جامعة أبو بكر بلقايد
UNIVERSITY OF TLEMCEN
كلية علوم الطبيعة والحياة علوم الارض والكون

UNIVERSITE TLEMCEN

جامعة أبي بكر بلقايد - تلمسان
UNIVERSITY OF BELKAID ABOU BEKR – TLEMCEN
كلية العلوم
FACULTY OF SCIENCES
قسم الرياضيات
DEPARTMENT OF MATHEMATICS

Cours de Biostatistique bidimensionnelle

Licence Biologie moléculaire

Année universitaire 2022/2023

Joanna Dib, Djilali Ameur, Majda Dali-Sahi

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Statistique descriptive bidimensionnelle

1.1 Représentation graphique : Nuage de points (Scatter plot)

Il s'agit d'un graphique très commode pour représenter les observations simultanées de deux variables quantitatives dans le but d'étudier la liaison entre les deux variables discrètes.

Rappelons ici que les données de base correspondent à des séries statistiques brutes contenant N lignes pour les N individus et 2 colonnes pour les deux variables X et Y. Une manière simple de visualiser les données, elle consiste à représenter chaque individu par un point dans le plan $\mathbb{R}^2$ en considérant deux axes perpendiculaires, l'axe horizontal représentant la variable X et l'axe vertical la variable Y. L'ensemble de ces points donne en général une idée assez bonne de la variation conjointe des deux variables.

Cette représentation graphique des données porte le nom de ***nuage de points***. Puisque ce graphique nous indique la façon dont les points se dispersent dans le plan, on lui donne aussi le nom de ***graphique ou diagramme de dispersion***, traduction plus fidèle de l'anglais ***scatter-plot***.

Remarque 1.1.1. *Dans le cas où l'une ou l'autre variable est quantitative continue, plusieurs représentations graphiques ont été proposées. Citons-en quelques-unes en guise d'exemple.*

1. *Nous construisons un nuage de points où ces derniers sont définis à partir des centres de classe.*

2. *Nous construisons un diagramme à trois dimensions, appelé* ***stéréogramme*** *et composé de parallélépipèdes rectangles dont la base est définie par les couples de classes et le volume par l'effectif (ou la fréquence) qui leur correspond.*

Les différentes observations des nuages de points permettent de détermine : Des tendances ou dépendances, des relations positives ou négatives, des répartitions plus ou moins homogènes ou encore des données aberrantes.

L'impression qu'on peut retirer d'un nuage de points dépend bien sûr des unités choisies le long de chaque axe. D'une façon générale, on distinguera deux cas :

1. Le cas de ***variables homogènes*** représentant la même grandeur et exprimées dans la même unité où on choisira la même échelle sur les deux axes orthonormés.

2. Celui des ***variables hétérogènes*** où il est recommandé soit de représenter les ***variables centrées et réduites*** sur des axes orthonormés.

1.1.1 Rappel : variables centrées et réduites

Si X est une variable quantitative de moyenne $\overline{X}$ et d'écart–type $\sigma(X)$, on appelle variable centrée associée à X la variable $X - \overline{X}$ et variable centrée et réduite (ou tout simplement variable réduite) associée à X la variable $\frac{X-\overline{X}}{\sigma(X)}$. Ainsi nous obtenons des données indépendantes de l'unité ou de l'échelle choisie et des variables ayant même moyenne et même dispersion.

Remarque 1.1.2. *Les valeurs des coefficients de corrélation entre variables centrées réduites restent identiques à ce qu'elles étaient avant le centrage et la réduction des variables .*

Exemple : On a mesuré la taille (variable X, en mètres) et le poids (variable Y, en kg) de 5 personnes. La série statistique bivariée obtenue est la suivante :

Individus	1	2	3	4	5
Taille	1.55	1.60	1.65	1.70	1.75
Poids	54	62	60	69	73

TABLE 1.1 – Répartition des 5 individus suivant leur taille et leur poids

Le nuage de points est présenté dans la figure ci-dessous :

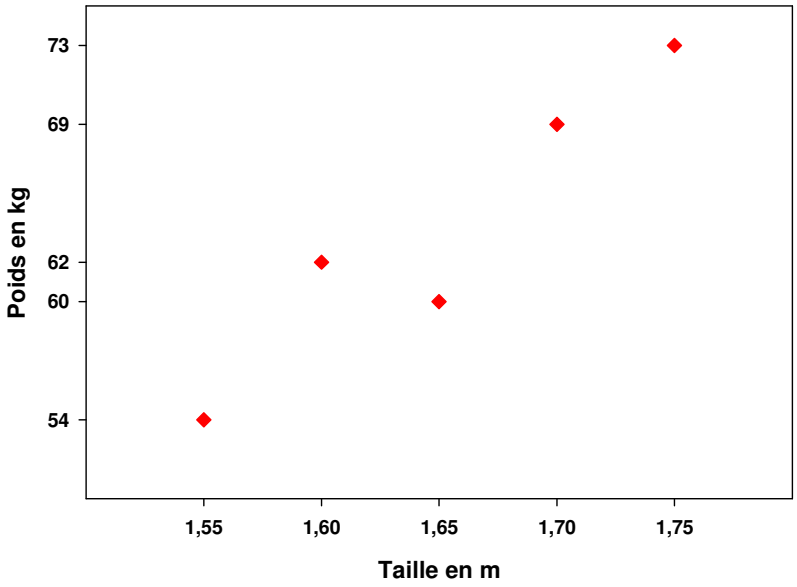

FIGURE 1.1 – Nuage de points (Taille ; Poids)

L'examen du nuage de points permet de révéler l'existence d'une ***structure d'association "positive"*** entre les deux variables considérées puisqu'on remarque que la variable "taille" a tendance à augmenter davantage que la variable "poids" augmente. Ce qui se traduit par le fait que les deux variables ont tendance à varier "dans le même sens".

Exemple : La variation de la pression atmosphérique avec l'altitude est le changement de la pression selon l'altitude au-dessus du sol. On a mesuré l'altitude (variable X, en mètres) et la pression atmo-

sphérique (variable Y, en hPa) de 10 différentes altitudes. La série statistique bivariée obtenue est la suivante :

Altitude en m	**Pression atmosphérique en hPa**
0	1 013.25
500	954.61
1 000	898.76
1 500	854.58
2 000	794.98
2 500	746.86
3 000	701.12
3 500	657.68
4 000	616.45
4 500	577.33
5 000	540.25
6 000	471.87
7 000	410.66
8 000	356.06
9 000	307.48
10 000	264.42
11 000	226.37

TABLE 1.2 – Tableau portant sur la variable statistique "variation de la pression atmosphérique selon l'altitude"

Le nuage de points est présenté dans la figure ci-dessous :

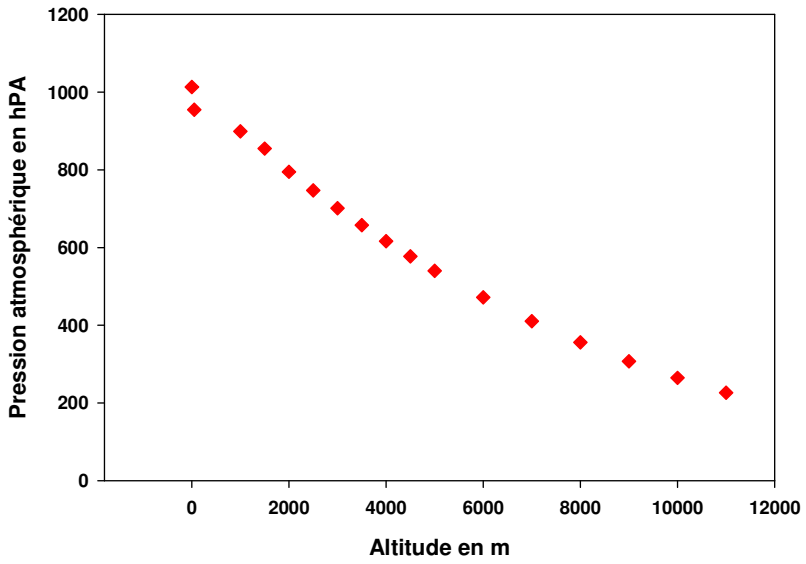

FIGURE 1.2 – Nuage de points (Altitude ; Pression)

Le nuage de points montre une ***structure "descendante"*** : La pression atmosphérique a tendance à diminuer davantage que l'altitude augmente. De manière générale, les deux variables étudiées semblent donc avoir tendance à varier dans des sens opposés. Il s'agit d'une "***association négative***" entre les deux variables.

1.2 Tableau de contingence

Un tableau de contingence est une méthode de représentation de données permettant de déceler une dépendance entre deux caractères. Elle consiste à croiser deux caractères d'une population en dénombrant l'effectif correspondant simultanément à la conjonction des deux caractères. Cette situation se présente surtout lorsque N est élevé, que les variables sont qualitatives ou quantitatives discrètes ou même quantitatives continues et que le nombre de valeurs distinctes de chaque variable est faible.

L'expression tableau de contingence a été introduite par le statisticien britannique ***Karl Pearson*** dans un essai intitulé *"On the Theory of Contingency and Its Relation to Association and Normal Correlation"*, en 1904.

Soit un échantillon de N individus ou unités statistiques décrites simultanément selon deux caractères : La variable statistique X prenant les modalités $x_1,\ x_2\ ,...,\ x_r$ et la variable statistique Y prenant les modalités $y_1,\ y_2\ ,...,\ y_s$. On désigne par :

◇ n_{ij} le nombre d'individus présentant les modalités x_i et y_j.

◇ N désigne l'effectif total de l'échantillon définie par :

$$N = \sum_i \sum_j n_{ij} = \sum_i n_{i.} = \sum_j n_{j.} \quad avec \quad i \in \{1, ..., r\} \quad et \quad j \in \{1, ..., s\}.$$

◇ $n_{i.}$ désigne l'effectif total de la ligne i, qui est le nombre total d'individus présentant la modalité x_i indépendamment du caractère Y, définie par :

$$n_{i.} = \sum_j n_{ij} = n_{i1} + n_{i2} + ... + n_{is} \quad avec \quad j \in \{1, ..., s\}.$$

◇ $n_{.j}$ désigne l'effectif total de la colonne j qui est le nombre total d'individus présentant la modalité y_j indépendamment du caractère X :

$$n_{.j} = \sum_i n_{ij} = n_{1j} + n_{2j} + ... + n_{rj} \quad avec \quad i \in \{1, ..., r\}.$$

◇ f_{ij} désigne la proportion d'individus présentant simultanément les deux modalités x_i et y_j définie par :

$$f_{ij} = \frac{n_{ij}}{N} \quad \forall i \in \{1, ..., r\} \quad et \quad \forall j \in \{1, ..., s\}$$

sachant que la somme de toutes les fréquences est égale à 1.

Le tableau de contingence est de la forme suivante :

Y / **X**	y_1	y_2	$\cdots$	y_j	$\cdots$	y_s	**Colonnes marginales**
x_1	n_{11}	n_{12}		n_{1j}		n_{1s}	$n_{1.}$
x_2	n_{21}	n_{22}		n_{2j}		n_{2s}	$n_{2.}$
$\vdots$							
x_i	n_{i1}	n_{i2}		n_{ij}		n_{is}	$n_{i.}$
$\vdots$							
x_r	n_{r1}	n_{r2}		n_{rj}		n_{rs}	$n_{r.}$
lignes marginales	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.s}$	$N = n_{..}$

TABLE 1.3 – Tableau de contingence d'un couple de variables statistiques (X,Y)

1.3 Séries et distributions marginales d'un couple de variables statistiques (X,Y)

1.3.1 Série et distribution marginale de la variable statistique X

Soit la ***série marginale univariée en*** X : $(x_i; n_{i.})$, $\forall i \in \{1, ..., r\}$ obtenues en ne considérant qu'une seule variable X.

Une ***distribution marginale en*** X est une distribution statistique à un seul caractère obtenue en associant à chacune des modalités x_i de la variable statistique X, l'effectif marginal correspondant $n_{i.}$ (la colonne marginale) représentant les individus présentant la modalité x_i indépendamment des modalités du second caractère Y.

Elle est définie par l'ensemble des couples $\{(x_i; n_{i.})\}$, $\forall i \in \{1, ..., r\}$ où l'on associe à chaque modalité x_i, l'effectif marginal $n_{i.}$ correspondant au nombre de fois que la valeur x_i apparaît dans la série marginale en X.

Nous pouvons aussi définir la fréquence marginale associée à chaque modalité x_i de la variable X par :

$$f_{i.} = \frac{n_{i.}}{N} \quad \forall i \in \{1, ..., r\}.$$

X	$n_{i.}$	f_i
x_1	$n_{1.}$	$f_{1.}$
x_2	$n_{2.}$	$f_{2.}$
$\vdots$	$\vdots$	$\vdots$
x_i	$n_{i.}$	$f_{i.}$
$\vdots$	$\vdots$	$\vdots$
x_r	$n_{r.}$	$f_{r.}$
	$N = n_{..}$	$f_{..}$

TABLE 1.4 – Tableau des effectifs et des fréquences marginaux de la variable X

Nous pouvons également considérer les effectifs cumulés marginaux ainsi que les fréquences cumulées marginales.

1.3.2 Distribution marginale de la variable statistique Y

De la même manière, on définit la distribution marginale de la variable statistique Y (en lui associant comme colonne des effectifs, la ligne marginale) et la fréquence marginale de la modalité y_j par :

$$f_{.j} = \frac{n_{.j}}{N} \quad \forall j \in \{1, ..., s\}.$$

Soit la ***série marginale de*** Y : $(y_j; n_{.j})$, $\forall j \in \{1, ..., s\}$:

Y	$n_{.j}$	$f_{.j}$
y_1	$n_{.1}$	$f_{.1}$
y_2	$n_{.2}$	$f_{.2}$
$\vdots$	$\vdots$	$\vdots$
y_i	$n_{.j}$	$f_{.j}$
$\vdots$	$\vdots$	$\vdots$
y_r	$n_{.s}$	$f_{.s}$
	$N = n_{..}$	$f_{..}$

TABLE 1.5 – Tableau des effectifs et des fréquences marginaux de la variable Y

S'agissant d'une distribution statistique unidimensionnelle, la distribution marginale de X est étudiée exactement de la même manière que dans le chapitre précédent.

1.3.3 Paramètres de position et de dispersion marginaux

1.3.3.1 Moyenne marginale de X

La moyenne marginale de la variable X, notée $\overline{X}$, correspond à la moyenne du nombre total d'individus présentant les modalités x_i indépendamment de la variable Y et est définie par :

$$\overline{X} = \frac{1}{N} \sum_i n_{i.}\, x_i = \sum_i f_{i.}\, x_i \quad \forall i \in \{1, ..., r\} \quad \textit{si } X \textit{ est une variable discrete}$$

et

$$\overline{X} = \frac{1}{N} \sum_i n_{i.}\, c_i = \sum_i f_{i.}\, c_i \quad \forall i \in \{1, ..., r\} \quad \textit{si } X \textit{ est une variable continue}$$

où c_i est le centre de classe $[x_i; x_{i+1}[$.

1.3.3.2 Moyenne marginale de Y

La moyenne marginale de la variable Y, notée $\overline{Y}$, correspond à la moyenne du nombre total d'individus présentant les modalités y_j indépendamment de la variable X et est définie par :

$$\overline{Y} = \frac{1}{N} \sum_j n_{.j}\, y_j = \sum_j f_{.j}\, y_j \quad \forall j \in \{1, ..., s\} \quad \textit{si } Y \textit{ est une variable discrete}$$

et

$$\overline{Y} = \frac{1}{N} \sum_j n_{.j}\, c'_j = \sum_j f_{.j}\, c'_j \quad \forall j \in \{1, ..., s\} \quad \textit{si } Y \textit{ est une variable continue}$$

où c'_j est le centre de classe $[y_j; y_{j+1}[$.

1.3.3.3 Variance marginale de X

La variance marginale de X, notée $V(X)$, est la mesure de la dispersion des individus de l'échantillon présentant les modalités x_i indépendamment de la variable Y et est définie par :

$$V(X) = \frac{1}{N}\sum_i n_{i.}\,(x_i - \overline{X})^2 = \sum_i f_{i.}\,(x_i - \overline{X})^2 \quad \forall i \in \{1,...,r\} \quad \textit{si } X \textit{ est une variable discrete}$$

et

$$V(X) = \frac{1}{N}\sum_i n_{i.}\,(c_i - \overline{X})^2 = \sum_i f_{i.}\left(c_i - \overline{X}\right)^2 \quad \forall i \in \{1,...,r\} \quad \textit{si } X \textit{ est une variable continue}$$

ou encore

$$V(X) = \frac{1}{N}\sum_i n_{i.}\,x_i - \overline{X}^2 = \sum_i f_{i.}\,x_i - \overline{X}^2 \quad \forall i \in \{1,...,r\} \quad \textit{si } X \textit{ est une variable discrete}$$

et

$$V(X) = \frac{1}{N}\sum_i n_{i.}\,c_i - \overline{X}^2 = \sum_i f_{i.}\,c_i - \overline{X}^2 \quad \forall i \in \{1,...,r\} \quad \textit{si } X \textit{ est une variable continue}$$

où c_i est le centre de classe $[x_i; x_{i+1}[$.

1.3.3.4 Variance marginale de Y

La variance marginale de Y, notée $V(Y)$, est la mesure de la dispersion des individus de l'échantillon présentant les modalités y_j indépendamment de la variable X et est définie par :

$$V(Y) = \frac{1}{N}\sum_j n_{.j}\,(y_j - \overline{Y})^2 = \sum_j f_{.j}\left(y_j - \overline{Y}\right)^2 \quad \forall j \in \{1,...,s\} \quad \textit{si } Y \textit{ est une variable discrete}$$

et

$$V(Y) = \frac{1}{N}\sum_j n_{.j}\,(c'_j - \overline{Y})^2 = \sum_j f_{.j}\left(c'_j - \overline{Y}\right)^2 \quad \forall j \in \{1,...,s\} \quad \textit{si } Y \textit{ est une variable continue}$$

ou encore

$$V(Y) = \frac{1}{N}\sum_j n_{.j}\,y_j - \overline{Y}^2 = \sum_j f_{.j}\,y_j - \overline{Y}^2 \quad \forall j \in \{1,...,s\} \quad \textit{si } Y \textit{ est une variable discrete}$$

et

$$V(Y) = \frac{1}{N}\sum_j n_{.j}\,c'_j - \overline{Y}^2 = \sum_j f_{.j}\,c'_j - \overline{Y}^2 \quad \forall j \in \{1,...,s\} \quad \textit{si } Y \textit{ est une variable continue}$$

où c'_j est le centre de classe $[y_j; y_{j+1}[$.

Exemple : Nous interrogeons 10 étudiants boursiers dans le but de préparer une enquête sur l'éventuelle liaison qui pourrait exister ente le montant mensuel de la bourse (y_j en DA) et les dépenses en biens ou services de loisirs journaliers (x_i en DA). Les résultats de cette (mini) pré-enquête sont présentés ans le tableau suivant :

Y / X	[1600 ;2400[	[2400 ;3200[	[3200 ;4000[
[250 ;450[	1	3	2
[450 ;650[	2	2	0

TABLE 1.6 – Tableau de contingence des caractères "montant mensuel de la bourse et les dépenses en biens ou services de loisirs journaliers"

1. Complétez le tableau de contingence avec les effectifs marginaux en X et en Y

Y / X	[1600 ;2400[	[2400 ;3200[	[3200 ;4000[	**Colonnes marginales** $n_{i.}$
[250 ;450[	1	3	2	6
[450 ;650[	2	2	0	4
Lignes marginales $n_{.j}$	3	5	2	10

2. Déterminez les moyennes et les variances des distributions marginales en X et Y et la médiane de la distribution marginale en Y

Reprenons dans le tableau ci-dessous la distribution marginale en X et Y et complétons-la avec les colonnes contenant les produits $n_{i.}c_i$, $(c_i - \overline{X})$, $(c_i - \overline{X})^2$ et $n_{i.}(c_i - \overline{X})^2$ (pour le calcul de la moyenne et de la variance).

Classe des dépenses X	**Centre de classe** c_i	**Nombre d'étudiants** $n_{i.}$	$n_{i.} \,.\, c_i$	$c_i - \overline{X}$	$(c_i - \overline{X})^2$	$n_{i.} \,.\, (c_i - \overline{X})^2$
[250 ; 450[	350	6	2 100	- 80	6 400	38 400
[450 ; 650[	550	4	2 200	120	14 400	57 600
Total		10	4 300			96 000

La moyenne est égale à

$$\overline{X} \;=\; \frac{1}{N}\sum_i n_{i.} \,.\, c_i \;=\; \frac{4300}{10} \;=\; 430\ DA$$

et la variance est égale à

$$Var(X) \;=\; \frac{1}{N}\sum_i n_{i.} \,.\, (c_i - \overline{X})^2 \;=\; \frac{96000}{10} \;=\; 9600\ DA.$$

Reprenons à présent dans le tableau ci-dessous la distribution marginale en Y et complétons-la avec les colonnes contenant les produits $n_{j.}c'_j$, $(c'_j - \overline{Y})$, $(c'_j - \overline{Y})^2$ et $n_{j.}(c'_j - \overline{Y})^2$ (pour le calcul de la moyenne et de la variance).

Classe des bourses Y	**Centre de classe** c'_j	**Nombre d'étudiants** $n_{.j}$	$n_{.j}\,.\,c'_j$	$c'_j - \overline{Y}$	$\left(c'_j - \overline{Y}\right)^2$	$n_{.j}\,.\,\left(c'_j - \overline{Y}\right)^2$
[1 600 ; 2 400[	2 000	3	6 000	-720	518 400	1 555 200
[2 400 ; 3 200[	2 800	5	14 000	80	6 400	32 000
[3 200 ; 4 000[	3 600	2	7 200	880	7740 400	1 548 800
Total		10	27 200			3 136 000

La moyenne est égale à

$$\overline{Y} \;=\; \frac{1}{N}\sum_j n_{.j}\,.\,c'_j \;=\; \frac{27200}{10} \;=\; 2720\;DA$$

et la variance est égale à

$$Var(Y) \;=\; \frac{1}{N}\sum_j n_{.j}\,.\,\left(c'_j - \overline{Y}\right)^2 \;=\; \frac{3136000}{10} \;=\; 313600\;DA.$$

Construisons le tableau des effectifs cumulés pour le calcul de la médiane, nous avons :

Classe des bourses Y	**Nombre d'étudiants** $n_{.j}$	**Effectif cumulé** $n_{.j}\nearrow$
[1 600 ; 2 400[	3	3
[2 400 ; 3 200[	5	8
[3 200 ; 4 000[	2	10
Total	10	

TABLE 1.7 – Tableau des effectifs cumulés de la variable "montant mensuel de la bourse"

L'effectif cumulé théoriquement associé à la médiane est égal à $N/2 = 10/2 = 5$. Il s'ensuit que $N/2$ appartient à la classe $[2400; 3200[$ puisque celle-ci est la première classe à avoir un effectif cumulé supérieur ou égal à 5.

La médiane de la distribution groupée des montants de bourses est dès lors égale à :

$$Me \;=\; 2400 \;+\; \frac{5-2}{8-2}\,(3200 - 2400) \;=\; 2800\;DA.$$

Nous pourrons en conclure qu'approximativement la moitié des étudiants enquêtées reçoivent un montant de bourse au plus 2800 DA.

3. Déterminez la classe modale des dépenses ainsi que la dépense médiane dans le groupe des étudiants enquêtés.

La classe modale des dépenses est celle contenant le plus grand nombre des étudiants de l'échantillon. Il s'agit de la classe $[250; 450[$ DA, l'effectif qui lui est associé est égal à 6.

Pour déterminer la dépense médiane de l'échantillon des étudiants enquêtés (ou du moins une valeur approchée de cet dépense médiane, la distribution des dépenses étant groupée), considérons la distribution marginale en X et complétons-la avec les effectifs cumulés.

Classe des dépenses X	**Nombre d'étudiants** $n_{i.}$	**Effectif cumulé** $n_{i.} \nearrow$
[250 ; 450[	6	6
[450 ; 650[	4	10
Total	10	

TABLE 1.8 – Tableau des effectifs cumulés de la variable "les dépenses en biens ou services de loisirs journaliers"

L'effectif cumulé théoriquement associé à la médiane est égal à $N/2 = 10/2 = 5$. Il s'ensuit que $N/2$ appartient à la classe $[250; 450[$ DA.

La médiane de la distribution groupée des dépenses est dès lors égale à :

$$Me \;=\; 250 \;+\; \frac{5-6}{10-6}\,(450-250) \;=\; 187.5\;DA.$$

Nous pourrons en conclure qu'approximativement la moitié des étudiants enquêtées dépensent au plus 187.5 DA.

1.4 Distributions conditionnelles d'un couple de variables statistiques (X,Y)

Les distributions conditionnelles sont obtenues en fixant l'un des deux caractères et en variant les modalités de l'autre. Ainsi on pourra définir :

◇ ***s-distributions conditionnelles de la variable*** X ***sachant*** $Y = y_j$. Nous définissons la distribution conditionnelle de X en Y, où Y est fixé à la valeur y_j, par $\{(x_i; n_{ij}),\ \forall i \in \{1, ..., r\}$ et j fixé$\}$ et remarquons que c'est une distribution observée univariée constituée de $n_{.j}$ observations.

Nous considérons pour chacune des distributions la j^ieme colonne (càd les n_{ij}, $\forall i \in \{1, ..., r\}$ et j fixé) comme étant la colonne des effectifs conditionnels et les fréquences associées représentant la proportion des individus présentant les modalités x_i et la modalité y_j simultanément, appelée ***fréquences conditionnelles*** , sont données par l'expression suivante :

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

sachant que

$$\sum_i f_{i_{/j}} \;=\; \sum_i \frac{n_{ij}}{n_{.j}} \;=\; \frac{n_{.j}}{n_{.j}} \;=\; 1.$$

$X/_{Y=y_j}$	$n_{.j}$	$f_{i/_j}$
x_1	n_{1j}	$f_{1/_j}$
x_2	n_{2j}	$f_{2/_j}$
$\vdots$	$\vdots$	$\vdots$
x_i	n_{ij}	$f_{i/_j}$
$\vdots$	$\vdots$	$\vdots$
x_r	n_{rj}	$f_{r/_j}$
		1

TABLE 1.9 – Tableau des effectifs et des fréquences conditionnels de la variable X

Les s-distributions conditionnelles de fréquences ainsi définies sont encore appelées les ***profils-colonnes*** associés au tableau de contingence.

Notons aussi que la distribution des fréquences marginales $\{(x_i; f_{ij}), \forall i \in \{1, ..., r\}$ et j fixé$\}$ est parfois appelée ***profil-colonne marginal***.

$\diamond$ ***r-distributions conditionnelles de la variable Y sachant $X = x_i$***. Nous définissons la distribution conditionnelle de Y en X, où X est fixé à la valeur x_i, par $\{(y_j; n_{ij}),\ \forall j \in \{1, ..., s\}$ et i fixé$\}$ et remarquons que c'est une distribution observée univariée constituée de $n_{i.}$ observations.

Nous considérons pour chacune des distributions la i^ieme colonne (càd les n_{ij}, i fixé et $\forall j \in \{1, ..., s\}$) comme étant la colonne des effectifs conditionnels et les fréquences associées représentant la proportion des individus présentant les modalités y_j et la modalité x_i simultanément, appelée ***fréquences conditionnelles*** , sont données par l'expression suivante :

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

sachant que

$$\sum_j f_{j/i} = \sum_j \frac{n_{ij}}{n_{i.}} = \frac{n_{i.}}{n_{i.}} = 1.$$

$Y/_{X=x_i}$	$n_{i.}$	$f_{i/_j}$
y_1	n_{i1}	$f_{i/_1}$
y_2	n_{i2}	$f_{i/_2}$
$\vdots$	$\vdots$	$\vdots$
y_i	n_{ij}	$f_{i/_j}$
$\vdots$	$\vdots$	$\vdots$
y_s	n_{is}	$f_{i/_s}$
		1

TABLE 1.10 – Tableau des effectifs et des fréquences conditionnels de la variable Y

Les r-distributions conditionnelles de fréquences ainsi définies sont encore appelées les ***profils-lignes*** associés au tableau de contingence.

Notons aussi que la distribution des fréquences marginales $\{(y_j; f_{ij}), \forall j \in \{1, ..., s\}$ et i fixé$\}$ est parfois appelée ***profil-ligne marginal***.

Les distributions marginales et conditionnelles sont des distributions observées univariées et sont analysées de la même manière que dans le chapitre de statistique descriptive univariée. Cette analyse peut s'avérer utile puisque la comparaison des profils-colonnes ou lignes entre eux et avec le profil-colonne ou ligne marginal permet de révéler une structure d'association entre les variables X et Y.

1.4.1 Paramètres de position et de dispersion conditionnels

1.4.1.1 Moyenne conditionnelle de X

Il y a s-moyennes conditionnelles de la variable X sachant $Y = y_j$, $\forall j \in \{1, .., s\}$. La moyenne conditionnelle de la variable X sachant $Y = y_j$, notée $\overline{X}_j$, correspond à la moyenne du nombre des individus, constituée de $n_{.j}$ observations, présentant la modalité x_i et la modalité y_j simultanément et est donnée par l'expression suivante :

$$\overline{X}_j \;=\; \frac{1}{n_{.j}} \sum_i n_{ij}\, x_i \;=\; \sum_i f_{ij}\, x_i \quad \textit{si } X \textit{ est une variable discrete}$$

et

$$\overline{X}_j \;=\; \frac{1}{n_{.j}} \sum_i n_{ij}\, c_i \;=\; \sum_i f_{ij}\, c_i \quad \textit{si } X \textit{ est une variable continue}$$

où c_i est le centre de classe $[x_i; x_{i+1}[$.

1.4.1.2 Moyenne conditionnelle de Y

Il y a r-moyennes conditionnelles de la variable Y sachant $X = x_i$, $\forall i \in \{1, .., r\}$. La moyenne conditionnelle de la variable Y sachant $X = x_i$, notée $\overline{Y}_i$, correspond à la moyenne du nombre des individus, constituée de $n_{i.}$ observations, présentant la modalité y_j et la modalité x_i simultanément et est donnée par l'expression suivante :

$$\overline{Y}_i \;=\; \frac{1}{n_{i.}} \sum_j n_{ij}\, y_j \;=\; \sum_j f_{ij}\, y_j \quad \textit{si } Y \textit{ est une variable discrete}$$

et

$$\overline{Y}_i \;=\; \frac{1}{n_{i.}} \sum_j n_{ij}\, c'_j \;=\; \sum_j f_{ij}\, c'_j \quad \textit{si } Y \textit{ est une variable continue}$$

où c'_j est le centre de classe $[y_j; y_{j+1}[$.

1.4.1.3 Variance conditionnelle de X

La variance conditionnelle de la variable X sachant $Y = y_j$, notée $V_j(X)$, est la mesure de la dispersion des individus de l'échantillon de $n_{.j}$ observations présentant la modalité x_i et la modalité y_j simultanément et est définie par :

$$V_j(X) \;=\; \frac{1}{n_{.j}} \sum_i n_{ij}\, (x_i \,-\, \overline{X}_j)^2 \;=\; \sum_i f_{ij} \Big(x_i - \, \overline{X}_j \Big)^2 \quad \textit{si } X \textit{ est une variable discrete}$$

et

$$V_j(X) \;=\; \frac{1}{n_{.j}} \sum_i n_{ij}\, (c_i - \, \overline{X}_j)^2 \;=\; \sum_i f_{ij} \left(c_i - \, \overline{X}_j \right)^2 \quad \textit{si } X \textit{ est une variable continue}$$

ou encore

$$V_j(X) \;=\; \frac{1}{n_{.j}} \sum_i n_{ij}\, x_i \;-\; \overline{X}_j^2 \;=\; \sum_i f_{ij}\, x_i - \; \overline{X}_j^2 \quad \textit{si } X \textit{ est une variable discrete}$$

et

$$V_j(X) \;=\; \frac{1}{n_{.j}} \sum_i n_{ij}\, c_i - \; \overline{X}_j^2 \;=\; \sum_i f_{ij}\, c_i - \; \overline{X}_j^2 \quad \textit{si } X \textit{ est une variable continue}$$

où c_i est le centre de classe $[x_i; x_{i+1}[$.

1.4.1.4 Variance conditionnelle de Y

La variance conditionnelle de la variable Y sachant $X = y_j$, notée $V_i(Y)$, est la mesure de la dispersion des individus de l'échantillon de $n_{.j}$ observations présentant la modalité x_i et la modalité y_j simultanément et est définie par :

$$V_i(Y) \;=\; \frac{1}{n_{i.}} \sum_j n_{ij}\, (y_j \;-\; \overline{Y}_i)^2 \;=\; \sum_j f_{ij} \Big(x_i - \; \overline{Y}_i \Big)^2 \quad \textit{si } Y \textit{ est une variable discrete}$$

et

$$V_i(Y) \;=\; \frac{1}{n_{i.}} \sum_j n_{ij}\, (c'_j - \; \overline{Y}_i)^2 \;=\; \sum_j f_{ij} \left(c'_j - \; \overline{Y}_i \right)^2 \quad \textit{si } Y \textit{ est une variable continue}$$

ou encore

$$V_i(Y) \;=\; \frac{1}{n_{i.}} \sum_j n_{ij}\, y_j \;-\; \overline{Y}_i^2 \;=\; \sum_j f_{ij}\, y_i - \; \overline{Y}_i^2 \quad \textit{si } Y \textit{ est une variable discrete}$$

et

$$V_i(Y) \;=\; \frac{1}{n_{i.}} \sum_j n_{ij}\, c'_j - \; \overline{Y}_i^2 \;=\; \sum_j f_{ij}\, c'_j - \; \overline{Y}_i^2 \quad \textit{si } Y \textit{ est une variable continue}$$

où c'_j est le centre de classe $[y_j; y_{j+1}[$.

1.4.1.5 Relation entre moyennes et variances

Nous avons :

$$\overline{X} \;=\; \frac{1}{N} \sum_j n_{.j}\, \overline{X}_j \quad \textit{et} \quad \overline{Y} \;=\; \frac{1}{N} \sum_i n_{i.}\, \overline{Y}_i$$

et

$$V(X) \;=\; \frac{1}{N} \sum_j n_{.j}\, V_j(X) \;+\; \frac{1}{N} \sum_j n_{.j} \left(\overline{X}_j - \overline{X} \right)^2,$$

$$V(Y) \;=\; \frac{1}{N} \sum_i n_{i.}\, V_i(Y) \;+\; \frac{1}{N} \sum_i n_{i.} \left(\overline{Y}_i - \overline{Y} \right)^2.$$

On retient donc :

Moyenne marginale = la moyenne des moyennes conditionnelles pondérées par les effectifs marginaux

1.5 Coefficient de corrélation d'un couple statistique (X,Y)

Lorsqu'une série statistique à deux variables quantitatives est représentée au moyen d'un graphique de dispersion ou un nuage de point, on s'intéresse toujours à déceler une structure d'association entre les deux variables. Le ***coefficient de corrélation*** permettra de quantifier l'intensité de la liaison qui pourra exister entre ces deux variables dans le cas où l'association est de nature linéaire.

La notion du coefficient de corrélation peut être attribuée au physicien français ***Auguste Bravais*** par le biais de ses travaux effectués dans l'étude des erreurs dans les tirs d'artillerie mais aussi à ***Francis Galton*** qui grâce à lui que la corrélation devient un concept statistique. C'est ensuite ***karl Pearson*** qui propose en 1986 la formulation mathématique actuelle.

Afin d'introduire le coefficient de corrélation, il nous faut considérer au préalable celui de la ***covariance***, qui est une généralisation bidimensionnelle de la variance, définie par :

Définition 1.5.1. *La covariance est désignée par $Cov(X,Y)$ et définie par l'expression :*

$$Cov(X,Y) \;=\; \frac{1}{N}\sum_{i=1} n_{ij}\,\left(x_i-\overline{X}\right)\left(y_i-\overline{Y}\right)$$

ou bien

$$Cov(X,Y) \;=\; \frac{1}{N}\sum_{i=1} n_{ij}\,x_i\,.\,y_i-\overline{X}\,.\,\overline{Y}$$

où $\overline{X}$ et $\overline{X}$ désignent les moyennes des séries marginales.

Nous retenons donc :

Covariance = moyenne des produits des écarts à la moyenne des deux variables X et Y,

ou bien encore

Covariance = moyenne des produits moins le produit des moyennes des deux variables X et Y.

1.5.1 Propriétés de la covariance

P1 - $Cov(X,X) \;=\; Var(X)$.

P2 - La covariance est une forme bilinéaire symétrique, i.e. :

$$Cov(X,Y) \;=\; Cov(Y,X).$$

P3 - $Cov(a\,X,Y) \;=\; a\,Cov(X,Y)$ et $Cov(X,b\,Y) \;=\; b\,Cov(X,Y)$.

P4 - $Cov(\sum_i X_i,\sum_j Y_j) \;=\; \sum_i\,\sum_j\,Cov(X_i,Y_j)$.

P5 - La covariance peut prendre toute valeur réelle et dont la variance est la forme quadratique associée. En particulier, on en déduit les deux formules suivantes :

$$Var(X+Y) \;=\; Var(X) \;+\; Var(Y) \;+\; Cov(X,Y).$$

P6 - $Var(a\,X+b\,Y) \;=\; a^2\,Var(X) \;+\; b^2\,Var(Y) \;+\; 2\,a\,b\,Cov(X,Y)$.

P7 - Signalons aussi une propriété fondamentale appelée l'***Inégalité de Cauchy-Schwartz*** et permettant de définir plutard le coefficient de corrélation linéaire :

$$Cov^2(X,Y) \leq Var(X)\, Var(Y).$$

Comme la covariance dépend des unités de mesure des deux variables X et Y, on doit la rendre intrinsèque en la divisant par le produit des écart-types des deux variables considérées.

Remarque 1.5.1. *Le concept de covariance se généralise à plusieurs variables (vecteur aléatoire) par la* ***matrice de covariance*** *(ou* ***matrice de variance-covariance****) qui, pour un ensemble de r variables aléatoires réelles $X_1, \ldots, X_r$, est la matrice carrée dont l'élément de la ligne i et de la colonne j est la covariance des variables X_i et X_j et on obtient :*

$$Var\Big(\sum_i X_i\Big) \;=\; \sum_i Var(X_i) \;+\; 2 \sum_{1<i<j<r} Cov(X_i, X_j).$$

1.5.2 Interprétation de la covariance

La covariance peut être positive ou négative selon la position des points par rapport au centre de gravité du nuage de points, dont les coordonnées sont $(\overline{X}, \overline{Y})$. Considérons à présent les trois sortes nuages de points :

◇ On remarque bien que si dans un nuage A, faisant apparaître une association linéaire, les points du nuage ont tendance à se concentrer autour d'une droite (c-à-d une structure positive, car lorsque la valeur de x augmente, celle de y a également tendance à augmenter), ainsi la plupart des points du nuage se trouvent dans les quadrants I et III et donnent donc lieu à des produits $\big(x_i - \overline{X}\big)\,\big(y_i - \overline{Y}\big)$ presque positifs, aussi bien que la covariance.

◇ Contrairement au nuage A, si un nuage B met en évidence une association linéaire et négative (puisque lorsque la valeur de x augmente, celle de y a au contraire tendance à diminuer), on remarque alors que presque tous les points du nuage occupent les quadrants II et IV, d'où tous les produits $\big(x_i - \overline{X}\big)\,\big(y_i - \overline{Y}\big)$ sont négatifs et la covariance est alors négative.

◇ Si un nuage C ne montre aucune structure particulière puisqu'il semble que les deux variables ne sont pas liées entre elles, alors les points du nuage se répartissent d'une manière équitable dans les quatre quadrants du plan et la somme des produits positifs compense la somme des produits négatifs, donnant ainsi lieu à une covariance pratiquement nulle.

Exemple : Sur une année glissante, on a mesuré pour un échantillon de 10 jeunes âgés de 11 à 16 ans, l'âge (variable X) et la durée journalière moyenne durée d'écoute de leur MP3 (variable Y exprimée en heure). La série statistique observée ainsi que sa représentation graphique au moyen d'un nuage de points sont présentées ci-après :

x_i	11	11	12	12	13	13	14	15	15	16
y_i	2	2.1	2.7	3	4	4.5	5	6.5	6.8	7.6

TABLE 1.11 – Répartition d'une population de 10 jeunes suivant l'âge et la durée journalière moyenne durée d'écoute de leur MP3

Le calcul de la covariance peut s'effectuer au moyen du tableau suivant :

x_i	y_i	$(x_i - \overline{X})$ avec $\overline{X} = 13.2$	$(y_i - \overline{Y})$ avec $\overline{Y} = 4.42$	$(x_i - \overline{X})^2$	$(y_i - \overline{Y})^2$	$(x_i - \overline{X})\ (y_i - \overline{Y})$
11	2	-2.2	-2.42	4.84	5.8564	5.324
11	2.1	-2.2	-2.32	4.84	5.3824	5.104
12	2.7	-1.2	-1.72	1.44	2.9584	2.064
12	3	-1.2	-1.42	1.44	2.0164	1.704
13	4	-0.2	-0.42	0.04	0.1764	0.084
13	4.5	-0.2	0.08	0.04	0.0064	-0.016
14	5	0.8	0.58	0.64	0.3364	0.464
15	6.5	1.8	2.08	3.24	4.3264	3.744
15	6.8	1.8	2.38	3.24	5.6644	4.284
16	7.6	2.8	3.18	7.84	10.1124	8.904
132	44.2	0	0	27.6	36.836	31.66

Par suite, on obtient :

$$Cov(X,Y) \ = \ \frac{1}{10} \ \sum_{i=1}^{10} (x_i - \overline{X}) \ (y_i - \overline{Y}) \ = \ \frac{31.66}{10} \ = \ 3.166.$$

Le ***coefficient de corrélation de Bravais-Pearson***, désigné par r ou r_{XY}, est défini comme suit :

$$r_{XY} \ = \ \frac{Cov(X,Y)}{\sigma_X \ \sigma_Y}.$$

Dans cette expression, $Cov(X,Y)$ est la covariance et σ_X et σ_Y sont les écarts-types des distributions marginales en X et en Y.

Le coefficient de corrélation a nécessairement le même signe que la covariance, puisque les écarts-types sont des nombres positifs. Ceci s'interprète de la même manière suivante :

$\diamond$ s'il existe une association linéaire et positive entre les deux variables, la covariance et le coefficient de corrélation sont tous deux positifs,

$\diamond$ s'il existe une association linéaire et négative entre les deux variables, la covariance et le coefficient de corrélation sont tous deux négatifs,

$\diamond$ s'il n'existe pas d'association entre les deux variables, la covariance et le coefficient de corrélation ont tous deux des valeurs proches de zéro.

Exemple : Reprenons l'exercice précédent portant sur les deux variables "âge et durée journalière moyenne durée d'écoute de leur MP3". Le coefficient de corrélation est égale à :

$$r \ = \ \frac{Cov(X,Y)}{\sigma_X \ \sigma_Y} \ = \ \frac{31.66)}{\sqrt{2.76}\ \sqrt{3.68}} \ = \ 0.9929.$$

1.6 Indépendance de deux variables X et Y

Deux variables statistiques sont ***statistiquement indépendantes*** si et seulement si l'une des propriétés suivantes est vérifiée :

P1 - Les distributions conditionnelles sont identiques à la distribution marginale correspondante, i.e. :

$$f_{i/j} = f_{i.} \quad ou \quad f_{j/i} = f_{.j}, \ \forall i \in \{1, ..., r\} \ et \ \forall j \in \{1, ..., s\}.$$

P2 - Les fréquences partielles sont égales au produit des fréquences marginales correspondantes, i.e. :

$$f_{ij} = f_{i.} \, . \, f_{.j}, \ \forall i \in \{1, ..., r\} \ et \ \forall j \in \{1, ..., s\}.$$

P3 - Les effectifs partiels sont égaux au produit des effectifs marginaux correspondants, i.e. :

$$n_{ij} = n_{i.} \, . \, n_{.j}, \ \forall i \in \{1, ..., r\} \ et \ \forall j \in \{1, ..., s\}.$$

P4 - Les moyennes conditionnelles sont égales entre elles d'une part et égales à la moyenne marginale d'autre part, i.e. :

$$\overline{X} = \overline{X}_1 = \overline{X}_2 = ... = \overline{X}_j = ... = \overline{X}_s \quad ou \quad \overline{Y} = \overline{Y}_1 = \overline{Y}_2 = ... = \overline{Y}_i = ... = \overline{Y}_r.$$

1.7 Droites de régression

Si l'examen du nuage de points représentant la série statistique bivariée décèle une relation de dépendance statistique linéaire de Y en X, il est naturel de représenter graphiquement cette relation particulière à l'aide d'une droite traversant le nuage de points, appelée ***droite de régression de*** Y ***en*** X qui ajuste au mieux le nuage. Le critère d'optimalité que nous allons considérer est celui des moindres carrés.

1.7.1 Détermination de la droite de de régression

Comme toute droite, la droite de régression de Y en X peut être définie au moyen d'une équation du premier degré de la forme :

$$y = a \, x \, + b.$$

Le principe de moindres carrés consiste à minimiser la fonctionnelle :

$$F(a, b) = \sum_i [y_i - (a \, x_i + b)]^2,$$

où $r_i = y_i - (a \, x_i + b)$ est le ***résidu*** ou l'***erreur d'ajustement*** entre la valeur réellement observée y_i pour la variable dépendante et la valeur ajustée fournie par la droite de régression $a \, x_i + b$. Nous pouvons dès lors considérer que le meilleur ajustement est fourni par la droite qui minimise globalement l'amplitude des erreurs d'ajustement r_i, $\forall i$. La détermination de a et b et est un problème classique de minimisation, dont la solution est :

$$a = \frac{Cov(X, Y)}{\sigma(X)^2} \quad et \quad b = \overline{Y} - a \, \overline{X}.$$

Nous constatons ainsi que la droite de régression de Y en X passe par le barycentre ou centre de gravité du nuage de points de coordonnées $(\overline{X}, \overline{Y})$ et que l'orientation de la droite de régression indique la nature de liaison entre les deux variables de telle manière que :

◇ si la covariance est positive, par suite le coefficient de corrélation et a le sont, ce qui signifie que la droite est ascendante,

◇ si la covariance est négative, par suite le coefficient de corrélation et a le sont, ce qui signifie que la droite est descendante.

Exemple : 100 salariés ont été observés selon les deux caractères salaires en 10^3 DA (X) et le nombre d'enfants (Y). Le tableau de contingence qui donne la distribution jointe de (X, Y) est le suivant :

Salaires \ **Nombre d'enfants**	0	1	2	3	4	**Colonnes marginales**
[0,5[	2	3	1	3	1	10
[05,10[	4	6	2	6	2	20
[10,14[	4	6	2	6	2	20
[14,16[	8	12	4	12	4	40
[16,20[	2	3	1	3	1	10
lignes marginales	20	30	10	30	10	100

TABLE 1.12 – Tableau de contingence des deux caractères "salaires et nombre d'enfants"

1. Montrer que X et Y sont indépendants.
2. Déterminer la covariance de X et Y.
3. Déterminer la droite de régression de Y sur X.

Solution :

1. Montrer que X et Y sont indépendants

1^{ere} méthode : X et Y sont statistiquement indépendants si et seulement si :

$$f_{ij} = f_{i.} \times f_{.j}, \quad \forall\ i, j \in \{1, .., 5\}.$$

Pour $i = 1$ et $j \in \{1, .., 5\}$, on a :

$$\begin{cases} f_{11} = \dfrac{2}{100} \text{ et } f_{1.} \times f_{.1} = \dfrac{10}{100} \times \dfrac{20}{100} \quad \text{d'où } f_{11} = f_{1.} \times f_{.1}, \\ f_{12} = \dfrac{3}{100} \text{ et } f_{1.} \times f_{.2} = \dfrac{10}{100} \times \dfrac{30}{100} \quad \text{d'où } f_{12} = f_{1.} \times f_{.2}, \\ f_{13} = \dfrac{1}{100} \text{ et } f_{1.} \times f_{.3} = \dfrac{10}{100} \times \dfrac{10}{100} \quad \text{d'où } f_{13} = f_{1.} \times f_{.3}, \\ f_{14} = \dfrac{4}{100} \text{ et } f_{1.} \times f_{.4} = \dfrac{10}{100} \times \dfrac{30}{100} \quad \text{d'où } f_{14} = f_{1.} \times f_{.4}, \\ f_{15} = \dfrac{1}{100} \text{ et } f_{1.} \times f_{.5} = \dfrac{10}{100} \times \dfrac{30}{100} \quad \text{d'où } f_{15} = f_{1.} \times f_{.5}. \end{cases}$$

Pour $i = 2$ et $j \in \{1, .., 5\}$, on a :

$$\begin{cases} f_{21} = \frac{4}{100} \text{ et } f_{2.} \times f_{.1} = \frac{20}{100} \times \frac{20}{100} \text{ d'où } f_{21} = f_{2.} \times f_{.1}, \\ f_{22} = \frac{6}{100} \text{ et } f_{2.} \times f_{.2} = \frac{20}{100} \times \frac{30}{100} \text{ d'où } f_{22} = f_{2.} \times f_{.2}, \\ f_{23} = \frac{2}{100} \text{ et } f_{2.} \times f_{.3} = \frac{20}{100} \times \frac{10}{100} \text{ d'où } f_{23} = f_{2.} \times f_{.3}, \\ f_{24} = \frac{6}{100} \text{ et } f_{2.} \times f_{.4} = \frac{20}{100} \times \frac{30}{100} \text{ d'où } f_{24} = f_{2.} \times f_{.4}, \\ f_{25} = \frac{2}{100} \text{ et } f_{2.} \times f_{.5} = \frac{20}{100} \times \frac{30}{100} \text{ d'où } f_{25} = f_{2.} \times f_{.5}. \end{cases}$$

De même, on démontre cette égalité pour les cas suivants : $i = 3$, $i = 4$, $i = 5$ et $j \in \{1, .., 5\}$ respectivement.

$\underline{2^{eme} \text{ méthode}}$: X et Y sont statistiquement indépendants si et seulement si les moyennes conditionnelles sont égales entre elles d'une part et égales à la moyenne marginale d'autre part :

$$\overline{X}_1 = \overline{X}_2 = \overline{X}_3 = \overline{X}_4 = \overline{X}_5 = \overline{X}$$

ou

$$\overline{Y}_1 = \overline{Y}_2 = \overline{Y}_3 = \overline{Y}_4 = \overline{Y}_5 = \overline{Y}.$$

Or

Salaires	**Centre de classe** c_i	$n_{i.}$	$n_{i.} \cdot c_i$	n_{i1}	$n_{i1} \cdot c_i$	n_{i2}	$n_{i2} \cdot c_i$	n_{i3}	$n_{i3} \cdot c_i$
[00,5[	2.5	10	25	2	5	3	7.5	1	2.5
[05,10[	7.5	20	150	4	30	6	45.0	2	15.0
[10,14[	12.0	20	240	4	48	6	72.0	2	24.0
[14,16[	15.0	40	600	8	120	12	180.0	4	60.0
[16,20[	18.0	10	180	2	36	3	54.0	1	18.0
Total		100	1195	20	239	30	358.5	10	119.5

Par suite, on a :

$$\overline{X} = \frac{1}{n_{i.}} \sum_{i=1}^{5} n_{i.}\, c_i = \frac{1195}{100} = 11.95 \,.10^3 DA$$

et

$$\overline{X}_1 = \frac{1}{n_{.1}} \sum_{i=1}{}^{5} n_{i1}\, c_i = \frac{239}{20} = 11.95 \,.10^3 DA,$$

$$\overline{X}_2 = \frac{1}{n_{.2}} \sum_{i=1}{}^{5} n_{i2}\, c_i = \frac{358.5}{30} = 11.95 \,.10^3 DA = \overline{X}_4,$$

$$\overline{X}_3 = \frac{1}{n_{.3}} \sum_{i=1}{}^{5} n_{i3}\, c_i = \frac{119.5}{10} = 11.95 \,.10^3 DA = \overline{X}_5.$$

D'où, X et Y sont statistiquement indépendants.

Ou bien,

Nombre d'enfants y_j	$n_{.j}$	$n_{.j}\ .\ y_j$	n_{1j}	$n_{1j}\ .\ y_j$	n_{2j}	$n_{2j}\ .\ y_j$	n_{4j}	$n_{4j}\ .\ y_j$
0	20	0	2	0	4	0	8	0
1	30	30	3	3	6	6	12	12
2	10	20	1	2	2	4	4	8
3	30	90	3	9	6	18	12	36
4	10	40	1	4	2	8	4	16
Total	100	180	10	18	20	36	40	72

Par suite, on a :

$$\overline{Y} \;=\; \frac{1}{n_{.j}}\sum_{i=1}^{5} n_{.j}\ y_i j \;=\; \frac{180}{100} \;=\; 1.8\ enfant$$

$$\overline{Y}_1 \;=\; \frac{1}{n_{1.}}\sum\nolimits_{j=1}{}^{5} n_{1j}\ y_j \;=\; \frac{18}{10} \;=\; 1.8\ enfant \;=\; \overline{Y}_5,$$

$$\overline{Y}_2 \;=\; \frac{1}{n_{2.}}\sum\nolimits_{j=1}{}^{5} n_{2j}\ y_j \;=\; \frac{36}{20} \;=\; 1.8\ enfant \;=\; \overline{Y}_3,$$

$$\overline{Y}_4 \;=\; \frac{1}{n_{3.}}\sum\nolimits_{j=1}{}^{5} n_{3j}\ y_j \;=\; \frac{72}{40} \;=\; 1.8\ enfant.$$

D'où, X et Y sont statistiquement indépendants.

2. Déterminer la covariance de X et Y La covariance de X et Y est donné par la formule suivante :

$$cov(X,Y) \;=\; \frac{1}{N}\sum_{i=1}^{5}\sum_{j=1}^{5} n_{ij}\ c_i\ y_j \;-\; \overline{X}\,\overline{Y}$$

y_j / c_i	0	1	2	3	4	**Total** $n_{ij}\ c_i\ y_j$
2.5	0	7.5	5	22.5	10	45
7.5	0	45.0	30	135.0	60	270
12.0	0	72.0	48	216.0	96	432
15.0	0	180.0	120	540.0	240	1080
18.0	0	54.0	36	162.0	72	324
Total	0	358.5	239	1075.5	478	2151

D'où

$$cov(X,Y) \;=\; \frac{2151}{100} \;-\; 11.95 \times 1.8 \;=\; 0.$$

3. Déterminer la droite de régression de Y sur X

D'après la méthode des moindres carrés ordinaire, on a l'estimation suivante :

$$Y = a\,x + b$$

avec $a = \dfrac{Cov(X,Y)}{\sigma(X)^2} = 0$ et $b = \overline{Y}$.

Chapitre 2

Exercices d'application avec corrigés sur la statistique descriptive bidimensionnelle

2.1 Exercice 1

On a examiné 1000 salariés d'une entreprise du point de vue leurs salaires en Dinars Algériens et leurs années d'expérience dans le secteur. Les résultats obtenus sont présentés dans le tableau suivant :

Années / **Salaires**	[0,10[	[10,15[	[15,20[	[20,30[	**Colonnes marginales**
[1000,2000[	220	100	90	50	460
[2000,3000[	120	180	80	40	420
[3000,4000[	40	40	30	10	120
lignes marginales	380	320	200	100	1000

TABLE 2.1 – Tableau de contingence des deux caractères "salaires et années d'expérience dans le secteur"

1. Lister les modalités de X et les modalités de Y.
2. Déterminer les distributions marginales correspondantes.
3. Déterminer les distributions des fréquences marginales correspondantes.
4. A partir de ce tableau, déterminer :
   a) Le nombre de salariés ayant des salaires inférieurs à 3 milles Dinars Algériens et des années d'expérience compris entre 15 et 30 ans.
   b) La proportion de salariés ayant moins que 4 milles Dinars Algériens comme salaires.
   c) La proportion de salariés ayant entre 2 et 4 milles Dinars Algériens de salaires et entre 10 et 20 ans d'expérience.
   d) Le nombre de salariés ayant 15 ans d'expérience et plus.
5. Calculer les moyennes, les variances et les écart-types marginales ainsi que la covariance.
6. Tracer l'histogramme de la distribution de Y.
7. Déterminer graphiquement et analytiquement le mode.
8. Tracer la courbe cumulative de la série statistique Y.
9. Déterminer graphiquement et analytiquement la médiane et interpréter le résultat.
10. Déterminer la droite de régression de Y sur X d'après la méthode de moindres carrés.

11. Calculer les moyennes, les variances et les écart-types conditionnelles de X et de Y.
12. Déterminer les distributions des fréquences conditionnelles.

Solution :

1. Lister les modalités de X et les modalités de Y

Les modalités de la variable statistique "salaires" sont : $[1000, 2000]$, $[2000, 3000]$ et $[3000, 4000]$ et ceux de la variable "années d'expérience" sont : $[0, 10]$, $[10, 15]$, $[15, 20]$ et $[20, 30]$.

2. Déterminer les distributions marginales correspondantes

Les distributions marginales correspondantes :

Années / **Salaires**	[00,10[	[10,15[	[15,20[	[20,30[	**Colonnes marginales**
[1000,2000[	$n_{11} = 220$	$n_{12} = 100$	$n_{13} = 90$	$n_{14} = 50$	$n_{1.} = 460$
[2000,3000[	$n_{21} = 120$	$n_{22} = 180$	$n_{23} = 80$	$n_{24} = 40$	$n_{2.} = 420$
[3000,4000[	$n_{31} = 040$	$n_{32} = 040$	$n_{33} = 30$	$n_{34} = 10$	$n_{3.} = 120$
lignes marginales	$n_{.1} = 380$	$n_{.2} = 320$	$n_{.3} = 200$	$n_{.4} = 100$	$n_{..} = 1000$

TABLE 2.2 – Tableau des distributions marginales des deux caractères "salaires et années d'expérience dans le secteur"

3. Déterminer les distributions marginales des fréquences correspondantes

Les distributions marginales des fréquences correspondantes :

Années / **Salaires**	[00,10[	[10,15[	[15,20[	[20,30[	**Colonnes marginales**
[1000,2000[	$f_{11} = 0.22$	$f_{12} = 0.10$	$f_{13} = 0.09$	$f_{14} = 0.05$	$f_{1.} = 0.46$
[2000,3000[	$f_{21} = 0.12$	$f_{22} = 0.18$	$f_{23} = 0.08$	$f_{24} = 0.04$	$f_{2.} = 0.42$
[3000,4000[	$f_{31} = 0.04$	$f_{32} = 0.04$	$f_{33} = 0.03$	$f_{34} = 0.01$	$f_{3.} = 0.12$
lignes marginales	$f_{.1} = 0.38$	$f_{.2} = 0.32$	$f_{.3} = 0.20$	$f_{.4} = 0.10$	$f_{..} = 1$

TABLE 2.3 – Tableau des fréquences marginales des deux caractères "salaires et années d'expérience dans le secteur"

4. A partir de ce tableau, déterminer

a). Le nombre de salariés ayant des salaires inférieurs à 3 milles Dinars Algériens et des années d'expérience compris entre 15 et 30 ans

b). La proportion de salariés ayant moins que 4 milles Dinars Algériens comme salaires

c). La proportion de salariés ayant entre 2 et 4 milles Dinars Algériens de salaires et entre 10 et 20 ans d'expérience

d). Le nombre de salariés ayant 15 ans d'expérience et plus

5. Calculer les moyennes, les variances et les écart-types marginales ainsi que la covariance

◇ La distribution marginale de X :

Salaires	**Centre de classe** c_i	**Nombre de salariés** $n_{i.}$	$n_{i.}\ .\ c_i$	$c_i - \overline{X}$	$\left(c_i - \overline{X}\right)^2$	$n_{i.}\ .\ \left(c_i - \overline{X}\right)^2$
[1000 ; 2000[	1500	460	690 000	- 660	435 600	200 376 000
[2000 ; 3000[	2500	420	1 050 000	340	115 600	48 552 000
[3000 ; 4000[	3500	120	420 000	1 340	1 795 600	215 472 000
Total		1000	2 160 000			464 400 000

TABLE 2.4 – Tableau de la distribution marginale de la variable X

La moyenne de la variable statistique "salaires" est donnée par la formule suivante :

$$\overline{X} \ = \ \frac{1}{N}\sum_{i=1}^{3} n_{i.}\ c_i \ = \ \frac{2160000}{1000} \ = \ 2160.$$

Par conséquent, le salaire moyen des salariés de cet entreprise est égal à 2160 euros.

L'écart-type est donné par la formule suivante :

$$\sigma(X) = \sqrt{V(X)} = \sqrt{\frac{1}{N}\sum\nolimits_{i}^{3} n_{i.}\ (c_i - \overline{X})^2} = \sqrt{\frac{464400000}{1000}} = 681.469\ euros.$$

◇ La distribution marginale de Y :

Années d'expérience	**Centre de classe** c'_j	**Nombre de salariés** $n_{.j}$	$n_{.j}\ .\ c'_j$	$c'_j - \overline{Y}$	$\left(c'_j - \overline{Y}\right)^2$	$n_{.j}\ .\ \left(c'_j - \overline{Y}\right)^2$
[00 ; 10[	05	380	1 900	- 6.9	47.61	18 091.8
[10 ; 15[	12.5	320	4 000	0.6	0.36	115.2
[15 ; 20[	17.5	200	3 500	5.6	31.36	6 272.0
[20 ; 30[	25	100	2 500	13.1	176.61	17 161.0
Total		1000	11 900			41 640

TABLE 2.5 – Tableau de la distribution marginale de la variable Y

La moyenne de la variable statistique "années d'expérience" est donnée par la formule suivante :

$$\overline{Y} \ = \ \frac{1}{N}\sum_{j=1}^{4} n_{.j}\ c'_j \ = \ \frac{11900}{1000} \ = \ 11.9.$$

Par conséquent, les salariés de cet entreprise ont en moyenne 11.9 années d'expérience.

L'écart-type est donné par la formule suivante :

$$\sigma(Y) = \sqrt{V(Y)} = \sqrt{\frac{1}{N}\sum_j^4 n_{.j}\ (c'_j - \overline{X})^2} = \sqrt{\frac{41640}{1000}} = 6.452\ annes.$$

◇ La covariance de X et Y est donné par la formule suivante :

$$cov(X,Y) \ = \ \frac{1}{N}\sum_{i=1}^{3}\sum_{j=1}^{4} n_{ij}\ c_i\ c'_j \ - \ \overline{X}\ \overline{Y}$$

avec

c'_j / c_i	5	12.5	17.5	25	**Total** $n_{ij}\ c_i\ c'_j$
1500	$n_{11}\ c_1\ c'_1$	$n_{12}\ c_1\ c'_2$	$n_{13}\ c_1\ c'_3$	$n_{14}\ c_1\ c'_4$	$\sum_{j=1}^{4} n_{1j}\ c_1\ c'_j$
2500	$n_{21}\ c_2\ c'_1$	$n_{22}\ c_2\ c'_2$	$n_{23}\ c_2\ c'_3$	$n_{24}\ c_2\ c'_4$	$\sum_{j=1}^{4} n_{2j}\ c_2\ c'_j$
3500	$n_{31}\ c_3\ c'_1$	$n_{32}\ c_3\ c'_2$	$n_{33}\ c_3\ c'_3$	$n_{34}\ c_3\ c'_4$	$\sum_{j=1}^{4} n_{3j}\ c_3\ c'_j$
Total	$\sum_{i=1}^{3} n_{i1}\ c_i\ c'_1$	$\sum_{i=1}^{3} n_{i2}\ c_i\ c'_2$	$\sum_{i=1}^{3} n_{i3}\ c_i\ c'_3$	$\sum_{i=1}^{3} n_{i4}\ c_i\ c'_4$	$\sum_{i=1}^{3}\sum_{j=1}^{4} n_{ij}\ c_i\ c'_j$

c'_j / c_i	5	12.5	17.5	25	**Total** $n_{ij}\ c_i\ c'_j$
1500	1 650 000	1 875 000	2 362 500	1 875 000	7 762 500
2500	1 500 000	5 625 000	3 500 000	2 500 000	13 125 000
3500	700 000	1 750 000	1 837 500	875 000	5 162 500
Total	3 850 000	9 250 000	7 700 000	5 250 000	26 050 000

D'où

$$cov(X,Y) \ = \ \frac{26050000}{1000} \ - \ 2160 \times 11.9 \ = \ 346.$$

6. Tracer l'histogramme de la distribution de Y

Dressons le tableau statistique attribué à cette distribution. Vu que les amplitudes des classes sont variables, il faut au préalable corriger les effectifs et les fréquences afin de tracer l'histogramme des fréquences.

Salaires	**Amplitude de classe** a_i	**Nombre de salariés** n_i	**Effectif corrigé** $(n_i)/(a_i)$
[00 ; 10[	10	380	38
[10 ; 15[	5	320	64
[15 ; 20[	5	200	40
[20 ; 30[	10	100	10
Total		1000	

Nous obtenons l'histogramme suivant :

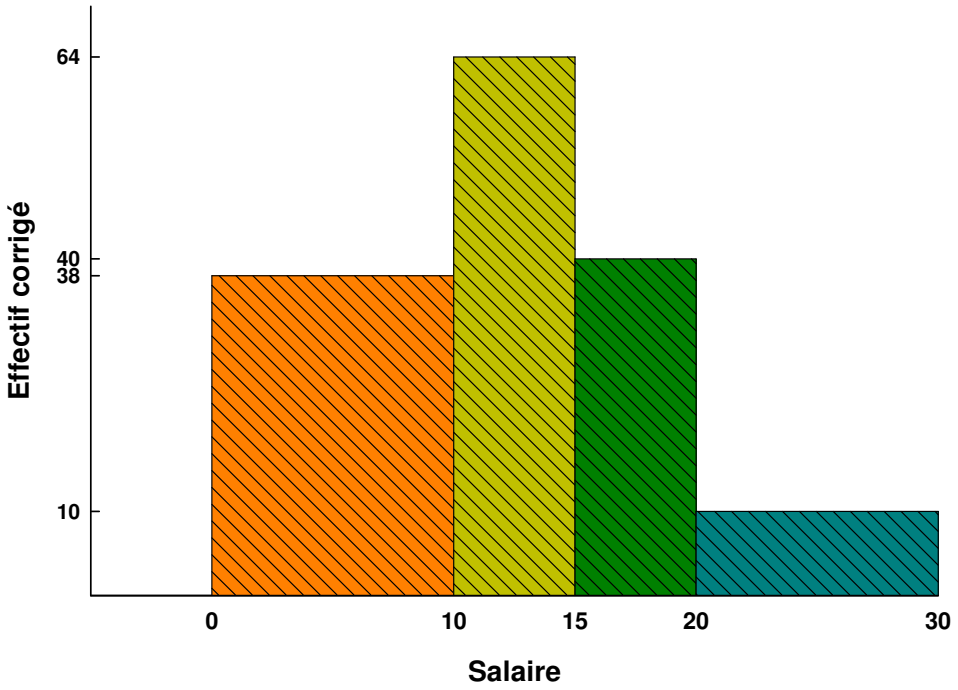

FIGURE 2.1 – Histogramme des effectifs corrigés de la variable "salaire"

7. Déterminer graphiquement et analytiquement le mode

La classe modale, à laquelle est associé l'effectif ou la fréquence corrigés les plus importants est la classe $[10; 15]$ année.

Graphiquement, la première méthode pour le calcul du mode consiste à joindre les segments [AB] et [CD] et déterminer le point d'intersection de ces deux segments. L'abscisse de ce point sera le mode $Mo =$ année. Une autre méthode graphique de calcul du mode consiste à prendre le centre de la classe modale $Mo = 12.5$ année.

Analytiquement, le mode est défini comme suit :

$$M_0 \;=\; 10 \;+\; \frac{(64-38)}{(64-38)+(64-40)}\,(15-10) \;=\; 12.6\; euros.$$

8. Tracer la courbe cumulative de la série statistique Y

Soit la colonne des fréquences cumulées croissantes de la variable Y :

Salaires	**Fréquence** f_i	**Fréquence cumulée** $\nearrow$ $f_i \nearrow$
[1000 ; 2000[	0.46	0.46
[2000 ; 3000[	0.42	0.88
[3000 ; 4000[	0.12	1.00
Total	1	

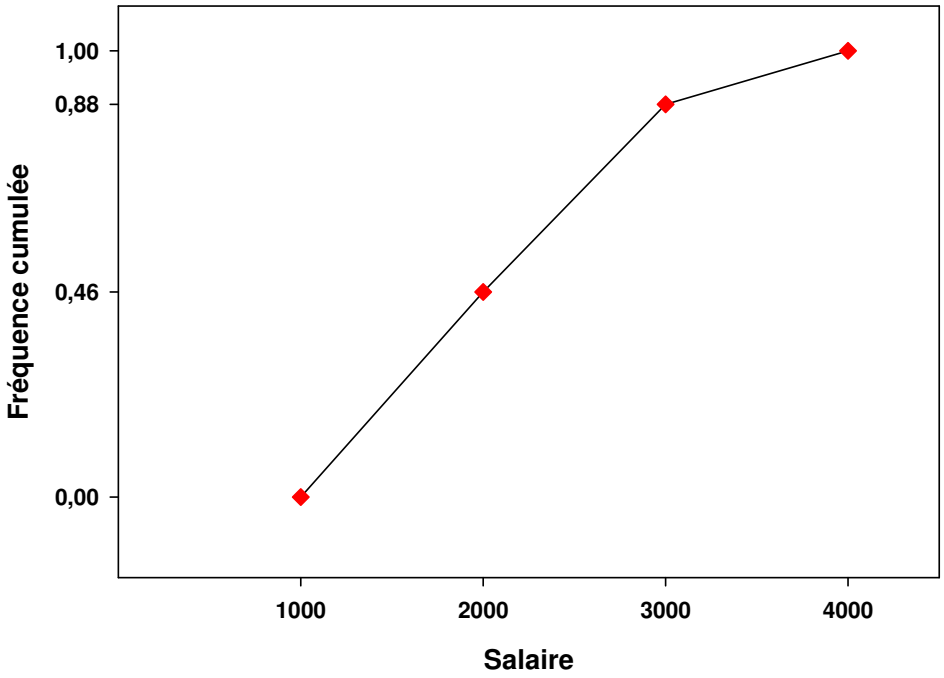

FIGURE 2.2 – Courbe cumulative des salaires

9. Déterminer graphiquement et analytiquement la médiane et interpréter le résultat

Graphiquement sur la courbe cumulative, la médiane est l'abscisse du point dont l'ordonnée est 0.5, d'où $Me =$ euros. Analytiquement d'après la colonne des fréquences cumulées, la classe médiane est $[2000; 3000]$ euros et la médiane est définie comme suit :

$$\frac{Me - 2000}{3000 - 2000} = \frac{0.50 - 0.46}{0.88 - 0.46}.$$

Soit $Me = 2095.23$ euros. On a donc parmi les 1000 salariés, autant de salariés dont le salaire est inférieur à 2095.23 euros, que de salariés dont le salaire est supérieur à cette valeur médiane.

10. Déterminer la droite de régression de Y sur X d'après la méthode de moindres carrés

D'après la méthode des moindres carrés ordinaire, on a l'estimation suivante :

$$Y = a\,x + b$$

avec $a = \dfrac{Cov(X,Y)}{\sigma(X)^2} = \dfrac{346}{464400} = 0.000745$ et $b = \overline{Y} - a\,\overline{X} = 11.9 - 0.000745\ 2160 = 10.2908$.
Par la suite $Y = 0.000745\ x + 10.2908$.

11. Calculer les moyennes, les variances et les écart-types conditionnelles de X et de Y

Calculons les moyennes, les variances et les écart-types conditionnelles de X :

◇ La distribution conditionnelle de X sachant $Y = [0;10]$ années est donnée par les deux premières colonnes du tableau ci-dessous :

$X/_{Y=[0;10]}$ X_1	**Centre de classe** c_i	**Nombre de salariés** n_{i1}	$n_{i1}\ .\ c_i$	$c_i - \overline{X}_1$	$(c_i - \overline{X}_1)^2$	$n_{i1}\ .\ (c_i - \overline{X}_1)^2$
[1000 ; 2000[	1500	220	330 000	- 526.315	277 007.479	60 941 645.429
[2000 ; 3000[	2500	120	300 000	473.685	224 377.479	26 925 297.507
[3000 ; 4000[	3500	40	140 000	1 473.685	2 171 747.479	86 869 899.169
Total		380	770 000			174 736 842.105

la moyenne conditionnelle de X sachant $Y = [0;10]$ années est donnée par la formule suivante :

$$\overline{X}_1 = \frac{1}{n_{.1}} \sum_{i=1}^{3} n_{i1}\ c_i = \frac{770000}{380} = 2026.315.$$

Par conséquent, le salaire moyen des salariés de cet entreprise ayant une expérience comprise entre 0 et 10 ans, est égal à 2026.315 euros.

L'écart-type est donné par la formule suivante :

$$V_1(X) = V(X/_{Y=[0;10]}) = \frac{1}{n_{.1}} \sum_{i}^{3} n_{i1}\ (c_i - \overline{X}_1)^2 = \frac{174736842.105}{380} = 459833.795,$$

d'où $\sigma_1(X) = 678.110$ euros.

◇ La distribution conditionnelle de X sachant $Y = [10;15]$ années est donnée par les deux premières colonnes du tableau ci-dessous :

$X/_{Y=[10;15]}$ X_2	**Centre de classe** c_i	**Nombre de salariés** n_{i2}	$n_{i2}\ .\ c_i$	$c_i - \overline{X}_2$	$(c_i - \overline{X}_2)^2$	$n_{i2}\ .\ (c_i - \overline{X}_2)^2$
[1000 ; 2000[	1500	100	150 000	- 812.5	660 156.25	66 015 625
[2000 ; 3000[	2500	180	450 000	187.5	35 156.25	6 328 125
[3000 ; 4000[	3500	40	140 000	1 187.5	1 410 156.25	56 406 250
Total		320	740 000			128 750 000

la moyenne conditionnelle de X sachant $Y = [10;15]$ années est donnée par la formule suivante :

$$\overline{X}_2 = \frac{1}{n_{.2}} \sum_{i=1}^{3} n_{i2}\ c_i = \frac{740000}{320} = 2312.5.$$

Par conséquent, le salaire moyen des salariés de cet entreprise ayant une expérience comprise entre 10 et 15 ans, est égal à 2026.315 euros.

L'écart-type est donné par la formule suivante :

$$V_2(X) = V(X/_{Y=[10;15]}) = \frac{1}{n_{.2}}\sum_i^3 n_{i2}\,(c_i - \overline{X}_2)^2 = \frac{128750000}{320} = 402343.75,$$

d'où $\sigma_2(X) = 634.305$ euros.

◇ La distribution conditionnelle de X sachant $Y = [15; 20]$ années est donnée par les deux premières colonnes du tableau ci-dessous :

$X/_{Y=[15;20]}$ X_3	**Centre de classe** c_i	**Nombre de salariés** n_{i3}	$n_{i3}\,.\,c_i$	$c_i - \overline{X}_3$	$(c_i - \overline{X}_3)^2$	$n_{i3}\,.\,(c_i - \overline{X}_3)^2$
[1000 ; 2000[	1500	90	135 000	- 700	490 000	44 100 000
[2000 ; 3000[	2500	80	200 000	300	90 000	7 200 000
[3000 ; 4000[	3500	30	105 000	1 300	1 690 000	50 700 000
Total		200	440 000			102 000 000

la moyenne conditionnelle de X sachant $Y = [15; 20]$ années est donnée par la formule suivante :

$$\overline{X}_3 \;=\; \frac{1}{n_{.3}}\sum_{i=1}^{3} n_{i3}\,c_i \;=\; \frac{440000}{200} \;=\; 2200.$$

Par conséquent, le salaire moyen des salariés de cet entreprise ayant une expérience comprise entre 15 et 20 ans, est égal à 2200 euros.

L'écart-type est donné par la formule suivante :

$$V_3(X) = V(X/_{Y=[15;20]}) = \frac{1}{n_{.3}}\sum_i^3 n_{i3}\,(c_i - \overline{X}_3)^2 = \frac{102000000}{200} = 510000,$$

d'où $\sigma_3(X) = 714.142$ euros.

◇ La distribution conditionnelle de X sachant $Y = [20; 30]$ années est donnée par les deux premières colonnes du tableau ci-dessous :

$X/_{Y=[20;30]}$ X_4	**Centre de classe** c_i	**Nombre de salariés** n_{i4}	$n_{i4}\,.\,c_i$	$c_i - \overline{X}_4$	$(c_i - \overline{X}_4)^2$	$n_{i4}\,.\,(c_i - \overline{X}_4)^2$
[1000 ; 2000[	1500	50	75 000	- 600	360 000	18 000 000
[2000 ; 3000[	2500	40	100 000	400	160 000	6 400 000
[3000 ; 4000[	3500	10	35 000	1 400	1 960 000	19 600 000
Total		100	210 000			44 000 000

la moyenne conditionnelle de X sachant $Y = [20; 30]$ années est donnée par la formule suivante :

$$\overline{X}_4 \;=\; \frac{1}{n_{.4}}\sum_{i=1}^{4} n_{i4}\,c_i \;=\; \frac{210000}{100} \;=\; 2100.$$

Par conséquent, le salaire moyen des salariés de cet entreprise ayant une expérience comprise entre 20 et 30 ans, est égal à 2100 euros.

L'écart-type est donné par la formule suivante :

$$V_4(X) = V(X/_{Y=[20;30]}) = \frac{1}{n_{.4}} \sum_i^4 n_{i3}\ (c_i - \overline{X}_4)^2 = \frac{44000000}{200} = 440000,$$

d'où $\sigma_4(X) = 663.324$ euros.

Calculons à présent les moyennes, les variances et les écart-types conditionnelles de Y :

◇ La distribution conditionnelle de Y sachant $X = [1000; 2000]$ euros est donnée par les deux premières colonnes du tableau ci-dessous :

$Y/_{X=[1000;2000]}$ Y_1	**Centre de classe** c'_j	**Nombre de salariés** n_{1j}	$n_{1j}\ .\ c'_j$	$c'_j - \overline{Y}_1$	$\left(c'_j - \overline{Y}_1\right)^2$	$n_{1j}\ .\ \left(c'_j - \overline{Y}_1\right)^2$
[00 ; 10[	5.0	220	1 100	- 6.25	39.06	8 593.75
[10 ; 15[	12.5	100	1 250	1.25	1.56	156.25
[15 ; 20[	17.5	90	1 575	6.25	39.06	3 515.62
[20 ; 30[	25.0	50	1 250	13.75	189.6	9 453.12
Total		460	5 175			21 718.75

la moyenne conditionnelle de Y sachant $X = [1000; 2000]$ euros est donnée par la formule suivante :

$$\overline{Y}_1 \ = \ \frac{1}{n_{1.}} \sum_{i=1}^{4} n_{1j}\ c'_j \ = \ \frac{5175}{460} \ = \ 11.25.$$

Par conséquent, le nombre des années d'expérience en moyenne des salariés de cet entreprise ayant un salaire compris entre 1000 et 2000 euros, est égal à 11.25 années.

L'écart-type est donné par la formule suivante :

$$V_1(Y) = V(Y/_{X=[1000;2000]}) = \frac{1}{n_{1.}} \sum_i^3 n_{1j}\ (c'_j - \overline{Y}_1)^2 = \frac{21718.75}{460} = 47.21,$$

d'où $\sigma_1(Y) = 6.871$ années.

◇ La distribution conditionnelle de Y sachant $X = [2000; 3000]$ euros est donnée par les deux premières colonnes du tableau ci-dessous :

$Y/_{X=[2000;3000]}$ Y_2	**Centre de classe** c'_j	**Nombre de salariés** n_{2j}	$n_{2j}\ .\ c'_j$	$c'_j - \overline{Y}_2$	$\left(c'_j - \overline{Y}_2\right)^2$	$n_{2j}\ .\ \left(c'_j - \overline{Y}_2\right)^2$
[00 ; 10[	5.0	120	600	- 7.5	56.25	6 750
[10 ; 15[	12.5	180	2 250	0.00	0	0
[15 ; 20[	17.5	80	1 400	5	25.00	2 000
[20 ; 30[	25.0	40	1 000	12.5	156.25	6 250
Total		420	5 250			15 000

la moyenne conditionnelle de Y sachant $X = [2000; 3000]$ euros est donnée par la formule suivante :

$$\overline{Y}_2 \;=\; \frac{1}{n_{2.}}\sum_{i=1}^{4} n_{2j}\; c'_j \;=\; \frac{5250}{420} \;=\; 12.5.$$

Par conséquent, le nombre des années d'expérience en moyenne des salariés de cet entreprise ayant un salaire compris entre 2000 et 3000 euros, s'établit à 12.5 années.

L'écart-type est donné par la formule suivante :

$$V_2(Y) = V(Y/_{X=[2000;3000]}) = \frac{1}{n_{2.}}\sum\nolimits_i^3 n_{2j}\;(c'_j - \overline{Y}_2)^2 = \frac{15000}{420} = 35.71,$$

d'où $\sigma_2(Y) = 5.976$ années.

◇ La distribution conditionnelle de Y sachant $X = [2000; 3000]$ euros est donnée par les deux premières colonnes du tableau ci-dessous :

$Y/_{X=[3000;4000]}$ Y_3	**Centre de classe** c'_j	**Nombre de salariés** n_{3j}	$n_{3j}\ .\ c'_j$	$c'_j - \overline{Y}_3$	$\left(c'_j - \overline{Y}_3\right)^2$	$n_{3j}\ .\ \left(c'_j - \overline{Y}_3\right)^2$
[00 ; 10[	5.0	40	200	- 7.29	53.144	2 125.764
[10 ; 15[	12.5	40	500	0.21	0.044	1.764
[15 ; 20[	17.5	30	525	5.21	27.144	814.323
[20 ; 30[	25.0	10	250	12.71	161.544	1 615.441
Total		120	1 475			4 557.292

la moyenne conditionnelle de Y sachant $X = [3000; 4000]$ euros est donnée par la formule suivante :

$$\overline{Y}_3 \;=\; \frac{1}{n_{3.}}\sum_{i=1}^{4} n_{3j}\; c'_j \;=\; \frac{1475}{120} \;=\; 12.29.$$

Par conséquent, le nombre des années d'expérience en moyenne des salariés de cet entreprise ayant un salaire compris entre 3000 et 4000 euros, s'établit à 12.29 années.

L'écart-type est donné par la formule suivante :

$$V_3(Y) = V(Y/_{X=[3000;4000]}) = \frac{1}{n_{3.}}\sum\nolimits_i^3 n_{3j}\;(c'_j - \overline{Y}_3)^2 = \frac{4557.292}{120} = 37.97,$$

d'où $\sigma_3(Y) = 6.162$ années.

12. Déterminer les distributions des fréquences conditionnelles

2.2 Exercice 2

On a examiné 60 salariés d'une entreprise du point de vue leurs salaires en 10^3 Dinars Algériens et leurs nombres d'enfants. Les résultats obtenus sont présentés dans le tableau de contingence suivant :

Nombre / Salaires d'enfant	[2,6[	[6,12[	[12,16[	Colonnes marginales
1	9	14	2	25
2	5	12	1	18
3	4	10	3	17
lignes marginales	18	36	6	60

1. Vérifier que $\overline{X} = \frac{1}{N} \sum_{j=1}^{3} n_{.j} \, \overline{X}_j$
2. Vérifier que $\overline{Y} = \frac{1}{N} \sum_{i=1}^{3} n_{i.} \, \overline{Y}_i$

Solution :

1. Vérifier que $\overline{X} = \frac{1}{N} \sum_{j=1}^{3} n_{.j} \, \overline{X}_j$

Calculons la moyenne et la variance de X. Soit la distribution marginale de X :

Nombre d'enfants x_i	**Nombre de salariés** $n_{i.}$	$n_{i.} \, . \, x_i$	$x_i - \overline{X}$	$\left(x_i - \overline{X}\right)^2$	$n_{i.} \, . \, \left(x_i - \overline{X}\right)^2$
1	25	25	-0.866	0.749	18.748
2	18	36	0.134	0.017	0.323
3	17	51	1.134	1.285	21.861
Total	60	112			40.932

La moyenne de la variable statistique "Nombre d'enfant" est donnée par la formule suivante :

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{3} n_{i.} \, x_i = \frac{112}{60} = 1.8666 \, enfants.$$

La variance est donné par la formule suivante :

$$V(X) = \frac{1}{N} \sum_{i}^{3} n_{i.} \, (x_i - \overline{X})^2 = \frac{40.932}{60} = 0.6822 \, enfants^2.$$

Calculons les moyennes et les variances conditionnelles de X :

◇ La distribution conditionnelle de X sachant $Y = [2;6].10^3$ DA est donnée par les deux premières colonnes du tableau ci-dessous :

$X/_{Y=[2;6]}$ X_1	**Nombre de salariés** n_{i1}	$n_{i1} \, . \, x_i$	$x_i - \overline{X}_1$	$\left(x_i - \overline{X}_1\right)^2$	$n_{i1} \, . \, \left(x_i - \overline{X}_1\right)^2$
1	9	9	-0.722	0.521	4.691
2	5	10	0.278	0.077	0.386
3	4	12	1.278	1.633	6.533
Total	18	31			11.61

la moyenne conditionnelle de X sachant $Y = [2;6].10^3$ DA est donnée par la formule suivante :

$$\overline{X}_1 \;=\; \frac{1}{n_{.1}}\sum_{i=1}^{3} n_{i1}\, x_i \;=\; \frac{31}{18} \;=\; 1.722\; enfants.$$

et

$$V_1(X) = V(X/_{Y=[2;6]}) = \frac{1}{n_{.1}}\sum\nolimits_i^3 n_{i1}\,\left(x_i - \overline{X}_1\right)^2 = \frac{11.61}{18} = 0.645\; enfant^2.$$

◇ La distribution conditionnelle de X sachant $Y = [6;12].10^3$ DA est donnée par les deux premières colonnes du tableau ci-dessous :

$X/_{Y=[6;12]}$ X_2	**Nombre de salariés** n_{i2}	$n_{i2}\,.\,x_i$	$x_i - \overline{X}_2$	$\left(x_i - \overline{X}_2\right)^2$	$n_{i2}\,.\,\left(x_i - \overline{X}_2\right)^2$
1	14	14	0.888	0.788	11.039
2	12	24	0.112	0.012	0.144
3	10	30	1.112	1.236	12.365
Total	36	68			23.548

la moyenne conditionnelle de X sachant $Y = [6;12].10^3$ DA est donnée par la formule suivante :

$$\overline{X}_2 \;=\; \frac{1}{n_{.2}}\sum_{i=1}^{3} n_{i2}\, x_i \;=\; \frac{68}{36} \;=\; 1.888\; enfants.$$

et

$$V_2(X) = V(X/_{Y=[6;12]}) = \frac{1}{n_{.2}}\sum\nolimits_i^3 n_{i2}\,\left(x_i - \overline{X}_2\right)^2 = \frac{23.548}{36} = 0.654\; enfant^2.$$

◇ La distribution conditionnelle de X sachant $Y = [12;16].10^3$ DA est donnée par les deux premières colonnes du tableau ci-dessous :

$X/_{Y=[12;16]}$ X_3	**Nombre de salariés** n_{i3}	$n_{i3}\,.\,x_i$	$x_i - \overline{X}_3$	$\left(x_i - \overline{X}_3\right)^2$	$n_{i3}\,.\,\left(x_i - \overline{X}_3\right)^2$
1	2	2	-1.166	1.359	2.719
2	1	2	-0.166	0.027	0.027
3	3	9	0.834	03695	2.086
Total	6	13			4.832

la moyenne conditionnelle de X sachant $Y = [12;16].10^3$ DA est donnée par la formule suivante :

$$\overline{X}_3 \;=\; \frac{1}{n_{.3}}\sum_{i=1}^{3} n_{i3}\, x_i \;=\; \frac{13}{6} \;=\; 2.166\; enfants.$$

et

$$V_3(X) = V(X/_{Y=[6;12]}) = \frac{1}{n_{.3}}\sum\nolimits_i^3 n_{i3}\,\left(x_i - \overline{X}_3\right)^2 = \frac{4.832}{6} = 0.805\; enfant^2.$$

Or

$$\frac{1}{N}\sum_{j=1}^{3} n_{.j}\,\overline{X}_j \;=\; n_{.1}\,\overline{X}_1 + n_{.2}\,\overline{X}_2 + n_{.3}\,\overline{X}_3 = \frac{18}{60}\times 1.722 + \frac{36}{60}\times 1.888 + \frac{6}{60}\times 2.166 \;=\; 1.866.$$

et

$$\begin{aligned}
&\frac{1}{N}\Big(n_{.1}\,V_1(X)\;+\;n_{.2}\,V_2(X)\;+\;n_{.3}\,V_3(X)\;+\;n_{.1}\,\big(\overline{X}_1-\overline{X}\big)^2\;+\;n_{.2}\,\big(\overline{X}_2-\overline{X}\big)^2\;+\;n_{.3}\,\big(\overline{X}_3-\overline{X}\big)^2\Big)\\
&\quad=\;\frac{1}{60}\Big(18\times 0.645\;+\;36\times 0.654\;+\;186\times 0.805\\
&\qquad+\,18\times(1.722-1.866)^2\;+\;36\times(1.888-1.866)^2\;+\;6\times(2.166-1.866)^2\Big)\\
&\qquad=\;0.6818.
\end{aligned}$$

Par suite, les formules de moyenne et de variance sont vérifiées.

2. Vérifier que $\overline{Y}\;=\;\dfrac{1}{N}\;\sum_{i=1}^{3} n_{i.}\;\overline{Y}_i$

Calculons la moyenne et la variance de Y. Soit la distribution marginale de Y :

Salaires	**Centre de classe** c'_j	**Nombre de salariés** $n_{.j}$	$n_{.j}\;.\;c'_j$	$c'_j-\overline{Y}$	$\big(c'_j-\overline{Y}\big)^2$	$n_{.j}\;.\;\big(c'_j-\overline{Y}\big)^2$
[02 ; 06[	4	18	72	4	16	288
[06 ; 12[	9	36	324	1	1	36
[12 ; 16[	14	6	84	6	36	216
Total		60	480			540

La moyenne de la variable statistique "Nombre d'enfant" est donnée par la formule suivante :

$$\overline{Y}\;=\;\frac{1}{N}\sum_{j=1}^{3} n_{.j}\;c'_j\;=\;\frac{480}{60}\;=\;8\;euros.$$

La variance est donné par la formule suivante :

$$V(Y)=\frac{1}{N}\sum_{j}^{3} n_{.j}\;(c'_j-\overline{Y})^2=\frac{540}{60}=9\;euros^2.$$

Calculons les moyennes et les variances conditionnelles de Y :

◇ La distribution conditionnelle de Y sachant $X=1$ est donnée par les deux premières colonnes du tableau ci-dessous :

$Y/_{X=1}$ Y_1	**Centre de classe** c'_j	**Nombre de salariés** n_{1j}	$n_{1j}\;.\;c'_j$	$c'_j-\overline{Y}_1$	$\Big(c'_j-\overline{Y}_1\Big)^2$	$n_{1j}\;.\;\Big(c'_j-\overline{Y}_1\Big)^2$
[02 ; 06[	4	9	36	-3.6	12.96	116.64
[06 ; 12[	9	14	126	1.4	1.96	27.44
[12 ; 16[	14	25	190	6.4	40.96	81.92
Total	25	190				226

la moyenne conditionnelle de Y sachant $X=1$ est donnée par la formule suivante :

$$\overline{Y}_1\;=\;\frac{1}{n_{1.}}\sum_{j=1}^{3} n_{1j}\;c'_j\;=\;\frac{190}{25}\;=\;7.6\;euros.$$

et

$$V_1(Y) = V(Y/_{X=1}) = \frac{1}{n_{1.}}\sum_j^3 n_{1j}\,\left(c'_j - \overline{Y}_1\right)^2 = \frac{226}{25} = 9.04\ euros^2.$$

◇ La distribution conditionnelle de Y sachant $X = 2$ est donnée par les deux premières colonnes du tableau ci-dessous :

$Y/_{X=2}$ Y_2	**Centre de classe** c'_j	**Nombre de salariés** n_{2j}	$n_{2j}\ .\ c'_j$	$c'_j - \overline{Y}_2$	$\left(c'_j - \overline{Y}_2\right)^2$	$n_{2j}\ .\ \left(c'_j - \overline{Y}_2\right)^2$
[02 ; 06[	4	5	20	3.88	15.12	75.61
[06 ; 12[	9	12	108	1.11	1.34	14.81
[12 ; 16[	14	1	14	6.11	37.34	37.34
Total	18	142				127.77

la moyenne conditionnelle de Y sachant $X = 2$ est donnée par la formule suivante :

$$\overline{Y}_2 \ = \ \frac{1}{n_{2.}}\sum_{j=1}^3 n_{2j}\ c'_j \ = \ \frac{142}{18} \ = \ 7.88\ euros.$$

et

$$V_2(Y) = V(Y/_{X=2}) = \frac{1}{n_{2.}}\sum_j^3 n_{2j}\,\left(c'_j - \overline{Y}_2\right)^2 = \frac{127.77}{18} = 7.09\ euros^2.$$

◇ La distribution conditionnelle de Y sachant $X = 3$ est donnée par les deux premières colonnes du tableau ci-dessous :

$Y/_{X=3}$ Y_3	**Centre de classe** c'_j	**Nombre de salariés** n_{3j}	$n_{3j}\ .\ c'_j$	$c'_j - \overline{Y}_3$	$\left(c'_j - \overline{Y}_3\right)^2$	$n_{3j}\ .\ \left(c'_j - \overline{Y}_3\right)^2$
[02 ; 06[	4	4	16	-4.70	22.13	88.54
[06 ; 12[	9	10	90	-0.29	0.08	0.87
[12 ; 16[	14	3	42	5.29	28.03	84.11
Total	17	148				173.52

la moyenne conditionnelle de Y sachant $X = 3$ est donnée par la formule suivante :

$$\overline{Y}_3 \ = \ \frac{1}{n_{3.}}\sum_{j=1}^3 n_{3j}\ c'_j \ = \ \frac{148}{17} \ = \ 8.70\ euros.$$

et

$$V_3(Y) = V(Y/_{X=3}) = \frac{1}{n_{3.}}\sum_j^3 n_{3j}\,\left(c'_j - \overline{Y}_3\right)^2 = \frac{173.52}{17} = 10.20\ euros^2.$$

Or

$$\frac{1}{N}\ \sum_{j=1}^3 n_{i.}\ \overline{Y}_i \ = \ n_{1.}\ \overline{Y}_1 + n_{2.}\ \overline{Y}_2 + n_{3.}\ \overline{Y}_3 = \frac{25}{60} \times 7.6 + \frac{18}{60} \times 7.88 + \frac{17}{60} \times 8.70 \ = \ 7.99 \ \approx \ 8.$$

et

$$\begin{aligned}
&\frac{1}{N}\Big(n_{1.}\, V_1(Y) \;+\; n_{2.}\, V_2(Y) \;+\; n_{3.}\, V_3(Y) \;+\; n_{1.}\,\big(\overline{Y}_1 - \overline{Y}\big)^2 \;+\; n_{2.}\,\big(\overline{Y}_2 - \overline{Y}\big)^2 \;+\; n_{3.}\,\big(\overline{Y}_3 - \overline{Y}\big)^2\Big) \\
&\quad = \frac{1}{60}\Big(25 \times 9.04 \;+\; 18 \times 7.09 \;+\; 17 \times 10.20 \\
&\qquad + 25 \times (7.6 - 8)^2 \;+\; 18 \times (7.88 - 8)^2 \;+\; 17 \times (8.70 - 8)^2\Big) \\
&\qquad\quad = 9.002 \;\approx\; 9.
\end{aligned}$$

Par suite, les formules de moyenne et de variance sont vérifiées.