

CH II – LES SÉRIES BIDIMENSIONNELLES À DONNÉES GROUPEES

OBJECTIFS :

- Les distributions marginales de x et de y
- Les distributions conditionnelles de x selon y et de y selon x
- Relations entre caractéristiques marginales et conditionnelles
- Le calcul de la covariance

Une série ou distribution bidimensionnelle s'obtient en croisant des modalités de deux variables x et y : on parle de « tri croisé ». A partir d'une série bidimensionnelle, on peut extraire différentes séries à une variable ou séries unidimensionnelles :

- Les deux distributions marginales de x et de y
- Les distributions conditionnelles de x selon y (c'est-à-dire de x avec une condition sur y) et de y selon x (c'est-à-dire de y avec une condition sur x)

I – La présentation des données

1 Le tableau à double entrée

Ce tableau fournit les effectifs obtenus lors du croisement des modalités des deux variables x et y .

		colonne $j =$					
		$j=1$	...	j	...	$j=q$	
ligne $i =$	y	y_1	...	y_j	...	y_q	TOTAL
x							
$i=1$	x_1	n_{11}	...	n_{1j}	...	n_{1q}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	n_{i1}	...	n_{ij}	...	n_{iq}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$i=p$	x_p	n_{p1}	...	n_{pj}	...	n_{pq}	$n_{p\bullet}$
	TOTAL	$n_{\bullet 1}$	...	$n_{\bullet j}$	...	$n_{\bullet q}$	$N = n_{\bullet\bullet}$

- Numérotation des lignes et des colonnes : on numérote les lignes de $i = 1$ à $i = p$ et les colonnes de $j = 1$ à $j = q$
- Modalités ou classes : les modalités $x_1, x_2, \dots, x_i, \dots, x_p$ du caractère x sont en ligne et les modalités $y_1, y_2, \dots, y_j, \dots, y_q$ du caractère y sont en colonne.

2 Les différents effectifs

➡ Effectif partiel

n_{ij} C'est l'effectif associé simultanément à la modalité (ou centre de classe) x_i de la variable x et à la modalité (ou centre de classe) y_j de la variable y .

Il est situé à l'intersection de la i ème ligne et de la j ème colonne.

➡ Effectifs marginaux

Ils sont ainsi appelés car ils apparaissent dans les *marges* du tableau à double entrée.

$n_{i\bullet}$ C'est l'effectif associé à la modalité (ou centre de classe) x_i de la variable x .
Il s'obtient comme le total des effectifs partiels de la ligne n° i : $n_{i\bullet} = \sum_j n_{ij}$

$n_{\bullet j}$ C'est l'effectif associé à la modalité (ou centre de classe) y_j de la variable y .
Il s'obtient comme le total des effectifs partiels de la colonne n° j : $n_{\bullet j} = \sum_i n_{ij}$

➡ Effectif total

$n_{\bullet\bullet}$ C'est l'effectif total ou l'effectif de la population étudiée.

3 Exemple

- 1. Numéroté les lignes et les colonnes du tableau à double entrée situé ci-dessous. Déterminer les centres de classe x_i du caractère x . Compléter alors le tableau.
- 2. Compléter les égalités ci-dessous.

Solution

- 1. Répartition de 400 individus selon l'âge x et le statut matrimonial y

			j = 1	j = 2	j = 3	j = 4	
			Statut matrimonial y				
			Célibataire	Marié(e)	Divorcé(e)	Veuf(ve)	$n_{i\bullet}$
Age x	x_i						
i = 1	[25 ; 35 [	30 ans	50	30	30	0	110
i = 2	[35 ; 45 [	40 ans	20	50	60	5	135
i = 3	[45 ; 55 [	50 ans	5	20	40	25	90
i = 4	[55 ; 65 [	60 ans	5	10	10	40	65
$n_{\bullet j}$			80	110	140	70	$N = n_{\bullet\bullet} = 400$

- 2. Compléter les égalités ci-dessous.
 $x_1 = 30 \text{ ans}$; $y_3 = \text{Divorcé(e)}$; $x_4 = 60 \text{ ans}$; $y_1 = \text{Célibataire}$
 $n_{32} = 20$; $n_{23} = 60$; $n_{43} = 10$; $n_{14} = 0$
 $n_{\bullet 4} = 70$; $n_{2\bullet} = 135$; $n_{\bullet 1} = 80$; $n_{\bullet\bullet} = 400$

II – Les deux distributions marginales

On les appelle ainsi car elles apparaissent dans les *marges* du tableau à double entrée.

Il s'agit, en fait :

- ➡ de la distribution de la variable x sur *la population étudiée, indépendamment* de la variable y .
- ➡ de la distribution de la variable y sur *la population étudiée, indépendamment* de la variable x .

1 La distribution marginale de x

a) Position dans le tableau à double entrée

Les effectifs $n_{i\bullet}$ associés aux modalités x_i sont situés dans la *marge verticale* du tableau à double entrée.

$y \backslash x$	y_1	y_2	...	y_j	...	y_q	TOTAL
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2\bullet}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i\bullet}$
...	...	...	...	...	...	...	...
x_p	n_{p1}	n_{p2}	...	n_{pj}	...	n_{pq}	$n_{p\bullet}$
TOTAL	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet j}$	...	$n_{\bullet q}$	$N = n_{\bullet\bullet}$

b) Éléments caractéristiques

- ➡ Modalités : x_i
- ➡ Effectif associé : $n_{i\bullet}$
- ➡ Effectif total : $N = n_{\bullet\bullet}$

A partir de ces éléments, on pourra calculer :

Fréquence marginale de la modalité x_i

$$f_{i\bullet} (\%) = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \times 100 \quad \text{C'est la part (en \%), dans la population, de la modalité } x_i \text{ du caractère } x.$$

Caractéristiques marginales de x

- ➡ Moyenne :
$$\bar{\bar{x}} = \frac{1}{n_{\bullet\bullet}} \left[\sum_i n_{i\bullet} x_i \right]$$
- ➡ Variance :
$$V(x) = \frac{1}{n_{\bullet\bullet}} \left[\sum_i n_{i\bullet} (x_i - \bar{\bar{x}})^2 \right] \quad (\text{formule non développée : démonstrations})$$

$$V(x) = \frac{1}{n_{\bullet\bullet}} \left[\sum_i n_{i\bullet} x_i^2 \right] - \bar{\bar{x}}^2 \quad (\text{formule développée : calculs})$$
- ➡ Ecart-type :
$$\sigma(x) = \sqrt{V(x)}$$

2 La distribution marginale de y

a) Position dans le tableau à double entrée

Les effectifs $n_{\bullet j}$ associés aux modalités y_j sont situés dans la *marge horizontale* du tableau.

x \ y	y						TOTAL
	y_1	y_2	...	y_j	...	y_q	
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$n_{2\bullet}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$n_{i\bullet}$
...	...	...	...	...	...	...	...
x_p	n_{p1}	n_{p2}	...	n_{pj}	...	n_{pq}	$n_{p\bullet}$
TOTAL	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet j}$	...	$n_{\bullet q}$	$N = n_{\bullet \bullet}$

b) Éléments caractéristiques

➡ Modalités : y_j

➡ Effectif associé : $n_{\bullet j}$

➡ Effectif total : $N = n_{\bullet \bullet}$

A partir de ces éléments, on pourra calculer :

Fréquence marginale de la modalité y_j

$$f_{\bullet j} (\%) = \frac{n_{\bullet j}}{n_{\bullet \bullet}} \times 100 \quad \text{C'est la part (en \%), dans la population, de la modalité } y_j \text{ du caractère } y.$$

Caractéristiques marginales de y

➡ Moyenne :

$$\bar{\bar{y}} = \frac{1}{n_{\bullet \bullet}} \left[\sum_j n_{\bullet j} y_j \right]$$

➡ Variance :

$$V(y) = \frac{1}{n_{\bullet \bullet}} \left[\sum_j n_{\bullet j} (y_j - \bar{\bar{y}})^2 \right] \quad (\text{formule non développée : démonstrations})$$

$$V(y) = \frac{1}{n_{\bullet \bullet}} \left[\sum_j n_{\bullet j} y_j^2 \right] - \bar{\bar{y}}^2 \quad (\text{formule développée : calculs})$$

➡ Ecart-type :

$$\sigma(y) = \sqrt{V(y)}$$

3 Intégration des calculs dans le tableau à double entrée

Caractéristiques marginales de x

Caractéristiques marginales de y

x \ y	y	Caractéristiques marginales de x		
	y_j	$n_{i\bullet}$	$n_{i\bullet} x_i$	$n_{i\bullet} x_i^2$
x_i	$\vdots$ $\dots n_{ij} \dots$ $\vdots$	$n_{i\bullet}$	$n_{i\bullet} x_i$	$n_{i\bullet} x_i^2$
$n_{\bullet j}$	$n_{\bullet j}$	$N = n_{\bullet\bullet}$	$\Sigma =$	$\Sigma =$
$n_{\bullet j} y_j$	$n_{\bullet j} y_j$	$\Sigma =$		
$n_{\bullet j} y_j^2$	$n_{\bullet j} y_j^2$	$\Sigma =$		

4 Exemple Caractère quantitatif continu et caractère quantitatif discret

Le tableau statistique situé ci-dessous relate la répartition de 400 individus d'une population selon l'âge x et le nombre d'enfants y.

- Détailler le calcul de **A**, **B**, **C**, **D**, **E** et **F**. Compléter alors le tableau statistique (?).
- Déterminer les caractéristiques marginales (moyenne, variance et écart-type) de l'âge x.
- Déterminer les caractéristiques marginales (moyenne, variance et écart-type) du nombre d'enfants y.

Solution

Nbre d'enfants y_j		0	1	2	3	$n_{i\bullet}$	$n_{i\bullet} x_i$	$n_{i\bullet} x_i^2$
Age x_i								
[25 ; 35 [	30	10	10	30	0	A = 50	B = 1 500	C = 45 000
[35 ; 45 [	40	30	40	150	30	250	10 000	400 000
[45 ; 55 [	50	10	50	20	20	100	5 000	250 000
$n_{\bullet j}$		50	100	D = 200	50	? 400	? 16 500	? 695 000
$n_{\bullet j} y_j$		0	100	E = 400	150	? 650		
$n_{\bullet j} y_j^2$		0	100	F = 800	450	? 1 350		

1. Calcul détaillé de A, B, C, D, E, F et G :

$$A = 10 + 10 + 30 + 0 = \boxed{50}$$

$$B = 50 \times 30 = \boxed{1500}$$

$$C = 50 \times 30^2 = \boxed{45\,000}$$

$$D = 30 + 150 + 20 = \boxed{200}$$

$$E = 200 \times 2 = \boxed{400}$$

$$F = 200 \times 2^2 = \boxed{800}$$

(?) : Cf. tableau statistique

2. Caractéristiques marginales de l'âge x

➔ Moyenne : $\bar{\bar{x}} = \frac{1}{n_{..}} \left[\sum_i n_{i.} x_i \right] = \frac{1}{400} \times 16\,500 = \boxed{41,25 \text{ ans}}$

➤ Variance : $V(x) = \frac{1}{n_{..}} \left[\sum_i n_{i.} x_i^2 \right] - \bar{\bar{x}}^2 = \frac{1}{400} \times 695\,000 - 41,25^2 = \boxed{35,9375}$

➡ Ecart-type : $\sigma(x) = \sqrt{V(x)} = \sqrt{35,9375} \approx \boxed{5,99 \text{ ans}}$

- ➡ Conclusion : ✓ Les 400 individus étudiés ont un âge moyen de 41,25 ans.
 ✓ L'écart-type, d'environ 6 ans, mesure la dispersion de l'âge par rapport à leur âge moyen 41,25 ans.

3. Caractéristiques marginales du nombre d'enfants y :

➔ Moyenne : $\bar{\bar{y}} = \frac{1}{n_{..}} \left[\sum_j n_{.j} y_j \right] = \frac{1}{400} \times 650 = \boxed{1,625 \text{ enfants}}$

➤ Variance : $V(y) = \frac{1}{n_{..}} \left[\sum_j n_{.j} y_j^2 \right] - \bar{\bar{y}}^2 = \frac{1}{400} \times 1\,350 - 1,625^2 = \boxed{0,734475}$

➡ Ecart-type : $\sigma(y) = \sqrt{V(y)} = \sqrt{0,734475} \approx \boxed{0,86 \text{ enfants}}$

- ➡ Conclusion : ✓ Les 400 individus étudiés ont, en moyenne, 1,625 enfants.
 ✓ L'écart-type, d'environ 0,86 enfants, mesure la dispersion du nombre d'enfants par rapport au nombre moyen d'enfants, soit 1,625 enfants.

III – Les distributions conditionnelles

Les effectifs associés à ces distributions apparaissent *à l'intérieur* du tableau à double entrée.

Il s'agit, en fait :

- ➡ de la distribution de la variable x *avec une condition* sur la variable y : $y = y_j$ (ou selon $j = \dots$)
- ➡ de la distribution de la variable y *avec une condition* sur la variable x : $x = x_i$ (ou selon $i = \dots$)

1 La distribution conditionnelle de x selon $y = y_j$ ou selon $j = \dots$

a) Position dans le tableau à double entrée

Les effectifs n_{ij} associés aux modalités x_i sont situés *à l'intérieur* du tableau à double entrée.

Condition							
y x	y ₁	y ₂	...	y _j	...	y _q	TOTAL
x₁	n ₁₁	n ₁₂	...	n_{1j}	...	n _{1q}	n _{1•}
x₂	n ₂₁	n ₂₂	...	n_{2j}	...	n _{2q}	n _{2•}
...	...	...	...	...	...	...	...
x_i	n _{i1}	n _{i2}	...	n_{ij}	...	n _{iq}	n _{i•}
...	...	...	...	...	...	...	...
x_p	n _{p1}	n _{p2}	...	n_{pj}	...	n _{pq}	n _{p•}
TOTAL	n _{•1}	n _{•2}	...	n_{•j}	...	n _{•q}	N = n _{••}

b) Éléments caractéristiques

- ➡ Modalités : x_i
- ➡ Effectif associé : n_{ij} (j fixé)
- ➡ Effectif total : $n_{•j}$

A partir de ces éléments, on pourra calculer :

Fréquence conditionnelle de la modalité x_i selon $y = y_j$ ou selon $j = \dots$

$$f_{i/j} (\%) = \frac{n_{ij}}{n_{•j}} \times 100 \quad \text{C'est la part (en \%), parmi les individus tels que } y = y_j, \text{ de la modalité } x_i \text{ de } x$$

Caractéristiques conditionnelles de x selon $y = y_j$ ou selon $j = \dots$

- ➡ Moyenne :
$$\bar{x}_j = \frac{1}{n_{•j}} \left[\sum_i n_{ij} x_i \right]$$
- ➡ Variance :
$$V_j(x) = \frac{1}{n_{•j}} \left[\sum_i n_{ij} (x_i - \bar{x}_j)^2 \right] \quad \text{(formule non développée : démonstrations)}$$

$$V_j(x) = \frac{1}{n_{•j}} \left[\sum_i n_{ij} x_i^2 \right] - \bar{x}_j^2 \quad \text{(formule développée : calculs)}$$
- ➡ Ecart-type :
$$\sigma_j(x) = \sqrt{V_j(x)}$$

2 La distribution conditionnelle de y selon x = x_i ou selon i = ...

a) Position dans le tableau à double entrée

Les effectifs n_{ij} associés aux modalités y_j sont situés *à l'intérieur* du tableau à double entrée.

x \ y	y ₁	y ₂	...	y _j	...	y _q	TOTAL
x ₁	n ₁₁	n ₁₂	...	n _{1j}	...	n _{1q}	n _{1•}
x ₂	n ₂₁	n ₂₂	...	n _{2j}	...	n _{2q}	n _{2•}
...	...	...	...	...	...	...	...
x _i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	n_{i•}
...	...	...	...	...	...	...	...
x _p	n _{p1}	n _{p2}	...	n _{pj}	...	n _{pq}	n _{p•}
TOTAL	n _{•1}	n _{•2}	...	n _{•j}	...	n _{•q}	N = n _{••}

Condition {

b) Éléments caractéristiques

- ➡ Modalités : y_j
- ➡ Effectif associé : n_{ij} (i fixé)
- ➡ Effectif total : $n_{i•}$

A partir de ces éléments, on pourra calculer :

Fréquence conditionnelle de la modalité y_j selon $x = x_i$ ou selon $i = ...$

$$f_{j/i} (\%) = \frac{n_{ij}}{n_{i•}} \times 100$$

C'est la part (en %), parmi les individus tels que $x = x_i$, de la modalité y_j de y

Caractéristiques conditionnelles de y selon $x = x_i$ ou selon $i = ...$

- ➡ Moyenne :

$$\bar{y}_i = \frac{1}{n_{i•}} \left[\sum_j n_{ij} y_j \right]$$

- ➡ Variance :

$$V_i(y) = \frac{1}{n_{i•}} \left[\sum_j n_{ij} (y_j - \bar{y}_i)^2 \right] \quad (\text{formule non développée : démonstrations})$$

$$V_i(y) = \frac{1}{n_{i•}} \left[\sum_j n_{ij} y_j^2 \right] - \bar{y}_i^2$$

(formule développée : calculs)

- ➡ Ecart-type :

$$\sigma_i(y) = \sqrt{V_i(y)}$$

3 Intégration des calculs dans le tableau à double entrée

Pour les *distributions marginales*, tous les *calculs intermédiaires* figurent dans le tableau à double entrée, ce qui ne pose pas de problème car il n'y a que deux distributions marginales.

Pour les *distributions conditionnelles*, beaucoup plus nombreuses, on ne fait figurer, dans le tableau à double entrée, que les résultats des totaux, c'est-à-dire des *sommes*.

		Caractéristiques conditionnelles de y selon x = x _i			
			$\begin{array}{cc} \bullet & \bullet \\ \updownarrow + \updownarrow & \\ \bullet & \bullet \end{array}$	$\begin{array}{cc} \bullet^2 & \bullet^2 \\ \updownarrow + \updownarrow & \\ \bullet & \bullet \end{array}$	
	y	y _j	n _{i•}	$\sum_j n_{ij} y_j$	$\sum_j n_{ij} y_j^2$
x	x _i	$\begin{array}{c} \vdots \\ \dots n_{ij} \dots \\ \vdots \end{array}$	n _{i•}	$\sum_j n_{ij} y_j$	$\sum_j n_{ij} y_j^2$
	n _{•j}	n _{•j}	n _{••}	$\Sigma =$	$\Sigma =$
Caractéristiques conditionnelles de x selon y = y _j	$\begin{array}{ccc} \bullet & \leftrightarrow & \bullet \\ + & & \\ \bullet & \leftrightarrow & \bullet \end{array}$	$\sum_i n_{ij} x_i$	$\sum_i n_{ij} x_i$	$\Sigma =$	
	$\begin{array}{ccc} \bullet^2 & \leftrightarrow & \bullet^2 \\ + & & \\ \bullet^2 & \leftrightarrow & \bullet^2 \end{array}$	$\sum_i n_{ij} x_i^2$	$\sum_i n_{ij} x_i^2$	$\Sigma =$	

4 Exemple

Le tableau, situé page 10, reprend et complète les données de l'exemple (page 5)

1. Numéroté les classes de x (i = 1 à i = 3) et les modalités de y (j = 1 à j = 4).
2. Détailler le calcul de **A**, **B**, **C** et **D**.
Compléter alors le tableau statistique (?). Flécher les totaux qui sont identiques.
3. Déterminer le nombre moyen d'enfants ainsi que l'écart-type du nombre d'enfants des individus âgés entre 25 et 35 ans.
4. Déterminer les caractéristiques conditionnelles de l'âge selon le nombre d'enfants :
 - détailler avec précision les calculs pour j = 1
 - achever les calculs et intégrer vos résultats dans le tableau récapitulatif

Solution

1. Cf. tableau statistique.
2. Calcul détaillé de **A**, **B**, **C** et **D** : Cf. tableau statistique.
(?) et fléchage des totaux identiques : Cf. tableau statistique.
3. Analyse de la question : on veut déterminer le nombre moyen d'enfants, ainsi que l'écart-type du nombre d'enfants des individus de cette population âgés entre 25 et 35 ans. Par suite :
la variable : c'est le nombre d'enfants y car on veut en calculer la moyenne et l'écart-type.
la condition : l'individu doit être âgé entre 25 et 35 ans : la condition s'écrit : $i = 1$

Résolution : les résultats numériques utiles aux calculs se trouvent tous sur la ligne $i = 1$

➡ Moyenne : $\bar{y}_1 = \frac{1}{n_{1\bullet}} \left[\sum_j n_{1j} y_j \right] = \frac{1}{50} \times 70 = \boxed{1,4 \text{ enfants}}$

➡ Variance : $V_1(y) = \frac{1}{n_{1\bullet}} \left[\sum_j n_{1j} y_j^2 \right] - \bar{y}_1^2 = \frac{1}{50} \times 130 - 1,4^2 = \boxed{0,64}$

➡ Ecart-type : $\sigma_1(y) = \sqrt{V_1(y)} = \boxed{0,8 \text{ enfant}}$

- ➡ Conclusion :
✓ Les 50 individus âgés entre 25 et 35 ans ont, en moyenne, 1,4 enfants.
✓ L'écart-type, de 0,8 enfant, mesure la dispersion du nombre d'enfants par rapport au nombre moyen d'enfants, soit 1,4.

4. Analyse de la question : on veut déterminer les caractéristiques conditionnelles de l'âge selon le nombre d'enfants, sans préciser celui-ci : il faut donc envisager les différentes valeurs possibles du nombre d'enfants. Par suite :

la variable : c'est l'âge x car on veut en calculer les caractéristiques.

la condition : il faut envisager les différentes valeurs possibles du nombre d'enfants :

$y = 0 \text{ enfant } (j = 1) ; y = 1 \text{ enfant } (j = 2) ; y = 2 \text{ enfants } (j = 3) ; y = 3 \text{ enfants } (j = 4)$

Résolution : on rédige avec précision le cas : $j = 1$ et on résume tous les calculs dans un tableau.

Cas $j = 1$: les résultats numériques utiles aux calculs se trouvent tous dans la colonne $j = 1$

➡ Moyenne : $\bar{x}_1 = \frac{1}{n_{\bullet 1}} \left[\sum_i n_{i1} x_i \right] = \frac{1}{50} \times 2\,000 = \boxed{40 \text{ ans}}$

➡ Variance : $V_1(x) = \frac{1}{n_{\bullet 1}} \left[\sum_i n_{i1} x_i^2 \right] - \bar{x}_1^2 = \frac{1}{50} \times 82\,000 - 40^2 = \boxed{40}$

➡ Ecart-type : $\sigma_1(x) = \sqrt{V_1(x)} \approx \boxed{6,32 \text{ ans}}$

- ➡ Conclusion :
✓ Les 50 individus n'ayant aucun enfant ont un âge moyen de 40 ans.
✓ L'écart-type, de 6,32 ans, mesure la dispersion de l'âge par rapport à l'âge moyen de 40 ans.

Autres cas : à effectuer directement à la calculatrice en faisant varier les valeurs de j de 2 à 4

Tableau récapitulatif (à compléter)

j	Effectif $n_{\bullet j}$	Moyenne $\bar{x}_j$	Variance $V_j(x)$	Ecart-type $\sigma_j(x)$ (à 10^{-2} près)
1	? 50	? 40 ans	? 40	? $\approx 6,32$ ans
2	? 100	? 44 ans	? 44	? $\approx 6,63$ ans
3	? 200	? 39,5 ans	? 24,75	? $\approx 4,97$ ans
4	? 50	? 44 ans	? 24	? $\approx 4,90$ ans
$n_{\bullet\bullet} = 400$				

IV – Relations entre les caractéristiques marginales et conditionnelles

1 Cas de la variable x

On subdivise la population P étudiée en q sous-populations :

sous-population	Dénomination	Effectif	Moyenne	Variance
P_1	ensemble des individus tels que : $y = y_1$	$n_{\bullet 1}$	$\bar{x}_1$	$V_1(x)$
P_2	ensemble des individus tels que : $y = y_2$	$n_{\bullet 2}$	$\bar{x}_2$	$V_2(x)$
...		...	...	...
P_j	ensemble des individus tels que : $y = y_j$	$n_{\bullet j}$	$\bar{x}_j$	$V_j(x)$
...		...	...	...
P_q	ensemble des individus tels que : $y = y_q$	$n_{\bullet q}$	$\bar{x}_q$	$V_q(x)$
		$n_{\bullet\bullet}$		

On admettra les résultats suivants, basés sur l'étude des *mélanges de populations*. On a :

➔ Moyennede x dans P = moyenne des moyennes de x dans les sous - populations P_j

Ici : Moyenne marginale de x = moyenne des moyennes conditionnelles de x selon $y = y_j$

$$\text{D'où : } \bar{\bar{x}} = \frac{1}{n_{\bullet\bullet}} \left[\sum_j n_{\bullet j} \bar{x}_j \right]$$

➔ Variance de x dans P = moyenne des variances de x dans P_j + variance des moyennes de x dans P_j

Ici : Variance marginale de x = moyenne des variances conditionnelles + variance des moyennes conditionnelles

On note : $V(x) = MDV + VDM$

Ou aussi : $V(x) = \overline{V_j(x)} + V(\bar{x}_j)$

On rappelle que :

$$\blacksquare MDV = \overline{V_j(x)} = \frac{1}{n_{\bullet\bullet}} \left[\sum_j n_{\bullet j} V_j(x) \right]$$

$$\blacksquare VDM = V(\bar{x}_j) = \frac{1}{n_{\bullet\bullet}} \left[\sum_j n_{\bullet j} \bar{x}_j^2 \right] - \bar{\bar{x}}^2$$

On aurait pu effectuer un raisonnement identique pour la variable y, en considérant les p sous-populations P_i regroupant les individus tels que : $x = x_i$

2 Propriétés

➔ Pour la variable x :

☑ **Moyenne** : $\bar{\bar{x}} = \frac{1}{n_{\bullet\bullet}} \left[\sum_j n_{\bullet j} \bar{x}_j \right]$

☑ **Variance** : $V(x) = MDV + VDM = \overline{V_j(x)} + V(\bar{x}_j)$

Avec :

$$MDV = \overline{V_j(x)} = \frac{1}{n_{\bullet\bullet}} \left[\sum_j n_{\bullet j} V_j(x) \right] \quad VDM = V(\bar{x}_j) = \frac{1}{n_{\bullet\bullet}} \left[\sum_j n_{\bullet j} \bar{x}_j^2 \right] - \bar{\bar{x}}^2$$

➔ Pour la variable y :

☑ **Moyenne** : $\bar{\bar{y}} = \frac{1}{n_{\bullet\bullet}} \left[\sum_i n_{i\bullet} \bar{y}_i \right]$

☑ **Variance** : $V(y) = MDV + VDM = \overline{V_i(y)} + V(\bar{y}_i)$

Avec :

$$MDV = \overline{V_i(y)} = \frac{1}{n_{\bullet\bullet}} \left[\sum_i n_{i\bullet} V_i(y) \right] \quad VDM = V(\bar{y}_i) = \frac{1}{n_{\bullet\bullet}} \left[\sum_i n_{i\bullet} \bar{y}_i^2 \right] - \bar{\bar{y}}^2$$

3 Exemple

Le tableau suivant reprend et complète les résultats obtenus en haut de la page 12.

j	Effectif $n_{\bullet j}$	Moyenne $\bar{x}_j$	Variance $V_j(x)$	$n_{\bullet j} \bar{x}_j$	$n_{\bullet j} \bar{x}_j^2$	$n_{\bullet j} V_j(x)$
1	50	40 ans	40	? 2 000	? 80 000	? 2 000
2	100	44 ans	44	4 400	193 600	4 400
3	200	39,5 ans	24,75	7 900	312 050	4 950
4	50	44 ans	24	2 200	96 800	1 200
	$n_{\bullet\bullet} = 400$			? 16 500	? 682 450	? 12 550

1. Compléter le tableau de calculs (?).
2. Calculer les caractéristiques marginales (moyenne et variance) de l'âge x . Conclusion ?

Solution

1. Cf. tableau de calculs (?).
2. Caractéristiques marginales de l'âge x :

➔ Moyenne : $\bar{\bar{x}} = \frac{1}{n_{\bullet\bullet}} \left[\sum_j n_{\bullet j} \bar{x}_j \right]$

$$\bar{\bar{x}} = \frac{1}{400} \times 16500$$

$$\bar{\bar{x}} = \boxed{41,25 \text{ ans}}$$

➡ Variance :

Dans un mélange de populations, on a : $V(x) = MDV + VDM$

Calcul de la MDV :

$$MDV = \overline{V_j(x)}$$

$$MDV = \frac{1}{n_{..}} \left[\sum_j n_{.j} V_j(x) \right]$$

$$MDV = \frac{1}{400} \times 12\,550$$

$$MDV = \boxed{31,375}$$

Calcul de la VDM :

$$VDM = V(\bar{x}_j)$$

$$VDM = \frac{1}{n_{..}} \left[\sum_j n_{.j} \bar{x}_j^2 \right] - \bar{\bar{x}}^2$$

$$VDM = \frac{1}{400} \times 682\,450 - 41,25^2$$

$$VDM = \boxed{4,5625}$$

$$\text{Finalement : } V(x) = MDV + VDM = 31,375 + 4,5625 = \boxed{35,9375}$$

➡ Conclusion :

On a : $\bar{\bar{x}} = 41,25$ ans et $V(x) = 35,9375$: ces deux résultats avaient déjà été obtenus précédemment page 6, d'une autre manière.

4 Conclusion : les différents modes de calcul des caractéristiques marginales

a) Les trois méthodes de calcul des caractéristiques marginales

Le calcul des *caractéristiques conditionnelles de x et de y*, on a pu le constater, s'effectue d'une seule façon possible.

En ce qui concerne les *caractéristiques marginales de x et de y*, il en va autrement.

Ainsi, il découle des différents résultats obtenus jusqu'à présent que l'on peut exprimer et donc déterminer les *caractéristiques marginales de x et de y* sous trois formes distinctes :

➡ Soit avec le seul indice i

➡ Soit avec les deux indices i et j

➡ Soit avec le seul indice j

On peut caractériser les méthodes de calculs de la façon suivante :

➡ 1^{ère} méthode : « *série à une variable* » faisant référence au semestre 1

➡ 2^{nde} méthode : « *à deux indices* », fondée sur les égalités de sommes (totaux identiques et vérification des calculs : Cf. exemple page 10)

➡ 3^{ème} méthode : « *mélange de populations* », basée sur la décomposition de la population P en q sous-populations pour la variable x et p sous-populations pour la variable y.

b) Cas de la variable x

Méthode	« série à une variable »	« à deux indices »	« mélange de populations »
Moyenne	$\bar{\bar{x}} = \frac{1}{n_{..}} \left[\sum_i n_{i.} x_i \right]$ un seul indice : i	$\bar{\bar{x}} = \frac{1}{n_{..}} \left[\sum_{i,j} n_{ij} x_i \right]$ deux indices : i et j	$\bar{\bar{x}} = \frac{1}{n_{..}} \left[\sum_j n_{.j} \bar{x}_j \right]$ un seul indice : j
Variance	$V(x) = \frac{1}{n_{..}} \left[\sum_i n_{i.} x_i^2 \right] - \bar{\bar{x}}^2$ un seul indice : i	$V(x) = \frac{1}{n_{..}} \left[\sum_{i,j} n_{ij} x_i^2 \right] - \bar{\bar{x}}^2$ deux indices : i et j	$V(x) = \overline{V_j(x)} + V(\bar{x}_j)$ un seul indice : j
Commentaire	c'est facile ; c'est la méthode la plus naturelle	c'est facile ; cette méthode équivaut en pratique à la précédente	c'est la méthode la plus difficile et la moins naturelle, surtout pour le calcul de la variance

c) Cas de la variable y

Méthode	« série à une variable »	« à deux indices »	« mélange de populations »
Moyenne	$\bar{\bar{y}} = \frac{1}{n_{..}} \left[\sum_j n_{.j} y_j \right]$ un seul indice : j	$\bar{\bar{y}} = \frac{1}{n_{..}} \left[\sum_{i,j} n_{ij} y_j \right]$ deux indices : i et j	$\bar{\bar{y}} = \frac{1}{n_{..}} \left[\sum_i n_{i.} \bar{y}_i \right]$ un seul indice : i
Variance	$V(y) = \frac{1}{n_{..}} \left[\sum_j n_{.j} y_j^2 \right] - \bar{\bar{y}}^2$ un seul indice : j	$V(y) = \frac{1}{n_{..}} \left[\sum_{i,j} n_{ij} y_j^2 \right] - \bar{\bar{y}}^2$ deux indices : i et j	$V(y) = \overline{V_i(y)} + V(\bar{y}_i)$ un seul indice : i
Commentaire	c'est facile ; c'est la méthode la plus naturelle	c'est assez facile ; cette méthode équivaut en pratique à la précédente	c'est la méthode la plus difficile et la moins naturelle, surtout pour le calcul de la variance

V – La covariance

1 Définition

La *covariance de x et de y* est le réel défini par :

$$\text{Cov}(x,y) = \frac{1}{n_{..}} \left[\sum_{i,j} n_{ij} (x_i - \bar{\bar{x}})(y_j - \bar{\bar{y}}) \right] \quad (\text{forme non développée de la covariance})$$

2 **Interprétation**

Contrairement à une variance, la **covariance de x et de y** peut être positive ou négative. Son signe donne une indication sur le sens de la liaison éventuelle entre x et y. On peut interpréter le tableau à double entrée de la façon suivante :

x \ y	y ₁	$\bar{y}$	y _q
x ₁	<u>Région I :</u> $\left. \begin{array}{l} x_i - \bar{x} < 0 \\ y_j - \bar{y} < 0 \end{array} \right\} \text{ par suite : } \\ n_{ij} (x_i - \bar{x})(y_j - \bar{y}) > 0$	<u>Région II :</u> $\left. \begin{array}{l} x_i - \bar{x} < 0 \\ y_j - \bar{y} > 0 \end{array} \right\} \text{ par suite : } \\ n_{ij} (x_i - \bar{x})(y_j - \bar{y}) < 0$	
$\bar{x}$	<u>Région III :</u> $\left. \begin{array}{l} x_i - \bar{x} > 0 \\ y_j - \bar{y} < 0 \end{array} \right\} \text{ par suite : } \\ n_{ij} (x_i - \bar{x})(y_j - \bar{y}) < 0$	<u>Région IV :</u> $\left. \begin{array}{l} x_i - \bar{x} > 0 \\ y_j - \bar{y} > 0 \end{array} \right\} \text{ par suite : } \\ n_{ij} (x_i - \bar{x})(y_j - \bar{y}) > 0$	
x _p			

- ➡ Si $Cov(x,y) > 0$: les couples (x_i , y_j) sont situés essentiellement dans les régions I et IV : les variables x et y varient globalement dans le même sens.
- ➡ Si $Cov(x,y) < 0$: les couples (x_i , y_j) sont situés essentiellement dans les régions II et III : les variables x et y varient globalement en sens inverse.

3 **Propriétés**

Soient x et y deux caractères quantitatifs quelconques. On a :

➡ Forme développée de la covariance :

$$Cov(x, y) = \frac{1}{n_{\bullet\bullet}} \left[\sum_{i,j} n_{ij} x_i y_j \right] - \bar{x} \bar{y}$$

➡ Changement de variable :

Pour a, b, c et d réels :

$$Cov(ax + b, cy + d) = a c Cov(x, y)$$

Pour a₁, a₂ réels,

$$Cov(a_1 x_1 + a_2 x_2, y) = a_1 Cov(x_1, y) + a_2 Cov(x_2, y)$$

➡ Symétrie :

$$Cov(x, y) = Cov(y, x)$$

➡ Lien avec la variance :

$$Cov(x, x) = V(x)$$

Remarques

- ➡ Comme pour la variance, la forme non développée de la **covariance** est en général utilisée pour les démonstrations, la forme développée pour les calculs numériques.
- ➡ Pour retenir l'expression de la forme développée de la **covariance**, on dit : « la **covariance** est égale à la moyenne des produits moins le produit des moyennes ».

4 Détermination pratique de la covariance de x et de y

$$\text{On a : } \text{Cov}(x,y) = \frac{1}{n_{\bullet\bullet}} \left[\sum_{i,j} n_{ij} x_i y_j \right] - \bar{x} \bar{y}$$

Problème : comment calcule-t-on $\sum_{i,j} n_{ij} x_i y_j$?

1^{ère} technique : *On effectue les calculs indiqués « à la main »*

Exemple

A partir du tableau à double entrée suivant, calculer : $\sum_{i,j} n_{ij} x_i y_j$

Nbre d'enfants y_j Age x_i		0	1	2	3	$n_{i\bullet}$
[25 ; 35 [	30	10	10	30	0	50
[35 ; 45 [	40	30	40	150	30	250
[45 ; 55 [	50	10	50	20	20	100
$n_{\bullet j}$		50	100	200	50	$n_{\bullet\bullet} = 400$

Solution

$$\sum_{i,j} n_{ij} x_i y_j = 10 \times 30 \times 0 + 10 \times 30 \times 1 + 30 \times 30 \times 2 + 0 \times 30 \times 3 + 30 \times 40 \times 0 + 40 \times 40 \times 1 + 150 \times 40 \times 2 +$$

$$30 \times 40 \times 3 + 10 \times 50 \times 0 + 50 \times 50 \times 1 + 20 \times 50 \times 2 + 20 \times 50 \times 3 = \boxed{26\,800}$$

On constate immédiatement que les calculs sont longs, avec un gros risque d'erreur.

2nde technique : On utilise des sommes déjà calculées pour les caractéristiques conditionnelles

On montre tout d'abord que :
$$\sum_{i,j} n_{ij} x_i y_j = \sum_i x_i \underbrace{\left(\sum_j n_{ij} y_j \right)}_{\text{Caractéristiques conditionnelles de y}} = \sum_j y_j \underbrace{\left(\sum_i n_{ij} x_i \right)}_{\text{Caractéristiques conditionnelles de x}}$$

Or, lors du calcul des *caractéristiques conditionnelles*, on a établi la valeur de $\sum_i n_{ij} x_i$ et $\sum_j n_{ij} y_j$, d'où le tableau de calculs suivant pour déterminer la valeur de $\sum_{i,j} n_{ij} x_i y_j$:

						• • ↕ + ↕ • •
y x		y_j		$n_{i\bullet}$	$\sum_j n_{ij} y_j$	$x_i \left(\sum_j n_{ij} y_j \right)$
x_i	⋮ ... n_{ij} ... ⋮		$n_{i\bullet}$	$\sum_j n_{ij} y_j$	$x_i \left(\sum_j n_{ij} y_j \right)$	
$n_{\bullet j}$	$n_{\bullet j}$	$n_{\bullet\bullet}$		S_5		
• ↔ • + • ↔ •	$\sum_i n_{ij} x_i$	$\sum_i n_{ij} x_i$				
$y_j \left(\sum_i n_{ij} x_i \right)$	$y_j \left(\sum_i n_{ij} x_i \right)$	S'_5	$\sum_{i,j} n_{ij} x_i y_j$			

d) Exemple

Le tableau, situé page 19, reprend une partie des données de l'exemple 1 (page 6).

- Détailler le calcul de **A** et **B**.
Compléter alors le tableau statistique (?). Flécher les totaux qui sont identiques.
- Préciser la valeur de : $\sum_{i,j} n_{ij} x_i y_j$
- On rappelle que : $\bar{x} = 41,25$ ans et $\bar{y} = 1,625$ enfants . En déduire la valeur de $\text{Cov}(x, y)$
- On pose $X = 2x + 15$ et $Y = 5y + 4$. Calculer $\text{Cov}(X, Y)$

Solution

Nbre d'enfants y_j		0	1	2	3	$n_{i\bullet}$	$\sum_j n_{ij} y_j$	$x_i \left(\sum_j n_{ij} y_j \right)$
Age	x_i							
[25 ; 35 [	30	10	10	30	0	50	70	A = 2 100
[35 ; 45 [	40	30	40	150	30	250	430	17 200
[45 ; 55 [	50	10	50	20	20	100	150	7 500
$n_{\bullet j}$		50	100	200	50	$n_{\bullet\bullet} = 400$	650	? 26 800
$\sum_i n_{ij} x_i$		2 000	4 400	7 900	2 200	16 500		
$y_j \left(\sum_i n_{ij} x_i \right)$		0	4 400	B=15800	6 600	? 26 800	$\sum_{i,j} n_{ij} x_i y_j = 26\,800$	

1. Calcul détaillé de A et B :

$$A = 30 \times 70 = \boxed{2\,100}$$

$$B = 2 \times 7\,900 = \boxed{15\,800}$$

(?) : Cf. tableau statistique

2. On obtient à partir du tableau statistique : $\sum_{i,j} n_{ij} x_i y_j = \boxed{26\,800}$

3. Calcul de Cov(x,y) :

$$\text{Cov}(x, y) = \frac{1}{n_{\bullet\bullet}} \left[\sum_{i,j} n_{ij} x_i y_j \right] - \bar{x} \bar{y}$$

$$\text{Cov}(x, y) = \frac{1}{400} \times 26\,800 - 41,25 \times 1,625$$

$$\text{Cov}(x, y) = \boxed{-0,03125}$$

4. Calcul de Cov(X,Y) :

$$\text{Cov}(X, Y) = \text{Cov}(2x + 15; 5y + 4)$$

$$\text{Cov}(X, Y) = 2 \times 5 \times \text{Cov}(x; y)$$

$$\text{Cov}(X, Y) = 2 \times 5 \times (-0,03125)$$

$$\text{Cov}(X, Y) = \boxed{-0,3125}$$

VI – Le tableau de calculs

L'ensemble des calculs peut être consigné dans un tableau unique. Pour vérifier si les calculs sont exacts, on doit retrouver les mêmes totaux en lignes et en colonnes :

$\begin{matrix} \diagdown \\ x \end{matrix} \backslash y$	y_j	$n_{i\bullet}$	$n_{i\bullet} x_i$	$n_{i\bullet} x_i^2$	$\sum_j n_{ij} y_j$	$\sum_j n_{ij} y_j^2$	$x_i \left(\sum_j n_{ij} y_j \right)$
x_i	$\begin{matrix} \vdots \\ \dots n_{ij} \dots \\ \vdots \end{matrix}$	$n_{i\bullet}$	$n_{i\bullet} x_i$	$n_{i\bullet} x_i^2$	$\sum_j n_{ij} y_j$	$\sum_j n_{ij} y_j^2$	$x_i \left(\sum_j n_{ij} y_j \right)$
$n_{\bullet j}$	$n_{\bullet j}$	$N = n_{\bullet\bullet}$	S_1	S_2	S_3	S_4	S_5
$n_{\bullet j} y_j$	$n_{\bullet j} y_j$	S'_1					
$n_{\bullet j} y_j^2$	$n_{\bullet j} y_j^2$	S'_2					
$\sum_i n_{ij} x_i$	$\sum_i n_{ij} x_i$	S'_3					
$\sum_i n_{ij} x_i^2$	$\sum_i n_{ij} x_i^2$	S'_4					
$y_j \left(\sum_i n_{ij} x_i \right)$	$y_j \left(\sum_i n_{ij} x_i \right)$	S'_5					

Remarques

- Ce tableau de calculs complet permet de vérifier la totalité des calculs. Cependant, on pourrait raisonner en affirmant que des lignes et des colonnes font double emploi. Pour obtenir un tableau de calculs « *a minima* », on peut supprimer :
 - les colonnes correspondant à : $n_{i\bullet} x_i$ et $n_{i\bullet} x_i^2$
 - les lignes correspondant à : $n_{\bullet j} y_j$ et $n_{\bullet j} y_j^2$
- Ce tableau correspond à la totalité des questions abordées (moyennes, variances, covariance), à l'exclusion d'autres caractéristiques (mode, médiane, quantiles, médiale, etc ...) nécessitant soit un tableau de calculs spécifique, soit au minimum des lignes ou colonnes supplémentaires à intégrer au tableau précédent. Cependant, dans un exercice, veiller à :
 - réserver une feuille à part pour le tableau de calculs
 - ne compléter le tableau de calculs qu'en fonction des questions posées (certaines parties du tableau de calculs ne sont peut-être pas nécessaires pour l'exercice traité).