STATISTIQUES

I. DISTRIBUTIONS MARGINALES

Activité 1 p 209

DEFINITION 1

Soit (X, Y) une série statistique double sur un échantillon de taille n et soit $(x_i, y_i)_{1\le i\le n}$ les valeurs numériques prises respectivement par les variables X et Y.

* La distribution marginale de la variable X est la distribution des valeurs $(x_i)_{1\le i\le n}$ prises par la variable X
* La distribution marginale de la variable Y est la distribution des valeurs $(y_i)_{1\le i\le n}$ prises par la variable Y

DEFINITION 2

* Soit X une série statistique sur un échantillon de taille n et soit $(x_i)_{1\le i\le p}$ les valeurs numériques prises par la variable X si elle est discrète, ou les centres des classes si la variable X est continue.
* L'entier n_i désigne l'effectif de la valeur x_i.
* Si $\overline{X}$, V(X) et σ_X désigne respectivement la moyenne, la variance et l'écart-type de X, alors :

* $\overline{X} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i$ * $V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \overline{X})^2$ * $\sigma_X = \sqrt{V(X)}$

* $V(X) = \overline{X^2} - \left(\overline{X}\right)^2$; avec $\overline{X^2} = \sum_{i=1}^{p} n_i(x_i)^2$

Activité 3 p 211

II. COVARIANCE D'UNE SERIE STATISTIQUE DOUBLE

1. CAS D'UN ECHANTILLON SIMPLE

EXERCICE (BAC. S. 2005)

Le tableau suivant donne la distance de freinage d (en mètres) d'une voiture, en fonction de sa vitesse v (en kilomètres par heure)

v (km/h)	30	40	50	60	70	80
d (mètres)	42	60	80	90	95	110

* On note $\overline{v}$ et $\overline{d}$ les moyennes respectives de v et d.
* On note V(v) et V(d) les variances respectives de v et d.
* On note cov(v, d) la covariance de v et d.

Calculer $\overline{v}$, $\overline{d}$, V(v), V(d) et $\frac{1}{6}\sum_{k=1}^{6} v_k.d_k - \overline{v}\,\overline{d}$.

DEFINITION

Soit (X, Y) une série statistique double sur un échantillon de taille n et soit $(x_i, y_i)_{1\le i\le p}$ les valeurs numériques prises respectivement par les variables X et Y si X et Y sont discrètes, ou les centres des classes si la variable X ou Y est continue.

* On appelle covariance de (X, Y) le réel, noté cov(X, Y) défini par :

* $cov(X, Y) = \frac{1}{n}\sum_{i=1}^{p} x_i y_i - \overline{X}.\overline{Y} = \overline{X.Y} - \overline{X}.\overline{Y}$

INTERPRETATION DE LA COVARIANCE

* La covariance mesure la tendance qu'on les variables X et Y de varier ensemble.
* La covariance est positive si X et Y ont une tendance à varier dans le même sens.
* La covariance est négative si X et Y ont une tendance à varier en sens contraire.

Activité 3 p 213

EXERCICE 1 (CONT. M. 95)

Dans le tableau statistique suivant, X désigne la température moyenne extérieure en 24 heures et Y désigne la consommation de pétrole de chauffage pour les mêmes 24 heures et pour une famille donnée.

X en degrés	-2	0	4	8	10
Y en litres	40	30	20	15	10

Calculer $\overline{X}$, $\overline{Y}$, V(X), V(Y) et cov(X,Y). Interpréter le résultat obtenu.

2. CAS D'UN ECHANTILLON GROUPE

DEFINITION

Soit (X, Y) une série statistique double sur un échantillon de taille n et soit $(x_i, y_j)_{1\leq i\leq p \text{ et } 1\leq i\leq q}$ les valeurs numériques prises respectivement par les variables X et Y si X et Y sont discrètes, ou les centres des classes si la variable X ou Y est continue. Soit n_{ij} l'effectif qui correspond au couple (x_i, y_j).

* On appelle covariance de (X, Y) le réel, noté cov(X, Y) défini par :
* $\text{cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_i y_j - \overline{X}.\overline{Y} = \overline{X.Y} - \overline{X}.\overline{Y}$

Exercice résolu p 213

III. AJUSTEMENT D'UNE SERIE STATISTIQUE DOUBLE

1. METHODE DE MAYER

ACTIVITE

Le rendement R d'une variété de blé (en quintaux par hectare) et la quantité E d'engrais azotés (en kilogrammes par hectare) utilisée pendant la culture sont indiqués dans le tableau suivant :

E(kg/ha)	30	40	50	60	70	80	90	100
R (q/ha)	28,7	34,4	35,7	41,4	45,7	47,2	50,8	54,7

1. Construire, dans un plan rapporté à un repère orthogonal, le nuage de points représentant la série statistique double donnée. Que peut-on remarquer ?
2. Calculer $\overline{X}$, $\overline{Y}$, V(X), V(Y), $\sigma(X)$, $\sigma(Y)$, $\overline{XY}$.
3. Représenter le point moyen G($\overline{X}$, $\overline{Y}$).
4. On partage le nuage de points en deux ensembles A et B
   a. Déterminer le point moyen G_1 de la partie A et le point moyen G_2 de la partie B.
   b. Donner une équation cartésienne de la droite (G_1G_2).
   c. Vérifier que la droite (G_1G_2) passe par le point moyen G du nuage de points.
5. Quel rendement peut-on prévoir pour une culture utilisant une quantité d'engrais azotés E = 120kg/ha ?

DEFINITION

Soit (X, Y) une série statistique double de valeurs $(x_i, y_i)_{1\leq i\leq n}$.

* L'ensemble des points Mi(xi, yi) dans un repère orthogonal est appelé nuage de points représentant la série statistique.
* Le point moyen du nuage est le point G($\overline{X}$,$\overline{Y}$).

PRINCIPE DE LA METHODE DE MAYER

Soit un nuage de points représentant une série statistique double (X, Y) et G son point moyen.

* On scinde le nuage de points de la série (X, Y) en deux parties contenant à peu près le même nombre de points.
* On considère alors les points moyens G_1 et G_2 des deux nages obtenus.
* La droite (G_1G_2) définie un ajustement affine du nuage de points représentant la série statistique double (X, Y).
* La droite (G_1G_2) est appelée droite de Mayer et passe par le point moyen G du nuage.

Activité 2 p 218

2. METHODE DES MOINDRES CARRES

THEOREME

Soit (X, Y) une série statistique double sur un échantillon de taille n et telle que $\sigma_X \neq 0$.

* Soit $(x_i, y_i)_{1 \leq i \leq n}$ les valeurs observées de la série. Alors la somme $\sum_{i=1}^{n} (ax_i + b - y_i)^2$ est minimale pour le couple (a, b) tel que $a = \frac{\text{cov}(X, Y)}{\sigma_X{}^2}$ et $b = \overline{Y} - a\overline{X}$.

DEFINITION 1

Soit (X, Y) une série statistique double sur un échantillon de taille n et telle que $\sigma_X \neq 0$.

* La droite d'équation y = ax + b ; avec $a = \frac{\text{cov}(X, Y)}{\sigma_X{}^2}$ et $b = \overline{Y} - a\overline{X}$; est appelée droite des moindres carrés de Y en X, ou droite de régression de Y en X.

DEFINITION 2

Soit (X, Y) une série statistique double sur un échantillon de taille n et telle que $\sigma_Y \neq 0$.

* La droite d'équation x = ay + b ; avec $a = \frac{\text{cov}(X, Y)}{\sigma_Y{}^2}$ et $b = \overline{X} - a\overline{Y}$; est appelée droite des moindres carrés de X en Y, ou droite de régression de X en Y.

CONSEQUENCE

* Les droites des moindres carrés de Y en X et de X en Y passent par le point moyen G du nuage associé à la série (X, Y).

Activité 2 p 220

3. COEFFICIENT DE CORRELATION LINEAIRE

DEFINITION

Soit (X, Y) une série statistique double.

* On appelle coefficient de corrélation linéaire le réel ρ_{XY} défini par : $\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X . \sigma_Y}$

PROPRIETES

Soit (X, Y) une série statistique double.

* $-1 \leq \rho_{XY} \leq 1$
* Le coefficient de corrélation linéaire est invariant par changement d'unité ou d'origine.

INTERPRETATION DU COEFFICIENT DE CORRELATION LINEAIRE

***** On convient que : si $|\,\rho_{XY}\,| > \dfrac{\sqrt{3}}{2}$ alors l'ajustement affine est justifié et les prédictions faites au moyen de cet ajustement sont raisonnables.