Statistiques

Lycée de Dindéfelo
Mr BA

17 mai 2024

I Rappels : Série statistique à une variable

Soit X une série statistique quantitative

Modalités x_i	x_1	x_2	x_3	$\dots$	x_p
Effectifs n_i	n_1	n_2	n_3	$\dots$	n_p

1. Fréquence :

La fréquence de la modalité x_i est le nombre $\frac{n_i}{N}$ ou n_i est l'effectif de la modalité x_i et N l'effectif total

$$N = n_1 + n_2 + \cdots + n_p = \sum_{i=1}^{p} n_i$$

Remarque :

$$\sum_{i=1}^{p} f_i = \frac{n_1 + n_2 + \cdots + n_p}{N} = 1$$

2. Moyenne :

La moyenne de cette série est le réel noté $\overline{x}$ définie par :

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \cdots + n_px_p}{N} = \frac{\sum_{i=1}^{p} n_ix_i}{N}$$

ou

$$\overline{x} = f_1x_1 + f_2x_2 + \cdots + f_px_p = \sum_{i=1}^{p} f_ix_i$$

3. Variance :

La variance de cette série est le réel positif $V(x)$ définie par :

$$V(x) = \frac{n_1(x_1 - \overline{x})^2 + n_2(x_2 - \overline{x})^2 \cdots + \ldots + n_p(x_p - \overline{x})^2}{N} = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{x})^2$$

ou

$$V(x) = f_1(x_1 - \overline{x})^2 + f_2(x_2 - \overline{x})^2 \cdots + ... + f_p(x_p - \overline{x})^2 = \sum_{i=1}^{p} f_i(x_i - \overline{x})^2$$

ou

$V(x) = \frac{1}{N}(n_1x_1^2 + n_2x_2^2 + \cdots + n_px_p^2)$ **moyenne des carrés – le carré de la moyenne**

Remarque :

Pour effectuer un calcul de la variance, la formule (3) qui porte le nom de formule de König est en général plus simple à utiliser.

4. Ecart type

l'écart type d'une série statistique est la racine carrée de la variance on le note

$\sigma_X = \sqrt{V(x)}$

5. Exercice d'application

On donne les notes de 20 élèves à un devoir de mathématiques : 12 ; 9 ; 11 ; 14 ; 9 ; 8 ; 15 ; 7 ; 4 ; 18 ; 12 ; 7 ; 14 ; 12 ; 15 ; 8 ; 15 ; 11 ; 12 ; 11.

1. Dresser le tableau des effectifs et des fréquences de cette série notée X.
2. Calculer la moyenne, la variance et l'écart type de cette série.
3. regrouper les notes par classe d'amplitude 4, puis calculer la moyenne, la variance et l'écart type correspondant

Solution :

1. Tableau des effectifs et des fréquences

Notes x_i	4	7	8	9	11	12	14	15	18
Effectifs n_i	1	2	2	2	3	4	2	3	1
Fréquence f_i	0.05	0.10	0.10	0.10	0.15	0.20	0.10	0.15	0.05

2. la moyenne de cette série est

Soit $\overline{x} = \frac{224}{20} = 11.2$
$\overline{x} = (41 + 72 + 82 + 92 + 113 + 124 + 142 + 153 + 181)$
Calcule de la variance

$$V(x) = \left(4^2 \times 1 + 7^2 \times 2 + 8^2 \times 2 + 9^2 \times 2 + 11^2 \times 3 + 12^2 \times 4 + 14^2 \times 2 + 15^2 \times 3 + 18^2 \times 1\right) - 11,$$

Soit

$$V(x) = \frac{2734}{20} - 11,2^2 = 11,26$$

L'écart type est $\sigma_X = \sqrt{11,26}$ soit $\sigma_X = 3.33$

Classe	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20[$
Effectifs	3	7	9	1
centre	6	10	14	18

Rappel : le centre de l'intervalle
$[4;8[$ est le nombre $\frac{a+b}{2}$
Les centres des intervalles seront considérés comme les modalités
On trouve $= \frac{232}{20} = 11,6$, $V(X) = \frac{2896}{20} - (11,6)^2$ soit $V(X) = 10,24$ et $\sigma_X = 3,2$

II. Série statistique double

L'étude simultanée de deux caractères X et Y sur une même population donne une série statistique double (x_i, y_j, n_{ij}) $1 \leq i \leq p$ et $1 \leq j \leq q$
Où les xi sont les modalités du caractère X et les y_j celles de Y et les nij l'effectif du couple (x_i, y_j).
Exemple : On relevé les notes sur 10 de dictée (X) et de calcul (Y) d'une classe de $CM2$ de 30 élèves. On a obtenu les résultats suivants

$X \setminus Y$	2	3	4	5	7	
0	1	1	2	1	1	$n_1 = 6$
1	3	1	1	1	1	$n_2 = 7$
3	0	0	1	1	1	$n_3 = 3$
4	0	1	0	0	1	$n_4 = 2$
5	0	0	1	0	1	$n_5 = 2$
6	1	0	1	1	0	$n_6 = 3$
7	1	2	0	1	0	$n_7 = 4$
8	1	0	0	1	0	$n_8 = 3$
	$n_1 = 7$	$n_2 = 5$	$n_3 = 7$	$n_4 = 6$	$n_5 = 5$	$N = 30$

Tableau de contingence, l'élément de la i^{ime} ligne et de la j^{ime} colonne est l'effectif n_{ij} du couple (x_i, y_j)

L'ensemble des triplets $\{(x_i, y_i, n_{ij})\}_{1\leq j\leq 8}^{1\leq i\leq 8}$

Première série marginale :

C'est l'ensemble des couples $(x_i, y_j) 1 \leq i \leq p$ où les $n_i = \sum_{j=1}^{q} n_{ij}$ (la somme des effectifs n_{ij} de la i^{ime} ligne) La première série marginale de l'exemple précédent est : $\{(x_1, n_1); (x_2, n_2); \ldots; (x_8, n_8)\}$ où

$\{(0, 6); (1, 7); (3, 3); (4, 2); (5, 2); (6, 3); (7, 4); (8, 3)\}$

x_i	0	1	3	4	5	6	7
n_i	6	7	3	2	2	3	3

Moyenne et variance de X :

on applique les formules

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i \quad ; \quad (\sigma_X)^2 = \frac{1}{N}\sum_{i=1}^{p} n_i (x_i - \overline{x})^2$$

Deuxième série marginale :

c'est l'ensemble des couples $(y_j, n_j) 1 \leq j \leq q$

$$n_j = \sum_{i=1}^{p} n_{ij} \textbf{(la somme des effectifs } n_{ij} \textbf{ de la } j^{ime} \textbf{ colonne)}$$

Moyenne et variance de Y :

on applique les formules

$$\overline{y} = \frac{1}{N}\sum_{j=1}^{q} n_j x_j \quad ; \quad (\sigma_y)^2 = \frac{1}{N}\sum_{j=1}^{q} n_j (y_j - \overline{y})^2$$

Remarque : série statistique double injective

Une série statistique est dite injective lors que :

$$n_{ij} = \delta_{ij} = \begin{cases} 0 & \textbf{si} \quad i \neq j \\ 1 & \textbf{si} \quad i = j \end{cases}$$

lors que X et Y ont le même nombre de valeur ont lui fait correspondre un tableau statistique du type

X	X_1	X_2	$\cdots$	X_n
Y	Y_1	Y_2	$\cdots$	Y_n

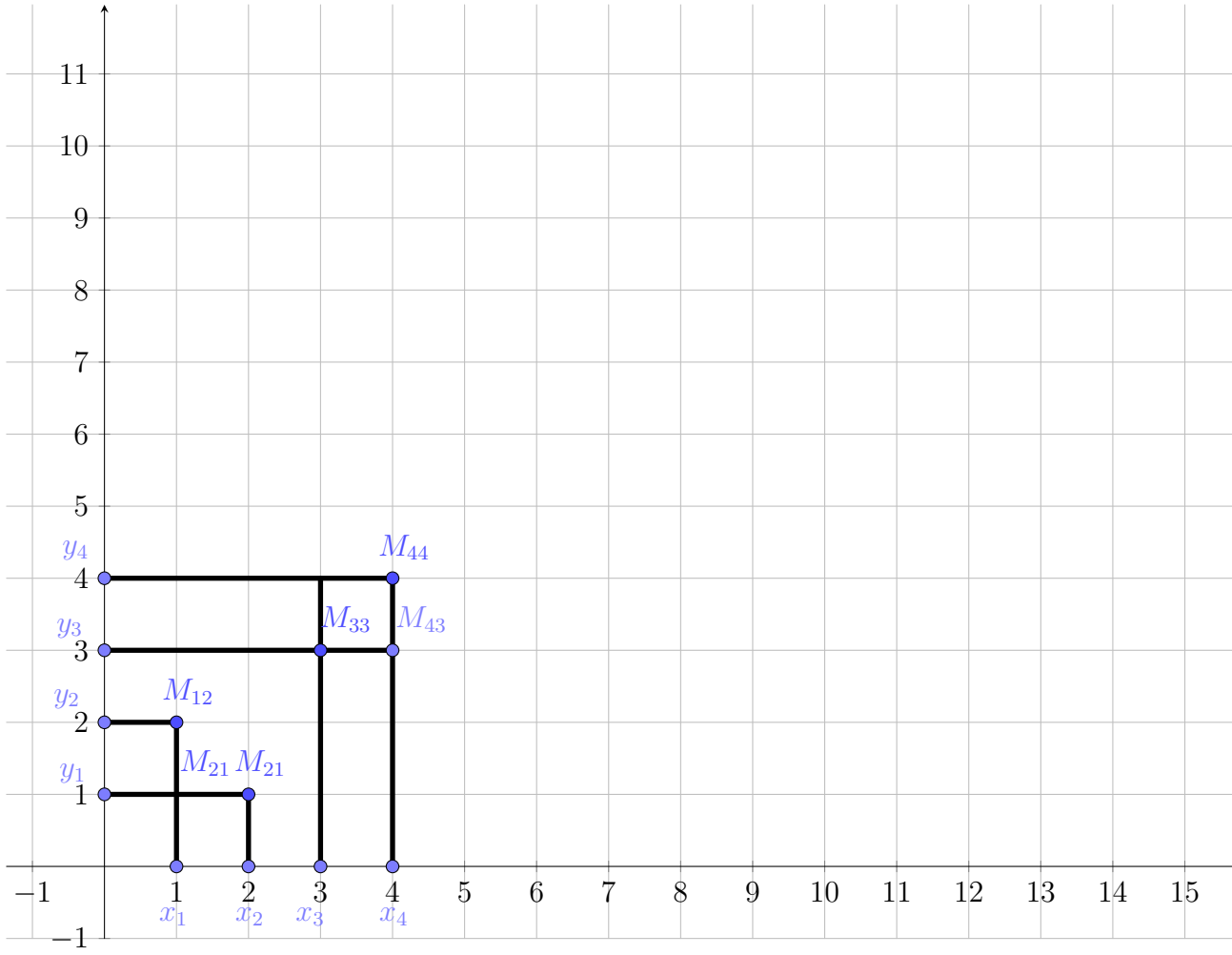
Dans ce tableau chaque couple (x_i, y_i) a pour effectif 1
Pour une série statistique injective on a :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} x_i \quad ; \quad (\sigma_X)^2 = \frac{1}{n}\sum_{i=1}^{p}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{p}(x_i^2 - \overline{x}^2)$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{p} y_i \quad ; \quad (\sigma_Y)^2 = \frac{1}{n}\sum_{i=1}^{p}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{p}(y_i^2 - \overline{y}^2)$$

Représentation graphique d'une série double

Dans un repère orthogonal, on porte en abscisse x_i et en ordonnées les y_j, les points $M_{ij}(x_i; y_j)$ auront un coefficient égal à n_{ij} et on note $M_{ij}(n_{ij})$



L'ensemble des points $M_{ij}(n_{ij})$ est appelé **nuage de points** de la série double $\{(x_i, y_j, n_{ij})\}$

Le barycentre des points pondérés (M_{ij}, n_{ij}) est appelé **point moyen** du nuage de point $G(,)$. Cas d'une série injective (x_i, y_i), le nuage de point est formé des points $M_i(x_i, y_i)$

Ajustement linéaire

• **la méthode des moindres carrés** Lors que les points M_i du nuage sont alignés on dit qu'il y'a une corrélation parfaite entre les grandeurs X et Y : ces deux grandeurs varient dans le même sens ou dans le sens contraire ;

il existe alors une relation affine du type $y = ax + b$ entre les valeurs de X et de Y permettant de prévoir l'une des grandeurs à partir de l'autre. Lors que les points M_i ne sont pas alignés, on essaie de trouver une droite (D) « la moins éloignée possible » des points Mi du nuage ; on dit qu'on fait un ajustement linéaire et la droite (D) sera appelée droite d'ajustement. Ceci se fait lors que les points du nuage semblent être distribués autour d'une droite.

Droite de régression$D_{Y/X}$

Graphe

Covariance

La covariance d'une série double (X, Y) est le réel noté $cov(X, Y)$ ou défini par σ_{XY}

$$\sigma_{XY} = \frac{n_{11}x_1y_1 + n_{12}x_1y_2 + \cdots}{N} = \frac{1}{N}\sum_{i=1}^{n} n_{ij}x_iy_j - \overline{xy}$$

Si la série est injective on aura

$$\sigma_{XY} = \frac{x_1y_1 + x_2y_2 + \cdots}{N} = \frac{1}{N}\sum_{i=1}^{n} x_iy_j - \overline{xy} = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

La droite $D_{Y/X}$ **a pour équation** $D_{Y/X}$ **:** $ax + (\overline{y} - a\overline{x}$ **) Soit** $D_{Y/X}$: $y - \overline{y} = a(x - \overline{x})$ **avec** $a = \frac{\sigma_{XY}}{\sigma_X^2}$ **ou** $y = ax + b$ **avec** $b = y - a\overline{x}$

Droite de régression de X **en** Y **:** $D_{X/Y}$

La droite $D_{X/Y}$: $x - \overline{x} = \alpha(y - \overline{y})$ ou $x = \alpha y + \beta$ avec $\alpha = \frac{\sigma_{XY}}{\sigma_Y^2}$ et $\beta = \overline{x} - \beta\overline{y}$

• **Méthode de MAYER**

Elle consiste à partager la série statistique double en deux séries doubles et de déterminer les points moyens $G1$ et $G2$ de chaque série.

La droite $(G1G2)$ est appelée droite d'ajustement de MAYER

Exemple

On a fait uine enquête sur cent familles portant sur le nombre d'enfants par famille (X) et le nombre de piéces d'habitation par famille (Y). Les résultats sont donnés dans le tableau ci-dessous :

$Y \setminus X$	0	1	2	3	4	5	Y
1	8	2	1	0	0	0	
2	3	11	10	5	1	0	
3	1	3	16	13	4	1	
4	0	1	3	5	8	4	
X							

1 .A l'aide de ce tableau reconstituer les séries statistiques marginales.
2 .On partage la série statistique en deux séries statistiques doubles $S1$ et $S2$

S1

$Y \setminus X$	0	1	2	Y
1	8	2	1	
2	3	11	10	
3	1	3	16	
4	0	1	3	
X				

S2

$Y \setminus X$	3	4	5	Y
1	0	0	0	
2	5	1	0	
3	13	4	1	
4	5	8	4	
X				

3 . déterminer leur point moyen respectif $G1$ et $G2$
4 . déterminer une équation de la droite $(G1G2)$
5 . estimer le nombre de pièces d'habitation d'une famille ayant 6 enfants

Coefficient de corrélation linéaire :

Le réel $r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ avec σ_X et σ_X non nuls est appelé coefficient de corrélation linéaire de X et de Y.

On a $a = \frac{r\sigma_Y}{\sigma_Y}$ et $\alpha = \frac{r\sigma_Y}{\sigma_Y}$

Remarque :

• Lors que r<0, X et Y varient en sens inverse, on dit qu'il y'a une corrélation négative entre X et Y.

• Lors que r<0 X et Y varient dans le même sens, on dit qu'il y'a une corrélation positive entre X et Y.

• Si r = 0, on ne peut trouver aucune relation affine entre X et Y ; on dit que les caractères X et Y sont indépendants.

• Si r est proche de 1 en valeur absolue on dit qu'il y'a une bonne corrélation

• Si r est proche de 0, on dit que la corrélation est faible.

Exercice 1

Dans une classe de première S de 20 élèves, on a relevé les poids (X) en kg et tailles (Y) en cm des élèves.

X (kg)	54	63	60	58	63	83	83	83	72	68
Y (cm)	165	183	178	168	175	180	173	180	173	179

TABLE 1 – Poids et tailles des élèves (suite)

X (kg)	66	72	66	67	72	68	83	76	76	68
Y (cm)	175	180	168	171	173	178	183	168	173	178

1. Dresser le tableau de contingence.
2. Calculer σ_x, σ_y et σ_{xy}.
3. En déduire le coefficient de corrélation.
4. (a) On s'intéresse maintenant à tous les élèves dont le poids est 83 kg. Combien y en a-t-il ?
   (b) Calculer les fréquences : $f(173/83)$ et $f(165/83)$.
5. (a) On s'intéresse maintenant à tous les élèves dont la taille est 180 cm. Combien y en a-t-il ?
   (b) Calculer les fréquences $f(54/180)$ et $f(58/180)$.

Exercice 2

Une entreprise a mis au point un nouveau produit et cherche à en fixer le prix de vente. Une enquête est réalisée auprès des clients potentiels ; les résultats sont donnés dans le tableau suivant où les y_i représentent le nombre d'exemplaires du produit que les clients sont disposés à acheter si le prix de vente, exprimé en milliers de francs, est x_i.

x_i	60	80	100	120	140	160	180	200
y_i	952	805	630	522	510	324	205	84

Questions

1. Calculer le coefficient de corrélation linéaire de x et y. La valeur trouvée justifie-t-elle la recherche d'un ajustement linéaire ?
2. Déterminer l'équation de la droite de régression de y en x.