Statistique Descriptive et Calcul de Probabilités

Antoine Ayache & Julien Hamonier

Université de Lille

Table des matières

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 29,1 k€ (signalons que 1 k€=mille euros), 29,12 k€, ..., 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 un individu 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.

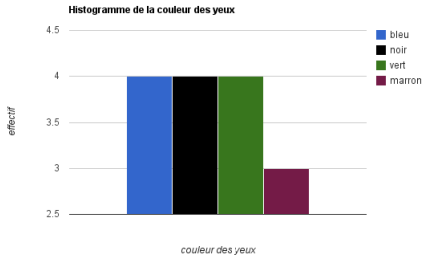
Remarque 2.1. *De façon générale, un individu appartient à une et une seule modalité d'une variable qualitative. Bien souvent, parmi les modalités d'une variable qualitative figure une modalité* ***Autres*** *(non répondants ou bien valeurs manquantes ou quelque chose dans ce genre-là) dans laquelle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

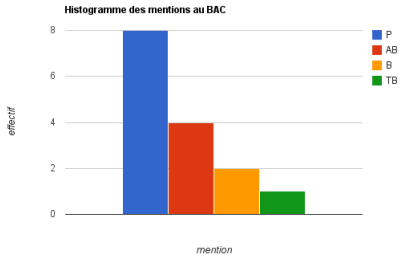
Étudions l'exemple de la variable **Couleurs des Yeux**. On commence d'abord par compter le nombre d'individus appartenant à chacune des modalités de cette variables : $n_B = 4$ individus ont les yeux bleus, $n_M = 3$ ont les yeux marrons, $n_N = 4$ ont les yeux noirs et $n_V = 4$ ont les yeux verts ; 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,

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

On a enfin

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

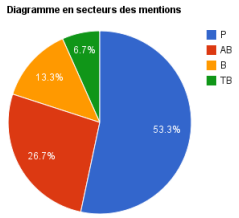
$$(\text{somme de tous les pourcentages des modalités d'une variable qualitative}) = 100.$$

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 :

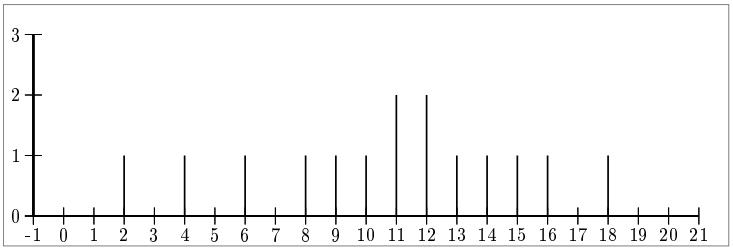
$$f_k = \frac{n_k}{N},$$

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 :

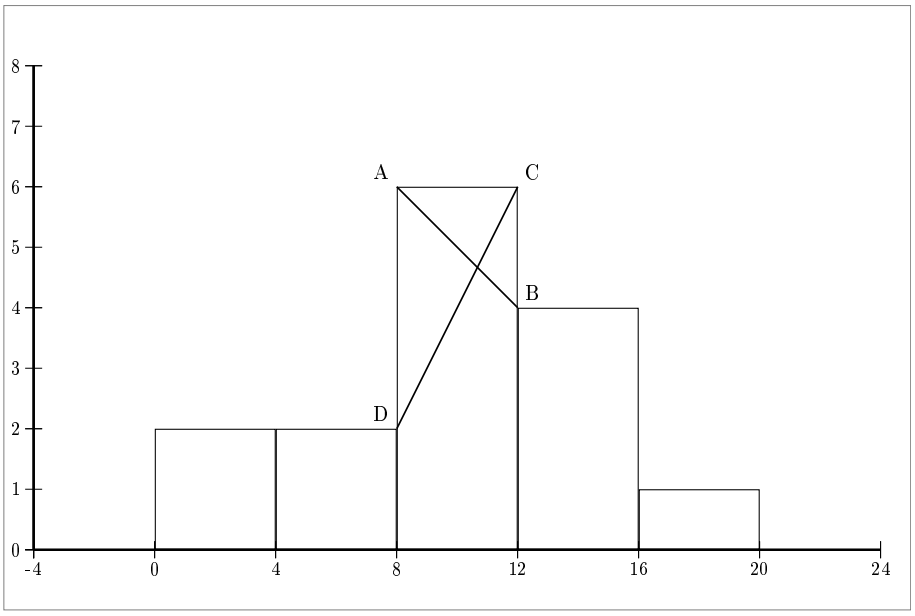


Néanmoins cette forme se prête difficilement à l'interprétation. Pour y remédier, il faut créer des **classes** de notes (nombre d'individus ayant obtenu des notes comprises entre 0 et 4, entre 4 et 8, ...) ; cette approche nous permet d'obtenir une variable dite **classée**. Il faut effectuer le **bornage** 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 Effectifs de 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*, c'est- à-dire la loi de probabilité sous-jacente. 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.*

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.

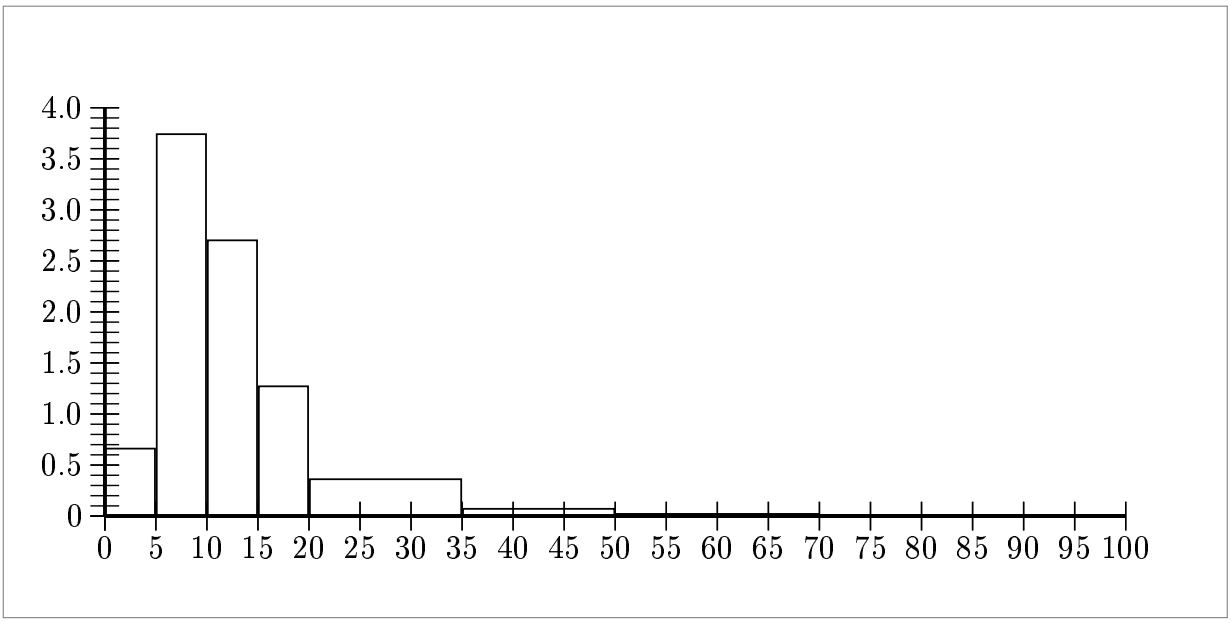
Lorsque les classes ont toutes, la même amplitude, les hauteurs des rectangles sont proportionnelles à leurs surfaces ; par conséquent les hauteurs des rectangles sont proportionnelles aux fréquences et aux effectifs. **Dans le cas où les classes sont d'amplitudes inégales, la hauteur du rectangle correspondant 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)**. Signalons que les deux quantités h_i et H_i sont parfois appelées *densités d'observation de la i-ème classe.*

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 × 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)



La forme de l'histogramme ressemble plus ou moins à celle d'une « cloche », ce qui laisse à penser que la loi de probabilité sous-jacente est *la loi normale*. Cette loi est extrêmement importante en Probabilités et en Statistique ; elle sera étudiée dans le prochain chapitre.

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 (c'est-à-dire dont la densité d'observation) 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.

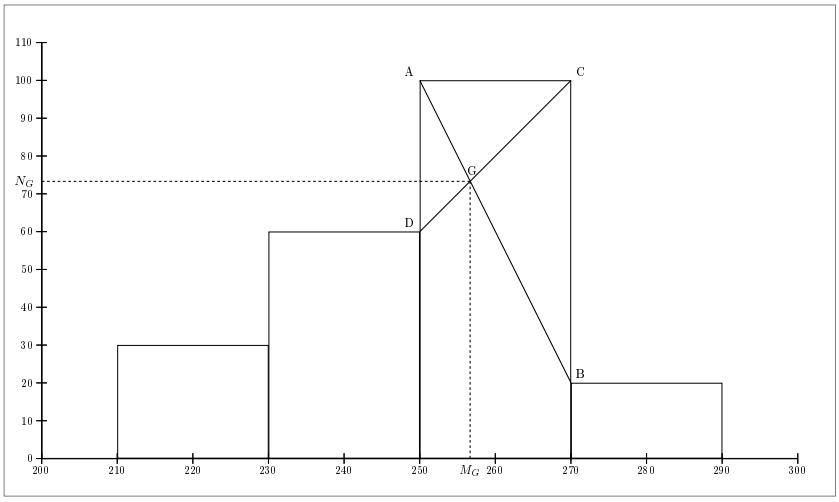
Lorsqu'on souhaite être plus précis, on peut déterminer à l'intérieur de la classe modale **la valeur exacte du mode** ; l'exemple suivant permet de comprendre la démarche à suivre.

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 autres méthodes, 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 l'effectif des consommateurs 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*

$$\begin{cases} 250a + b = 100 \\ 270a + b = 20 \end{cases} \Leftrightarrow \begin{cases} 250a + b = 100 \\ -20a = 80 \end{cases} \Leftrightarrow \begin{cases} b = 100 - 250 \times (-4) = 1100 \\ a = -4 \end{cases}$$

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*

$$\begin{cases} 250a + b = 60 \\ 270a + b = 100 \end{cases}$$

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*

$$\begin{cases} 250a + b = 60 \\ 270a + b = 100 \end{cases} \Leftrightarrow \begin{cases} 250a + b = 60 \\ 20a = 40 \end{cases} \Leftrightarrow \begin{cases} b = 60 - 250 \times 2 = -440 \\ a = 2 \end{cases}$$

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*

$$\begin{cases} N_G = -4M_G + 1100 \\ N_G = 2M_G - 440 \end{cases}$$

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

$$\begin{cases} N_G = -4M_G + 1100 \\ N_G = 2M_G - 440 \end{cases} \Leftrightarrow \begin{cases} -6M_G + 1540 = 0 \\ N_G = 2M_G - 440 \end{cases} \Leftrightarrow \begin{cases} M_G = \frac{770}{3} \simeq 256,66 \\ \\ N_G = 2 \times \frac{770}{3} - 440 \simeq 73,33 \end{cases}$$

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)

Nous allons décrire une méthode [1] (parmi d'autres) qui permet d'obtenir la médiane. On attribue d'abord à chacun des individus un rang en partant de l'individu (ou des individus) pour

1. Cette méthode (ainsi que d'autres méthodes) doit être utilisée avec prudence, notamment lorsque certaines valeurs de la variable statistique discrète se répètent ; en effet, dans ce cas particulier, cette méthode pourrait alors conduire à un résultat qui n'est guère satisfaisant.

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. **La médiane de la variable statistique est alors la moyenne de ses valeurs qui correspondent aux profondeurs maximales.**

Pour illustrer cette méthode étudions deux exemples concrets :

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*	$\boxed{7}$
Agnès	*10*	*10*	*6*	*6*
Nadine	*9*	*11*	*5*	*5*
Étienne	*16*	*2*	*14*	*2*
Gilles	*14*	*4*	*12*	*4*
Aurélie	*11*	*8*	*7*	$\boxed{7}$
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

$$M_e = \frac{11+11}{2} = 11\,.$$

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*	$\boxed{6}$
Jean	*8*	*12*	*3*	*3*
Stéphane	*13*	*5*	*10*	*5*
Charles	*11*	*8*	*6*	$\boxed{6}$
Agnès	*10*	*10*	*5*	*5*
Nadine	*9*	*11*	*4*	*4*
Étienne	*16*	*2*	*13*	*2*
Gilles	*14*	*4*	*11*	*4*
Aurélie	*11*	*8*	*6*	$\boxed{6}$
Stéphanie	*15*	*3*	*12*	*3*
Marie-Claude	*4*	*14*	*1*	*1*
Anne	*18*	*1*	*14*	*1*
Christophe	*12*	*6*	*8*	$\boxed{6}$
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 $[x_0, x_1],]x_1, x_2], \ldots,]x_{k-1}, x_k]$. Les effectifs correspondant à ces classes sont notés « n_1 », « n_2 », ..., « 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

$$N_2 = n_1 + n_2 = N_1 + n_2.$$

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

$$N_3 = n_1 + n_2 + n_3 = N_2 + n_3.$$

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

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

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.*

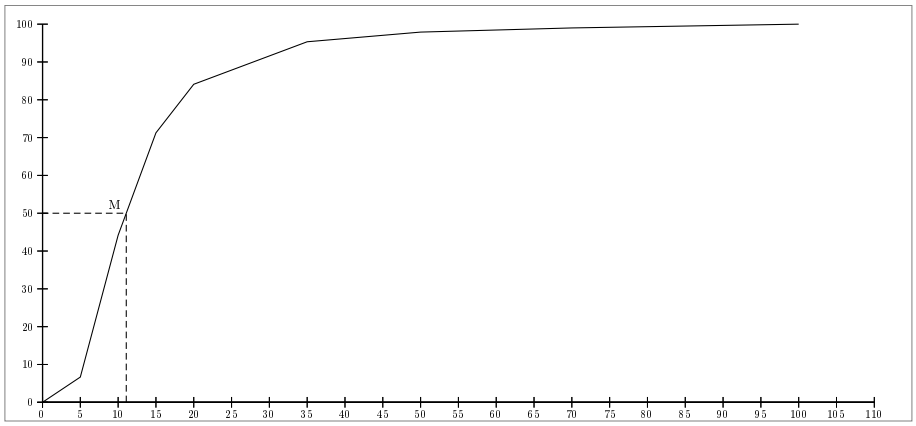
Remarque 2.5. *De façon générale, la médiane notée par M_e d'une variable statistique continue X de fonction cumulative F est telle que*

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

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 équation est vérifiée par $(M_e\,;50)$, les coordonnées du point M, on obtient :

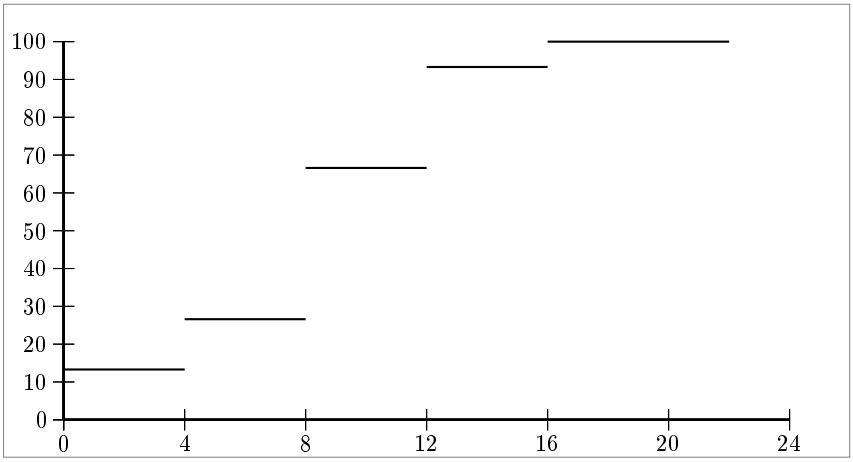
$$50 = 5,416M_e - 9,98 \Leftrightarrow M_e = \frac{50 + 9,98}{5,416} \simeq 11,075 \text{ milliers de Francs.}$$

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 » dont il est question 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, c'est-à-dire

$$IIQ = Q_3 - Q_1.$$

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.

Exercice 2.3. *Déterminer Q_1, Q_3 et IIQ dans le cas de la variable continue « Revenus des Contribuables ».*

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*[2] *a un impact plus significatif sur la 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 valeur 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 (dans l'Exemple 2.3 on a $N = 15$ et $K = 13$; dans l'Exemple 2.4 on a $N = 14$ et $K = 12$).

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* $[y_0, y_1]$, $]y_1, y_2]$, ... , $]y_{k-1}, y_k]$;
- *Y est une variable discrète classée dont les classes sont* $[y_0, y_1]$, $]y_1, y_2]$, ... , $]y_{k-1}, y_k]$.

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 2.4. (*a*) *Calculer la moyenne arithmétique de la variable continue « Revenu des Contribuables »(on trouve* $14292,5$ *Francs).* (*b*) *Calculer la moyenne arithmétique de la variable classée « Note à l'Examen de Statistique » dont il est question dans l'Exercice 2.2 (on trouve* $10,008$*).*

b) Moyenne quadratique

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

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

2. C'est la seule différence entre l'Exemple 2.3 et l'Exemple 2.4.

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

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

Exercice 2.5. *Calculer la moyenne quadratique de la variable « Note à l'Examen de Statistique » dont il est question dans l'Exemple 2.3, et de celle dont il est question dans l'Exemple 2.4.*

c) Moyenne harmonique

Attention : on ne peut définir cette moyenne que lorsque les observations $x_1, \ldots, x_N$ sont tous des nombres réels strictement positifs. Si tel est le cas, la moyenne harmonique de ces observations est notée par m_{-1} et elle est définie par :

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

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 deux 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 2.6. *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 strictement positifs. Si tel est le cas, la moyenne géométrique de ces observations est notée par M_g, et elle est définie par :

$$\begin{aligned} M_g &= \left(x_1 \times x_2 \times \ldots \times x_N\right)^{1/N} = \exp\left(\frac{1}{N} \ln\left(x_1 \times x_2 \times \ldots \times x_N\right)\right) \\ &= x_1^{f_1} \times \ldots \times x_K^{f_K}, \end{aligned}$$

où exp désigne la fonction exponentielle et ln la fonction logarithme népérien.

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 strictement positifs, on a alors :*

$$\min_{1 \leq i \leq N} x_i \leq m_{-1} \leq M_g \leq \overline{x} \leq m_2 \leq \max_{1 \leq i \leq N} x_i \,.$$

Autrement dit, on a :

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

2.4 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 réécrire 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 elle aussi le calcul de la variance, est :

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

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 précédent) ; ainsi en utilisant la formule (2.1), on obtient :*

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

Présentons maintenant **la seconde méthode**. *La somme des carrés des observations vaut* 2001 *(voir la dernière colonne du tableau précédent) ; 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 des 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* $\sqrt{18,27} \simeq 4,27$.

Une question : *A votre avis, que vaut la somme des nombres qui se trouvent dans la troisième colonne du tabeau précédent ? (justifier votre réponse)*

Exercice 2.7. *Les revenus en milliers d'euros de quatre salariés sont : 3,1 ; 2,6 ; 1,7 ; 1,9 . La variable statistique correspondante est notée par R.* (*a*) *Calculer la variance de R d'abord au moyen de la formule (2.1) puis au moyen de la formule (2.2).* (*b*) *Calculer l'écart-type de R.*

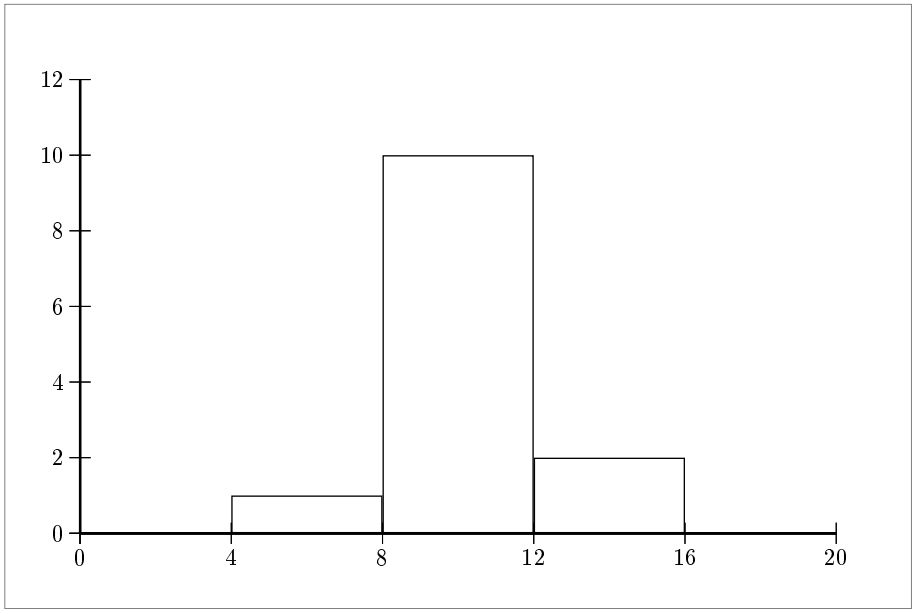
Exemple 2.12 (Illustration de l'utilité de l'écart-type)**.** *Les 25 étudiants d'un Master sont répartis en deux groupes, 13 étudiants sont dans le groupe 1 et les 12 restant dans le groupe 2. Ces 25 étudiants ont passé un examen ; le tableau suivant donne un descriptif de la répartition des notes obtenues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

Tableau de répartition des notes dans chacun des deux groupes

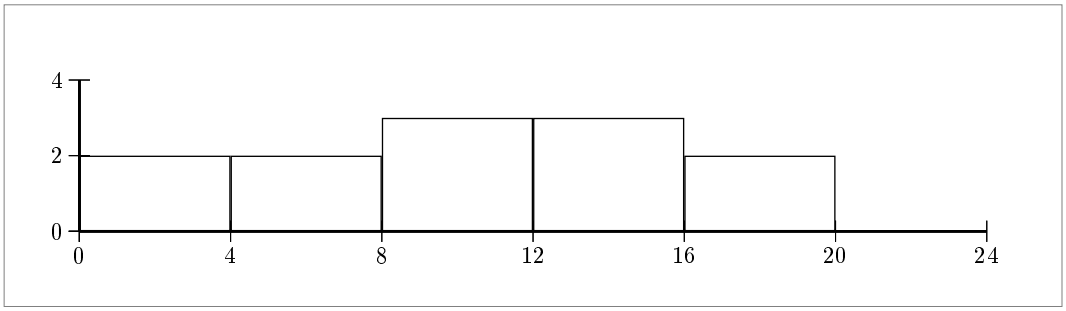
Centres des Classes	*Classes de Note*	*Effectifs du groupe 1*	*Effectifs du groupe 2*
2	*[0,4]*	*0*	*2*
6	*]4,8]*	*1*	*2*
10	*]8,12]*	*10*	*3*
14	*]12,16]*	*2*	*3*
18	*]16,20]*	*0*	*2*
		Total $= N_1 =$*13*	*Total* $= N_2 =$*12*

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 arithmétiques 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 :*

$$V_1 = \frac{1 \times 6^2 + 10 \times 10^2 + 2 \times 14^2}{13} - \left(\frac{134}{13}\right)^2 \simeq 3,60$$

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 (utilisés 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 :*

$$\sigma_1 = \sqrt{3,60} \simeq 1,90$$

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 et 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 :

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

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

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

Ainsi,

$$V_T \simeq \left(\frac{13}{25} \times 3,60 + \frac{12}{25} \times 27,96\right) + \left(\frac{13}{25}\left(10,31-10,32\right)^2 + \frac{12}{25}\left(10,33-10,32\right)^2\right) \simeq 15,30$$

et donc l'écart-type de X vaut $\sqrt{15,30} \simeq 3,91$.

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

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

Remarque 2.9. *On a toujours que $e_{\overline{x}} \leq \sigma_X$. Autrement dit, on a toujours*

(Ecart aboslu moyen **à la moyenne***) $\leq$ (Ecart-type).*

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 2.8. *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 Lois normales et lois dérivées

3.1 Notions de base de probabilités

Le concept de **variable aléatoire**, qui sera défini dans la section suivante, permet d'obtenir des modèles mathématiques pour des variables statistiques quantitatives ; **dans ce chapitre nous nous restreignons à des variables continues**, par exemple la variable « Revenus des Contribuables » qui a été étudiée précédemment. Dans ce cadre, l'ensemble des individus associés à la variable (en l'occurrence l'ensemble de tous les contribuables soumis à l'impôt sur le revenu

en 1965) est supposé être très grand (infiniment grand en théorie). Il est noté par Ω « grand omega » et il est appelé **espace de probabilité**. Chaque élément de Ω, c'est-à-dire chaque individu (en l'occurence chaque contribuable), est noté par ω « petit omega ».

Très très souvent on est amené à s'intéresser à certaines parties (c'est-à-dire à certains sous-ensembles) de Ω ; par exemple, le sous-ensemble noté par A des contribuables de sexe féminin, celui noté par B des contribuables de sexe masculin, celui noté par C des contribuables (femmes ou hommes) âgés entre 35 et 45 ans, etc.. Les sous-ensembles A, B, C, etc. de Ω dont on peut mesurer les probabilités (désignées par $\mathbb{P}(A)$, $\mathbb{P}(B)$, $\mathbb{P}(C)$, etc.) sont appelés les **événements** de Ω. **La probabilité d'un événement est toujours comprise entre** 0 **(zéro) et** 1 **(un).** L'ensemble Ω lui-même est un événement ; il est appelé **l'événement certain** car $\mathbb{P}(\Omega) = 1$. L'ensemble vide, autrement dit l'ensemble qui ne contient aucun élément (par exemple l'ensemble des contribuables âgés de plus de 200 ans), qui est noté par $\emptyset$, est lui aussi un événement ; il est appelé **l'événement impossible** car $\mathbb{P}(\emptyset) = 0$.

Soient A et B deux événements arbitraire de Ω.

— On dit que A est **inclus** dans B, ce que l'on note par $A \subset B$, lorsque **tout** ω **qui fait partie de** A **fait aussi partie de** B.

— **L'intersection de** A **et** B est l'événement, noté par $A \cap B$ (ou $B \cap A$), qui est défini comme étant l'ensemble des ω **qui font partie de** A **<u>et</u>** B **à la fois**. Lorsque $A \subset B$, on a alors $A \cap B = A$. Par ailleurs, A et B sont dits **incompatibles** lorsque leur intersection ne contient aucun éléments ; autrement dit $A \cap B = \emptyset$ (c'est par exemple le cas lorsque A est formé par les contribuables de sexe féminin, et B ceux de sexe masculin). Par ailleurs, A et B sont dits **indépendants** lorsque $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$. Les notions d'événements incompatibles et d'événements indépendants sont en général exclusives l'une de l'autre et ne doivent donc surtout pas être confondues !

— **La réunion (ou union) de** A **et** B est l'événement, noté par $A \cup B$ (ou $B \cup A$), qui est défini comme étant l'ensemble des ω **qui font partie de** A **<u>ou</u> de** B (le « ou » est inclusif, c'est-à-dire que $A \cap B \subset A \cup B$). Lorsque $A \subset B$, on a alors $A \cup B = B$.

— **La différence** $A \setminus B$ (à ne pas confondre avec $B \setminus A$) est l'événement qui est défini comme étant l'ensemble des ω **qui font partie de** A **<u>mais</u> ne font pas partie de** B ; la différence $\Omega \setminus B$ (il s'agit du cas particulier où $A = \Omega$) est appelée **l'événement contraire de** B.

— **La différence symétrique** $A \Delta B$ est l'événement défini par :

$$A \Delta B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B)\,.$$

Remarque 3.1. *Soient A, B et C trois événements de Ω. On a*

$$C \cap (A \cup B) = (C \cap A) \cup (C \cap B)\,, \tag{3.1}$$

$$C \setminus (A \cup B) = (C \setminus A) \cap (C \setminus B)\,, \tag{3.2}$$

et

$$C \setminus (A \cap B) = (C \setminus A) \cup (C \setminus B)\,. \tag{3.3}$$

Remarque 3.2. (Probabilité d'une réunion de deux événements) *On a*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \tag{3.4}$$

Notons que, **dans le cas où** A **et** B **sont incomptibles** ($A \cap B = \emptyset$)*, la formule (3.4) se réduit à*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B), \tag{3.5}$$

puisque dans ce cas on a $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$.

Remarque 3.3. (Probabilité d'une différence de deux événements) *On a*

$$\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B), \tag{3.6}$$

et ainsi,

$$\mathbb{P}(\Omega \setminus B) = 1 - \mathbb{P}(B), \tag{3.7}$$

puisque $\mathbb{P}(\Omega) = 1$ *et* $\Omega \cap B = B$.

Exercice 3.1. *Soient A, B et C trois événements de Ω. Montrer que l'on a*

$$\mathbb{P}(A \cup B \cup C) \quad = \quad \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)\,.$$

Exercice 3.2. *Belmont Construction débute la construction de deux nouveaux projets domiciliaires, indépendants l'un de l'autre, que nous notons projet I et projet II. Toutefois, il existe une certaine incertitude quant à l'échéancier pour compléter ces projets : dans un an, la probabilité que le projet I soit complété à 100% vaut 1/4, la probabilité qu'il soit complété à 85% vaut 1/2 et la probabilité qu'il soit complété à 50% vaut 1/4 ; dans un an, la probabilité que le projet II soit complété à 100% vaut 1/3, la probabilité qu'il soit complété à 85% vaut 1/2 et la probabilité qu'il soit complété à 50% vaut 1/6.*
1) Quelle est la probabilité que le projet I ne soit pas complété à 100% dans un an ? Quelle est celle que le projet II ne soit pas complété à 100% dans un an ? Quelle est celle qu'aucun de ces deux projets ne soit complété à 100% dans un an ?
2) Quelle est la probabilité qu'au moins l'un des deux projets soit complété à 100% dans un an ?
3) Quelle est la probabilité que l'un de ces deux projets soit complété à 100% dans un an et l'autre ne le soit pas ?

3.2 Variables aléatoires continues

Définition 3.1. *Une variable aléatoire notée par X (on aurait pu choisir une autre lettre Y, Z, etc.) est une grandeur numérique attachée à une expérience aléatoire ; par exemple X désigne le revenu d'un contribuable choisi au hasard. Plus précisément une variable aléatoire est une application de l'espace de probabilité Ω dans l'ensemble nombres réels $\mathbb{R}$.*

Remarque 3.4. *Comme nous l'avons déjà indiqué, dans ce chapitre nous nous restreignons uniquement à des variables aléatoires continues ; c'est-à-dire qu'elles peuvent prendre toute valeur d'un intervalle. Lorsque X est une variable aléatoire de ce type, déterminer $\mathbb{P}(X = t)$, la probabilité que X vaille exactement une certaine valeur t, n'est pas d'un grand intérêt, parce qu'en général $\mathbb{P}(X = t) = 0$. Il est beaucoup plus intéressant de déterminer $\mathbb{P}(a \leq X \leq b)$, la probabilité que X soit comprise entre deux valeurs quelconques a et b, où $a < b$. Dans tout ce chapitre, on suppose systématiquement que cette probabilité est donnée par l'intégrale :*

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\,dx, \tag{3.8}$$

où f_X est une fonction (dépendant de la variable aléatoire X) qui est continue (c'est-à-dire « dont on peut tracer le graphe d'un seul trait ») et **à valeurs positives**, *appelée* **densité** *(de probabilité) de la variable aléatoire continue X. On suppose aussi systématiquement que les deux intégrales*

$$\int_{-\infty}^{+\infty} x f_X(x)\,dx \qquad \text{et} \qquad \int_{-\infty}^{+\infty} x^2 f_X(x)\,dx$$

existent, et que la valeur de chacune d'elles est un nombre fini. Signalons que la forme du graphe de la densité f_X reproduit à peu près celle de l'histogramme des fréquences de la variable statistique qui est modélisée par la variable aléatoire X. Signalons aussi que le fait que X possède une densité nous assure que $\mathbb{P}(X = t) = 0$, pour tout nombre réel t, et par conséquent que

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X < b)\,. \tag{3.9}$$

Remarque 3.5. (Interprétation graphique d'une intégrale d'une fonction positive) *Désignons par f une fonction continue arbitraire* **à valeurs positives**.

- *L'intégrale $\int_a^b f(x)\,dx$ est l'aire (c'est-à-dire la superficie) de la surface qui est comprises entre les deux droites verticales d'équations $x = a$ et $x = b$, qui, de plus, est délimitée par en haut par le graphe de f et par en bas par l'axe des abscisses.*
- *L'intégrale $\int_{-\infty}^a f(x)\,dx$ est l'aire de la surface qui est située à gauche de la droite verticale d'équation $x = a$, qui, de plus, est délimitée par en haut par le graphe de f et par en bas par l'axe des abscisses.*
- *L'intégrale $\int_b^{+\infty} f(x)\,dx$ est l'aire de la surface qui est située à droite de la droite verticale d'équation $x = b$, qui, de plus, est délimitée par en haut par le graphe de f et par en bas par l'axe des abscisses.*
- *L'intégrale $\int_{-\infty}^{+\infty} f(x)\,dx$ est l'aire de la surface qui est délimitée par en haut par le graphe de f et par en bas par l'axe des abscisses.*

On a donc

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{a} f(x)\,dx + \int_a^b f(x)\,dx + \int_b^{+\infty} f(x)\,dx\,.$$

Définition 3.2. (Fonction de répartition) *On appelle* **fonction de répartition** *d'une variable aléatoire X, la fonction, notée par F_X, à valeurs dans l'intervalle $[0\,;1]$, qui est définie, pour nombre réel t, par*

$$F_X(t) = \mathbb{P}(X \leq t)\,. \tag{3.10}$$

Remarque 3.6. (Propriétés importantes d'une fonction de répartition)

(i) La probabilité $\mathbb{P}(a \leq X \leq b)$ se calcule au moyen de la fonction de répartition F_X à l'aide de la formule suivante :

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a). \tag{3.11}$$

(ii) La fonction de répartition F_X est la primitive de la densité f_X qui est donnée, pour tout nombre réel t, par

$$F_X(t) = \int_{-\infty}^{t} f_X(x)\,dx\,.$$

Ainsi, la dérivée de la fonction de répartition F_X n'est rien d'autre que la densité f_X, autrement dit, pour tout nombre réel t, on a

$$F'_X(t) = f_X(t)\,.$$

(iii) F_X est une fonction croissante, continue[3]*, dont la limite en $-\infty$ est égale à 0 (zéro), et dont la limite en $+\infty$ est égale à 1 (un).*

3. Lorsque X n'admet pas de densité (ce qui sort du cadre de ce cours), F_X n'est plus nécessairement une fonction continue mais une fonction continue à droite de tout point et qui de plus admet une limite finie à gauche de tout point.

Définition 3.3. (Espérance) *L'espérance de la variable aléatoire X est notée par $\mathbb{E}(X)$, et elle est définie par*

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x f_X(x)\, dx\,. \tag{3.12}$$

Intuitivement parlant, $\mathbb{E}(X)$ est interprétée comme « la valeur que prend en moyenne la variable aléatoire X ».

Définition 3.4. (Variance et Ecart-type) *La variance de la variable aléatoire X est notée par $\mathbb{V}(X)$, et elle est définie par*

$$\mathbb{V}(X) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)^2\Big] = \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2 = \int_{-\infty}^{+\infty} x^2 f_X(x)\, dx - \bigg(\int_{-\infty}^{+\infty} x f_X(x)\, dx\bigg)^2. \tag{3.13}$$

L'écart-type de X est noté par $\sigma(X)$, et il est définie par

$$\sigma(X) = \sqrt{\mathbb{V}(X)}\,. \tag{3.14}$$

Intuitivement parlant, $\sigma(X)$ est interprétée comme « l'écart moyen de la variable aléatoire X par rapport à son espérance $\mathbb{E}(X)$ ».

Définition 3.5. (Variables aléatoires indépendantes) *Deux variables aléatoires X_1 et X_2 sont dites* **indépendantes**, *lorsque pour tous nombres réels a_1, b_1 et a_2, b_2, vérifiant $a_1 < b_1$ et $a_2 < b_2$, les deux événements $\{a_1 \leq X_1 \leq b_1\}$ et $\{a_2 \leq X_2 \leq b_2\}$ sont indépendants.*

Plus généralement, soit un entier $n \geq 2$, n variables aléatoires $X_1, X_2, \ldots, X_n$ sont dites **indépendantes**, *lorsque pour tous nombres réels a_1, b_1 et a_2, b_2 et $\ldots$ et a_n, b_n, vérifiant $a_1 < b_1$ et $a_2 < b_2$ et $\ldots$ et $a_n < b_n$, les n événements*

$$\{a_1 \leq X_1 \leq b_1\} \text{ et } \{a_2 \leq X_2 \leq b_2\} \text{ et } \ldots \text{ et } \{a_n \leq X_n \leq b_n\}$$

sont mutuellement indépendants.

3.3 Variables aléatoires de lois normales

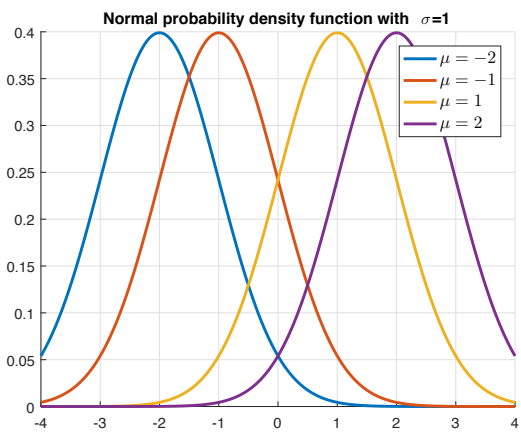
Lorsque les valeurs prises par une variable aléatoire X résultent d'un très grand nombre de causes aléatoires qui s'ajoutent, qui sont indépendantes l'une de l'autre et dont aucune ne domine, on peut considérer que X suit une loi normale. C'est par exemple le cas lorsque X désigne le poids aléatoire (exprimé en grammes) de riz contenu dans un paquet choisi au hasard dans un hypermarché. En effet, ce poids global X résulte de la somme des très nombreux poids des grains de riz se trouvant dans le paquet.

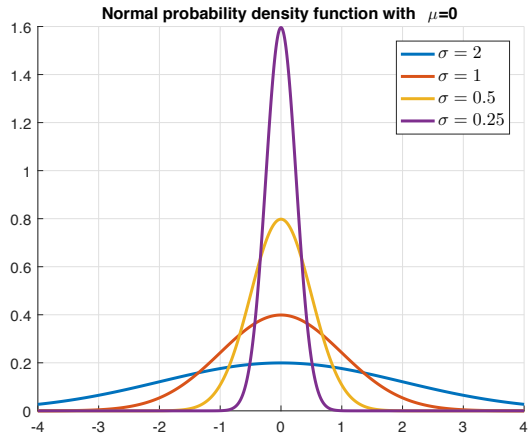
Définition 3.6. *Soit μ (« mu ») un nombre réel arbitraire. Soit σ (« sigma ») un nombre réel strictement positif arbitraire. On dit qu'une variable aléatoire continue X suit* **une loi normale de moyenne (c'est-à-dire d'espérance)** μ **et d'écart-type** σ *(ou encore de variance σ^2), expression qui se note par $X \hookrightarrow \mathcal{N}(\mu, \sigma^2)$, si X admet comme densité la fonction f_X qui vaut, pour tout nombre réel x,*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\bigg(-\frac{(x-\mu)^2}{2\sigma^2}\bigg)\,. \tag{3.15}$$

Signalons que dans le cas particulier où l'on a $\mu = 0$ et $\sigma = 1$ la loi normale est dite **centrée et réduite**.

Remarque 3.7. *Comme le montrent les deux figures suivantes : l'allure du graphe de f_X est celle du courbe en cloche qui admet comme axe de symétrie la droite d'équation $x = \mu$; plus le paramètre σ est grand plus cette courbe est aplatie et plus il est petit plus elle est pointue ; la fonction f_X atteint son maximum lorsque $x = \mu$ et on a alors $f_X(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$*





Remarque 3.8. *Si une variable aléatoire X suit une loi normale de moyenne μ et d'écart-type σ, alors la variable aléatoire*

$$Z = \frac{X - \mu}{\sigma} \tag{3.16}$$

suit une loi normale centrée et réduite ; de plus, F_X et F_Z, les fonctions de répartition de X et Z, vérifient, pour tout nombre réel t,

$$F_X(t) = F_Z\left(\frac{t - \mu}{\sigma}\right). \tag{3.17}$$

Réciproquement, si une variable aléatoire T suit une loi normale centrée et réduite, alors, pour tout nombre réel μ et pour tout nombre réel strictement positif σ, la variable aléatoire

$$Y = \sigma T + \mu \tag{3.18}$$

suit une loi normale de moyenne μ et d'écart-type σ.

Remarque 3.9. *Soit Z une variable aléatoire de loi normale centrée et réduite, alors sa fonction de répartition F_Z vérifie, pour tout nombre réel t,*

$$F_Z(t) = 1 - F_Z(-t)\,, \tag{3.19}$$

et il en résulte, entre autres, que $F_Z(0) = 1/2$. Signalons que l'égalité (3.19) est très utile, et qu'elle provient de la symétrie par rapport à l'axe des ordonnées (c'est-à-dire la droite d'équation $x = 0$) du graphe de f_Z la densité de Z.

Exemple 3.1. *La variable aléatoire X, qui est exprimée en milliers d'euros, désigne le résultat net (la différence entre les produits et les charges) d'une certaine PME pour le prochain mois ; ainsi $X > 0$ signifie que la PME aura réalisé un bénéfice, en revanche $X < 0$ signifie qu'elle aura subi une perte. On admet que X suit une loi normale de moyenne 10 et d'écart-type 4. Au moyen de la table de loi normale centrée et réduite calculons les cinq probabilités suivantes : $\mathbb{P}(X \leq 12)$, $\mathbb{P}(X < 9)$, $\mathbb{P}(X > 11)$, $\mathbb{P}(X \geq 6)$, et $\mathbb{P}(-2 < X < 13)$.*

D'après la Remarque 3.8, nous savons que la variable aléatoire

$$Z = \frac{X - 10}{4}$$

suit une loi normale centrée et réduite, et que sa fonction de répartition F_Z, vérifie, pour tout nombre réel t,

$$F_X(t) = F_Z\left(\frac{t - 10}{4}\right), \tag{3.20}$$

où F_X désigne la fonction de répartition de X. En utilisant (3.10), (3.20) et la table de la loi normale, on trouve que

$$\mathbb{P}(X \leq 12) = F_X(12) = F_Z\left(\frac{12 - 10}{4}\right) = F_Z(0,5) = 0,6915\,.$$

En utilisant l'égalité $\mathbb{P}(X = 9) = 0$, (3.10), (3.20), (3.19) et la table de la loi normale, on trouve que

$$\mathbb{P}(X < 9) = F_X(9) = F_Z\left(\frac{9 - 10}{4}\right) = F_Z(-0,25) = 1 - F_Z(0,25) = 1 - 0,5987 = 0,4013\,.$$

En utilisant l'égalité $\{X \leq 11\} = \Omega \setminus \{X > 11\}$ (autrement dit le fait que $\{X \leq 11\}$ est l'événement contraire de $\{X > 11\}$), (3.7), (3.10), (3.20) et la table de la loi normale, on obtient que

$$\mathbb{P}(X > 11) = 1 - F_X(11) = 1 - F_Z\left(\frac{11 - 10}{4}\right) = 1 - F_Z(0,25) = 0,4013\,.$$

En utilisant l'égalité $\{X < 6\} = \Omega \setminus \{X \geq 6\}$, l'égalité $\mathbb{P}(X = 6) = 0$, (3.10), (3.20), (3.19) et la table de la loi normale, on trouve que

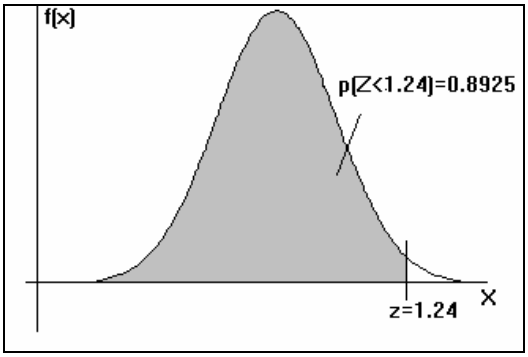
$$\mathbb{P}(X \geq 6) = 1 - \mathbb{P}(X < 6) = 1 - F_X(6) = 1 - F_Z\left(\frac{6 - 10}{4}\right) = 1 - F_Z(-1) = F_Z(1) = 0,8413\,.$$

En utilisant (3.9), (3.11), (3.20), (3.19) et la table de la loi normale, on obtient que

$$\begin{aligned}\mathbb{P}(-2 < X < 13) &= F_X(13) - F_X(-2) = F_Z\left(\frac{13 - 10}{4}\right) - F_Z\left(\frac{-2 - 10}{4}\right)\\ &= F_Z(0,75) - F_Z(-3) = F_Z(0,75) - 1 + F_Z(3) = 0,7734 - 1 + 0,99865 = 0,77205\,.\end{aligned}$$

TABLE DE LA LOI NORMALE CENTREE REDUITE

Lecture de la table: Pour z=1.24 (intersection de la ligne 1.2 et de la colonne 0.04), on a la proportion P(Z < 1,24) = 0.8925



P(Z > 1,96) = 0,025
P(Z > 2,58) = 0,005
P(Z > 3,29) = 0,0005

Rappels:

1/ P(Z > z) = 1 - P(Z < z) et 2/ P(Z < -z) = P(Z > z)

Exemple: Sachant P(Z < 1,24) = 0,8925, on en déduit:

1/ (P(Z > 1,24) = 1 - P(Z < 1,24) = 1- 0,8925 = 0,1075

2/ P(Z < -1,24) = P(Z > 1,24) = 0,1075

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997
4,0	0,99997	0,99997	0,99997	0,99997	0,99997	0,99997	0,99998	0,99998	0,99998	0,99998

Remarque 3.10. (Somme de variables aléatoires indépendantes de lois normales) *Soient deux variables aléatoires indépendantes X_1 et X_2 qui sont respectivement de lois normales $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$. Soient trois nombres réels a_1, a_2 et b. Alors, la variable aléatoire $Y = a_1X_1 + a_2X_2 + b$ est de loi normale de moyenne $\mu = a_1\mu_1 + a_2\mu_2 + b$ et de variance $\sigma^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2$.*

Plus généralement, soient n variables aléatoires indépendantes $X_1, X_2, \ldots, X_n$ qui sont respectivement de lois normales $\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2), \ldots, \mathcal{N}(\mu_n, \sigma_n^2)$. Soient $n+1$ nombres réels $a_1, a_2, \ldots, a_n$ et b. Alors, la variable aléatoire

$$Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n + b$$

est de loi normale de moyenne

$$\mu = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n + b$$

et de variance

$$\sigma^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_n^2\sigma_n^2\,.$$

Exemple 3.2. *On transporte des paquets par lot de 500 paquets qui sont placés dans un carton qui lui-même pèse 30 kg. Le poids (en grammes) d'un paquet est une variable aléatoire qui suit une loi normale d'espérance 1000 g et d'écart-type 100 g. On suppose les poids des paquets indépendants. La livraison des lots se fait par un monte-charge qui ne fonctionne pas si le poids du lot dépasse 535 kg. Nous allons déterminer la probabilité qu'un lot ne puisse pas être livré.*

Nous notons par X_1, $X_2, \ldots, X_{500}$ les poids (en grammes) des paquets d'un lot dont le poids global (en grammes) est noté par Y. Etant donné que les variables aléatoires X_1, $X_2, \ldots, X_{500}$ sont indépendantes et de même loi normale $\mathcal{N}(1000\,, 100^2)$, grâce à la Remarque 3.10, nous pouvons affirmer que la variable aléatoire

$$Y = X_1 + X_2 + \ldots + X_{500} + 30000$$

suit une loi normale avec une moyenne égale à $500 \times 1000 + 30000 = 530000$ et une variance égale à 500×100^2, c'est-à-dire que l'écart-type vaut $1000\sqrt{5}$. Ainsi, en notant par Z la variable aléatoire normale centrée et réduite définie par

$$Z = \frac{Y - 530000}{1000\sqrt{5}}$$

et par F_Z sa fonction de répartition, on trouve que $\mathbb{P}(Y > 535000)$, la probabilité que le lot de poids Y ne puisse pas être livré, vaut

$$\mathbb{P}(Y > 535000) = 1 - F_Z\left(\frac{535000 - 530000}{1000\sqrt{5}}\right) \simeq 1 - F_Z(2,24) \simeq 1 - 0,9875 = 0,0125\,.$$

3.4 Variables aléatoires de lois du χ^2 (« Chi2 »)

La loi du χ^2 est très utile en statistique, elle permet entre autres de tester s'il existe ou non une liaison significative entre deux variables statistiques qualitatives (voir la Remarque 4.5).

Définition 3.7. *Soit un entier naturel strictement positif ν (« nu »). On dit qu'une variable aléatoire Y suit une loi du χ^2 (ou loi de Pearson) à ν degrés de liberté, expression qui se note par $Y \hookrightarrow \chi^2(\nu)$, lorsque Y s'écrit sous la forme :*

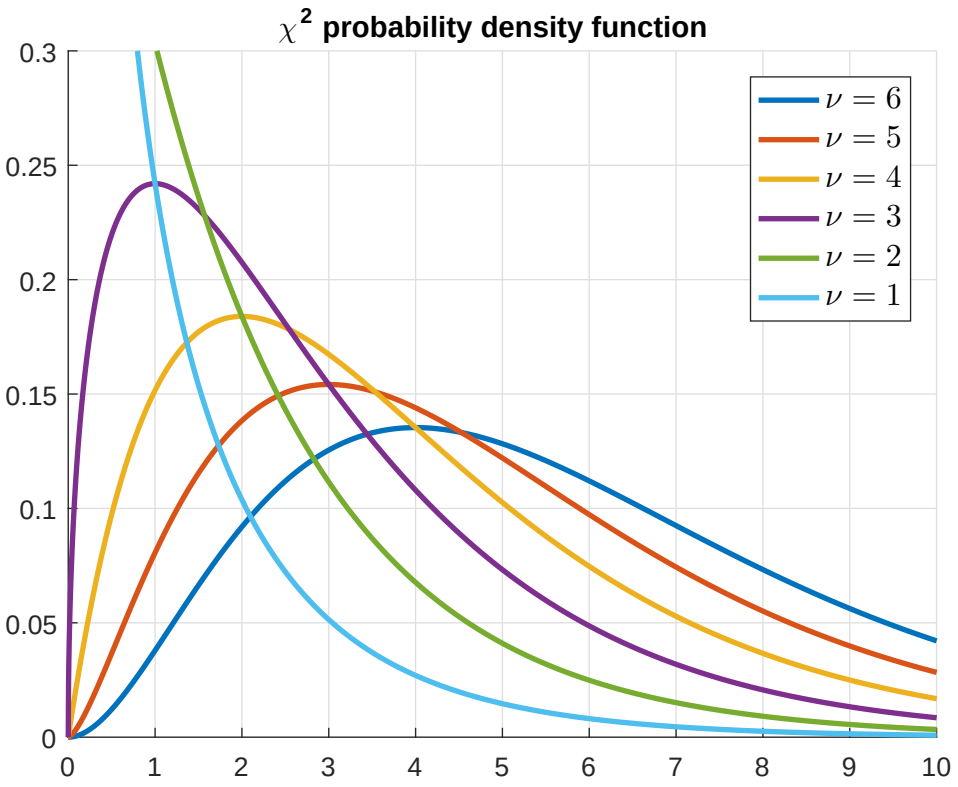
$$Y = Z_1^2 + \ldots + Z_\nu^2 = \sum_{n=1}^{\nu} Z_n^2\,,$$

où $Z_1, \ldots, Z_\nu$ sont ν **variables aléatoires indépendantes de même loi normale centrée et réduite**. *Alors l'espérance et la variance de Y sont telles que $\mathbb{E}(Y) = \nu$ et $\mathbb{V}(Y) = 2\nu$. De plus f_Y, la densité de Y, vérifie $f_Y(x) = 0$, pour nombre réel $x \leq 0$, et vérifie aussi*

$$f_Y(x) = c_\nu \, x^{\frac{\nu}{2}-1} \exp\left(-\frac{x}{2}\right), \qquad \textit{pour nombre réel } x > 0,$$

où c_ν est une constante strictement positive de normalisation telle que $\int_{-\infty}^{+\infty} f_Y(x)\,dx = 1$.

L'allure du graphe de f_Y est la suivante :



Remarque 3.11. (Propriétés importantes)

(*i*) *Soient ν' et ν'' deux entiers vérifiant $1 \leq \nu' \leq \nu''$. Soient Y' et Y'' deux variables aléatoires telles que $Y' \hookrightarrow \chi^2(\nu')$ et $Y'' \hookrightarrow \chi^2(\nu'')$. On a alors, pour tout nombre réel $s \geq 0$, $\mathbb{P}(Y' > s) \leq \mathbb{P}(Y'' > s)$.*

(*ii*) *Soient ν_1 et ν_2 deux entiers arbitraires strictement positifs, soient Y_1 et Y_2 deux variables aléatoires* **indépendantes** *telles que $Y' \hookrightarrow \chi^2(\nu_1)$ et $Y'' \hookrightarrow \chi^2(\nu_2)$; alors la variable aléatoire $U = Y_1 + Y_2$ suit une loi du χ^2 à $\nu_1 + \nu_2$ degrés de liberté. Plus généralement, soient $\nu_1, \nu_2, \ldots, \nu_n$ n entiers arbitraires strictement positifs, soient $Y_1, Y_2, \ldots, Y_n$ n variables aléatoires* **indépendantes** *qui sont telles que $Y_1 \hookrightarrow \chi^2(\nu_1)$, $Y_2 \hookrightarrow \chi^2(\nu_2), \ldots, Y_n \hookrightarrow \chi^2(\nu_n)$; alors la variable aléatoire $U' = Y_1 + Y_2 + \ldots + Y_n$ suit une loi du χ^2 à $\nu_1 + \nu_2 + \ldots + \nu_n$ degrés de liberté.*

(*iii*) *Soient* X_1, $X_2, \ldots,$ X_ν ν *variables aléatoires* **indépendantes** *qui sont* **toutes de même loi normale** $\mathcal{N}(\mu, \sigma^2)$ *; alors la variable aléatoire*

$$\begin{aligned} S &= \left(\frac{X_1 - \mu}{\sigma}\right)^2 + \left(\frac{X_2 - \mu}{\sigma}\right)^2 + \cdots + \left(\frac{X_\nu - \mu}{\sigma}\right)^2 \\ &= \frac{1}{\sigma^2}\Big(\big(X_1 - \mu\big)^2 + \big(X_2 - \mu\big)^2 + \cdots + \big(X_\nu - \mu\big)^2\Big) \end{aligned} \tag{3.21}$$

suit une loi du χ^2 *à* ν *degrés de liberté.*

(*iv*) *Soit un entier* $\nu \geq 2$*, soient* X_1, $X_2, \ldots,$ X_ν ν *variables aléatoires* **indépendantes** *qui sont* **toutes de même loi normale** $\mathcal{N}(\mu, \sigma^2)$*, nous désignons par* $\overline{X}_\nu$ *leur* **moyenne empirique (ou moyenne arithmétique)** *définie par*

$$\overline{X}_\nu = \frac{X_1 + X_2 + \cdots + X_\nu}{\nu}\,; \tag{3.22}$$

alors, la variable aléatoire

$$\begin{aligned} V &= \left(\frac{X_1 - \overline{X}_\nu}{\sigma}\right)^2 + \left(\frac{X_2 - \overline{X}_\nu}{\sigma}\right)^2 + \cdots + \left(\frac{X_\nu - \overline{X}_\nu}{\sigma}\right)^2 \\ &= \frac{1}{\sigma^2}\Big(\big(X_1 - \overline{X}_\nu\big)^2 + \big(X_2 - \overline{X}_\nu\big)^2 + \cdots + \big(X_\nu - \overline{X}_\nu\big)^2\Big) \end{aligned} \tag{3.23}$$

suit une loi du χ^2 *à* $\nu - 1$ *degrés de liberté (attention il y a* 1 *degré de liberté en moins par rapport au cas précédent).*

Table de lois du χ^2

Cette table a été téléchargée sur Internet (http ://www.math.univ-metz.fr/ bonneau/STAT0607/)

Loi de Khi-deux

Le tableau donne **x** tel que **P(K > x) = p**

p	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001
ddl														
1	0,0000	0,0000	0,0002	0,0006	0,0039	0,0158	0,0642	1,6424	2,7055	3,8415	5,4119	6,6349	7,8794	10,8276
2	0,0020	0,0100	0,0201	0,0404	0,1026	0,2107	0,4463	3,2189	4,6052	5,9915	7,8240	9,2103	10,5966	13,8155
3	0,0243	0,0717	0,1148	0,1848	0,3518	0,5844	1,0052	4,6416	6,2514	7,8147	9,8374	11,3449	12,8382	16,2662
4	0,0908	0,2070	0,2971	0,4294	0,7107	1,0636	1,6488	5,9886	7,7794	9,4877	11,6678	13,2767	14,8603	18,4668
5	0,2102	0,4117	0,5543	0,7519	1,1455	1,6103	2,3425	7,2893	9,2364	11,0705	13,3882	15,0863	16,7496	20,5150
6	0,3811	0,6757	0,8721	1,1344	1,6354	2,2041	3,0701	8,5581	10,6446	12,5916	15,0332	16,8119	18,5476	22,4577
7	0,5985	0,9893	1,2390	1,5643	2,1673	2,8331	3,8223	9,8032	12,0170	14,0671	16,6224	18,4753	20,2777	24,3219
8	0,8571	1,3444	1,6465	2,0325	2,7326	3,4895	4,5936	11,0301	13,3616	15,5073	18,1682	20,0902	21,9550	26,1245
9	1,1519	1,7349	2,0879	2,5324	3,3251	4,1682	5,3801	12,2421	14,6837	16,9190	19,6790	21,6660	23,5894	27,8772
10	1,4787	2,1559	2,5582	3,0591	3,9403	4,8652	6,1791	13,4420	15,9872	18,3070	21,1608	23,2093	25,1882	29,5883
11	1,8339	2,6032	3,0535	3,6087	4,5748	5,5778	6,9887	14,6314	17,2750	19,6751	22,6179	24,7250	26,7568	31,2641
12	2,2142	3,0738	3,5706	4,1783	5,2260	6,3038	7,8073	15,8120	18,5493	21,0261	24,0540	26,2170	28,2995	32,9095
13	2,6172	3,5650	4,1069	4,7654	5,8919	7,0415	8,6339	16,9848	19,8119	22,3620	25,4715	27,6882	29,8195	34,5282
14	3,0407	4,0747	4,6604	5,3682	6,5706	7,7895	9,4673	18,1508	21,0641	23,6848	26,8728	29,1412	31,3193	36,1233
15	3,4827	4,6009	5,2293	5,9849	7,2609	8,5468	10,3070	19,3107	22,3071	24,9958	28,2595	30,5779	32,8013	37,6973
16	3,9416	5,1422	5,8122	6,6142	7,9616	9,3122	11,1521	20,4651	23,5418	26,2962	29,6332	31,9999	34,2672	39,2524
17	4,4161	5,6972	6,4078	7,2550	8,6718	10,0852	12,0023	21,6146	24,7690	27,5871	30,9950	33,4087	35,7185	40,7902
18	4,9048	6,2648	7,0149	7,9062	9,3905	10,8649	12,8570	22,7595	25,9894	28,8693	32,3462	34,8053	37,1565	42,3124
19	5,4068	6,8440	7,6327	8,5670	10,1170	11,6509	13,7158	23,9004	27,2036	30,1435	33,6874	36,1909	38,5823	43,8202
20	5,9210	7,4338	8,2604	9,2367	10,8508	12,4426	14,5784	25,0375	28,4120	31,4104	35,0196	37,5662	39,9968	45,3147
21	6,4467	8,0337	8,8972	9,9146	11,5913	13,2396	15,4446	26,1711	29,6151	32,6706	36,3434	38,9322	41,4011	46,7970
22	6,9830	8,6427	9,5425	10,6000	12,3380	14,0415	16,3140	27,3015	30,8133	33,9244	37,6595	40,2894	42,7957	48,2679
23	7,5292	9,2604	10,1957	11,2926	13,0905	14,8480	17,1865	28,4288	32,0069	35,1725	38,9683	41,6384	44,1813	49,7282
24	8,0849	9,8862	10,8564	11,9918	13,8484	15,6587	18,0618	29,5533	33,1962	36,4150	40,2704	42,9798	45,5585	51,1786
25	8,6493	10,5197	11,5240	12,6973	14,6114	16,4734	18,9398	30,6752	34,3816	37,6525	41,5661	44,3141	46,9279	52,6197
26	9,2221	11,1602	12,1981	13,4086	15,3792	17,2919	19,8202	31,7946	35,5632	38,8851	42,8558	45,6417	48,2899	54,0520
27	9,8028	11,8076	12,8785	14,1254	16,1514	18,1139	20,7030	32,9117	36,7412	40,1133	44,1400	46,9629	49,6449	55,4760
28	10,3909	12,4613	13,5647	14,8475	16,9279	18,9392	21,5880	34,0266	37,9159	41,3371	45,4188	48,2782	50,9934	56,8923
29	10,9861	13,1211	14,2565	15,5745	17,7084	19,7677	22,4751	35,1394	39,0875	42,5570	46,6927	49,5879	52,3356	58,3012
30	11,5880	13,7867	14,9535	16,3062	18,4927	20,5992	23,3641	36,2502	40,2560	43,7730	47,9618	50,8922	53,6720	59,7031
40	17,9164	20,7065	22,1643	23,8376	26,5093	29,0505	32,3450	47,2685	51,8051	55,7585	60,4361	63,6907	66,7660	73,4020
50	24,6739	27,9907	29,7067	31,6639	34,7643	37,6886	41,4492	58,1638	63,1671	67,5048	72,6133	76,1539	79,4900	86,6608
60	31,7383	35,5345	37,4849	39,6994	43,1880	46,4589	50,6406	68,9721	74,3970	79,0819	84,5799	88,3794	91,9517	99,6072
70	39,0364	43,2752	45,4417	47,8934	51,7393	55,3289	59,8978	79,7146	85,5270	90,5312	96,3875	100,4252	104,2149	112,3169
80	46,5199	51,1719	53,5401	56,2128	60,3915	64,2778	69,2069	90,4053	96,5782	101,8795	108,0693	112,3288	116,3211	124,8392
90	54,1552	59,1963	61,7541	64,6347	69,1260	73,2911	78,5584	101,0537	107,5650	113,1453	119,6485	124,1163	128,2989	137,2084
100	61,9179	67,3276	70,0649	73,1422	77,9295	82,3581	87,9453	111,6667	118,4980	124,3421	131,1417	135,8067	140,1695	149,4493
120	77,7551	83,8516	86,9233	90,3667	95,7046	100,6236	106,8056	132,8063	140,2326	146,5674	153,9182	158,9502	163,6482	173,6174
140	93,9256	100,6548	104,0344	107,8149	113,6593	119,0293	125,7581	153,8537	161,8270	168,6130	176,4709	181,8403	186,8468	197,4508
160	110,3603	117,6793	121,3456	125,4400	131,7561	137,5457	144,7834	174,8283	183,3106	190,5165	198,8464	204,5301	209,8239	221,0190
180	127,0111	134,8844	138,8204	143,2096	149,9688	156,1526	163,8682	195,7434	204,7037	212,3039	221,0772	227,0561	232,6198	244,3705
200	143,8428	152,2410	156,4320	161,1003	168,2786	174,8353	183,0028	216,6088	226,0210	233,9943	243,1869	249,4451	255,2642	267,5405
250	186,5541	196,1606	200,9386	206,2490	214,3916	221,8059	231,0128	268,5986	279,0504	287,8815	298,0388	304,9396	311,3462	324,8324
300	229,9634	240,6634	245,9725	251,8637	260,8781	269,0679	279,2143	320,3971	331,7885	341,3951	352,4246	359,9064	366,8444	381,4252
400	318,2596	330,9028	337,1553	344,0781	354,6410	364,2074	376,0218	423,5895	436,6490	447,6325	460,2108	468,7245	476,6064	493,1318
500	407,9470	422,3034	429,3875	437,2194	449,1468	459,9261	473,2099	526,4014	540,9303	553,1268	567,0698	576,4928	585,2066	603,4460
600	498,6229	514,5289	522,3651	531,0191	544,1801	556,0560	570,6680	628,9433	644,8004	658,0936	673,2703	683,5156	692,9816	712,7712
700	590,0480	607,3795	615,9075	625,3175	639,6130	652,4973	668,3308	731,2805	748,3591	762,6607	778,9721	789,9735	800,1314	821,3468
800	682,0665	700,7250	709,8969	720,0107	735,3623	749,1852	766,1555	833,4557	851,6712	866,9114	884,2789	895,9843	906,7862	929,3289
900	774,5698	794,4750	804,2517	815,0267	831,3702	846,0746	864,1125	935,4987	954,7819	970,9036	989,2631	1001,6296	1013,0364	1036,8260

3.5 Variables aléatoires de lois de Student

Définition 3.8. *Soit un entier naturel strictement positif ν. On dit qu'une variable aléatoire T suit une loi de Student à ν degrés de liberté, lorsque T s'écrit sous la forme :*

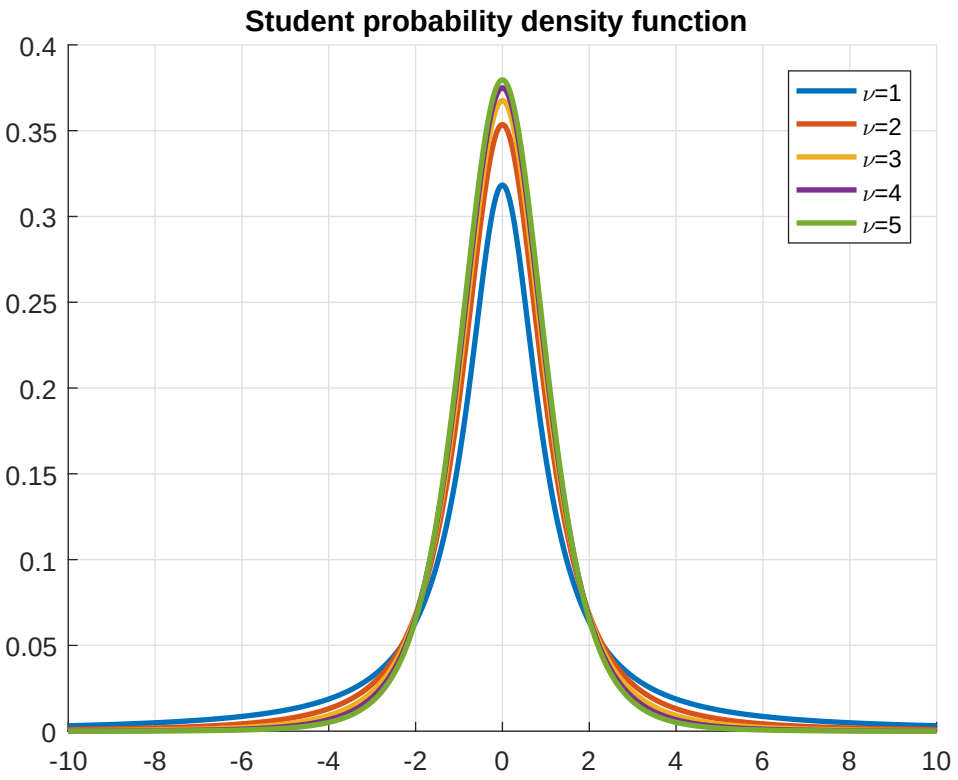
$$T = \frac{Z}{\sqrt{Y/\nu}}$$

où **la variable aléatoire Z qui est de loi normale centrée et réduite est indépendante de la variable aléatoire Y qui suit une loi du χ^2 à ν degrés de liberté**. *Alors l'espérance de T, qui n'existe que si $\nu \geq 2$, vaut zéro et sa variance, qui n'existe que si $\nu \geq 3$, vaut $\nu/(\nu-2)$. De plus, f_T, la densité de T, est une fonction paire qui vérifie*

$$f_T(x) = a_\nu \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \qquad \textit{pour tout nombre réel } x,$$

où a_ν est une constante strictement positive de normalisation telle que $\int_{-\infty}^{+\infty} f_T(x)\,dx = 1$.

L'allure du graphe de f_T est la suivante :



Remarque 3.12. *Soit un entier naturel strictement positif ν et soit T une variable aléatoire qui suit une loi de Student à ν degrés de liberté. Alors F_T la fonction de répartition de T vérifie, pour tout nombre réel t,*

$$F_T(t) = 1 - F_T(-t)\,, \tag{3.24}$$

et il en résulte, entre autres, que $F_T(0) = 1/2$. Signalons que l'égalité (3.24) est très utile, et qu'elle provient de la symétrie par rapport à l'axe des ordonnées (c'est-à-dire la droite d'équation $x = 0$) du graphe de f_T la densité de T.

Table de lois de Student

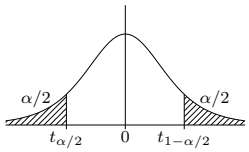
Cette table a été téléchargée sur Internet (http ://wwwmathlabo.univ-poitiers.fr/∼phan/)

A.3. Lois de Student

Si T est une variable aléatoire suivant la loi de Student à ν degrés de liberté, la table donne, pour α fixé, la valeur $t_{1-\alpha/2}$ telle que

$$\mathbb{P}\{|T| \geqslant t_{1-\alpha/2}\} = \alpha.$$

Ainsi, $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à ν degrés de liberté.



ν \ α	0,900	0,500	0,300	0,200	0,100	0,050	0,020	0,010	0,001
1	0,1584	1,0000	1,9626	3,0777	6,3138	12,7062	31,8205	63,6567	636,6193
2	0,1421	0,8165	1,3862	1,8856	2,9200	4,3027	6,9646	9,9248	31,5991
3	0,1366	0,7649	1,2498	1,6377	2,3534	3,1824	4,5407	5,8409	12,9240
4	0,1338	0,7407	1,1896	1,5332	2,1318	2,7764	3,7469	4,6041	8,6103
5	0,1322	0,7267	1,1558	1,4759	2,0150	2,5706	3,3649	4,0321	6,8688
6	0,1311	0,7176	1,1342	1,4398	1,9432	2,4469	3,1427	3,7074	5,9588
7	0,1303	0,7111	1,1192	1,4149	1,8946	2,3646	2,9980	3,4995	5,4079
8	0,1297	0,7064	1,1081	1,3968	1,8595	2,3060	2,8965	3,3554	5,0413
9	0,1293	0,7027	1,0997	1,3830	1,8331	2,2622	2,8214	3,2498	4,7809
10	0,1289	0,6998	1,0931	1,3722	1,8125	2,2281	2,7638	3,1693	4,5869
11	0,1286	0,6974	1,0877	1,3634	1,7959	2,2010	2,7181	3,1058	4,4370
12	0,1283	0,6955	1,0832	1,3562	1,7823	2,1788	2,6810	3,0545	4,3178
13	0,1281	0,6938	1,0795	1,3502	1,7709	2,1604	2,6503	3,0123	4,2208
14	0,1280	0,6924	1,0763	1,3450	1,7613	2,1448	2,6245	2,9768	4,1405
15	0,1278	0,6912	1,0735	1,3406	1,7531	2,1314	2,6025	2,9467	4,0728
16	0,1277	0,6901	1,0711	1,3368	1,7459	2,1199	2,5835	2,9208	4,0150
17	0,1276	0,6892	1,0690	1,3334	1,7396	2,1098	2,5669	2,8982	3,9651
18	0,1274	0,6884	1,0672	1,3304	1,7341	2,1009	2,5524	2,8784	3,9216
19	0,1274	0,6876	1,0655	1,3277	1,7291	2,0930	2,5395	2,8609	3,8834
20	0,1273	0,6870	1,0640	1,3253	1,7247	2,0860	2,5280	2,8453	3,8495
21	0,1272	0,6864	1,0627	1,3232	1,7207	2,0796	2,5176	2,8314	3,8193
22	0,1271	0,6858	1,0614	1,3212	1,7171	2,0739	2,5083	2,8188	3,7921
23	0,1271	0,6853	1,0603	1,3195	1,7139	2,0687	2,4999	2,8073	3,7676
24	0,1270	0,6848	1,0593	1,3178	1,7109	2,0639	2,4922	2,7969	3,7454
25	0,1269	0,6844	1,0584	1,3163	1,7081	2,0595	2,4851	2,7874	3,7251
26	0,1269	0,6840	1,0575	1,3150	1,7056	2,0555	2,4786	2,7787	3,7066
27	0,1268	0,6837	1,0567	1,3137	1,7033	2,0518	2,4727	2,7707	3,6896
28	0,1268	0,6834	1,0560	1,3125	1,7011	2,0484	2,4671	2,7633	3,6739
29	0,1268	0,6830	1,0553	1,3114	1,6991	2,0452	2,4620	2,7564	3,6594
30	0,1267	0,6828	1,0547	1,3104	1,6973	2,0423	2,4573	2,7500	3,6460
40	0,1265	0,6807	1,0500	1,3031	1,6839	2,0211	2,4233	2,7045	3,5510
60	0,1262	0,6786	1,0455	1,2958	1,6706	2,0003	2,3901	2,6603	3,4602
80	0,1261	0,6776	1,0432	1,2922	1,6641	1,9901	2,3739	2,6387	3,4163
120	0,1259	0,6765	1,0409	1,2886	1,6577	1,9799	2,3578	2,6174	3,3735
∞	0,1257	0,6745	1,0364	1,2816	1,6449	1,9600	2,3263	2,5758	3,2905

Lorsque $\nu = \infty$, $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.

Remarque 3.13. *Soit un entier $\nu \geq 2$ et soient $X_1, X_2, \ldots, X_\nu$ ν variables aléatoires* **indépendantes** *qui sont* **toutes de même loi normale** $\mathcal{N}(\mu, \sigma^2)$. *Nous désignons par $\overline{X}_\nu$ la* **moyenne empirique (ou moyenne arithmétique)** *de $X_1, X_2, \ldots, X_\nu$ qui est définie, de la même façon que dans (3.22), par*

$$\overline{X}_\nu = \frac{1}{\nu} \sum_{i=1}^{\nu} X_i \,.$$

Nous désignons par V_ν la **variance empirique** *de $X_1, X_2, \ldots, X_\nu$ qui est définie par*

$$V_\nu = \frac{1}{\nu} \sum_{i=1}^{\nu} \left(X_i - \overline{X}_\nu\right)^2 = \left(\frac{1}{\nu} \sum_{i=1}^{\nu} X_i^2\right) - \left(\overline{X}_\nu\right)^2 . \tag{3.25}$$

Alors, la variable aléatoire

$$W = \sqrt{\nu - 1} \left(\frac{\overline{X}_\nu - \mu}{\sqrt{V_\nu}}\right) \tag{3.26}$$

suit une loi de Student à $\nu - 1$ degrés de liberté.

3.6 Construction d'intervalles de confiance

On s'intéresse à une variable statistique quantitative continue qui est modélisée par une variable aléatoire, notée par exemple par X, de loi normale $\mathcal{N}(\mu, \sigma^2)$ dont la vraie valeur [4] de l'un des deux paramètres μ ou σ est inconnue ; il se peut aussi que les vraies valeurs de μ et de σ soient toutes les deux inconnues. A partir de l'observation d'un nombre de valeurs pas très grand de cette variable statistique on souhaite construire un intervalle, appelé *intervalle de confiance*, qui soit le moins large possible et qui en même temps a de grandes chances de contenir la vraie valeur inconnue de μ ou de σ. La probabilité que cette valeur inconnue appartiennent à l'intervalle de confiance est appelée le *niveau de confiance.* Grâce à l'Exemple 3.3, qui sera présenté par la suite, nous étudierons les trois cas figure suivant :

1. La vraie valeur de l'écart-type σ est connue, mais celle de la moyenne μ est inconnue ; on cherche alors à construire un intervalle de confiance pour la vraie valeur inconnue de μ en tirant profit du fait que la vraie valeur de σ est connue.
2. La vraie valeur de la moyenne μ est connue, mais celle de l'écart-type σ est inconnue ; on cherche alors à construire un intervalle de confiance pour la vraie valeur inconnue de σ en tirant profit du fait que la vraie valeur de μ est connue.
3. Les vraies valeurs de μ et de σ sont toutes les deux inconnues ; on cherche alors à construire deux intervalles de confiance, l'un pour la vraie valeur de μ et l'autre pour la vraie valeur de σ.

Exemple 3.3. (important) *On admet que la détente sèche, mesurée en centimètres (cm), d'un basketteur, choisi au hasard, est une variable aléatoire, désignée par X, qui suit une loi normale de moyenne $\mu > 0$ et d'écart-type $\sigma > 0$. Les détentes sèches de six basketteurs sont : 59,4 57,7 60,5 58,2 58,6 61,0. On peut considérer que ces six valeurs sont une réalisation de six variables aléatoires $X_1, \ldots, X_6$ indépendantes et de même loi que X.*
1) On suppose que la vraie valeur de σ est $2,3$ cm, et on cherche à construire un intervalle de confiance pour la vraie valeur de μ au niveau de confiance 98%.
2) On suppose que la vraie valeur de μ est $58,5$ cm, et on cherche à construire un intervalle de confiance pour la vraie valeur de σ au niveau de confiance 98%.

4. C'est-à-dire la valeur qui rend compte le mieux possible des observations.

3) On suppose que les vraies valeurs de μ et de σ sont toutes les deux inconnues, et on cherche à construire deux intervalles de confiance au niveau de confiance 98%, l'un pour la vraie valeur de μ et l'autre pour la vraie valeurs de σ.

Etude de l'Exemple 3.3 1) Désignons par $\overline{X}_6$ la moyenne empirique de $X_1, \ldots, X_6$ (voir (3.22)), c'est-à-dire que

$$\overline{X}_6 = \frac{1}{6} \sum_{k=1}^{6} X_k \,. \tag{3.27}$$

Signalons que compte tenu des données dont on dispose, $\overline{X}_6$ prend la valeur

$$\overline{X}_6(\omega_0) = \frac{59,4 + 57,7 + 60,5 + 58,2 + 58,6 + 61,0}{6} = 59,23 \text{ cm.} \tag{3.28}$$

D'après la Remarque 3.10, nous savons que $\overline{X}_6$ est de loi normale de moyenne inconnue μ et d'écart-type connu égal à $2,3/\sqrt{6}$. Ainsi, il résulte de la Remarque 3.8 que la variable aléatoire

$$Z = \frac{\overline{X}_6 - \mu}{2,3/\sqrt{6}} \tag{3.29}$$

est de loi normale centrée et réduite. La fonction de répartition de Z est notée par F_Z. Cherchons le nombre positif a qui vérifie

$$\mathbb{P}(-a \leq Z \leq a) = 0,98 \tag{3.30}$$

D'après (3.11) et (3.19), on a

$$\mathbb{P}(-a \leq Z \leq a) = F_Z(a) - F_Z(-a) = 2F_Z(a) - 1$$

Ainsi, (3.30) se ramène à $F_Z(a) = 0,99$, et en utilisant la table de loi normale centrée et réduite on trouve que $a \simeq 2,33$. Par ailleurs, il résulte de (3.29) que

$$\begin{aligned} &-a \leq Z \leq a \Longleftrightarrow -a \leq \frac{\mu - \overline{X}_6}{2,3/\sqrt{6}} \leq a \Longleftrightarrow -a \times (2,3/\sqrt{6}) \leq \mu - \overline{X}_6 \leq a \times (2,3/\sqrt{6}) \\ &\Longleftrightarrow \overline{X}_6 - a \times (2,3/\sqrt{6}) \leq \mu \leq \overline{X}_6 + a \times (2,3/\sqrt{6}) \,. \end{aligned}$$

Ainsi, l'intervalle de confiance pour la vraie valeur de μ qu'on recherche est

$$\Big[\overline{X}_6(\omega_0) - a \times (2,3/\sqrt{6})\,;\, \overline{X}_6(\omega_0) + a \times (2,3/\sqrt{6})\Big] \,.$$

Compte tenu des données dont on dispose cet intervalle est $\big[57,04\,;61,42\big]$.

2) Désignons par H_6 la variable aléatoire définie par

$$H_6 = \sum_{k=1}^{6} \big(X_k - 58,5\big)^2 \,. \tag{3.31}$$

Signalons que compte tenu des données dont on dispose, H_6 prend la valeur

$$\begin{aligned} H_6(\omega_0) = &(59,4 - 58,5)^2 + (57,7 - 58,5)^2 + (60,5 - 58,5)^2 \\ &+(58,2 - 58,5)^2 + (58,6 - 58,5)^2 + (61,0 - 58,5)^2 = 11,8 \,. \end{aligned}$$

D'après la Remarque 3.11 (*iii*), nous savons que la variable aléatoire H_6/σ^2 suit une loi du χ^2 à 6 degrés de liberté. Cherchons deux nombres positifs b et c qui vérifient

$$\mathbb{P}\left(c \leq \frac{H_6}{\sigma^2} \leq b\right) = 0,98\,. \tag{3.32}$$

Il résulte de (3.11) que

$$\mathbb{P}\left(c \leq \frac{H_6}{\sigma^2} \leq b\right) = \mathbb{P}\left(\frac{H_6}{\sigma^2} \leq b\right) - \mathbb{P}\left(\frac{H_6}{\sigma^2} \leq c\right)\,.$$

Ainsi, en utilisant le fait que, pour tout nombre réel t, on a

$$\mathbb{P}\left(\frac{H_6}{\sigma^2} \leq t\right) = 1 - \mathbb{P}\left(\frac{H_6}{\sigma^2} > t\right)\,,$$

(3.32) se ramène à

$$\mathbb{P}\left(\frac{H_6}{\sigma^2} > c\right) - \mathbb{P}\left(\frac{H_6}{\sigma^2} > b\right) = 0,98\,. \tag{3.33}$$

Remarquons alors que pour obtenir (3.33), il suffit que l'on ait

$$\mathbb{P}\left(\frac{H_6}{\sigma^2} > c\right) = 0,99 \quad \text{et} \quad \mathbb{P}\left(\frac{H_6}{\sigma^2} > b\right) = 0,01\,.$$

Ainsi, en utilisant la table de lois du χ^2 on trouve que $c \simeq 0,872$ et $b \simeq 16,812$. Remarquons enfin que

$$c \leq \frac{H_6}{\sigma^2} \leq b \Longleftrightarrow \frac{H_6}{b} \leq \sigma^2 \leq \frac{H_6}{c} \Longleftrightarrow \sqrt{\frac{H_6}{b}} \leq \sigma \leq \sqrt{\frac{H_6}{c}}\,.$$

Ainsi, l'intervalle de confiance pour la vraie valeur de σ qu'on recherche est :

$$\left[\sqrt{\frac{H_6(\omega_0)}{b}}\,;\,\sqrt{\frac{H_6(\omega_0)}{c}}\,\right]\,.$$

Compte tenu des données dont on dispose, cet intervalle est $[0,837\,;\,3,679]$.

3) Nous allons d'abord déterminer l'intervalle de confiance pour la vraie valeur de σ. Désignons par K_5 la variable aléatoire définie par

$$K_5 = \sum_{k=1}^{6} \left(X_k - \overline{X}_6\right)^2\,. \tag{3.34}$$

Signalons que compte tenu des données dont on dispose et compte tenu de (3.28), K_5 prend la valeur

$$\begin{aligned} K_5(\omega_0) = {} & (59,4 - 59,23)^2 + (57,7 - 59,23)^2 + (60,5 - 59,23)^2 \\ & + (58,2 - 59,23)^2 + (58,6 - 59,23)^2 + (61,0 - 59,23)^2 \simeq 8,574\,. \end{aligned}$$

D'après la Remarque 3.11 (*iv*), nous savons que la variable aléatoire K_5/σ^2 suit une loi du χ^2 à 5 degrés de liberté. Cherchons deux nombres positifs d et e qui vérifient

$$\mathbb{P}\left(e \leq \frac{K_5}{\sigma^2} \leq d\right) = 0,98\,. \tag{3.35}$$

Il résulte de (3.11) que

$$\mathbb{P}\left(e \leq \frac{K_5}{\sigma^2} \leq d\right) = \mathbb{P}\left(\frac{K_5}{\sigma^2} \leq d\right) - \mathbb{P}\left(\frac{K_5}{\sigma^2} \leq e\right).$$

Ainsi, en utilisant le fait que, pour tout nombre réel t, on a

$$\mathbb{P}\left(\frac{K_5}{\sigma^2} \leq t\right) = 1 - \mathbb{P}\left(\frac{K_5}{\sigma^2} > t\right),$$

(3.35) se ramène à

$$\mathbb{P}\left(\frac{K_5}{\sigma^2} > e\right) - \mathbb{P}\left(\frac{K_5}{\sigma^2} > d\right) = 0,98. \tag{3.36}$$

Remarquons alors que pour obtenir (3.36), il suffit que l'on ait

$$\mathbb{P}\left(\frac{K_5}{\sigma^2} > e\right) = 0,99 \quad \text{et} \quad \mathbb{P}\left(\frac{K_5}{\sigma^2} > d\right) = 0,01.$$

Ainsi, en utilisant la table de lois du χ^2 on trouve que $e \simeq 0,554$ et $d \simeq 15,086$. Remarquons enfin que

$$e \leq \frac{K_5}{\sigma^2} \leq d \Longleftrightarrow \frac{K_5}{d} \leq \sigma^2 \leq \frac{K_5}{e} \Longleftrightarrow \sqrt{\frac{K_5}{d}} \leq \sigma \leq \sqrt{\frac{K_5}{e}}.$$

Ainsi, l'intervalle de confiance pour la vraie valeur de σ qu'on recherche est :

$$\left[\sqrt{\frac{K_5(\omega_0)}{d}}\,;\,\sqrt{\frac{K_5(\omega_0)}{e}}\right].$$

Compte tenu des données dont on dispose, cet intervalle est $[0,753\,;\,3,934]$.

Nous allons maintenant déterminer l'intervalle de confiance pour pour la vraie valeur de μ. Désignons par V_6 la variance empirique de $X_1, \ldots, X_6$ (voir (3.25)), c'est-à-dire que

$$V_6 = \frac{1}{6}\sum_{k=1}^{6}\left(X_k - \overline{X}_6\right)^2 = \frac{K_5}{6}, \tag{3.37}$$

où la dernière égalité résulte de (3.34). Signalons que compte tenu des données dont on dispose et compte tenu de (3.34) V_6 prend la valeur $V_6(\omega_0) \simeq 8,574/6 = 1,429$. Désignons par T_5 la variable aléatoire définie par

$$T_5 = \sqrt{5}\left(\frac{\overline{X}_6 - \mu}{\sqrt{V_6}}\right). \tag{3.38}$$

D'après la Remarque 3.13, nous savons que T_5 suit une loi de Student à 5 degrés de liberté. Cherchons un nombre positif f qui vérifie

$$\mathbb{P}(|T_5| \leq f) = 0,98. \tag{3.39}$$

En utilisant le fait que, pour tout nombre réel t, on a

$$\mathbb{P}(|T_5| \leq t) = 1 - \mathbb{P}(|T_5| > t) = 1 - \mathbb{P}(|T_5| \geq t),$$

(3.39) se ramène à $\mathbb{P}(|T_5| \geq f) = 0,02$. Ainsi, au moyen de la table de lois de Student, on trouve que $f \simeq 3,365$. Par ailleurs, il résulte de (3.38) que

$$|T_5| \leq f \iff -f \leq -T_5 \leq f \iff -f \leq \sqrt{5}\left(\frac{\mu - \overline{X}_6}{\sqrt{V_6}}\right) \leq f \iff -\frac{f}{\sqrt{5}} \leq \frac{\mu - \overline{X}_6}{\sqrt{V_6}} \leq \frac{f}{\sqrt{5}}$$

$$\iff -\frac{f\sqrt{V_6}}{\sqrt{5}} \leq \mu - \overline{X}_6 \leq \frac{f\sqrt{V_6}}{\sqrt{5}} \iff \overline{X}_6 - \frac{f\sqrt{V_6}}{\sqrt{5}} \leq \mu \leq \overline{X}_6 + \frac{f\sqrt{V_6}}{\sqrt{5}}$$

Ainsi, l'intervalle de confiance pour pour la vraie valeur de μ qu'on recherche est :

$$\left[\overline{X}_6(\omega_0) - \frac{f\sqrt{V_6(\omega_0)}}{\sqrt{5}}\,;\, \overline{X}_6(\omega_0) + \frac{f\sqrt{V_6(\omega_0)}}{\sqrt{5}}\right].$$

Compte tenu des données dont on dispose, cet intervalle est $[57,43\,;\,61,03]$.

4 Analyse bivariée

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

4.1 Liaison entre deux variables quantitatives

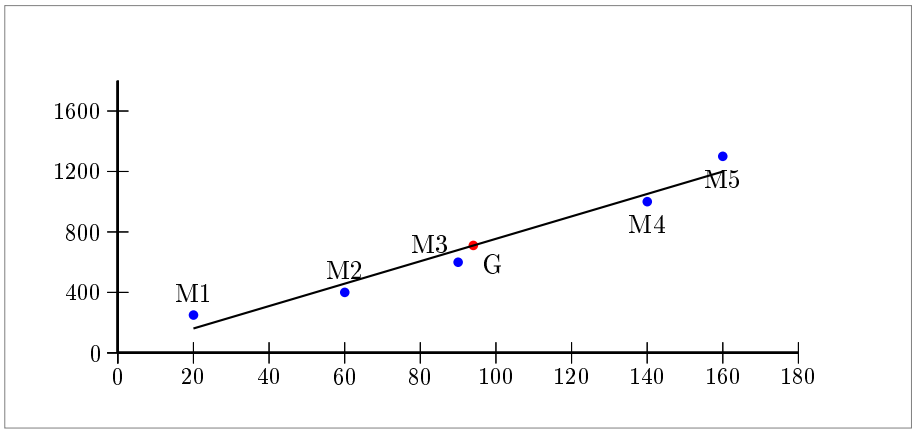
4.1.1 La régression linéaire simple

Exemple 4.1. *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 points** : 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 (qu'on appelle aussi droite de régression)**, c'est-à-dire de la droite qui permet d'ajuster au mieux, au sens du critère des moindres carrés, le nuage de points $M_1 = (x_1, y_1)$; $M_2 = (x_2, y_2)$; ... ; $M_N = (x_N, y_N)$ 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 :

$$a = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \ldots + (x_N - \overline{x})(y_N - \overline{y})}{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_N - \overline{x})^2} \tag{4.1}$$

et

$$b = \overline{y} - a\overline{x}\,. \tag{4.2}$$

La formule (4.2) signifie que la droite des moindres carrés passe par **le centre de gravité** du nuage de points $M_1 = (x_1, y_1)$; $M_2 = (x_2, y_2)$; ... ; $M_N = (x_N, y_N)$, 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 :

$$\widehat{y}_i = ax_i + b\,; \tag{4.3}$$

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 (c'est-à-dire la valeur observée) de Y lorsque X vaut x_i.

Exercice 4.1. *Montrer que, de façon générale, la moyenne arithmétique de $y_1, y_2, \ldots, y_n$ est égale à celle de $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_n$.*

Appliquons maintenant dans le cas de l'Exemple 4.1 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 associés 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

(4.4)

ainsi, grâce aux formules (4.1) et (4.2), on trouve que :

$$a = \frac{97300}{13120} \simeq 7,416 \quad \text{et} \quad b = 710 - 7,416 \times 94 \simeq 12,896\,; \tag{4.5}$$

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 (4.3) et à (4.5), 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 (c'est-à-dire les prix observés) 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 (observée) 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)	$\widehat{y}_1 = 161$	$\widehat{y}_2 = 458$	$\widehat{y}_3 = 680$	$\widehat{y}_4 = 1051$	$\widehat{y}_5 = 1199$

Remarque 4.1. *Signalons que de façon générale $VT(Y)$, $VE(Y)$ et $VR(Y)$,* **la variation totale, la variation expliquée et la variation résiduelle de variable** Y*, sont définies par :*

$$VT(Y) = (y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \ldots + (y_N - \overline{y})^2 = N \times \mathrm{Var(Y)}\,, \tag{4.6}$$

$$VE(Y) = (\widehat{y}_1 - \overline{y})^2 + (\widehat{y}_2 - \overline{y})^2 + \ldots + (\widehat{y}_N - \overline{y})^2\,, \tag{4.7}$$

et

$$VR(Y) = (y_1 - \widehat{y}_1)^2 + (y_2 - \widehat{y}_2)^2 + \ldots + (y_N - \widehat{y}_n)^2\,. \tag{4.8}$$

L'importante égalité suivante, qui rappelle le théorème de Pythagore, est toujours vérifiée :

$$VT(Y) = VE(Y) + VR(Y)\,. \tag{4.9}$$

Le coefficient de détermination, *qui est noté par $R^2(X,Y)$, est défini par le ratio :*

$$R^2(X,Y) = \frac{VE(Y)}{VT(Y)}\,. \tag{4.10}$$

Il résulte de (4.9) que $R^2(X,Y)$ est **toujours compris entre** 0 **et** 1**.**

Exercice 4.2. *Calculer $VT(Y)$, $VE(Y)$, $VR(Y)$ et $R^2(X,Y)$ dans le cas de l'Exemple 4.1.*

4.1.2 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 $\mathrm{cov(X, Y)}$, (ou $\mathrm{Cov(X, Y)}$) est définie par :

$$\mathrm{cov(X, Y)} = \frac{(\mathrm{x_1} - \overline{\mathrm{x}})(\mathrm{y_1} - \overline{\mathrm{y}}) + (\mathrm{x_2} - \overline{\mathrm{x}})(\mathrm{y_2} - \overline{\mathrm{y}}) + \ldots + (\mathrm{x_N} - \overline{\mathrm{x}})(\mathrm{y_N} - \overline{\mathrm{y}})}{\mathrm{N}}\,, \tag{4.11}$$

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) :

$$\mathrm{cov(X, Y)} = \left(\frac{\mathrm{x_1y_1 + x_2y_2 + \ldots + x_Ny_N}}{\mathrm{N}}\right) - \overline{\mathrm{x}}\,\overline{\mathrm{y}}\,; \tag{4.12}$$

en fait la formule (2.2) n'est rien d'autre que la formule (4.12) dans le cas particulier où $X = Y$.

Exemple 4.2. *Soient X et Y les variables « Superficie » et « Prix », dont il est question dans l'Exemple 4.1 (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 (4.11), et la seconde consiste à à utiliser la formule (4.12).*

Présentons d'abord la première méthode. On a déjà vu que (voir le tableau (4.4)) :

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

ainsi, il résulte de la formule (4.11) 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 :*

$$\frac{x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + x_5y_5}{5} = \frac{431000}{5} = 86200\,. \tag{4.13}$$

D'autre part, dans la Sous-section 4.1.1, on a vu que $\overline{x} = 94$ *et* $\overline{y} = 710$ *; on a par conséquent :*

$$\overline{x}\,\overline{y} = 94 \times 710 = 66740\,. \tag{4.14}$$

Finalement, en utilisant la formule (4.12), ainsi que (4.13) et (4.14), on obtient :

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

Remarque 4.2. (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 particulier de l'Exemple 4.1 (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 (4.4)) :

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

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 4.1, 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

$$R(X,Y) = \frac{\mathrm{cov(X,Y)}}{\sigma_X\sigma_Y}\,. \tag{4.15}$$

Ainsi, dans le cas de l'Exemple 4.1, on a

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

Remarque 4.3. (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*) *Lorsqu'on multiplie par lui-même le coefficient de corrélation linéaire (défini par (4.15)) on obtient le coefficient de détermination (défini par (4.10)) ; c'est ce qui justifie la notation* $R^2(X,Y)$ *pour ce dernier coefficient.*

(*iii*) *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 4.4. (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 reposant notamment sur le test Student et celui Fisher-Snedecor, qui dépasse le cadre de ce cours, s'impose.*

4.2 Liaison entre deux variables qualitatives

4.2.1 Tableau de contingence

Exemple : On dispose d'une enquête de l'INSEE sur les établissements industriels et commerciaux de plus de 10 salariés en 1986. On cherche s'il existe un lien entre la taille d'un établissement (c'est-à-dire son effectif, autrement dit son nombre de salariés) et sa localisation géographique (c'est-à-dire la région où un établissement est situé).

On considère que la variable « Classe d'Effectif des Etablissements » est qualitative ordinale, et que ses 5 modalités sont les classes : 10-49, 50-199, 200-499, 500-1999 et plus de 2000 salariés.

La variable « Régions » est clairement qualitative nominale ; elle dispose de 22 modalités qui correspondent aux 22 régions métropolitaines à l'époque de l'enquête.

Les 218645 établissements industriels et commerciaux de plus de 10 salariés recensés à l'époque par l'INSEE se répartissent en fonction de leur localisation géographique et de leur classe d'effectif comme l'indique le Tableau 1 ci-après. Un tel tableau s'appelle **tableau de contingence** ou encore **tableau croisé**.

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

REGIONS	TABLEAU DES EFFECTIFS OBSERVES Classe d'effectif des établissements 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

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 trouve 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 Ile de France. 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 : $80,5\%$ ont moins de 50 salariés(puisque $f_{1\bullet} = \frac{176004}{218645} \simeq 0,805$) et seul $0,1\%$ en ont plus de 2000 (puisque $f_{5\bullet} = \frac{233}{218645} \simeq 0,001$).

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

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

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

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

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 $f_{1|i} + f_{2|i} + \ldots + f_{l|i} = 1$. Signalons au passage que $f_{j|i}$ 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 ci-après 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.

Tableau 2 (Profils lignes)

REGIONS	PROFILS LIGNES Classe d'effectif des établissements 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)

REGIONS	PROFILS COLONNES Classe d'effectif des établissements 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%	

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

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 Ile 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 $n_{i_0j_0}$ est <u>plus grand que</u> l'effectif théorique $n_{i_0j_0}^*$ ($n_{i_0j_0} > n_{i_0j_0}^*$) on dit qu'il y a <u>attraction</u> 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 $n_{i_1j_1}$ est <u>plus petit</u> que l'effectif théorique $n_{i_1j_1}^*$ ($n_{i_1j_1} < n_{i_1j_1}^*$) on dit qu'il y a <u>répulsion</u> 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 Ile 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 variable 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)

REGIONS	EFFECTIFS THEORIQUES Classe d'effectif des établissements 10-49	50-199	200-499	500-199	+2000	Profil 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)

REGIONS	TABLEAU DES χ2 Classe d'effectif des établissements 10-49	50-199	200-499	500-199	+2000	Total 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 la distance du 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 la distance du 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 se trouvant dans la ligne Total.

La méthode permettant de savoir s'il existe ou non une liaison significative entre les deux variables qualitatives Z et T repose sur la remarque fondamentale suivante :

Remarque 4.5. (Test du χ^2 d'indépendance) *De façon générale, considérons un tableau de contingence à k lignes et à l colonnes, où $k \geq 2$ et $l \geq 2$. Nous posons*

$$\nu = (k-1) \times (l-1)\,,$$

et nous nous plaçons, théoriquement parlant, **sous l'hyposthèse, désignée par $\mathcal{H}_0$, et appelée hypothèse nulle, suivante : *il n'y a pas de liaison significative entre les deux variables qualitatives associées au tableau de contingence.***

Alors, **la distance du χ^2 suit approximativement une loi du χ^2 à ν degrés de liberté (voir la Définition 3.7), sous réserve que les trois conditions suivantes soient remplies** *:*

1. *on a $n \geq 20$ (n est l'effectif total) ;*
2. *on a $n_{i\bullet} \geq 5$, pour tout i, et $n_{\bullet j} \geq 5$, pour tout j ;*
3. *on a $n^*_{ij} \geq 5$ pour au moins 80% des cases (i, j) du tableau des effectifs théoriques.*

Remarque 4.6. *Le contraire de l'hypothèse nulle $\mathcal{H}_0$, c'est-à-dire l'affirmation* il existe une liaison significative entre les deux variables qualitatives associées au tableau de contingence*, est appelé* **l'hypothèse alternative $\mathcal{H}_1$**.

Dans le cas de notre exemple, les trois conditions de la Remarque 4.5 sont clairement remplies ; ainsi, en se plaçant sous l'hypothèse [5] nulle $\mathcal{H}_0$: *absence d'une liaison significative entre la variable « Classe d'Effectif des Etablissements » et la variable « Région »*, nous pouvons alors supposer que la distance du χ^2 suit approximativment une loi du χ^2 à $(22-1)\times(5-1)=84$ degrés de liberté. D'autre part, la valeur de la distance du χ^2 qui a été obtenue au moyen des données dont nous disposons est $842,0$. Si cette distance suivait vraiment une loi du χ^2 à 84 degrés de liberté, ou même à 90 degrés de liberté, alors la probabilité qu'elle dépasse la valeur seuil $137,2084$ est moins que $0,001$, c'est-à-dire moins qu'un millième (voir la table de lois du χ^2 et la Remarque 3.11). Au vu de cette analyse, nous sommes conduits à rejeter l'hypothèse nulle $\mathcal{H}_0$, et donc à accepter l'hypothèse alternative $\mathcal{H}_1$: *il existe une liaison significative entre la variable « Classe d'Effectif des Etablissements » et la variable « Région »*.

Avant de clore ce chapitre, nous allons examiner de façon plus attentive 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	$\mathrm{CTR_{11}}$	$\cdots$	$\mathrm{CTR_{1j}}$	$\cdots$	$\mathrm{CTR_{1l}}$	$\mathrm{CTR_{1\bullet}}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_i	$\mathrm{CTR_{i1}}$	$\cdots$	$\mathrm{CTR_{ij}}$	$\cdots$	$\mathrm{CTR_{il}}$	$\mathrm{CTR_{i\bullet}}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
z_k	$\mathrm{CTR_{k1}}$	$\cdots$	$\mathrm{CTR_{kj}}$	$\cdots$	$\mathrm{CTR_{kl}}$	$\mathrm{CTR_{k\bullet}}$
Total	$\mathrm{CTR_{\bullet 1}}$	$\cdots$	$\mathrm{CTR_{\bullet j}}$	$\cdots$	$\mathrm{CTR_{\bullet l}}$	100%

5. En fait la démarche consiste à chercher à confronter la validité de cette hypothèse avec la réalité des données dont nous disposons.

(4)

Tableau 6 (tableau des contributions)

REGIONS	TABLEAU DES CONTRIBUTIONS Classe d'effectif des établissements 10-49	50-199	200-499	500-1999	+2000	Contributio 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%

5 Exercices sur l'analyse statistique univariée

Exercice 5.1. *Suite à une enquête concernant les revenus mensuels en euros de* 105 *ménages français on a obtenu le tableau de répartition suivant :*

Classe de Revenus	$[1000; 1300]$	$]1300; 1600]$	$]1600; 1900]$	$]1900; 2200]$	$]2200; 2500]$	$]2500; 2800]$
Effectifs	$n_1 = 15$	$n_2 = 21$	$n_3 = 28$	$n_4 = 21$	$n_5 = 12$	$n_6 = 8$

1) Calculer la fréquence de chaque classe de la variable « Revenu ».
2) Tracer l'histogramme des fréquences de la variable « Revenu ».
3) Calculer la moyenne arithmétique de la variable « Revenu ».
4) Calculer la fréquence cumulée de chaque classe de la variable « Revenu ».

Exercice 5.2. *Une compagnie de vérificateurs-comptables, durant la période des vérifications annuelles, a dressé une statistique des durées de vérification de 50 bilans et comptes de fin d'année. Les résultats sont résumés dans le tableau suivant*

variable classée "durée de vérification" (en minutes)	*[10, 20]*	*]20,30]*	*]30,40]*	*]40,50]*	*]50,60]*
Effectifs	*3*	*5*	*10*	*12*	*20*

1) Calculer les fréquences, les effectifs cumulés et les fréquences cumulées correspondants aux différentes classes de la variable « durée de vérification ».
2) Représenter l'histogramme des fréquences correspondant à cette variable et déterminer sa classe modale.
3) Calculer la moyenne arithmétique de cette variable.
4) Calculer la variance et l'écart-type de cette variable.
5) a) Expliquer la méthode qui permet de tracer le graphe de la fonction cumulative associée à cette variable puis tracer ce graphe.
b) Au moyen d'une méthode graphique, déterminer une valeur approximative de la médiane de cette variable.
6) Au moyen de calculs, déterminer une valeur plus précise de cette médiane.

Exercice 5.3. *La distribution des revenus mensuels nets, exprimés en centaines d'euros, d'une promotion de* 153 *étudiants d'une certaine école de commerce, est synthétisée par le tableau suivant :*

variable classée "revenus"	*[20, 25]*	*]25,30]*	*]30,35]*	*]35,40]*	*]40,45]*	*]45,50]*
effectifs	*13*	*33*	*45*	*39*	*14*	*9*

1) Calculer les fréquences, les effectifs cumulés et les fréquences cumulées correspondants aux différentes classes de la variable classée « revenus ».
2) a) Représenter l'histogramme des effectifs correspondant à cette variable.
b) Déterminer la classe modale.
c) Déterminer graphiquement la valeur exacte du mode.
3) Calculer la moyenne arithmétique de la variable classée « revenus ».

4) Calculer la variance et l'écart-type de cette variable.
5) a) Expliquer la méthode qui permet de tracer le graphe de la fonction cumulative associée à cette variable puis tracer ce graphe.
b) Au moyen d'une méthode graphique, déterminer une valeur approximative de la médiane de cette variable.
6) Au moyen de calculs, déterminer une valeur plus précise de cette médiane.

Exercice 5.4. *1) On considère les quatre nombres suivants :* $x_1 = 2,1$ *;* $x_2 = 4,3$ *;* $x_3 = 5,7$ *et* $x_4 = 7,3$*. Calculer leur moyenne géométrique, leur moyenne quadratique et leur moyenne harmonique.*
2) Un étudiant a trouvé que la moyenne géométrique de six nombres vaut $10,22$ *et que leur moyenne arithmétique est égale à* $7,32$*. L'étudiant a-t-il fait une erreur de calcul (justifier votre réponse) ?*
3) Le chiffre d'affaires de l'entreprise « Durand » a augmenté de 80% au cours de l'année 2002 et de 30% au cours de l'année 2003. Que vaut le taux de croissance annuel moyen du chiffre d'affaires de cette entreprise pour la période de ces deux ans ?

Exercice 5.5. (issu de l'examen de L2 (2014)) *Une enquête a été réalisée sur un échantillon de 170 voyageurs, concernant le temps d'attente (mesuré en minutes) de chacun d'entre eux, pour accéder à un guichet d'une certaine gare, à une heure de pointe. Cette enquête a permis d'obtenir le tableau suivant qui fournit la répartition des effectifs des voyageurs selon les classes de la variable continue classée temps d'attente que l'on note par* T*.*

Classe de T	$0 \leq T \leq 5$	$5 < T \leq 10$	$10 < T \leq 15$	$15 < T \leq 20$
Effectif	$n_1 = 45$	$n_2 = 51$	$n_3 = 54$	$n_4 = 20$

1) a) Quelle est la classe modale de la variable T *?*
b) Tracer l'histogramme des effectifs de la variable T*.*
c) Déterminer graphiquement la valeur exacte du mode.
2) Calculer la moyenne arithmétique de la variable T*, sa variance et son écart-type.*
3) a) Calculer les fréquences des classes de la variable T*.*
b) Calculer les effectifs cumulés des classes de la variable T*.*
c) Calculer les fréquences cumulées des classes de la variable T*.*
4) Expliquer la méthode qui permet de tracer le graphe de la fonction cumulative associée à la variable T*, puis tracer ce graphe.*
5) Au moyen d'une méthode graphique, déterminer approximativement la médiane et le troisième quartile de la variable T*.*
6) Au moyen de calculs, déterminer de façon plus précise la médiane et le troisième quartile de la variable T*.*

Exercice 5.6. (issu de l'examen de L2 (2015)) *Les deux questions de cet exercice sont complètement indépendantes l'une de l'autre (on vous demande de donner les résultats de vos calculs avec, au plus, 2 chiffres après la virgule).*
1) Un automobiliste parcourt d'abord 50 kilomètres à 60 km/h, il parcourt ensuite 90 kilomètres à 120 km/h, et enfin il parcourt 10 kilomètres à 50 km/h. On note par v_m *sa vitesse moyenne en km/h sur l'ensemble de ce trajet de 150 kilomètres ; calculer* v_m*.*
2) Soit X *une variable quantitative discrète dont la moyenne arithmétique vaut* $10,34$ *et dont l'écart-type vaut* $4,51$*. Calculer la moyenne quadratique de* X*.*

Exercice 5.7. (issu de l'examen de L2 (2016)) *Une entreprise de fabrication d'appareils électroménagers a effectué une étude statistique concernant les durées de vie (mesurées en nombre*

d'années) d'un échantillon de 133 machines à laver d'un certain modèle. Cette étude a permis d'obtenir le tableau suivant qui fournit la répartition des effectifs des machines à laver selon les classes de la variable continue classée « durée de vie », que l'on note par D.

Classe de D	$0 \leq D \leq 2$	$2 < D \leq 4$	$4 < D \leq 6$	$6 < D \leq 8$	$8 < D \leq 10$
Effectif	$n_1 = 12$	$n_2 = 15$	$n_3 = 61$	$n_4 = 27$	$n_5 = 18$

1) Calculer les effectifs cumulés des classes de la variable D.
2) Calculer les fréquences des classes de la variable D, puis calculer les fréquences cumulées (on donnera les résultats sous forme de pourcentages).
3) Calculer la moyenne arithmétique de la variable D.
4) Calculer la variance de la variable D et son écart-type.

Exercice 5.8. (issu de l'examen de L2 (2017)) *Une entreprise de fabrication d'appareils électroménagers a effectué une étude statistique concernant les durées de vie (mesurées en nombre d'années) d'un échantillon de 233 machines à laver d'un certain modèle. Cette étude a permis d'obtenir le tableau suivant qui fournit la répartition des effectifs des machines à laver selon les classes de la variable continue classée « durée de vie », que l'on note par V.*

Classe de V	$0 \leq V \leq 2$	$2 < V \leq 4$	$4 < V \leq 6$	$6 < V \leq 8$	$8 < V \leq 10$
Effectif	$n_1 = 32$	$n_2 = 35$	$n_3 = 81$	$n_4 = 47$	$n_5 = 38$

1) Déterminer la classe modale de V.
2) Calculer les quartiles de V puis calculer l'intervalle interquartile (IIQ).

Exercice 5.9. (issu de l'examen de L2 (2019)) *Entre le 1 janvier et le 30 avril 2018, Jean a relevé tous les jours, à 9 heures très précisément, la température en degrés Celsius affichée par son thermomètre placé à l'entrée de son jardin. Ces 120 relevés de températures sont résumés dans le tableau suivant qui fournit la répartition des effectifs (c'est-à-dire les nombres de jours) selon les classes de la variable continue classée température (en degrés Celsius) que l'on note par T*

Classes	$-5 \leq T \leq 0$	$0 < T \leq 5$	$5 < T \leq 10$	$10 < T \leq 15$	$15 < T \leq 20$
effectifs	*11*	*20*	*46*	*24*	*19*

1) Calculer les fréquences, les effectifs cumulés et les fréquences cumulées correspondants aux différentes classes de T.
2) Expliquer la méthode qui permet de tracer le graphe de la fonction cumulative associée à T, puis tracer ce graphe.
3) Au moyen d'une méthode graphique, déterminer approximativement chacun des trois quartiles Q_1, Q_2 et Q_3 de T.
4) Au moyen de calculs, déterminer de façon plus précise Q_1, Q_2 et Q_3, puis calculer l'intervalle interquartile (IIQ).
5) Calculer la moyenne arithmétique, la variance et l'écart-type de T.

Exercice 5.10. (issu de l'examen de L3M2S (2019)) *La distribution des montants en dizaines d'euros des factures de 1756 personnes pour leurs achats dans un hypermarché est synthétisée par le tableau suivant :*

variable classée "facture"	*[0 ; 30]*	*]30 ; 60]*	*]60 ; 100]*	*]100 ; 200]*	*]200 ; 300]*
effectifs	*503*	*640*	*414*	*135*	*64*

1) Calculer les trois quartiles et l'intervalle interquartile de la variable classée "facture".
2) Calculer la moyenne arithmétique, la variance et l'écart-type de cette variable.
3) On désigne par h_1 la hauteur du rectangle de l'histogramme des fréquences de cette variable

qui correspond à la classe $[0\,;30]$*, et on désigne par* h_5 *la hauteur de celui qui correspond à la classe* $[200\,;300]$*. Calculer le ratio* h_1/h_5*.*

6 Exercices sur le calcul de probabilités

Exercice 6.1. *Dans un centre de vacances accueillant 120 personnes, 2 sports (tennis et canoë) sont proposés aux vacanciers. On sait que 24 personnes font du tennis et 15 du canoë. En outre, 6 pratiquent à la fois le tennis et le canoë.*
1) Combien de personnes ne pratiquent aucun de ces 2 sports ?
2) On interroge au hasard une personne de ce centre. Quelle est la probabilité d'avoir choisi :
a) une personne faisant du tennis ?
b) une personne faisant uniquement l'un des 2 sport ?

Exercice 6.2. *Sur 100 personnes qui ont posé leur candidature à un poste de direction d'une importante société industrielle, 55 ont de l'expérience professionnelle mais ne possèdent pas de diplôme de deuxième cycle, 35 possèdent un tel diplôme mais n'ont pas d'expérience professionnelle et les 10 restant possèdent un tel diplôme ainsi que de l'expérience professionnelle. On choisit au hasard l'une de ces 100 personnes.*
1) Quelle est la probabilité que cette personne ne possède pas de l'expérience professionnelle. Quelle la probabilité que cette personne ne possède pas de diplôme de deuxième cycle ?
2) Quelle est la probabilité que cette personne possède soit de l'expérience professionnelle soit un diplôme de deuxième cycle (mais pas les deux à la fois) ?

Exercice 6.3. *Un groupe de touristes est composé de 20 hommes (dont 10 européens) et de 30 femmes (dont 20 européennes). On choisit au hasard une personne dans ce groupe, déterminer la probabilité pour qu'elle soit :*
1) du sexe masculin,
2) européenne,
3) un homme non européen.

Exercice 6.4. *Deux chasseurs Antoine et Bernard aperçoivent au même moment un lapin et font feu simultanément. On désigne par A l'événement « Antoine tue le lapin » et par B l'événement « Bernard tue le lapin ». On admet que ces deux événements sont indépendants.* $\mathbb{P}(A)$ *la probabilité de A vaut 0,5 et* $\mathbb{P}(B)$ *la probabilité de B vaut 0,8. Calculer la probabilité de l'événement E : « le lapin est tué ».*

Exercice 6.5. *Le gestionnaire de la restauration scolaire d'une certaine commune a observé que 78% des enfants sont demi-pensionnaires, 10% mangent tous les midis chez une assistante maternelle, les autres rentrent déjeuner chez eux. On choisit au hasard un enfant de cette commune.*
1) Les événements « l'enfant choisi déjeune chez une assistante maternelle » et « l'enfant choisi déjeune chez lui » sont-ils incompatibles ?
2) Quelle est la probabilité des événements suivants :
- *l'enfant choisi est demi-pensionnaire ;*
- *l'enfant choisi déjeune chez lui ;*
- *l'enfant choisi ne déjeune pas chez lui ;*
- *l'enfant choisi déjeune chez une assistante maternelle ou chez lui.*

3) Les événements « l'enfant choisi déjeune chez une assistante maternelle »et « l'enfant choisi déjeune chez lui » sont-ils indépendants ?

Exercice 6.6. *Une ville de 100000 habitants compte trois journaux locaux désignés par A, B et C. 10% des habitants lisent A, 30% lisent B, 5% lisent C, 8% lisent à la fois A et B, 2% lisent*

à la fois A et C, 4% lisent à la fois B et C, 1% lisent à la fois A et B et C.
1) Trouver le nombre de personnes ne lisant qu'un journal.
2) Combien de personnes lisent au moins deux journaux ?
3) B est un quotidien du soir, tandis que A et C sortent le matin. Combien de personnes lisent-elles au moins un journal du matin plus celui du soir ?
4) Combien de personnes lisent-elles un journal du matin seulement et le journal du soir ?

Exercice 6.7. *Un petit atelier comporte trois machines M_1, M_2 et M_3. Pour $i = 1, 2, 3$ on désigne par D_i l'événement « la machine M_i tombe en panne ». On sait que les événements D_1, D_2 et D_3 sont indépendants et que leurs probabilités sont respectivement de : 0,1 ; 0,2 et 0,3. Quelle est la probabilité qu'une seule machine tombe en panne ?*

Exercice 6.8. *L'un des slogans publicitaire concernant un nouveau véhicule automobile est « avec une telle voiture pas de grosse réparation avant 100000 km ». Le service des études techniques du constructeur a cependant fourni au service commercial les probabilités p_1, p_2, p_3, p_4, p_5 d'occurrence avant 100000 km des 5 grosses pannes classiques (ces pannes sont généralement indépendantes l'une de l'autre) : pour les cadrans $p_1 = 0,001$, pour le moteur $p_2 = 0,05$, pour l'embrayage $p_3 = 0,01$, pour les freins $p_4 = 0,013$ et pour la boîte à vitesses $p_5 = 0,03$. Quelle est la probabilité pour que le banc d'essai des revues spécialisées de l'automobile, ou des associations de consommateurs, prennent à défaut la publicité après étude d'une seule voiture ?*

Exercice 6.9. *On choisit au hasard un jeune diplômé d'une certaine formation. La variable aléatoire Y désigne le temps mis par ce jeune diplômé, pour trouver son premier emploi ; Y est exprimée en nombre de mois, et des études statistiques ont montré qu'elle suit une loi normale de moyenne $\mu = 3$ et d'écart-type $\sigma = 0,75$. Calculer les probabilités suivantes : $\mathbb{P}(1,5 \leq Y \leq 4,5)$, $\mathbb{P}(Y \geq 1,5)$, $\mathbb{P}(Y > 4,5)$ et $\mathbb{P}(Y = 3)$.*

Exercice 6.10. *Dans une station de métro, aux heures de pointe, lorsqu'une rame de métro vient de partir, le temps d'attente (exprimé en minute(s)) de la prochaine rame dans la même direstion, est une variable aléatoire désignée par T qui suit une loi normale de moyenne 1,5 et d'écart-type 0,1.*
1) Quelle est la valeur de la probabilité $\mathbb{P}(T = 1,5)$?
2) En utilisant la table de la loi normale centrée et réduite, calculer les probabilités suivantes : a) $\mathbb{P}(T < 1,65)$; b) $\mathbb{P}(T \geq 1,37)$; c) $\mathbb{P}(1,65 \geq T > 1,37)$.

Exercice 6.11. *Lors d'un jour ouvrable choisi au hasard, le temps aléatoire, noté par X et exprimé en minutes, que met Sophie pour se rendre de son domicile à son lieu de travail, suit une loi normale de moyenne 50 minutes et d'écart-type 10 minutes. Calculer les probabilités suivantes : $\mathbb{P}(X \geq 75)$, $\mathbb{P}(40 < X < 55)$, $\mathbb{P}(X = 50)$, $\mathbb{P}(X < 30)$ et $\mathbb{P}(70 \geq X \geq 50)$.*

Exercice 6.12. *La variable aléatoire X désigne l'âge d'une personne choisie au hasard parmi les habitants d'une certaine ville. Les statistiques ont montré que X suit une loi normale de moyenne 32 et d'écart-type 18.*
1) En utilisant la table de la loi normale centrée et réduite, calculer les probabilités suivantes : a) $\mathbb{P}(32 < X \leq 59)$; b) $\mathbb{P}(X > 50)$; c) $\mathbb{P}(X < 18)$.
2) Les propriétaires des salles de cinéma de la ville désirent offrir une réduction aux 5% plus âgés de la ville. En utilisant la table de la loi normale centrée et réduite, trouver l'âge à partir duquel ils offriront cette réduction.

Exercice 6.13. *Une ouvrière couturière travaille dans l'industrie du prêt-à-porter, et fabrique, en série, des tabliers de cuisine. On admet que la durée de la confection d'un tel tablier, en minutes, est une variable aléatoire X suivant une loi normale de moyenne 22 et d'écart-type 2.*

1) Calculer la probabilité que la confection d'un tablier de cuisine par cette couturière mette : a) moins de 23 minutes ; b) moins de 20 minutes ; c) plus de 25 minutes ; d) entre 20 et 25 minutes.
2) La couturière doit produire 8 tabliers par demi-journée. Entre chaque tablier, elle prend 5 minutes de pause. On note X_i la variable aléatoire qui désigne la durée de la confection du i-ème tablier d'une demi-journée. On suppose que, grâce aux pauses, il n'y a pas d'effet de fatigue. Les variables aléatoires X_1, X_2, ..., X_8 peuvent donc être supposées mutuellement indépendantes et de même loi que X.

a) On considère la variable aléatoire $Y = X_1+X_2+\ldots+X_8+35$. Comment peut-on interpréter cette variable ?

b) Trouver la loi de probabilités de Y (justifier votre réponse).

c) L'après-midi la couturière commence à 14h ; quelle est la probabilité qu'elle ait encore à travailler après 17h30 ?

d) Quelle est la probabilité que la couturière consacre plus de 20 minutes à chacun des 8 tabliers confectionnés au cours d'une demi-journée

e) Quelle est la probabilité qu'il y ait plus de 2 minutes d'écart entre le temps consacré à la confection du premier tablier d'une demi-journée et le temps consacré à la confection du dernier ?

Exercice 6.14. *On suppose que pour une machine à laver d'un certain modèle, choisie au hasard, la durée de vie, mesurée en années, est une variable aléatoire, désignée par D, qui suit une loi normale de moyenne $\mu > 0$ et d'écart-type $\sigma > 0$. On a observé que pour sept machines à laver de ce même modèle, qui ont été choisies au hasard, les durées de vie sont : 7,1 8,9 5,2 9,6 4,5 6,3 8,4. On peut considérer que ces sept valeurs sont une réalisation de sept variables aléatoires $D_1, \ldots, D_7$ indépendantes et de même loi que D.*
1) Dans cette question on suppose que la vraie valeur de σ est $2,7$. Construiser un intervalle de confiance pour pour la vraie valeur de μ au niveau de confiance 95%.
2) Dans cette question on suppose que la vraie valeur de μ est $6,4$. Construiser un intervalle de confiance pour pour la vraie valeur de σ au niveau de confiance 95%.
3) Dans cette question on suppose que les vraies valeurs de μ et de σ sont toutes les deux inconnues. Construiser pour chacune d'elles un intervalle de confiance au niveau de confiance de 95%.

Exercice 6.15. *Dans un cabinet médical, la durée aléatoire en minutes que passe un patient dans la salle d'attente est une variable aléatoire, désignée par D, de loi normale de moyenne $\mu > 0$ et d'écart-type sigma > 0. Sophie s'est déjà rendue huit fois dans ce cabinet, et à chaque fois elle a noté avec précision la durée en minutes qu'elle a passée dans la salle d'attente. Les huit durées sont les suivantes : 15 17 27 11 10 8 25 22. On peut considérer que ces huit valeurs sont une réalisation de huit variables aléatoires $D_1, \ldots, D_8$ indépendantes et de même loi que D.*
1) Dans cette question on suppose que la vraie valeur de σ est 5. Construiser un intervalle de confiance pour la vraie valeur de μ au niveau de confiance de 98%.
2) Dans cette question on suppose que la vraie valeur de μ est 15. Construiser un intervalle de confiance pour la vraie valeur de σ au niveau de confiance de 80%.
3) Dans cette question on suppose que les vraies valeurs de μ et de σ sont toutes les deux inconnues. Construiser pour chacune d'elles un intervalle de confiance au niveau de confiance de 80%.

Exercice 6.16. *On admet que pour un jour ouvrable choisi au hasard, le chiffre d'affaires journalier, exprimé en milliers d'euros, d'un certain hypermarché est une variable aléatoire, désignée par X, de loi normale dont la vraie valeur de la moyenne $\mu > 0$ et celle de l'écart-type $\sigma > 0$ sont inconnus. Huit chiffres d'affaires journaliers (en milliers d'euros) de cet hypermarché pour huit*

jours ouvrables du mois dernier sont les suivants : 10,71 8,66 11,02 9,45 7,79 9,98 6,69 11,48. Signalons que l'on peut considérer que ces huit valeurs sont une réalisation de huit variables aléatoires $X_1, \ldots, X_8$ indépendantes et de même loi que X. A partir de ces huit chiffres d'affaires journaliers, construiser des intervalles de confiance pour les vraies valeurs de μ et σ au niveau de confiance 90%.

Exercice 6.17. *Le temps aléatoire X, mesuré en minutes, que met Jean pour faire le trajet de son nouveau domicile à son lieu de travail suit une loi normale dont la vraie valeur de la moyenne μ et celle de l'écart-type σ sont inconnus. Cependant, il a déjà fait dix fois ce trajet et les dix durées correspondantes, exprimées en minutes, sont : 22 17 25 19 27 21 18 20 23 19. Signalons que l'on peut considérer que ces dix durées sont une réalisation de dix variables aléatoires $X_1, \ldots, X_{10}$ indépendantes et de même loi que X. A partir de ces dix durées construiser des intervalles de confiance pour les vraies valeurs de μ et σ, au niveau de confiance* 95%.

Exercice 6.18. (issu de l'examen de L2 (2018)) *La variable aléatoire D désigne, pour un jour ouvrable choisi au hasard, la durée en minutes que met Jean pour se rendre le matin de son domicile à son lieu de travail. On admet que D suit une loi normale de moyenne* 30 *et d'écart-type* 5.
1) Calculer (sous forme de pourcentages) les trois probabilités suivantes :
$\mathbb{P}(D > 28)$ *;* $\mathbb{P}(D < 19)$ *;* $\mathbb{P}(20 < D \leq 40)$.
2) Demain Jean a un important rendez-vous à 9 heures du matin très précisément à son lieu de travail. A quelle heure au plus tard devra-t-il partir de son domicile pour que le risque d'arriver en retard à son rendez-vous soit moins que 2% ?

Exercice 6.19. (issu de l'examen de L3M2S (2018)) *Au tout début de l'année 2018, Jean a placé la somme de 40000 euros dont il dispose de la manière suivante : 25000 euros ont été déposés dans un fonds garanti (sans risque) d'un contrat d'assurance vie dont le taux d'intérêt pour l'année 2018 est de 1,4% ; les 15000 euros qui restent ont été répartis entre trois placements risqués, indépendant l'un de l'autre, désignés par "Placement 1", "Placement 2" et "Placement 3". Les variables aléatoires indépendantes X_1, X_2 et X_3 désignent, pour l'ensemble de l'année 2018, les trois montants globaux en euros des gains (ou pertes) de Jean associés à ces trois placements risqués. On admet que : X_1 suit une loi normale de moyenne 500 et d'écart-type 300 ; X_2 suit une loi normale de moyenne 700 et d'écart-type 450 ; X_3 suit une loi normale de moyenne 800 et d'écart-type 550.*
1) En utilisant la table de la loi normale centrée et réduite, calculer les cinq probabilités suivantes :
a) $\mathbb{P}(X_1 > 1000)$ *; b)* $\mathbb{P}(X_1 \leq -100)$ *; c)* $\mathbb{P}(150 \leq X_1 \leq 500)$ *; d)* $\mathbb{P}(X_2 < -150)$ *;*
e) $\mathbb{P}(X_3 < -200)$.
2) On considère la variable aléatoire $Y = X_1 + X_2 + X_3 + 350$.
a) Comment peut-on interpréter concrètement Y ?
b) Trouver la loi de probabilité de Y (justifier votre réponse).
c) Au moyen de la table de la loi normale centrée et réduite, calculer la probabilité $\mathbb{P}(Y > 560)$.
Est-ce que la façon dont Jean a placé ses 40000 euros vous semble pertinente ? (justifier votre réponse)
3) Donner une interprétation concrète de la probabilité $\mathbb{P}(X_1 > X_3)$, puis calculer cette probabilité au moyen de la table de la loi normale centrée et réduite.

Exercice 6.20. (issu de l'examen de L3M2S (2018)) *Dans un groupe de 30 étudiants, 21 personnes utilisent le métro pour venir de leur domicile à l'université, 9 personnes utilisent le bus pour faire ce trajet, et 5 personnes utilisent à la fois le métro et le bus pour le faire.*
1) Combien d'étudiants dans ce groupe n'utilisent ni le métro ni le bus pour venir de leur domicile à l'université ? (justifier votre réponse)

2) Combien d'étudiants dans ce groupe utilisent uniquement l'un des deux moyens de transport (métro et bus), sans utiliser l'autre moyen de transport, pour se rendre de leur domicile à l'université ? (justifier votre réponse)

Exercice 6.21. (issu de l'examen de L2 (2019)) *On admet que la vitesse en km/h à laquelle roule un véhicule, choisi au hasard, sur une certaine portion d'une route départementale est une variable aléatoire, désignée par V, de loi normale de moyenne* 65 *et d'écart-type* 10.
1) Calculer (sous forme de pourcentages) les trois probabilités suivantes :
$\mathbb{P}(V \leq 50)$ *;* $\mathbb{P}(V > 75)$ *;* $\mathbb{P}(65 \geq V > 55)$.
2) En fait sur cette portion de route la vitesse limite autorisée est 80 km/h. Calculer (sous forme de pourcentage) la probabilité qu'un véhicule y roulant, choisi au hasard, dépasse cette limitation de vitesse.

Exercice 6.22. (issu de l'examen de L3M2S (2019)) *Stéphane est serveur d'un restaurant du Vieux-Lille ; son revenu net mensuel est constitué d'un salaire fixe de 1565 euros et de pourboires dont le montant en euros est une variable aléatoire X_1 de loi normale de moyenne 190 et d'écart-type 36. Sa femme Virginie est serveuse d'un restaurant à Lambersart ; son revenu net mensuel est constitué d'un salaire fixe de 1460 euros et de pourboires dont le montant en euros est une variable aléatoire X_2 de loi normale de moyenne 244 et d'écart-type 22. On admet que les deux variables aléatoires X_1 et X_2 sont indépendantes.*
1) En utilisant la table de la loi normale centrée et réduite, calculer sous forme de pourcentages les six probabilités suivantes :
a) $\mathbb{P}(X_1 < 125)$ *; b)* $\mathbb{P}(X_1 \geq 265)$ *; c)* $\mathbb{P}(265 > X_1 \geq 125)$ *;*
d) $\mathbb{P}(211 \geq X_2)$ *; e)* $\mathbb{P}(X_2 \geq 280)$ *; f)* $\mathbb{P}(211 \leq X_2 \leq 280)$.
2) On considère la variable aléatoire $Y = X_2 - X_1$. Trouver la loi de probabilité de Y (justifier votre réponse).
3) En utilisant votre réponse à la question 2), et la table de la loi normale centrée et réduite, calculer sous forme de pourcentage la probabilité que le revenu net mensuel de Virginie dépasse celui de son mari.
4) La variable aléatoire T désigne le revenu net mensuel de l'ensemble du couple Virginie et Stéphane. Trouver la loi de probabilité de T (justifier votre réponse), puis au moyen de la table de la loi normale centrée et réduite calculer sous forme de pourcentage la probabilité $\mathbb{P}(T > 3500)$.

7 Exercices sur la droite des moindres carrés

Exercice 7.1. *Afin de déterminer le prix de vente (exprimé en Euros) d'un nouveau produit un magasin a fait une enquête auprès d'un échantillon représentatif de ses clients et celle-ci à révéler les résultats suivants :*

X	3	$3,5$	4	$4,5$	5
Y	20	18	13	9	7

La variable X désigne le prix (exprimé en Euros) qui a été proposé aux clients et la variable Y désigne le nombre de clients qui sont prêts à acheter le produit à ce prix.
1) Calculer $\overline{x}$ et $\overline{y}$ les moyennes arithmétiques de X et de Y.
2) Calculer la covariance entre X et Y.
3) a) Calculer les moyennes quadratiques de X et Y.
b) Calculer les variances de X et Y.
c) Calculer les écarts-type de X et Y.
d) Calculer le coefficient de corrélation linéaire entre X et Y et commenter votre résultat.

3) Déterminer l'équation de la droite des moindres carrés (on ne vous demande pas de tracer cette droite).

Exercice 7.2. *Les ventes (exprimées en milliers d'articles) d'un produit P ont été observées durant 4 trimestres consécutifs et on cherche à quantifier la relation de causalité entre ces ventes et les visites des représentants chez les clients commerçants.*

	Trimestre 1	*Trimestre 2*	*Trimestre 3*	*Trimestre 4*
Visites	26	27	31	30
Ventes	53	68	79	69

La variable statistique qui correspond aux « visites » est notée par X et celle qui correspond aux « ventes » est notée par Y.
1) a) Représenter les variables X et Y sous la forme d'un nuage de points.
b) Déterminer les coordonnées du centre de gravité (noté par G) de ce nuage de points et représenter G.
2) a) Calculer $Cov(X, Y)$ la covariance entre X et Y.
b) Calculer σ_X l'écart-type de X et σ_Y l'écart-type de Y.
c) Calculer $r(X, Y)$ le coefficient de corrélation linéaire entre X et Y et commenter votre résultat.
3) a) Est-ce que la droite des moindres carrés passe nécessairement par le centre de gravité G de nuage de point associé à X et Y ?
b) Déterminer l'équation de cette droite et tracer cette droite.

Exercice 7.3. *Un exploitant agricole souhaite quantifier pour une certaine parcelle de terrain dont la superficie est de 1 hectare (c'est-à-dire 10000 m^2) la relation de causalité entre la quantité d'engrais utilisée notée par X (mesurée en kilogrammes) et la quantité de production agricole obtenue notée par Y (mesurée en quintaux, 1 quintal est égal à 100 kilogrammes). Il dispose des données statistiques suivantes :*

X	*100*	*200*	*300*	*400*	*500*	*600*	*700*
Y	*40*	*50*	*50*	*70*	*65*	*65*	*80*

1) a) Représenter les variables X et Y sous la forme d'un nuage de points.
b) Donner les coordonnées du centre de gravité, noté par G, de ce nuage puis représenter G.
2) a) Calculer la covariance de X et Y.
b) Calculer la variance de X puis celle de Y.
c) Calculer l'écart-type de X puis celui de Y.
d) Calculer le coefficient de corrélation linéaire de X et Y et interpréter votre résultat.
e) Trouver l'équation de la droite des moindres carrés puis représenter cette droite. Est-ce que la droite des moindres carrés passe nécessairement par G ?
f) Pour chacune des valeurs x_i de la variable X calculer $\widehat{y_i}$ la valeur correspondante de Y estimée par la droite des moindres carrés.
3) a) Calculer l'écart absolu moyen à la moyenne de la variable X.
b) Calculer l'écart absolu moyen à la moyenne de la variable Y.

Exercice 7.4. *On dispose d'un échantillon de 8 couples de mariés ; ces couples sont désignés par les lettres A, B, C, D, E, F, G et H. La variable X correspond l'âge de l'époux et la variable Y à celui de l'épouse, X et Y sont exprimées en nombre d'années. Les valeurs prises par ces 2 variables pour chacun des 8 couples sont données dans le tableau suivant :*

couple	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
X (âge de l'époux)	*33*	*43*	*23*	*37*	*35*	*40*	*28*	*26*
Y (âge de l'épouse)	*30*	*39*	*22*	*34*	*31*	*32*	*25*	*30*

1) a) Calculer la médiane de la variable X.
b) Calculer l'écart absolu moyen à la médiane de la variable X.
2) a) Représenter les variables X et Y sous la forme d'un nuage de points.
b) Donner les coordonnées du centre de gravité, noté par G, de ce nuage puis représenter G.
3) Calculer les écarts-types de X et Y.
4) Calculer le coefficient de corrélation linéaire de X et Y et interpréter votre résultat.
5) Trouver l'équation de la droite des moindres carrés puis représenter cette droite. Est-ce qu'elle passe nécessairement par G ?
6) Pour le couple D la variable X vaut 37 ans, quelle est la valeur correspondante de Y estimée par la droite des moindres carrés ?

Exercice 7.5. (issu de l'examen de L2 (2014)) *Dans l'objectif de déterminer le prix optimal de vente d'un dentifrice d'une certaine marque, une étude statistique a été menée par un grand magasin, auprès d'un échantillon de 100 consommateurs. Cette étude a permis d'aboutir au tableau suivant, où la variable X désigne le prix en euros du dentifrice, et la variable Y le nombre de consommateurs qui se sont déclarés prêts à acheter ce produit à ce prix.*

X	3,5	2,5	1	1,5	3	2
Y	3	21	96	63	8	31

1) Déterminer les médianes des variables X et Y.
2) Déterminer les étendues des variables X et Y.
3) Calculer les écarts absolus moyens à la médiane des variables X et Y.
4) a) Représenter les variables X et Y sous la forme d'un nuage de points.
b) Donner les coordonnées de G le centre de gravité de ce nuage, puis représenter G.
5) Calculer les moyennes quadratiques des variables X et Y.
6) Calculer les variances et les écarts-types des variables X et Y.
7) Calculer la covariance des variables X et Y.
8) Calculer le coefficient de corrélation linéaire des variables X et Y, puis interpréter votre résultat.
9) Trouver l'équation de la droite des moindres carrés puis représenter cette droite. Est-ce qu'elle passe nécessairement par le centre de gravité G ?
10) Pour un prix du dentifrice de 1,75 euros, à combien peut-on estimer le nombre des consommateurs (dans l'échantillon) se déclarant prêts à acheter le dentifrice à ce prix ? (justifier clairement votre réponse)

Exercice 7.6. (issu de l'examen de L2 (2015)) *La variable X désigne « le nombre des années d'études après le baccalauréat », et la variable Y désigne « le salaire » exprimé en milliers d'euros. Les valeurs de ces deux variables ont été observées sur un échantillon de huit employés désignés par les lettres : A, B, C, D, E, F, H et I. Le tableau suivant donne les valeurs observées de X et Y.*

Employé	*A*	*B*	*C*	*D*	*E*	*F*	*H*	*I*
X (Nbre années d'études après bac)	4	5	3	0	8	2	1	5
Y (Salaire en milliers d'euros)	1,93	2,35	1,75	1,39	2,99	1,64	1,54	2,13

On vous demande de donner les résultats de vos calculs avec, au plus, 2 chiffres après la virgule.
1) a) Trouver la médiane et la moyenne arithmétique de X.

b) Trouver la médiane et la moyenne arithmétique de Y.
2) Calculer la covariance de X et Y.
3) Déterminer l'équation de la droite des moindres carrés.
4) On désigne par M le point dont l'abscisse est la médiane de X et dont l'ordonnée est la médiane de Y ; la droite des moindres carrés passe-t-elle par M ?
5) Représenter le nuage de points associé à X et Y, puis représenter la droite des moindres carrés.
6) Calculer pour chacun des huits employés la valeur estimée de son salaire par la droite des moindres carrés.
7) Calculer le coefficient de corrélation linéaire de X et Y, puis interpréter votre résultat.

Exercice 7.7. (issu de l'examen de L2 (2016)) *La durée, mesurée en mois, qui sépare la date de la naissance d'un bébé de la date à laquelle il commence un peu à parler est appelée* l'âge du premier mot de ce bébé. *Dans l'objectif de déterminer s'il existe une relation entre l'âge du premier mot d'un bébé et le nombre de points qu'il a obtenu à un test d'habiletés mentales, appelé* le test de Gesell, *on s'intéresse à un échantillon de 9 bébés désignés par A, B, C, D, E, F, G, H et I. La variable quantitative X correspond aux âges du premier mot de ces bébés. La variable quantitative Y correspond aux résultats qu'ils ont obtenus au test de Gesell. Le tableau suivant donne pour chacun des 9 bébés les valeurs de X et Y.*

Bébé	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
X	*15*	*26*	*20*	*9*	*12*	*10*	*17*	*8*	*18*
Y	*95*	*71*	*87*	*96*	*105*	*100*	*121*	*104*	*93*

1) Déterminer les médianes des variables X et Y.
2) Calculer les écarts absolus moyens à la médiane des variables X et Y.
3) Calculer les moyennes arithmétiques des variables X et Y.
4) Calculer la covariance des variables X et Y.
5) Calculer les moyennes quadratiques des variables X et Y.
6) Calculer les variances et les écarts-types des variables X et Y.
7) Calculer le coefficient de corrélation linéaire des variables X et Y.
8) Trouver l'équation de la droite des moindres carrés.

Exercice 7.8. (issu de l'examen de L2 (2017)) *On s'intéresse à 7 liaisons en bus proposées par une compagnie de transports. Ces liaisons sont désignées par les lettres A, B, C, D, E, F et K. La variable quantitative X donne pour chacune d'elles la distance correspondante mesurée en kilomètres. La variable quantitative Y donne pour chacune d'elles le prix en euros du billet aller-retour.*

Liaison	*A*	*B*	*C*	*D*	*E*	*F*	*K*
X	*75*	*15*	*117*	*93*	*41*	*25*	*35*
Y	*25*	*5*	*35*	*30*	*20*	*10*	*15*

1) a) Représenter les variables X et Y sous la forme d'un nuage de points.
b) Donner les coordonnées du centre de gravité, noté par G, de ce nuage puis représenter G.
2) a) Calculer la covariance de X et Y.
b) Calculer la moyenne quadratique de X puis celle de Y.
c) Calculer la variance de X puis celle de Y.
d) Calculer l'écart-type de X puis celui de Y.
e) Calculer le coefficient de corrélation linéaire de X et Y et interpréter votre résultat.
3) Trouver l'équation de la droite des moindres carrés puis représenter cette droite. Est-ce que la droite des moindres carrés passe nécessairement par G ?

4) Pour chacune des valeurs x_i de la variable X calculer $\widehat{y}_i$ la valeur correspondante de Y estimée par la droite des moindres carrés.
5) Déterminer la médiane de X puis celle de Y.
6) Calculer pour chacune des variables X et Y son écart absolu moyen à la médiane.

Exercice 7.9. (issu de l'examen de L2 (2018)) *On retrouve dans le tableau suivant des données relatives à 8 monteurs en électronique ; ces monteurs sont désignés par les lettres A, B, C, D, E, F, G et H. La variable X correspond au nombre de semaines d'expérience acquises par chacun d'eux et la variable Y correspond au nombre de montages défectueux qui ont été faits par chacun d'eux.*

Monteur	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
X (nombre de semaines d'expérience)	*7*	*9*	*6*	*14*	*4*	*2*	*1*	*8*
Y (nombre de montages défectueux)	*26*	*20*	*28*	*16*	*26*	*38*	*32*	*25*

1) Calculer les moyennes arithmétiques des variables X et Y.
2) Calculer les moyennes quadratiques de ces deux variables.
3) En utilisant la notion de "profondeur", calculer les médianes de ces deux variables.
4) Calculer les variances et les écarts-types de ces deux variables.
5) a) Calculer l'écart absolu moyen à la moyenne de la variable X.
b) Calculer l'écart absolu moyen à la médiane de la variable X.
c) Calculer l'écart absolu moyen à la moyenne de la variable Y.
d) Calculer l'écart absolu moyen à la médiane de la variable Y.
6) a) Calculer la covariance des deux variables X et Y.
b) Calculer le coefficient de corrélation linéaire des deux variables X et Y, puis interpréter votre résultat.
7) a) Représenter les variables X et Y sous la forme d'un nuage de points et donner les coordonnées du centre de gravité de ce nuage.
b) Trouver l'équation de la droite des moindres carrés puis tracer cette droite.
8) Pour chacune des valeurs x_i de la variable X calculer $\widehat{y}_i$ la valeur correspondante de Y estimée par la droite des moindres carrés.
9) Calculer la variation totale de Y en faisant le moins de calculs possible.
10) Calculer la variation expliquée de Y.
11) Calculer la variation résiduelle de Y en faisant le moins de calculs possible.

Exercice 7.10. (issu de l'examen de L3M2S (2018)) *On retrouve dans le tableau suivant des données relatives à 6 magasins désignés par les lettres A, B, C, D, E et F. La variable quantitative X correspond à l'augmentation en milliers d'euros du budget publicitaire de chacun d'eux, et la variable quantitative Y correspond à l'augmentation en milliers d'euros du chiffre d'affaires de chacun d'eux qui en a résulté.*

Magasin	*A*	*B*	*C*	*D*	*E*	*F*
X (Augmentation du budget publicitaire)	*20*	*28*	*10*	*32*	*15*	*22*
Y (Augmentation du chiffre d'affaires)	*60*	*85*	*50*	*90*	*55*	*80*

1) Calculer les médianes des variables X et Y.
2) Calculer les écarts absolus moyens à la médiane des variables X et Y.
3) Représenter les variables X et Y sous la forme d'un nuage de points et donner les coordonnées du centre de gravité de ce nuage.
4) Calculer les moyennes quadratiques des variables X et Y.
5) Calculer les variances et les écarts-types des variables X et Y.

6) Calculer le coefficient de corrélation linéaire des variables X et Y, puis interpréter votre résultat.
7) Trouver l'équation de la droite des moindres carrés puis tracer cette droite.
8) Pour chacune des valeurs x_i de la variable X calculer $\widehat{y}_i$ la valeur correspondante de Y estimée par la droite des moindres carrés.

Exercice 7.11. (issu de l'examen de L2 (2019)) *Le tableau suivant concerne 8 pays d'Amérique Centrale. La variable quantitative X fournit pour l'année 1985 le taux d'urbanisation de chacun d'eux, c'est-à-dire le pourcentage de la population vivant dans des villes de plus de 100000 habitants. La variable quantitative Y fournit le taux de natalité de chacun d'eux, c'est-à-dire le nombre de naissances pour 1000 habitants pendant l'année 1985.*

Pays	*Mexique*	*Cuba*	*Salvador*	*Haïti*	*Honduras*	*Trinité-et-Tobago*	*Panama*	*Nicaragua*
X	*43,2*	*33,3*	*11,5*	*13,9*	*19,0*	*6,8*	*37,7*	*28,5*
Y	*33,9*	*16,9*	*40,2*	*41,3*	*43,9*	*24,6*	*28,0*	*44,2*

1) Représenter les variables X et Y sous la forme d'un nuage de points et donner les coordonnées du centre de gravité de ce nuage.
2) Calculer les moyennes quadratiques de ces deux variables.
3) Calculer les variances et les écarts-type de ces deux variables.
4) Calculer la covariance de ces deux variables.
5) Calculer le coefficient de corrélation linéaire de ces deux variables, puis interpréter votre résultat.
6) Trouver l'équation de la droite des moindres carrés puis tracer cette droite.
7) Au moyen du coefficient de corrélation linéaire calculer le coefficient de détermination.
8) En utilisant le coefficient de détermination et la variance de Y, calculer la variation expliquée de Y, puis calculer la variation résiduelle de Y.

Exercice 7.12. (issu de l'examen de L3M2S (2019)) *On s'intéresse à 8 quartiers d'une certaine ville qui sont désignés par les lettres A, B, C, D, E, F, H et I. La variable quantitative X donne pour chacun d'eux la distance moyenne, en centaines de mètres, qui le sépare de l'hypercentre de la ville. La variable quantitative Y donne pour chacun d'eux, en centaines d'euros, le prix de vente moyen d'un appartement au mètre carré.*

Quartier	*A*	*B*	*C*	*D*	*E*	*F*	*H*	*I*
X	*0*	*3*	*6*	*9*	*12*	*15*	*18*	*21*
Y	*137*	*135*	*127*	*105*	*85*	*55*	*38*	*26*

1) Représenter les variables X et Y sous la forme d'un nuage de points et donner les coordonnées du centre de gravité G de ce nuage.
2) Calculer les variances et les écarts-types des variables X et Y.
3) Calculer le coefficient de corrélation linéaire de X et Y, puis interpréter votre résultat.
4) Trouver l'équation de la droite des moindres carrés puis tracer cette droite.
5) Pour chacune des valeurs x_i de la variable X calculer $\widehat{y}_i$ la valeur correspondante de Y estimée par la droite des moindres carrés.
6) Déterminer les médianes des variables X et Y.
7) Est-ce que la droite des moindres carrés passe par le point dont l'abscisse est la médiane de X et dont l'ordonnée est la médiane de Y ? (justifier votre réponse)

8 Exercices sur les tableaux de contingence

Exercice 8.1. *une enquête aux USA concernant le revenu (exprimé en milliers de dollars) et la localisation géographique de 400 familles, a permis d'obtenir le tableau de contingence suivant :*

	[0,5]	]5,10]	]10,15]	Plus de 15	Total
Sud	28	42	30	24	124
Nord	44	78	78	76	276
Total	72	120	108	100	400

[0,5] ;]5,10] ;]10,15] ; et « Plus de 15 » désignent les différentes classes de revenu. « Sud » et « Nord » désignent les deux modalités de la variable localisation géographique. Donner les tableaux des profils lignes et des profils colonnes.

Exercice 8.2. *Une enquête effectuée pendant l'année* 1957 *auprès de* 1800 *ménages d'ouvriers et de* 600 *ménages d'employés concernant leurs conditions de logement a permis d'obtenir le tableau de contingence suivant :*

	bien logés	mal logés	total
ouvriers	1170	630	1800
employés	450	150	600
total	1620	780	2400

On note par Z la variable qualitative dont les modalités sont « ouvriers » et « employés ». On note par T la variable qualitative dont les modalités sont « bien logés » et « mal logés ».
1) Déterminer le tableau des profils lignes, et celui des profils colonnes.
2) Expliquer ce que sont les effectifs croisés théoriques puis déterminer le tableau des effectifs théoriques.
3) Y-a-t-il attraction entre la modalité « ouvriers » de la variable Z et la modalité « bien logés » de la variable T, ou bien y-a-t-il répulsion (justifier votre réponse au moyen d'un argument statistique) ?

Exercice 8.3. *Une enquête a été réalisée auprès d'un échantillon de 975 personnes qui sont amenées à conduire régulièrement leurs voitures. L'objet de cette enquête était de savoir si pour cet échantillon « le goût, plus ou moins prononcé, pour la vitesse » lors de la conduite est lié au « sexe » de l'individu ; elle a permis d'obtenir le tableau de contingence suivant :*

	homme	femme	total
faible	150	107	257
moyen	180	96	276
fort	320	122	442
total	650	325	975

On note par Z la variable qualitative qui désigne le « sexe », ses deux modalités sont « homme » et « femme ». On note par T la variable qualitative qui désigne « le goût, plus ou moins prononcé, pour la vitesse », ses trois modalités sont « faible », « moyen » et « fort ».
1) Déterminer le tableau des profils lignes et celui des profils colonnes.
2) Expliquer ce que sont les effectifs croisés théoriques puis déterminer le tableau des effectifs théoriques.
3) Y a-t-il des modalités de T qui sont « attirées » par la modalité « femme » de Z (justifier votre réponse au moyen d'un argument statistique) ?
4) Déterminer le tableau des χ^2 « chi2 » partiels puis calculer la distance du χ^2 « chi2 ».

Exercice 8.4. (issu de l'examen de L2 (2015)) *Une enquête a été réalisée sur un échantillon de 298 voyageurs qui se sont rendus à Paris, par le train ou l'avion, et y ont séjourné à l'hôtel. On a demandé à chacun des voyageurs de préciser lequel de ces deux modes de transport il a utilisé ; on lui a aussi demandé de donner la catégorie de son hôtel. Ainsi, on a obtenu le tableau de contingence suivant :*

	Train	*Avion*	*Total*
0 et 1 étoile	*54*	*8*	*62*
2 et 3 étoiles	*140*	*64*	*204*
4 étoiles et plus	*13*	*19*	*32*
Total	*207*	*91*	*298*

On note V la variable qualitative désignant « le mode de transport », ses deux modalités sont « Train » et « Avion ». On note S la variable qualitative désignant « la catégorie de l'hôtel », ses trois modalités sont « 0 et 1 étoile », « 2 et 3 étoiles » et « 4 étoiles et plus ».
1) Déterminer le tableau des profils lignes et celui des profils colonnes (on donnera les fréquences sous forme de pourcentages avec 2 chiffres après la virgule).
2) Expliquer ce que sont les effectifs croisés théoriques puis déterminer le tableau des effectifs théoriques.
3) Y a-t-il une modalité de V qui est « repoussée » par la modalité « 4 étoiles et plus » de S (justifier votre réponse au moyen d'un argument statistique) ?
4) Déterminer le tableau des χ^2 « chi2 » partiels puis calculer la distance du χ^2 « chi2 » (on donnera les résultats des calculs avec, au plus, 2 chiffres après la virgule).

Exercice 8.5. (issu de l'examen de L2 (2016)) *Une enquête concernant à la fois le niveau d'études (primaire, secondaire, ou supérieur) et le secteur d'activité (public, privé, ou autre) a été réalisée en Belgique sur un échantillon de 200 individus. Cette enquête a permis d'obtenir le tableau de contingence suivant :*

	Public	*Privé*	*Autre*	*Total*
Primaire	*10*	*4*	*30*	*44*
Secondaire	*25*	*16*	*15*	*56*
Supérieur	*35*	*60*	*5*	*100*
Total	*70*	*80*	*50*	*200*

On note A la variable qualitative désignant « le secteur d'activité », ses trois modalités sont « Public », « Privé » et « Autre ». On note E la variable qualitative désignant « le niveau d'études », ses trois modalités sont « Primaire », « Secondaire » et « Supérieur ».
1) Déterminer le tableau des profils lignes et celui des profils colonnes (on donnera les résultats sous forme de pourcentages).
2) Expliquer ce que sont les effectifs croisés théoriques, puis déterminer le tableau des effectifs théoriques.
3) Quelles sont les modalités de A et E pour lesquelles il y a attraction (justifier votre réponse au moyen d'un argument statistique) ?
4) Déterminer le tableau des χ^2 « chi2 »partiels, puis calculer la distance du χ^2 « chi2 ».
5) Calculer la contribution relative de la case (Supérieur, Privé) à la valeur de χ^2 « chi2 ».

Exercice 8.6. (issu de l'examen de L2 (2017)) *On s'intéresse à un groupe de 1000 salariés d'une certaine entreprise. Le tableau de contingence suivant donne leur répartition à la fois selon le dernier diplôme (bac, licence, master, ou doctorat) et selon le sexe (femme ou homme) :*

	Femme	*Homme*	*Total*
Bac	*70*	*112*	*182*
Licence	*152*	*289*	*441*
Master	*115*	*168*	*283*
Doctorat	*16*	*78*	*94*
Total	*353*	*647*	*1000*

On note S la variable qualitative désignant « sexe », ses deux modalités sont « Femme » et « Homme ». On note D la variable qualitative désignant « le dernier diplôme », ses quatre modalités sont « Bac », « Licence », « Master » et « Doctorat ».
1) Déterminer le tableau des profils lignes et celui des profils colonnes (on donnera les résultats sous forme de pourcentages).
2) Expliquer ce que sont les effectifs croisés théoriques, puis déterminer le tableau des effectifs théoriques.
3) Quelles sont les modalités de S et D pour lesquelles il y a répulsion (justifier votre réponse au moyen d'un argument statistique) ?
4) Déterminer le tableau des χ^2 « chi2 » partiels, puis calculer la distance du χ^2 « chi2 ».
5) En affirmant « qu'il y a une liaison entre les variables S et D » peut-on dire que la probabilité de se tromper est moins que 0,1% ? (justifier de façon très précise et détaillée votre réponse)

Exercice 8.7. (issu de l'examen de L2 (2018)) *Une enquête a porté à la fois sur l'âge A (exprimé en nombre d'années) et le salaire S (exprimé en centaines d'euros) des 657 salariés d'une entreprise. Cette enquête a permis d'obtenir le tableau de contingence suivant :*

	$20 \leq A < 35$	$35 \leq A < 50$	$50 \leq A < 65$	*Total*
$15 \leq S < 25$	*139*	*95*	*84*	*318*
$25 \leq S < 40$	*82*	*74*	*67*	*223*
$40 \leq S < 60$	*35*	*34*	*47*	*116*
Total	*256*	*203*	*198*	*657*

Précisons que la variable qualitative ordinale qui concerne l'âge est notée par T, et que ces trois modalités sont : $20 \leq A < 35$, $35 \leq A < 50$ et $50 \leq A < 65$. Précisons aussi que la variable qualitative ordinale qui concerne le salaire est notée par Z, et que ces trois modalités sont : $15 \leq S < 25$, $25 \leq S < 40$ et $40 \leq S < 60$.
1) Déterminer le tableau des "profils lignes" et celui des "profils colonnes" (on donnera les résultats sous forme de pourcentages).
2) Expliquer ce que sont les effectifs croisés théoriques, puis déterminer le tableau des "effectifs théoriques".
3) Y a-t-il une modalité de Z qui est "attirée" par la modalité $50 \leq A < 65$ de T ?
4) Déterminer le tableau des χ^2 "chi2" partiels, puis calculer la distance du χ^2 "chi2".
5) En affirmant "qu'il y a une liaison entre les variables T et Z", peut-on dire que la probabilité de se tromper est moins que 5% ? peut-on dire que cette probabilité est moins que 1% ? (justifier de façon très précise et détaillée vos deux réponses)
6) Calculer la contribution de la case $\big(15 \leq S < 25\,,\, 20 \leq A < 35\big)$ à la valeur de χ^2 "chi2".

Exercice 8.8. (issu de l'examen de L3M2S (2018)) *Une certaine agence immobilière souhaite savoir s'il y a une liaison significative entre la tranche d'âge d'un client et le fait qu'il préfère être locataire ou propriétaire de son logement. A partir d'un échantillon de 500 clients dans sa base de données l'agence a pu obtenir le tableau de contingence suivant :*

	[25,35]	*]35,45]*	*]45,55]*	*]55,65]*	*Plus de 65*	*Total*
Locataire	*85*	*79*	*47*	*36*	*27*	*274*
Propriétaire	*52*	*52*	*61*	*40*	*21*	*226*
Total	*137*	*131*	*108*	*76*	*48*	*500*

Précisons que la variable qualitative ordinale qui concerne l'âge, exprimé en nombre d'années, est notée par T, et que ces cinq modalités sont les tranches d'âge : "[25,35]", "]35,45]", "]45,55]", "]55,65]", et "Plus de 65". Précisons aussi que la variable qualitative nominale qui concerne le logement est notée par Z, et que ces deux modalités sont : "Locataire" et "Propriétaire".
1) Déterminer le tableau des "profils lignes" et celui des "profils colonnes" (on donnera les résultats sous forme de pourcentages).
2) Expliquer ce que sont les effectifs croisés théoriques, puis déterminer le tableau des "effectifs théoriques".
3) Y a-t-il des modalités de T qui sont "repoussées" par la modalité "Locataire" de Z ? (justifier votre réponse)
4) Déterminer le tableau des χ^2 "chi2" partiels, puis calculer la distance du χ^2 "chi2".
5) On note par p la probabilité de se tromper en affirmant : "qu'il y a une liaison significative entre les variables T et Z". Au moyen de la table de lois du χ^2 "chi2" déterminer un intervalle contenant p qui soit le plus petit possible (justifier de façon très précise et détaillée votre réponse)

Exercice 8.9. (issu de l'examen de L2 (2019)) *Une enquête concernant le maire d'une certaine ville a été réalisée sur un échantillon de 1230 de ses habitants. L'objet de l'enquête était de savoir si le fait d'habiter le centre de la ville ou sa périphérie avait une influence sur le degré de satisfaction qu'on éprouve à l'égard du maire. Cette enquête a permis d'obtenir le tableau de contingence suivant :*

	Pas Bien	*Moyen*	*Assez Bien*	*Bien*	*Très Bien*	*Total*
Centre	*83*	*141*	*134*	*147*	*30*	*535*
Périphérie	*140*	*141*	*209*	*152*	*53*	*695*
Total	*223*	*282*	*343*	*299*	*83*	*1230*

Précisons que la variable qualitative ordinale qui concerne la satisfaction éprouvée à l'égard du maire est notée par T, et que ces cinq modalités sont : "Pas Bien", "Moyen", "Assez Bien", "Bien" et "Très Bien". Précisons aussi que la variable qualitative nominale qui concerne le lieu d'habitation est notée par Z, et que ces deux modalités sont : "Centre" et "Périphérie".
1) Déterminer le tableau des "profils lignes" et celui des "profils colonnes" (on donnera les résultats sous forme de pourcentages).
2) Expliquer ce que sont les effectifs croisés théoriques, puis déterminer le tableau des "effectifs théoriques".
3) Quelles sont les modalités de la variable T qui sont attirées par la modalité "Centre" de la variable Z (justifiez votre réponse au moyen d'un argument statistique) ?
4) Déterminer le tableau des χ^2 "chi2" partiels, puis calculer la distance du χ^2 "chi2".
5) En affirmant "qu'il y a une liaison entre les variables T et Z", peut-on dire que la probabilité de se tromper est moins que 1% ? peut-on dire que cette probabilité est moins que 0,1% ? (justifiez de façon très précise et détaillée vos deux réponses)

Exercice 8.10. (issu de l'examen de L3M2S (2019)) *Une enquête concernant un roman, paru récemment, a été réalisée auprès de 750 étudiants qui l'ont lu ; certains d'entre eux sont en droit, d'autres en management et d'autres en psychologie. L'objet de cette enquête était de savoir si le type de la formation suivie par l'un de ces étudiants était lié à l'intérêt plus ou moins grand qu'il a éprouvé à la lecture de ce roman. Cette enquête a permis d'obtenir le tableau de contingence suivant :*

	Droit	*Management*	*Psychologie*	*Total*
Pas Intéressant	*77*	*115*	*75*	*267*
Intéressant	*93*	*104*	*100*	*297*
Très Intéressant	*52*	*64*	*70*	*186*
Total	*222*	*283*	*245*	*750*

Précisons que la variable qualitative ordinale qui concerne le roman est notée par Z, et que ces trois modalités sont : "Pas Intéressant", "Intéressant", et "Très Intéressant". Précisons aussi que la variable qualitative nominale qui concerne la formation est notée par T, et que ces trois modalités sont : "Droit", "Management" et "Psychologie".
1) Déterminer le tableau des "profils lignes" et celui des "profils colonnes" (on donnera les résultats sous forme de pourcentages).
2) Expliquer ce que sont les effectifs croisés théoriques, puis déterminer le tableau des "effectifs théoriques".
3) Y a-t-il une modalité de Z qui est "attirée" par la modalité "Management" de T ? (justifier votre réponse)
4) Déterminer le tableau des χ^2 "chi2" partiels, puis calculer la distance du χ^2 "chi2".
5) Calculer la contribution de la case ("Pas Intéressant" ; "Management") à la valeur de χ^2 "chi2".
6) On note par p la probabilité de se tromper en affirmant : "qu'il y a une liaison significative entre les variables T et Z". Au moyen de la table de lois du χ^2 "chi2" déterminer un intervalle contenant p qui soit le plus petit possible (justifier de façon très précise et détaillée votre réponse)