République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Ghardaïa

Cours de Biostatistique 01

au profil des étudiants de Tronc Commun (L3 écologie et environnement)

Sciences de la Nature et de la Vie

Auteur : **HADDAD Soumia**

Enseignant-Chercheur | Département de Biologie | Faculté SNV-ST | Université de Ghardaïa.

Science de l'Environnement.

Laboratoire Valorisation et Conservation des Ecosystèmes Arides (LVCEA). Faculté des Sciences de la Nature et de la Vie et Sciences de la terre, Université de Ghardaïa, Algeria;

Table des matières

Chers étudiants,

Bienvenue dans le monde passionnant de l'analyse des données écologiques !

Comprendre comment les êtres vivants interagissent avec leur environnement ne se limite pas à l'observation sur le terrain : cela passe aussi par une lecture fine et rigoureuse des chiffres. C'est là que les bio-statistiques entrent en jeu.

Chers étudiants, bienvenue dans le cours de Biostatistique !

La biostatistique est un outil fondamental pour analyser les phénomènes biologiques à partir de données mesurées sur le terrain ou en laboratoire. Elle vous permet de transformer ces données en informations claires et interprétables, essentielles pour la recherche, le diagnostic, la gestion écologique ou l'amélioration des pratiques agronomiques.

Ce cours vous apportera les principaux concepts statistiques appliqués aux sciences de la vie, ainsi que des méthodes concrètes pour analyser vos propres données. Il est organisé en trois grandes parties :

1. Statistique descriptive

- ✓ Identifier et classer les types de variables biologiques (qualitatives ou quantitatives)
- ✓ Calculer et interpréter les paramètres de position : moyenne, médiane, mode
- ✓ Représenter les données à l'aide de tableaux, histogrammes, diagrammes
- ✓ Évaluer la dispersion : étendue, variance, écart-type, coefficient de variation

2. Statistique inférentielle

- ✓ Comprendre les lois de probabilité utilisées en biostatistique (notamment la loi normale)
- ✓ Apprendre les principaux tests statistiques, avec leurs hypothèses et interprétations :
- ✓ Test de conformité
- ✓ Comparaison de deux moyennes :
    - ➢ Séries indépendantes (Student, Z)
    - ➢ Séries appariées
    - ➢ Comparaison à une moyenne de référence
- ✓ Analyse de la variance (ANOVA) :
    - ➢ À un facteur

- À deux facteurs
- À deux facteurs avec mesures répétées

3. Corrélation et régression

- ✓ Déterminer s'il existe une relation entre deux variables biologiques
- ✓ Calculer et interpréter un coefficient de corrélation
- ✓ Modéliser cette relation à l'aide de la régression linéaire simple
- ✓ Interpréter la pente et la qualité de la droite de régression

Chers étudiants, pour renforcer vos acquis et faciliter la compréhension ce cours repose sur une alternance entre explications théoriques, exemples concrets, QCM et devoirs. Il vise à rendre la biostatistique accessible, utile et directement applicable dans vos études et vos futures recherches:

- ❖ Des QCM seront proposés tout au long du module, pour vous aider à assimiler la théorie et les logiques des tests statistiques.
- ❖ Des devoirs seront régulièrement donnés, sous forme d'exercices d'interprétation ou d'analyse, pour mettre en pratique les connaissances acquises.

Bonne découverte, et n'oubliez pas : la statistique est une alliée de la science, pas un obstacle !.

I. Notions statistiques

1. Statistique

C'est la science dont l'objet est de recueillir, de traiter et d'analyser des données issues de l'observation de phénomènes aléatoires « hasards »

- **But :** Décrire les phénomènes mathématiquement et étudies des prévisions
- **Pourquoi ? :** Pour faire des prévisions et prendre des décisions à leur sujet.
- **Donc :** les données sont entachées **d'incertitude** et présentent **des variations**
  - ✓ Le déroulement des phénomènes observés n'est pas prévisibles à l'avance avec certitude
  - ✓ Toute mesure est entachée d'erreur
  - ✓ Seuls quelques individus sont observés et/ou doit extrapole les conclusions de l'étude d'une population.

Donc il y a toujours le hasard et la probabilité

2.Population statistique

Est un ensemble d'éléments ou d'événements similaires sur lesquels porte une étude statistique, et elle est utilisée pour étudier des caractéristiques de cette population.

Exemple

- **L'ensemble des étudiants d'une université**. Si nous voulons étudier la taille moyennes des étudiants, tous les étudiants inscrits dans cette université constituent la population statistique.
- **Tous les élèves d'une école**. Si l'on veut savoir leur taille moyenne, alors tous les élèves de cette école forment la population étudiée.

3.Un échantillon statistique

Est un ensemble d'individus représentatifs d'une population, utilisé pour étudier des caractéristiques de cette population.

Exemple

- **Un groupe de 100 étudiants** choisi au hasard dans une université de 1000000 étudiants pour étudier leur taille moyenne. Cet échantillon représente la population totale des élèves de l'école.

4.Un individu statistique (unité statistique)

Est un élément d'une population sur laquelle une étude statistique est menée. Il peut s'agir d'une personne, d'un ménage, d'une entreprise, d'un établissement, d'une commune, d'un département, d'une région ou encore d'un pays. Les individus statistiques sont les éléments constitutifs d'une population, et c'est sur ces individus que les mesures et les analyses statistiques sont effectuées pour en tirer des conclusions.

Exemple

- **Un étudiant spécifique** dans une classe, par exemple, "Mohamed". Chaque étudiant de la classe est un individu statistique, et ses caractéristiques (comme l'âge, la taille, ou la note) sont des observations collectées pour l'analyse.

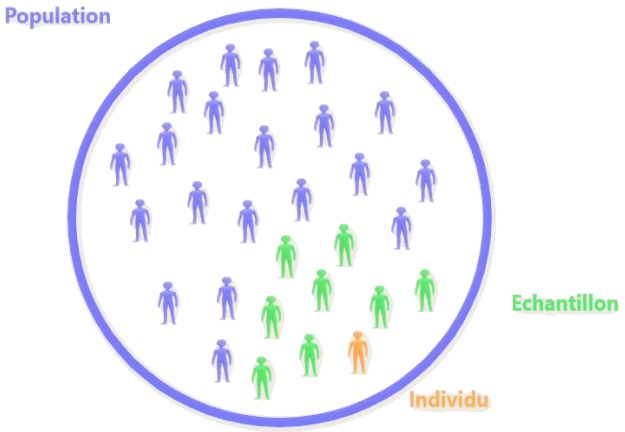


Figure 01 : Population, échantillon et individu statistique

5. Une variable statistique

Est une caractéristique qui peut être mesurée ou observée et qui peut prendre différentes valeurs.

Exemple

✧ **La taille des étudiants** dans une classe. La taille est une caractéristique qui peut varier d'un étudiant à l'autre, et il peut être mesuré et analysé dans le cadre d'une étude statistique.

6. Une observation statistique

Est une mesure ou une donnée collectée pour une variable donnée. Elle peut être exprimée sous forme de nombre ou de qualité.

Exemple

✧ La taille de 1,75 m d'un étudiant. C'est la donnée spécifique collectée pour la variable "taille" d'un individu particulier dans une population.

7. Série statistique

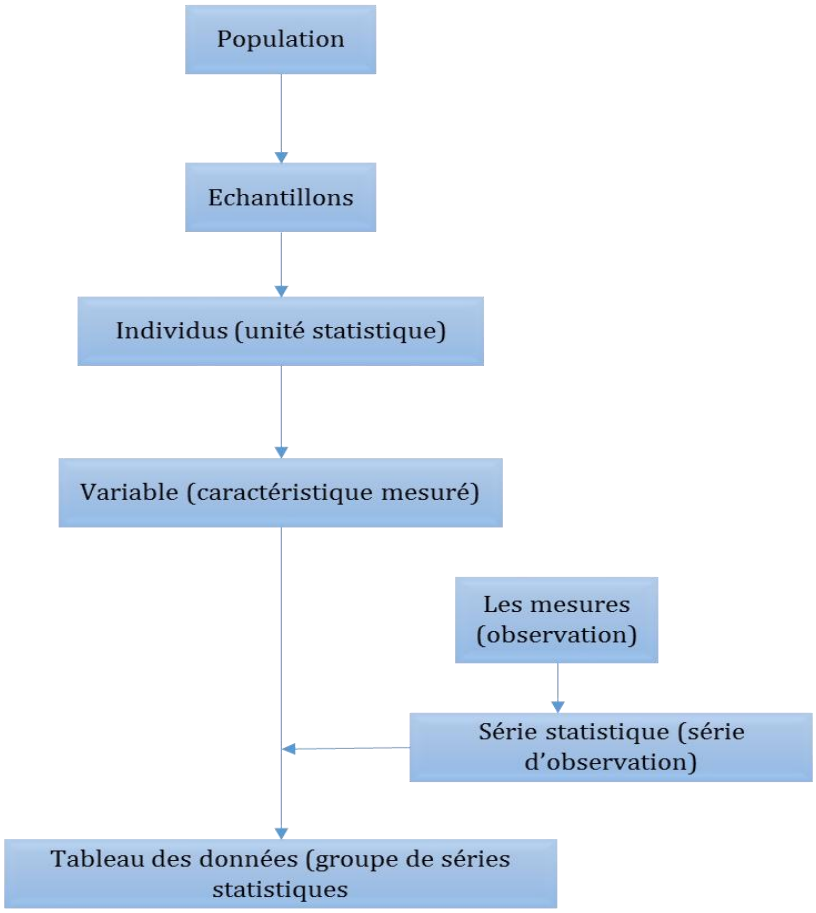
C'est un ensemble des observations

✧ **Série statistique quantitative :** Lorsque le caractère est mesurable et exprimé en nombre.

✓ **Quantitative discrète :** Nombre fini de valeurs (ex. nombre d'enfants dans une famille).

✓ **Quantitative continue :** Nombre infini de valeurs dans un intervalle donné (ex. taille des individus).

✧ **Série statistique qualitative :** Lorsque le caractère n'est pas mesurable numériquement (ex. couleur des yeux, genre, ...etc)



8. Tableaux des données (Matrice)

Englobe des groupes de séries statistiques

Exemple

Le tableau suivant présente un échantillon de cinq étudiants (individus), pour lesquels trois variables quantitatives ont été relevées :

- ✓ **Taille (variable 1, en mètres) :** mesure continue représentant la hauteur de chaque étudiant.
- ✓ **Notes de travaux dirigés – TD (variable 2, en points) :** variable discrète exprimant la performance académique.
- ✓ **Poids (variable 3, en kilogrammes) :** mesure continue du poids corporel.

Chaque ligne correspond à un individu de l'échantillon et chaque colonne indique la valeur observée pour la variable considérée. Ainsi, par exemple, l'étudiant 1 a une taille de 1,77 m, une note de TD égale à 13, et un poids de 65 kg.

Tableau 01 : Tableau des données statistiques

Étudiant (échantillon)	**Taille** (variable 1)	**Notes TD** (variable 2)	**Poids** (variable 3)
1 (individu)	**1,77** (observation)	**13** (observation)	**65** (observation)
2	**1,65**	**14**	**53**
3	**1,45**	**12**	**72**
4	**1,50**	**10**	**61**
5	**1,62**	**8**	**50**

9. Démarche expérimentale

C'est un processus itératif qui permet aux chercheurs de mieux comprendre les phénomènes complexes et de réfuter ou de confirmer leurs hypothèses. Voici les étapes d'une démarche expérimentale :

9.1. L'observation

C'est l'action de recueillir des informations sur un phénomène biologique, chimique ou physique, sans le modifier.

9.2. Problème

Observer un phénomène naturel et identifier une question, un problème scientifique est une question précise qui surgit d'une observation et qui nécessite une explication rationnelle.Il doit être formulable sous forme d'hypothèse et testable par une expérience.

9.3. Hypothèse

Une hypothèse scientifique est une explication provisoire à un problème scientifique. Elle doit être formulée de manière claire, testable et falsifiable (c'est-à-dire qu'on peut la prouver vraie ou fausse par une expérience ou des observations).

9.4. Expérience

Une expérience scientifique est une démarche méthodique visant à tester une hypothèse en contrôlant les variables et en observant les résultats. Elle permet de vérifier si une relation existe entre une cause (variable indépendante) et un effet (variable dépendante).

9.5. Résultats

Les résultats d'une expérience sont les données obtenues après l'observation et la mesure des variables étudiées.

9.6. Test statistique

Un test statistique est une méthode mathématique permettant d'analyser des données afin de déterminer si une différence observée entre des groupes ou des variables est significative ou simplement due au hasard.

9.7. Interprétation

L'interprétation est l'analyse des résultats d'une expérience ou d'une étude afin d'en tirer des conclusions significatives. Elle permet de donner du sens aux données recueillies et de vérifier si elles confirment ou réfutent une hypothèse.

Exemple

a) Observation

Pendant les examens, certains étudiants ont de très bonnes notes, d'autres ont des notes plus faibles. En parlant avec eux, on voit que ceux qui réussissent ont révisé régulièrement. Les autres ont révisé à la dernière minute ou pas du tout.

b) Problème Scientifique

Pourquoi certains étudiants ont-ils de meilleures notes que d'autres aux examens ?

c) Hypothèse

Les étudiants qui révisent régulièrement obtiennent de meilleures notes que ceux qui révisent à la dernière minute.

- **Hypothèse Générale**

"Le temps de révision influence les performances scolaires des étudiants."

- **Hypothèse Opérationnelle**

"Les étudiants qui révisent au moins 10 heures par semaine obtiennent une note moyenne supérieure à 14/20."

- **Hypothèse Nulle (H_0) et Hypothèse Alternative (H_1)**

H_0 : "Le temps de révision n'a pas d'effet significatif sur les notes des étudiants."

H_1 : "Les étudiants qui révisent plus longtemps ont de meilleures notes que ceux qui révisent moins."

d) Expérience :

Séparer deux groupes d'étudiants :

Groupe 1 : Étudiants qui révisent au moins 10 heures par semaine.

Groupe 2 : Étudiants qui révisent moins de 3 heures par semaine.

Comparer leurs moyennes aux examens.

e) Mesures :

Temps de révision (en heures/semaine).

Notes obtenue aux examens (sur 20).

f) Analyse statistique :

Calculer la moyenne des notes pour chaque groupe.

Vérifier s'il y a une différence significative entre les deux groupes.

g) Résultat Attendu et Conclusion

Si l'analyse montre que les étudiants qui révisent plus ont des notes plus élevées, alors « il y a une différence significative entre les deux groupes et cette différence est liée au temps de révision pas au hasard donc l'hypothèse nulle est rejeté .

Si aucune différence significative n'est trouvée, alors l'hypothèse nulle est accepté .

II. Statistiques descriptives

1. Definition

La statistique descriptive est une branche des statistiques qui regroupe de nombreuses techniques utilisées pour **décrire** un ensemble relativement important de données.

- **Objectif :** résumer ou représenter les données disponibles quand elles sont nombreuses, en utilisant des statistiques telles que la moyenne, la médiane, le mode, l'écart-type, la variance, les quartiles, les déciles, les histogrammes, les diagrammes en boîte, les nuages de points, les diagrammes circulaires, etc.

2. Nature de variables statistiques

Les variables statistiques peuvent être de nature qualitative ou quantitative.

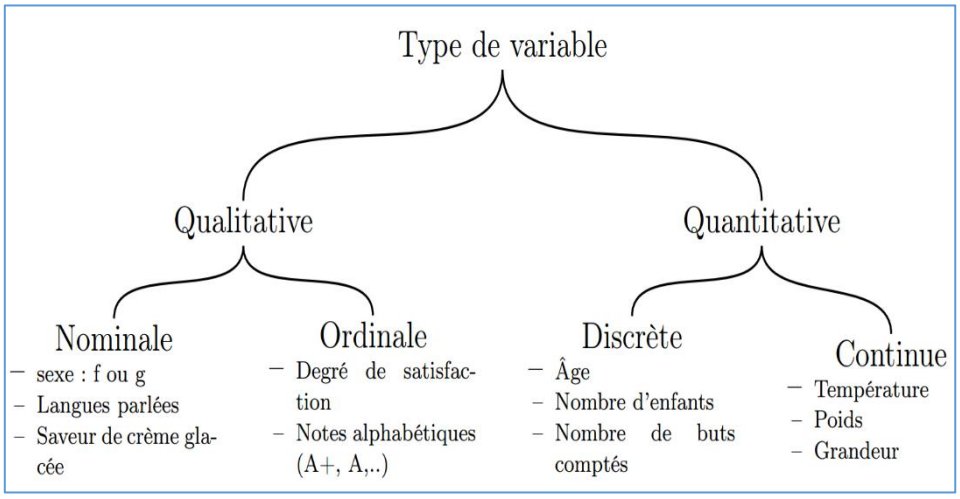


Figure 02 : Types de variables

On distingue donc deux catégories de variables :

2.1 Qualitatives

Variable portant sur des grandeurs non numériques

- **Nominales :** Variable correspond à des noms sans ordre précis (sexe : femme/homme ; saveur : bonne/mauvaise ... etc)
- **Ordinales :** Variable possédant un ordre (très bien , bien , passable...)

2.2 Quantitatives

Variable portant sur des grandeurs numériques

- **Discrètes :** Variable dont les valeurs sont énumérables (1;2;3;4....etc)
- **Continues :** Variable dont les valeurs sont tellement nombreuses qu'elles en deviennent non-énumérable (1,55; 2,33 ; 10,05 ; ..... etc)

3. Paramètres caractéristiques

3.1 Paramètre de position

3.1.1 Mode /classe modale

Il s'agit de la valeur ou la classe la plus fréquente (leur effectif est le plus élevé). Dans un ensemble de données ; le mode est utile pour décrire ce qui apparaît le plus souvent ,il permet donc d'identifier ce qui est le plus courant ou le plus typique dans un échantillon.

Exemple :

- ✓ Dans l'ensemble de données (2, 3, 3, 5, 6, 6, 6), le mode est 6 car c'est la valeur qui apparaît le plus fréquemment.
- ✓ Si, dans une classe, la majorité des étudiants ont eu la note de 12, alors le mode est 12.

3.1.2 Moyenne

Il s'agit de la valeur centrale de la distribution, calculée en sommant toutes les valeurs d'une variable et en divisant par le nombre d'observations

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$$

La moyenne permet de résumer un ensemble de données par une seule valeur représentative, donnant une idée générale du niveau global, utile pour comparer des groupes ou observer des variations dans le temps

Exemples :

- ✓ Si cinq oiseaux ont pondu 2, 3, 3, 4 et 8 œufs respectivement , la moyenne est :

$$(2+3+3+4+8)\div 5=20\div 5=4 \text{ œufs.}$$

Cela signifie qu'en moyenne, chaque oiseau a pondu 4 œufs.

3.1.3 La médiane

Il s'agit de la valeur qui sépare l'ensemble des observations en deux parties égales, lorsque ces observations sont classées dans l'ordre croissant.

Exemple :

S1: 17,6,9,13,14,5,11,15,3 9 N **obs. impairs**

S2: 2, 14, 6, 11, 16,5,18,18,9 8 N **obs pairs**

1. Ordonner

S1 : 3, 5, 6, 9, 11, 13, 14, 15,17

S2 : 2, 5, 6, 11, 14, 16, 18,19

2. Calculer la position de la médiane

Si N est impair → la médiane est la valeur à la position N/2

Si N est pair → la médiane est la moyenne des deux valeurs aux positions N/2 et (N+1)/2

S1 : 3, 5, 6, 9, 11, 13, 14, 15,17 11=> médiane

S2 : 2, 5, 6, 11, 14, 16, 18,19 (11+ 14)/2 =12.5=> médiane

NB : La moyenne résume les données en calculant une valeur centrale basée sur l'ensemble des valeurs, tandis que la médiane indique la valeur du milieu et est plus fiable quand les données sont très dispersées ou contiennent des valeurs extrêmes.

3.2 Paramètres de dispersion « de variation »

3.2.1 Etendue

C'est la différence entre la plus grande et la plus petite valeur s de la variable

- **Discrets :** $W=X_{Max} - X_{Min}$
- **Continues :** W= Borne supérieur de la dernière classe – Borne inférieur de la première classe.

Exemple :

Si une série de notes est : 5, 9, 12, 14, 16

Alors : Etendue=16−5=11

Cela signifie que les notes s'étalent sur 11 points.

3.2.2 Ecart semi interquartile

- **La médiane Q2 :** Partage la série en deux distributions
- **Le quartile Q3 :** Partage la série en quatre distributions

Exemple :

S_1 : 3, 5, 6, 9, 11, 13, 14, 15, 17 9 obs. impairs

Q_1= N× (1/4) => je cherche la valeur dans S1 (25%) c'est (5+6)/2

Q_3= N× (3/4) => je cherche la valeur dans S1 (75%) c'est (14+15)/2

Q_1 : La deuxième valeur c'est 5,5

Q_3 : la sixième valeur c'est 14,5

S_2 : 3, 5, 6, 9, 11, 13, 14, 15 8 obs. impairs

Q_1 : La deuxième valeur c'est 5

Q_3 : la sixième valeur c'est 14

=> L'écart semi interquartile

$$Q_{SI} = (Q_3-Q_1) / 2 ; Q_{SI} = (14,5-5,5) / 2=4,5$$

$$Q_{S2} = (Q_3-Q_1) / 2 ; Q_{S2} = (14-5) / 2=4,5$$

3.2.3 L'écart absolu moyen : EAM

L'écart c'est l'observation –La moyenne

$$EAM = \sum \frac{ni|xi-\bar{x}|}{N} \quad \text{pour les variables discontinue}$$

$$EAM = \sum \frac{ni|ci-\bar{x}|}{N} \quad \text{pour les variables continue.}$$

3.2.4 La variance et l'écartype

L'illustration suivante permet de visualiser la relation entre la moyenne, la variance et l'écart-type de façon intuitive :

La moyenne ($\bar{x}$) est représentée comme le point d'équilibre de la série.C'est la valeur centrale de référence.

Les «xi» sont les valeurs mesuré gravitent autour de la moyenne .

Les flèches rouges indiquent les écarts entre chaque donnée et la moyenne$(xi - \bar{x})$. Ces écarts sont appelés écarts à la moyenne.

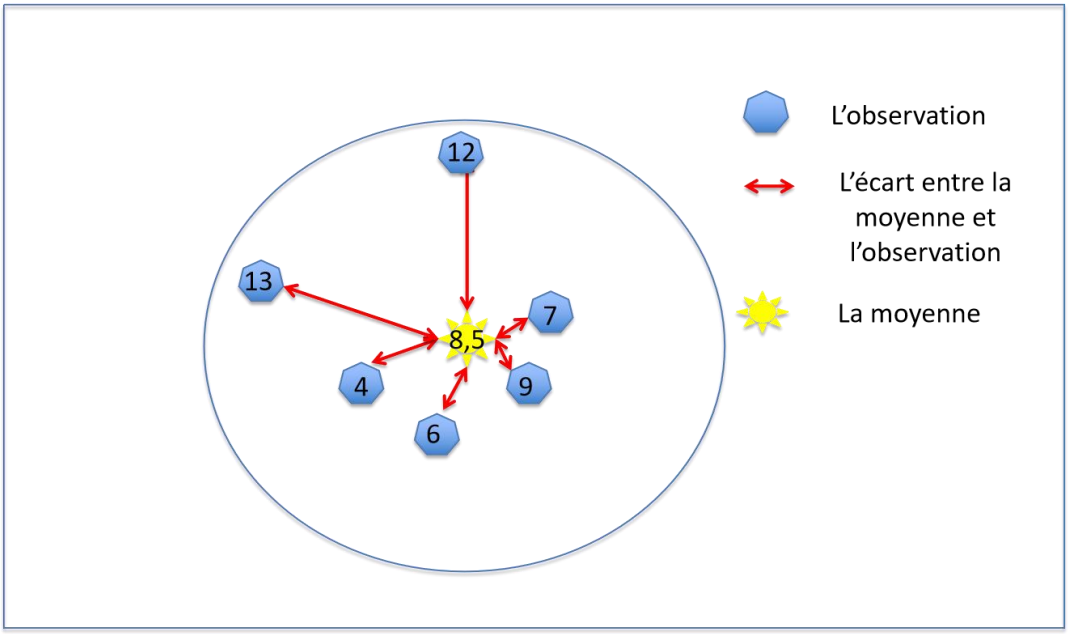


Figure 03 : Relation visuel entre moyenne et observations

- ✓ Plus une flèche est longue, plus l'observation (xi) est éloignée de la moyenne $(\bar{x})$.Cela reflète une plus grande contribution à la variance.
- ✓ **La variance** est la moyenne des carrés de ces flèches : elle mesure donc la dispersion totale des données autour de la moyenne.

$$\partial^2 = \sum_{i=1}^{n} \frac{(xi-\bar{x})^2}{N}$$

Avec effectif $\partial^2 = \sum_{i=1}^{n} \frac{ni(xi-\bar{x})^2}{N}$

Avec fréquence $\partial^2 = \sum_{i=1}^{n} fi(xi-\bar{x})^2$

- ✓ **L'écart-type** est la racine carrée de la variance : il donne une valeur moyenne des écarts, exprimée dans la même unité que les données.

$$\sqrt{\sum_{i=1}^{n} \frac{(xi-\bar{x})^2}{N}} = \sqrt{\partial^2} = \sqrt{\text{variance}}$$

Avec effectif $\sqrt{\sum_{i=1}^{n} \frac{ni(xi-\bar{x})^2}{N}}$

4. Analyse Comparative des Paramètres de Position et de Dispersion : Comprendre le rôle :

4.1. Moyenne vs Médiane

Salaires mensuels (en milliers de dinars) :

30, 32, 31, 33, 35, 34, **150**

Moyenne = **49,3** (très influencée par la valeur extrême 150)

Médiane = **33** (plus stable face aux valeurs extrêmes)

Conclusion : La médiane est plus fiable en cas de données extrêmes.

4.2. Moyenne, Médiane, Mode

Notes de deux classes (sur 20)

Classe 1 : 10, 11, 11, 12, 12, 12, 13

→ **Moyenne** = 11,6 ; **Médiane** = 12 ; **Mode** = 12

Classe 2 : 8, 9, 10, 12, 17, 18, 19

→ **Moyenne** = 13,3 ; **Médiane** = 12 ; **Mode** = aucun

- ✓ La **moyenne** est la somme de toutes les valeurs divisée par le nombre total. Elle donne une idée générale du niveau moyen. Mais elle est **sensible aux valeurs extrêmes** (comme 17, 18, 19 dans la classe 2 qui augmentent fortement la moyenne).
- ✓ La **médiane** est la valeur du milieu quand les données sont rangées. Elle est **moins influencée par les valeurs extrêmes**, donc plus fiable quand les données sont déséquilibrées.
- ✓ Le **mode** est la valeur qui apparaît le plus souvent. Dans la classe 1, le mode est 12 (elle apparaît 3 fois), ce qui montre une **concentration autour de cette note**. Dans la classe 2, il n'y a **aucune note répétée**, donc **pas de mode**.

Donc :

Ces trois indicateurs aident à comprendre **où se situe le centre des données** et si les résultats sont **regroupés ou dispersés**. Comparer ces valeurs permet de savoir si les données sont **symétriques ou déséquilibrées (asymétriques)**.

4.3. Moyenne & Écart-type ; Variance

Deux classes ont passé le même examen (sur 20). Voici leurs notes :

Classe A : 12, 13, 12, 13, 12 **Moyenne = 12.4** **Écart-type ≈ 0.55**	**Classe B :** 8, 10, 12, 14, 16 **Moyenne = 12.0** **Écart-type ≈ 2.83**

Les deux classes ont des **moyennes similaires** (12.4 et 12.0).

Mais l'écart-type montre que :

- ✓ En **Classe A**, les notes sont proches de la moyenne (écart-type ≈ 0.55) élèves ont des résultats **homogènes** (peu de différence entre eux)

✓ En **Classe B**, les notes sont plus dispersées autour de la moyenne (écart-type ≈ 2.83) résultats **très variés** (certains très bons, d'autres faibles)

Donc :

Même si Les **deux classes ont presque la même moyenne**, l'écart-type permet de voir si les données sont régulières ou dispersées.

✓ Écart-type = **0.49 ;** Variance = $(\mathbf{0.49})^2 = \mathbf{0.24}$
✓ Écart-type ≈ **2.83 ;** Variance = $(\mathbf{2.83})^2 \approx \mathbf{9}$

3.4. Différence entre Variance et Écart-type :

✓ La **variance** donne une mesure **mathématique** de la dispersion mais en **unités au carré** (ex : points^2), donc moins intuitive.
✓ L'**écart-type** est la **racine carrée de la variance**, exprimée **dans la même unité que les données**, donc plus facile à **interpréter concrètement**.

Si les notes sont en points :

✓ Variance = 9 (en **points^2**) → on comprend mal ce que ça signifie concrètement.
✓ Écart-type = 3 (en **points**) → on peut dire que les notes s'écartent en moyenne de 3 points par rapport à la moyenne.

Donc :

L'écart-type est plus facile à interpréter car il est **dans la même unité** que les données, contrairement a la **variance qui** donne une idée globale, mais est moins intuitive car elle est en unité **au carré**.

QCM statistique descriptive :

I.1. Quelle est la nature d'une variable mesurant le sexe d'un individu ?

A) Quantitative discrète

B) Qualitative nominale

C) Qualitative ordinale

D) Quantitative continue

2. Une variable qui prend comme valeurs : 'faible', 'moyen', 'élevé' est :

A) Nominale

B) Ordinale

C) Discrète

D) Continue

3. Parmi les mesures suivantes, laquelle est la plus sensible aux valeurs extrêmes ?

A) Mode

B) Médiane

C) Moyenne

D) Ecart interquartile

4. L'écart-type est défini comme :

A) La moyenne des valeurs

B) La racine carrée de la variance

C) Le carré de la moyenne

D) L'écart à la moyenne

5. Si la moyenne et la médiane sont très différentes, on peut dire que :

A) La série est homogène

B) Il n'y a pas de dispersion

C) Il y a une valeur extrême

D) La moyenne est plus fiable

6. La variance est exprimée :

A) Dans la même unité que les données

B) En unité au carré

C) En pourcentage

D) En fréquence

7. Pour une série symétrique :

A) Moyenne > Médiane > Mode

B) Moyenne < Médiane < Mode

C) Moyenne = Médiane = Mode

D) Mode > Médiane > Moyenne

II. Voici les notes obtenues par 10 étudiants à un examen : {12, 14, 15, 10, 8, 17, 15, 10, 18, 11}

1. Quelle est la moyenne des notes ?

A) 11,5

B) 13

C) 13,5

D) 12

2. Quelle information donne la moyenne ?

A) La note la plus fréquente

B) La note la plus élevée

C) Le centre de la distribution des notes

D) L'écart entre les notes

3. Quelle est la médiane ?

A) 12

B) 13

C) 15

D) 10

4. À quoi sert la médiane ?

A) À connaître la note maximale

B) À savoir si les données sont symétriques

C) À identifier la valeur au centre d'un ensemble de notes

D) À calculer la dispersion

5. Quelle est la note la plus fréquente (le mode) ?

A) 10

B) 15

C) 12

D) 17

6. À quoi sert le mode ?

A) À mesurer l'écart entre les extrêmes

B) À savoir la note la plus obtenue

C) À remplacer la moyenne

D) À trier les données

7. Quelle est l'étendue des notes ?

A) 10

B) 9

C) 8

D) 6

8. À quoi sert l'étendue ?

A) À calculer la moyenne

B) À savoir le nombre d'étudiants

C) À mesurer la différence entre la meilleure et la moins bonne note

D) À vérifier si les données sont normales

9. Si l'écart-type est grand, cela signifie que :

A) Les notes sont très proches

B) Il y a beaucoup de notes identiques

C) Les notes sont très dispersées

D) La moyenne est basse

10. Si deux groupes ont la même moyenne mais pas le même écart-type, alors :

A) Les deux groupes sont identiques

B) Le groupe avec l'écart-type le plus grand a des notes plus variées

C) Cela veut dire qu'il y a une erreur

D) Le mode est aussi différent

11. Que mesure la variance ?

A) Le nombre total d'étudiants

B) Le carré de la moyenne

C) La dispersion des notes autour de la moyenne

D) La note la plus éloignée de la moyenne

12. On analyse les notes d'étudiants (de 0 à 20). On a calculé la moyenne = 10, et deux indicateurs de dispersion : la variance et l'écart-type.

Que peut-on dire sur leur rôle ?

A) La variance et l'écart-type donnent la même information, mais l'écart-type est plus facile à comprendre car il est dans la même unité que les notes.

B) On utilise toujours la variance, jamais l'écart-type.

C) L'écart-type sert à mesurer la moyenne.

D) La variance est plus petite que l'écart-type.

Devoir :

1. Objectif du devoir :

Construire un tableau de statistiques descriptives à partir de vos propres notes, puis interpréter les résultats obtenus.

2. Instructions :

- ✓ Ouvrez le logiciel : Lancez Statistica 10.
- ✓ Saisissez vos notes du semestre précédent (par exemple : les résultats obtenus dans vos examens, sur 20).
- ✓ Générez les statistiques descriptives automatiques, en utilisant la fonction "Basic Statistics" ou "Descriptive Statistics".
- ✓ Relevez et organisez les résultats obtenus dans le tableau suivant :

3. Tableau des statistiques descriptives

Paramètre	**Valeur obtenue**
N (effectif totale)	
Moyenne	
Médiane	
Premier quartile (Q1)	
Troisième quartile (Q3)	
Écart-type	
Variance	
Note minimale	
Note maximale	

4. Interprétation attendue :

- ✓ Que représente la moyenne dans votre cas ?

.....

.....

- ✓ La médiane est-elle proche de la moyenne ? Que cela indique-t-il ?

.....

.....

- ✓ Les quartiles montrent-ils une répartition équilibrée de vos notes ?

.....

.....

- ✓ La variance et l'écart-type sont-ils élevés ? Vos résultats sont-ils homogènes ou dispersés ?

.....

.....

- ✓ Quelle est l'amplitude totale de vos notes (différence entre note max et min) ?

.....

.....

- ✓ Concluez en quelques lignes sur la régularité de vos performances et ce que vous apprennent ces chiffres.

.....

.....

5. Représentation graphique des séries de distribution

Les représentations graphiques facilitent grandement l'interprétation et la communication des résultats statistiques, en offrant une visualisation claire des données.

Les principales représentations graphiques utilisées en biostatistique descriptive sont :

5.1 Les histogrammes

Qui représentent la distribution d'une variable quantitative

Exemple :

Supposons que nous ayons collecté des données sur les poids (en kg) d'un échantillon de patients.

Intervalle de poids	Effectif
[50 - 55[	5
[55 - 60[	12
[60 - 65[	18
[65 - 70[	10
[70 - 75[	5

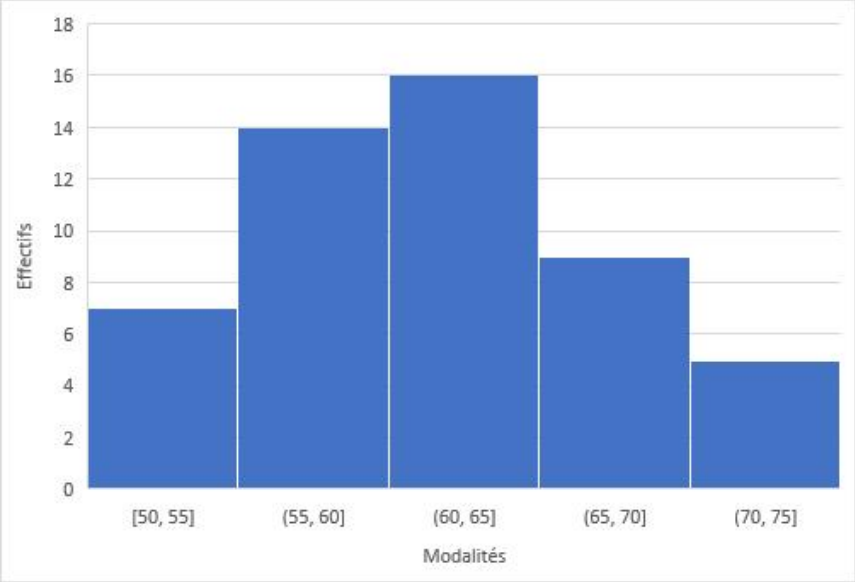


Figure 04 : Histogramme

Cet histogramme permet de visualiser la distribution des poids dans notre échantillon. On peut ainsi observer que la majorité des patients ont un poids compris entre 60 et 65 kg.

L'histogramme est particulièrement utile pour représenter la distribution d'une variable quantitative continue, comme le poids, la taille, l'âge, etc.

Il permet d'identifier facilement les valeurs les plus fréquentes, ainsi que l'étendue, la symétrie et l'éventuelle présence de plusieurs modes dans la distribution.

Cette représentation graphique complète utilement les informations numériques du tableau statistique et facilite l'interprétation des résultats en biostatistique descriptive.

5.3 Les diagrammes circulaires ou "camemberts"

Qui montrent la répartition en pourcentages d'une variable catégorielle.

Exemple :

Supposons que nous ayons collecté des données sur le statut tabagique d'un échantillon de patients dans un service hospitalier.

Statut tabagique	**Effectif**
Fumeur	25
Non-fumeur	45
Ancien fumeur	30

- ✓ Ancien fumeur 30%
- ✓ Fumeur 25%
- ✓ Non-fumeur 45%

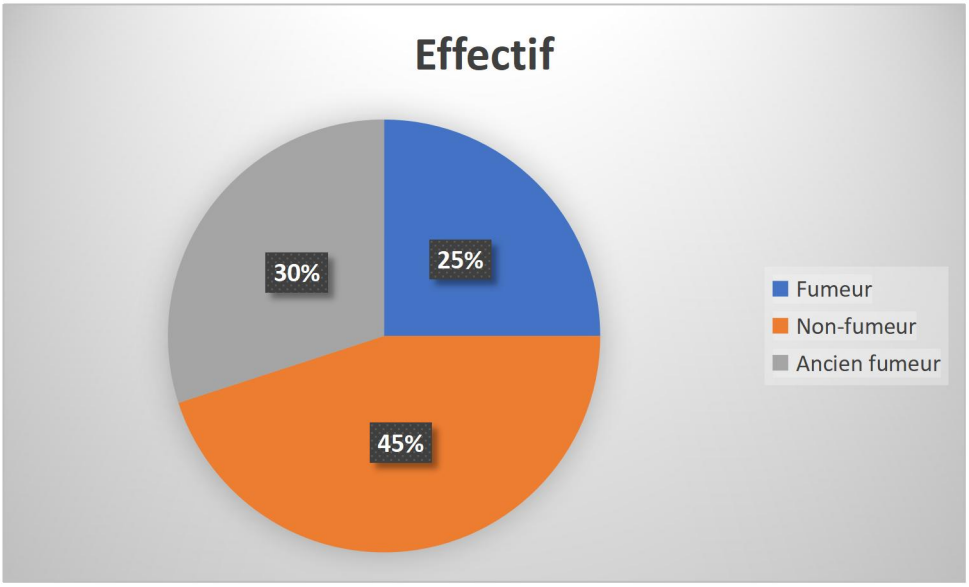


Figure 06 : Diagramme circulaire ou "camembert"

Ce diagramme circulaire permet de visualiser facilement la répartition des patients selon leur statut tabagique. On peut ainsi observer que la majorité des patients (45%) sont non-fumeurs, contre 25% de fumeurs et 30% d'anciens fumeurs.

Les diagrammes circulaires, également appelés "camemberts", sont particulièrement adaptés pour représenter la distribution d'une variable qualitative, car ils permettent de visualiser rapidement les proportions relatives de chaque modalité.

5.2 Les diagrammes en bâtons ou à barres

Qui permettent de visualiser la répartition d'une variable catégorielle

Exemple :

Supposons que nous ayons collecté des données sur le sexe de patients dans un service hospitalier. Nous pourrions représenter ces données à l'aide d'un diagramme en barres :

Sexe	Effectif
Homme	48
Femme	52

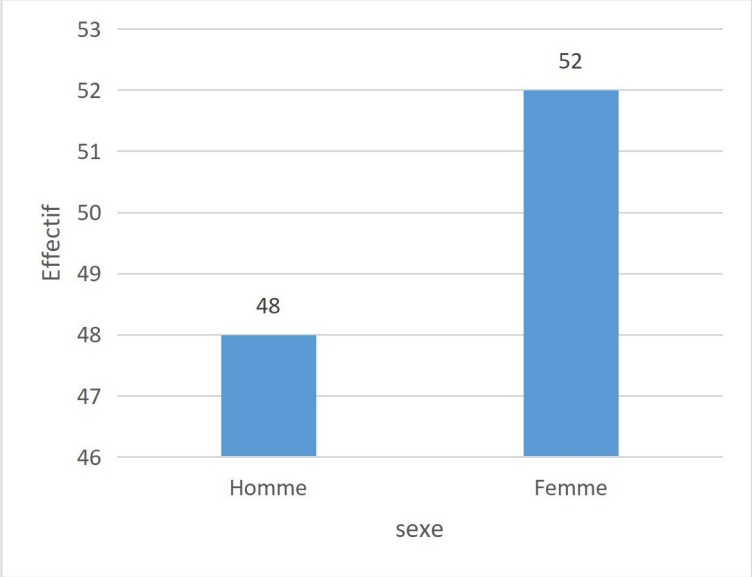


Figure 07 : Diagramme en bâtons

Ce diagramme en barres permet de visualiser facilement la répartition des patients par sexe. On peut ainsi voir que dans cet échantillon, il y a 48 hommes et 52 femmes.

Ce type de représentation graphique est très utile pour comprendre rapidement la composition d'une population selon une variable catégorielle comme le sexe, l'origine, la profession, etc

III. Statistique inférentielle

1. Définition des statistiques inférentielles

L'inférence statistique est un ensemble de techniques permettant d'induire les caractéristiques d'une population à partir d'un échantillon. Elle utilise des méthodes statistiques pour extrapoler des informations à partir de la théorie des probabilités et des modèles statistiques afin d'estimer les paramètres de la population. Contrairement aux statistiques descriptives, qui résument les propriétés d'une collection de données, les statistiques inférentielles donnent des estimations et des probabilités sur une population, ainsi que des intervalles de confiance et des tailles d'effet.

Les étapes clés de l'inférence statistique comprennent la formulation d'une hypothèse, la sélection d'une statistique d'écart et d'un seuil de décision, le calcul de la valeur de la statistique d'écart pour les valeurs observées, la comparaison à la valeur théorique de la statistique d'écart seuil choisi, et enfin, le calcul ou la lecture de la "p-value" associée pour prendre une décision concernant l'hypothèse nulle.

Les statistiques inférentielles donc sont essentielles pour tirer des conclusions fiables sur une population à partir d'un échantillon, en utilisant des méthodes probabilistes et statistiques pour estimer les paramètres de la population et tester des hypothèses.

2. Types de statistiques inférentielles

Les statistiques inférentielles sont divisées en deux catégories :

- **Tests d'hypothèses.**
- **Analyse de régression**.

Les chercheurs utilisent fréquemment ces méthodes pour généraliser les résultats obtenus sur de petits échantillons à des populations plus importantes. Examinons quelques-unes des méthodes disponibles en matière de statistiques inférentielles.

3. Test d'hypothèse

Un test d'hypothèse est une procédure statistique qui permet de décider entre deux hypothèses :

Hypothèse nulle (H_0) : suppose qu'il n'y a pas d'effet réel, toute différence observée est due au hasard.

Hypothèse alternative (H_1) : suppose qu'il existe un effet réel ou une différence.

Le test se fait à partir d'un échantillon de données, en suivant trois étapes principales :

a. Formuler H_0 et H_1.

b. Calculer une statistique de test (et une p-value).

c. Comparer le résultat au seuil de décision (α) pour accepter ou rejeter

3.1 Hypothèse nulle H_0

$$H_0 = ?$$

H_0 part de l'idée qu'il n'existe aucune différence réelle, donc toute différence observée est due **au hasard** d'échantillonnage.

Paramètre 1 = Paramètre 2

Exemple :

L'utilisation de deux antibiotiques et voire la plus efficace pénicilline, spiromycine

La question qui se pose c'est : Es ce que la pénicilline est plus efficace que la spiromycine ?

La réponse : L'hypothèse nulle (H_0) affirme qu'il n'existe pas de différence d'efficacité entre les deux antibiotiques. Toute différence observée serait alors simplement due au hasard d'échantillonnage.

Donc H_0 selon les tests sera

Comparaison	H_0	Exemple
Deux pourcentages	$p_1 = p_2$ (Les deux pourcentages sont pareils)	Le pourcentage d'oiseaux malades est le même dans 2 habitat différents (ville et foret)
Deux moyennes	$\mu_1 = \mu_2$ (Les deux moyennes sont les mêmes)	Poids moyen des oiseaux dans deux habitats différents est le même.
Un pourcentage obs. Avec H étendu (une valeurs fixe)	$p = \rho$ (Le % observé est égal au % attendu)	Le pourcentage observé des oiseaux infecté égale a la référence que 20 % des oiseaux soit infectés
Une moyenne obs. /a H étendu(valeur fixe)	$m = \mu$ (La moyenne observée = moyenne attendue)	Taille moyenne des œufs observé égale a 25 mm (valeur connue).
Plusieurs pourcentages (Chi ^{2})	$p_1 = p_2 = p_3 = p_4 = \cdots$ (Tous les % sont les mêmes)	Répartition des espèces est la même dans tous les types d'environnements.
Plusieurs moyennes	$\mu_1 = \mu_2 = \mu_3 = \ldots$ (Toutes les moyennes sont égales.)	Température des nids est identique pour tous les matériaux.
Corrélation	$r = 0$ (Il n'y a pas de relation entre les 2 variables)	Pas de lien entre taille du nid et nombre d'œufs.

3.2 L'hypothèse alternative H_1

H_1 suppose qu'il existe une différence réelle entre les groupes ou une relation entre variables.

H_1 : Les deux antibiotiques ayant une efficacité différente, un est meilleur par rapport a l'autre

Paramètre 1 ≠ Paramètre 2

On peut savoir mathématiquement exactement cette différence.

- ✓ Si $P_1 \neq P_2$ => bilatérale soit $P_1 > P_2$ ou $P_1 < P_2$
- ✓ Si $P_1 > P_2$ => unilatérale
- ✓ Si $P_1 < P_2$ => unilatérale

Donc : Dans un test statistique on compare toujours H_0 avec H_1 soit on accepte H_0 ou H_1

Conclusion H_0 et H_1:

Quand on compare deux groupes

On observe presque toujours une petite différence (parce que dans les données réelles, les valeurs ne tombent jamais parfaitement identiques).

Mais la vraie question est : D'où vient cette différence ?

Soit :

- ✓ **Du hasard**

Si elle est petite et expliquée par le hasard→ Dans ce cas, on garde H_0 : pas de vraie différence.

- ✓ **D'un effet réel**

Si la différence est trop grande pour être expliquée uniquement par le hasard, on conclut qu'il y a un vrai effet.

→ On rejette H_0 et on accepte qu'il existe une différence significative.

3.3 Le risque α

Si je dois étudier une comparaison entre deux hypothèses il faut savoir qu'il y a un risque d'erreur car :

- ✓ L'échantillon ne présent pas toute la population .
- ✓ L'observation est entachée d'erreur.

α Donc est la probabilité d'erreur lors de la réalisation d'un test statistique

C'est un risque d'erreur et c'est une probabilité

$$\mathbf{0 < \alpha < 1}$$

$$\mathbf{0\% < \alpha < 100\%}$$

Exemple :

A $>$ B avec α =0.03 c'est-à-dire il y a 3 chance sur 100 que A $>$ B soit faut

Le risque α c'est quand on affirme une différence alors qu'en réalité il n'y en a pas

⇨ α=probabilité de rejeté H_0, si H_0 est vrais

Donc : Le seuil α (alpha) est une limite fixée à l'avance (souvent 0,05 = 5 %) qui sert de critère de décision.

3.4 *p* value « petit p »

La *p*-value mesure à quel point ton résultat peut être expliqué par le hasard « Si en réalité il n'y a aucune différence (H_0), est-ce que le résultat que j'ai trouvé peut s'expliquer juste par hasard ? »

La *p*-value mesure donc la probabilité que la différence observée vienne seulement du hasard si H_0 est vraie.

p aussi le risque de ce trempé mais il est calculé après le test.

p est aussi une probabilité donc

$$\mathbf{\mathit{0<p<1}}$$

$$\mathbf{\mathit{0<p<100\%}}$$

Règle de décision :

α=5% c'est le plus grand risque d'erreur ont rejetant l'hypothèse nulle H_0

Si p<5% => H_0 rejeté , la différence est trop grande pour être expliquée par le hasard → on rejette H_0 (différence significative).

Si p>5% => la différence peut s'expliquer par le hasard

- ✓ P<0.05 } différence significatif
- ✓ P<0.0001 } différence très significatif } =>α=5%
- ✓ P<0.049 } différence significatif

3.5 Un degré de liberté (ddl)

L'idée des degrés de liberté (ddl), c'est de savoir combien de choix libres il nous reste quand certaines conditions sont imposées.

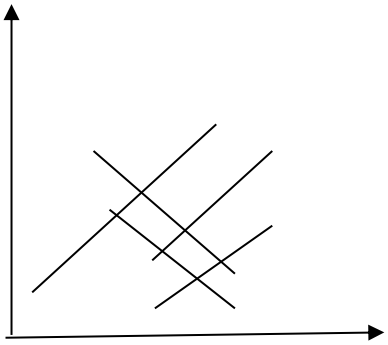
Il peut être défini de deux manières différentes :

- Le nombre de variables aléatoires qui ne peuvent être déterminées ou fixées par une équation.
- Le nombre d'observations moins le nombre de relations nécessaires entre ces observations

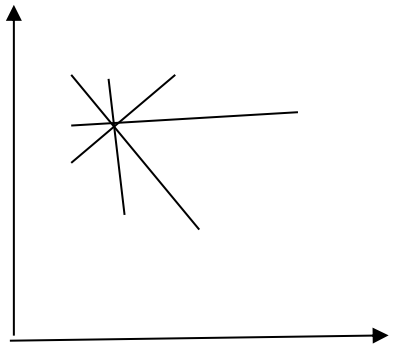
Exemple 01:

On veut comprendre comment les conditions influencent le nombre de droites possibles.

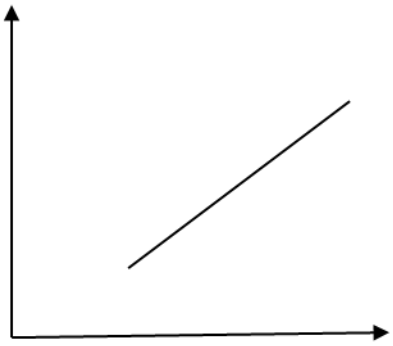
(1) Aucune condition ou limite pour tracé des lignes ; donc je peux tracer une droite ou plusieurs



(2) Un grand Nbr de possibilité mais nous avons perdu un degré de liberté « le point »



(3) On n'a pas d'autre possibilité une seul droite qui passe par deux points fixe



Exemple 2 :

(1) Si un effectif totale est égale a 56 ; somme nous libre pour proposer les effectifs de différentes modalité ?

Modalité	Effectifs
A	22
B	11
C	15
D	6
Totale	56

Je suis libre

Je ne suis pas libre

Valeur totale fixe

1) *ddl* = (n-1)= 4-1 =3

(2) Si les totaux des lignes et des colonnes sont déjà fixés, sommes-nous libres de remplir toutes les cases des tableaux suivants ?Combien de cases peut-on remplir librement ?

Modalité	Modalité	Totale
5	19	24
32	20	52
37	39	76

ddl =(2-1)(2-1)=**1**

Modalité	Modalité	Totale
15	9	24
19	33	52
3	73	76
37	115	152

ddl = (3-1)(2-1) = **2**

2) *ddl* = (C-1) (L-1)

(3) Si j'ai deux séries de valeurs $(ddl_{s1} + ddl_{s2})=(n_1-1)+(n_2-1) =n_1+n_2-2.$

QCM Comprendre les bases des tests statistiques:

Un chercheur veut savoir si un nouvel engrais augmente la croissance des plantes. Il réalise une expérience :

Il utilise deux groupes de plantes :

- Groupe A (sans engrais)
- Groupe B (avec engrais)

Après 4 semaines, il mesure la taille moyenne des plantes dans chaque groupe.

Il veut comparer les moyennes pour savoir si l'engrais a un effet réel.

Questions : Une seule bonne réponse par question

1. Quelle est l'hypothèse H_0 dans cette étude ?

A. L'engrais améliore la croissance
B. Il y a une différence entre les deux groupes
C. L'engrais n'a pas d'effet
D. L'engrais est meilleur que l'eau

2. Que cherche à montrer le chercheur avec l'hypothèse H_1 ?

A. Que les plantes sont malades
B. Que l'engrais a un effet sur la croissance
C. Que les plantes n'ont pas poussé
D. Que l'eau est suffisante

3. Le niveau alpha ($\alpha = 0,05$) signifie que :

A. On fait l'étude 5 fois
B. On accepte de se tromper 5 fois sur 100 si on rejette H_0
C. Il faut que p soit égal à 0,05
D. On teste 5 plantes dans chaque groupe

4. Si la valeur p obtenue est 0,02, que doit faire le chercheur ?

A. Rejeter H_0 : l'engrais a probablement un effet
B. Garder H_0 : il n'y a pas d'effet
C. Refuser de conclure
D. Augmenter alpha à 0,10

5. Pourquoi peut-on rejeter H_0 ici ?

A. Parce que p > α
B. Parce que p < α
C. Parce que les plantes sont grandes
D. Parce que H_1 est toujours vraie

6. Si p avait été de 0,10 au lieu de 0,02, que se passerait-il ?

A. On ne rejette pas H_0
B. On rejette H_1
C. On conclut qu'il y a un effet
D. L'expérience est annulée

7. Le chercheur utilise un test statistique. À quoi servent les degrés de liberté dans ce test ?

A. À choisir les plantes
B. À mesurer la croissance
C. À trouver la bonne valeur dans la table du test
D. À déterminer la dose d'engrais

8. Quel est le rôle de la valeur p dans ce test ?

A. Elle mesure la taille des plantes
B. Elle indique si les plantes sont bien arrosées
C. Elle sert à comparer avec alpha pour rejeter ou non H_0
D. Elle remplace les moyennes

9. Si on avait choisi α = 0,01, le résultat (p = 0,02) serait-il significatif ?

A. Oui, car 0,02 est petit
B. Non, car 0,02 est plus grand que 0,01
C. Oui, car p est toujours significatif
D. Oui, si les plantes sont assez nombreuses

10. Pourquoi est-il important de formuler clairement H_0 et H_1 au début ?

A. Pour décorer l'étude
B. Pour avoir plus de données
C. Pour savoir ce qu'on teste et comment interpréter les résultats
D. Pour connaître la taille des plantes à l'avance

4. Introduction aux lois de distribution : loi normale

4.1 Loi normale

Également connue sous le nom de distribution normale, est l'une des lois de probabilité les plus utilisées en statistique et en théorie des probabilités pour modéliser des phénomènes naturels issus de plusieurs événements aléatoires. Elle est caractérisée par sa forme en cloche, sa symétrie et le fait que sa moyenne et sa médiane sont égales. La courbe de densité de probabilité de la loi normale est définie par deux paramètres : sa moyenne (μ) et son écart type (σ). Lorsqu'une variable aléatoire suit une loi normale, elle est dite gaussienne ou normale.

4.2 Fonction de Laplace-Gauss

Aussi appelée fonction de répartition de la loi normale, est la fonction qui permet de calculer la probabilité qu'une variable aléatoire suivant une loi normale prenne une valeur inférieure ou égale à une valeur donnée.

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

x est la variable aléatoire suivant une loi normale

π est le constant mathématique pi (environ 3,14159)

e est la base des logarithmes népériens (environ 2,71828)

a) Démontrer que la fonction *f* est paire

f(x) est paire si *f (-x)=f(x)*

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

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}}$$

$$= f(x)$$

Donc paire => la courbe *f(x)* est symétrique par rapport à l'axe des ordonnées.

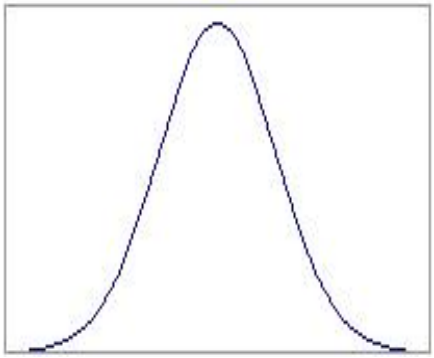


Figure 08 : Courbe cloche symetrique

b) La variation de *f* sur $\mathbb{R}$

f est dérivable sur $\mathbb{R}$ =>e^{μ} est dérivable sur $\mathbb{R}$

$$f(x)= \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}} \quad => f(x) = \frac{1}{\sqrt{2\pi}} . \frac{-2x}{2} e^{-\frac{(x)^2}{2}} = - \frac{1}{\sqrt{2\pi}} x . e^{-\frac{(x)^2}{2}}$$

$$f(0)= \frac{1}{\sqrt{2\pi}} . 1 = \frac{1}{\sqrt{2\pi}}$$

$$\lim_{x\to+\infty} \frac{-x^2}{2} = -\infty$$

$$\lim_{X\to-\infty} e^X = 0$$

Donc $\lim_{x\to+\infty} e^{\frac{-x^2}{2}} = 0$

x	$-\infty$		$+\infty$
$\sqrt{2\pi}$	+		+
$-x$	+		-
$\frac{-x^2}{2}$	+		+
f'	+		-
f	0 ↗	$\frac{1}{\sqrt{2\pi}}$	↘ 0

4.3 La fonction de densité de la loi normale

Une variable X qui suit une loi normale

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

x est la variable aléatoire

μ est la moyenne de la loi normale

σ est l'écart-type de la loi normale

π est le constant mathématique pi (environ 3,14159)

e est la base des logarithmes népériens (environ 2,71828)

Cette formule permet de définir complètement une loi normale, en fonction de ses deux paramètres : la moyenne μ et l'écart-type σ.

a) Quelques propriétés de cette fonction de densité :

- Elle est positive sur tout $\mathbb{R}$
- Son maximum est atteint pour $x = \mu$ et vaut $1/(\sigma\sqrt{(2\pi)})$
- Elle est symétrique par rapport à la droite $x = \mu$
- L'intégrale de $f(x)$ sur tout $\mathbb{R}$ vaut 1

b) Démontré que la fonction f est paire

$f(x)$ est paire => $f(-x)=f(x)$

$$f(-x)=\frac{1}{\sqrt{2\pi}}e^{\frac{-(-x)^2}{2}} \quad =\frac{1}{\sqrt{2\pi}}e^{\frac{-(x)^2}{2}} \quad =f(x)$$

Donc paire =>la courbe a *f* est symétrique par rapport à l'axe des ordonnées.

c) La variation de *f* sur IR

f est dérivable sur IR => e^{μ} est dérivable sur IR

$$f(x)= \frac{1}{\sqrt{2\pi}} e^{\frac{-(x)^2}{2}} \Rightarrow f'(x)= \frac{1}{\sqrt{2\pi}}\left(-\frac{1}{2}.2x\right)e^{\frac{-(-x)^2}{2}}= \frac{1}{\sqrt{2\pi}}(-x)e^{\frac{-(-x)^2}{2}}$$

$$f(0)= \frac{1}{\sqrt{2\pi}}.1 = \frac{1}{\sqrt{2\pi}}$$

$$\lim_{x\to+\infty} -\frac{x^2}{2} = -\infty$$

$$\lim_{x\to-\infty} e^x = 0$$

- μ => 50% la loi normale est symétrique autrement dit, la probabilité que la variable aléatoire X soit inférieure ou égale à la moyenne μ est de 50%, tout comme la probabilité qu'elle soit supérieure ou égale à μ.
- μ => 95% des valeurs comprise entre -2σ et +2σ Cela signifie que la probabilité que la variable aléatoire X soit comprise dans l'intervalle allant de μ - 2σ à μ + 2σ est d'environ 95%.
- μ => 99.7% des valeurs sont comprises entre -3σ et +3σ
- Il y peu de chance qu'un individu s'écarte de l moyenne +3σ

5. Vérification de l'adéquation à la loi normale

La loi normale est un concept en statistiques qui décrit comment certaines valeurs ou événements se distribuent autour d'une moyenne. On l'appelle aussi la courbe en cloche parce que, si tu dessines la distribution des données, elle ressemble à une cloche : la majorité des valeurs se trouve autour de la moyenne, et plus tu t'éloignes de cette moyenne, moins il y a de valeurs.

Exemples :

- ✓ La taille des êtres humains suit souvent une distribution normale : la plupart des gens mesurent autour de la taille moyenne (par exemple, 170 cm), et plus on s'éloigne de cette moyenne (soit très grand, soit très petit), moins il y a de gens.

- ✓ Les notes d'examens dans une grande classe peuvent aussi suivre une loi normale : beaucoup d'étudiants ont une note proche de la moyenne, et peu d'étudiants ont des notes extrêmement basses ou extrêmement hautes

5.1 Pourquoi utilise-t-on la loi normale ?

La loi normale est utilisée parce qu'elle modélise bien de nombreux phénomènes naturels et sociaux. Voici quelques raisons :

- **Facilité de calcul :** Elle permet de faire des prédictions faciles sur les données.
- **Applications pratiques :** Beaucoup de choses dans la nature, les sciences sociales et l'économie suivent approximativement une loi normale. Par exemple, les mesures de poids, de taille, ou les notes scolaires.
- **Théorème central limite :** Ce théorème dit que si tu fais la moyenne d'un grand nombre de variables aléatoires indépendantes, cette moyenne va suivre une loi normale, même si les données de départ ne sont pas normales. C'est un principe fondamental en statistiques.

5.2 Comment vérifier si nos données suivent une loi normale ?

Pour savoir si tes données suivent une loi normale (c'est-à-dire si elles sont bien réparties autour d'une moyenne), on peut faire ce qu'on appelle un test de conformité.

Le test de conformité à la loi normale, également connu sous le nom de test de normalité, est une méthode statistique utilisée pour déterminer si un échantillon de données suit ou non une distribution normale (ou gaussienne)

5.2.1 Par les graphiques

L'histogramme est un graphique qui montre comment tes données sont réparties. S'il ressemble à une cloche symétrique, tes données sont probablement normales.

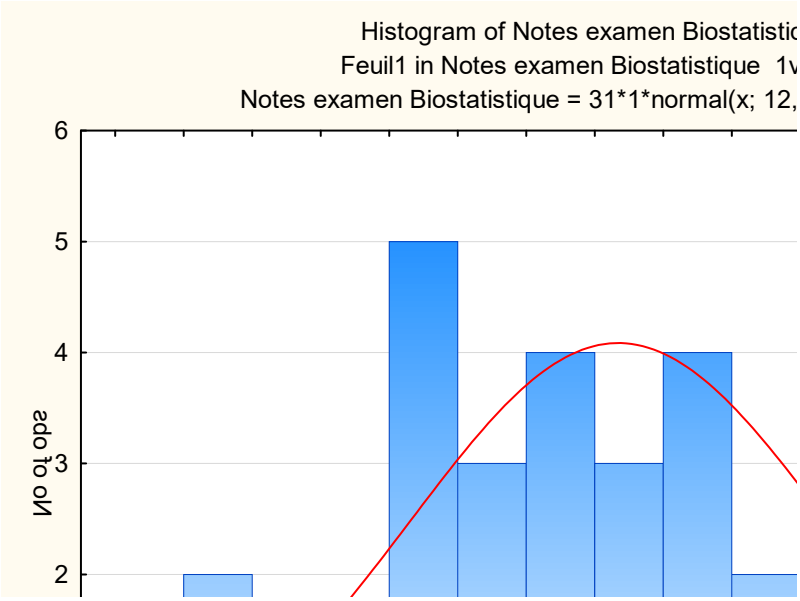


Figure 09 : Histogramme présente des données suivent une loi normale

5.2.2 Par les tests statistiques :

- **Formuler l'hypothèse nulle H_0: les données suivent une loi normale**
- **Calculer la valeur de test et la p-value associée**

A) Test de Shapiro-Wilk

Le test de Shapiro-Wilk est un test simple et très souvent utilisé pour vérifier si tes données suivent une loi normale. C'est particulièrement utile pour les petits ensembles de données (typiquement moins de 50 observations).

✓ **Comment ça marche ?**

Ce test calcule une statistique (qu'on appelle W).

A) Si W est proche de 1, cela signifie que les données sont très proches d'une loi normale.

B) Si W est loin de 1, cela signifie que les données ne suivent probablement pas une loi normale.

C) ***p*-valeur :** Si la p-valeur est supérieure à 0,05 on accepte donc H_0 et les données suivent une loi normale. Si elle est inférieure à 0,05 on rejette H_0 et donc les données ne suivent probablement pas une loi normale.

Exemple

Tu fais le test de Shapiro-Wilk sur un ensemble de données avec 30 mesures de poids d'étudiants :

P-valeur = 0,07 : Tu peux dire que les poids des étudiants suivent probablement une loi normale (car 0,07 > 0,05).

P-valeur = 0,02 : Tu rejettes l'idée que les poids des étudiants suivent une loi normale (car 0,02 < 0,05).

B) Test de Kolmogorov-Smirnov (K-S)

L'un des tests les plus utilisés pour vérifier l'adéquation d'une variable statistique à une loi de probabilité théorique, en particulier la loi normale.

En 1933, le mathématicien Kolmogorov a proposé une méthode pour vérifier si des données suivent une distribution donnée, comme la loi normale.

En 1939, Smirnov, sans savoir ce que Kolmogorov avait fait, a proposé une méthode très similaire, mais pour comparer deux groupes de données.

Ils se sont rencontrés par hasard... et ont fusionné leurs idées ! Le test de Kolmogorov-Smirnov est né.

Voici les principales caractéristiques de ce test :

- ✓ **Hypothèses :**

H_0 : la variable suit la loi normale $f_0(x)$

H_1 : la variable ne suit pas la loi normale $f_0(x)$

✓ **Statistique de test :**

$$D = sup\ |f_n(x) - f_0(x)|$$

D la mesure de la plus grande différence entre les deux courbes

La DCE (distribution cumulative empirique) vient des données réelles.

La DCT (distribution cumulative théorique) vient d'une loi mathématique (ex. : normale).

$f_n(x)$ étant la fonction de répartition empirique de l'échantillon

$f_0(x)$ étant la fonction de répartition théorique de la loi normale

Ce test mesure donc la plus grande différence entre la distribution réelle de tes données et la distribution théorique (ici, une loi normale).

S'il y a une grande différence entre tes données et la loi normale, le test te dira que tes données ne suivent pas la loi normale.

✓ **Règle de décision :**

Si la p-value associée à la valeur de D est inférieure au seuil choisi (généralement 5%), alors comme dans tout test statistique :

H_0 : il n'y a pas de différence entre les distributions

H_1 : il y a une différence significative

- Si $p > 0,05 \rightarrow$ on accepte H_0, les distributions sont proches
- Si $p < 0,05 \rightarrow$ on rejette H_0, les distributions sont différentes

✓ **Avantage :**

Ne nécessite pas de calcul préalable des paramètres de la loi normale, robuste aux valeurs aberrantes, applicable à des échantillons de petite taille

Ce test de Kolmogorov-Smirnov est donc un outil puissant et très utilisé en pratique pour vérifier si une variable suit bien une loi normale.

Exemple

Tu fais le test de Kolmogorov-Smirnov sur un grand ensemble de données (par exemple, les revenus de 1 000 personnes) :

- *p*-valeur = 0,08 : Les revenus suivent probablement une loi normale.
- *p*-valeur = 0,01 : Les revenus ne suivent pas une loi normale.

C) Test du Chi-deux (ou Chi-carré)

Le test du Chi-deux est un peu différent des deux autres. Il ne compare pas directement tes données avec une loi normale, mais il regarde si la fréquence des classes dans un tableau correspond aux fréquences attendues d'une loi normale. C'est très utile pour les données catégoriques.

✓ **Comment ça marche ?**

On divise les données en classes : Par exemple, on peut diviser les tailles d'étudiants en différentes classes (entre 150 cm et 160 cm, entre 160 cm et 170 cm, etc.).

On calcule combien de valeurs tombent dans chaque classe.

On compare le nombre d'observations réelles dans chaque classe au nombre que l'on s'attendrait à avoir si les données suivaient une loi normale.

✓ **Règle de décision :**

Statistique du Chi-deux : Plus cette statistique est élevée, plus il y a de différences entre tes données et la loi normale attendue.

***p*-valeur :** Comme pour les autres tests, si la p-valeur est supérieure à 0,05, tes données suivent probablement une loi normale. Si elle est inférieure à 0,05, tes données ne suivent probablement pas une loi normale.

Exemple

Tu fais le test du Chi-deux sur la distribution des tailles d'étudiants dans une classe :

P-valeur = 0,10 : Les tailles suivent probablement une loi normale.

P-valeur = 0,03 : Les tailles ne suivent probablement pas une loi normale.

Exemple générale :

Imaginons que nous disposons d'une liste des notes d'examen en biostatistique de 31 étudiants (le tableau des données présenté dans la vidéo des travaux pratiques)

- **Hypothèse nulle (H_0) :** Les notes d'examen en biostatistique suivent une distribution normale.
- **Hypothèse alternative (H_1) :** Les notes d'examen en biostatistique ne suivent pas une distribution normale.

En appliquant les tests, nous obtenons les résultats suivants dans le tableau.

Caractéristiques du test	Shapiro-Wilk	Kolmogorov-Smirnov	Chi-deux
Statistique de test	W=0.9844	D=0.0673	χ^2=0 ,82780
P value	0.9190	0.9972	0,36291
Interprétation	• W est proche de 1, cela signifie que les données sont très proches d'une loi normale. • La valeur p est bien supérieure à 0.05, ce qui signifie qu'on accepte l'hypothèse nulle. **Les données suivent la loi normale**.	• D=0.0673 cela signifie que la différence maximale entre les deux distributions est de 6,73 %. Donc D est faible, les données sont proches de la distribution Normale • La valeur p est bien supérieure à 0.05, ce qui signifie qu'on accepte l'hypothèse nulle. **Les données suivent la loi normale**	• Une valeur de χ^2 faible (ici, 0.82780) signifie que les fréquences observées sont très proches des fréquences attendues. • La valeur p est bien supérieure à 0.05, ce qui signifie qu'on accepte l'hypothèse nulle. **Les données suivent la loi normale**

6. Test du chi-deux

Le test du Khi-deux est un outil statistique utilisé quand on travaille avec des catégories (comme : homme/femme, espèce A/espèce B, oui/non...).

Il sert à répondre à une question simple :

Est-ce que les résultats qu'on observe sont dus au hasard ou à une vraie différence ?

Exemple : Imaginons que tu observes 30 papillons sur 3 types de fleurs différentes :

Fleurs	Nombre de papillons
Fleur A	10
Fleur B	10
Fleur C	10

Ici, les papillons semblent aimer toutes les fleurs pareil.

Mais maintenant regarde cet autre cas :

Fleurs	Nombre de papillons
Fleur A	20
Fleur B	5
Fleur C	5

On se demande alors :

- ✓ Est-ce que les papillons préfèrent vraiment la fleur A ?
- ✓ Ou est-ce juste une coïncidence ?

Le test du Khi-deux permet donc de tester statistiquement si cette différence est significative (réelle) ou si elle peut être expliquée par le hasard.

Le test du Khi-deux (χ^2) est un test statistique utilisé pour vérifier :

- ✓ L'indépendance entre deux variables (test d'indépendance),
- ✓ La conformité d'une distribution observée à une distribution théorique (test d'ajustement),
- ✓ La similarité entre des distributions de groupes différents (test d'homogénéité).

La valeur du khi-deux est calculée de la manière suivante:

Formule :

$$\chi^2 = \sum \frac{(oi - ei)^2}{ei}$$

o_i sont les effectifs observés

e_i sont les effectifs théoriques attendus sous H_0.

Le test du khi-deux a plusieurs applications. Il peut être utilisé pour répondre aux questions suivantes:

- ✓ **Test d'indépendance :** Deux variables catégorielles sont-elles indépendantes l'une de l'autre?
- ✓ **Test de distribution :** Les valeurs observées de deux variables catégorielles sont-elles égales aux valeurs attendues?
- ✓ **Test d'homogénéité :** Deux échantillons ou plus sont-ils issus de la même population?

6.1 Test d'indépendance du Chi-deux

Des chercheurs vont dans 50 zones de sol (25 calcaire, 25 acide). Dans chaque zone, ils regardent s'ils trouvent ou non une plante rare.

Ils notent pour chaque zone :

- ✓ Si la plante est là → "Plante présente"
- ✓ Si la plante n'est pas là → "Plante absente"

C'est ce qu'on appelle une variable catégorielle binaire (oui/non).

	Sol calcaire	**Sol acide**	**Total**
Plante présente	8	12	20
Plante absente	17	13	30
Total	25	25	50

Cela signifie :

- ✓ Sur 25 zones de sol calcaire, ils ont trouvé 8 fois la plante et 17 fois non.
- ✓ Sur 25 zones de sol acide, ils ont trouvé 12 fois la plante et 13 fois non.

Pour répondre à la question :

Est-ce que la présence de la plante dépend du type de sol ?

- ✓ Si oui → alors certaines plantes préféreraient un type de sol.
- ✓ Si non → alors la plante pousse au hasard, sans préférence de sol.

Le test du Khi-deux compare ce qu'on observe avec ce qu'on attendrait par hasard si le type de sol n'avait aucun effet.

A. Étapes du calcul :

H_0 : Il n'y a pas de relation entre le type de sol et la présence de la plante.

Autrement dit, la présence (ou l'absence) de la plante est indépendante du type de sol.

- **Calcul des fréquences attendues**

Formule :

$$Frequence\ attendue = \frac{(Totale\ ligne \times Totale\ colonne)}{Totale\ générale}$$

$$E\ (plante\ présente - sol\ calcaire) = \frac{(20\times25)}{50} = 10;$$

$$E\ (plante\ présente - sol\ acide) = \frac{(20\times25)}{50} = 10;$$

$$E\ (plante\ absente - sol\ calcaire) = \frac{(30\times25)}{50} = 15;$$

$$E\ (plante\ absente - sol\ acide) = \frac{(30\times25)}{50} = 15;$$

Donc :

	Sol calcaire	**Sol acide**
Plante présente	10	10
Plante absente	15	15

- **Calcul de la statistique du Chi-deux**

Formule :

$$\chi^2 = \sum \frac{(oi - ei)^2}{ei}$$

$$\chi^2 = \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(17-15)^2}{15} + \frac{(13-15)^2}{15} = 1.34$$

- **Interprétation :**

On compare cette valeur à une valeur critique du Khi-deux (selon le nombre de degrés de liberté).

✓ Si χ^2 est grand, il y a une différence significative.

✓ Si χ^2 est petit (comme ici), alors pas de preuve d'une relation entre sol et plante.

Pour calculer le degré de liberté (ddl) dans un test d'indépendance du Khi-deux, on utilise la formule suivante :

$$\boldsymbol{ddl = (L-1)\ (C-1)}$$

L = nombre de lignes du tableau (hors totaux): Il y a 2 lignes (plante présente / plante absente)

C = nombre de colonnes du tableau (hors totaux): Il y a 2 colonnes (sol calcaire / sol acide)

Donc : $\boldsymbol{ddl = (2-1)\ (2-1) = 1}$

On compare la valeur de χ^2 calculé ou dite observé à la valeur critique dans la table du khi-deux à 1 degré de liberté et un seuil α = 0,05 qui est 3,48 (Tableau (I) annexe)

$1{,}34 < 3{,}84$, on ne rejette pas H_0

Donc dans cet exemple, le test du Khi-deux montre que la présence de la plante rare ne dépend pas significativement du type de sol (calcaire ou acide).

Autrement dit, elle pousse à peu près autant sur les deux types de sol.

6.2 Test de distribution

Le test de distribution du Khi-deux, ou test d'adéquation, vérifie si les fréquences des valeurs caractéristiques individuelles dans l'échantillon correspondent aux fréquences d'une distribution définie. Dans la plupart des cas, cette distribution définie correspond à celle de la population. Dans ce cas, on vérifie si l'échantillon provient de la population correspondante.

Exemple :

Répartition d'insectes dans une prairie

Insecte	Observé	Proportion attendue	Fréquence attendue
Coccinelle	30	0.25	25
Criquet	20	0.20	20
Punaise	15	0.25	25
Papillon	35	0.30	30
Total	100		100

- **Calcul de la statistique du Chi-deux**

Formule :

$$\chi^2 = \sum \frac{(oi - Ei)^2}{Ei}$$

où :

oi : fréquence observée pour la catégorie i

$\boldsymbol{Ei = n \times pi}$: fréquence attendue selon la proportion pi

$$\chi^2 = \frac{(30-25)^2}{25} + \frac{(20-20)^2}{20} + \frac{(15-25)^2}{25} + \frac{(35-30)^2}{30} = 5{,}83$$

- **Interprétation :**

$$\boldsymbol{ddl = n\text{-}1} = 4\text{-}1 = 3 \; ; \; \boldsymbol{\alpha = 0{,}05} \rightarrow \chi^2 \text{ critique} = 7.815$$

5,83 <7.815 H_0 accepté il n'y a pas une différence entre les fréquence observé et les fréquence attendu

Donc l'échantillon provient de la population correspondante.

6.3 Test d'homogénéité du Chi-deux

Le test d'homogénéité du Khi-deux peut être utilisé pour vérifier si deux échantillons ou plus proviennent de la même population ou pour comparer les distributions de plusieurs groupes pour savoir si elles sont identiques.

Exemple : Type de nid selon l'habitat

Habitat	Nid au sol	Nid en arbre	Nid en bâtiment	Total
Forêt	12	18	0	30
Agricole	6	10	4	20
Urbain	0	5	25	30
Total	18	33	29	80

✓ **Fréquences attendues :**

(ex. Forêt / Nid au sol) : $Frequence\ attendue = \frac{(Totale\ ligne \times Totale\ colonne)}{Totale\ générale}$

$$E = \frac{(30 \times 18)}{80}$$

= 6,75

on fait ça pour chaque cellule

Habitat	Nid au sol	Nid en arbre	Nid en bâtiment	Total
Forêt	**6,75**	**12,375**	**0**	30
Agricole	**4,5**	**8,25**	**7,25**	20
Urbain	**0**	**12,375**	**10,875**	30
Total	18	33	29	80

- **Calcul de la statistique du Chi-deux**

Formule :

$$\chi^2 = \sum \frac{(oi - Ei)^2}{Ei}$$

Cellule	O	E	$\chi^2 = \sum \frac{(oi - Ei)^2}{Ei}$
Forêt / Sol	12	6.75	$\chi^2 = \sum \frac{(12-6,75)^2}{6,75} = 4.07$
Forêt / Arbre	18	12.375	$\chi^2 = \sum \frac{(18-12,375)^2}{12,375} = 2.55$
Forêt / Bâtiment	0	10.875	$\chi^2 = \sum \frac{(0-10,875)^2}{10,875} = 10.89$
Agricole / Sol	6	4.5	$\chi^2 = \sum \frac{(oi-Ei)^2}{Ei} = 0.5$
Agricole / Arbre	10	8.25	$\chi^2 = \sum \frac{(10-8,25)^2}{8,25} = 0.37$
Agricole / Bâtiment	4	7.25	$\chi^2 = \sum \frac{(4-7,25)^2}{7,25} = 1.46$
Urbain / Sol	0	6.75	$\chi^2 = \sum \frac{(0-6,75)^2}{6,75} = 6.75$
Urbain / Arbre	5	12.375	$\chi^2 = \sum \frac{(5-12,375)^2}{12,375} = 4.4$
Urbain / Bâtiment	25	10.875	$\chi^2 = \sum \frac{(25-10,875)^2}{10,875} = 17.61$

- **Addition finale**

$$\chi 2 = 4.07 + 2.55 + 10.89 + 0.5 + 0.37 + 1.46 + 6.75 + 4.4 + 17.61 = 48.6$$

- **Degrés de liberté**

$$ddl=(L-1)\ (C-1)$$

$$ddl=(3-1)\times(3-1)=2\times2=4$$

- **Interprétation :**

$$ddl = 4 ; \alpha = 0,05 \rightarrow \chi^2 \text{ critique} = 9,488$$

48.6 > 9,488 H_0 rejeté il y a une différence entre les fréquence observé et les fréquence attendu

Donc : La répartition des types de nids dépend de l'habitat

6.4 Taille de l'effet dans le test du Chi-deux

Jusqu'à présent, nous savons seulement si nous pouvons rejeter l'hypothèse nulle ou non, mais il est souvent très intéressant de savoir quelle est la force de la relation entre les deux variables. On peut répondre à cette question à l'aide de l'intensité de l'effet.

Dans le test du Khi-deux, le V de Cramers peut être utilisé pour calculer l'ampleur de l'effet. Ici, une valeur de 0,1 est petite, une valeur de 0,3 est moyenne et une valeur de 0,5 est grande.

QCM – Loi normale, Tests de normalité et Chi-deux

1. Une variable suit une loi normale si :

A. Sa moyenne est égale à 0

B. Sa distribution est symétrique et en forme de cloche

C. Sa médiane est toujours supérieure à sa moyenne

D. Elle est qualitative

2. Le test de Shapiro-Wilk est utilisé pour :

A. Comparer deux moyennes

B. Vérifier si une variable suit une loi normale

C. Tester l'égalité de deux variances

D. Tester une corrélation

3. Quel test est recommandé pour vérifier la normalité sur un grand échantillon (> 50 individus) ?

A. Test de Mann-Whitney

B. Test de Shapiro-Wilk

C. Test de Kolmogorov-Smirnov

D. Test de Wilcoxon

4. Le test du Khi-deux peut-il être utilisé pour tester la normalité ?

A. Oui, si on regroupe les données en classes

B. Non, jamais

C. Oui, mais seulement pour des variables qualitatives

D. Oui, uniquement si la moyenne est nulle

5. Le test du Khi-deux d'ajustement est utilisé pour :

A. Comparer deux moyennes

B. Tester si une variable suit une distribution théorique

C. Comparer des variances

D. Tester la relation entre deux variables

6. Le test du Khi-deux d'indépendance permet de :

A. Savoir si deux variables quantitatives sont corrélées

B. Savoir si deux variables qualitatives sont liées

C. Tester la normalité d'une variable

D. Vérifier la variance

7. Le test du Khi-deux d'homogénéité permet de :

A. Comparer des distributions entre plusieurs groupes

B. Tester une moyenne

C. Évaluer la normalité

D. Mesurer une corrélation

8. Quel est le principe général du test du Khi-deux ?

A. Comparer les variances observées

B. Comparer les fréquences observées et attendues

C. Comparer les moyennes

D. Estimer une pente de régression

9. Quelle mesure est utilisée pour évaluer la taille de l'effet du Khi-deux ?

A. r de Pearson

B. d de Cohen

C. V de Cramer

D. t de Student

10. Que signifie une valeur de V de Cramer proche de 0 ?

A. L'effet est fort

B. Il n'y a pas de relation entre les variables

C. L'échantillon est trop grand

D. Les fréquences attendues sont erronées

Devoir :

✓ **Objectifs du Devoir**

✧ Comprendre le concept de normalité dans les distributions de données.
✧ Appliquer le test de normalité à un ensemble de données en utilisant le logiciel Statistica.
✧ Interpréter les résultats du test de normalité et tirer des conclusions basées sur ces résultats.

✓ **Instructions:**

✧ Choix des Données : vos notes de semestre précédent
✧ Réalisation du Test de Normalité : Utilisez le test de Shapiro-Wilk ou le test de Kolmogorov-Smirnov pour vérifier la conformité à la loi normale.

✓ **Interprétation des Résultats :**

✧ Examinez les résultats fournis par Statistica, notamment la valeur p et les statistiques de test.
✧ Rédigez une brève interprétation de ces résultats : Que signifie la valeur p obtenue ? Êtes-vous en mesure de rejeter ou de ne pas rejeter l'hypothèse nulle de normalité ?

7. Comparaison de deux moyennes

NB : tous les tests statistiques qu'ils que soit calculent la probabilité de se tromper en rejetant H_0 et le petit p

7.1 Comparaison de deux moyennes de deux séries indépendantes

7.1.1 Séries indépendantes

A) Test Z « ε »

Le test Z est un test statistique utilisé pour comparer les moyennes de deux groupes pour déterminer si les deux groupes présentent des différences significatives.

Exemple : La concentration sanguine en vitamine D entre les fumeurs et les non-fumeurs

Non-fumeurs σ_1=8.3 n_1=97 m_1=23.6mg/ml

Fumeur σ_2=7.6 n_2=85 m_2=20.9mg/ml

⇨ **Problème :** la moyenne de la [D] est-elle réellement différente entre fumeurs et non-fumeurs ?

H_0 : vit D_f= vit D_{nf}

H_1 : vit $D_f \neq$ vit D_{nf}

Si H_0 est vrai => m_1=m_2 =>Δ= m_1-m_2 ≈0

Δ=23.6-20.9=2.7

Est-ce que 2.7 est suffisamment éloigné de « 0 » ?

Pour qu'on rejet H_0 :

Théorème : La différence de deux variables normales est une variable normale

1) La variable Δ= m1-m2 suit une loi normale
2) ε suit une loi normale centrée réduite

-1.96< ε<1.96 => si| ε | >1.96 ou rejet H_0 ; P<5%

$$\boldsymbol{\varepsilon=\frac{\Delta}{écartype(m_1-m_2)}=\frac{|\ m_1-m_2\ |}{écartype(m_1-m_2)}}$$

Théorème : la variance de la différence de deux variables est égale à la somme des deux variances.

Donc :

$$\sigma_\Delta{}^2=\sigma_{m1}{}^2+\sigma_{m2}{}^2$$

$$\sigma_{m1}{}^2=\frac{\sigma_1{}^2}{n_1}$$

$$\sigma_{moyenne}{}^2 => \ \sigma_{m2}{}^2=\frac{\sigma_2{}^2}{n_2} \qquad =>\sigma_\Delta{}^2=\frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2}$$

$$\varepsilon=\frac{\Delta}{2\sigma_\Delta}=\frac{|\ m_1-m_2|}{\sqrt{\frac{\sigma_1{}^2}{n1}+\frac{\sigma_2{}^2}{n2}}}=\frac{|23.6-20.9|}{\sqrt{\frac{(8.3)^2}{9.7}+\frac{(7.6)^2}{85}}}=2.29$$

2.29>1.96=> H_0 rejeté H_1 accepté (significatif ou non ?); p<3% donc significatif

Table de |Z| alpha Bilatérale

$\lvert\varepsilon\rvert$	2.5	2.33	2.17	2.05	1.96	1.64	1.28
α	1%	2%	3%	4%	5%	10%	20%

3) La concentration [D]différente significativement entre fumeur et non-fumeur

Conditions d'application :

4) Taille de deux échantillons >30
5) Taille de deux échantillons < 30=> t student

B) Test t de student

On estime la variance des moyennes par la variance commune des 2 échantillons

$$\sigma^2 = \frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{n_1+n_2-2}$$

$$=>t = \frac{|m_1-m_2|}{\sqrt{\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}}} \qquad ddl = n_1+n_2-2$$

Exemple :

$n_1=15$ $m_1=1.81$ $\sigma_1=0.50$ non traité

$n_2=12$ $m_2=1.41$ $\sigma_2=0.39$ traité

$\sigma^2=\frac{(15-1)(0.5)^2+(12-1)(0.39)^2}{15+12+2}=0.21$

$t=\frac{|1.81-1.41|}{\sqrt{\frac{0.21}{15}+\frac{0.21}{12}}}=2.27$; ddl=15+112-2=25 →tableau => p<0.04 (Tableau II annexe)

$t_{obs} \leq t_t$ => donc H_1 accepté =>≠ significatif

Application t student : variance égales ($v_1/v_2>3$)

7.1.2 Comparaison de deux moyennes de deux séries appariées

Des séries de même taille il s'agit des mesures sur les mêmes individus

Exemple :

Volume des GR avant et après le travail dans un usine de production des produits chimique moyenne de GR de chaque sérié

Avant travaille = $m_A=94.7\mu^3$

Après travaille = $m_b=93.6\ \mu^3$ => période bien définie

La différence des différences de chaque individu $m_d=1.14$

Problème => es ce que le volume de GR avant et après le travail est-il en moyenne identique pour chaque individus ?

H_0 : la moyenne de différence est identique

H_1 : la moyenne de différence est différente

C'est-à-dire

$$H_0 = md = 0$$

$$H_1 = md \neq 0$$

Donc dans notre exemple es ce que 1.14 est suffisamment éloigne de zéro ?

Si oui => accepte H_1

Si non => rejet H_1

Les mêmes tests mais calculé diffirament

A) Test Z « ε »

$Z = \frac{|m_d - 0|}{écortype\ de\ m_d}$ suit un loi normale centrée réduit

$\sigma^2_{md} = \frac{\sigma^2_d}{n}$ d = valeur avant – valeur aprée

$$\sigma^2_d = \frac{\sum {d_i}^2 - \frac{(\sum d_i)}{n}}{n-1}$$

⇨ $Z = \frac{|m_d - 0|}{\sqrt{\frac{\sigma^2_d}{n}}}$ => Z >1.96 pour accepté H_1

B) Test t de student

Même condition

$t = \frac{|m_d - 0|}{\frac{\sigma_d}{\sqrt{n}}}$ $t_{obs} \leq t_{tab}$ pour accepter H_1

ddl = n-1

7.1.3 Comparaison d'une moyenne observée et une moyenne attendue

m=22.3→echantillon

μ=20.9 →papulation de ref

H_0=m= μ

$$\Delta = 22.3 - 20.9 = 1.4$$

H_1=m≠ μ

Est-elle suffisamment éloignée de 0 ?

A) Test Z « ε »

⇨ Z =$\frac{|m-\mu|}{écartype\ \mu}$ suit une loi N C R

⇨ Ecartype μ=$\frac{\sigma}{\sqrt{n}}$ => Z =$\frac{|m-\mu|}{\frac{\sigma}{\sqrt{n}}}$

z>1.96 =>H_1 accepté

B) Test t de student

→t= $\frac{|m-\mu|}{écarrtype\ \mu}$ loi student centré sur 0

=> $t = \frac{|m-\mu|}{\frac{\sigma}{\sqrt{n}}}$ ddl =n-1

QCM test t student et test Z

1. Le test t de Student est utilisé pour :
a) Comparer des proportions entre deux groupes
b) Comparer des moyennes quand la variance est inconnue
c) Tester la relation entre deux variables qualitatives
d) Étudier une fréquence observée

2. Le test Z est approprié :
a) Pour des échantillons < 30 avec variance inconnue
b) Pour des échantillons > 30 avec variance connue
c) Pour vérifier si deux groupes appariés ont les mêmes moyennes
d) Pour tester la normalité des données

3. Les tests t et Z servent à :
A) Comparer des médianes
B) Évaluer des corrélations entre variables ordinales
C) Comparer des moyennes entre groupes
D) Calculer une variance

4. On utilise un test t pour :
A) Évaluer une fréquence observée
B) Comparer la moyenne d'un échantillon à une valeur théorique
C) Comparer deux groupes indépendants
D) Comparer des moyennes avant/après traitement sur les mêmes individus

5. On utilise un test Z pour :
A) Comparer deux proportions entre groupes indépendants
B) Comparer deux médianes
C) Tester si une proportion observée correspond à une proportion théorique
D) Évaluer la normalité

6. Limites du test t :
A) Suppose la normalité des données, surtout pour petits échantillons
B) Utilisable seulement si la moyenne est connue
C) Sensible aux outliers
D) Utilisable uniquement avec variables nominales

7. Limites du test Z :
A) Nécessite une grande taille d'échantillon
B) Demande que la variance soit connue ou estimée avec précision
C) Moins fiable avec petits échantillons
D) Peut être utilisé pour toutes sortes de données, y compris qualitatives

8. Le test t est utilisé quand :
A) La variance est connue
B) La variance est inconnue
C) L'échantillon est toujours très grand
D) Les données suivent une loi de Student

9. Le test Z est utilisé quand :
A) La variance est connue et n > 30
B) Les deux séries sont appariées
C) Les données suivent une loi de Poisson
D) On travaille avec des proportions sur un grand échantillon

10. Comparaison de taille moyenne entre 20 hommes et 20 femmes :
A) Test t pour séries appariées
B) Test t pour séries indépendantes
C) Test t pour une seule moyenne
D) Test du Khi-deux

11. Mesure de la pression artérielle avant et après traitement sur les mêmes patients :
A) Test t pour séries indépendantes
B) Test t pour séries appariées
C) Test t pour une seule moyenne
D) Test du Khi-deux

12. Consommation d'eau moyenne comparée à la norme de 2 litres (30 personnes interrogées) :

A) Test t pour séries indépendantes
B) Test t pour une seule moyenne
C) Test t pour séries appariées
D) Test du Khi-deux

13. Comparaison de la proportion de fumeurs entre hommes et femmes :

A) Test t pour séries indépendantes
B) Test Z pour deux proportions
C) Test t pour séries appariées
D) Test du Khi-deux

14. Vérifier si 30 % des étudiants ont un ordinateur portable (100 étudiants interrogés) :

A) Test t pour une seule moyenne
B) Test Z pour une proportion
C) Test t pour séries appariées
D) Test du Khi-deux

Devoir :

À travers ces trois sections, vous aurez l'opportunité d'acquérir des compétences pratiques en statistiques. Vous apprendrez à effectuer des tests t dans le logiciel Statistica et à développer votre capacité à interpréter les résultats dans le contexte de vos données. Cela vous permettra de mieux comprendre les méthodes statistiques et d'appliquer vos connaissances à des situations réelles.

Bonne chance dans vos analyses !

1. Test *t* pour deux échantillons indépendants

- ✓ **Objectif** : Comparer les moyennes de deux groupes indépendants (par exemple, hommes et femmes) sur une variable donnée (par exemple, taille).
- ✓ **Structure des données**

Participant	Groupe (Sexe)	Taille (cm)
1	Homme	180
2	Femme	165
3	Homme	172
4	Femme	160
5	Homme	175
6	Femme	162

✓ **Analyse dans Statistica**

1. Sélectionner le test *t* pour deux échantillons indépendants.
2. Indiquer "Groupe (Sexe)" comme variable de regroupement.
3. Choisir "Taille (cm)" comme variable à analyser.
4. Exécuter le test et interpréter les résultats pour déterminer s'il y a une différence significative entre les tailles des hommes et des femmes

2. Test *t* pour échantillons appariés

✓ **Objectif** : Comparer les moyennes de deux mesures prises sur le même groupe (par exemple, score de stress avant et après une thérapie).

✓ **Structure des données**

Participant	Stress Avant (Score)	Stress Après (Score)
1	15	10
2	18	12
3	20	15
4	14	10
5	16	11

✓ **Analyse dans Statistica**

1. Sélectionner le test *t* pour échantillons appariés.
2. Choisir "Stress Avant (Score)" et "Stress Après (Score)" comme variables à comparer.
3. Exécuter le test et interpréter les résultats pour déterminer si la thérapie a eu un effet significatif sur le niveau de stress.

3. Test *t* à un échantillon

✓ **Objectif** : Comparer la moyenne d'un seul échantillon à une valeur hypothétique fixe (par exemple, taille moyenne nationale).

✓ **Structure des données**

Étudiant	Taille (cm)
1	172
2	168
3	174
4	169
5	171
6	175
7	170
8	168
9	173
10	169

✓ **Analyse dans Statistica**

1. Sélectionner le test *t* à un échantillon.
2. Choisir "Taille (cm)" comme variable à analyser.
3. Indiquer la valeur hypothétique (par exemple, 170 cm).
4. Exécuter le test et interpréter les résultats pour déterminer si la taille moyenne de l'échantillon diffère significativement de la valeur hypothétique.

8. Analyse de la variance (comparaison de plusieurs moyennes)

8.1 Analyse de la variance à un facteur (ANOVA à un facteur)

8.1.1. Définition

L'analyse de la variance à un facteur (ANOVA à un facteur) est une technique statistique utilisée pour comparer les moyennes de plusieurs populations ou groupes, lorsqu'il n'y a qu'un seul facteur étudié.

Cette méthode permet de déterminer si les différences observées entre les moyennes des groupes sont significatives ou si elles peuvent être attribuées au hasard.

L'ANOVA à un facteur répond à la question suivante : Est-ce que les moyennes des différents groupes sont significativement différentes les unes des autres ?

Pour cela, elle compare la variabilité entre les groupes à la variabilité à l'intérieur des groupes. Si la variabilité entre les groupes est suffisamment grande par rapport à la variabilité intra-groupe, alors on peut conclure que les moyennes diffèrent significativement.

Les résultats de l'ANOVA à un facteur sont généralement présentés sous forme d'un tableau d'analyse de variance, indiquant la source de variation, les degrés de liberté, la somme des carrés, le carré moyen, la valeur du test F et la p-value associée.

Cette analyse permet donc de comparer de manière robuste les moyennes de plusieurs populations ou groupes, et constitue un outil essentiel en statistique pour l'expérimentation et l'analyse de données.

8.1.2. Etapes principale de calcul pour l'analyse de la variance à un facteur

a) Calcul de la variance inter-groupes :

Somme des carrés inter-groupes

$$(SC_{inter\ groupe}) = \Sigma\, n_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2$$

$\bar{x}_i$ est la moyenne du groupe j

b) Calcul de la variance intra-groupes :

Somme des carrés intra-groupes

$$(SC_{intra\ groupe}) = \Sigma\left(x_{ij} - \bar{x_j}\right)^2$$

x_{ij} représente les valeurs individuelles de chaque groupe.

$\bar{x_j}$ la moyenne du groupe .

c) Calcul de la variance totale :

Somme des carrés totale (SCT) = $\Sigma(x_i - \bar{\bar{x}})^2$

x_i représente les valeurs individuelles

$\bar{\bar{x}}$ la moyenne générale

d) Calcul des degrés de liberté :

- ✓ **Degrés de liberté totaux**

$$ddl_{Totale} = N - 1$$

- ✓ **Degrés de liberté inter-groupes**

$$ddl_{inter\ groupe} = k - 1 \text{ (k = nombre de groupes)}$$

- ✓ **Degrés de liberté intra-groupes**

$$ddl_{intra\ groupe} = N - k$$

e) Calcul des carrés moyens :

Carré moyen inter-groupes = $SC_{inter\ groupe}$ / (k - 1)

Carré moyen intra-groupes = $SC_{intra\ groupe}$ / (N - k)

f) Calcul du test F :

F = Carré moyen inter-groupes / Carré moyen intra-groupes

g) Interprétation du résultat :

Si la p-value associée au test F est inférieure au seuil de significativité choisi, on conclut qu'il y a un effet significatif du facteur.

Cette analyse permet de décomposer la variance totale en une part expliquée par les différences entre les groupes et une part résiduelle liée à la variabilité intra-groupes. Le test F permet alors de déterminer si les différences observées entre les groupes sont statistiquement significatives.

✓ **Exemple calculé :** Le nombre d'œufs pondus par des hirondelles pendant 3 saisons successives

Saison 01	Saison 02	Saison 03
1	2	5
4	2	6
5	3	4
3	1	7
6	4	5
4	2	

Moyenne saison 01 $(\overline{x_{\text{Saison 01}}})$= **3,83**

Moyenne saison 02 $(\overline{x_{\text{Saison 02}}})$= **2,33**

Moyenne saison 03 $(\overline{x_{\text{Saison 03}}})$ = **5,4**

Moyenne Totale $(\overline{\overline{x}})$= **3,76**

1) Calcul de la variance pour le facteur Saison (inter-groupes):

Somme des carrés inter-groupes

$$(SC_{\text{inter groupe}}) = \Sigma\, n_j \left(\overline{x_j} - \overline{\overline{x}}\right)^2 = n_{\text{Saison 01}}(\overline{x_{\text{Saison 01}}} - \overline{\overline{x}})^2 + n_{\text{Saison 02}}(\overline{x_{\text{Saison 02}}} - \overline{\overline{x}})^2 + n_{\text{Saison 03}}(\overline{x_{\text{Saison 03}}} - \overline{\overline{x}})^2$$

$$= 6(3,83 - 3,76)^2 + 6(2,33 - 3,76)^2 + 5(5,4 - 3,76)^2$$

0,0294 + 12,2694 + 13,448= 25,7468

2) Calcul de la variance résiduelle (intra groupe):

$$(SC_{intra\ groupe}) = \Sigma\left(x_{ij} - \bar{x_j}\right)^2$$

$$[(1-3{,}83)^2+(4-3{,}83)^2+(5-3{,}83)^2+(3-3{,}83)^2+(6-3{,}83)^2+(4-3{,}83)^2]_{\text{Saison 0}}$$
$$+[(2-2{,}33)^2+(2-2{,}33)^2+(3-2{,}33)^2+(1-2{,}33)^2+(4-2{,}33)^2+(2-2{,}33)^2]_{\text{Saison 02}}$$
$$+[(5-5{,}4)^2+(6-5{,}4)^2+(4-5{,}4)^2+(7-5{,}4)^2+(5-5{,}4)^2]_{\text{Saison 03}}$$

$$= 14{,}8334 + 5{,}3334 + 5{,}2$$

$$= 25{,}3668$$

3) Calcul de la variance totale :

Somme des carrés totale

$$(SCT) = \Sigma(x_i - \bar{\bar{x}})^2 = [(1-3{,}76)^2 + (4-3{,}76)^2 + (5-3{,}76)^2 + (3-3{,}76)^2 + (6-3{,}76)^2 + (4-3{,}76)^2]_{\text{Saison 01}} + [(2-3{,}76)^2 + (2-3{,}76)^2 + (3-3{,}76)^2 + (1-3{,}76)^2 + (4-3{,}76)^2 + (2-3{,}76)^2]_{\text{Saison 02}} + [(5-3{,}76)^2 + (6-3{,}76)^2 + (4-3{,}76)^2 + (7-3{,}76)^2 + (5-3{,}76)^2]_{\text{Saison 03}}$$

$$= 51{,}11$$

4) Calcul des degrés de liberté :

- ✓ **Degrés de liberté totaux**

$$ddl_{Totale} = N - 1 = 15-1 = 14$$

- ✓ **Degrés de liberté inter-groupes (inter saison)**

$$ddl_{inter\ groupe} = k - 1 = 3-1 = 2$$

- ✓ **Degrés de liberté intra-groupes**

$$ddl_{intra\ groupe} = N - k = 15-3 = 12$$

5) Calcul des carrés moyens :

Carré moyen inter-groupes = $SC_{\text{inter groupe}}$ / $ddl_{inter\ groupe}$

= 25,7468/2 = **12,8734**

Carré moyen intra-groupes = *($SC_{intra\ groupe}$) / $ddl_{intra\ groupe}$*

= 25,3668/12 =**2,1139**

6) Calcul du test F :

F = Carré moyen inter-groupes / Carré moyen intra-groupes

= 12,8734/2,1139 = **6,089**

7) Interprétation des résultats :

$F_{observé}$ = **6,089**

- Degrés de liberté entre les groupes $ddl_{inter\ groupe}$ = 2
- Degrés de liberté dans les groupes $ddl_{intra\ groupe}$= 12

Nous allons comparer la valeur observée $F_{observé}$=6,089 à une valeur critique $F_{critique}$ à un seuil de signification donné (α=0,05) obtenue apartir du tableau de ficher par le croisement de ***$ddl_{inter\ groupe}$*** et ***$ddl_{intra\ groupe}$***

Dans notre cas $F_{critique}$ = 3,89 (Tableau III annexe)

Si $F_{observé}$ est supérieur à $F_{critique}$, alors **H_0** rejeté c à dire il y a une différence significative et l'effet est significatif à α=0.05

Si $F_{observé}$ est inférieur à $F_{critique}$, alors **H_0** accépté c à dire il n' y a pas une différence significative et il n' ya pas un effet est significatif à α=0.05

Dans notre cas $F_{observé}$ **(6,089)** est supérieur à $F_{critique}$ ***(3,89)*** alors **H_0** rejeté c à dire il y a une différence significative et l'effet de saison est significatif à α=0.05

Donc le nombre d'œufs pondus par les hirondelles de notre exemple est diffèrent pour chaque saison autrement dit Il existe une différence significative du nombre d'œufs pondus entre les saisons

8.2 Analyse de la variance à 2 facteurs

8.2.1 Définition

L'analyse de la variance à deux facteurs, également connue sous le nom de "two-way ANOVA", est une technique statistique utilisée pour évaluer l'influence de deux variables indépendantes catégoriel (ou "facteurs") sur une variable dépendante généralement quantitative.

Cette méthode permet d'analyser les effets des modalités de chacun des deux facteurs, ainsi que leur effet d'interaction potentiel. Elle est une extension de l'ANOVA à un facteur, permettant d'étudier l'impact combiné de deux variables sur une même mesure.

✓ **L'ANOVA à deux facteurs répond à trois questions principales :**

- Le premier facteur a-t-il un effet significatif sur la variable dépendante ?
- Le deuxième facteur a-t-il un effet significatif ?
- Y a-t-il une interaction significative entre les deux facteurs ?

Les résultats de l'ANOVA à deux facteurs sont généralement présentés sous forme d'un tableau d'analyse de variance, indiquant la source de variation, les degrés de liberté, la somme des carrés, le carré moyen, la valeur du test F et la p-value associée.

Cette analyse permet donc de mieux comprendre la façon dont deux facteurs influencent conjointement une variable d'intérêt, et constitue un outil puissant pour l'expérimentation et la recherche dans de nombreux domaines.

8.2.2 Etapes principale de calcul pour l'analyse de la variance à deux facteurs

✓ **Hypothéses statistiques**

Test	Hypothèse nulle (H_0)	Hypothèse alternative (H_1)
A	Pas d'effet du facteur A	Le facteur A a un effet
B	Pas d'effet du facteur B	Le facteur B a un effet
A×B	Pas d'interaction entre A et B	Interaction significative entre A et B

✓ **Tableau de calcule de l'ANOVA a deux facteurs**

Source de variation	Somme des carrés (SC)	ddl	Carré moyen (CM)	Valeur F
Facteur A	**SCA**	n_A-1	SCA/(n_A-1)	CM_A / CM_erreur
Facteur B	**SCB**	n_B-1	SSB/(n_B-1)	CM_B / CM_erreur
Interaction A×B	**SCAB**	(n_A-1)(n_B-1)	SCAB/(n_A-1)(n_B-1)	CM_AB / CM_erreur
Erreur (résidus)	**SCE**	AB(N-1)		
Total	**SCT**	AB (N– 1)		

Exemple : Voici un exemple d'analyse de la variance à deux facteurs sur le nombre d'oeufs pondus par des hirondelles selon deux facteurs : la saison et le type de nid.

Saison	Nids	Nombre d'œufs pondus
A	Ancien	4
A	Ancien	5
A	Ancien	3
A	Ancien	6

A	Nouveau	3
A	Nouveau	4
A	Nouveau	2
A	Nouveau	5
A	Nouveau	3
B	Ancien	5
B	Ancien	6
B	Ancien	4
B	Ancien	7
B	Nouveau	2
B	Nouveau	3
B	Nouveau	1
B	Nouveau	4
B	Nouveau	2
C	Ancien	6
C	Ancien	7
C	Ancien	5
C	Ancien	8
C	Ancien	6
C	Nouveau	5
C	Nouveau	6
C	Nouveau	4
C	Nouveau	7
C	Nouveau	5

1. **Variable dépendante (quantitative mesurée) :** Nombre d'œufs pondus

2. **Facteurs (variables indépendantes catégorielles) :**

✓ Facteur A : Saison (avec 3 modalités : A, B, C)

✓ Facteur B : Type de nid (avec 2 modalités : Ancien, Nouveau)

3. **Hypothèses nulles (H_0)**

Effet testé	**Hypothèse nulle (H_0)**
Effet de la saison	Le nombre moyen d'œufs pondus ne diffère pas selon la saison.
Effet du type de nid	Le nombre moyen d'œufs pondus ne diffère pas selon le type de nid.
Interaction saison × type de nid	Il n'y a pas d'interaction entre la saison et le type de nid sur le nombre d'œufs pondus.

4. **Tableau de calcule de l'ANOVA a deux facteurs**

Source de variation	**Somme des carrés (SC)**	**ddl**	**Carré moyen (CM)**	**Valeur F**	**Valeur p**
Saison	**SC_{Saison} =24,3360**	**n_{Saison}-1 = (3-1) = 2**	**24,3360/2 =12,1680**	**8,6914**	**0,001653**
Type de nid	**$SC_{Type\ de\ nid}$ =20,8000**	**$n_{Type\ de\ nid}$-1= (2-1) = 1**	**20,8000/1 = 20,8000**	**14,8571**	**0,000859**
Interaction Saison×Type de nid	**$SC_{Saison \times Type\ de\ nid}$ =6,3804**	**(n_{Saison}-1)($n_{Type\ de\ nid}$-1) = (3-1)(2-1) = 2**	**6,3804/2 = 3,1902**	**2,2787**	**0,126066**
Erreur (résidus)	**SCE =30,8000**	**(23-1)= 22**	**1,4000**		
Total	**SCT**	**AB (N– 1)**			

5. Interpretation des résultats :

- ✓ **Effet principal de la Saison**

F = 8,6914, p = 0,001653

La saison a un effet significatif sur le nombre d'œufs pondus. Cela signifie que le nombre moyen d'œufs varie de manière statistiquement significative entre les saisons A, B et C.

- ✓ **Effet principal du Type de nid**

F = 14,8571, p = 0,000859

Le type de nid (ancien vs. nouveau) a aussi un effet significatif sur le nombre d'œufs pondus. Les femelles pondent donc un nombre d'œufs significativement différent selon le type de nid.

- ✓ **Interaction Saison × Type de nid**

F = 2,2787, p = 0,126066

L'interaction entre la saison et le type de nid n'est pas significative. Cela signifie que l'effet du type de nid sur le nombre d'œufs pondus ne dépend pas significativement de la saison, et inversement. Les effets sont plutôt additifs, et non combinés.

- ✓ **Erreur résiduelle :**

SCE=30.8000, ddl=22, CM=1.4000

Les erreurs représentent la variabilité inexpliquée.

- ✓ **Analyse globale :**

Les deux facteurs principaux (**Saison** et **Type de nid**) ont des effets significatifs sur la variable dépendante.

L'interaction entre **Saison** et **Type de nid** n'est pas significative, ce qui signifie que l'effet d'un facteur est relativement indépendant de l'autre.

8.3 Analyse de la variance à mesures répétées

8.3.1. L'analyse de la Variance à Mesures Répétées (ou RM-ANOVA, For Repeated Measures ANOVA)

Est une méthode statistique utilisée pour comparer des résultats obtenus plusieurs fois sur les mêmes individus, soit à différents moments, soit dans différentes conditions.

Elle permet de savoir si les variations observées sont dues aux effets étudiés (comme le temps, le traitement...) ou simplement au hasard.Cette méthode permet d'étudier l'effet de la répétition des mesures , et la principale question auxquelles l'ANOVA avec répétition permet de répondre est

Les mesures répétées ont-elles un effet significatif ?

L'analyse de variance à un facteur avec mesures répétées est donc l'extension du test t pour les échantillons dépendants (appariés) mais avec plus de deux groupes.

A) Exemple : ANOVA a mesures répétées à un facteurs

Un chercheur souhaite évaluer l'impact d'un traitement sur la performance cognitive (la capacité du cerveau à effectuer différentes tâches mentales.) de 6 participants mesurée à 3 moments différents (avant, après 1 mois, et après 2 mois).

Patients	Temps 1 (Avant)	Temps 2 (1 Mois)	Temps 3 (2 Mois)
1	12	15	18
2	14	16	20
3	13	17	19

a) Hypothèse nulle :

Il n’y a pas de différence significative de la performance cognitive entre les trois moments de mesure.

(Autrement dit : le traitement n'a eu aucun effet mesurable dans le temps.)

b) Résultat :

Les calculs sont résumés dans le tableau suivant :

Source de variation	Somme des carrés (SC)	ddl	Carré moyen (CM)	Valeur F	Valeur p
Intercept	**2304,000**	**1**	**2304,000**	**987,4286**	**0,001011**
Facteur Temps	**54,000**	**2**	**27,000**	**81,0000**	**0,000581**
Error	**1,333**	**4**	**0,333**		

c) Interpretation :

- ✓ **Intercept** :

SC=2304, CM=2304, F=987.43, p=0.0010 donc **H_0 rejeté**

Cela signifie qu'il existe une **différence significative** au niveau global de la variable dépendante par rapport à une valeur de référence (souvent la moyenne générale ou une constante).

- ✓ **Facteur Temps (Effet principal) :**

SC=54, CM=27, F=81, p=0.00058 donc **H_0 rejeté**

Le facteur Temps (différents moments de traitement) a un effet **très significatif** sur la variable dépendante, car $p<0.05$.

- ✓ **Erreur (Error) :**

Erreur associée au facteur Temps : CM=0.333

Cette composante reflète la variance résiduelle non expliquée par le modèle.

8.3.2. L'analyse de la variance à deux facteurs avec mesures répétées

Est une extension de l'ANOVA à deux facteurs qui prend en compte le fait que les mesures sont répétées sur les mêmes individus ou unités expérimentales.

Cette méthode permet de répondre a ces hypothèses :

- ✓ **Le premier facteur a-t-il un effet significatif ?**

- ✓ **Le deuxième facteur a-t-il un effet significatif ?**
- ✓ **Y a-t-il une interaction significative entre les deux facteurs ?**
- ✓ **Les mesures répétées ont-elles un effet significatif ?**

Cette méthode est particulièrement utile lorsqu'on souhaite étudier l'effet de deux facteurs sur une variable dépendante, tout en tenant compte du fait que les mesures sont répétées sur les mêmes individus. Cela permet d'augmenter la puissance statistique de l'analyse.

Exemple 02 : ANOVA a mesures répétées à deux facteurs :

Un chercheur souhaite évaluer l'impact d'un traitement sur la performance cognitive de 6 participants mesurée à 3 moments différents (avant, après 1 mois, et après 2 mois) entre deux groupes (traitement et contrôle).

Patients	Groupe	Temps 1 (Avant)	Temps 2 (1 Mois)	Temps 3 (2 Mois)
1	Traitement	12	15	18
2	Traitement	14	16	20
3	Traitement	13	17	19
4	Contrôle	10	11	12
5	Contrôle	9	10	11
6	Contrôle	8	9	10

a) Hypothèse nulle :

Le traitement n'a pas d'effet sur la performance cognitive, quel que soit le moment.

Et les deux facteurs (groupe et moment) n'interagissent pas entre eux.

Facteurs	Hypothèse nulle H_0
Temps	Les scores ne changent pas significativement dans le temps.
Groupe	Pas de différence significative entre le groupe traitement et le groupe contrôle.
Interaction	L'évolution dans le temps est la même pour les deux groupes.

Facteurs	Hypothèse nulle H_0
Temps×Groupe	

b) Résultats :

Les calculs sont résumés dans le tableau suivant :

Source de variation	SC	ddl	CM	F	p	Interprétation
Intercept	**3042**	**1**	**3042**	**1140,75**	**0,0000046**	**Moyenne générale très significative**
Groupe	**162**	**1**	**162**	**60,75**	**0,00146**	**Effet significatif du traitement**
Temps (R1)	**48**	**2**	**24**	**144**	**0,00000053**	**Effet significatif du temps**
Temps × Groupe	**12**	**2**	**6**	**36**	**0,0001**	**Interaction significative**
Erreur intra-sujets	**1,33**	**8**	**0,17**			

c) Interpretation :

- ✓ **Effet du groupe :**

$F(1,4) = 60,75$; $p = 0,00146$ donc H_0 rejeté

Le groupe ayant reçu le traitement a obtenu des scores cognitifs plus élevés que le groupe contrôle.

- ✓ **Effet du temps (mesures répétées) :**

F(2,8) = 144 ; p = 0,00000053 donc H0 rejeté

Les scores cognitifs évoluent dans le temps, ce qui suggère une amélioration ou un changement significatif des performances entre les 3 moments.

- ✓ **Interaction Groupe × Temps :**

F(2,8) = 36 ; p = 0,0001 donc H0 rejeté

L'évolution des scores n'est pas la même dans les deux groupes.

Le groupe traitement s'améliore plus rapidement que le groupe contrôle.

Le traitement améliore la performance cognitive au fil du temps, contrairement au groupe contrôle.

QCM test ANOVA (un facteur , plusieurs facteurs , a mesures répétées)

1. Un chercheur mesure le nombre de poussins éclos dans 3 types de nids (bois, plastique, métal).Quelle est la variable dépendante dans cette étude ?
A. Le type de nid
B. Le nombre de poussins éclos
C. Le lieu de l'étude
D. Le nombre de nids

2. Quel est le facteur (variable indépendante) dans l'exemple précédent ?
A. Le climat
B. Le nombre de poussins
C. Le type de nid
D. La taille des œufs

3. Si on mesure la croissance d'une plante dans 3 types de sol différents. Quel test statistique utiliser ?
A. Test t pour échantillons appariés
B. Corrélation
C. ANOVA à un facteur
D. Régression linéaire

4. Dans une ANOVA à deux facteurs, on peut évaluer :
A. L'effet d'un seul facteur
B. L'effet combiné de deux facteurs et leur interaction
C. Une relation linéaire entre deux variables
D. La corrélation entre des variables continues

5. On étudie le succès de nidification selon la saison (printemps/été/automne) et le type d'arbre (pin/chêne). Quel test utiliser ?
A. ANOVA à un facteur
B. Régression multiple
C. ANOVA à deux facteurs
D. Test du Chi^2

6. Dans un test ANOVA, une valeur p < 0,05 signifie :
A. Il n'y a pas de différence entre les groupes
B. Il y a une différence significative entre au moins deux groupes
C. Tous les groupes sont égaux
D. Les données ne sont pas valides

7. Qu'est-ce qu'une mesure répétée ?
A. Quand on mesure plusieurs variables sur un individu
B. Quand on mesure un même individu à différents moments
C. Quand on compare plusieurs groupes
D. Quand on répète une analyse plusieurs fois

8. On suit la croissance de 6 oiseaux mesurée à 3 moments (semaines 1, 2 et 3). Quel test appliquer ?
A. ANOVA à un facteur
B. Test de Mann-Whitney
C. ANOVA à mesures répétées
D. Corrélation de Pearson

9. Dans ce cas, quelle est la variable dépendante ? Croissance des oiseaux mesurée à plusieurs semaines
A. L'espèce
B. Le poids ou la taille mesurée
C. La semaine
D. Le lieu d'observation

10. Quel est l'effet d'une interaction significative dans une ANOVA à deux facteurs ?
A. Aucun des facteurs n'a d'effet
B. Les deux facteurs ont le même effet
C. L'effet d'un facteur dépend du niveau de l'autre
D. L'analyse est faussée

11. Les températures corporelles varient selon l'espèce (moineau, mésange) et selon l'heure de la journée (matin, midi, soir). Que teste-t-on dans une ANOVA à deux facteurs avec interaction ?
A. Si l'espèce et l'heure influencent séparément la température
B. Si le type d'espèce modifie l'effet de l'heure

C. Les deux précédentes
D. Ni l'une ni l'autre

12. Un chercheur mesure la performance cognitive de 6 individus à 3 moments (avant, après 1 mois, après 2 mois), dans 2 groupes (traitement vs contrôle). Quel test choisir ?

A. ANOVA à un facteur
B. ANOVA à deux facteurs avec mesures répétées
C. Régression multiple
D. Test de Wilcoxon

13. En résumé, que permet l'ANOVA avec mesures répétées ?

A. Éviter les erreurs de mesure
B. Comparer plusieurs mesures sur les mêmes individus
C. Comparer deux groupes indépendants
D. Étudier la corrélation entre deux facteurs

Devoir

Dans cette étude, nous allons explorer trois types d'ANOVA pour mieux comprendre comment les différents facteurs peuvent influencer les performances des étudiants :

- ✓ **ANOVA à un facteur**
- ✓ **ANOVA à deux facteurs :**
- ✓ **ANOVA à mesures répétées**

1.ANOVA à un facteur :

Cet exemple teste l'effet du nombre d'heures d'étude sur les performances des étudiants (notes entre 1 et 20).

Heures d'étude (x)	Note finale (y)
4	10
4	12
4	11
6	13
6	14
6	15
8	16

Heures d'étude (x)	Note finale (y)
8	17
8	15
10	18
10	19
10	17
12	19
12	20
12	18

Explication : Vous allez comparer les moyennes des notes des étudiants en fonction des différentes catégories d'heures d'étude (4 heures, 6 heures, 8 heures, 10 heures, 12 heures). Cette analyse permettra de voir si plus d'heures d'étude mènent à de meilleures performances.

2. ANOVA à deux facteurs :

Cet exemple teste l'effet combiné du nombre d'heures d'étude et de la méthode d'enseignement (en ligne ou en personne) sur les performances des étudiants (notes entre 1 et 20).

Heures d'étude (x)	Méthode d'enseignement	Note finale (y)
4	En ligne	10
4	En ligne	12
4	En personne	13
4	En personne	11

Heures d'étude (x)	Méthode d'enseignement	Note finale (y)
6	En ligne	14
6	En ligne	16
6	En personne	17
6	En personne	15
8	En ligne	17
8	En ligne	18
8	En personne	19
8	En personne	17
10	En ligne	18
10	En ligne	19
10	En personne	20
10	En personne	18

Explication : Cette analyse vous permet d'examiner l'effet de deux facteurs (le nombre d'heures d'étude et la méthode d'enseignement) simultanément. Vous pourrez également tester si l'interaction entre ces deux facteurs influence les résultats des étudiants.

3.ANOVA à mesures répétées avec un seul facteur :

Cet exemple examine l'évolution des performances des étudiants avant, pendant, et après le cours, sans prendre en compte d'autres facteurs comme la méthode d'enseignement.

Étudiant	Avant le cours (T1)	Pendant le cours (T2)	Après le cours (T3)
Étudiant 1	8	12	16
Étudiant 2	10	14	18
Étudiant 3	7	11	15
Étudiant 4	9	13	17
Étudiant 5	6	10	14
Étudiant 6	8	12	16
Étudiant 7	7	11	15
Étudiant 8	10	14	18

Explication: Dans cet exemple, vous avez mesuré les performances des étudiants à trois moments différents :

- **Avant le cours (T1)** : Note avant que le cours ne commence.
- **Pendant le cours (T2)** : Note pendant que le cours est en cours.
- **Après le cours (T3)** : Note après la fin du cours.

L'ANOVA à mesures répétées permet d'analyser la variance des scores de ces étudiants à travers ces trois moments pour vérifier s'il y a une différence significative dans l'évolution des performances au fil du temps.

Tableaux à remplir dans l'activité dans un fichier PDF :

- Vous allez utiliser les données de chaque tableau pour effectuer les tests ANOVA dans un logiciel statistica
- Pour chaque test, vous devrez indiquer les résultats du test (valeur F, valeur p) et interpréter les conclusions basées sur la valeur p (si $p < 0,05$, il y a une différence significative).

IV. Corrélation et régression linéaire de deux variables

1. Corrélation

1.1 Défintion

La corrélation est une mesure statistique qui exprime la notion de liaison linéaire entre deux variables qui évoluent ensemble à une vitesse constante. Elle est utilisée pour tester l'hypothèse d'une relation entre deux variables et pour évaluer la force de cette relation

Il existe deux types de corrélation :

- **Corrélation simple :** Elle se réfère à la liaison entre deux caractères.
- **Corrélation multiple :** Elle concerne plusieurs variables et peut être représentée dans une matrice de corrélation.

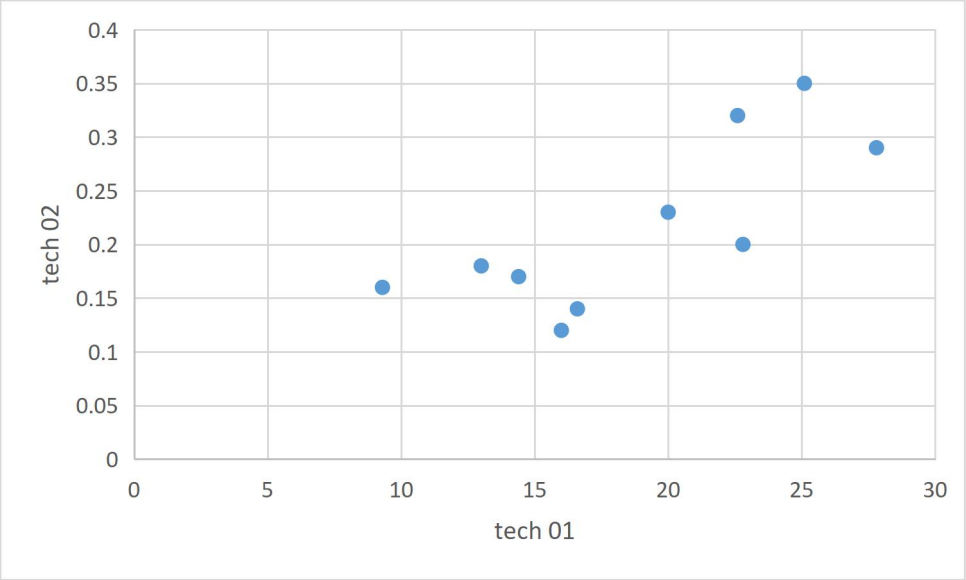
Exemple : Dosage d'une protéine par deux techniques

Technique 01	Technique 02
9.30	0.16
13	0.18
14.40	0.17
16	0.12
16.60	0.14
20	0.23
22.60	0.32
22.80	0.20
25.10	0.35
27.80	0.29

Nous avons deux variables quantitatives appariées

- ⇨ Es ce que le résultat élevé avec la technique 01 est aussi élevé par la technique 02 ?
- ⇨ Es ce que le résultat bas avec la technique 01 est aussi bas par la technique 02 ?

Pour visualiser la relation on va présenter graphiquement les couple de résultats (x.y)



Lorsque x est élevé es ce que y est élevé ?

On grosomodo oui mais quelques couples non

Donc le nuage de point peut visualiser la relation entre les deux résultats

1.2 Mesure de la liaison entre les deux variables quantitatives :

Le coefficient de corrélation de Pearson est un indicateur statistique utilisé pour mesurer la relation linéaire entre deux variables quantitatives. Il varie entre -1 et 1, où 1 indique une corrélation positive parfaite, -1 une corrélation négative parfaite et 0 l'absence de corrélation.

1.2.1 Coefficient de corrélation r Pearson :

Le coefficient de corrélation de Pearson, noté r, est une mesure statistique qui permet d'évaluer la force et la direction de la relation linéaire entre deux variables quantitatives.

$$r = \frac{COV(x,y)}{\sigma(x) \times \sigma(y)}$$

$$COV(x,y) = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{(n-1)}$$

$$\sigma(x) = \sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$$

$$\sigma(y) = \sqrt{\frac{\sum(y-\bar{y})^2}{(n-1)}}$$

$$Donc\ r = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{\sum[(x-\bar{x})^2(y-\bar{y})^2]}$$

Où :

x et y sont les valeurs des deux variables

$\bar{x}$ et $\bar{y}$ sont leurs moyennes respectives

Σ représente la somme des produits ou des carrés

✓ **Interprétation des résultats :**

r = 1 corrélation positif parfaite

r=0.4 une corrélation faible médiocre.

r=0 pas de corrélation

r= -0.9 corrélation négatif forte

1.2.2 Test d'un coefficient de corrélation

Exemple : *r=0.83, n = 30*

r =0.83 est-elle significatif ?

H_0 : il n'existe aucune corrélation

H_1 : il y a une corrélation significative

✓ **Test student :**

$$t = \frac{|r - 0|}{\sigma r}$$

$$\sigma r = \sqrt{\frac{(1 - r^2)}{(n - 2)}} = \sqrt{\frac{(1 - 0{,}83^{\,2})}{(30 - 2)}} = 0{,}11$$

$$Donc\ t = \frac{|0{,}83 - 0|}{0{,}11} = 7{,}5$$

ddl = n-2 ; α=0.05 ; $t_{tableaux}$ = 1,7011 ; $p<0.0000001$

Donc La corrélation est significative.

2. Régression linéaire simple

2.1 Défintion

La régression linéaire simple est un outil statistique qui permet de modéliser la relation linéaire entre deux variables quantitatives. L'objectif de la régression linéaire est d'obtenir une équation mathématique qui permet de prédire la valeur d'une variable (y) en fonction de la valeur d'une autre variable (x).

2.2 Présentation graphique des données

Pour afficher une droite de régression et son équation dans Excel, il est possible d'utiliser des outils graphiques pour tracer la droite et obtenir son équation mathématique.

La présentation graphique de la droite de régression et du nuage de points est essentielle pour analyser visuellement la relation entre les variables mesurées et interpréter les résultats d'une expérience ou d'une étude statistique (fig 01)

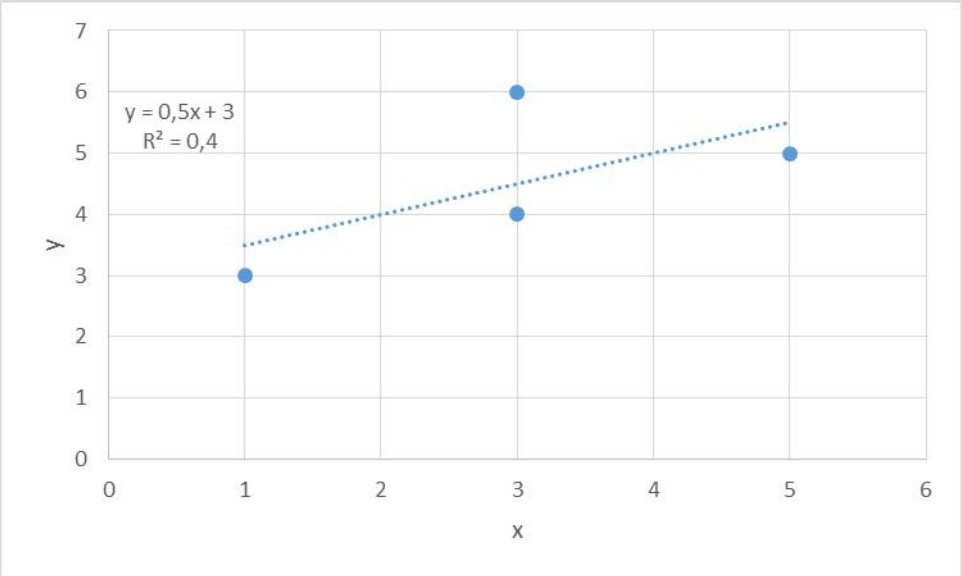


Figure 01 : Présentation graphique des données (nuage de point et droite de détermination)

La régression selon « a» comme la corrélation

$a > 0$ régression positif

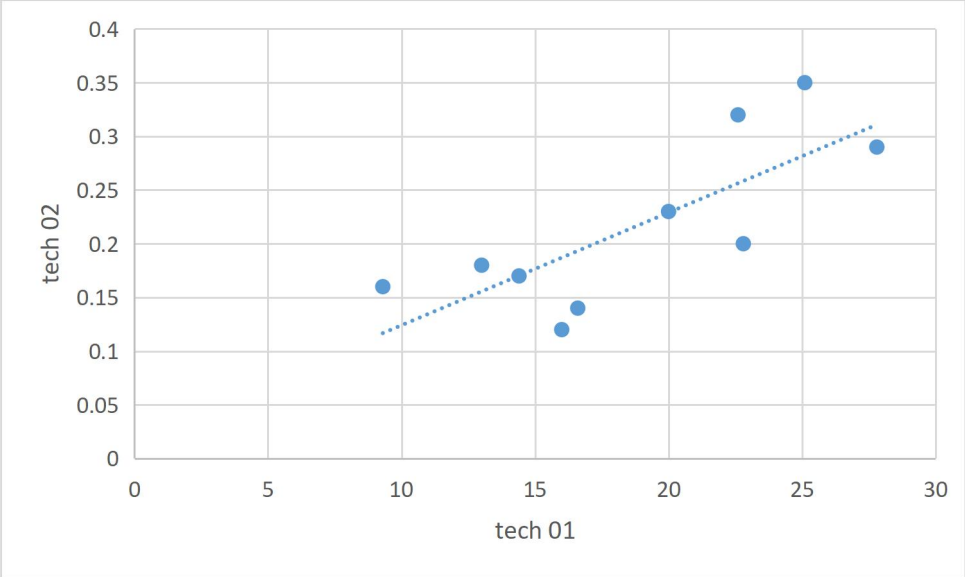
$a < 0$ régression négative

a = 0 nulle pas de régression

2.3 Détermination de l'équation de la droite D

Tracé une droite $y=ax+b$

Le choix de la droite : Cherchons la droite qui se rapproche au nombre plus élevé des points.



La somme de distances entre les points et la droite [somme carrée car il y a des distances au-dessous (+) et autre au-dessus de la droite (-)]

La droite la plus approprier ou la somme des distances est la plus petite

L'équation de la droite de régression est de la forme : $y= ax +b$

$$a = \frac{cov(x,y)}{var(x)} = \frac{\overline{xy} - \bar{x}.\bar{y}}{\overline{x^2} - \bar{x}}$$

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

- **Exemple :** soit deux variable x et y comme suit dans le tableau

	x	y	xy	x^2	y^2
	1	3	3	1	9
	3	4	12	9	16
	3	6	18	9	36
	5	5	25	25	25
Somme	**12**	**18**	**58**	**44**	**86**
Moyenne	**3**	**4.5**	**14.5**	**11**	**21.5**

A) Calcul « a »

$$a = \frac{\overline{xy} - \bar{x}.\bar{y}}{\overline{x^2} - (\bar{x})2}$$

$$a = \frac{14.5-(3)(4.5)}{11-9} = 0.5$$

B) Calcul « b »

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

$$b = 4.5 - (0.5)(3)$$

C) Equation de la droite D

$$y=0.5\,x+3$$

2.4 Coefficient de détermination R^2

Le coefficient de détermination, noté R^2, est une mesure de la qualité de la prédiction d'une régression linéaire. Il mesure la proportion de la variance de la variable dépendante (y) qui est expliquée par la variable indépendante (x) dans le modèle de régression. Le coefficient de détermination est compris entre 0 et 1, et plus il est proche de 1, plus la régression linéaire est en adéquation avec les données observées. Si le coefficient de détermination est nul, cela signifie que l'équation de la droite de régression ne permet pas de prédire la distribution des points, tandis qu'un coefficient de détermination de 1 indique que l'équation de la droite de régression est capable de prédire 100% de la distribution des points.

Le coefficient de détermination est calculé en utilisant la formule suivante :

$$R^2 = 1 - \frac{SCE_D}{SCE_{\bar{y}}}$$

Le coefficient de détermination est une mesure importante pour évaluer la qualité de la régression linéaire, mais il ne doit pas être utilisé seul pour interpréter les résultats. Il est important de considérer également la signification statistique de la pente et de l'intercept de la droite de régression, ainsi que la distribution des résidus et la normalité des données.

SCE_D Elle mesure la distance de la droite de régression aux points du nuage de points qui est minimale au sens des moindres carrés (fig 02..).

$$\left.\begin{aligned} SCE_D &= \sum_{n}^{1} e_i \\ &= \sum (y_i - \hat{y})^2 \\ \hat{y} &= ax_i + b \end{aligned}\right\} \quad SCE_D = \sum [y_i - (ax_i + b)]^2$$

$$SCE_D = [y_1 - (ax_1 + b)]^2 + [y_2 - (ax_2 + b)]^2 + \ldots + [y_i - (ax_i + b)]^2$$

Si SCE_Dest petite la droite est bien ajustée car la somme des erreurs est proche de 0 et R^2 proche de 1.

Si SCE_Dest grand la droite est mal ajustée car la somme des erreurs est proche de 1 et R^2 proche de 0.

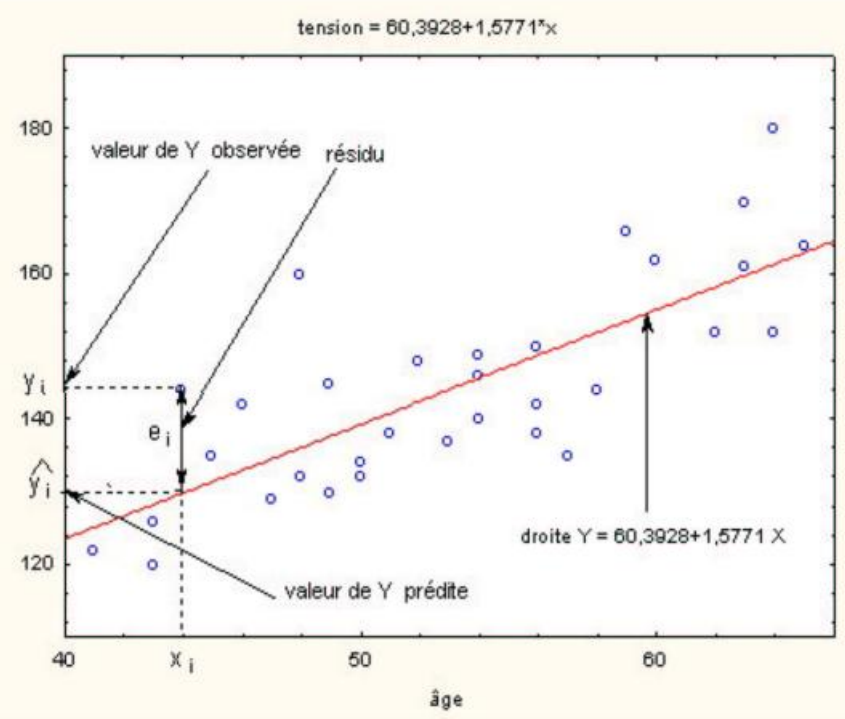


Figure 02 : Droite des moindres carrées

$SCE_{\bar{y}}$: Mesure la variation des valeurs ajustées autour de la moyenne y c'est-à-dire c'est la somme des erreurs carré par rapport à la moyenne (fig 03)

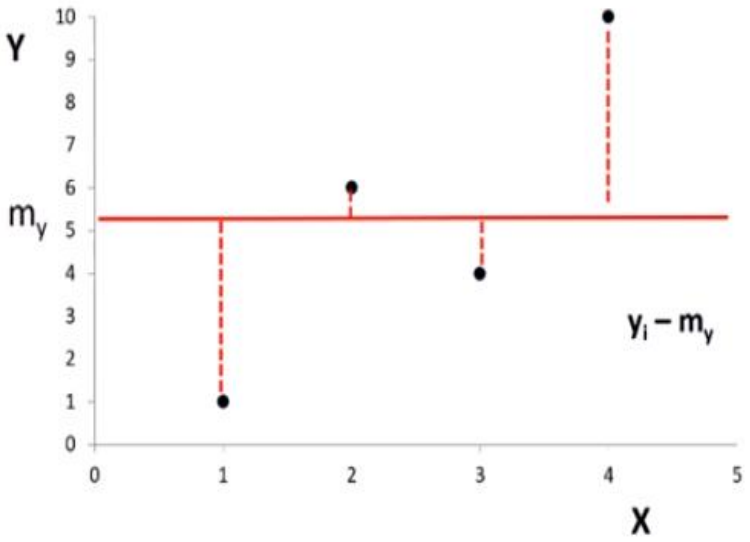


Figure 03 : Somme des erreurs carrés par rapport à la moyenne

$SCE_{\bar{y}} = \sum_{n}^{1}(y_i - \bar{y})^2 = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2$

- **Exemple :** le calcul du coefficient de détermination R^2 pour le même exemple précédent

	x	**y**	$\hat{y}$**=0.5 x+3**	$y_i - \hat{y}$	$(y_i - \hat{y})^2$	$y_i - \bar{y}$	$(y_1 - \bar{y})^2$
	1	3	3.5	0.5	0.25	1.5	2.25
	3	4	4.5	0.5	0.25	0.5	0.25
	3	6	4.5	1.5	2.25	1.5	2.25
	5	5	5.5	0.5	0.25	0.5	0.25
Somme					$SCE_D = 3$		$SCE_{\bar{y}} = 5$
Moyenne		$\bar{y}$ =**4.5**					

$$R^2 = 1 - \frac{SCE_D}{SCE_{\bar{y}}}$$

$$R^2 = 1 - \frac{3}{5}$$

$$R^2 = 0.40$$

Cela signifie que 40% de la variation observée dans le modèle calculé peut être expliquée par les points. Plus précisément, un R carré de 0.40 indique que 40% de la distribution des points est déterminée par la régression.

QCM corrélation et régression linéaire entre deux variable :

1. Que mesure le coefficient de corrélation de Pearson ?

A. La moyenne des deux variables
B. Le degré de dépendance non linéaire entre deux variables
C. L'intensité et la direction de la relation linéaire entre deux variables
D. Le rapport entre les écarts-types

2. Quelle est la valeur possible du coefficient de corrélation de Pearson (r) ?

A. Entre 0 et 1
B. Entre -1 et 0
C. Entre -1 et +1
D. Supérieure à 1

3. Que signifie un coefficient de corrélation r = 0,85 ?

A. Relation faible et négative
B. Relation forte et positive
C. Relation nulle
D. Corrélation non significative

4. Si r = 0,02, cela indique :

A. Une forte corrélation positive
B. Une corrélation modérée
C. Une très faible corrélation
D. Une forte corrélation négative

5. Quel test permet de savoir si le coefficient de corrélation est significatif ?

A. Test t de Student
B. Test du Chi^2
C. Test de Shapiro-Wilk
D. Test de Kruskal-Wallis

6. Quelle hypothèse est testée pour la significativité du coefficient de corrélation ?

A. $H_0 : r \neq 0$
B. $H_0 : r = 0$
C. $H_0 : r > 1$
D. $H_0 : r < -1$

7. Dans une régression linéaire simple, quelle est la forme de l'équation de la droite ?

A. $y = x + b$
B. $y = ax + b$
C. $y = ax^2 + b$
D. $y = bx^2 + a$

8. Dans l'équation y = ax + b, que représente le coefficient a ?

A. L'ordonnée à l'origine
B. Le point moyen
C. La pente de la droite (variation de y pour 1 unité de x)
D. L'écart-type de x

9. Que signifie un coefficient a = -2 dans une droite de régression ?

A. Quand x augmente de 1, y augmente de 2
B. Quand x augmente de 1, y diminue de 2
C. Quand y augmente de 2, x augmente de 1
D. y est constant

10. Que mesure le coefficient de détermination R^2 ?

A. Le degré d'erreur de mesure
B. Le pourcentage de la variance de x expliquée par y
C. Le pourcentage de la variance de y expliquée par x
D. L'inverse du coefficient de corrélation

11. Si $R^2 = 0,81$, que peut-on dire ?

A. 81 % de la variation de y est expliquée par x
B. 81 % de la variation de x est expliquée par y
C. La corrélation est nulle
D. Il n'y a pas de relation linéaire

12. Un chercheur mesure la température moyenne mensuelle (x) et l'abondance d'une espèce d'oiseau (y). Quel test appliquer ?

A. ANOVA
B. Régression linéaire simple
C. Corrélation de Spearman
D. Test de Kruskal-Wallis

13. Dans une étude, on obtient : r = 0,92 et p < 0,001. Quelle est la meilleure interprétation ?

A. La corrélation est faible et non significative
B. La relation est forte et significative
C. Il n'y a pas de relation entre les variables
D. La variance est nulle

14. Un chercheur obtient l'équation de régression : *y = 0,5x + 3*. *Que signifie le coefficient 0,5 ?*

A. Quand x augmente de 0,5, y augmente de 1
B. Quand x augmente de 1, y augmente de 0,5
C. C'est la moyenne de y
D. C'est le test de normalité

15. Quel graphique est le plus approprié pour représenter une régression linéaire simple ?

A. Histogramme
B. Courbe de survie
C. Nuage de points avec droite ajustée
D. Boîte à moustaches

16. Un écologue étudie la relation entre la température moyenne mensuelle (°C) et le nombre de cigales observées dans une région donnée pendant 10 mois.

✓ **Quelle analyse est la plus appropriée ?**

A. Test du Chi^2
B. Corrélation de Pearson
C. Régression multiple
D. ANOVA

✓ **Quelle hypothèse teste-t-on ?**

A. H_0 : r = 0 (absence de corrélation linéaire)
B. H_0 : la température est plus élevée dans un mois précis
C. H_0 : toutes les températures ont le même effet
D. H_0 : y = a + bx

17. On veut prédire le taux de germination (%) d'une plante en fonction de la concentration en sel du sol (g/L), mesurée dans 20 sites.

✓ **Quel test convient ici ?**

A. Corrélation de Spearman
B. Régression linéaire simple
C. Test de Mann-Whitney
D. ANOVA à un facteur

✓ **Quelle est la variable dépendante ?**

A. Concentration en sel
B. Taux de germination
C. Le site
D. Aucune

✓ **L'équation obtenue est : germination = -2 × sel + 80 . Que signifie la pente -2 ?**

A. Quand le sel augmente de 1 g/L, la germination baisse de 2 %
B. Quand le sel augmente de 2 g/L, la germination augmente
C. La germination diminue de 80 %
D. Le sel n'a aucun effet sur la germination

18. Un chercheur étudie l'association entre la taille des arbres (en m) et la richesse en oiseaux dans 30 parcelles forestières.

✓ **Si l'objectif est uniquement de mesurer l'association, on utilise :**

A. Régression linéaire
B. Corrélation de Pearson
C. ANOVA à deux facteurs
D. Analyse en composantes principales

✓ **Si l'objectif est de prédire la richesse en oiseaux selon la taille des arbres, on utilise :**

A. Corrélation de Pearson
B. Régression linéaire simple
C. Test du Chi^2
D. Régression logistique

19. Un modèle de régression entre la hauteur des plantes et le nombre de pollinisateurs donne : $R^2 = 0,64$.

✓ **Comment interpréter ce résultat ?**

A. 64 % de la variation du nombre de pollinisateurs est expliquée par la hauteur des plantes
B. Il y a une corrélation nulle
C. La hauteur est responsable de 64 % des erreurs de mesure
D. La relation est non significative

20. On observe r = 0,22 entre le pH du sol et la densité de vers de terre, avec p = 0,04.

✓ **Quelle conclusion tirer ?**

A. Il existe une forte corrélation positive
B. Il n'y a pas de corrélation
C. La corrélation est faible mais significative
D. La variable dépendante est le pH

Devoir

Exercice 1 : Corrélation entre deux variables

Vous disposez des données suivantes :

Étudiant	Temps de révision (heures)	Note obtenue (sur 20)
1	2	8
2	3	10
3	5	13
4	7	15
5	9	18

1. Quel est le coefficient de corrélation entre le temps de révision et la note obtenue ?
2. Quelle est la force et la direction de la relation (faible, modérée, forte ; positive ou négative) ?
3. Que pouvez-vous conclure sur la relation entre ces deux variables ?

Exercice 2 : Régression linéaire simple

Remarque: le même exemple de corrélation

1. Quelle est l'équation de la droite de régression ?
2. Que signifie le coefficient «a»(pente) dans ce contexte ?
3. Quelle proportion de la variance de la note obtenue est expliquée par le temps de révision (R^2) ?

Interprétation et rapport :

Rédigez un court rapport décrivant :

1. La relation entre le temps de révision et la note obtenue.
2. Les résultats de l'analyse de corrélation (r) et leur interprétation.
3. Les résultats de la régression (a,b,R^2) et ce qu'ils signifient dans ce contexte.
4. Ajoutez le graphique de dispersion à votre rapport.

Répondez à la question : **Un étudiant augmentera-t-il toujours sa note de manière proportionnelle en révisant plus longtemps ? Expliquez.**

(1) TERRY ANCELLE, Statistique Epidémiologie 4 eme édition MALOINE

(2) Xavier Nogues, André Garenne, Xavier Bouteiller, Virgil Fiévet LE COURS DE BIOSTATISTIQUE 2eme edition Dunod

Cité numiqo: numiqo Team (2025). numiqo: Online Statistics Calculator. DATAtab e.U. Graz, Austria. URL https://numiqo.com

Tableau I :

TABLE DU χ^2

La table donne la probabilité α pour que χ^2 égale ou dépasse une valeur donnée, en fonction du nombre de degrés de liberté ν.
Exemple : avec $\nu = 3$, pour $\chi^2 = 0{,}11$ la probabilité $\alpha = 0{,}99$.

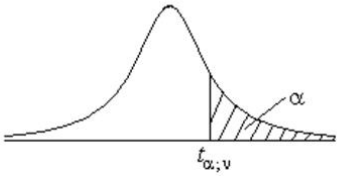
ν \ α	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,001
1	0,0002	0,001	0,004	0,016	2,71	**3,84**	5,02	**6,63**	10,83
2	0,02	0,05	0,10	0,21	4,61	**5,99**	7,38	**9,21**	13,82
3	0,11	0,22	0,35	0,58	6,25	**7,81**	9,35	**11,34**	16,27
4	0,30	0,48	0,71	1,06	7,78	**9,49**	11,14	**13,28**	18,47
5	0,55	0,83	1,15	1,61	9,24	**11,07**	12,83	**15,09**	20,51
6	0,87	1,24	1,64	2,20	10,64	**12,59**	14,45	**16,81**	22,46
7	1,24	1,69	2,17	2,83	12,02	**14,07**	16,01	**18,48**	24,32
8	1,65	2,18	2,73	3,49	13,36	**15,51**	17,53	**20,09**	26,12
9	2,09	2,70	3,33	4,17	14,68	**16,92**	19,02	**21,67**	27,88
10	2,56	3,25	3,94	4,87	15,99	**18,31**	20,48	**23,21**	29,59
11	3,05	3,82	4,57	5,58	17,28	**19,68**	21,92	**24,73**	31,26
12	3,57	4,40	5,23	6,30	18,55	**21,03**	23,34	**26,22**	32,91
13	4,11	5,01	5,89	7,04	19,81	**22,36**	24,74	**27,69**	34,53
14	4,66	5,63	6,57	7,79	21,06	**23,68**	26,12	**29,14**	36,12
15	5,23	6,26	7,26	8,55	22,31	**25,00**	27,49	**30,58**	37,70
16	5,81	6,91	7,96	9,31	23,54	**26,30**	28,85	**32,00**	39,25
17	6,41	7,56	8,67	10,09	24,77	**27,59**	30,19	**33,41**	40,79
18	7,01	8,23	9,39	10,86	25,99	**28,87**	31,53	**34,81**	42,31
19	7,63	8,91	10,12	11,65	27,20	**30,14**	32,85	**36,19**	43,82
20	8,26	9,59	10,85	12,44	28,41	**31,41**	34,17	**37,57**	45,31
21	8,90	10,28	11,59	13,24	29,62	**32,67**	35,48	**38,93**	46,80
22	9,54	10,98	12,34	14,04	30,81	**33,92**	36,78	**40,29**	48,27
23	10,20	11,69	13,09	14,85	32,01	**35,17**	38,08	**41,64**	49,73
24	10,86	12,40	13,85	15,66	33,20	**36,42**	39,36	**42,98**	51,18
25	11,52	13,12	14,61	16,47	34,38	**37,65**	40,65	**44,31**	52,62
26	12,20	13,84	15,38	17,29	35,56	**38,89**	41,92	**45,64**	54,05
27	12,88	14,57	16,15	18,11	36,74	**40,11**	43,19	**46,96**	55,48
28	13,56	15,31	16,93	18,94	37,92	**41,34**	44,46	**48,28**	56,89
29	14,26	16,05	17,71	19,77	39,09	**42,56**	45,72	**49,59**	58,30
30	14,95	16,79	18,49	20,60	40,26	**43,77**	46,98	**50,89**	59,70

Tables statistiques

Tableau II :

Table of the Student's *t*-distribution

The table gives the values of $t_{\alpha;\nu}$ where
$\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom



ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Tableau II :

QUANTILES D'ORDRE 0.95 DE LA LOI DE FISHER

Degrés de liberté du numérateur sur la première ligne
Degrés de liberté du dénominateur sur la colonne de gauche

	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927
150	3.904	3.056	2.665	2.432	2.274	2.160	2.071	2.001	1.943	1.894
200	3.888	3.041	2.650	2.417	2.259	2.144	2.056	1.985	1.927	1.878
400	3.865	3.018	2.627	2.394	2.237	2.121	2.032	1.962	1.903	1.854