne). Cela signifie que les résidus doivent être approximativement centrés autour de zéro et être répartis de manière symétrique.

(d) **Indépendance des Résidus** : Les résidus doivent être indépendants les uns des autres. Les erreurs d'une observation ne doivent pas être corrélées avec les erreurs des autres observations.

(e) **Absence de Multicolinéarité** : Dans le cas où il y a plusieurs variables indépendantes, elles ne doivent pas être fortement corrélées entre elles (phénomène de multicolinéarité). Une forte corrélation entre les variables indépendantes peut rendre les estimations des coefficients de régression instables et difficiles à interpréter.

Ces hypothèses sont importantes car elles garantissent la validité et la fiabilité des résultats de la régression linéaire. Si ces hypothèses ne sont pas satisfaites, les conclusions tirées du modèle pourraient être biaisées ou peu fiables. Avant d'interpréter les résultats d'une régression linéaire simple, il est crucial de vérifier si ces hypothèses sont respectées.

③ **Paramètres et Estimateurs** : Les paramètres du modèle, β_0 et β_1, sont estimés à partir des données disponibles. Les estimateurs des paramètres sont obtenus en minimisant la somme des carrés des résidus (erreurs) entre les valeurs observées et les valeurs prédites par le modèle.

④ **Méthode des Moindres Carrés (MCO)** : La méthode des moindres carrés est utilisée pour estimer les paramètres du modèle. Elle consiste à minimiser la somme des carrés des résidus :

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum e_i^2 \tag{3.2}$$

Où, n est le nombre d'observations. y_i sont les valeurs observées de la variable dépendante. $\hat{y}_i$ sont les valeurs prédites par le modèle.
Pour un modèle de régression linéaire simple, les paramètres estimés β_0 (l'intercept) et β_1 (la pente) peuvent être calculés comme suit :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.3}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \times \bar{x} \tag{3.4}$$

Où x_i sont les valeurs de la variable indépendante, $\bar{x}$ est la moyenne des valeurs de x, y_i sont les valeurs observées de la variable dépendante, $\bar{y}$ est la moyenne des valeurs de y.

Supposons que vous collectiez des données sur le rendement de cultures de blé en fonction de la quantité d'engrais utilisée. Voici un exemple de données fictives :

Engrais :[50,100,150,200,250](kg par hectare)

Rendement :[1000,1200,1400,1600,1800](kg par hectare)

En utilisant la régression linéaire simple, vous pouvez estimer les paramètres β_0 et β_1 pour le modèle. Ces paramètres vous donneront une équation de la forme $Rendement = \beta_0 + \beta_1 \times Engrais$.

Une fois que vous avez ajusté le modèle aux données, vous pouvez l'utiliser pour prédire le rendement en fonction de différentes quantités d'engrais. Cela pourrait vous aider à déterminer la quantité optimale d'engrais à utiliser pour obtenir le rendement maximal de la culture de blé.

```r
# Étape 1 : Créer les données
engrais <- c(50, 100, 150, 200, 250)  # Quantité d'engrais en kg/ha
rendement <- c(1000, 1200, 1400, 1600, 1800)  # Rendement en kg/ha

# Étape 2 : Ajuster le modèle de régression linéaire
model <- lm(rendement ~ engrais)

# Étape 3 : Estimer les paramètres
coefficients <- coef(model)
beta0 <- coefficients[1]  # Ordonnée à l'origine (intercept)
beta1 <- coefficients[2]  # Pente (slope)

# Afficher les résultats
cat("Équation du modèle : Rendement =", round(beta0, 2), "+",
    round(beta1, 2), "* Engrais\n")

# Étape 4 : Prédire le rendement pour différentes quantités d'
    engrais
# Exemple de prédiction pour 300 kg/ha d'engrais
engrais_pred <- 300
rendement_pred <- beta0 + beta1 * engrais_pred
```

```r
cat("Pr diction du rendement pour", engrais_pred, "kg/ha d'engrais
   :", round(rendement_pred, 2), "kg/ha\n")
```

Listing 3.1 – Script R pour la Régression Linéaire Simple

```
 quation  du mod le : Rendement = 800 + 4 * Engrais
Pr diction du rendement pour 300 kg/ha d'engrais : 2000 kg/ha
```

Listing 3.2 – Régression Linéaire Simple

```r
#  tape  1 : Cr er les donn es
engrais <- c(50, 100, 150, 200, 250)  # Quantit  d'engrais en kg/ha
rendement <- c(1000, 1200, 1400, 1600, 1800)  # Rendement en kg/ha

#  tape  2 : Ajuster le mod le de r gression lin aire
model <- lm(rendement ~ engrais)

#  tape  3 : Afficher le r sum  complet du mod le
summary(model)
```

Listing 3.3 – Script R pour la Régression Linéaire Simple

```
Call:
lm(formula = rendement ~ engrais)

Residuals:
    1    2    3    4    5
 0.00  0.00  0.00  0.00  0.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    800.0        0.0       Inf   <2e-16 ***
engrais          4.0        0.0       Inf   <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
    0.1         1

Residual standard error: 0 on 3 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
```

```
F-statistic:    Inf on 1 and 3 DF,  p-value: < 2.2e-16
```

Listing 3.4 – Régression Linéaire Simple

Interprétation des résultats :

— Équation du modèle : L'équation ajustée pour prédire le rendement en fonction de la quantité d'engrais est $Rendement = 800 + 4 \times Engrais$. Cela signifie que pour chaque augmentation de 1 kg/ha d'engrais, le rendement augmente de 4 kg/ha, et le rendement de base (sans engrais) serait de 800 kg/ha.

— Prédiction du rendement : Pour une application de 300 kg/ha d'engrais, le rendement attendu serait de 2000 kg/ha.

— Residuals : Montre les résidus du modèle, qui sont les différences entre les valeurs observées et les valeurs ajustées.

— Coefficients : Donne les estimations des coefficients du modèle, l'erreur standard, les valeurs t, et les p-values associées.

— Residual standard error : L'écart-type des résidus, qui donne une mesure de la dispersion des résidus.

— R-squared : Le coefficient de détermination, indiquant la proportion de la variance dans la variable dépendante qui est prédite par le modèle.

— F-statistic : Mesure de la signification globale du modèle.

Avec ce modèle, vous pouvez maintenant prédire le rendement attendu pour différentes quantités d'engrais et ajuster vos pratiques agricoles pour optimiser la production. Vous pourriez également utiliser le modèle pour déterminer la quantité d'engrais qui maximise le rendement, en effectuant des calculs supplémentaires ou en utilisant l'outil d'optimisation de R.

3.1.2 Régression Linéaire Multiple

La "Régression Linéaire Multiple" est une extension de la Régression Linéaire Simple, qui permet de modéliser la relation entre une variable dépendante et plusieurs variables indépendantes. En agronomie, cela peut être utilisé pour étudier comment plusieurs facteurs influencent le rendement d'une culture.

① **Modèle** : Le modèle de Régression Linéaire Multiple s'exprime comme suit :

$$y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + ... + \beta_p \times X_p + \varepsilon \tag{3.5}$$

Où, y représente la variable dépendante (par exemple, le rendement de la culture). x_i pour $i \in \{1, 2, ..., p\}$ sont les variables indépendantes (par exemple, la quantité d'engrais, l'irrigation, la température, etc.). β_i pour $i \in \{0, 1, 2, ..., p\}$ sont les coefficients de régression correspondants. ε représente le terme d'erreur.

② **Hypothèse** : L'hypothèse de la régression linéaire simple repose sur plusieurs suppositions qui doivent être vérifiées pour que les résultats et les interprétations du modèle soient fiables et valides. Voici les principales hypothèses de la régression linéaire simple :

(a) **Linéarité** : La relation entre la variable indépendante x et la variable dépendante y est linéaire. Cela signifie que le modèle de régression linéaire est approprié pour décrire la relation entre ces deux variables.

(b) **Homoscédasticité (Homogénéité de la variance)** : Les résidus (écarts entre les valeurs observées y et les valeurs prédites $\hat{y}$) doivent avoir une variance constante à tous les niveaux de la variable indépendante x. En d'autres termes, la dispersion des résidus ne doit pas changer de manière significative à mesure que x varie.

(c) **Normalité des Résidus** : Les résidus doivent suivre une distribution normale (gaussienne). Cela signifie que les résidus doivent être approximativement centrés autour de zéro et être répartis de manière symétrique.

(d) **Indépendance des Résidus** : Les résidus doivent être indépendants les uns des autres. Les erreurs d'une observation ne doivent pas être corrélées avec les erreurs des autres observations.

(e) **Absence de Multicolinéarité** : Dans le cas où il y a plusieurs variables indépendantes, elles ne doivent pas être fortement corrélées entre elles (phénomène de

multicolinéarité). Une forte corrélation entre les variables indépendantes peut rendre les estimations des coefficients de régression instables et difficiles à interpréter.

③ **Paramètres et Estimateurs** : Les paramètres β_i pour $i \in \{0, 1, 2, ..., p\}$ sont estimés à partir des données en utilisant la Méthode des Moindres Carrés, tout comme dans le cas de la Régression Linéaire Simple.

④ **Méthode des Moindres Carrés** :

$$SSE = ee^{tr} \tag{3.6}$$

$$\hat{\beta} = (xx^{tr})^{-1}x^{tr}y \tag{3.7}$$

Supposons que vous ayez collecté des données sur le rendement d'une culture en fonction de la quantité d'engrais et de la quantité d'irrigation. Voici un exemple fictif de données :
x_1 :Quantité d'engrais (kg par hectare)
x_2 :Quantité d'irrigation (mm par semaine)
y :Rendement de la culture (kg par hectare)
En utilisant la Régression Linéaire Multiple, vous pouvez estimer les coefficients β_0, β_1 et β_2 du modèle, créer l'équation du modèle et prédire le rendement en fonction de la quantité d'engrais et d'irrigation pour différentes parcelles de terre.

```r
# Étape 1 : Créer les données
engrais <- c(50, 100, 150, 200, 250)          # Quantité d'engrais en kg/ha
irrigation <- c(10, 20, 30, 40, 50)           # Quantité d'irrigation en mm/semaine
rendement <- c(1100, 1300, 1500, 1700, 1900)  # Rendement en kg/ha

# Étape 2 : Ajuster le modèle de régression linéaire multiple
model <- lm(rendement ~ engrais + irrigation)

# Étape 3 : Afficher le résumé complet du modèle
summary(model)
```

Listing 3.5 – Script R pour la Régression Linéaire Multiple

```
Call:
lm(formula = rendement ~ engrais + irrigation)

Residuals:
      1       2       3       4       5
 7.546e-14 -7.389e-14  4.038e-14  1.516e-13 -9.438e-14

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   900.00       0.00     Inf   <2e-16 ***
engrais         2.00       0.00     Inf   <2e-16 ***
irrigation      2.00       0.00     Inf   <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
    0.1         1

Residual standard error: 0 on 2 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic:    Inf on 2 and 2 DF,  p-value: < 2.2e-16
```

Listing 3.6 – Régression Linéaire Multiple

Interprétation des résultats :

— Residuals : Montre les résidus du modèle, représentant les différences entre les valeurs observées et les valeurs ajustées.

— Coefficients : Donne les estimations des coefficients du modèle (Intercept, engrais, irrigation), l'erreur standard, les valeurs t, et les p-values associées.

— Residual standard error : L'écart-type des résidus, qui mesure la dispersion des résidus.

— R-squared : Le coefficient de détermination, indiquant la proportion de la variance dans le rendement qui est expliquée par le modèle.

— F-statistic : Mesure la signification globale du modèle, testant si au moins un des coefficients de régression est statistiquement significatif.

Le modèle de régression linéaire multiple a été ajusté pour prédire le rendement en fonction de la quantité d'engrais et de la quantité d'irrigation. Les coefficients estimés montrent l'impact de chaque variable indépendante sur le rendement, permettant ainsi de comprendre l'effet combiné de l'engrais et de l'irrigation sur la culture.

3.1.3 Régression Probit

La Régression Probit est un modèle statistique qui est utilisé pour analyser les relations entre une variable binaire dépendante (0 ou 1) et des variables indépendantes continues ou catégorielles. Elle est souvent utilisée dans le contexte de la modélisation de probabilités de résultats binaires, tels que la présence ou l'absence d'un événement.

① **Modèle** : Dans la régression Probit, le modèle est généralement exprimé sous forme d'une fonction de répartition cumulative de la distribution normale, également appelée fonction Probit :

$$P(y = 1|x) = \Phi(\beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + ... + \beta_p \times x_p) \tag{3.8}$$

Où $P(y = 1|x)$ est la probabilité conditionnelle que la variable binaire y prenne la valeur 1 compte tenu des variables explicatives x. Φ est la fonction de répartition cumulative de la distribution normale standard. β_i pour $i \in \{0, 1, ..., p\}$ sont les coefficients de régression correspondants. x_i pour $i \in \{1, 2, ..., p\}$ sont les variables explicatives.

② **Paramètres et Estimateurs** : Les coefficients β_i pour $i \in \{0, 1, ..., p\}$ sont estimés à partir des données en utilisant des méthodes d'optimisation qui maximisent la fonction de vraisemblance. Les estimations des paramètres sont obtenues de manière à maximiser la probabilité d'observer les données réelles en fonction du modèle Probit.

③ **Méthode de maximum de vraissemblance** : La technique d'estimation par vraisemblance est couramment utilisée pour estimer les paramètres d'un modèle statistique, y compris les modèles de régression comme la régression Probit.
La fonction de vraisemblance est la probabilité d'observer les données observées compte tenu des paramètres du modèle. Pour un modèle Probit, la vraisemblance pour une seule observation ii peut être exprimée comme suit :

$$L_i(\beta) = (\Phi(\beta_0+\beta_1\times x_{i1}+\beta_2\times x_{i2}+...+\beta_p\times x_{ip})^{y_i}\times(1-\Phi(\beta_0+\beta_1\times x_{i1}+\beta_2\times x_{i2}+...+\beta_p\times x_{ip})^{1-y_i} \tag{3.9}$$

L'objectif de l'estimation par vraisemblance est de trouver les valeurs des paramètres β_i pour $i \in \{0, 1, ..., p\}$ qui maximisent la fonction de vraisemblance globale, qui est la multiplication des vraisemblances individuelles pour toutes les observations :

$$L(\beta) = \prod_{i=1}^{n} L_i(\beta) \tag{3.10}$$

Où n est le nombre d'observations.

Supposons que vous souhaitez étudier la probabilité de réussite d'une culture donnée en fonction de variables telles que la quantité d'engrais utilisée, la température, et la quantité d'eau d'irrigation. Vous pourriez collecter des données sur la présence (1) ou l'absence (0) de réussite de la culture, ainsi que sur les variables explicatives.
En utilisant la régression Probit, vous pouvez estimer les coefficients de régression et prédire la probabilité de réussite de la culture en fonction des valeurs des variables explicatives.

```r
# Étape 1 : Créer les données fictives
set.seed(123)  # Pour la reproductibilité

# Variables explicatives
engrais <- c(50, 100, 150, 200, 250)        # Quantité d'engrais en kg/ha
temperature <- c(15, 18, 20, 22, 25)        # Température en degrés Celsius
irrigation <- c(10, 20, 30, 40, 50)         # Quantité d'eau d'irrigation en mm/semaine

# Variable dépendante (succès ou échec de la culture)
reussite <- c(0, 0, 1, 1, 1)  # 0 = échec, 1 = réussite

# Étape 2 : Ajuster le modèle de régression Probit
model_probit <- glm(reussite ~ engrais + temperature + irrigation, family = binomial(link = "probit"))

# Étape 3 : Afficher le résumé complet du modèle
summary(model_probit)
```

Listing 3.7 – Script R pour la Régression Probit

```
Call:
glm(formula = reussite ~ engrais + temperature + irrigation, family
    = binomial(link = "probit"))

Deviance Residuals:
       1         2         3         4         5
-0.59457  -0.65529   0.60315   0.65489   0.59451

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -25.81230   20.68292  -1.248    0.212
engrais       0.30134    0.24818   1.214    0.225
temperature   1.72049    1.40300   1.226    0.220
```

```
irrigation    1.72062    1.40309   1.226    0.220

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.4334  on 4  degrees of freedom
Residual deviance: 1.0816  on 1  degrees of freedom
AIC: 11.816

Number of Fisher Scoring iterations: 6
```

Listing 3.8 – Régression Linéaire Probit

Interprétation des résultats :

- Coefficients : Les estimations des coefficients de régression pour chaque variable explicative (engrais, température, irrigation), accompagnées de leur erreur standard, valeur z, et p-value.
- Deviance Residuals : Les résidus de déviance, indiquant l'écart entre les valeurs observées et celles prédites par le modèle.
- Null deviance : La déviance du modèle null (modèle avec seulement l'interception), qui sert de base de comparaison.
- Residual deviance : La déviance résiduelle du modèle ajusté, utilisée pour évaluer l'ajustement du modèle.
- AIC : Le critère d'information d'Akaike, une mesure de la qualité relative du modèle (plus faible est meilleur).

Le modèle de régression Probit a été ajusté pour prédire la probabilité de réussite d'une culture en fonction de la quantité d'engrais, de la température, et de la quantité d'irrigation. Les coefficients obtenus permettent de comprendre comment chaque variable explicative influence la probabilité de réussite. Vous pouvez également utiliser ce modèle pour prédire la probabilité de réussite pour de nouvelles combinaisons de variables explicatives.

3.1.4 Régression Logit

La "Régression Logistique" (ou "Régression Logit") est une méthode statistique utilisée pour modéliser des relations entre des variables binaires (0/1) et des variables explicatives continues. Elle est couramment utilisée en sciences sociales, en épidémiologie et en agronomie pour analyser des situations où la variable dépendante est binaire, comme la présence ou l'absence d'un événement.

① **Modèle** : Le modèle de régression logistique modélise la probabilité conditionnelle que la variable binaire yy prenne la valeur 1 compte tenu des variables explicatives xx. La relation entre les variables explicatives et la probabilité de succès est modélisée à l'aide de la fonction logistique (ou sigmoid) :

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1\times x_1+\beta_2\times x_2+...+\beta_p\times x_p)}} \tag{3.11}$$

Où $P(y = 1|x)$ est est la probabilité conditionnelle que la variable binaire y prenne la valeur 1 compte tenu des variables explicatives x. e est est la base du logarithme naturel. β_i pour $i \in \{0, 1, ..., p\}$ les coefficients de régression correspondants. x_i pour $i \in \{1, 2, ..., p\}$ sont les variables explicatives.

② **Paramètres et Estimateurs** : Les coefficients β_i pour $i \in \{0, 1, ..., p\}$ sont estimés à partir des données en utilisant des méthodes d'optimisation qui maximisent la fonction de vraisemblance. Les estimations des paramètres sont obtenues de manière à maximiser la probabilité d'observer les données réelles en fonction du modèle logistique.

Supposons que vous souhaitez modéliser la probabilité de succès d'une culture en fonction de la quantité d'engrais utilisée et de la température. Vous auriez des données sur la présence (1) ou l'absence (0) de succès de la culture, ainsi que sur les valeurs de la quantité d'engrais et de la température.
En utilisant la régression logistique, vous pourriez estimer les paramètres β_i pour $i \in \{0, 1, 2\}$ du modèle logistique pour déterminer comment la quantité d'engrais et la température influencent la probabilité de succès de la culture.

```r
# Étape 1 : Créer les données fictives
set.seed(123)  # Pour la reproductibilité

# Variables explicatives
engrais <- c(50, 100, 150, 200, 250)        # Quantité d'engrais en kg/ha
temperature <- c(15, 18, 20, 22, 25)        # Température en degrés Celsius

# Variable dépendante (succès ou échec de la culture)
reussite <- c(0, 0, 1, 1, 1)  # 0 = échec, 1 = réussite

# Étape 2 : Ajuster le modèle de régression logistique
model_logit <- glm(reussite ~ engrais + temperature, family = binomial(link = "logit"))

# Étape 3 : Afficher le résumé complet du modèle
summary(model_logit)
```

Listing 3.9 – Script R pour la Régression Logit

```r
# Nouvelle donnée pour prédiction
nouvelle_donnee <- data.frame(engrais = 180, temperature = 21)

# Prédire la probabilité de succès
probabilite_succes <- predict(model_logit, nouvelle_donnee, type = "response")
print(probabilite_succes)
```

Listing 3.10 – Utilisation du Modèle pour la Prédiction

```
Call:
glm(formula = reussite ~ engrais + temperature, family = binomial(
    link = "logit"))

Deviance Residuals:
       1         2         3         4         5
```

```
-1.1700  -1.5243   0.8124   0.8123   0.8123

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -36.2415    38.8726  -0.932    0.351
engrais        0.3351     0.3767   0.890    0.374
temperature    2.4308     2.6082   0.932    0.351

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.9188  on 4  degrees of freedom
Residual deviance: 3.0249  on 2  degrees of freedom
AIC: 9.0249

Number of Fisher Scoring iterations: 6
```

Listing 3.11 – Régression Linéaire Logit

Interprétation des résultats :

— Coefficients : Les estimations des coefficients de régression pour chaque variable explicative (engrais, température), accompagnées de leur erreur standard, valeur z, et p-value.

— Deviance Residuals : Les résidus de déviance, indiquant l'écart entre les valeurs observées et celles prédites par le modèle.

— Null deviance : La déviance du modèle null (modèle avec seulement l'interception), qui sert de base de comparaison.

— Residual deviance : La déviance résiduelle du modèle ajusté, utilisée pour évaluer l'ajustement du modèle.

— AIC : Le critère d'information d'Akaike, qui est une mesure de la qualité relative du modèle (plus faible est meilleur).

Le modèle de régression logistique a été ajusté pour prédire la probabilité de succès d'une culture en fonction de la quantité d'engrais et de la température. Les coefficients obtenus indiquent comment chaque variable explicative influence la probabilité de succès. Ce modèle peut également être utilisé pour prédire la probabilité de succès pour différentes combinaisons de quantité d'engrais et de température.

3.2 Correlation

Deux caractères quantitatifs, X et Y, qui décrivent le même ensemble d'unités sont considérés comme ayant une relation lorsque l'attribution de leurs modalités n'est pas aléatoire. En d'autres termes, cela signifie que les valeurs de X ne sont pas indépendantes des valeurs de Y, ou vice versa. Lorsque l'on dit que Y dépend de X, cela implique que connaître les valeurs de X peut, dans une certaine mesure, nous aider à prédire les valeurs de Y. En d'autres termes, si Y dépend de X, il existe une fonction f telle que :

$$Y = f(X) \tag{3.12}$$

Supposons que vous ayez collecté des données sur la quantité d'azote (X, en kilogrammes par hectare) appliquée à une culture de blé et les rendements de cette culture de blé (Y, en tonnes par hectare) sur différentes parcelles.

Parcelle	Quantité d'Azote (X)	Rendement de Blé (Y)
Parcelle 1	50	3.2
Parcelle 2	75	3.5
Parcelle 3	100	4.1
Parcelle 4	125	4.3
Parcelle 5	150	4.0

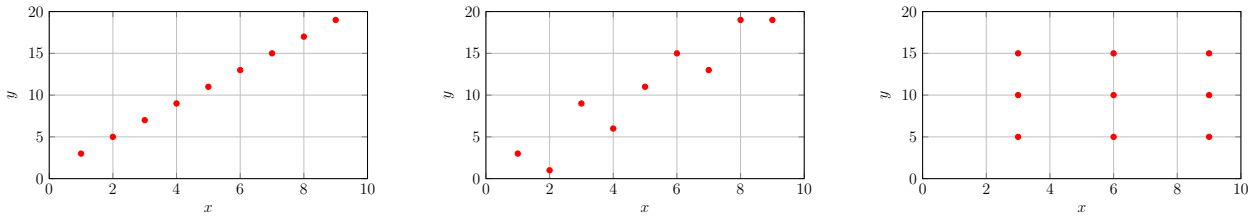
Vous soupçonnez qu'il existe une relation entre la quantité d'azote (X) et le rendement de blé (Y). Plus précisément, vous pensez que l'ajout d'azote peut avoir un impact sur le rendement du blé.

$$Y = 0.02X + 2.5$$

La notion de dépendance entre deux variables n'est pas symétrique, ce qui signifie que la dépendance de la variable Y par rapport à la variable X peut être différente de la dépendance de la variable X par rapport à la variable Y. En d'autres termes, la relation entre deux variables n'implique pas nécessairement une relation équivalente dans les deux sens.

Supposons que vous étudiez la relation entre la quantité d'eau d'irrigation (X) et la croissance des plantes (Y) dans une serre. Vous pourriez constater que lorsque vous augmentez la quantité d'eau d'irrigation (X), la croissance des plantes (Y) augmente de manière significative. Cela suggère une dépendance de Y par rapport à X, car X semble influencer Y.

Cependant, si vous inversiez le raisonnement et observiez que la croissance des plantes (Y) n'influence pas la quantité d'eau d'irrigation (X), la dépendance de X par rapport à Y n'existerait pas dans ce cas.



(a) Relation forte (b) Relation faible (c) Absence de relation

FIGURE 3.1 – Intensité de la relation

3.2.1 Les types de relations entre deux caractéres quantitatifs

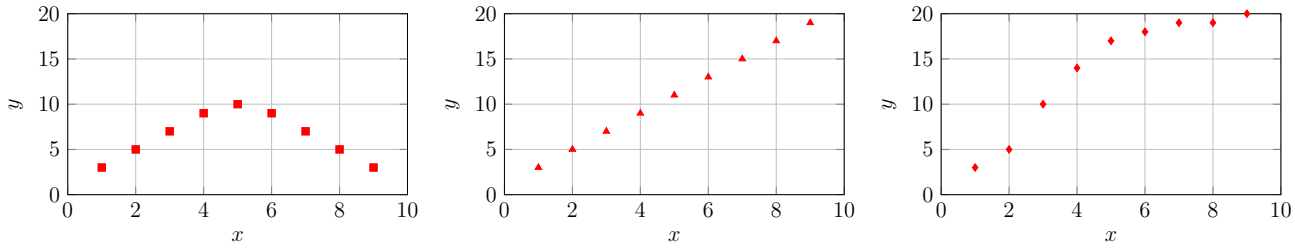
Pour déterminer si une relation existe entre deux caractères, nous construisons un diagramme de corrélation, qui représente graphiquement les modalités de X et de Y. Chaque point individuel, noté (X_i, Y_i), est positionné sur ce diagramme. L'ensemble de ces points forme un nuage de points, dont l'aspect global offre des indices pour caractériser la relation à travers trois aspects clés :

① **Intensité de la relation** : L'intensité de la relation entre deux variables peut être évaluée en observant comment les valeurs de ces variables se comportent ensemble (voir figure 3.1). Une relation est considérée comme forte lorsque les unités ayant des valeurs proches sur la variable X ont également des valeurs proches sur la variable Y. En d'autres termes, si vous avez Xi proche de Xj, alors vous pouvez vous attendre à ce que Yi soit également proche de Yj. Dans ce cas, le nuage de points prend généralement la forme d'une ligne ou d'une courbe, et les points sont relativement proches les uns des autres.

D'un autre côté, une relation est qualifiée de faible lorsque les unités ayant des valeurs proches sur la variable X peuvent avoir des valeurs éloignées sur la variable Y. En d'autres termes, deux valeurs proches de X peuvent correspondre à deux valeurs très différentes de Y. Dans cette situation, le nuage de points ne présente pas de forme nette de ligne ou de courbe, ou si une tendance existe, elle est très grossière.

Enfin, une relation est nulle lorsque les valeurs de X ne fournissent aucune indication permettant de prédire les valeurs de Y. Dans ce cas, le nuage de points peut avoir une forme indéfinie, ressemblant à un carré, un cercle, ou même à une dispersion aléatoire sans direction claire.

② **Forme de la relation** : La forme de la relation entre deux variables peut être caractérisée de la manière suivante (voir figure 3.2) :



(a) Non monotone (b) Linéaire (c) Non linéaire

FIGURE 3.2 – La forme de la relation

— Une relation est considérée comme linéaire lorsqu'il est possible d'établir une relation mathématique de la forme $Y = aX + b$, où a et b sont des constantes. Autrement dit, le nuage de points peut être parfaitement ajusté à une droite. Dans ce cas, la relation entre les variables est régulière et directement proportionnelle.

— Une relation est qualifiée de non linéaire lorsque la relation entre X et Y ne peut pas être exprimée sous la forme $Y = aX + b$, mais plutôt sous une forme différente, telle qu'une parabole, une hyperbole, une sinusoïde, etc. Le nuage de points prend alors une forme complexe avec des courbures.

— Une relation non linéaire est dite monotone si elle présente une tendance stricte à la croissance ou à la décroissance, c'est-à-dire qu'elle ne comporte pas de minima ou de maxima. Cela signifie que la relation ne s'inverse pas en passant par des valeurs extrêmes.

Il est important de noter que toutes les relations linéaires sont également monotones, mais l'inverse n'est pas nécessairement vrai.

③ **Sens de la relation** : Le sens de la relation entre deux caractères, qu'ils soient liés de manière monotone (linéaire ou non), peut être interprété comme suit (voir figure 3.3) :

— Une relation monotone est qualifiée de positive lorsque les deux caractères varient dans le même sens, c'est-à-dire que généralement, lorsque X_i est plus grand que X_j, on observe que Y_i est également plus grand que Y_j. Autrement dit :

- Les valeurs élevées de X correspondent généralement à des valeurs élevées de Y.
- Les valeurs moyennes de X correspondent généralement à des valeurs moyennes de Y.
- Les valeurs faibles de X correspondent généralement à des valeurs faibles de Y.

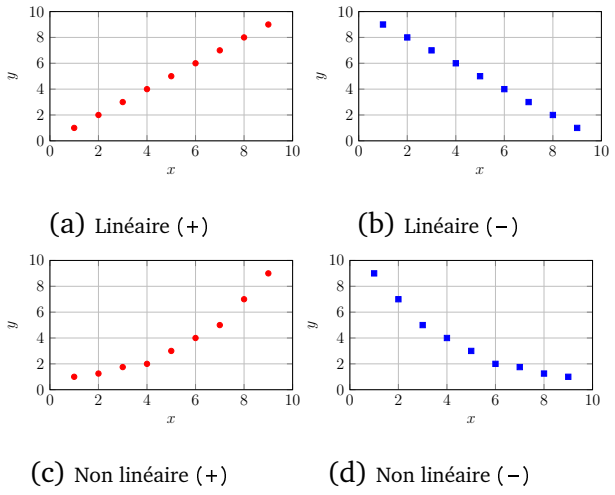
— En revanche, une relation monotone est dite négative lorsque les deux caractères varient en sens inverse, c'est-à-dire que généralement, lorsque X_i est plus grand que X_j, Y_i est généralement plus petit que Y_j. En d'autres termes :

- Les valeurs élevées de X correspondent généralement à des valeurs faibles de Y.
- Les valeurs moyennes de X correspondent généralement à des valeurs moyennes de Y.
- Les valeurs faibles de X correspondent généralement à des valeurs élevées de Y.

La compréhension du sens de la relation est essentielle pour déterminer comment les variables interagissent. Une relation positive suggère une co-variation dans le même sens, tandis qu'une relation négative suggère une co-variation dans des directions opposées. Cette information peut être cruciale pour la prise de décision et l'interprétation des résultats dans divers domaines, y compris en agronomie pour évaluer les relations entre les facteurs agronomiques et les performances des cultures.

3.2.2 Le calcul des coefficients de correlation

Les coefficients de corrélation fournissent une mesure synthétique de l'intensité et du sens de la relation entre deux caractères, en particulier lorsqu'il s'agit d'une relation monotone. Dans ce contexte, le coefficient de corrélation de Pearson est utilisé pour analyser les relations linéaires, tandis que le coefficient de corrélation de Spearman est adapté aux relations non linéaires monotones. Bien qu'il existe d'autres coefficients pour les relations non linéaires et non monotones, nous nous concentrerons ici sur ces deux.



(a) Linéaire (+) (b) Linéaire (−)

(c) Non linéaire (+) (d) Non linéaire (−)

FIGURE 3.3 – Le sens de la relation

Supposons que nous menions une étude agronomique visant à évaluer la relation entre la quantité de nutriments dans le sol (X) et le rendement en graines de maïs par hectare (Y) (voir tableau 3.1). Nous collectons des données sur plusieurs parcelles agricoles et souhaitons déterminer si une corrélation existe entre ces deux variables.

Voici un tableau de données fictif représentant la quantité de nutriments dans le sol (exprimée en unités d'azote par hectare) et le rendement en graines de maïs (exprimé en kilogrammes par hectare) pour 10 parcelles différentes :

Parcelle	Quantité de Nutriments (X)	Rendement en Maïs (Y)
Parcelle 1	40	2500
Parcelle 2	60	2800
Parcelle 3	45	2600
Parcelle 4	35	2400
Parcelle 5	70	3200
Parcelle 6	30	2300
Parcelle 7	75	3300
Parcelle 8	68	3150
Parcelle 9	33	2450
Parcelle 10	55	2750

TABLE 3.1 – La quantité de nutriments dans le sol

Dans ce contexte agronomique, nous utiliserions le coefficient de corrélation de Pearson pour évaluer s'il existe une relation linéaire entre la quantité de nutriments dans le sol

et le rendement en maïs. De plus, nous pourrions également calculer le coefficient de corrélation de Spearman pour examiner si une relation monotone, qu'elle soit linéaire ou non, est présente.

Ces coefficients de corrélation nous aideraient à quantifier l'intensité de la relation entre ces deux variables agronomiques et à déterminer si une variation dans la quantité de nutriments dans le sol est associée à une variation correspondante dans le rendement en maïs. Ces informations seraient précieuses pour les agriculteurs et les chercheurs en agronomie afin de mieux comprendre les facteurs qui influencent les récoltes.

① **Le coefficient de corrélation linéaire de Bravais-Pearson** : Le coefficient de corrélation linéaire de Bravais-Pearson, communément appelé le coefficient de corrélation de Pearson, est une mesure statistique qui quantifie la force et la direction d'une relation linéaire entre deux variables continues. Il a été développé par Francis Galton, puis perfectionné par Karl Pearson, d'où son nom.

La formule mathématique du coefficient de corrélation de Pearson entre deux variables X et Y est la suivante :

$$\rho_{(X,Y)} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum \left(\left(X_i - \bar{X}\right) \left(Y_i - \bar{Y}\right) \right)}{\sqrt{\sum \left(\left(X_i - \bar{X}\right)^2 \left(Y_i - \bar{Y}\right)^2 \right)}} \tag{3.13}$$

Où :

$\rho_{(X,Y)}$ est le coefficient de corrélation de Pearson.

X_i et Y_i sont les observations individuelles des variables X et Y.

$\bar{X}$ est la moyenne des observations de X.

$\bar{Y}$ est la moyenne des observations de Y.

Le coefficient de corrélation de Pearson peut prendre des valeurs dans l'intervalle de -1 à 1 :

- Si $\rho_{(X,Y)} = 1$, cela signifie une corrélation linéaire parfaite positive, ce qui indique que les deux variables sont parfaitement corrélées de manière positive, c'est-à-dire qu'elles augmentent ensemble de manière linéaire.
- Si $\rho_{(X,Y)} = -1$, cela signifie une corrélation linéaire parfaite négative, ce qui indique que les deux variables sont parfaitement corrélées de manière négative, c'est-à-dire qu'elles diminuent ensemble de manière linéaire.
- Si $\rho_{(X,Y)} = 0$, cela signifie qu'il n'y a aucune corrélation linéaire entre les deux variables.

Le coefficient de corrélation de Pearson est une mesure puissante pour quantifier les relations linéaires entre deux variables continues. Cependant, il présente certaines limites et précautions à prendre en compte :

- **Linearité requise** : Le coefficient de Pearson suppose une relation linéaire entre les variables. Il peut ne pas détecter des relations non linéaires, même si elles existent. Si la relation entre les variables est courbe ou complexe, Pearson peut sous-estimer la corrélation.
- **Sensibilité aux valeurs aberrantes** : Le coefficient de Pearson est sensible aux valeurs aberrantes. Une valeur extrême peut considérablement influencer la corrélation. Par conséquent, il est essentiel de vérifier la présence de valeurs aberrantes et de décider de les inclure ou de les exclure judicieusement de l'analyse.

- **Distribution requise** : Il suppose que les variables suivent une distribution normale ou approximativement normale. Si les données ne sont pas normalement distribuées, cela peut biaiser les résultats.
- **Non-robustesse** : Le coefficient de Pearson n'est pas robuste aux violations de ses hypothèses. Si les données ne satisfont pas les conditions requises, il peut conduire à des estimations incorrectes de la corrélation.

- **Effet de taille d'échantillon** : Pour de petits échantillons, le coefficient de Pearson peut être moins fiable pour estimer la corrélation réelle dans la population.
- **Ne mesure que les relations linéaires** : Comme mentionné précédemment, Pearson ne détecte que les relations linéaires. Si la relation entre les variables est courbe, sinusoïdale, ou tout autre modèle non linéaire, Pearson ne sera pas approprié.
- **Absence de causalité** : La corrélation n'implique pas la causalité. Une forte corrélation entre deux variables ne signifie pas nécessairement que l'une cause l'autre. Cela peut être dû à une troisième variable non identifiée.
- **Échantillonnage biaisé** : Si l'échantillonnage n'est pas représentatif de la population sous-jacente, les résultats peuvent être biaisés.
- **Dépendance du domaine** : L'interprétation de la force de la corrélation dépend du domaine d'application. Une corrélation modérée peut être significative dans un contexte, mais moins importante dans un autre.

② **Le coefficient de corrélation de rang de Spearman** : Le coefficient de corrélation de rang de Spearman est une mesure statistique de la dépendance monotone entre deux variables continues ou ordinales. Contrairement au coefficient de corrélation de Pearson, qui évalue la dépendance linéaire, Spearman se concentre sur la dépendance monotonique, ce qui signifie qu'il détecte également des relations non linéaires si elles sont monotoniques. Il est nommé d'après le statisticien britannique Charles Spearman.

La formule mathématique du coefficient de corrélation de rang de Spearman, notée ρ, est la suivante :

$$\rho = 1 - \frac{6 \sum d_i^2}{n\left(n^2 - 1\right)} \tag{3.14}$$

Où :

ρ est le coefficient de corrélation de rang de Spearman.

d_i est la différence entre les rangs des paires d'observations appariées pour les deux variables :

$$r(X_i) - r(Y_i)$$

n est la taille de l'échantillon.

Le coefficient de Spearman varie de -1 à 1, où :

- ρ = 1 indique une corrélation parfaite positive monotonique, ce qui signifie que lorsque les valeurs d'une variable augmentent, les valeurs de l'autre variable augmentent également de manière monotone.
- $\rho = -1$ indique une corrélation parfaite négative monotonique, ce qui signifie que lorsque les valeurs d'une variable augmentent, les valeurs de l'autre variable diminuent de manière monotone.
- ρ = 0 indique qu'il n'y a pas de corrélation monotone entre les deux variables.

Le coefficient de corrélation de rang de Spearman est non paramétrique, ce qui signifie qu'il ne repose pas sur des hypothèses spécifiques sur la distribution des données. Il est souvent utilisé lorsque les données ne satisfont pas les conditions d'application du coefficient de corrélation de Pearson ou lorsque l'on souhaite évaluer des relations non linéaires.

```r
# Étape 1 : Créer les données
quantite_nutriments <- c(40, 60, 45, 35, 70, 30, 75, 68, 33, 55)
rendement_mais <- c(2500, 2800, 2600, 2400, 3200, 2300, 3300, 3150, 2450, 2750)

# Étape 2 : Calculer le coefficient de corrélation linéaire de Bravais-Pearson
cor_pearson <- cor(quantite_nutriments, rendement_mais, method = "pearson")
print(paste("Le coefficient de corrélation de Pearson est :", cor_pearson))

# Étape 3 : Calculer le coefficient de corrélation de rang de Spearman
```

```r
cor_spearman <- cor(quantite_nutriments, rendement_mais, method = "spearman")
print(paste("Le coefficient de corr lation de Spearman est :", cor_spearman))
```

Listing 3.12 – Script R pour Le calcul des coefficients de correlation

```
[1] "Le coefficient de corr lation de Pearson est : 0.955687009213691"
[1] "Le coefficient de corr lation de Spearman est : 0.963636363636364"
```

Listing 3.13 – Le calcul des coefficients de correlation

Interprétation des résultats :

— Coefficient de corrélation de Pearson (0.956) : Il y a une forte corrélation positive linéaire entre la quantité de nutriments et le rendement en maïs. Cela signifie que, généralement, à mesure que la quantité de nutriments augmente, le rendement en maïs augmente également de manière linéaire.

— Coefficient de corrélation de Spearman (0.964) : Il y a une forte corrélation positive monotone entre la quantité de nutriments et le rendement en maïs. Ce coefficient est particulièrement utile si la relation entre les variables n'est pas strictement linéaire mais monotone (c'est-à-dire toujours croissante ou toujours décroissante).

3.3 Correlation et causalité

La corrélation est une mesure statistique qui quantifie la relation entre deux variables, indiquant comment elles évoluent conjointement. Cependant, il est crucial de comprendre que la corrélation ne détermine pas la causalité. En d'autres termes, le fait qu'il existe une corrélation entre deux variables ne signifie pas nécessairement qu'une variable en cause provoque l'autre :

i. **Corrélation ne signifie pas Causalité** : La présence d'une corrélation entre deux variables X et Y indique simplement qu'elles varient ensemble, mais cela ne révèle pas la nature de la relation. Les deux variables peuvent être liées par une troisième variable non observée qui les influence toutes les deux, ou il peut s'agir d'une coïncidence.

ii. **Troisième Variable Confondante** : Parfois, une troisième variable, appelée variable confondante, peut influencer à la fois X et Y, créant ainsi une corrélation apparente. Il est important d'examiner attentivement les données pour identifier de telles variables confondantes.

iii. **Direction de la Causalité** : La corrélation seule ne permet pas de déterminer la direction de la causalité. Dans de nombreux cas, il est difficile de savoir si X cause Y, si Y cause X, ou si une autre variable non observée influence les deux.

iv. **Relation Bidirectionnelle** : Il existe des situations où X et Y s'influencent mutuellement, créant une relation bidirectionnelle. Par exemple, la corrélation entre l'exercice physique et la santé cardiovasculaire est bidirectionnelle, car l'exercice peut améliorer la santé cardiovasculaire, mais une meilleure santé cardiovasculaire peut également encourager l'exercice.

v. **Causalité vs. Corrélation** : Pour établir la causalité, des méthodes de recherche spécifiques, telles que les expériences contrôlées, sont nécessaires. Ces méthodes visent à démontrer que la modification d'une variable (X) entraîne un changement dans une autre variable (Y) tout en éliminant d'autres explications possibles.

vi. **Prudence dans l'Interprétation** : Lorsqu'une corrélation est observée, il est important d'exercer la prudence dans l'interprétation des résultats. Il peut être utile de considérer la plausibilité biologique ou théorique de la causalité, mais cela ne suffit pas pour prouver la causalité.

Le paradoxe de la Grenouille d'Albert Simon est une histoire intrigante qui met en évidence la distinction entre corrélation et causalité, ainsi que les pièges potentiels de l'interprétation des données. L'histoire va comme suit :

Imaginez que vous êtes un chercheur qui observe deux phénomènes dans un marais : le croassement des grenouilles et le nombre de moustiques. Vous recueillez des données pendant plusieurs semaines et constatez une forte corrélation entre le croassement des grenouilles et le nombre de moustiques. Chaque fois que le croassement des grenouilles augmente, le nombre de moustiques diminue, et vice versa.

Vous pourriez conclure que le croassement des grenouilles semble être un excellent moyen de réduire le nombre de moustiques. Vous pourriez même conseiller aux habitants du marais d'élever davantage de grenouilles pour lutter contre les moustiques.

Cependant, voici le paradoxe : Albert Simon, un autre chercheur, décide de mener une expérience. Il place des enregistreurs de son dans le marais pour surveiller le croassement des grenouilles et compte le nombre de moustiques. Mais au lieu d'observer une réduction des moustiques lorsque le croassement des grenouilles augmente, il découvre que les moustiques sont en réalité attirés par le bruit des grenouilles. Plus il y a de croassements, plus il y a de moustiques.

Le paradoxe de la Grenouille d'Albert Simon met en évidence le danger de tirer des conclusions hâtives sur la causalité à partir de la seule corrélation. Dans cet exemple, la corrélation positive entre le croassement des grenouilles et le nombre de moustiques ne signifie pas que le croassement des grenouilles cause une diminution des moustiques. Au contraire, la relation est inverse : plus de croassements attirent plus de moustiques.

L'exemple de Grenouille d'Albert Simon illustre pourquoi il est essentiel de mener des expériences contrôlées et d'examiner attentivement les mécanismes sous-jacents pour comprendre la vraie nature des relations entre les variables. Se fier uniquement à la corrélation peut conduire à des conclusions trompeuses et à des actions inappropriées. Dans la recherche scientifique, il est crucial de ne pas confondre la corrélation avec la causalité et d'approfondir notre compréhension des relations entre les variables.

Bibliographie

HOFF, Katharina J, LA HOTHORN et Universitetslektor J-E ENGLUND (2005). « R-Manual for Biometry ». In : *Univeristy of Hannover, Bachelor Thesis*.

MONTGOMERY, Douglas C, Elizabeth A PECK et G Geoffrey VINING (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

STEEL, Robert G d et James H TORRIE (1986). *Principles and procedures of statistics : a biometrical approach*. McGraw-Hill New York, NY, USA.