Année Universitaire 2021-2022

Partie 2. Séries statistiques à deux dimensions (Bidimensionnelle)

Pr. Aicha EL ALAOUI

Université Sultan Moulay Slimane

Faculté d'Economie et de Gestion

Béni Mellal

جامعة السلطان مولاي سليمان

الكلية الاقتصاد والتدبير

بني ملال

Plan

- **Chapitre I : Caractéristiques marginales et conditionnelles d'une distribution à deux variables**
- **Chapitre II : Régression, ajustement linéaire et corrélation**

Chapitre I: Caractéristiques marginales et conditionnelles d'une distribution à deux variables

Rappel

- Une population peut être classée non seulement par rapport à un seul caractère, mais également selon deux ou plusieurs caractères.
- Lorsque l'on décrit une population selon deux caractères, le tableau correspondant est un **tableau croisé à deux dimensions (double entrée)**.
- Les tableaux croisés sont appelés ainsi car ils « croisent » deux **distributions au sein d'un même tableau.**
- Ainsi, on étudiera, par exemple, un ensemble de salariés non plus seulement selon le salaire, mais encore selon l'âge ou bien le nombre d'enfants, etc.

- Soit une population pour laquelle on observe le caractère X et le caractère Y.
- Le caractère X peut prendre les k modalités $x_1, x_2, \ldots, x_k$.
- Le caractère Y peut prendre les l modalités $y_1, y_2, \ldots, y_l$.
- L'effectif total de la population est noté « n ».
- Le nombre d'individus présentant la modalité x_i ET la modalité y_j est noté n_{ij} .
- Le tableau à double entrée représentant les séries statistiques à deux variables, prennent le nom de **tableaux de contingence**.

Tableau de contingence

- Considérons la distribution, selon les deux variables statistiques X et Y, d'une population P d'effectif total n (Tableau 1).

Variable Y Variable X	y_1	y_2	…	y_j	…	y_l	Totaux
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1l}	$n_{1.}$
x_2	n_{21}	n_{22}		n_{2j}		n_{2l}	$n_{2.}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{il}	$n_{i.}$
…	…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{kl}	$n_{k.}$
Totaux	$n_{.1}$	$n_{.2}$	…	$n_{.j}$	…	$n_{.l}$	**n**

Distribution de 17500 jeunes selon leurs âges et leurs salaires nets (10^3DH). Tableau 2

Salaire (y_j) Âge (x_i)	[5,6[	[6,7[	[7,8[	Total
[20,22[	1 200	500	100	1 800
[22,24[	2 500	3 500	600	6 600
[24,26[	1 800	5 000	2 300	9 100
Total	5 500	9 000	3 000	17 500

- Les modalités du caractère Y apparaissent en ligne, celles du caractère X apparaissent en colonne.
- Les effectifs apparaissant à l'intérieur du tableau (n_{ij}) sont toujours notés en prenant l'indice de x « i » d'abord, et celui de y « j » ensuite.
- Exemple :
- n_{21} du tableau est égal à 2500
- n_{12} du tableau est égal à 500
- $n_{kl} = n_{33} = 2300$
- La ligne et la colonne appelées ici « Total », prennent le nom de « marges ».

Caractéristiques marginales

- Le nombre d'individus présentant la modalité x_i (quelle que soit la modalité y_j) s'écrit:
- $n_{i\cdot} = \sum_{j=1}^{l} n_{ij}$
- Le nombre d'individus présentant la modalité y_j (quelle que soit la modalité x_i) s'écrit:
- $n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$
- Cette notation applique la convention selon laquelle une sommation sur l'une ou l'autre des dimensions se représente en remplaçant l'indice correspondant par un point.

- Dans les marges, ainsi que pour toute somme, on remplace l'indice qui varie par un point.
- Exemple : « $n_{1.}$ » signifie que l'on a sommé les effectifs de la première ligne. C'est l'indice j qui varie de 1 à l ; on le remplace par « • ». Dans le tableau 2, $n_{1.}$ = 1800.
- De la même manière « $n_{.j}$ » représente l'effectif de la modalité y_j, pour l'ensemble des modalités de x. Ici c'est l'indice « i » qui varie de 1 à k ; on le remplace par « • ».
- « $n_{.1}$ » signifie que l'on a sommé les effectifs de la première colonne.
- Les effectifs de la distribution marginale de X (Y) seront repérés par le terme générique « $\mathbf{n_{i.}}$ » (« $\mathbf{n_{.j}}$ »). L'effectif total s'écrit donc « $\mathbf{n_{..}}$ », c'est le nombre d'individus de la population étudiée.

9

Distributions marginales : effectifs et fréquences

- La somme des effectifs des lignes fournit la distribution marginale du caractère X, comme si l'on étudiait seulement le caractère X.
- La somme des effectifs des colonnes fournit la distribution marginale du caractère Y, comme si l'on étudiait seulement le caractère Y.
- Les distributions marginales de X et de Y représentent des distributions à une seule dimension où les individu sont distribués suivant le seul caractère X ou selon le seul caractère Y.
- Les effectifs marginaux : Ce sont les effectifs lus dans les deux marges du tableau (marge du bas « ligne total » et celle de droite « colonne total »):
- $n_{1.}, n_{2.}, \ldots, n_{i.}, \ldots, n_{k.}$ pour x
- $n_{.1}, n_{.2}, \ldots, n_{.j}, \ldots, n_{.l}$ pour y
- Il y aura autant d'effectifs marginaux que de modalités (« k » pour X et « l » pour Y)
- La fréquence marginale, $f_{i.}$, représente la fréquence de la modalité x_i
- La fréquence marginale, $f_{.j}$, représente la fréquence de la modalité y_j

10

Fréquences

- $f_{ij} = n_{ij}/n$: la proportion d'individus présentant à la fois la modalité x_i et la modalité y_j.
- $f_{i.} = \sum_{j=1}^{l} f_{ij} = n_{i.}/n$ est la somme des fréquences de la ligne i.

$f_{i.}$ représente la proportion des individus présentant la modalité x_i quelle que soit la modalité y_j (fréquences marginales du caractère X).

- $f_{.j} = \sum_{i=1}^{k} f_{ij} = n_{.j}/n$ est la somme des fréquences de la colonne j ;

$f_{.j}$ représente la proportion des individus présentant la modalité y_j quelle que soit les modalités x_i (fréquences marginale du caractère Y).

11

- Si l'on reprend le tableau 2 on a par exemple :
- $\mathbf{f_{22}}$ = n_{22}/n.. = 3500/17500 = 0,2 ⟹ 20% des individus étudiés ont entre 22 et 24 ans, **et** gagnent entre 6000 et 7000 DH de salaire net.
- $\mathbf{f_{2.}}$ = $n_{2.}$/n.. = 6600/17500 = 0,377 ⟹ 37,7% des individus étudiés sont âgé de 22 à 24 ans.
- $\mathbf{f_{.2}}$ = $n_{.2}$/n.. = 9000/17500 = 0,514 ⟹ 51,4% des individus étudiés touchent un salaire net compris entre 6000 et 7000 DH.

12

Exemple

Âge (x_i)	$n_{i.}$	$f_{i.}=n_{i.}/n$ (%)
[20,22[	$n_{1.}$ = 1 800	10,29
[22,24[	$n_{2.}$ = 6 600	37,71
[24,26[	$n_{3.}$ = 9 100	52
Total	**n = 17 500**	**100**

Salaire (y_j)	$n_{.j}$	$f_{.j}= n_{.j}/n$ (%)
[5,6[	$n_{.1}$ = 5 500	31,43
[6,7[	$n_{.2}$ = 9 000	51,43
[7,8[	$n_{.3}$ = 3 000	17,14
Total	**17 500**	**100**

Distributions conditionnelles de X liées par y_j

- Si on ne s'intéresse qu'aux individus présentant la modalité y_j , l'effectif de ce groupe est $n_{.j}$;
- La fréquence conditionnelle de la modalité x_i liée par y_j est:

$$f_{i\,|\,j} = f x_i \,|\, y_j = \frac{n_{ij}}{n.j}$$

- Les k fréquences conditionnelles pour la modalité y_j forment la distribution conditionnelle de X liée par y_j.
- Dans un tableau, il y a « l » **distributions conditionnelles en colonne** (l = modalités du caractère Y).

Exemple

Distribution conditionnelle de l'âge liée par la classe de salaire [5,6[

$x_i \mid y_1$

$x_i \mid y_1$ y_1 = [5,6[	$n_{i/j} = n_{ij}$	$f_{i\|1}= n_{ij}/n_{.1}$ (%)
[20,22[	n_{11} = 1 200	21,82
[22,24[	n_{21} = 2 500	45,45
[24,26[	n_{31} = 1 800	32,73
Total	**n.1 = 5 500**	100

Distribution conditionnelle de l'âge liée par la classe de salaire [6,7[

$x_i \mid y_2$

$x_i \mid y_2$ y_2 = [6,7[	$n_{i/j} = n_{ij}$	$f_{i\|2}= n_{ij}/n_{.2}$ (%)
[20,22[	n_{12} = 500	5,55
[22,24[	n_{22} = 3 500	38,89
[24,26[	n_{32} = 5 000	55,56
Total	**$n_{.2}$ = 9 000**	100

Distributions conditionnelles de Y liées par x_i

- De même, si on ne s'intéresse qu'aux individus présentant la modalité x_i, l'effectif de ce groupe est $n_{i.}$;
- La fréquence conditionnelle de la modalité y_j liée par x_i est:

$$f_{j\,|\,i} = f y_j \,|\, x_i = \frac{n_{ij}}{n_{i.}}$$

- Les « l » fréquences conditionnelles pour la modalité x_i forment la distribution conditionnelle de Y liée par x_i.
- Dans un tableau, **il y a k distributions conditionnelles en ligne (k = nombre de modalités du caractère X**).

Exemple

Distribution conditionnelle du salaire liée par la classe d'âge [20,22[

$y_j | x_1$

$y_j \mid x_1$ x_1 = [20,22[	$n_{j/i} = n_{ij}$	$f_{j/1} = n_{ij}/n_{1.}$ en %
[5,6[	n_{11} = 1 200	66,67
[6,7[	n_{12} = 500	27,78
[7,8[	n_{13} = 100	5,55
Total	**$n_{1.}$ = 1 800**	100

Distribution conditionnelle du salaire liée par la classe d'âge [22,24[

$y_j | x_2$

$y_j \mid x_2$ x_2 = [22,24[	$n_{j/i} = n_{ij}$	$f_{j/2} = n_{ij}/n_{2.}$ en %
[5,6[	n_{21} = 2 500	37,88
[6,7[	n_{22} = 3 500	53,03
[7,8[	n_{23} = 600	9,09
Total	**$n_{2.}$ = 6 600**	100

17

- La somme des fréquences est égale à l'unité :
- Fréquences marginales :

$$\sum_{j=1}^{l} f_{.j} = \sum_{j=1}^{l} \frac{n_{.j}}{n_{..}} = \frac{1}{n_{..}} \sum_{j=1}^{l} n_{.j} = \frac{n_{..}}{n_{..}} = 1$$

$$\sum_{i=1}^{k} f_{i.} = \sum_{i=1}^{k} \frac{n_{i.}}{n_{..}} = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} = \frac{n_{..}}{n_{..}} = 1$$

- Fréquences conditionnelles :

$$\sum_{i=1}^{k} f_{i/j} = \sum_{i=1}^{k} \frac{n_{ij}}{n_{.j}} = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} = \frac{n_{.j}}{n_{.j}} = 1$$

$$\sum_{j=1}^{l} f_{j/i} = \sum_{j=1}^{l} \frac{n_{ij}}{n_{i.}} = \frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} = \frac{n_{i.}}{n_{i.}} = 1$$

18

Application

- Le directeur d'un établissement privé veut étudier la liaison entre l'âge de ses cadres et le niveau du salaire mensuel. L'enquête portant sur un échantillon de 195 cadres a donné les résultats suivants :

		y_j : salaire mensuel en 10^2 DH							
		[10,16[	[16,22[	[22,28[	[28,34[	[34,40[	[40,46[	[46,52[	[52,58[
Âge x_i	[22,28[	5	4	1	0	0	0	0	0
	[28,34[	4	5	8	6	3	2	1	0
	[34,40[	0	2	8	10	9	6	4	2
	[40,46[	1	2	6	8	10	8	6	2
	[46,52[	0	0	2	4	8	6	8	4
	[52,58[	0	0	1	7	10	10	8	4

19

1. Déterminer les distributions marginales (en termes d'effectifs) des deux caractères : âge et salaire mensuel

2. Donner la signification des valeurs des $n_{.5}$; $n_{4.}$; n_{45} ; n_{54}

3. Dégager les distributions conditionnelles (en termes de fréquences) de x liée à $y = y_3$ et à $y = y_7$. De même donner les distributions conditionnelles (en termes de fréquences) de y liées à $x = x_4$ et à $x = x_6$

4. Trouver :

- La proportion des salariés âgés entre 46 et 52 ans quel que soit leur salaire

- La proportion des salariés qui gagnent entre 4600 et 5200 DH quel que soit leur âge

- La proportion des salariés âgés entre 28 et 34 ans <u>et</u> dont le salaire est compris entre 2800 et 3400 DH.

- La proportion des salariés qui gagnent entre 2200 et 2800 DH <u>sachant</u> qu'ils sont âgés entre 40 et 46 ans

- La proportion des salariés âgés entre 34 et 40 ans <u>sachant</u> que leur salaire est compris entre 2800 et 3400 DH.

20

Relations entre fréquences marginales et conditionnelles

- On démontre également que le produit des fréquences marginales par les fréquences conditionnelles est égal aux fréquences conjointes :

$$f_{i.} \times f_{j/i} = \frac{n_{i.}}{n_{..}} \times \frac{n_{ij}}{n_{i.}} = \frac{n_{ij}}{n_{..}} = f_{ij}$$

$$f_{.j} \times f_{i/j} = \frac{n_{.j}}{n_{..}} \times \frac{n_{ij}}{n_{.j}} = \frac{n_{ij}}{n_{..}} = f_{ij}$$

$$\boldsymbol{f_{i.} \times f_{j/i} = f_{ij} \text{ et } f_{.j} \times f_{i/j} = f_{ij}}$$

21

Remarques sur l'effectif total

- $n_{..}$, l'effectif total, est égal, soit à la somme des $n_{i.}$, soit à celle des $n_{.j}$ (effectifs marginaux), qui sont eux-mêmes des sommes. Donc :

$$n_{..} = \sum_{i=1}^{k} n_{i.} = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}$$

$$n_{..} = \sum_{j=1}^{l} n_{.j} = \sum_{j=1}^{l} \sum_{i=1}^{k} n_{ij}$$

- On voit donc que l'on peut inverser l'ordre des sommes (commutativité) : $\sum_i \sum_j = \sum_j \sum_i$
-

22

Caractéristiques des séries à deux caractères

- On se limitera aux moyennes arithmétiques et aux variances. Puisqu'il existe des distributions marginales et conditionnelles, il est logique de déterminer des moyennes et des variances marginales et conditionnelles.
- Si nous reprenions l'exemple de la distribution de 17500 jeunes selon leurs âges et leurs salaires nets, l'âge moyen et la variance des âges seront calculés à partir de la distribution marginale de l'âge (caractère x). De la même manière le salaire moyen et la variance des salaires seront calculés à partir de la distribution marginale du salaire (caractère y).
- L'âge moyen et la variance des âges pour chaque classe de salaire y_j, seront tirés des distributions conditionnelles de x liées par y (« i » varie et « j » est fixé), et inversement.
-

23

Caractéristiques des distributions marginales

Caractéristiques marginales du caractère x

Caractéristiques marginales du caractère y

24

Caractéristiques marginales du caractère x

- Les caractéristiques marginales de x (moyenne et variance) s'obtiennent à partir de la distribution marginale de x, c'est-à-dire à partir des « k » modalités (x_i) et des « k » effectifs marginaux ($n_{i.}$).
- **La moyenne marginale** $\bar{\bar{x}}$
- Il s'agit d'une moyenne arithmétique, que l'on note $\bar{\bar{x}}$ (lue : « x double barre »).
- Sa détermination relève de la même logique que celle utilisée pour les séries à une seule dimension. On a :

$$\bar{\bar{x}} = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} x_i = \frac{1}{n} \sum_{i=1}^{k} n_{i.} x_i$$

25

- L'effectif marginal $n_{i.}$ étant lui-même une somme, on peut écrire :

$$\bar{\bar{x}} = \frac{1}{n_{..}} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} x_i$$

- Selon les données du tableau 2 :

$$\bar{\bar{x}} = \frac{1}{17500} [(1800 \times 21) + (6600 \times 23) + (9100 \times 25)] = 23{,}83$$

- L'âge moyen des individus est donc de 23,83 ans (soit 23 ans, 9 mois et 29 jours)

26

La variance marginale de x : V(x)

- La formule de définition est la même que pour les séries à une seule dimension :

$$V(x) = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} (x_i - \bar{\bar{x}})^2$$

- La formule développée est :

$$V(x) = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} x_i^2 - \bar{\bar{x}}^2$$

- L'écart-type est égal à la racine carrée de la variance.
- Dans le tableau 2 :

$$\frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} x_i^2 = \frac{1}{17500} [(1800 \times 21^2) + (6600 \times 23^2) + (9100 \times 25^2)] = 569{,}869$$

$$\bar{\bar{x}}^2 = (23{,}8)^2 = 566{,}44$$

$$Et\ V(x) = 569{,}869 - 566{,}44 = 3{,}43$$

$$\sigma_x = 1{,}85\ an$$

27

Caractéristiques marginales du caractère y

- La démarche est la même que précédemment : les caractéristiques marginales de y s'obtiennent à partir des « l » modalités (y_j) et des « l » effectifs marginaux ($n_{.j}$).
- La moyenne s'écrit :

$$\bar{\bar{y}} = \frac{1}{n_{..}} \sum_{j=1}^{l} n_{.j} y_j = \frac{1}{n_{..}} \sum_{j=1}^{l} \sum_{i=1}^{k} n_{ij} y_j$$

- La variance s'écrit :

$$V(y) = \frac{1}{n_{..}} \sum_{j=1}^{l} n_{.j} (y_j - \bar{\bar{y}})^2$$

- Soit, par la formule développée :

$$V(y) = \frac{1}{n_{..}} \sum_{j=1}^{l} n_{.j} y_j^2 - \bar{\bar{y}}^2 \qquad \sigma_y = \sqrt{V(y)}$$

28

Application numérique

- Selon les données du tableau 2 :

$$\bar{y} = \frac{1}{17500}[(5500 \times 5{,}5) + (9000 \times 6{,}5) + (3000 \times 7{,}5)] = 6{,}36$$

$$\frac{1}{n}\sum_{j=1}^{l} n_{.j} y_j^2 = \frac{1}{17500}[(5500 \times 5{,}5^2) + (9000 \times 6{,}5^2) + (3000 \times 7{,}5^2)] = 40{,}88$$

$$\bar{\bar{y}}^2 = (6{,}36)^2 = 40{,}45$$

$$Et\ V(y) = 40{,}88 - 40{,}45 = 0{,}43$$

$$\sigma_y = 0{,}65\ soit\ 650\ DH$$

29

Caractéristiques conditionnelles des distributions

Caractéristiques conditionnelles de x lié par y

Moyennes conditionnelles de x lié par y

30

Caractéristiques conditionnelles de x lié par y

- Il existe deux types de distributions conditionnelles : celle de x lié par y (où x varie et y est fixe) ; celle de y lié par x (où y varie et x est fixe).
- Les caractéristiques de x lié par y (moyenne, variance) s'obtiennent à partir des distributions conditionnelles de x lié par y, c'est-à-dire à partir des « k » modalités de x et des « l » colonnes d'effectifs, à l'intérieur du tableau de contingence.
- Il y a en effet « l » distributions conditionnelles de x lié par y. Les distributions conditionnelles de x lié par y sont les colonnes d'effectifs du tableau de contingence associées aux modalités de x.

31

Moyennes conditionnelles de x lié par y

- Les « l » moyennes conditionnelles de x lié par y (ou « de x si y ») sont de la forme :

$$\bar{x}_j = \frac{1}{n_{.j}}\sum_{i=1}^{k} n_{ij} x_i$$

- La notation $\bar{x}_j$ signifie : « moyenne conditionnelle de x, si y = y_j ».
- La sommation se fait par rapport à « i », puisque l'on somme « en colonne ». Les effectifs par lesquels on multiplie chaque valeur du caractère $x(x_i)$ sont les effectifs conjoints n_{ij} du tableau de contingence. L'effectif par lequel on divise la somme pondérée est $n_{.j}$, c'est-à-dire la somme des effectifs conjoints de la colonne « j ».

32

- Par exemple:
- Sur les données chiffrées du tableau 2, la moyenne conditionnelle $\bar{x}_3$ est :

$$\bar{x}_3 = \frac{1}{3000}[(100 \times 21) + (600 \times 23) + (2300 \times 25)] = 24{,}47$$

- L'âge moyen des individus qui gagnent entre 7000 et 8000 DH est donc de 24,47 ans (environ 24 ans et 6 mois)

33

Variances conditionnelles de x lié par y

- Les « l » variances conditionnelles de x lié par y (ou « de x si y ») se notent $V_j(x)$. Elles sont de la forme :

$$V_j(x) = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij}(x_i - \bar{x}_j)^2$$

- Ou bien par la formule développée :

$$V_j(x) = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij}x_i^2 - \bar{x}_j{}^2$$

- Les sommations se font « par rapport à i » puisque l'on somme, en colonne, des effectifs se rapportant aux valeurs du caractère x. Les écarts que l'on élève au carré, sont des écarts entre la variable (x_i) et la moyenne conditionnelle correspondante $\bar{x}_j$.

34

- En utilisant les données du tableau 1, la variance conditionnelle de x si y = y_3 est V3(x) :

$$\frac{1}{n_{.3}} \sum_{i=1}^{k} n_{i3}x_i^2 = \frac{1}{3000}[(100 \times 21^2) + (600 \times 23^2) + (2300 \times 25^2)] = 599{,}67$$

$$\bar{x}_3{}^2 = (24{,}47)^2 = 598{,}78$$

$$Donc : V_3(x) = 599{,}67 - 598{,}78 = 0{,}89$$

- Donc l'écart-type des âges des individus gagnant entre 7000 et 8000 DH de salaire est de 0,94 année.

35

Moyennes conditionnelle de y lié par x

- Elles s'obtiennent à partir des « l » modalités du caractère y et à partir des « k » lignes d'effectifs conjoints du tableau de contingence.
- Il y a donc « k » distributions, « k » moyennes et « k » variances, conditionnelles de y lié par x.

$$\bar{y}_i = \frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} y_j$$

- La notation $\bar{y}_i$ signifie : « moyenne conditionnelle de y, si x = x_i ».
- La sommation se fait par rapport à « j », puisque l'on somme « en ligne ». Les effectifs par lesquels on multiplie chaque valeur du caractère y (y_j) sont les effectifs conjoints n_{ij} du tableau de contingence. L'effectif par lequel on divise la somme pondérée est $n_{i.}$, c'est-à-dire la somme des effectifs conjoints de la ligne « i ».

36

Variances conditionnelles de y lié par x

- Les « k » variances conditionnelles de y lié par x (ou « de y si x ») se notent $V_i(y)$. Elles sont de la forme :

$$V_i(y) = \frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} (y_j - \bar{y}_i)^2$$

- Ou bien par la formule développée :

$$V_i(y) = \frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} y_j^2 - \bar{y}_i{}^2$$

- Les sommations se font « par rapport à j » puisque l'on somme, en ligne, des effectifs se rapportant aux valeurs du caractère x. Les écarts que l'on élève au carré, sont des écarts entre la variable (x_i) et la moyenne conditionnelle correspondante.

37

Exemple

- Combien gagnent en moyenne les individus de l'échantillon sachant qu'ils sont âgés de 22 à 24 ans, et avec quelle dispersion?
- Pour répondre à cette question, il faudrait calculer la moyenne et variance de la distribution conditionnelle de y lié à $x = x_2$.

$$\bar{y}_2 = \frac{1}{6600}[(2500 \times 5{,}5) + (3500 \times 6{,}5) + (600 \times 7{,}5)] = 6{,}21$$

$$\frac{1}{n_{2.}} \sum_{j=1}^{3} n_{2j} y_j^2 = \frac{1}{6600}[(2500 \times 5{,}5^2) + (3500 \times 6{,}5^2) + (600 \times 7{,}5^2)] = 38{,}98$$

$$Donc : V_2(y) = 38{,}98 - 38{,}56 = 0{,}42$$

- En moyenne, les jeunes âgés entre 22 et 24 ans gagent 6210 DH avec une dispersion de 420 DH.

38

Relations entre caractéristiques marginales et conditionnelles

- La population P peut être considérée comme composée :
- Soit des l sous-population $P_1, P_2, \ldots, P_l$ d'effectifs $n_{.1}, n_{.2}, \ldots, n_{.l}$, correspondant aux distributions conditionnelles de X lié par Y ;
- Soit des k sous-population $P'_1, P'_2, \ldots, P'_k$ d'effectifs $n_{1.}, n_{2.}, \ldots, n_{k.}$, correspondant aux distributions conditionnelles de Y lié par X.
- **Expression de la moyenne marginale en fonction des moyennes conditionnelles :**
- La moyenne relative à l'ensemble de la population est égale à la moyenne pondérée des moyennes des sous-populations. Les coefficients de pondération sont les proportions des sous-populations dans la population totale.

39

Relations entre les moyennes marginales et les moyennes conditionnelles

- On a par définition :

$$\bar{\bar{x}} = \frac{1}{n_{..}} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} x_i$$

- Ou :

$$\bar{\bar{x}} = \frac{1}{n_{..}} \sum_{j=1}^{l} \left(\sum_{i=1}^{k} n_{ij} x_i \right)$$

- Or :

$$\bar{x}_j = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} x_i \Rightarrow \sum_{i=1}^{k} n_{ij} x_i = n_{.j} \times \bar{x}_j$$

- Donc :

$$\bar{\bar{x}} = \frac{1}{n_{..}} \sum_{j=1}^{l} n_{.j} \bar{x}_j$$

- La démonstration est faite pour x ; elle est identique pour y :

$$\bar{\bar{y}} = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} \bar{y}_i$$

40

- La moyenne marginale est égale à la moyenne des moyennes conditionnelles, pondérée par les effectifs marginaux.

Relations entre les variances

- La variance relative à l'ensemble d'une population est égale à la somme de deux termes, la moyenne pondérée des variances des sous-populations (variances conditionnelles) et la variance pondérée des moyennes des sous-populations (moyennes conditionnelles).
- Par conséquent, en considérant la distribution marginale de X comme formée de distributions conditionnelles de X lié par Y, on obtient :

$$V(x) = \frac{1}{n_{..}}\sum_{j=1}^{l} n_{.j}\left(\bar{x}_j - \bar{\bar{x}}\right)^2 + \frac{1}{n_{..}}\sum_{j=1}^{l} V_j(x) \times n_{.j}$$

- Autre notation :

$$V(x) = V\left(\bar{x}_j\right) + \overline{V_j(x)}$$

Démonstration

- Nous démontrons ici pour x ; la démonstration pour y est identique. On a :

$$\sum_{j=1}^{l} n_{ij}(x_i - \bar{\bar{x}})^2$$

- Ajoutons et retranchons la même valeur, la moyenne conditionnelle de x, dans le terme entre parenthèses :

$$V(x) = \frac{1}{n_{..}}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}\left[(x_i - \bar{x}_j) + (\bar{x}_j - \bar{\bar{x}})\right]^2$$

$$= \frac{1}{n_{..}}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}\left(x_i - \bar{x}_j\right)^2 + \frac{2}{n_{..}}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}\left(x_i - \bar{x}_j\right)\left(\bar{x}_j - \bar{\bar{x}}\right) + \frac{1}{n_{..}}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}\left(\bar{x}_j - \bar{\bar{x}}\right)^2$$

$$\text{Soit } A = \frac{1}{n_{..}}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}\left(x_i - \bar{x}_j\right)^2 = \frac{1}{n_{..}}\sum_{j=1}^{l}\sum_{i=1}^{k} n_{ij}\left(x_i - \bar{x}_j\right)^2$$

$$\text{Or : } V_j(x) = \frac{1}{n_{.j}}\sum_{i=1}^{k} n_{ij}\left(x_i - \bar{x}_j\right)^2$$

$$\Rightarrow \sum_{i=1}^{k} n_{ij}\left(x_i - \bar{x}_j\right)^2 = n_{.j}V_j(x)$$

$$Donc\ A = \frac{1}{n_{..}}\sum_{j=1}^{l} n_{.j}V_j(x)$$

$$\text{Soit } B = \frac{2}{n_{..}}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}\left(x_i - \bar{x}_j\right)\left(\bar{x}_j - \bar{\bar{x}}\right)$$

$$= \frac{2}{n_{..}} \sum_{j=1}^{l} \left(\sum_{i=1}^{k} n_{ij}(x_i - \bar{x}_j)(\bar{x}_j - \bar{\bar{x}}) \right) = \frac{2}{n_{..}} \sum_{j=1}^{l} (\bar{x}_j - \bar{\bar{x}}) \sum_{i=1}^{K} n_{ij}(x_i - \bar{x}_j)$$

$$\text{Or} \sum_{i=1}^{k} n_{ij} x_i - \sum_{i=1}^{k} n_{ij} \bar{x}_j = n_{.j}\bar{x}_j - n_{.j}\bar{x}_j = 0$$

$$\text{Soit } C = \frac{1}{n_{..}} \sum_{j=1}^{l} \sum_{i=1}^{k} n_{ij}(\bar{x}_j - \bar{\bar{x}})^2 = \frac{1}{n_{..}} \sum_{j=1}^{l} (\bar{x}_j - \bar{\bar{x}})^2 \sum_{i=1}^{K} n_{ij}$$

$$C = \frac{1}{n_{..}} \sum_{j=1}^{l} n_{.j}(\bar{x}_j - \bar{\bar{x}})^2$$

$$\textit{En définitive} : V(x) = \frac{1}{n_{..}} \sum_{j=1}^{l} n_{.j}(\bar{x}_j - \bar{\bar{x}})^2 + \frac{1}{n_{..}} \sum_{j=1}^{l} V_{j(x)}\, n_{.j}$$

45

- La variance marginale est égale à la moyenne des variances conditionnelles augmentée de la variance des moyennes conditionnelles.

$$V(y) = \frac{1}{n_{..}} \sum_{j=1}^{l} n_{i.}(\bar{y}_i - \bar{y})^2 + \frac{1}{n_{..}} \sum_{j=1}^{l} V_i(y) \times n_{i.}$$

- Autre notation :

$$V(y) = V(\bar{y}_i) + \overline{V_i(y)}$$

46

- La variance marginale est égale à la somme de la moyenne pondérée des variances conditionnelles et de la variance pondérée des moyennes conditionnelles.
- La dispersion de la distribution marginale résulte donc de deux facteurs :
- - La dispersion de chacune des distributions conditionnelles autour de leur moyenne ;
- - La dispersion des moyennes conditionnelles entre elles.
- Ainsi une partie de la variance totale de X (ou de Y) peut être expliquée par la variance des moyennes conditionnelles.
- La variance moyenne des hétérogénéités propres à chacune des distributions conditionnelles (Moyenne des variances conditionnelles) apparaît alors come une variance résiduelle. C'est sur cette décomposition de la variance totale que sera fondée la définition du rapport de corrélation.

47

Covariance

- La notion de covariance nous servira pour étudier la corrélation. On appelle covariance entre deux variables statistiques x et y [notée cov(x,y)] l'expression définie par :

$$\boldsymbol{Cov(x,y) = \frac{1}{n_{..}} \sum_{i} \sum_{j} n_{ij}(x_i - \bar{x})(y_j - \bar{y})}$$

- On remarquera que si l'on pose y = x, on retrouve la formule de la variance.
- La covariance sert à vérifier l'existence de **lien de dépendance** ou pas entre deux variables. La covariance est nulle si les deux variables indépendantes. Cette grandeur interviendra dans l'étude de la liaison entre deux variables et, notamment, dans celle de la corrélation linéaire.

48

Calcul pratique

- Pour simplifier le calcul de la covariance, on utilise des méthodes analogues à celles employées pour le calcul de la variance : formule développée et changement de variable.
- **Formule développée**:
- Il est possible de développer la formule de définition pour obtenir une expression mieux adaptée au calcul numérique :

$$Cov(x,y) = \frac{1}{n_{..}}\sum_i\sum_j n_{ij}(x_i - \bar{\bar{x}})(y_j - \bar{\bar{y}})$$

$$= \frac{1}{n_{..}}\sum_i\sum_j n_{ij}x_iy_j - \bar{\bar{y}}\frac{1}{n_{..}}\sum_i\sum_j n_{ij}x_i - \bar{\bar{x}}\frac{1}{n_{..}}\sum_i\sum_j n_{ij}y_j + \bar{\bar{x}}\,\bar{\bar{y}}\frac{1}{n_{..}}\sum_i\sum_j n_{ij}$$

- Or, par définition :

49

$$\sum_i\sum_j n_{ij} = 1$$

$$\sum_i\sum_j n_{ij}x_i = \sum_i x_i\sum_j n_{ij} = \sum_i n_{i.}x_i = n_{..}\bar{\bar{x}}$$

$$\sum_i\sum_j n_{ij}y_j = \sum_j y_j\sum_i n_{ij} = \sum_i n_{.j}y_j = n_{..}\bar{\bar{y}}$$

- D'où :

$$Cov(x,y) = \frac{1}{n_{..}}\sum_i\sum_j n_{ij}x_iy_j - \bar{\bar{x}}\,\bar{\bar{y}}$$

50

Grille du calcul des caractéristiques des séries à deux dimensions

51

y(j) / x(i)	y_1	y_2	y_3	y_4	(7) $n_{i.}$	(8) $n_{i.}x_i$	(9) $n_{i.}x_i^2$	(10) $\Sigma n_{ij}y_j$	(11) $\Sigma n_{ij}y_j^2$	(12) $\Sigma n_{ij}x_iy_j$
x_1	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$	$n_{1.}x_1$	$n_{1.}x_1^2$	$\Sigma n_{1j}y_j$	$\Sigma n_{1j}y_j^2$	$\Sigma n_{1j}x_1y_j$
x_2	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$	$n_{2.}x_2$	$n_{2.}x_2^2$	$\Sigma n_{2j}y_j$	$\Sigma n_{2j}y_j^2$	$\Sigma n_{2j}x_2y_j$
x_3	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$	$n_{3.}x_3$	$n_{3.}x_3^2$	$\Sigma n_{3j}y_j$	$\Sigma n_{3j}y_j^2$	$\Sigma n_{3j}x_3y_j$
x_4	n_{41}	n_{42}	n_{43}	n_{44}	$n_{4.}$	$n_{4.}x_4$	$n_{4.}x_4^2$	$\Sigma n_{4j}y_j$	$\Sigma n_{4j}y_j^2$	$\Sigma n_{4j}x_4y_j$
x_5	n_{51}	n_{52}	n_{53}	n_{54}	$n_{5.}$	$n_{5.}x_5$	$n_{5.}x_5^2$	$\Sigma n_{5j}y_j$	$\Sigma n_{5j}y_j^2$	$\Sigma n_{5j}x_5y_j$
(1) $n_{.j}$	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{..}$	$\Sigma =$	$\Sigma =$			$\Sigma\Sigma =$
(2) $n_{.j}y_j$	$n_{.1}y_1$	$n_{.2}y_2$	$n_{.3}y_3$	$n_{.4}y_4$	$\Sigma n_{.j}y_j$					
(3) $n_{.j}y_j^2$	$n_{.1}y_1^2$	$n_{.2}y_2^2$	$n_{.3}y_3^2$	$n_{.4}y_4^2$	$\Sigma n_{.j}y_j^2$					
(4) $\Sigma n_{ij}x_i$	$\Sigma n_{i1}x_i$	$\Sigma n_{i2}x_i$	$\Sigma n_{i3}x_i$	$\Sigma n_{i4}x_i$						
(5) $\Sigma n_{ij}x_i^2$	$\Sigma n_{i1}x_i^2$	$\Sigma n_{i2}x_i^2$	$\Sigma n_{i3}x_i^2$	$\Sigma n_{i4}x_i^2$						
(6)	$\Sigma n_{i1}x_iy_1$	$\Sigma n_{i2}x_iy_2$	$\Sigma n_{i3}x_iy_3$	$\Sigma n_{i4}x_iy_4$	$\Sigma\,\Sigma=$					

Calcul pratique des caractéristiques marginales et conditionnelles

- Le principe de calcul consiste à ajouter des colonnes et des lignes, en bon ordre, au tableau de contingence.
- Les variables x_i et y_j seront représentées par leurs centres de classe, si elles sont continues.
- Pour le calcul de la moyenne marginale de x on doit calculer la colonne : $n_{i.}x_i$, que l'on ajoutera à droite du tableau de contingence. Pour la moyenne marginale de y il nous faudra la ligne $n_{.j}y_j$. On la rajoutera en bas du tableau de contingence.
- Pour les calculs des variances marginales (de x et de y) il nous faudra calculer respectivement la colonne $n_{i.}x_i^2$ et la ligne $n_{.j}y_j^2$.
- Pour le calcul des moyennes et variances conditionnelles de y il nous faudra ajouter les deux colonnes $\Sigma n_{ij}y_j$ et $\Sigma n_{ij}y_j^2$. De même pour obtenir les moyennes et variances conditionnelles de x, on ajoute les deux lignes $\Sigma n_{ij}x_i$ et $\Sigma n_{ij}x_i^2$

- Le calcul de la covariance dans cette présentation, nécessite le calcul des « nombres en coin » ; chaque « coin » des cases n_{ij} du tableau de contingence correspond au calcul des produits $n_{ij}x_iy_j$; on somme des nombres en coins dans la dernière colonne ou la dernière ligne. Si l'on retranche de cette somme, divisée par $n_{..}$, le produit $\bar{\bar{x}}\bar{\bar{y}}$ on obtient la covariance.
- Les trois premières lignes (1,2,3) et colonnes (7,8,9) ajoutées au tableau de contingence servent pour le calcul des moyennes et des variances marginales, respectivement de x et de y.
- Les 2 lignes (4,5) et colonnes (10, 11) d'après servent au calcul des moyennes et des variances conditionnelles, respectivement de x et de y. La dernière ligne (6) et la dernière colonne (12) dans le tableau servent à calculer la covariance.
- Chaque case des 3 dernières lignes et colonnes correspond à la somme d'un produit. La 1ère case de la ligne 4 :
- $\Sigma n_{i1}x_i = n_{11}x_1 + n_{21}x_2 + n_{31}x_3 + n_{41}x_4 + n_{51}x_5$
- La 1ère case de la colonne 11 :
- $\Sigma n_{1j}y_j^2 = n_{11}y_1^2 + n_{12}y_2^2 + n_{13}y_3^2 + n_{14}y_4^2$

Exemple 1

	2	4	6	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$	$\Sigma n_{ij}yj$	$\Sigma n_{ij}y_j^2$	$\Sigma n_{ij}x_iy_j$
2	0 0	1 8	1 12	2	4	8	$\Sigma n_{1j}y_j = 10$	$\Sigma n_{1j}y_j^2 = 52$	$\Sigma n_{1j}x_1y_j = 20$
4	2 16	3 48	0 0	5	20	80	$\Sigma n_{2j}y_j = 16$	$\Sigma n_{2j}y_j^2 = 56$	$\Sigma n_{2j}x_2y_j = 64$
6	1 12	1 24	1 36	3	18	108	$\Sigma n_{3j}y_j = 12$	$\Sigma n_{3j}y_j^2 = 56$	$\Sigma n_{3j}x_3y_j = 72$
$n_{.j}$	3	5	2	$\Sigma = 10$	$\Sigma = 42$	$\Sigma = 196$			$\Sigma\Sigma = 156$
$n_{.j}y_j$	6	20	12	$\Sigma = 38$					
$n_{.j}y_j^2$	12	80	72	$\Sigma = 164$					
$\Sigma n_{ij}x_i$	$\Sigma n_{i1}x_i = 14$	$\Sigma n_{i2}x_i = 20$	$\Sigma n_{i3}x_i = 8$						
$\Sigma n_{ij}x_i^2$	$\Sigma n_{i1}x_i^2 = 68$	$\Sigma n_{i2}x_i^2 = 88$	$\Sigma n_{i3}x_i^2 = 40$						
$\Sigma n_{ij}x_iy_j$	$\Sigma n_{i1}x_iy_1 = 28$	$\Sigma n_{i2}x_iy_2 = 80$	$\Sigma n_{i3}x_iy_3 = 48$	$\Sigma\Sigma = 156$					

Caractéristiques marginales

$$\bar{\bar{x}} = \frac{1}{n_{..}}\sum_{i=1}^{k} n_{i.}x_i = \frac{42}{10} = 4,2$$

$$\bar{\bar{y}} = \frac{1}{n_{..}}\sum_{j=1}^{l} n_{.j}y_j = \frac{38}{10} = 3,8$$

$$V(x) = \frac{1}{n_{..}}\sum_{i=1}^{k} n_{i.}x_i^2 - \bar{\bar{x}}^2 = \frac{196}{10} - (4,2)^2 = 1,96$$

$$V(y) = \frac{1}{n_{..}}\sum_{j=1}^{l} n_{.j}y_j^2 - \bar{\bar{y}}^2 = \frac{164}{10} - (3,8)^2 = 1,96$$

Caractéristiques conditionnelles

$$\bar{x}_1 = \frac{1}{n_{.1}}\sum_{i=1}^{3} n_{i1}x_i = \frac{14}{3} = 4{,}67 \qquad \bar{y}_1 = \frac{1}{n_{1.}}\sum_{j=1}^{3} n_{1j}y_j = \frac{10}{2} = 5$$

$$\bar{x}_2 = \frac{1}{n_{.2}}\sum_{i=1}^{3} n_{i2}x_i = \frac{20}{5} = 4 \qquad \bar{y}_2 = \frac{1}{n_{2.}}\sum_{j=1}^{3} n_{2j}y_j = \frac{16}{5} = 3{,}2$$

$$\bar{x}_3 = \frac{1}{n_{.3}}\sum_{i=1}^{3} n_{i3}x_i = \frac{8}{2} = 4 \qquad \bar{y}_3 = \frac{1}{n_{3.}}\sum_{j=1}^{3} n_{3j}y_j = \frac{12}{3} = 4$$

57

$$V_1(x) = \frac{1}{n_{.1}}\sum_{i=1}^{3} n_{i1}x_i^2 - \bar{x}_1{}^2 = \frac{68}{3} - (4{,}66)^2 = 0{,}96$$

$$V_2(x) = \frac{1}{n_{.2}}\sum_{i=1}^{3} n_{i2}x_i^2 - \bar{x}_2{}^2 = \frac{88}{5} - (4)^2 = 1{,}6$$

$$V_3(x) = \frac{1}{n_{.3}}\sum_{i=1}^{3} n_{i3}x_i^2 - \bar{x}_3{}^2 = \frac{40}{2} - (4)^2 = 4$$

$$V_1(y) = \frac{1}{n_{1.}}\sum_{j=1}^{3} n_{1j}y_j^2 - \bar{y}_1{}^2 = \frac{52}{2} - (5)^2 = 1$$

$$V_2(y) = \frac{1}{n_{2.}}\sum_{j=1}^{3} n_{2j}y_j^2 - \bar{y}_2{}^2 = \frac{56}{5} - (3{,}2)^2 = 0{,}96$$

$$V_3(y) = \frac{1}{n_{3.}}\sum_{j=1}^{3} n_{3j}y_j^2 - \bar{y}_3{}^2 = \frac{56}{3} - (4)^2 = 2{,}67$$

$$Cov(x,y) = \frac{1}{n_{..}}\sum_{i=1}^{3}\sum_{j=1}^{3} n_{ij}x_iy_j - \bar{x}\bar{y} = \left(\frac{1}{10}.156\right) - (4{,}2)(3{,}8) = -0{,}36$$

58

Calcul de la covariance et changement de variable

- Un changement de variable judicieux appliqué à x et à y permettra souvent une simplification supplémentaire des calculs. Choisissons pour x et y de nouvelles origines x_o et y_o et de nouvelles unités de mesure $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$, de façon à ce que les nouvelles variables x' et y' soient des nombres entiers plus simples que les valeurs originelles de x et y :

$$x_i' = \frac{x_i - x_0}{\alpha}, \qquad y_j' = \frac{y_j - y_0}{\beta}$$

59

$$x_i = \alpha x_i' + x_0, \qquad y_j = \beta y_j' + y_0$$

- En vertu des propriétés de la moyenne arithmétique, les mêmes relations existent entre les moyennes :

$$\bar{\bar{x}} = \alpha\bar{\bar{x}}' + x_0 \;;\; \bar{\bar{y}} = \beta\bar{\bar{y}}' + y_0$$

- Par soustraction :

$$x_i - \bar{\bar{x}} = \alpha\left(x'_i - \bar{\bar{x}}'\right) \quad ; y_j - \bar{\bar{y}} = \beta\left(y'_j - \bar{\bar{y}}'\right)$$

- Par conséquent :

$$\begin{aligned} Cov(x,y) &= \frac{1}{n_{..}}\sum_i\sum_j n_{ij}(x_i - \bar{\bar{x}})(y_j - \bar{\bar{y}}) \\ &= \alpha\beta\,\frac{1}{n_{..}}\sum_i\sum_j n_{ij}(x_i' - \bar{\bar{x}}')\beta(y_j' - \bar{\bar{y}}') \\ &= \alpha\beta\, Cov(x',y') \end{aligned}$$

60

Exemple 2

- Soit la distribution à deux caractères x et y suivante :

y / x	40	60	80
50	0	1	1
60	2	3	0
70	1	1	1

- En se servant des résultats de l'exemple 1, montrez que la covariance de x et de y est égale à - 18.
- On remarque que chaque valeur de x et de y peut s'écrire comme une combinaison linéaire des valeurs de x et de y dans l'exemple 1 :
- 50 = (5 x 2) + 40   40 = (10 x 2) + 20
- 60 = (5 x 4) + 40   60 = (10 x 4) + 20
- 70 = (5 x 6) + 40   80 = (10 x 6) + 20

- Par suite, les valeurs de x et de y dans l'exemple 1 (x_1, y_1)constituent des changements de variables.
- Ainsi, on peut écrire :
  - $x_2 = \alpha x_1 + x_0 \Longrightarrow x_1 = (x_2 - x_0)/\alpha$
  - $y_2 = \beta y_1 + y_0 \Longrightarrow y_1 = (y_2 - y_0)/\beta$
- avec $\alpha = 5$ et $\beta = 10$
- $x_0 = 40$ et $y_0 = 20$
- Par conséquent :
- $Cov(x_2,y_2) = \alpha\beta\, Cov(x_1,y_1) = 50.(-0,36) = -18.$