République Algérienne Démocratique et populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université 20 aout 1955 – Skikda

Cours

Statistiques Expérimentales

3 année Licence (LMD)

Filière : sciences agronomiques

Spécialité : Amélioration de la production animale

<u>Dr. C. Benatmane</u>

2023/2024

Objectifs

L'étudiant approfondira ses connaissances dans le traitement des résultats expérimentaux en se concentrant sur plusieurs aspects clés :

- Explorer des techniques de modélisation, comme la régression, pour établir des relations entre les variables et prédire les résultats futurs.
- Apprendre à utiliser des outils de visualisation pour représenter les données de manière claire et informative, facilitant ainsi la communication des résultats.
- Développer la capacité à interpréter les résultats statistiques, en comprenant les intervalles de confiance et les implications pratiques des analyses.
- Apprendre à appliquer différentes méthodes statistiques, telles que les tests t et l'ANOVA, pour évaluer les différences entre les groupes et déterminer la signification des résultats.

Contenu de la matière

1. Introduction générale

Les **statistiques expérimentales** jouent un rôle essentiel dans la conception et l'analyse des expériences agricoles. Elles permettent de garantir la validité des résultats et d'interpréter les données de manière rigoureuse. Voici les principaux concepts et méthodes associés aux statistiques expérimentales dans le contexte de l'agriculture :

1.1 Conception expérimentale

- **Essais randomisés** : Chaque traitement est attribué de manière aléatoire pour éviter les biais.
- **Blocs randomisés** : Les essais sont organisés en blocs pour contrôler les variations non contrôlées (par exemple, la qualité du sol).
- **Plans factoriels** : Étudier plusieurs facteurs simultanément pour observer les interactions entre eux.

1.2. Échantillonnage.

- Choix de méthodes d'échantillonnage (aléatoire, stratifié, systématique) pour obtenir des données représentatives de la population cible.
- Détermination de la taille d'échantillon nécessaire pour assurer une puissance statistique suffisante.

1.3. Analyse des données

- **Tests d'hypothèses** : Utilisation de tests statistiques pour évaluer si les différences observées entre les groupes sont significatives (tests t, ANOVA).
- **Analyse de la variance (ANOVA)** : Comparaison de plusieurs groupes pour déterminer s'il existe des différences significatives.
- **Régression linéaire** : Modélisation de la relation entre une variable dépendante et une ou plusieurs variables indépendantes.

1.4. Interpretation des résultats

- Compréhension des concepts tels que la valeur p, les intervalles de confiance, et les erreurs de type I et II.
- Évaluation de l'importance pratique des résultats statistiques, en tenant compte de leur signification agronomique.

1.5. Visualisation des données

- Utilisation de graphiques (histogrammes, box plots, diagrammes de dispersion) pour représenter visuellement les résultats et faciliter l'interprétation.

1.6. Utilisation de logiciels statistiques

- Familiarisation avec des outils comme R, SAS ou SPSS pour effectuer des analyses statistiques et gérer des bases de données.

1.7.Applications dans l'agriculture

Les statistiques expérimentales sont essentielles pour :

- Évaluer l'impact de différents traitements sur les rendements.
- Comparer l'efficacité des variétés de cultures ou des pratiques agricoles.
- Prédire les résultats en fonction des variables manipulées.

Les statistiques expérimentales fournissent les outils nécessaires pour concevoir, analyser et interpréter les expériences agricoles, contribuant ainsi à des pratiques de recherche rigoureuses et basées sur des données probantes.

2 Introduction à la statistique descriptive

2.1 Statistique descriptive

La **statistique descriptive** est une branche de la statistique qui se concentre sur la collecte, l'organisation, le résumé et la présentation des données. Contrairement à la statistique inférentielle, qui vise à tirer des conclusions sur une population à partir d'un échantillon, la statistique descriptive sert à décrire et analyser des ensembles de données en fournissant des informations claires et synthétiques.

2.2 Objectifs de la statistique descriptive

L'objectif principal de la statistique descriptive est de rendre les données plus compréhensibles en :

- **Résumant les informations** à l'aide de mesures statistiques simples.
- **Présentant les données** sous forme de tableaux, de graphiques ou de diagrammes.
- **Identifiant les tendances principales** au sein des données, comme la moyenne, la dispersion, etc.

2.3. Importance de la statistique descriptive

La statistique descriptive est essentielle car elle permet de :

- **Simplifier l'analyse de grandes quantités de données** en les résumant de manière concise.
- **Identifier des patterns ou tendances** qui peuvent ensuite orienter des analyses plus approfondies.
- **Faciliter la prise de décision** dans de nombreux domaines, tels que les sciences, l'ingénierie, l'économie et l'agriculture.

2.3 Types de mesures en statistique descriptive

2.3.1. Mesures de tendance centrale :

Moyenne :

La **moyenne**, également appelée **moyenne arithmétique**, est une mesure statistique qui résume un ensemble de données numériques en un seul nombre. Elle représente la tendance centrale des données, permettant d'avoir une idée de la valeur "typique" d'un ensemble. C'est l'une des mesures les plus couramment utilisées en statistiques.

Calcul de la Moyenne

Pour calculer la moyenne d'un ensemble de valeurs, suivez ces étapes :

1. **Additionnez toutes les valeurs** de l'ensemble de données.
2. **Divisez la somme obtenue** par le nombre total de valeurs dans l'ensemble.

Formule de la moyenne :

$$m = \frac{\sum_1^k n_i x_i}{N}$$

Où :

- x_i = chaque valeur individuelle de l'ensemble de données
- N = nombre total de valeurs

Si vous ayez les rendements de cinq champs en tonnes par hectare : 10, 12, 15, 20, et 18 t/ha.

$$m = \frac{10 + 12 + 15 + 20 + 18}{5} = 15t/ha$$

Dans le cadre d'une étude sur les rendements agricoles, la moyenne est souvent utilisée pour donner un aperçu général de la production des champs. Elle permet aux agriculteurs et aux chercheurs de :

- Évaluer les performances générales des cultures.
- Comparer les rendements entre différentes variétés de cultures.
- Prendre des décisions basées sur des données quantitatives.

Médiane :

La valeur qui sépare la moitié supérieure des données de la moitié inférieure. Elle est utile lorsque les données contiennent des valeurs extrêmes ou des outliers.

La **médiane** est une mesure de tendance centrale qui représente la valeur qui sépare un ensemble de données en deux parties égales. Elle est particulièrement utile pour décrire des données, car elle est moins influencée par les valeurs extrêmes (outliers) que la moyenne. La médiane est souvent utilisée dans des analyses statistiques pour des ensembles de données asymétriques.

Calcul de la Médiane

Le calcul de la médiane varie selon que le nombre d'observations est impair ou pair.

1. **Données avec un nombre impair d'observations** :
   - Rangez les valeurs dans l'ordre croissant.
   - La médiane est la valeur du milieu.
2. **Données avec un nombre pair d'observations** :
   - Rangez les valeurs dans l'ordre croissant.
   - La médiane est la moyenne des deux valeurs du milieu.

Dans le cadre de l'agriculture, la médiane peut être utilisée pour :

- Analyser les rendements des cultures, surtout si les données contiennent des valeurs extrêmes (par exemple, un champ exceptionnellement productif ou peu productif).
- Comparer les rendements entre différentes variétés de cultures de manière à obtenir une estimation plus précise de la performance "typique".

- Évaluer la distribution des rendements dans différentes conditions de culture.

Mode :

La valeur qui apparaît le plus souvent dans un ensemble de données. C'est une mesure utile pour les variables qualitatives ou pour détecter les pics dans les données.

Le **mode** est une mesure de tendance centrale qui représente la valeur la plus fréquemment observée dans un ensemble de données. Contrairement à la moyenne et à la médiane, le mode peut être utilisé pour des données qualitatives ou quantitatives. Un ensemble de données peut avoir un seul mode (unimodal), plusieurs modes (multimodal) ou aucun mode (si toutes les valeurs sont uniques).

Quartiles :

Les quartiles sont des mesures statistiques qui divisent un ensemble de données en quatre parties égales. Chaque quartile représente 25 % de l'ensemble des données. Ils sont particulièrement utiles pour analyser la dispersion et la distribution des données, en identifiant les valeurs extrêmes et les tendances centrales. Les quartiles se notent généralement comme suit:

- Q1 : Le premier quartile, qui correspond à 25 % des données.
- Q2 : Le deuxième quartile, qui est également la médiane (50 % des données).
- Q3 : Le troisième quartile, qui correspond à 75 % des données.

Exemple :

Un agriculteur mesure la hauteur de 12 plants de blé après trois mois. Voici les données (en cm) :

50, 55, 58, 60, 62, 65, 70, 72, 75, 77, 80, 82

1- Calculez les quartiles Q1, Q2, et Q3.

2- Que peuvent indiquer ces quartiles sur la croissance des plants de blé ?

2.3.2. Mesures de dispersion :

Écart-type :

Une mesure qui quantifie la variation ou la dispersion des valeurs autour de la moyenne. Plus l'écart-type est élevé, plus les données sont dispersées.

L'écart-type, noté généralement **σ** (sigma) pour une population et **s** pour un échantillon, est la **racine carrée de la variance**. Il mesure à quel point les valeurs d'un ensemble de données s'éloignent, en moyenne, de la valeur moyenne.

Formule de l'écart-type pour un échantillon :

$$\mathbf{s} = \sqrt{\frac{\sum_{1}^{k} \mathbf{n_i}(\mathbf{x_i} - \bar{\mathbf{x}})^2}{\mathbf{n}}}$$

Formule de l'écart-type pour une population:

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum_{1}^{k} \mathbf{n_i}(\mathbf{x_i} - \mathbf{m})^2}{\mathbf{n}}}$$

Variance :

La moyenne des carrés des écarts par rapport à la moyenne. C'est une mesure de la dispersion similaire à l'écart-type mais exprimée au carré.

La variance est particulièrement utile dans le cadre de l'**amélioration de la production animale** et des **sciences agronomiques** pour évaluer la variabilité dans les mesures importantes, telles que le poids des animaux, la production laitière, ou les rendements agricoles. Elle permet de comprendre l'homogénéité ou l'hétérogénéité des résultats et d'identifier des situations nécessitant des ajustements.

Exemple

Un agronome suit la croissance en hauteur (en centimètres) de cinq plants de maïs après trois semaines de croissance. Les mesures sont les suivantes :

150, 155, 160, 165, 170

3- Calculez la moyenne de la hauteur des plants.
4- Utilisez la formule de la variance pour évaluer la dispersion des hauteurs des plants autour de la moyenne.
5- Que signifie la variance obtenue ? Une faible variance indiquerait-elle que les plants ont poussé de manière homogène ?

Étendue :

La différence entre la valeur la plus élevée et la plus basse d'un ensemble de données.

L'**étendue** est une mesure de dispersion simple en statistique descriptive. Elle indique la différence entre la plus grande et la plus petite valeur dans un ensemble de données. L'étendue donne une idée générale de la **variabilité des données**, mais elle est très sensible aux valeurs extrêmes et ne prend pas en compte la distribution des autres données.

$$\text{Etendue} = \text{Valeur maximale} - \text{Valeur minimale}$$

Un agriculteur surveille les températures quotidiennes à l'intérieur d'une serre pendant une semaine. Les températures mesurées (en °C) sont :

Températures (en °C) :22, 24, 25, 23, 27, 21, 26.

1. Calculez l'étendue des températures.
2. Que révèle l'étendue sur les conditions climatiques à l'intérieur de la serre ?

2.4 Représentation graphique

La statistique descriptive utilise aussi divers outils graphiques pour illustrer les données, tels que :

2.4.1 Histogrammes :

Pour représenter la distribution des fréquences d'un ensemble de données.

Un **histogramme** est un type de graphique utilisé en **statistique descriptive** pour représenter la distribution des données numériques continues. Il permet de visualiser la fréquence de chaque classe de valeurs dans un ensemble de données, en affichant des barres verticales dont la hauteur correspond au nombre d'observations (ou fréquence) dans chaque intervalle.

Caractéristiques d'un histogramme.

1. **Axes** :
   - **Axe des abscisses (horizontal)** : Il représente les **intervalles de valeurs** ou les **classes** (appelées "bins") dans lesquelles les données sont regroupées.
   - **Axe des ordonnées (vertical)** : Il représente la **fréquence** ou l'**effectif** des données dans chaque intervalle.
2. **Barres** :
   - Les barres dans un histogramme sont **contiguës** (sans espace entre elles), car elles représentent des données continues.
   - La **hauteur** des barres indique la fréquence d'observations dans chaque intervalle.
   - La **largeur** des barres est la même pour tous les intervalles, bien que les limites des classes puissent varier en fonction des besoins.
3. **Classes ou intervalles** :
   - Un histogramme regroupe les données en **classes**, qui sont des intervalles de valeurs numériques.

- Le nombre de classes peut influencer l'apparence de l'histogramme : trop peu de classes peuvent masquer des détails importants, tandis que trop de classes peuvent rendre le graphique difficile à interpréter.

Exemple d'utilisation d'un histogramme

Si vous avez des données sur la taille des plantes dans une expérience agricole, un histogramme pourrait regrouper les plantes en fonction de leur taille en plusieurs intervalles, par exemple :

0-10 cm, 10-20 cm, 20-30 cm, etc. La hauteur des barres montrera combien de plantes se trouvent dans chaque intervalle de taille.

Avantage des histogrammes :

Visualisation rapide : Un histogramme permet d'identifier facilement la distribution des données, comme la tendance centrale, la dispersion, et la forme (symétrique, asymétrique, multimodale).

Détection des outliers : Les valeurs qui sortent de l'ordinaire, comme les données extrêmes, sont faciles à repérer.

Facilité d'interprétation : Ce type de graphique est simple à lire et est très utilisé dans les analyses descriptives pour présenter les données d'une manière claire.

2.4.2. Diagrammes en bâtons

Les **diagrammes en bâtons** (ou **diagrammes à barres**) sont des graphiques utilisés en **statistique descriptive** pour représenter des données **discrètes** ou **catégorielles**. Ils permettent de comparer visuellement des groupes ou des catégories en représentant chaque catégorie par une barre, dont la longueur ou la hauteur est proportionnelle à sa fréquence ou à sa valeur.

Caractéristiques des diagrammes en bâtons

1. **Axes** :
   - **Axe des abscisses (horizontal)** : Représente les **catégories** ou les groupes distincts (par exemple, différents produits, types de cultures, ou classes de données).
   - **Axe des ordonnées (vertical)** : Représente la **valeur** ou la **fréquence** associée à chaque catégorie.
2. **Barres** :
   - Les barres peuvent être **verticales** ou **horizontales**.
   - Chaque barre représente une catégorie, et sa **longueur** ou **hauteur** est proportionnelle à la valeur ou à la fréquence de cette catégorie.
   - Contrairement aux **histogrammes**, les barres sont **séparées** les unes des autres, car les données sont discrètes, sans continuité entre les catégories.

Types de diagrammes en bâtons

1. **Diagramme en bâtons simple** :
   - Chaque barre représente une catégorie unique. Par exemple, si vous analysez les rendements de différentes cultures (blé, maïs, riz), chaque culture aura une barre indiquant son rendement en tonnes par hectare.
2. **Diagramme en bâtons groupé** :
   - Utilisé pour comparer plusieurs ensembles de données au sein de chaque catégorie. Par exemple, comparer les rendements de plusieurs cultures (blé, maïs, riz) sur plusieurs années (2021, 2022, 2023). Chaque groupe de catégories aura plusieurs barres côte à côte pour représenter les différentes périodes.
3. **Diagramme en bâtons empilé** :
   - Chaque barre est divisée en segments représentant plusieurs sous-catégories. Par exemple, un diagramme pourrait représenter le rendement global de chaque culture, avec des segments dans chaque barre indiquant les contributions de différentes variétés de chaque culture.

Exemple d'utilisation d'un diagramme en bâtons

Dans une étude agricole, vous pourriez vouloir comparer le rendement de différentes cultures. Un diagramme en bâtons montrerait la fréquence ou la valeur de chaque culture :

- Catégories : Blé, Maïs, Riz.
- Valeur sur l'axe vertical : Rendement en tonnes par hectare. Chaque barre montrerait le rendement pour chaque culture, permettant de visualiser les différences entre elles.

2.4.3.: Diagrammes circulaires (camemberts)

Les **diagrammes circulaires**, également appelés **camemberts** ou **diagrammes en secteurs**, sont des outils visuels utilisés en statistique pour représenter la répartition des données en différentes catégories, sous forme de parts d'un cercle. Chaque secteur du cercle représente une catégorie, et la taille de chaque secteur est proportionnelle à la part qu'elle occupe dans l'ensemble des données.

En **sciences agronomiques**, plus spécifiquement dans le domaine de l'**amélioration de la production animale**, les **diagrammes circulaires** peuvent être utilisés pour visualiser la répartition et l'importance de divers facteurs qui influencent la productivité. Cette méthode de représentation est particulièrement utile pour montrer comment certaines variables se partagent un ensemble global, facilitant ainsi l'interprétation des données en termes de proportion.

Exemple : Dans une ferme polyculture-élevage, il est courant de produire plusieurs types de produits animaux comme le lait, la viande, et les œufs. Un diagramme circulaire peut représenter la répartition des différents types de production dans le chiffre d'affaires global de l'exploitation.

Lait : 50%
Viande : 30%
Œufs : 15%

Autres produits (ex. : laine, peaux) : 5%

2.4.3 Diagrammes en boîte (boxplots) :

Les **diagrammes en boîte** (ou **boxplots**) sont des outils statistiques puissants qui permettent de visualiser la distribution d'un ensemble de données en résumé. Ils montrent la médiane, les quartiles, et les valeurs aberrantes, et sont très utiles pour comparer plusieurs distributions. Dans le domaine des **sciences agronomiques** et de l'**amélioration de la production animale**, les boxplots peuvent être utilisés pour visualiser la variabilité des paramètres de production, comme le poids des animaux, la production laitière, ou la quantité de nourriture consommée.

Un diagramme en boîte est constitué des éléments suivants :

1. Médiane (ligne centrale) : Représente la valeur médiane des données (50e percentile).
2. Quartile inférieur (Q1) : Représente la valeur en dessous de laquelle se trouvent 25 % des données.
3. Quartile supérieur (Q3) : Représente la valeur en dessous de laquelle se trouvent 75 % des données.
4. Étendue interquartile (IQR) : Distance entre le premier quartile (Q1) et le troisième quartile (Q3). Elle mesure la dispersion des données centrales.
5. "Whiskers" (Moustaches) : Elles s'étendent généralement jusqu'à 1,5 fois l'IQR. Au-delà, on trouve des valeurs potentiellement aberrantes.
6. Valeurs aberrantes (outliers) : Points situés en dehors des "whiskers", ce qui signifie qu'ils sont très éloignés de la majorité des données.

2.5. Exercices

Exercice 1 : Comparaison de rendements agricoles

Un agriculteur cultive quatre types de céréales : blé, maïs, riz et avoine. Il a obtenu les rendements suivants (en tonnes par hectare) :

- Blé : 6 t/ha
- Maïs : 9 t/ha
- Riz : 5 t/ha
- Avoine : 4 t/ha

1. Représentez ces données à l'aide d'un **diagramme en bâtons**.
2. Comparez les rendements entre les différentes cultures et identifiez celle qui a eu le meilleur rendement.
3. Quel est l'écart entre le rendement le plus faible et le rendement le plus élevé ?

Exercice 2 production laitière annuelle

Vous avez les données suivantes sur la production laitière annuelle (en litres) de cinq fermes dans une région donnée :

- **Ferme A** : 5000 litres
- **Ferme B** : 8000 litres
- **Ferme C** : 4000 litres
- **Ferme D** : 6000 litres
- **Ferme E** : 7000 litres

1. Représentez ces données sous forme d’un diagramme en bâtons.
2. Quelle ferme a la plus grande production laitière ? Laquelle a la plus faible ?
3. Quelle est la différence de production entre la ferme qui produit le plus et celle qui produit le moins ?

4. Quels enseignements tirez-vous de cette représentation en bâtons concernant la répartition de la production laitière dans la région ?

Exercice 3: Distribution des rendements

Les rendements de 25 champs agricoles (en tonnes par hectare) sont enregistrés comme suit :

2, 5, 6, 8, 10, 11, 12, 15, 17, 18, 20, 21, 22, 23, 25, 26, 28, 30, 32, 33, 35, 36, 37, 40, 42

1. Divisez les rendements en classes d'intervalle de 10.
2. Construisez un histogramme des rendements.
3. Quelle est la classe avec le rendement le plus fréquent ?
4. Déterminez l'étendue des rendements.

Exercice 4 : Le poids moyen des veaux

Le poids moyen des veaux dans une ferme a été mesuré et regroupé dans les catégories suivantes (en kg) :

- **40-50 kg** : 5 veaux
- **50-60 kg** : 12 veaux
- **60-70 kg** : 8 veaux
- **70-80 kg** : 10 veaux
- **80-90 kg** : 3 veaux

1. Représentez ces données sous forme d'un histogramme.
2. Quelle est la classe de poids la plus représentée ? Et la moins représentée ?
3. Quelle est la différence entre le nombre de veaux dans la classe 50-60 kg et celle de 80-90 kg ?
4. Que révèle cet histogramme sur la répartition des poids des veaux dans la ferme ? Y a-t-il une tendance vers des veaux plus lourds ou plus légers ?
5. Si vous prévoyez une augmentation de la moyenne des poids dans l'élevage, à quelles classes l'augmentation de veaux serait-elle probablement concentrée ?

Exercice 5 : Temps de Croissance des Plantes

Dans une étude sur la croissance des plantes, un chercheur a mesuré le temps (en jours) que prennent plusieurs variétés de maïs pour atteindre la maturité :

80, 75, 82, 78, 85, 90

1. Trouvez la médiane du temps de croissance.
2. Quel est l'impact d'une variété ayant une maturation particulièrement rapide sur la médiane ?

Exercice 6 : Teneur en Sucres dans les Fruits

Un laboratoire a mesuré la teneur en sucre (en grammes par 100g) de différents fruits. Voici les résultats pour 8 fruits:

12, 14, 9, 11, 15, 13, 10, 12

1. Calculez la moyenne de la teneur en sucre des fruits.
2. Calculez la médiane de la teneur en sucre.
3. Déterminez le mode des valeurs de sucre.
4. Que représentent ces trois mesures dans l'analyse nutritionnelle des fruits ?

Exercice 7 :

Dans une ferme spécialisée dans l'élevage de bovins, ovins, caprins et volailles, les propriétaires cherchent à mieux comprendre la répartition de leurs ressources et de leur production pour améliorer la gestion de l'exploitation. Les données suivantes ont été collectées sur la répartition de la production animale (en tonnes par an) et des coûts d’alimentation pour chaque type d'animal.

Données de production annuelle (en tonnes) :

Bovins : 40 tonnes
Ovins : 25 tonnes

Caprins : 15 tonnes
Volailles : 20 tonnes

Données de coûts d'alimentation (en euros) :

- **Bovins** : 40,000 €
- **Ovins** : 20,000 €
- **Caprins** : 15,000 €
- **Volailles** : 25,000 €

    - Représentez les données de **production animale annuelle** sous forme d'un **diagramme circulaire**. Chaque secteur du diagramme doit correspondre à un type d'animal, et la taille de chaque secteur doit être proportionnelle à la quantité produite.
    - Représentez ensuite les **coûts d'alimentation** sous forme d'un **deuxième diagramme circulaire**, avec chaque secteur représentant un type d'animal et les coûts correspondants.
    - Quelle est l'espèce animale qui contribue le plus à la production totale dans la ferme ?
    - Quelle espèce représente le plus grand coût d'alimentation ?
    - Comparez la proportion de production de chaque espèce à sa part de coûts d'alimentation. Y a-t-il un déséquilibre entre la production et les coûts pour une espèce ?
    - À partir des deux diagrammes circulaires, quelles actions recommanderiez-vous pour optimiser la rentabilité de la ferme ? Par exemple, certaines espèces coûtent-elles plus cher à nourrir que ce qu'elles rapportent en production ?
    - Proposez des solutions pour réduire les coûts ou augmenter la production d'une espèce animale spécifique, en vous basant sur l'analyse des données.

Exercice 8 : pH de Différentes Boissons

Un chimiste a mesuré le pH de diverses boissons pour analyser leur acidité :

3.2, 4.5, 3.8, 3.0, 4.0, 3.5, 4.2, 3.3

1. Calculez la moyenne des valeurs de pH.
2. Calculez la médiane des valeurs de pH.
3. Déterminez le mode des valeurs de pH.
4. Que peut-on conclure sur l'acidité moyenne des boissons à partir de ces mesures ?

Exercice 9 : Durée de Conservation des Produits Laitiers

Un fabricant de produits laitiers a enregistré la durée de conservation (en jours) de 7 lots de yaourts :

12, 15, 14, 10, 17, 13, 12

1. Calculez la moyenne de la durée de conservation.
2. Calculez la médiane de la durée de conservation.
3. Déterminez le mode de la durée de conservation.
4. Comment ces mesures peuvent-elles être utilisées pour améliorer les conditions de stockage des produits ?

Exercice 10:

Vous avez collecté des données sur la production laitière quotidienne (en litres) de vaches laitières dans trois fermes différentes (Ferme A, Ferme B, Ferme C) sur une semaine.

- **Ferme A** : [20, 22, 21, 23, 25, 24, 26]
- **Ferme B** : [18, 19, 18, 17, 20, 21, 18]
- **Ferme C** : [25, 30, 29, 27, 31, 28, 26]

  - Tracez un **diagramme en boîte** pour chaque ferme afin de visualiser la répartition de la production laitière.
  - Quelle ferme a la production laitière la plus homogène ? Comment pouvez-vous le déterminer à partir des box-plots ?

- Quelle est la ferme avec la plus grande variabilité dans la production laitière ? Que signifie cette variabilité en termes de gestion de l'exploitation ?
- Y a-t-il des valeurs aberrantes (outliers) ? Si oui, dans quelle ferme apparaissent-elles ?
- Comparez la médiane de la production laitière des trois fermes. Quelle ferme a la production laitière médiane la plus élevée ?
- En fonction de ces résultats, quelle ferme semble être la plus performante en termes de production régulière et élevée ? Quelles recommandations feriez-vous pour améliorer la production dans les autres fermes ?

Exercice 11 : Analyse de la Variabilité du Poids des Porcs

Un fermier mesure le poids de sept porcs d'engraissement (en kg) avant la vente. Les poids mesurés sont: 90, 95, 100, 85, 92, 97, 88

1. Calculez l'écart-type des poids.
2. Que révèle l'écart-type sur l'homogénéité des porcs ?

Exercice 12

Un biologiste effectue des mesures de température (en °C) dans un habitat aquatique pendant une semaine. Voici les données :

15, 16, 15, 14, 20, 19, 22, 21, 23, 18

1. Calculez la **moyenne**, la **médiane** et le **mode** des températures.
2. Calculez l'**étendue**, la **variance** et l'**écart-type**.

3. Régression linéaire

La régression linéaire est une méthode statistique utilisée pour modéliser la relation entre une variable dépendante (ou réponse) et une ou plusieurs variables indépendantes (ou prédicteurs). Elle est particulièrement utile en sciences agricoles et biologiques pour analyser comment différentes variables influencent un résultat spécifique.

3.1 Modèle de régression linéaire simple

Le modèle de régression linéaire simple peut être exprimé par l'équation :

$$y = bx + a + \epsilon$$

Où :

- y: variable dépendante
- a: ordonnée à l'origine (intercept)
- b: coefficient de régression (pente)
- x: variable indépendante
- ϵ: terme d'erreur

Estimation des Coefficients :

Les coefficients a et b sont estimés à l'aide de la méthode des moindres carrés, qui minimise la somme des carrés des résidus (écarts entre les valeurs observées et prédites).

- Le coefficient b indique la variation de y pour chaque unité de changement dans$x$.
- Un coefficient positif indique une relation directe, tandis qu'un coefficient négatif indique une relation inverse.

Évaluation du Modèle :

On utilise des mesures telles que le coefficient de détermination R^2 pour évaluer la qualité du modèle. Un R^2 proche de 1 indique que le modèle explique une grande proportion de la variance dans la variable dépendante.

Calcul des Coefficients

Le coefficient b (pente) est calculé à l'aide de la formule suivante :

$$b = \frac{Cov(x, y)}{\sigma_x^2}$$

Une fois que b a été calculé, l'intercept aaa peut être déterminé par la formule suivante

$$a = \bar{y} - a\bar{x}$$

Étapes de Calcul

i	(Xi)	(Yi)	$(Xi - mX)$	$(Yi - mY)$	$(Xi - mX)(Yi - mY)$
1					
2					
3					
4					
5					
6					
7					
8					
moyenne	mX	mY =	0	0	0
écart-type			-	-	-

3.2 Exercice

Exercice 1

Vous avez collecté des données sur l'effet d'un pesticide sur le nombre de parasites dans un champ.

Pesticide (X)	Nombre de Parasites (Y)
0	50
1	35
2	25
3	15

1-Établissez un modèle de régression linéaire pour prédire le nombre de parasites en fonction de la quantité de pesticide utilisée.

2-Interprétez les coefficients du modèle.

3-Calculez R^2 et discutez de la qualité du modèle.

Exercice 2 Relation entre le Nombre de Semis et le Rendement des Céréales

Nombre de Semis (X, unités)	**Rendement (Y, quintaux)**
50	15
60	20
70	25
80	30
90	35

1. Effectuez les calculs pour
2. Calculez a et b pour la régression linéaire.
3. Quelle conclusion pouvez-vous tirer concernant l'impact du nombre de semis sur le rendement des céréales ?

Exercice 3 : Relation entre la Taille des Animaux et leur Poids

Un éleveur de bovins souhaite établir une relation entre la taille (en cm) et le poids (en kg) de ses animaux. Il recueille les informations suivantes :

Taille (X, cm)	Poids (Y, kg)
120	300
130	320
140	360
150	400
160	420

1. Tracer un graphique de dispersion des données.
2. Calculez la pente b et l'intercept a de la régression linéaire en utilisant la méthode des moindres carrés.
3. Prédisez le poids d'un animal mesurant 145 cm à l'aide de l'équation de régression.
4. Interprétez les coefficients dans ce contexte : Quelle est la signification de la pente et de l'intercept pour les animaux ?

Exercice 6 : Comparaison de la Régression Linéaire dans Différentes Fermes

Deux fermes mesurent la relation entre la quantité d'eau utilisée pour l'irrigation et le rendement. Voici leurs données:

<u>Ferme A</u>:

Eau (x, m³)	Rendement (y, quintaux)
100	40

Eau (x, m³)	Rendement (y, quintaux)
150	50
200	60
250	70

Ferme B:

Eau (X, m³)	Rendement (Y, quintaux)
100	30
150	45
200	55
250	65

1. Calculez la régression linéaire pour chaque ferme.
2. Comparez les pentes b des deux fermes : Quelle ferme semble avoir un meilleur rendement en fonction de l'eau utilisée ?
3. Prédisez le rendement des deux fermes si 180 m³ d'eau sont utilisés.
4. Discutez des différences entre les deux fermes en termes d'efficacité d'irrigation.

4. Echantillonnage

4.1

L'échantillonnage est une étape clé dans les études scientifiques,. Il s'agit de la méthode par laquelle on sélectionne un sous-ensemble représentatif (échantillon) d'une population plus large pour tirer des conclusions sur cette dernière.

L'échantillonnage permet de:

- Réduire les coûts et le temps : Il est souvent impossible d'analyser une population entière en raison de contraintes en temps, en coûts ou en ressources. L'échantillonnage permet d'obtenir des résultats fiables avec moins de données.
- Obtenir des conclusions généralisables : Si l'échantillon est représentatif, les conclusions peuvent être extrapolées à la population entière.
- Améliorer la précision : Bien que l'analyse d'un échantillon soit moins précise qu'une analyse complète, de bonnes méthodes d'échantillonnage permettent d'obtenir des résultats avec un degré élevé de précision.

La taille de l'échantillon détermine la précision des résultats. Plus l'échantillon est grand, plus les conclusions seront proches de la réalité de la population. Cependant, un échantillon trop grand peut être coûteux.

4.2 Types d'Échantillonnage

Le choix de la méthode d'échantillonnage dépend de la nature de l'étude, des ressources disponibles et de la structure de la population. Un échantillonnage bien conçu permet de réaliser des études efficaces tout en minimisant les biais et en maximisant la représentativité des résultats.

4.2.1. Échantillonnage aléatoire simple :

- Chaque élément de la population a une chance égale d'être sélectionné.

- Exemple : Sélectionner 50 parcelles de terrain au hasard parmi 1 000 pour mesurer leur rendement en blé.

4.2.2. Échantillonnage systématique :

- Les éléments sont sélectionnés à intervalles réguliers dans une liste de la population.
- Exemple : Dans une ferme, on mesure tous les 10 mètres les niveaux de nutriments dans le sol.

4.2.3 Échantillonnage stratifié :

- La population est divisée en sous-groupes (strates), puis des échantillons sont tirés de chaque strate proportionnellement.
- Exemple : Diviser une population d'animaux en groupes d'âge (jeunes, adultes, âgés), puis échantillonner chaque groupe pour analyser la production de lait.

4.2.4 Échantillonnage en grappes (ou par grappe) :

- L'échantillonnage se fait en sélectionnant d'abord des groupes ou des grappes d'éléments, puis on sélectionne des éléments à l'intérieur de ces groupes.
- Exemple : Sélectionner 5 fermes au hasard, puis échantillonner des plantes dans ces fermes.

4.3. Exercices

Exercice 1 : Échantillonnage pour Estimer le Rendement d'une Culture

Un agriculteur possède un champ de 100 parcelles et souhaite estimer le rendement moyen en blé (en quintaux) par parcelle sans tester toutes les parcelles.

L'agriculteur choisit un échantillon de 10 parcelles au hasard, et il obtient les rendements suivants (en quintaux) :

Parcelle	1	2	3	4	5	6	7	8	9	10
Rendement	35	40	42	38	50	36	44	41	39	37

- Calculez la moyenne de l'échantillon.
- Estimez la variance de l'échantillon et l'écart-type.
- Quelle est l'estimation du rendement moyen du champ entier sur la base de cet échantillon ? Quelle est l'incertitude associée à cette estimation ?

Exercice 2 : Comparaison de la Production Animale entre Deux Fermes

Deux fermes d'élevage utilisent des techniques d'alimentation différentes et souhaitent comparer le poids moyen des animaux. Un échantillonnage est réalisé dans chaque ferme.

- **Ferme A** (50 animaux) : Échantillon de 5 animaux avec les poids suivants (en kg) : 550, 580, 570, 560, 590.
- **Ferme B** (80 animaux) : Échantillon de 8 animaux avec les poids suivants (en kg) : 600, 620, 610, 605, 615, 595, 590, 620.
- Calculez la moyenne et l'écart-type pour chaque ferme.

- En utilisant l'échantillonnage, comparez les poids moyens des deux fermes. Quel échantillon semble avoir un poids moyen plus élevé ?

- Proposez une méthode d'échantillonnage (systématique, stratifié, etc.) qui pourrait être utilisée pour améliorer la précision de l'estimation du poids moyen dans ces fermes.

Exercice 3 : Échantillonnage pour Estimer la Croissance des Plantes sous Différentes Températures

Un chercheur étudie la croissance des plantes sous différentes conditions de température dans une serre. Il divise la serre en quatre zones de températures : **15°C**, **20°C**, **25°C**, et **30°C**. Dans chaque zone, il prélève un échantillon de 5 plantes pour mesurer leur croissance (en cm).

- 15°C : 10, 12, 11, 13, 14
- 20°C : 15, 18, 16, 19, 17
- 25°C : 20, 21, 22, 23, 21
- 30°C : 25, 27, 26, 28, 29

1. Calculez la moyenne et l'écart-type pour chaque zone de température.
2. Utilisez un diagramme en boîte (boxplot) pour représenter graphiquement les données.
3. Proposez une méthode d'échantillonnage pour une future expérience où le chercheur souhaiterait réduire le nombre de mesures tout en conservant des données représentatives.

5. Lois de probabilités usuels continues.

5.1.

Les lois de probabilités continues sont des distributions de probabilité qui s'appliquent à des variables aléatoires continues, c'est-à-dire celles qui peuvent prendre une infinité de valeurs dans un intervalle donné.

Les lois de probabilités continues sont des outils essentiels pour modéliser des phénomènes aléatoires dans différents domaines. En sciences agricoles et en production animale, ces lois permettent d'estimer des paramètres tels que la production moyenne, la durée de vie des équipements, ou les rendements des cultures. L'utilisation de chaque loi dépend de la nature des données et du phénomène observé.

- Loi Normale (ou Gaussienne) **Exemple** : Estimer la taille moyenne des épis de maïs dans une région.
- Loi Uniforme (continue) **Exemple** : La probabilité qu'une température soit comprise entre 15°C et 30°C de manière égale dans une serre.
- Loi Exponentielle **Exemple** : Le temps avant la défaillance d'un équipement dans une installation agricole.
- Loi de Weibull **Exemple** : Modéliser la durée de vie d'une machine dans une usine de transformation de produits alimentaires
- Loi Gamma **Exemple** : Modéliser le temps nécessaire pour récolter une quantité donnée dans une ferme.
- Loi du Khi-Deux (ou Chi-carré) **Exemple** : Utilisée pour tester l'adéquation d'un modèle de répartition des récoltes dans des parcelles agricoles.
- Loi Beta Exemple : Modéliser le taux de germination des graines dans différentes conditions environnementales.

(Cette partie est distribuée aux étudiants sous forme de recherche)

5.2. Exercices

Exercice 1 : Loi Normale (Gaussienne)

Vous travaillez dans une ferme spécialisée dans la production laitière. La production quotidienne de lait suit une loi normale avec une moyenne de 25 litres par vache et un écart-type de 3 litres.

1. Quelle est la probabilité qu'une vache produise entre 22 et 28 litres de lait en une journée ?
2. Quelle est la probabilité qu'une vache produise plus de 30 litres de lait en une journée ?
3. Si vous sélectionnez 5 vaches au hasard, quelle est la probabilité que la production moyenne de ces 5 vaches soit supérieur à 27 litres ?

Exercice 2 : Loi Exponentielle

Vous étudiez le temps entre deux pannes de machines dans une ferme automatisée. Ce temps suit une loi exponentielle de paramètre λ=0.05\lambda = 0.05λ=0.05 pannes par heure.

- Quelle est la probabilité qu'une panne survienne dans les 10 prochaines heures ?
- Quelle est la probabilité qu'une panne survienne après plus de 20 heures ?
- Calculez l'espérance et la variance de la durée entre deux pannes.

Exercice 3 : Loi Uniforme Continue

La température dans une serre dédiée à la culture des tomates varie entre 15°C et 30°C de manière uniforme.

- Quelle est la probabilité que la température soit comprise entre 18°C et 25°C à un moment donné ?
- Quelle est la température médiane de la serre ?
- Quelle est la probabilité que la température dépasse 28°C ?

Exercice 4 : Loi de Weibull

La durée de vie des équipements de mesure utilisés dans une station d'expérimentation agricole suit une loi de Weibull avec un paramètre de forme k=1.5k = 1.5k=1.5 et un paramètre d'échelle λ=2000 heures.

- Quelle est la probabilité qu'un équipement dure plus de 2500 heures ?
- Quelle est la durée médiane de vie d'un équipement ?
- Quelle est la probabilité qu'un équipement tombe en panne avant 1500 heures ?

Exercice 5 : Loi Gamma

Vous surveillez le nombre de jours avant que certaines maladies des plantes n'apparaissent dans des cultures. Le temps d'apparition suit une loi Gamma avec un paramètre de forme α=2et un paramètre d'échelle β=10 jours.

- Quelle est la probabilité que la maladie apparaisse dans les 15 premiers jours ?
- Quelle est la probabilité que la maladie apparaisse après 20 jours ?
- Calculez l'espérance et la variance du temps d'apparition de la maladie.

Exercice 6 : Loi Beta

Vous mesurez le taux de germination de certaines graines sous différentes conditions. Le taux de germination suit une loi Beta avec des paramètres α=3et β=2.

- Quelle est la probabilité que le taux de germination soit supérieur à 0.8 ?
- Quelle est la valeur médiane du taux de germination ?
- Quelle est la probabilité que le taux de germination soit compris entre 0.4 et 0.6 ?

6. Estimation paramétrique

6.1

L'estimation paramétrique est une méthode statistique qui consiste à estimer les paramètres d'une distribution de probabilité en se basant sur un échantillon de données observées. Cette méthode repose sur l'hypothèse que les données suivent une distribution spécifique (par exemple, la distribution normale) et cherche à déterminer les paramètres de cette distribution (comme la moyenne et l'écart-type pour une distribution normale).

L'estimation paramétrique est essentielle pour modéliser et faire des prédictions à partir de données échantillonnées, particulièrement utile dans des domaines comme l'agronomie, où il est souvent difficile ou coûteux d'obtenir des informations complètes sur une population entière.

6.2 Objectifs de l'estimation paramétrique :

- **Estimer les paramètres** : L'objectif principal est de fournir des estimations des paramètres d'une distribution à partir des données échantillonnées.

Exemple : estimer la moyenne et l'écart-type d'une population à partir d'un échantillon.

- **Améliorer la compréhension des processus sous-jacents** : Les estimations paramétriques permettent de modéliser et de mieux comprendre les processus ou phénomènes observés dans des domaines tels que l'agriculture, la biologie, et l'ingénierie.
- **Faire des prédictions** : Une fois les paramètres estimés, le modèle peut être utilisé pour prédire de nouvelles observations.

6.3. Principes :

6.3.1 Modèle de probabilité :

On suppose que les données proviennent d'une distribution connue, mais les paramètres de cette distribution ne sont pas connus.

Par exemple, on suppose que les rendements agricoles suivent une distribution normale, mais on ne connaît pas la moyenne ni l'écart-type.

6.3.2 Méthodes d'estimation :

Maximum de vraisemblance (MLE) : Cette méthode consiste à choisir les valeurs des paramètres qui maximisent la probabilité (ou vraisemblance) d'observer les données collectées.

Soit $X1, X2, \dots, Xn$ un échantillon aléatoire issu d'une distribution de probabilité paramétrée par un vecteur de paramètres θ. La fonction de vraisemblance $L(\theta)$ est définie comme la probabilité de l'observation de ces données en fonction des paramètres θ. En pratique, on cherche le θ qui maximise cette vraisemblance.

La fonction de vraisemblance est donnée par :

$$L(\theta) = \prod_{1}^{n} f(x_i, \theta)$$

où $f(xi \mid \theta)$est la fonction de densité de probabilité de la variable Xi étant donné les paramètres θ.

Pour simplifier les calculs, on utilise généralement la **log-vraisemblance**, qui est simplement le logarithme de la fonction de vraisemblance.

L'estimation du maximum de vraisemblance consiste à maximiser la fonction log-vraisemblance.

Méthode des moments :

Une autre approche consiste à estimer les paramètres en faisant correspondre les moments théoriques de la distribution (par exemple, la moyenne, la variance) aux moments calculés à partir des données observées.

La **méthode des moments** est une technique d'estimation statistique qui consiste à estimer les paramètres d'une distribution de probabilité en égalisant les moments théoriques de cette distribution avec les moments empiriques (calculés à partir des données). C'est une méthode simple et intuitive utilisée lorsque les moments d'une distribution sont faciles à calculer et à manipuler.

Les moments théoriques d'une distribution dépendent des paramètres du modèle. Le k-ième moment théorique est défini comme :

$$\mu_k = E(X^k)$$

Les moments empiriques sont calculés directement à partir des données observées. Le k-ième moment empirique est donné par :

$$\widetilde{\mu_k} = \sum_{1}^{n} {x_i}^k$$

On égale les moments théoriques aux moments empiriques pour obtenir des estimations des paramètres du modèle. Si un modèle a p paramètres, on utilise les p premiers moments pour obtenir un système de p équations, puis on résout ce système pour les paramètres.

Exemples d'applications en sciences agricoles et biologiques :

- **Estimation de la moyenne de rendement** : Supposons que vous ayez un échantillon de rendements agricoles (en quintaux/ha) et que vous souhaitiez estimer la moyenne de rendement de toute la population.
- **Estimation du taux de croissance** des plantes ou des animaux en fonction des paramètres environnementaux comme la température ou l'humidité.

6.4. Exercices

Exercice 1:

Une étude cherche à estimer la production moyenne de maïs par hectare dans une région. Un échantillon de 10 parcelles donne les rendements suivants (en quintaux/ha) : 85, 90, 88, 92, 89, 87, 93, 91, 86, 89.

- Estimez la moyenne et l'écart-type du rendement de la population à partir de cet échantillon.
- Calculez l'intervalle de confiance pour la moyenne avec un niveau de confiance de 95 % en supposant que les rendements suivent une distribution normale.

Exercice 2 : Estimation de la moyenne et de la variance dans un échantillon de rendement agricole

Dans une étude sur le rendement de différentes parcelles de maïs, les données suivantes (en tonnes par hectare) ont été collectées sur un échantillon de 10 parcelles :

8.2,7.9,9.1,8.7,8.5,9.0,8.6,9.2,8.8,8.38.2, 7.9, 9.1, 8.7, 8.5, 9.0, 8.6, 9.2, 8.8, 8.38.2,7.9,9.1,8.7,8.5,9.0,8.6,9.2,8.8,8.3

- Calculez l'estimation de la moyenne du rendement par hectare à partir de cet échantillon.
- Calculez l'estimation de la variance du rendement.

Exercice 3 : Estimation d'un paramètre d'une loi exponentielle

On modélise le temps de germination d'une variété de blé par une loi exponentielle de paramètre λ. Voici les temps de germination observés (en jours) pour un échantillon de 7 graines :

4,6,5,7,5,4,6

Utilisez la méthode des moments pour estimer λ\lambdaλ, le paramètre de la loi exponentielle.

Exercice 3 : Estimation dans une loi binomiale pour la survie animale

Dans une étude sur la survie d'un groupe d'animaux soumis à un traitement, on note que 15 animaux sur 20 ont survécu après 10 jours.

- Modélisez la survie des animaux par une loi binomiale $B(n = 20, p)$, où p est la probabilité de survie. (Utilisez la méthode des moments pour estimer p).

Exercice 4 : Estimation dans une loi de Poisson pour le nombre de parasites

Un biologiste observe que le nombre de parasites sur les plantes suit une loi de Poisson. Voici les nombres de parasites observés sur un échantillon de 6 plantes :

3,4,2,5,4,3

Estimez le paramètre λ de la loi de Poisson, qui représente le nombre moyen de parasites par plante.

Référence

- **Montgomery, D. C.** (2019). *Design and Analysis of Experiments*. 10th Edition. Wiley.
- **Goupy, J.** (2005). *Plans d'expériences: Applications à l'entreprise*. Dunod.
- **Steel, R. G., Torrie, J. H., & Dickey, D. A.** (1997). *Principles and Procedures of Statistics: A Biometrical Approach*. 3rd Edition. McGraw-Hill.
- **Piegorsch, W. W., & Bailer, A. J.** (2005). *Analyzing Environmental Data*. Wiley.
- **Freund, R. J., Wilson, W. J., & Sa, P.** (2010). *Regression Analysis: Statistical Modeling of a Response Variable*. 2nd Edition. Academic Press.
- **Cochran, W. G., & Cox, G. M.** (1992). *Experimental Designs*. 2nd Edition. Wiley. **Draper, N. R., & Smith, H.** (1998). *Applied Regression Analysis*. Wiley.
- **Gomez, K. A., & Gomez, A. A.** (1984). *Statistical Procedures for Agricultural Research*. 2nd Edition. Wiley.
- **Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O.** (2006). *SAS for Mixed Models*. 2nd Edition. SAS Institute.
- **Everitt, B. S., & Hothorn, T.** (2010). *A Handbook of Statistical Analyses Using R*. 3rd Edition. Chapman and Hall/CRC.

Remarques

1. Dans l'élaboration de ce cours, nous avons largement recours à l'intelligence artificielle, ce qui nous a permis d'améliorer la clarté du langage utilisé et d'assurer une cohérence et une harmonie entre les différents thèmes abordés."
2. L'intelligence artificielle a joué un rôle crucial dans l'organisation de ce module, facilitant ainsi l'accès aux concepts pour nos étudiants en sciences agricoles, qui ne sont pas toujours spécialisés en mathématiques et en statistiques."
3. En nous appuyant sur l'intelligence artificielle, nous avons pu éviter les répétitions, simplifier les notions complexes et présenter un contenu qui soit à la fois pédagogique et adapté aux besoins de nos étudiants."
4. L'utilisation de l'intelligence artificielle dans la création de ce cours nous a permis d'offrir une expérience d'apprentissage enrichissante, en veillant à la précision du langage et à la structuration logique des sujets, ce qui est essentiel pour des étudiants non spécialisés en statistiques."
5. Nous avons adopté une approche basée sur l'intelligence artificielle pour concevoir ce cours, ce qui a contribué à la clarté et à la fluidité des idées présentées, tout en rendant les concepts d'importance statistique plus accessibles aux étudiants en sciences agricoles."