CHAPITRE I : LES SERIES STATISTIQUES A DEUX DIMENSIONS : DISTRIBUTIONS MARGINALES ET CONDITIONNELLES

Il est très courant que l'étude statistique ne porte pas que sur un seul caractère, mais plusieurs caractères simultanément pour une même population.

On étudie par exemple, un ensemble de salariés non plus seulement selon leur salaire, mais encore selon leur ancienneté. Les tableaux de données seront alors des tableaux à deux dimensions.

L'étude statistique peut porter en même temps sur deux caractères qui peuvent être de même nature ou de nature différente (qualitatif-quantitatif discret, qualitatif- quantitatif continu……)

Dans ce chapitre on présentera ces tableaux, ainsi que le traitement statistique associé,

I. Tableaux à double entrée

Les tableaux à double entrée présentant pour chaque couple de modalités des deux caractères étudiées, l'effectif des individus, présentant ces deux modalités simultanément.

1. Construction d'un tableau de contingence

Considérons une population de N unités statistiques décrites simultanément selon deux variables statistiques X et Y prenant les modalités (ou valeurs) suivantes :

X : x_1 , x_2 , ... , x_p

Y : y_1 , y_2 , ... , y_q .

Le nombre d'individus présentant les modalités x_i et y_j simultanément est noté : n_{ij} .

Le tableau est de la forme :

X \ Y	y_1	y_2	.....	y_j	.....	y_q	Colonne marginale $n_{i.}$
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}	$n_{p.}$
Ligne marginale $n_{.j}$	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.q}$	N=n..

On désigne par N l'effectif total de la population de référence.

$$N = \sum_i \sum_j n_{ij.} \qquad \begin{array}{l} i \in 1,\ p \\ j \in 1,\ q \end{array}$$

L'effectif $n_{i.}$ (total de la ligne i) est le nombre total d'individus présentant la modalité x_i du caractère X indépendamment des modalités de Y.

En effectuant la somme des termes de chaque ligne, on définit les termes ni.de la colonne marginale

$\sum_{j=1}^{q} nij = n_{i1}+n_{i2}+\ldots+n_{ij}+\ldots+n_{iq} = n_{i}$.

n_{i} . est l'effectif de la population qui présente les modalités x_i du caractère X, $\forall$ les modalités du caractère Y.

L'effectif $n_{.j}$ (total de la colonne j) est le nombre total d'individus présentant la modalité y_j du caractère Y indépendamment des modalités de X.

En effectuant la somme des termes de chaque colonne, on définit les termes n.j de la ligne marginale

$\sum_{i=1}^{p} nij = n._j = n_{1j} + n_{2j} + \ldots + n_{ij} + \ldots + n_{pj}$: c'est l'effectif de la population qui présente les modalités y_j du caractère Y, indépendamment du caractère X.
De plus :

$$\sum_{i=1}^{p} ni. = \sum_{j=1}^{q} n.j = n.. = N$$

On appelle fréquence du couple de modalités (x_i , y_j) (ou encore fréquence totale), la proportion f_{ij} d'individus présentant simultanément les deux modalités :

$$\left[f_{ij} = \frac{n_{ij}}{N} \right].$$

<u>Remarque</u> : la somme de toutes les fréquences de couples = 1.

<u>EXEMPLE</u>

Soit la répartition des salariés d'une entreprise selon le nombre d'enfant(X) et le salaire mensuel (Y) en 10^3DH.

Y (salaire) X (nb enfants)	2 - 6	6 - 10	10 - 16	Colonne marginale $n_{i.}$
1	15	8	2	25
2	13	4	1	18
3	11	3	3	17
Ligne marginale $n_{.j}$	39	15	6	60

n..= 60 Le nombre total des salariés de l'entreprise

n_{22}= 4 Salariés ont chacun 2 enfants et gagnent un salaire compris entre 6000 et 10000 DH

n_{13}= 2 Salariés ont chacun 1 enfant et gagnent un salaire compris entre 10000 et 16000 DH

$n_{2.}$=18 Salariés ont 2 enfants quelque soit leurs salaires

$n_{.2}$= 15 salariés ont un salaire compris entre 6000 et 10000 DH quelque soit leurs nombre d'enfants

$f_{22}=\frac{n22}{n_{..}}$ =4/60 ; $f_{2.}=\frac{n2.}{n_{..}}$ =18/60 ;$f_{.2}=\frac{n.2}{n_{..}}$ =15/60

2. Distributions marginales

Considérons la colonne de droite du tableau de contingence. Les effectifs $n_{i.}$ représentent les individus présentant la modalité x_i indépendamment des modalités du second caractère étudié Y.

On dit qu'ils définissent la **distribution marginale de X**.

(Cette série statistique est une série statistique à un seul caractère).

Série marginale de X : $(x_i\ ,\ n_{i.})i\in[1,\ p]$.

X	$n_{i.}$	$f_{i.}$
x_1	$n_{1.}$	$f_{1.}$
x_2	$n_{2.}$	$f_{2.}$
.	.	.
.	.	.
x_i	$n_{i.}$	$f_{i.}$
.	.	.
.	.	.
x_p	$n_{p.}$	$f_{p.}$
$\sum$	$n_{..}$	$f_{..}=1$

On définit alors la fréquence marginale de la modalité x_i par :

$\left[f_{i.} = \frac{n_{i.}}{N} \right]$. (On a : $\sum_{i=1}^{p} f_{i.} = 1$).

Exemple

Nombre d'enfant (X)	$n_{i.}$	$f_{i.}$
1	*25*	*0,417*
2	*18*	*0 ,3*
3	*17*	*0,283*
$\sum$	*60*	*1*

De la même façon on définit la **distribution marginale de Y** en considérant la dernière ligne du tableau.

Série marginale de Y : $(y_j\ ,\ n_{.j})i\in[1,\ q]$.

Y	n.j	$f_{.j}$
y_1	$n_{.1}$	$f_{.1}$
y_2	$n_{.2}$	$f_{.2}$
.	.	.
.	.	.
y_j	$n_{.j}$	$f_{.j}$
.	.	.
.	.	.
y_q	$n_{.q}$	$f_{.q}$
$\sum$	n..	f..=1

On définit la **fréquence marginale** de la modalité y_j :

$\left[f_{.j} = \frac{n_{.j}}{N} \right]$. (on a, bien évidemment, $\sum_{j=1}^{q} f_{.j} = 1$).

Exemple

salaires en 10^3DH(Y)	n.j	$f_{.j}$
2-6	*39*	*0,65*
6-10	*15*	*0 ,25*
10-16	*6*	*0,1*
$\sum$	*60*	*1*

3. Distributions conditionnelles

C'est une distribution suivant l'un des deux caractères (X,Y) liée par une modalité de l'autre caractère. Il y a autant de distributions conditionnelles X qu'il y a de j , et il y a autant de distributions conditionnelles Y qu'il y a de i.

- Considérons les $n_{.j}$ individus qui présentent la modalité y_j du caractère Y ; cette population est décrite par la $j^{ième}$ colonne du tableau suivant les modalités du caractère X.

On dit qu'elle définit la "**distribution conditionnelle de X sachant que Y = y_j**". Ou une distribution de X liée (conditionnée par) une valeur y_j de Y.

X/(Y= y_j)	$n_{i/j}$	$f_{i/j}$
x_1	n_{1j}	f_{1j}
x_2	n_{2j}	f_{2j}
.	.	.
.	.	.
x_i	n_{ij}	f_{ij}
.	.	.
.	.	.
x_p	n_{pj}	f_{pj}
$\sum$	n.j	f.j

On peut définir les **fréquences conditionnelles** associées par : $\left[f_{i/j} = \frac{n_{ij}}{n._j} \right]$; $f_{i/j}$ représente la proportion d'individus présentant la modalité x_i parmi l'ensemble des individus présentant la modalité y_j de Y.

$$\sum_{i=1}^{p} fi/j = 1$$

Exemple

Distribution conditionnelle de X sachant que Y = y_2 (distribution conditionnelle de X liée par y_2)

Nombre d'enfant (X)	$n_{i/2}$ (y_2(6-10))	$f_{i/2}$
1	*8*	*0,53*
2	*4*	*0 ,27*
3	*3*	*0,2*
$\sum$	*15*	*1*

- De façon analogue, en considérant les $n_{i.}$ individus présentant la modalité x_i du caractère X on définit la "**distribution conditionnelle de Y sachant que X = x_i**".

(ième ligne du tableau). Ou une distribution de Y liée (conditionnée par) une valeur x_i de X

Y	$n_{j/i}$	$f_{j/i}$
y_1	n_{i1}	f_{i1}
y_2	n_{i2}	f_{i2}
.	.	.
.	.	.
y_j	n_{ij}	f_{ij}
.	.	.
.	.	.
y_q	n_{iq}	f_{iq}
$\sum$	ni.	fi.

Les fréquences conditionnelles associées sont définies par :

$\left[f_{j/i} = \frac{n_{ij}}{n_{i.}} \right]$; $f_{j/i}$ représente la proportion d'individus présentant la modalité y_j parmi l'ensemble des individus présentant la modalité x_i de X.

$$\sum_{j=1}^{q} fj/i = 1$$

Exemple :

Distribution conditionnelle de Y sachant que X = x_3 (distribution conditionnelle de Y liée par x_3)

salaires en 10^3DH(Y)	$n_{j/3}$	$f_{j/3}$
2-6	*11*	*0,65*
6-10	*3*	*0 ,175*
10-16	*3*	*0,175*
$\sum$	*17*	*1*

Remarque : Les séries conditionnelles ; comme les séries marginales, sont des séries statistiques à un seul caractère.

- Il est possible de synthétiser dans un seul tableau toutes les séries conditionnelles et marginales. Dans les cases centrales du tableau de contingence, on indique en plus des effectifs n_{ij} les fréquences conditionnelles $f_{i/j}$ et $f_{j/i}$ (appelées parfois "fréquences en colonnes et "fréquences en

lignes"). Dans les marges du tableau, on indique en plus des effectifs marginaux $n_{i.}$ et $n_{.j}$, les fréquences marginale $f_{i.}$ et $f_{.j}$.

Exemple : Considérons la répartition de 2.000 salariés d'une entreprise selon leur âge (en années) (X) et leur salaire mensuel net (en milliers de dirhams) (Y) :

Salaire Y Âge X	[1 1,2[	[1,2 1,4[	[1,4 1,6[	Marge
Effectif [20 25[% ligne % colonne	250 62,5% 41,7%	100 25% 11,6%	50 12,5% 9,3%	400 20%
[25 35[	200 25% 33,3%	400 50% 46,5%	200 25% 37%	800 40%
[35 45[	140 20% 23,3%	350 50% 40,7%	210 30% 38,9%	700 35%
[45 60[	10 10% 1,7%	10 10% 1,2%	80 80% 14,8%	100 5%
Marge	600 30%	860 43%	540 27%	2.000

On peut lire par exemple : (2ème ligne, 3ème colonne) 25% des gens âgés de 25 à 35 ans, ont un salaire compris entre 1 400 DH et 1 600 DH.
37% des gens ayant un salaire compris entre 1 400 DH et 1 600 DH ont entre 25 et 35 ans.

Remarque : Relations entre les fréquences

On a défini les fréquences de couples, marginales, et conditionnelles de la façon suivante :

$$f_{ij} = \frac{n_{ij}}{N}; f_{i.} = \frac{n_{i.}}{N}; f_{.j} = \frac{n_{.j}}{N}$$

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

On remarque alors les relations suivantes :

$$\left[f_{ij} = f_{i/j} \times f_{.j} = f_{j/i} \times f_{i.} \right].$$

II. caractéristiques des séries à deux variables

Dans le cas où les variables X et Y sont des variables quantitatives, on peut associer à chacune des séries conditionnelles et des séries marginales définies par le tableau de contingence des caractéristiques de tendance centrale et de dispersion

On considère un tableau de contingence comme celui défini au paragraphe **I**:

X prend les valeurs x_1 , x_2 , ... , x_p ;

Y prend les valeurs y_1 , y_2 , ... , y_q .

(Les x_i ou y_i sont les centres de classes dans le cas où X et Y sont des variables quantitatives continues).

1. Caractéristiques Marginales

1-1 Moyennes marginales

Moyenne marginale de X

La moyenne marginale de X notée $\bar{\bar{X}}$ correspond à la valeur moyenne du caractère X possédée par les individus de la population indépendamment du caractère Y

la moyenne marginale : $$\bar{\bar{X}} = \frac{\sum_{i=1}^{p} n_{i..} x_i}{N} = \frac{1}{n..} \sum_{i=1}^{p} n_{i.} x_i = \sum_{i=1}^{P} f_{i.} x_i \ ;$$

Exemple :

x_i	$n_{i.}$	$n_{i.}x_i$
1	25	25
2	18	36
3	17	51
	60	112

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{p} n_{i..} x_i}{N} = \frac{1}{n..} \sum_{i=1}^{p} n_{i.} x_i = \frac{112}{60} = 1{,}86$$

Moyenne marginale de Y

La moyenne marginale de Y notée $\bar{\bar{Y}}$ correspond à la valeur moyenne du caractère Y possédées par les individus de la population indépendamment du caractère X

la moyenne marginale : $$\bar{\bar{Y}} = \frac{\sum_{J=1}^{q} n_{.j} y_j}{N} = \frac{1}{n..} \sum_{J=1}^{q} n_{.j} y_j = \sum_{j=1}^{q} f_{.j} y_j$$

Exemple :

y_j	$n_{.j}$	$n_{.j}y_j$
4	39	156
8	15	120
13	6	78
	60	354

$$\overline{\overline{Y}} = \frac{\sum_{J=1}^{q} n_{.j} y_j}{N} = \frac{1}{n_{..}} \sum_{J=1}^{q} n_{.j} y_j = \frac{354}{60} = 5{,}9$$

1-2 Variances marginales

Variance marginale de X

La variance marginale de X notée V(x) est une mesure de la dispersion des individus de la population selon X et indépendamment de Y

- la **variance marginale** : $V(X) = \frac{\sum_{i=1}^{p} n_{i.}\left(x_i - \overline{\overline{X}}\right)^2}{N} = \frac{\sum_{i=1}^{p} ni.x_i^{\,2}}{N} - \overline{\overline{X}}^2$.

L'écart type marginale de X

$$\left[\sigma_X = \sqrt{V(X)}\right]$$

Exemple

x_i	$n_{i.}$	x_i^2	$n_{i.}x_i^2$
1	25	1	25
2	18	4	72
3	17	9	153
	60	-	250

Variance marginale de X $V(X) = \frac{\sum_{i=1}^{p} ni.x_i^{\,2}}{N} - \overline{\overline{X}}^2 = 250/60 - 1{,}86^2 = 0{,}71$

L'écart type marginale de X $\sigma_X =$

Variance marginale de Y

La variance marginale de Y notée V(y) est une mesure de la dispersion des individus de la population selon Y et indépendamment de X

- la **variance marginale** : $V(Y) = \frac{\sum_{j=1}^{q} n_{.j}\left(y_j - \overline{\overline{Y}}\right)^2}{N} = \frac{\sum_{j=1}^{q} n.jy_j^{\,2}}{N} - \overline{\overline{Y}}^2$.

Exemple

y_j	$n_{.j}$	y_j^2	$n_{.j}y_j^2$
4	39	16	624
8	15	64	960
13	6	169	1014
	60	-	2598

Variance marginale de Y $V(Y) = \dfrac{\sum_{j=1}^{q} n.jy_j^{\,2}}{N} - \overline{\overline{Y}}^{2} = 2598/60 - 5{,}9^2 = 8{,}49$

L'écart type marginale de Y

$\left[\sigma_Y = \sqrt{V(Y)}\right]$

2. Caractéristiques Conditionnelles

2-1 Moyennes Conditionnelles

On peut, de la même façon définir les caractéristiques de toutes les séries conditionnelles (toujours dans le cas où X et Y sont quantitatives).

Il y a (p + q) séries conditionnelles associées au tableau : p séries conditionnelles "Y sachant X = x_i" et q séries "X sachant Y = y_j".

Moyennes conditionnelles de X sachant Y=y_j

la **moyenne conditionnelle de "X sachant Y = y_j"** : $\overline{X}_j = \dfrac{\sum_{i=1}^{p} n_{ij}\, x_i}{n._j}$;

Exemple : calculons les moyennes conditionnelles $\overline{X}_1$ $\overline{X}_2$ $\overline{X}_3$

x_i	$Y=y_1$	$n_{i1}x_i$
1	15	15
2	13	26
3	11	33
	39	74

$$\overline{X}_1 = \frac{\sum_{i=1}^{p} n_{i1} x_i}{n._1} = 74/39 = 1{,}89$$

x_i	$Y=y_2$	$n_{i2}x_i$
1	8	8
2	4	8
3	3	9
	15	25

$$\overline{X}_2 = \frac{\sum_{i=1}^{p} n_{i2} x_i}{n._2} = 25/15 = 1{,}66$$

x_i	$Y=y_3$	$n_{i3}x_i$
1	2	2
2	1	2
3	3	9
	6	13

$$\overline{X}_3 = \frac{\sum_{i=1}^{p} n_{i3} x_i}{n_{.3}} = 13/6 = 2{,}16$$

Moyennes conditionnelles de Y sachant X=x_i

- la **moyenne conditionnelle de "Y sachant X = x_i"** : $\overline{Y_i} = \frac{\sum_{j=1}^{q} n_{ij}\ Y_j}{n_{i.}}$.

Exemple : calculons les moyennes conditionnelles $\overline{Y_1}$ $\overline{Y_2}$ $\overline{Y_3}$

y_j	$X=x_1$	$n_{1j}\ y_j$
4	15	60
8	8	64
13	2	26
	25	150

$$\overline{Y_1} = \frac{\sum_{j=1}^{q} n_{1j} y_j}{n1.} = \frac{150}{25} = 6$$

y_j	$X=x_2$	$n_{2j}\ y_j$
4	13	52
8	4	32
13	1	13
	18	97

$$\overline{Y_2} = \frac{\sum_{j=1}^{q} n_{2j} y_j}{n2.} = \frac{97}{18} = 5{,}38$$

y_j	$X=x_3$	$n_{3j}\ y_j$
4	11	44
8	3	24
13	3	39
	17	107

$$\overline{Y_3} = \frac{\sum_{j=1}^{q} n_{3j} y_j}{n3.} = \frac{107}{17} = 6{,}29$$

2-2 Variances Conditionnelles

Variances conditionnelles associées à X

$$V_j(X) = \frac{\sum_{i=1}^{p} n_{ij}\left(x_i - \overline{X}_j\right)^2}{n_{.j}} = \frac{\sum_{i=1}^{p} n_{ij} x_i^{\,2}}{n_{.j}} - \overline{X}_j^{\,2}$$

On peut donc calculer j variances conditionnelles associées à X

Exemple

Calculons $V_2(X)$

x_i	$Y=y_2$	x_i^2	$n_{i21}x_i^2$
1	8	1	8
2	4	4	16
3	3	9	27
	15		51

$$V_2(X) = \frac{\sum_{i=1}^{P} n_{i2}xi^2}{n.2} - \bar{X}_2^{\,2} = \frac{51}{15} - 1{,}66^2 = 0{,}645$$

Variances conditionnelles associées à Y

$$V_i(Y) = \frac{\sum_{j=1}^{q} n_{ij}\left(y_j - \bar{Y}_i\right)^2}{n_{i.}} = \frac{\sum_{j=1}^{q} n_{ij}\, y_j^{\,2}}{n_{i.}} - \bar{Y}_i^{\,2} \ .$$

On peut donc calculer i variances conditionnelles associées à Y

Exemple

Calculons $V_2(Y)$

y_j	$X=x_2$	y_j^2	$n_{2j}\, y_j^2$
4	13	16	208
8	4	64	256
13	1	169	169
	18		633

$$V_2(Y) = \frac{\sum_{j=1}^{q} n_{2j}yj^2}{n2.} - \bar{Y}_2^{\,2} = \frac{633}{18} - 5{,}38^2 = 6{,}21$$

Remarque : Relations entre caractéristiques marginales et conditionnelles.

La distribution marginale de X résulte en fait du mélange des distributions conditionnelles "X sachant Y = y_j" représentées en effectifs $n_{.j}$ et la distribution marginale de Y résulte du mélange des distributions conditionnelles "Y sachant X = x_i" représentées en effectifs $n_{i.}$.

On en déduit alors les **relations** suivantes **entre les moyennes** :

La moyenne marginale est égale à la moyenne des moyennes conditionnelles pondérées par les effectifs marginaux :

$$\left[\ \bar{\bar{X}} = \frac{\sum_{j=1}^{q} n_{.j}\,\bar{X}_j}{N} \quad ; \quad \bar{\bar{Y}} = \frac{\sum_{i=1}^{p} n_{i.}\,\bar{Y}_i}{N}\ \right]$$

et les **relations** suivantes **entre les variances** :

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

$$\left[V(X) = \frac{\sum_{j=1}^{q} n_{.j} V_j(X)}{N} + \frac{\sum_{j=1}^{q} n_{.j} \left(\overline{X}_j - \overline{\overline{X}}\right)^2}{N} \right] . \left[V(Y) = \frac{\sum_{i=1}^{p} n_{i.} V_i(Y)}{N} + \frac{\sum_{i=1}^{p} n_{i.} \left(\overline{Y}_i - \overline{\overline{Y}}\right)^2}{N} \right] .$$

On peut traduire ces dernières relations en disant que l'hétérogénéité de la distribution marginale résulte à la fois de :

- l'hétérogénéité propre à chaque distribution conditionnelle ;
- l'hétérogénéité des moyennes conditionnelles entre elles.

Indépendance et liaisons de deux caractères (X,Y) :

- Indépendance : les caractères X et Y sont indépendants si les moyennes conditionnelles $\overline{X}_j$ sont égales entre elles d'une part et égales à la moyenne marginale $\overline{\overline{X}}$ ($\overline{X}_1 = \overline{X}_2 = \ldots = \overline{X}_j = \overline{\overline{X}}$) ou ($\overline{Y}_1 = \overline{Y}_2 = \ldots = \overline{Y}_i = \overline{\overline{Y}}$)
  De même, ils sont indépendants si on vérifie : $n_{ij} = \frac{n_{i.} X n_{.j}}{n_{..}}$ ou $f_{ij} = f_{i.} \times f_{.j}$
- Liaison entre X et Y : les deux caractères sont liés si les égalités précédentes ne sont pas vérifiées

3. La Covariance

On associe aux deux caractères quantitatifs X et Y une caractéristique globale appelée la covariance et définie par :

$$\left[Cov(X,Y) = \frac{\sum_i \sum_j n_{ij} \left(x_i - \overline{\overline{X}}\right)\left(y_j - \overline{\overline{Y}}\right)}{N} \right] .$$

On montre que l'on peut calculer la covariance par la formule :

$$Cov(X,Y) = \frac{\sum_i \sum_j n_{ij} x_i y_j}{N} - \overline{\overline{X}}.\overline{\overline{Y}}$$