3.7 Caractéristiques de dispersion

3.7.1 L'étendue

L'étendue E est la longueur de l'intervalle sur laquelle sont réparties les valeurs de la série, c'est donc la différence entre la plus grande et la plus petite valeur de la série.

$$E = Max\,(x_i) - Min\,(x_i) \tag{3.6}$$

.

Si la série est représentée en classes de valeurs, on prend comme étendue la différence entre la borne supérieure e_p de la dernière classe $[e_{p-1}; e_p[$ et la borne inférieure de la première classe $[e_1; e_2[$

$$E = e_p - e_1 \tag{3.7}$$

3.7.2 Quantile

Un quantile d'ordre $0 < \alpha < 1$, noté x_α, d'une série quantitative, formée de N observations, est un nombre noté x_α, tel que, à peu prés, αN des observations formant la suite sont inférieures à ce quantile x_α, et le reste, c'est-à-dire à peu prés $(1-\alpha)\,N$ observations, sont supérieures à x_α.

1er quartile

Le premier quartile, noté Q_1,est le quantile d'ordre $\alpha = \frac{1}{4}$.

3eme quartile

Le troisième quartile noté Q_3 est un nombre tel que 25% des observations lui sont supérieures (donc situées à sa droite).

L'étendue interquartile

L'étendu interquartile noté I_Q est la différence entre le troisième quartile Q_3 et le premier quartile Q_1 :

$$I_Q = Q_3 - Q_1 \tag{3.8}$$

Décile

Le premier décile est le quantile d'ordre $\alpha = \frac{1}{10}$, le deuxième décile est le quantile d'ordre, $\frac{2}{10}, ...$

Centiles

Le premier centile est le quantile d'ordre $\alpha = \frac{1}{100}$, le deuxième décile est le quantile d'ordre, $\frac{2}{100}, ...$

3.7.3 Calcul des quartiles : Données discrètes

L'effectif total N n'est pas toujours un multiple de quatre, et ainsi le un quart de l'effectif total n'existe pas toujours. Ainsi, on calcule approximativement le premier et le troisième quartile de la façon suivante.

Calcul de Q_1.

Si $\frac{N}{4}$ est un entier naturel, Q_1 est l'observation numéro $\frac{N}{4}$

$$Q_1 = \left(\frac{N}{4}\right)^{ème} \ obs. \tag{3.9}$$

Si $\frac{N}{4}$ n'est pas un entier naturel, Q_1 est l'observation correspondant au numéro qui vient juste aprés le nombre décimal $\frac{N}{4}$.

Calcul de Q_3.

Si $\frac{3N}{4}$ est un entier naturel, Q_3 est l'observation numéro $\frac{3N}{4}$

$$Q_3 = \left(\frac{3N}{4}\right)^{ème} \ obs. \tag{3.10}$$

Si $\frac{3N}{4}$ n'est pas un entier naturel, Q_3 est l'observation correspondant au numéro qui vient juste aprés le nombre décimal $\frac{3N}{4}$.

3.7.4 Exemple 1. (données en vrac).

1)- Soit la série

$$2\ 2\ 3\ 4\ 5\ 6\ 6\ 6\ 7\ 7\ 7\ 8\ 8\ 10$$

L'effectif total est

$$N = 14 \text{ est pair.}$$

La médiane M_e est la moyenne des deux observations $\left(\frac{N}{2}\right)^{ème}$ et $\left(\frac{N}{2}+1\right)^{ème}$:

$$M_e = \frac{6+6}{2} = 6.$$

Puisque $\frac{N}{4} = \frac{14}{4} = 3,5$ n'est pas un nombre entier et également $\frac{3N}{4}$, le premier quartile Q_1 et le troisième quartile Q_3 sont

$$Q_1 = 4^{ème} \text{ obs.} = 4. \text{ et } Q_3 = 11^{ème} \text{ obs.} = 7$$

et ainsi

$$I_Q = 7 - 4 = 3.$$

3.7.5 Exemple 2. (données groupées en valeurs).

Soit la série

Valeurs (x_i)	-2	0	1,5	4	5	10	11	15
effectifs (n_i)	10	2	40	8	5	15	5	15

Pour trouver les valeurs d'observations des rangs concernés, on doit calculer les effectifs cummulés, en ajoutant une autre ligne au tableau des données :

Valeurs (x_i)	-2	0	1,5	4	5	10	11	15	total
effectifs (n_i)	10	2	40	8	5	15	5	15	100
effectifs cummulés (n_i cum)	10	12	52	60	65	80	85	100	

L'effectif total est

$$N = 100, \text{ il est pair.}$$

la mediane M_e est

$$M_e = \frac{1,5 + 1,5}{2} = 1,5.$$

Le premier quartile Q_1 est

$$Q_1 = \left(\frac{100}{4} = 25\right)^{ème} obs. = 1,5.$$

Le troisième quartile Q_3 est

$$Q_3 = \left(\frac{3 \times 100}{4} = 75\right)^{ème} obs. = 10.$$

L'intervalle interquartile I_Q est

$$I_Q = Q_3 - Q_1 = 10 - 1,5 = 8,5.$$

3.7.6 Calcul des quartiles : Données groupées en classes de valeurs.

Le premier et le troisième quartiles se calcule, comme la médiane, par la méthode d'interpolation linéaire.

Calcul de Q_1.

On détermine la classe $[e_{i-1}, e_i[$ du premier quartile Q_1, qui est la classe contenant l'observation numéro $\frac{N}{4}$ (ou le numéro qui vient juste aprés le nombre décimal $\frac{N}{4}$) et on applique la formule

$$Q_1 = e_{i-1} + k\frac{\frac{N}{4} - n_{i-1}cum}{n_i cum - n_{i-1}cum}, \tag{3.11}$$

où

e_{i-1} est la borne inférieure de la classe du premier quartile,
k est l'amplitude de la même classe.

Calcul de Q_3.

On détermine la classe $[e_{i-1}, e_i[$ du troisième quartile Q_3, qui est la classe contenant l'observation numéro $\frac{3N}{4}$ (ou le numéro qui vient juste aprés le nombre dicimal $\frac{3N}{4}$) et on applique la formule

$$Q_3 = e_{i-1} + k\frac{\frac{3N}{4} - n_{i-1}cum}{n_i cum - n_{i-1}cum}, \tag{3.12}$$

où

e_{i-1} est la borne inférieure de la classe du troisième quartile,
k est l'amplitude de la même classe.

3.7.7 Exemple 3.

On considère la série définie par le tableau suivant.

Classe C_i	$[-4;-2[$	$[-2;0[$	$[0;2[$	$[2;4[$	$[4;6[$	$[6;8[$	total
effectif n_i	6	4	8	5	10	1	34

L'effectif total est N=34.
L'etendu E est

$$E = 8 - (-4) = 12.$$

Pour les quartiles, et les quantiles en général, on ajoute au tableau une ligne pour les effectifs cummulés :

Classe C_i	$[-4;-2[$	$[-2;0[$	$[0;2[$	$[2;4[$	$[4;6[$	$[6;8[$	total
effectif n_i	6	4	8	5	10	1	34
effectifs cummulés $n_i cum$	6	10	18	23	33	34	

a)-La médiane :
$\frac{N}{2} = \frac{34}{2} = 17$, la classe médiane est la classe qui contient l'observation numéro 17, c'est donc la classe $[e_{i-1}; e_i[= [0;2[$. L'amplitude de classe est $k = 2$.

La médiane M_e est

$$\begin{aligned} M_e &= e_{i-1} + k\frac{\frac{N}{2} - n_{i-1}cum}{n_i cum - n_{i-1}cum} \\ &= 0 + 2\frac{17-10}{18-10} \\ &= 1.75. \end{aligned}$$

b)- Le premier quartile Q_1 :

$\frac{N}{4} = 8.5$, la classe du premier quartile est celle contenant l'observation numéro 9, c'est la classe $[-2; 0[$. Le premier quartile Q_1 est

$$\begin{aligned} Q_1 &= e_{i-1} + k\frac{\frac{N}{4} - n_{i-1}cum}{n_i cum - n_{i-1}cum} \\ &= -2 + 2\frac{8.5-6}{10-6} \\ &= -0.75. \end{aligned}$$

c)- Le troisième quartile Q_3 :

$\frac{3N}{4} = 25.5$, la classe du troisième quartile est celle contenant l'observation numéro 26, c'est la classe $[4; 6[$. Le premier quartile Q_3 est

$$\begin{aligned} Q_3 &= e_{i-1} + k\frac{\frac{3N}{4} - n_{i-1}cum}{n_i cum - n_{i-1}cum} \\ &= 4 + 2\frac{25.5-23}{33-23} \\ &= 4.5. \end{aligned}$$

d)- L'intervalle interquartile I_Q

$$\begin{aligned} I_Q &= Q_3 - Q_1 \\ &= 4.5 - (-0.75) \\ &= 5.25. \end{aligned}$$

3.7.8 Boite à moustache

Définition.

La boite à moustache (ou box plot) est un graphique qui résume la dispersion d'une série statistique. Elle est définie par cinq nombres, la valeur minimale et la valeur maximale (les moustaches), les deux quartiles Q_1, Q_2 et la médiane M_e (la moustache).

Exemple.

Considérons les trois séries statistiques A, B et C suivantes.
A : 1, 1.5, 1.5, 2, 2.5, 3, 3, 3, 3, 3.4, 4.5, 4, 4.8, 5, 5, 5.
B : 1, 1, 1, 1, 1, 1.2, 1.2, 1.3, 2, 2.3, 2.4, 2.5, 3, 4, 4, 5, 5.
C : 1, 3, 3, 3, 3.2, 3.2, 3.4, 3.5, 4.5, 4.6, 4.7, 5, 5, 5, 5, 5.
Les boites à moustaches des trois séries sont représentées dans (Figure 1)
.

Utilité de la boite à moustache.

La boite à moustache peut être utilisée pour comparer la dispersion de séries, relativement à la médiane.

On obtient de la boite à moustaches des trois séries A, B, C, représenté dans Figure 1, une idée sur la dispersion des observations de chaque série, relativement à la médiane.

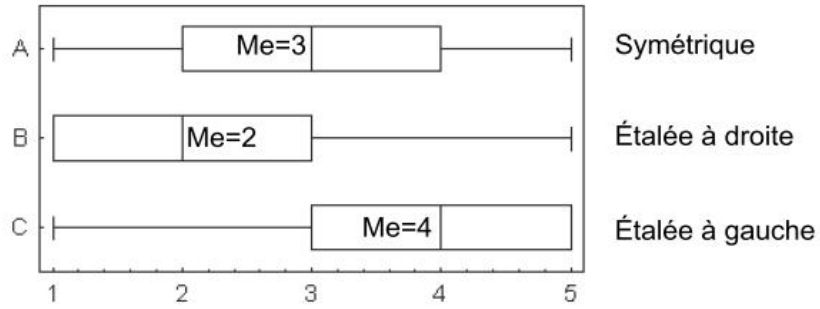

FIG. 3.1 – Les boites à moustaches des trois séries statistiques A, B et C.

3.7.9 Variance et Ecart-type

La variance et l'écart-type donne une idée sur l'éloignement des valeurs par rapport à la valeur moyenne.

Définition. La variance σ^2 d'une série statistique $\{x_1, ...x_N\}$ de moyenne m est définie par

$$\sigma^2 = \frac{\sum_{i-1}^{N} (x_i - m)^2}{N}. \tag{3.13}$$

L'écart-type σ est défini par

$$\sigma = \sqrt{\sigma^2}. \tag{3.14}$$

Si les données de la série (on dit aussi la distribution statistique) sont groupées en valeurs $\{(n_{1;}x_1), ..., (n_p; x_p)\}$, la définition (3.13) devient

$$\sigma^2 = \frac{\sum_{i-1}^{p} n_i (x_i - m)^2}{\sum_{i-1}^{p} n_i}. \tag{3.15}$$

Si les données de la série (on dit aussi la distribution statistique) sont groupées en classes de valeurs, la variance et l'écart-type sont définies par les mêmes formules où x_i est le centre de la classe d'effectif $n_{i.}$.

3.7.10 Théorème de Konig-Huygens

On peut démontrer que l'expression (3.15) définissant la variance est algébriquement équivalente à l'expression suivante, appelée formule de Konig-Huygens.

$$\sigma^2 = \frac{\sum_{i-1}^{p} n_i x_i^2}{\sum_{i-1}^{p} n_i} - m^2. \tag{3.16}$$

Coefficient de variation

Le coefficient de variation C_v est le rapport entre l'écart-type et la moyenne.

$$C_v = \frac{\sigma}{m}. \tag{3.17}$$

Le coefficient de variation donne une meilleure idée de la dispersion des valeurs de la série que l'écart-type, car c'est une valeur relative et non une valeur absolue.

3.7.11 Exemples

Exemple 1.
Soit la série statistique
1.5 ; 2.0 ; 3 ; 4 ; 5 ;8 ; 8 ; 9 ; 10 :12 ; 12.5 ; 15.
L'effectif total est $N = 12$.

Puisque les valeurs des mêmes observations ne se répète pas beaucoup, il n'est pas necéssaire d'utiliser la représentation des données groupées en valeurs.

$$\begin{aligned} m &= \frac{\sum_{i=1}^{12} x_i}{N} \\ &= \frac{1.5 + 2.0 + ... + 12.5 + 15}{12} \\ &= \frac{90}{12} = 7.5. \end{aligned}$$

$$\begin{aligned} \sigma^2 &= \frac{\sum_{i=1}^{12} x_i^2}{N} - m^2 \\ &= \frac{1.5^2 + 2.0^2 + ... + 12.5^2 + 15^2}{12} - (7.5)^2 \\ &\simeq 17.96. \end{aligned}$$

$$\begin{aligned} \sigma &= \sqrt{\sigma^2} \\ &= \sqrt{17.96} \\ &\simeq 4.24. \end{aligned}$$

$$\begin{aligned} C_V &= \frac{\sigma}{m} \simeq \frac{4.24}{7.5} \\ &\simeq 0.56. \end{aligned}$$

Exemple 2.
Soit la série
-5 ; -8 ; -3 ; 0 ; 1 ; 0 ; 1 ; 2 ; 2 ; -3 ; 4 ; 0 ; 1 ; 2 ; -3 ; -3 ; 1 ;0 ; 0 ; 0 ; 0 ; 0 ; 0
-8 ; -8 ; -3 ; 2 ; 4 ; 4 ; 2 ; 3 ; -5 ; -5 ; 2 ; 2 ; 2 ; 4 ; 2 ; 3 ; 1 ; 5 ; 0 ; 0
Puisque beaucoup d'observations ont la même valeur, on utilise la représentations des données groupées en valeurs

x_i	-8	-5	-3	0	1	2	3	4	5	total
n_i	3	3	5	11	5	9	2	4	1	43

Le nombre des valeurs est $p = 9$.
Le nombre d'observations (c'est-à-dire l'effectif total) est $N = 43$.
La moyenne m est donnée par

$$\begin{aligned} m &= \frac{\sum_{i=1}^{9} n_i x_i}{\sum_{i=1}^{9} n_i} \\ &= \frac{3(-8) + 3(-5) + ... + 4(4) + 1(5)}{3 + 3 + ... + 4 + 1} \\ &= -\frac{4}{43} \\ &\simeq -0.093. \end{aligned}$$

La variance σ^2 et l'écart type σ sont donnés par

$$\begin{aligned} \sigma^2 &= \frac{\sum_{i=1}^{9} n_i x_i^2}{\sum_{i=1}^{9} n_i} - m^2 \\ &= \frac{3(-8)^2 + 3(-5)^2 + ... + 4(4)^2 + 1(5)^2}{3 + 3 + ... + 4 + 1} - (-0.093)^2 \\ &\simeq 10.689. \end{aligned}$$

$$\sigma = \sqrt{\sigma^2} \simeq \sqrt{10.689} \simeq 3.269$$

Le coefficient de variation C_V est égale à

$$C_V = \frac{\sigma}{m} \simeq \frac{3.269}{-0.093} = -35.146.$$

On peut utiliser un tableau statistique pour y faire les calculs, de la façon suivante.

x_i	-8	-5	-3	0	1	2	3	4	5	*total*
n_i	3	3	5	11	5	9	2	4	1	43
$n_i x_i$	-24	-15	-15	0	5	18	6	16	5	-4
$n_i x_i^2$	192	75	45	0	5	36	18	64	25	460

Et ainsi,

$$m = \frac{-4}{43} \simeq -0.093,$$
$$\sigma^2 \simeq \frac{460}{43} - (-0.093)^2 \simeq 10.689.$$

Exemple 3.
Soit la série suivante, dont les données sont groupées en classes de valeurs.

classe C_i	$[-2;0[$	$[0;2[$	$[2;4[$	$[4;6[$	$[6,8[$	
effectif n_i	5	0	6	2	4	

Pour calculer la moyenne m, la variance σ^2, l'écart type σ et le coefficient de variation C_V, on calcule les centres x_i des classes C_i et on procède de la même façon que dans le cas des données groupées en valeurs, c'est-à-dire comme dans l'exemple précédent. On peut également faire les calculs sur le tableau statstique :

classe C_i	$[-2;0[$	$[0;2[$	$[2;4[$	$[4;6[$	$[6,8[$	total
effectif n_i	5	0	6	2	4	17
centre de classe x_i	-1	1	3	5	7	
$n_i x_i$	-5	0	18	10	28	51
$n_i x_i^2$	5	0	54	50	196	305

Ainsi, on a

$$m = \frac{\sum_{i=1}^{5} n_i x_i}{\sum_{i=1}^{5} n_i} = \frac{51}{17} = 3.$$
$$\sigma^2 = \frac{\sum_{i=1}^{5} n_i x_i^2}{\sum_{i=1}^{5} n_i} - m^2 = \frac{305}{17} - (3)^2 \simeq 8.94.$$

Chapitre 4

Séries statistiques Doubles

4.1 Vocabulaire

4.1.1 Definition

une série statistique double intervient quand on considère deux variables statistiques X et Y en même temps dans le but de chercher une relation les liant. Les éléments d'une série statistique double sont des binômes (x, y), où x et y sont des valeurs de X et Y respectivement. Une série double est quantitative si les deux variables statistiques X et Y sont quantitatives, elles est qualitative si X et Y sont qualitatives et elle est mixte si l'une des deux variables est quantitative et l'autre et qualitative.

Pratiquement, on considère un echantillon de N individus et on mesure simultanément la valeur x du caractère X et la valeur y du caractère Y sur chaque individu de l'échantillon, on obtient ainsi N couples (x_1, y_1), (x_2, y_2), ..., (x_N, y_N). On note qu'un couple peut être obtenu plus qu'une fois.

4.1.2 Notation

On note $x_1, x_2, ..., x_k$ les k valeurs ou modalités du caractère X et y_1, y_2, ... , y_l les l valeurs ou modalités du caractère Y.

On note n_{ij} l'effectif du couple (x_i, y_j).

4.1.3 Exemple

Dans le but d'étudier la relation entre la taille et le poids d'un groupe d'adolescents, on a pris un échantillon de 18 adolescents et on a mésuré la taille et le poids de chacun. On a obtenu la série quantitative suivante, où la première composante est la taille et la deuxième est le poids.

$\{(155; 46, 1)$;(140 ; 38,2) ; (161 ; 44,3) ; (148; 38, 2) ; (155; 50, 5) ; (123; 22, 4) ; (160; 40, 4) ;

(140; 34, 7) ; (165; 50, 5) ; (172; 50, 5) ; (155; 38, 1) ; (160; 57, 3) ; (142; 39, 3) ; (157; 46, 1) ;

(142; 37, 1) ; (167; 60) ;(148; 45, 9) ; (165; 50, 5)$\}$.

Cette série double est quantitative, puisque ses deux composantes sont quantitatives.

4.1.4 Exemple

Les données concernant le sexe et l'activité professionnelle de 20 personnes sont présentées en couples. La première composante est le sexe avec deux modalités mâle (M) ou femelle (F), la deuxième est l'activité professionnelle avec trois modalités chomeur (C), actif occupé (AO) et inactif (I). Ces données obtenues par questionnaire sont comme suit

{{F ; AO} ; {M ; I} ; {F ; C} ; {F ; C} ; {M ; AO} ; {M ; AO} ; {M ; C} ; {F ; I} ; {F ;I } ; {F ;I } ; {M ; C} ;

{F ; AO} ; {F ; AO} ; {F ; AO} ; {M ; AO} ; {M ; C} ; {M ; AO} ; {F ; I} ; {F , C} ; {M , AO}}.

Cette série double est qualitative, puisque ses deux composantes sont qualitatives.

4.2 Variable indépendante Versus Variable dépendante

4.2.1 Définition.

La notion de variable statistique indépendante et variable statistique dépendante est évoquée lorsqu'on veut étudier l'influence d'une variable (ou caractère) statistique sur une autre variable statistique. Généralement, l'expérimentateur agit sur les variations d'une variable statistique dite indépen-

dante et mesure l'effet de ces variations sur une deuxième variable statistique, dite dépendante. Dans une série statistique double (X, Y), généralement, la première composante X est la variable indépendante et la deuxième composante est la variable dépendante.

4.2.2 Exemple.

Si on compare les hommes et les femmes quant à leur satisfaction au travail dans une usine, la variable indépendante c'est le sexe et la variable dépendante c'est la satisfaction au travail.

La question de recherche statistique est : Quele est l'effet de l'appartenance à un sexe sur la satisfaction au travail ?

Le sexe et la satisfaction au travail sont deux variables qui existent par elles-mêmes. Le rôle de variable indépendante et variable dépendante sera fixé selon la question de recherche statistique.

4.2.3 Remarque.

Une variable dépendante dans une situation peut être une variable indépendante dans une autre situatiation. Par exemple la satisfaction au travail (dépendante) peut être analysée selon le sexe (indépendante). Il serait aussi possible d'analyser l'absentéisme (dependante) selon la satisfaction au travail (indépendante).

4.3 Représentations

On représente une série double de plusieurs façons :

4.3.1 Représentation brute

Les couples (de valeurs) de la série double sont écrites, côte à côte. Notons qu'un couple de valeurs peut apparaitre plusieurs fois.

Dans l'exemple de la taille et du poid, mesurés sur 18 adolescents, on a utilisé une représentation brute.

4.3.2 Représentation dans un tableau à deux lignes

On dessine un tableau à deux lignes, la première ligne est réservée à la première composante noté x_i et la deuxième est réservée à la deuxième composante noté y_i. Généralement on commence par les paires dont les premières composantes sont les plus petites. Ce mode convient pour les séries quantitatives à petit effectif.

Ce mode de représentation donne pour la série double

$\{(155; 46, 1)$;$(140\,; 38, 2)\,;\,(161\,; 44, 3)\,;\,(148; 38, 2)\,;\,(155; 50, 5)\,;\,(123; 22, 4)\,;$ $(160; 40, 4)\,;$

$(140; 34, 7)\,;\,(165; 50, 5)\,;\,(148; 38, 2)\}$, le tableau suivant.

x_i	123	140	140	148	148	155	155	160	161	165
y_i	22,4	34,7	38,2	38,2	38,2	46,1	50,5	40,4	44,3	50,5

4.3.3 Représentation dans un tableau à trois lignes

S'il y a des couples (x_i, y_i) de la série double se répètant plus qu'une fois, on représente la série par un tableau à trois lignes, en ajoutant une troixièème ligne pour les effectifs.

Par exemple, la série double

$$(1.1)\,;(1,1)\,;(1,3)\,;(3;2)\,;(1;3)\,;(1,5)\,;(3,2)\,;(3;2)\,;(4;2)\,;(1;5)\,;(7,8)$$

est représentée par le tableau à trois lignes

x_i	1	1	1	3	4	7
y_i	1	3	5	2	2	8
n_i	2	2	2	3	1	1

4.3.4 Représentation dans un tableau de contingence

Ce mode de représentation est plus pratique.

La série est représentée par un tableau dont, le nombre de lignes est égale au nombre de valeurs (ou de classes de valeurs) du caractère X et le nombre de colonnes est égale au nombre de valeurs (ou de classes de valeurs)

du caractère Y. L'effectif de chaque couple est inscrit dans la cellule du tableau, intersection de la colonne contenant la premiere composante et la ligne contenant la deuxième composante.

Exemple 1 la série qualitative de l'exemple 2.1.3 est représentée dans le tableau de contigence suivant.

Sexe\ *Statut*	AO	C	I	Totaux
M	5	3	1	9
F	4	3	4	11
Totaux	9	6	5	20

L'effectif du couple (M ;AO) est 5, puisque ce couple apparait 5 fois dans cette série. Cet effectif est inscrit dans la cellule intersection de la ligne contenant la première composante (ici M) et de La colonne contenant la deuxième composante (ici OA).

Exemple 2 Une étude sur 5761 femmes de la survenue d'accouchement prématuré et de l'exposition à des évènements stressants a donné les résultats suivants.

X : type d'accouchement, variable qualitative à 2 modalités

Y : score sur une échelle allant de 0 à 3, variable quantitative discrète à 4 valeurs.

La représentation de cette série double dans un tableau de contingence a donné le tableau suivant.

$X \backslash Y$	0	1	2	3	totaux
A terme	4698	413	250	197	5558
Prématuré	165	16	12	10	203
Totaux	4863	429	262	207	5761

4.3.5 Notation et complétition d'un tableau de contingence

On utilise les notations suivantes pour la représentation d'une série double dans un tableau de contingence

$X \diagdown Y$	y_1	...	y_j	...	y_l	Totaux
x_1	n_{11}	...	n_{1j}	...	n_{1l}	$n_{1\bullet}$
...	...					
x_i	n_{i1}	...	n_ij		n_{il}	$n_{i\bullet}$
...	...					
x_k	n_{k1}		n_{kj}		n_{kl}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet l}$	$n_{\bullet\bullet}$

4.4 Fréquences associées à une série double

4.4.1 Définition

1)- Pour chaque i et j, le nombre n_{ij} est l'effectif du couple (x_i, y_j), dont la première composante est x_i, se trouvant dans la ième ligne, et la deuxième composante est y_j se trouvant dans la jème colonne. L'effectif n_{ij} s'appelle aussi fréquence absolue du couple (x_i, y_j), ou également fréquence absolue jointe de x_i et y_j.

Le nombre

$$f_{ij} = \frac{n_{ij}}{N} \tag{4.1}$$

s'appelle fréquence (ou fréquence relative, s'il ya ambiguité) du couple (x_i, y_j), ou également fréquence jointe de x_i et y_j..

2)- Pour chaque i, le nombre $n_{i\bullet}$ est l'effectif de x_i, c'est-à-dire, le nombre de couples dont la première composante est x_i, ou en d'autres termes, le nombre d'observations de x_i. Leffectif $n_{i\bullet}$ s'appelle aussi fréquence absolue marginale de x_i.

Le nombre

$$f_{i\bullet} = \frac{n_{i\bullet}}{N} \tag{4.2}$$

s'appelle fréquence marginale de x_i.

3)-Pour chaque j, le nombre $n_{\bullet j}$ est l'effectif de y_j, c'est-à-dire, le nombre de couples dont la deuxième composante est y_j, ou en d'autres termes, le nombre d'observation de y_j. Leffectif $n_{\bullet j}$ s'appelle aussi fréquence absolue marginale de y_j.

Le nombre

$$f_{\bullet j} = \frac{n_{\bullet j}}{N} \tag{4.3}$$

s'appelle fréquence marginale de y_j.

4.4.2 Remarques.

1)- Ici on a utilisé les notations suivantes

$$n_{i\bullet} = \sum_{j=1}^{l} n_{ij}, \qquad n_{\bullet j} = \sum_{i=1}^{k} n_{ij}, \quad n_{\bullet\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} \tag{4.4}$$

2)-$n_{\bullet\bullet}$ est l'effectif total de la série double, c'est-à-dire

$$n_{\bullet\bullet} = \sum_{n=1}^{k}\sum_{m=1}^{l} n_{nm} = N \tag{4.5}$$

3)-

$$f_{\bullet\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = 1, \quad \sum_{i=1}^{k} f_{i\bullet} = 1, \quad \sum_{j=1}^{l} f_{\bullet j} = 1 \tag{4.6}$$

4.4.3 Tableau de contingence des fréquences.

Si on remplace dans le tableau de contigence les effectifs par les fréquences correspondantes, on obtient le tableau de contingence (des fréquences) :

$X \diagdown Y$	y_1	...	y_j	...	y_l	Totaux
x_1	f_{11}	...	f_{1j}	...	f_{1l}	$f_{1\bullet}$
...	...					
x_i	f_{i1}	...	f_ij		f_{il}	$f_{i\bullet}$
...	...					
x_k	f_{k1}		f_{kj}		f_{kl}	$f_{k\bullet}$
Totaux	$f_{\bullet 1}$		$f_{\bullet j}$		$f_{\bullet l}$	$f_{\bullet\bullet} = 1$

4.4.4 Exemple

Considérons le tableau de contingence

$X \diagdown Y$	0	1	2	3	Totaux
A terme	4698	413	250	197	5558
Prématuré	165	16	12	10	203
Totaux	4863	429	262	207	5761

On a

$$
\begin{aligned}
n_{11} &= 4698 \text{ est l'effectif du couple (A terme ;0),} \\
n_{21} &= 165 \text{ est l'effectif du couple (Prématuré ;0),} \\
n_{23} &= 12 \text{ est l'effectif du couple (Prématuré ;2),} \\
n_{1\bullet} &= 5558 \text{ est l'effectif marginale de "A terme",} \\
n_{\bullet 2} &= 429 \text{ est l'effectif marginale de "1"} \\
n_{\bullet\bullet} &= 5761 \text{ est l'effectif total de la série double,} \\
f_{11} &= \frac{4698}{5761} \text{ est la fréquence du couple (A terme ;0),} \\
f_{1\bullet} &= \frac{5558}{5761} \text{ est la fréquence marginale de "A terme",} \\
f_{\bullet 2} &= \frac{429}{5761} \text{ est la fréquence marginale de "1".}
\end{aligned}
$$

4.5 Différentes distributions

4.5.1 Distribution jointe des effectifs de X et Y

On appellera distribution jointe des effectifs de X et Y l'ensemble des informations (x_i, y_j, n_{ij}) pour $i = 1, ..., k$ et
$j = 1, ..., \ell$.

Dans l'exemple ci-dessus, la distribution jointe est représentée par le tableau de contingence, elle est représentée aussi par la suite des huit triplets :
(A terme, 0, 4698), (A terme, 1, 413), (A terme, 2, 250), (A terme, 3, 197),
(Prématuré, 0, 165), (Prématuré,1 , 16), (Prématuré, 2,12), (Prématuré, 3,10).

4.5.2 Distributions marginales

Considérons le tableau de contingence

$X \backslash Y$	y_1	...	y_j	...	y_l	Totaux
x_1	n_{11}	...	n_{1j}	...	n_{1l}	$n_{1\bullet}$
...	...					
x_i	n_{i1}	...	n_{ij}		n_{il}	$n_{i\bullet}$
...	...					
x_k	n_{k1}		n_{kj}		n_{kl}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet l}$	$n_{\bullet\bullet}$

A)- En marge à droite (totaux en ligne), pour chaque indice i, l'effectif $n_{i\bullet}$ est le nombre total d'observations de la modalité x_i de X, quelle que soit la modalité de Y . C'est-'a-dire

$$n_{i\bullet} = \sum_{j=1}^{l} n_{ij} = \text{ total de la ligne } i. \tag{4.7}$$

Definition. Les k couples $(x_i, n_{i\bullet})$ définissent la distribution marginale des effectifs de la variable X.

Remarque.

$$\sum_{i=1}^{k} n_{i\bullet} = N \tag{4.8}$$

B)- En marge en bas (totaux en colonne), pour chaque indice j, l'effectif $n_{\bullet j}$ est le nombre total d'observations de la modalité y_j de Y quelle que soit la modalité de X. C'est-à-dire

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij} = \text{ total de la colonne } j.$$

Définition. Les l couples $(y_j, n_{\bullet j})$ définissent la distribution marginale des effectifs de la variable Y.

Remarque.

$$\sum_{j=1}^{l} n_{\bullet j} = N \tag{4.9}$$

4.5.3 Exemple.

Considérons la série double représentée par le tableau de contingence

$X \diagdown Y$	0	1	2	3	Totaux
A terme	4698	413	250	197	5558
Prématuré	165	16	12	10	203
Totaux	4863	429	262	207	5761

Alors,

A)- la distribution marginale des effectifs de la variable X est

X	à terme	prématuré	effectif total
effectifs	5558	203	5761

B)- la distribution marginale des effectifs de la variable Y est

Y	0	1	2	3	Totaux
effectifs	4863	429	262	207	5761

4.5.4 Distributions conditionnelles

Soit une série statistique double dont les données sont représentées par le tableau de contingence, à k lignes et l colonnes, suivant :

$X \diagdown Y$	y_1	...	y_j	...	y_l	Totaux
x_1	n_{11}	...	n_{1j}	...	n_{1l}	$n_{1\bullet}$
...	...					
x_i	n_{i1}	...	n_{ij}		n_{il}	$n_{i\bullet}$
...	...					
x_k	n_{k1}		n_{kj}		n_{kl}	$n_{k\bullet}$
Totaux	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet l}$	$n_{\bullet\bullet}$

A)- pour chaque $1 \leq j \leq l$, la distribution conditionnelle des fréquences du caractère X sachant $Y = y_j$ est représentée par le tableau à 2 lignes

X	x_1	...	x_i	...	x_k	$total$
$f(X \diagup Y = y_j)$	$\frac{n_{1j}}{n_{\bullet j}}$	...	$\frac{n_{ij}}{n_{\bullet j}}$	...	$\frac{n_{kj}}{n_{\bullet j}}$	1

On utilise la notation

$$f(X = x_i \diagup Y = y_j) = \frac{n_{ij}}{n_{\bullet j}}. \tag{4.10}$$

et on lit la fréquence de $X = x_i$, sachant $Y = y_j$, est égale à $\frac{n_{ij}}{n_{\bullet j}}$.

B)- pour chaque $1 \leq i \leq k$, la ditribution conditionnelle (des fréquences) du caractère Y sachant $X = x_i$ est représentée par le tableau à 2 lignes

Y	y_1	...	y_j	...	y_l	Total
$f(Y \diagup X = x_i)$	$\frac{n_{i1}}{n_{i\bullet}}$	...	$\frac{n_{ij}}{n_{i\bullet}}$	...	$\frac{n_{il}}{n_{i\bullet}}$	1

On note

$$f(Y = y_j / X = x_i) = \frac{n_{ij}}{n_{i\bullet}} \tag{4.11}$$

et on lit, la fréquence de $Y = y_j$, sachant $X = x_i$, est égale à $\frac{n_{ij}}{n_{i\bullet}}$.

Exemple

Soit le tableau de contingence suivant

$X \backslash Y$	-2	1	4	totaux
a	2	1	0	3
b	0	0	2	2
c	4	5	3	12
totaux	6	6	5	17

A)- Les distributions conditionnelles des fréquences du caractère X sont :

1)- La distribution de X sachant $Y = -2$, représentée par le tableau à deux lignes suivant

X	a	b	c	total
$f(X/Y = -2)$	$\frac{n_{11}}{n_{\bullet 1}} = \frac{2}{6}$	$\frac{n_{21}}{n_{\bullet 1}} = \frac{0}{6}$	$\frac{n_{31}}{n_{\bullet 1}} = \frac{4}{6}$	1

La distribution de X sachant $Y = 1$, représentée par le tableau à deux lignes suivant

X	a	b	c	total
$f(X/Y = 1)$	$\frac{n_{12}}{n_{\bullet 2}} = \frac{1}{6}$	$\frac{n_{22}}{n_{\bullet 2}} = \frac{0}{6}$	$\frac{n_{32}}{n_{\bullet 2}} = \frac{5}{6}$	1

La distribution de X sachant $Y = 4$, représentée par le tableau à deux lignes suivant

X	a	b	c	total
$f(X/Y = 4)$	$\frac{n_{13}}{n_{\bullet 3}} = \frac{0}{5}$	$\frac{n_{23}}{n_{\bullet 3}} = \frac{2}{5}$	$\frac{n_{33}}{n_{\bullet 3}} = \frac{3}{5}$	1

B)- Les distributions conditionnelles des fréquences du caractère Y sont :

1)- La distribution de Y sachant $X = a$, représentée par le tableau à deux lignes suivant

X	a	b	c	total
$f(Y/X = a)$	$\frac{n_{11}}{n_{1\bullet}} = \frac{2}{3}$	$\frac{n_{12}}{n_{1\bullet}} = \frac{1}{3}$	$\frac{n_{13}}{n_{1\bullet}} = \frac{3}{3}$	1

La distribution de Y sachant $X = b$, représentée par le tableau à deux lignes suivant

X	a	b	c	total
$f(Y/X = b)$	$\frac{n}{n_{2\bullet}} = \frac{0}{2}$	$\frac{n_{22}}{n_{2\bullet}} = \frac{0}{2}$	$\frac{n_{23}}{n_{2\bullet}} = \frac{2}{2}$	1

La distribution de Y sachant $X = c$, représentée par le tableau à deux lignes suivant

X	a	b	c	total
$f(X/Y = 4)$	$\frac{n_{31}}{n_{\bullet 3}} = \frac{4}{12}$	$\frac{n_{32}}{n_{\bullet 3}} = \frac{5}{12}$	$\frac{n_{33}}{n_{\bullet 3}} = \frac{3}{12}$	1

Remarque.

Les distributions conditionnelles des fréquences du caractère X peuvent être représentées dans un même tableau comme suit :

X	a	b	c	total
$f(X/Y = -2)$	$\frac{n_{11}}{n_{\bullet 1}} = \frac{2}{6}$	$\frac{n_{21}}{n_{\bullet 1}} = \frac{0}{6}$	$\frac{n_{31}}{n_{\bullet 1}} = \frac{4}{6}$	1
$f(X/Y = 1)$	$\frac{n_{12}}{n_{\bullet 2}} = \frac{1}{6}$	$\frac{n_{22}}{n_{\bullet 2}} = \frac{0}{6}$	$\frac{n_{32}}{n_{\bullet 2}} = \frac{5}{6}$	1
$f(X/Y = 4)$	$\frac{n_{13}}{n_{\bullet 3}} = \frac{0}{5}$	$\frac{n_{23}}{n_{\bullet 3}} = \frac{2}{5}$	$\frac{n_{33}}{n_{\bullet 3}} = \frac{3}{5}$	1

Ce tableau s'appelle aussi tableau de la distribution (au lieu de tableau des distributions) de X sachant Y.

De la même façon on construit le tableau de la distribution de Y sachant X.

4.5.5 Indépendance de variables statistiques

Définition.

Les deux variables statistiques X et Y d'une série statistiques doubles, représentée par un tableau de contingence, sont dites indépendantes si

$$n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n_{\bullet\bullet}}, \text{ pour tous } i \text{ et } j. \tag{4.12}$$

Conséquence.

L'indépendance des caratères statistiques X et Y signifie que les distributions conditionnelles des fréquences sont identiques aux distributions marginales des fréquences, c'est-à-dire

$$\frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n_{\bullet\bullet}} \text{ et } \frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{n_{\bullet\bullet}}, \text{ pour tous } i \text{ et } j. \tag{4.13}$$

Exemple 1.
Soit une série double représentée par le tableau de contingence suivant

$X \diagdown Y$	y_1	y_2	y_3	totaux
x_1	2	5	6	13
x_2	6	15	18	39
Totaux	8	20	24	52

Les deux caractères X et Y sont indépendants, car la condition

$$n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n_{\bullet\bullet}},$$

est vérifiée pour tous $i = 1, 2$ et $j = 1, 2, 3$. En effet,
$i = 1$ et $j = 1$ donne

$$2 = \frac{13 \times 8}{52},$$

$i = 1$ et $j = 2$ donne

$$5 = \frac{13 \times 20}{52},$$

$i = 1$ et $j = 3$ donne

$$6 = \frac{13 \times 24}{52},$$

$i = 2$ et $j = 1$ donne

$$6 = \frac{39 \times 8}{52},$$

$i = 2$ et $j = 2$ donne

$$15 = \frac{39 \times 20}{52},$$

$i = 2$ et $j = 3$ donne

$$18 = \frac{39 \times 24}{52}.$$

Exemple 2.
Soit une série double représentée par le tableau de contingence suivant

$X \diagdown Y$	y_1	y_1	totaux
x_1	3	4	7
x_2	6	8	14
x_3	9	10	19
totaux	18	22	40

Les caractères X et Y ne sont pas indépendants. Pour le prouver, il suffit de montrer qu'il existe au moins un indice i et un indice j, tels que

$$n_{ij} \neq \frac{n_{i\bullet} \times n_{\bullet j}}{n_{\bullet\bullet}}.$$

C'est le cas, car, par exemple pour $i = 1$ et $j = 1$, on a

$$n_{11} = 3 \text{ et } \frac{n_{1\bullet} \times n_{\bullet 1}}{n_{\bullet\bullet}} = \frac{7 \times 18}{40} = 3.15.$$

4.6 Séries doubles quantitatives

4.6.1 Passage d'une représentation à une autre

Exemple 1.

Soit une série quantitative double dont les données sont représentées par le tableau à deux lignes suivant :

x_i	2	4	5	5	6	6	7	7	10	10	10	
y_i	3	3	6	6	2	5	1	7	8	8	8	

Ces données donnent le tableau à trois lignes suivant :

x_i	2	4	5	6	6	7	7	10
y_i	3	3	6	2	5	1	7	8
n_i	1	1	2	1	1	1	1	3

Ces mêmes données donnent le tableau de contingence suivant :

$X \backslash Y$	1	2	3	5	6	7	8
2	0	0	1	0	0	0	0
4	0	0	1	0	0	0	0
5	0	0	0	0	2	0	0
6	0	1	0	1	0	0	0
7	1	0	0	0	0	1	0
10	0	0	0	0	0	0	3

Exemple 2.
Le tableau de contingence

$X \backslash Y$	4	8	10
-1	2	0	3
0	1	2	1

Equivaut au tableau à deux lignes suivant

x_i	-1	-1	-1	-1	-1	0	0	0	0
y_i	4	4	10	10	10	4	8	8	10

Ce même tableau équivaut au tableau à trois lignes

x_i	-1	-1	0	0	0
y_i	4	10	4	8	10
n_i	2	3	1	2	1

Remarque.

1)- Si les valeurs des caractères quantitatives X et Y sont données individuellement, alors $x_1, x_2, ...x_k$ et $y_1, y_2, ...y_l$ désigneront les valeurs de X et Y respectivement. Ces valeurs sont représentées dans le tableau de contingence dans l'ordre croissant, c'est-à-dire du plus petit au plus grand.

2)- Si les valeurs des caractères quantitatives X et Y sont données en classes, alors $x_1, x_2, ...x_k$ et $y_1, y_2, ...y_l$ désigneront les centres de classes des valeurs de X et Y respectivement.

4.6.2 Exemple

Une entreprise employant 100 femmes relève pour chaque femme son âge, noté X, et le nombre de journées d'absence durant le mois de janvier, noté Y. Le résultat de cette opération est représenté dans le tableau de contingence suivant.

$X \diagdown Y$	0	1	2	3
$[20; 30[$	0	0	5	15
$[30; 40[$	0	15	20	0
$[40; 50[$	15	10	5	0
$[50; 60[$	0	5	5	5

Pour faire des calculs sur ce tableau, on remplace les classes par leurs centres et on le complète par les totaux marginaux. On obtient alors le tableau suivant.

$X \diagdown Y$	0	1	2	3	Totaux
25	0	0	5	15	20
35	0	15	20	0	35
45	15	10	5	0	30
55	0	5	5	5	15
Totaux	15	30	35	20	100

4.6.3 Moyennes des distributions marginales :

A)- Moyenne de X, noté $\mu(X)$ ou $\bar{X}$:

$$\mu(X) = \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet}x_i = \sum_{i=1}^{k} p_{i\bullet}x_i \tag{4.14}$$

B)- Moyenne de Y, noté $\mu(Y)$ ou $\bar{Y}$:

$$\mu(Y) = \frac{1}{N}\sum_{j=1}^{l} n_{\bullet j}y_j = \sum_{j=1}^{l} p_{\bullet j}y_j \tag{4.15}$$

L'application de cette definition à notre exemple donne

$$\mu(X) = \frac{1}{100}(20 \times 25 + 35 \times 35 + 30 \times 45 + 15 \times 55) = 39$$

$$\mu(Y) = \frac{1}{100}(15 \times\ 0 + 30 \times\ 1 + 35 \times\ 2 + 20 \times\ 3) = 1.6$$

4.6.4 Variances des distributions marginales :

A)-Variance (notée V_X ou σ_X^2) et écart-type (noté σ_X) de X :

$$V_X = \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet}x_i^2 - \mu(X)^2 \tag{4.16}$$

$$\sigma_X = \sqrt{V_X}$$

B)-Variance (notée V_Y ou σ_Y^2) et écart-type (noté σ_Y) de Y :

$$V_Y = \frac{1}{N}\sum_{j=1}^{l} n_{j\bullet}y_j^2 - \mu(Y)^2 \tag{4.17}$$

$$\sigma_Y = \sqrt{V_Y} \tag{4.18}$$

L'application de cette définition à notre exemple donne

$$V_X = \frac{1}{100}(20 \times 25^2 + 35 \times 35^2 + 30 \times 45^2 + 15 \times 55^2) - 39^2 = 94$$

$$\sigma_X = \sqrt{94} = 9,70$$

$$V_Y = \frac{1}{100}(15 \times\ 0^2 + 30 \times\ 1^2 + 35 \times\ 2^2 + 20 \times\ 3^2) - 1,6^2 = 0,94$$

$$\sigma_Y = \sqrt{0,94} = 0,97$$

Covariance - Corrélation

Definition.

La covariance d'une série quantitative double $(x_i, y_j), 1 \leq i \leq l, 1 \leq j \leq k$, qu'on note ici $Cov(X,Y)$ ou V_{XY} , est le nombre réel, défini par

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}(x_i - \overline{x})(y_j - \overline{y}). \tag{4.19}$$

Propriétés.

On a

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{l}\sum_{j=1}^{k} n_{ij}x_i y_j - \overline{x}\ \overline{y} \tag{4.20}$$

$$Cov(X,Y) = \sum_{i=1}^{l}\sum_{j=1}^{k} f_{ij}x_i y_j - \overline{x}\ \overline{y}, \tag{4.21}$$

où f_{ij} est la fréquence du couple (x_i, y_j).

$$Cov(X,Y) = Cov(Y,X) \ \text{ et } \ cov(X,X) = V_X.$$

L'application de la définition de la covariance à l'exemple précédent donne

$$\begin{aligned} Cov(X,Y) &= \frac{1}{N}\sum_{i=1}^{l}\sum_{j=1}^{k} n_{ij}x_i y_j - \overline{x}\ \overline{y} \\ &= \frac{1}{100}\left[(0 \times 25 \times 0) + (0 \times 25 \times 1) + ...(5 \times 55 \times 2) + (5 \times 55 \times 3)\right] - (39 \times 1.6) \\ &= \frac{5850}{100} - 39 \times 1.6 = -3.9. \end{aligned}$$

Dans le calcul de $Cov(X,Y)$ ci-dessus, on a commencé par la première ligne jusqu'à la quatrième ligne

Definition.

Le coefficient de corrélation, qu'on note ici $Cor\,(X,Y)$ ou r, est le nombre réel, défini par

$$Cor\,(X,Y) = \frac{Cov\,(X,Y)}{\sigma_X \sigma_Y}. \tag{4.22}$$

L'application de cette définition à l'exemple précédent donne

$$\begin{aligned} Cor\,(X,Y) &= \frac{-3.9}{9.70 \times 0.97} \\ &= 0.41. \end{aligned}$$

Propriétés.

1)-

$$-1 \leq Cor\,(X,Y) \leq 1. \tag{4.23}$$

2)-

$$Cor\,(X,Y) = Cor\,(Y,X)\,. \tag{4.24}$$

3)-

$$Cor\,(X,X) = 1. \tag{4.25}$$

Utilité du coefficient de corrélation

Le coefficient de corrélation mesure l'existance d'une liaison linéaire entre X et Y et le degré de cette liaison:

1. $Cor(X,Y) = 1$: liaison linéaire exacte, $Y = aX + b,$ avec $a > 0.$
2. $Cor(X,Y) = -1:$ liaison linéaire exacte, $Y = aX + b,$ avec $a < 0.$
3. $Cor(X,Y) = 0:$ pas de corrélation.
4. $Cor(X,Y) > 0:$ X et Y ont tendance à varier dans le même sens.
5. $Cor(X,Y) < 0:$ X et Y ont tendance à varier dans le sens contraire.
6. $|Cor(X,Y)| > 0,8:$ X et Y ont une forte corrélation (liaison) linéaire.

4.6.5 Ajustement linéaire d'une série quantitative double

L'ajustement linéaire signifie la recherche d'une droite qui représente le mieux, une liaison linéaire entre X et Y.

Ajustement graphique.

L'ajustement graphique d'une série double (x_i, y_i) se fait par la visualisation du nuage de points (x_i, y_i) de la série et le traçage d'une droite, appelée droite d'ajustement graphique, de sorte que les points (x_i, y_i) du nuage soient le plus proches possible de cette droite. L'ajustement graphique sert à donner rapidement une première idée sur la relation entre l'évolution du caractère X et l'évolution du caractère Y.

L'avantage de l'ajustement graphique réside dans sa simplicité et son utilisation immédiate. L'inconvénient vient de sa subjectivité : il n'est pas unique car il dépend de la personne qui l'a réalisée.

Nuage de points.

Pour représenter le nuage de points, on dessine un repère orthogonal. Le nuage de points associé à la série double est l'ensemble des points (x_i, y_i) représentés dans ce repère.

Ajustement par la méthode des moindres carrés.

Principe.

Soit une série double représentée par un tableau de contingence.

A)- On montre qu'il existe une droite unique D_y d'équation

$$y = ax + b$$

qui minimise la somme

$$\sum_{j=1}^{l} \sum_{i=1}^{k} n_{ij} \left(y_j - (ax_i + b)\right)^2 .$$

Les coefficients a et b se calculent par les formules

$$a = \frac{cov\,(X, Y)}{\sigma_X^2} \text{ et } b = \overline{y} - a\overline{x}. \tag{4.26}$$

La droite D_y s'appelle droite de regression de y en x. Cette droite est utilisée pour prévoir les valeurs de Y pour des valeurs de X, non trouvées expérimentalement,...

B)- On montre qu'il existe une droite unique D_x d'équation

$$x = a'y + b'$$

qui minimise la somme

$$\sum_{j=1}^{l}\sum_{i=1}^{k} n_{ij}\left(x_i - (a'y_j + b')\right)^2.$$

Les coefficient a et b se calculent par les formules

$$a' = \frac{cov\,(X,Y)}{\sigma_y^2} \text{ et } b' = \overline{x} - a'\overline{y}. \tag{4.27}$$

La droite D_x s'appelle droite de regression de x en y. Cette droite est utilisée pour prévoir les valeurs de X pour des valeurs de Y, non données expérimentalement,...

Autres formules pour a, a′et Cor (X,Y).

A partir des formules précédentes, on obtient les formules suivantes.

$$\begin{aligned}
a &= \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_iy_j - \overline{x}\sum_{j=1}^{l} n_{\bullet j}y_j}{\sum_{i=1}^{k} n_{i\bullet}x_i^2 - \overline{x}\sum_{i=1}^{k} n_{i\bullet}x_i} \\
&\quad \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_iy_j - N\overline{x}\ \overline{y}}{\sum_{i=1}^{k} n_{i\bullet}x_i^2 - N\overline{x}^2} \\
a' &= \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_iy_j - \overline{x}\sum_{j=1}^{l} n_{\bullet j}y_j}{\sum_{j=1}^{l} n_{\bullet j}y_j^2 - \overline{y}\sum_{j=1}^{l} n_{\bullet j}y_j} \\
&= \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_iy_j - N\overline{x}\ \overline{y}}{\sum_{j=1}^{l} n_{\bullet j}y_j^2 - N\overline{y}^2} \\
Cor\,(X,Y) &= \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_iy_j - \overline{x}\sum_{j=1}^{l} n_{\bullet j}y_j}{\sqrt{\left(\sum_{i=1}^{k} n_{i\bullet}x_i^2 - \overline{x}\sum_{i=1}^{k} n_{i\bullet}x_i\right)\left(\sum_{j=1}^{l} n_{\bullet j}y_j^2 - \overline{y}\sum_{j=1}^{l} n_{\bullet j}y_j\right)}}
\end{aligned} \tag{4.28}$$

Le coefficient de corrélation $\mathbf{Cor}\,(X,Y)$, $\mathbf{a}$ et $\mathbf{a}'$ ont même signe et on a

$$\mathbf{Cor}\,(X,Y)^2 = \mathbf{a}\mathbf{a}'. \tag{4.29}$$

Exemple.

On considère la série double représentée par le tableau de contingence suivant

$X \diagdown Y$	2	3	Totaux
25	5	15	20
35	20	5	25
45	5	0	5
Totaux	30	20	50

On a

$$\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{3} n_{i\bullet} x_i}{N} \\
&= \frac{20 \times 25 + 25 \times 35 + 5 \times 45}{50} \\
&\quad \frac{1600}{50} \\
&= 32.
\end{aligned}$$

$$\begin{aligned}
\overline{y} &= \frac{\sum_{j=1}^{2} n_{\bullet j} x_j}{N} \\
&= \frac{30 \times 2 + 20 \times 3}{50} \\
&\quad \frac{120}{50} \\
&= 2.4
\end{aligned}$$

$$\begin{aligned}
&\sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} x_i y_j \\
&= (5 \times 25 \times 2) + (15 \times 25 \times 3) + (20 \times 35 \times 2) \\
&\quad + (5 \times 35 \times 3) + (5 \times 45 \times 2) + (0 \times 45 \times 3) \\
&= 3750
\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{k} n_{i\bullet}x_i^2 &= \left(20\times 25^2\right)+\left(25\times 35^2\right)+\left(5\times 45^2\right)\\ &= 53250.\end{aligned}$$

$$\begin{aligned}\sum_{j=1}^{2} n_{\bullet j}y_j^2 &= 30\times 2^2+20\times 3^2\\ &= 300.\end{aligned}$$

$$\begin{aligned}a &= \frac{3750-50\times 32\times 2.4}{53250-50\times 32^2}\\ &= -0.044.\end{aligned}$$

$$\begin{aligned}a' &= \frac{3750-50\times 32\times 2.4}{300-50\times 2.4^2}\\ &= -7.5.\end{aligned}$$

$$\begin{aligned}Cor\,(X,Y) &= \frac{3750-50\times 32\times 2.4}{\sqrt{\left(53250-50\times 32^2\right)\left(300-50\times 2.4^2\right)}}\\ &= -0.57.\end{aligned}$$

$$\begin{aligned}b &= \overline{y}-a\overline{x}\\ &= 2.4-(-0.044)\times 32\\ &= 3.808.\end{aligned}$$

$$\begin{aligned}b' &= \overline{x}-a'\overline{y}\\ &= 32-(-7.5)\times 2.4\\ &= 50.\end{aligned}$$

L'équation de la droite de regression de y en x est

$$y=-0.044x+3.808.$$

L'équation de la droite de regression de x en y est

$$x=-7.5y+50.$$

Puisque la valeur absolue ($=0.57$) du coefficient de corrélation n'est pas proche de un, la corrélation linéaire entre x et y n'est pas bonne.

Cas de représentation par un tableau à deux lignes

Pour une série double représentée par un tableau à deux lignes.

x_1	x_2	...	x_N
y_1	y_2	...	y_N

On a les formules simples suivantes

$$\begin{aligned} \overline{x} &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ \overline{y} &= \frac{1}{N}\sum_{i=1}^{N} y_i \\ V(X) &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \overline{x}^2 \\ V(Y) &= \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \overline{y}^2 \\ Cov(X,Y) &= \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{x}\ \overline{y} \end{aligned} \tag{4.30}$$

A partir de ces formules on peut également vérifier les formules suivantes.

$$\begin{aligned} a &= \frac{\sum_{i=1}^{N} x_i y_i - \overline{x}\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i^2 - \overline{x}\sum_{i=1}^{N} x_i} \\ a' &= \frac{\sum_{i=1}^{N} x_i y_i - \overline{x}\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} y_i^2 - \overline{y}\sum_{i=1}^{N} y_i} \\ Cor(X,Y) &= \frac{\sum_{i=1}^{N} x_i y_i - \overline{x}\sum_{i=1}^{N} y_i}{\sqrt{\left(\sum_{i=1}^{N} x_i^2 - \overline{x}\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i^2 - \overline{y}\sum_{i=1}^{N} y_i\right)}} \end{aligned} \tag{4.31}$$

Exemple.

Le tableau ci-dessous liste les classements de salaires et de stress pour des emplois sélectionnés aléatoirement. Le rang 1 pour les salaires correspond au salaire le plus bas et le rang 1 pour le stress correspond au stress le plus faible.

Emploi	Rang du salaire	Rang du stress
agent de change	9	9
zoologiste	5	4
ingénieur en électricité	8	5
CPE	6	7
gérant d'hôtel	4	6
employé de banque	1	3
inspecteur de sécurité	2	2
économiste	3	1
psycologue	7	8
pilote de l'air	10	11
trader à wall street	11	10

1)- Calculez les moyennes marginales $\overline{x}$, $\overline{y}$, de X et Y.

2)- Calculer les sommes des carrés x_i^2, les sommes des carrés y_i^2 et les sommes des produits $x_i y_i$

3)- Calculez le coefficient de corrrélation linéaire $Cor\,(X,Y)$ et estimez la qualité de la corrélation linéaire entre le rang X du salaire et le rang Y du stress.

4)- Calculez les coefficients a et b de la droite de regression de Y en X : $y = ax + b$.

On peut faire les calculs sur un tableau statistique :

	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
	9	9	81	81	81
	5	4	20	25	16
	8	5	40	64	25
	6	7	42	36	49
	4	6	24	16	36
	1	3	3	1	9
	2	2	4	4	4
	3	1	3	9	1
	7	8	56	49	64
	10	11	110	100	121
	11	10	110	121	100
Total	66	66	493	506	506

1)-

$$\overline{x} = \frac{\sum_{i=}^{N} x_i}{N} = \frac{\sum_{i=}^{11} x_i}{11} = \frac{66}{11} = 6$$

$$\overline{y} = \frac{\sum_{i=}^{N} y_i}{N} = \frac{\sum_{i=}^{11} y_i}{11} = \frac{66}{11} = 6$$

2)-

$$\sum_{i=}^{N} x_i^2 = 506, \ \sum_{i=1}^{11} y_i^2 = 506, \ \sum_{i=}^{11} x_i y_i = 493.$$

3)-

$$\begin{aligned} Cor\,(X,Y) &= \frac{\sum_{i=1}^{11} x_i y_i - \overline{x}\sum_{i=1}^{11} y_i}{\sqrt{\left(\sum_{i=1}^{11} x_i^2 - \overline{x}\sum_{i=1}^{11} x_i\right)\left(\sum_{i=1}^{11} y_i^2 - \overline{y}\sum_{i=1}^{11} y_i\right)}} \\ &= \frac{493-(6\times 66)}{\sqrt{[506-(6\times 66)]\,[506-(6\times 66)]}} \\ &\quad \frac{97}{\sqrt{12100}} \\ &= 0.88 \end{aligned}$$

Il y a une forte corrélation linéaire entre les deux caractères X et Y.

4)-

$$\begin{aligned} a &= \frac{\sum_{i=1}^{11} x_i y_i - \overline{x}\sum_{i=1}^{11} y_i}{\sum_{i=1}^{11} x_i^2 - \overline{x}\sum_{i=1}^{11} x_i} \\ &= \frac{97}{[506-(6\times 66)]} \\ &= 0.88 \end{aligned}$$

$$\begin{aligned} b &= \overline{y} - a\overline{x} \\ &= 6-(0.88\times 6) \\ &= 0.72. \end{aligned}$$

Cas de représentation par un tableau à trois lignes

Pour une série double représentée par un tableau à trois lignes.

x_1	x_2	...	x_r
y_1	y_2	...	y_r
n_1	n_2	...	n_r

On a les formules simples suivants

$$\begin{aligned} \overline{x} &= \frac{1}{N}\sum_{i=1}^{r} n_i x_i \\ \overline{y} &= \frac{1}{N}\sum_{i=1}^{r} n_i y_i \\ V(X) &= \frac{1}{N}\sum_{i=1}^{r} n_i x_i^2 - \overline{x}^2 \\ V(Y) &= \frac{1}{N}\sum_{i=1}^{r} n_i y_i^2 - \overline{y}^2 \\ Cov(X,Y) &= \frac{1}{N}\sum_{i=1}^{r} n_i x_i y_i - \overline{x}\,\overline{y} \end{aligned} \tag{4.32}$$

Il faut noter que dans ce cas, l'entier r est le nombre des couples (x_i, y_i), n_i est le nombre de répétition du couple (x_i, y_i) et l'effectif total N est égal à la somme des n_i.

Pour cette représentation, on peut également vérifier les formules suivantes.

$$\begin{aligned} a &= \frac{\sum_{i=1}^{r} n_i x_i y_i - \overline{x}\sum_{i=1}^{r} n_i y_i}{\sum_{i=1}^{N} n_i x_i^2 - \overline{x}\sum_{i=1}^{N} n_i x_i} \\ a' &= \frac{\sum_{i=1}^{r} x_i y_i - \overline{x}\sum_{i=1}^{r} n_i y_i}{\sum_{i=1}^{r} n_i y_i^2 - \overline{y}\sum_{i=1}^{r} n_i y_i} \\ Cor(X,Y) &= \frac{\sum_{i=1}^{r} n_i x_i y_i - \overline{x}\sum_{i=1}^{r} n_i y_i}{\sqrt{\left(\sum_{i=1}^{r} n_i x_i^2 - \overline{x}\sum_{i=1}^{r} n_i x_i\right)\left(\sum_{i=1}^{r} n_i y_i^2 - \overline{y}\sum_{i=1}^{r} n_i y_i\right)}} \end{aligned} \tag{4.33}$$

Exemple.

on considère la série double suivante

$$\begin{array}{ccccc} \{(-1;0), & (1;0), & (2;1), & (-1;0), & (2;-1), \\ (2;5) & (3;2), & (-1;0), & (2;5), & (2;-1), \\ (-1;0), & (-1;0), & (3;2), & (-1;3), & (-1;0) \\ (2;5), & (2;-1), & (2;-1), & (-1;3), & (0;0)\}. \end{array}$$

1)-Représenter les données de cette série double dans un tableau statistique.

2)- Calculer les paramètres suivants : $\overline{x}, \overline{y}, V_X, V_Y, V_{XY}, Cor(X,Y)$.

Solution.

1)- On représente cette série double dans le tableau statistique (à trois lignes) suivant :

x_i	-1	-1	0	1	2	2	2	3
y_i	0	3	0	0	-1	1	5	2
n_i	6	2	1	1	4	1	3	2

2)- Pour faire des calculs sur le tableau, on ajoute des lignes :

									Total
x_i	-1	-1	0	1	2	2	2	3	
y_i	0	3	0	0	-1	1	5	2	
n_i	6	2	1	1	4	1	3	2	20
$n_i x_i$	-6	-2	0	1	8	2	6	6	15
$n_i y_i$	0	6	0	0	-4	1	15	4	22
$n_i x_i y_i$	0	-6	0	0	-8	2	30	12	30
$n_i x_i^2$	6	2	0	1	16	4	12	18	59
$n_i y_i^2$	0	18	0	0	4	1	75	8	106

D'aprés le tableau, on a

$$\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{8} n_i x_i}{\sum_{i=1}^{8} n_i} = \frac{15}{20} = \mathbf{0.75}, \\
\overline{y} &= \frac{\sum_{i=1}^{8} n_i y_i}{\sum_{i=1}^{8} n_i} = \frac{22}{20} = \mathbf{1.1}, \\
V_X &= \frac{\sum_{i=1}^{8} n_i x_i^2}{\sum_{i=1}^{8} n_i} - \overline{x}^2 = \frac{59}{20} - 0.75^2 = \mathbf{2.3875}, \\
V_Y &= \frac{\sum_{i=1}^{8} n_i y_i^2}{\sum_{i=1}^{8} n_i} - \overline{y}^2 = \frac{106}{20} - 1.1^2 = 4.\mathbf{09}, \\
V_{XY} &= \frac{\sum_{i=1}^{8} n_i x_i y_i}{\sum_{i=1}^{8} n_i} - \overline{xy} = \frac{30}{20} - 0.75 \times 1.1 = \mathbf{0.675}, \\
Cor\,(X,Y) &= \frac{V_{XY}}{V_X V_Y} = \frac{0.675}{2.3875 \times 6.19} = \mathbf{0.216}.
\end{aligned}$$

Remarque importante

Dans les exemples précédents, on a utilisé des applications disponibles dans les calculatrices ordinaires (KENKO,...). Ces applications donnent les valeurs des paramètres directement. Si on fait les calculs à la main, les résultats obtenus peuvent être légèrement différents, à cause des arrondis.