Cours de Statistique Descriptive

Antoine Ayache & Julien Hamonier

1 Un peu d'histoire

L'objectif de la Statistique Descriptive est de décrire de façon synthétique et parlante des données observées pour mieux les analyser. Le terme « statistique » est issu du latin « statisticum », c'est-à-dire qui a trait à l'État. Ce terme a été utilisé, semble-t-il pour la première fois, à l'époque de Colbert, par Claude Bouchu, intendant de Bourgogne, dans une « Déclaration des biens, charges, dettes et statistiques des communautés de la généralité de Bourgogne de 1666 à 1669 ».

Par contre, l'apparition du besoin « statistique » de posséder des données chiffrées et précises, précède sa dénomination de plusieurs millénaires. À son origine, il est le fait de chefs d'États (ou de ce qui en tient lieu à l'époque) désireux de connaître des éléments de leur puissance : population, potentiel militaire, richesse, ...

2 Analyse descriptive univariée

2.1 Vocabulaire

1. On appelle **population** un ensemble d'éléments homogènes auxquels on s'intéresse. Par exemple, les étudiants d'une classe, les contribuables français, les ménages lillois ...
2. Les éléments de la population sont appelés **les individus** ou **unités statistiques**.
3. **Des observations** concernant un thème particulier ont été effectuées sur ces individus. La série de ces observations forme ce que l'on appelle **une variable statistique**. Par exemple, les Notes des Etudiants à l'Examen de Statistique, les Mentions qu'ils ont obtenues à leur Bac, leur Sexe, les Couleurs de leurs Yeux, le Chiffre d'Affaire par PME, le Nombre d'Enfants par Ménage, ...
4. Une variable statistique est dite :
   (i) **quantitative** : lorsqu'elle est mesurée par un nombre (les Notes des Etudiants à l'Examen de Statistique, le Chiffre d'Affaire par PME, le Nombre d'Enfants par Ménage, ...). On distingue 2 types de variables quantitatives : les variables quantitatives **discrètes** et les variables quantitatives **continues**. Les variables discrètes (ou discontinues) ne prennent que des valeurs isolées. Par exemple le nombre d'enfants par ménage ne peut être que 0, ou 1, ou 2, ou 3, ... ; il ne peut jamais prendre une valeur strictement comprise entre 0 et 1, ou 1 et 2, ou 2 et 3, ... C'est aussi le cas de la note à l'examen de statistique (on suppose que les notations sont entières sans possibilités de valeurs décimales intermédiaires). Les variables quantitatives continues peuvent prendre toute valeur dans un intervalle. Par exemple, le chiffre d'affaire par PME peut être 29000,1€, 29000,12€, ..., même si dans la pratique il faut l'arrondir.
   (ii) **qualitative** : lorsque les modalités (ou les valeurs) qu'elle prend sont désignées par des noms. Par exemples, les modalités de la variable Sexe sont : Masculin et Féminin ;

les modalités de la variable Couleur des Yeux sont : Bleu, Marron, Noir et Vert ; les modalités de la variable Mention au Bac sont : TB, B, AB et P. On distingue deux types de variables qualitatives : les variables qualitatives **ordinales** et les variables qualitatives **nominales**. Plus précisément une variable qualitative est dite ordinale, lorsque ses modalités peuvent être classées dans un certain ordre naturel (c'est par exemple le cas de la variable Mention au Bac) ; une variable qualitative est dite nominale, lorsque ses modalités ne peuvent être classées de façon naturelle (c'est par exemple le cas de la variable Couleur des Yeux ou encore de la variable Sexe).

2.2 Représentation graphique d'une variable

Pour un groupe de 15 étudiants, on a observé les valeurs des variables : Couleur des Yeux, Sexe, Mention au Bac et Note à l'Examen de Statistique ; ainsi le tableau de données suivant a été obtenu. Ces données seront souvent utilisées dans ce chapitre.

Tableau de Données

Individu	Couleur des Yeux	Sexe	Mention au Bac	Note à l'Examen de Statistique
Michel	V	H	P	12
Jean	B	H	AB	8
Stéphane	N	H	P	13
Charles	M	H	P	11
Agnès	B	F	AB	10
Nadine	V	F	P	9
Étienne	N	H	B	16
Gilles	M	H	AB	14
Aurélie	B	F	P	11
Stéphanie	V	F	B	15
Marie-Claude	N	F	P	4
Anne	B	F	TB	18
Christophe	V	H	AB	12
Pierre	N	H	P	6
Bernadette	M	F	P	2

2.2.1 Variables qualitatives (ordinales et nominales)

On représente les variables Couleurs des Yeux, Sexe et Mention au Bac par **des diagrammes en bâtons**. On notera que chacun des individus appartient à une seule modalité de chacune de ces 3 variables. En effet, on ne peut avoir des individus dont les yeux possèdent plusieurs couleurs (on exclut les cas d'hétérochromie). On ne peut pas avoir non plus un individu qui soit à la fois Homme et Femme (on exclut les cas d'hermaphrodisme). Enfin, un même individu ne peut obtenir plusieurs mentions au Bac.

Remarque 2.1. *De façon générale, un individu appartient à une et une seule modalité d'une variable qualitative. Bien souvent, parmi les modalités d'une variable qualitative figure une modalité* ***Autres*** *(non répondants ou bien valeurs manquantes ou quelque chose dans ce genre-là) dans laquelle on place les individus qu'on n'arrive pas à caser dans une autre modalité de cette variable.*

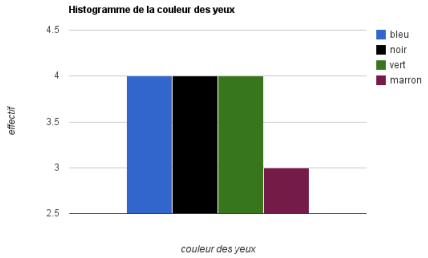
Étudions l'exemple de la variable **Couleurs des Yeux**. On commence d'abord par compter le nombre d'individus appartenant à chacune des modalités de cette variables : $n_B = 4$ individus

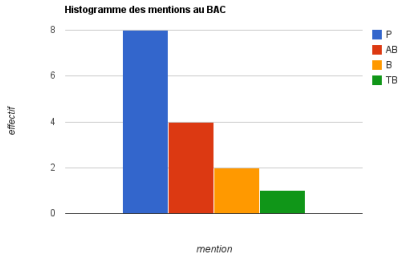
ont les yeux bleus, $n_M = 3$ ont les yeux marrons, $n_N = 4$ ont les yeux noirs et $n_V = 4$ ont les yeux verts ; on peut résumer tout cela dans le tableau récapitulatif suivant :

Couleur	Bleu	Marron	Noir	Vert
Effectif	4	3	4	4

Faisons de même avec la variable **Mention au Bac** ; on obtient le tableau récapitulatif suivant :

mention	P	AB	B	TB
effectif	8	4	2	1





On constate que les étudiants sont répartis inégalement entre les différentes modalités de la variable Mention au Bac. Une première façon d'apprécier la répartition d'une variable est de construire **un tableau de répartition des effectifs et des fréquences** entre les différentes valeurs possibles de la variable. De façon générale, la fréquence d'une modalité « M » d'une variable qualitative se calcule au moyen de la formule suivante :

$$f_M = (\text{ fréquence de la modalité « M » d'une variable qualitative}) = \frac{(\text{effectif correspondant à « M »})}{(\text{effectif total})}.$$

On a de plus,

$$p_M = (\text{pourcentage des individus correspondant à la modalité « M »}) = f_M \times 100.$$

On a enfin

$$(\text{somme des fréquences de toutes les modalités d'une variable qualitative}) = 1$$

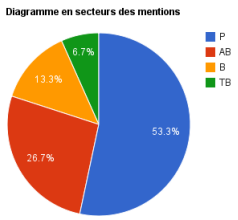
$$(\text{somme de tous les pourcentages correspondant aux modalités d'une variable qualitative}) = 100.$$

Tableau de Répartition de la variable Mention au Bac

Mention au Bac	Effectifs	Fréquences	Pourcentages
P	$n_P = 8$	$f_P = 8/15 = 0.533$	53.3%
AB	$n_{AB} = 4$	$f_{AB} = 4/15 = 0.267$	26.7%
B	$n_B = 2$	$f_B = 2/15 = 0.133$	13.3%
TB	$n_{TB} = 1$	$f_{TB} = 1/15 = 0.067$	6.7%
	effectif total $N = 15$	$f_P + f_{AB} + f_B + f_{TB} = 1$	Total = 100%

Notons que dans ce tableau les pourcentages sont donnés au dixième près, c'est-à-dire avec un chiffre après la virgule.

Avant de finir cette sous-section, signalons que la répartition des fréquences (ou pourcentages) entre les différentes modalités d'une variable qualitative, peut non seulement être représentée au moyen d'un diagramme en bâtons, mais aussi à l'aide d'un **diagramme en secteurs**. Dans le cas de la variable Mention au Bac, on obtient :



2.2.2 Variable quantitative discrète

De façon générale à chaque valeur k d'une variable quantitative discrète correspond un effectif, noté par n_k ; il s'agit en fait du nombre des individus pour lesquels on a observé la valeur k. La fréquence f_k de la valeur k, se calcule au moyen de la formule :

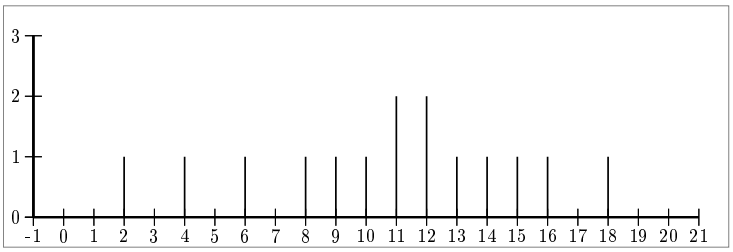
$$f_k = \frac{n_k}{N},$$

où n_k désigne l'effectif correspondant à la valeur k et N l'effectif total ; tout comme dans le cas des variables qualitatives, en multipliant les fréquences par 100, on obtient les pourcentages correspondants.

Tableau de Répartition de la variable
Note à l'Examen de Statistique

Note à l'Examen de Statistique	Effectifs	Fréquences
k=0	0	0
k=1	0	0
k=2	1	1/15
k=3	0	0
k=4	1	1/15
k=5	0	0
k=6	1	1/15
k=7	0	0
k=8	1	1/15
k=9	1	1/15
k=10	1	1/15
k=11	2	2/15
k=12	2	2/15
k=13	1	1/15
k=14	1	1/15
k=15	1	1/15
k=16	1	1/15
k=17	0	0
k=18	1	1/15
k=19	0	0
k=20	0	0

De façon générale, Pour représenter le tableau ci-dessus, on pourrait utiliser un diagramme en bâtons :

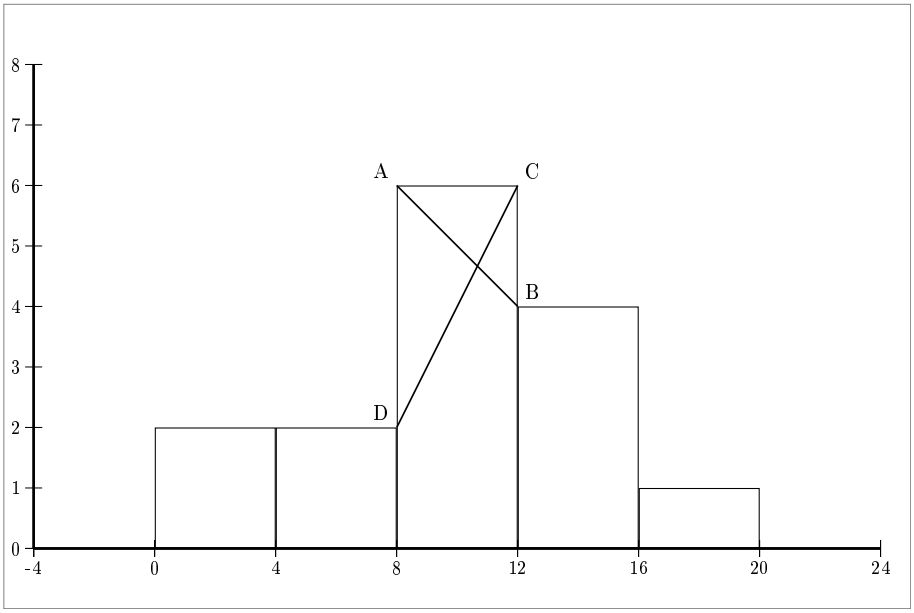


Néanmoins cette forme se prête difficilement à l'interprétation. Pour y remédier, il faut créer des **classes** de notes (nombre d'individus ayant obtenu des notes comprises entre 0 et 4, entre 4 et 8, ...) ; cette approche nous permet d'obtenir une variable dite **classée**. Il faut effectuer le **bornage** des classes en excluant et incluant les valeurs en début et fin de classe.

Tableau de Répartition de la variable classée Note à l'Examen de Statistique

variable classée	Effectifs	Fréquences
$[0, 4]$	2	2/15
$]4, 8]$	2	2/15
$]8, 12]$	6	6/15
$]12, 16]$	4	4/15
$]16, 20]$	1	1/15

Histogramme des Effectifsde la variable classée Note à l'Examen de Statistique



La représentation graphique des effectifs de chaque classe s'appelle **l'histogramme des effectifs** ; on peut de la même façon réaliser **l'histogramme des fréquences**.

En créant des classes, *on agglomère* des informations ; on perd de l'information mais en contrepartie, on fait ressortir la structure de *la distribution statistique*. Pour une série d'observations relatives à une variable quantitative X, discrète, discrète classée ou continue classée, la donnée des classes (ou encore des valeurs) et de leurs fréquences (ou encore de leur effectif) est appelée *distribution statistique de la variable X*.

Dans le cas de la variable Note à l'Examen de Statistique, on voit que la majeure partie de l'effectif se situe autour de la moyenne ; une telle distribution est appelée *loi normale*. On retrouve souvent la loi normale en statistique ; sa forme caractéristique est celle d'une « cloche ».

2.2.3 Variable quantitative continue

L'infinité des valeurs observables d'une variable quantitative continue ne rend pas possible la généralisation du diagramme en bâtons. L'établissement d'un tableau de répartition exige que l'on

découpe l'intervalle de variation d'une telle variable, en k sous-intervalles $[x_0, x_1],]x_1, x_2], \ldots,]x_{k-1}, x_k]$. Chacun de ces intervalles est appelé **classe** ; l'idée étant que chaque classe forme **une entité homogène** qui se distingue des autres classes. Le nombre de classes k doit être modéré (une dizaine au maximum). L'amplitude de la classe $[x_0, x_1]$, c'est-à-dire sa « largeur », est égale à $a_1 = x_1 - x_0$, de même pour tout $i = 2, \ldots, k$ l'amplitude de la classe $]x_{i-1}, x_i]$ est égale à $a_i = x_i - x_{i-1}$. Lorsque la dernière classe est définie par « plus de ... » son amplitude est alors indéterminée.

L'histogramme des fréquences d'une telle variable est constitué de la juxtaposition de rectangles dont les bases représentent les différentes classes, et dont **les surfaces** sont proportionnelles aux fréquences des classes et par conséquent à leurs effectifs. Ainsi, à la i-ème classe correspond un rectangle dont la base est l'intervalle $]x_{i-1}, x_i]$ (dans le cas particulier $i = 1$, la base est l'intervalle $[x_0, x_1]$), et dont la surface est proportionnelle à la fréquence f_i et à l'effectif n_i.

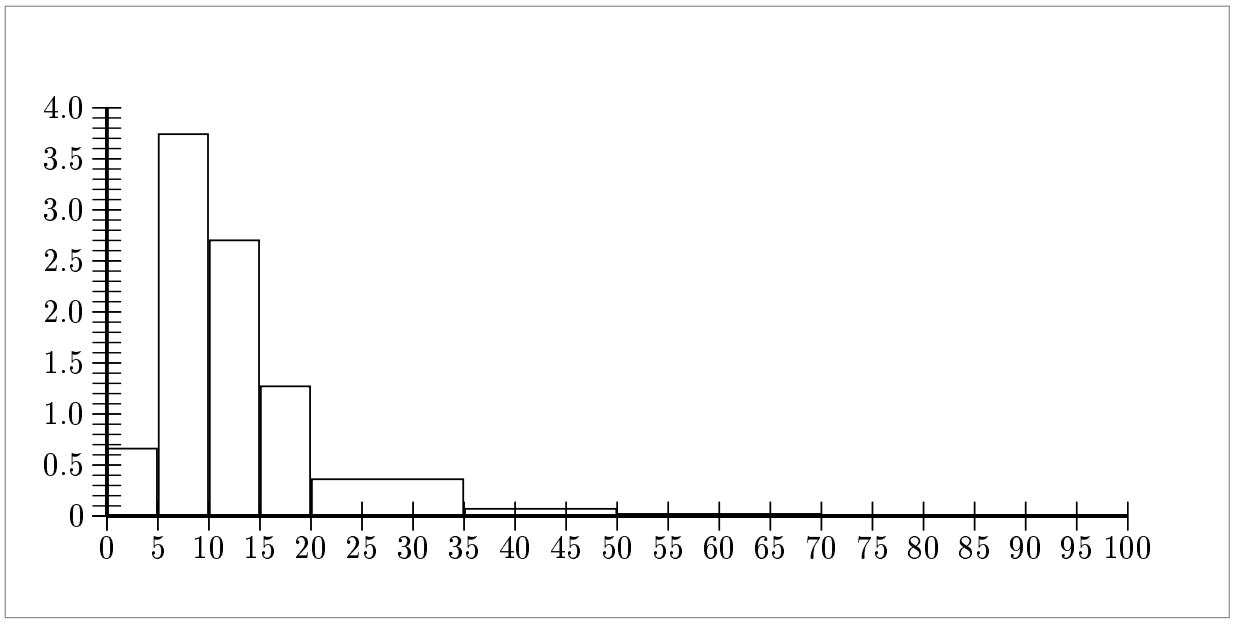
Lorsque les classes ont toutes, la même amplitude, les hauteurs des rectangles sont proportionnelles à leurs surfaces ; par conséquent les hauteurs des rectangles sont proportionnelles aux fréquences et aux effectifs. Dans le cas où les classes sont d'amplitudes inégales, la hauteur du rectangle correspondant à la i-ème classe sera $h_i = f_i/a_i$ (c'est-à-dire la fréquence par unité d'amplitude) ou encore $H_i = n_i/a_i$ (c'est-à-dire l'effectif par unité d'amplitude).

Etudions maintenant un exemple concret :

Tableau de Répartition de la variable quantitative continue
« Revenus des Contribuables soumis à l'impôt sur le revenu en 1965 » (source DGI)

Classe de revenus en Francs	Effectif en milliers d'individus	Fréquence	Amplitude en Francs	Hauteur $\times$ 50000 $= \frac{\text{Fréquence}}{\text{Amplitude}} \times 50000$
$[0, 5000]$	549,3	$6,67.10^{-2}$	5000	0,67
$]5000, 10000]$	3087,4	$37,51.10^{-2}$	5000	3,75
$]10000, 15000]$	2229,0	$27,08.10^{-2}$	5000	2,71
$]15000, 20000]$	1056,7	$12,84.10^{-2}$	5000	1,28
$]20000, 35000]$	925,0	$11,24.10^{-2}$	15000	0,37
$]35000, 50000]$	211,0	$2,56.10^{-2}$	15000	0,09
$]50000, 70000]$	90,8	$1,1.10^{-2}$	20000	0,03
$]70000, 100000]$	81,6	$0,99.10^{-2}$	30000	0,02
	Effectif total $= 8230,8$			

Histogramme des Fréquences de la variable
« Revenus des Contribuables »
(L'échelle sur l'axe des abscisses est 1 millier de Francs
et l'échelle sur l'axe des ordonnées est 1/50000)



2.3 Valeurs centrales

2.3.1 Le mode

a) Variable quantitative discrète (non classée)

Le **mode** correspond à la valeur de la variable pour laquelle l'effectif (ou la fréquence) est le plus grand.

Exemple 2.1. *Recensement des familles dans une population régionale dont le nombre d'enfants de moins de 14 ans est le suivant :*

Nombre d'enfants	*Nombre de familles*
0	*2601*
1	*6290*
2	*2521*
3	*849*
4	*137*
	Total = 12398

Ici le mode correspond à la valeur de 1 enfant.

Remarque 2.2. *Certaines variables peuvent présenter plusieurs modes. Par exemple, dans le cas de la variable « note à l'examen » l'effectif maximum correspond aux valeurs 11 et 12 de la variable ; étant donné que ces deux valeurs se suivent, on dit qu'il y a un intervalle modal.*

b) Variable quantitative continue ou discrète classée

La **classe modale** est la classe dont la fréquence par unité d'amplitude est la plus élevée ; cette classe correspond donc au rectangle le plus haut de l'histogramme des fréquences. Par exemple, dans le cas de la variable « Revenu des Contribuables » $]5000, 10000]$ est la classe modale. Signalons au passage que certaines variables peuvent avoir plusieurs classes modales.

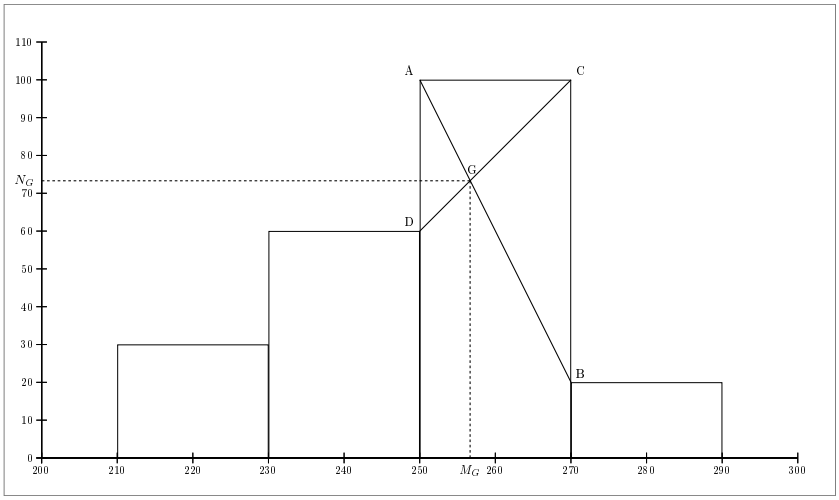
Lorsqu'on souhaite être plus précis, on peut déterminer à l'intérieur de la classe modale **la valeur exacte du mode** ; l'exemple suivant permet de comprendre la démarche à suivre.

Exemple 2.2. *On désire lancer un nouveau produit sur le marché ; on recherche le prix psychologique nous permettant d'attirer le plus de consommateurs possible. La détermination du mode peut, entre autre méthode, nous permettre d'approcher au mieux le prix psychologique de lancement du produit. Présentant le produit à un échantillon représentatif de la population étudiée, nous observons pour chaque classe de prix, les effectifs prêts à faire l'acquisition du produit. Nous obtenons les résultats suivants :*

Prix (en Euros)	*Effectifs*
$[210, 230]$	*30*
$]230, 250]$	*60*
$]250, 270]$	*100*
$]270, 290]$	*20*
	Total = 210

Les classes de prix étant toutes de même amplitude (égale à 20), les hauteurs des rectangles de l'histogramme des effectifs seront donc égales aux effectifs.

Histogramme des effectifs



La classe modale est $]250, 270]$. *La projection du point d'intersection G des segments [AB] et [CD] sur l'axe Prix correspond à la valeur exacte du mode,* $M_G \simeq 257$ *Euros. Si on souhaite davantage de précisons, on peut calculer* (M_G, N_G) *les coordonnées de G. Pour ce faire il faut d'abord trouver les équations des droites (AB) et (CD). Rappelons que de façon générale, l'équation d'une droite qui n'est pas verticale, s'écrit de la forme* $y = ax + b$. *Pour déterminer les valeurs des paramètres a et b dans le cas de la droite (AB), il faut résoudre le système d'équations*

$$\begin{cases} 250a + b = 100 \\ 270a + b = 20 \end{cases}$$

qui traduit le fait que cette droite passe par le point A de coordonnées $(250, 100)$ *et le point B de coordonnées* $(270, 20)$*. On a*

$$\left\{ \begin{array}{l} 250a + b = 100 \\ 270a + b = 20 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} 250a + b = 100 \\ -20a = 80 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} b = 100 - 250 \times (-4) = 1100 \\ a = -4 \end{array} \right.$$

ainsi la droite (AB) admet pour équation $y = -4x + 1100$*. Pour déterminer les valeurs des paramètres a et b dans le cas de la droite (CD), il faut résoudre le système d'équations*

$$\left\{ \begin{array}{l} 250a + b = 60 \\ 270a + b = 100 \end{array} \right.$$

qui traduit le fait que cette droite passe par le point D de coordonnées $(250, 60)$ *et le point C de coordonnées* $(270, 100)$*. On a*

$$\left\{ \begin{array}{l} 250a + b = 60 \\ 270a + b = 100 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} 250a + b = 60 \\ 20a = 40 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} b = 60 - 250 \times 2 = -440 \\ a = 2 \end{array} \right.$$

ainsi la droite (CD) admet pour équation $y = 2x - 440$*. Finalement les coordonnées* (M_G, N_G) *du point G sont obtenues en résolvant le système d'équations*

$$\left\{ \begin{array}{l} N_G = -4M_G + 1100 \\ N_G = 2M_G - 440 \end{array} \right.$$

qui traduit le fait que ces coordonnées vérifient à la fois l'équation de la droite (AB) et celle de la droite (CD). On a

$$\left\{ \begin{array}{l} N_G = -4M_G + 1100 \\ N_G = 2M_G - 440 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} -6M_G + 1540 = 0 \\ N_G = 2M_G - 440 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} M_G = \frac{770}{3} \simeq 256.66 \\ \\ N_G = 2 \times \frac{770}{3} - 440 \simeq 73.33 \end{array} \right.$$

2.3.2 Médiane et Quantile

La **médiane** (notée M_e) d'une variable quantitative est la valeur de cette variable qui permet de scinder la population étudiée en deux sous-populations de même effectif. Plus précisément, il y a autant d'individus pour lesquels on a observé une valeur supérieure à M_e que d'individus pour lesquels on a observé une valeur inférieure à M_e.

a) Variable quantitative discrète (non classée)

On attribue d'abord à chacun des individus un rang, en partant de l'individu (ou des individus) pour lequel (lesquels) on a observé la valeur la plus forte. On attribue ensuite à chacun des individus un autre rang, en partant, cette fois, de l'individu (ou des individus) pour lequel (lesquels) on a observé la valeur la plus faible. On attribue enfin à chacun des individus une quantité appelée « profondeur » qui est le minimum de ses deux rangs.

→ **Dans le cas où la population est formée par un nombre impair des individus,** la médiane de la variable statistique est alors sa valeur qui corresponds aux profondeurs maximales.

Etudions un exemple concret :

Exemple 2.3.

Individu	*Note à l'Examen de Statistique*	*Rang (haut)*	*Rang (bas)*	*Profondeur*
Michel	*12*	*6*	*9*	*6*
Jean	*8*	*12*	*4*	*4*
Stéphane	*13*	*5*	*11*	*5*
Charles	*11*	*8*	*7*	*7*
Agnès	*10*	*10*	*6*	*6*
Nadine	*9*	*11*	*5*	*5*
Étienne	*16*	*2*	*14*	*2*
Gilles	*14*	*4*	*12*	*4*
Aurélie	*11*	*8*	*7*	*7*
Stéphanie	*15*	*3*	*13*	*3*
Marie-Claude	*4*	*14*	*2*	*2*
Anne	*18*	*1*	*15*	*1*
Christophe	*12*	*6*	*9*	*6*
Pierre	*6*	*13*	*3*	*3*
Bernadette	*2*	*15*	*1*	*1*

La médiane vaut $M_e = 11$.

→ **Dans le cas où la population est formée par un nombre pair d'individus,** la médiane de la variable statistique est alors la moyenne de ses valeurs qui correspondent aux profondeurs maximales.

Etudions un exemple concret :

Exemple 2.4. *Il s'agit du même exemple que celui qu'on vient de voir, sauf que l'on suppose ici que Bernadette n'a pas participé l'examen*

Individu	*Note à l'Examen de Statistique*	*Rang (haut)*	*Rang (bas)*	*Profondeur*
Michel	*12*	*6*	*8*	*6*
Jean	*8*	*12*	*3*	*3*
Stéphane	*13*	*5*	*10*	*5*
Charles	*11*	*8*	*6*	*6*
Agnès	*10*	*10*	*5*	*5*
Nadine	*9*	*11*	*4*	*4*
Étienne	*16*	*2*	*13*	*2*
Gilles	*14*	*4*	*11*	*4*
Aurélie	*11*	*8*	*6*	*6*
Stéphanie	*15*	*3*	*12*	*3*
Marie-Claude	*4*	*14*	*1*	*1*
Anne	*18*	*1*	*14*	*1*
Christophe	*12*	*6*	*8*	*6*
Pierre	*6*	*13*	*2*	*2*

La médiane M_e *vaut*

$$M_e = \frac{11 + 11 + 12 + 12}{4} = 11,5$$

Exercice 2.1. *(a) Supposons que Agnès et Stéphanie n'ont pas passé l'examen. Déterminer la médiane. (b) Supposons que Jean et Agnès n'ont pas passé l'examen. Déterminer la médiane.*

b) Variable quantitative continue et variable discrète classée

Commençons d'abord par introduire les notions **d'effectif cumulé**, **de fréquence cumulée**, et **de fonction cumulative**. X désigne une variable quantitative continue, ou encore une variable discrète classée, dont l'intervalle de variation a été divisé en « k » classes disjointes $[x_0, x_1], \ldots,]x_{k-1}, x_k]$. Les effectifs correspondant à ces classes sont notés « n_1 », « n_2 », $\ldots$, « n_k ». **L'effectif cumulé de la 1-ère classe** (c'est-à-dire de la classe $[x_0, x_1]$) est le nombre « N_1 » d'individus pour lesquels *la variable X prend une valeur au plus égale à* x_1 ; on a donc

$$N_1 = n_1.$$

L'effectif cumulé de la 2-ème classe (c'est à dire de la classe $]x_1, x_2]$) est le nombre « N_2 » d'individus pour lesquels *la variable X prend une valeur au plus égale à* x_2 ; on a donc

$$N_2 = n_1 + n_2.$$

L'effectif cumulé de la 3-ème classe (c'est à dire de la classe $]x_2, x_3]$) est le nombre « N_3 » d'individus pour lesquels *la variable X prend une valeur au plus égale à* x_3 ; on a donc

$$N_3 = n_1 + n_2 + n_3.$$

Plus généralement, **l'effectif cumulé de la i-ème classe** (c'est-à-dire de la classe $]x_{i-1}, x_i]$) où $i = 1, 2, \ldots, k$ est le nombre « N_i » d'individus pour lesquels *la variable X prend une valeur au plus égale à* x_i ; on a donc

$$N_i = n_1 + n_2 + \ldots + n_i = \sum_{l=1}^{i} n_l.$$

La fréquence cumulée de la i-ème classe est désignée par F_i et elle est définie par

$$F_i = \frac{N_i}{N} = \sum_{l=1}^{i} f_l,$$

où f_l est la fréquence de la l-ème classe et N est l'effectif total. Ainsi, on a $F_1 = f_1$ et $F_i = F_{i-1} + f_i$ pour tout $i = 2, \ldots, k$.

Exemple 2.5. *Construisons le tableau des effectifs cumulés et des fréquences cumulés de la variable « Revenu des Contribuables »*

Classes des revenus	*Effectifs*	*Effectifs Cumulés*	*Fréquences*	*Fréquences Cumulées*
$[0, 5000]$	*549,3*	*549,3*	*0,0667*	*0,0667*
$]5000, 10000]$	*3087,4*	*3636,7*	*0,3751*	*0,4418*
$]10000, 15000]$	*2229,0*	*5865,7*	*0,2708*	*0,7126*
$]15000, 20000]$	*1056,7*	*6922,4*	*0,1284*	*0,841*
$]20000, 35000]$	*925,0*	*7847,4*	*0,1124*	*0,9534*
$]35000, 50000]$	*211,0*	*8058,4*	*0,0256*	*0,979*
$]50000, 70000]$	*90,8*	*8149,2*	*0,011*	*0,99*
$]70000, 100000]$	*81,6*	*8230,8*	*0,0099*	$0,9999 \simeq 1$

Exercice 2.2. *Construisez le tableau des effectifs cumulés et des fréquences cumulées de la variable discrète classée « Note à l'Examen de Statistique » dont il est question dans l'Exemple 2.3.*

Correction de l'Exercice 2.2

Note à l'Examen de Statistique	*Effectifs*	*Effectifs Cumulés*	*Fréquences*	*Fréquences Cumulées*
$[0, 4]$	*2*	*2*	*0.133*	*0.133*
$]4, 8]$	*2*	*4*	*0.133*	*0.266*
$]8, 12]$	*6*	*10*	*0.4*	*0.666*
$]12, 16]$	*4*	*14*	*0.267*	*0.933*
$]16, 20]$	*1*	*15*	*0.067*	*1*

La fonction cumulative (qu'on appelle aussi fonction de répartition) est souvent notée par F ; cette fonction donne, pour tout nombre réel t, le pourcentage, noté par $F(t)$, des individus de la population pour lesquels on a observé une valeur de la variable X plus petite ou égale à t.

Remarque 2.3. (Propriétés importantes de la fonction cumulative F)

1. *Elle est croissante, c'est-à-dire que pour tous nombres réels t_1 et t_2, vérifiant $t_1 \leq t_2$, on a $F(t_1) \leq F(t_2)$.*
2. *Elle est nulle pour tout nombre réel t inférieur à x_0, où x_0 désigne la borne de gauche de la première classe c'est-à-dire $[x_0, x_1]$.*
3. *Elle est égale à 1 pour tout nombre réel t supérieur à x_k, où x_k désigne la borne de droite de la dernière classe c'est-à-dire $]x_{k-1}, x_k]$.*

Remarque 2.4. *Lorsque X est une variable continue, sa fonction cumulative F n'est connue que pour les valeurs de X égales aux extrémités des classes c'est-à-dire pour $t = x_0, t = x_1, \ldots, t = x_k$. On peut considérer que F est linéaire (fonction affine) entre ces valeurs, parce qu'on suppose que les classes forment des entités homogènes.*

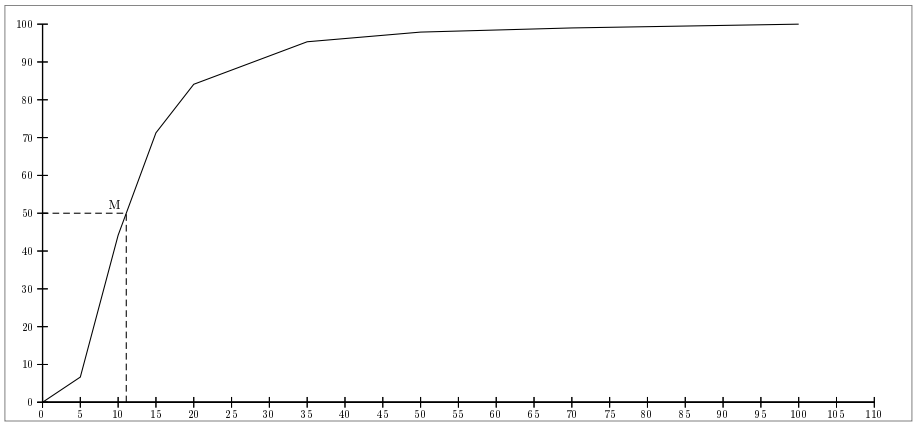
Remarque 2.5. *De façon générale, la médiane notée par M_e d'une variable statistique continue X de fonction cumulative F est telle que*

$$F(M_e) = 50\% ;$$

on peut déterminer M_e au moyen de la représentation graphique de F.

Exemple 2.6. *Traçons le graphe de la fonction cumulative de la variable continue « Revenu des Contribuables », puis déterminons la médiane de cette variable.*

Graphe de la fonction cumulative F de la variable continue « Revenu des Contribuables »



l'unité sur l'axe des abscisses est 1 *millier de Francs, l'axe des ordonnées représente les pourcentages cumulés*

Graphiquement on trouve que la médiane M_e de cette variable vaut $M_e \simeq 11.1$ milliers de Francs.

Si on souhaite obtenir M_e avec davantage de précision on peut procéder de la façon suivante. On commence d'abord par déterminer l'équation de la droite sur laquelle se trouve le point M ; il s'agit en fait de la droite passant par le point de coordonnées $(10,\ 44.18)$ *et le point de coordonnées* $(15,\ 71.26)$ *; ainsi il faut résoudre le système d'équation*

$$\begin{cases} 10a + b = 44.18 \\ 15a + b = 71.26 \end{cases}$$

On a

$$\begin{cases} 10a + b = 44.18 \\ 15a + b = 71.26 \end{cases} \Leftrightarrow \begin{cases} 10a + b = 44.18 \\ 5a = 71.26 - 44.18 = 27.08 \end{cases} \Leftrightarrow \begin{cases} b = 44.18 - 10 \times 5.416 = -9.98 \\ a = \frac{27.08}{5} = 5.416 \end{cases}$$

L'équation qu'on cherche à déterminer est donc $y = 5.416x - 9.98$. Finalement, en traduisant le fait que cette vérifiée par $(M_e, 50)$ les coordonnées du point M, on obtient $50 = 5.416M_e - 9.98$, d'où

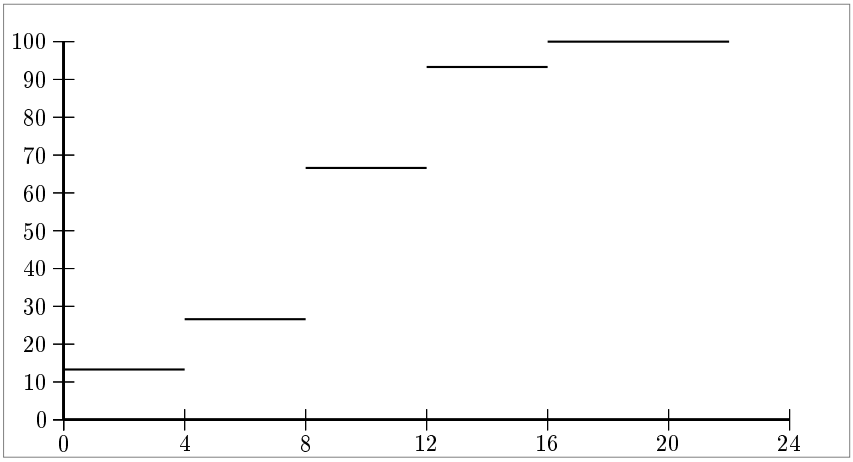
$$M_e = \frac{50 + 9.98}{5.416} \simeq 11.075 \text{ milliers de Francs.}$$

Une autre méthode de calcul de M_e, consiste à utiliser le Théorème de Thalès :

$$\frac{50 - 44.18}{71.26 - 44.18} = \frac{M_e - 10}{15 - 10} \Leftrightarrow M_e = (15-10) \times \Big(\frac{50 - 44.18}{71.26 - 44.18}\Big) + 10 \simeq 11.075 \text{ milliers de Francs.}$$

Remarque 2.6. *Lorsque X est* **une variable discrète classée** *(par exemple la variable « Note à l'Examen » dans l'Exercice 2.2), le graphe de sa fonction cumulative présente des sauts et a l'allure de marches d'escalier ; ainsi,* **en général, il n'existe pas une valeur médiane M_e pour laquelle la fonction cumulative vaut** 50% **exactement.** *Il faut donc dans ce cas utiliser d'autres valeurs typiques pour caractériser la tendance centrale de cette variable.*

Graphe de la fonction cumulative de la variable discrète classée « Note à l'Examen »



La notion de **quantile d'ordre** α $(0 \leq \alpha \leq 1)$, encore appelée **fractile d'ordre** α, généralise la notion de médiane. Le quantile d'ordre α d'une variable quantitative X, est la valeur x_α de cette variable qui permet de scinder la population étudiée en deux sous-populations dont les effectifs respectifs sont égaux à α et $1-\alpha$ de l'effectif de la population initiale. Lorsque X est continue, on peut déterminer x_α au moyen de l'égalité

$$F(x_\alpha) = \alpha.$$

Les quartiles de X sont ses trois quantiles $x_{0,25}$, $x_{0,5}$ et $x_{0,75}$. $Q_1 = x_{0,25}$, s'appelle le premier quartile ; un quart des valeurs prises par X sont inférieures ou égales à Q_1. $Q_2 = x_{0,5} = M_e$ est la médiane. $Q_3 = x_{0,75}$ s'appelle le troisième quartile ; un quart des valeurs prises par X sont supérieures ou égales à Q_3.

L'intervalle interquartile (IIQ) est la différence entre le troisième quartile et le premier quartile ; il s'écrit :

$$IIQ = Q_3 - Q_1.$$

L'intervalle interquartile sert à apprécier la dispersion de X, de façon absolue, ou bien par comparaison avec une autre variable quantitative, à condition que cette dernière soit exprimée dans la même unité que X. En effet, les valeurs Q_1 et Q_3 délimitent une plage au sein de laquelle 50% des valeurs de X sont concentrées. Plus IIQ est grand, plus X est dispersée.

2.3.3 Moyennes

On dispose d'une population de N individus et on observe $x_1, x_2, \ldots, x_N$ les valeurs d'une variable quantitative discrète X pour ces individus.

a) Moyenne arithmétique

Elle est notée par $\overline{x}$ et elle est définie de la manière suivante :

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Exemple 2.7. *La moyenne arithmétique de la variable « Note à l'Examen de Statistique », dont il est question dans l'Exemple 2.3), vaut* $\frac{161}{15} \simeq 10,73$ *; dans le cas de l'Exemple 2.4, la moyenne*

arithmétique devient $\frac{159}{14} \simeq 11,36$. *Notons que le fait que Bernadette ne participe pas à l'examen (c'est la seule différence entre l'Exemple 2.3 et l'Exemple 2.4), a un impact plus significatif sur moyenne arithmétique que sur la médiane ; rappelons que cette dernière augmente de* 11 *à* $11,5$. *De façon générale, la moyenne arithmétique est davantage sensible aux valeurs extrêmes que la médiane.*

Désignons par n_i le nombre de fois où la valeur x_i de la variable X est observée (par exemple dans le cas de la variable « Note à l'Examen de Statistique », la valeur 18 est observée 1 fois, tandis que la aleur 11 est observée 2 fois) ; ainsi, étant donné que $\underbrace{x_i + x_i + \ldots + x_i}_{n_i \text{ fois}} = n_i x_i$, la formulation précédente de $\overline{x}$, peut aussi s'écrire

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{K} n_i x_i = \sum_{i=1}^{K} f_i x_i,$$

où K désigne le nombre de valeurs *distinctes* de X et $f_i = n_i/N$ est la fréquence de la valeur x_i.

La formulation $\sum_{i=1}^{K} f_i x_i$ est appelée **moyenne arithmétique pondérée de X**, car l'on pondère chacune des valeurs distinctes de X par la fréquence correspondante.

Exemple 2.8. *Une étude statistique menée sur une population de ménages a montré que* 30% *de ces ménages ont 1 enfants,* 40% *2 enfants,* 15% *3 enfants,* 10% *4 enfants, et* 5% *5 enfants.*

Le nombre moyen d'enfants par ménage vaut :

$$\overline{x} = 0,3 \times 1 + 0,4 \times 2 + 0,15 \times 3 + 0,1 \times 4 + 0,05 \times 5 \simeq 2,2 \text{ enfants.}$$

Remarque 2.7. *Plaçons nous dans l'un ou l'autre des deux cas suivants :*

- *Y est une variable quantitative continue, dont l'intervalle de variation a été divisé en k classes jointives $[y_0, y_1]\,,\,]y_1, y_2]\,,\, \ldots\,,\,]y_{k-1}, y_k]$;*
- *Y est une variable discrète classée dont les classes sont $[y_0, y_1]\,,\,]y_1, y_2]\,,\, \ldots\,,\,]y_{k-1}, y_k]$.*

Alors, $\overline{y}$ la moyenne arithmétique de Y, est définie comme la moyenne arithmétique des centres des classes de Y pondérées par les fréquences correspondantes *; plus précisément :*

$$\overline{y} = \sum_{i=1}^{k} f_i\Big(\frac{y_{i-1} + y_i}{2}\Big) = \frac{1}{N}\sum_{i=1}^{k} n_i\Big(\frac{y_{i-1} + y_i}{2}\Big),$$

où, pour tout i, f_i et n_i désignent respectivement la fréquence et l'effectif de la i-ème classe, $N = \sum_{i=1}^{k} n_i$ étant l'effectif total.

Exercice 2.3. (*a*) *Calculer la moyenne arithmétique de la variable continue « Revenu des Contribuables ».* (*b*) *Calculer la moyenne arithmétique de la variable classée « Note à l'Examen de Statistique » dont il est question dans l'Exercice 2.2.*

b) Moyenne quadratique

Elle est notée par m_2 et elle est définie de la manière suivante :

$$m_2 = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} = \sqrt{\sum_{i=1}^{K} f_i x_i^2}.$$

Ainsi, la moyenne quadratique de la variable « Nombre d'Enfants par Ménage », dont il est question dans l'Exemple 2.8, vaut :

$$m_2 = \big(0,3 \times 1^2 + 0,4 \times 2^2 + 0,15 \times 3^2 + 0,1 \times 4^2 + 0,05 \times 5^2\big)^{1/2} \simeq 2,47.$$

c) Moyenne harmonique

Elle est notée par m_{-1} et elle est définie de la manière suivante :

$$m_{-1} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} = \frac{1}{\sum_{i=1}^{K} \frac{f_i}{x_i}}.$$

La moyenne harmonique peut être utilisée chaque fois qu'il est possible d'attribuer un sens réel aux inverses des données (taux d'équipement, pouvoir d'achat, calcul d'indice, ...).

Exemple 2.9. *On achète des Dollars une première fois pour* 100 *Euros au cours de* $0,87$ *Euro le Dollar, puis on en achète une seconde fois pour* 100 *Euros également mais au cours de* $0,71$ *Euro le Dollars ; ainsi le montant total des Dollars achetés lors de ces deux opérations est :*

$$\frac{100}{0,87} + \frac{100}{0,71} \simeq 255,79 \text{ Dollars.}$$

Le cours moyen du Dollar pour l'ensemble de ces opérations *est, par définition, le cours de* c_m *Euro le Dollar, qui aurait permis l'achat, en une seule fois, de* $255,79$ *Dollars pour* 200 *Euros ; ainsi*

$$\frac{200}{c_m} = \frac{100}{0,87} + \frac{100}{0,71} \simeq 255,79$$

d'où

$$c_m = \frac{200}{\frac{100}{0,87} + \frac{100}{0,71}} = \frac{2}{\frac{1}{0,87} + \frac{1}{0,71}} \simeq 0,78$$

Il apparait donc que c_m *est la moyenne harmonique des deux cours correspondant aux deux opérations ; aussi, il est important de noter que* c_m *est différent (strictement plus petit) de la moyenne arithmétique de ces deux cours, en effet cette dernière moyenne vaut* $(0,87+0,71)/2 = 0,79$.

Exercice 2.4. *Un automobiliste parcourt 40 kilomètres à 60 km/h puis 40 autres kilomètres à 120km/h ; on note par* v_m *sa vitesse moyenne en km/h sur l'ensemble de ce trajet de 80 kilomètres. Calculer* v_m.

d) Moyenne géométrique

Attention : on ne peut définir cette moyenne que lorsque les observations $x_1, \ldots, x_N$ sont tous des nombres réels positifs. Si tel est le cas, la moyenne géométrique de ces observations est notée par M_g, et elle est définie par :

$$M_g = \sqrt[N]{x_1 x_2 \ldots x_N} = \sqrt[N]{x_1^{n_1} \ldots x_K^{n_K}} = x_1^{f_1} \ldots x_K^{f_K}.$$

Exemple 2.10. *Supposons que pendant une décennie, les salaires aient été multipliés par* 2 *et que pendant la décennie suivante ils aient été multipliés par* 4 *; alors pour la période de l'ensemble de ces deux décennies le coefficient multiplicateur est* $2 \times 4 = 8$. Le coefficient multiplicateur moyen par décennie *pour cette période de vingt ans est, par définition, le coefficient* μ *qui ne change pas d'une décennie à l'autre, et qui permet une multiplication par* 8 *des salaires entre le début et la fin de la période. On a donc* $\mu^2 = 8 = 2 \times 4$, *d'où* $\mu = \sqrt{2 \times 4} \simeq 2,83$. *Ainsi, il apparait que* μ *est la moyenne géométrique des deux coefficients multiplicateurs correspondant aux deux décennies ; aussi, il est important de noter que* μ *est différent (strictement plus petit) de la moyenne arithmétique de ces deux coefficients, en effet cette dernière moyenne vaut* $(2+4)/2 = 3$.

Remarque 2.8. *Lorsque les observations $x_1, \ldots, x_N$ sont tous des nombres réels positifs, alors*

$$\min_{1\leq i\leq N} x_i \leq m_{-1} \leq M_g \leq \overline{x} \leq \max_{1\leq i\leq N} x_i$$

autrement dit

(Le minimum des observations)
$\leq$ *(La moyenne harmonique des observations)*
$\leq$ *(La moyenne géométrique des observations)*
$\leq$ *(La moyenne arithmétique des observations)*
$\leq$ *(Le maximum des observations)*

Grâce à ces inégalités, on peut se rendre compte de certaines erreurs qui seraient commises lors du calcul de ces moyennes.

2.3.4 Indicateurs de dispersion

On dispose d'une population de N individus, et on observe $x_1, \ldots, x_N$ les valeurs d'une variable quantitative discrète X pour ces individus.

a) L'étendue

L'étendue e_X de la variable quantitative discrète X est *la différence entre la plus grande et la plus petite des valeurs observées :*

$$e_X = \max_{1\leq i\leq N} x_i - \min_{1\leq i\leq N} x_i.$$

Dans le cas de la variable « Note à l'Examen de Statistique », l'étendue vaut $18 - 2 = 16$.

b) Variance et Écart-type

La variance de la variable quantitative X, notée par $\mathrm{Var}(X)$, est, par définition, *la moyenne arithmétique des carrés des écarts à la moyenne arithmétique* :

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2\,; \tag{2.1}$$

cette formule peut également se réecrire sous la forme :

$$\mathrm{Var}(X) = \sum_{i=1}^{K} f_i(x_i - \overline{x})^2,$$

où K désigne le nombre de valeurs *distinctes* de X et $f_i = n_i/N$ est la fréquence de la valeur x_i. Une autre formule importante (parfois désignée par formule de Huygens) permettant le calcul de la variance, est :

$$\begin{aligned}\mathrm{Var(X)} &= \left(\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\mathrm{x_i^2}\right) - (\overline{\mathrm{x}})^2 = \left(\sum_{\mathrm{i}=1}^{\mathrm{K}}\mathrm{f_i x_i^2}\right) - (\overline{\mathrm{x}})^2 \\ &= \big(\text{Moyenne quadratique de } X\big)^2 - \big(\text{Moyenne arithmétique de } X\big)^2 \end{aligned} \tag{2.2}$$

L'écart-type de la variable X, noté par σ_X, est, par définition, *la racine carrée de la variance de cette variable* :

$$\sigma_X = \sqrt{\mathrm{Var(X)}}.$$

Signalons au passage que l'écart-type est la mesure de la dispersion la plus couramment utilisée.

Exemple 2.11. *Déterminons la variance et l'écart-type de la variable « Note à l'Examen de Statistique » désignée par X ; rappelons que $\overline{x}$, la moyenne arithmétique de cette variable, vaut* $= 10,73$

Individu	*Note à l'Examen de Statistique*	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	x_i^2
Michel	*12*	*1,27*	*1,61*	*144*
Jean	*8*	*-2,73*	*7,45*	*64*
Stéphane	*13*	*2,27*	*5,15*	*169*
Charles	*11*	*0,27*	*0,07*	*121*
Agnès	*10*	*-0,73*	*0,53*	*100*
Nadine	*9*	*-1,73*	*2,99*	*81*
Étienne	*16*	*5,27*	*27,77*	*256*
Gilles	*14*	*3,27*	*10,69*	*196*
Aurélie	*11*	*0,27*	*0,07*	*121*
Stéphanie	*15*	*4,27*	*18,23*	*225*
Marie-Claude	*4*	*-6,73*	*45,29*	*16*
Anne	*18*	*7,27*	*52,86*	*324*
Christophe	*12*	*1,27*	*1,61*	*144*
Pierre	*6*	*-4,73*	*22,37*	*36*
Bernadette	*2*	*-8,73*	*76,21*	*4*
			Total=272,9	*Total=2001*

Nous allons calculer Var(X) *au moyen de deux méthodes, la première d'entre elles consiste à utiliser la formule (2.1) et la seconde la formule (2.2).*

Présentons d'abord **la première méthode**. *La somme des carrés des écarts à la moyenne arithmétique vaut* $272,9$ *(voir l'avant dernière colonne du tableau) ; ainsi en utilisant la formule (2.1), on obtient :*

$$\mathrm{Var(X)} = \frac{272,9}{15} \simeq 18,19 \tag{2.3}$$

Présentons maintenant **la seconde méthode**. *La somme des carrés des observations vaut* 2001 *(voir la dernière colonne du tableau) ; ainsi*

$$\big(\textit{Moyenne quadratique de } X\big)^2 = \frac{2001}{15} = 133,4$$

et d'après la formule (2.2),

$$\mathrm{Var(X)} = 133,4 - (10,73)^2 \simeq 18,27 \tag{2.4}$$

Signalons que la légère différence entre le résultat (2.3) et le résultat (2.4), s'explique par les erreurs d'arrondi. D'ailleurs cette petite différence devient presque inexistante, lorsqu'on calcule l'écart-type correspondant à chacun de ces deux résultats ; en effet on a $\sqrt{18,19} \simeq 4,26$ *et* $\sqrt{18,27} \simeq 4,27$.

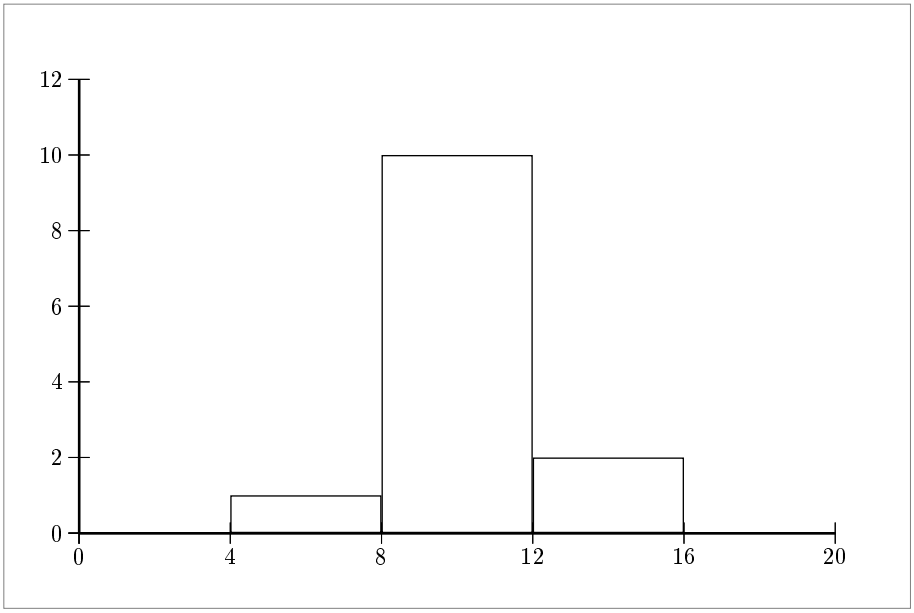
Exemple 2.12 (Illustration de l'utilité de l'écart-type)**.** *Les 25 étudiants d'un Master sont répartis en deux groupes, 13 étudiants sont dans le groupe 1 et les 12 restant dans le groupe 2. Ces 25 étudiants ont passé un examen ; le tableau suivant donne un descriptif de la répartition des notes obtenues dans chacun de ces deux groupes :*

Tableau de répartition des notes dans chacun des deux groupes

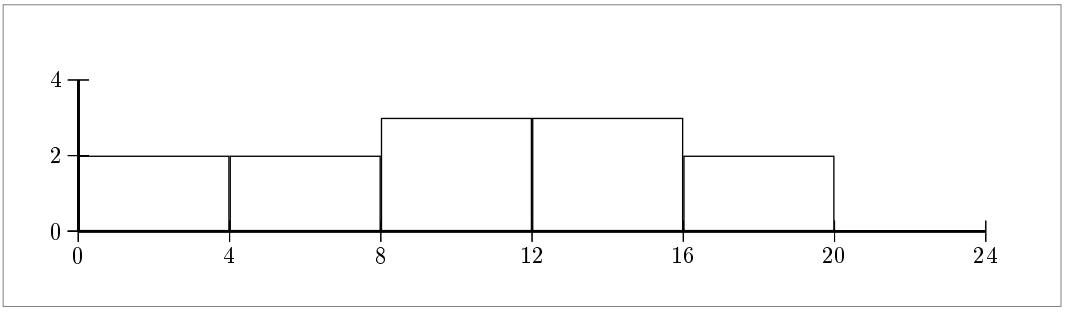
Centres des Classes	*Classes de Note*	*Effectifs du groupe 1*	*Effectifs du groupe 2*
2	*[0,4]*	*0*	*2*
6	*]4,8]*	*1*	*2*
10	*]8,12]*	*10*	*3*
14	*]12,16]*	*2*	*3*
18	*]16,20]*	*0*	*2*
		$Total = N_1 = 13$	$Total = N_2 = 12$

Nous souhaitons comparer les répartitions des notes, dans chacun de ces deux groupes.

Histogramme des effectifs du groupe 1



Histogramme des effectifs du groupe 2



Nous constatons graphiquement que les notes des étudiants du groupe 1 sont très resserrées, alors que celles des étudiants du groupe 2 sont dispersées. Le calcul, pour chacun des deux groupes, de la moyenne arithmétique des notes ainsi que leur écart-type, va nous permettre de préciser cette constatation graphique. Commençons d'abord par $\overline{x}_1$ et $\overline{x}_2$ les moyennes respectives des deux groupes ; la variable « Note » (désignée par X_1 pour le groupe 1, et par X_2 pour le groupe 2) étant classée, sa moyenne, dans chacun des deux groupes, est égale à la moyenne des centres des classes pondérés par les fréquences correspondantes. On a donc pour le groupe 1,

$$\overline{x}_1 = \frac{1 \times 6 + 10 \times 10 + 2 \times 14}{13} = \frac{134}{13} \simeq 10,31$$

et pour le groupe 2

$$\overline{x}_2 = \frac{2 \times 2 + 2 \times 6 + 3 \times 10 + 3 \times 14 + 2 \times 18}{12} = \frac{124}{12} \simeq 10,33$$

Calculons maintenant V_1 et V_2 les variances respectives de la variable « Note » dans chacun des deux groupes. En utilisant la formule (2.2), on obtient :

$$V_1 = \frac{1 \times 6^2 + 10 \times 10^2 + 2 \times 14^2}{13} - \left(\frac{134}{13}\right)^2 \simeq 3,76$$

et

$$V_2 = \frac{2 \times 2^2 + 2 \times 6^2 + 3 \times 10^2 + 3 \times 14^2 + 2 \times 18^2}{12} - \left(\frac{124}{12}\right)^2 \simeq 27,96\,;$$

notons que les carrés des moyennes quadratiques (utilisé dans les calculs de V_1 et V_2), sont les moyennes arithmétiques des carrés des centres des classes pondérés par les fréquences correspondantes. *Enfin, σ_1 et σ_2, les écarts-type respectifs de la variable « Note » dans chacun des deux groupes, valent :*

$$\sigma_1 = \sqrt{3,76} \simeq 1,94.$$

et

$$\sigma_2 = \sqrt{27,96} \simeq 5,29.$$

Conclusion : *L'écart-type des notes du groupe 1 est modéré, cela signifie que les notes dans ce groupe sont homogènes et concentrées autour de la moyenne. En revanche, avec une moyenne pratiquement identique, les notes dans le groupe 2 présentent un écart-type nettement plus important, ce qui réflète leur hétérogénéité.*

c) Variance Totale, Variance Intra-groupe, Variance Inter-groupe

L'Exemple 2.12, qu'on vient d'étudier, permet d'introduire brièvement les notions de Variance Totale, Variance Intra-groupe, Variance Inter-groupe. Intéressons-nous à présent à la répartition des notes des 25 étudiants du Master, dans leur ensemble ; le tableau suivant donne un descriptif de celle-ci :

Tableau de répartition des notes de l'ensemble des étudiants du Master

Centres des Classes	Classes de Note	Effectifs
2	[0,4]	2
6	]4,8]	3
10	]8,12]	13
14	]12,16]	5
18	]16,20]	2
		$Total = N$ =25

Dans ce cadre la variable classée « Note » est désignée par X. La moyenne arithmétique de X est appelée *la moyenne arithmétique totale* (puisqu'il s'agit de la moyenne pour les deux groupes à la fois), et elle est notée par $\overline{x}_T$. Cette moyenne totale est intimement liée à $\overline{x}_1$ et $\overline{x}_2$, les moyennes respectives dans chacun des deux groupes ; plus précisément $\overline{x}_T$ est la moyenne arithmétique de $\overline{x}_1$ et $\overline{x}_2$, pondérée par les "poids" des deux groupes :

$$\overline{x}_T = \Big(\frac{N_1}{N_1+N_2}\Big)\,\overline{x}_1 + \Big(\frac{N_2}{N_1+N_2}\Big)\,\overline{x}_2$$

Ainsi,

$$\overline{x}_T \simeq \frac{13}{25} \times 10,31 + \frac{12}{25} \times 10,33 \simeq 10,32$$

La variance de X est appelée *la variance totale* et elle est notée par V_T. Rappelons que V_1 et V_2 désignent les variances au sein de chaque groupe ; on peut montrer que

$$V_T = \overbrace{\underbrace{\Big(\frac{N_1}{N_1+N_2}\Big)V_1 + \Big(\frac{N_2}{N_1+N_2}\Big)V_2}_{\text{Variance Intra-groupe}}}^{\text{moyenne des variances des deux groupes}} + \overbrace{\underbrace{\Big(\frac{N_1}{N_1+N_2}\Big)\big(\overline{x}_1-\overline{x}\big)^2 + \Big(\frac{N_2}{N_1+N_2}\Big)\big(\overline{x}_2-\overline{x}\big)^2}_{\text{Variance Inter-groupe}}}^{\text{variance des moyennes des deux groupes}}$$

Ainsi,

$$V_T \simeq \left(\frac{13}{25} \times 3,76 + \frac{12}{25} \times 27,96\right) + \left(\frac{13}{25}\big(10,31-10,32\big)^2 + \frac{12}{25}\big(10,33-10,32\big)^2\right) \simeq 15,38$$

et donc l'écart-type de X vaut $\sqrt{15,38} \simeq 3,92$.

d) L'écart absolu moyen

L'écart absolu moyen à la moyenne de la variable quantitative discrète X est la moyenne arithmétique des valeurs absolues des écarts à la moyenne arithmétique :

$$e_{\overline{x}} = \frac{1}{N}\sum_{i=1}^{N} |x_i - \overline{x}| = \sum_{i=1}^{K} f_i |x_i - \overline{x}|,$$

où K désigne le nombre de valeurs distinctes de X et f_i la fréquence de x_i

Exemple 2.13. *Calculons $e_{\overline{x}}$, l'écart absolu moyen à la moyenne, de la variable quantitative « Note à l'Examen de Statistique », dont il est question dans l'Exemple 2.3 ; rappelons que $\overline{x}$, la moyenne arithmétique de cette variable, vaut à peu près* $10,73$. *On a*

Individu	*Note à l'Examen de Statistique*	$(x_i - \overline{x})$	$\lvert x_i - \overline{x}\rvert$
Michel	*12*	*1,27*	*1,27*
Jean	*8*	*-2,73*	*2,73*
Stéphane	*13*	*2,27*	*2,27*
Charles	*11*	*0,27*	*0,27*
Agnès	*10*	*-0,73*	*0,73*
Nadine	*9*	*-1,73*	*1,73*
Étienne	*16*	*5,27*	*5,27*
Gilles	*14*	*3,27*	*3,27*
Aurélie	*11*	*0,27*	*0,27*
Stéphanie	*15*	*4,27*	*4,27*
Marie-Claude	*4*	*-6,73*	*6,73*
Anne	*18*	*7,27*	*7,27*
Christophe	*12*	*1,27*	*1,27*
Pierre	*6*	*-4,73*	*4,73*
Bernadette	*2*	*-8,73*	*8,73*
			Total=50,81

Ainsi, on trouve que

$$e_{\overline{x}} \simeq \frac{50,81}{15} \simeq 3,39$$

L'écart absolu moyen à la médiane de la variable quantitative discrète X est la moyenne arithmétique des valeurs absolues des écarts à la médiane M_e.

$$e_{M_e} = \frac{1}{N}\sum_{i=1}^{N} |x_i - M_e| = \sum_{i=1}^{K} f_i |x_i - M_e|.$$

Exercice 2.5. *Calculer e_{M_e} l'écart absolu moyen à la médiane de la variable « Note à l'Examen de Statistique », dont il est question dans l'Exemple 2.3 ; rappelons que M_e, la médiane de cette variable, vaut* 11.

3 Analyse bivariée

L'objectif de l'analyse bivariée est d'étudier les éventuelles relations entre deux variables statistiques.

3.1 Liaison entre deux variables quantitatives

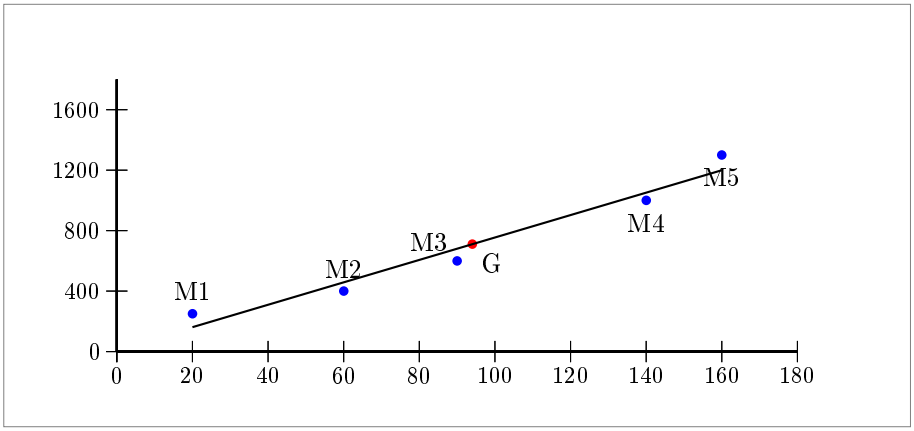
3.1.1 La régression linéaire simple

Exemple 3.1. *On souhaite étudier la relation superficie-prix de* 5 *appartements à Paris ; la variable quantitative X désigne la surface en m^2, et la variable quantitative Y le prix de vente en milliers d'Euros. Le tableau suivant donne les valeurs de ces deux variables, pour les 5 appartements :*

Tableau de Données

X (en m^2)	$x_1 = 20$	$x_2 = 60$	$x_3 = 90$	$x_4 = 140$	$x_5 = 160$
Y (en milliers d'Euros)	$y_1 = 250$	$y_2 = 400$	$y_3 = 600$	$y_4 = 1000$	$y_5 = 1300$

On commence par visualiser les variables X et Y en les représentant sous la forme **d'un nuage de point** : dans un repére cartésien, chaque observation (x_i, y_i) est figurée par le point M_i de coordonnées (x_i, y_i). On cherche une approximation de ce nuage dans un but de simplification ; sa forme donne une information sur le type d'une éventuelle liaison entre les variables X et Y.



Dans l'exemple étudié, on observe un nuage **oblong** (allongé), nous permettant d'envisager **une liaison linéaire** entre la surface d'un appartement et son prix. Plus précisément, il semble raisonnable de considérer que la relation entre la surface x_i d'un appartement et son prix y_i, est à peu près de la forme $y_i = ax_i + b$. Les coefficients (ou paramètres) a et b seront choisis de la sorte que la droite d'équation $y = ax + b$ passe **« le plus près possible de l'ensemble des points du nuage »** ; nous allons maintenant formaliser cette idée.

Considérons une droite D d'équation $y = ax + b$ et soit Δ la droite parallèle à l'axe des ordonnées et passant par le point M_i. Les droites Δ et D se coupent en un point $M_i^{'}$; la distance de M_i à $M_i^{'}$ vaut $|y_i - ax_i - b|$. Les coefficients a et b seront choisis de sorte que la quantité :

$$(y_1 - ax_1 - b)^2 + (y_2 - ax_2 - b)^2 + (y_3 - ax_3 - b)^2 + (y_4 - ax_4 - b)^2 + (y_5 - ax_5 - b)^2,$$

soit minimale.

Plus généralement, soient $x_1, x_2, \ldots, x_N$ et $y_1, y_2, \ldots, y_N$, les valeurs observées de deux variables quantitatives X et Y, pour un échantillon de N individus. Les coefficients de la **droite des moindres carrés**, c'est-à-dire de la droite qui permet d'ajuster au mieux, au sens du critère des moindres carrés, le nuage de points $M_1 = (x_1, y_1)$; $M_2 = (x_2, y_2)$; $\ldots$; $M_N = (x_N, y_N)$ sont les nombres a et b qui rendent minimale la quantité

$$(y_1 - ax_1 - b)^2 + (y_2 - ax_2 - b)^2 + \ldots + (y_N - ax_N - b)^2.$$

Ils sont donnés par les deux formules :

$$a = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \ldots + (x_N - \overline{x})(y_N - \overline{y})}{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_N - \overline{x})^2} \tag{3.1}$$

et

$$b = \overline{y} - a\overline{x}\,. \tag{3.2}$$

La formule (3.2) signifie que la droite des moindres carrés passe par **le centre de gravité** du nuage de points $M_1 = (x_1, y_1)$; $M_2 = (x_2, y_2)$; $\ldots$; $M_N = (x_N, y_N)$, c'est-à-dire par le point G de coordonnées $(\overline{x}, \overline{y})$ où, $\overline{x}$ et $\overline{y}$ sont les moyennes arithmétiques des variables X et Y.

Une fois qu'on a déterminé a et b, pour tout $i = 1, 2, \ldots, N$, on pose :

$$\widehat{y}_i = ax_i + b\,; \tag{3.3}$$

cette quantité $\widehat{y}_i$ est appelée **la valeur estimée de Y, par la droite des moindres carrés, lorsque X vaut x_i**. Quand l'ajustement est de bonne qualité, cette valeur estimée $\widehat{y}_i$ est assez proche de y_i la valeur réelle de Y lorsque X vaut x_i.

Appliquons maintenant, à l'exemple qui nous intéresse, les formules qu'on vient de donner dans un cadre général.

La moyenne arithmétique $\overline{x}$ des surfaces des 5 appartements vaut $\overline{x} = \frac{470}{5} = 94\,m^2$, la moyenne arithmétique $\overline{y}$ de leurs prix vaut $\overline{y} = \frac{3550}{5} = 710$ milliers d'Euros ; ainsi, G le centre gravité du nuage des 5 points associé aux variables X et Y, admet pour coordonnées $(94, 710)$.

Le tableau suivant va nous permettre de calculer les valeurs de a et b :

$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
-74	-460	34040	5476	211600
-34	-310	10540	1156	96100
-4	-110	440	16	12100
46	290	13340	2116	84100
66	590	38940	4356	348100
		Total = 97300	Total = 13120	Total = 752000

(3.4)

ainsi, grâce aux formules (3.1) et (3.2), on trouve que :

$$a = \frac{97300}{13120} \simeq 7,416 \quad \text{et} \quad b = 710 - 7,416 \times 94 \simeq 12,896\,, \tag{3.5}$$

donc la droite des moindres carrés admet pour équation :

$$y = 7,416\,x + 12,896\,.$$

Calculons enfin, $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_5$, les prix estimés en milliers d'Euros des 5 appartements. Grâce à (3.3) et à (3.5), on trouve que : $\widehat{y}_1 = 7,416 \times 20 + 12,896 \simeq 161$; $\widehat{y}_2 = 7,416 \times 60 + 12,896 \simeq 458$; $\widehat{y}_3 = 7,416 \times 90 + 12,896 \simeq 680$; $\widehat{y}_4 = 7,416 \times 140 + 12,896 \simeq 1051$ et $\widehat{y}_5 = 7,416 \times 160 + 12,896 \simeq 1199$.

Le tableau suivant permet de comparer les prix réels des appartements à leurs prix estimés au moyen de droite des moindres carrés :

X (en m^2)	$x_1 = 20$	$x_2 = 60$	$x_3 = 90$	$x_4 = 140$	$x_5 = 160$
Valeur réelle de Y (en milliers d'Euros)	$y_1 = 250$	$y_2 = 400$	$y_3 = 600$	$y_4 = 1000$	$y_5 = 1300$
Valeur estimée de Y (en milliers d'Euros)	$\widehat{y}_1 = 161$	$\widehat{y}_2 = 458$	$\widehat{y}_3 = 680$	$\widehat{y}_4 = 1051$	$\widehat{y}_5 = 1199$

3.1.2 Covariance et coefficient de corrélation

Il est toujours possible de tracer la droite des moindres carrés quelle que soit la forme du nuage de points $M_1 = (x_1, y_1)$; $M_2 = (x_2, y_2)$; … ; $M_N = (x_N, y_N)$. L'approximation de ce nuage par cette droite est-elle pour autant légitime ?

Un premier élément de réponse à cette question est donné par l'examen de $R(X, Y)$ **le coefficient de corrélation linéaire des variables X et Y** (parfois on dit le coefficient de corrélation

linéaire entre les variables X et Y). Pour pouvoir définir ce coefficient, il faut d'abord définir **la covariance de X et Y** (parfois on dit la covariance entre X et Y).

$x_1, x_2, \ldots, x_N$ et $y_1, y_2, \ldots, y_N$ désignent les valeurs prises par X et Y pour une population de N individus. **La covariance de X et Y**, notée par $\mathrm{cov(X, Y)}$, est définie par :

$$\mathrm{cov(X, Y)} = \frac{(\mathrm{x_1} - \overline{\mathrm{x}})(\mathrm{y_1} - \overline{\mathrm{y}}) + (\mathrm{x_2} - \overline{\mathrm{x}})(\mathrm{y_2} - \overline{\mathrm{y}}) + \ldots + (\mathrm{x_N} - \overline{\mathrm{x}})(\mathrm{y_N} - \overline{\mathrm{y}})}{\mathrm{N}}\,, \tag{3.6}$$

où $\overline{x}$ et $\overline{y}$ désignent les moyennes arithmétiques de X et Y ; notons que

$$\mathrm{cov(X, X)} = \mathrm{Var(X)}\,.$$

La covariance de X et Y peut aussi être calculée au moyen de la formule (parfois désignée par formule de Huygens) :

$$\mathrm{cov(X, Y)} = \left(\frac{\mathrm{x_1 y_1} + \mathrm{x_2 y_2} + \ldots + \mathrm{x_N y_N}}{\mathrm{N}}\right) - \overline{\mathrm{x}}\,\overline{\mathrm{y}}\,; \tag{3.7}$$

en fait la formule (2.2) n'est rien d'autre que la formule (3.7), dans le cas où $X = Y$.

Exemple 3.2. *Soient X et Y les variables « Superficie » et « Prix », dont il est question dans l'Exemple 3.1 (l'exemple des appartements). Nous allons calculer* $\mathrm{cov(X, Y)}$ *au moyen de deux méthodes : la première d'entre elles consiste à utiliser la formule (3.6), et la seconde consiste à à utiliser la formule (3.7).*

Présentons d'abord la première méthode. On a déjà vu que (voir le tableau (3.4)) :

$$(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + (x_3 - \overline{x})(y_3 - \overline{y}) + (x_4 - \overline{x})(y_4 - \overline{y}) + (x_5 - \overline{x})(y_5 - \overline{y}) = 97300\,;$$

ainsi, il résulte de la formule (3.6) que :

$$\mathrm{cov(X, Y)} = \frac{97300}{5} = 19460\,.$$

Présentons maintenant la seconde méthode. Pour calculer $x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + x_5y_5$, nous utilisons le tableau suivant :

x_i	y_i	x_iy_i
20	*250*	*5000*
60	*400*	*24000*
90	*600*	*54000*
140	*1000*	*140000*
160	*1300*	*208000*
		total = 431000

qui nous permet de trouver que : $x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + x_5y_5 = 431000$; ainsi, on obtient que :

$$\frac{x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + x_5y_5}{5} = \frac{431000}{5} = 86200\,. \tag{3.8}$$

D'autre part, dans la Sous-section 3.1.1, on a vu que $\overline{x} = 94$ et $\overline{y} = 710$; on a par conséquent :

$$\overline{x}\,\overline{y} = 94 \times 710 = 66740\,. \tag{3.9}$$

Finalement, en utilisant la formule (3.7), ainsi que (3.8) et (3.9), on obtient :

$$\mathrm{cov(X, Y)} = 86200 - 66740 = 19460\,.$$

Remarque 3.1. (Inégalité de Cauchy-Schwarz) *La valeur absolue de la covariance de deux variables quantitatives X et Y, est toujours inférieure ou égale au produit de leurs écarts-types :*

$$|\mathrm{cov(X,Y)}| \leq \sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}\,;$$

cette inégalité peut aussi s'écrire sous la forme

$$-\sigma_X\sigma_Y \leq \mathrm{cov(X,Y)} \leq \sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}.$$

Ecrivons l'inégalité de Cauchy-Schwarz dans le cas de l'Exemple 3.1 (l'exemple des appartements). Pour cet exemple, on a déjà montré que $\mathrm{cov(X,Y)} = 19460$; il nous reste à calculer les écarts-types σ_X et σ_Y. On a déjà vu que (voir le tableau (3.4)) :

$$(x_1-\overline{x})^2+(x_2-\overline{x})^2+(x_3-\overline{x})^2+(x_4-\overline{x})^2+(x_5-\overline{x})^2 = 13120$$

et

$$(y_1-\overline{y})^2+(y_2-\overline{y})^2+(y_3-\overline{y})^2+(y_4-\overline{y})^2+(y_5-\overline{y})^2 = 752000\,;$$

on obtient donc, au moyen de la formule (2.1), que $\mathrm{Var(X)} = \frac{13120}{5} = 2624$ et $\mathrm{Var(Y)} = \frac{752000}{5} = 150400$, d'où $\sigma_X = \sqrt{2624} \simeq 51,22$ et $\sigma_Y = \sqrt{150400} \simeq 387,81$. Ainsi, dans le cas de l'Exemple 3.1, l'inégalité de Cauchy-Schwarz s'écrit :

$$19460 = |\mathrm{cov(X,Y)}| \leq \sigma_{\mathrm{X}}\sigma_{\mathrm{Y}} \simeq 51,22 \times 387,81 \simeq 19863,63\,.$$

Le coefficient de corrélation linéaire des deux variables X et Y, noté $R(X,Y)$, est défini par

$$R(X,Y) = \frac{\mathrm{cov(X,Y)}}{\sigma_X\sigma_Y}\,. \tag{3.10}$$

Ainsi, dans le cas de l'Exemple 3.1, on a

$$R(X,Y) \simeq \frac{19460}{51,22 \times 387,81} \simeq 0,979.$$

Remarque 3.2. (Propriétés importantes du coefficient de corrélation linéaire)

(*i*) *Il résulte de l'inégalité de Cauchy-Schwarz que $R(X,Y)$ est toujours compris entre -1 et $+1$.*

(*ii*) *Le coefficient directeur a (la pente) de la droite des moindres carrés vérifie :*

$$a = \frac{\mathrm{cov(X,Y)}}{\mathrm{Var(X)}} = \frac{\sigma_Y}{\sigma_X}R(X,Y)\,;$$

par conséquent a et $R(X,Y)$ sont toujours de même signe.

Remarque 3.3. (Interprétation du coefficient de corrélation linéaire)

(*i*) ***Lorsque** $R(X,Y)$ **est voisin de** 0, il y a absence de corrélation entre les variables X et Y ; l'approximation du nuage de points par la droite des moindres carrés est alors illégitime et il faut rejeter l'ajustement linéaire.*

(*ii*) ***Lorsque** $R(X,Y)$ **est voisin de** $+1$, il y a une corrélation directe entre les variable X et Y ; cela signifie grosso modo que Y augmente lorsque X augmente, et que X augmente lorsque Y augmente.*

(*iii*) ***Lorsque** $R(X,Y)$ **est voisin de** -1, il y a une corrélation inverse entre les variables X et Y ; cela signifie grosso modo que Y augmente lorsque X diminue, et X diminue lorsque Y augmente.*

Avant de conclure cette section, il convient de souligner que : *pour que l'ajustement d'un nuage de points par la droite des moindres carrées soit de bonne qualité, il est indispensable que le coefficient de corrélation linéaire soit voisin de* $+1$*, ou encore de* -1 *; cependant cela, à lui tout seul, ne suffit pas pour garantir la bonne qualité de cet ajustement, une étude complémentaire, qui dépasse le cadre de ce cours, s'impose.*

3.2 Liaison entre deux variables qualitatives

3.2.1 Tableau de contingence

Sources (Christophe Benavent, note pédagogique de l'IAE)

218645 établissements industriels et commerciaux de plus de 10 salariés recensés par l'INSEE se répartissent en fonction de leur localisation géographique et de leur taille de la manière suivante

Tableau 1

	TABLEAU DES EFFECTIFS OBSERVES Classe d'effectif des établissements					
REGIONS	10-49	50-199	200-499	500-199	+2000	Total
ILE DE FRANCE	43943	8825	1812	668	101	55349
RHONE ALPES	18055	3453	569	188	15	22280
PROVENCE COTE D'AZUR	12174	1930	284	108	16	14512
NORD-PAS DE CALAIS	10307	2362	487	157	8	13318
PAYS DE LOIRE	8131	1665	312	89	5	10206
BRETAGNE	7841	1609	246	48	5	9749
AQUITAINE	7935	1308	203	67	7	9520
CENTRE	7348	1545	286	85	5	9269
MIDI PYRENEE	6978	1018	179	61	4	8240
LORRAINE	6258	1332	251	86	15	7942
ALSACE	5670	1025	231	82	6	7014
HAUTE-NORMANDIE	5113	1130	209	74	5	6531
PICARDIE	4843	1075	203	88	5	6214
LANGUEDOC ROUSSILLON	5058	795	121	28	4	6006
BOURGOGNE	4772	937	171	60	7	5947
CHAMPAGNE ARDENNE	4088	897	194	56	4	5239
POITOU CHARENTES	4256	732	126	48	2	5164
BASSE-NORMANDIE	3807	790	122	34	5	4758
AUVERGNE	3821	572	87	40	5	4525
FRANCHE COMTE	3152	618	114	26	7	3917
LIMOUSIN	1894	356	63	13	1	2327
CORSE	560	51	4	3	0	618
Total	176004	34025	6274	2109	233	218645

Exemple : On dispose d'une enquête de l'INSEE sur les établissements industriels et commerciaux en 1986 et on cherche s'il existe un lien entre la taille d'un établissement (c'est-à-dire l'effectif, le nombre de salariés d'un établissement) et sa localisation géographique. On considère que la variable « Classe d'Effectif des Etablissements » est qualitative ordinale, et que ses modalités sont les classes : 10-49, 50-199, 200-499, 500-1999 et plus de 2000 salariés. La variable « Régions » est clairement qualitative nominale.

Les 218645 établissements industriels et commerciaux de plus de 10 salariés recensés par l'INSEE se répartissent en fonction de leur localisation et de leur classe d'effectif comme l'indique le Tableau 1. Un tel tableau s'appelle **tableau de contingence** ou encore **tableau croisé**. Le nombre 2362 se trouve sur la ligne Nord-Pas de Calais et sur la colonne 50-199 ; cela signifie que sur les 218645 établissements recensés 2362 se trouvent dans la région NPdC et possèdent chacun un effectif compris entre 50 et 199 salariés.

Le nombre 13318 qui se trouve sur la colonne Total et sur la ligne NPdC signifie que sur les 218645 établissements recensés 13318 se trouvent dans la région NPdC ; ce nombre est donc égal à la somme de tous les autres nombres qui se trouvent sur la ligne NPdC.

Le nombre 34025 qui se trouve sur la ligne Total et sur la colonne 50-199 signifie que sur les 218645 établissements recensés 34025 possèdent un effectif compris entre 50 et 199 salariés ; ce nombre est donc égal à la somme de tous les autres nombres qui se trouvent sur la colonne 50-199.

Le nombre qui se trouvent sur la ligne Total et sur la colonne Total correspond au total des établissements recensés c'est-à-dire 218645 ; ce nombre est donc égal à la somme de tous les autres nombres qui se trouvent sur la ligne Total, il est aussi égal à la somme de tous les autres nombres qui se trouvent sur la colonne Total.

De façon générale, soient Z et T deux variables qualitatives dont les modalités sont respectivement $z_1, \ldots, z_i, \ldots, z_k$ et $t_1, \ldots, t_j, \ldots, t_l$. Les valeurs de ces variables ont été observées sur une population de n individus.

La répartition des effectifs suivant les modalités de Z et de T, se présente sous forme d'un tableau à double entrée, appelé tableau de contingence ou encore tableau croisé :

$Z \setminus T$	t_1	$\cdots$	t_j	$\cdots$	t_l	Total
z_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet l}$	n

L'effectif n_{ij} qui se trouve sur la i-ème ligne et la j-ème colonne du tableau de contingence, est le nombre d'individus qui possèdent à la fois la modalité z_i de la variable Z et la modalité t_j de la variable T. Les effectifs $n_{ij}, i = 1, \ldots, k$ et $j = 1, \ldots, l$ sont appelés **les effectifs croisés observés**.

L'effectif $n_{i\bullet}$ qui se trouve sur la i-ème ligne et la colonne Total est le nombre d'individus qui possèdent la modalité z_i de la variable Z ; on a donc

$$n_{i\bullet} = n_{i1} + n_{i2} + \ldots + n_{il}\,.$$

L'effectif $n_{\bullet j}$ qui se trouve sur la j-ème colonne et la ligne Total est le nombre d'individus qui possèdent la modalité t_j de la variable T ; on a donc

$$n_{\bullet j} = n_{1j} + n_{2j} + \ldots + n_{kj}\,.$$

L'effectif n qui se trouve sur la ligne Total et la colonne Total est le nombre d'individus de la population étudiée ; on a donc

$$n = n_{1\bullet} + n_{2\bullet} + \ldots + n_{k\bullet}$$

et

$$n = n_{\bullet 1} + n_{\bullet 2} + \ldots + n_{\bullet l}\,.$$

La fréquence de la modalité z_i de la variable Z est donnée par :

$$f_{i\bullet} = \frac{n_{i\bullet}}{n}\,.$$

Ainsi sur les 218645 établissements recensés $f_{1\bullet} = \frac{55349}{218645} \simeq 0,253$ (soit $25,3\%$) c'est-à-dire plus d'un établissement sur 4 se trouve dans la région parisienne. Trois autres régions concentrent les établissements, Rhônes-Alpes ($f_{2\bullet} = \frac{22280}{218645} \simeq 0,102$ soit $10,2\%$), Provence Côte d'Azur ($f_{3\bullet} = \frac{14512}{218645} \simeq 0,066$ soit $6,6\%$) et Nord-Pas de Calais ($f_{4\bullet} = \frac{13318}{218645} \simeq 0,061$ sot $6,1\%$).

La fréquence de la modalité t_j de la variable T est donnée par

$$f_{\bullet j} = \frac{n_{\bullet j}}{n}.$$

Dans notre exemple, il ressort de l'étude des fréquences $f_{\bullet j}$, une répartition asymétrique des entreprises en fonction de leurs effectifs ($f_{1\bullet} = \frac{176004}{218645} \simeq 0,805$ soit $80,5\%$) ont moins de 50 salariés et seuls ($f_{5\bullet} = \frac{233}{218645} \simeq 0,001$ soit $0,1\%$) en ont plus de 2000.

La donnée des modalités z_i de la variable Z et des fréquences correspondantes $f_{i\bullet}$ (ou encore des effectifs correspondant $n_{i\bullet}$) est appelée **distribution marginale** de la variable Z.

La donnée des modalités t_j de la variable T et des fréquences correspondantes $f_{\bullet j}$ (ou encore des effectifs correspondant $n_{\bullet j}$) est appelée **distribution marginale** de la variable T.

La fréquence conditionnelle de z_i sachant que $T = t_j$ est donnée par

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

$f_{i|j}$ se lit « f indice i si j ». On a donc $f_{1|j} + f_{2|j} + \ldots + f_{k|j} = \frac{n_{\bullet j}}{n_{\bullet j}} = 1$.

Le tableau suivant est appelé **tableau des profils colonnes**

$Z \setminus T$	t_1	$\cdots$	t_j	$\cdots$	t_l	Distribution marginale de Z
z_1	$f_{1\|1}$	$\cdots$	$f_{1\|j}$	$\cdots$	$f_{1\|l}$	$f_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	$f_{i\|1}$	$\cdots$	$f_{i\|j}$	$\cdots$	$f_{i\|l}$	$f_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	$f_{k\|1}$	$\cdots$	$f_{k\|j}$	$\cdots$	$f_{k\|l}$	$f_{k\bullet}$
Total	1	$\cdots$	1	$\cdots$	1	1

$f_{i|j}$ se trouve sur la i-ème ligne et la j-ème colonne du tableau. De façon général, ce tableau permet de comparer les profils colonnes (les colonnes) au profil marginal colonne (dernière colonne) et de les comparer entre eux. Dans le cas de notre exemple, au moyen du Tableau 3 (voir un peu plus loin), on peut capter pour chaque classe d'effectif la répartition géographique des entreprises correspondantes. On se rend compte notamment que la concentration dans la région Île de France des grandes entreprises est nettement plus forte que celle des petites.

La fréquence conditionnelle de t_j sachant que $Z = z_i$ est donnée par

$$f_{j|i} = \frac{n_{ij}}{n_{i\bullet}}$$

On a donc

$$f_{1|i} + f_{2|i} + \ldots + f_{l|i} = 1.$$

$f_{j|i}$ se lit « f indice j si i ».

Le tableau suivant est appelé **tableau des profils lignes**

$Z \setminus T$	t_1	$\cdots$	t_j	$\cdots$	t_l	Total
z_1	$f_{1\|1}$	$\cdots$	$f_{j\|1}$	$\cdots$	$f_{l\|1}$	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	$f_{1\|i}$	$\cdots$	$f_{j\|i}$	$\cdots$	$f_{l\|i}$	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	$f_{1\|k}$	$\cdots$	$f_{j\|k}$	$\cdots$	$f_{k\|l}$	1
Distribution marginale de T	$f_{\bullet 1}$	$\cdots$	$f_{\bullet j}$	$\cdots$	$f_{\bullet l}$	1

$f_{j|i}$ se trouve sur la i-ème ligne et la j-ème colonne du tableau. De façon générale, ce tableau permet de comparer les profils lignes (les lignes) au profil marginal ligne (dernière ligne) et de les comparer entre eux. Dans le cas de notre exemple, le Tableau 2 (voir un peu plus loin) donne pour chaque région la répartition des entreprises par classe d'effectif. On se rend compte qu'il n'y a guère de différence entre les régions. Dans chaque région, les petites entreprises sont largement majoritaires alors que les grandes sont largement minoritaires.

(2)

Tableau 2 (Profils lignes)

PROFILS LIGNES						
	Classe d'effectif des établissements					
REGIONS	10-49	50-199	200-499	500-199	+2000	Fréquence
ILE DE FRANCE	79,4%	15,9%	3,3%	1,2%	0,2%	100,0%
RHONE ALPES	81,0%	15,5%	2,6%	0,8%	0,1%	100,0%
PROVENCE COTE D'AZ	83,9%	13,3%	2,0%	0,7%	0,1%	100,0%
NORD-PAS DE CALAIS	77,4%	17,7%	3,7%	1,2%	0,1%	100,0%
PAYS DE LOIRE	79,7%	16,3%	3,1%	0,9%	0,0%	100,0%
BRETAGNE	80,4%	16,5%	2,5%	0,5%	0,1%	100,0%
AQUITAINE	83,4%	13,7%	2,1%	0,7%	0,1%	100,0%
CENTRE	79,3%	16,7%	3,1%	0,9%	0,1%	100,0%
MIDI PYRENEE	84,7%	12,4%	2,2%	0,7%	0,0%	100,0%
LORRAINE	78,8%	16,8%	3,2%	1,1%	0,2%	100,0%
ALSACE	80,8%	14,6%	3,3%	1,2%	0,1%	100,0%
HAUTE-NORMANDIE	78,3%	17,3%	3,2%	1,1%	0,1%	100,0%
PICARDIE	77,9%	17,3%	3,3%	1,4%	0,1%	100,0%
LANGUEDOC ROUSSIL	84,2%	13,2%	2,0%	0,5%	0,1%	100,0%
BOURGOGNE	80,2%	15,8%	2,9%	1,0%	0,1%	100,0%
CHAMPAGNE ARDENN	78,0%	17,1%	3,7%	1,1%	0,1%	100,0%
POITOU CHARENTES	82,4%	14,2%	2,4%	0,9%	0,0%	100,0%
BASSE-NORMANDIE	80,0%	16,6%	2,6%	0,7%	0,1%	100,0%
AUVERGNE	84,4%	12,6%	1,9%	0,9%	0,1%	100,0%
FRANCHE COMTE	80,5%	15,8%	2,9%	0,7%	0,2%	100,0%
LIMOUSIN	81,4%	15,3%	2,7%	0,6%	0,0%	100,0%
CORSE	90,6%	8,3%	0,6%	0,5%	0,0%	100,0%
Fréquence	80,5%	15,6%	2,9%	1,0%	0,1%	100,0%

Tableau 3 (Profils colonnes)

PROFILS COLONNES						
	Classe d'effectif des établissements					
REGIONS	10-49	50-199	200-499	500-1999	+2000	Fréquence
ILE DE FRANCE	25,0%	25,9%	28,9%	31,7%	43,3%	25,3%
RHONE ALPES	10,3%	10,1%	9,1%	8,9%	6,4%	10,2%
PROVENCE COTE D'AZ	6,9%	5,7%	4,5%	5,1%	6,9%	6,6%
NORD-PAS DE CALAIS	5,9%	6,9%	7,8%	7,4%	3,4%	6,1%
PAYS DE LOIRE	4,6%	4,9%	5,0%	4,2%	2,1%	4,7%
BRETAGNE	4,5%	4,7%	3,9%	2,3%	2,1%	4,5%
AQUITAINE	4,5%	3,8%	3,2%	3,2%	3,0%	4,4%
CENTRE	4,2%	4,5%	4,6%	4,0%	2,1%	4,2%
MIDI PYRENEE	4,0%	3,0%	2,9%	2,9%	1,7%	3,8%
LORRAINE	3,6%	3,9%	4,0%	4,1%	6,4%	3,6%
ALSACE	3,2%	3,0%	3,7%	3,9%	2,6%	3,2%
HAUTE-NORMANDIE	2,9%	3,3%	3,3%	3,5%	2,1%	3,0%
PICARDIE	2,8%	3,2%	3,2%	4,2%	2,1%	2,8%
LANGUEDOC ROUSSIL	2,9%	2,3%	1,9%	1,3%	1,7%	2,7%
BOURGOGNE	2,7%	2,8%	2,7%	2,8%	3,0%	2,7%
CHAMPAGNE ARDENN	2,3%	2,6%	3,1%	2,7%	1,7%	2,4%
POITOU CHARENTES	2,4%	2,2%	2,0%	2,3%	0,9%	2,4%
BASSE-NORMANDIE	2,2%	2,3%	1,9%	1,6%	2,1%	2,2%
AUVERGNE	2,2%	1,7%	1,4%	1,9%	2,1%	2,1%
FRANCHE COMTE	1,8%	1,8%	1,8%	1,2%	3,0%	1,8%
LIMOUSIN	1,1%	1,0%	1,0%	0,6%	0,4%	1,1%
CORSE	0,3%	0,1%	0,1%	0,1%	0,0%	0,3%
Fréquence	100,0%	100,0%	100,0%	100,0%	99,6%	

3.2.2 Test d'une éventuelle liaison (test du χ^2 « chi 2 »)

Il n'y a pas de liaison entre les variables Z et T, lorsque tous les profils colonnes sont identiques au profil marginal colonne. Autrement dit, pour tout $i = 1, \ldots, k$ et tout $j = 1, \ldots, l$ $f_{i|j}$ la fréquence conditionnelle de z_i sachant $T = t_j$ est égale à $f_{i,\bullet}$, la fréquence de z_i. Cette égalité est équivalente à l'égalité

$$\frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{n}$$

ou encore à l'égalité

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

Il n'y a également pas de liaisons entre les variables Z et T, lorsque tous les profils lignes sont identiques au profil marginal ligne. Autrement dit, pour tout $i = 1, \ldots, k$ et tout $j = 1, \ldots, l$ $f_{j|i}$ la fréquence conditionnelle de t_j sachant $Z = z_i$ est égale à $f_{\bullet j}$, la fréquence de t_j. Cette égalité est équivalente à l'égalité

$$\frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n}$$

ou encore à l'égalité

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

qu'on a déjà vue plus haut.

Dans le cas de notre exemple, les profils colonnes ne sont pas identiques au profil marginal colonne. Cela signifie qu'il existe une liaison entre la variable « Régions » et la variable « Classe d'Effectif des Etablissements ». Pour tout $i = 1, \ldots, k$ et tout $j = 1, \ldots, l$ on pose

$$n^*_{ij} = \frac{n_{i\bullet}\, n_{\bullet j}}{n}.$$

Les quantités n^*_{ij} sont appelées **les effectifs (croisés) théoriques** ; il s'agit en fait des effectifs qu'on aurait obtenus **s'il n'y avait pas eu de liaison** entre les variables Z et T. Par exemple, l'effectif théorique croisé Ile de France, Classe d'effectif 10-49 vaut $n^*_{11} = \frac{55349 \times 176004}{218645} \simeq 44555$ et l'effectif théorique croisé Nord-Pas de Calais, Classe d'effectif 200-499 vaut $n^*_{43} = \frac{13318 \times 6274}{218645} \simeq 382$.

Le tableau suivant est appelé tableau des effectifs théoriques

$Z \setminus T$	t_1	$\cdots$	t_j	$\cdots$	t_l	Total
z_1	n^*_{11}	$\cdots$	n^*_{1j}	$\cdots$	n^*_{1l}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	n^*_{i1}	$\cdots$	n^*_{ij}	$\cdots$	n^*_{il}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	n^*_{k1}	$\cdots$	n^*_{kj}	$\cdots$	n^*_{kl}	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$\cdots$	$n^*_{\bullet j}$	$\cdots$	$n^*_{\bullet l}$	n

n^*_{ij} se trouve sur la i-ème ligne et la j-ème colonne du tableau. Plus la différence entre le tableau de contingence (le tableau des effectifs croisés observés) et le tableau des effectifs théoriques est grande, plus grande est la probabilité d'existence d'une liaison significative entre les variables Z et T. Pour formaliser cette idée, il convient d'introduire la quantité suivante appelée distance du χ^2 (« chi 2 »).

$$\begin{aligned}\chi^2 = &\frac{(n_{11}-n_{11}^*)^2}{n_{11}^*} + \frac{(n_{12}-n_{12}^*)^2}{n_{12}^*} + \ldots + \frac{(n_{1l}-n_{1l}^*)^2}{n_{11}^*} \\ &+ \frac{(n_{21}-n_{21}^*)^2}{n_{21}^*} + \frac{(n_{22}-n_{22}^*)^2}{n_{22}^*} + \ldots + \frac{(n_{2l}-n_{2l}^*)^2}{n_{21}^*} \\ &\vdots \\ &+ \frac{(n_{k1}-n_{k1}^*)^2}{n_{k1}^*} + \frac{(n_{k2}-n_{k2}^*)^2}{n_{k2}^*} + \ldots + \frac{(n_{kl}-n_{kl}^*)^2}{n_{kl}^*}\end{aligned}$$

La distance du χ^2 mesure l'écart entre le tableau de contingence et la tableau des effectifs théoriques. Plus elle est grande, plus cet écart est important. Lorsqu'il n'y a pas de liaisons en tre Z et T, comme on l'a vu précédemment, les effectifs croisés observés sont égaux aux effectifs théoriques (pour tout $i = 1, \ldots, k$ et pour tout $j = 1 \ldots l$ $n_{ij} = n_{ij}^*$) et cela est équivalent à $\chi^2 = 0$.

Les χ^2 partiels sont les quantités χ_{ij}^2 définies pour tout $i = 1, \ldots, k$ et tout $l = 1, \ldots, l$ par

$$\chi_{ij}^2 = \frac{(n_{ij}-n_{ij}^*)^2}{n_{ij}^*}.$$

χ_{ij}^2 mesure le carré de l'écart entre l'effectif observé n_{ij} et l'effectif théorique n_{ij}^* relativement à l'effectif théorique n_{ij}^*. Par exemple, le χ^2 partiel Île de France, Classe d'effectif 10-49 vaut $\chi_{11}^2 = \frac{(43943-44555)^2}{44555} \simeq 8,4$ et le χ^2 partiel Nord-Pas de Calais, Classe d'effectif 200-499 vaut $\chi_{43}^2 = \frac{(487-382)^2}{382} \simeq 28,86$.

Lorsque pour un certain i_0 et un certain j_0 l'effectif observé $n_{i_0 j_0}$ est plus grand que l'effectif théorique $n_{i_0 j_0}^*$ ($n_{i_0 j_0} > n_{i_0 j_0}^*$) on dit qu'il y a attraction entre la modalité z_{i_0} de la variable Z et la modalité t_{j_0} de la variable T. Lorsque pour un certain i_1, et un certain j_1, l'effectif observé $n_{i_1 j_1}$ est plus petit que l'effectif théorique $n_{i_1 j_1}^*$ ($n_{i_1 j_1} < n_{i_1 j_1}^*$) on dit qu'il y a répulsion entre la modalité z_{i_1} de la variable Z et la modalité t_{j_1} de la variable T.

Dans le cas de notre exemple, il y a répulsion entre la modalité Île de France de la variable Région et la modalité 10-49 de la variable classe d'effectif (car $n_{11} = 43943 < 44555 = n_{11}^*$). En revanche, il y a attraction entre la modalité Nord-Pas de Calais de la classe Région et la modalité 200-499 de la variable classe d'effectif (car $n_{43} = 487 > 382 = n_{43}^*$).

Il résulte de ce qui précède que la distance du χ^2 est égale à la somme de tous les χ^2 partiels

$$\begin{aligned}\chi^2 = &\chi_{11}^2 + \chi_{12}^2 + \ldots + \chi_{1l}^2 \\ &+ \chi_{21}^2 + \chi_{22}^2 + \ldots + \chi_{2l}^2 \\ &\vdots \\ &+ \chi_{k1}^2 + \chi_{k2}^2 + \ldots + \chi_{kl}^2\end{aligned}$$

Tableau 4 (effectifs théoriques)

EFFECTIFS THEORIQUES						
	Classe d'effectif des établissements					Profil
REGIONS	10-49	50-199	200-499	500-199	+2000	colonne
ILE DE FRANCE	44555	8613	1588	534	59	55349
RHONE ALPES	17935	3467	639	215	24	22280
PROVENCE COTE D'AZ	11682	2258	416	140	15	14512
NORD-PAS DE CALAIS	10721	2073	382	128	14	13318
PAYS DE LOIRE	8216	1588	293	98	11	10206
BRETAGNE	7848	1517	280	94	10	9749
AQUITAINE	7663	1481	273	92	10	9520
CENTRE	7461	1442	266	89	10	9269
MIDI PYRENEE	6633	1282	236	79	9	8240
LORRAINE	6393	1236	228	77	8	7942
ALSACE	5646	1092	201	68	7	7014
HAUTE-NORMANDIE	5257	1016	187	63	7	6531
PICARDIE	5002	967	178	60	7	6214
LANGUEDOC ROUSSIL	4835	935	172	58	6	6006
BOURGOGNE	4787	925	171	57	6	5947
CHAMPAGNE ARDENN	4217	815	150	51	6	5239
POITOU CHARENTES	4157	804	148	50	6	5164
BASSE-NORMANDIE	3830	740	137	46	5	4758
AUVERGNE	3643	704	130	44	5	4525
FRANCHE COMTE	3153	610	112	38	4	3917
LIMOUSIN	1873	362	67	22	2	2327
CORSE	497	96	18	6	1	618
Profil ligne	176004	34025	6274	2109	233	218645

Tableau 5 (des X^2 partiels)

TABLEAU DES χ^2						
	Classe d'effectif des établissements					Total
REGIONS	10-49	50-199	200-499	500-199	+2000	c2
ILE DE FRANCE	8,40	5,20	31,53	33,69	29,93	108,75
RHONE ALPES	0,80	0,06	7,74	3,37	3,22	15,19
PROVENCE COTE D'AZ	20,74	47,73	42,11	7,31	0,02	117,90
NORD-PAS DE CALAIS	15,96	40,43	28,76	6,34	2,70	94,20
PAYS DE LOIRE	0,87	3,71	1,25	0,91	3,17	9,91
BRETAGNE	0,01	5,57	4,07	22,54	2,80	34,97
AQUITAINE	9,63	20,31	18,03	6,71	0,97	55,66
CENTRE	1,72	7,30	1,51	0,22	2,41	13,15
MIDI PYRENEE	17,94	54,47	13,96	4,30	2,60	93,27
LORRAINE	2,86	7,47	2,34	1,15	5,05	18,87
ALSACE	0,10	4,05	4,39	3,04	0,29	11,88
HAUTE-NORMANDIE	3,96	12,71	2,49	1,92	0,55	21,63
PICARDIE	5,06	12,06	3,42	13,14	0,40	34,08
LANGUEDOC ROUSSIL	10,31	20,86	15,30	15,47	0,90	62,84
BOURGOGNE	0,05	0,14	0,00	0,12	0,07	0,38
CHAMPAGNE ARDENN	3,96	8,19	12,68	0,59	0,45	25,88
POITOU CHARENTES	2,36	6,38	3,32	0,07	2,23	14,36
BASSE-NORMANDIE	0,14	3,32	1,55	3,08	0,00	8,09
AUVERGNE	8,75	24,81	14,14	0,30	0,01	48,00
FRANCHE COMTE	0,00	0,12	0,02	3,67	1,91	5,73
LIMOUSIN	0,23	0,10	0,21	3,97	0,88	5,41
CORSE	7,86	21,22	10,64	1,47	0,66	41,84
Total c2	121,71	306,22	219,44	133,38	61,23	842,0

Le tableau des χ^2 partiels est le tableau suivant :

$Z \setminus T$	t_1	$\cdots$	t_j	$\cdots$	t_l	Total
z_1	χ^2_{11}	$\cdots$	χ^2_{1j}	$\cdots$	χ^2_{1l}	$\chi^2_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	χ^2_{i1}	$\cdots$	χ^2_{ij}	$\cdots$	χ^2_{il}	$\chi^2_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	χ^2_{k1}	$\cdots$	χ^2_{kj}	$\cdots$	χ^2_{kl}	$\chi^2_{k\bullet}$
Total	$\chi^2_{\bullet 1}$	$\cdots$	$\chi^2_{\bullet j}$	$\cdots$	$\chi^2_{\bullet l}$	χ^2

χ^2_{ij} se trouve sur la i-ème ligne et la j-ème colonne. Pour tout $i = 1, \ldots, k$, $\chi^2 i\bullet$ désigne la somme des χ^2 partiels se trouvant sur la i-ème ligne du tableau :

$$\chi^2_{i\bullet} = \chi^2_{i1} + \chi^2_{i2} + \ldots + \chi^2_{il}.$$

Pour tout $j = 1, \ldots, l$, $\chi^2_{\bullet j}$ désigne la somme des χ^2 partiels se trouvant sur la j-ème ligne du tableau :

$$\chi^2_{\bullet j} = \chi^2_{1j} + \chi^2_{2j} + \ldots + \chi^2_{kj}.$$

D'après ce qui précède, on a

$$\chi^2 = \chi^2_{1\bullet} + \ldots + \chi^2_{k\bullet} = \chi^2_{\bullet 1} + \ldots + \chi^2_{\bullet l}.$$

Pour calculer χ^2, on commence, par calculer, pour chaque ligne du tableau, la somme des nombres s'y trouvant et on reporte les résultats dans la colonne Total. Ensuite, on calcule la somme de nombres se trouvant dans la colonne Total.

On peut également, pour calculer χ^2, commencer par calculer pour chaque colonne du tableau, la somme des nombres s'y trouvant, reporter le résultat dans la ligne Total puis faire la somme des nombres s'y trouvant dans la ligne Total.

De façon générale, lorsque la valeur de la distance du χ^2 est plus grande qu'un certain seuil (la méthode permettant de d{eterminer ce seuil dépasse le cadre de ce cours), on accepte l'hypothèse d'existence d'une liaison entre le variables Z et T. Dans le cas de notre exemple, on trouve que $\chi^2 = 842$ et cela nous amène à accepter l'hypothèse de l'existence d'un lien entre la taille (l'effectif) d'un établissement industriel ou commercial et sa localisation géographique.

En examinant plus attentivement le tableau des χ^2 partiels, on s'aperçoit que dans certaines cases les valeurs sont sensiblement plus élevées qu'ailleurs. On est tenté de considérer que ce sont les cases les plus importantes, que ce sont ces situations qu'il faut interpréter. C'est notamment le cas des cases (Midi-Pyrénées, 50-199) ; (PACA ; 50-199) ; (PACA, 200-499) ; (NPdC, 50-199) ; (IdF, 500-1999) ; (IdF, 200-499) ...

Pour pouvoir identifier de façon précise les cases $(\cdot, \cdot)$ les plus importantes du tableau des χ^2 partiels, on est amené à considérer, pour tout $i = 1, \ldots, k$ et tout $j = 1, \ldots, l$ la quantité

$$\mathrm{CTR_{ij}} = \frac{\chi^2_{\mathrm{ij}}}{\chi^2} \times 100$$

Cette quantité est appelée **contribution relative de la case** (i, j) à la valeur de χ^2. Dans le cas de la case (Midi-Pyrénées, 50-199), on trouve que $\mathrm{CTR}_{92} = \frac{54,47}{842} \times 100 = 6,47\%$

Le tableau des contributions est le tableau suivant :

$Z \setminus T$	t_1	$\cdots$	t_j	$\cdots$	t_l	Total
z_1	CTR_{11}	$\cdots$	$\mathrm{CTR}_{1\mathrm{j}}$	$\cdots$	$\mathrm{CTR}_{1\mathrm{l}}$	$\mathrm{CTR}_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	$\mathrm{CTR}_{\mathrm{i}1}$	$\cdots$	$\mathrm{CTR}_{\mathrm{ij}}$	$\cdots$	$\mathrm{CTR}_{\mathrm{il}}$	$\mathrm{CTR}_{\mathrm{i}\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	$\mathrm{CTR}_{\mathrm{k}1}$	$\cdots$	$\mathrm{CTR}_{\mathrm{kj}}$	$\cdots$	$\mathrm{CTR}_{\mathrm{kl}}$	$\mathrm{CTR}_{\mathrm{k}\bullet}$
Total	$\mathrm{CTR}_{\bullet 1}$	$\cdots$	$\mathrm{CTR}_{\bullet\mathrm{j}}$	$\cdots$	$\mathrm{CTR}_{\bullet\mathrm{l}}$	100%

(4)

Tableau 6 (tableau des contributions)

TABLEAU DES CONTRIBUTIONS						
	Classe d'effectif des établissements					Contributi
REGIONS	10-49	50-199	200-499	500-1999	+2000	Ligne
ILE DE FRANCE	1,00%	0,62%	3,74%	4,00%	3,55%	12,9%
RHONE ALPES	0,10%	0,01%	0,92%	0,40%	0,38%	1,8%
PROVENCE COTE D'AZ	2,46%	5,67%	5,00%	0,87%	0,00%	14,0%
NORD-PAS DE CALAIS	1,90%	4,80%	3,42%	0,75%	0,32%	11,2%
PAYS DE LOIRE	0,10%	0,44%	0,15%	0,11%	0,38%	1,2%
BRETAGNE	0,00%	0,66%	0,48%	2,68%	0,33%	4,2%
AQUITAINE	1,14%	2,41%	2,14%	0,80%	0,12%	6,6%
CENTRE	0,20%	0,87%	0,18%	0,03%	0,29%	1,6%
MIDI PYRENEE	2,13%	6,47%	1,66%	0,51%	0,31%	11,1%
LORRAINE	0,34%	0,89%	0,28%	0,14%	0,60%	2,2%
ALSACE	0,01%	0,48%	0,52%	0,36%	0,03%	1,4%
HAUTE-NORMANDIE	0,47%	1,51%	0,30%	0,23%	0,07%	2,6%
PICARDIE	0,60%	1,43%	0,41%	1,56%	0,05%	4,0%
LANGUEDOC ROUSSIL	1,23%	2,48%	1,82%	1,84%	0,11%	7,5%
BOURGOGNE	0,01%	0,02%	0,00%	0,01%	0,01%	0,0%
CHAMPAGNE ARDENNI	0,47%	0,97%	1,51%	0,07%	0,05%	3,1%
POITOU CHARENTES	0,28%	0,76%	0,39%	0,01%	0,26%	1,7%
BASSE-NORMANDIE	0,02%	0,39%	0,18%	0,37%	0,00%	1,0%
AUVERGNE	1,04%	2,95%	1,68%	0,04%	0,00%	5,7%
FRANCHE COMTE	0,00%	0,01%	0,00%	0,44%	0,23%	0,7%
LIMOUSIN	0,03%	0,01%	0,03%	0,47%	0,10%	0,6%
CORSE	0,93%	2,52%	1,26%	0,17%	0,08%	5,0%
Contribution colonne	14,5%	36,4%	26,1%	15,8%	7,3%	100,00%