I- Covariance :

Introduction :

En deuxième et en troisième année on a vu que la variance permet une mesure de l'écart à la moyenne des valeurs de la variable d'une série statistique simple. On peut se demander : existe-t-il un paramètre qui permet de mesurer la dispersion des points du nuage par rapport au point moyen dans le cas d'une série double ?

Activité (1 , 1)

On considère les deux séries statistiques doubles suivantes :
La série A présente le taux global X (en %) de la population active en Tunisie et le taux Y (en %) de la population masculine active.
La série B présente le taux global X' (en %) de la population active en Tunisie et le taux Y' (en %) de la population féminine active

SERIE A	1966	1975	1984	1994	2004
Taux global X	46	50	51	48	46
Taux masculine Y	86	81	80	74	68

SERIE B	1966	1975	1984	1994	2004
Taux global X'	46	50	51	48	46
Taux masculine Y'	6	19	22	33	24

1) Construire dans deux repère différents, les nuages des points des deux séries A et B
2) Placer les points moyens G_A de la série A et G_B de la série B
3) Calculer le réel $C_A = \frac{1}{5}\left(\sum_{i=1}^{5} x_i . y_i\right) - \bar{X}.\bar{Y}$, où $\bar{X}$ et $\bar{Y}$ sont respectivement les moyennes arithmétiques des variables X et Y
4) Calculer $cov(X', Y')$
5) Quelle est la série dont les points sont plus dispersés par rapport à son point moyen ?

Définition :

Soit (X , Y), une série statistique double sur un échantillon de taille n.
On appelle covariance de (X , Y) le réel noté $cov(X , Y)$ défini par
$cov(X , Y) = \frac{1}{n}(\sum_{i=1}^{n} x_i . y_i) - \bar{X}.\bar{Y}$ où (x_i , y_i) est la valeur observée pour l'individu i si X et Y sont discrètes, ou bien le centre de la classe si l'une des variables est continue.

Conséquence : On a : $cov(X , Y) = cov(Y , X)$

Remarque :

La variance permet une mesure de l'écart à la moyenne des valeurs de la variable d'une série statistique simple :

1) La covariance permet une mesure de la dispersion des points du nuage par rapport au point moyen
2) La covariance est positive si X et Y ont tendance à varier dans le même sens
3) La covariance est négative si X et Y ont tendance à varier en sens contraire

Propriétés :

Soient (x_i , y_i) avec $1 \leq i \leq n$, une série statistique doubles, $\alpha \in IR$ et $\beta \in IR$ on a :

$$cov(x + \alpha , y + \beta) = cov(x , y) \quad \text{et} \quad cov(\alpha x , \beta y) = \alpha.\beta cov(x, y)$$

Activité (1 , 2) :

On a relevé dans le tableau suivant le nombre de logements (en milliers) et le nombre de logements modernes (villa, appartement) durant quelque années

SERIE A	1992	1998	2001	2008	2013
X : nombre de logements	1313	1512	1870	2204	2501
Y : Nombre de logements modernes	265	343	630	848	1128

Calculer $\bar{X}$ et $\bar{Y}$ puis $cov(X , Y)$. Interpréter le résultat

Définition :

Soit (X , Y), une série statistique double sue un échantillon de taille n
Soit n_i le nombre de fois qu'apparaît le couple (x_i , y_i)
$cov(X , Y) = \frac{1}{n}\left(\sum_{j=1}^{q} \sum_{i=1}^{p} n_{ij}\, x_i y_i\right) - \bar{X}.\bar{Y}$

Activité (1 , 3) :

Le tableau ci-dessous donne le poids Y (en kg) de 63 nouveaux nés ainsi que le poids maternel X

Y \ X	[40 , 50[	[50 , 60[	[60 , 70[	[70 , 80[
[1,5 ; 2,5[	1	0	1	0
[2,5 ; 3,5[	11	17	13	2
[3,5 ; 4,5[	4	4	8	2

1) Calculer $\bar{X}$ et σ_X de X , ainsi que $\bar{Y}$ et σ_Y de Y
2) Calculer $cov(X , Y)$; Interpréter le résultat

1) <u>Etude de la variable X :</u>

x_i : centres des classes	45	55	65	75	
n_i	16		22		$\sum_{i=1}^{4} n_i =$
x_i^2	2025				
$n_i x_i$	720				$\sum_{i=1}^{4} n_i x_i =$
$n_i x_i^2$	32400				$\sum_{i=1}^{4} n_i x_i^2 =$

Le calcule donne :

$\bar{X} = \frac{1}{63}\sum_{i=1}^{4} n_i x_i =$; $V(X) = \left[\frac{1}{63}\sum_{i=1}^{4} n_i x_i^2\right] - (\bar{X})^2 =$ et $\sigma(X) = \sqrt{V(X)} =$

2) <u>Etude de la variable Y :</u>

y_j : centre des classes	2	3	4	
n_j	2			$\sum_{j=1}^{3} n_j =$
y_j^2	4			
$n_j y_j$	4			$\sum_{j=1}^{3} n_j y_j =$
$n_j y_j^2$	8			$\sum_{j=1}^{3} n_j y_j^2 =$

Le calcule donne :

$\bar{Y} = \frac{1}{63}\sum_{j=1}^{3} n_j y_j =$; $V(Y) = \left[\frac{1}{63}\sum_{j=1}^{3} n_j y_j^2\right] - (\bar{Y})^2 =$ et $\sigma(Y) = \sqrt{V(Y)} =$

3) <u>Dressons les couples distincts des valeurs observées et leurs effectifs :</u>

Couples (x_i , y_i)	(45,2)	(45,3)	(45,4)	(55,3)	(55,4)	(65,2)	(65,3)	(65,4)	(75,3)	(75,4)
Effectifs n_{ij}	1	11	4	17	4	1	13			
$n_{ij} . x_i . y_i$	90	1485								

Le calcule donne : $\sum_{j=1}^{3}\sum_{i=1}^{4} n_{ij}\, x_i y_i =$

D'où $cov(X , Y) = \frac{1}{63}\left(\sum_{j=1}^{3}\sum_{i=1}^{4} n_{ij}\, x_i y_i\right) - \bar{X}.\bar{Y} = \frac{1}{63}\text{x}11775 - \bar{X}.\bar{Y} =$

<u>Utilisation d'une calculatrice :</u>

Pour choisir le mode de fonctionnement en statistiques, appuyer sur : « MODE » , « 1 » puis appuyer sur : « 1 » pour sélectionner le sous mode statistique à deux variables.

- Pour entrer les données, taper : «x_i » ; « STO » ; « y_i » ; « STO » ; « n_{ij} » ; « M+ »
  Par exemple pour le couple (45,2) taper : « 45 » ; « STO » ; « 2 » ; « STO » ; « 1 » ; « M+ » et ainsi de suite pour tout les autres couples.
  *** On appuie sur : « RCL » ; « n ». La calculatrice affiche : 63
  *** On appuie sur : « RCL » ; « $\bar{X}$ ». La calculatrice affiche : …
  *** On appuie sur : « RCL » ; « σ_x ». La calculatrice affiche : …
  *** On appuie sur : « RCL » ; « σ_x » ; « x^2 ». La calculatrice affiche : … V(X)
  *** On appuie sur : « RCL » ; « $\sum xy$ ». La calculatrice affiche : … $\sum_{j=1}^{3}\sum_{i=1}^{4} n_{ij}\, x_i y_i$
  *** On appuie sur : « RCL » ; « $\sum xy$ » ; « ÷ » ; « 63 » ; «-» ;« RCL » ;« $\bar{X}$ » ; « x » ;« RCL » ; « $\bar{Y}$ »
  La calculatrice affiche : … (la valeur de $cov(X , Y)$)

II- Ajustement :

Introduction :

L'analyse d'un nuage de point $M_i(x_i , y_i)$ représentant une série statistique double (x_i , y_i) peut conduire à la recherche d'une liaison entre les deux variables x et y . Cette liaison aide, entre autre, à faire des prévisions et à répondre à des questions parfois décisives.
Une question s'impose alors : peut-on trouver une formule mathématique qui exprime le lien entre les deux variables ? la réponse à cette question conduit à étudier le type de relation entre les deux variables (affine, polynomiale, homographique, logarithmique, exponentiel). On parle d'ajustement

Ajustement affine d'une série statistique double :

1) **Méthode de Mayer :**

Activité (2 , 1)

Le tableau ci-dessus donne le relevé des valeurs d'une action (en DT) sur 15 jours consécutifs d'une bourse.

Jour X	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Valeur Y	18.8	18.9	18.9	19.5	19.2	19	19.2	19.6	19.5	19.7	19.2	19.7	19.8	20	20.5

On note par le nuage N_1 des points associé à la série (x_i , y_i) avec $1 \leq i \leq 8$ et N_2 le nuage des points restant.

1) Déterminer le point moyen G_1 de la première série
2) Déterminer le point moyen G_2 de la deuxième série
3) Déterminer l'équation de la droite (G_1G_2)
4) La droite (G_1G_2) passe telle par le point moyen de la série totale ?

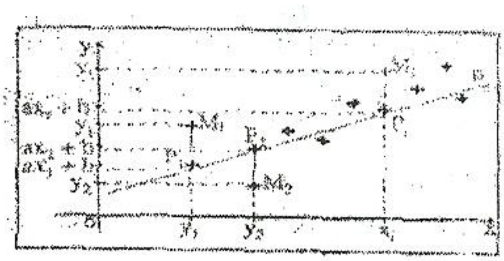
Définition :

Le principe de l'ajustement par la méthode de Mayer consiste à partager le nuage associé à une série (x_i , y_i) en deux nuages dont le nombre de points diffère d'au plus un. On désigne par G_1 et G_2 les points moyens respectifs du premier et du deuxième nuage, la droite (G_1G_2) est appelée **droite de Mayer** on a $G \in (G_1G_2)$

2) **Méthode d'ajustement par les moindres carrés :**

Définition :

Le principe de l'ajustement par la méthode des **moindres carrés** consiste à déterminer les réels a et b tels que la somme $\sum_{i=1}^{n}(M_iH_i)^2$ soit minimale avec $M_i(x_i , y_i)$, $1 \leq i \leq n$



Le nuage de points d'une série statistique double, ainsi $D : y = ax + b$ et $H_i(x_i , y_i)$ le point de la droite D de même abscisse que M_i . **On admet qu'une telle droite existe et qu'elle est unique. On l'appelle droite de régression de y en x .**

Théorème :

La droite de régression de y en x dans un repère orthogonal associée à la série statistique double (X , Y) est la droite qui passe par le point moyen $G(\bar{X} , \bar{Y})$ et de coefficient directeur le réel $a = \frac{cov(X,Y)}{V(X)}$

Définition :

Soit (X , Y) une série statistique double sur un échantillon de taille n

1) La droite d'équation $y = \frac{cov(X,Y)}{V(X)}.(x - \bar{X}) + \bar{Y}$ est appelée droite des moindres carrés de Y en X, ou droite de régression de Y en X.
2) La droite d'équation $x = \frac{cov(X,Y)}{V(Y)}.(y - \bar{Y}) + \bar{X}$ est appelé droite des moindres carrées de X en Y, ou droite de régression de X en Y.

Activité (2 , 2) :

X :(en degré C°)	-2	0	4	8	10
Y :(en litres)	40	30	20	15	10

Dans le tableau ci-contre
X désigne la température moyenne extérieur en 24 heures et
Y désigne la consommation de pétrole de chauffage pour les mêmes 24 heures et pour une famille donnée

1) Déterminer le point moyen G de la série (X , Y)
2) Représenter, dans un repère orthogonal le nuage de points $M_i(x_i\ , y_i)$; L'ajustement affine est-il possible ?
3) Donner une équation de la droite de régression de Y en X
4) Quelle prévision (en litres) sur sa consommation de pétrole peut faire la famille considérée, si une vague de froid persiste pendant 48 heures avec une température moyenne de (-4) C° ?

3) Coefficient de corrélation linéaire :

On peut toujours au vu des formules précédentes construire une droite de régression. Mais parfois cette dernière n'est d'aucune efficacité G, dans la mesure où les prédictions que l'on fait à partir de cette droite ne sont pas raisonnables. C'est le cas lorsqu'il n'existe pas réellement de corrélation entre les deux variables. Pour savoir si a est pertinent d'ajuster un nuage de point par les moindres carrés, on calcule un réel appelé coefficient de corrélation linéaire

Définition :

Soit (X , Y) une série statistique double. On appelle coefficient de corrélation linéaire le réel noté r(X , Y) défini par : $r(X,Y) = \frac{cov(X,Y)}{\sigma_X.\sigma_Y}$

Remarque :

1) On a : $-1 \leq r(X,Y) \leq 1$
2) Si $|r(X,Y)| = 1$ alors il y a une dépendance totale, l'une est une fonction affine de l'autre.
3) Si $|r(X,Y)| \in [0\ ; 0{,}7]$ alors la corrélation entre X et Y est faible.
4) Si $|r(X,Y)| \in]0{,}7\ ; 0{,}95]$ alors la corrélation entre X et Y est forte.
5) Si $|r(X,Y)| \in]0{,}95\ ; 1]$ alors la corrélation entre X et Y est très forte.

Activité (2 , 3) :

Le tableau suivant donne l'effectif de la population scolaire de la $4^{ème}$ année de l'enseignement secondaire du mois d'octobre 2008 au mois d'octobre 2013

X : (année)	2008	2009	2010	2011	2012	2013
Y : (population scolaire en $4^{ème}$)	77755	84581	89266	86138	90123	100087

1) Calculer le coefficient de corrélation linéaire
2) Déterminer un ajustement par les moindres carrés de la série double puis donner une estimation de la population scolaire en $4^{ème}$ année secondaire au mois d'octobre 2015

4)Exemples d'ajustements non affines d'une série double :

Activité (2 , 4)

Le tableau suivant donne l'évolution de salaires nets en indice, de base 100 en 2002 dans un pays industrialisé :

Année	2002	2003	2004	2005	2006	2007	2008	2009
X : (Rang)	0	1	2	3	4	5	6	7
Y : (Indice)	100	97.6	96.8	98.4	98.3	99.8	103.3	106.7

1) a- Représenter la série double (X, Y) dans un repère orthogonal
   b- Un ajustement affine est-il justifier ?
2) On se propose de faire un ajustement par une fonction polynôme f de la forme $f(x) = ax^2 + bx + c$
   a- Déterminer les réels a, b et c pour avoir f $(0) = 100$, $f(5) = 100$ et $f(7) = 107$
   b- Construire C_1 dans le repère précédent
   c- A l'aide de cet ajustement calculer la prévision de l'indice des salaires en 2015

Activité (2 , 5) : (ajustement logarithmique)

Le tableau ci-dessous donne la production de pétrole de 1987 à 1997 suivant L'OPEP , x : le rang de l'année et y : la production (en millions de tonnes). On pose $X = \ln(x)$, les valeurs arrondies à 10^{-2} près

Année	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
Y : production	944	1065	1137	1232	1231	1297	1332	1333	1368	1408	1423
Rang : X	1	2	3	4	5	6	7	8	9	10	11
$X = \ln(x_i)$	0	0.69	1.1	1.39	1.61	1.79	1.95	2.08	2.2	2.3	2.4

1) a- Donner une équation de la droite de régression de Y en X de la série double (X , Y) sous la forme $y = \alpha x + \beta$ avec α et β arrondis à l'unité
   b- En déduire une relation entre x et la production y : $y = f(x)$
2) a- Dans un même repère orthogonal, placer le nuage de points $M_i(x_i\,, y_i)$ et représenter la fonction f définie sur $[1\,, +\infty[$
   b- A l'aide de cet ajustement, donner une estimation de la production de pétrole en 2015, si cette politique se poursuit

Activité (2 , 6) : (ajustement exponentiel)

Le tableau suivant donne l'effet de la pollution sur la population piscicole d'une rivière de 2006 et 2011Soit un repère orthogonal, on pose $Z = \ln(Y)$, les valeurs arrondies de Z à 10^{-2} près

Année	2006	2007	2008	2009	2010	2011
X : (Rang)	1	2	3	4	5	6
Y : (Nombre de poissons)	951.3	106.7	96.5	63.2	21	9.4
$Z = \ln(Y)$	6.86	4.67	4.57	4.15	3	2.24

1) Représenter le nuage de points $M_i(x_i\,, \ln(y_1))$, dans ce repère
2) a- Calculer le coefficient de corrélation de (X , Z) et justifier que l'on peut procéder à un ajustement affine par les moindres carrés.
   b- Donner une équation de la droite de régression de Z en X, sous la forme $Z = \alpha X + \beta$, en arrondissant α et β au centième
   c- En déduire en utilisant l'égalité $Z = \ln(Y)$, un ajustement exponentiel de Y en X sous la forme $Y = A.e^{B.X}$
3) On suppose que l'évolution de cette population se poursuit sur le même modèle
   a- A partir de quelle année cette population sera-t-elle inférieur à 1000 ?
   b- Donner une estimation de la population de cette rivière en l'an 2014 ?

Activité (2 , 7) : (ajustement homographique)

Le tableau suivant donne le taux d'équipement des ménages en automobile de 1969 à 2000 en France :

Année	1969	1973	1977	1979	1980	1984	1986	1988	1990	1991	1992	1993	1996	1998	2000
Taux	55.4	61.6	66.1	68.6	70	72.9	73.4	74.6	76.5	76.8	77	78	78.9	79.4	80

1) a- Déterminer l'ajustement affine par la méthode des moindres carrés de la série double (X , Y) en prenant : x : (Année – 1900) et Y : Taux et on donnera le coefficient α à 10^{-3} près et β à 10^{-1} près
   b- En déduire la valeur du taux d'équipement en 2000 à l'aide de cet ajustement
   Comparer le résultat trouvé à la valeur réel
2) On se propose de faire un ajustement par une fonction homographique f de la forme $f(x) = \frac{kx+m}{x-50}$ pour $x \geq 60$
   a- Déterminer les réels k et m pour avoir $f(80) = 70$ et $f(100) = 80$
   b- Etudier la fonction f
   c- Calculer le taux en 2000 à l'aide de cet ajustement ; Comparer le résultat trouvé à la valeur réel
   d- A l'aide de cet ajustement calculer la prévision du taux d'équipement en 2014