Chapitre 8 : SERIE STATISTIQUE DOUBLE

8.1. RESUME DU COURS

Définition

L'étude simultanée de deux caractères x et y d'une population donne une série statistique double.

8.1.1. SERIE STATISTIQUE DOUBLE

Définition

Soit x et y deux caractères observés sur une population, $x_1, x_2, \ldots, x_p$ les valeurs prises par x et $y_1, y_2, \ldots, y_q$ les valeurs prises par y ; soit n_{ij} le nombre de fois qu'on observe le couple $(x_i, y_j)_{1\leq i\leq p \text{ et } 1\leq j\leq q}$ dans la population.

- n_{ij} est appelé effectif du couple (x_i, y_j).
- $\{(x_i, y_j, n_{ij})\}_{1\leq i\leq p \text{ et } 1\leq j\leq q}$ est appelée série statistique double.

Exemple

L'étude simultanée des notes de sciences physiques(x) et de mathématiques (y) de 11 élèves d'une classe de Terminale S1 donne les résultats suivants, présentés dans le tableau à double entrée ci-dessous appelé tableau de contingence ou de corrélation:

y / x	**10**	**12**	**15**	**18**	**Totaux**
11	1	2	1	0	$n_{1.}$ = ..
15	2	3	1	0	$n_{2.}$ = ..
19	0	0	0	1	$n_{3.}$ = ..
Totaux	$n_{.1}$= ..	$n_{.2}$= ..	$n_{.3}$= ..	$n_{.4}$= ..	N= 11

- Le nombre 2 se trouvant dans la case grisée signifie que 2 élèves ont obtenu 11 en Sciences physiques et 12 en Maths.
- Effectifs partiels

$n_{1.} = n_{11} + n_{12.} + n_{13} + n_{14}$

= 1+ 2+ 1+ 0 = 4 ; c'est le nombre d'élèves ayant obtenu 11 en Sciences Physiques.

$n_{2.} = n_{21} + n_{22.} + n_{23} + n_{24}$

= 2 + 3+ 1+ 0 = 6 ; c'est le nombre d'élèves ayant obtenu 12 en Sciences Physiques.

$n_{.1} = n_{11} + n_{21} + n_{31}$

= 1 + 2 + 0 = 3 ; c'est le nombre d'élèves ayant obtenu 10 en Mathématiques.

$n_{.4} = n_{14} + n_{24} + n_{34} + n_{14}$

= 0 + 0 + 1 = 1 ; c'est le nombre d'élèves ayant obtenu 18 en Mathématiques.

Effectifs partiels – Effectif total

- $n_{i.} = \sum_{j=1}^{q} n_{ij}$; $n_{.j} = \sum_{i=1}^{p} n_{ij}$
- L'effectif total N = $\sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{q} n_{.j}$
  $= \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$

Fréquences

- $f_{ij} = \frac{n_{ij}}{N}$; $f_{i.} = \frac{n_{i.}}{N}$; $f_{.j} = \frac{n_{.j}}{N}$.
- Fréquences conditionnelles : $f_{x_i/y_j} = \frac{n_{ij}}{n_{.j}}$; $f_{y_j/x_i} = \frac{n_{ij}}{n_{i.}}$

Série statistique marginale

- La série statistique simple$\{(x_i, n_{i.})\}_{1 \le i \le p}$, appelée première série statistique marginale est présentée à l'aide du tableau ci-dessous :

x	x_1	x_2	----------	x_p
$n_{i.}$	$n_{1.}$	$n_{2.}$	----------	$n_{p.}$

Cette série a pour moyenne $\bar{x} = \frac{1}{N} \sum_{i=1}^{p} n_{i.} x_i$, pour variance V($x$) $= \frac{1}{N} \sum_{i=1}^{p} n_{i.} x_i^2 - \bar{x}^2$ et pour écart-type $\sigma(x) = \sqrt{V(x)}$.

- La série statistique simple$\{(y_j, n_{.j})\}_{1 \le j \le q}$, appelée deuxième série statistique marginale est présentée à l'aide du tableau ci-dessous :

y	y_1	y_2	----------	y_q
$n_{.j}$	$n_{.1}$	$n_{.2}$	…….	$n_{.q}$

Cette série a pour moyenne $\bar{y} = \frac{1}{N} \sum_{j=1}^{q} n_{.j} y_j$, pour variance V(y) $= \frac{1}{N} \sum_{j=1}^{q} n_{.j} y_j^2 - \bar{y}^2$ et pour écart-type $\sigma(y) = \sqrt{V(y)}$.

Covariance de x et y

La covariance de x et y est

cov(x, y) $= \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\, x_i y_j - \bar{x}\,.\bar{y}$.

Représentation graphique

- Pour représenter la série double $\{(x_i, y_j, n_{ij})\}_{1\leq i\leq p \text{ et } 1\leq j\leq q}$, on place dans un repère orthogonal les x_i en abscisse et les y_j en ordonnée. Le triplet (x_i, y_j, n_{ij}) est représenté dans le repère par le point pondéré $M_{ij}(n_{ij})$ de coordonnées $(x_i,\ y_j)$. L'ensemble des points $M_{ij}(n_{ij})$ constitue la représentation graphique de la série ; cette représentation est appelée nuage de points.
- On appelle point-moyen du nuage de points, le barycentre G des points pondérés $M_{ij}(n_{ij})$.

8.1.2. CAS PARTICULIER : SERIE DOUBLE INJECTIVE

Définition

Si les caractères *x* et *y* prennent le même nombre de valeurs (p = q= n) et si $n_{ij} = \begin{cases} 1 \ si\ i = j \\ 0 \ si\ i \neq j \end{cases}$ alors la série double est dite injective ; elle est notée $\{\{(x_i,\ y_i)\}_{1\leq i\leq n}$ et ces résultats sont présentés dans un tableau de cette forme :

x	x_1	x_2	----------	x_n
y	y_1	y_2	----------	y_n

Effectif total

L'effectif total N de la série est égal au nombre de colonnes du tableau ou bien au nombre de valeurs du caractère x ou du caractère y.

Série marginale

- La première série marginale $\{(x_i, n_i)\}_{1\leq i\leq n}$ a pour moyenne $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i$, pour variance V(x) = $\frac{1}{N}\sum_{i=1}^{n} x^2{}_i$ - $\bar{x}^2$ et pour écart-type $\sigma(x) = \sqrt{V(x)}$.

- La deuxième série marginale $\{(y_i, n_i)\}_{1\leq i\leq n}$ a pour moyenne $\bar{y} = \frac{1}{N}\sum_{i=1}^{n} y_i$, pour variance V(y) = $\frac{1}{N}\sum_{i=1}^{n} y^2{}_i$ - $\bar{y}^2$ et pour écart-type $\sigma(y) = \sqrt{V(y)}$.

Covariance

La covariance de x et y est cov(x, y) = $\frac{1}{N}\sum_{i=1}^{n} x_i y_i$ - $\bar{x}\,\bar{y}$.

Représentation graphique

- Le couple (x_i, y_i) est représenté dans un repère orthogonal par le point $M_i(x_i, y_i)$. L'ensemble des points M_i constitue le nuage de points de la série double.
- Le point G de coordonnées $(\bar{x}, \bar{y})$ est le point-moyen du nuage.

Ajustement linéaire

Si les points du nuage ont l'allure d'une droite, on peut trouver une droite « trés proche » de ces points.

Par la méthode des moindres carrés on trouve deux droites, l'une appelée droite de régression de y en x, notée $d_{y/x}$ et l'autre appelée droite de régression de x en y, notée $d_{x/y}$.

Equations des droites de régression

- $d_{y/x}$: $y - \bar{y} = a(x - \bar{x})$ où $a = \frac{cov(x,\ y)}{V(x)}$.
- $d_{x/y}$: $x - \bar{x} = \alpha(y - \bar{y})$ où $\alpha = \frac{cov(x,\ y)}{V(y)}$.

Remarque :
Les droites de régression passent par le point-moyen $G(\bar{x}, \bar{y})$.

Coefficient de corrélation linéaire

Le coefficient de corrélation linéaire des caractères x et y est $r = \frac{cov(x,\ y)}{\sigma(x)\sigma(y)}$.

Remarques :

- $r^2 = \alpha a$; $-1 \leq r \leq 1$.
- Si $|r|$ très proche de 1, on a une bonne ou forte corrélation entre les caractères x et y.

Estimation

Pour estimer la valeur de y (respectivement x) connaissant celle de x (respectivement y) on remplace x (respectivement y) par sa valeur dans l'une des équations des droites de régression.