1

Statistique descriptive

2 Introduction

Le premier souci de l'expérimentateur sera de chercher à mettre de l'ordre dans ses observations, de façon à pouvoir les comprendre et les synthétiser pour pouvoir ensuite les comparer et les transmettre.

La statistique descriptive peut être définie comme un instrument statistique qui permet de donner un sens, une expression à l'information recueillie. Elle rend plus intelligible une série d'observations en permettant de dégager les caractéristiques essentielles qui dissimilent dans une masse de données. Nous obtenons donc par la statistique descriptive une image concise et simplifiée de la réalité ; un résumé statistique qui caractérise l'essentiel.

Le recueil statistique est devenu une activité indispensable. La statistique descriptive est très utilisée en biologie, médecine, génétique, biométrie (lois statistiques de Mendel sur l'hérédité), étude de l'état sanitaire d'une population, étude de l'efficacité de nouveaux traitements.

En agronomie : agro-alimentaire (étude de l'efficacité d'un nouvel engrais, de nouvelles méthodes de culture, recherche de meilleurs variétés).

Etude descriptive des poids des étudiants inscrits en première année de biologie à l'université de guelma.

Une étude statistique sur l'ensemble des exploitations agricoles dans la région de guelma.

Malgré la terminologie une population n'est pas nécessairement humaine.

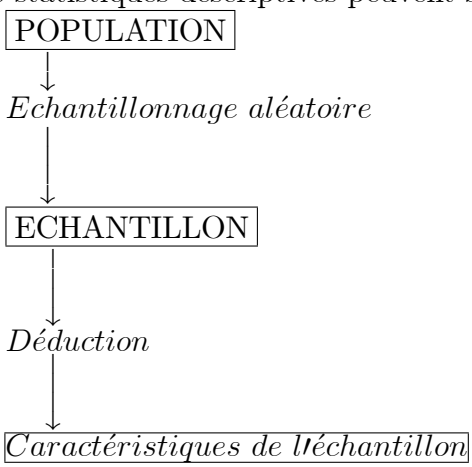
Attention aux fausses variables numériques (exp:N° de tél).

Remarque Un relevé statistique peut fournir plusieurs variables que l'on peut voir comme vecteur $i \rightarrow \begin{pmatrix} x_i \\ y_i \end{pmatrix}$

Definition 1 *:Une variable est dite discrète si elle peut prendre un nombre fini ou dénombrable (cest-à-dire que l'on peut numéroter)de valeurs ; dans le cas contraire la variables est dite continue.*

3 Echantillonnage statistique

Les statistiques descriptives peuvent se résumer par le schéma suivant :



Pour recueillir des informations sur une population statistique, l'on dispose de deux méthodes :

- la **méthode exhaustive** ou recensement où chaque individu de la population est étudié selon le ou les caractères étudiés.
- la **méthode des sondages** ou échantillonnage qui conduit à n'examiner qu'une fraction de la population, un **échantillon.**

Definition 2 *L'échantillonnage représente l'ensemble des opérations qui ont pour objet de prélever un certain nombre d'individus dans une population donnée.*

Pour que les résultats observés lors d'une étude soient généralisables à la population statistique, l'échantillon doit être représentatif de cette dernière, c'est à dire qu'il doit refléter fidèlement sa composition et sa complexité. Seul l'échantillonnage aléatoire assure la représentativité de l'échantillon.

Definition 3 *Un échantillon est qualifié d'aléatoire lorsque chaque individu de la population a une probabilité connue et non nulle d'appartenir à l'échantillon.*

Le cas particulier le plus connu est celui qui affecte à chaque individu la même probabilité d'appartenir à l'échantillon.

4 Séries simples

4.1 Objectifs

Dans ce chapitre on devra être capable :

1) de préciser ce qu'on entend par :

a) population, b) unité statistique, c) caractère, d) modalité, e) variable statistique, f) échantillon, g) fréquence relative et fréquence absolue.

2) de ranger les observations d'une série statistique par valeurs non décroissantes et d'en établir la distribution des fréquences.

3) de tracer les principales représentations graphiques associées aux distributions des

fréquences, notamment le diagramme en bâtons, L'histogramme et le polygone des fréquences.

4) de dresser le tableau des fréquences cumulées croissantes ou décroissantes et d'en tracer les courbes correspondantes.

4.2 Paramètres d'une distribution

4.2.1 Introduction

Pour résumer une longue série de données, nous pouvons les rassembleren des tableaux et en faire des représentations graphiques

Nous pouvons également résumer les données à l'aide de paramètres.

Les plus importants paramètres décrivant les distributions de fréquence sont:

- **Paramètres de position:**

valeurs centrales autour desquelles se groupent les valeurs observées.

- **Paramètres de dispersion:**

ils renseignent quant à l'étalement de la distribution des valeurs autour des valeurs centrales.

Il existe d'autres types de paramètres des distributions: Scherrer p. 132.

4.2.2 Les paramètres de position (série statistique simple)

Moyenne arithmétique (mean) E(x) = moyenne de la distribution théorique des éléments x.

$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$ désigne la moyenne arithmétique d'une population finie comportant N éléments (N = effectif). Mêmes unités physiques que x.

$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ désigne la moyenne arithmétique de n éléments (n = effectif) tirés d'une population finie ou infinie. Mêmes unités que x.

$\overline{x} = \frac{1}{n}\sum_{i=1}^{K} f_i x_i$désigne la moyenne arithmétique calculée pour n éléments divisés en k classes. f_i représente la fréquence de la classe i et x_i représente la valeur centrale de cette même classe (indice de classe).

Pour les variables quantitatives continues, la moyenne arithmétique estimée à partir des données brutes est toujours plus précise que la moyenne estimée à partir des données en classes. Ces deux valeurs peuvent différer légèrement:

Médiane (median) Symbole: Me_x; Mêmes unités physiques que x.

La médiane est la valeur de la variable qui se situe au centre de la série statistique, classée en ordre croissant. La médiane sépare la série en deux groupes d'égale importance.

• S'il y a un nombre impair d'observations, Me est une observation de la série.

Example 4 *: pour la série: [1, 32, 128, 129, 1000235], Me = 128.*

• S'il y a un nombre pair d'observations, la médiane est située entre les deux observations centrales de la série. Par convention, on utilise la moyenne de ces deux valeurs.

Exemple: pour la série [1, 32, 128, 129, 532, 1000235], Me = 128,5.

Mode (mode) Symbole: Mo_x Mêmes unités physiques que x.

Le mode est la valeur d'une variable ayant la plus forte fréquence.

• Pour une variable méristique comportant naturellement peu de classes, on trouve la classe la plus fréquente. Sa valeur est le mode.

Exemple: nombre de rayons de la nageoire anale chez le cyprinidés (Nom de famille de poissons des eaux douces, souvent appelés « poissons blancs »,) (Scott & Crossman 1974):

Nombre de rayons	10	11	12	13	14	15
Nombre de spécimens	1	2	39	20	1	1

Le mode est 12 rayons. La classe modale comporte 39 observations.

• Pour une variable quantitative continue, on divise celle-ci en classes.

• Pour les variables qualitatives, le mode correspond à la classe ayant la plus forte fréquence.

• On dit qu'une distribution de fréquences a plusieurs modes si on veut mettre en évidence le fait qu'elle a plusieurs classes non contiguës dont la fréquence est nettement plus élevée que celle des autres classes.

Comparaison entre la moyenne, la médiane et le mode Pour les distributions de fréquence symétriques et unimodales, les trois paramètres de position ont la même valeur. Ce n'est pas le cas pour les distributions asymétriques.

La moyenne obéit au principe des moindres carrés. On peut en effet montrer (Scherrer p. 146) que c'est la moyenne qui minimise la somme des carrés des écarts entre les valeurs observées et le paramètre de tendance centrale ($\sum_i (x_i - paramètre)^2$). Pour la moyenne, cette somme est toujours inférieure ou égale à la somme des carrés des écarts entre les valeurs observées et la médiane ou le mode:

$\sum_i (x_i - \overline{x})^2 \leq \sum_i (x_i - Me_x)^2$ et

$\sum_i (x_i - \overline{x})^2 \leq \sum_i (x_i - Mo_x)^2$

La moyenne, la médiane et le mode se confondent si la distribution est symétrique. Si la distribution est asymétrique à droite, la moyenne est plus décalée vers la droite que la médiane et, a fortiori, que le mode. L'inverse se produit si l'asymétrie est à gauche.

Comparaison des indicateurs de position

	Avantages
arithmétique Moyenne	*Facile à calculer *Répond au principe des moindres carrés
Médiane	*Pas influencée par les valeurs extrêmes de la v.a *Peu sensible aux variations d'amplitude *Calculable sur des caractères cycliques (saison, etc.) où la moyenne a peu de signification
Mode	*Pas influencée par les valeurs extrêmes de la v.a *Calculable sur des caractères cycliques des classes, (saison, etc.) où la moyenne a peu de signification *Bon indicateur de population hétérogène.

	Inconvenients
Moyenne arithmétique	*Fortement influencée par les valeurs extrêmes de la v.a. *Représente mal une population hétérogène (polymodale).
Médiane	*Se prête mal aux calculs statistiques, *Suppose l'équi-répartition des données *Ne représente que la valeur qui sépare des classes, l'échantillon en 2 parties égales.
Mode	*Se prête mal aux calculs statistiques, *Très sensible aux variations d'amplitude *Son calcul ne tient compte que des individus dont les valeurs se rapprochent de la classe modale .

4.2.3 Paramètres de dispersion (série statistique simple)

Les paramètres de dispersion renseignent sur l'étalement de la distribution de fréquence autour de la moyenne.

1-**Étendue de variation (range)**

Synonyme: plage de variation; Mêmes unités physiques que x.

2- Variance (variance)

Symboles: s_x^2 pour un échantillon σ^2 ("sigma^2") ou Var(x) pour une population ou distribution théorique

• Pour une population statistique d'effectif N dont la moyenne vraie μ est connue par théorie ou par hypothèse, on utilise la formule suivante: $\sigma^2 = \frac{1}{N}\sum_i^N (x_i - \overline{\mu})^2$

• Pour un échantillon d'effectif n ou pour une population d'effectif N dont on doit estimer la moyenne à l'aide de , on utilise la formule:

$s_x^2 = \frac{1}{n-1}\sum_i^n (x_i - \overline{x})^2$ ou $s_x^2 = \frac{1}{N-1}\sum_i^N (x_i - \overline{x})^2$

Unités physiques: celles de la variable au carré.

La valeur (n – 1) s'appelle le nombre de degrés de liberté. On soustrait 1 pour éliminer le biais dû au fait qu'on doit utiliser les données x une première fois pour calculer la moyenne ($= \frac{1}{n}\sum_i^n x_i$), avant le calcul de la variance. On peut montrer que, sans cette correction, la variance serait toujours sous-estimée.

Biais d'un estimateur statistique Un estimateur statistique est non biaisé si la moyenne des valeurs de cet estimateur pour tous les sous-ensembles possibles de taille n est égale à la valeur de l'estimateur pour toute la population.

Les estimateurs de la moyenne et de la variance: un exemple Soit un ensemble de 4 nombres {1, 2, 4, 5}. Considérons tous les sous ensembles possibles de 2 de ces nombres pour les estimateurs

Moyenne $= \frac{1}{n}\sum_i^n x_i$

Variance $= s^2_{x(n)} = \frac{1}{n}\left[\sum_i^N (x_i - \overline{x})^2\right]$

"Estimateur du maximum de vraisemblance"

Variance $= s^2_{x(n-1)} = \frac{1}{n-1}\sum_i^n (x_i - \overline{x})^2$

	Moyenne	$s^2_{x(n)}$	$s^2_{x(n-1)}$
{1, 2, 4, 5}	3	2,5	3,3333'
{1, 2}	1,5	0,25	0,5
{1, 4}	2,5	2,25	4,5
{1, 5}	3,0	4,00	8,0
{2, 4}	3,0	1,00	2,0
{2, 5}	3,5	2,25	4,5
{4, 5}	4,5	0,25	0,5
Moyenne	3	1,6666	3,3333

Conclusion: les estimateurs

Moyenne $= \frac{1}{n}\sum_i^n x_i$ et variance $= s^2_{x(n-1)} = \frac{1}{n-1}\sum_i^n (x_i - \overline{x})^2$ ne sont pas biaisés. L'estimateur variance $= s^2_{x(n)} = \frac{1}{n}\left[\sum_i^N (x_i - \overline{x})^2\right]$ est biaisé.

4.2.4 Propriétés de la variance

• Si tous les x_i sont égaux, la variance est nulle puisque tous les termes $(x_i - \overline{x})^2$ composant la somme sont nuls.

• s^2_x augmente à mesure que la variabilité augmente. La variance mesure donc bien la variabilité des données.

• Pour estimer la variance, on doit disposer au moins de deux observations. Avec une seule observation, $(n-1) = 0$ et la valeur de s^2_x devient

indéterminée. La formule correspond bien à notre intuition: on ne peut rien conclure quant à la variabilité d'une variable à partir d'une seule observation.

• Les unités physiques de la variance sont celles de la variable x au carré.

Lorsqu'on calcule la variance d'un ensemble de données, on ne veut pas avoir à manipuler les données deux fois: une première fois pour calculer la moyenne et une seconde pour calculer les écarts à la moyenne. On peut montrer que les formules suivantes donnent exactement le même résultat (Scherrer p. 159):

$s_x^2 = \frac{1}{n-1}\left[\sum_i^n x_i^2 - \frac{\left(\sum_i^n x_i\right)^2}{n}\right]$ et $s_{x(n-1)}^2 = \frac{1}{n\,(n-1)}\left[\sum_i^n x_i^2 - \left(\sum_i^n x_i\right)^2\right]$

Attention: $\sum_i^n x_i^2 \neq \left(\sum_i^n x_i\right)^2$

4.2.5 L'écart type (standard deviation)

Symboles: σ pour une population ou une distribution théorique s_x pour un échantillon

Formule: $s_x = \sqrt{s_x^2}$; Unités physiques: celles de la variable x.

4.2.6 Coefficient de variation (coefficient of variation)

Symbole: C.V., CV ou V

Formule: $CV = \frac{100 * s_x}{\overline{x}}$ **

Unités physiques: aucune, puisque les unités du numérateur annulent celles du dénominateur. Le coefficient de variation permet donc de comparer la variation de variables exprimées originellement dans des unités physiques différentes.

L'équation ** n'a de sens que pour les variables quantitatives à échelle de variation relative à un vrai zéro.

Il estutilisé dès qu'il faut comparer deux séries statistiques exprimées dans deux systèmes d'unités différents, cet indice est sans dimension et donc indépendant des unités choisies.

4.2.7 Les moments d'une distribution (moments)

Il est souvent utile de centrer les valeurs d'une variable sur la moyenne.

L'écart de la moyenne pour une observation i,$(x_i - \overline{x})$ s'appelle aussi un moment central, ou moment par rapport à la moyenne.

La moyenne possède la propriété que la somme des moments, $\sum (x_i - \overline{x})$, ou somme des écarts à la moyenne, est 0. Cette notion sert de base à la définition d'une série de statistiques des moments.

Le moment de premier ordre d'une distribution est $m_1 = \frac{1}{N} \sum (x_i - \overline{x}) = 0$
Le moment de deuxième ordre d'une distribution est $m_2 = \frac{1}{N} \sum (x_i - \overline{x})^2$
Le moment de deuxième ordre, m_2, est la variance d'une distribution théorique.
La variance d'un échantillon d'effectif n, corrigée pour le biais d'estimation de
la moyenne $\overline{x}$, est dérivée de m_2:$s_2^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2$
Moment de troisième ordre: $m_3 = \frac{1}{N} \sum (x_i - \overline{x})^3$
Moment de quatrième ordre: $m_4 = \frac{1}{N} \sum (x_i - \overline{x})^4$

4.3 Notions fondamentales et vocabilaire de base

4.3.1 Ensemble statistique, population statistique, unité statistique.

Un des objectifs statistiques est d'étudier les propriétés numériques d'ensembles comportant de nombreux individus ou unité statistiques. Ainsi la réunion de toutes les unités statistiques possibles (ou éléments ou individus) constitue l'ensemble statistique ou la population statistique (c'est l'ensemble étudier par le statisticien) ; ce sont sur les unités statistiques (ou individus) que sont recueillies les observations.

Remark 5 Il est important que la population étudiée soit définie correctement pour Que l'on puisse dire si une unité appartient ou non à la population.

Example 6 *Dans l'exemple 1 l'ensemble des étudiants inscrits en troisième année chimie industrielle constitue la population statistique, chaque étudiant est une unité statistique.*

Remark 7 On constate qu'une population

présente des caractères propres qui retrouvent chez toutes les unités qui la composent.

L'ensemble des exploitations fruitières dans la région de constantine est une population où chaque fruit est une unité statistique.

Les unités statistiques sont : (pomme, poire, cerise, pêche, abricot, amande, oronge, citron, coin, raisin, noix, fraise, mures).

4.3.2 Caractères:

Definition 8 *Dans une étude spécifique on peut s'intéresser à certaines particularités des individus, ces particularités qu'on appelle caractères*

Example 9 *Lecaractères peuvent être : la taille, l'age, lepoids, le sexe, l'état matrimoniale, lieu d'habitation , le nombre d'arbres, la quantité de production par an, la couleur du fruit,...*

Dans les listes précédentes on remarque que certains caractères sont mesurables (l'age, le poids, le nombre d'arbres, la quantité de production par an) d'autre non (Le sexe, l'état matrimonial, le lieu d'habitation, la couleur du fruit). Ceci nous amène à apporter la distinction suivante:

Caractère quantitatif, Caractère qualitatif. Les résultas de l'observation d'un caractère peuvent s'exprimer d'une manière quantitative ou qualitative selon qu'ils sont mesurables ou non.

Example 10 *le poids est une variable quantitative ; l'état matrimonial est une variable qualitative.*

Modalités des caractères : Les caractères peuvent présenter plusieurs modalités. C'est-à-dire des spécificités qui leurs sont propres. Les modalités d'un caractère doivent être définies de telle sorte que toute unité statistique appartienne à une seule, Il est donc nécessaire que les modalités que peut présenter un caractère soient incompatibles et exhaustives.

Example 11 *Modalités pour les caractères mentionnés précédemment*

a- Pour la taille des étudiants on pouvait envisager les modalités suivantes : petit, moyen, grand. Puisque ce caractère est mesurable, on pourrait toutefois préciser les modalités d'une manière quantitative : 1,4 mais moins de 1,6, de 1,6 mais moins de 1,8, de 1,8 mais moins de 2 m. Dans ce cas nous avons trois modalités.

b- Etat matrimonial peut présenter les modalités suivantes : célibataire, marié, veuf, divorcé. Ce caractère présente quatre modalités.

Remark 12 On pourrait donc préciser qu'un caractère est quantitatif si ses modalités sont mesurables, et qualitatif si ne le sont pas.

Example 13 *L'age, le poids, la taille, le nombre d'enfants à charge, le taux de cholestérol, le nombre de globules blancs, le gain moyen quotidien d'animaux à l'engraissement sont des variables quantitatives ; Le sexe, la nationalité, la couleur des yeux sont des caractères qualitatifs.*

Definition 14 *Une variable statistique est un caractère statistique quantitatif.*

4.3.3 Variables discrètes et continues

Variables discrètes : Une variable statistique est dite discrète si ses modalités ne prennent que des valeurs isolées.

Example 15 *Le nombre d'enfant à charge, la pointure sont des variables discrètes.*

Remark 16 Il convient d'éviter la confusion classique entre discret et entier, une variable prenant ses valeurs dans $\blacksquare$ est une variable discrète, mais la réciproque n'est pas forcément vraie.

Variables continues : Une variable statistique est continue si elle peut prendre n'importe qu'elle valeur entre deux bornes données.

D'une façon générale liées à des mesures où interviennent le temps, l'espace, la masse sont des variables continues.

Example 17 *L'age, la taille, le poids, la surface agricole exploitée sont des variables continues.*

4.4 Mise en forme des informations

4.4.1 Echantillon aléatoire:

Definition 18 *Un échantillon est un groupe restreint (ou sous ensemble) d'unités statistiques tirées d'une population de manière telle que les résultats de l'analyse pourront être étendus à la population.*

4.4.2 Fréquences, Effectifs

Definition 19 *La fréquence associée à une valeur d'une variable statistique est le nombre de fois que cette valeur se rencontre dans l'échantillon observé (ou dans la population). On utilise également les termes « effectif » ou « Fréquence absolue » pour identifier cette fréquence.*

Dans le cas d'une distribution par classe, la fréquence d'une classe correspondra au nombre de mesures dont les résultats appartiennent à cette classe particulière.

Definition 20 *La fréquence relative associée à une valeur d'une variable statistique est le rapport entre la fréquence correspondante à cette valeur et le nombre total de valeurs qui ont été observées sur les unités statistiques.*

Dans le cas d'une distribution par classe, la fréquence relative sera le rapport entre la fréquence d'une classe et la somme des fréquences de toutes les classes (le nombre total d'observations).

Example 21 *:Considérons à titre d'exemple une étude sur la croissance des peupliers nous avons observé un échantillon de 35 arbres âgés de 5 ans, pour lesquels on a mesuré la circonfirance à 13 cm du sol.*

Tableau des données initiales:

28.9	28.2	28.7	28.0	29.0	29.1	28.6
28.7	28.4	28.1	28.5	28.6	28.8	28.77
29.0	29.2	28.5	28.7	27.8	28.6	28.5
28.6	28.0	28.4	28.8	29.0	28.7	28.4
28.9	28.8	28.4	28.6	28.5	28.7	28.2

Les valeurs seront rangées par ordre croissant: $x_1 = 27.8$, $n_1 = 1$, $x_5 = 28.4$ et $n_5 = 4$. $f_5 = \frac{n_5}{n} = \frac{4}{35} = 0.1143$.

Multipliée par 100, cette fréquence relative donne le pourcentage d'apparition de la valeur considérée.

Valeur du caractère x_i	*Eff* n_i	*Fré* f_i	*%* f_i*100
27,8	1	0,0286	2,86
28,0	2	0,0571	5,71
28,1	1	0,0286	2,86
28,2	2	0,0571	5,71
28,4	4	0,1143	11,43
28,5	4	0,1143	11,43
28,6	5	0,1429	14,29
28,7	6	0,1714	17,14
28,8	3	0,0857	8,57
28,9	2	0,0571	5,71
29,0	3	0,0857	8,57
29,1	1	0,0286	2,86
29,2	1	0,0286	2,86

4.5 Résumés d'une série par ses paramètres

Le choix d'un résumé d'une série statistisque par ses paramètres n'est pas des compétences du mathématicien, ce sont celles des statisticiens, des économistes... suivant ce qu'ils veulent en faire. En tous cas, une étude statistique est accompagnéé de commentaires qui justifient la méthode employée et les choix faits. On peut cependant indiquer les résumés possibles suivants :

Le couple (médiane ; étendue)

Le couple (moyenne ; étendue)

Ces deux couples sont simples à obtenir mais ils ne permettent pas de positionner le maximum et le minimum de la série. De plus l'étendue est un caractère de dispersion très grossier car sensible aux valeurs extrêmes.

Le couple (médiane ; intervalle interquartile) Il est insensible aux valeurs extrêmes.

L'ensemble {minimum, premier quartile, médiane, troisième quartile, maximum}.

Il permet de construire un diagramme en boîte et donc de mieux visualiser le comportement d'une série (notamment sa dispersion) et de comparer des séries. Il présente un inconvénient : la connaisance de

ces paramètres pour deux séries ne permet pas de calculer les paramètres du regroupement des deux séries.

Enfin, le couple (moyenne, écart-type).

Ce couple permet de faire des calculs sur des regroupements ... et il permet à l'aide de l'inégalité de Bienaymé-Tchebychev d'avoir une idée assez précise de la répartition de la série.

Par exemple on sait que pour une série quelconque la proportion des valeurs de la série en dehors de l'intervalle $[\overline{x} - 2s_x;\ \overline{x} + 2s_x]$ est inférieur à 25% et la proportion des valeurs de la série en dehors de l'intervalle $[\overline{x} - 3s_x;\ \overline{x} + 3s_x]$ est inférieur à 12%

5

Questions de cours

1- Préciser en quelques mots ce qu'est la statistique descriptive.

2- Les éléments sur lesquels porte une étude statistique sont appelés...

3- La réunion de toutes les unités statistique possible constitue la ...

4- Les particularités que peut présenter une unité statistique sont appelés ...

5- Une unité statistique peur-elle comporter plusieurs caractères ?.

6- Si dans une enquête on s'intéresse à un groupe d'étudiants étranges de cette université (ceci définit la population statistique),

a)- Quelle est l'unité statistique ?.

b)- Pouvez-vous énoncer trois caractères associés à l'unité statistique décrite en a).

7- Est-ce que les caractères précisés en 6, sont de nature quantitative ou Qualitative ?.

8- Un caractère est également connu sous le nom de ...

9- Est-ce que une variable qualitative peut être dénombrable ou mesurable ?.

10- Une variable est dite .....si elle peut être exprimée numériquement.

11- Si certaines valeurs seulement sont possibles pour une variable quantitative, nous disons alors qu'elle est .....

12- Une variable quantitative qui ne peut prendre que des valeurs entières est dite .....

13- Une variable qui peut prendre toutes les valeurs comprises dans un intervalle fini ou infini est dite .....

14- Un groupe restreint d'unités statistiques tirées d'une population est appelé ... Le nombre d'unités statistiques qui le constitue correspond à la ...

15- La grandeur finie servant de base à la mesure de toutes les unités statistiques de même espèce s'appelle ...

16- Le nombre de fois que se rencontre la même valeur dans un échantillon s'appelle ... de cette valeur. D'autres termes sont également employés soit ... ou...

17- Le rapport entre la fréquence d'une classe et le nombre total d'observations s'appelle ....

18- U n ensemble d'observations associé à un caractère s'appelle ...

19- La première étape dans le dépouillement d'une série d'observations consiste à les ... par valeurs ...

20- Le résultat de la répartition des observations en classes accompagnées des

fréquences respectives s'appelle ...

21- Une façon pratique de déterminer le nombre de classe nécessaire au dépouillement d'une série numérique est d'utiliser ...

22- Dans la limite du possible, chaque classe devrait avoir la ...

23- L'écart entre la plus grande et la plus petite valeur dans une série s'appelle ...

24- Les nombres entre lesquelles sont classées les observations s'appellent ...

25- La somme des fréquences relatives de toutes les classes égale ...

26- Une distribution des fréquences peut se présenter avec des classes .....ou des classes ...

27-La représentation graphique d'une distribution des fréquences dans le cas d'une variable discontinue se présente sous forme ... ; Dans le cas d'une variable continue, on utilise ....

28- Dans la représentation graphique mentionnée en 27, quel avantage y a-t-il à utiliser en ordonnée, lez fréquences relatives au lieu des fréquences absolues ?.

29- Dans le tracé d'un histogramme, la hauteur de chaque rectangle est telle que sa surface est ... à la fréquence (absolue ou relative).

30 Lorsque les classe ont la même amplitude, chaque rectangle aura comme hauteur le nombre correspondant à ...

31- Si la distribution des fréquences se présente avec des classes d'amplitude inégales, il faut ... les fréquences pour que la surface de chaque rectangle soit toujours proportionnelle à la ...

32- Pour représenter la distribution de fréquences sous forme de courbe, on utilise un ...

33- Le nombre de cotés d'un polygone de fréquences est égal au nombre de classes de la distribution de fréquences plus ...

34- Si l'on veut des histogrammes constitués à partir d'échantillons de tailles différentes, il est alors préférable d'utiliser, en ordonnées, les ... au lieu des ...

35- Dans une distribution de fréquences dont les classes extrêmes sont ouvertes, on représente ces classes dans le tracé de

l'histogramme, avec une amplitude égale à ...

36- Les résultats qu'on obtient avec les courbes cumulatives sont entachés d'une erreur

d'approximation due à ... à l'intérieure des classes.

5.1 Réponses

1-Instrument statistique qui permet de donner un sens, une expression à l'information recueillie.—2- unités statistiques..3- population statistique.; 4-caractères.; 5-oui; 6 et 7-a: etudiant etranger.; b: ntionalité (quali), etat matrimonial (quali), âge.(quant); 8-variable statistique.; 9-ni l'un ni l'autre; 10-quantitative.; 11-discontinue.; 12-discrète.; 13-continue.; 14–echantillon.; 15-l'unité de mesure.; 16 fréquence; effectife ou fréquence absolue.; 17-fréquence relative.; 18-série numérique. ; 19-ranger; non décroissante.; 20-distribution de fréquences.; 21-règle de Sturge.; 22-même amplitude. ; 23-l'étendue.; 24-limites des classe.; 25-1.; 26-fermées; ouvertes.; 27-d'un diagramme en bâton; un histogramme.; 28- les nombres en ordonnée seront toujours compris entre 0 et un.; 29-proportionnelle.; 30-la fréquence.;

31-rectifier; fréquence.; 32-polygone des

fréquences.; 33-1.; 34-fréquences relatives,

fréquences absolues.; 35-l'amplitude de base.; 36-l'interpolation linéaire.

– – – – – – – – – – – –

6

Séries doubles

6.1 caractères

Definition 22 *On appelle caractère statistique double un couple (X,Y), où X et Y sont deux caractères.*

X s'appelle premier caractère marginal.
Y s'appelle deuxième caractère marginal.
On appelle :
1- Modalité toute valeur (x_i,y_i) , valeur prise par les caractères.
2- Effectif de la modalité (x_i,y_i) le cardinal n_i= nombre d'individus caractérisés par cette modalité.
3- Fréquence de la modalité (x_i,y_i) le réel f_i=n_i/N , avec :

$$N = \sum_j \sum_i n_{ij} = \sum_i \sum_j n_{ji}$$

Remark 23 *Les effectifs et fréquence cumulés ne sont pas définis, car il n'existe pas d'ordre dans* $\mathbb{R}^2$.

On peut aussi procéder à des regroupements en classes :
- du premier caractère marginal
- du deuxième caractère marginal.
- des deux.

6.2 Analyse simultanée de deux variables mesurées sur les mêmes éléments.

• Un élément peut représenter un objet physique; on devra mesurer les deux variables sur chaque objet.

Example 24 *Élément: une oie. Variables: longueur de l'aile, longueur du torse.*

L'élément peut représenter une autre forme d'appariement des mesures.

Example 25 *Élément: une période d'observation. Variables: numéro du jour d'observation, nombre d'espèces d'oiseaux observées à une mangeoire entre 7h00 et 8h00 am.*

Élément	Var. 1	Var. 2
él. 1	x_1	y_1
él. 2	x_2	y_2
él. 3	x_3	y_3
.		
.		
.		
él. n	x_n	y_n

Pour qu'il y ait série statistique, il faut qu'au moins l'une des deux variables soit aléatoire.

Cas 1: Une variable aléatoire et une variable contrôlée

Example 26 *intensité de l'assimilation chlorophyllienne (variable aléatoire) en fonction de l'éclairement (contrôlé par l'expérimentateur).*

Cas 2: Deux variables aléatoires

Example 27 *Abondance de la récolte (aléatoire) en fonction du nombre de jours d'ensoleillement dans l'année (aléatoire).*

Paramètres d'une série statistique double **Objectif:** décrire la position et la forme de la distribution conjointe de deux variables.

Paramètre de position: le point moyen (= centre de gravité, centroïde). Point moyen = $(\overline{x}, \overline{y})$

Paramètres de dispersion: Les variances estimées s_x^2 et s_y^2

où $s_x^2 = \dfrac{1}{n-1} \sum (x_i - \overline{x})^2$

La covariance qui est également estimée à partir des variables centrées $(x_i - \overline{x})$ et $(y_i - \overline{y})$:

$$s_{xy} = \frac{1}{n-1} \sum (x_i - \overline{x})(y_i - \overline{y})$$

Il s'agit d'une formule de variance généralisée. Si $x = y$, alors$s_{xy} = s_x^2$

Comme pour la variance, il existe une formule qui ne requiert pas le calcul préalable des moyennes:

$$s_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n(n-1)}$$

Contrairement à la variance qui ne peut être négative, la covariance peut être positive, nulle ou négative.

Interprétation du signe de la covariance à partir du graphique desdonnées centrées:

Dans quels quadrats se trouve la majorité des points? Fig. 4.18

La covariance nous renseigne sur l'inclinaison du nuage de points, mais elle ne nous donne aucune idée de l'intensité de la liaison existant entre les variables x et y. En effet, la covariance peut augmenter alors que la liaison entre x et y reste constante.

• La covariance dépend de la dispersion des points.

• Il faudra trouver une autre mesure qui nous renseignera sur l'intensité de la relation.B

6.2.1 Paramètre de liaison linéaire entre deux variables: la corrélation linéaire de Pearson

• **Covariance:** mesure de la dispersion conjointe de deux variables quantitatives (x et y) autour de leur moyenne.

• **Corrélation:** mesure de la liaison entre deux variables x et y.

• **Corrélation linéaire de Pearson:** mesure de la liaison linéaire entre deux variables quantitatives x et y.

La corrélation se calcule à partir de séries statistiques (doubles, triples, etc.) En d'autres termes, au moins l'une des variables doit être aléatoire.

La corrélation linéaire (r de Pearson) est la covariance de deux variables centrées réduites.

Cette mesure de liaison a été développée par Francis Galton et Karl Pearson.

Étapes de calcul:

1- Transformation de x et y en variables centrées réduites:

$x'_i = \dfrac{x_i - \overline{x}}{s_x}$ et $y'_i = \dfrac{y_i - \overline{y}}{s_y}$

2- Calcul de la covariance entre ces nouvelles variables:

$$s_{x'y'} = \frac{1}{n-1} \sum \left(x'_i - \overline{x'}\right) \left(y'_i - \overline{y'}\right)$$

$$r_{xy} = \frac{1}{n-1} \sum \left(\frac{x_i - \overline{x}}{s_x} - 0\right) \left(\frac{y_i - \overline{y}}{s_y} - 0\right)$$

On soustrait 0 parce que la moyenne des variables centrées réduites est nulle

On peut donc calculer la corrélation linéaire de Pearson directement à l'aide des formules suivantes:

$$r_{xy} = \frac{\dfrac{1}{n-1} \sum (x_i - \overline{x}) (y_i - \overline{y})}{s_x s_y}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}, \; r_{xy} = \frac{s_x^2}{s_x s_x} = 1$$

$$r_{xy} = \frac{\sum (x_i - \overline{x}) (y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}$$

Propriété 1: $-1 \leq r \leq +1$

$r = +1$ ou $r = -1$ si les points forment une ligne droite dans le diagramme de dispersion.

Propriété2: Le signe de r est le même que le signe de la covariance. Il indique si la relation est de pente positive (croissante) ou négative (décroissante).

Part I

CHAPITRE II

Part II

REGRESSION ET CORRELATION

7

PARTIE 1 : Régression simple

– –

8 1-*Définitions et notions de base*

8.1 1-1-Observation:

Lors d'une expérimentation, on peut être amené à differents mements à mesurer differents paramètres liés au phénomène ou processus étudié.

Ainsi des medecins scolaires noterent pour chaque élève d'une école, leur taille et leur poids. A chaque élève est associée une observation. L'ensemble des élèves est parfois appelé aussi une population et chaque élève est un individu.

8.1.1 Exemple 1:

On a mesuré les hauteurs en centimètres de plants de blé d'un même champ dont on a tiré un echantillon chaque semaine; Les observations sont données dans le tableau suivant:

âge en semaines x_i	1	2	3	4	5	6	7	8
hauteur en cm y_i	5	12	15	22	34	35	41	48

Les variables observées simultanément sont donc:
X: âge en semaines et Y: hauteur en centimètres.
Cette série peut être representée par
le graphique ci- contre:

8.1.2 1-2-Relation fonctionnelle:

Etant données deux variables statistiques X et Y, on dit qu'il existe une relation fonctionnelle entre de X vers Y si à chaque valeur de la variable X est associée une valeur et une seule de la variable Y.

8.1.3 1-3-Modèle:

S'il existe une relation fonctionnelle entre les variable X et Y et si f est la fonction donnant pour toute valeur x de X la valeur correspondante y de Y, $(Y = f(x))$ on dit que f est un modèle du phénomène étudié.

8.1.4 1-4-Valeur observée:

Etant donnée une variable statistique X on appelle valeur (obsevée de cette variable, toute valeur de cette variable relevée au cours d'une observation.

8.1.5 1-5-Valeur théorique:

S'il existe une relation fonctionnelle entre les variable X et Y de X vers Y et si f un modèle de cette relation $(Y = f(X))$.
A toute valeurs x_i de X est associée une valeur $y_i = f(x_i)$.
y_i est la valeur théorique ou valeur expliquée par le modèle.

8.1.6 1-6-Variable statistique centrée réduite:

Soit X une variable statistique et $\overline{X}$ sa moyenne, S_X son écart-type.
On appelle varible statistique centrée la varible $X - \overline{X}$.
On appelle varible statistique centrée réduite la varible $Z = \frac{X-\overline{X}}{S_X}$

8.1.7 1-7-Varible explicative et variable expliquée.

S'il existe une relation fonctionnelle entre les variable X et Y de X vers Y et si f un modèle de cette relation $(Y = f(X))$ on dit que X est la variable explicative et Y est la variable expliquée.

8.1.8 1-8-Nuage de points :

Soit R un repère cartésien du plan.

Si au cours de n observations w_1 , w_2 , ..., w_n.on relève les vaaleurs $X(w_i)$ et $Y(w_i)$ prises par deux variables statistiques X et Y On appelle nuage de points l'ensemble des points du plan dont les cordonnées dans le repère R sont les couples $(X(w_i)$, $Y(w_i))$.

8.1.9 1-9-Point moyen:

On appelle point moyen d'un nuage de points, le point dont les cordonnées sont les moyennes arithmetiques des cordonnées des points du nuage.

8.2 2-Principe d'ajustement:

Lorsqu'on procède à un ajustement, on recherche à partir des observations une fonction notée f telle que les deux variables X et Y soient liées par la relation $Y = f(X)$.

8.3 3-Intérêt de l'ajustement:

Connaissant l'équation d'ajustement de tendance générale, il est alors aisé de prévoire les résultas des autres observations. Par exemple le résultat de notre experience à une date ulterieur si notre experience dépond du temps. On en distingue deux familles:

8.4 4-L'ajustement linéaire:

Dans ce cas le nuage statistique est approché par une doite, la fonction f est donc donnée par $y = f(x) = ax + b$; Le nombre a est appelé pente de la droite.

8.4.1 4-1-La méthode des points moyens (méthode de Mayer)

Après classement ordonné selon le x_i croissants, on partage le nuage en deux sous nuages égaux (ou égaux à une unité près si le nombre d'obsevations est impair). On considèr pour chaque sous nuage un point appelé point moyen.

Les cordonnées du premier point sont $(\overline{x_1}, \overline{y_1})$.

$\overline{x_1}$: Moyen arithmétique des abscisses x_i du premier sous nuage.

$\overline{y_1}$: Moyen arithmétique des ordonnées y_j du premier sous nuage.

On note ce point $G_1(\overline{x_1}, \overline{y_1})$.

On procède en suite de même pour le reste des valeurs du second sous nuage, on obtient un second point moyen noté $G_2(\overline{x_2}, \overline{y_2})$.

La droite de Mayer est l'unique droite qui passe par ces deux points, son équation est de la forme $y = ax + b$.

Or $G_1(\overline{x_1}, \overline{y_1})$ appartient à la droite; Ces cordonnées vérifient donc l'équation et en conséquence: $\overline{y_1} = a\overline{x_1} + b$.

Il en est de même pour $G_2(\overline{x_2}, \overline{y_2})$ et donc $\overline{y_2} = a\overline{x_2} + b$.

Pour déterminer a et b, on résoud le système de deux équations à deux inconues donnés par: $\begin{cases} \overline{y_1} = a\overline{x_1} + b \\ \overline{y_2} = a\overline{x_2} + b \end{cases}$

On trouve, en retranchant membre à membre la première équation de la deuxième:

$$\overline{y_2} - \overline{y_1} = a\left(\overline{x_2} - \overline{x_1}\right) \quad Soit\ a = \frac{\overline{y_2} - \overline{y_1}}{\overline{x_2} - \overline{x_1}}$$

Pour trouver b il suffit de roporter la valeur de a dans l'une des deux équations précédentes.

8.4.2 Application à l'exemple 1:

Le premier sous nuage de la distribution des couples $(x_i; y_i)$ est l'ensemble des points $(1; 5)$ $(2; 12)$ $(3; 15)$ $(4; 22)$.

$$On\ calcule\ \overline{x_1} = \frac{1+2+3+4}{4} = 2.5, \quad \overline{y_1} = \frac{5+12+15+22}{4} = 13.5$$

$$d'où \quad G_1(\overline{x_1}, \overline{y_1}) = G_1(2.5, 13.5)$$

$$De\ même \overline{x_2} = \frac{5+6+7+8}{4} = 6.5, \quad \overline{y_2} = \frac{34+35+41+48}{4} = 39.5$$

$$G_2(\overline{x_2}, \overline{y_2}) = G_1(6.5, 39.5)$$

Pour déterminer a et b, on résoud le système $\begin{cases} 13.5 = 2.5a + b \\ 39.5 = 6.5a + b \end{cases}$

d'où $a = 6.5$; En portant cette valeur dans la première équationon obtient:

$b = -2.75$.

L'équation de la droite de Mayer est donc: $Y = 6.5x - 2.75$.

8.4.3 4-2-<u>La méthode des moindres carrés.</u>

Les points du nuage sont numérotés $M_1, M_2, ..., M_n$.

L'orsque les points du nuage semblent à peut- près alignés, on essaie de trouver l'équation d'une droite Δ approchant mieux le nuage.

Le point $G(\overline{x}, \overline{y})$ s'appelle point myen du nuage.

Soit Δ la droite d'équation $Y = ax + b$.

Chercher l'équation $Y = ax + b$ de la droite d'ajustement (Δ) de y en x par la méthode des moindes carrés revient à chercher a et b telles que $\sum_k^n n_k (d_k)^2$ soit minimale.

$$Comme \quad (d_k)^2 = [(y_k - ax) - b]^2$$

alors $S = \sum_k^n n_k (d_k)^2 = \sum_k^n n_k (y_k - ax)^2 - 2b \sum n_k (y_k - ax) + b^2 \sum_k^n n_k$.

On remarque que S est un polynome du second degré en b dont le coeffecient de b^2 vaut $\sum_k^n n_k = N > 0$.

Il sera minimum si sa dérivée par rapport à b est nulle.

soit: $-2 \sum n_k (y_k - ax) + 2Nb = 0 \Longrightarrow b = \frac{1}{N} \sum n_k (y_k - ax)$.

Ou encore $b = \overline{Y} - A\overline{X}$, n a:

$$a = \frac{\frac{1}{N} \sum n_k \left(x_k - \overline{X}\right)\left(y_k - \overline{Y}\right)}{\frac{1}{N} \sum n_k \left(x_k - \overline{X}\right)^2} = \frac{COV(X,Y)}{varX} = \frac{\sum x_k y_k - n\overline{XY}}{\sum x_k^2 - \left(\overline{X}\right)^2}$$

Si a et b sont les coefficients directeurs des droites d'ajustement de Y en X et de X en Y, on a : $aa' = r^2$où r est le coefficient de corrélation de X et Y.

8.4.4 4-3-Relation entre le coefficient de corrélation et l'ajustement mathématique:

1-Le coefficient de corrélation mesure la qualité d'ajustement affine.

2- L'ajustement est d'autant meilleur que $|r| \approx 1$ (en fait en considère que l'ajustement est correcte si $r \geqslant 0.8$)

3-L'ajustement est d'autant plus mauvais que $|r| \approx 0$.

4-Si $r = 0$, X et Y sont dits non corrélés ou indépendants.

5-$|r| = 1$ si et seulement si $Y = ax + b$.

REMARQUE :

Une relation statistique, détectée par le coefficient de corréltion ou par un grapphique, ne montre jamais de relation causale entre deux variables. La causalité ne peut être déduite que d'une analyse non statistique des données.

8.5 Conséquence importante:

La droite d'ajustement toujours passe par le point moyen: $\left(\overline{X}\ ;\overline{Y}\right)$.

8.5.1 Application à l'exemple 1:

On calcule dans un premier temps les résultas intermédiaires en utilisant le tableau suivant:

x_i	y_i	$x_i y_i$	x_i^2
1	5	5	1
2	12	24	4
3	15	45	9
4	34	136	16
5	25	225	25
6	32	192	36
7	41	287	49
8	48	386	64
$X = 4.5$	$Y = 26.5$		204

$a = \frac{1213 - 8*4.5*26.5}{204 - 8*20.5} = 6.116$

$b = \overline{Y} - a\overline{X} = -1.25$

$soit\ Y = 6.166x - 1.25$

La dixième semaine correspond à $x = 10$

donc l'estimation fournie par cette équation est

$y = 6.166 * 10 - 1.25 = 60.41$

On cherche la droite d'ajustement de X en Y.

$$\sum_{1}^{8} y_i^2 = 7244; a' = \frac{259}{7244 - 8 * 702.25} = 0.159$$
$$b' = \overline{Y} - a'\overline{X} = 26.5 - 0.59 * 4.5 = 25.78$$

$x = 0.159y + 25.78;\ r = (0.159 * 6.116)^{\frac{1}{2}} = 0.99.$

8.6 5-L'ajustement exponentiel:

S'il existe une relation fonctionnelle entre les variables X et Y et si f est un modèle de cette relation ($Y = f(X)$), on dit que f est un modèle exponentiel ou un ajustement exponentiel de Y en X si la fonction est de la forme $f(X) = ka^x$.

Dans le cas où $a = \varrho$ on aura le modèle classique $y = k\varrho^x$.

Dans un ajustement exponentiel on cherche a et b pour l'équation $y = ba^x$. Il est facil de trover un changement de variable qui permet de revenir à un modèle linéair, en prenant le logarithme de chaque mombre, on a: $\ln y = \ln b + x \ln a$.

donc; En posant $Y = \ln y$; $B = \ln b$; $A = \ln a$

L'équation devient $Y = B + Ax1213$

On sait trouver A et B depuis les paragrahes précedents (méthode des moindres carrés), et on reviendra aux valeurs initiales en considérant que : $a = \varrho^A \ \ et \ \ b = \varrho^B$

8.7 6-L'ajustement logarithmique:

Dans un ajustement logarithmique on cherche a et b pour l'équation $y = alnx + b$, en posant $X = lnx$, on obtient une équation de la forme : $Y = aX + b$.

Ici encore, on est ramené à traiter un modèle linéaire.

8.8 7-L'ajustement parabolique:

c'est un cas que l'on ne peut pas ramener à un modèle linéaire.

Il faut rendre minimale la quantité $\sum_{i=1}^{n} (y_i - y_i')^2$ pour déterminer a, b et c en fonction des x_i et y_i.

8.8.1 APPLICATION 1:

Pour tester l'éfficacité d'un fumier on a procédé à une série d'essais sur différentes parcelles, l'expérimentation a donnée les résultats suivants:

x_i	5	6	7	7
y_i	38.04	40.5	41.8	42.3

x_i =la quantité du fumier en Kg par m^2

y_i =le rendement en quantaux par hectar

1-Représenter graphiquement les points $(x_i; y_i)$.

2-Calculer les valeurs $z_i = ln(43 - x_i)$.

3-Représenter graphiquement les points $(x_i; z_i)$.

4-Déterminer l'equation de la droite de z_i en fonction des x_i par la méthode des moindres carrés.

5-Déduire l'expréssion de Y en fonction des x.

8.8.2 APPLICATION 2:

Le tableau ci après donne les résultats de 7 déterminations de la distance nécessaire à l'arrêt d'une voiture automobile (y en m) suivant sa vitesse (x en Km/h):

vitesse (x)	33	49	65	33	79	49	93
distance de freinage (y)	5.3	14.45	20.21	6.5	38.45	11.23	50.42

1-Représenter le nuage des 7 points de cordonnées $(x_i; y_i)$.

2- On fait le changement de variable $z = (y)^{\frac{1}{2}}$.

a- Représenter le nuage des 7 points de cordonnées $(x_i; z_i)$.

b- Calculer le coefficient de corrélation linéaire entre x et z.

c- Que peut-on conclure?.

d-Ajuster par la méthode des moindres carrés z en fonction de x.

e- Donner l'équation $z = ax + b$.

f- En déduire $y = f(x)$.

g- Utiliser la relation entre x et y pour déterminer la distance nécissaire à l'arrêt d'une voiture lancée à 120 Km/h.

9

PARTIE 2: REGRESSION LINEAIRE MULTIPLE

– –

10

I – LES MATRICES

1 – *DEFINITIONS*

Définition 1: On appelle matrice Aou A_{pxq} ou $[a_{ij}]$ untableau rectangulaire constitué de $p * q$ éléments $(a_{ij})_{i=1,p}^{j=1,q}$

$$A = A_{pxq} = [a_{ij}] = \begin{bmatrix} a_{11} & . & . & . & a_{1q} \\ . & . & . & . & . \\ . & . & a_{ij} & . & . \\ . & . & . & . & . \\ a_{p1} & . & . & . & a_{pq} \end{bmatrix} ; \textbf{Exemple:} \begin{bmatrix} 3 & 2 & 5 \\ 6 & \text{-1} & 1 \\ 9 & 3 & -4 \\ 5 & 0 & 0 \end{bmatrix}$$

Le nombre de ligne p et le nombre de colonne q constituent les dimensions ou l'ordre de la matrice.

Définition 2 : Une matrice qui ne comporte qu'une ligne ou une colonne est appelée vecteur (vecteur ligne ou vecteur colonne).

Enparticulier, une matrice qui ne possède qu'un seul élément, est considérée comme égale à cet élément $c_{11} = [c] = c$.

Définition 3 : Une matrice dont le nombre de lignes est égal au nombre $(p = q)$ de colonnes est dite carrée.

Définition 4 : Une matrice carrée est dite symétrique lorsque tous les éléments occupant de positions symétriques par rapport à la diagonale discendante sont égaux.

Définition 5 : Une matrice est dite triangulaire lorsque les éléments situés d'un même coté de cette diagonale sont nuls ($a_{ij} = 0$ pour tout $i > j$ ou $i < j$, $i \neq j$).

Définition 6 : Matrice identité ou unité si $a_{ii} = 1$, et $a_{ij} = 0$ si $i \neq j$.

Exmples:

$$I_{3x3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \quad A_{3x3} = \begin{bmatrix} 2 & 5 & 4 \\ 5 & 1 & 7 \\ 4 & 7 & 3 \end{bmatrix} \begin{array}{l} est \text{ une} \\ \text{matrice carrée} \\ \text{symétrique} \end{array};$$

$$B_{3x3} = \begin{bmatrix} 4 & 5 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{array}{l} \text{est une} \\ \text{matrice} \\ \text{triangulaire} \end{array}$$

Définition 7 : Transposée d'une matrice

La transopsée d'une matrice A_{pxq} est la mtrice A_{qp} obtenue en permutant les lignes et les colonnes, elle est telle que $a'_{ij} = a_{ji}$ $\forall i$, $\forall j$.

En particulier : La transposée d'une matrice carée symétrique est identique à elle même.

Tandis que le transposé d'un vecteur ligne est un vecteur colonne et vice-versa.

exemple: $A'_{3x3} = \begin{bmatrix} 2 & 3 & 5 \\ 1 & 1 & 4 \\ 3 & 6 & 2 \end{bmatrix}' = \begin{bmatrix} 2 & 1 & 3 \\ 3 & 1 & 6 \\ 5 & 4 & 2 \end{bmatrix};$

$$(2,6,4)' = \begin{pmatrix} 2 \\ 6 \\ 4 \end{pmatrix}; \begin{pmatrix} 5 \\ 4 \\ 8 \\ 0 \end{pmatrix}' = (5,4,8,0)$$

Définition 8 : Déterminant d'une matrice carrée.

A toute matrice carrée A correspond un nombre $|A|$ qui peut être obtenu comme suit:

$|a_{11}| = a_{11}$;

Pour une matrice d'ordre deux $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{11}$.

D'une façon générale, le déterminant d'une matrice carée quelconque peut être calculé grace au notion du mineur.

Définition 9: Le mineur de l'élément a_{ij} est le déterminent que l'on obtient en éliminant la i ème ligne et la j ième colonne.

Définition 10: Le cofacteur de a_{ij} est égal au mineur de a_{ij} multiplié par$(-1)^{i+j}$.

Définition 11: Le déterminant $|A|$ d'une matrice A peut être calculé en effectuant la somme desproduit des differents éléments d'une même ligne ou d'une même colonne par leurs cofacteurs respéctifs: $|A| = \sum_{i=1}^{p} a_{ij}A_{ij}$ pour tout j ou $\sum_{j=1}^{p} a_{ij}A_{ij}$ pour tout i.

Exemple:A= $\begin{bmatrix} 2 & 5 & 4 \\ 3 & 1 & -6 \\ -2 & 7 & 3 \end{bmatrix}$, $|A| = \begin{vmatrix} 2 & 5 & 4 \\ 3 & 1 & -6 \\ -2 & 7 & 3 \end{vmatrix} =$

$2 * \begin{vmatrix} 1 & -6 \\ 7 & 3 \end{vmatrix} - 5 * \begin{vmatrix} 3 & -6 \\ -2 & 3 \end{vmatrix} + 4 * \begin{vmatrix} 3 & 1 \\ -2 & 7 \end{vmatrix} =$

$\{2 * [(1 * 3) - (-6) * (7)]\} - \{5 * [3 * 3 - (-6) * (-2)]\} + \{4 * [3 * 7 - 1 * (-2)]\} =$

11

$2 - Les\ principales\ opérations$

Définition 1: Deux matrices de même dimensins sont dite égales ou équivalente lorsque les éléments de l'une sont tous égaux aux élément correspondants de l'autre $A = B$ si et seulement si $(a_{ij} = b_{ij}, \forall i, \ \forall j)$

Définition 2: Addition et soustraction des deux matrices

La somme de deux matrices de même dimension est obtenue en additionnant deus à deux les éléments correspondants des deux matrices: $A + B = [a_{ij} + b_{ij}]$

De même la différence de deux matrices de même dimensions est obtenue en soustrayant l'un de l'autre les éléments correspondants des deux matrices.

Définition 3: Multiplication d'une matrice par un nombre et par une matrice.

a- Le produit d'une matrice par un nombre est obtenu en multipliant chacun des éléments de la matrice par ce nombre $c\,[a_{ij}] = [ca_{ij}]$.

Exemple: $5 * \begin{bmatrix} 2 & 5 & 4 \\ 3 & 1 & -6 \\ -2 & 7 & 3 \end{bmatrix} = \begin{bmatrix} 10 & 25 & 20 \\ 15 & 5 & -30 \\ -10 & 35 & 15 \end{bmatrix}$

Quant au produit de deux matrice A et B n'est défini que si le nombre de colonnes de la première est égal au nombre de ligne de la deuxième.

Dans ce cas si les dimensions des deux matrices sont respectivement pxq et qxn le produit C est la matrice de dimension pxn.$C_{pxn} = A_{pxq} * B_{qxn}$; $c_{ik} = \sum_{j=1}^{q} a_{ij} b_{jk} \quad (\forall j, \quad \forall k)$

En général: $A_{pxp} * B_{pxp} \neq B_{pxp} * A_{pxp}$

Exemple: produit de deux matrices $\begin{bmatrix} 2 & 5 & 4 \\ 3 & 1 & -6 \\ -2 & 7 & 3 \end{bmatrix} * \begin{bmatrix} 2 & 5 \\ 3 & 1 \\ -2 & 7 \end{bmatrix} =$

$\begin{bmatrix} 4+15-8 & 10+5+28 \\ 6+3+12 & 15+1-42 \\ -4+21-6 & -10+7+21 \end{bmatrix} = \begin{bmatrix} 11 & 43 \\ 21 & -26 \\ 11 & 18 \end{bmatrix}$

Définition 4: Produit de deux vecteurs

Par application de la règle enoncée ci-dessus , en particulier $A_{1xp} * B_{px1} = C_{1x1}$

Exemple: produit de deux vecteurs $(1,3,5,2) * \begin{pmatrix} -3 \\ 0 \\ 1 \\ -2 \end{pmatrix} = 1*(-3)+3*0+5*1+2*(-2) = 2$

Définition 5 : Inverse d'une matrice

La matrice inverse d'une matrice carrée A non singulière est telle que le résultat de sa multiplication par la matrice initiale est une matrice unité ou identité.

A^{-1} est la matrice inverse de la matrice A si est seulement si $A^{-1}A = I$

$A^{-1} = \dfrac{adj(A)}{|A|}$, où $adj(A)$ est la matrice de cofacteurs

Trouver l'inverse de la matrice $\begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix}$, si $\begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} x & y \\ z & t \end{bmatrix}$ alors $\begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix} * \begin{bmatrix} x & y \\ z & t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2x+z & 2y+t \\ 4x+z & 2y+2t \end{bmatrix}$ doù lesystème d'équations suivant :

2x+z=1; 2y+t=0; 4x+z=0; 2y+2t=1

la solution de ce sytème d'équations donne: x=? y=?, z=? et t=?.

12

II – REGRESSION LINEAIRE MULTIPLE

12.1 1-Présentation

La régression linéaire multiple est une généralisation, à p variables explicatives, de la régression linéaire simple.

Nous sommes toujours dans le cadre de la régression mathématique : nous cherchons à prédire, avec le plus de précision possible, les valeurs prises par une variable y, dite endogène, à partir d'une série de variables explicatives $x_1, x_2, \ldots, x_p$.

Dans le cas de la régression linéaire multiple, la variable endogène et les variables exogènes sont toutes quantitatives (continues) ; et le modèle de prédiction est linéaire.

12.2 2- Equation de régression et objectifs

Nous disposons de n observations (i = 1,..., n). L'équation de régression s'écrit

$y_i = a_0 + a_1 \times x_{i,1} + \ldots + a_p \times x_{i,p} + \blacksquare_i$

où $\blacksquare_i$ est l'erreur du modèle, elle exprime, ou résume, l'information manquante dans l'explication linéaire des valeurs de y à partir des x_j (problème de spécifications, variables non prises en compte, etc.) ; $a_0, a_1, \ldots, a_p$ sont les coefficients (paramètres) du modèle à estimer.

La problématique reste la même que pour la régression simple :

1-estimer les paramètres a_j en exploitant les observations.

2-évaluer la précision de ces estimateurs ;

3-mesurer le pouvoir explicatif du modèle. ;

4-évaluer l'influence des variables dans le modèle :

5-globalement (les p variables en bloc) et, individuellement (chaque variable) ;

6-évaluer la qualité du modèle lors de la prédiction (intervalle de prédiction) ;

7-détecter les observations qui peuvent influencer exagérément les résultats (points atypiques).

12.3 3-Notation matricielle

Nous pouvons adopter une écriture condensée qui rend la lecture et la manipulation de l'ensemble plus facile. Les équations suivantes: $\begin{cases} y_1 = a_0 + a_1 \times x_{1,1} + \ldots + a_p \times x_{1,p} \\ y_2 = a_0 + a_1 \times x_{2,1} + \ldots + a_p \times x_{2,p} \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = a_0 + a_1 \times x_{n,1} + \ldots + a_p \times x_{n,p} \end{cases}$

peuvent être résumées avec la notation matricielle $\mathrm{Y} = \mathrm{Xa} + \blacksquare$

avec

Y est de dimension $(n, 1)$;

X est de dimension $(n,\ p+1)$;

a est de dimension $(p+1, 1)$;

et $\blacksquare >$: $(n, 1)$;

la matrice X est égale à $\begin{pmatrix} 1 & \mathrm{x}_{1,1} & . & . & . & \mathrm{x}_{1,p} \\ . & . & & . & & . \\ . & . & & & . & . \\ . & . & & & . & . \\ 1 & \mathrm{x}_{n,1} & . & . & . & \mathrm{x}_{np} \end{pmatrix}$

la première colonne sert à indiquer que nous procédons à une régression avec constante.

12.4 4- Hypothèses

Les hypothèses permettent de déterminer : les propriétés des estimateurs (biais, convergence) ; et leurs lois de distributions (pour les estimations par intervalle et les tests d'hypothèses).

Il existe principalement deux catégories d'hypothèses :

Hypothèses stochastiques

H1 : les X_j sont non aléatoires, j = 1, ..., p ;

H2 : $E(\blacksquare_i) = 0$, le modèle est bien spécifié en moyenne ;

H3 : $E(\blacksquare_i^2) = \blacksquare_{\blacksquare^2}$, homoscédasticité ;

H4 : $COV(\blacksquare_i, \blacksquare_j)$ pour $i \neq j$, absence d'autocorrélation des erreurs ;

H5 : $COV(X_{i,j}, \blacksquare_i) = 0$, les erreurs sont indépendantes des variables exogènes ;

12.5 5-Hypothèses structurelles

H7 : absence de colinéarité entre les variables explicatives, i.e. $X\prime X$ est régulière, $\det(X\prime X) \neq 0$ et $(X\prime X)^{-1}$ existe (remarque : c'est la même chose, $\text{rang}(X) = \text{rang}(X\prime X) = p + 1$) ;

H8 : $\dfrac{X\prime X}{n}$ tend vers une matrice finie non singulière lorsque $n \to +\infty$;

H9 : n > p + 1, le nombre d'observations est supérieur au nombre de variables + 1. Notons que s'il y avait égalité, le nombre d'équations serait égal au nombre d'inconnues a_j, la droite de régression passe par tous les points, nous sommes face à un problème d'interpolation linéaire

13 6- La méthode des moindres carrés ordinaires

13.1 6-1- Estimateur des moindres carrés ordinaires (EMCO)

Le principe des moindres carrés consiste à rechercher les valeurs des paramètres qui minimisent la somme des carrés des erreurs, à savoir $S = \sum_{i=1}^{n} \blacksquare_i^2$

En adoptant l'écriture matricielle, nous minimisons donc $S = \blacksquare\prime\blacksquare$.

Ce qui revient à rechercher les solutions de . Nous disposons de p + 1 équations, dites équations normales, à résoudre.

La solution obtenue est l'estimateur des moindres carrés ordinaires, il s'écrit : $\hat{a} = (X\prime X)^{-1} X\prime Y$

où $X\prime$ est la transposée de X ;

si les x_j sont centrés, $X\prime X$ correspond à la matrice de variance covariance des exogènes ; s'ils sont centrés et réduits, $X\prime X$ correspond à la matrice de corrélation.

13.2 6-2-Propriétés des estimateurs

Si les hypothèses initiales sont respectées, cet estimateur des MCO (Moindres Carrés Ordinaires) possède d'excellentes propriétés :

il est sans biais, c.-à-d. $E(\hat{a}) = a$;

il est convergent, c.-à-d. la variance des estimateurs tend vers zéro lorsque le nombre des observations n tend vers l'infini ;

on peut même prouver que l'EMCO est le meilleur estimateur linéaire sans biais c-à.-d. il n'existe pas d'estimateur sans biais de a qui ait une variance plus petite.

13.3 6-3-Évaluation

13.3.1 a- Écart-type de l'erreur et matrice de variance covariance des coefficients

Pour réaliser les estimations par intervalle et les tests d'hypothèses, la démarche est presque toujours la même en statistique paramétrique :

définir l'estimateur (â dans notre cas) ;

calculer son espérance mathématique (ici $E(\hat{a}\,) = a$) ;

calculer sa variance (ou sa matrice de variance co-variance) et produire son estimation ;

et enfin déterminer sa loi de distribution (en général et sous l'hypothèse nulle des tests).

Matrice de variance co-variance de $\hat{a}$

La matrice de variance co-variance est définie par : $\Omega_{\hat{a}} = E[(\hat{a}-a)(\hat{a}-a)\prime]$

En développant cette expression, nous obtenons : $\Omega_{\hat{a}} = \blacksquare_{\blacksquare}{}^{2} \cdot (X\prime X)^{-1}$

13.3.2 b- Evaluation globale de la régression — Tableau d'analyse de variance et coefficient de détermination

L'évaluation globale de la pertinence du modèle de prédiction s'appuie sur l'équation d'analyse de variance $SCT = SCE + SCR$, où

SCT, somme des carrés totaux, traduit la variabilité totale de l'endogène ;

SCE, somme des carrés expliqués, traduit la variabilité expliquée par le modèle ;

SCR, somme des carrés résiduels correspond à la variabilité non-expliquée par le modèle.

Toutes ces informations sont résumés dans un tableau, le tableau d'analyse de variance.

Source de var	Somme des carrés	D de liberté	Carrés moyens
Expliquée	$SCE = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$	p	$CME = \frac{SCE}{p}$
Résiduelle	$SCR = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$	$n-p-1$	$CMR = \frac{SCR}{n-p-1}$
Totale	$SCT = \sum_{i=1}^{n} (y_i - \overline{y})^2$	$n-1$	

Dans le meilleur des cas, $SCR = 0$, le modèle arrive à prédire exactement toutes les valeurs de y à partir des valeurs des x_j. Dans le pire des cas, $SCE = 0$, le meilleur prédicteur de y est sa moyenne .

Un indicateur spécifique permet de traduire la variance expliquée par le modèle, il s'agit du coefficient de détermination. Sa formule est la suivante :$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$. ; $R = \sqrt{R^2}$ est le coefficent de corrélation multiple.

Dans une régression avec constante, nous avons forcément $0 \leq R^2 \leq 1$.

Enfin, si le $\mathbf{R^2}$ est certes un indicateur pertinent, il présente un défaut parfois ennuyeux, il a tendance à mécaniquement augmenter à mesure que l'on ajoute des variables dans le modèle. De fait, il est inopérant si l'on veut comparer des modèle comportant un nombre différent de variables. Il est conseillé dans ce cas d'utiliser le coefficient de détermination ajusté qui est corrigé des degrés de libertés :

$$1 - \frac{SCR/(n-p-1)}{SCT/(n-1)} = 1 - \frac{(n-p-1)}{(n-1)} \left(1 - R^2\right)$$

13.3.3 c-Significativité globale du modèle

Le R^2 est un indicateur simple, on comprend aisément que plus il s'approche de la valeur 1, plus le modèle est intéressant. En revanche, il ne permet pas de savoir si le modèle est statistiquement pertinent pour expliquer les valeurs de y.

La formulation du test d'hypothèse qui permet d'évaluer globalement le modèle est le suivant :

$H0 : a_1 = a_2 = \ldots = a_p = 0$;

$H1$: un des coefficients au moins est non nul.

La statistique dédiée à ce test s'appuie (parmi les différentes formulations possibles) sur le R^2, il s'écrit: $F_{calc} = \frac{R^2/p}{1 - R^2/(n-p-1)}$ et suit une loi de Fisher à $(p, n-p-1)$ degrés de liberté.

La région critique du test est donc : rejet de $H0$ si et seulement si $F_{calc} > F_{1-\blacksquare}(p, n-p-1)$, où $\blacksquare$ est le risque de première espèce.

Une autre manière de lire le test est de comparer la p-value (probabilité critique du test) avec $\blacksquare$: si elle est inférieure, l'hypothèse nulle est rejetée.

13.4

TP1 (pollution atmosphirique)

Objectif: Nous somme toujours dans le cadre de la régression mathématique, on cherche à analyser y (le nombre de morts) en fonction de deux varibles atmosphiriques.

Les modèles proposés:1- modèle de régression linéaire multiple. **2-** modèle plynomial

Pendant la première quinzaine du mois de décembre 1952, il y eut dans la région de Londres une période de brouillard très intense constituant un record, et on remarque une mortalité accrue pendant cette période. On note pendant ces quinze jours le nombre de morts y et la teneur atmosphérique moyenne en fumée x_1, mesurée en mg par mètre cube, et en dioxyde de soufre x_2, mesurée par nombre de particules par million. La variable y est la variable expliquée et la pollution atmosphirique $x = (x_{1,}x_2)$ la variable explicative, ici à deux dimensions.

Date	1	2	3	4	5	6	7	8	9	10
Nombre de morts y	112	140	143	120	196	294	513	518	430	274
Fumée (x_1)	0.30	0.49	0.61	0.61	0.49	2.64	3.45	4.46	1.22	1.22
SO_2 (x_2)	0.9	0.16	0.22	0.14	0.75	0.86	1.34	1.34	0.47	0.47

LANGAGE R ET REGRESSION MULTIPLE

vecteurs des données
>#y=nombre de morts
>#x1=teneur de la fumée (mg/m3)
>#x2=dioxyde de soufre(nombre de particules par million)
>y<-c(112,140,...,213)
>x1<-c(0.30,0.49,...,0.32)
>x2<-c(0.9,0.16,...,0.16)
>id<-(1,1,...,1)
#matrice combinant par colonnes les éléments (id,x1,x2)
>x<-cbind(id,x1,x2)
>x
#arrondi les éléments de x avec 2 chiffres après la virgule
>round(x,2)
>x<-round(x,2)

>x
>b<-solve(t(x)%*%x)%*%t(x)%*%y
>b
>summary(x)
>cor(x)
>cov(x)

LANGAGE R ET MODELE POLYNOMIALE

>x2<-x1*x1 ou >x1<-x2*x2

Utiliser le même programme (régression linéaire multiple).

13.5

TP2 de biostatistique
Régression multiple

1-Objectif: Etude de la relation entre le rendement d'une culture de blé et certains facteurs météorologiques.

Nous nous limiterons à l'examen de cinq variables (p=5), à savoire:

x1=précipitation des mois de novembre et décembre (mm)

x2=tepérature moyenne du mois de juillet (degrès centigrades)

x3=précipitation du mois de juillet (mm)

x4=radiation du mois de juillet (ml d'alcool mesurés à l'actinomètre de BELLANI)

y=rendement moyen des cultures (quintaux par hectare)

2-Les modèles proposés:

a- modèle de régression linéaire multiple: $y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4$

b-modèle plynomial : $y = b_0 + b_1x_i + b_2x_i^2 + b_3x_i^3 + b_4x_i^4$

c-l'ajustement d'une équation quadratique à deux variables $y = c_0 + c_1x_i + c_2x_i^2 + c_3x_j + c_4x_j^2 + c_5x_ix_j$

3-Le choix des variables: Principe général

D'une façon générale, le but poursuivi lors du choix des variables explicatives et d'assurer une précision maximum de l'équation de régression, c'est-à-dire une variance résiduelle minimum. Cet objectif conduit à donner la préférence, apriori,à des variables éxplicatives fortement corrélées avec la variable dépendante et faiblement corrélées entre elles.

4-Données: Le tableau suivant donne les rendements observés au cours de 11 années successives et les données metéorologiques correspondantes.

année	y	x1	x2	x3	x4
1995	28,37	87,9	19,6	81,0	1661
1996	23,77	89,9	15,2	90,1	0968
1997	26,04	153,0	19,7	56,6	1353
1998	25,74	132,1	17,0	91,0	1293
1999	26,68	88,8	18,3	93,7	1153
2000	24,29	220,9	17,8	106,9	1286
2001	28,00	117,7	17,8	65,5	1104
2002	28,37	109,0	18,3	41,8	1574
2003	24,96	156,1	17,8	57,4	1222
2004	21,66	181,5	16,8	140,6	0902
2005	24,37	181,4	17,0	74,3	1150

14

EXERCICES

Exercice 1- On a mesuré les hauteurs en centimètres de plants de blé d'un même champ dont on a tiré un echantillon chaque semaine; Les observations sont données dans le tableau suivant:

âge en semaines x_i	1	2	3	4	5	6	7	8
hauteur en cm y_i	5	12	15	22	34	35	41	48

Les variables observées simultanément sont donc:

X: âge en semaines et Y: hauteur en centimètres.

Cette série peut être representée par

le graphique ci- contre:

Exercice 2- *On a étudié l'hydrolyse enzymatique du saccharose.*

L'invertase catalyse l'hydrolyse du saccharose suivant la réaction:

$saccharose + eau \leftrightarrows\leftrightarrows glucose + fructose$

En théorie la vitesse initiale V de la réaction pour une concentration initiale S en saccharose vérifie: $\frac{1}{V_I} = \frac{K_M}{V_{\max}} * \frac{1}{S_i} + \frac{1}{V_{\max}}$

Une série de cinétiques enzymatiques, réalisée avec des condition physico-chimiques

identiques (PH,température,...) mais des concentrations initiales en saccharose différentes, a donnée les résultats suivants:

$S_i(mmol.dm^{-3})$	2	4	5	8	10
$V_i(UI.dm^{-3})$	50	89	114	155	200

.

On pose $X_I = \frac{1}{S_I}$ *et* $Y_i = \frac{1}{V_I}$.

1) *Calculer à* 10^{-4} *près les valeurs de* X *et de* Y *pour les valeurs de* S_i *et* V_i.

2) *Calculer le coefficient de correlation linéaire entre* X *et* Y *.Qu'en déduit-on?.*

3) *Déterminer par laméthode des moindres carrés la droite d'ajustement linéaire de* Y *par rapport à* X .

4) *Estimer la vitesse initiale de la réaction pour une concentration nitiale en saccharose de 16* $mmol.dm^{-3}$ *et déter* $\min$ *er les valeurs des paramètres* $V_{\max}$ *et* K_M .

$$Rép : \frac{1}{V_i} \approx 0.037\frac{1}{S_i} + 0.0016;\ V_{\max} \approx 626;\ K_M \approx 23.125$$

Exercice 3- (Cinétique chimique)
Dimérisation du butadiène
la bitadiène se démirise selon le schéma suivant:

$$2C_4H_6(gaz) - 326°C - - \ C_8H_{12}\,(gaz)\,.$$

La préssion initiale dans le réacteur contenant le réactif est de 632 mmHg et sa variation dans le temps est donnée par le tableau suivant :

temps $t_i(mn)$	0	10	20	50	75	100
préssion P_{t_i}	632	591	557	497	466	446

Donner une représentation graphique du processus chimique de cette réaction dan un repère orthogonale.

2°) On pose $P^*_{t_i} = lnP_{t_i}$.

a) Calculer les valeurs $P^*_{t_i}$ pour $i = 1; ...; 6$.

b) Déterminer une fonctionnelle linéaire entre les deux variables t et lnP.

c) Calculer l'intensité de liaison linéaire entre les deux variables.

d) Déduire de la question b l'expression de la pression en fonction du temps.

d) Tracer le graphique de la fonction $P = f(t)$

e) Calculer la pression dans le réacteur après trois heurs.

Indications:$\overline{T} = 57.287; Var(T) = 520.57; \overline{P^*_{t_i}} = 4.257;$
$Var(P^*_{t_i}) = 3.28; Cov(T, P^*_{t_i}) = 35.12$

Exercice 4- Isolation thermique:

Le mur d'une habitation est constitué par une couche de béton et une couche de polystyrène d'épaisseur variable x (encm), on a mesuré les résistances thermiques R de ce mur pour diverses valeurs de x,et on a obtenu les résultats suivants:

x	*2*	*4*	*6*	*8*	*10*	*12*	*15*	*20*
R	*0.83*	*1.34*	*1.63*	*2.29*	*2.44*	*2.93*	*4.06*	*4.48*

On demande d'arrondir les résultats au centième le plus proche.

1)Métode de Mayer:

a)Calculer les cordonnées du point moyen (G_1) associé associé au points du nuage ayant les quatres plus petites abscisses et les cordonnées du point moyen (G_2)associé au quatres autres points.

b)On choisit la droite (G_1G_2) comme droite d'ajustement. En déterminer une équation.

c)Quelle résistance thermique peut-on espérer obtenir avec une épaisseur depolystérene de 25 cm?.; La résistance est exprimée en m.dC/W.

2) Méthodedes moindres carrés.

a) Clauculer le coefficient de correlation linéaire entre x et R. Conclusion?.

b) Déterminer par laméthode des moindres carrés la droite de regression de R en x.

c) Quelle résistance thermique peut-on espérer obtenir avec une épaisseur de polystérene de 25 cm?.

R:droite de Mayer:R=0.21x+0.48

droite des moindres carrés:R=0.21x+0.45

Exercice 5: Test à l'effort:

Les relvés de l'intensité du travail fourni x exprimée en Kilojoules par miniute et de la fréquence quardiaque y exprimée en nombre de battement par minute d'une personne pendant un test à l'efforty sont donnés par le tableau suivant:

x_i	9.6	12.8	18.4	31.2	36.8	47.2	49.6	56.8
y_i	70	86	90	104	120	128	144	154

1°) Calculer le coefficient de corrélation de cette série statistique.

2°) déterminer par leurs équations les droites d'justement linéaire de X en Y et de Y en X.(On arrondira les résultas au centième le plus proche).

3°) Lorsque l'intensité du travail fourni est de $65KJ/min$*, estimer la fréquence quardiaque.*

4°)Lorsue la fréquence quardiaque est de 100 battement par minute, esimer

l'intensité du travail fourni.

Rép : $r \approx 0.985$; $y = 1.63x + 58.57; x = 0.6y - 33.93$; *165 battement /m; 26* KJ/min.

Exercice 6: Ajustement exponentiel :

Atmosphère, atmosphère...:Le nombre de moles par unité de volume, appelé concentration molaire, varie en fonction de l'altitude.

On souhaite vérifier expérimentalement la formule donnant la concentration molaire c en fonction de l'altitude z. Pour cela on mesure cette concentration pour divèrses altitudes au moyen de ballons- sondes. Ls résultats obtenus sont consignés dans le tableau suivant où l'altitude z_i *est éxprimée en mèttre et la concentration* c_i *en* $mol.l^{-1}$

z_i	0	1000	2000	5000	10000	15000	20000
c_i	0.042	0.037	0.033	0.023	0.012	0.0065	0.0035

1°) On pose $y_i = ln10^4.c_i$.

Dreseer un tableau précisant les valeurs de y_i *et celles des* z_i *correspondantes. on donnera les valeurs de y sous forme décimale arrondie au centième le plus proche.*

2°) Calculer le coefficient de corrélation entre Y *et* Z *. Un ajustement affine est-il justifié ?.*

3°) Déterminer une équation de la droite d'ajustement de Y *en* Z *par la méthode des moindres carrés. On écrit cette équation sous la forme* $Y = aZ + b$*, où* a *est éxprimé sous forme décimale arrondie à* 10^{-6} *et* b *à* 10^{-3} *près.*

4°) Déduire du résultat précédent, une expréssion de la concentration molaire $c(z)$ *en fonction de l'altitude* z *sous la forme* $c(z) = k\varrho^{\beta z}$ *où* k *et* β *sont expéimés en fonction de* a *et* b*. Donner ensuite une valeur approchée sous forme décimale arrondie à* 10^{-3} *près de* K *et à* 10^{-6} *près de* β*. Estimer la concentration molaire à l'altitude* $1500m$.

Réponse $::r(Z,Y) = \approx 0.999$; $Y = aZ + b = -1.24 * 10^{-4}Z + 6.044$; $C = K\varrho^{\beta z} = \left[10^{-4}\varrho^{b}\right]\varrho^{aZ}$; $K \approx 0.042 à 10^{-3}$ *près et* $\beta = -1.24 * 10^{-4}$*près.* $4°) C \approx 0.0349 mol.l^{-1}$. $C(Z) = A * B^Z$ *avec* : $A = 10^{-4}\varrho^b$ *et* $B = \varrho^\beta$*, soit* $A = 0.042$ *à* 10^{-3} *près et* $B = 0.999876$ *à* 10^{-6} *près.*

14.1 Données

Tableau 01: données sur le rendement photosynthétique en $\mu molm^{-2}s^{-1}$(y) et mesure de la radiation en $molm^{-2}s^{-1}$,

observées sur une éssence forestiére

radiation ($molm^{-2}s^{-1}$)	0.7619	0,7684	0,7961	0,8380	0,8381	0,8435	0,8599
rendement photosynthétique ($\mu molm^{-2}s^{-1}$)	7,58	9,76	10,76	11,51	11,68	12,68	12,76

0,9209	0,9993	1,0041	1.0089	1,0137	1,0184	1,0232	1,0280
13,73	13,89	13,97	14,05	14,13	14,20	14,28	14,36

14.2 Analyse Descriptive des données:

14.3 Modélisation: Modèle Linéair

14.3.1 Modéle exponentiel

14.3.2 Modéle Démographique

le modéle démographique. qui explique le rendement photosynthétique en fonction de la radiation par la fonction

$$f(radiation) = \frac{1}{rendementPH} = b + a\frac{1}{radiation}$$

14.3.3 Tests statistiques

15 APPLICATION 2:

plusieurs entreprises d'agriculture utilisent la pulvérisation pour but de distribuer uniformement

une fongicide, insecticide, miticide, ou régulateur de croissance sur toutes les parties de l'arbre.

Le tableau ci-dessous énumère les gallons de pulvirisation diluée par acre nécessaire pour fournir

une couverture pour arbres adultes de différentes tailles et des espacements.

15.1 Données:

tableau 02: les observation de la pulvérisation par acre nécéssaire pour fornir une couvérture

pour des arbre adultesde différentes tailles et d'éspacement.

Distance entre les lignes (*pied*)	la hauteur des arbres (*peid*)	la largeur des arbres (*pied*)	volume maximum d'arbre/acre ($1.000 piedcubes$)	pulvérisation maximum ($gallons/acre$)
30	20	15	436	300
26	16	12	354	225
24	14	10	254	180
22	14	10	272	200
20	12	10	261	185
18	10	10	242	175
16	8	8	174	125
14	6	8	149	105
12	6	6	131	90

15.2 Données:

tableau 03: Valeurs du pH et de la teneur en carbone organique obsevées dans des

échantillons de terrain prélevés dans des forêts naturelles.

pH	5.7	6.1	5.2	5.7	5.36	5.1	5.8
Carbone organique (C) (%)	2.10	2.17	1.97	1.39	2.26	1.29	1.17
5.5	5.4	5.9	5.3	5.4	5.1	5.1	5.2
1.14	2.09	1.01	0.89	1.06	0.90	1.01	1.21