Table des matières

Chapitre 1

STATISTIQUE

1.1 GÉNÉRALITÉS

1.1.1 La statistique

La statistique est une science qui a pour objet de collecter, classer, analyser et présenter de façon compréhensible un ensemble de données. Ces données peuvent prévenir de plusieurs domaines comme : la médecine, l'éducation, l'économie,...

1.1.2 Les statistiques

Les statistiques sont les résultats produit par la statistique entant que science.

1.1.3 Collecte des données

Elle se fait au moyen de deux approches : le recensement et le sondage.

a) Le recensement

Il consiste à interroger chaque individu de la population cible.

b) Le sondage

Il consiste à interroger une partie de la population qu'on appelle échantillon.

1.2 Série statistique à un seul caractère

1.2.1 Définition

On appelle série statistique à un seul caractère ou à une variable X, l'ensemble des couples (x_i, n_i) souvent données sous la forme suivante :

x_i	x_1	x_2	x_3	x_4	x_5	....	x_k
n_i	n_1	n_2	n_3	n_4	n_5	...	n_k

1.2.2 Effectif relatif

On appelle effectif relatif noté n_i, le nombre d'individus d'une population présentant la modalité x_i du caractère X.

1.2.3 Effectif total

L'effectif total noté N est la des effectifs relatifs.
On a : $N = \sum_{k=1}^{p} n_i = n_1 + n_2 + n_3 + ... + n_p$.

1.2.4 Fréquence relative

On appelle fréquence relative, le quotient entre l'effectif relatif et l'effectif total.
On a : $f_i = \frac{n_i}{N} \times 100$.
N.B : La somme de toutes les fréquences d'une série statistique est égale à 1.

1.2.5 Moyenne arithmétique

On appelle moyenne arithmétique le nombre réel noté $\overline{x}$ défini par : $\overline{x} = \frac{1}{N} \sum_{i=1}^{p} n_i x_i$.

1.2.6 Variance

On appelle variance le nombre réel positif noté $V(x)$ défini par :
$V(x) = \frac{1}{N} \sum_{i=1}^{p} n_i x_i^2 - \overline{x}^2$.

1.2.7 Écart-type

C'est la racine carré de la variance.
On a : $\Gamma(x) = \sqrt{V(x)}$.

Exercice

Soit la série statistique suivante :

x_i	59	62	65	68	71	74	77
n_i	1	4	6	7	5	5	2

1. Donner l'effectif total de cette série.
2. Calculer la moyenne, la variance et l'écart-type de cette série.

1.3 Série statistique à deux variables

1.3.1 Définition

On appelle série statistique double ou à deux variables (X, Y), l'ensemble des triplets $(x_i; y_j; n_{ij})$. On parle aussi de distribution double, avec n_{ij} l'effectif du couple $(x_i; y_j)$.
On distingue deux types de séries statistiques à double caractères :
▷ série statistique double linéaire ;
▷ série statistique à double entrée où pondérée.

1.3.2 Série statistique double linéaire

Définition

Une série statistique double est dite linéaire lorsque une seule valeur du caractère x correspond à une seule valeur du caractère y et inversement.

Elle se présente sous la forme suivante :

x_i	x_1	x_2	x_3	...	x_p
y_i	y_1	y_2	y_3	...	y_p

N.B : Dans cette série les effectifs partiels sont partout identiques et égaux à 1.

1.3.3 Série statistique à double entrée

Définition

Une série statistique est dite à double entrée lorsque une valeur du caractère X correspond à plusieurs valeurs du caractère Y et inversement.
Elle se présente sous la forme suivante :

y_j \ x_i	x_1	x_2	...	x_p
y_1	n_{11}	n_{21}	...	n_{p1}
y_2	n_{12}	n_{22}	...	n_{p2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_j	n_{1q}	n_{2q}	...	n_{pq}

1.3.4 Nuage des points associés à une série double

Soit $(x_i; y_j; n_{ij})$ une série statistique à deux caractères quantitatifs.
On appelle nuage des points l'ensemble des points $M_{ij}(x_i, y_i)$.
N.B : Dans le cas d'une série double linéaire, le nuage des points est l'ensemble des points $M_i(x_i, y_i)$.

1.3.5 Représentation graphique du nuage des points

Le nuage des points est représenté de deux manières :

a) Représentation par points pondérés

On indique à côté de chaque point M_{ij} l'effectif n_{ij}.

b) Représentation par tâches

Chaque point M_{ij} est remplacé par un disque dont l'aire est proportionnelle à n_{ij}.

Exemple

Soit la série double suivante :

y_j \ x_i	0	2	3	4
1	1	2	3	5
3	5	2	8	6

1. Déterminer le nuage des points de cette série statistique.
2. Représenter le nuage des points dans le plan muni du repère orthonormé $(O; \vec{i}, \vec{j})$.

1.3.6 Lois marginales ou distributions marginales

On appelle loi marginale de X ou de Y, la loi de X ou Y extraite d'un tableau statistique à double entrée.

a) Loi marginale dans le cas d'un tableau à double entrée

Exemple

Soit la série double suivante :

Y \ X	0	2	3	4
1	1	2	3	5
3	5	2	8	6

Donner les lois marginales de X et Y.

Solution

Donnons les lois marginales de X et Y

▷ Pour X

x_i	0	2	3	4	$\sum$
n_i	6	4	11	11	$N = 32$

▷ Pour Y

y_j	1	3	$\sum$
n_j	11	21	$N = 32$

b) Loi marginale dans le cas d'un tableau linéaire

Exercice

Soit la série statistique suivante :

X	59	62	65	68	71	74	77
Y	12	43	46	54	55	60	62

Donner les lois marginales de X et Y.

Solution

Donnons les lois marginales de X et Y

▷ Pour X

x_i	59	62	65	68	71	74	77	$\sum$
n_i	1	1	1	1	1	1	1	$N = 7$

$\triangleright$ Pour Y

y_i	12	43	46	54	55	60	62	$\sum$
n_i	1	1	1	1	1	1	1	$N = 7$

1.3.7 Point moyen

Définitions

$\triangleright$ Cas d'une série statistique à double entrée

Soit $(x_i; y_j; n_{ij})$ une série statistique à deux caractères quantitatifs.

On appelle point moyen du nuage le point $G(\overline{x}, \overline{y})$ barycentre de tous les points du nuage de cette série, avec $\overline{x} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{p} n_i x_i$ et $\overline{y} = \dfrac{1}{N}\displaystyle\sum_{j=1}^{q} n_j x_j$.

$\triangleright$ Cas d'une série statistique double linéaire

Soit $(x_i; y_j)$ une série statistique à deux caractères quantitatifs.

On appelle point moyen du nuage le point $G(\overline{x}, \overline{y})$ barycentre de tous les points du nuage de cette série, avec $\overline{x} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{p} x_i$ et $\overline{y} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{q} y_i$

1.3.8 Transformation des tableaux statistiques doubles

a) Passage d'un tableau à double entrée à un tableau linéaire

Exercice

Soit la série double suivante :

y_j \ x_i	0	2	3	4
1	1	2	3	0
3	0	2	1	1

1. Donner l'effectif total de cette série statistique.
2. Transformer ce tableau à un tableau linéaire.

Solution

1. Donnons l'effectif total de cette série statistique.

   $N = 1 + 2 + 3 + 0 + 0 + 2 + 1 + 1 = 10$

2. Transformons ce tableau à un tableau linéaire.

   $(0, 1) \longrightarrow 1$, $(0, 3) \longrightarrow 0$, $(2, 1) \longrightarrow 2$, $(2, 3) \longrightarrow 2$, $(3, 1) \longrightarrow 3$, $(3, 3) \longrightarrow 1$, $(4, 1) \longrightarrow 0$ et $(4, 3) \longrightarrow 1$.

   On a le tableau suivant :

X	0	2	2	2	2	3	3	3	3	4
Y	1	1	1	3	3	1	1	1	3	3

b) Passage d'un tableau linéaire à un tableau à double entrée

Exercice

Soit la série double suivante :

X	0	2	1	O	2	3	3	3
Y	1	1	0	1	1	1	2	1

1. Donner l'effectif total de cette série statistique.
2. Transformer ce tableau à un tableau à double entrée.

Solution

1. Donnons l'effectif total de cette série statistique.

   $N = 8$

2. Transformons ce tableau à un tableau double entrée.

Y \ X	0	1	2	3
0	0	1	0	0
1	2	0	2	2
2	0	0	0	1

1.4 Inerties du nuage des points

1.4.1 Inertie minimale

Définition

On appelle inertie minimale, l'inertie par rapport au point moyen $G(\overline{x}, \overline{y})$.
On le note par I_G et est définie par : $I_G = N\,[V(x) + V(y)]$.

1.4.2 Inertie par rapport à un point

Définition

Soit $A(a,b)$ un point du plan. On appelle inertie du nuage par rapport au point A, le nombre réel positif noté I_A défini par :
$I_A = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\|\overrightarrow{BM_{ij}}\|^2$ ou $I_A = I_G + NAG^2$ (Théorème de Huyguens).

1.4.3 Propriétés

▷ L'inertie du nuage par rapport au point A est dite minimale, si et seulement si $A = G$.
▷ L'inertie du nuage par rapport au point A est dite maximale, si et seulement si $A \neq G$.

1.5 Ajustement linéaire : Méthode de moindres carrées

1.5.1 Covariance

Soit $(x_i; y_j; n_{ij})$ une série statistique à deux caractères x et y d'effectif total N.
On appelle covariance du couple (x,y) le nombre réel noté $cov(x,y)$ tel que :
$cov(x,y) = \dfrac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$.

Formule de Koenig

On a : $cov(x,y) = \dfrac{1}{N}\sum_{i,j=1}^{N} n_{ij}x_iy_j - \overline{x}\times\overline{y}$.

1.5.2 Droite de régression de y en x

La droite de régression de y en x est la droite (D) passant par le point moyen G associé à la série statistique double (x_i, y_j, n_{ij}) et dont le coefficient directeur est $a = \dfrac{cov(x,y)}{V(x)}$.
Une équation cartésienne de cette droite est : $(D) : y - \overline{y} = \dfrac{cov(x,y)}{V(x)}(x - \overline{x})$.
Cette droite peut s'écrire sous la forme $(D) : y = ax + b$ avec $a = \dfrac{cov(x,y)}{V(x)}$ et $b = \overline{y} - a\overline{x}$.

1.5.3 Droite de régression de x en y

La droite de régression de x en y est la droite (D') passant par le point moyen G associé à la série statistique double (x_i, y_j, n_{ij}) et dont le coefficient directeur est $a' = \dfrac{cov(x,y)}{V(y)}$.
Une équation cartésienne de cette droite est : $(D') : x - \overline{x} = \dfrac{cov(x,y)}{V(y)}(y - \overline{y})$. Cette droite peut s'écrire sous la forme $(D') : x = a'y + b'$ avec $a' = \dfrac{cov(x,y)}{V(y)}$ et $b' = \overline{x} - a'\overline{y}$.

1.5.4 Coefficient de corrélation

a) Définition

Soit $(x_i; y_j; n_{ij})$ une série statistique à deux caractères x et y telles que $V(x)$ et $V(y)$ non nulles. Le coefficient de corrélation linéaire de cette série est le nombre réel noté r tel que : $r = \dfrac{cov(x,y)}{\sqrt{V(x) \times V(y)}} = \dfrac{cov(x,y)}{\Gamma(x) \times \Gamma(y)}$.

b) Propriétés

$\triangleright$ $r = a.a' \Longrightarrow |r| = \sqrt{a.a'}$;

$\triangleright$ $-1 \leq r \leq 1$, c'est-à- dire $0 \leq |r| \leq 1$;

$\triangleright$ Si $|r| = 1$, alors tous les points du nuage sont alignés. L'ajustement est dit parfait (les résultats sont fiables), droites de régressions (D) et (D') sont confondues ;

$\triangleright$ Si $0,87 \leq |r| \leq 1$, on dit qu'il y a une forte corrélation entre x et y (l'ajustement se justifie) ;

$\triangleright$ Si $|r|$ est très voisin de 0, alors il y a indépendance linéaire statistique ;

$\triangleright$ Si $|r| < 0,87$, on dit qu'il y a une faible corrélation entre x et y
(l'ajustement n'est pas justifie).

Exercices d'applications

Exercice 1

Les résultats d'une étude statistique effectuée sur une population féminine sont confinés dans le tableau ci-dessous :

Age : x	36	42	48	54	60
Tension artérielle : y	11,7	14	12,5	15	15,6

Les résultats des calculs seront donnés sous forme décimale et à 10^{-2} près.

1. Représenter le nuage de points de cette série statistique double.
   On prendra $1,5cm$ pour 12 an et $0,5cm$ pour l'unité de tension artérielle.
2. Calculer l'âge moyen et la tension moyenne de cette série statistique.
3. Déterminer l'équation de la droite de régression de y en x.
4. Calculer le coefficient de corrélation linéaire entre x et y.
5. Si l'évolution de la valeur de la tension artérielle se poursuit de la même manière, une personne âgée de 65 ans pourrait-elle avoir une tension artérielle de 17 ? Justifier votre réponse.

Exercice 2

Soit le tableau statistique à double entrée suivant :

Y \ X	0	2
-1	3	2
α	1	4

avec α un entier naturel non nul.

1. Donner les lois marginales de X et Y.
2. Déterminer le réel α pour que le point moyen G ait pour coordonnées $\left(\frac{4}{5}; \frac{1}{2}\right)$.
3. Calculer la covariance de (X, Y).
4. Calculer le coefficient de corrélation linéaire.

Exercice 3

La série statistique ci-dessous concerne les ventes annuelles exprimées en milliers de francs CFA de 2014 à 2019 d'un magasin de pagnes, x_i en année et y_i vente en milliers de francs CFA.

x_i	2014	2015	2016	2017	2018	2019
y_i	3400	2800	3200	3800	4300	4700

1. Déterminer les coordonnées du point moyen G.
2. Déterminer la variance de X et Y puis la covariance de (X, Y).
3. Déterminer la droite de régression de y en x.
4. Quelle prévision de vente annuelle le magasin de pagne peut-il espérer atteindre en 2025 ?

Exercice 4

A l'oral d'un examen, chaque candidat est interrogé en première langue où il obtient la note X et en seconde langue où il obtient la note Y (notes sur 20). Les résultats obtenus par 100 candidat sont donnes ci-dessous :

Y \ X	[0, 4[	[4, 8[	[8, 12[	[12, 16[	[16, 20[
[0, 4[	2	5	2	0	0
[4, 8[	1	12	10	3	0
[8, 12[	0	3	28	12	0
[12, 16[	0	1	5	10	2
[16, 20[	0	0	0	1	2

1. Représenter graphiquement la statistique double (X, Y) par un nuage de points pondérés.
2. (a) Déterminer les distributions marginales de X et Y.
   (b) Calculer les notes moyennes de X et Y.
   (c) Calculer la variance de X et Y.
   (d) En déduire l'inertie minimale de cette série statistique.
3. (a) Calculer la covariance de couple (X, Y).
   (b) Déterminer les droites de régressions linéaires de y en x et de x en y
   (c) Déterminer le coefficient de corrélation linéaire puis interpréter.

Exercice 5

On considère la série statistique à deux caractères présentée dans le tableau suivant :

x_i	4	20	1	10	9	5	2	2	6	5	11	15
y_i	3,3	0,6	10,5	1,3	1,3	2,2	5	8	1,8	2	1,2	0,9

1. Représenter le nuage de points associés à cette série.
2. La forme du nuage suggère-t-elle un ajustement linéaire ?
3. On pose $z_i = \dfrac{1}{y_i}$.
   (a) Déterminer le coefficient de corrélation linéaire de la série statistique du couple (x_i, z_i).
   (b) La valeur de ce coefficient justifie-t-elle un ajustement linéaire ?
   (c) Déterminer une équation de la droite de régression de z en x.

Exercice 6

Soit le tableau statistique linéaire suivant :

x	0	1	0	1	0	-1	1	0
y	0	0	α	α	0	0	0	0

où α est un entier nature non nul.

1. Convertir ce tableau en un tableau à double entrée.
2. Déterminer α sachant que $G\left(\dfrac{1}{4}; \dfrac{1}{2}\right)$ est un point moyen de cette série.
3. Dans la suite de l'exercice, on donne $\alpha = 2$.
   (a) Déterminer l'inertie minimale.
   (b) En déduire l'inertie du nuage par rapport au point $P(1, 1)$.
4. Déterminer la droite (D) de régression de y en x.
5. Déterminer le coefficient de corrélation linéaire de la série.
6. Donner une interprétation de cette corrélation.

Exercice 7

Une entreprise achète, utilise et vend des machines après un certain nombres x_i d'années. Après six années, l'évolution y_i en milliers de $FCFA$ du prix de vente d'une machine en fonction du nombre d'années d'utilisation, se présente comme suite :

x_i	1	2	3	4	5	6
y_i	150	125	90	75	50	45

1. Déterminer les coordonnées du point moyen G, puis les variances de X et Y
2. Calculer l'inertie par rapport au point moyen G, puis déduire celui du point $A(0, -1)$.
3. Déterminer une équation de la droite de régression de y en x.
4. En déduire une estimation du prix de vente d'une machine après 7 ans d'utilisation.

Exercice 8

Une série statistique est distribuée selon le tableau ci-dessous.

Une mauvaise manipulation a effacé une donne et on se propose de reconstituer la série en déterminant la donnée manquante.

$x \backslash y$	-1	1	2
1	3	2	1
-1	β	1	1

où β est entier naturel.

1. Déterminer en fonction de β les moyennes $\overline{x}$ et $\overline{y}$, les variances $V(x)$ et $V(x)$ respectives des caractères X et Y.
2. Déterminer β sachant que la covariance de la série est $cov(x, y) = -0,5$.
3. Déterminer le coefficient de corrélation linéaire de la série statistique.

Exercice 9

Une série statistique double d'effectif total 100 est ajustée par la méthode des moindres carrés par les droites de régression dont les équations sont les suivantes :
$(D_{y/x}) : y = \frac{1}{14}x + \frac{3}{14}$ et $(D_{x/y}) : x = \frac{1}{4}y - \frac{1}{4}$.

1. Déterminer les coordonnées du point moyen G.
2. Sachant que la variance de la variable X est : $V(X) = \frac{14}{25}$, déterminer la covariance de (X, Y).
3. Déterminer la variance de la variable Y et l'inertie minimale du nuage des points de la série.

Exercice 10

Le tableau à double entrée ci-dessous présente les résultats d'une étude faite sur un échantillon des ménages d'une ville.

Les x et y sont respectivement les revenus et épargnes mensuels de ces ménages. x et y sont en milliers de francs cfa.

x \ y	45	75	125
14	4	24	2
25	4	36	0
40	0	12	18

1. Donner les deux séries marginales associées à ce tableau.
2. Calculer le revenu moyen et l'épargne moyenne.
3. Calculer les variances de x et de y puis déduire l'inertie minimale.
4. (a) Déterminer l'équation de la droite (D) de régression qui permet d'estimer le revenu mensuel à partir de l'épargne mensuelle.

   (b) En déduire une estimation du revenu mensuel d'un ménage ayant réalisé une épargne mensuelle de $50.000 fcfa$.
5. Calculer le coefficient de corrélation linéaire r, puis interpréter le résultat.

Exercice 11

Dans une classe de terminale au lycée Emery Patrice LUMUMBA, deux élèves de la série D se disputent sur un problème statistique à deux variable X et Y de dix individus. Dans ce problème, il est question de calculer les variances de X et Y connaissant la droite de régression linéaire de y par rapport à x $(D) : y = \frac{1}{2}x$, l'inertie par rapport au point $A(1, 2)$ et l'inertie par rapport au point G qui sont respectivement $I_A = 100$ et $I_G = 50$ et le coefficient de corrélation linéaire de X et Y est $cov(x, y) = \frac{9}{10}$.

Pour les aider, votre enseigne vous demande de :

1. Déterminer les coordonnées du point moyen $G(\overline{x}, \overline{y})$ où $\overline{x} \in \mathbb{R}^*$ et $\overline{y} \in \mathbb{R}$.
2. Soit (D') la droite de régression linéaire de x par rapport à y d'équation cartésienne $(D') : x = a'y + b'$ où a' et b' sont des réels.

   (a) Déterminer les réels a' et b'.

   (b) Montrer que $V(x) - 2a'V(y) = 0$ où $V(x)$ et $V(y)$ les variances de X et Y et a' le coefficient directeur de la droite (D').

(c) Calculer les variances de X et Y.

Exercice 12

Le tableau ci-dessous représente le couple (X, Y) des deux caractères d'une série statistique.

X est le nombre de jours et Y le poids en mg d'une larve.

X	1	2	3	4	5	6
Y	0,2	1,4	1,8	2	2,6	3

1. Calculer les coordonnées du point moyen G.
2. Calculer les variances de X et Y puis la covariance de (X, Y).
3. Déterminer le coefficient de corrélation linéaire puis interpréter.
4. Calculer l'inertie minimale puis déduire l'inertie par rapport au point $B(1; -1)$.
5. Déterminer l'équation de la droite (D) de régression linéaire de y en x.
6. Estimer le poids de la larve au septième jour.

Exercice 13

Le directeur des ressources humaines de l'entreprise $\ll$ *EMERGENCE* 2025 $\gg$ doit embaucher des ouvriers. Lors de la précédente campagne de recrutement pour les postes analogues, il fait une étude statistique sur le nombre des candidatures y en fonction des salaires proposés x. Il a eu les résultats suivants :

▷ Salaire moyen : $\overline{X} = 660.000 fcfa$.

▷ Variance de X : $V(X) = 20.000$.

▷ Équation de la droite de régression de y en x : $y = 0,001125x - 56$.

▷ Coefficient de corrélation linéaire : $r = 0,922$.

1. Déterminer le nombre moyen de candidatures $\overline{Y}$.
2. Déterminer la covariance de (X, Y) de la série.
3. Déterminer l'équation de la droite de régression de x en y.
4. En déduire une estimation de salaire que doit proposer le Directeur s'il veut embaucher 30 ouvriers.