Introduction à la méthode statistique

Manuel et exercices corrigés

Bernard Goldfarb
Catherine Pardoux

6e édition

DUNOD

ISBN 978-2-10-055892-6

Table des matières

Avant-propos

Tout le monde sait et dit que celui qui observe sans idée, observe en vain.
Éléments de philosophie, Alain (1868 – 1951)

Le recueil, le traitement et l'analyse de l'information sont au cœur de tous les processus de gestion et de décision. Les méthodes de description, de prévision et de décision se sont considérablement enrichies et développées, ce qui place la statistique appliquée[1] au carrefour de l'observation et de la modélisation.

L'utilisation des méthodes statistiques s'est généralisée avec le développement et l'interprétation de logiciels et progiciels (généralistes ou spécialisés), assurant la gestion des données, les calculs, les représentations graphiques…

Plusieurs générations de logiciels statistiques [2] se sont succédé en modifiant considérablement, d'abord, l'analyse des données statistiques et maintenant, l'enseignement de la statistique. Sous peine d'être noyé, non plus dans les calculs mais dans les résultats, l'utilisateur doit disposer d'idées précises sur les outils, leurs fonctions et leurs champs d'application.

Nous avons ainsi voulu guider les futurs consommateurs et utilisateurs de données vers les descriptions statistiques majeures et les représentations courantes des phénomènes rencontrés dans tous les domaines de l'activité humaine.

La visualisation par tableaux et graphiques [3] est une clef indispensable pour traiter et comprendre efficacement les multiples ensembles de données statistiques ; l'usage généralisé qui en est fait pour tous les publics et par de nombreux médias confirme son importance.

Dans cette sixième édition, nous avons maintenu toute notre attention sur les visualisations, ainsi que sur la pratique et l'utilisation du tableur Excel ® largement répandu.

1. À laquelle les programmes, tant de l'enseignement secondaire que de l'enseignement supérieur, accordent une place de plus en plus importante.
2. Sans compter les versions évoluées des langages de programmation scientifique qui mettent l'application de traitements très sophistiqués à la portée du plus grand nombre.
3. La représentation visuelle est remarquablement mise en valeur dans le très bel ouvrage de Edward R. Tufte (1991) : *The Visual Display of Quantitative Information* , Graphics Press, Cheshire, Connecticut

La théorie reste volontairement limitée pour donner toute son importance à l'approche interprétative des données. Le lecteur, selon ses connaissances préalables et son intérêt pour la formalisation, pourra en première lecture passer outre la présentation de certains supports théoriques. Ce livre n'est qu'une introduction à la méthode statistique, et nous donnons quelques références d'ouvrages pour élargir idées et connaissances.

Dans cette sixième édition, nous avons remis à jour, à partir des recueils les plus récents, les données de nombreux exemples et des exercices (complétés et enrichis). Nous avons également inclus une très brève introduction illustrée à la pratique et à l'usage de la simulation, outil de plus en plus incontournable dans des secteurs tels que la logistique, la stratégie, ou encore l'analyse financière…

Issu de nombreuses expériences d'enseignement en formation initiale comme en formation continue pour des étudiants en sciences économiques, en sciences de gestion et en informatique de gestion, ce livre tient compte de leurs besoins et des dernières évolutions. Nous pensons qu'il correspond bien aux exigences actuelles. Nous remercions par avance les lectrices et les lecteurs qui voudront bien nous faire part de leurs remarques ou suggestions.

Bernard Goldfarb
Catherine Pardoux

1. Distributions statistiques à un caractère

Le savant doit ordonner ; on fait la science avec des faits
comme une maison avec des pierres ;
mais une accumulation de faits n'est pas plus une science
qu'un tas de pierres n'est une maison.

La Science et l'hypothèse, Henri Poincaré (1854-1912)

La statistique descriptive est un ensemble de méthodes permettant de décrire, présenter, résumer des données souvent très nombreuses. Ces méthodes peuvent être numériques (tris, élaboration de tableaux, calcul de moyennes...) et/ou mener à des représentations graphiques.

I. Définitions

A. Population, individu, échantillon

Une *population* est l'ensemble des éléments auxquels se rapportent les *données* étudiées. En statistique, le terme « population » s'applique à des ensembles de toute nature : étudiants d'une académie, production d'une usine, poissons d'une rivière, entreprises d'un secteur donné…

Des enquêtes de l'Office statistique des communautés européennes donnent la durée hebdomadaire moyenne du travail des salariés à temps complet pour 15 pays membres. Les résultats de ces enquêtes ne donnent pas d'information « atomisée » à un niveau plus bas que le pays ; la population de référence n'est donc pas ici l'ensemble (plusieurs millions) de tous les salariés des 15 pays. L'étude de ces 15 observations concerne un ensemble

de 15 *unités* (*statistiques*), les 15 pays sélectionnés qui constituent la *population* de l'étude.

Dans une population donnée, chaque élément est appelé « individu » ou « unité statistique ».

La collecte d'informations sur une population peut être effectuée sur la totalité des individus ; on parle alors d'enquêtes *exhaustives*. Lorsque la taille de la population étudiée est élevée, de telles enquêtes sont fort coûteuses ou impossibles, et le cas échéant, leurs résultats souvent très longs à rassembler peuvent être dépassés avant même la fin de l'enquête. C'est la raison pour laquelle on a souvent recours aux enquêtes par *sondage* qui portent sur une partie de la population appelée *échantillon.* Les observations obtenues sur une population ou sur un échantillon constituent un ensemble de données auxquelles s'appliquent les méthodes de la statistique descriptive dont le but est de décrire le plus complètement et le plus simplement l'ensemble des observations qu'elles soient relatives à toute la population ou seulement à un sous-ensemble.

B. Variables

Chaque individu d'une population peut être décrit selon une ou plusieurs *variables* qui peuvent être des caractéristiques qualitatives ou prendre des valeurs numériques.

Une variable est dite *qualitative* si ses différentes réalisations (modalités) ne sont pas numériques. Ainsi : le sexe, la situation matrimoniale, la catégorie socioprofessionnelle… sont des variables qualitatives. On peut toujours rendre numérique une telle variable en associant un nombre à chaque modalité ; on dit alors que les modalités sont codées. Bien entendu, les valeurs numériques n'ont dans ce cas aucune signification particulière, et effectuer des opérations algébriques sur ces valeurs numériques n'a pas de sens.

Une variable est dite *quantitative* lorsqu'elle est intrinsèquement numérique : effectuer des opérations algébriques (addition, multiplication…) sur une telle variable a alors un sens. Une variable quantitative peut être une variable statistique discrète ou continue.

Les *variables statistiques discrètes* sont des variables qui ne peuvent prendre que des valeurs isolées, discrètes. Le nombre d'enfants d'une famille, le nombre de pétales d'une fleur, le nombre de buts marqués lors d'une rencontre de football… sont des variables quantitatives discrètes. Le plus fréquemment, les valeurs possibles sont des nombres entiers.

Les *variables statistiques continues* peuvent prendre toutes les valeurs numériques possibles d'un ensemble inclus dans $\mathbb{R}$: le revenu, la taille, le taux de natalité sont des variables continues.

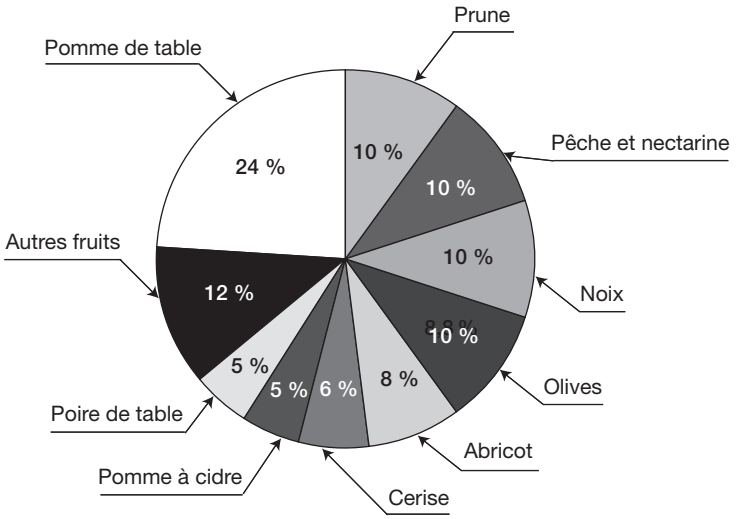
La distinction entre variables quantitatives discrètes et continues peut paraître factice, car toute mesure est discrète en raison d'une précision toujours limitée ; et inversement, lorsqu'une variable discrète peut prendre un grand nombre de valeurs et que la taille de la population (ou de l'échantillon) étudiée est élevée, on regroupera des valeurs voisines et la variable sera, par extension, traitée comme une variable continue. En pratique, lorsque les valeurs d'une variable sont regroupées en k classes, la variable est traitée comme une variable quantitative continue, mais elle peut aussi être envisagée comme une variable qualitative à k modalités.

Les données dont on dispose sont les modalités ou valeurs prises par plusieurs variables qualitatives ou quantitatives sur les individus d'une population ou d'un échantillon ; pour une population d'entreprises, on peut disposer, par exemple, de données sur le chiffre d'affaire, le bénéfice net, le nombre d'employés, la masse salariale annuelle, le secteur d'activité principale…

On peut, dans un premier temps, décrire chaque variable séparément, puis ensuite, étudier les relations ou liaisons existantes entre elles. Ainsi, dans ce livre, nous envisagerons d'abord les populations statistiques décrites selon une seule variable, puis selon deux variables. L'étude des populations caractérisées par plus de deux variables n'est pas abordée dans cet ouvrage.

II. Représentations graphiques

Deux méthodes de représentation des données vont être exposées. On commencera par celles adaptées aux données nombreuses et/ou anonymes, c'est-à-dire pour lesquelles l'identité des individus n'a pas été relevée ou ne présente pas d'intérêt à être conservée pour l'interprétation. Ceci n'est pas le cas lorsque les individus sont peu nombreux (régions, pays…), où on définira un nouveau mode de représentation graphique dû à J.W. Tukey (§ II.B.). L'étude d'une population selon une variable sera restreinte au cas des variables quantitatives, car la description d'une population selon une variable qualitative est totalement résumée dans un tableau de pourcentages ou dans un diagramme circulaire, appelé aussi diagramme en « camembert » (*cf.* figure 1.1).



Extrait de Agreste, *GraphAgri 2006*,
Ministère de l'Agriculture et de la Pêche.

Figure 1.1 – Surface du verger français en 2005

A. Distributions statistiques et représentations graphiques

Considérons une variable observée sur une population $\mathbb{P}$ de n individus. Si la variable X prend k valeurs ou ensembles de valeurs (appelés dans ce qui suit, modalités), le premier traitement des données brutes consiste à compter le nombre n_i d'individus qui présentent la i^{e} modalité ($i = 1, 2, \ldots, k$).

1) Variables statistiques discrètes

Les résultats concernant les observations de la variable X dont l'ensemble des valeurs est $\{x_i, i = 1, \ldots, k\}$, sont présentés dans le tableau des effectifs (x_i , n_i) ou dans le tableau des fréquences (x_i , f_i) avec $f_i = n_i/n$ (on utilise souvent le pourcentage $100 \cdot f_i$). Il est préférable de calculer les fréquences à partir des effectifs cumulés (§ II.A.3) afin que des erreurs successives d'arrondis ne donnent pas une somme totale de fréquences différente de 1.

Tableau des effectifs

Modalité	Effectif
x_1	n_1
$\vdots$	$\vdots$
x_i	n_i
$\vdots$	$\vdots$
x_k	n_k
	$\sum_{i=1}^{k} n_i = n$

Tableau des fréquences

Modalité	Fréquence
x_1	$f_1 = n_1/n$
$\vdots$	$\vdots$
x_i	$f_i = n_i/n$
$\vdots$	$\vdots$
x_k	$f_k = n_k/n$
	$\sum_{i=1}^{k} f_i = 1$

On présente logiquement les modalités numériques en ordre croissant. On peut associer à ces tableaux une représentation graphique appelée « diagramme en bâtons ».

Un *diagramme en bâtons* (*cf.* figure 1.2) est construit dans un système d'axes rectangulaires ; les valeurs de la variable statistique X sont portées en abscisse ; à partir de chaque valeur x_i, on trace un segment de droite vertical et dont la hauteur est proportionnelle à l'effectif correspondant. On peut retenir indifféremment une échelle qui explicite les effectifs n_i, ou une échelle qui explicite les fréquences f_i. Pour les distributions du tableau 1.1, on pourrait représenter sur le même graphique les diagrammes en bâtons de plusieurs pays avec des couleurs différentes, chaque couleur correspondant à un pays, ce qui permettrait de comparer les distributions du nombre de personnes par ménage.

Tableau 1.1 – Ménages suivant le nombre de personnes du ménage dans quelques pays en 1995 (%)

	Allemagne	Espagne	Finlande	France	Grèce	Irlande	Italie	Pays-Bas	Portugal
Ménages de :									
– 1 personne	34,4	12,7	37,4	29,2	20,7	22,8	22,7	30,6	13,7
– 2 personnes	32,3	24,5	31,0	31,8	28,9	23,1	23,1	34,0	26,4
– 3 personnes	16,0	21,8	14,4	16,8	19,8	15,6	15,6	13,4	24,7
– 4 personnes	12,6	24,0	11,9	14,2	21,7	17,1	17,1	15,9	22,8
– 5 personnes et plus	4,7	17,0	5,3	8,0	8,9	21,4	21,4	6,2	12,4
Ensemble (en milliers)	34 413	12 112	2 222	23 126	3 756	1 146	1 146	6 425	3 275

Source : Tableaux de l'Économie Française 1999-2000, INSEE.

Nombre de personnes	f_i (%)
1	29,2
2	31,8
3	16,8
4	14,2
5 ou plus	8,0
	100 %

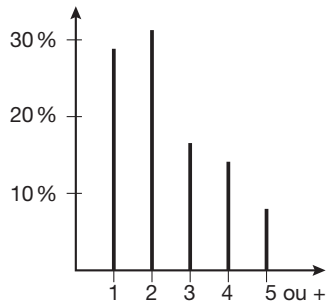


Figure 1.2 – Diagramme en bâtons – *Nombre de personnes par ménage en France en 1995*

2) Variables statistiques continues

L'infinité des valeurs observables ne rend pas possible la généralisation du diagramme en bâtons. *Le domaine de variation* d'une *variable statistique continue X* est partagé en *k parties*. L'intervalle [x_{i-1} , x_i [fermé à gauche, ouvert à droite, est appelé i^e classe ($i = 1,2, \ldots,k$) ; son amplitude est égale à :

$$a_i = x_i - x_{i-1}$$

Il arrive que l'amplitude des classes extrêmes soit indéterminée : la première classe étant définie par « *moins de…* », et la dernière par « *plus de…* » (*cf.* tableau 1.2).

Le choix des *extrémités* des classes se fait à partir des données brutes ; le nombre *k* de classes doit être modéré (usuellement entre 4 et 10). Le découpage en classes est assez souvent choisi tel que l'amplitude des classes soit constante, ou tel que les effectifs des classes soient constants (par exemple, 10 % de la population dans chaque classe, *cf.* tableau 1.6).

Le classement d'une série statistique correspond à une perte d'information par rapport aux données initiales puisque seuls les effectifs des classes sont retenus. Le travail sur une telle série impose alors l'hypothèse que les données sont réparties *uniformément* à l'intérieur de chacune des classes. On parle aussi d' *équirépartition* des individus ou encore d' *homogénéité* dans chacune des classes. Chaque partie de la classe correspond alors à un effectif proportionnel à sa longueur. L'idée est, bien sûr, que chaque classe représente une entité qui doit se distinguer par rapport aux autres classes. Comme précédemment, les résultats sont présentés dans un tableau d'effectifs ou de fréquences. On associe à un tel tableau un *histogramme* qui est une représentation graphique très répandue. L'histogramme est constitué de la juxtaposition de rectangles (pour respecter l'hypothèse d'équirépartition) dont les bases représentent les différentes classes et dont *les surfaces sont proportionnelles aux effectifs des classes* (*cf.* figure 1.3).

On verra par la suite qu'une difficulté du travail avec des séries classées est le choix des limites pour les classes extrêmes, indispensable aussi pour le tracé de l'histogramme.

À la i^e classe, correspond un rectangle dont la base est l'intervalle $[\ x_{i-1}, x_i\ [$ et dont la surface est proportionnelle à la fréquence f_i (ou à l'effectif n_i). Si les classes ont toutes la même amplitude, les hauteurs des rectangles sont proportionnelles aux fréquences. Dans le cas où les classes sont d'amplitudes inégales, la hauteur du rectangle correspondant à la i^e classe d'amplitude a_i sera $h_i = f_i/a_i$. La surface du rectangle représentant la i^e classe sera ainsi égale à f_i

Pour une série d'observations relatives à une variable statistique X discrète ou continue classée, la donnée des modalités et de leurs fréquences est appelée « *distribution statistique* » de la variable X.

Tableau 1.2 – Chômeurs BIT selon le sexe et l'ancienneté de chômage en septembre 2006

	Distribution en milliers		Distribution en pourcentage	
Ancienneté d'inscription	Hommes	Femmes	Hommes	Femmes
Moins d'un mois	180,3	181,0	16,5	16,8
D'un à moins de trois mois	203,9	204,9	18,6	19,0
De trois à moins de six mois	169,3	163,1	15,5	15,1
De six mois à moins d'un an	202,1	191,1	18,5	17,7
D'un à moins de deux ans	197,3	199,3	18,0	18,5
De deux à moins de trois ans	74,5	75,4	6,8	7,0
Trois ans ou plus	67,1	62,9	6,1	5,8
Ensemble	1 094,5	1 077,7	100	100
Ancienneté moyenne en jours	341	334		

Source : Bulletin Mensuel des Statistiques du Travail, www.travail.gouv.fr, octobre 2006.

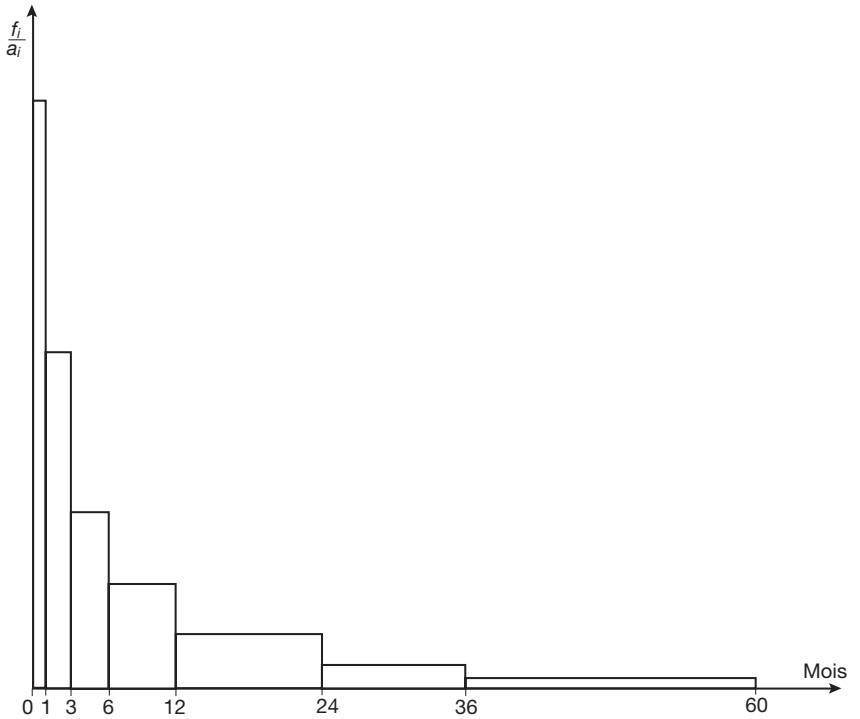


Figure 1.3 – Histogramme de la distribution des chômeurs « Femmes » selon l'ancienneté (voir tableau 1.2)

La classe « Trois ans ou plus » est supposée bornée supérieurement par 5 ans (60 mois).

3) Fréquences cumulées et courbe cumulative

a) Tableau des fréquences cumulées

Les tableaux de fréquences (ou d'effectifs) qui viennent d'être définis peuvent être modifiés de facon à présenter un résumé des données sous une forme différente.

On appelle *effectif cumulé* de la i^e classe, le nombre d'individus N_i pour lesquels la variable prend une valeur inférieure à x_i :

$$N_i = \sum_{j \leq i} n_j \text{ pour } i = 1, 2, \ldots, k$$

On définit de même F_i, la *fréquence cumulée* de la i^e classe : $F_i = N_i/n$

Les tableaux d'effectifs cumulés ou de fréquences cumulées se déduisent des tableaux d'effectifs ou de fréquences (non cumulés) en substituant aux effectifs ou fréquences non cumulés les effectifs ou fréquences cumulés. Les deux types de tableaux sont donc équivalents (cf. figures 1.2 et 1.4).

b) Fonction cumulative et courbe cumulative

La *courbe cumulative* ou courbe des fréquences cumulées est la représentation graphique des fréquences cumulées. Plus précisément, la courbe cumulative est la représentation graphique de la proportion $F(t)$ des individus de la population dont le caractère prend une valeur inférieure à t. Cette fonction, appelée *fonction cumulative* ou *fonction de répartition*, est :

1. définie pour tout $t \in \mathbb{R}$

2. croissante (mais non strictement croissante)

3. nulle pour t inférieur à $\min_{1 \le i \le n} x_i$

4. égale à 1 pour t au moins égal à $\max_{1 \le i \le n} x_i$

Pour une variable statistique *discrète*, cette fonction est une *fonction en escalier*, présentant en chacune des valeurs possibles x_i, un saut égal à la fréquence correspondante f_i (*cf.* figure 1.4).

Dans le cas d'une variable statistique *continue*, la fonction cumulative n'est connue que pour les valeurs de X égales aux extrémités des classes. L'hypothèse d'équirépartition (§ II.A.2) implique que la fonction F est linéaire entre ces valeurs (*cf.* figure 1.5). *Cette fonction est donc continue et linéaire par morceaux.* Ici encore, il est nécessaire de choisir des limites pour les classes extrêmes.

t	$F(t)$ (%)
< 1	0
[1 ; 2[	29,2
[2 ; 3[	61,0
[3 ; 4[	77,8
[4 ; 5[	92,0
≥ 5	100

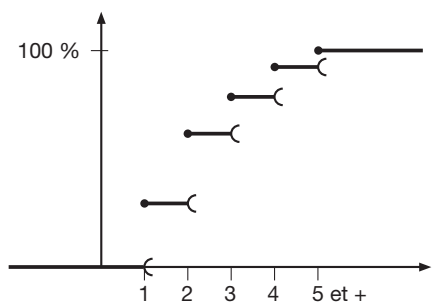


Figure 1.4 – Graphe des fréquences cumulées de la distribution représentée à la figure 1.2

Ces fréquences cumulées sont des fréquences cumulées *ascendantes*, car elles ont été obtenues en calculant les fréquences F_i d'individus pour lesquelles le caractère étudié X est *au plus* égal à x_i ; on peut aussi définir les fré-

t	*F(t)* (%)
0	0
1	16,8
3	35,8
6	50,9
12	68,7
24	87,2
36	94,2
60	100

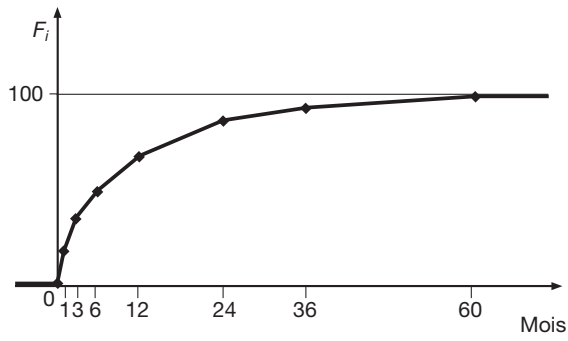


Figure 1.5 – Courbe cumulative de la distribution représentée à la figure 1.3

quences cumulées *descendantes*, c'est-à-dire les fréquences pour lesquelles le caractère étudié X est supérieur à x_i. Quand on ne spécifie pas le type de fréquences cumulées, on sous-entend qu'il s'agit des fréquences cumulées ascendantes.

B. Le diagramme « branche et feuille »

Lorsque la taille de la population étudiée n'est pas trop élevée (inférieure à la centaine), il est intéressant d'utiliser la représentation en ***diagramme « branche et feuille »*** due à J. W. Tukey[1]. Ce diagramme tient à la fois du *tableau* et de la *représentation graphique* et donne une vision d'ensemble des données *sans perdre* l'information numérique valeur par valeur.

1) Profondeur d'une observation

Selon qu'on range les valeurs observées de la variable statistique X de la plus faible à la plus élevée ou de la plus élevée à la plus faible, on associe à chaque observation x_i deux rangs, croissant et décroissant. On dit alors que la distribution est ordonnée.

On appelle ***profondeur*** de x_i le nombre égal au *plus petit des deux rangs*.

Les durées hebdomadaires du travail des salariés à temps complet dans les pays de l'Union européenne (*cf.* tableau 1.3) peuvent être ordonnées, et on en déduit la profondeur de chaque valeur de chacune des séries.

1. J. W. Tukey, *Exploratory Data Analysis* (EDA), Addison-Wesley, 1977.

Tableau 1.3 – Durée hebdomadaire du travail des salariés à temps complet dans l'Union européenne (heures)

	1990	1995	2000
Allemagne	39,9	39,7	40,1
Autriche	40,1	39,3	40,1
Belgique	38	38,4	38,5
Danemark	39	39	39,3
Espagne	40,7	40,7	40,6
Finlande	38,4	38,6	39,3
France	39,6	39,9	38,9
Grèce	40,2	40,3	40,9
Irlande	40,4	40,2	39,9
Italie	38,6	38,4	38,6
Luxembourg	39,9	39,5	39,8
Pays-Bas	39	39,5	39
Portugal	41,9	41,2	40,3
Royaume-Uni	43,7	43,9	43,6
Suède	40,7	40	40

Source : Tableaux de l'Économie Française, INSEE.

Le nombre de pays étant impair et égal à 15, il y a deux valeurs de profondeur 1, 2, 3, 4, 5, 6, 7 et une seule valeur de profondeur 8 (*cf.* tableau 1.4).

Tableau 1.4 – Pays ordonnés selon la durée hebdomadaire du travail des salariés à temps complet en 2000

Rang croissant	Rang décroissant	Profondeur	Durée (heures)	Pays
1	15	1	38,5	Belgique
2	14	2	38,6	Italie
3	13	3	38,9	France
4	12	4	39,0	Pays-Bas
5	11	5	39,3	Danemark
6	10	6	39,3	Finlande
7	9	7	39,8	Luxembourg
8	8	**8**	**39,9**	**Irlande**
9	7	7	40,0	Suède
10	6	6	40,1	Allemagne
11	5	5	40,1	Autriche
12	4	4	40,3	Portugal
13	3	3	40,6	Espagne
14	2	2	40,9	Grèce
15	1	1	43,6	Royaume-Uni

2) La représentation en diagramme « branche et feuille »

Son principe consiste à distinguer deux parties pour tout nombre : les chiffres de plus « *faible poids* », la *feuille*, et les chiffres de plus « haut poids », la *branche*.

La figure 1.6 reproduit les *diagrammes « Branche et feuille »* donnés par le logiciel SPSS pour les séries du tableau 1.3.

1990 Frequency Stem & Leaf	1995 Frequency Stem & Leaf	2000 Frequency Stem & Leaf
3,00 38 . 046 5,00 39 . 00699 5,00 40 . 12477 1,00 41 . 9 1,00 Extrêmes (>=43,7)	3,00 38 . 446 6,00 39 . 035579 4,00 40 . 0237 1,00 41 . 2 1,00 Extrêmes (>=43,9)	0,00 38 . 3,00 38 . 569 3,00 39 . 033 2,00 39 . 89 4,00 40 . 0113 2,00 40 . 69 1,00 Extrêmes (>=43,6)
Stem width : 1,0 Each leaf : 1 case(s)	Stem width : 1,0 Each leaf : 1 case(s)	Stem width : 1,0 Each leaf : 1 case(s)

Figure 1.6 – « Branche et feuille » *(logiciel SPSS) pour les séries du tableau 1.3*

Par exemple, pour le diagramme de l'année 1995 de la figure 1.6, en se référant aux valeurs ordonnées :

– la valeur 38,4 est représentée par la branche 38 et la feuille 4 (pour les deux observations) ;

– la valeur 38,6 est représentée par la branche 38 et la feuille 6.

Ces trois observations conduisent à l'écriture : 3,00 38. 446

La valeur 43,9 est beaucoup plus élevée que les autres ; elle est mentionnée comme valeur « *extrême* ». On verra comment une valeur est ainsi classée (§ IV.B). Le nombre de feuilles de chaque branche donnant l'effectif, un histogramme à classes égales d'amplitude 1 donne une représentation similaire, mais l'avantage du diagramme branche et feuille est de conserver ici l'information donnée par le premier chiffre décimal, donc de garder l'information de la répartition à l'intérieur des classes.

Les logiciels choisissent, selon la structure des données, des « amplitudes » égales à 1, 0,5 ou 0,25. La plage des valeurs étant plus restreinte en 2000 qu'en 1990 et 1995, le logiciel SPSS a choisi des amplitudes égales à 1 pour les années 1990 et 1995, et des amplitudes égales à 0,5 pour l'année 2000.

On peut compléter ce type de diagramme pour garder l'identité des individus en indiquant symétriquement l'identité de chaque feuille (*cf.* figure 1.7). On pourrait aussi représenter *dos à dos* les distributions correspondant à deux années différentes pour suivre l'évolution de la durée hebdomadaire du travail.

Frequency			Stem & Leaf
	3,00	Fin It Bel	38 . 446
	6,00	Fr All P.Bas Lux Aut Dk	39 . 035579
	4,00	Esp Gr Irl Suèd	40 . 0237
	1,00	Por	41. 2
	1,00	R-U	Extremes (> = 43,9)

Stem width : 1,0
Each leaf : 1 case(s)

Figure 1.7 – Diagramme « Branche et feuille » complété par l'identité des pays (1995)

III. Les indicateurs statistiques

Le tableau de distribution d'une variable statistique présente l'information recueillie sur cette variable. Une représentation graphique en fournit un portrait pour appréhender plus facilement la globalité de l'information. On peut désirer aller plus loin en cherchant à caractériser la représentation visuelle par des éléments synthétiques sur :

- la valeur de la variable située au « centre » de la distribution : la *tendance centrale* et, plus généralement, un *indicateur de position* non nécessairement centrale, liée à un rang donné ;
- la variation des valeurs : la *dispersion* ;
- la *forme* de la distribution ;
- les aspects particuliers : valeurs *extrêmes*, *groupes* de valeurs…

Ces indicateurs étant exprimés dans les unités de la variable étudiée, on verra qu'il peut être intéressant pour comparer plusieurs distributions entre elles de calculer des *caractéristiques de dispersion relative.*

A. Conditions de Yule

Le statisticien britannique Yule [1] a énoncé un certain nombre de *propriétés* souhaitées pour les indicateurs des séries statistiques ; ceux-ci doivent être d'une part, des résumés « maniables » et d'autre part, les plus exhaustifs possibles relativement à l'information contenue dans les données.

1. G. Udny Yule et M. G. Kendall, *An Introduction to the Theory of Statistics*, Charles Griffin & Co, 14e édition, 1950.

Dans son schéma, une caractéristique statistique doit être une valeur-type :

1. définie de façon objective et donc indépendante de l'observateur ;

2. dépendante de toutes les observations ;

3. de signification concrète pour être comprise par des non-spécialistes ;

4. simple à calculer ;

5. peu sensible aux fluctuations d'échantillonnage ;

6. se prêtant aisément aux opérateurs mathématiques classiques.

En réalité, on ne dispose pas de caractéristiques répondant simultanément à ces six conditions. Le choix d'un indicateur sera l'objet d'un compromis guidé par la spécificité de l'étude en cours.

B. Les indicateurs de tendance centrale et de position

Selon l'usage courant, toutes les mesures de tendance centrale méritent le nom de « moyenne ». Lorsqu'on parle de moyenne, on pense à la moyenne arithmétique ; mais il existe d'autres types de moyennes, chacune d'entre elles ayant la propriété de *conserver une caractéristique* de l'ensemble quand on remplace chaque élément de l'ensemble par cette valeur unique ; chaque moyenne n'a donc d'intérêt que pour autant que cette propriété soit utile [1].

Les « moyennes » sont des valeurs abstraites qui, sauf par hasard, ne correspondent à aucune réalisation concrète.

1) La moyenne arithmétique

On appelle *moyenne arithmétique* la somme de toutes les données statistiques divisée par le nombre de ces données. La moyenne arithmétique *conserve la somme totale des valeurs* observées : si on modifie les valeurs de deux observations d'une série statistique tout en conservant leur somme, la moyenne de la série sera inchangée.

Soit la série statistique de données brutes : $x_1, \ldots, x_i, \ldots, x_n$, sa moyenne arithmétique a pour expression :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Bien entendu, si une valeur x_i de X est observée n_i fois, comme $\underbrace{x_i + x_i + \ldots + x_i}_{n_i \text{ fois}} = n_i x_i$, la formule précédente devient :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i$$

1. Ch. Antoine, « Les moyennes au quotidien », dans *Les Moyennes*, Que Sais-je, PUF, n° 3383, 1998, p. 107.

où k désigne le nombre de valeurs *distinctes* de X et $f_i = \frac{n_i}{n}$

Lorsqu'on a une variable statistique continue, on ne connaît pas les valeurs exactes prises par la variable, mais seulement le nombre d'observations à l'intérieur de chaque classe. Pour calculer la moyenne arithmétique d'une telle variable, on ramène *chaque observation au centre de sa classe,* ceci en raison de l'hypothèse d'équirépartition à l'intérieur des classes, et celà revient à considérer la moyenne des individus de la i^e classe égale à $(x_{i-1} + x_i)/2$.

Dans le cas des classes extrêmes non limitées, le choix des limites de ces classes influe évidemment sur la valeur de la moyenne arithmétique. Ces limites devront être choisies en fonction des connaissances sur les données et en n'oubliant pas l'hypothèse de base : l'homogénéité à l'intérieur des classes. Pour une classe extrême dans laquelle on sait qu'il n'y a pas équirépartition, les observations étant vraisemblablement en majorité regroupées sur une partie de la classe, il conviendra de choisir la borne extrême :

- moins faible que la borne réelle (supposée) s'il s'agit de la première classe ;
- plus faible que la borne réelle (supposée) s'il s'agit de la dernière classe.

C'est ce qui a été fait pour la série présentée au tableau 1.2 et à la figure 1.3, l'ancienneté moyenne du chômage a été considérée égale à 48 mois pour ceux dont l'ancienneté était au moins égale à 36 mois et la borne supérieure de la dernière classe a été de ce fait fixée à 60 mois (l'hypothèse d'équirépartition amène à considérer que la moyenne des observations d'une classe est égale au centre de la classe).

Propriétés

1. La moyenne est une caractéristique qui satisfait à toutes les conditions de Yule, sauf à la conditions 5 : une observation « extrême » (exceptionnellement élevée ou faible) peut avoir une forte incidence sur sa valeur.

2. La somme algébrique des écarts des valeurs d'une variable statistique à sa moyenne arithmétique est nulle : $\sum_{i=1}^{k} f_i(x_i - \bar{x}) = 0$

3. Lorsqu'on fait subir à une variable statistique X une transformation affine, c'est-à-dire un changement d'origine et d'unité $\{ Y = aX + x_0 \}$, sa moyenne arithmétique subit la même transformation : $\bar{y} = a\bar{x} + x_0$

4. Soit une population $\mathbb{P}$ de taille n partagée en deux sous-populations $\mathbb{P}_1$ de taille n_1 et $\mathbb{P}_2$ de taille n_2.

Soit X, une variable statistique observée sur la population $\mathbb{P}$, on peut exprimer sa moyenne $\bar{x}$ en fonction de ses moyennes $\overline{x_1}$ sur $\mathbb{P}_1$ et $\overline{x_2}$ sur

$\mathbb{P}_2$ en remarquant que la somme totale $n\bar{x}$ s'obtient en additionnant $n_1\overline{x_1}$ et $n_2\overline{x_2}$:

$$\bar{x} = \frac{1}{n}(n_1\overline{x_1} + n_2\overline{x_2})$$

Ce résultat se généralise à une partition en k sous-populations ($k \geq 2$) :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i\overline{x_i}$$

▶ **Exemple**

L'ancienneté moyenne d'inscription au chômage pour hommes et femmes réunis en septembre 2006 est égale à (*cf.* tableau 1.2 pour les données) :

$$\bar{x} = \frac{1}{2\,172{,}2}\,(1\,094{,}5 \cdot 341 + 1\,077{,}7 \cdot 334) \simeq 338 \text{ jours}$$

2) D'autres moyennes

a) La moyenne géométrique

C'est la moyenne applicable à des mesures de grandeurs dont la croissance est géométrique ou exponentielle.

La *moyenne géométrique conserve le produit des* x_i : si on modifie les valeurs de deux observations tout en conservant leur produit, la moyenne géométrique sera inchangée.

La moyenne géométrique G de la série de valeurs $x_1, \ldots, x_i, \ldots, x_n$ supposées toutes positives (strictement), est définie ainsi :

$$G = \sqrt[n]{\prod_{i=1}^{n} x_i} \quad \Rightarrow \quad \ln(G) = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i)$$

Lorsque la distribution de la variable statistique est donnée par les k couples (x_i, n_i), les x_i étant tous positifs ; la moyenne géométrique a pour expression :

$$G = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}} = \prod_{i=1}^{k} x_i^{f_i} \quad \Rightarrow \quad \ln(G) = \sum_{i=1}^{k} f_i \ln(x_i)$$

▶ **Exemple**

Supposons que pendant une décennie, les salaires aient été multipliés par 2 et que pendant la décennie sui vante, ils aient été multipliés par 4 ; le coefficient multiplicateur moyen par décennie est égal à :

$$\sqrt{2 \cdot 4} = \sqrt{8} \approx 2{,}83$$

La moyenne arithmétique (= 3) n'est pas égale au coefficient demandé.

Prenons, par exemple, un salaire de 300 € au début de la première décennie, il sera de 300 · 2 · 4 = 2 400 € au bout des vingt ans, ce qui équivaut à $300 \cdot (2{,}83)^2$, soit un coefficient multiplicateur moyen de 2,83 par décennie.

b) La moyenne harmonique

La *moyenne harmonique* est l'inverse de la moyenne arithmétique des inverses des valeurs. L'*inverse de la moyenne harmonique conserve ainsi la somme des inverses des* x_i : si on modifie les valeurs de deux observations tout en conservant la somme de leurs inverses, la moyenne harmonique sera inchangée.

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \quad \text{ou} \quad H = \frac{1}{\sum_{i=1}^{k} \frac{f_i}{x_i}}$$

La moyenne harmonique peut être utilisée lorqu'il est possible d'attribuer un sens réel aux inverses des données en particulier pour les taux de change, les taux d'équipement, le pouvoir d'achat, les vitesses. Elle est notamment utilisée dans les calculs d' *indices*.

▶ **Exemple**

On achète des dollars une première fois pour 100 € au cours de 1,23 € le dollar, une seconde fois pour 100 € au cours de 0,97 € le dollar.

Le cours moyen du dollar pour l'ensemble de ces deux opérations est égal à :

$$\frac{200}{\frac{100}{1{,}23} + \frac{100}{0{,}97}} \approx 1{,}085 \text{ €}$$

La moyenne arithmétique (= 1,1) ne représente pas le cours moyen du dollar.

Comparaison des 3 moyennes étudiées

On montre que si les x_i sont tous positifs :

$$\min_{1 \le i \le n} x_i \le H \le G \le \bar{x} \le \max_{1 \le i \le n} x_i$$

L'égalité de deux de ces moyennes entre elles entraîne leur égalité dans leur ensemble, et dans ce cas, toutes les valeurs x_i sont égales.

3) Le mode

Pour obtenir une mesure de la tendance centrale non influencée par les valeurs extrêmes de la distribution, on peut prendre la valeur – ou la classe de valeurs – du caractère pour laquelle le diagramme en bâtons – respectivement l'histogramme – présente son *maximum* : c'est le *mode* – respectivement l'*intervalle modal* – de la distribution ; dans le cas où le diagramme en bâtons – ou l'histogramme – présente aussi un maximum local, il y a deux modes – respectivement deux classes modales.

Lorsque la variable statistique est discrète, le mode se définit donc à l'aide du tableau de distribution ou du diagramme en bâtons. Pour la distribution présentée à la figure 1.2, le mode est égal à 2. Si la fréquence maximum correspond à deux valeurs successives de la variable, il y a un *intervalle modal.*

Lorsqu'une distribution présente plusieurs modes auxquels correspondent (généralement) des fréquences différentes, c'est souvent l'indice du mélange de deux ou plusieurs populations ayant chacune leur mode propre (*cf.* figure 1.8). Un exemple peut en être la distribution des pointures de chaussures des hommes et femmes réunies.

Lorsque la variable statistique est continue, la *classe modale* est la classe dont la fréquence par unité d'amplitude est la plus élevée. Pour la distribution présentée à la figure 1.3, la classe modale est la classe [1, 3[. Mais cette détermination n'est absolument pas précise, car elle dépend du découpage en classes retenu ; son intérêt est limité par cette imprécision.

Dans le cas d'une distribution discrète, le mode satisfait aux conditions 1, 3, 4 et 5 de Yule. Dans le cas de la distribution du nombre d'enfants par famille, le mode est réellement une valeur typique et paraît mieux correspondre à la réalité que la moyenne arithmétique qui est rarement un nombre entier et qui est sensiblement influencée par un nombre relativement petit de familles très nombreuses. À l'inverse de la moyenne arithmétique, le mode néglige délibérement la précision numérique au profit de la représentativité. Dans un tel cas, il est souvent souhaitable de disposer de ces deux mesures de la tendance centrale.

Le mode, historiquement l'un des premiers paramètres de position utilisés, est un peu moins employé aujourd'hui.

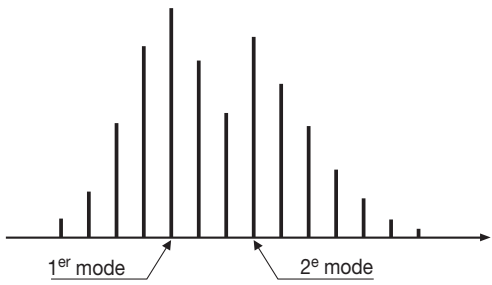


Figure 1.8 – Exemple de distribution bimodale d'une variable discrète

4) La médiane et les quantiles

Bien qu'homogènes dans leur composition, de nombreuses distributions présentent de très grands écarts entre les valeurs extrêmes de leurs éléments.

De plus, elles ont souvent un manque de symétrie prononcé, les éléments ayant tendance à s'agglomérer plus près d'un extrême que de l'autre. Les

distributions de salaires ou de revenus en donnent des exemples typiques. Il est évident que, dans de tels cas, nous avons besoin d'une mesure de la tendance centrale qui ne soit pas influencée par un nombre relativement petit de valeurs extrêmes se situant en « queue » de la distribution.

a) La médiane

La *médiane* est la valeur de la variable statistique telle qu'il y ait autant d'observations supérieures et d'observations inférieures à cette valeur. Elle partage la série statistique en deux parties d'égal effectif. Elle se détermine soit à partir de la série des valeurs ordonnées, soit à partir de la fonction cumulative (§ II.A.3).

Pour les ***variables statistiques discrètes***, la médiane est déterminée à l'aide de la « profondeur ».

Dans le cas où la série comporte un nombre impair n d'observations, la médiane est égale à la valeur de profondeur maximum $(n + 1)/2$: pour la série des 15 valeurs du tableau 4, la médiane est égale à la valeur de profondeur 8, soit 39,9 *h*.

Dans le cas où la série comporte un nombre pair n d'observations, la médiane est la moyenne arithmétique des deux valeurs de profondeur $n/2$ et est ainsi définie comme la valeur de profondeur $(n + 1)/2$.

La *médiane* est ainsi dans tous les cas *la valeur de* ***profondeur*** $(n + 1)/2$.

Lorsque les données d'une variable statistique discrète sont classées, il n'existe généralement pas une valeur médiane *Me* pour laquelle la fonction cumulative vaut 50 %. Il faut dans ce cas utiliser d'autres valeurs typiques pour caractériser la tendance centrale de la série : ceci est le cas pour la distribution du nombre de personnes par ménage dont la fonction cumulative est représentée à la figure 1.4.

Pour les ***variables statistiques continues***, la valeur médiane *Me* est telle que $\boldsymbol{F(Me) = 50\%}$. On commence par chercher la *classe médiane* à l'aide des fréquences cumulées, la classe médiane $[x_{i-1}, x_i[$ étant telle que $F_{i-1} < 50\%$ et $F_i > 50\%$. La valeur de la médiane s'obtient ensuite par *interpolation linéaire* en raison de l'hypothèse d'équirépartition à l'intérieur des classes. Cette détermination peut se faire par le calcul ou graphiquement (*cf.* figure 1.9) :

$$\frac{Me - x_{i-1}}{x_i - x_{i-1}} = \frac{0,5 - F_{i-1}}{f_i} \quad \Rightarrow \quad Me = x_{i-1} + (x_i - x_{i-1}) \cdot \frac{0,5 - F_{i-1}}{f_i}$$

Pour la distribution de l'ancienneté du chômage des femmes (tableau 1.2 et figure 1.5), la médiane appartient à la classe [3 ; 6[:

$$Me = 3 + 3 \cdot \frac{50 - 35,8}{15,1} \approx 5,8 \text{ mois}$$

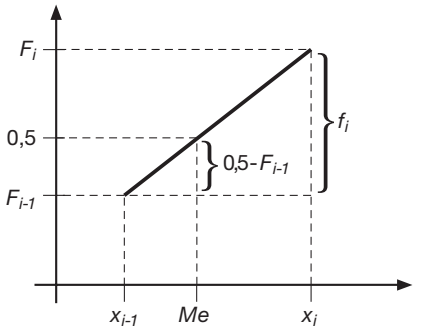


Figure 1.9 – Détermination graphique de la médiane pour une variable continue

La médiane peut aussi être déterminée à partir de la courbe des fréquences cumulées comme l'abscisse du point d'ordonnée 50 %.

Une *seule* observation très élevée (ou très faible) peut influencer fortement la moyenne, alors que la médiane peut supporter sans être modifiée qu'une moitié des observations soit très élevée (ou très faible) : on dit que la médiane est *résistante.* La médiane satisfait aux conditions 1, 3, 4 et 5 de Yule.

Dans le cas de distribution unimodale, la médiane est fréquemment comprise entre la moyenne arithmétique et le mode, et plus près de la moyenne que du mode. Si la distribution est symétrique, ces *trois caractéristiques* de tendance centrale sont *confondues* (*cf.* figure 1.10).

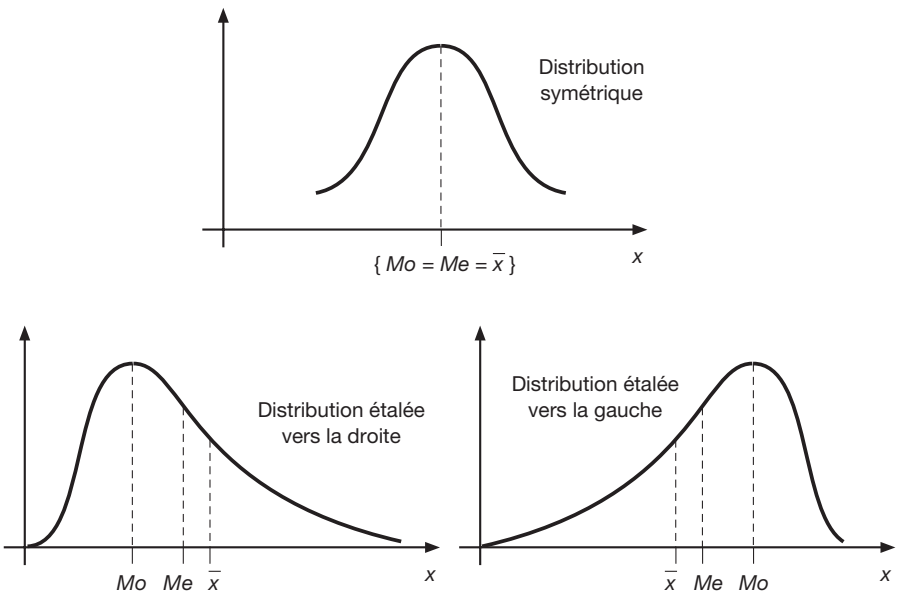


Figure 1.10 – Positions respectives du mode, de la médiane et de la moyenne

b) Les quantiles

Les *quantiles* sont des *indicateurs de position*.

Le *quantile d'ordre* α ($0 \leq \alpha \leq 1$), noté x_α, est tel qu'une proportion α des individus ait une valeur du caractère X inférieure ou égale à x_α

Le quantile $x_{0,5}$ est égal à la médiane.

On utilise couramment les quantiles d'ordre 1/4, 1/2 et 3/4. Ils sont ainsi notés et nommés :

$$Q_1 = \textit{premier quartile} = x_{0,25}$$
$$Q_2 = \textit{deuxième quartile} = \textit{médiane} = x_{0,5}$$
$$Q_3 = \textit{troisième quartile} = x_{0,75}$$

Les quartiles se déterminent, comme la médiane, à l'aide de la profondeur (variable discrète), ou à l'aide des fréquences cumulées (variable continue).

Dans le cas d'une variable statistique *discrète*, le premier quartile Q_1 et le troisième quartile Q_3 sont des éléments de *même profondeur égale à* $(m + 1)/2$ où m désigne la *partie entière* de la profondeur de la médiane. On peut aussi considérer Q_1 comme la médiane des m premières valeurs de la série et Q_3 comme la médiane des m dernières valeurs. Ainsi par exemple, pour une série de 39 observations, la médiane a une profondeur égale à 20, et les quartiles Q_1 et Q_3 sont de profondeur 10,5 ; pour une série de 50 observations, la médiane a une profondeur de 25,5 et la partie entière de cette profondeur étant 25, les quartiles Q_1 et Q_3 sont de profondeur 13.

La pratique de la détermination des quartiles ne respecte pas toujours la définition précédente due à Tukey. Ainsi les calculatrices de poche (TI, Casio,…) déterminent le 1er quartile (resp. le 3e quartile) comme la médiane des valeurs de profondeur inférieure (resp. supérieure) à la profondeur de la médiane. Le résultat diffère de celui calculé avec la définition de Tukey dans le cas d'un nombre impair d'observations. Le logiciel SPSS détermine deux types de quartiles : « Valeurs charnières » selon la définition de Tukey, et « Moyenne pondérée » à l'aide d'une formule d'interpolation linéaire [Dodge, 1993]. La détermination des premier et troisième quartiles n'est pas standardisée.

Pour la distribution de la durée hebdomadaire du travail dans les 15 pays de l'Union européenne en 2000 (*cf.* tableau 1.4), les premier et troisième quartiles sont les valeurs de profondeur 4,5 :

$$Q_1 = 39{,}15\ h \quad \text{et} \quad Q_3 = 40{,}2\ h$$

Dans le cas d'une variable statistique *continue*, on a $F(Q_1) = 0{,}25$ et $F(Q_3) = 0{,}75$ et on calcule les quartiles par *interpolation linéaire*, en raison de l'hypothèse d'équirépartition. Pour la distribution de l'ancienneté du chômage des femmes (*cf.* figure 1.5) :

$$Q_1 = 1 + 2 \cdot \frac{25 - 16{,}8}{19} \approx 1{,}9 \text{ mois}$$

$$Q_3 = 12 + 12 \cdot \frac{75 - 68{,}7}{18{,}5} \approx 16{,}1 \text{ mois}$$

On peut définir à partir des quartiles Q_1 et Q_3 le paramètre de tendance centrale $(Q_1 + Q_3)/2$, égal à la médiane dans le cas d'une distribution symétrique, ainsi que l'intervalle interquartile $[Q_1, Q_3]$ qui contient 50 % des observations.

Plus généralement, deux quantiles d'ordres complémentaires x_α et $x_{1-\alpha}$ définissent un intervalle dont le milieu peut être considéré comme un paramètre de tendance centrale.

De la même façon, on définit les *déciles* D_1, D_2,… , D_9 qui sont les quantiles $x_{i/10}$ (i = 1 à 9), les *vingtiles*, quantiles $x_{i/20}$ (i = 1 à 19), les *centiles,* etc.

Les classes d'une variable statistique continue sont souvent définies à l'aide des déciles. Dans ce cas, on a 10 classes contenant chacune 10 % de l'effectif total (*cf.* tableau 1.5 et figure 1.11).

Tableau 1.5 – Distribution des salaires annuels nets de tous prélèvements pour les salariés à temps complet du secteur privé et semi-public

Déciles* (en euros courants)	Ensemble		Hommes		Femmes	
	2000	2006	2000	2006	2000	2006
D_1	10 790	12 718	11 230	13 181	10 190	12 075
D_2	12 220	14 219	12 760	14 776	11 420	13 431
D_3	13 520	15 545	14 140	16 209	12 500	14 531
D_4	14 910	16 977	15 580	17 729	13 710	15 715
Médiane	**16 500**	**18 631**	**17 270**	**19 466**	**15 130**	**17 141**
D_6	18 410	20 685	19 330	21 657	16 810	18 924
D_7	20 890	23 430	22 170	24 734	18 850	21 300
D_8	24 780	27 826	26 660	29 787	21 620	24 590
D_9	32 810	36 941	35 020	40 305	26 950	30 962
D_9/D_1	3	2,9	3,2	3,1	2,6	2,6
Salaire moyen	20 400	23 292	21 890	24 912	17 510	20 232

* En 2006, 10 % des salariés à temps complet du secteur privé et semi-public gagnent un salaire annuel net inférieur à 12 718 euros, 20 % inférieur à 14 219 euros…

Source : INSEE.

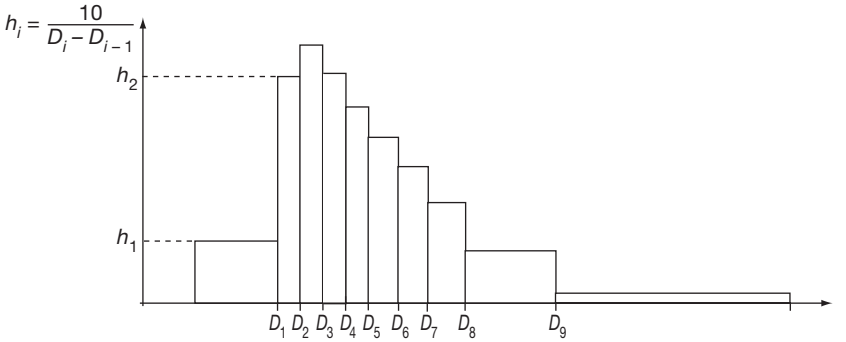


Figure 1.11 – Histogramme de la distribution des salaires « Ensemble » en 2000 (voir tableau 1.5)

C. Les indicateurs de dispersion

1) L'étendue

L'*étendue* est la différence entre la plus grande et la plus petite des valeurs observées :

$$\text{Étendue} = \max_{1 \le i \le n} x_i - \min_{1 \le i \le n} x_i$$

Cette mesure de la dispersion ne dépend que des *valeurs extrêmes* souvent exceptionnelles ; elle ne satisfait pas aux conditions 2 et 5 de Yule. Il faut remarquer aussi que la forme de la distribution entre les valeurs extrêmes n'influe pas sur l'étendue. Cependant, cette caractéristique, étant facile à calculer et ayant une signification concrète facile à comprendre, est fréquemment utilisée en contrôle industriel de fabrication.

2) L'étendue interquartile

De par la définition des quartiles, l'intervalle interquartile $[Q_1, Q_3]$ contient 50 % des observations. Sa longueur, notée EIQ (Étendue InterQuartile), est un indicateur de dispersion :

$$EIQ = Q_3 - Q_1$$

Le calcul de l'étendue interquartile a l'avantage par rapport à celui de l'étendue d'écarter les valeurs extrêmes, souvent sans signification.

Plus généralement, les longueurs des fourchettes définies par les déciles extrêmes, les centiles extrêmes constituent des indicateurs de dispersion contenant respectivement 80 % et 98 % des observations.

3) L'écart absolu moyen

On peut définir une caractéristique de dispersion d'une distribution statistique en calculant les écarts des observations à une tendance centrale C. La tendance centrale de la série $(x_i - C)$ ne peut pas être une mesure de dispersion puisque les écarts positifs sont compensables par les écarts négatifs.

Par contre, la série $|x_i - C|$ définit une variable statistique positive dont les valeurs centrales constituent une mesure de dispersion.

L'écart absolu moyen à la médiane est la moyenne arithmétique des valeurs absolues des écarts à la médiane ; on démontre que c'est *le plus petit écart absolu moyen* :

$$e_{Me} = \frac{1}{n}\sum_{i=1}^{n}|x_i - Me| \quad \text{ou} \quad e_{Me} = \sum_{i=1}^{k} f_i|x_i - Me|$$

L'écart absolu moyen à la moyenne est la moyenne arithmétique des valeurs absolues des écarts à la moyenne arithmétique :

$$e_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \quad \text{ou} \quad e_{\bar{x}} = \sum_{i=1}^{k} f_i|x_i - \bar{x}|$$

Dans le cas d'une variable continue classée, on considère, comme pour le calcul de la moyenne, que chaque individu a sa valeur égale au milieu de sa classe d'affectation.

4) L'écart-type

L'*écart-type* s_X d'une variable statistique X est la mesure de dispersion la plus couramment utilisée.

Algébriquement, il se définit comme la *racine carrée de la variance,* et la variance est la *moyenne arithmétique des carrés des écarts à la moyenne arithmétique* :

$$\text{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{ou} \quad \text{var}(X) = \sum_{i=1}^{k} f_i(x_i - \bar{x})^2 \Rightarrow s_X = \sqrt{\text{var}(X)}$$

Il est possible de développer la formule de la variance pour obtenir une expression mieux adaptée au calcul (mais cette formule devient inusitée de par la diffusion des calculatrices munies des fonctions statistiques [1]) :

1. Les calculatrices munies des fonctions statistiques donnent les valeurs de la moyenne et de l'écart-type d'une variable statistique dont on a saisi la distribution. Certaines calculatrices (dont les calculatrices de marque CASIO®) proposent deux écarts-types : σ_n et σ_{n-1}. La valeur de σ_n correspond à celle de l'écart-type s_X défini ici et utilisé en statistique descriptive ; quant à celle de σ_{n-1}, elle est utilisée en inférence statistique et se déduit de σ_n par la formule suivante : $\sigma_{n-1}^2 = \frac{n}{n-1}\sigma_n^2$

$$\text{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2$$

$$\text{ou} \quad \text{var}(X) = \sum_{i-1}^{k} f_i(x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i x_i^2 - (\bar{x})^2$$

Dans le cas d'une variable statistique continue, on ramène la valeur de chaque individu au milieu de sa classe d'affectation. Là encore, le choix des bornes des classes extrêmes non limitées doit être fait avec précaution.

Mais, alors que pour le calcul de la moyenne, l'erreur liée à ce choix était faible dans le cas de distributions approximativement symétriques autour de la moyenne, il n'en est pas de même pour le calcul de la variance où les erreurs s'ajoutent et ne peuvent pas se compenser.

L'écart-type est exprimé dans la *même unité* que les observations, alors que la variance s'exprime dans le carré de cette unité.

On démontre que l'écart-type, donnant plus de poids aux observations extrêmes que l'écart absolu moyen à la moyenne, lui est toujours supérieur : $s_X \geq e_{\bar{x}}$

Propriétés

1. L'écart-type satisfait aux conditions 1, 2 et 6 de Yule ; l'écart-type est plus *sensible* aux fluctuations d'échantillonnage et aux valeurs extrêmes que la moyenne, en raison des élévations au carré.

2. On montre que la variance est le plus petit écart quadratique moyen, c'est-à-dire :

$$var(X) \leq \frac{1}{n}\sum_{i=1}^{n}(x_i - C)^2 \text{ pour tout } C$$

3. Lorsque deux variables X et Y sont en correspondance par le changement d'origine x_0 et le changement d'échelle a, les écart-types se correspondent par le seul changement d'échelle a pris en valeur absolue :

$$Y = aX + x_0 \implies s_Y = |a| s_X$$

4. Soit une population $\mathbb{P}$ de taille n composée de deux sous-populations $\mathbb{P}_1$ de taille n_1 et $\mathbb{P}_2$ de taille n_2. Soit X, une variable statistique observée sur la population $\mathbb{P}$, on peut exprimer sa variance var(X) en fonction de $\bar{x}$, $\overline{x_1}$ $\overline{x_2}$ var(X_1) et var(X_2) :

$$\text{var}(X) = \frac{1}{n}\left(n_1 var(X_1) + n_2 var(X_2) + n_1(\overline{x_1} - \bar{x})^2 + n_2(\overline{x_2} - \bar{x})^2 \right)$$

Il faut bien remarquer que la variance de X sur $\mathbb{P}$ est la somme pondérée des variances de X sur $\mathbb{P}_1$ et $\mathbb{P}_2$ augmentée de la somme pondérée des carrés des différences entre la moyenne de X sur $\mathbb{P}$ et les moyennes sur $\mathbb{P}_1$ et $\mathbb{P}_2$. Ce résultat se généralise à une partition en k sous-populations ($k \geq 2$).

5. Les distributions statistiques symétriques telles qu'environ :

• 2/3 de la distribution se situent à moins d'un écart-type de $\bar{x}$;

• 95 % de la distribution se situent à moins de deux écarts-types de $\bar{x}$

sont dites normales (chapitre 7, § II).

Le *triplet* $(n, \bar{x}, s_X)$ est un *résumé exhaustif* des distributions de ce type. Dans de nombreux cas, la normalité étant approximative, $(n, \bar{x}, s_X)$ est alors un résumé (quasi-exhaustif) qui présente un intérêt primordial.

D'autres mesures de la dispersion peuvent être envisagées. On peut calculer un *écart médian,* égal à la médiane de la série des valeurs absolues des écarts à une valeur centrale choisie. On peut aussi calculer la *différence moyenne* égale à la moyenne arithmétique des valeurs absolues des différences entre les observations prises deux à deux. C'est cet indicateur de dispersion qui est utilisé pour le calcul de l'indice de concentration de Gini (§ III.E) et qui, ne mesurant pas la dispersion par rapport à la moyenne, est adapté aux distributions non symétriques.

D. Les caractéristiques de forme

La plupart des distributions statistiques sont unimodales. En complément de l'étude de la tendance centrale et de la dispersion, il est intéressant de repérer la forme (déjà mise en évidence par une représentation graphique) par des mesures de son *asymétrie* (en anglais, *skewness*) et de son *aplatissement* (*kurtosis*).

La symétrie est un concept important pour plusieurs raisons. Tout d'abord, la définition de la tendance centrale est sans ambiguité pour une distribution symétrique puisque pour une telle distribution, la médiane est égale à la moyenne et à $(x_\alpha + x_{1-\alpha})/2$ pour tout α compris entre 0 et 0,5, et la dispersion des observations est symétrique par rapport à la moyenne. D'autre part, de nombreuses méthodes statistiques reposent sur une hypothèse de distribution(s) normale(s) ou s'en approchant (chapitre 7). Le caractère de symétrie d'une distribution apparaît donc particulièrement important.

Les mesures de la forme sont indépendantes des unités de mesure de la variable étudiée.

1) Définition des moments centrés

Le *moment centré* d'ordre r d'une distribution est égal à la moyenne arithmétique des puissances d'ordre r des écarts $(x_i - \bar{x})$:

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_1 - \bar{x})^r \qquad \text{ou} \qquad \mu_r = \sum_{i=1}^{k} f_i(x_i - \bar{x})^r$$

→ ***Remarque***

Le moment centré μ_1 est nul, et le moment centré μ_2 n'est autre que la variance et ne peut être nul, comme tous les moments centrés d'ordre pair, que si toutes les observations ont la même valeur.

2) L'asymétrie

Pour une distribution symétrique, la moyenne arithmétique est égale à la médiane et à $(x_\alpha + x_{1-\alpha})/2$ pour α compris entre 0 et 0,5. D'autre part, les moments centrés d'ordre impair sont nuls pour une distribution symétrique, négatifs pour une distribution unimodale étalée à gauche, positifs pour une distribution unimodale étalée à droite. Ces propriétés sont utilisées pour diagnostiquer et mesurer l'asymétrie.

a) Diagnostic et mesure de l'asymétrie à l'aide des quantiles

Dans un cas d'asymétrie, la *comparaison des quantités* $(x_\alpha + x_{1-\alpha})/2$, milieux des intervalles $[x_\alpha, x_{1-\alpha}]$, pour différentes valeurs de α $(0 \leq \alpha \leq 0{,}5)$ donne une indication rapide sur le type de l'asymétrie. Certains logiciels donnent la représentation graphique de ces quantités en fonction des amplitudes $(x_{1-\alpha} - x_\alpha)$. Pour une distribution symétrique, on obtient une droite parallèle à l'axe des abscisses puisque les termes $(x_\alpha + x_{1-\alpha})/2$ sont tous égaux à la médiane (et à la moyenne !).

Pour la distribution des salariés masculins en 2000 (*cf.* tableau 1.5), la comparaison des milieux des intervalles des déciles symétriques par rapport à la médiane montre qu'il s'agit d'une distribution étalée vers la droite :

$$D_5 = 17\,270 < \frac{D_6 + D_4}{2} = 17\,455 < \frac{D_7 + D_3}{2} = 18\,155 < \frac{D_8 + D_2}{2} = 19\,710 < \frac{D_9 + D_1}{2} = 23\,125$$

Le quotient suivant définit un coefficient d'asymétrie, appelé coefficient de Yule et Kendall :

$$\frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

Ce coefficient, compris entre – 1 et + 1, est nul pour une distribution symétrique, positif pour une distribution unimodale étalée vers la droite et négatif dans le cas contraire, et il est, de plus invariant par changement d'origine et d'échelle.

On obtient des variantes de ce coefficient en remplaçant les quartiles par les déciles. Pour les distributions des salaires présentées dans le tableau 1.5, on peut calculer le coefficient d'asymétrie suivant :

$$\frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$$

qui vaut respectivement 0,49 et 0,41 pour les distributions des salaires masculins et féminins en 2000 ; ces valeurs indiquent des distributions asymétriques, étalées vers la droite.

b) Le coefficient d'asymétrie de Fisher

Le *coefficient d'asymétrie de Fisher*, noté γ_1, est ainsi défini :

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} \qquad \text{pour} \qquad \mu_2 \neq 0$$

Comme tout coefficient d'asymétrie, il est *nul* pour une distribution *symétrique*, négatif pour une distribution unimodale étalée vers la gauche, positif pour une distribution unimodale étalée vers la droite (figure 1.12).

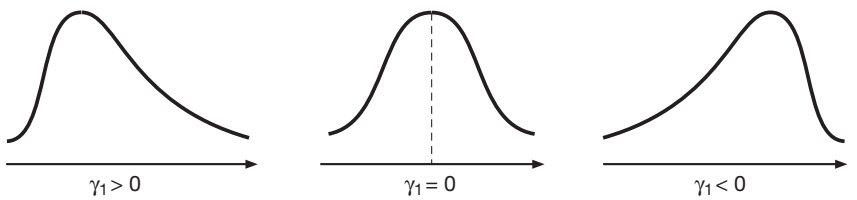


Figure 1.12 – Signe du coefficient d'asymétrie

Les coefficients calculés par les logiciels statistiques sont soit celui de Fisher, soit des variantes de même l'interprétation. Par exemple, le logiciel *SPSS* donne un coefficient d'asymétrie légèrement modifié :

$$\frac{n}{(n-1)\cdot(n-2)}\gamma_1 \qquad \text{pour} \qquad n \geq 3$$

3) L'aplatissement

Les coefficients d'aplatissement mesurent l'aplatissement d'une distribution ou l'importance des « queues » d'une distribution. Le *coefficient d'aplatissement de Fisher,* noté γ_2 , est ainsi défini :

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 \qquad \text{pour} \qquad \mu_2 \neq 0$$

Ce coefficient est *nul* pour une *distribution normale* (chapitre 7), positif ou négatif selon que la distribution est plus ou moins aplatie que la distribution normale de même moyenne et de même écart-type.

Les coefficients calculés par les logiciels sont celui de Fisher ou des variantes de même interprétation.

Ces coefficients d'asymétrie et d'aplatissement sont invariants par changement d'origine et d'échelle, mais ils sont sensibles aux fluctuations d'échantillonnage puisqu'ils font intervenir des moments d'ordre élevé.

E. Les caractéristiques de dispersion relative

Ces caractéristiques permettent de *comparer* les distributions statistiques de plusieurs sous-ensembles d'une même population, ou de faire des comparaisons dans le temps ou dans l'espace.

1) Le coefficient de variation et l'interquartile relatif

Supposons que nous sachions que l'écart-type de poids d'une certaine population est de 8 kg, l'importance du degré de variabilité que cela suggère dépend de la valeur du poids moyen : 10 kg, 50 kg ou plusieurs centaines de kg…

Pour remédier à cette difficulté d'interprétation, il est naturel d'examiner le rapport $s_X / \bar{x}$ appelé *coefficient de variation* et défini en général pour des variables *positives*.

C'est un nombre *sans dimension*, invariant si on effectue un changement d'unité de mesure.

Plus le coefficient de variation est élevé, plus la dispersion autour de la moyenne est élevée.

Ce coefficient permet de comparer les dispersions de distributions qui ne sont pas exprimées dans la même unité (comme des distributions de salaires de pays différents) ou de distributions dont les moyennes sont différentes (comme des distributions de salaires pour différentes qualifications).

On peut construire d'autres coefficients de ce type en utilisant les statistiques d'ordre comme les quartiles et les déciles ; citons l' *interquartile relatif* : $\frac{Q_3 - Q_1}{Q_2}$ et l' *interdécile relatif* : $\frac{D_9 - D_1}{D_5}$

Pour les distributions des salaires « Hommes » et « Femmes » en 2001 (*cf.* tableau 1.5), les interdéciles relatifs valent respectivement 1,45 et 1,12.

2) Les caractéristiques de concentration

La notion de *concentration* a été introduite à propos des distributions de salaires et de revenus. Cette notion est apparentée à celle de dispersion puisqu'elle concerne l'intensité du groupement des données.

Elle ne s'applique qu'à des variables *continues* à valeurs *positives*, et pour des ensembles statistiques dont chaque élément est affecté d'un caractère susceptible d'addition :

- un ensemble de ménages classés selon le revenu, l'épargne, le patrimoine… ;
- un ensemble d'entreprises classées selon le chiffre d'affaire, le nombre de salariés, les montants des factures… ;
- un ensemble d'exploitations agricoles classées selon la surface agricole utilisée.

Il est clair que la notion de concentration ne peut pas s'appliquer, par exemple, à des ensembles d'individus classés selon l'âge, la taille ou le poids, puisque la somme des âges, des tailles ou des poids d'une population est sans signification.

La concentration peut se caractériser, soit par un *procédé graphique,* soit par le *calcul*.

a) Construction de la courbe de concentration

Considérons la distribution des exploitations agricoles par classes de grandeurs des régions Provence-Alpes-Côte d'Azur (PACA) et Midi-Pyrénées en 2005 (*cf.* tableau 1.6). L'intervalle de variation de la *SAU* (superficie agricole utilisée) est partagé en k classes (ici, $k = 9$) dont les bornes supérieures sont notées dans l'ordre : x_1 ,… , x_i , … , x_k

On calcule pour chaque classe ($i = 1$ à k) :

- la *proportion cumulée* p_i des exploitations de *SAU* inférieure à x_i
- la *proportion cumulée* q_i de la *SAU* totale des exploitations de *SAU* inférieure à x_i

Sur un diagramme cartésien, on représente les k points de coordonnées (p_i , q_i). Ces points s'inscrivent dans un carré OABC dont la longueur des côtés est égale à 1 (ou 100 si les proportions sont exprimées en pourcentage).

La courbe qui joint les points successifs est la *courbe de concentration* ou *courbe de Lorenz* (*cf.* figure 1.13). La courbe, toujours en-dessous de la bissectrice, permet de lire que les α % des exploitations les moins bien loties cultivent β % de la *SAU* totale. Si toutes les exploitations ont une part égale de *SAU*, la courbe se confond avec la bissectrice OB. La courbe s'en éloigne lorsque l'inégalité s'accroît.

Tableau 1.6 – Distribution des exploitations agricoles par classes de grandeurs en régions PACA et Midi-Pyrénées

	Midi-Pyrénées		*PACA*		*Midi-Pyrénées*		*PACA*	
	f_i	Proportion *SAU*	f_i	Proportion *SAU*	p_i	q_i	p_i	q_i
Moins de 5 ha	15,5	0,8	44,9	2,6	15,5	0,8	44,9	2,6
5 à moins de 10 ha	9,0	1,4	12,5	3,1	24,6	2,2	57,4	5,7
10 à moins de 20 ha	13,2	4,2	14,8	7,6	37,7	6,4	72,2	13,2
20 à moins de 35 ha	15,7	9,2	9,3	8,6	53,4	15,7	81,5	21,9
35 à moins de 50 ha	12,2	11,1	5,1	7,4	65,6	26,8	86,6	29,3
50 à moins de 100 ha	23,1	35,1	7,2	17,6	88,7	61,9	93,8	46,9
100 à moins de 200 ha	9,6	27,5	3,7	18,1	98,2	89,4	97,5	65,0
200 à moins de 300 ha	1,3	6,6	1,4	11,5	99,5	96,0	98,9	76,5
300 ha ou plus	0,5	4,0	1,1	23,5	100	100	100	100
	100	100	100	100				

Source : agreste.agriculture.gouv.fr

Ceci suggère d'utiliser l'aire, dite *aire de concentration* , comprise entre la courbe et la bissectrice OB comme indicateur d'inégalité.

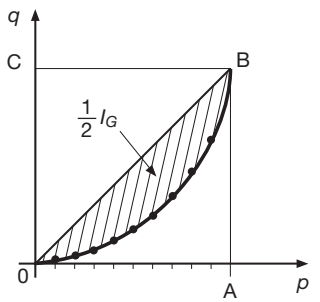


Figure 1.13 – Courbe de Lorenz

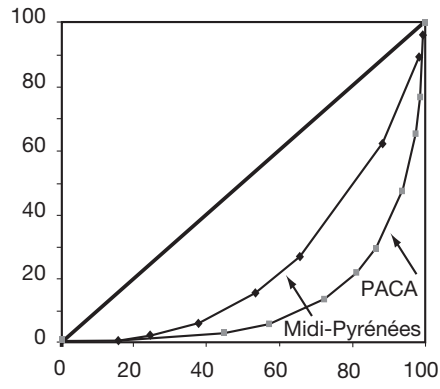
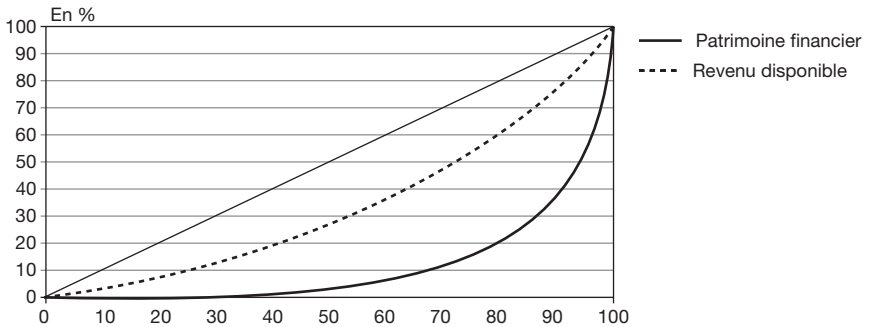


Figure 1.14 – Courbes de concentration des SAU dans les régions PACA et Midi-Pyrénées

On peut *comparer* la concentration de deux ou plusieurs populations selon un même caractère en représentant sur un même graphique leurs courbes de Lorenz. Les terres agricoles sont plus concentrées dans la région PACA que dans la région Midi-Pyrénées puisque la courbe de Lorenz de la *SAU* de la région Midi-Pyrénées est incluse dans celle de la région *PACA* (*cf.* figure 1.14).

On peut aussi comparer la concentration de deux caractères sur une même population : sur la figure 1.15, on constate que la concentration du patrimoine financier des ménages est plus forte que celle des revenus.

Dans les cas où les courbes se coupent, on ne peut pas comparer les degrés d'inégalité.



Lecture : plus la courbe s'éloigne de la diagonale, plus la distribution de la variable considérée est concentrée. La moitié des ménages les moins riches possède 27 % de la masse des revenus disponibles tandis que la moitié des ménages les moins bien dotés possède environ 4 % de la masse totale de patrimoine financier. Les 10 % les mieux dotés en patrimoine financier en possèdent environ 63 %.

Champ : ménages dont la personne de référence n'est pas étudiante et dont le revenu déclaré est positif ou nul.

Sources : enquête *Revenus fiscaux 2003, Insee-DGI* pour le revenu disponible et enquête *Patrimoine 2004, Insee,* montants de patrimoine financier recalés sur les données de la Comptabilité nationale.

Source : INSEE, Économie et Statistique, n° 414, 2008.

Figure 1.15 – Courbes de concentration

b) Détermination de l'indice de concentration ou indice de Gini

L'indice I_G de Gini est égal au double de l'aire de concentration (*cf.* figure 1.13). Cet indice, compris entre 0 et 1, a une valeur d'autant plus élevée que la répartition est plus inégalitaire, et peut être évalué selon la formule [1] :

$$I_G = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} |x_i - x_j|}{n(n-1) \cdot \bar{x}}$$

les x_i ($i = 1,\ldots, n$) désignant ici les valeurs prises (supposées toutes distinctes) par la variable sur chacun des n individus de la population étudiée.

1. Le statisticien italien Corrado Gini a proposé cette mesure de la concentration en 1912 et a montré deux années plus tard que son indice était égal au double de l'aire comprise entre la droite d'équirépartition et la courbe proposée par Max Otto Lorenz en 1905.

Cet indice s'apparente donc bien à la notion de dispersion relative des éléments d'une série. C'est un *nombre sans dimension*. Cette caractéristique de dispersion ne fait pas appel au calcul d'écarts à la moyenne. Elle est ainsi particulièrement *bien adaptée à l'étude de distributions très dissymétriques pour lesquelles la notion d'écart à la moyenne est sans grande signification.*

IV. La boîte de distribution

La ***boîte de distribution*** (*box-plot* en anglais, ou encore « *boîte-à-pattes* », « *boîte à moustaches* », « *boîte de dispersion* » en français) est un outil privilégié de l' *analyse exploratoire des données* . Elle fournit en un seul coup d'oeil des informations sur sa tendance centrale, sa dispersion, son asymétrie, l'importance des valeurs extrêmes. Elle est aussi particulièrement intéressante pour la comparaison de distributions sur plusieurs de ces critères.

A. Résumé d'une distribution par des quantiles

Les *trois quartiles* Q_1, Q_2 et Q_3 et les *deux valeurs extrêmes* fournissent pour une distribution des informations sur sa *tendance centrale* par les quantités Q_2, $\frac{1}{2}(Q_1 + Q_3)$ et $\frac{1}{2}\left(\min_{1\leq i\leq n} x_i + \max_{1\leq i\leq n} x_i\right)$, sur sa *dispersion* par l'étendue et l'étendue interquartile, et sur sa *forme* par la comparaison des trois indicateurs de tendance centrale.

En analyse exploratoire des données, ces cinq valeurs sont présentés avec leur profondeur dans un tableau. Pour la distribution de la durée hebdomadaire du travail en 2000 (*cf.* tableau 1.4) :

$n = 15$	Durée hebdomadaire	
8	$Me = 39{,}9$ h	
4,5	$Q_1 = 39{,}15$	$Q_3 = 40{,}2$
1	$\min_{1\leq i\leq n} x_i = 38{,}5$	$\max_{1\leq i\leq n} x_i = 43{,}6$

On peut compléter ce tableau en indiquant l'étendue interquartile, le milieu de l'intervalle interquartile, l'étendue et le milieu de l'intervalle déterminé par les deux valeurs extrêmes. On obtient ainsi un résumé des informations sur la dispersion et l'asymétrie :

$n = 15$	Durée hebdomadaire		*Dispersion*	*Position*
8	39,9 h			
4,5	39,15	40,2	$EIQ = 1{,}05$	$\frac{1}{2}(Q_1 + Q_3) = 39{,}615$
1	38,5	43,6	*Étendue* = 5,1	$\frac{1}{2}\left(\min_{1\le i\le n} x_i + \max_{1\le i\le n} x_i\right) = 41{,}05$

B. Représentation d'une boîte de distribution

Dans une *boîte de distribution* , la boîte représente l'intervalle interquartile, et à l'intérieur, la médiane la sépare en deux parties. Les lignes qui partent du bord de la boîte s'étendent jusqu'aux valeurs les plus extrêmes qui ne sont pas considérées comme éloignées. Le logiciel SPSS note « valeur éloignée » (o), les points situés à plus de 1,5 fois l'étendue interquartile par rapport aux bords de la boîte, et « valeur extrême » (*), les points situés à plus de 3 fois l'étendue interquartile (*cf.* figure 1.17).

Ainsi, la taille de la boîte représente l'étendue interquartile, la position de la médiane est un bon indicateur de la symétrie de la distribution, la taille des lignes de part et d'autre de la boîte traduit la dispersion, et les valeurs éloignées ou extrêmes sont immédiatement repérées.

On représente une *boîte de distribution* de la façon suivante (*cf.* figure 1.16) :

a) on trace un rectangle de largeur fixée à priori et de longueur $EIQ = (Q_3 - Q_1)$, et on y situe la médiane par un segment positionné à la valeur Q_2, par rapport à Q_3 et Q_1 ; on a alors la *boîte*,

b) on calcule ($Q_3 + 1{,}5 \cdot EIQ$) et ($Q_1 - 1{,}5 \cdot EIQ$) et on cherche :

- la dernière observation x_h en deçà de la limite ($Q_3 + 1{,}5 \cdot EIQ$) soit

$$x_h = \max\{x_i \mid x_i \le Q_3 + 1{,}5 \cdot EIQ\}$$

- la première observation x_b au delà de la limite ($Q_1 - 1{,}5 \cdot EIQ$) soit

$$x_b = \min\{x_i \mid x_i \ge Q_1 - 1{,}5 \cdot EIQ\}$$

c) on trace deux lignes allant des milieux des largeurs du rectangle aux valeurs x_b et x_h

Ainsi, pour la distribution représentée à la figure 1.16, la valeur « éloignée » associée au Royaume-Uni et mise en évidence sur le diagramme *Branche et feuille* de la figure 1.6, est à l'extérieur de la boîte de distribution.

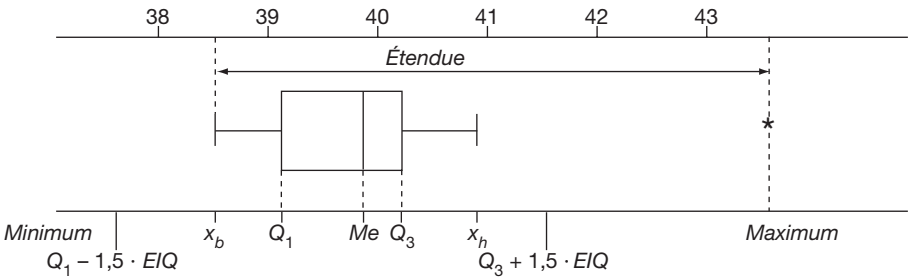


Figure 1.16 – Construction de la boîte de distribution de la durée du travail en 2000 (tableau 1.4)

Ce type de diagramme permet aussi de *comparer* facilement plusieurs distributions en terme de médiane, quartiles et valeurs éloignées ou extrêmes.

On peut représenter en parallèle les boîtes de distribution de la durée hebdomadaire du travail des salariés à temps complet de l'Union européenne en 1990,1995 et 2000, et comparer les trois distributions (*cf.* figure 1.17).

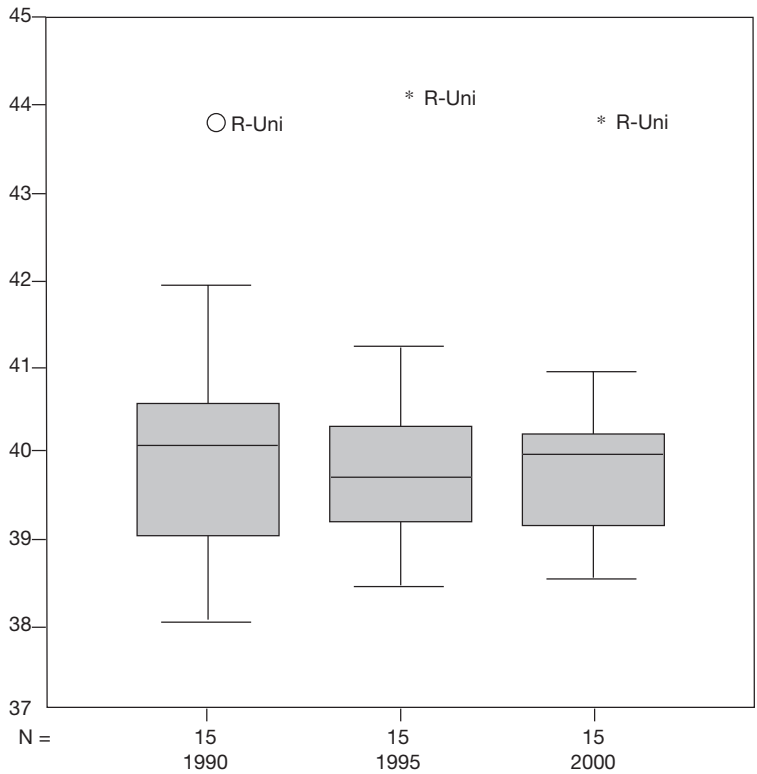


Figure 1.17 – Représentation SPSS des boîtes de distribution du tableau 1.3

La médiane n'évolue pas de façon monotone, la dispersion diminue, le Royaume-Uni passe de « valeur éloignée » en 1990 à « valeur extrême » en 1995 et 2000.

Pour les distributions présentées par leurs déciles (*cf.* tableau 1.5), on ne connaît pas les valeurs individuelles. Dans ce cas, on peut convenir de considérer *valeurs éloignées* les valeurs inférieures au premier décile ou supérieures au neuvième décile.

La représentation des boîtes de distribution des distributions de salaires en 2000 permet de comparer les salaires selon le sexe (*cf.* figure 1.18). La représentation par des histogrammes (*cf.* figure 1.11) ne permettrait pas de comparer aussi aisément les distributions, les histogrammes ne pouvant pas être superposés si on veut conserver la lisibilité, mais seulement juxtaposés.

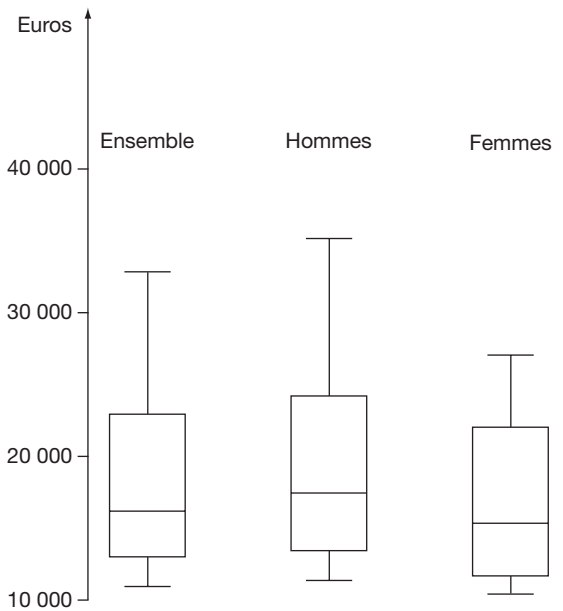


Figure 1.18 – Représentation des boîtes de distribution des salaires en 2000

C. Interprétation d'une boîte de distribution

Une boîte de distribution rend compte de la tendance centrale, de la dispersion, des valeurs éloignées ou extrêmes et de la forme de la distribution (*cf.* figure 1.19), même si d'autre modes de représentation (histogramme, branche et feuille) peuvent apporter un complément d'information sur la forme.

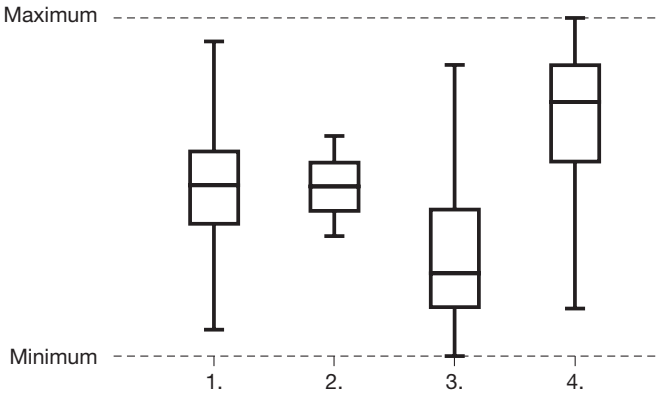


Figure 1.19 – Quelques types de boîtes de distribution :
1. *Distribution symétrique*
2. *Distribution peu dispersée*
3. *Distribution étalée vers les valeurs élevées*
4. *Distribution étalée vers les valeurs faibles*

En statistique descriptive, on a vu l'importance du *triplet* (n , $\bar{x}$ s_X). Pour la distribution de la durée hebdomadaire du travail du tableau 1.4, ce triplet prend les valeurs (15 ; 39,93 ; 1,2) pour l'année 2000. La *boîte de distribution* (*cf.* figures 1.15 et 1.16) est un complément qui se révèle intéressant puisqu'elle permet de détecter l'asymétrie, les valeurs extrêmes, et de repérer la médiane et l'intervalle interquartile qui contient la moitié des observations.

Dans le cas d'une asymétrie, l'écart-type qui mesure la dispersion *symétriquement* par rapport à la moyenne n'est pas la mesure de dispersion la mieux adaptée, et peut être complété par l'étendue interquartile. D'autre part, si la boîte de distribution indique des valeurs éloignées ou extrêmes, on sait que la moyenne et l'écart-type sont particulièrement influencés par ces valeurs.

V. Bilan

Avant toute étude formelle, il est nécessaire de procéder à une évaluation descriptive des données. Cette approche descriptive présente deux difficultés, l'une liée aux calculs, l'autre à la diversité des indicateurs. Si les calculatrices de poche ont permis depuis longtemps déjà de rendre aisés les calculs de moyenne et écart-type, il a fallu attendre la généralisation des moyens de calcul

automatique (en particulier, des logiciels statistiques sur m icro-ordinateurs) pour que tous les indicateurs basés sur la notion de profondeur, et en particulier la médiane, soient facilement accessibles. C'est aussi l'environnement récent des micro-ordinateurs qui a permis de développer les modes de représentation graphique par lesquels on peut appréhender des indicateurs très divers. L'approche descriptive des données trouve dans la représentation graphique un enrichissement et une aide à l'interprétation. Simplicité et inter activité de cette démarche en font une première étape maintenant indispensable à toute étude statistique.

Testez-vous *(les réponses sont données page 283)*

Il y a *au moins* une réponse exacte par question.

1. Pour une série d'observations d'une variable statistique :

a) on peut calculer quatre quartiles

b) l'intervalle interquartile contient 50 % des observations

c) le cinquième décile est égal à la médiane

d) 50 % des observations sont supérieures au premier quartile

2. Pour une variable statistique de distribution symétrique :

a) la moyenne est égale à la médiane

b) 50 % des observations sont supérieures à la moyenne

c) la boîte de distribution contient toutes les observations

d) $(Q_3 - Q_1) = 2(Me - Q_1)$

3. Pour comparer des distributions de variables statistiques exprimées dans des unités différentes (par exemple des distributions de salaires exprimés dans des monnaies différentes), on peut utiliser les caractéristiques suivantes :

a) la médiane

b) l'étendue interquartile

c) le coefficient de variation

d) le rapport D_9 / D_1

4. Pour une série d'observations d'une variable statistique :

a) la somme des écarts à la moyenne est nulle

b) l'écart absolu moyen à la moyenne est un indicateur de dispersion

c) la médiane de la série des écarts absolus à la moyenne est une mesure de l'asymétrie

d) les trois quartiles sont des indicateurs de tendance centrale

5. Une étude des notes obtenues par deux classes d'une école à un test commun a fourni les résultats suivants :

Classe	Classe 1	Classe 2
Effectif	20	30
Moyenne	12	10
Écart-type	4	6
Médiane	12	12

a) la note moyenne des deux classes réunies est égale à 11

b) l'écart-type des notes des deux classes réunies est égal à 5

c) la médiane des notes des deux classes réunies est égale à 12

d) l'écart absolu moyen des notes à la médiane est inférieur ou égal à 4 pour la classe 1

6. Si on veut minimiser l'influence des valeurs extrêmes :

a) on préfère la médiane à la moyenne

b) on préfère l'écart-type à l'écart absolu moyen à la moyenne

c) on préfère l'étendue à l'étendue interquartile

d) on préfère l'étendue interdécile ($D_9 - D_1$) à l'étendue

7. Soit une grandeur dont le taux de croissance au cours de 3 années successives a été de 0,5 % pour les 2 premières années et de 2 % pour la dernière année. Le taux annuel moyen de croissance pendant ces 3 années est égal à :

a) $\left((0{,}005)^2 \cdot (0,\, 02)\right)^{1/3}$

b) $\frac{1}{3}(2 \cdot 0{,}005 + 0{,}02)$

c) $\left((1{,}005)^2 \cdot 1{,}02\right)^{1/3} - 1$

d) une moyenne harmonique

8. Pour la distribution d'une variable statistique continue (ou supposée continue) :

a) l'histogramme est la représentation graphique des fréquences cumulées

b) 15 % des observations sont comprises entre le troisième quartile et le neuvième décile

c) la médiane peut se déterminer à l'aide de la courbe cumulative

d) l'étendue interdécile ($D_9 - D_1$) contient 90 % des observations

9. Si les notes (comprises entre 4 et 16) obtenues à une épreuve de statistique dans une classe de 30 élèves sont toutes augmentées de 2 points :

a) la moyenne sera augmentée de 2 points

b) l'écart-type sera augmenté de 2 points

d) la médiane sera augmentée de 2 points

d) l'étendue sera augmentée de 2 points

Exercices *(corrigés page 289)*

Exercice 1.1

Le tableau suivant donne la répartition des familles selon le nombre d'enfants et leur âge de 1968 à 1999 :

Enfants de 0 à 18 ans (milliers)

	1968	1975	1982	1990	1999
Ensemble	12 063	13 176	14 119	15 391	16 097
sans enfant	5 302	5 836	6 508	7 900	8 679
avec enfants	6 760	7 340	7 610	7 491	7 418
1 enfant	2 723	3 110	3 303	3 281	3 317
2 enfants	2 052	2 374	2 734	2 756	2 772
3 enfants	1 063	1 088	1 081	1 063	1 008
4 enfants	481	427	310	259	230
5 enfants ou plus	441	342	183	132	91
Nombre total d'enfants	14 569	14 826	14 294	13 748	13 308

Sources : Recensements de la population, *INSEE*

1. Définir les populations étudiées, l'unité statistique, le caractère étudié et sa nature.

2. Examinez l'évolution du nombre total de familles sans enfant, du nombre de familles avec enfants, avec un enfant, avec deux enfants…

3. On considère dans cette dernière question les familles avec enfant(s).

3.1. Après avoir calculé les fréquences, tracez les diagrammes en bâtons de ces distributions, et indiquez le mode.

3.2. Pour chacune des cinq années, calculez le nombre moyen d'enfants par famille et l'écart-type (on considérera le nombre moyen d'enfants des familles ayant cinq enfants ou plus égal à 6). Commentez les résultats.

Exercice 1.2

Le tableau suivant donne la distribution du niveau de l'indice de la qualité de l'air *ATMO* en agglomération parisienne de 2000 à 2006 (en nombre de jours par an).

1. Définir les populations étudiées, l'unité statistique, le caractère étudié et sa nature.

2. Tracez le diagramme en bâtons de la distribution en 2006, et indiquez le mode.

3. Calculez les niveaux annuels moyens de 2000 à 2006.

Niveau	Qualité	2000	2001	2002	2003	2004	2005	2006	Total
1	Très bon	0	0	0	0	0	0	0	0
2	Très bon	8	15	9	15	23	23	25	118
3	Bon	206	190	183	138	186	188	177	1 268
4	Bon	99	97	111	109	96	99	106	717
5	Moyen	36	33	45	47	39	34	26	260
6	Médiocre	13	13	8	30	19	11	16	110
7	Médiocre	2	14	7	16	2	6	11	58
8	Mauvais	2	3	2	10	1	4	4	26
9	Mauvais	0	0	0	0	0	0	0	0
10	Très mauvais	0	0	0	0	0	0	0	0
	Total	366	365	365	365	366	365	365	2 557

Source : AIRPARIF.

Exercice 1.3

On a relevé pendant 50 quinzaines successives les niveaux de ventes, exprimés en milliers d'unités de produit, de deux présentations notées *G* (Gel) et *P* (Poudre) d'un même produit. Les résultats sont les suivants :

Présentation *G*				
Niveau de vente	< 5	[5-10[	[10-12[	[12-20]
Nombre de quinzaines	5	20	15	10

Présentation *P*				
Niveau de vente	< 10	[10-12[	[12-16[	[16-20]
Nombre de quinzaines	10	25	10	5

1. Calculez les moyennes, écarts-types et médianes des niveaux de ventes pour chacune des deux présentations.

Quelle est la condition nécessaire sur la moyenne et la médiane d'une distribution pour que celle-ci soit symétrique ?

2. Sur l'ensemble des points de vente pour toute la période de l'étude, on disposait de 30 % du produit en gel, et de 70 % du produit en poudre.

Quel a été le niveau de ventes moyen pour l'ensemble des deux présentations du produit ?

3. Les niveaux de ventes étant maintenant exprimés en centaines d'unités de produit, donnez les nouvelles valeurs des moyennes, écarts-types et médianes calculées à la 1re question.

Exercice 1.4

Afin d'étudier les disparités de salaires entre hommes et femmes, une enquête a été réalisée auprès du personnel ouvrier d'un secteur industriel. Les résultats concernant les salaires annuels nets en euros sont résumés dans les deux tableaux suivants :

Tableau 1. Hommes

Effectif	180
Salaire moyen	15 400
Écart-type	3 620
1er décile	10 950
1er quartile	12 750
Médiane	14 800
3e quartile	17 660
9e décile	20 220

Tableau 2. Femmes

Salaire annuel (en milliers d'€)	Nombre d'ouvrières
[10 ; 12[	82
[12 ; 14[	34
[14 ; 16[	12
[16 ; 20]	n_4
Total	N

1. Définir la population étudiée, l'unité statistique, le caractère étudié et sa nature.

2. Proposez pour la distribution du salaire des hommes en précisant les valeurs correspondantes :

- trois indicateurs de tendance centrale ;
- deux indicateurs de dispersion ;
- deux indicateurs de dispersion relative.

3. Sachant que le salaire annuel moyen des femmes enquêtées est égal à 12 000 €, déterminez l'effectif n_4 de la dernière classe de la distribution du salaire des femmes, ainsi que l'effectif total N.

4. Déterminez l'écart-type et le coefficient de variation de la distribution des femmes.

5. Déterminez le salaire annuel moyen de l'ensemble des ouvriers hommes et femmes de l'enquête.

Exercice 1.5

Dans un atelier, le coût horaire de la main d'oeuvre est de 8 € (base 35 h par semaine). Une heure supplémentaire revient à 10 €, et le service de paie indique que le coût total des heures supplémentaires représente 30 % du coût total de la main d'oeuvre.

Calculez le coût horaire moyen et indiquez le type de moyenne utilisée.

Exercice 1.6

Une même somme S a été confiée à deux banques B_1 et B_2 pour une durée de 10 ans. Les rendements successifs des placements effectués par les deux banques ont été les suivants :

- Banque B_1 : 12 % pendant 2 ans, puis 8 % pendant 4 ans, puis 6 % pendant 4 ans ;
- Banque B_2 : 10 % pendant 3 ans, puis 8 % pendant 3 ans, puis 7 % pendant 4 ans.

Les intérêts sont toujours capitalisés en fin d'année.

1. Calculez le taux moyen de croissance du placement dans chaque banque.

2. À quel taux la moins performante des deux banques aurait-elle dû placer l'argent pendant la troisième période pour égaler l'autre ?

Exercice 1.7

Le tableau ci-après donne des caractéristiques des 30 premiers groupes français de l'industrie et des services selon leur chiffre d'affaires en 2001 (Source : *Tableaux de l'Économie Française 2003-2004*, INSEE) :

Société	*CAHT* (millions d'€)	Effectif	Société	*CAHT* (millions d'€)	Effectif
TotalFinaElf	105 318	122 025	Aventis	22 941	91 729
Carrefour	69 486	382 821	Groupe Casino (Rallye)	21 984	106 736
Vivendi Universal	57 360	321 000	Bouygues	20 473	126 560
PSA Peugeot Citroën	51 663	192 500	Airbus (EADS)	20 427	2 000
France Telecom	43 026	206 184	SNCF	20 129	220 747
Suez	42 359	188 050	Vonci	17 172	129 499
EDF	40 716	161 738	La poste	17 028	313 854
Les Mousquetaires	37 200	112 000	Publicis Groupe	16 667	20 592
Renault	36 351	140 417	Michelin	15 775	127 467
Saint-Gobain	30 390	173 329	Havas	14 950	20 373
Pinault-Printemps-La Redoute	27 799	115 935	Usinor (Arcelor)	14 523	59 516
Groupe Auchan	26 200	136 000	Groupe Danone	14 470	100 560
Alcatel Alsthom	25 353	99 314	Gaz de France	14 357	36 451
Galec (Leclerc)	25 000	75 000	L'Oréal (Gespartal)	13 740	49 150
Alstom	23 453	118 995	Lafarge	13 698	82 892

1. Définir la population étudiée, l'unité statistique et les caractères étudiés.

2. Calculez la moyenne et l'écart-type du chiffre d'affaires et de l'effectif.

3. Étude du chiffre d'affaires des 30 premiers groupes français.

3.1. Déterminez les trois quartiles.

3.2. Représentez le diagramme *branche et feuille* de cette distribution.

3.3. Représentez la *boîte de distribution*.

4. Quel est l'intérêt de chacune de ces deux représentations graphiques comparativement à un histogramme ?

5. Reprendre la question 3 pour l'étude de l'effectif.

Exercice 1.8

Le tableau suivant donne le revenu annuel moyen des ménages, en euros, pour les dix intervalles définis par les déciles, et la part de chaque intervalle dans le revenu total.

Valeur des déciles (euros)	Intervalle	Revenu moyen dans l'intervalle	% de la masse totale des revenus dans l'intervalle
$D_1 = 7\,304$	$< D_1$	3 845	2
$D_2 = 11\,091$	$[D_1 ; D_2[$	9 318	3
$D_3 = 14\,099$	$[D_2 ; D_3[$	12 601	5
$D_4 = 17\,219$	$[D_3 ; D_4[$	15 640	6
$D_5 = 20\,631$	$[D_4 ; D_5[$	18 863	7
$D_6 = 24\,653$	$[D_5 ; D_6[$	22 579	9
$D_7 = 29\,361$	$[D_6 ; D_7[$	26 904	11
$D_8 = 35\,757$	$[D_7 ; D_8[$	32 324	13
$D_9 = 46\,642$	$[D_8 ; D_9[$	40 548	16
	$\geq D_9$	69 930	28

Source : INSEE, Revenus fiscaux 1999, hors revenus du patrimoine.

1. Définir la population, l'unité statistique, le caractère étudié et sa nature.

2. Calculez le revenu annuel moyen des ménages.

3. Est-il légitime de faire l'hypothèse d'équirépartition dans les classes définies par les déciles ?

4. Proposez trois indicateurs de tendance centrale, un indicateur de dispersion et un indicateur de dispersion relative. Donnez les valeurs de ces indicateurs.

5. Cette distribution de revenus est-elle symétrique ? (justifiez votre réponse)

6. Proposez un indicateur de disparité des revenus, et donnez sa valeur. Interprétez.

7. Quelle est la part de l'ensemble des revenus perçus par les 4 dixièmes des ménages aux revenus les plus faibles ?

8. Soit $F_1 = 10\ \%$, $F_2 = 20\ \%$, … , $F_{10} = 100\ \%$, et R_i la part de l'ensemble des revenus perçus par l'ensemble des F_i ménages aux revenus les plus faibles.

8.1. Tracez la courbe joignant, dans l'ordre, les points (F_i , R_i). Comment s'appelle cette courbe ?

8.2. Rappelez l'interprétation graphique de l'indice de concentration de Gini ?

8.3. Quelles sont les valeurs minimum et maximum de cet indice ?

8.4. À quelles situations correspondent-elles ?

D'après examen de juin 2004, GEA 1re année Paris-Dauphine.

Exercice 1.9

Le tableau suivant donne le nombre (en milliers) et la superficie agricole utilisée (*SAU*, en milliers d'ha) des exploitations agricoles en France métropolitaine par classes de grandeur pour les années 1979, 1988, 2000 et 2005.

	1979		1988		2000		2005	
	Nombre	*SAU*	Nombre	*SAU*	Nombre	*SAU*	Nombre	*SAU*
Moins de 5 ha	357	677	278	519	193	362	132	262
5 à moins de 20 ha	410	4 778	279	3 238	132	1 464	104	1 163
20 à moins de 50 ha	347	10 962	288	9 348	138	4 666	109	3 714
50 à moins de 100 ha	114	7 683	128	8 709	122	8 662	113	8 083
100 à moins de	29	3 798	37	4 864	64	8 655	70	9 486
200 ha	6	1 598	7	1 918	15	4 047	17	4 762
200 ha ou plus								
Ensemble	1 263	29 496	1 017	28 596	664	27 856	545	27 470

Source : INSEE.

1. Définir la population, l'unité statistique, le caractère étudié et sa nature.

2. Calculez, en pourcentage, les taux annuels moyens de variation du nombre des exploitations agricoles de 1979 à 1988, de 1988 à 2000, de 2000 à 2005.

Exprimez le taux annuel moyen de variation de 1979 à 2005 en fonction de ces 3 taux, de quel type de moyenne s'agit-il ?

Calculez sa valeur.

3. Pour les années 1979, 1988, 2000 et 2005, calculez la *SAU* moyenne et la *SAU* moyenne des exploitations de 50 ha ou plus.

4. Pour l'année 2005, représentez l'histogramme de la distribution des exploitations agricoles, ainsi que la courbe de concentration de la *SAU*.

2. Indices statistiques

Pour l'étude des problèmes économiques et sociaux, on a souvent besoin de décrire les variations de grandeurs simples telles que le prix du baril de pétrole, la production de blé, le taux de fécondité… Ces comparaisons dans le temps (ou dans l'espace) se font généralement en effectuant le rapport des valeurs de la grandeur considérée à deux dates différentes (ou en deux lieux distincts) ; on parle d'*indice statistique élémentaire.*

Mais, il est important d'être en mesure de suivre les évolutions de grandeurs complexes telles que le niveau général des prix, la production industrielle, les exportations… Celles-ci peuvent être résumées par une caractérisque de tendance centrale d'indices élémentaires, ce qui amène à la construction d' *indices synthétiques.*

Toute caractéristique de tendance centrale, notamment les différents types de moyennes, présentant à la fois des avantages et des inconvénients, il n'est pas possible de proposer une méthode unique de construction des indices synthétiques. Il existe différentes formules. On va exposer les plus utilisées.

De par l'importance que revêtent ces indicateurs d'évolution dans les discussions économiques et politiques, il est nécessaire de bien comprendre leur élaboration, d'analyser leurs modes de construction et d'étudier leurs propriétés.

I. Indices élémentaires

A. Définition

On appelle indice élémentaire de la grandeur simple x à la date (ou période) t, dite *date courante* , par rapport à la date 0, dite *date de référence* , le rapport :

$$I_{t/0}(x) = \frac{x_t}{x_0}$$

On a l'habitude, pour éviter de traiter des valeurs d'indice avec trop de chiffres après la virgule de multiplier le résultat par 100 et de laisser un chiffre après la virgule. Une variation négative est repérée par une valeur inférieure à 100.

▶ **Exemple**

La population de la France métropolitaine est passée de 53 731 milliers d'habitants au 1er janvier 1980 à 56 577 milliers d'habitants au 1er janvier 1990 et à 58 749 milliers d'habitants au 1er janvier 2000 (*source : Tableaux de l'Économie française 2003-2004*, INSEE) :

$$I_{1990/1980}(P) = 100 \cdot \frac{56\,577}{53\,731} \approx 105,3$$

$$I_{2000/1980}(P) = 100 \cdot \frac{58\,749}{53\,731} \approx 109,3$$

$\Rightarrow$ La population française a augmenté de 5,3 % de 1980 à 1990 et de 9,3 % de 1980 à 2000.

B. Propriétés

1) Circularité (ou transitivité ou transférabilité)

$$I_{t/0}(x) = I_{t/t'}(x) \cdot I_{t'/0}(x)$$

Cette formule permet de changer de base en passant de la date de référence 0 à la date de référence t' :

$$I_{t/t'}(x) = \frac{I_{t/0}(x)}{I_{t'/0}(x)}$$

L'utilisateur a en effet souvent besoin de mesurer l'évolution d'une grandeur entre deux dates différentes de la date de référence.

De cette propriété, résulte la propriété d'enchaînement :

$$I_{t/0}(x) = I_{t/t-1}(x) \cdot \ldots \cdot I_{1/0}(x)$$

2) Réversibilité

$$I_{0/t}(x) = \frac{1}{I_{t/0}(x)}$$

Cette propriété est intéressante dans le cas de comparaison géographique, car le choix du lieu de référence est arbitraire.

3) Multiplication

Si une grandeur simple z est le produit de deux grandeurs x et y, l'indice élémentaire de la grandeur produit est égal au produit des indices des grandeurs facteurs :

$$\text{quel que soit } t : z_t = x_t \cdot y_t \Rightarrow I_{t/0}(z) = I_{t/0}(x) \cdot I_{t/0}(y)$$

Cas particulier fondamental :

Valeur = Prix · Quantité ou encore : **Dépense = Prix · Volume**

Cette égalité entraîne :

Indice élémentaire de valeur =
Indice élémentaire de prix · Indice élémentaire de quantité

Ces propriétés immédiates d'un indice élémentaire ne sont généralement pas satisfaites par un indice synthétique.

II. Indices synthétiques

Les indices élémentaires retracent l'évolution d'une seule grandeur parfaitement définie et homogène.

Mais, le plus souvent, l'économiste ou le dirigeant d'entreprise, si ce n'est le citoyen désire suivre les variations de grandeurs complexes telles que les prix, la production industrielle…

Ces grandeurs complexes sont composées d'un nombre plus ou moins important de grandeurs simples dont l'évolution est décrite par un indice élémentaire.

On appelle indice synthétique, un indice faisant intervenir dans son calcul plusieurs grandeurs intéressant un même phénomène économique. Ce type d'indice résulte d'un calcul de moyenne .

Il est impossible de proposer une méthode unique et incontestable permettant de décrire l'évolution d'une grandeur complexe.

Les indices synthétiques ont l'inconvénient de ne pas présenter généralement les propriétés de circularité et réversibilité. Or, ces propriétés seraient très utiles au calcul économique ; les changements de base et les raccordements d'indices ne peuvent être effectués de façon rigoureuse que sur des indices possédant la *propriété de circularité.*

A. Indices synthétiques de Laspeyres et Paasche : premières formules

Soient deux dates 0 et *t*, la situation à chaque date est caractérisée par les quantités disponibles de *n* biens physiques hétérogènes q_0^i $(i = 1, 2, \ldots, n)$ – respectivement q_t^i – non sommables, le prix de chaque unité étant p_0^i – respectivement p_t^i

Seules les valeurs des divers biens sont sommables. On peut définir un *indice élémentaire de valeur* qui retrace l'évolution de la valeur sous l'influence simultanée des variations de prix et de quantité :

$$V_{t/0} = \frac{v_t}{v_0} = \frac{\sum_i p_t^i q_t^i}{\sum_i p_0^i q_0^i}$$

Pour séparer les deux influences et chiffrer les variations « moyennes » des prix et celles des quantités, il est nécessaire de recourir à des indices synthétiques. Le problème est de décomposer la variation entre la situation 0 et la situation t en ce qui est dû à la variation des prix et en ce qui est dû à la variation des quantités vendues.

Première idée :

Quelle aurait été la recette (ou la dépense) si les prix étant restés ce qu'ils étaient à la date 0, les ventes (ou les achats) avaient été celles (ou ceux) de la date t ?

Cela revient à mesurer seulement l'effet de la variation des quantités :

$$L_{t/0}(q) = \frac{\sum_i p_0^i q_t^i}{\sum_i p_0^i q_0^i}$$

On définit $\Pi_{t/0}(p)$ tel que :

$$V_{t/0} = L_{t/0}(q) \cdot \Pi_{t/0}(p) \quad \Rightarrow \quad \Pi_{t/0}(p) = \frac{\sum_i p_t^i q_t^i}{\sum_i p_0^i q_t^i}$$

On peut aussi proposer l'*autre solution* suivante :

$$V_{t/0} = L_{t/0}(p) \cdot \Pi_{t/0}(q) = \frac{\sum_i p_t^i q_0^i}{\sum_i p_0^i q_0^i} \cdot \frac{\sum_i p_t^i q_t^i}{\sum_i p_t^i q_0^i}$$

$L(p)$ et $L(q)$ sont les indices de Laspeyres des prix et des quantités, $\Pi(p)$ et $\Pi(q)$ sont les indices de Paasche des prix et des quantités [1].

Essayons d'exprimer littérairement la différence entre l'indice de Laspeyres et l'indice de Paasche. Pour un indice des prix par exemple :

- indice de Laspeyres : on fige le *panier*[2] dans sa composition de la période de base et on compare la valeur qu'il aurait à la période courante avec sa valeur réelle à la période de base ;
- indice de Paasche : on fige le *panier* dans sa composition de la période courante, on calcule rétrospectivement ce qu'aurait été sa valeur à la période de base et on la compare avec sa valeur actuelle.

B. Formules développées

1) Indice de Laspeyres

$$L_{t/0}(p) = \frac{\sum_i q_0^i p_t^i}{\sum_i q_0^i p_0^i} = \sum_i \frac{q_0^i p_0^i}{\sum_i q_0^i p_0^i} \cdot \frac{p_t^i}{p_0^i}$$

La pondération $k_0^i = \dfrac{q_0^i p_0^i}{\sum_i q_0^i p_0^i} = \dfrac{q_0^i p_0^i}{v_0}$ s'interprète dans un indice des prix de détail, comme le *coefficient budgétaire* (structure de valeurs) du produit « i », c'est-à-dire la part de dépense totale qui lui est consacrée, à la période de base. On constate que la somme de ces pondérations est égale à 1. L'indice de Laspeyres des prix apparaît comme *une moyenne arithmétique pondérée* des indices élémentaires des prix des biens individuels.

On montre de même :

$$L_{t/0}(q) = \frac{\sum_I p_0^i q_t^i}{\sum_I p_0^i q_0^i} = \sum_i \frac{p_0^i q_0^i}{\sum_i p_0^i q_0^i} \cdot \frac{q_t^i}{q_0^i}$$

1. Étienne Laspeyres (économiste et statisticien allemand d'origine française) et Hermann Paasche (statisticien allemand) proposèrent ces formules respectivement en 1864 et 1874.
2. *Panier* : expression INSEE, le « panier » par rapport à celui de la ménagère a la particularité de contenir aussi des services immatériels (tickets d'autobus, mois de loyer, biens durables comme appareils ménagers…).

Ces formules développées apparemment plus compliquées que les premières, sont plus pratiques à appliquer ; c'est sous cette dernière forme que les instituts de statistique calculent les indices de Laspeyres, les plus fréquemment utilisés. Ils déterminent d'abord les coefficients de pondération, structure des valeurs de la période de base, et les appliquent aux indices élémentaires de prix ou de quantités relevés mois après mois.

2) Indice de Paasche

$$\Pi_{t/0}(p) = \frac{\sum_i q_t^i p_t^i}{\sum_i q_t^i p_0^i} = \frac{\sum_i q_t^i p_t^i}{\sum_i q_t^i p_t^i \cdot \frac{p_0^i}{p_t^i}} \quad \Rightarrow \quad \frac{1}{\Pi_{t/0}(p)} = \sum_i \frac{q_t^i p_t^i}{\sum_i q_t^i p_t^i} \cdot \frac{p_0^i}{p_t^i}$$

L'indice de Paasche des prix ou des quantités est *la moyenne harmonique* des indices élémentaires (de prix et de quantités) *pondérée* par les structures de valeurs de la période courante.

C. Comparaison des indices de Laspeyres et de Paasche

On sait que la moyenne harmonique est inférieure à la moyenne arithmétique, mais on ne peut comparer les indices de Laspeyres et de Paasche que si les coefficients de pondération sont les mêmes.

L'indice de Paasche est souvent plus petit que l'indice de Laspeyres. En effet, si les coefficients ne changeaient pas entre la date de base et la date courante, l'indice de Paasche, moyenne harmonique, serait inférieur à celui de Laspeyres qui est une moyenne arithmétique. Pour que l'indice de Paasche dépasse l'indice de Laspeyres, il faut que les pondérations des indices élémentaires tendent à se modifier dans le sens d'un accroissement pour ceux qui sont élevés, et d'une diminution pour ceux qui sont faibles.

Conformément à la loi économique de l'offre et de la demande, les consommateurs ont tendance à acheter moins lorsque les prix sont élevés et à acheter davantage quand les prix baissent. Ce phénomène, appelé parfois la demande élastique, n'est valable que dans le cas où les biens ne servent pas de façon essentielle.

Dans le cas de l'indice de Laspeyres, le numérateur $\sum_i q_0^i p_t^i$ (*cf.* premières formules) est un peu plus fort qu'il ne devrait l'être, car, conformément à la loi de l'offre et de la demande, les consommateurs ont tendance à acheter

moins de biens de prix élevés et davantage de biens bon marché. Il en résulte que le coût total sera inférieur à celui donné par $\sum_i q_0^i p_t^i$. Ainsi, l'*indice de Laspeyres* a tendance à *surévaluer* une hausse.

Dans le cas de l'indice de Paasche, les rôles joués par les quantités consommées pendant l'année de référence et les quantités consommées pendant l'année considérée sont diamétralement opposés de ceux joués par ces mêmes quantités dans le cas de l'indice de Laspeyres. L'*indice de Paasche* a donc tendance à *sous-évaluer* une hausse.

▶ **Exemple**

Entre janvier 2006 et janvier 2010, l'évolution des prix et du nombre d'exemplaires de journaux vendus en un mois par une société de presse éditant trois journaux mensuels *A*, *B* et *C* a été la suivante :

	Janvier 2006		Janvier 2010	
	Prix (en euros)	Quantité	Prix (en euros)	Quantité
Journal *A*	2,5	8 000	3	6 500
Journal *B*	4	4 000	4,5	5 000
Journal *C*	5	2 000	6	1 500

i) La variation des recettes de la société de presse entre janvier 2006 et janvier 2010 est de 10,9 %, en effet :

$$V_{2010/2006} \cdot 100 = \frac{51\,000}{46\,000} \cdot 100 \approx 110{,}9$$

ii) Cette variation fait intervenir un effet-quantité et un effet-prix qu'on peut évaluer en calculant les indices des prix et des quantités de Laspeyres et de Paasche :

$$L_{2010/2006}(p) = 117{,}4 \qquad \Pi_{2010/2006}(p) = 116{,}6$$
$$\Rightarrow \qquad L_{2010/2006}(p) > \Pi_{2010/2006}(p)$$
$$L_{2010/2006}(q) = 95{,}1 \qquad \Pi_{2010/2006}(q) = 94{,}4$$
$$\Rightarrow \qquad L_{2010/2006}(q) > \Pi_{2010/2006}(q)$$

iii) La variation de la valeur globale peut être décomposée en ses deux effets prix et quantité. En effet, à partir de la formule :

$$V_{2010/2006} = L_{2010/2006}(p) \cdot \Pi_{2010/2006}(q) = L_{2010/2006}(q) \cdot \Pi_{2010/2006}(p)$$

On peut établir le schéma de décomposition donné à la figure 2.1.

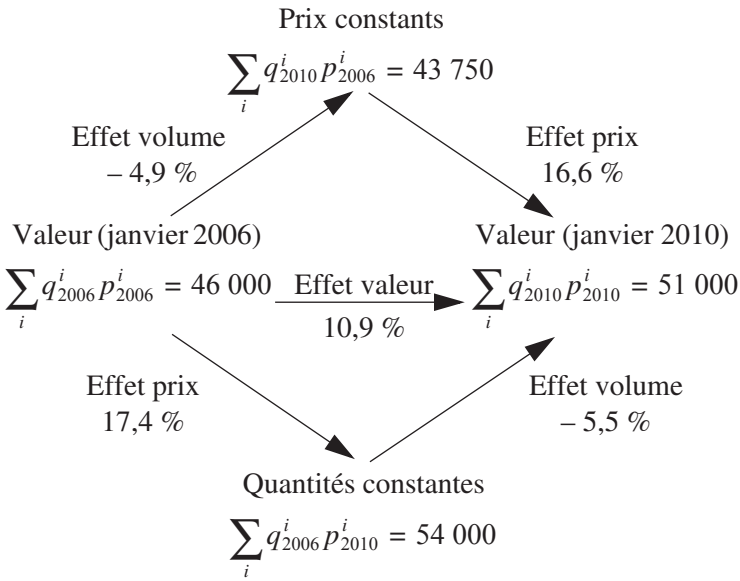


Figure 2.1 – Schéma de décomposition de l'évolution d'un indice de valeur

D. Indice de Fisher

Cet indice a été construit à la suite de la recherche d'un indice *idéal*.

Définition

$$F_{t/0}(p) = \sqrt{L_{t/0}(p) \cdot \Pi_{t/0}(p)}$$

Cette définition provient du développement suivant :

$V_{t/0} = L_{t/0}(p) \cdot \Pi_{t/0}(q) = L_{t/0}(q) \cdot \Pi_{t/0}(p)$

Les indices de Laspeyres et de Paasche étant des nombres positifs, on peut écrire :

$$V^2_{t/0} = L_{t/0}(p) \cdot \Pi_{t/0}(p) \cdot L_{t/0}(q) \cdot \Pi_{t/0}(q) \Rightarrow V_{t/0} = F_{t/0}(p) \cdot F_{t/0}(q)$$

Moyenne géométrique des indices de Laspeyres et de Paasche, la valeur de l'indice de Fisher est comprise entre les valeurs de ces deux indices.

Comme l'indice de Laspeyres a tendance à surestimer une hausse de prix, tandis que l'indice de Paasche a tendance à la sous-estimer, on en déduit que l'indice de Fisher doit donner une meilleure estimation d'une hausse des prix.

E. Propriétés des indices de Fisher, Laspeyres et Paasche

• Les indices de Laspeyres et de Paasche ne sont pas réversibles, mais :

$$\Pi_{t/0} = \frac{1}{L_{0/t}} \quad \Rightarrow \quad \Pi_{t/0} \cdot L_{0/t} = 1$$

L'indice de Fisher est donc *réversible*, ce qui en fait un outil privilégié dans les comparaisons géographiques.

• Ces trois indices ne sont pas transitifs.

• *Agrégation*

Les indices de Laspeyres et de Paasche ont des structures de moyenne. On peut calculer la moyenne arithmétique d'un ensemble à partir des moyennes des sous-ensembles qui le composent. Il en résulte que l'indice de Laspeyres (resp. de Paasche) d'un ensemble peut s'obtenir à partir des indices des groupes formant cet ensemble en leur appliquant la formule de Laspeyres (resp. de Paasche).

Les 303 postes de dépenses, répartis en 159 groupes, servant aux calculs des indices actuels des prix à la consommation, base 100 en 1998, font l'objet de regroupements en 12 fonctions (ex : 01 produits alimentaires et boissons non alcoolisées) et 37 sous-fonctions (ex : 01.1 produits alimentaires)[1]. C'est la formule de Laspeyres qui est utilisée. On commence par calculer l'indice de Laspeyres de chacun des regroupements. On obtient ensuite l'indice d'ensemble en appliquant à nouveau la formule de Laspeyres à ces sous-indices, avec des coefficients de pondération égaux aux parts de chacun des regroupements dans la valeur de la consommation totale. Cette propriété permet de publier non seulement un indice global, mais aussi des sous-indices correspondant aux groupes et sous-groupes.

Qualité	Laspeyres	Paasche	Fisher
Réversibilité	non mais : $L_{0/t} = \frac{1}{\Pi_{t/0}}$	non mais : $\Pi_{0/t} = \frac{1}{L_{t/0}}$	oui
Transitivité	non	non	non
Agrégation	oui	oui	non
Emploi	couramment utilisé	peu utilisé	quasiment inusité

1. « Le nouvel indice des prix à la consommation, année de base 1998 », *Bulletin Mensuel de la Statistique,* n° 2-1999, INSEE.

F. Utilisation de ces trois indices

L'indice de Laspeyres est le plus commode à utiliser ; la plupart des indices courants établis par les instituts du monde entier sont du type « Laspeyres ».

L'indice de Paasche, symétrique de celui de Laspeyres quant à sa signification, présente des inconvénients pratiques à cause de la mise à jour permanente de ses pondérations. Il n'est, de ce fait, pas utilisé dans le calcul direct des indices courants. Son calcul est néanmoins intéressant pour obtenir avec l'indice de Laspeyres une *fourchette* d'estimation.

L'indice de Fisher est quasiment inusité, car son calcul ne peut pas se faire par *agrégation progressive.*

Lorsqu'on divise un indice de valeur par un indice de Laspeyres de prix (resp. de quantités), on obtient un indice de Paasche de quantités (resp. de prix). Si on *déflate*[1] l'indice rendant compte de l'évolution de la masse salariale (indice de valeur) par un indice de Laspeyres des prix (se rapportant évidemment aux mêmes dates), on obtient un indice de pouvoir d'achat de la masse salariale qui est un indice de Paasche des quantités consommables.

On dispose assez souvent de séries de valeur totale : chiffre d'affaire, montant des investissements… Pour obtenir les indices de volume correspondants représentatifs de l'évolution réelle compte tenu des variations des prix, il faut diviser les indices de valeur par les indices de prix correspondants. Mais, on n'obtient pas un indice de Paasche de volume puisque l'indice des prix utilisé en France – et dans la plupart des pays étrangers – n'est pas un indice de Laspeyres, mais un indice-chaîne de Laspeyres.

III. Indices-chaînes

A. Raccord d'indices

Les indices ont une durée de vie limitée en raison de l'évolution des structures économiques. Lorsqu'on veut décrire l'évolution d'une grandeur complexe

1. *Déflater :* annuler la hausse due à l'effet de l'inflation.
La *déflation du revenu nominal par l'indice des prix à la consommation* permet de raisonner en revenus constants en évitant l'illusion monétaire, et de comparer les niveaux de vie à des périodes différentes sans tenir compte d'une augmentation du revenu ne compensant que la hausse des prix.

sur une longue période, on est amené à se poser le problème du raccord de deux séries d'indices synthétiques consécutives.

Soit un indice I, base 100 à la date 0, calculé jusqu'à la date t où il a été remplacé par un indice I^*. La valeur de I à une date t' postérieure à la date t s'évalue en multipliant l'indice $I^*_{t'/t}$ par l'indice $I_{t/0}$:

$$I_{t'/0} = I^*_{t'/t} \cdot I_{t/0}$$

Cette formule, obtenu par un raccord d'indice, n'est qu'une approximation, car:

- les indices synthétiques ne possèdent pas la propriété de circularité ;
- il est fréquent que les indices I et I^* n'aient ni le même champ, ni la même composition (changement du nombre d'articles dû à l'introduction de produits nouveaux…).

B. Les indices-chaînes

Pour évaluer l'évolution d'une grandeur complexe sur une longue période, l'emploi de la formule de Laspeyres présente un inconvénient, car la pondération vieillit. Les préférences des consommateurs comme les procédés auxquels recourent les producteurs se modifient : les articles choisis pour représenter l'évolution de certaines catégories de biens cessent d'être bien adaptés à cet objectif et les pondérations de la période de base et de la période courante deviennent trop différentes pour que la comparaison reste valable.

On a donc proposé de calculer des indices dont la base changerait à chaque période.

Mais, comment comparer alors la situation entre deux dates où ont été calculés deux ou plusieurs indices ayant des bases différentes ? On adopte une solution parfaitement empirique : le raccordement entre ces indices intermédiaires.

Les *indices-chaînes* résultent de la généralisation de l'opération de raccord de deux indices. Ce sont des indices définis à partir du produit des indices ayant pour base l'année précédente. L'indice-chaîne de Laspeyres est un produit d'indices de Laspeyres, mais n'est pas un indice de Laspeyres :

$$CL_{t/0} = \prod_{i=1}^{t} L_{i/i-1} \quad \Rightarrow \quad CL_{t/0} = L_{t/t-1} \cdot CL_{t-1/0}$$

On définit de même l'indice-chaîne de Paasche.

L'indice-chaîne permet, mieux que les indices de Laspeyres ou de Paasche, de suivre l'évolution de la grandeur étudiée entre deux dates successives. Si chaque maillon est calculé selon la formule de Laspeyres :

$$\frac{CL_{t/0}}{CL_{t-1/0}} = L_{t/t-1} \qquad \text{alors que :} \qquad \frac{L_{t/0}}{L_{t-1/0}} \neq L_{t/t-1}$$

On est donc dans d'excellentes conditions pour comparer deux périodes successives.

On a la même propriété si chaque maillon est un indice de Paasche. Par contre:

- toute erreur sur l'un des éléments de la chaîne se retrouve dans tous les indices suivants ;
- l'indice obtenu n'a pas une signification bien précise, le résultat dépendant des modifications des pondérations d'une période à l'autre.

Un indice-chaîne sera donc moins bien adapté qu'un indice de Laspeyres ou de Paasche pour étudier les variations survenues depuis la période de base.

C. Indices publiés par l'INSEE

Les principaux indices publiés par l'INSEE [1] sont les suivants :

- indices des prix : prix à la consommation, prix de gros ;
- indices du commerce extérieur ;
- indices de la production industrielle ;
- indices boursiers…

Les indices des prix à la consommation des ménages (IPC) calculés par l'INSEE sont des indices-chaînes de Laspeyres. L'INSEE publie chaque mois plusieurs indices des prix, base 1998. L'indice des ménages urbains dont le chef est ouvrier ou employé (métropole et DOM) sert, dans sa version « hors tabac », à l'indexation du SMIC. Les autres indices concernant l'ensemble des ménages ont un usage *économique* dans leur version « y compris tabac » et un usage *indexation* dans leur version « hors tabac ».

L'indice des prix à la consommation harmonisé (IPCH) sert aux comparaisons internationales.

IV. Traitement statistique des indices

Pour représenter certains phénomènes, on peut être amené à graduer les axes selon des échelles particulières. Le papier semi-logarithmique est particulièrement adapté à certains types de séries chronologiques, et les séries économiques sont souvent des séries d'indices.

1. www.insee.fr/fr/themes

A. Échelle logarithmique

Le papier semi-logarithmique comporte un axe des abscisses à échelle arithmétique et un axe des ordonnées à échelle logarithmique. Sur l'axe des abscisses, on peut choisir l'origine et une unité de longueur quelconque. Mais pour l'axe des ordonnées, on utilise une échelle logarithmique ; la place des nombres est fixée par leur ***logarithme décimal*** (*cf.* figure 2.2) :

Nombre	1	2	3	4	5	6	7	8	9	10
log	0	0,301	0,477	0,602	0,699	0,778	0,845	0,903	0,954	1

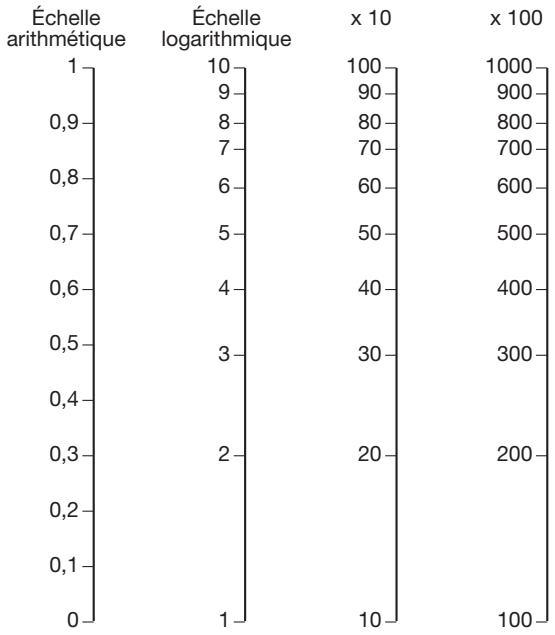


Figure 2.2 – Construction d'une échelle logarithmique

Sur une échelle logarithmique, la distance séparant *deux multiples successifs de dix* est toujours la même puisque :

$\log 10^k - \log 10^{k-1} = \log 10 \qquad \log 10^{k+1} - \log 10^k = \log 10 \qquad \ldots$

L'intervalle entre deux puissances successives de 10 s'appelle un ***module*** et à l'intérieur d'un module, la place des nombres est donc fixée par leur ***logarithme décimal*** (*cf.* figures 2.2 et 2.4).

Les papiers semi-logarithmiques ont habituellement 2, 3 ou 4 modules ; un papier à trois modules permet de représenter des séries temporelles x_t dont le rapport entre la plus grande et la plus petite valeur est au plus de 10^3. Les représentations graphiques des logiciels usuels (Excel ®…) offrent directement la possibilité d'utiliser les échelles logarithmiques.

B. Propriétés d'un graphique à ordonnée logarithmique

Une grandeur dont le taux d'accroissement (ou de diminution) est *constant* sur des laps de temps égaux a son évolution représentée sur du papier à ordonnée logarithmique par une suite de *points alignés* (*cf.* figure 2.3).

En effet, si une grandeur x a un taux de variation annuel i constant, la valeur x_0 de x à la date initiale prend, après t années, la valeur x_t telle que :

$$x_t = x_0\,(1 + i)^t \quad \Rightarrow \quad \log x_t = \log x_0 + t \log(1 + i)$$

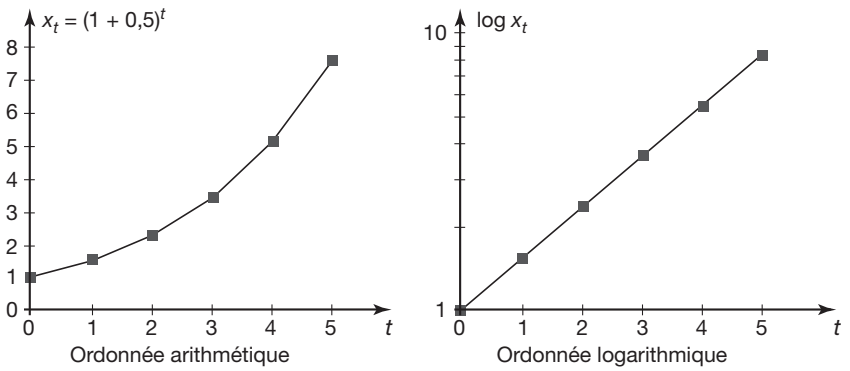


Figure 2.3 – Grandeur à taux de croissance annuel constant

Une représentation avec une ordonnée logarithmique permet :

• la *détermination graphique du rapport* entre deux valeurs de la variable pour en déduire le taux de variation entre les deux dates considérées ; une différence de logarithme représentant un rapport, celui-ci est égal à la différence des ordonnées entre les deux valeurs de la variable ;

• la *détermination graphique du taux moyen* de variation i, la pente de la droite joignant les deux points extrêmes (*cf.* figure 2.4a) étant égale à $(1 + i)$;

• la *comparaison* graphique entre les taux de variation de deux grandeurs représentées sur le même graphique à ordonnée logarithmique ; deux droites parallèles indiquent des taux de variation égaux ;

• la représentation des séries aux *variations importantes* puiqu'avec quatre modules, on peut représenter une série variant de 1 à 10^4.

V. Bilan

Un indice n'est ni parfait, ni rigoureux, ni parfaitement représentatif ; en fait, il existe autant d'indices que le statisticien veut en construire, et chacun a la signification qui résulte de son calcul même. Parmi tous ces indices, l'économiste choisira celui qui lui paraît le mieux correspondre à l'usage qu'il veut en faire.

Pour construire un indice synthétique, on est amené à faire quatre choix :

• *deux choix d'ordre économique* :

– choix des grandeurs entrant dans la composition de l'indice,

– choix de la période de référence ;

• *deux choix d'ordre statistique* :

– choix de la moyenne à utiliser pour le calcul de l'indice à partir des grandeurs composantes,

– choix de la pondération à appliquer aux valeurs des grandeurs afin de tenir compte de leur importance relative.

Testez-vous *(les réponses sont données page 284)*

Il y a *au moins* une réponse exacte par question.

1. L'indice de Laspeyres des prix est :

a) un indice des dépenses

b) une moyenne arithmétique d'indices élémentaires

c) l'indice des prix actuellement calculé par l'INSEE

d) s'exprime dans une unité monétaire

2. L'indice de Paasche est :

a) n'est pas un indice des prix

b) transitif

c) une moyenne harmonique d'indices élémentaires

d) au plus égal à l'indice de Laspeyres

3. Un indice des dépenses est :

a) un indice de valeur

b) réversible

c) transitif

d) un indice de prix

4. Une grandeur mesurée tous les ans :

a) est représentée sur un papier semi-logarithmique par une suite de points alignés si le taux annuel de variation est constant

b) a un taux annuel moyen de variation qui peut être déterminé graphiquement

c) a un taux annuel moyen de variation égal à la moyenne arithmétique des taux annuels de variation

d) est représentée sur un papier à échelles arithmétiques par une suite de points alignés si l'accroissement annuel est constant

5. Les taux annuels moyens de croissance du *PIB* en volume en France ont été les suivants de 1997 à 2002 (*source* : *Tableaux de l'Économie française*, INSEE) :

Année	1997	1998	1999	2000	2001	2002
Taux annuel moyen de croissance (en %)	2,0	3,2	2,9	3,8	2,1	1,2

a) le taux de croissance sur les cinq années est la somme des cinq taux de croissance

b) le taux annuel moyen de croissance sur la période 1997 – 2002 est égal à la moyenne arithmétique des taux annuels moyens de croissance

c) le taux annuel moyen de croissance sur la période 1997 – 2002 se calcule à l'aide d'une moyenne géométrique

d) pour la période 1999 – 2001, le taux de croissance du *PIB* en volume a été de 9 %

Exercices *(corrigés page 297)*

Exercice 2.1

Une entreprise utilise pour ses fabrications trois types de matières premières qui sont notées respectivement *A*, *B* et *C*.

En 2000 et 2004, les prix observés et les quantités achetées par cette entreprise ont été les suivants :

Matières premières	Prix par tonne en euros 2000	Quantités achetées en tonnes en 2000	Prix par tonne en euros 2004	Quantités achetées en tonnes en 2004
A	800	10	900	6
B	500	4	700	4
C	600	5	600	8

1. Calculez les indices élémentaires rendant compte de l'évolution des prix de chacune des matières premières entre 2000 et 2004.

2. Calculez la moyenne arithmétique des indices élémentaires précédents pondérée par la part des dépenses engagées par l'entreprise pour chacune de ces matières premières en 2000. De quel indice s'agit-il ?

3. Effectuez le même calcul pour rendre compte de l'évolution des quantités entre 2000 et 2004.

4. Calculez l'indice mesurant l'évolution globale des dépenses de matières premières entre 2000 et 2004.

5. Déterminez, en utilisant les résultats des questions précédentes, les taux de variation (exprimés en pourcentage) des prix, des quantités et de la dépense totale. Comment s'explique l'évolution de la dépense totale ?

Exercice 2.2

Entre 1980 et 2000, les quantités de sel extraites d'une mine ont été multipliées par 1,5 entre 1980 et 1985, sont passées de l'indice 130 en 1985 à l'indice 168 en 1992 avant d'augmenter de 6 % par an entre 1992 et 2000.

1. Quel est le taux annuel moyen de variation des quantités de sel extraites entre 1980 et 2000 ?

2. Au cours de la même période, le taux de variation annuel moyen du prix du sel a été de – 5 %. Quelle est la valeur de l'indice du chiffre d'affaire en 2000, base 1980 ?

Exercice 2.3

Ce tableau donne les indices trimestriels des salaires horaires de base de l'ensemble des ouvriers (secteurs non agricoles), base 100 au 4e trimestre 2008. La série est rétropolée depuis le 4e trimestre 1998 (*Source :* INSEE) :

	1999	2000	2001	2002	2003
31 mars	72,3	76,1	79,4	82,5	84,8
30 juin	72,8	76,8	80	83	85,3
30 septembre	73,7	77,6	80,8	83,7	86,1
31 décembre	74,4	78,1	81,3	84,1	86,4

	2004	2005	2006	2007	2008	2009
31 mars	87,1	89,7	92,5	95,2	97,9	100,8
30 juin	87,6	90,2	93,1	95,8	99	101,2
30 septembre	88,6	91,4	94	96,6	99,7	101,7
31 décembre	88,9	91,8	94,3	97	100	

Sachant que cet indice vaut 71,9 au 31 décembre 1998, calculez le taux trimestriel moyen de croissance entre le 31 décembre 1998 et le 30 septembre 2009, et le taux annuel moyen de croissance entre le 31 décembre 1998 et le 31 décembre 2008.

Exercice 2.4

Le tableau suivant est un extrait du tableau « Production et valeur ajoutée de l'agriculture » :

	2008	2008/2007 en %		
	En Mds d'euros	Volume	Prix	Valeur
Produits végétaux	38,2	3,6	?	– 0,3
Céréales	10,7	19,2	– 21,3	– 6,2
Oléagineux, protéagineux	2,4	4,8	?	3,2
Betteraves industrielles	0,8	– 7,2	– 3,4	– 10,3
Autres plantes industrielles*	0,3	– 2,9	13,5	10,3
Fruits, légumes, pommes de terre	7,4	– 3,1	6,3	3,0
Vins	9,4	?	3,7	– 2,1
Fourrages, plantes, fleurs	7,4	– 1,2	9,7	8,4

* Tabac, lin textile, houblon, canne à sucre, etc.

Source : Tableaux de l'Économie française, édition 2010.

1. Donnez l'indice de valeur de la production des « Produits végétaux » en 2008, base 100 en 2007. Même question pour « Oléagineux, protéagineux » et pour « Vins ».
2. Calculez l'évolution 2008/2007 (en pourcentage) des prix à la production des « Produits végétaux ». Même question pour « Oléagineux, protéagineux ».
3. Calculez l'évolution 2008/2007 (en pourcentage) du volume de la production des « Vins ».
4. Commentez les résultats obtenus.

Exercice 2.5

Considérons la consommation médicale totale en France (en milliards d'euros courants) de 1970 à 2000 (*Source* : *Tableaux de l'Économie française* , INSEE).

Année	CM (milliards d'euros)	Année	CM (milliards d'euros)	Année	CM (milliards d'euros)
1970	6,494				
1971	7,516	1981	35,399	1991	87,430
1972	8,568	1982	41,146	1992	93,482
1973	9,833	1983	46,848	1993	98,665
1974	11,586	1984	52,000	1994	101,866
1975	14,452	1985	57,046	1995	106,257
1976	16,815	1986	61,711	1996	109,245
1977	18,812	1987	64,776	1997	111,059
1978	22,547	1988	70,447	1998	112,731
1979	26,084	1989	76,377	1999	117,093
1980	30,215	1990	81,911	2000	123,545

1. Calculez la variation relative (en %) de la consommation médicale entre 1970 et 2000.
2. Calculez la série des indices de la consommation médicale base 1970.
3. Représentez la série des indices sur un graphique à ordonnée logarithmique, et calculez le taux annuel de croissance de cet indice pendant la période 1970-1982.
4. Représentez la série des indices sur un graphique à ordonnée arithmétique, et calculez l'augmentation annuelle moyenne entre 1982 et 2000.
5. Étude de l'évolution de l'indice en volume :
   - sachant que l'indice des prix $I_{82/70}$ est égal à 318,7, calculez la variation de l'indice en volume entre 1970 et 1982, et en déduire le taux annuel moyen de variation de cet indice entre 1970 et 1982 ;

– sachant qu'on utilise le coefficient de raccordement [1] de 5,584 pour calculer un prix en 2000 à partir d'un prix en 1970, calculez le taux annuel moyen de croissance de l'indice en volume entre 1982 et 2000.

6. Conclusion.

1. http://www.insee.fr/fr/indicateur/achatfranc.htm

3. Distributions statistiques à deux caractères

Lorsque les observations portent simultanément sur deux caractères, et lorsqu'elles sont trop nombreuses pour qu'on les cite une à une, on les présente sous la forme d'un *tableau à double entrée*. On définit alors la distribution conjointe, les distributions marginales et les distributions conditionnelles. L'étude de la distribution de deux variables se pousuit par celle de leur *liaison*.
L'étude de la liaison entre les variables observées, appelée communément l'étude des corrélations, dépend de leur nature. On envisagera les trois cas suivants : *deux variables quantitatives, une variable quantitative et une variable qualitative, deux variables qualitatives* . Lorsque le domaine de variation d'une variable quantitative a été découpé en classes et que les observations sont présentées dans un tableau à double entrée, alors cette variable peut être traitée comme une variable qualitative et dans ce cas, on a plusieurs méthodes pour l'étude de la liaison.

I. Distributions statistiques à deux variables

A. Distribution conjointe

Désignons par X et Y les deux variables qui peuvent être qualitatives ou quantitatives, et qui peuvent ne pas être de même nature. Les k modalités de X sont désignées par $x_1, \ldots, x_i, \ldots, x_k$; les l modalités de Y sont désignées par $y_1, \ldots, y_j, \ldots, y_l$. La i^e modalité d'une variable désigne le centre de la i^e classe dans le cas d'une variable quantitative continue.

La répartition des n observations, ou *distribution conjointe*, suivant les modalités de X et Y se présente sous forme d'un tableau à double entrée, appelée ***tableau de contingence*** (*cf.* tableaux 3.1 et 3.2).

Tableau 3.1 – Tableau de contingence : distribution conjointe de deux variables X et Y

Modalité de Y / Modalité de X	y_1	…	y_j	…	y_l	Total
x_1	n_{11}	…	n_{1j}	…	n_{1l}	$n_{1\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	n_{i1}	…	n_{ij}	…	n_{il}	$n_{i\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_k	n_{k1}	…	n_{kj}	…	n_{kl}	$n_{k\bullet}$
Total	$n_{\bullet 1}$	…	$n_{\bullet j}$	…	$n_{\bullet l}$	n

Tableau 3.2 – Exemple de tableau de contingence : distribution des notes de 100 étudiants à une épreuve d'un concours selon leur filière d'origine

Classe de notes Y / Filière d'origine X	[0 ; 6[3	[6 ; 10[8	[10 ; 14[12	[14 ; 20] 17	Total
Filière *A*	26	6	4	1	37
Filière *B*	12	9	3	1	25
Filière *C*	1	4	5	6	16
Filière *D*	10	8	3	1	22
Total	49	27	15	9	100

L'effectif n_{ij} désigne le nombre de fois où la modalité x_i de la variable X et la modalité y_j de la variable Y ont été observées simultanément.

L'effectif $n_{i\bullet}$ est le *nombre total d'observations de la modalité* x_i *de X, quelle que soit la modalité de Y* :

$$n_{i\bullet} = \sum_{j=1}^{l} n_{ij}$$

De même, l'effectif $n_{\bullet j}$ est le *nombre total d'observations de la modalité* y_j *de Y, quelle que soit la modalité de X* :

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$$

On a évidemment : $\sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j} = n$

La distribution conjointe peut aussi être définie par les fréquences :

$$f_{ij} = \frac{n_{ij}}{n}$$

B. Distributions marginales

Les k couples $(x_i , n_{i\bullet})$ forment la *distribution marginale* de la *variable X*.
Les l couples $(y_j , n_{\bullet j})$ forment la *distribution marginale* de la *variable Y*.

Les distributions marginales peuvent aussi être données sous forme de fréquences :

$$f_{i\bullet} = \frac{n_{i\bullet}}{n} \quad \text{et} \quad f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

Disposant d'une distribution conjointe, on peut déduire les distributions marginales qui permettent d'étudier séparément chaque variable en représentant graphiquement sa distribution et s'il s'agit d'une variable quantitative, en calculant ses caractéristiques de tendance centrale, de dispersion, de forme…

C. Distributions conditionnelles

La distribution de la variable Y, la variable X étant égale à x_i, est appelée *distribution conditionnelle de Y pour* $X = x_i$:

$Y/X = x_i$	y_1	…	y_j	…	y_l	Total
Effectif	n_{i1}	…	n_{ij}	…	n_{il}	$n_{i\bullet}$

Cette distribution des $n_{i\bullet}$ observations, satisfaisant à la condition $X = x_i$, est présentée sous la forme de fréquences conditionnelles :

$$f_{j/i} = \frac{n_{ij}}{n_{i\bullet}} \quad \text{avec :} \quad \sum_{j=1}^{l} f_{j/i} = 1$$

$Y/X = x_i$	y_1	…	y_j	…	y_l	Total
Fréquence	$f_{1/i}$	…	$f_{j/i}$	…	$f_{l/i}$	1

La fréquence [1] $f_{j/i}$ se lit « f indice j si i », c'est-à-dire *fréquence* de y_j si $X = x_i$. Il y a k distributions conditionnelles de Y pour $X = x_i$ $(i = 1,\ldots, k)$.

Lorsque la variable Y est quantitative, on peut calculer pour chaque valeur x_i sa *moyenne conditionnelle* $\overline{y_i}$ et son *écart-type conditionnel* s_i :

$$\overline{y_i} = \sum_{j=1}^{l} f_{j/i}\, y_j \qquad \text{et} \qquad s_i^2 = \sum_{j=1}^{l} f_{j/i}\, ((y_j - \overline{y_i})^2$$

Les k modalités de X induisant une partition des observations en k sous-groupes, la moyenne $\overline{y}$ peut s'exprimer comme somme pondérée des k moyennes $\overline{y_i}$ (chapitre 1) :

$$\overline{y} = \sum_{i=1}^{k} f_{i\bullet}\, \overline{y_i}$$

Symétriquement, on a l distributions conditionnelles de X et on définit les fréquences conditionnelles *f indice i si j* :

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}} \qquad \text{avec :} \qquad \sum_{i=1}^{k} f_{i/j} = 1$$

$X/Y = y_j$	x_1	…	x_i	…	x_k	Total
Fréquence	$f_{1/j}$	…	$f_{i/j}$	…	$f_{k/j}$	1

Lorsque la variable X est quantitative, on peut calculer pour chaque valeur y_j sa moyenne conditionnelle $\overline{x_j}$ et son écart-type conditionnel s_j :

$$\overline{x_j} = \sum_{i=1}^{k} f_{i/j}\, x_i \qquad \text{et} \qquad s_j^2 = \sum_{i=1}^{k} f_{i/j}\, (x_i - \overline{x_j})^2$$

et on a la relation suivante entre la moyenne $\overline{x}$ et les l moyennes conditionnelles $\overline{x_j}$:

$$\overline{x} = \sum_{j=1}^{l} f_{\bullet j}\, \overline{x_j}$$

Lorsqu'on dispose d'observations portant simultanément sur deux variables, il est fréquent de les présenter dans un tableau donnant l'ensemble des distributions conditionnelles de Y, et on a alors un tableau dont toutes les

1. Les fréquences $f_{j/i}$ sont aussi parfois notées f_j^i

sommes en ligne sont égales à 100 % ; ce tableau est appelé *tableau des profils en ligne* (*cf.* tableau 3.3).

Tableau 3.3 –Tableau des profils en ligne correspondant au tableau de contingence 3.2

Classe de notes Y / Filière d'origine X	[0 ; 6[3	[6 ; 10[8	[10 ; 14[12	[14 ; 20] 17	Total
Filière *A*	70,3	16,2	10,8	2,7	100
Filière *B*	48,0	36,0	12,0	4,0	100
Filière *C*	6,3	25,0	31,2	37,5	100
Filière *D*	45,5	36,4	13,6	4,5	100
Distribution marginale de *Y*	49,0	27,0	15,0	9,0	100

Bien évidemment, on définit d'une façon symétrique le *tableau des profils en colonne* qui est le tableau des distributions conditionnelles de X avec des sommes en colonne égales à 1 (*cf.* tableau 3.4).

Tableau 3.4 –Tableau des profils en colonne

Modalité de Y / Modalité de X	y_1	…	y_j	…	y_l	Distribution marginale de X
x_1	$f_{1/1}$	…	$f_{1/j}$	…	$f_{1/l}$	$f_{1\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	$f_{i/1}$	…	$f_{i/j}$	…	$f_{i/l}$	$f_{i\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_k	$f_{k/1}$	…	$f_{k/j}$	…	$f_{k/l}$	$f_{k\bullet}$
Total	1		1		1	1

D. Dépendance et indépendance statistique

Si tous les profils en colonne du tableau 3.4 sont identiques, cela signifie que la distribution de la variable X ne dépend pas de la variable Y, on dit alors que les variables X et Y sont *statistiquement indépendantes* dans l'ensemble des n individus considérés, et dans ce cas toutes les distributions conditionnelles de X sont identiques à la distribution marginale de X.

On peut écrire en termes d'effectifs ou de fréquences ce que signifie l'indépendance statistique entre X et Y ; pour tout couple (i, j) :

$$f_{i/j} = f_{i\bullet} \Leftrightarrow \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet} \Leftrightarrow f_{ij} = f_{i\bullet} \cdot f_{\bullet j} \Leftrightarrow n_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

Par raison de symétrie, l'indépendance statistique entre X et Y implique aussi des profils en ligne identiques à la distribution marginale de Y : $f_{j/i} = f_{\bullet j}$ pour tout couple (i , j).

Lorsque deux variables dépendent statistiquement l'une de l'autre, on cherche à évaluer l'intensité de leur liaison et dans le cas de deux variables quantitatives, on examine si on peut les considérer liées par une relation linéaire.

II. Deux variables quantitatives

Si les observations de deux variables statistiques X et Y sont connues individuellement, on commence par les visualiser en les représentant sous la forme d'un *nuage de points* (*cf.* figure 3.1) : dans un repère cartésien, chaque observation (x_i , y_i) est figurée par le point M_i de coordonnées (x_i , y_i), et la forme du nuage donne une information sur le type d'une éventuelle liaison.

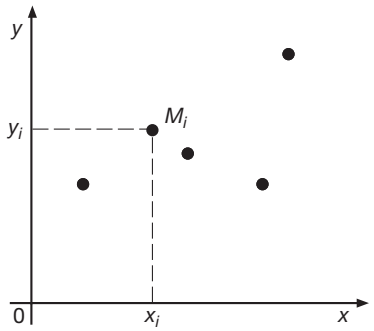


Figure 3.1 – Nuage de points

Supposons que l'examen du nuage de points conduise à rechercher une droite d'ajustement. Le calcul des coefficients de cette droite va être exposé dans le cas où les observations sont connues *individuellement*. La généralisation des résultats au cas d'une distribution résumée dans un tableau de contingence se fait sans difficulté.

A. Caractéristiques d'un couple de deux variables quantitatives

1) Moyenne d'une somme de deux variables statistiques

On montre sans difficulté le résultat suivant : $\overline{x+y} = \bar{x} + \bar{y}$

$$\Rightarrow \quad \forall a, b, c \in \mathbb{R} \quad \overline{ax+by+c} = a\bar{x} + b\bar{y} + c$$

2) Covariance entre deux variables statistiques

Cas de *données individuelles* :

$$\text{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}$$

Cas de *données groupées* dans un tableau de contingence (covariance pondérée) :

$$\text{cov}(X,Y) = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij}\, x_i y_i - \bar{x}\cdot\bar{y}$$

Propriétés de la covariance

1. $\text{cov}(X,Y) = \text{cov}(Y, X)$
2. $\text{cov}(X, X) = \text{var}(X)$
3. $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\,\text{cov}(X, Y)$
4. $\forall a, b, c, x_0, y_0 \in \mathbb{R} : \text{cov}(aX + x_0, bY + y_0) = ab\ \text{cov}(X,Y)$
$\Rightarrow \text{var}(aX + bY + c) = a^2\,\text{var}(X) + b^2\,\text{var}(Y) + 2ab\ \text{cov}(X,Y)$
5. $|\text{cov}(X, Y)| \leq \sqrt{\text{var}(X)\cdot\text{var}(Y)}$

Les propriétés 1 et 2 sont évidentes. Montrons la propriété 3 dans le cas de données individuelles, la démonstration pour des données groupées dans un tableau de contingence se faisant de la même façon en utilisant les formules pondérées par les fréquences :

$$\begin{aligned}\text{var}(X+Y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i + y_i - \overline{x+y})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i + y_i - \bar{x} + \bar{y})^2 \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right) \\ &= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)\end{aligned}$$

La propriété 4 se démontre sans difficulté si on se souvient que $\overline{ax + x_0} = a\bar{x} + x_0$. Quant à la propriété 5, elle sera démontrée au § II.C.1.

3) Coefficient de corrélation linéaire

On appelle *coefficient de corrélation linéaire* entre deux variables statistiques X et Y, le rapport de leur covariance par le produit de leurs écarts-types :

$$r(X, Y) = \frac{\text{cov}(X, Y)}{s_X \cdot s_Y}$$

Propriétés du coefficient de corrélation linéaire

On a pour tout $a, b, x_0, y_0 \in \mathbb{R}$:

$$r(aX + x_0\ ,\ bY + y_0) = \frac{\text{cov}(aX + x_0\ ,\ bY + y_0)}{s_{aX + x_0} \cdot s_{bY + y_0}} = \frac{ab\,\text{cov}(X, Y)}{|ab| s_X \cdot s_Y}$$

$$= \begin{cases} + r(X, Y) & \text{si } a \text{ et } b \text{ de même signe} \\ - r(X, Y) & \text{si } a \text{ et } b \text{ de signe opposé} \end{cases}$$

Ce coefficient, *invariant par changement d'origine et d'échelle*, est un nombre *sans dimension* qui, d'après la propriété 5 de la covariance, varie entre – 1 et + 1. On montrera que s'il est égal à ± 1, les n points $(x_i\ ,\ y_i)$ sont alignés.

B. Ajustement linéaire d'un nuage de points

Les points $(x_i\ ,\ y_i)$ forment un *nuage* dont on cherche une *approximation* dans un but de *simplification*. Mais qui dit simplification dit *déformation* : nous voudrions qu'elle soit minimale ; encore faut-il préciser ce que l'on entend par là. Disons tout de suite que le choix du critère sera *arbitraire* même si l'on tente de le justifier par des considérations plus ou moins « intuitives ». On peut vouloir par exemple :

- préserver *au mieux* les distances entre points ;
- préserver *au mieux* les angles des droites joignant les points…

Il n'existe pas de moyen de satisfaire à toutes ces exigences à la fois. Il nous faut donc choisir.

Nous allons chercher la ***meilleure droite au sens des moindres carrés***, c'est-à-dire telle que : $\sum_{i=1}^{n} |M_i H_i|^2$ soit minimum (*cf.* figure 3.2) :

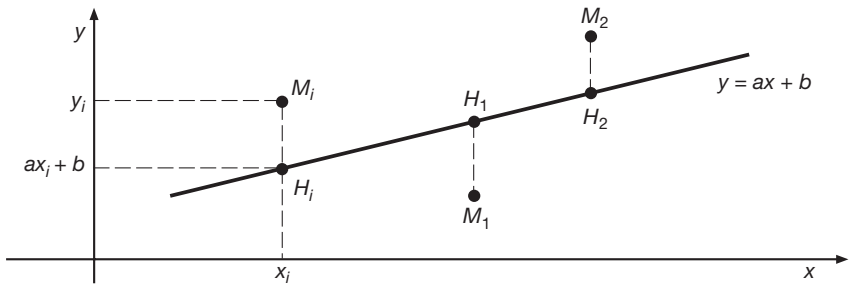


Figure 3.2 – Interprétation géométrique de la droite des moindres carrés

Les *distances* sont comptées *parallèlement* à l'un des axes des coordonnées ; nous avons choisi ici l'axe des ordonnées (*cf.* figure 3.2).

Il s'agit de déterminer la droite $\mathbb{D}$ d'équation $\{y = ax + b\}$ telle que :

$$F(a, b) = \sum_{i=1}^{n} \Big(y_i - (ax_i + b) \Big)^2 \text{ soit minimum}$$

Nos *inconnues* sont a et b.

Commençons par chercher le minimum de $F(a, b)$ relativement à b lorsque a est fixé. On peut écrire $F(a, b)$ comme un trinôme du second degré en b :

$$F(a, b) = \sum_{i=1}^{n} \Big((y_i - ax_i) - b \Big)^2 = \sum_{i=1}^{n} \Big((y_i - ax_i)^2 - 2b(y_i - ax_i) + b^2 \Big)$$

$$= \sum_{i=1}^{n} (y_i - ax_i)^2 - 2b \sum_{i=1}^{n} (y_i - ax_i) + nb^2$$

Quand a est fixé, le dernier membre constitue une fonction de b qui atteint son minimum pour $b = \hat{b}$ tel que $\dfrac{\partial F}{\partial b}(a, \hat{b}) = 0$, soit :

$$\frac{\partial F}{\partial b}(a, \hat{b}) = -2 \left(\sum_{i=1}^{n} (y_i - ax_i) - n\hat{b} \right) = 0$$

$$\Rightarrow \hat{b} = \frac{1}{n} \sum_{i=1}^{n} (y_i - ax_i) = \overline{y} - a\overline{x}$$

• 1re *conséquence* : la droite des moindres carrés passe par le point de coordonnées $(\overline{x}, \overline{y})$ qu'on appelle parfois *le centre de gravité* ou *point moyen du nuage.*

Notre problème est maintenant de trouver le minimum de $F(a, \hat{b})$ relativement à a :

$$F(a, \hat{b}) = \sum_{i=1}^{n} \Big((y_i - \overline{y}) - a(x_i - \overline{x}) \Big)^2$$

$$= \sum_{i=1}^{n} (y_i - \overline{y})^2 - 2a \sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x}) + a^2 \sum_{i=1}^{n} (x_i - \overline{x})^2$$

ce qui peut encore s'écrire :

$$F(a, \hat{b}) = n\Big(a^2 \operatorname{var}(X) - 2a \operatorname{cov}(X, Y) + \operatorname{var}(Y) \Big)$$

Le coefficient de a^2 étant positif ou nul, ce trinôme du second degré en a atteint son *minimum* relativement à a pour $a = \hat{a}$ avec :

$$\hat{a} = \frac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)}$$

Ainsi le couple $(\hat{a}, \hat{b})$ avec $\hat{b} = \overline{y} - \hat{a}\overline{x}$ réalise le minimum de la fonction F

• 2e conséquence : la droite des moindres carrés a pour équation $y = \hat{a}x + b$ soit

$$y - \overline{y} = \frac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)} \cdot (x - \overline{x})$$

On posera pour tout i variant de 1 à n : $\hat{y}_i = \hat{a}x_i + \hat{b}$, $\hat{y}_i$ est la *valeur estimée* de Y par la droite des moindres carrés lorsque $X = x_i$

C. Interprétation du coefficient de corrélation linéaire

1) Interprétation à l'aide de la droite 𝔻

Il est toujours possible de tracer la droite des moindres carrés précédente quelle que soit la forme du nuage. L'approximation du nuage par cette droite est-elle *légitime* ? Quel sens, quelle signification donner à cette droite ?

C'est là une autre question, et fort importante. On pourra dire qu'il est d'autant plus légitime de remplacer le nuage par la droite trouvée que la dispersion du nuage de points par rapport à la droite des moindres carrés :

$$\sum_{i=1}^{n} |M_i H_i|^2 = F(a, \hat{b}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \text{ sera plus faible}$$

En remplaçant a par son estimation $\hat{a}$, on obtient :

$$F(a, \hat{b}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = n\left(\frac{(\text{cov}(X, Y))^2}{\text{var}(X)} - 2\frac{(\text{cov}(X, Y))^2}{\text{var}(X)} + \text{var}(Y)\right)$$

$$= n\left(\text{var}(Y) - \frac{(\text{cov}(X, Y))^2}{\text{var}(X)}\right)$$

et comme : $$r^2 = \frac{(\text{cov}(X, Y))^2}{\text{var}(X) \cdot \text{var}(Y)}$$

on a :

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = n \text{ var}(Y) \cdot (1 - r^2) \Leftrightarrow \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2 (1 - r^2)$$

ce qui implique :

$$1 - r^2 \geq 0 \quad \Rightarrow \quad |r| \leq +1 \quad \text{et} \quad |\text{cov}(X, Y)| \leq \sqrt{\text{var}(X) \cdot \text{var}(Y)}$$

La quantité $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, appelée *Somme des Carrés Résiduelle* ($SC_{\text{rés}}$), est d'autant plus faible que r^2 est proche de 1.

Elle est nulle pour $r = \pm 1$ et dans ce cas, on a *une liaison linéaire entre X et Y*, car si $\{ \hat{y}_i = y_i \text{ pour tout } i\}$, alors les n points (x_i, y_i) sont alignés.

La quantité $\sum_{i=1}^{n} (y_i - \overline{y})^2$ étant appelée *Somme des Carrés Totale* (SC_{tot}) *de Y,* il s'ensuit :

$$1 - r^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \frac{SC_{\text{rés}}}{SC_{\text{tot}}}$$

$\Rightarrow$ la quantité $\{1 - r^2\}$ est égale à la *proportion de variation* de *Y non expliquée* par la droite des moindres carrés $\mathbb{D}$ (*cf.* figures 3.3 et 3.4).

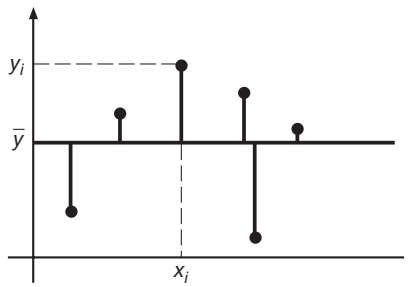


Figure 3.3 – $\sum_{i=1}^{n} (y_i - \bar{y})^2 = SC_{\text{tot}}$

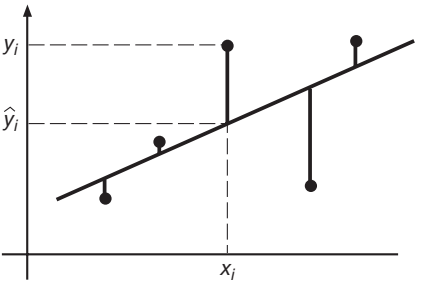


Figure 3.4 – $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SC_{\text{rés}}$

Notons que la somme des écarts à la droite $\mathbb{D}$ est nulle :

$$y = \hat{a}x + \hat{b} \quad \Rightarrow \quad \sum_{i=1}^{n} (y_i - \hat{a}x_i - \hat{b}) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

ce qui implique aussi que *les moyennes des* $\hat{y}_i$ *et des* y_i *sont égales* : $\bar{\hat{y}} = \bar{y}$ et ceci est dû au fait que la droite des moindres carrés passe par le point moyen $(\bar{x}, \bar{y})$ du nuage des n points.

La décomposition de la *variation totale de Y* permet une autre interprétation de r^2 :

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Montrons que le 3 e terme du dernier membre est nul. On peut écrire :

$$\hat{y}_i - \overline{y} = \hat{a}(x_i - \overline{x}) \quad \text{et} \quad y_i - \hat{y}_i = y_i - \overline{y} - (\hat{y}_i - \overline{y}) = y_i - \overline{y} - \hat{a}(x_i - \overline{x})$$

ce qui donne une nouvelle expression de ce 3 e terme :

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \hat{a} \sum_{i=1}^{n} \Big(y_i - \overline{y} - \hat{a}(x_i - \overline{x}) \Big)(x_i - \overline{x})$$

$$= \hat{a} \left(\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x}) - \hat{a} \sum_{i=1}^{n} (x_i - \overline{x})^2 \right)$$

$$\Rightarrow \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = n\hat{a} \Big(\operatorname{cov}(X, Y) - \hat{a} \operatorname{var}(X) \Big) = 0$$

puisque $\hat{a} = \dfrac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)}$

La quantité $\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$ étant appelée *Somme des Carrés Expliquée* (SC_{expl}), on obtient l'***équation de la décomposition de la variation totale de Y*** :

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Leftrightarrow SC_{\text{tot}} = SC_{\text{expl}} + SC_{\text{rés}}$$

et une *autre interprétation* de r^2, complémentaire à celle de $(1 - r^2)$:

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \frac{SC_{\text{expl}}}{SC_{\text{tot}}}$$

$\Rightarrow$ le carré r^2 du coefficient de corrélation linéaire est égal à la *proportion de la variation* de *Y expliquée* par la droite des moindres carrés

Conclusion sur l'interprétation de la valeur du coefficient de corrélation linéaire :

$$r = \pm 1 \quad \Leftrightarrow \quad y_i = \hat{y}_i = \hat{a}x_i + \hat{b} \quad \forall i$$

$r = \pm 1$	$\Leftrightarrow$	les n points (x_i, y_i) sont alignés
$r = 0$	$\Leftrightarrow$	pas de liaison linéaire, mais possibilité d'une liaison d'un autre type

Voici un exemple de deux variables X et Y non indépendantes avec $r(X, Y) = 0$:

X	-2	-1	0	1	2
Y	4	1	0	1	4

$$n = 5,\ \bar{x} = 0,\ \bar{y} = 2 \text{ et } \sum_{i=1}^{n} x_i y_i = 0 \Rightarrow r(X, Y) = 0 \text{ et } Y = X^2$$

$\Rightarrow$ *Le coefficient de corrélation linéaire entre deux variables quantitatives indépendantes est nul, mais la réciproque n'est pas vraie* :

$$X \text{ et } Y \text{ indépendantes} \quad \underset{\not\Leftarrow}{\Rightarrow} \quad r(X, Y) = 0$$

2) Droite des moindres carrés $\mathbb{D}'$

Dans toute l'étude précédente, on a fait jouer des rôles non symétriques à X et à Y. On a procédé comme si la variable X pouvait être mesurée, et qu'on cherchait à *prévoir* la variable Y.

Inversement, la droite $\mathbb{D}'$ des moindres carrés pour laquelle les distances sont comptées parallèlement à l'axe des abscisses (*cf.* figure 3.5) a pour équation :

$$x - \bar{x} = \frac{\text{cov}(X, Y)}{\text{var}(Y)} \cdot (y - \bar{y}) \qquad \Rightarrow \qquad y - \bar{y} = \frac{\text{var}(Y)}{\text{cov}(X, Y)} \cdot (x - \bar{x})$$

Mais, dans certains cas, comme celui où la variable X désigne le temps, seule la droite $\mathbb{D}$ a un sens.

Le coefficient r étant symétrique par rapport à X et à Y, la *Somme des Carrés Résiduelle* associée à la droite $\mathbb{D}'$ est égale à :

$$\sum_{i=1}^{n} |M_i G_i|^2 = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 = n \text{ var } (X) \cdot (1 - r^2)$$

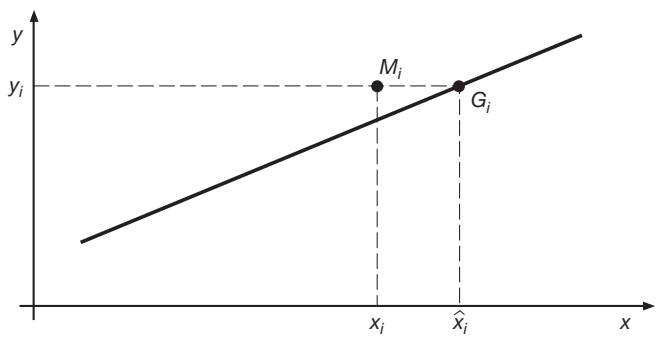


Figure 3.5 – Interprétation géométrique de la droite des moindres carrés $\mathbb{D}'$

D. Comparaison des deux droites des moindres carrés

Les deux droites $\mathbb{D}$ et $\mathbb{D}'$ sont généralement distinctes. Elles se *coupent au point moyen* du nuage, et leurs coefficients directeurs sont de même signe et du signe de r :

$$\frac{\text{cov}(X, Y)}{\text{var}(X)} = r\sqrt{\frac{\text{var}(Y)}{\text{var}(X)}} \quad \text{et} \quad \frac{\text{var}(Y)}{\text{cov}(X, Y)} = \frac{1}{r}\sqrt{\frac{\text{var}(Y)}{\text{var}(X)}}$$

De plus, la valeur absolue du coefficient de corrélation r étant comprise entre 0 et 1, la valeur absolue de la pente de la droite $\mathbb{D}$ est toujours inférieure ou égale à celle de la droite $\mathbb{D}'$ (*cf.* figure 3.6).

Ces deux droites seront confondues si et seulement si les variables X et Y sont liées par une relation linéaire :

$$r = 1/r \quad \Rightarrow \quad r = \pm 1$$

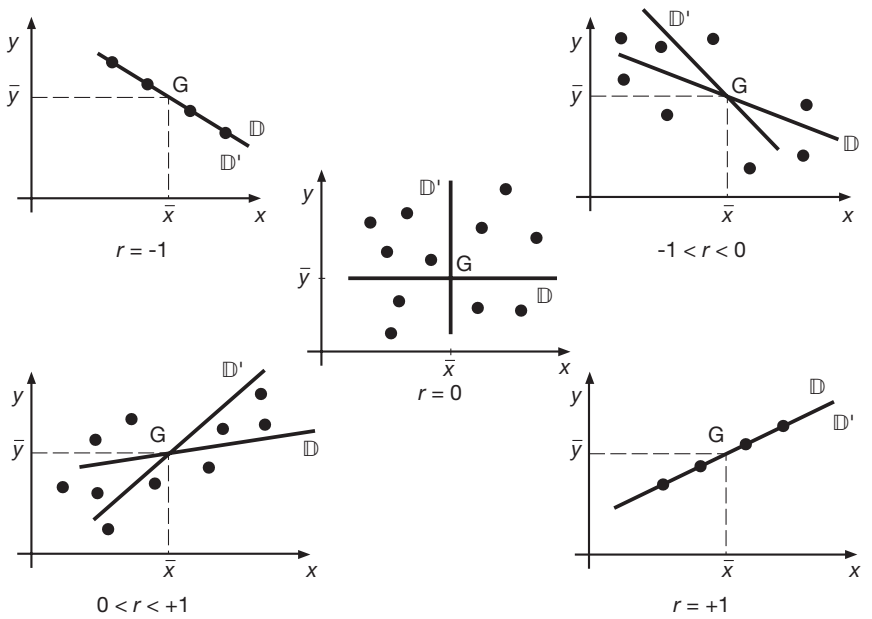


Figure 3.6 – Positions respectives des droites des moindres carrés selon les valeurs de r

E. Le coefficient *r* et la qualité de l'ajustement linéaire

Comment juger la qualité de l'ajustement linéaire ? Il est clair que si le coefficient *r* est *voisin de 0*, il faut *rejeter* l'ajustement linéaire, mais pour quelles valeurs de *r*, le considère-t-on de *bonne* qualité ? C'est une question importante, et beaucoup d'exemples montrent qu'on ne peut pas établir de règles de décision à partir du seul examen de la valeur de *r*. Moyennant certaines hypothèses dont il ne faut pas oublier d'examiner la validité, la formalisation du modèle linéaire (qui dépasse le cadre de ce livre) répond partiellement à la question.

Un résumé numérique est insuffisant pour rendre compte de la pertinence d'une liaison linéaire. Pour s'en convaincre, on se reportera aux résultats de F. J. Anscombe (*cf.* figure 3.7) : pour quatre séries de 11 observations simultanées de deux variables *X* et *Y*, on obtient la même valeur du coefficient de corrélation linéaire { $r = 0{,}82$} et la même droite des moindres carrés {$y = 3 + 0{,}5x$}, mais l'examen graphique montre que l'ajustement linéaire n'est adapté qu'au premier cas.

I		II		III		IV	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	8,0	12,50
12,0	10,84	12,0	9,13	12,0	8,15	19,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89

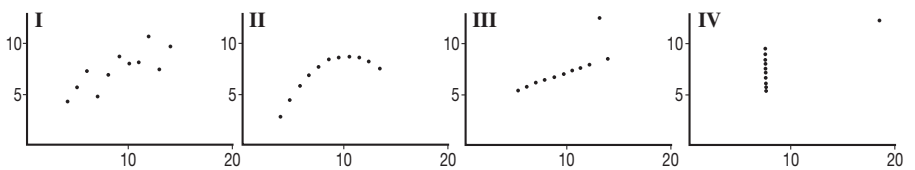


Figure 3.7 – Extrait de F. J. Anscombe : Graphs in Statistical Analysis*, adapté avec la permission de The American Statistician, 27 (February 1973), 17-21, American Statistician Association*

L'ajustement linéaire de la série de la composition minérale en fluorures et sodium (mg/l) de 21 eaux minérales gazeuses (*cf.* tableau 3.5) ne peut que renforcer l'idée de la nécessité d'une étude graphique.

Tableau 3.5 – Données extraites du journal Que Choisir ?, *n° 422 bis, 2005*

Eau minérale	Fluorures	Sodium
Arcens	1,3	439
Arvie	0,9	650
Badoit	1	150
Beckerich	0,6	34
Châteauneuf	3	651
Eau de Perrier	0,05	11,5
Faustine	2	230
La Salvetat	0,25	7
Perrier	0,05	11,5
Puits St-Georges	0,5	434
Pyrénées	0,05	31
Quézac	2,1	255
San Pellegrino	0,6	35
St-Diéry	0,3	385
St-Jean	1,1	228
St-Pierre	1,7	383
St-Yorre	9	1 708
Vernet	1,3	120
Vernière	0,05	154
Vichy-Célestins	5	1 172
Wattwiller	1,6	3
Moyenne	*1,55*	*338*
Écart-type	*2,03*	*417*

Le coefficient de corrélation linéaire entre les deux composants minéraux est égal à 0,90. Cette valeur assez proche de 1 peut conduire à considérer que la droite des moindres carrés (*cf.* figure 3.8) permet d'évaluer approximativement la teneur Y en sodium en fonction de la teneur X en fluorures :

$$Y \approx 185X + 51 \quad \text{puisque} \quad r\frac{s_Y}{s_X} \simeq 185 \quad \text{et} \quad \bar{y} - 185\bar{x} \simeq 51$$

Mais la représentation graphique du nuage des 21 points (*cf.* figure 3.8) montre deux points caractérisés par une minéralité particulièrement élevée : « Vichy-Célestins » et « Saint-Yorre ».

La représentation des boîtes de distribution des deux variables « Fluorures » et « Sodium » (*cf.* figure 3.9) confirme que ces deux eaux minérales ont respectivement des valeurs « éloignée » et « extrême » pour les deux composants minéraux (chapitre 1, § IV).

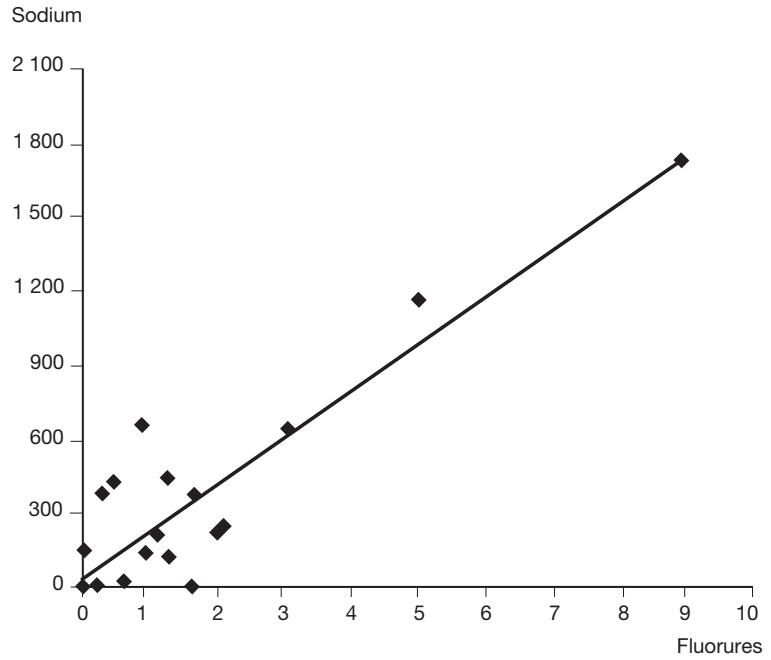


Figure 3.8 – Nuage des 21 eaux minérales gazeuses et droite des moindres carrés

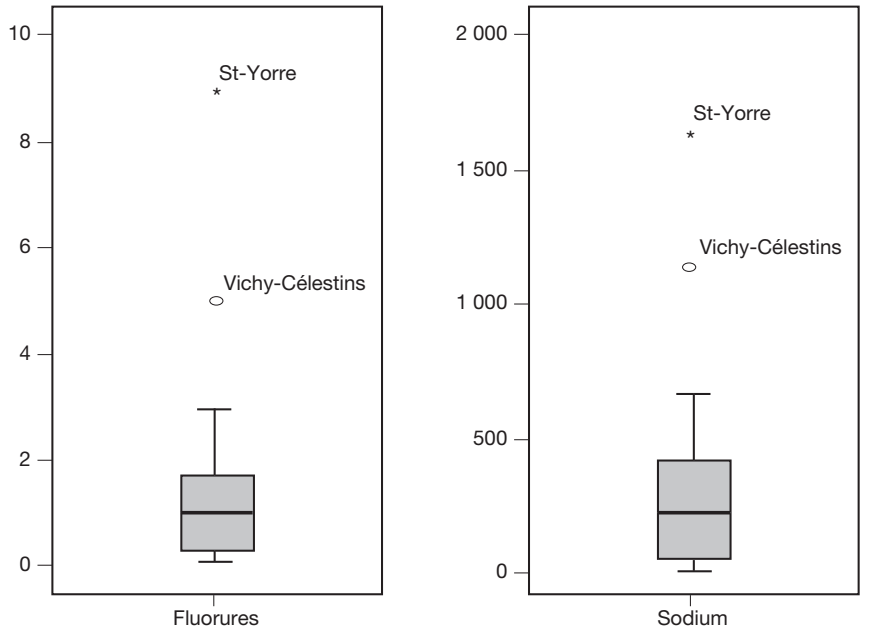


Figure 3.9 – Boîtes de distribution des deux composants « Fluorures » et « Sodium »

En supprimant ces deux points et en réalisant l'ajustement sur les 19 autres points, on obtient :

$$r \approx 0{,}50 \quad \hat{a} = 129 \quad \text{et} \quad \hat{b} = 96$$

Le coefficient r est passé de 0,9 à 0,5, et il faut aussi remarquer que les coefficients de la droite des moindres carrés sont passés respectivement de 185 à 129 et de 51 à 96.87

Quel crédit apporter à un ajustement pour lequel deux points ont une telle influence ? On est donc obligé d'abandonner l'idée d'une relation linéaire entre les deux composants minéraux.

Cet exemple nous montre que le calcul du coefficient de corrélation linéaire doit toujours être complété par un examen graphique.

L'*analyse exploratoire des données* propose d'autres méthodes et d'autres coefficients pour l'ajustement linéaire. Voici un exemple de coefficient proposé pour la mesure de la qualité de l'ajustement et pouvant être considéré comme un équivalent du carré du coefficient de corrélation linéaire qui, rappelons-le, peut être ainsi défini :

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Le deuxième terme de cette égalité peut être interprété comme le rapport de la variance des écarts $(y_i - \hat{y}_i)$, puisque ceux-ci sont de moyenne nulle, à la variance des y_i. L'analyse exploratoire des données propose de mesurer les dispersions de ces quantités par leur *étendue interquartile* , d'où le coefficient :

$$1 - \frac{EIQ(y_i - \hat{y}_i)}{EIQ(y_i)}$$

Si les points du nuage sont alignés, ce coefficient est égal à 1, et plus la dispersion des écarts à la droite sera faible (c'est le cas lorsque l'ajustement linéaire du nuage est adapté), plus il sera proche de 1. Pour l'ajustement des données « Eaux minérales gazeuses » réalisé par la méthode des moindres carrés, ce coefficient est égal à : $1 - 221/400 \simeq 0{,}45$

Au cas où l'examen graphique n'aurait pas été fait, cette valeur très différente de 1 doit amener à remettre en cause l'ajustement linéaire.

Tous ces résultats montrent qu'il ne faut jamais conclure sur la dépendance entre deux variables quantitatives au seul examen de la valeur du coefficient de corrélation linéaire.

D'autre part, lorsqu'une liaison linéaire entre deux variables a été mise en évidence par l'étude d'une série de n observations sur ce couple, il faut bien se garder de conclure à une relation de *cause à effet* entre ces variables sans en avoir examiné attentivement la signification : une corrélation voisine de 1 entre la taille (en cm) et la note à un contrôle de mathématiques pour un groupe de 12 élèves ne doit pas amener à conclure que plus on est grand, mieux on réussit en mathématiques !

L'*examen graphique*, ainsi que celui de la *signification des variables*, sont des compléments indispensables à l'information donnée par la valeur du coefficient de corrélation linéaire.

Dans le cas d'observations non connues individuellement et dont la distribution est donnée dans un *tableau de contingence*, le coefficient de corrélation linéaire et les droites des moindres carrés sont calculés à partir des formules pondérées. Cependant, si le groupement de données quantitatives en classes a l'avantage de permettre de présenter la distribution sous une forme synthétique et de pouvoir en déduire des profils en ligne ou en colonne, il constitue une perte d'information qu'il est préférable d'éviter de répercuter sur les calculs du coefficient de corrélation linéaire et des coefficients des droites des moindres carrés.

Nous avons exposé la méthode des moindres carrés pour l'ajustement d'un nuage de points par une droite qui est la fonction analytique la plus simple, mais cette méthode peut se généraliser à un ajustement par d'autres fonctions analytiques. Les logiciels proposent des ajustements par un polynôme du second degré, une fonction exponentielle… C'est l'examen graphique qui donne une indication sur le type de fonction à adopter. On peut aussi dans certains cas transformer une des deux variables ou les deux variables avant d'envisager une relation linéaire.

III. Une variable qualitative et une variable quantitative

Soient n observations portant simultanément sur une variable qualitative X à k modalités $\{x_1,\dots, x_i, \dots, x_k\}$ et sur une variable quantitative Y à l modalités $\{y_1,\dots, y_j, \dots, y_l\}$.

A. Mesure de la liaison par le rapport de corrélation

1) Définition du rapport de corrélation

Pour les n_i ($i = 1,\ldots, k$) observations de chaque modalité x_i de la variable X, on calcule la moyenne conditionnelle $\overline{y_i}$ et la somme des carrés des écarts à la moyenne (*cf.* tableau 3.6). On supposera tous les effectifs n_i (ou les fréquences $f_i = n_i/n$) non nuls, cette hypothèse impliquant la suppression des modalités pour lesquelles on ne dispose pas d'observations.

Tableau 3.6 – Caractéristiques de Y conditionnellement à X pour les données des tableaux 2 et 3

Modalité de X	n_i	$\overline{y_i}$	$\sum_{j=1}^{n} n_{ij}(y_{ij} - \overline{y_i})^2$
Filière *A*	37	5,16	496,91
Filière *B*	25	6,44	368,25
Filière *C*	16	12,31	293,44
Filière *D*	22	6,68	340,78

La moyenne $\overline{y_i}$ étant la moyenne de Y pour $X = x_i$, on a $\overline{y} = \sum_{i=1}^{k} f_i \overline{y_i}$ (§I.C), et pour notre exemple, $\overline{y} = 6{,}96$

On définit la *Somme des Carrés Intraclasse* , la *Somme des Carrés Interclasse* et la *Somme des Carrés Totale* :

$$SC_{\text{intra}} = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}(x_{ij} - \overline{y_i})^2 \qquad SC_{\text{inter}} = \sum_{i=1}^{k} n_i(y_i - \overline{y})^2$$

$$SC_{\text{tot}} = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}(y_{ij} - \overline{y})^2$$

On montre que : $SC_{\text{tot}} = SC_{\text{intra}} + SC_{\text{inter}}$

Le ***rapport de corrélation*** $\eta^2_{Y/X}$ de Y en x est ainsi défini :

$$\eta^2_{Y/X} = \frac{SC_{\text{inter}}}{SC_{\text{tot}}}$$

2) Interprétation du rapport de corrélation

Ce rapport est toujours positif et inférieur ou égal à 1. Il est *égal à 0* si la somme des carrés interclasse est nulle, c'est-à-dire si *les moyennes conditionnelles* $\overline{y_i}$ *sont toutes égales à* $\overline{y}$, mais cette condition n'est pas suffisante à l'indépendance des variables X et Y.

Si une variable quantitative Y est indépendante d'une variable qualitative X, alors leur *rapport de corrélation est nul*, mais la réciproque n'est pas vraie :

$$X \text{ et } Y \text{ indépendantes} \quad \begin{matrix}\Rightarrow \\ \not\Leftarrow\end{matrix} \quad \eta^2_{Y/X} = 0$$

Le rapport de corrélation est *égal à* 1 si la somme des carrés intraclasse est nulle, donc si *à chaque modalité x_i de X, correspond une* ***seule*** *valeur de Y égale à* $\overline{y_i}$

Dans ce cas, la variable Y est liée *fonctionnellement* à la variable X.

$$\eta^2_{Y/X} = 1 \iff \text{à chaque } x_i, \text{ correspond une } \textit{seule} \text{ valeur de } Y$$
$$\iff Y \textit{ liée fonctionnellement à } X$$

Pour les données du tableau 3.6 :

$$SC_{\text{inter}} = SC_{\text{tot}} - SC_{\text{intra}} = 2\,086 - 1\,499{,}38 = 586{,}62 \quad \Rightarrow \quad \eta^2_{Y/X} = 0{,}28$$

L'examen du tableau 3.3 des profils en ligne montre la dépendance entre la filière d'origine et la note, résultat en accord avec la valeur 0,28

Considérons maintenant une variable qualitative X à 3 modalités et une variable quantitative Y (discrète ou continue) rapportée à 2 valeurs (*cf.* tableau 3.7). Les observations portent sur n individus : $n_{11} + n_{22} + n_{31} = n$:

Tableau 3.7 –Tableau de contingence avec calculs des moyennes conditionnelles de Y

X \ Y	y_1	y_2	$\overline{y_i}$
x_1	n_{11}	0	y_1
x_2	0	n_{22}	y_2
x_3	n_{31}	0	y_1

Pour ces données :

$$\{ X = x_i \quad \Rightarrow \quad Y = \overline{y_i} \quad \text{pour } i = 1, 2, 3 \} \quad \Rightarrow \quad \eta^2_{Y/X} = 1$$

et ce résultat ne dépend pas des valeurs de y_1, y_2, n_{11}, n_{22} et n_{31}. Quelles que soient ces valeurs, la variable Y est liée fonctionnellement à la variable X.

Supposons maintenant que X soit une variable quantitative rapportée à 3 valeurs ainsi définies :

$$x_1 = 1 \quad x_2 = 4 \quad x_3 = 6$$

et que les effectifs soient les suivants (*cf.* tableau 3.8) :

$$n_{11} = 20 \quad n_{22} = 50 \quad n_{31} = 30$$

Tableau 3.8 –Valeurs particulières pour les effectifs du tableau 3.7

X \ Y	y_1	y_2	$\overline{y_i}$
1	20	0	y_1
4	0	50	y_2
6	30	0	y_1
$\overline{x_j}$	4	4	

La variable X étant quantitative, on peut aussi calculer le rapport de corrélation de X en y. Les moyennes conditionnelles de X étant égales, la somme des carrés interclasse est nulle et le rapport de corrélation $\eta^2_{Y/X}$ est nul.

Il y a donc *absence de corrélation* entre la variable X et toute fonction de Y. Cet exemple montre qu'on peut avoir à la fois Y lié fonctionnellement à X et absence de corrélation entre X et toute fonction de Y.

On remarquera que le rapport de corrélation $\eta^2_{Y/X}$ de cet exemple est nul quelles que soient les valeurs n_{11}, n_{22}, n_{31} et x_1, x_2, x_3 telles que les moyennes $\overline{x_1}$ et $\overline{x_2}$ soient égales, c'est-à-dire si :

$$\frac{n_{11}x_1 + n_{31}x_3}{n_{11} + n_{31}} = x_2$$

B. Comparaison du coefficient de corrélation linéaire et des rapports de corrélation

Si la variable X est une variable quantitative à k modalités, on peut représenter graphiquement les moyennes conditionnelles $\overline{y_i}$ en fonction des modalités de la variable X. On obtient k points qu'on peut joindre, dans l'ordre, par des segments de droite. On appelle la ligne brisée obtenue « courbe de régression de Y en x » (*cf.* figure 3.10).

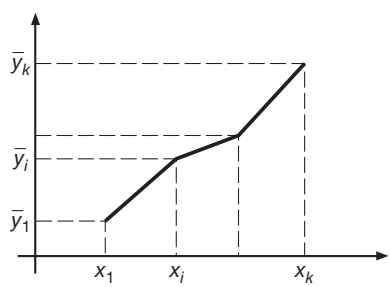


Figure 3.10 – Courbe de régression de Y en x

Dans ce cas où X et Y sont toutes les deux des variables quantitatives, on peut calculer le rapport de corrélation $\eta^2_{X/Y}$ de X en y (généralement non égal à $\eta^2_{Y/X}$) et tracer la « courbe de régression de X en y ».

Disposant de n observations portant simultanément sur deux variables quantitatives, on peut aussi calculer leur coefficient de corrélation linéaire et on montre qu'on a toujours la relation suivante entre les rapports de corrélation et le coefficient de corrélation linéaire :

$$0 \leq r^2 \leq \min(\eta^2_{X/Y};\eta^2_{Y/X}) \leq \max(\eta^2_{X/Y};\eta^2_{Y/X}) \leq 1$$

Si l'un des rapports de corrélation est nul, alors le coefficient de corrélation linéaire l'est aussi. Le lecteur peut vérifier que c'est le cas pour l'exemple du tableau 3.8, et il peut constater un nouvel exemple de deux variables non indépendantes avec un coefficient de corrélation linéaire nul.

- Si $\eta^2_{Y/X} = 0$, toutes les moyennes conditionnelles de Y sont égales et la courbe de régression de Y en x est une droite parallèle à l'axe des abscisses ; et réciproquement, si $\eta^2_{X/Y} = 0$, les moyennes conditionnelles de X sont égales et la courbe de régression de X en y est une droite parallèle à l'axe des ordonnées.
- Si $r^2 = \eta^2_{Y/X}$, alors les moyennes conditionnelles $\overline{y_i}$ sont liées aux modalités x_i par *une relation linéaire*, et la courbe de régression de Y en x est une droite qui n'est autre que la droite des moindres carrés $\mathbb{D}$ de Y en x :

$$r^2 = \eta^2_{Y/X} \quad \Leftrightarrow \quad \overline{y_i} = a + bx_i$$

et symétriquement, si $r^2 = \eta^2_{X/Y}$, alors la courbe de régression de X en y n'est autre que la droite des moindres carrés $\mathbb{D}'$ de X en y.

IV. Deux variables qualitatives

Les données relatives aux observations portant simultanément sur deux variables qualitatives X et Y sont généralement présentées dans un tableau de contingence (*cf.* tableau 3.1), ou dans un tableau de profils en ligne ou en colonne (*cf.* tableaux 3.3 et 3.4).

À condition de disposer des *effectifs marginaux*, on peut retrouver le tableau de contingence à partir d'un tableau de profils en ligne ou en colonne.

La question qui se pose est celle de l'existence d'une *liaison* entre les deux caractères X et Y. On a vu que s'ils sont *statistiquement indépendants* dans l'ensemble des n individus considérés (§ I.D) :

$$f_{ij} = f_{i\bullet} \cdot f_{\bullet j} \quad \Leftrightarrow \quad n_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

pour tous les couples (i, j) tels que $i = 1,\ldots, k$ et $j = 1,\ldots, l$

Dans le cas où les observations ne portent pas sur la population totale, mais sur une partie de la population appelée échantillon, on ne peut pas conclure à l'indépendance de X et Y par le seul examen des relations d'indépendance, leur non-vérification sur un échantillon pouvant être due au fait que les observations ne sont pas exhaustives ; autrement dit, il faut tenir compte des *fluctuations d'échantillonnage*.

La comparaison des effectifs *théoriques* (ou « attendus ») sous l'hypothèse d'indépendance $\left(n_{ij}^{*} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} \right)$ et des effectifs *observés* n_{ij} donne une idée de la *dépendance* entre X et Y. Mais pour être plus précis, il convient de calculer l'écart entre ces effectifs théoriques et observés.

Pour des raisons théoriques, la mesure usuellement adoptée est celle du χ^2 (khi-deux) qui peut être considérée comme un *coefficient d'association* entre deux variables :

$$\chi^2 = \sum_{i,\,j} \frac{(n_{ij} - n_{ij}^{*})^2}{n_{ij}^{*}} = n \sum_{i,\,j} \frac{(f_{ij} - f_{ij}^{*})^2}{f_{ij}^{*}} \quad \text{avec :} \quad f_{ij} = f_{i\bullet} \cdot f_{\bullet j} = \frac{n_{ij}^{*}}{n}$$

Le χ^2 est nul lorsque les effectifs théoriques et observés coïncident, et plus les effectifs théoriques et observés diffèrent, plus sa valeur est élevée.

Une autre mesure de la dépendance est le *coefficient d'association* Φ^2 (phi-deux) de Pearson égal à $\frac{\chi^2}{n}$. Ce coefficient ne dépend donc pas de la taille n de la population :

$$\Phi^2 = \sum_{i,\,j} \frac{(f_{ij} - f_{ij}^{*})^2}{f_{ij}^{*}}$$

Les valeurs de ces mesures d'*association* entre deux variables peuvent permettre de comparer plusieurs groupes d'observations sur un même couple de variables.

Reprenons les données du tableau 3.8 en considérant maintenant les variables X et Y comme des variables qualitatives et calculons les effectifs théoriques (ceux-ci sont écrits entre parenthèses dans le tableau 3.9) :

Tableau 3.9 – Calcul des effectifs théoriques du tableau 3.8

X \ Y	y_1	y_2	$n_{i\bullet}$
x_1	20 (10)	0 (10)	20
x_2	0 (25)	50 (25)	50
x_3	30 (15)	0 (15)	30
$n_{\bullet j}$	50	50	100

Les valeurs des mesures d'association χ^2 et Φ^2 sont les suivantes :

$$\chi^2 = \sum_{i,\,j} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}} = 100 \quad \text{et} \quad \Phi^2 = \frac{\chi^2}{n} = \sum_{i,\,j} \frac{(f_{ij} - f^*_{ij})^2}{f^*_{ij}} = 1$$

V. Bilan

La mesure de l'association de deux variables dépend de leur nature. Lorsque les observations de deux variables quantitatives sont suffisamment nombreuses pour être présentées dans un tableau de contingence, on peut traiter l'une d'elles comme une variable qualitative ou même les deux variables comme des variables qualitatives. Leur association peut se mesurer par le *coefficient de corrélation linéaire*, les *rapports de corrélation* et le *khi-deux*.

Pour les données du tableau 3.8, les deux variables X et Y ne sont pas indépendantes, mais :

- elles sont *linéairement* indépendantes puisque $r = 0$
- il y a *absence de corrélation* entre X et toute fonction de Y puisque $\eta^2_{X/Y} = 0$
- la variable Y est *liée fonctionnellement* à la variable X puisque $\eta^2_{Y/X} = 1$

L'analyse conjointe de deux variables est un problème très délicat ; il faut bien examiner les données avant de conclure à l'indépendance, et en cas de liaison, il convient de ne pas conclure hâtivement à une relation de cause à effet sans s'être penché sur sa signification concrète.

On n'oubliera pas les différents modes d'études de la liaison de deux variables selon leur nature

Nature des variables et présentation des données	Étude de la liaison entre deux variables X et Y
• X et Y quantitatives : n couples (x_i, y_i), ou tableau de contingence	• Calcul du coefficient de corrélation linéaire : $r = \dfrac{\text{cov}(X, Y)}{s_X \cdot s_Y}$ avec : $-1 \leq r \leq +1$ • Calcul et représentation graphique des *deux droites des moindres carrés* : $y - \overline{y} = r \cdot \dfrac{s_Y}{s_X} \cdot (x - \overline{x})$ $\qquad$ $y - \overline{y} = \dfrac{1}{r} \cdot \dfrac{s_Y}{s_X} \cdot (x - \overline{x})$ Elles se coupent au point moyen $(\overline{x}, \overline{y})$
• Y quantitative et X qualitative à k modalités (ou quantitative avec k classes de valeurs) Pour chaque modalité x_i de X, on dispose de : $n_{i\bullet}$ = nbre de valeurs de Y associées à $\{X = x_i\}$ moyenne conditionnelle $\overline{y_i}$ pour $\{X = x_i\}$	• Calcul du *rapport de corrélation* de Y en x : $\eta^2_{Y/X} = \dfrac{\sum_{i=1}^{k} n_i (y_i - \overline{y})^2}{SC_{\text{tot}}} = \dfrac{SC_{\text{inter}}}{SC_{\text{tot}}}$ • Si X est une variable quantitative classée, graphique de la *courbe de régression de Y en x* qui joint les points $(x_i, \overline{y_i})$
• X et Y quantitatives classées : tableau de contingence	• Calcul des *rapports de corrélation* de Y en x et de X en y : $\eta^2_{Y/X}$ et $\eta^2_{X/Y}$ • Graphiques de la *courbe de régression de Y en x* qui joint les points $(x_i, \overline{y_i})$, les valeurs x_i étant ordonnées, et de la *courbe de régression de X en y* qui joint les points $(\overline{x_j}, y_j)$, les valeurs y_j étant ordonnées.
• X qualitative, Y qualitative : tableau de contingence	• Calcul du *khi-deux* : $\chi^2 = \sum_{i,j} \dfrac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}} = n \sum_{i,j} \dfrac{(f_{ij} - f^*_{ij})^2}{f^*_{ij}}$

Testez-vous *(les réponses sont données page 284)*

Il y a *au moins* une réponse exacte par question.

1. Le coefficient de corrélation linéaire entre deux variables statistiques :

a) ne peut être calculé que si les deux variables sont quantitatives

b) est un nombre positif ou nul

c) est égal à 0 si les variables sont indépendantes

d) est un nombre sans dimension

2. Deux variables statistiques X et Y sont liées par la relation $X + Y = 2$, alors :

a) la somme de leur moyenne est égale à 2

b) les écarts-types des deux variables sont égaux

c) les deux droites des moindres carrés sont confondues

d) les deux droites des moindres carrés ont une pente positive

3. On dispose de deux observations {(2, 3) et (– 3, 1)} sur un couple (X, Y) de variables quantitatives :

a) les deux droites des moindres carrés sont confondues

b) le coefficient de corrélation linéaire entre X et Y est égal à + 1

c) la pente de la droite des moindres carrés de Y en x est négative

d) on peut calculer le rapport de corrélation de X en y

4. On dispose pour 10 années du nombre X d'abonnés au téléphone et du nombre Y de boîtes d'antalgiques (médicament contre la douleur) vendues dans une ville moyenne ; le coefficient de corrélation linéaire calculé à partir de ces 10 couples d'observations est égal à 0,996 :

a) les deux variables X et Y sont liées par une relation linéaire

b) pour diminuer la consommation d'antalgiques, il suffit de refuser des abonnements téléphoniques

c) les deux droites des moindres carrés sont quasi-confondues

d) les droites des moindres carrés ont des pentes négatives

5. Sur une population, on a observé une variable quantitative X et une variable qualitative Y à trois modalités. La distribution est la suivante :

X \ Y	y_1	y_2	y_3	
0	75	40	100	50
1	25	60	0	50
	100	100	100	100

a) ce tableau est un tableau de contingence

b) la variable X a une moyenne égale à 0,5

c) on peut mesurer la liaison entre X et Y par un rapport de corrélation

d) si les profils en colonne étaient tous identiques, alors X et Y seraient indépendantes

6. Le tableau suivant donne la distribution de deux variables statistiques X et Y :

X \ Y	0	3	4
0	20	20	0
1	10	40	10

a) la moyenne conditionnelle $\overline{x_1}$ est égale à 1/3

b) les moyennes conditionnelles de X s'obtiennent à partir du tableau des profils en colonnes

c) la moyenne $\bar{x}$ est égale à la somme des moyennes conditionnelles $\overline{x_1}$

d) les moyennes conditionnelles de Y s'obtiennent à partir du tableau des profils en lignes

7. Le tableau suivant donne la distribution conjointe de deux variables quantitatives X et Y :

X \ Y	0	1
– 1	a	10
1	10	b

a) si $a = 20$ et $b = 5$, alors le coefficient de corrélation linéaire r est nul

b) si $a = 0$ et $b = 0$, alors $r = -1$

c) si $a = 0$ et $b = 10$, alors $r = -1$

d) si $a = 10$ et $b = 10$, alors $r = 0$

8. Pour définir un tableau de contingence d'effectif total n à k lignes et l colonnes :

a) il suffit de connaître les effectifs marginaux

b) il suffit de connaître $k \cdot (l - 1)$ éléments du tableau

c) il suffit de connaître $k \cdot (l - 1)$ éléments du tableau et les sommes en lignes

d) il suffit de connaître $(k - 1) \cdot (l - 1)$ éléments du tableau et ses marges

9. Parmi un groupe de 100 malades qui se plaignent de ne pas bien dormir, certains ont pris un somnifère sous forme de cachet, d'autres ont pris un cachet de sucre ; tous pensaient prendre un somnifère. Après la nuit, on leur a demandé si le cachet avait été efficace. Le tableau suivant donne la répartition des réponses (on suppose que tous les malades ont dit la vérité) :

	Ont bien dormi	N'ont pas bien dormi
Somnifère	26	6
Sucre	48	20

a) ce tableau est un tableau de contingence

b) parmi les malades qui ont pris un somnifère, 26 % ont bien dormi

c) pour calculer le χ^2, il faut calculer les effectifs marginaux

d) le χ^2 est égal à 1,284

10. Ce tableau donne la répartition des salariés et non-salariés par sexe pour les actifs de 15 ans ou plus ayant un emploi et vivant en France métropolitaine :

	Hommes	Femmes
Non-salariés	**13,4**	**7,3**
Salariés	**86,6**	**92,7**
Intérimaires	2,8	1,4
Apprentis	1,7	0,9
Contrats à durée déterminée	6,0	10,8
Contrats à durée indéterminée	76,1	79,6
	100,0	100,0
Total des emplois (milliers)	13 670	12 243

Source : INSEE, enquêtes Emploi du 1er au 4e trimestre 2008.

a) les femmes plus souvent salariées que les hommes

b) la répartition entre les statuts « salariés » et « non-salariés » est indépendante du sexe

c) pour l'ensemble des hommes et des femmes, il y a 20,7 % de non-salariés

d) pour l'ensemble des hommes et des femmes, il y a 89,5 % de salariés

Exercices *(corrigés page 300)*

Exercice 3.1

Une étude menée par un groupe de compagnies d'assurances auprès de 30 000 assurés pour le risque « véhicules à moteur » a permis de déterminer les proportions (en pourcentage) d'assurés correspondant à la *puissance fiscale*, notée *X*, du véhicule assuré et au *kilométrage parcouru au cours de la dernière année*, noté *Y*. Les résultats sont reportés dans le tableau suivant :

X (chevaux fiscaux) \ *Y* (milliers de km)	< 10	[10 ; 20[	[20 ; 30[	[30 ; 40[	≥ 40
≤ 4	4,4	1,6			
5 – 6	7,2	8,2	4,0	2,6	
7 – 8	2,4	7,2	13,6	14,4	4,4
9 – 10			2,4	11,6	6,0
> 10				4,4	5,6

1. Précisez la population étudiée, les caractères étudiés et leur nature.

2. Donnez la distribution du kilométrage parcouru. Comment s'appelle cette distribution ? Calculez sa moyenne et son écart-type en supposant que tous les assurés ont fait au moins 2 000 km et au plus 50 000 km. Déterminez la médiane.

3. Donnez la distribution, en pourcentage, du kilométrage parcouru par les possesseurs d'une voiture d'une puissance fiscale d'au plus 6 CV. Quel est le type de cette distribution ?

Calculez sa moyenne et son écart-type.

Exercice 3.2

Dans une entreprise, on étudie la répartition de 100 salariées femmes (*cf.* tableau 1) et 140 salariés hommes (*cf.* tableau 2) selon le salaire mensuel brut *X* exprimé en euros et l'ancienneté *Y* exprimée en années.

Tableau 1 – Salariées femmes

X \ *Y*	[0 ; 4[	[4 ; 8[	[8 ; 12[	[12 ; 20[	[20 ; 28]
[1 200 ; 1 800[	12	10	10	8	
[1 800 ; 2 200[	8	14	5	4	4
[2 200 ; 3 000[		6	5	6	3
[3 000 ; 4 200]				2	3

Tableau 2 – Salariés hommes

X \ Y	[0 ; 4[	[4 ; 8[	[8 ; 12[	[12 ; 20[	[20 ; 28]
[1 200 ; 1 800[	10	6			
[1 800 ; 2 200[	4	9	18	8	8
[2 200 ; 3 000[	4	8	16	12	4
[3 000 ; 4 200]		5	8	8	12

1. Définissez la population étudiée, l'unité statistique, les caractères étudiés et leur nature.

2. Quel pourcentage de femmes gagnent moins de 2 200 € parmi les femmes qui ont moins de 8 ans d'ancienneté ?

3. Calculez la moyenne et l'écart-type du salaire des femmes, ainsi que la moyenne et l'écart-type du salaire des hommes. En déduire le salaire moyen de l'ensemble des 240 salariés.

4. Calculez la moyenne et l'écart-type de l'ancienneté des femmes.

5. Représentez le graphe des fréquences cumulées de la distribution marginale de l'ancienneté des femmes.

6. Calculez la distribution (en pourcentage) de l'ancienneté des femmes gagnant au moins 1 800 €.

7. On considère la distribution conjointe du salaire et de l'ancienneté des cent salariés femmes. Sachant que le coefficient de corrélation entre X et Y est égal à 0,45 pour cette distribution, donnez l'équation de la droite des moindres carrés de Y en X. Quel est le point d'intersection de cette droite avec l'autre droite des moindres carrés de X en Y ?

Exercice 3.3

Le tableau suivant donne les pourcentages de variation par rapport à la période précédente du produit intérieur brut (prix constants) et de la consommation finale privée (prix constants) en France (*source :* http://stats.oecd.org/)

Année	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
PIB	2,6	1,0	1,4	– 0,9	2,2	2,1	1,1	2,2	3,5	3,3
Consommation	2,5	0,6	1,0	– 0,4	1,4	1,7	1,6	0,4	3,9	3,5
Année	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**
PIB	3,9	1,9	1,0	1,1	2,5	1,9	2,2	2,3	0,4	– 2,2
Consommation	3,6	2,6	2,4	2,0	2,5	2,6	2,4	2,5	1,0	0,8

1. Calculez les séries des indices, base 1989, du PIB et de la Consommation privée.

– Calculez le coefficient de corrélation linéaire entre les indices du PIB et de la consommation privée.

2. Peut-on considérer qu'il y a approximativement une liaison linéaire entre les indices de volume du PIB et de la Consommation privée ? Calculez l'équation de la droite des moindres carrés expliquant l'indice de la Consommation privée en fonction de l'indice du PIB.

Représentez le nuage des 21 points avec la droite des moindres carrés.

Quelle est la part de variation de l'indice de la consommation privée expliquée par la relation linéaire ?

3. Calculez le coefficient de corrélation linéaire entre les variations du PIB et de la Consommation privée. Calculez l'équation de la droite des moindres carrés expliquant la variation de la Consommation privée en fonction de la variation du PIB.

Représentez le nuage des 20 points avec la droite des moindres carrés.

4. Vous semble-t-il plus intéressant d'analyser la liaison entre les variations du PIB et celles de la Consommation privée qu'entre les indices du PIB et de la Consommation privée ? Si oui, pourquoi ?

Exercice 3.4

Une entreprise a effectué un sondage auprès de sa clientèle pour connaître son appréciation sur le service livraison. Les résultats ont été les suivants :

	Pas du tout satisfait	Plutôt pas satisfait	Plutôt satisfait	Très satisfait
Clients de plus de 2 ans d'ancienneté	10	50	245	195
Clients d'au plus 2 ans d'ancienneté	40	90	205	165

1. Calculez le pourcentage total de clients plutôt satisfaits ou très satisfaits.

2. Calculez le pourcentage de clients de plus de 2 ans d'ancienneté parmi les clients plutôt satisfaits ou très satisfaits.

3. Donnez le tableau des profils en ligne.

4. Donnez le tableau de contingence obtenu en regroupant :
– d'une part les clients pas du tout satisfaits et plutôt pas satisfaits ;
– et d'autre part les clients plutôt satisfaits et très satisfaits.

5. Si les 2 caractères étaient indépendants, combien aurait-on de clients de plus de 2 ans d'ancienneté dans la catégorie plutôt satisfait ou très satisfait ?

D'après examen de juin 2001, GEA 1re année Paris IX-Dauphine.

Exercice 3.5

L'observation des quantités offertes sur un marché de raisin de table et des prix de vente a donné les résultats suivants :

Quantité X à la vente (tonnes)	100	120	84	78	87	80	110	95
Prix moyen Y par kg (euros)	1,60	1,40	1,95	2,10	1,75	2,25	1,50	1,80

1. Calculez le coefficient de corrélation linéaire entre X et Y.

2. Déterminez l'équation de la droite des moindres carrés de Y en X. Sans faire de calcul, donnez le signe de la pente de la droite des moindres carrés de X en Y.

3. On admet que la valeur du prix moyen Y par kg en fonction de la quantité à la vente X est déterminée par l'équation trouvée à la question 2.
La recette globale correspondant à la vente de la totalité du raisin est-elle une fonction constamment croissante de x ?
Sinon, quelle est la valeur critique x_c que les producteurs ont intérêt à ne pas dépasser ?

Exercice 3.6

Le tableau suivant, extrait de la revue *Synthèses*, « Revenus et patrimoine des ménages » (INSEE, n° 19, 1998), donne la répartition (en %) des ménages selon leur niveau de vie et leur type socio-économique.

Niveau de vie (en F/uc/mois) / Type socio-économique	Inférieur au 1er décile < 3 700	Du 1er décile au 3e quartile [3 700 ; 9 933[	Du 3e quartile au 9e décile [9 933 ; 13 900[	Au moins égal au 9e décile ≥ 13 900	Ensemble
Communes agricoles	22 18	13 70	7 8	5 4	12 100
Communes et quartiers ouvriers	41 11	41 71	31 12	22 6	37 100
Communes et quartiers des classes moyennes tertiaires	28 8	34 65	39 17	32 9	34 100
Communes et quartiers techniques très qualifiés	3 5	5 51	11 25	13 19	7 100
Quartiers huppés	6 6	6 44	12 20	28 30	9 100
Ensemble	100 10	100 65	100 15	100 10	100 100

« uc » : unité de consommation.
Lecture : 30 % des habitants des « quartiers huppés » appartiennent au 10e décile de niveau de vie (c'est-à-dire parmi les 10 % des ménages les plus aisés). Et 28 % des ménages du 10e décile habitent dans des quartiers huppés.
Champ : ménages hors étudiants.

Source : Enquête Logement 1996, INSEE.

1. Précisez la population étudiée, l'unité statistique, les caractères et leur nature.

2. Quels types de distributions avez-vous dans ce tableau ? Écrire les deux tableaux de distributions conditionnelles.

3. Donnez la valeur médiane du niveau de vie en F/uc/mois des ménages appartenant aux « Quartiers huppés ».

4. Proposez un indicateur de disparité des niveaux de vie pour l'ensemble des ménages. Donnez sa valeur.

5. Parmi les ménages ayant un niveau de vie supérieur au 3e quartile (ménages qui se situent parmi les 25 % ayant le niveau de vie le plus élevé, soit plus de

9 933 F/uc/mois), quel pourcentage habite dans un quartier « huppé » ou dans un quartier « technique très qualifié ».

6. Calculez la distribution (en %) du niveau de vie en F/uc/mois des ménages appartenant aux « communes et quartiers techniques très qualifiés » ou aux « quartiers huppés ».

Exercice 3.7

Le tableau suivant donne la distribution de 200 étudiants selon leur note d'examen X en Économie et leur note d'examen Y en Statistique.

X \ Y	[5 , 7[	[7 , 9[	[9 , 11[	[11 , 13[	[13 , 15[	[15 , 17[	[17 , 19]
[5 , 7[	7	3	2				
[7 , 9[	2	12	12	2			
[9 , 11[	1	10	18	8	2		
[11 , 13[		7	15	21	10	1	
[13 , 15[			11	12	13	5	
[15 , 17[			1	3	10	7	1
[17 , 19]					1	1	2

1. Calculez les rapports de corrélation de Y en x, et de X en y.

2. Tracez la courbe de régression de Y en x.

3. Peut-on calculer une autre mesure de la liaison des variables X et Y ?

Exercice 3.8

Reprenons les données relatives aux 21 eaux minérales gazeuses (*cf.* tableau 3.5). On recode la variable X (fluorures) en trois classes et la variable Y (sodium) en quatre classes, de la façon suivante :

$C1_X$	[0 ; 1[
$C2_X$	[1 ; 2[
$C3_X$	[2 ; 9]

$C1_Y$	[0 ; 100[
$C2_Y$	[100 ; 300[
$C3_Y$	[300 ; 500[
$C4_Y$	[500 ; 2 000]

1. Écrire un tableau qui a pour première colonne les eaux minérales, pour deuxième colonne la variable X^C (variable X recodée) égale au numéro de classe dans le recodage de X, et pour troisième colonne la variable Y^C (variable Y recodée) égale au numéro de classe dans le recodage de Y.

2. Construire le tableau de la distribution conjointe des variables X^C et Y^C (X^C en ligne et Y^C en colonne). Donnez le tableau des profils en ligne associé.

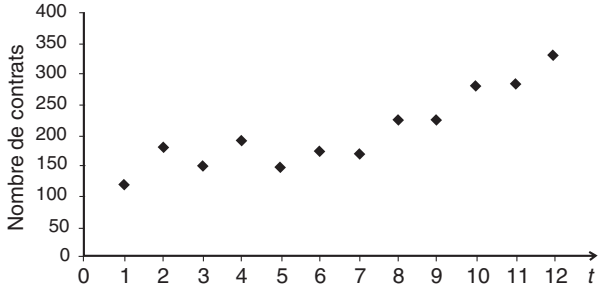
3. Les variables X^C et Y^C sont-elles indépendantes ? (justifiez votre réponse)

4. Donnez le tableau de distribution de la variable X^C sachant que Y est supérieur à 300 mg/l. Comment s'appelle cette distribution ?

Exercice 3.9

Soit les données trimestrielles suivantes relatives à des souscriptions de contrats d'assurance vie de fin mars 2000 à fin décembre 2002 :

t	Y = Nombre de contrats souscrits
1	117
2	178
3	149
4	189
5	145
6	173
7	170
8	223
9	223
10	281
11	285
12	339



Représentation graphique

Tous les résultats de cet exercice seront donnés avec une précision de deux décimales.

1. Calculez le taux trimestriel moyen de croissance du nombre de contrats souscrits.

2. *On ajuste cette série par le modèle linéaire : $Y = a\,t + b$.*

2.1. Calculez les coefficients de ce modèle par la méthode des moindres carrés.

2.2. Quelle est la part de variation de Y non expliquée par le modèle ?

3. *On ajuste maintenant cette série par le modèle quadratique : $Y = at^2 + b$.*

3.1. Calculez les coefficients de ce modèle par la méthode des moindres carrés.

3.2. Quelle est la part de variation de Y non expliquée par ce nouveau modèle ?

4. Quel modèle choisissez-vous ? (justifiez votre réponse)

D'après examen de juin 2006, DUGEAD 1re année Paris-Dauphine.

4. Séries chronologiques et prévision

Une série chronologique – ou chronique – est constituée par une suite ordonnée d'observations d'une grandeur au cours du temps. L'étude de ces séries intéresse tous ceux qui désirent décrire, expliquer, contrôler, prévoir des phénomènes évoluant au cours du temps.

I. Éléments constitutifs d'une série chronologique

L'étude d'une série chronologique $\{ x_t , t = 1,\dots , T\}$ consiste à dissocier les différents mouv ements qui la composent et à les analyser . Cette décomposition est une construction de l'esprit puisque les séries composantes sont des concepts abstraits et ne peuv ent pas être directement observées. Une représentation graphique s'impose en début d'analyse de toute chronique afin de faire apparaître les éléments fondamentaux.

Les intervalles entre deux observations successives sont supposés de même longueur. Dans la pratique, cette hypothèse est rarement réalisée. Pour des séries mensuelles de productions, de ventes…, le nombre de jours ouvrables de chaque mois varie : le nombre de dimanches dans le mois, les jours de certaines fêtes mobiles ne sont pas les mêmes chaque année. Pour que ces variations ne soient pas intégrées dans la composante résiduelle du modèle, on corrige les données en adoptant une correction proportionnelle qui consiste pour des données mensuelles, par exemple, à ramener chaque mois à un même nombre théorique de jours.

A. La tendance à long terme

La *tendance à long terme* ou *trend,* notée f_t , est le f acteur représentant l'évolution à long terme de la grandeur, et traduit l'aspect général de la série :

croissance de la consommation d'électricité, croissance du trafic aérien, diminution de la population rurale, par exemple.

Pour de longues séries, un *mouvement cyclique* peut se superposer à la tendance. La composante cyclique liée à la succession des phases du cycle économique (prospérité, dépression, reprise), a donné lieu jusqu'au milieu du XX^{e} siècle à de multiples travaux, mais n'est plus actuellement l'objet d'un intérêt aussi marqué.

B. Le mouvement saisonnier

Le *facteur saisonnier*, noté s_t, se répète à intervalles de temps égaux avec une forme à peu près constante. Il peut être dû au rythme des saisons ou à des facteurs humains. Sa période est de 12 pour des séries mensuelles, de 4 pour des séries trimestrielles…

Si p désigne la période du mouvement saisonnier : $s_t = s_{t+p} = s_{t+2p} = \ldots$

Le facteur saisonnier est donc totalement déterminé par p coefficients saisonniers :

$$s_1, \ldots, s_j, \ldots, s_p$$

C. Les irrégularités

Cette composante, appelée aussi *mouvement résiduel* et notée e_t, regroupe tout ce qui n'a pas été pris en compte par la tendance et le facteur saisonnier. Elle est la résultante de fluctuations irrégulières et imprévisibles dues à des facteurs perturbateurs non permanents ; ces fluctuations sont supposées de *faible amplitude* et de *moyenne nulle* sur un petit nombre d'observations consécutives.

D. Les perturbations

Les *perturbations* sont des fluctuations ponctuelles de forte amplitude. Elles sont dues, par exemple, à une grève, à des conditions météorologiques exceptionnelles pour l'agriculture, à un krach financier… Il convient de les *éliminer* avant tout traitement de la série. Les méthodes pour le faire sont simples ; pour *faire comme si* ces événements n'avaient pas eu lieu, les instruments privilégiés sont *l'interpolation* et *la règle de trois*. La représentation de la série chronologique des « Voyageurs RATP » de 1995 à 2002 (*cf.* figure 4.1) montre une baisse importante du nombre de voyageurs en décembre 1995 due à une longue grève. Avant d'estimer les composantes de cette chronique, il est nécessaire de corriger la valeur 0,19 milliard de voyageurs-km de ce mois de décembre en la remplaçant, par exemple par la

moyenne des mois de décembre 1994 et 1996 (resp. 1,06 et 0,95 milliards de voyageurs-km), soit 1 milliard de voyageurs-km.

On traite généralement des séries à deux composantes : tendance et mouvement résiduel, ou à trois composantes : tendance, mouvement saisonnier et mouvement résiduel. Les observations d'une chronique possédant une composante saisonnière peuvent être disposées dans un tableau selon les deux dimensions du temps, annuelle et mensuelle (ou trimestrielle), comme pour les tableaux 4.1, 4.2 et 4.4. Cette présentation, introduite par C. Buys-Ballot en 1847, est appelée « table de Buys-Ballot ».

II. Les modèles de composition d'une série chronologique

La décomposition d'une série chronologique possédant un mouvement saisonnier peut s'effectuer selon trois types de modèles :

- modèle additif $x_t = f_t + s_t + e_t \qquad t = 1,\dots, T$
- modèle multiplicatif $x_t = f_t \cdot (1+ s_t) \cdot (1+ e_t) \qquad t = 1,\dots, T$
- modèle mixte $x_t = f_t \cdot (1+ s_t) + e_t \qquad t = 1,\dots, T$

On choisit un modèle multiplicatif ou mixte si le mouvement saisonnier présente des amplitudes proportionnelles à la tendance.

Notons qu'une transformation logarithmique du modèle multiplicatif ramène au modèle additif :

$$\log(x_t) = \log\Big(f_t \cdot (1 + s_t) \cdot (1 + e_t)\Big) \approx \log(f_t) + \log(1 + s_t) + e_t$$

puisque $\log(1 + e_t) \approx e_t$

Nous n'envisagerons de méthodes de décomposition que pour les modèles additif et multiplicatif.

Pour le mouvement saisonnier de période p, on fait l'hypothèse d'une compensation exacte sur une période entre les variations saisonnières positives et les variations saisonnières négatives, sinon, le partage entre le facteur saisonnier et la tendance serait indéterminé :

$$\sum_{j=1}^{p} s_j = 0$$

Quand on analyse une série chronologique, le premier problème est le suivant : la série présente-t-elle des variations saisonnières et si oui, quel est le schéma de composition le mieux adapté ?

On commence par représenter la série graphiquement. Si la série présente des variations saisonnières, les points hauts (maxima) ainsi que les points bas (minima), sont toujours distants du même nombre de dates, ce nombre étant la période du mouvement saisonnier. La chronique représentée à la figure 4.1 a une composante saisonnière de période 12 (série mensuelle), et la chronique représentée la figure 4.2 a une composante saisonnière de période 4 (série trimestrielle).

Tableau 4.1 – Voyageurs RATP (milliards de voyageurs/km)

Année / Mois	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	Moyenne mensuelle
Janvier	1,04	0,91	0,98	1,01	1,05	1,09	1,14	1,18	1,22	1,21	1,08
Février	0,93	0,95	0,89	0,91	0,98	0,99	1,00	1,09	1,06	1,12	0,99
Mars	1,06	0,94	1,02	1,07	1,13	1,17	1,19	1,23	1,24	1,31	1,14
Avril	0,89	0,93	0,96	0,98	1,01	1,00	1,02	1,11	1,08	1,15	1,01
Mai	0,98	0,92	0,94	0,94	0,99	1,12	1,10	1,12	1,01	1,18	1,03
Juin	1,01	0,94	0,97	1,01	0,99	1,03	1,12	1,16	1,04	1,26	1,05
Juillet	0,79	0,85	0,86	0,88	0,90	0,99	0,99	1,08	1,01	1,07	0,94
Août	0,65	0,62	0,65	0,67	0,71	0,76	0,79	0,80	0,76	0,84	0,73
Septembre	0,87	0,92	0,93	1,00	1,02	1,04	1,05	1,12	1,14	1,2	1,03
Octobre	0,98	1,07	1,08	1,10	1,14	1,20	1,21	1,28	1,27	1,31	1,16
Novembre	0,83	0,96	0,99	1,04	1,05	1,14	1,14	1,16	1,16	1,24	1,07
Décembre	0,19	0,95	1,00	1,08	1,07	1,14	1,09	1,18	1,23	1,28	1,02
Moyenne annuelle	0,85	0,91	0,94	0,97	1,00	1,06	1,07	1,13	1,10	1,18	1,02

Source : www.insee.fr

Tableau 4.2 – Indices de valeur des produits alimentaires (base 2000)

Année / Mois	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	Moyenne mensuelle
Janvier	62,2	68,4	69,4	76,1	78,5	75,9	87,0	87,8	90,3	96,0	79,2
Février	64,7	67,2	70,7	79,3	76,3	79,1	82,5	89,1	90,1	93,9	79,3
Mars	81,3	77,9	78,9	92,4	99,0	99,7	98,7	99,9	102,7	117,5	94,8
Avril	72,4	83,3	87,1	92,7	90,1	88,5	89,6	103,0	108,9	118,2	93,4
Mai	85,3	85,0	84,6	91,8	90,8	103,8	100,7	100,0	103,1	108,1	95,3
Juin	84,5	85,8	86,3	98,3	100,5	98,6	102,6	103,6	116,4	133,6	101,0
Juillet	89,0	90,4	95,0	99,9	102,9	95,0	101,4	110,8	125,2	130,9	104,1
Août	82,5	81,1	88,6	93,3	102,4	108,1	107,7	107,6	117,6	125,0	101,4
Septembre	89,1	86,5	98,0	102,7	110,4	113,9	105,9	112,4	121,7	130,3	107,1
Octobre	85,1	92,9	101,7	96,0	104,0	105,3	111,0	119,8	125,8	118,4	106,0
Novembre	91,9	90,9	96,2	106,3	118,6	119,7	122,8	126,9	127,8	141,5	114,3
Décembre	88,5	98,5	101,5	107,3	111,9	112,6	107,8	122,5	134,8	142,3	112,8
Moyenne annuelle	81,4	84,0	88,2	94,7	98,8	100,0	101,5	107,0	113,7	121,3	99,0

Source : www.insee.fr

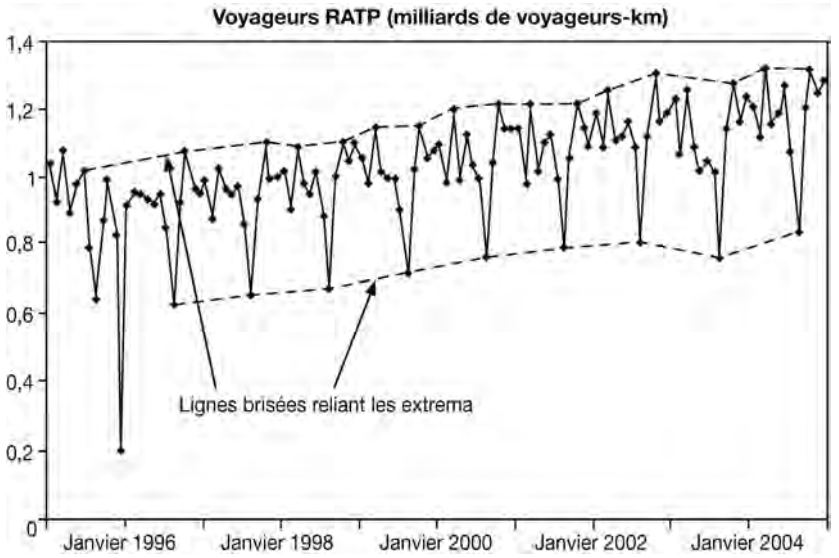


Figure 4.1 – Représentation graphique de la chronique du tableau 4.1

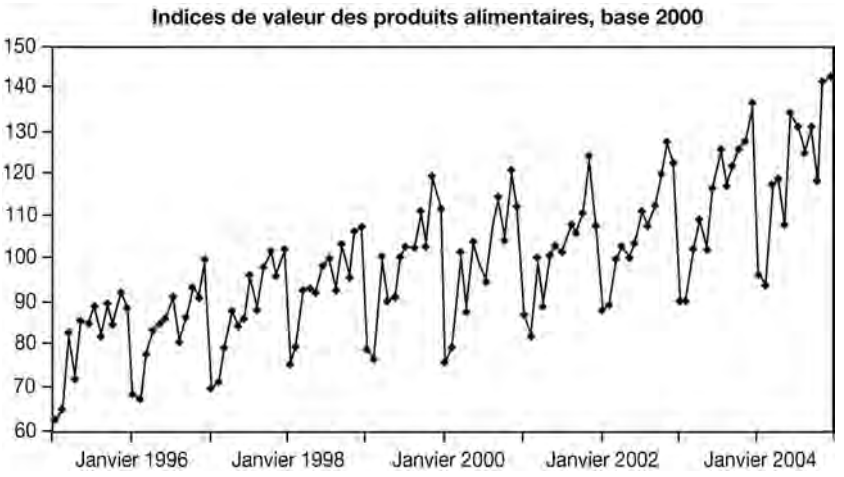


Figure 4.2 – Représentation graphique de la chronique du tableau 4.2

Pour *choisir le modèle de composition* , on peut relier par une courbe (ou plutôt par une ligne brisée) les maxima distants d'une période p et faire de même avec les minima.

• Si ces deux courbes sont à peu près parallèles, alors le facteur saisonnier a des amplitudes à peu près constantes, c'est-à-dire qu'il affecte la tendance indépendamment de son niveau, et le schéma additif est adapté.

C'est le cas de la chronique des « Voyageurs RATP » de 1995 à 2004 (*cf.* figure 4.1).

• Sinon, on représente la chronique sur un papier à *ordonnée logarithmique* (chapitre 2, § IV.A). Si les deux courbes reliant les extrema sont à peu près parallèles, alors le facteur saisonnier a des amplitudes à peu près proportionnelles à la tendance, c'est-à-dire que les effets des variations saisonnières sont *proportionnels* au niveau atteint par la tendance, et le schéma multiplicatif est adapté. C'est le cas de la chronique des « Indices de valeur des produits alimentaires » de 1995 à 2004 (*cf.* figures 4.2 et 4.3).

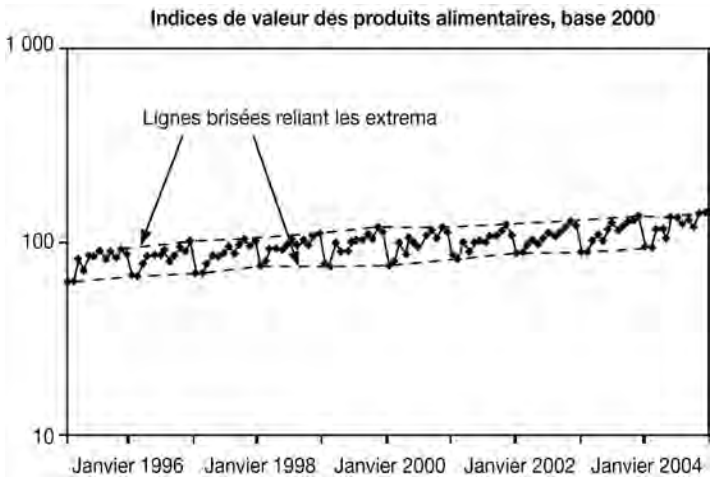


Figure 4.3 – Représentation de la chronique du tableau 4.2 avec une ordonnée logarithmique

Le modèle multiplicatif convient dans la plupart des cas puisque d'une part, l'effet saisonnier est généralement proportionnel à la tendance, et que d'autre part, dans le cas d'une chronique à tendance faiblement croissante ou faiblement décroissante, les deux schémas sont quasiment équivalents. C'est la raison pour laquelle on n'évoque bien souvent que le modèle multiplicatif.

III. Analyse de la tendance

A. Ajustement de la tendance par une fonction analytique

Les logiciels spécialisés (SPSS…), mais aussi les tableurs (Excel ®…), proposent des fonctions analytiques pour ajuster la tendance, l'ajustement se

faisant par la méthode des moindres carrés (méthode qui minimise les écarts quadratiques entre modèle et observ ations). Citons quelques-uns de ces modèles :

- modèle *linéaire* : $y(t) = a + b \cdot t$
- modèle *quadratique* : $y(t) = a + b \cdot t + c \cdot t^2$
- modèle *exponentiel* : $y(t) = \exp(a + b \cdot t)$, ce type de modèle convient à des quantités évoluant à taux constant puisque $y(t + 1) = \exp(b) \cdot y(t)$ ($\Leftrightarrow y(t + 1) = c \cdot y(t)$ où c est constant)
- modèle *logarithmique* : $y(t) = a + b \cdot \ln(t)$
- modèle S (courbe « sigmoïde ») : $y(t) = \exp(a + b/t)$, ce type d'ajustement convient à la description du cycle de vie de certains produits.

Ces méthodes analytiques sont simples, mais reposent sur l'hypothèse d'une tendance évoluant selon une fonction analytique déterminée, hypothèse qu'on ne peut pas fréquemment faire, même à la suite d'une transformation de variable.

En l'absence de référence à un modèle précis pour la tendance, on préférera utiliser une méthode non-paramétrique qui filtre la tendance en éliminant le facteur saisonnier tout en réduisant les irrégularités. Dans la suite, nous appellerons *filtre* une sorte de « boîte noire » régularisant une chronique X en la transformant en une chronique Y qui est une approximation de la composante tendancielle de la chronique X :

Nous étudierons deux des principaux filtres linéaires qui sont la moyenne mobile et le lissage exponentiel simple. Un filtre linéaire est une application linéaire de l'ensemble des chroniques dans lui-même transformant la chronique X en une nouvelle chronique Y de la façon suivante :

$$y_t = \sum_{k \in K} \alpha_k x_{t+k} \quad \text{avec} \quad K \subset \mathbb{Z} \quad \text{et} \quad \sum_{k \in K} \alpha_k = 1$$

Le choix du filtre linéaire approprié à certains objectifs se fait par l'intermédiaire du choix de ses coefficients α_k

B. Définition d'une moyenne mobile

On appelle ***moyennes mobiles centrées*** de longueur p ($p < T$) de la série $\{x_t, t = 1, \ldots, T\}$ les moyennes successives calculées en fonction de la parité de p selon les formules qui suivent.

• Premier cas, *p impair*, $p = 2m + 1$: $M_p(t) = \frac{1}{p} \sum_{k=-m}^{+m} x_{t+k}$

Il y a $(T - p + 1)$ moyennes mobiles centrées de longueur impaire p.

• Deuxième cas, *p pair*, $p = 2m$:

$$M_p(t) = \frac{1}{p}\left(\frac{x_{t-m}}{2} + \sum_{k=-m+1}^{m-1} x_{t+k} + \frac{x_{t+m}}{2} \right)$$

La moyenne mobile centrée $M_{2m}(t)$ apparaît comme la moyenne pondérée de valeurs de la série encadrant la date t avec les coefficients de pondération égaux à $\frac{1}{2p}$ pour les deux valeurs extrêmes x_{t-m} et x_{t+m} et égaux à $\frac{1}{p}$ pour les $(p - 2)$ valeurs intermédiaires x_{t-m+1} à x_{t+m-1}.

Elle comporte donc $(p + 1)$ termes :

Valeurs	x_{t-m}	x_{t-m+1}	....	x_t	....	x_{t+m-1}	x_{t+m}
Pondérations	$\frac{1}{2p}$	$\frac{1}{p}$	....	$\frac{1}{p}$	....	$\frac{1}{p}$	$\frac{1}{2p}$

Il y a $(T - p)$ moyennes mobiles centrées de longueur paire p.

Pour simplifier, la longueur p de la moyenne mobile étant fixée, on notera désormais y_t la moyenne mobile centrée de longueur p à la date t.

C. Détermination de la tendance par la méthode des moyennes mobiles

Si une série X est périodique de période p, c'est-à-dire si la série rede vient identique à elle-même tous les p termes, alors toute suite de mo yennes mobiles de longueur p' (différente de p) a pour période p.

Démontrons cette propriété dans le cas où p' est impair ($p' = 2m + 1$). Soit y_t la moyenne mobile centrée de longueur p' à la date t de la série X, montrons que la série Y est de période p :

$$y_{t+p} = \frac{1}{2m+1} \sum_{k=-m}^{+m} x_{t+p+k} = \frac{1}{2m+1} \sum_{k=-m}^{+m} x_{t+k} = y_t$$

La démonstration de cette propriété est laissée au lecteur pour le cas où p est pair, et celui-ci pourra montrer en sus que lorsque la période de la série

X est égale à la longueur de la moyenne mobile ($p' = p$), les moyennes mobiles forment alors une suite de termes constants égaux à la moyenne des termes de la série X sur une période.

La moyenne mobile centrée de longueur p rend *constantes* les séries périodiques de période p.

$\Rightarrow$ Deux chroniques ont la même suite de moyennes mobiles centrées de longueur p si leur différence est une série périodique de période p dont la somme des termes sur une période est nulle.

▶ **Exemple**

La chronique $\{x_t\ ,\ t = 1,\ldots,12\}$ du tableau 4.3 est périodique de période $p = 4$; les suites des moyennes mobiles de longueur 2, 3, 5 sont aussi de période 4, et la suite des moyennes mobiles de période 4 est une suite de termes constants égaux à -1/4, moyenne des termes sur une période.

Tableau 4.3 – Calcul de moyennes mobiles

	A	B	C	D	E	F	G	H	I	J
1	t	x_t	$M2(t)$	$M3(t)$	$M4(t)$	$M5(t)$				
2	1	2								
3	2	0	0,25	0,33						
4	3	-1	-1	-1,00	-0,25	0,2				
5	4	-2	-0,75	-0,33	-0,25	-0,2				
6	5	2	0,5	0,00	-0,25	-0,4				
7	6	0	0,25	0,33	-0,25	-0,6				
8	7	-1	-1	-1,00	-0,25	0,2				
9	8	-2	-0,75	-0,33	-0,25	-0,2				
10	9	2	0,5	0,00	-0,25	-0,4				
11	10	0	0,25	0,33	-0,25	-0,6				
12	11	-1	-1	-1,00						
13	12	-2								
14										
15										
16										

=(B2/2+SOMME(B3:B5)+B6/2)/4

=SOMME(B3:B7)/5

=SOMME(B5:B7)/3

=(B5/2+B6+B7/2)/2

Soit C, la courbe joignant les points (t , x_t). Si la concavité de C est tournée vers le haut, alors y_t est supérieur à x_t pour tout t ; dans le cas contraire, y_t est inférieur à x_t pour tout t. Si C est une droite, y_t est égal à x_t pour tout t.

En conclusion, la *moyenne mobile centrée transforme une série alignée en elle-même* et plus généralement, une série monotone à faible courbure en une série peu différente.

La moyenne mobile transforme des écarts dus à des irrégularités – indépendantes, de moyenne nulle sur un petit nombre de dates successives (par hypothèse) et de même variance – en écarts de variance plus faible; on dit qu'elle a un effet de « rabot », ou aussi qu'elle « lisse » la chronique, en ce sens que la série Y est moins dispersée que la série initiale X. Mais les nouvelles irrégularités qui sont corrélées entre elles, peuvent faire apparaître des oscillations parasites qui ne figuraient pas dans la série initiale (effet de Slutsky-Yule).

⇒ Si la période du mouvement saisonnier est égale à p, alors la moyenne mobile centrée de longueur p est un filtre linéaire qui élimine le mouvement saisonnier tout en réduisant l'amplitude du mouvement résiduel. De plus, on montre que sa valeur y_t à la date t peut être assimilée à la tendance f_t si celle-ci est à faible courbure – à faible variation dans le cas d'un schéma multiplicatif – sur p dates consécutives.

D. Inconvénients de la méthode des moyennes mobiles

Un changement de niveau ou de pente de la tendance à une date t entraîne une mauvaise approximation de cette composante pendant toute une période précédant et sui vant cette date (figure 4.4). C'est la raison pour laquelle on fait l'hypothèse d'une tendance monotone à faible courbure.

t	x_t	$M_4(t)$
1	130	
2	121	
3	120	123,50
4	127	123,50
5	122	125,00
6	129	124,63
7	124	125,88
8	120	128,88
9	139	131,13
10	136	134,25
11	135	135,63
12	134	134,88
13	136	134,88
14	133	135,13
15	138	
16	133	

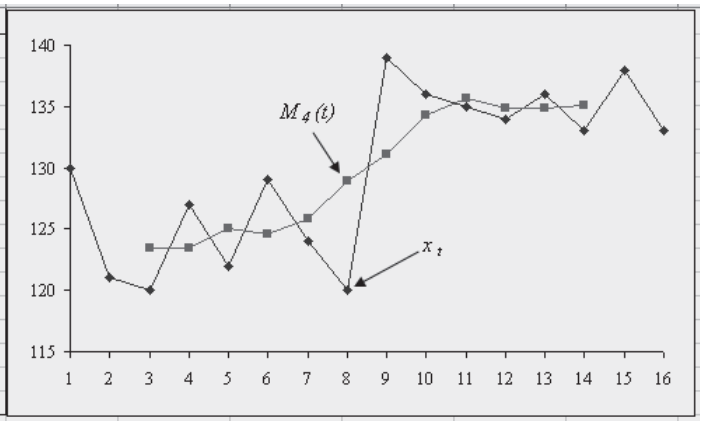


Figure 4.4 – Représentation d'une chronique et de ses moyennes mobiles centrées de longueur 4

Si on dispose de $T = np$ observations (n = nombre d'années et p = période du mouvement saisonnier) et si p est pair, on ne peut calculer que ($T - p$) moyennes mobiles de longueur p. On ne disposera pas de valeurs pour la tendance sur les $p/2$ dernières dates qui ne pourront pas être prises en compte pour une prévision.

Malgré ces inconvénients, on admettra que dans la plupart des cas, la valeur f_t de la tendance s'évalue par la moyenne mobile centrée y_t de longueur égale à la période du mouvement saisonnier.

IV. Correction des variations saisonnières

Si on étudie une chronique avec variations saisonnières, l'évaluation de la tendance à chaque date t par la moyenne mobile centrée de longueur adéquate, conduit pour chaque coefficient saisonnier à plusieurs valeurs qu'il faut résumer. Pour bien comprendre toutes les opérations successives pour la détermination des coefficients saisonniers et de la série corrigée des variations saisonnières, on se reportera à l'exemple traité au paragraphe 5.

A. Modèle additif

Le modèle est le suivant : $x_t = f_t + s_t + e_t$

On approxime la tendance f_t par la moyenne mobile centrée y_t.

Soient n le nombre d'années et p la période du facteur saisonnier :

$T = np$ observations $\Rightarrow$ $np - p = p\,(n-1)$ moyennes mobiles si p est pair (*cf.* tableau 4.5).

Les coefficients saisonniers étant périodiques de période p, on dispose pour chacun des p coefficients saisonniers de $(n-1)$ valeurs qui sont $(n-1)$ différences $\{x_t - y_t\}$. On résume ces $(n-1)$ valeurs par leur moyenne arithmétique, ou leur médiane, ou leur moyenne arithmétique après élimination de la valeur la plus faible et de la valeur la plus élevée (le logiciel SPSS utilise ce dernier résumé).

Si la somme des coefficients saisonniers n'est pas nulle sur une période, on corrige les coefficients saisonniers obtenus de façon à avoir une somme nulle :

$$s_t \quad \rightarrow \quad s_t^* = s_t - \bar{s} \quad \text{avec} \quad \bar{s} = \frac{1}{p}\sum_{t=1}^{p} s_t$$

On appelle *série corrigée des variations saisonnières* (série *CVS*) la série des différences :

$$x_t^* = x_t - s_t^*$$

Pour toutes les dates pour lesquelles on dispose de la valeur de la moyenne mobile, et donc d'une évaluation de la tendance, on peut calculer l'écart entre le modèle et l'observation :

$$e_t = x_t - y_t - s_t^* = x_t^* - y_t$$

Si le modèle est adapté, les valeurs absolues des écarts ne doivent pas être élevées, et leur somme voisine de zéro.

B. Modèle multiplicatif

Le modèle est le suivant : $x_t = f_t \cdot (1 + s_t) \cdot (1 + e_t)$

Comme précédemment, on approxime la tendance f_t par la moyenne mobile centrée y_t.

Les coefficients saisonniers étant périodiques de période p, on dispose pour chacun des p coefficients saisonniers de $(n - 1)$ valeurs qui sont $(n - 1)$ quotients $\{x_t / y_t\}$. On résume ces $(n - 1)$ valeurs par leur moyenne arithmétique, ou leur médiane, ou leur moyenne arithmétique après élimination de la valeur la plus faible et de la valeur la plus élevée (le logiciel *SPSS* utilise ce dernier résumé).

Si la somme des $(1 + s_t)$ n'est pas égale à p sur une période, on fait une correction proportionnelle :

$$1 + s_t \quad \rightarrow \quad 1 + s_t^* = \frac{1 + s_t}{1 + \bar{s}} \quad \text{avec} \quad \bar{s} = \frac{1}{p}\sum_{t=1}^{p} s_t$$

On établit ensuite la série corrigée des variations saisonnières :

$$x_t^* = \frac{x_t}{1 + s_t^*}$$

Dans le cas du modèle multiplicatif, les coefficients saisonniers s'expriment en pourcentage de la tendance. Ils ont une interprétation plus concrète que ceux du modèle additif.

Le modèle multiplicatif prédit ainsi des valeurs $y_t \cdot (1 + s_t^*)$ et il est alors naturel, pour toutes les dates auxquelles on dispose de la valeur de la moyenne mobile, et donc d'une évaluation de la tendance, de considérer les résidus et sous la forme :

$$e_t = \frac{x_t}{y_t \cdot \left(1 + s_t^*\right)} - 1 = \frac{x_t^*}{y_t} - 1$$

Les écarts entre le modèle et les observations sont égaux à :

$$x_t - y_t \cdot (1 + s_t^*) = y_t \cdot (1 + s_t^*) \cdot e_t$$

Si le modèle est adapté, les valeurs absolues des écarts ne doivent pas être élevées, et leur somme voisine de zéro.

C. Autres approches

On peut chercher à améliorer l'évaluation de la tendance en repassant un filtre moyenne mobile sur la série *CVS*. On choisit généralement une longueur assez faible pour cette nouvelle suite de moyennes mobiles : 5 ou 7 dans le cas d'une série de période 12, et 3 dans le cas d'une série de période 4. Avec cette nouvelle évaluation de la tendance, on détermine de nouveaux coefficients saisonniers et une nouvelle série *CVS*. Cette méthode itérative pourrait évidemment être poursuivie, mais le gain devient à peu près nul au-delà de deux étapes.

On peut aussi remplacer la moyenne mobile centrée par la médiane mobile centrée qui est un filtre non linéaire : au lieu de synthétiser une suite de valeurs de la série par une moyenne pondérée, on les résume par leur médiane (particulièrement aisée à déterminer à la main avec $p = 3$). Les médianes mobiles, développées par Tukey, sont robustes puisqu'étant fondées sur l'utilisation de statistiques d'ordre, elles éliminent les valeurs « singulières » (chapitre 1, § III.B.4). Elles constituent des lisseurs aux propriétés complémentaires des moyennes mobiles. Certaines méthodes de désaisonnalisation reposent sur une association de ces deux types de lisseurs.

Disposant des coefficients saisonniers, on peut ajuster la série *CVS* par une fonction, faire une prévision pour la tendance en extrapolant cette fonction d'ajustement ou en utilisant une méthode de lissage exponentiel sur la série *CVS* (§ VI). Mais, il ne faut pas oublier que ce mode de prévision ne peut être envisagé que sur du court terme puisqu'il suppose une évolution future non perturbée par des changements sur l'environnement.

V. Un exemple de décomposition d'une série chronologique

Pour déterminer la tendance et les coefficients saisonniers d'une chronique, on peut actuellement utiliser un logiciel ou un tableur.

Néanmoins, une bonne compréhension des méthodes demande de les avoir appliquées. On va montrer les étapes successives du traitement de la chronique des ventes trimestrielles en France d'essences aviation (*cf.* tableau 4.4).

Tableau 4.4 – Ventes en France d'essence aviation (en milliers de tonnes)

Trimestre / Année	1er trimestre	2e trimestre	3e trimestre	4e trimestre	Moyenne annuelle
2005	3,6	7,0	7,6	3,7	5,5
2006	3,6	6,7	7,4	3,9	5,4
2007	3,7	6,4	7,1	4,1	5,3
2008	3,6	5,7	7,1	3,7	5
Moyenne trimestrielle	3,7	6,5	7,6	3,9	5,3

Source : Comité Professionnel du Pétrole

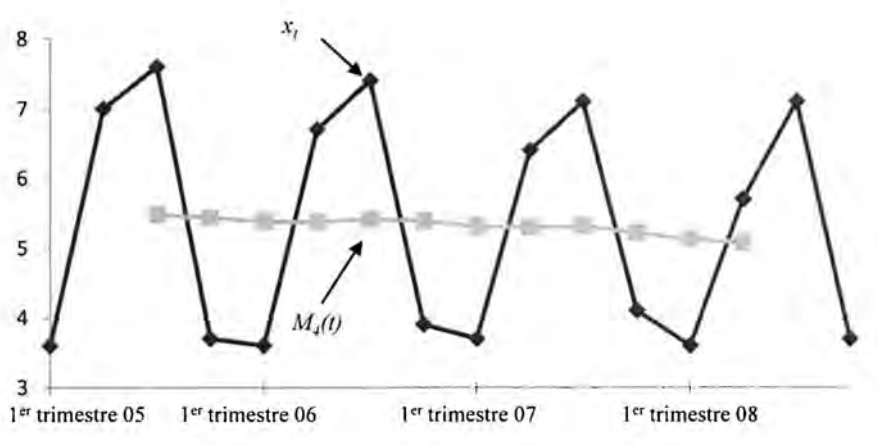


Figure 4.5 – Chronique du tableau 4.4 et suite des moyennes mobiles de longueur 4

Une saisonnalité de période 4 (nombre de trimestres dans l'année) apparaît sur la représentation graphique (*cf.* figure 4.5), ce qui explique que la suite des moyennes mobiles de longueur 4 filtre la tendance.

Pour une décomposition de cette chronique, nous allons envisager successivement le modèle additif et le modèle multiplicatif.

A. Schéma additif

Pour obtenir la série *CVS* et la série des résidus, les calculs ont été réalisés à l'aide du tableur Excel® selon les étapes indiquées (*cf.* tableau 4.5). Dans cet exemple, la synthèse des coef ficients saisonniers a été réalisée par la moyenne.

Tableau 4.5 – Décomposition de la chronique du tableau 4.4 avec le schéma additif

C4 : =(B2/2+B3+B4+B5+B6/2)/4

	A	B	C	D	E	F	G	H
1	t	x_t	$M_4(t)$	$x_t - M_4(t)$	s_t	s^*_t	*CVS*	*Ecart*
2	1	3,6			-1,64	-1,61	5,21	
3	2	7			1,02	1,04	5,96	
4	3	7,6	5,48	2,13	1,97	1,99	5,61	0,13
5	4	3,7	5,44	-1,74	-1,45	-1,42	5,12	-0,32
6	5	3,6	5,38	-1,78	-1,64	-1,61	5,21	-0,16
7	6	6,7	5,38	1,33	1,02	1,04	5,66	0,28
8	7	7,4	5,41	1,99	1,97	1,99	5,41	0,00
9	8	3,9	5,39	-1,49	-1,45	-1,42	5,32	-0,07
10	9	3,7	5,31	-1,61	-1,64	-1,61	5,31	0,00
11	10	6,4	5,30	1,10	1,02	1,04	5,36	0,06
12	11	7,1	5,31	1,79	1,97	1,99	5,11	-0,20
13	12	4,1	5,21	-1,11	-1,45	-1,42	5,52	0,31
14	13	3,6	5,13	-1,53	-1,64	-1,61	5,21	0,09
15	14	5,7	5,08	0,62	1,02	1,04	4,66	-0,42
16	15	7,1			1,97	1,99	5,11	
17	16	3,7			-1,45	-1,42	5,12	

colonne C : moyennes mobiles de longueur 4 évaluant la tendance

C4 =(B2/2 + SOMME(B3 : B5) + B6/2)/4, puis « recopier vers le bas »

colonne D : différence entre valeurs observées et tendance

D4 = B4 – C4, puis « recopier vers le bas »

colonne E : E4 = (D4 + D8 + D12) / 3

E5 = (D5 + D9 + D13) / 3

E6 = (D6 + D10 + D14) / 3

E7 = (D7 + D11 + D15) / 3

$\Rightarrow$ premières valeurs des 4 coefficients saisonniers qu'on reporte sur toute la colonne à l'aide du « collage spécial », option « coller valeurs », puis « recopier vers le bas »

colonne F : calcul des coefficients saisonniers « normalisés » : F2 = E2 – $\bar{s}$ puis « recopier vers le bas »

colonne G : calcul de la série *CVS*

G2 = B2 – F2, puis « recopier vers le bas »

colonne H : calcul de la série des écarts (§ IV.A)

H4 = G4 – C4, puis « recopier vers le bas »

B. Schéma multiplicatif

Comme pour le modèle additif, les calculs ont été réalisés à l'aide du tableur Excel® (*cf.* tableau 4.6). La synthèse des coef ficients saisonniers a aussi été réalisée par la mo yenne. Les différences entre certains résultats donnés dans le tableau 4.6 avec ceux obtenus par calcul direct, sont à e xpliquer par le f ait que Excel® utilise pour les calculs un grand nombre de décimales.

Tableau 4.6 – Décomposition de la chronique du tableau 4.4 selon le schéma multiplicatif

D4 : =B4/C4

	A	B	C	D	E	F	G	H	I	J
1	t	x_t	$M_4(t)$	$x_t/M_4(t)$	$1+s_t$	$1+s^*_t$	CVS	$1+e_t$	e_t	*Ecart*
2	1	3,6			0,69	0,70	5,18			
3	2	7			1,19	1,20	5,84			
4	3	7,6	5,48	1,39	1,36	1,37	5,55	1,01	0,01	0,10
5	4	3,7	5,44	0,68	0,73	0,74	5,03	0,92	-0,08	-0,30
6	5	3,6	5,38	0,67	0,69	0,70	5,18	0,96	-0,04	-0,14
7	6	6,7	5,38	1,25	1,19	1,20	5,59	1,04	0,04	0,26
8	7	7,4	5,41	1,37	1,36	1,37	5,40	1,00	0,00	-0,01
9	8	3,9	5,39	0,72	0,73	0,74	5,30	0,98	-0,02	-0,07
10	9	3,7	5,31	0,70	0,69	0,70	5,32	1,00	0,00	0,01
11	10	6,4	5,30	1,21	1,19	1,20	5,34	1,01	0,01	0,05
12	11	7,1	5,31	1,34	1,36	1,37	5,18	0,98	-0,02	-0,18
13	12	4,1	5,21	0,79	0,73	0,74	5,57	1,07	0,07	0,26
14	13	3,6	5,13	0,70	0,69	0,70	5,18	1,01	0,01	0,04
15	14	5,7	5,08	1,12	1,19	1,20	4,76	0,94	-0,06	-0,38
16	15	7,1			1,36	1,37	5,18			
17	16	3,7			0,73	0,74	5,03			

colonne C : moyennes mobiles de longueur 4 évaluant la tendance

C4 = (B2/2 + SOMME(B3:B5) + B6/2)/4, puis « recopier vers le bas »

colonne D : quotient entre valeurs observées et tendance

D4 = B4 / C4, puis « recopier vers le bas »

colonne E : E4 = (D4 + D8 + D12) / 3

E5 = (D5 + D9 + D13) / 3

E6 = (D6 + D10 + D14) / 3

E7 = (D7 + D11 + D15) / 3

$\Rightarrow$ premières valeurs des 4 coefficients $(1 + s_t)$ qu'on reporte sur toute la colonne à l'aide du « collage spécial », option « coller valeurs »

colonne F : calcul des coefficients saisonniers « normalisés » : F2 = E2 – $\bar{s}$,
puis « recopier vers le bas »
colonne G : calcul de la série *CVS*
G2 = B2 / F2, puis « recopier vers le bas »
colonne H : calcul de la série $(1 + e_t)$
H4 = G4 / C4, puis « recopier vers le bas »
colonne I : calcul de la série e_t
I4 = H4 – 1, puis « recopier vers le bas »
colonne J : calcul de la série des écarts (§ IV.B)
J4 = C4 · F4 · I4, puis « recopier vers le bas »

Les séries *CVS* induites par les deux modèles de composition sont presque confondues (*cf.* figure 4.6).

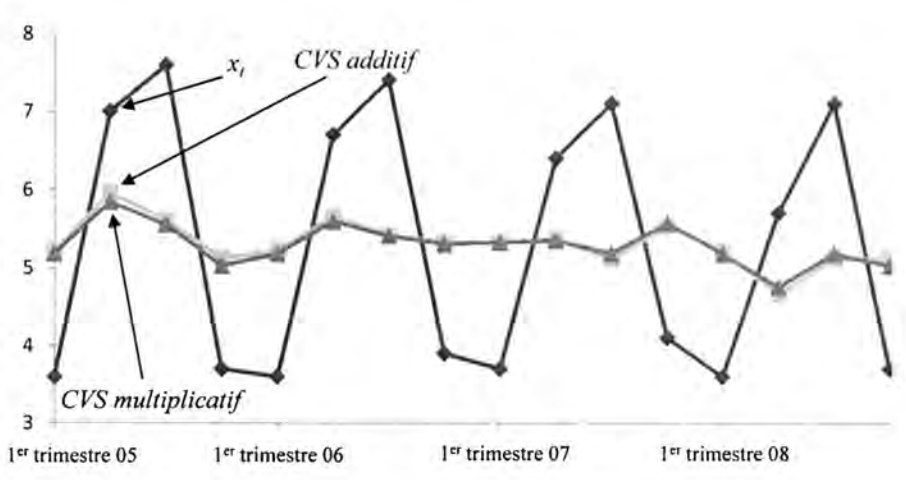


Figure 4.6 – Séries CVS

La représentation des séries des écarts (colonne H du tableau 4.5 et colonne J du tableau 4.6) permet de comparer les ajustements entre les deux modèles et les observations (*cf.* figure 4.7). On constate que les deux séries des écarts sont presque confondues.

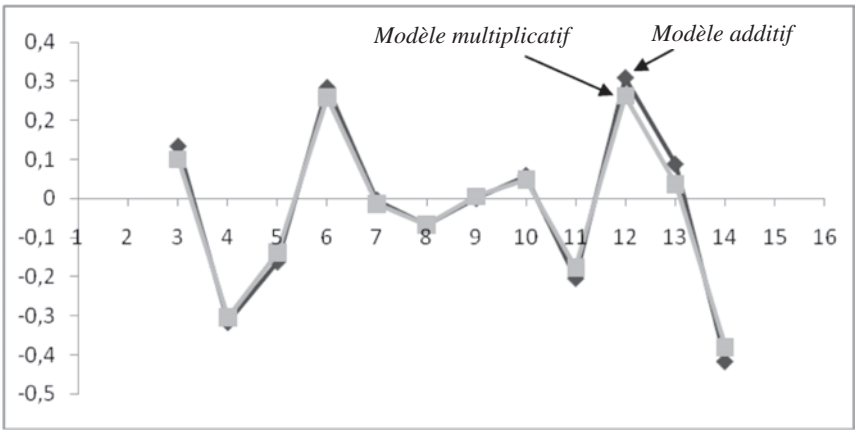


Figure 4.7 – Écarts entre les modèles et les observations

VI. Les méthodes de lissage exponentiel

Les méthodes de lissage exponentiel, développées par R. G. Brown dans les années 60 [1], sont des méthodes d'extrapolation qui donnent un poids prépondérant aux valeurs récentes. Elles se caractérisent, en outre, par la simplicité des calculs et le petit nombre des données à garder en mémoire.

A. Le lissage exponentiel simple

Cette méthode de prévision s'applique à des ***chroniques sans variations saisonnières et à tendance localement constante***. On suppose la grandeur observée caractérisée par des variations irrégulières autour de la moyenne :

$$x_t = a + e_t \qquad t = 1, \ldots, T$$

Les séries économiques présentent souvent un niveau moyen qui évolue à travers le temps. Pour la chronique représentée à la figure 4.4, il est clair

1. R. G. Brown, *Smoothing, forecasting and prediction of discrete time series*, Prentice Hall, 1962.

que le recours à la moyenne arithmétique des observations conduirait à sous-évaluer les valeurs futures. Il convient de donner aux observations les plus récentes un poids prépondérant.

La prévision $\hat{x}_T(h)$ faite par la méthode de lissage exponentiel simple à la date T pour l'horizon h, c'est-à-dire pour la date $T + h$, est la suivante :

$$\hat{x}_T(h) = \alpha \sum_{i=0}^{T-1} (1-\alpha)^i \cdot x_{T-i} \quad \text{avec} \quad 0 < \alpha < 1$$

Le paramètre α est la constante de lissage. Si T est élevé, la somme des pondérations est peu différente de 1, en effet :

$$\alpha \sum_{i=0}^{T-1} (1-\alpha)^i = \alpha \cdot \frac{1-(1-\alpha)^T}{\alpha} = 1-(1-\alpha)^T \approx 1$$

et la prévision $\hat{x}_T(h)$ apparaît comme la moyenne pondérée des valeurs $x_1, \ldots, x_T$. Cette prévision ne dépendant pas de l'horizon h, nous la noterons désormais $\hat{x}_T$

Cette méthode de prévision repose sur l'idée que les observations influencent d'autant moins la prévision qu'elles sont éloignées de la date T. En outre, on suppose cette décroissance exponentielle. Plus la constante de lissage α est proche de 0, plus l'influence des observations passées remontera loin dans le temps et plus la prévision sera « rigide », c'est-à-dire peu sensible aux fluctuations conjoncturelles. Au contraire, plus la constante de lissage α est voisine de 1, plus la prévision sera « souple », c'est-à-dire principalement influencée par les observations récentes.

1) Autres interprétations de la méthode

• On voit aisément que :

$$\hat{x}_T = (1-\alpha) \cdot \hat{x}_{T-1} + \alpha \cdot x_T \qquad (1)$$

La prévision apparaît comme la moyenne pondérée entre la prévision $\hat{x}_{T-1}$ faite à la date $T-1$ et la dernière observation x_T, le poids donné à cette observation étant d'autant plus fort que α est plus élevé.

Dans le cas où α est égal à 1 : $\hat{x}_T = s_T$, ce qui signifie que la prévision est égale à la dernière valeur observée, on parle de prévision « naïve ».

• On peut encore écrire :

$$\hat{x}_T = \hat{x}_{T-1} + \alpha \cdot (x_T - \hat{x}_{T-1}) \qquad (2)$$

La prévision apparaît alors comme égale à la prévision à la date précédente corrigée d'un terme proportionnel à la dernière erreur de prévision.

Dans ces deux formules qui fournissent des méthodes élémentaires de mise à jour de la prévision, l'information apportée par le passé est résumée dans $\hat{x}_{T-1}$

• On peut montrer que la valeur de a qui minimise la quantité :

$$\sum_{i=0}^{T-1} (1-\alpha)^i \cdot (x_{T-1} - a)^2 \qquad (3)$$

est la suivante :

$$\hat{a} = \alpha \cdot \frac{\sum_{i=0}^{T-1} (1-\alpha)^i \cdot x_{T-i}}{1-(1-\alpha)^T} \approx \hat{x}_T$$

La prévision s'interprète alors comme la constante qui s'ajuste le mieux à la série « au voisinage » de T, l'expression « au voisinage » traduisant le fait que dans la minimisation, l'influence des observations décroît lorsqu'on s'éloigne de la date T.

Cette dernière interprétation montre clairement que le lissage exponentiel simple ne s'applique que si la chronique peut être approchée par une droite horizontale au voisinage de T, ce qui implique une tendance localement constante.

2) Propriétés du lissage exponentiel simple

1. La chronique lissée $\{\hat{x}_t, t = 1,\ldots,T\}$ a une variance inférieure à celle de la chronique initiale $\{x_t, t = 1,\ldots,T\}$ Comme tout filtre, le lissage exponentiel simple réalise un « écrêtage » des irrégularités de la série.

2. Le lissage exponentiel simple est un filtre linéaire.

3. De même que la moyenne mobile, le lissage exponentiel simple s'adapte avec retard à un changement de niveau de la chronique (*cf.* figures 4.4 et 4.8). C'est de la valeur de la constante de lissage α que dépendent la stabilité et le taux de réponse de la série lissée, ces deux caractéristiques ayant un aspect complémentaire.

3) Mise en œuvre de la méthode

a) Initialisation

La méthode du *LES* utilisée à l'aide des formules (1) ou (2) nécessite l'initialisation de l'algorithme. On prend généralement $\hat{x}_1$ égal à x_1 ou $\hat{x}_1$ égal à $\bar{x}$ (initialisation par déf aut du logiciel SPSS), et il est clair que la v aleur choisie pour $\hat{x}_1$ aura d'autant moins d'influence sur que T sera grand.

b) Choix de la constante de lissage

Ce choix peut se f aire selon des critères subjectifs de « rigidité » ou de « souplesse » de la prévision. Mais une méthode plus objecti ve consiste à choisir α minimisant :

- soit l'*Erreur Quadratique Moyenne de prévision* :

$$EQM = \frac{1}{T-1}\sum_{t=1}^{T-1}(x_{t+1} - \hat{x}_t)^2$$

- soit l'*Erreur Absolue Moyenne de prévision* :

$$EAM = \frac{1}{T-1}\sum_{t=1}^{T-1}|x_{t+1} - \hat{x}_t|$$

Il ne faut pas manquer d'examiner aussi l' *Erreur Moyenne de prévision* qui peut indiquer dans certains cas une sous-évaluation ou une surévaluation systématique de la prévision qui s'observe d'ailleurs à l'examen des graphiques des séries initiales et lissées :

$$EM = \frac{1}{T-1}\sum_{t=1}^{T-1}(x_{t+1} - \hat{x}_t)$$

La minimisation de ces critères peut être faite sur toute la série des erreurs de prévision ou sur un pourcentage donné de ses derniers termes (dans ce cas, on prend souvent le dernier tiers de la série, tableau 4.7). Certains logiciels proposent actuellement les méthodes de lissage avec une constante α déterminée par la minimisation d'un critère. Le logiciel SPSS calcule la constante optimale en minimisant l'Erreur Quadratique Moyenne de prévision.

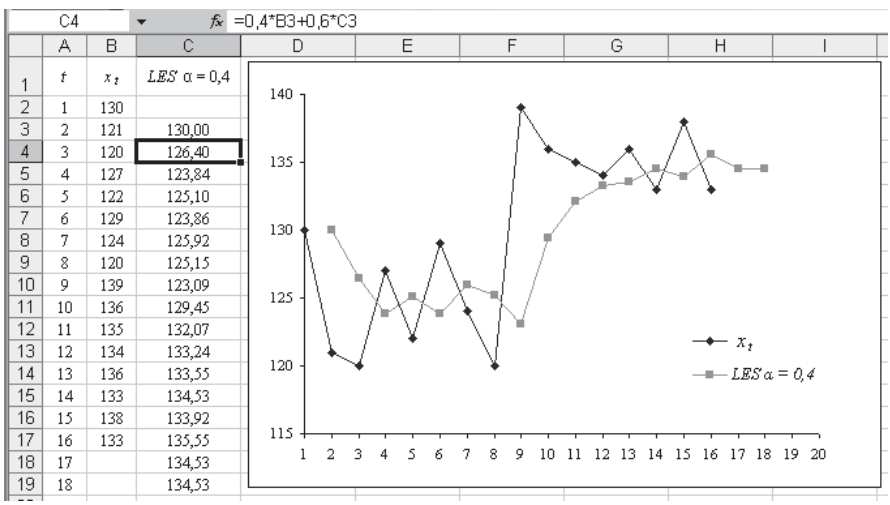


Figure 4.8 – Chronique du tableau 4.6 et série obtenue par LES avec $\alpha = 0,4$

Tableau 4.7 – Présentation des calculs du LES avec les critères calculés sur le dernier tiers de la série

		$\alpha = 0,4$				$\alpha = 0,5$			
t	x_t	*LES*	e_t	*ABS* (e_t)	$(e_t)^2$	*LES*	e_t	*ABS* (e_t)	$(e_t)^2$
1	130								
2	121	130,00	– 9,00	9,00	81,00	130,00	– 9,00	9,00	81,00
3	120	126,40	– 6,40	6,40	40,96	125,50	– 5,50	5,50	30,25
4	127	123,84	3,16	3,16	9,99	122,75	4,25	4,25	18,06
5	122	125,10	– 3,10	3,10	9,63	124,88	– 2,88	2,88	8,27
6	129	123,86	5,14	5,14	26,39	123,44	5,56	5,56	30,94
7	124	125,92	– 1,92	1,92	3,68	126,22	– 2,22	2,22	4,92
8	120	125,15	– 5,15	5,15	26,53	125,11	– 5,11	5,11	26,11
9	139	123,09	15,91	15,91	253,12	122,55	16,45	16,45	270,45
10	136	129,45	6,55	6,55	42,85	130,78	5,22	5,22	27,28
11	135	132,07	2,93	2,93	8,57	133,39	1,61	1,61	2,60
12	134	133,24	0,76	0,76	0,57	134,19	– 0,19	0,19	0,04
13	136	133,55	2,45	2,45	6,02	134,10	1,90	1,90	3,62
14	133	134,53	– 1,53	1,53	2,33	135,05	– 2,05	2,05	4,20
15	138	133,92	4,08	4,08	16,67	134,02	3,98	3,98	15,81
16	133	135,55	– 2,55	2,55	6,50	136,01	– 3,01	3,01	9,07
17		134,53				134,51			
			EM = 0,64	*EAM* = 2,27	*EQM* = 6,42		*EM* = 0,12	*EAM* = 2,23	*EQM* = 6,55

Le tableau 4.8 donne, selon la constante de lissage α variant par pas de 0,1, les valeurs des critères *EM*, *EQM* et *EAM* pour le *LES* appliquées à la série de la figure 4.8, ces critères ayant été calculés sur le dernier tiers de la série, c'est-à-dire avec les cinq dernières erreurs de prévision.

Le critère *EQM* est minimum pour $\alpha = 0,4$, le critère *EAM* pour $\alpha = 0,5$ et la valeur absolue de l'erreur moyenne est minimum pour $\alpha = 0,5$

Tableau 4.8 – Valeurs des critères calculés sur le dernier tiers de la série du tableau 4.7

Valeur de α	*EM*	*EQM*	*EAM*
0,1	4,548	25,311	4,548
0,2	2,931	14,068	3,101
0,3	1,545	8,151	2,495
0,4	0,643	6,421	2,274
0,5	0,125	6,547	2,227
0,6	– 0,148	7,361	2,449
0,7	– 0,280	8,436	2,648
0,8	– 0,339	9,670	2,833
0,9	– 0,369	11,095	3,012

B. Le lissage exponentiel double

Le lissage e xponentiel double est une généralisation du lissage e xponentiel simple au cas d'une ***chronique à tendance localement linéaire*** ; on suppose que la série peut être ajustée par une droite au voisinage de T :

$$x_t = a_1(T) + a_2(T) \cdot (t - T)$$

Les coefficients $a_1(T)$ et $a_2(T)$ sont choisis de façon à minimiser la quantité suivante qui est l'analogue de la quantité (3) minimisée pour le lissage exponentiel simple :

$$\sum_{i=0}^{T-1} (1-\alpha)^i \left(x_{T-i} - \Big(a_1(T) + a_2(T) \cdot (-i) \Big) \right)^2$$

On obtient la solution suivante :

$$\begin{cases} \hat{a}_1(T) = 2S_1(T) - S_2(T) \\ \hat{a}_2(T) = \dfrac{\alpha}{1-\alpha} \cdot \Big(S_1(T) - S_2(T) \Big) \end{cases} \text{avec} \begin{cases} S_1(T) = \alpha \displaystyle\sum_{i=0}^{T-1} (1-\alpha)^i \cdot x_{T-i} \\ S_2(T) = \alpha \displaystyle\sum_{i=0}^{T-1} (1-\alpha)^i \cdot S_1(T-i) \end{cases}$$

ce qui conduit à la prévision : $\hat{x}_T(h) = \hat{a}_1(T) + \hat{a}_2(T) \cdot h$

La quantité $S_1(T)$ résultant du lissage exponentiel simple de la série $\{x_t,\ t = 1,\ldots,T\}$ et la quantité $S_2(T)$ du lissage exponentiel simple de la série $\{S_1(t),\ t = 1,\ldots,T\}$ – d'où le nom de lissage exponentiel double, on dispose pour leurs calculs des formules de mise à jour du *LES* :

$$\begin{cases} S_1(T) = \alpha \cdot x_T + (1-\alpha) \cdot S_1(T-1) \\ S_2(T) = \alpha \cdot S_1(T) + (1-\alpha) \cdot S_2(T-1) \end{cases}$$

L'initialisation de ces formules de mise à jour peut être :

$$\begin{cases} S_1(1) = x_1 \\ S_2(2) = S_1(2) \end{cases}$$

En développant les égalités ci-dessus, on obtient les formules de mise à jour des coefficients $\hat{a}_1(T)$ et $\hat{a}_2(T)$:

$$\begin{cases} \hat{a}_1(T) = \hat{a}_1(T-1) + \hat{a}_2(T-1) + (1-(1-\alpha)^2) \cdot \left(x_T - \hat{x}_{T-1}(1) \right) \\ \qquad\quad = x_T - (1-\alpha)^2 \cdot \left(x_T - \hat{x}_{T-1}(1) \right) \\ \hat{a}_2(T) = \hat{a}_2(T-1) + \alpha^2 \cdot \left(x_T - \hat{x}_{T-1}(1) \right) \end{cases}$$

L'initialisation de ces formules peut être : $\begin{cases} \hat{a}_1(2) = x_2 \\ \hat{a}_2(2) = x_2 - x_1 \end{cases}$

Comme pour le lissage exponentiel simple, le choix de la constante de lissage α peut se faire par la minimisation d'un critère choisi.

La méthode de Holt-Winters étend les méthodes de lissage exponentiel aux séries saisonnières. C'est une méthode de prévision très utilisée.

Testez-vous *(les réponses sont données page 286)*

Il y a *au moins* une réponse exacte par question.

1. Pour une chronique à 12 termes :

a) on peut calculer 8 moyennes mobiles centrées de longueur 4

b) on peut calculer une moyenne mobile centrée de longueur 12

c) on peut calculer 10 médianes mobiles centrées de longueur 3

d) on peut calculer 2 moyennes mobiles centrées de longueur 11

2. Identification du modèle de décomposition adapté :

a) si le facteur saisonnier est proportionnel à la tendance, on choisit le modèle additif

b) si les deux courbes joignant respectivement les maxima et les minima sont quasi-parallèles sur un graphique à ordonnée logarithmique, on choisit le modèle multiplicatif

c) si les maxima de la courbe représentative de la chronique sont distants de 5 dates, on choisit le modèle additif

d) on peut toujours ramener un modèle multiplicatif à un modèle additif

3. Si une chronique X a une composante saisonnière de période p, alors :

a) les moyennes mobiles centrées de longueur 2 p éliminent la saisonnalité

b) on peut approximer la tendance par la suite des moyennes mobiles centrées de longueur p

c) la somme de p termes successifs de X donne une approximation de la moyenne de la tendance

d) on peut toujours calculer $(T - p)$ moyennes mobiles centrées de longueur p si elle a T termes

4. Une prévision par lissage exponentiel simple :

a) tient d'autant plus compte des valeurs récentes de la série que la constante α est faible

b) peut s'envisager pour une chronique possédant une composante saisonnière

c) ne peut pas s'envisager pour une chronique possédant une tendance à la hausse

d) s'adapte d'autant plus rapidement à un changement de niveau de la chronique que α est élevée

Exercices *(corrigés page 309)*

Exercice 4.1

On dispose aussi de la répartition mensuelle du niveau de l'indice de la qualité de l'air ATMO dans l'agglomération parisienne selon trois classes de niveau pour les six années agrégées.

Fréquences mensuelles d'apparition des indices de 1998 à 2003

Niveau	1 à 4	5 à 7	8 à 10	Nombre total de jours
Janvier	164	22	0	186
Février	136	29	4	169
Mars	151	35	0	186
Avril	152	28	0	180
Mai	132	54	0	186
Juin	115	65	0	180
Juillet	123	59	4	186
Août	93	83	10	186
Septembre	155	25	0	180
Octobre	155	31	0	186
Novembre	172	8	0	180
Décembre	177	9	0	186
Nombre total de jours	1725	448	18	2191

Légende : Niveau 1 à 4 : très bon à bon.
Niveau 5 à 7 : moyen à médiocre.
Niveau 8 à 10 : mauvais à très mauvais.

On s'intéresse à la classe de niveau « 5 à 7 ».

1. Représentez graphiquement son évolution au cours des 12 mois.

2. Calculez la suite des moyennes mobiles de longueur 3 et représentez-la sur le même graphique. Quelle propriété de la moyenne mobile venez-vous d'illustrer ?

Exercice 4.2

t	1	2	3	4	5	6	7	8	9	10	11	12
x_t	3	–	1	5	1	3	–	15	1	3	–	1

1. Calculez les suites des moyennes mobiles de longueurs 2, 3, 4 et 5.

Quelles sont les propriétés de la moyenne mobile qui sont illustrées par cet exemple ?

2. Soit la chronique $z_t = 10 - 2t + x_t$, calculez la suite des moyennes mobiles de longueur 4 de la nouvelle série z_t

Exercice 4.3

Le tableau suivant donne la série chronologique bimestrielle du transport des voyageurs sur le réseau Air France International (en milliards de passagers-km) de 2002 à 2005.

	Janv.-Fév	Mars-Avril	Mai-Juin	Juil.-Août	Sept.-Oct.	Nov.-Déc.
2002	13,3	15,1	14,8	16,3	14,8	14,2
2003	13,8	14,2	14,1	17,0	15,2	14,8
2004	14,4	16,0	16,2	18,5	16,2	15,3
2005	15,4	16, 8	17,4	19,9	17,9	17,4

Source : www.insee.fr

1. On choisit de modéliser cette chronique par un schéma additif. Justifiez ce choix.

2. Déterminez la tendance de cette chronique par la suite des moyennes mobiles de longueur adaptée, et représentez-la sur le même graphique que la série initiale.

3. Calculez les coefficients saisonniers.

4. Calculez la série corrigée des variations saisonnières. Ajustez cette chronique par une droite en utilisant la méthode des moindres carrés.

5. Au vu des résultats, quelles prévisions pouvait-on faire fin 2005 pour janvier-février, mars-avril et mai-juin 2006 ?

6. Sachant qu'on a observé 17,2 milliards de passagers-km en janvier-février 2006, 18,5 en mars-avril et 18,6 en mai-juin, calculez l'erreur absolue moyenne de prévision.

Exercice 4.4

1. Voici pour ses trois premiers mois d'ouverture, le nombre de places x_t vendues par semaine par le cinéma PARADISO (t désignant le numéro de la semaine varie de 1 à 12) :

t	1	2	3	4	5	6	7	8	9	10	11	12
x_t	3 428	3 295	3 376	3 195	3 573	3 334	3 434	3 300	3 703	3 411	3 545	3 327

1. Représentez cette chronique graphiquement. A-t-elle une composante saisonnière ? Si oui, de quelle période ?

2. Calculez la suite des moyennes mobiles de longueur appropriée pour évaluer la tendance de la série chronologique. Représentez cette suite sur le graphique précédent.

3. On choisit un modèle multiplicatif. Évaluez les coefficients saisonniers.

4. Calculez la série corrigée des variations saisonnières (série *CVS*) et représentez-la sur le graphique précédent. Calculez la série des résidus.

5. Ajustez la série *CVS* par une droite en utilisant la méthode des moindres carrés. Représentez cette droite sur le graphique précédent.

6. Donnez une prévision pour le nombre de places vendues pendant les deux premières semaines du quatrième mois.

Exercice 4.5

La demande d'un certain article a été relevée au cours de 15 mois consécutifs :

Mois	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Demande	37	41	39	40	42	39	41	39	42	41	40	42	43	40	42

1. Appliquez un lissage exponentiel simple à cette série chronologique en prenant $\alpha = 0{,}6$ jusqu'au 6e mois inclus et $\alpha = 0{,}3$ pour les mois suivants. Tracez sur le même graphique la chronique initiale et la série lissée.

2. Justifiez le changement de valeur de la constante de lissage α.

3. Calculez l'erreur moyenne, l'erreur absolue moyenne et l'erreur quadratique moyenne.

4. Donnez les prévisions de la demande pour les trois mois suivants.

Exercice 4.6

Le tableau ci-dessous donne les valeurs des indices trimestriels (base 2000) de la production industrielle des boissons pour les années 2002 à 2005 :

	1er trimestre	2e trimestre	3e trimestre	4e trimestre
2002	94,2	105,3	103,5	127,5
2003	98,3	103,8	115,7	125,5
2004	100,9	110,7	106,6	126,3
2005	98,7	109,8	110,8	129,4

Source : www.insee.fr

1. Commentez l'évolution de cette série chronologique en utilisant sa représentation graphique. Justifiez le recours à un schéma de composition additif.

2. Déterminez la tendance de cette chronique par la suite des moyennes mobiles de longueur adaptée, et représentez-la sur le même graphique que la série initiale.

3. Calculez les coefficients saisonniers et la série corrigée des variations saisonnières.

4. Appliquez un lissage exponentiel simple à la série *CVS* avec $\alpha = 0{,}3$.

5. Quelles prévisions pouvait-on faire au dernier trimestre 2005 pour les deux premiers trimestres 2006 ? Sachant que cet indice a pris les valeurs 100,8 et 110,8 pour les 1 er et 2 e trimestres 2006, calculez l'erreur moyenne et l'erreur absolue moyenne de prévision.

5. Modèle probabiliste et variable aléatoire

Il ne faut donc pas se demander si nous percevons vraiment un monde, il faut se dire au contraire : le monde est cela que nous percevons.

Phénoménologie de la perception, Maurice Merleau-Ponty (1908-1961)

La statistique descriptive permet de résumer les mesures d'une ou plusieurs grandeurs obtenues sur les individus d'un échantillon ou d'une population par un classement (tri simple dans le cas d'une seule variable, tri croisé dans le cas de plusieurs variables). Une grandeur est alors décrite par sa distribution, qui est déterminée à partir des observations, justifiant ainsi le nom de *distribution empirique* de la grandeur (ou de la variable). C'est la représentation « de base » pour apprécier une grandeur après qu'elle a été classée.

L'observation de nombreuses distributions empiriques montre pour certaines d'entre elles des analogies de formes, et des caractéristiques voisines. Ceci conduit à définir des *distributions théoriques* afin de disposer d'instruments plus formels regroupant les propriétés. Ces distributions théoriques sont une abstraction destinée non pas simplement à présenter les données, mais à les interpréter ou à les expliquer. Ce parallélisme entre l'observation et la représentation théorique se retrouve également au niveau de l'observation individuelle qu'on replace dans un ensemble potentiel d'observations supposées homogènes. Les variations entre différentes observations sont considérées comme des fluctuations non attribuables à une cause identifiée (contrôlable ou non), et on dit alors qu'elles sont le résultat du *hasard*.

Il est nécessaire de disposer d'un outil théorique permettant de considérer globalement les différentes observations provenant d'une même population en tenant compte d'une part, de l'homogénéité liée à leur origine commune et d'autre part, des fluctuations entre observations. C'est le concept de *variable aléatoire* qui remplit ce double rôle. Son intérêt dépend des propriétés générales qu'on pourra lui associer, et de leur fiabilité.

Le *calcul des probabilités* (puis son *axiomatique*) est le support formel de cette représentation. Il a été introduit initialement au XVII^e siècle pour

étudier les jeux de société (dés, cartes, roulette), et son développement a permis une approche intégrant les éléments fluctuants et non permanents des phénomènes physiques, sociaux ou encore psychologiques. Les probabilités se fondent sur la notion d' *expérience aléatoire*, c'est-à-dire dont les conditions d'exécution bien que parfaitement définies, ne permettent pas de s'assurer à priori de l'issue de l'expérience.
À contrario les expériences, dites *déterministes*, celles dont le résultat est maîtrisé par les conditions initiales, ont un seul résultat possible, en négligeant les éventuelles erreurs de mesure. Ces situations s'opposent à celles où le hasard intervient dans le résultat attendu et pour lesquelles on parle d' *incertitude*. C'est pour ce dernier contexte qu'a été fondé le calcul des probabilités. Sous le terme de hasard, on a longtemps résumé les facteurs considérés comme mineurs [1] dans l'étude d'un phénomène. On pourrait plutôt dire actuellement qu'entre déterminisme et incertitude repose toute la notion de ce qui échappe au contrôle, ou encore de l'information non disponible.
Le caractère aléatoire d'une grandeur peut être partie intrinsèque du phénomène étudié. C'est le cas du résultat d'un jet de dé, ou bien de la quantité de fuel consommé annuellement en France. Dans d'autres cas, il n'en est pas ainsi. Si on s'intéresse à la distance moyenne parcourue sur une autoroute par les automobilistes passant à un poste de péage donné, un certain jour, on peut interroger tous les automobilistes se présentant au péage et calculer la moyenne. On peut aussi chercher cette information en interrogeant un échantillon d'automobilistes se présentant à ce péage. La valeur moyenne observée sur cet échantillon va dépendre de l'échantillon retenu qui n'est pas lui-même fixé à l'avance (il y a beaucoup d'échantillons possibles), et peut être considéré comme le résultat d'une expérience aléatoire (le choix de l'échantillon d'automobilistes). Ainsi, alors qu'initialement le problème se posait en termes déterministes, la procédure surajoutée de choix de l'échantillon introduit un élément aléatoire. La grandeur étudiée (moyenne) n'est pas elle-même aléatoire, mais les données recueillies sur l'échantillon le sont puisque le contenu de l'échantillon n'est pas déterminé par son mode de tirage.
L'objectif du calcul des probabilités est l'analyse et l'explication des phénomènes non déterministes. Ses fondements théoriques, et en particulier l'*axiomatique de Kolmogorov*, lui donnent une valeur scientifique relativisée toutefois par la signification de la notion de probabilité.

1. Historiquement, ces facteurs « mineurs » ont été d'abord restreints à la notion d'erreur.

I. Éléments de calcul des probabilités

Le calcul des probabilités est basé, comme nous l'avons déjà vu, sur la notion d'expérience aléatoire. On associe à une telle expérience $\mathscr{E}$ l'ensemble de tous les résultats possibles, appelé *ensemble fondamental*, qu'on désigne communément par Ω. Chaque résultat possible est une partie de Ω.

L'ensemble fondamental associé à une épreuve non déterministe peut contenir un nombre fini d'éléments (de la forme $\{\omega_1, \omega_2, \ldots, \omega_n\}$) ou bien être de type infini dénombrable[1] ou enfin être de type infini non dénombrable. On assimile, comme dans la théorie classique des ensembles, un événement, donc une partie de Ω, à la propriété qui le caractérise au sein de l'ensemble fondamental, c'est-à-dire à la réalisation de cette propriété. On désigne par ω le résultat qui sera observé à l'issue de l'expérience aléatoire, et on écrit $\omega \in G$ (avec $G \subset \mathbb{P}(\Omega)$) lorsque la situation décrite par G est le résultat de la réalisation de l'épreuve E. La non-réalisation de l'événement G est notée $\omega \in \overline{G}$ (où $\overline{G}$ désigne l'ensemble complémentaire[2] de G dans Ω).

La notation et le vocabulaire ensembliste sont tout à fait adaptés à la description des situations aléatoires, et l'événement dont on a la certitude de la réalisation est désigné par Ω (événement certain), tandis que l'événement dont on sait qu'il ne se produira pas est désigné par $\varnothing$ (événement impossible). La réunion ensembliste $G \cup H$ correspond à la réalisation d'au moins un des événements G et H. L'intersection ensembliste $G \cap H$ correspond à la réalisation des deux événements G et H. L'incompatibilité de G et H se traduit par $G \cap H = \varnothing$. Enfin, la différence ensembliste $G - H$ correspond à la réalisation de G et à la non-réalisation de H, ou encore à l'intersection $G \cap \overline{H}$.

La représentation ensembliste justifie la caractérisation des issues d'une épreuve aléatoire au sein d'une structure mathématique sur laquelle on pourra définir une probabilité. Cette structure est celle d'une algèbre de Boole (cas fini), ou d'une σ-algèbre (cas infini).

A. Notion de probabilité

Soit Ω l'ensemble fondamental associé à une épreuve aléatoire et $\mathscr{P}(\Omega)$ l'ensemble des parties de Ω.

1. C'est-à-dire pouvant être mis en bijection avec tout ou partie de l'ensemble $\mathbb{N}$ des entiers naturels.
2. Le complémentaire de l'ensemble G est noté indifféremment $\overline{G}$ ou G^c.

On dit que $\mathcal{I} \subset \mathcal{P}(\Omega)$ est une *algèbre de Boole* si les deux conditions suivantes sont vérifiées :

• C1 : $G \in \mathcal{I} \Rightarrow \overline{G} \in \mathcal{I}$

• C2 : $G \in \mathcal{I}$ et $H \in \mathcal{I} \Rightarrow G \cup H \in \mathcal{I}$

Cette structure d'algèbre de Boole correspond à la traduction ensembliste de la logique des événements dans le cas où l'ensemble fondamental est fini.

On dit que $\mathcal{A} \subset \mathcal{P}$ (Ω) est une σ-*algèbre de Boole* (ou plus simplement une σ-algèbre) si les deux conditions suivantes sont vérifiées :

• C1 : $G \in \mathcal{A} \Rightarrow \overline{G} \in \mathcal{A}$

• C2 : $G_i \in \mathcal{A}$ pour $i \in \mathbb{N} \Rightarrow \bigcup_{i \in \mathbb{N}} G_i \in \mathcal{A}$

• La notion de σ-algèbre étend la correspondance entre la logique événementielle et la théorie des ensembles au cas d'épreuves aléatoires dont l'ensemble fondamental est infini.

On notera que si E est un événement quelconque de Ω, $\mathcal{A} = \{\varnothing, \Omega, E, \overline{E}$ est une σ-algèbre.

Lorsqu'on a défini une σ-algèbre $\mathcal{A}$ d'événements de Ω, on dit que le couple (Ω, $\mathcal{A}$) est un espace probabilisable dans le sens où il ne reste plus qu'à préciser la probabilité de chaque événement de $\mathcal{A}$.

Les premières fondations de la notion de probabilité[1] visaient à définir une échelle ordonnée des chances de réussite à certains jeux. L'approche fréquentiste qui en a résulté est construite sur l'observation et le dénombrement de situations dites élémentaires, c'est-à-dire représentant toutes les issues différentes de l'épreuve aléatoire. Ce point de vue ne peut s'appliquer qu'à des cas où l'ensemble fondamental associé est fini.

On suppose tout d'abord que les événements élémentaires ont une chance égale de réalisation, contexte dit d' *équiprobabilité*, ce qui implique l'additivité des chances. Pour cette hypothèse et pour un ensemble fondamental de type fini, on définit la probabilité d'un événement comme l'analogue d'une fréquence relative afin d'avoir une échelle de valeurs comprise entre 0 et 1, et de disposer d'une mesure additive : pour des événements élémentaires équiprobables, la probabilité d'un événement quelconque $\mathcal{A}$ est sa fréquence relative d'apparition dans l'ensemble fondamental. Ce point est connu sous le nom de règle de Laplace.

1. Blaise Pascal et Pierre de Fermat correspondent en 1654, sur la répartition équitable des enjeux dans les jeux de hasard ; et Christian Huyghens, en 1657, formule et résout le problème dit de la *ruine du joueur*.

Donnons-en un exemple d'application, en calculant la probabilité de faire apparaître les chiffres 4, 2 et 1 en lançant trois dés parfaitement équilibrés. Chaque dé possède 6 faces, ce qui implique que l'ensemble fondamental Ω possède $6^3 = 216$ éléments. Dans cet ensemble fondamental formé des figures à 3 chiffres entre 1 et 6, celles qui permettent de reconstituer « 421 » sont toutes les 3 ! = 6 permutations des trois chiffres 1, 2 et 4. La fréquence relative de la figure « 421 » ou probabilité de l'événement « obtention des chiffres 4, 2, 1 » est égale à 1/36

Dès lors que l'ensemble fondamental n'est pas fini et/ou que l'équiprobabilité n'est pas assurée sur les événements élémentaires de Ω, on ne peut plus appliquer cette règle du nombre de cas favorables sur nombre de cas possibles. On doit généraliser cette démarche et définir abstraitement la probabilité pour qu'elle coïncide avec la règle de Laplace lorsque cette dernière s'applique. On utilise alors la représentation ensembliste des événements pour définir une probabilité sur un espace probabilisable (Ω, $\mathcal{A}$).

Soit (Ω, $\mathcal{A}$) un espace probabilisable. Une probabilité[1] P sur cet espace est une application de $\mathcal{A}$ à valeurs dans l'intervalle [0;1] vérifiant :
i) $P(\Omega) = 1$
ii) pour des événements $\{G_i \in \mathcal{A}, i \in \mathbb{N}\}$ incompatibles ($i \neq j \Rightarrow G_i \cap G_j = \varnothing$) :

$$P\left(\bigcup_{i \in \mathbb{N}} G_i\right) = \sum_{i \in \mathbb{N}} P(G_i)$$

On dit alors que le triplet (Ω, $\mathcal{A}$, P) est un *espace probabilisé*. Il est construit sur une épreuve aléatoire dont on se donne l'ensemble fondamental Ω, tous les événements simples ou complexes étant décrits par $\mathcal{A}$, sur laquelle on se donne l'échelle des chances P.

De cette définition, ou axiomatique de Kolmogorov, on déduit les propriétés suivantes :

1. Si $G \in \mathcal{A}$, alors $P(\overline{G}) = 1 - P(G)$

En effet, on a : $G \cap \overline{G} = \varnothing$ et $G \cup \overline{G} = \Omega$, ce qui donne :

$$P(\Omega) = 1 = P(G \cup \overline{G}) = P(G) + P(\overline{G})$$

2. La probabilité de l'événement impossible est nulle : $P(\varnothing) = 0$

Il suffit d'appliquer la propriété précédente en posant $G = \Omega$

1. On dit encore une mesure de probabilité pour bien faire référence aux qualités métrologiques de cette application. On désignera indifféremment par la suite la probabilité par Pr ou par P.

3. Si $G \in \mathcal{A}$ et $H \in \mathcal{A}$ sont tels que $G \subset H$, alors $P(G) \leq P(H)$

Puisque $H = G \cup (\overline{G} \cap H)$ et que $G \cap (\overline{G} \cap H) = \varnothing$, on a :

$P(H) = P(G) + P(\overline{G} \cap H)$, et $P(\overline{G} \cap H) \geq 0$ implique $P(H) \geq P(G)$

Il est important de noter que l'inégalité entre les probabilités est au sens large.

4. Si G et H sont deux éléments quelconques de $\mathcal{A}$:

$$P(G \cup H) = P(G) + P(H) - P(G \cap H)$$

En effet, on a : $G \cup H = G \cup (\overline{G} \cap H)$ avec $G \cap (\overline{G} \cap H) = \varnothing$

donc $P(G \cup H) = P(G) + P(\overline{G} \cap H)$

De même $H = (G \cap H) \cup (\overline{G} \cap H)$ avec $(G \cap H) \cap (\overline{G} \cap H) = \varnothing$

donc $P(H) = P(G \cap H) + P(\overline{G} \cap H)$

En combinant les deux résultats, on obtient :

$$P(G \cup H) = P(G) + P(H) - P(G \cap H)$$

Ce dernier résultat est connu sous le nom de *théorème des probabilités totales*.

B. Probabilités conditionnelles

Nous avons évoqué en introduction de ce chapitre le lien particulier entre l'information disponible, le contrôle des facteurs déterminants d'un phénomène et l'importance de sa partie aléatoire, donc de sa probabilité de réalisation. Nous allons retrouver ceci au travers de la notion de *probabilité conditionnelle*.

Soit une épreuve aléatoire donnée, munie de son ensemble fondamental Ω, de la σ-algèbre des événements, et de la probabilité P associée à chacun de ces derniers, en d'autres termes, nous supposons donné un espace probabilisé $(\Omega, \mathcal{A}, P)$. La connaissance d'une information complémentaire sur le déroulement de l'épreuve équivaut à la modification des probabilités définies sur les éléments de $\mathcal{A}$. En effet, cette information acquise n'est autre qu'une condition désormais supposée réalisée quel que soit le résultat de l'expérience aléatoire. Prenons-en un exemple. Nous avons vu que la probabilité de réaliser la « figure 421 » lors du jet de 3 dés était de 1/36. Supposons maintenant que le premier dé soit lancé avant les deux autres, et qu'il fasse apparaître le chiffre 2. L'ensemble fondamental associé au jet des 2 dés restant contient 36 événements élémentaires, mais parmi ceux-ci, seuls les couples (4 ;1) et (1 ; 4) permettent de compléter la configuration « 421 ». On en déduit donc que si on sait que le premier dé a affiché la valeur 2, la probabilité de réaliser un 421 est de 1/18.

On remarque dans cet exemple que l'ensemble fondamental a été modifié, et donc aussi la σ-algèbre des événements, ainsi que la mesure de probabilité P.

Cette modification s'appelle un conditionnement, car elle correspond à la prise en compte d'une condition supplémentaire sur la réalisation de l'épreuve aléatoire (ici le fait que le premier dé doit être lancé séparément et qu'il affichera la valeur 2). On est ainsi conduit à définir les probabilités conditionnelles.

Définition 1

Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé et soit $C \in \mathcal{A}$ un événement particulier, appelé *condition*, de probabilité non nulle. Pour tout événement $A \in \mathcal{A}$, on appelle *probabilité conditionnelle de A sachant C*, notée $P(A|C)$, la quantité :

$$P(A|C) = \frac{P(A \cap C)}{P(C)}$$

Cette définition est bien évidemment dérivée de l'approche fréquentiste des probabilités puisqu'en raisonnant avec la règle de Laplace, on pourrait dire que les cas favorables sont ceux où les événements A et C sont tous deux réalisés, alors que les cas possibles sont ceux pour lesquels de toutes façons l'événement C est observé. Il faut noter qu'une probabilité conditionnelle n'a de sens que si la condition est réalisable (de probabilité non nulle). La notion de probabilité conditionnelle, ou encore de conditionnement des probabilités, revient à modifier l'ensemble fondamental puisque l'événement $C \in \mathcal{A}$ se trouve être rapporté à une probabilité égale à un. Ainsi, sur la figure 5.1, par conditionnement la probabilité de A devient ramenée à la seule part de A incluse dans C.

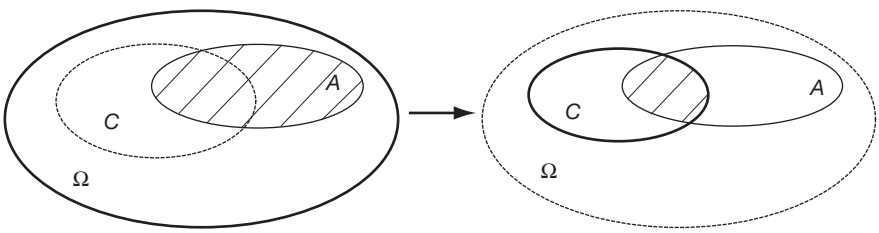


Figure 5.1 – Conditionnement d'une probabilité

On peut vérifier que l'application qui à tout $A \in \mathcal{A}$ associe $P(A|C)$ est bien une mesure de probabilité.

Propriété

Si $A_1, A_2, \ldots, A_n$ sont n événements quelconques d'une σ-algèbre $\mathcal{A}$ d'un espace probabilisé $(\Omega, \mathcal{A}, P)$, on peut écrire :

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) =$$
$$P(A_1) \cdot P(A_2|A_1) \cdot \ldots \cdot P(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1})$$

En effet, si $n = 2$, cette propriété n'est autre que la formule de définition des probabilités conditionnelles. Supposons cette propriété vraie à l'ordre $n - 1$:

$$P(A_1 \cap A_2 \cap \ldots \cap A_{n-1}) = P(A_1) \cdot P(A_2|A_1) \cdot \ldots \cdot P(A_{n-1}|A_1 \cap A_2 \cap \ldots \cap A_{n-2})$$

et montrons qu'elle est encore vraie à l'ordre n.

On peut écrire $A_1 \cap A_2 \ldots \cap A_n = (A_1 \cap A_2 \ldots \cap A_{n-1}) \cap A_n$

On pose : $B = A_1 \cap A_2 \cap \ldots \cap A_{n-1}$ et on obtient :

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_n \cap B) = P(A_n|B) \cdot P(B)$$

soit :

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \cdot P(A_1) \cdot P(A_2|A_1) \cdot \ldots \cdot P(A_{n-1}|A_1 \cap A_2 \cap \ldots \cap A_{n-2})$$

Définition 2

Deux événements A et B d'un espace probabilisé (Ω, $\mathcal{A}$, P) sont dits indépendants en probabilité si la réalisation de l'un d'eux ne modifie pas la probabilité de survenue de l'autre.

Il s'agit d'une relation symétrique. On parle également d'événements stochastiquement indépendants. Dans la suite de ce livre, on écrira toutefois simplement événements indépendants.

On voit alors que si A et B sont deux événements indépendants, on a :

$$P(A|B) = P(A)$$

$$P(B|A) = P(B)$$

et chacune de ces égalités montre que :

$$A \text{ et } B \text{ indépendants} \Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$$

D'autre part, il est important de ne pas confondre les notions d'indépendance et d'incompatibilité. Dans le premier cas, si les deux événements A et B sont de probabilité non nulles, alors la probabilité $P(A \cap B)$ est aussi non nulle. Dans le second cas, même si A et B sont de probabilité non nulles, l'intersection $(A \cap B)$ est de probabilité nulle. Il s'ensuit que deux événements à la fois indépendants et incompatibles sont tels qu'au moins l'un d'eux est un événement impossible (c'est-à-dire de probabilité nulle).

Notons encore que si A et B sont deux événements indépendants, alors :

$$P(A|B) = P(A|\overline{B}) = P(A)$$

$$P(B|A) = P(B|\overline{A}) = P(B)$$

Cette notion d'indépendance s'étend à plus de deux événements.

Définition 3

Soient n événements d'un espace probabilisé $(\Omega, \mathcal{A}, P)$. On dit qu'ils sont mutuellement indépendants si quels que soient $A_1, A_2, \ldots, A_k$ choisis parmi ces n événements, on a :

$$P(A_1 \cap A_2 \cap \ldots \cap A_k) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_k)$$

Des événements mutuellement indépendants sont indépendants deux à deux (simple application particulière de la définition), mais des événements $A_1, A_2, \ldots, A_n$ qui sont indépendants deux à deux ne sont pas toujours mutuellement indépendants.

Cette notion d'indépendance sera étendue plus loin au cas des variables aléatoires, et peut aussi être généralisée à plusieurs épreuves (ou expériences aléatoires).

La mise en œuvre des probabilités conditionnelles a conduit à une réflexion très importante sur le concept de probabilité lui-même, ce que nous verrons plus loin. C'est certainement l'apport de Thomas Bayes [1] qui en a représenté le point de départ. Nous donnerons donc d'abord le résultat connu sous le nom de théorème de Bayes, pour examiner ensuite le débat sur la notion de probabilité.

Théorème de Bayes

Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé, et soient $A_1, A_2, \ldots, A_n$ un ensemble d'événements deux à deux incompatibles vérifiant $\bigcup_{k=1}^{n} A_k = \Omega$ (on dit que les A_k forment un système complet d'événements). Pour tout événement B, on a alors : $P(A_i|B) = \dfrac{P(B|A_i) \cdot P(A_i)}{\sum_{k=1}^{n} P(B|A_k) \cdot P(A_k)}$ pour $i = 1, 2, \ldots, n$

En effet, on sait que :

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{P(B)}$$

et d'autre part que $B = B \cap \Omega = B \cap \left(\bigcup_{k=1}^{n} A_k \right) = \bigcup_{k=1}^{n} (B \cap A_k)$

1. Le révérend Thomas Bayes (1701-1761) est l'auteur de *An Essay Towards Solving a Problem in the Doctrine of Chances* qui ne fut publié qu'en 1763, après sa mort.

Les événements $B \cap A_k$ étant incompatibles deux à deux puisque les A_k le sont, on en déduit que :

$$P(B) = \sum_{k=1}^{n} P(B \cap A_k) = \sum_{k=1}^{n} P(B|A_k) \cdot P(A_k)$$

et la formule de Bayes est démontrée.

On dit fréquemment que les événements A_i, qui forment une partition de Ω, sont les causes. Une autre dénomination, très courante, consiste à nommer probabilités *a priori* les valeurs $P(A_k)$, et probabilités *a posteriori* les valeurs $P(A_k|B)$. En effet, la formule de Bayes permet d'obtenir une valeur révisée des probabilités des événements A_i une fois connue la réalisation de l'événement B.

On notera que l'application de la formule de Bayes demande l'évaluation des probabilités dites *a priori* $P(A_k)$ ainsi que des probabilités $P(B|A_k)$ de l'effet B connaissant chacune des causes.

▶ **Exemple**

Pour un système de crédit à la clientèle on distingue trois types de dossiers : les dossiers aboutissant en contentieux, les dossiers à difficultés temporaires ou légères et les dossiers sans difficultés de paiement. On a évalué sur la base d'expériences antérieures les proportions respectives des trois catégories à 1/5, 3/10 et 1/2. D'autre part, on dispose pour chaque dossier d'un score d'appréciation global du client rapporté à l'une des deux modalités suivantes : élevé ou bas. Enfin, on sait que 90 % des dossiers en contentieux correspondaient à un score bas, que 60 % des dossiers à difficultés légères correspondaient à un score bas, et que 85 % des dossiers sans difficultés correspondaient à un score élevé. Si on tire un dossier au hasard pour lequel le score est bas, quelle est la probabilité qu'il ait abouti en contentieux ? (resp. qu'il n'ait donné lieu à aucune difficulté de paiement ? qu'il ait engendré des difficultés légères ?)

Les trois événements A_1 = « aboutir en contentieux », A_2 = « difficultés légères » et A_3 = « aucune difficulté » forment un système complet. On dispose des probabilités *a priori* :

$$P(A_1) = 0{,}2 \qquad P(A_2) = 0{,}3 \qquad P(A_3) = 0{,}5$$

ainsi que des probabilités conditionnelles pour les événements

B = « score bas » et $\overline{B}$ = « score élevé »

$$P(B|A_1) = 0{,}9 \qquad P(B|A_2) = 0{,}6 \qquad P(B|A_3) = 0{,}15$$

d'où :

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) \\ &= P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) + P(B|A_3) \cdot P(A_3) \\ &= 0{,}435 \end{aligned}$$

On en déduit :

$$P(A_1|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_1) \cdot P(A_1)}{P(B)} = \frac{0{,}9 \cdot 0{,}5}{0{,}435} = 0{,}414$$

ainsi que : $P(A_2|B) = 0{,}414$ et $P(A_3|B) = 0{,}172$

Ce calcul montre que l'information complémentaire « *le dossier étudié correspond à un score bas* » a permis une augmentation de la probabilité associée au de venir « *contentieux* » (0,414 au lieu de 0,2) et de la probabilité associée au de venir « *difficultés légères* » (0,414 au lieu de 0,3), et une forte diminution de la probabilité associée au de venir « aucune difficulté » (0,172 au lieu de 0,5).

On nomme aussi probabilités révisées, les probabilités *a posteriori* $P(A_k|B)$.

Le théorème de Bayes est à l'origine de nombreux développements formant ce qu'on a appelé la statistique bayésienne. Les domaines d'application sont très variés : gestion financière, prévisions, diagnostic,…

Cependant, son utilisation est très contestée, notamment en raison de la nécessité d'une évaluation *a priori*, subjective, de probabilités. De plus les « causes » A_k se trouvent affectées de probabilités, ce qui peut apparaître paradoxal si on se réfère à la notion déterministe de causalité. Pour ceux qui contestent la statistique bayésienne, un phénomène est, ou n'est pas, cause (éventuellement partielle) d'un autre, et ne saurait donc être muni d'une probabilité sur cette causalité [1].

De nombreuses difficultés persistent autour de la notion de probabilité, et particulièrement celle du choix des probabilités qu'on affecte aux événements rapportés à une épreuve aléatoire. L'analyse combinatoire et l'approche fréquentiste offrent une solution, dite « objectiviste ». Cependant, ce point de vue se heurte à :

- quelques contradictions logiques : le lien entre la probabilité et la fréquence relative, qui permet d'évaluer une probabilité, est à relativiser par la loi faible des grands nombres (*cf. infra*), donc par une probabilité ; on définit concrètement une probabilité en se basant sur une autre probabilité qui demande à être évaluée, et ainsi de suite… ;
- quelques paradoxes : le paradoxe de Bertrand [2] montre 3 solutions distinctes, 1/4, 1/3 et 1/2 (toutes par l'approche fréquentiste) au calcul de la probabilité que la longueur d'une corde d'un cercle soit supérieure au côté du triangle équilatéral inscrit dans ce cercle ; le paradoxe de St Peters-

1. On ne vise pas, dans ce livre, à prendre parti pour ou contre l'optique bayésienne, mais à donner au lecteur des éléments simples sur les arguments en présence. Le débat n'est pas encore clos !
2. Présenté en détail, par exemple, dans le livre de G. Saporta, pages 11, *op. cit*.

bourg[1] montre que la notion fréquentiste de probabilité (donc « objective ») peut parfois demander des appréciations complémentaires (utilité) très subjectives ;

• et quelques limites (comment évaluer les probabilités pour une épreuve qui n'est pas répétable ?).

L'analyse bayésienne cherche à contourner ces obstacles, surtout ceux liés à l'approche fréquentiste, mais au prix de valeurs subjectives pour certaines probabilités, ainsi que de quelques difficultés mathématiques de mise en œuvre.

II. Variables aléatoires à une dimension

A. Définitions

Étant donné un espace probabilisé $(\Omega, \mathcal{A}, P)$, une *variable aléatoire* (v.a. en abrégé) est une application X définie sur l'ensemble fondamental Ω et à valeurs réelles :

$$\begin{array}{ccccc} X : & \Omega & \rightarrow & \mathbb{R} \\ & \omega & & X(\omega) \end{array}$$

À tout événement élémentaire ω, l'application X associe une valeur numérique $X(\omega)$; c'est pourquoi on précise parfois en parlant de *variable aléatoire réelle*[2] nommée aussi « aléa » ou « aléa numérique ».

On observe que la terminologie utilisée peut paraître abusive, car X est une application, donc une fonction de Ω dans $\mathbb{R}$. Les variables aléatoires seront notées par des lettres majuscules telles que X, Y, Z,… pour les distinguer des valeurs qu'elles sont susceptibles de prendre (ou réalisations), généralement notées en minuscules.

1. Jacques et Pierre jouent avec une pièce. Pierre paie à Jacques 1 € si pile sort dès le premier jet, 2 € si pile sort seulement au 2e jet, 4 € s'il ne sort qu'au 3e jet et ainsi de suite en doublant la somme payée par Pierre à Jacques à chaque jet supplémentaire où pile n'est pas sorti. On cherche à savoir quelle somme Jacques devrait accepter de payer à Pierre pour jouer à ce jeu si l'on veut qu'il soit équilibré, c'est-à-dire que leurs espoirs de gain soient égaux; le paradoxe de cette situation provient du fait qu'on peut montrer que le prix alors à payer par Jacques devrait être infini. Ce paradoxe a longuement été étudié par Daniel et Nicolas Bernoulli, puis par Buffon, Laplace, Poisson entre autres ; il a contribué à dégager la notion d'utilité.

2. Il faut distinguer une variable aléatoire à laquelle est associée une loi, appelée aussi « distribution », de probabilité (théorique) d'une variable statistique quantitative à laquelle est associée une distribution statistique (observée), chapitre 1, § II.A.

▶ **Exemple**

On jette deux dés non pipés ; l'ensemble fondamental associé à cette expérience aléatoire est formé de 36 événements élémentaires équiprobables :

$$\Omega = (\{1,1\} ; \{1,2\} ; \{2,1\} ; \ldots ; \{6,6\})$$

Si on s'intéresse à la somme des points marqués par les deux dés, on définira sur cet espace probabilisé une v.a. X égale à cette somme ; l'ensemble de ses valeurs possibles est :

$$\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Pour obtenir la probabilité d'une valeur quelconque de X, il suffit de dénombrer les événements élémentaires de Ω qui réalisent cette valeur; ainsi :

$$\begin{aligned} P(X = 4) &= P(\{1,3\} \cup \{2,2\} \cup \{3,1\}) \\ &= P(\{1,3\}) + P(\{2,2\}) + P(\{3,1\}) = 3/36 = 1/12 \end{aligned}$$

On dit que la *variable aléatoire* X est :

• *discrète* finie si l'ensemble $X(\Omega)$ est fini, discrète infinie si l'ensemble $X(\Omega)$ est infini dénombrable,

• *continue* si l'ensemble $X(\Omega)$ est un intervalle de $\mathbb{R}$ non réduit à un point (ou une réunion d'intervalles de $\mathbb{R}$).

On retrouve une classification analogue à celle rencontrée pour les variables statistiques (chapitre 1), la notion de probabilité remplaçant la notion de fréquence ; la loi des grands nombres (§ V) permet d'établir un lien entre ces deux notions.

⟶ ***Remarque***

Pour une variable aléatoire *continue* X, il faut compléter la définition en ajoutant que l'image réciproque de tout intervalle $]-\infty, x]$ doit appartenir à la σ-algèbre $\mathcal{A}$:

$$\forall x \in \mathbb{R} \quad X^{-1}(]-\infty,x]) = (\omega \in (\Omega|X)(\omega) \leq x) \in \mathcal{A}$$

La probabilité étant définie sur la famille des parties de Ω formant une σ-algèbre, cette condition permet de déterminer la probabilité de tout intervalle de $\mathbb{R}$.

Notons que cette condition est générale puisqu'elle est réalisée pour les variables aléatoires discrètes ; pour ces variables aléatoires, l'image réciproque de tout intervalle de $\mathbb{R}$ est une partie de Ω à laquelle est associée une probabilité.

B. Loi de probabilité d'une variable aléatoire

1) Fonction de répartition d'une variable aléatoire

La fonction de répartition F_X (ou F) d'une variable aléatoire X à valeurs dans l'intervalle [0, 1] est définie par : $F_X(x) = P(X \leq x)$

Propriétés caractéristiques d'une fonction de répartition d'une variable aléatoire

F est une fonction de répartition si :

1. F est croissante (au sens large)
2. $\lim_{x \to -\infty} F(x) = 0$ et $\lim_{x \to +\infty} F(x) = 1$
3. F est continue à droite

Compte tenu de la propriété ii) d'une probabilité (§ I.A), on peut écrire pour tout $a, b \in \mathbb{R}$, $a < b$:

$$F(b) = F(a) + P(a < X \leq b) \Rightarrow P(a < X \leq b) = F(b) - F(a)$$

$\Rightarrow$ la probabilité pour que X appartienne à un intervalle de $\mathbb{R}$ pouvant se calculer à partir de sa fonction de répartition F_X, cette fonction caractérise la loi de X.

2) Loi de probabilité d'une variable aléatoire discrète

La fonction de répartition d'une telle v.a. est une fonction constante par intervalle (ou « en escalier ») continue à droite, les points de discontinuité correspondant à des valeurs possibles de X (*cf.* figure 5.2) ; sa courbe représentative s'appelle la *courbe de répartition* ou *courbe cumulative* ; on peut remarquer que cette fonction présente une identité formelle avec la fonction de répartition d'une variable statistique discrète.

Considérons le cas d'une v.a. *X discrète finie* ; ses différentes valeurs possibles, en nombre fini, sont supposées distinctes et rangées dans l'ordre croissant :

$$X(\Omega) = \{x_1, \ldots, x_i, \ldots, x_k\}$$

Connaissant la fonction de répartition de X, on peut calculer la probabilité p_i de réalisation de toute valeur $x_i (1 \leq i \leq k)$:

$$p_i = P(X = x_i) = \begin{cases} F(x_1) & \text{pour } i = 1 \\ F(x_i) - F(x_{i-1}) & \text{pour } i = 2, \ldots, k \end{cases}$$

Une telle distribution de probabilité peut se représenter par un diagramme en bâtons (*cf.* figure 5.3).

Valeur de X	x_1	…	x_i	…	x_k
Probabilité	p_1	…	p_i	…	p_k

$$\sum_{i=1}^{k} p_i = 1$$

Exemple 1

Loi de probabilité de la v.a. discrète finie X égale à la somme des points marqués lors du lancer de deux dés non pipés :

Valeur de X	2	3	4	5	6	7	8	9	10	11	12
Probabilité	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

Inversement, on calcule aisément la fonction de répartition à partir de la connaissance des k couples (x_i , p_i) :

$$F(x) = \begin{cases} 0 & \text{si } x < x_1 \\ \sum_{j=1}^{i} p_j & \text{si } x_i \le x < x_{i+1} \quad \text{pour } (1 \le i \le k-1) \\ 1 & \text{si } x \ge x_k \end{cases}$$

Lorsque la v .a. est *discrète infinie* , l'ensemble $X(\Omega)$ est infini dénombrable, et on peut, comme dans le cas fini, calculer les probabilités de chaque v aleur possible à partir de la fonction de répartition ; en sens inverse, on peut déduire la fonction de répartition de la connaissance des valeurs possibles et des probabilités associées.

Exemple 2

Loi de probabilité de la v .a. discrète infinie X égale au nombre de jets nécessaires d'une pièce de monnaie non pipée pour obtenir la f ace « pile » :

Valeur de X	1	2	3	…	i	…
Probabilité	$\frac{1}{2}$	$\frac{1}{2^2}$	$\frac{1}{2^3}$	…	$\frac{1}{2^i}$	…

$$\sum_{i=1}^{\infty} \frac{1}{2^i} = 1$$

On verra au chapitre 6 (§ II.C) que X suit une loi géométrique de paramètre 0,5

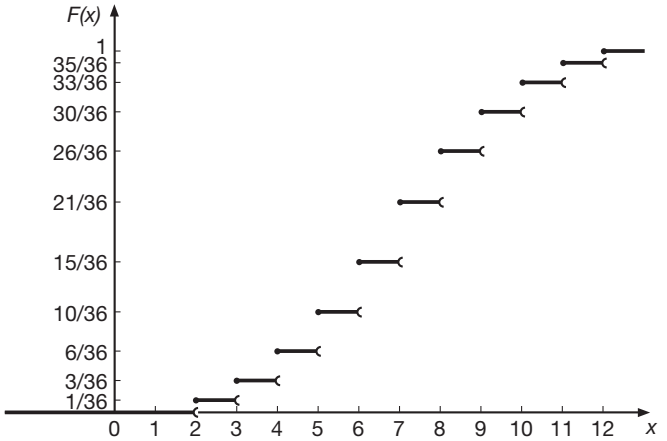


Figure 5.2 – Fonction de répartition (exemple 1)

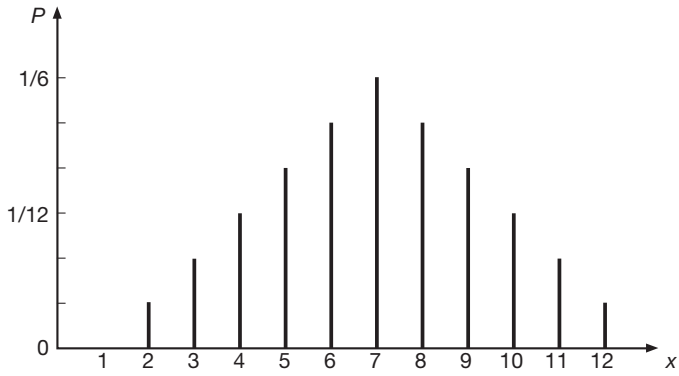


Figure 5.3 – Diagramme en bâtons (exemple 1)

3) Loi de probabilité d'une variable aléatoire continue

Une variable aléatoire X absolument continue est une variable aléatoire dont la *fonction de répartition* F_X possède en sus des trois propriétés déjà énoncées, les deux propriétés suivantes :

4. F_X est une fonction continue sur tout $\mathbb{R}$
5. F_X est dérivable presque partout [1]

1. C'est-à-dire que la fonction Fx peut ne pas être dérivable sur un ensemble dénombrable de points de $\mathbb{R}$

Toute fonction vérifiant ces cinq propriétés peut être considérée comme la fonction de répartition d'une variable aléatoire absolument continue.

La dérivée de F_X, notée f_X, est appelée *densité de probabilité* de la variable aléatoire X.

Une fonction f, définie sur tout $\mathbb{R}$, peut être considérée comme la *densité de probabilité* d'une variable aléatoire absolument continue si elle possède les trois propriétés suivantes :

1. $f(x) \geq 0 \ \forall x \in \mathbb{R}$
2. f continue presque partout [1]
3. $\int_{-\infty}^{+\infty} f(x)dx = 1$

La fonction de densité est une représentation très utile de la loi de probabilité d'une variable aléatoire continue. On peut définir la loi de probabilité d'une variable aléatoire continue, soit par sa fonction de répartition, soit par sa fonction de densité, et on a la relation fondamentale suivante :

$$\forall x \in \mathbb{R} \quad F(x) = \int_{-\infty}^{x} f(t)dt$$

La probabilité relative à un intervalle se calcule à l'aide de la fonction de répartition ou de la fonction de densité (*cf.* figure 5.4) :

$$P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)dx$$

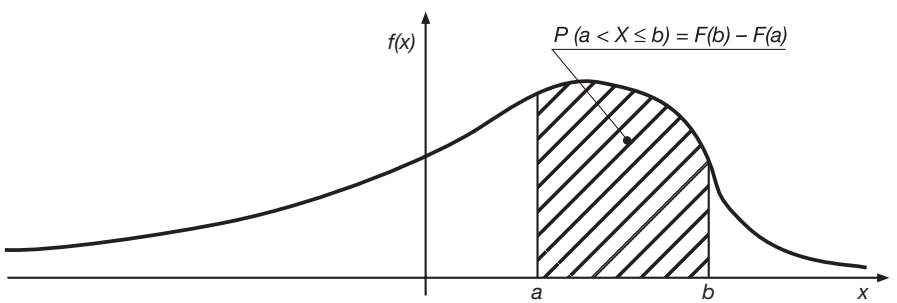


Figure 5.4 – Représentation de la probabilité d'un intervalle

1. C'est-à-dire que la fonction f peut ne pas être continue sur un ensemble dénombrable de points de $\mathbb{R}$; on dit encore que f est continue par morceaux ; mentionnons que les points de non-dérivabilité de F correspondent aux points de discontinuité de f

Probabilité attachée à un point x

Soient deux nombres réels a et b positifs :

$$0 \le P(X = x) \le P(x - a < X \le x + b) = F(x + b) - F(x - a)$$

$$= \Big(F(x + b) - F(x)\Big) + \Big(F(x) - F(x - a)\Big)$$

la fonction F étant continue : $\Big(F(x + b) - F(x)\Big) \to 0$ si $b \to 0$

$$\Big(F(x) - F(x - a)\Big) \to 0 \quad \text{si } a \to 0$$

d'où : $P(X = x) = 0$

$\Rightarrow$ la probabilité qu'une v.a. continue X prenne une valeur donnée x est nulle, on dit que la loi de X est *diffuse* (ou *continue*).

Par conséquent, pour une *variable aléatoire continue* :

$$F(x) = P(X \le x) = P(X < x) \quad \Rightarrow \quad \forall\, a, b \in \mathbb{R},\ a < b :$$

$$P(a < X < b) = P(a \le X \le b) = P(a < X \le b)$$

$$= P(a \le X < b) = F(b) - F(a) = \int_a^b f(x)dx$$

▶ **Exemple**

Soit la fonction f définie par :

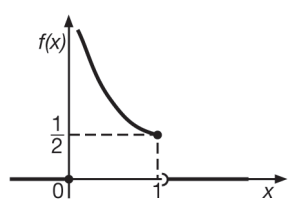
$$f(x) = \begin{cases} 0 & \text{pour} \quad x \le 0 \\ \dfrac{1}{2\sqrt{x}} & \text{pour} \quad 0 < x \le 1 \\ 0 & \text{pour} \quad x > 1 \end{cases}$$

Montrons que cette fonction peut être considérée comme la fonction de densité d'une v.a. continue :

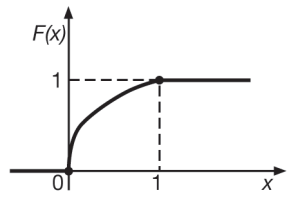
1. $f(x) \ge 0\ \forall x \in \mathbb{R}$

2. f continue sauf en $x = 0$ et $x = 1$

3. $\displaystyle\int_{-\infty}^{+\infty} f(x)dx = \frac{1}{2}\int_0^1 x^{-1/2}dx = \Big[x^{1/2}\Big]_0^1 = 1$



Calculons sa fonction de répartition :



pour $x \leq 0$ $\quad F(x) = \int_{-\infty}^{x} 0 \cdot dt = 0$

pour $0 < x \leq 1$ $\quad F(x) = \int_{-\infty}^{0} 0 \cdot dt + \int_{0}^{1} \frac{1}{2\sqrt{t}} dt = \left[\sqrt{t}\right]_{0}^{x} = \sqrt{x}$

pour $x > 1$ $\quad F(x) = \int_{-\infty}^{0} 0 \cdot dt + \int_{0}^{1} \frac{1}{2\sqrt{t}} dt + \int_{0}^{x} 0 dt = 1$

On vérifie aisément que cette fonction F possède les propriétés de la fonction de répartition d'une v.a. continue[1].

On peut calculer la probabilité de tout interv alle ou réunion d'interv alles, par exemple :

$$P(0{,}16 < X < 0{,}25) = F(0{,}25) - F(0{,}16) = 0{,}5 - 0{,}4 = 0{,}1$$

C. Loi d'une fonction de variable aléatoire

Si φ est une fonction définie sur $\mathbb{R}$ à valeurs dans $\mathbb{R}$, l'application $\varphi \circ X$, notée $Y = \varphi(X)$ est une variable aléatoire dont on peut déterminer la fonction de répartition – et donc la loi de probabilité – à partir de celle de X.

1) Changement de variable $Y = aX + b$

Les paramètres a ($a \neq 0$) et b sont des nombres réels. Connaissant la fonction de répartition de X, on peut calculer la fonction de répartition F_Y de la v.a. Y :

• pour $a > 0$:

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$$

1. On peut remarquer que les deux points de discontinuité de la fonction de densité { $x = 0$} et {$x = 1$} correspondent aux deux points de non-dérivabilité de la fonction de répartition.

• pour $a < 0$:

$$F_Y(y) = P(Y \le y) = P\left(X \ge \frac{y-b}{a}\right)$$

$$= \begin{cases} 1 - F_X\left(\dfrac{y-b}{a}\right) & \text{si } X \text{ est une v.a. continue} \\ 1 - P\left(X < \dfrac{y-a}{a}\right) & \text{si } X \text{ est une v.a. discrète} \end{cases}$$

Lorsque la variable aléatoire X est continue, on obtient la fonction de densité f_Y par dérivation de la fonction F_Y

2) Autres types de fonctions

• φ *bijective* (donc monotone)

φ croissante : $F_Y(y) = P(Y \le y) = P(X \le \varphi^{-1}(y)) = F_X(\varphi^{-1}(y))$

φ décroissante :

$$F_Y(y) = P(Y \le y) = P(X \ge \varphi^{-1}(y))$$

$$= \begin{cases} 1 - F_X(\varphi^{-1}(y)) & \text{si } X \text{ est une v.a. continue} \\ 1 - P(X < \varphi^{-1}(y)) & \text{si } X \text{ est une v.a. discrète} \end{cases}$$

Si X est une v.a. continue et si la fonction φ est dérivable, on obtient la fonction de densité f_Y par dérivation de la fonction F_Y

▶ **Exemple**

Soit une v.a. continue X, on peut calculer les fonctions de répartition et de densité de $Y = \exp(X)$, la fonction exponentielle étant croissante :

$$F_Y(y) = \begin{cases} 0 & \text{pour} \quad y \le 0 \\ F_X(\ln y) & \text{pour} \quad y > 0 \end{cases} \Rightarrow f_Y(y) = \begin{cases} 0 & \text{pour} \quad y \le 0 \\ \dfrac{1}{y} f_X(\ln y) & \text{pour} \quad y > 0 \end{cases}$$

• φ *quelconque*

Le principe consiste toujours à identifier la fonction de répartition F_Y en recherchant l'antécédent pour X de l'événement $\{ Y \le y = \varphi(x) \}$.

Par exemple, pour $Y = X^2$:

$$F_Y(y) = \begin{cases} 0 & \text{si } y < 0 \\ P(-\sqrt{y} \le X \le +\sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{si } y \ge 0 \end{cases}$$

III. Couple de variables aléatoires

A. Fonction de répartition d'un couple aléatoire

Soient deux variables aléatoires X et Y définies sur le même espace probabilisé $(\Omega, \mathcal{A}, P)$; on appelle *fonction de répartition du couple aléatoire (X,Y)*, la fonction F définie sur $\mathbb{R}^2$ par :

$$\forall\ (x\ ,\ y) \in \mathbb{R}^2 \qquad F(x\ ,\ y)\ =\ P\Big((X \leq x) \cap (Y \leq y)\Big)$$

Caractérisation d'une fonction de répartition d'un couple aléatoire (X,Y)

1. F croissante par rapport à chacune des variables x et y

2. $\lim\limits_{\substack{x \to +\infty \\ y \to +\infty}} F(x\ ,\ y) = 1$ et $\lim\limits_{\substack{x \to -\infty \\ y \to -\infty}} F(x\ ,\ y) = 0$

3. Continuité à droite : $\lim\limits_{\substack{x \to x_0^+ \\ y \to y_0^+}} F(x\ ,\ y) = F(x_0\ ,\ y_0)$

B. Loi d'un couple aléatoire discret

Les variables aléatoires discrètes finies X et Y sont définies sur le même espace probabilisé $(\Omega, \mathcal{A}, P)$. Leurs valeurs, supposées distinctes, sont rangées dans l'ordre croissant :

$$X(\Omega) = \{\, x_1, \ldots\ ,\ x_i\ ,\ \ldots\ ,\ x_k\ \} \text{ et } Y(\Omega) = \{\, y_1, \ldots\ ,\ y_j\ ,\ \ldots\ ,\ y_l\ \}$$

La loi du couple aléatoire (X,Y) est définie par les probabilités p_{ij} associées à tout couple de valeurs possibles $(x_i\ ,\ y_j)$ (*cf.* tableau 5.1) :

$$p_{ij} = P(X = x_i\ ,Y = y_j) \qquad \Rightarrow \qquad \sum_{j=1}^{l} \sum_{i=1}^{k} p_{ij} = 1$$

Tableau 5.1 –Distribution de probabilité d'un couple aléatoire (X,Y)

Valeur de Y / Valeur de X	y_1	…	y_j	…	y_l	Loi marginale de X
x_1	p_{11}	…	p_{1j}	…	p_{1l}	$p_{1\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	p_{i1}	…	p_{ij}	…	p_{il}	$p_{i\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_k	p_{k1}	…	p_{kj}	…	p_{kl}	$p_{k\bullet}$
Loi marginale de Y	$p_{\bullet 1}$	…	$p_{\bullet j}$	…	$p_{\bullet 1}$	1

On notera l'analogie du tableau 5.1 avec le tableau d'une distribution conjointe en statistique descriptive (chapitre 3, § I.A).

On a :
$$p_{i\bullet} = \sum_{j=1}^{l} p_{ij} = P(X = x_i)$$

$$p_{\bullet j} = \sum_{i=1}^{k} p_{ij} = P(X = x_i)$$

Les couples (x_i, $p_{i\bullet}$) constituent la *loi marginale de X* et les couples (y_j, $p_{\bullet j}$) constituent la *loi marginale de Y.*

Si la probabilité que X prenne la valeur x_i n'est pas nulle ($p_{i\bullet} \neq 0$), on peut calculer la *probabilité conditionnelle* $p_{j/i}$ de $Y = y_j$ sachant que $X = x_i$:

$$p_{j/i} = P(Y = y_j | X = x_i) = \frac{p_{ij}}{p_{i\bullet}}$$

Les couples (y_j, $p_{j/i}$) constituent la *loi conditionnelle* de Y liée par $X = x_i$ On note cette v.a. { $Y|X = x_i$ }, et on présente sa distribution comme celle de toute v.a. à une dimension :

Valeur de Y	y_1	….	y_j	….	y_l
$P(Y = y_j \mid X = x_i)$	$p_{1/i}$	….	$p_{j/i}$	….	$p_{l/i}$

$$\sum_{j=1}^{l} p_{j/i} = 1$$

Il y a k lois conditionnelles de Y sachant que X prend une valeur donnée.

De même, si la probabilité $p_{\bullet j}$ n'est pas nulle :

$$p_{i/j} = P(X = x_i | Y = y_j) = \frac{p_{ij}}{p_{\bullet j}}$$

Les couples $(x_i , p_{i/j})$ constituent la *loi conditionnelle* de X liée par $Y = y_j$:

$$\{ Y | X \quad y_j \}=$$

Il y a l lois conditionnelles de X sachant que Y prend une valeur donnée.

Les deux formules précédentes entraînent [1] :

$$p_{ij} = p_{i\bullet} \cdot p_{j/i} = p_{\bullet j} \cdot p_{i/j}$$

Indépendance

Les variables aléatoires X et Y sont indépendantes si *pour tout couple* (x_i , y_j), on a la relation :

$$P((X = x_i) \cap (Y = y_j)) = P(X = x_i) \cdot P(Y = y_i)$$

X et Y indépendantes $\Leftrightarrow$ $p_{ij} = p_{i\bullet} \cdot p_{\bullet j}$ pour tout couple (i , j)

En cas d'indépendance, les lois conditionnelles sont égales à la loi marginale correspondante :

$$p_{j/i} = \frac{p_{ij}}{p_{i\bullet}} = p_{\bullet j} \quad \text{et} \quad p_{i/j} = \frac{p_{ij}}{p_{\bullet j}} = p_{i\bullet}$$

ce qui signifie que la connaissance de la valeur prise par X n'apporte aucune information sur la valeur de Y, et inversement.

La loi de probabilité d'un couple aléatoire (X,Y) permet de calculer les lois marginales des deux variables X et Y. En revanche, la connaissance de ces lois ne permet pas de déterminer la loi conjointe, sauf si les variables X et Y sont indépendantes.

Mentionnons l'analogie existant entre les notions de lois de probabilité marginales et conditionnelles définies pour un couple aléatoire et celles de distributions marginales et conditionnelles rencontrées en statistique descriptive (chapitre 3).

Toutes les notions développées pour les couples de variables aléatoires discrètes finies peuvent être généralisées à des variables aléatoires discrètes infinies.

La loi de probabilité d'un couple aléatoire discret peut aussi être définie par sa fonction de répartition.

Pour $\{x_i \leq x < x_{i+1}\}$ et $\{y_j \leq y < y_{j+1}\}$:

$$F(x , y) = P((X \leq x) \cap (Y \leq y)) = \sum_{n=1}^{j} \sum_{m=1}^{i} p_{mn}$$

1. Les probabilités conditionnelles $p_{j/i}$ et $p_{i/j}$ sont aussi parfois notées p_j^i et p_i^j

C. Loi d'un couple de variables aléatoires continues

La fonction de répartition d'un couple (X,Y) de variables aléatoires continues possède en sus des trois propriétés déjà énoncées, les deux propriétés suivantes :

4. F est une fonction continue sur $\mathbb{R}^2$
5. F est dérivable presque partout

Toute fonction vérifiant les cinq propriétés peut être considérée comme la fonction de répartition d'un couple de variables aléatoires continues.

La *densité f* du couple (X, Y) est donnée par : $f(x\ ,\ y) = \dfrac{\partial^2 F}{\partial x \partial y}(x\ ,\ y)$

La loi de probabilité d'un couple de variables aléatoires continues peut être définie, soit par la fonction de répartition, soit par la fonction de densité, et on a la relation fondamentale suivante :

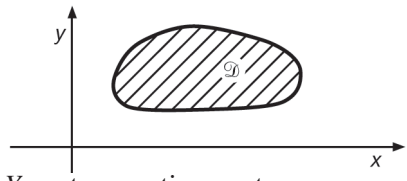
$$\forall (x\ ,\ y) \in \mathbb{R}^2 \qquad F(x\ ,\ y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u\ ,\ v)\,du\,dv$$

La probabilité relative à un sous-ensemble de $\mathbb{R}^2$ du type $[a\ ;\ b] \times [c\ ;\ d]$ est égale à :

$$P\Big((X, Y) \in [a\ ;\ b] \times [c\ ;\ d]\Big) = \int_{c}^{d} \int_{a}^{b} f(u\ ,\ v)\,du\,dv$$

Plus généralement, la probabilité que le couple aléatoire (X,Y) appartienne à un « domaine » $\mathcal{D} \subset \mathbb{R}^2$ est égale à :

$$P\Big\{(X, Y) \in \mathcal{D}\Big\} = \int\int_{\mathcal{D}} f(x\ ,\ y)\,dx\,dv$$



Les densités marginales g de X et h de Y sont respectivement :

$$g(x) = \int_{-\infty}^{+\infty} f(x\ ,\ y)\,dy \qquad \text{et} \qquad h(y) = \int_{-\infty}^{+\infty} f(x\ ,\ y)\,dx$$

► **Exemple**

Un couple (X, Y) de variables aléatoires continues suit une *loi uniforme* sur $[0\ ;\ 1] \times [0\ ;\ 1]$ si sa densité de probabilité est la suivante :

$$F(x, y) = \begin{cases} 0 & \text{pour tout} \quad (x\ ,\ y) \notin [0\ ;\ 1] \times [0\ ;\ 1] \\ 1 & \text{pour tout} \quad (x\ ,\ y) \in [0\ ;\ 1] \times [0\ ;\ 1] \end{cases}$$

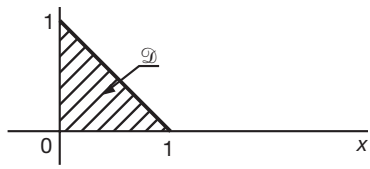
Connaissant la fonction de densité, on peut calculer la probabilité de tout sous-ensemble de $\mathbb{R}^2$:

$$P\Big((X < 0{,}3) \cap (0{,}1 \leq Y < 0{,}8)\Big) = \int_{0{,}1}^{0{,}8} \int_{0}^{0{,}3} 1 \cdot dxdy = 0{,}21$$

Considérons le domaine $\mathcal{D}$ de $\mathbb{R}^2$ ainsi défini :
$\mathcal{D} = \{(x\,,\,y) \in \mathbb{R}^2 \mid x > 0,\, y > 0 \text{ et } x + y < 1\}$,

alors $P\{X, Y \in \mathcal{D}\} = \int_{0}^{1} \int_{0}^{1-u} 1 \cdot dudv = 0{,}5$

Le lecteur peut vérifier que les lois marginales de X et Y sont des lois uniformes continues sur [0 ; 1] – (chapitre 7, §I.A).



Indépendance

Les variables aléatoires X et Y sont indépendantes si et seulement si $\forall(x\,,\,y) \in \mathbb{R}^2$:

$$f(x\,,\,y) = g(x) \cdot h(y)$$

Plus généralement, un n-uplet de variables aléatoires $(X_1, X_2, \ldots, X_n)$ de densité de probabilité f est un n-uplet de variables aléatoires indépendantes si et seulement si la densité f du n-uplet est le produit des n densités marginales f_i :

$$f(x_1\,,\,x_2\,,\ldots,\,x_n) = f_1(x_1) \cdot f_2(x_2) \cdot \ldots \cdot f_n(x_n)$$

IV. Indicateurs des variables aléatoires

Une différence entre la statistique descriptive et la théorie des probabilités réside dans le fait que la première discipline vise à représenter les données de façon à les rendre plus « lisibles », tandis que la seconde a pour objectif de fournir des modèles adaptés au traitement mathématique, donc abstraits, qui se veulent des images, à la fois idéales et approchées de ces données.

L'utilisation simultanée de ces deux démarches doit permettre de faire apparaître les lois susceptibles de régir les phénomènes dont proviennent les données, puis de les exprimer de manière plus précise et maniable grâce au formalisme mathématique qui en dégage les propriétés essentielles.

Il est naturel, comme on l'a fait en statistique descriptive, de définir et d'étudier des indicateurs (ou caractéristiques) des variables aléatoires. La motivation est la même : la loi de probabilité constitue une grande quantité d'informations, et est souvent trop riche pour être appréhendée dans sa globabilité. Il est donc utile d'en résumer certains aspects (les mêmes que ceux envisagés en statistique descriptive) par des valeurs numériques convenablement choisies.

Des indicateurs relatifs aux trois aspects principaux des lois de probabilité sont définis, à savoir :

- la tendance centrale ;
- la dispersion ;
- la forme (asymétrie et aplatissement).

Les outils mathématiques qui interviennent dans la définition de ces indicateurs varient d'un type de loi à l'autre. Les lois discrètes finies utilisent les mêmes outils que ceux nécessaires à la définition des indicateurs des variables statistiques. Pour les lois discrètes infinies, quelques connaissances sur les séries numériques (et parfois sur les séries entières) sont utiles. Pour les lois absolument continues, des notions sur l'intégration sont utilisées. Cependant, la signification de ces indicateurs ne dépend pas du type de loi de probabilité considérée, ni des techniques mathématiques utilisées.

A. Mode

Le *mode* d'une variable aléatoire est la valeur pour laquelle le diagramme en bâtons ou la courbe de densité présente son maximum. On appelle *mode relatif* une valeur correspondant à un maximum local du diagramme en bâtons ou de la courbe de densité, mais en général, le mode est unique. Le mode est un indicateur de tendance centrale.

B. Espérance mathématique

L'espérance mathématique d'une variable aléatoire X est aussi appelée « moyenne » ou « valeur moyenne » de X. Elle est généralement notée m.

1) Cas discret

Soit X une variable aléatoire discrète finie :

Valeur de X	x_1	.....	x_i	.....	x_k
Probabilité	p_1	.....	p_i	.....	p_k

On appelle *espérance mathématique* ou *moyenne* $E(X)$ de X :

$$E(X) = \sum_{i=1}^{k} x_i p_i$$

À titre d'illustration, le lecteur peut vérifier que la v.a. de l'exemple 1 du § II.B a une moyenne égale à 7. On peut remarquer la similitude des définitions de l'espérance mathématique et de la moyenne arithmétique d'une variable statistique discrète. On a remplacé dans la formulation de cette dernière les fréquences par les probabilités.

L'espérance mathématique est un nombre réel, mais souvent, pour une variable aléatoire discrète, sa valeur ne correspond pas à une des valeurs possibles de cette variable aléatoire.

▶ **Exemple**

Une loterie comporte 1 000 billets et un seul lot de 10 000 €. Si tous les billets ont été v endus et si le tirage se f ait « au hasard », l'espérance mathématique de la valeur V d'un billet sera :

$$E(V) = 10\,000 \cdot \frac{1}{1\,000} + 0 \cdot \frac{999}{1\,000} = 10 \text{ €}$$

Mais, en f ait, aucun billet ne rapporte 10 € : chacun rapporte 0 € ou 10 000 €. Cependant, si on achète un billet à chaque tirage de cette loterie (en supposant qu'elle ait lieu régulièrement dans les mêmes conditions), la moyenne des gains sera « voisine » de 10 € au bout d'un grand nombre de tirages ; ce résultat qui f ait l'importance du concept d'espérance mathématique se réfère à la loi des grands nombres (§ V).

La moyenne d'une variable aléatoire X a ainsi la signification d'un indicateur de « tendance centrale » de X.

Dans le cas d'une variable aléatoire X discrète infinie : $E(X) = \sum_{i=1}^{\infty} x_i p_i$ sous réserve que la série de terme général $x_i p_i$ soit *absolument convergente*[1], sinon, et même si elle est simplement convergente, on dira que la v.a. X n'a pas d'espérance mathématique.

1. La série $\sum_{i=1}^{+\infty} x_i p_i$ est absolument convergente si la série $\sum_{i=1}^{+\infty} |x_i p_i| = \sum_{i=1}^{+\infty} |x_i||p_i|$ est convergente.

L'espérance mathématique de la v.a. discrète conditionnelle $\{ Y|X = x_i\}$, définie au § III.B. – est appelée *espérance conditionnelle* de Y sachant que $X = x_i$. Elle a pour expression :

$$E\{Y|X = x_i\} = \sum_{j=1}^{l} y_j p_{j/i}$$

De même : $$E\{X|Y = y_i\} = \sum_{i=1}^{k} x_i p_{i/j}$$

2) Cas continu

La variable aléatoire X étant continue de densité f, on appelle espérance mathématique $E(X)$ de X :

$$E(X) = \int_{-\infty}^{+\infty} x\ f(x)\ dx$$

Cette définition suppose l'intégrale du second membre *absolument convergente*[1], sinon – et même si elle est simplement convergente, on dira que la v.a. X n'a pas d'espérance mathématique.

➞ ***Remarque***

L'espérance mathématique d'une variable *certaine* X, c'est-à-dire une v.a. ne prenant qu'une seule valeur, notée b, avec la probabilité 1, est égale à cette valeur : $E(b) = b$. Notons qu'une telle variable ne mérite pas exactement le nom de variable aléatoire puisqu'elle peut être identifiée à la constante b.

On appelle *variable aléatoire centrée* une variable aléatoire dont l'espérance mathématique est nulle.

3) Propriétés de l'espérance mathématique

1. Si a et b sont deux nombres réels : $E(aX + b) = a\,E(X) + b$

$\Rightarrow$ si une v.a. X possède une espérance mathématique m, alors la variable aléatoire $Y = X - m$ est la variable aléatoire centrée associée à X.

---

1. L'intégrale $\int_{-\infty}^{+\infty} x\ f(x)\ dx$ est absolument convergente si l'intégrale $\int_{-\infty}^{+\infty} |x|\ f(x)\ dx$ est convergente.

2. Soit φ une fonction définie sur $\mathbb{R}$ à valeurs dans $\mathbb{R}$, alors si X est une v.a., $\varphi(X)$ est une v.a. (§ II.C) dont on peut calculer l'espérance sans avoir à déterminer sa loi.

Cas discret

$E(\varphi(X)) = \sum_i \varphi(x_i) p_i$ en supposant toujours que la série du second membre est absolument convergente. En particulier :

$$E(X^2) = \sum_i x^2 p_i$$

Cas continu

$E(\varphi(X)) = \int_R \varphi(x) f(x)\, dx$ en supposant toujours l'intégrale du second membre absolument convergente. En particulier :

$$E(X^2) = \int_R x^2 f(x)\, dx$$

3. L'*espérance d'une somme* de variables aléatoires est égale à la somme des espérances :

$$E(X + Y) = E(X) + E(Y)$$

1re conséquence :

$$E(X - Y) = E(X) + E(-Y) = E(X) - E(Y)$$

2de conséquence :

Soient n variables aléatoires $X_1, X_2, \ldots, X_n$ ayant la même espérance mathématique m. L'espérance de leur somme est égale à :

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = n \cdot m$$

si on désigne par $\overline{X}$ leur moyenne : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, on obtient : $E(\overline{X}) = m$

4. *Espérance d'un produit de variables aléatoires*

Cas discret

Soit (X, Y) un couple de v.a. discrètes, on a, si la série du second membre est absolument convergente :

$$E(X \cdot Y) = \sum_{i,j} x_i y_i p_{ij}$$

Cas continu

Soit (X, Y) un couple de v.a. continues, on a, si l'intégrale du second membre est absolument convergente :

$$E(X \cdot Y) = \int_{R^2} xy \; f(x, y) \; dxdy$$

Si X et Y sont *indépendantes*, de la propriété $\{ \quad p_{ij} = p_{i\bullet} \cdot p_{\bullet j}$ ou $f(x,y) = g(x) \cdot h(y)\}$, on déduit $E(X \cdot Y) = E(X) \cdot E(Y)$, mais la réciproque n'est pas vraie :

$$X \text{ et } Y \text{ indépendantes} \quad \begin{matrix} \Rightarrow \\ \not\Leftarrow \end{matrix} \quad E(X \cdot Y) = E(X) \cdot E(Y)$$

C. Variance

La variance d'une variable aléatoire X est l'espérance mathématique du carré de la v.a. centrée associée à X (si elle existe) :

$$\text{var}(X) = E(X - m)^2$$

La variance est un nombre positif ou nul ; sa racine carrée, notée σ, est appelée *écart-type*[1] :

$$\sigma = \sqrt{\text{var}(X)}$$

L'écart-type d'une v.a. X, exprimé dans les mêmes unités que la variable X, a la signification d'un indicateur de « *dispersion* » autour de la moyenne m de X. Illustrons cette idée par un exemple. Soient les variables aléatoires X et Y :

Valeur de X	2	4	8
Probabilité	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

$E(X) = 4$

$\sigma_X = \sqrt{6}$

Valeur de Y	-6	2	30
Probabilité	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$

$E(Y) = 4$

$\sigma_Y = \sqrt{148}$

Ces deux variables aléatoires ont même espérance. Cette grandeur typique ne permet pas de les distinguer. Cet exemple montre bien que l'écart-type d'une variable aléatoire est un indicateur de dispersion autour de sa moyenne.

1. Les calculs de la moyenne et de l'écart-type des v.a. discrètes finies peuvent se faire avec des calculatrices possédant les fonctions statistiques, les fréquences étant remplacées par les probabilités.

Propriétés de la variance

1. var(X) est la *valeur minimale* de $E\Big((X-b)^2\Big)$, car on montre que :

$$E\Big((X-b)^2\Big) = \text{var}(X) + \Big(E(X)-b\Big)^2$$

De cette formule, on déduit :

• $\{\text{var}(X) = 0 \Leftrightarrow X$ est une variable *certaine*$\}$

• la relation usuelle : $\text{var}(X) = E(X^2) - \Big(E(X)\Big)^2$

2. À l'aide de cette dernière expression de la variance, on montre sans difficulté :

$$\forall a \text{ et } b \in \mathbb{R},\ \text{var}(aX+b) = a^2\,\text{var}(X) \Rightarrow \sigma_{aX+b} = |a|\sigma_X$$

3. La variance d'une somme de deux variables aléatoires *indépendantes* X et Y est égale à la somme des variances :

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$$

en effet :

$$\begin{aligned}\text{var}(X+Y) &= E((X+Y-E(X+Y))^2)\\ &= E(X-E(X)+(Y-E(Y))^2\\ &= E((X-E(X))^2) + E((Y-E(Y))^2)\\ &\quad + 2E((X-E(X))(Y-E(Y)))\\ &= \text{var}(X) + \text{var}(Y) + 2E(X-E(X))(Y-E(Y))\end{aligned}$$

pour deux variables *indépendantes*, le dernier terme est nul
$\Rightarrow \text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$

1re conséquence

X et Y indépendantes $\Rightarrow \text{var}(X-Y) = \text{var}(X) + \text{var}(-Y) = \text{var}(X) + \text{var}(Y)$

2de conséquence

Soient n variables aléatoires $X_1, X_2, \ldots, X_n$ mutuellement indépendantes et de même variance σ^2, la variance de leur somme est égale à $n\sigma^2$:

$$\text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{var}(X_i) = n\sigma^2$$

Si on désigne par $\overline{X}$ leur moyenne : $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, on obtient :

$$\text{var}(\overline{X}) = \frac{\sigma^2}{n}$$

Une variable aléatoire *réduite* est une variable aléatoire dont l'écart-type est égal à 1. Si X a une moyenne m et un écart-type σ, on peut lui associer une variable aléatoire Y *centrée réduite* :

$$Y = \frac{X - m}{\sigma}$$

D. Covariance de deux variables aléatoires, coefficient de corrélation linéaire

On appelle *covariance* d'un couple de variables aléatoires X et Y la quantité :

$$\text{cov}(X, Y) = E\{(X - E(X))(Y - E(Y))\} = E(XY) - E(X) \cdot E(Y)$$

$$\Rightarrow \text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$$

On déduit de la propriété 4 de l'espérance mathématique :

X et Y indépendantes $\Rightarrow$ $\not\Leftarrow$ **$\text{cov}(X, Y) = 0$**

Propriétés de la covariance

1. $\text{cov}(X,Y) = \text{cov}(Y, X)$

2. $\text{cov}(X, X) = \text{var}(X)$

3. $\forall a, b, c$ et $d \in \mathbb{R}$:

$$\text{cov}(aX + b, cY + d) = ac \cdot \text{cov}(X, Y)$$

$$\Rightarrow \text{var}(aX + bY + c) = a^2\text{var}(X) + b^2\text{var}(Y) + 2ab \cdot \text{cov}(X, Y)$$

4. $|\text{cov}(X, Y)| \leq \sqrt{\text{var}(X) \cdot \text{var}(Y)}$, cette inégalité est une conséquence de l'inégalité de Schwarz.

On appelle *coefficient de corrélation linéaire* entre X et Y le rapport :

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Des propriétés de la covariance, on déduit que le coefficient de corrélation linéaire est *invariant par changement d'origine et d'échelle* et qu'il est compris entre – 1 et + 1. On peut montrer qu'il est égal à + 1 si et seulement si X et Y sont liées par une relation linéaire. D'autre part, si X et Y sont indépendantes, leur coefficient de corrélation linéaire est nul, mais la réciproque n'est pas vraie. On retrouve l'analogie de ce coefficient ρ avec le coefficient de corrélation linéaire r défini entre deux variables statistiques au chapitre 3, § II.A.

E. Moment, fonction génératrice des moments

1) Moment

On appelle *moment* m_k d'ordre k (k entier positif) d'une variable aléatoire X l'espérance mathématique de X^k si elle existe :

$$m_k = E(X^k)$$

L'espérance mathématique n'est autre que le moment d'ordre 1.

On appelle *moment centré* μ_k d'ordre k (k entier positif) d'une variable aléatoire X l'espérance mathématique de $(X - E(X))^k$, si elle existe :

$$\mu_k = E\Big(X - E(X)\Big)^k$$

La variance n'est autre que le moment centré d'ordre 2 ; le moment centré d'ordre 1 est toujours nul.

2) Moment factoriel

On appelle *moment factoriel* $\mu_{[k]}$ d'ordre k (k entier positif) d'une variable aléatoire X l'espérance mathématique de $X(X-1).....(X-k+1)$ – si elle existe :

$$\mu_{[k]} = E\Big(X(X-1).....(X-k+1)\Big)$$

Le moment factoriel d'ordre k est une combinaison linéaire des moments non centrés m_1 , m_2 , … , m_k

Relations entre moments et moments factoriels jusqu'à l'ordre 4 :

$$\begin{cases} \mu_{[1]} = m_1 \\ \mu_{[2]} = m_2 - m_1 \\ \mu_{[3]} = m_3 - 3m_2 + 2m_1 \\ \mu_{[4]} = m_4 - 6m_3 + 11m_2 - 6m_1 \end{cases} \Leftrightarrow \begin{cases} m_1 = \mu_{[1]} \\ m_2 = \mu_{[2]} + \mu_{[1]} \\ m_3 = \mu_{[3]} + 3\mu_{[2]} + \mu_{[1]} \\ m_4 = \mu_{[4]} + 6\mu_{[3]} + 7\mu_{[2]} - 11\mu_{[1]} \end{cases}$$

3) Fonction génératrice des moments

La fonction génératrice des moments va être présentée en se restreignant à une variable aléatoire discrète à valeurs possibles entières non négatives. Cette fonction caractérise la loi d'une variable aléatoire, et elle permet de plus d'obtenir les moments factoriels par dérivation.

On appelle *fonction génératrice* g_X des moments d'une variable aléatoire X discrète, à valeurs possibles *entières non négatives*, l'espérance mathématique de u^X, u étant compris entre 0 et +1 :

$$g_X(u) = E(u^X) = \sum_{i=0}^{+\infty} u^i p_i \quad \text{pour} \quad 0 \leq u \leq 1$$

Propriétés de la fonction génératrice des moments

1. Pour $0 \leq u \leq 1$, la fonction g_X est continue, car la série qui la définit est uniformément convergente en u :

$$g_X(u) = \sum_{i=0}^{+\infty} u^i p_i \leq \sum_{i=0}^{+\infty} p_i = 1$$

2. $g_X(0) = 0$ et $g_X(1) = 1$

3. Si le moment factoriel d'ordre k de X existe, on montre que pour u compris entre 0 et 1:

$$g_X^{(k)}(u) = \sum_{i=k}^{+\infty} [i(i-1).....(i-k+1)u^{i-k}]\ p_i \quad \Rightarrow \quad g_X^{(k)}(1) = \mu_{[k]}$$

en notant $g_X^{(k)}$ la dérivée d'ordre k de la fonction g_X

Cette propriété de la fonction génératrice est utilisée pour le calcul des moments factoriels qui permettent de calculer les moments non centrés, puis centrés.

F. Indicateurs de forme

Ces indicateurs donnent des informations sur la forme de la loi de X, et en particulier, ils la comparent à la loi normale (chapitre 7, § II.B). Ils sont directement inspirés des coefficients d'asymétrie (en anglais *skewness*) et d'aplatissement (*kurtosis*) définis en statistique descriptive.

Fisher a défini les coefficients d'asymétrie et d'aplatissement d'une variable aléatoire X, dont les premiers moments existent, par :

- *coefficient d'asymétrie* $\gamma_1 = \dfrac{\mu_3}{\mu_2^{3/2}}$
- *coefficient d'aplatissement* $\gamma_2 = \dfrac{\mu_4}{\mu_2^2} - 3$

Les moments centrés d'ordre impair étant nuls pour une distribution symétrique, γ_1 est nul si la distribution de X est symétrique par rapport à la

moyenne m, mais la réciproque n'est pas vraie : γ_1 peut être nul sans que la loi de X soit symétrique. Si la distribution de X est unimodale étalée vers la droite, γ_1 est positif. Dans le cas contraire, γ_1 est négatif.

Le coefficient d'aplatissement γ_2 est nul pour une variable distribuée selon une loi normale, mais là encore, la réciproque n'est pas vraie. Selon que la loi de X est plus ou moins aplatie que la loi normale, γ_2 sera positif ou négatif. Plus que l'aplatissement, ce coefficient mesure l'importance des « queues » d'une distribution.

Ces coefficients γ_1 et γ_2 sont invariants par changement d'origine et d'échelle.

G. Quantiles

Comme pour les variables statistiques, on définit pour les variables aléatoires les *quantiles*, encore appelés *fractiles*, qui sont indicateurs de position à partir desquels on peut définir des indicateurs de tendance centrale et de dispersion.

On appelle *quantile d'ordre* α $(0 \leq \alpha \leq 1)$ d'une variable aléatoire X de fonction de répartition F toute valeur x_α telle que : $F(x_\alpha) = \alpha$ $(\Leftrightarrow P(X \leq x_\alpha) = \alpha)$

Notons que si F est continue et strictement croissante, le quantile x_α, pour α donné, existe et est unique. Si F n'est pas continue et strictement croissante, il peut ne pas exister ou il peut y avoir plusieurs solutions possibles.

La *médiane Me* d'une v.a. X est le quantile d'ordre 1/2 : $Me = x_{0,5}$

Le *premier quartile*, noté Q_1, est le quantile d'ordre 1/4. Le *troisième quartile*, noté Q_3, est le quantile d'ordre 3/4. La médiane est le *second quartile*. On définit aussi les *déciles* : le $i^{\text{ème}}$ décile D_i est le quantile d'ordre i /10 $(1 \leq i \leq 9)$.

Comme en statistique descriptive, on peut définir plusieurs indicateurs à partir des quantiles :

• des *indicateurs de tendance centrale* comme par exemple, la médiane *Me* ou encore le milieu de l'intervalle interquartile :

$$\frac{1}{2}(Q_1 + Q_3)$$

• des *indicateurs de dispersion* comme, par exemple, l'étendue interquartile $(Q_3 - Q_1)$ ou l'espérance mathématique des écarts absolus à la médiane :

$$E|X - Me| \ (= \min_b E|X - b|)$$

• des *indicateurs de forme* comme, par exemple :

$$\frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

Ce coefficient, compris entre – 1 et + 1, est nul pour une distribution symétrique, positif pour une distribution unimodale étalée vers la droite, et négatif dans le cas contraire.

V. Convergence des variables aléatoires réelles

Les variables aléatoires réelles sont des applications de l'ensemble fondamental Ω dans $\mathbb{R}$. Dans certains cas, il est utile (voire nécessaire) de considérer des suites de v.a. correspondant à des suites d'épreuves aléatoires (ou comme on le verra aux chapitres suivants, à des suites de modèles). Pour ces suites, on va définir plusieurs notions de convergence, visant toutes à définir un comportement (ou une distribution) limite. Chacune correspond à des conditions différentes, mais bien entendu, plus ces conditions seront générales, moins les propriétés qui s'en déduisent seront puissantes. Nous partirons du mode de convergence le plus général, donc le plus faible.

Soit (X_n) une suite de v.a. réelles, de fonctions de répartition F_n. On dit qu'elle **converge en loi** vers la v.a. X de fonction de répartition F si on a :

$\lim_{n \to \infty} F_n(x) = F(x)$ en tout point $x \in \mathbb{R}$, sauf aux points de discontinuité de F

On écrit alors $X_n \xrightarrow{L} X$, et on parle aussi de **convergence faible**.

Cette notion de convergence est particulièrement simple en pratique. En effet, pour des variables aléatoires $\{ X_n \}$ et X discrètes, elle revient à :

$$\lim_{n \to \infty} P(X_n = x) = P(X = x)$$

et dans tous les autres cas, elle met en œuvre les critères classiques de convergence des fonctions.

Pour le cas des variables aléatoires discrètes, cette notion de convergence est utilisée par exemple pour l'approximation d'une loi binomiale par une loi de Poisson (à l'aide de la formule de Stirling), ce que nous voyons au chapitre 6, § III.E. On notera cependant qu'il est possible par la convergence en loi, qui ne fait intervenir que les fonctions de répartition, de rechercher (et/ou de poser) la convergence de v.a. discrètes vers une v.a. continue.

D'autre part, si on suppose la convergence en loi des v.a. $\{ X_n \}$ vers X, on peut approcher F_n par F, et si leurs densités existent on peut approximer f_n par f, ce qui est pratiqué dans les chapitres suivants.

Plus restrictive que la convergence en loi, la convergence en probabilité est définie ainsi :

Une suite X_n de v.a. réelles **converge en probabilité** vers la v.a. X, si on a :

$$\lim_{n \to \infty} P(|X_n - X| > \varepsilon) = 0 \text{ pour tout } \varepsilon > 0$$

On écrit alors $X_n \xrightarrow{P} X$ ou également plim $X_n = X$

Dans le cas de la convergence en probabilité vers une v.a. certaine, on peut se ramener à des conditions portant sur les deux premiers moments des v.a. X_n. Pour passer au cas général de la convergence de X_n vers X, on considère alors la convergence vers 0 de la suite $\{X_n - X\}$.

Pour le montrer, nous allons d'abord établir un résultat intermédiaire.

Inégalité de Bienaymé-Tchébychev

Soit Z une v.a. de moyenne μ et d'écart-type σ, on a alors pour tout nombre réel k :

$$P(|Z - \mu| > k \cdot \sigma) \leq \frac{1}{k^2}$$

Supposant la variable Z continue, soit $\mathbb{D}$ l'ensemble des valeurs z de Z tels que $|Z - \mu| > k \cdot \sigma$, on peut écrire, en partant de la définition de la variance de Z :

$$\sigma^2 = \text{var}(Z) = \int_R (z - \mu)^2 \cdot f(z) \cdot dz > \int_{\mathbb{D}} (z - \mu)^2 \cdot f(z) \cdot dz$$

$$\sigma^2 \geq \int_{\mathbb{D}} k^2 \cdot \sigma^2 \cdot f(z) \cdot dz = k^2 \cdot \sigma^2 \cdot \int_{\mathbb{D}} f(z) \cdot dz = k^2 \cdot \sigma^2 \cdot P(|Z - \mu| > k \cdot \sigma)$$

et l'inégalité s'en déduit. La démonstration pour une v.a. discrète, identique dans son principe, est laissée au lecteur.

Appliquons maintenant ce résultat dans le contexte de la convergence en probabilité d'une suite de v.a. Z_n vers 0. En posant alors $k \cdot \sigma = \varepsilon$, l'inégalité de Bienaymé-Tchebychev s'écrit :

$$P\left(|Z_n - E(Z_n)| > \varepsilon\right) \leq \frac{\text{var}(Z_n)}{\varepsilon^2}$$

On voit par conséquent que si la suite des moyennes $E(Z_n)$ converge vers 0, et si la suite des variances var(Z_n) converge aussi vers 0, alors on a :

$$\lim_{n \to \infty} P(|Z_n| > \varepsilon) = 0$$

ce qui montre que la suite de v.a. $\{Z_n\}$ converge en probabilité vers la v.a. certaine 0.

Il suffit alors de prendre $\{Z_n = X_n - a\}$ pour voir que si la suite des moyennes $E(X_n)$ converge vers a et la suite des variances var(X_n) converge vers 0, alors la suite des v.a. { X_n} converge en probabilité vers a.

Une suite de v.a. dont la suite des moyennes converge vers une limite a et dont la suite des variances converge vers 0, converge en probabilité vers$a$.

On notera cependant que ce dernier résultat n'est pas équivalent à la définition, et que l'on peut parler de convergence en probabilité sans supposer l'existence des moments d'ordre 1 et 2.

On peut montrer que la convergence en probabilité implique la convergence en loi, mais que la réciproque n'est pas exacte.

Le troisième mode de convergence que nous présenterons, la convergence en moyenne quadratique, est très utilisé dans les problèmes d'estimation statistique.

Soit X_n une suite de v.a. réelles de moyennes et de variances finies. On dit que la suite X_n **converge en moyenne quadratique** vers X si

$$\lim_{n \to \infty} E\left((X_n - X)^2 \right) = 0$$

Il s'agit en fait d'un cas particulier de la convergence dite **en moyenne d'ordre *p***, et définie pour des v.a. X_n telles que $E\left(|X_n - X|^p \right)$ existe, par :

$$\lim_{n \to \infty} E\left(|X_n - X|^p \right) = 0$$

Dans la convergence en moyenne d'ordre p de la suite X_n vers X, on notera l'hypothèse d'existence de :

$$E\left(|X_n - X|^p \right)$$

On montre que si la suite X_n converge en moyenne quadratique vers X, et que si la suite Y_n converge en moyenne quadratique vers Y, alors la suite X_nY_n converge en moyenne d'ordre 1 vers XY, c'est-à-dire que la suite des moyennes $E(X_nY_n)$ converge vers $E(XY)$. Ce résultat est évidemment particulièrement intéressant dans l'étude des liaisons entre variables aléatoires. Plus généralement, on montre que si $g(x, y)$ est une fonction continue en x et en y, et si X_n (resp Y_n) converge en probabilité vers X (resp. vers Y), alors $g(X_n, Y_n)$ converge en probabilité vers $g(X, Y)$.

Il s'agit d'un mode de convergence fort qui implique la convergence en probabilité.

D'autre part, il est important de noter que la convergence en moyenne d'ordre p implique la convergence en moyenne d'ordre q pour tout $q < p$. On

notera aussi que la convergence en probabilité n'implique pas la convergence en moyenne d'ordre 1, c'est-à-dire la convergence des moyennes.

L'ensemble de ces trois modes de convergence est donc hiérarchiquement ordonné.

Mais il existe d'autres modes de convergence, qu'il est plus difficile de placer dans une telle séquence hiérarchique.

Ainsi, la convergence presque sûre est définie comme suit.

La suite de v.a. réelles X_n **converge presque sûrement** vers la v.a. réelle X si on a :

$$P\left(\lim_{n \to \infty} (X_n - X) = 0\right) = 1$$

Ce mode de convergence implique aussi la convergence en probabilité, donc également la convergence en loi. Il n'est pas lié à la convergence en moyenne d'ordre p, mais les deux modes de convergence peuvent cependant exister simultanément pour une suite de v.a. réelles X_n.

Le diagramme de la figure 5.9 montre les relations que l'on peut établir entre les différents modes de convergence.

D'autres modes de convergence (dont l'étude est en dehors du cadre de cet ouvrage) sont utilisés pour obtenir certaines propriétés en théorie des probabilités, parmi lesquelles on citera :

– la convergence complète ;

– la convergence uniforme presque sûre.

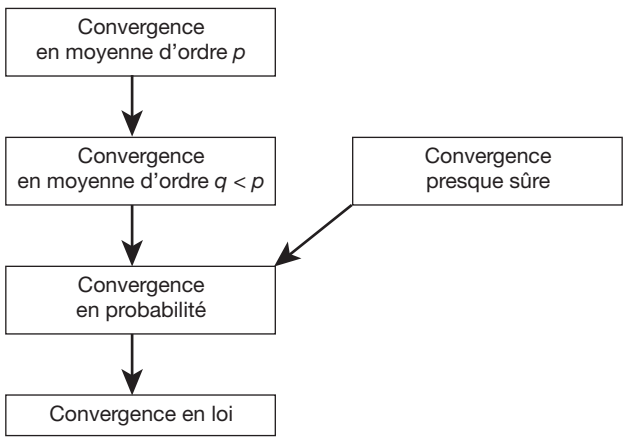


Figure 5.9 – Hiérarchie des différents modes de convergence

L'application majeure des convergences de v.a., et particulièrement de la convergence en probabilité est la **loi faible des grands nombres** :

Soient X_i $(i = 1, 2, \ldots, n)$ n v.a. réelles indépendantes d'espérances m_i et d'écarts-types σ_i toutes finies, telles que :

$$\lim_{n \to \infty} \frac{1}{n} \cdot \sum_{i=1}^{n} m_i = m \quad \text{et} \quad \lim_{n \to \infty} \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 = 0$$

alors on a : $$\frac{1}{n} \cdot \sum_{i=1}^{n} X_i \xrightarrow{P} m$$

Soit alors une épreuve aléatoire à l'issue de laquelle un résultat R peut se produire avec la probabilité p. On répète cette épreuve indépendamment n fois, et l'on désigne par F_n la variable aléatoire égale à la proportion d'épreuves ayant donné le résultat R. Ces variables sont appelées fréquences empiriques.

On applique directement la loi des grands nombres pour montrer la convergence en probabilité des fréquences empiriques vers la probabilité p. C'est le théorème de De Moivre-Laplace.

À partir de ce résultat, toute l'approche fréquentiste des probabilités (*supra*, § I) s'est développée sur l'évaluation de la probabilité d'un événement par la limite de la fréquence relative d'apparition de cet événement lorsqu'on répète indéfiniment l'épreuve aléatoire lors de laquelle il peut se réaliser.

On peut aussi démontrer un résultat plus général.

Loi forte des grands nombres

Soient X_i $(i = 1, 2, \ldots, n)$ n variables aléatoires réelles indépendantes d'espérances m_i et d'écarts-types σ_i tous finis, telles que :

$$\lim_{n \to \infty} \frac{1}{n} \cdot \sum_{i=1}^{n} m_i = m \quad \text{et} \quad \lim_{n \to \infty} \sum_{i=1}^{n} \frac{\sigma_i^2}{i^2} < \infty$$

alors on a : $$\frac{1}{n} \cdot \sum_{i=1}^{n} X_i \xrightarrow{p.s} m$$

L'une des applications de ce résultat est la généralisation du théorème central-limite sous la condition de Lindeberg (chapitre 7, § II.E).

Au total, ce sont donc toutes les bases des applications du calcul des probabilités en statistique classique qui reposent sur ces deux résultats, loi faible et loi forte des grands nombres, donc sur les différentes notions de convergence des suites de variables aléatoires.

On n'oubliera pas :

	Cas discret	Cas continu
Loi ou distribution de probabilité d'une variable aléatoire réelle	Les événements élémentaires sont : $\{X = x_i\}, x_i \in \mathbb{R}, i \in I, I \subseteq \mathbb{N}$ avec : $P(X = x_i) = p_i > 0$ et $\sum_{i \in I} p_i = 1$ La loi de probabilité est définie par les couples : $\{x_i, p_i\}, i \in I$ $\forall a, b \in \mathbb{R}, a < b$: $P(X \in [a\,;b]) = \sum_{i \in I^*} p_i$ avec $I^* = (i \in I \mid x_i \in [a\,;b])$ $P(X \in]a;b]) =$ $P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$	X prend ses valeurs dans un intervalle (ou une réunion d'intervalles) de $\mathbb{R}$, ou dans $\mathbb{R}$ tout entier *Densité de probabilité* f_X : $\begin{cases} 1.\ f(x) \geq 0 \qquad \forall x \in \mathbb{R} \\ 2.\ f \text{ presque continue partout} \\ 3.\ \int_{\mathbb{R}} f(x)dx = 1 \end{cases}$ $P(X = x) = 0$ $\forall a, b \in \mathbb{R}, a < b$: $P(X \in [a\,;b]) = P(X \in [a\,;b[) = P(X \in]a\,;b]) = P(X \in]a\,;b[)$ $= F_X(b) - F_X(a) = \int_a^b f(x)dx$
Espérance mathématique	$E(X) = \sum_{i \in I} x_i p_i$ si $\sum_{i \in I} \lvert x_i \rvert p_i < +\infty$	$E(X) = \int_{\mathbb{R}} x f(x)dx$ si $\int_{\mathbb{R}} \lvert x \rvert f(x)dx < +\infty$
Moment d'ordre k : $m_k = E(X^k)$	$m_k = \sum_{i \in I} x_i^k p_i$ si $\sum_{i \in I} \lvert x_i \rvert^k p_i < +\infty$	$m_k = \int_{\mathbb{R}} x^k f(x)dx$ si $\int_{\mathbb{R}} \lvert x \rvert^k f(x)dx < +\infty$

Testez-vous *(les réponses sont données page 286)*

Il y a *au moins* une réponse exacte par question.

1. Dans l'ensemble des classes « Terminales » d'un lycée, 14 % des élèves étudient le russe, 68 % n'étudient ni le russe ni l'espagnol, 2 % étudient ces deux langues :

a) 20 % des élèves étudient au moins l'espagnol

b) 16 % des élèves étudient seulement le russe

c) 18 % des élèves étudient seulement l'espagnol

d) 32 % des élèves étudient le russe ou l'espagnol

2. Soient deux événements A et B d'un même espace de probabilité tels que :

$$A \cap B = \varnothing$$

a) $P(A \cap B) = 0$

b) A et B sont deux événements incompatibles

c) A et B sont deux événements indépendants

d) A et B sont à la fois incompatibles et indépendants

3. Soient deux événements A et B d'un même espace de probabilité tels que :

$$P(A) = 0{,}3 \quad P(B) = 0{,}2 \quad \text{et} \quad P(A \cap B) = 0{,}09$$

a) $P(A|B) = 1{,}50$ et $P(B|A) = 0{,}60$

b) $P(A|B) = 0{,}30$ et $P(B|A) = 0{,}45$

c) $P(A|B) = 0{,}45$ et $P(B|A) = 0{,}30$

d) $P(A|B) = 0{,}27$ et $P(B|A) = 0{,}18$

4. Soient deux événements indépendants A et B d'un même espace de probabilité tels que : $P(A) = 0{,}3$ **et** $P(B) = 0{,}2$

a) $P(A \cap B) = 0{,}5$

b) $P(A \cap B) = 0{,}06$

c) $P(A \cup B) = 0{,}06$

d) $P(A \cup B) = 0{,}44$

5. Trois chasseurs visent simultanément un même lièvre et tirent en même temps. Soient p_1, p_2, p_3 les probabilités respectives de toucher le lièvre pour chaque chasseur, alors la probabilité que le lièvre soit touché par au moins un des chasseurs :

a) peut être inférieure à p_1

b) est égale à $(p_1 + p_2 + p_3)$

c) est égale à $(1 - (1 - p_1)(1 - p_2)(1 - p_3))$

d) est comprise entre $(p_1 \cdot p_2 \cdot p_3)$ et $(p_1 + p_2 + p_3)$

6. Soient deux événements A et B d'un même espace de probabilité tels que :

$$P(A) = 0{,}6 \quad P(B) = 0{,}5 \quad \text{et} \quad P(A \cap B) = 0{,}1$$

a) les événements A et B sont indépendants

b) les événements A et B sont incompatibles

c) l'événement $A \cup B$ est certain

d) $P(A \mid B) = 0{,}2$

7. Si X est une variable aléatoire continue, on a, quelque soient les nombres réels a et b :

a) $P(X = a) = 0$

b) $P(a < X < b) = P(a < X \leq b)$

c) $P(a < X < b) \neq P(a \leq X < b)$

d) $P(X > a) = 1 - P(X < a)$

8. Une fonction de répartition :

a) est une fonction strictement croissante

b) est définie sur tout $\mathbb{R}$

c) prend ses valeurs dans l'intervalle [0 ; 1]

d) est toujours continue et dérivable

9. La loi de probabilité d'une variable aléatoire :

a) est entièrement définie par la fonction de répartition

b) est entièrement définie par la fonction de densité

c) est entièrement définie par l'espérance mathématique et la variance

d) est associée à un espace probabilisé

10. L'espérance mathématique d'une variable aléatoire réelle :

a) est toujours égale à l'une des valeurs possibles de la variable aléatoire

b) est un nombre réel

c) est égale à la médiane si la distribution de probabilité est symétrique

d) existe toujours si la variable aléatoire est discrète

11. Soient X une variable aléatoire, a et b deux nombres réels :

a) $E(aX + b) = aE(X) + b$

b) $\text{var}(-X + b) = -\text{var}(X) + b$

c) $P(X > E(X)) = 0{,}5$

d) $Y = aX + b \implies F_Y(y) = F_X\left(\dfrac{y - b}{a}\right)$

12. La loi jointe des deux variables aléatoires X et Y est donnée dans le tableau suivant :

X \ Y	0	1	2
1	0,10	0,20	0,10
2	0,15	0,30	0,15

a) X et Y sont indépendantes

b) $P(Y = 2 \mid X = 1) = 1/4$

c) $\rho(X,Y) = +1$

d) $E(Y) = 1$

13. Soient deux variables aléatoires X et Y liées par la relation $X - 2Y = 1$:

a) $E(X) = 2E(Y) + 1$

b) $\text{var}(X) = 2\text{var}(Y)$

c) $\rho(X,Y) = +1$

d) X et Y sont indépendantes

14. Soient deux variables aléatoires X et Y telles que $\text{var}(X) = 144$, $\text{var}(Y) = 81$ et $\text{var}(X + Y) = 25$

a) $\text{cov}(X,Y) = -100$

b) $\rho(X,Y) = 0$

c) $\text{var}(X - Y) = 425$

d) X et Y sont liées par une relation linéaire

15. Soit un couple de v.a. (X,Y) pour lequel on dispose des lois conditionnelles de X pour chaque valeur possible de Y et de la loi marginale en Y :

X \ Y	1	2	3
1	0,2	0	0,5
2	0,8	1	0,5

et de la loi marginale de Y :

Valeur de Y	1	2	3
Probabilité	0,5	0,3	0,2

a) disposant de cette information, on peut calculer la loi du couple

b) la distribution de la v.a. $\{ X \mid Y = 3 \}$ est symétrique

c) la v.a. $\{ X \mid Y = 2 \}$ est certaine

d) $P(X = 1) = 0,7$

16. La loi jointe des deux variables aléatoires X et Y est donnée dans le tableau suivant :

X \ Y	**0**	**1**	**2**
0	**0,15**	**0,20**	**0,15**
1	**0,10**	**0,25**	**0,15**

a) le coefficient d'asymétrie de la v.a. X est nul

b) var(X) = 0,25

c) $E(Y|X = 0) = 1$

d) X et Y sont indépendantes

17. La loi jointe d'un couple (X, Y) de variables aléatoires discrètes finies :

a) est entièrement spécifiée par le tableau donnant les x_i , y_j et p_{ij}

b) est entièrement spécifiée par les k lois conditionnelles $\{ Y|X = x_i\}$

c) est entièrement spécifiée par les l lois conditionnelles $\{ X|Y = y_j\}$

d) est entièrement spécifiée par les espérances, variances de X et Y et leur covariance

Exercices *(corrigés page 315)*

Exercice 5.1

Dans un club sportif, il y a 75 adultes (dont 45 femmes) et 45 enfants (dont 25 filles). On interroge au hasard un adhérent du club. Quelle est la probabilité que cet adhérent :

– soit un adulte ;

– soit de sexe masculin ;

– soit une femme adulte ;

– soit un adulte ou soit de sexe féminin.

Exercice 5.2

Après une enquête auprès d'une population, on sait que 40 % des individus ne sont jamais allés en Espagne et que 55 % des individus n'ont jamais pris l'avion, mais que 25 % ont été en Espagne et ont déjà pris l'avion.

Quelle est la probabilité qu'un individu tiré au hasard dans cette population ne soit pas allé en Espagne et n'ait jamais pris l'avion ?

Exercice 5.3

Une enquête exhaustive sur un campus universitaire montre que sur les 32 564 étudiants, 23 522 lisent la revue *Notre campus* publiée par l'Université, 18 859 lisent la revue *La Vie étudiante* publiée par le BDE, et 11 422 étudiants lisent *Notre campus* et *La Vie étudiante*.

1. On interroge au hasard un étudiant du campus. Calculez la probabilité que cet étudiant :

– ne lise ni *Notre campus*, ni *La Vie étudiante* ;

– lise *Notre campus* et ne lise pas *La Vie étudiante*.

2. On interroge au hasard deux étudiants du campus et on admet que leurs réponses sont indépendantes. Calculez la probabilité

– que les deux étudiants ne lisent aucune des deux revues ;

– qu'un étudiant lise les deux revues et que le second n'en lise aucune.

Exercice 5.4

On lance n fois une pièce de monnaie, on suppose que la probabilité d'obtenir pile est égale à la probabilité d'obtenir face. Soient A et B les événements suivants :

A = « obtenir au plus une fois pile »

B = « obtenir au moins une fois pile et au moins une fois face »

1. Calculez $P(A)$, $P(B)$ et $P(A \cap B)$ pour $n = 2$; A et B sont-ils indépendants pour $n = 2$?

2. Même question pour $n = 3$.

Exercice 5.5

Calculez la probabilité qu'il y ait 3 filles et 2 garçons dans une famille de 5 enfants :

1. Si on suppose la probabilité de naissance d'une fille égale à la probabilité de naissance d'un garçon

2. Si on suppose la probabilité de naissance d'une fille égale à 0,48

Exercice 5.6

La demande journalière X d'un bien fabriqué par une entreprise est une v.a. qui suit la loi suivante : $P(X = 0) = 1/6$ $P(X = 1) = 1/6$ $P(X = 2) = 1/2$ $P(X = 3) = 1/6$. On suppose que le profit, fonction de la demande et du coût, vérifie la relation : $\Pi(X) = p.X - C$, p étant le prix unitaire du bien fixé à 600 €, C étant le coût supposé indépendant de la demande et égal à 800 €.

1. Calculez l'espérance et l'écart-type du profit. Quelle est la signification de l'espérance du profit ?

2. Déterminez la fonction de répartition du profit et tracez son graphe.

Exercice 5.7

Une compagnie d'assurances admet pour l'année à venir et pour un certain type de contrat, que 60 % des assurés n'auront pas de sinistre. Par ailleurs on suppose que le coût moyen de règlement des accidents est de 500 € avec une probabilité de 0,25, de 1 500 € avec une probabilité de 0,1, de 2 500 € avec une probabilité de 0,05. Un assuré déclare au plus un sinistre de ce type dans l'année.

1. Pour espérer un bénéfice moyen de 50 € par assuré, quel doit être le montant de la cotisation ?

2. Quelle est la probabilité pour que le coût de règlement total de deux assurés pris au hasard n'excède pas le montant encaissé de leurs cotisations (au tarif déterminé au 1) ?

Exercice 5.8

Dans une banque, un système de guichet automatique a été mis en place et permet de faire des opérations bancaires courantes : extrait de compte, remise de chèque, retrait. Le nombre de clients utilisant le guichet automatique dans un intervalle de temps de 5 minutes est une v.a. X telle que :

$$P(X = 0) = 0{,}3,\ P(X = 1) = 0{,}3 \text{ et } P(X = 2) = 0{,}4$$

1. Calculez $E(X)$ et var(X).

2. On suppose que les nombres de clients utilisant le guichet automatique sur deux périodes de 5 minutes ne se chevauchant pas sont indépendants. Soit Y la v.a. égale au nombre de clients utilisateurs sur une période d'une heure. La v.a. Y peut s'écrire :

$$Y = \sum_{i=1}^{12} X_i$$

où X_i désigne le nombre de clients utilisateurs au cours de i^e intervalle de 5 minutes lorsqu'on découpe l'heure en 12 intervalles de 5 minutes ; chaque X_i suit la même loi que X.

Quelles sont les valeurs possibles de Y ?

Calculez $E(Y)$, var(Y) et $P(Y = 0)$.

3. Chaque client ne peut effectuer plus de 2 opérations au guichet automatique. La banque a constaté que chaque client effectue :

3 fois sur 10 : 2 opérations

6 fois sur 10 : 1 opération

1 fois sur 10 : 0 opération (compte non approvisionné, par exemple)

Soit Z, le nombre d'opérations effectuées dans un intervalle de temps de 5 minutes.

3.1. Donnez dans un tableau à double entrée l'ensemble des probabilités conditionnelles de Z sachant X.

3.2. Quelle est la loi de Z ? Calculez $E(Z)$ et var(Z).

Exercice 5.9

Une usine de pellicules de photo dispose de trois machines A, B et C qui fabriquent respectivement 20 %, 50 % et 30 % de la production totale. Les proportions de pellicules défectueuses fabriquées par les machines A, B ou C sont respectivement égales à 6 %, 5 % et 3 %.

On tire au hasard une pellicule dans la production, calculez :

– la probabilité que cette pellicule soit défectueuse ;

– la probabilité qu'elle provienne de la machine A sachant qu'elle est défectueuse ;

– la probabilité qu'elle provienne de la machine A sachant qu'elle est non défectueuse.

Exercice 5.10

Un couple (X, Y) de variables aléatoires suit la loi jointe donnée dans le tableau suivant :

X \ Y	u	0	1
0	1/4	a	1/8
1	1/5	b	1/10

u, a et b étant des valeurs réelles.

1. Pouvez-vous déterminer a et b de telle sorte que les variables aléatoires X et Y soient indépendantes en probabilité ?

2. Dans ces conditions, déterminez la loi marginale de X, et les lois conditionnelles de X pour les différentes valeurs de Y.

3. Si $a = 1/5$, existe-t-il une valeur de u telle que le coefficient de corrélation linéaire $\rho(X, Y)$ soit nul ? Les variables aléatoires X et Y sont-elles alors indépendantes en probabilité ?

Exercice 5.11

Soient deux variables aléatoires X et Y : X prend les valeurs 0 et 1 avec les probabilités 1/2 et 1/2, Y prend les valeurs 0 et 2 avec les probabilités 1/3 et 2/3. On note : $P(X = 0 \text{ et } Y = 0) = p$.

1. Calculez, en fonction de p, les probabilités suivantes :

$P(X = 0 \text{ et } Y = 2)$ $P(X = 1 \text{ et } Y = 0)$ et $P(X = 1 \text{ et } Y = 2)$

Entre quelles limites peut varier p ?

2. Calculez, en fonction de p, le coefficient de corrélation linéaire $\rho(X, Y)$.

6. Les principaux modèles statistiques discrets

Notion de modèle

Par modèle on entend une représentation simplifiée d'un processus, d'un système.
Dans les domaines des sciences économiques et de gestion, on cherche à disposer de modèles pour analyser, prévoir et décider. La nature même des facteurs intervenant en gestion et en économie explique le caractère aléatoire, c'est-à-dire non déterministe, donc non contrôlable totalement du modèle qu'on cherche à définir pour représenter le système étudié.
Dans la plupart des cas, on dispose d'un ensemble fragmentaire de données à partir desquelles on cherche une représentation globale. C'est là une des démarches classiques en statistique, déduire des informations fournies par un échantillon une ou plusieurs caractéristiques concernant la population d'où l'on extrait l'échantillon ; il s'agit là de l'*inférence statistique*.
La construction d'un modèle est destinée donc à analyser, prévoir ou décider à partir d'un support rigoureux et fiable ; sa recherche est ainsi un travail formel. Pour l'aborder il est nécessaire de définir avec précaution tous les éléments dont on dispose :
– la *population* pour laquelle le modèle est destiné ;
– l'*individu*, ou unité élémentaire de la population ;
– le *caractère* étudié sur chacun des individus, et qui définit le phénomène étudié ;
– la nature de ce caractère (qualitatif, quantitatif, discret ou continu).
À partir de là, on peut associer – par une démarche analogue à celle vue en statistique descriptive – une variable aléatoire à chaque individu de la population. C'est cette variable aléatoire et sa distribution de probabilité qui vont constituer les éléments du modèle ; on dit que cette variable aléatoire est la variable générique de la population (on dit aussi

variable parente) puisque tout individu – tant qu'on ne connaît pas ses caractéristiques individuelles – peut être représenté par une variable aléatoire de même loi qu'elle. Il sera alors possible d'étudier un ensemble d'individus extrait de la population générale comme un ensemble de variables aléatoires ayant toutes comme loi, la loi de la variable *générique* de la population. Lorsque ces variables sont indépendantes entre elles, on dit qu'elles forment un échantillon de la variable *parente* ; cette condition d'indépendance est équivalente à un tirage avec remise des individus formant l'échantillon au sein de la population.

Modèles empiriques (ou expérimentaux)

Ce sont des modèles qui sont construits sur l'observation d'une série statistique. Leur validité dépend tout particulièrement de la taille de la série statistique des observations. On recherche ici les caractéristiques essentielles de la série observée (moyenne, médiane, mode, quartiles, symétrie ou non…). Parmi les représentations en lois de probabilité connues, on en cherche une qui soit cohérente avec les données observées, du point de vue de ces caractéristiques. On procède par analogie.

Modèles théoriques (ou analytiques)

On étudie le phénomène en essayant de le décomposer en composantes élémentaires directement représentées et de façon naturelle par une loi de probabilité (telle que la loi de Bernoulli ou la loi uniforme).
Le schéma binomial comme le schéma hypergéométrique (*infra* § II.B et II.C), ou encore la loi géométrique (§ II.D) sont des exemples de cette approche.

Classification des modèles

On doit distinguer les *modèles discrets* pour lesquels les diverses occurences sont ponctuelles et parfaitement bien isolées (séparées) les unes des autres, des *modèles continus* pour lesquels les occurences sont beaucoup trop nombreuses pour pouvoir être isolées ponctuellement et ne peuvent être étudiées que par classes de valeurs. À l'intérieur des modèles discrets, on distingue encore les modèles discrets finis (c'est-à-dire dont le domaine des valeurs est de cardinal fini) des modèles discrets infinis dénombrables.
Il existe d'autres classifications mais qui concernent des modèles qui ne sont pas abordés dans ce cours du fait de leur plus grande complexité et de leur utilisation moins fréquente.
De très nombreux modèles (discrets ou continus) ont été construits pour correspondre à des situations pratiques déterminées. Nous présentons dans ce chapitre et dans le suivant ceux qui sont le plus fréquemment utilisés, mais bien entendu il ne faudra pas croire que tout phénomène puisse être rapporté aux quelques modèles décrits ici.

I. Les modèles élémentaires

A. Le schéma de Bernoulli

Toute épreuve aléatoire n'ayant que deux résultats possibles peut être considérée comme une situation d'alternative : si l'un des deux résultats ne se réalise pas, c'est que l'autre le sera obligatoirement. En d'autres termes, dans une telle situation, les deux résultats possibles sont complémentaires l'un de l'autre, la somme de leurs probabilités étant égale à 1.

Il s'agit là d'une situation extrêmement fréquente puisque dès qu'on cherche à mettre en évidence la présence d'un caractère particulier pour les individus d'une population, tout individu de cette population peut être décrit selon une telle alternative : ou bien il présente ce caractère ou bien il ne le présente pas.

Ainsi par exemple lorsqu'on cherche à évaluer l'impact d'une campagne publicitaire sur les achats d'un nouveau produit, on peut associer à chaque individu sondé (parmi ceux ayant acquis ce produit après la campagne publicitaire) trois variables aléatoires :

- la première met en évidence si l'individu possédait déjà auparavant ce produit ;
- la seconde met en évidence si l'individu a été touché par la campagne publicitaire ;
- la troisième décrit si l'acquisition du produit a été induite par la campagne publicitaire.

Il s'agit là d'une possibilité de formalisation (et bien entendu ce n'est pas la seule !), mais chacune de ces trois variables correspond bien à une situation d'alternative. L'étude des effets éventuels de cette campagne publicitaire met en œuvre les outils appropriés de l'analyse statistique.

Dans ces situations de dualité, l'une des deux issues est celle que privilégie l'étude, elle correspond à la positivité d'un index, à la présence du caractère pour chaque individu de la population faisant l'objet de l'étude, par opposition à son absence. Les aléas qu'on peut définir dans ces cas étant des aléas qualitatifs, il faut trouver le codage le plus approprié. C'est cet aspect de présence/absence qui l'impose, et on code par 0 et 1 les deux issues possibles, *celle qu'on cherche à mettre en évidence étant codée* 1.

On définit ainsi une variable aléatoire qui ne peut prendre que deux valeurs, à savoir 0 et 1. Elle porte alors le nom de *variable aléatoire de Bernoulli*[1], et possède alors une loi de probabilité très simple pour laquelle *p*

1. Jacques Bernoulli (1654-1705), scientifique suisse a beaucoup contribué au développement du calcul des probabilités (loi des grands nombres) et aux statistiques.

représente la probabilité de l'issue qu'on veut mettre en évidence (notation conventionnelle). On note souvent $q = 1 - p$ la probabilité de l'autre terme de l'alternative. Le terme de variable aléatoire de Bernoulli est synonyme de celui de *variable aléatoire indicatrice* (indiquant la réalisation éventuelle de l'événement de probabilité p). Il faut bien se souvenir qu'une variable de Bernoulli est définie par les 2 valeurs 0 et 1 (et celles-là seulement ; toute autre paire de valeurs ne permet plus l'appellation de variable de Bernoulli ; ceci se justifie comme on le verra dans la suite pour la construction des modèles binomial, hypergéométrique et de Pascal). Le tableau de la loi de probabilité d'une telle variable est parfaitement connu dès que p l'est. *La loi de Bernoulli dépend du seul paramètre* p.

Valeur de X	0	1
Probabilité	$q = 1 - p$	p

Le diagramme en bâtons et le graphe de la fonction de répartition d'une variable de Bernoulli (*cf.* figure 6.1) sont particulièrement simples.

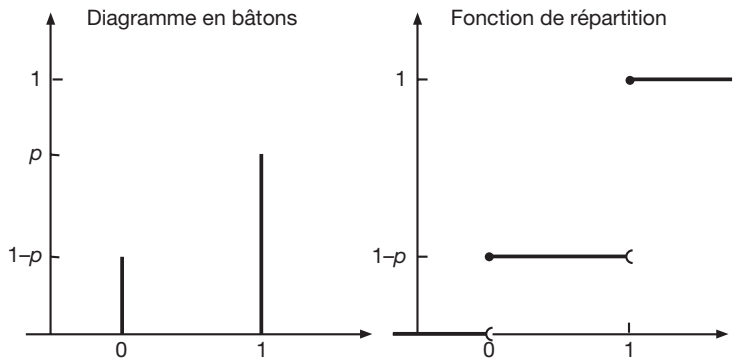


Figure 6.1 – Loi de Bernoulli

L'espérance d'une variable de Bernoulli de paramètre p est égale à p. En effet :

$$E(X) = 0 \cdot (1 - p) + 1 \cdot p = p$$

Le moment d'ordre 2 est égal aussi à p, puisque :

$$E(X^2) = 0^2 \cdot (1 - p) + 1^2 \cdot p = p$$

Par conséquent, la variance est égale à pq :

$$\text{var}(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1 - p) = pq$$

On remarquera au passage que la fonction $x(1 - x)$ dont la dérivée est égale à $(1 - 2x)$ a un maximum pour $x = 1/2$, maximum égal à 1/4. Par conséquent,

la variance d'une variable de Bernoulli est au plus égale à 1/4, et l'écart-type est au plus égal à 1/2.

En conclusion, on retiendra que toute situation aléatoire d'alternative peut être représentée par une variable de Bernoulli dont le paramètre p, égal à la probabilité de l'issue qu'on cherche à mettre en évidence, est égal à l'espérance, la variance étant égale à $p(1-p)$.

B. La loi uniforme discrète

Un exemple particulier de loi de Bernoulli est celui pour lequel $p = q = 1/2$. Dans ce cas, les deux termes de l'alternative pouvant se présenter à l'issue de l'épreuve aléatoire sont équiprobables. *Cette situation d'équiprobabilité correspond souvent à des situations dans lesquelles on ne dispose d'aucune information permettant de mieux appréhender l'événement auquel on s'intéresse.*

La loi uniforme discrète en est la généralisation. On suppose cette fois que l'expérience aléatoire possède k issues distinctes, possédant chacune la même chance d'être réalisée. On définit alors dans ce contexte une variable aléatoire X pouvant prendre toutes les valeurs entières comprises entre 1 et k, chacune de ces valeurs étant associée à l'une des k issues de l'épreuve aléatoire. On peut donc écrire d'une part :

$$\sum_{i=1}^{k} P(X = i) = P\left(\bigcup_{i=1}^{k} (X = i)\right) = 1$$

et d'autre part, $P(X = i)$ étant constante, on peut la désigner par p.

On en déduit :

$$1 = \sum_{i=1}^{k} P(X = i) = \sum_{i=1}^{k} p = k \cdot p$$

et la probabilité commune p est égale à $1/k$

La loi de probabilité de cette variable aléatoire est résumée dans le tableau suivant :

Valeur de X	1	2	…	k
Probabilité	$1/k$	$1/k$	…	$1/k$

On déduit les caractéristiques essentielles :

$$E(X) = \sum_{i=1}^{k} i \cdot \frac{1}{k} = \frac{1}{k} \sum_{i=1}^{k} i = \frac{1}{k} \cdot \frac{k(k+1)}{2} = \frac{k+1}{2}$$

autrement dit, l'espérance de cette variable aléatoire se situe à l'exact milieu des valeurs possibles. Ce résultat est tout à fait naturel compte tenu de l'équiprobabilité.

D'autre part :

$$E(X^2) = \sum_{i=1}^{k} i^2 \cdot \frac{1}{k} = \frac{1}{k}\sum_{i=1}^{k} i^2 = \frac{1}{k} \cdot \frac{k(k+1)(2k+1)}{6} = \frac{(k+1)(2k+1)}{6}$$

d'où l'expression de la variance :

$$\text{var}(X) = E(X^2) - [E(X)]^2 = \frac{(k+1)(2k+1)}{6} - \frac{(k+1)^2}{4} = \frac{k^2-1}{12}$$

En ce qui concerne ces résultats, on notera qu'ils sont obtenus pour une loi uniforme discrète dont les valeurs sont les entiers compris entre 1 et k (au sens large). Par conséquent, ils peuvent s'appliquer à toute situation aléatoire à k issues équiprobables dès que celles-ci peuvent être codées par les nombres 1, 2,… ,k. Si on doit adopter un autre codage, les valeurs de l'espérance et de la variance (comme de tout autre moment) seront modifiées puisqu'elles dépendent des valeurs possibles de la variable aléatoire.

On peut encore donner la fonction génératrice des moments de cette variable aléatoire uniforme discrète. On a :

$$g_X(u) = E(u^X) = \sum_{i=1}^{k} u^i \cdot \frac{1}{k}$$

ce qui permet d'obtenir les moments successifs, mais aussi de retrouver les résultats précédents. On constate aussi par ailleurs que les moments factoriels d'ordre strictement supérieur à k sont nuls : $\mu_{[n]}(X)$ dès que $n > k$

On ajoutera simplement pour terminer que le diagramme en bâtons de cette loi est formé de bâtons de même hauteur, et que le graphe de la fonction cumulative est formé de marches d'escalier également espacées (l'espace entre deux d'entre elles étant de $1/k$) et de même largeur (l'unité).

II. Les schémas de Bernoulli itératifs

Le schéma de Bernoulli est le plus simple des modèles probabilistes, cependant il est fondamental. Ceci est dû au fait que le plus grand nombre de situations aléatoires peuvent se décomposer en successions d'épreuves élémentaires de Bernoulli. On n'envisagera ici que la situation où le résultat du phénomène complexe initial est égal à la somme des résultats des épreuves élémentaires de Bernoulli.

Dans ce cadre, on étudiera une succession d'épreuves de Bernoulli :

– en nombre fixé, et avec indépendance $\Rightarrow$ schéma binomial,

– en nombre fixé et sans indépendance $\Rightarrow$ schéma hypergéométrique,

– en nombre aléatoire, jusqu'à ce que l'on ait obtenu pour la 1re fois l'issue recherchée de l'alternative ainsi répétée $\Rightarrow$ schéma géométrique,

– en nombre aléatoire, jusqu'à ce que l'on ait obtenu pour la $k^{\text{ème}}$ fois l'issue recherchée de l'alternative ainsi répétée $\Rightarrow$ schéma de Pascal.

Les deux premiers cas sont de nature totalement différente des deux derniers, car le nombre des itérations du modèle élémentaire de Bernoulli est, pour les premiers, connu au départ, alors qu'il est la quantité aléatoire pour les derniers.

A. Le schéma binomial

Définition

Une variable aléatoire est dite suivre une loi binomiale de paramètres n et p, notée $\mathcal{B}(n\,;\,p)$, si elle peut être considérée comme la somme de n variables aléatoires de Bernoulli, indépendantes et de même paramètre p.

Soit par exemple, une population dans laquelle une proportion p d'individus présente un caractère donné. On se pose la question de savoir si un échantillon[1] de n individus choisis au hasard dans la population a de grandes chances de contenir k individus ayant le caractère.

Chaque individu de la population (et donc de l'échantillon) est présent dans ce problème par une alternative : il possède le caractère étudié ou non. Il est parfaitement justifié de lui associer une variable de Bernoulli prenant la valeur 1 s'il a le caractère étudié, et la valeur 0 sinon. Cette variable ainsi définie pour chaque individu est la *variable générique* de la population (ou encore la *variable parente*). Si on la note X, on a $P(X = 1) = p$, et donc aussi $P(X = 0) = 1 - p = q$. Les n individus $(1, 2\ldots, n)$ de l'échantillon seront ainsi représentés par n variables de Bernoulli $X_1, X_2,\ldots, X_n$ ayant toutes la même loi de probabilité, celle de X, une loi de Bernoulli de paramètre p. On peut supposer toutes ces variables indépendantes pour la simplicité du problème, ce qui correspond par exemple à un tirage des n individus avec remise, ou bien à un taux de sondage n / N inférieur à 10 %, N étant la taille de la population (ce point important sera revu au § II.C avec la loi hypergéométrique).

1. Ce terme d'échantillon se réfère à la fois au sens usuel, et également à une collection de variables aléatoire indépendantes et de même distribution.

Considérons la variable aléatoire Y, somme des n v.a. X_i :

$$Y = X_1 + X_2 + \ldots\ldots + X_n$$

Les réalisations de cette variable aléatoire étant des sommes de 0 et de 1, sont des nombres entiers compris entre 0 et n.

La réalisation de la v.a Y associée à un échantillon donné représente le nombre d'individus qui possèdent le caractère étudié dans l'échantillon .

C'est ce type de construction par itération d'un processus de Bernoulli, le nombre d'itérations étant fixé et les épreuves étant indépendantes, qu'on appelle schéma binomial.

La loi de probabilité de la variable somme Y est définie par :

– les valeurs susceptibles d'être prises, ici les valeurs entières comprises entre 0 et n

– les probabilités correspondant à ces valeurs :

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

En effet, à chaque groupe de k individus associés à la valeur 1 (possédant le caractère étudié), correspond un groupe formé de ($n-k$) individus associés à la valeur 0. La probabilité de réalisation d'une telle situation (k fois présence du caractère et ($n-k$) fois son absence) s'obtient en multipliant les probabilités associées aux réalisations des variables de Bernoulli correspondant à chaque individu (ces variables étant indépendantes, les événements le sont aussi) :

$$\underbrace{p \cdot p \ldots p}_{k\ fois} \cdot \underbrace{(1-p) \cdot (1-p) \ldots (1-p)}_{(n-k)\ fois} = p^k (1-p)^{n-k}$$

Il y a exactement $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ façons d'isoler k individus parmi les n de l'échantillon (les k premiers, les ($k-1$) premiers et le dernier, …, les k derniers), donc d'obtenir une somme égale à k, chacun de ces « assemblages » étant incompatible avec l'un quelconque des autres puisqu'au moins une paire d'individus passe d'un état à l'autre. La probabilité que la somme Y prenne la valeur k, sans tenir compte du rang des X_i prenant la valeur 1 à condition qu'il y en ait k et k seulement, est ainsi l'addition de $\binom{n}{k}$ fois la probabilité $p^k (1-p)^{n-k}$

Le tableau suivant présente la loi de probabilité binomiale $\mathcal{B}(n\ ;\ p)$:

Valeur de Y	0	1	2	…	k	…	n
Probabilité	$(1-p)^n$	$np(1-p)^{n-1}$	$\binom{n}{2} p^2 (1-p)^{n-2}$	…	$\binom{n}{k} p^k (1-p)^{n-k}$	…	p^n

Les caractéristiques d'une loi binomiale $\mathcal{B}(n\,;\,p)$ sont très faciles à calculer si on utilise la décomposition en somme de variables de Bernoulli indépendantes. En effet :

$$\begin{aligned} E(Y) &= E(X_1 + X_2 + \ldots + X_n) \\ &= E(X_1) + E(X_2) + \ldots + E(X_n) \\ &= n \cdot E(X) \end{aligned}$$

et par conséquent :

$$E(Y) = np$$

Pour le calcul de la variance, la propriété d'additivité, toujours vraie pour l'espérance, suppose que les variables de Bernoulli X_i sont indépendantes, et cette hypothèse est fondamentale pour la validité du résultat :

$$\begin{aligned} \text{var}(Y) &= \text{var}(X_1 + X_2 + \ldots + X_n) \\ &= \text{var}(X_1) + \text{var}(X_2) + \ldots + \text{var}(X_n) \\ &= n \cdot \text{var}(X) = np(1-p) \end{aligned}$$

On obtient le résultat :

$$\text{var}(Y) = npq$$

On pourra comparer ce dernier résultat avec celui du § II.B obtenu pour une loi hypergéométrique, cas d'une somme de variables aléatoires de Bernoulli non indépendantes.

La propriété suivante est intéressante en pratique.

Propriété 1

Si Y et Z sont deux variables aléatoires indépendantes, respectivement distribuées selon des lois binomiales $\mathcal{B}(n_1\,;\,p)$ et $\mathcal{B}(n_2\,;\,p)$, leur somme $Y + Z$ suit une loi binomiale $\mathcal{B}(n_1 + n_2\,;\,p)$

En effet, Y étant la somme de n_1 variables de Bernoulli indépendantes de même paramètre p, et Z étant la somme de n_2 variables de Bernoulli indépendantes de même paramètre p, la v.a. $Y + Z$ est la somme de $(n_1 + n_2)$ variables aléatoires de Bernoulli indépendantes de même paramètre p, et suit une loi binomiale $\mathcal{B}(n_1 + n_2\,;\,p)$.

Une seconde propriété très utilisée est la suivante.

Propriété 2

Si Y suit une loi $\mathcal{B}(n\,;\,p)$, alors $n - Y$ suit une loi $\mathcal{B}(n\,;\,1-p)$

En effet, nous avons vu que Y représente le nombre de fois, sur n individus, où on a observé l'un des termes de l'alternative, celui de probabilité p. Il s'ensuit que $(n - Y)$ est le nombre des autres résultats, ceux correspondant à une probabilité élémentaire $(1 - p)$. Or, il serait tout à fait possible de recoder les deux termes de l'alternative, en définissant une nouvelle variable de Bernoulli prenant la valeur 1 avec la probabilité $(1 - p)$ et la valeur 0 avec la probabilité p. La somme de ces n nouvelles variables aléatoires représente de nombre d'épreuves (parmi les n réalisées) qui donnent le résultat de probabilité $(1 - p)$, c'est-à-dire la variable aléatoire $(n - Y)$ que nous étudions. C'est une somme de variables aléatoires de Bernoulli, indépendantes et de même paramètre $(1 - p)$. D'où le résultat annoncé.

La construction du schéma binomial par les variables de Bernoulli justifie d'autre part la notation $\mathcal{B}(1 ; p)$ adoptée parfois pour désigner un aléa de Bernoulli de paramètre p.

Une variante de la loi binomiale est la loi dite *binomiale en proportion*. On a vu qu'une loi binomiale caractérise le nombre de résultats codés 1 dans une succession d'épreuves de Bernoulli (dont le nombre est fixé à l'avance) indépendantes. Dans un certain nombre de circonstances, on s'intéresse plutôt à la proportion des résultats codés 1. Or si Y est le nombre des résultats codés 1 dans une suite de n épreuves de Bernoulli indépendantes, Y/n est la *fréquence relative* ou *proportion*.

Lorsque Y prend une valeur quelconque k comprise entre 0 et n, Y/n prend la valeur k/n et réciproquement. Les deux événements équivalents $\{Y = k\}$ et $\{Y/n = k/n\}$ ont ainsi la même probabilité. La loi de Y/n est définie par ses valeurs et les probabilités correspondantes :

Valeur de Y/n	0	$1/n$	$2/n$	…	k/n	…	1
Probabilité	$(1-p)^n$	$np(1-p)^{n-1}$	$\binom{n}{2}p^2(1-p)^{n-2}$	…	$\binom{n}{k}p^k(1-p)^{n-k}$	…	p^n

Le tableau de cette loi de probabilité se déduit de celui d'une loi binomiale en divisant simplement chaque valeur possible par n.

Le diagramme en bâtons et la fonction de répartition d'une loi $\mathcal{B}(n ; p)$ dépendent des 2 paramètres n et p. Le cas particulier où $p = 0{,}5$ correspond à l'équiprobabilité des deux termes de l'alternative de base (présence/absence) et se traduit graphiquement par une symétrie du diagramme en bâtons (*cf.* figure 6.2).

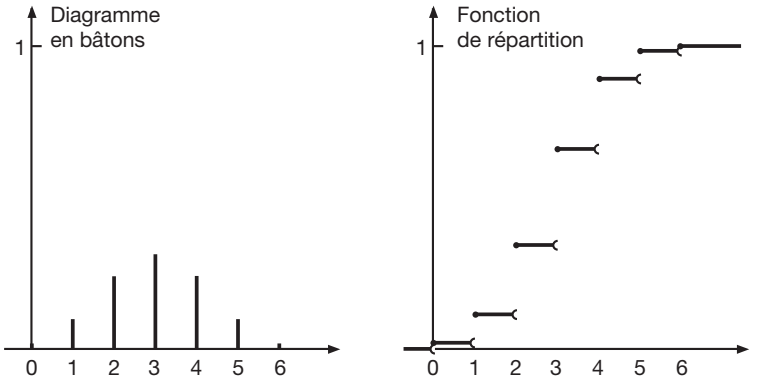


Figure 6.2 – Loi binomiale $\mathcal{B}(6\ ;\ 0{,}5)$

Les calculs relatifs aux distributions binomiales peuvent se faire à l'aide de tables statistiques (*cf.* annexe IV). Ces tables donnent pour quelques valeurs de n et de p, les probabilités cumulées de telles répartitions. Le nombre des valeurs de n et de p envisagées est forcément très limité. Grâce à la propriété 2, on peut déduire les probabilités d'une loi $\mathcal{B}(n\ ;\ 1-p)$ de celles d'une loi $\mathcal{B}(n\ ;\ p)$. Au lieu de recourir à des interpolations linéaires (parfois causes d'importantes erreurs d'approximation), on utilisera plutôt la formule de récurrence suivante (rappelée à l'annexe II), entre les probabilités de deux valeurs successives k et $(k+1)$ d'une distribution binomiale $\mathcal{B}(n\ ;\ p)$:

$$\frac{P(X=k+1)}{P(X=k)} = \frac{\binom{n}{k+1}p^{k+1}(1-p)^{n-k-1}}{\binom{n}{k}p^{k}(1-p)^{n-k}} = \frac{\dfrac{n!}{(k+1)!(n-k-1)!}p}{\dfrac{n!}{k!(n-k)!}(1-p)}$$

$$= \frac{(n-k)p}{(k+1)(1-p)}$$

Cette formule permet de calculer successivement les probabilités individuelles, en partant de $P(X=0) = (1-p)^n$

▶ **Exemple**

Après une élection à deux candidats A et B, c'est A qui l'emporte avec un score de 52 %. On suppose que le nombre d'électeurs qui se sont exprimés est élevé.

On cherche à déterminer la probabilité qu'un sondage préélectoral portant sur 50 électeurs ait donné une majorité de suffrages pour B (c'est-à-dire un résultat à l'opposé de la réalité des intentions de vote de la population).

Ce problème doit être modélisé en définissant la population, celle des électeurs, puis l'individu, un électeur et le caractère étudié, le bénéficiaire du vote : soit A, soit B (B équivalant à « non A »).

Le caractère est un caractère qualitatif à deux modalités, et à chaque électeur est associée une variable de Bernoulli qui prend la valeur 1 s'il vote pour A et la valeur 0 sinon. La variable générique de la population est une variable de Bernoulli de paramètre 0,52 puisque chaque électeur tiré au hasard vote pour A avec une probabilité égale à la proportion de votants en faveur de A (*cf.* l'analogie entre probabilité et fréquence relative vue au chapitre 5).

Les 50 électeurs interrogés avant le scrutin forment un ensemble de 50 variables de Bernoulli de même paramètre, à savoir 0,52. De plus ces variables sont indépendantes si le tirage est effectué avec remise ou si le taux de sondage est inférieur à 10 % (ce qui est supposé ici compte tenu de la taille de l'échantillon).

La somme de ces 50 variables $Y = X_1 + X_2 + \ldots + X_{50}$ contient autant de 1 que d'électeurs favorables à A, et représente le nombre d'électeurs, parmi les 50 sondés, favorables à A. Cette somme de variables de Bernoulli suit une loi binomiale $\mathcal{B}(50 ; 0{,}52)$.

La probabilité que cet échantillon donne une majorité pour B est égale à la probabilité que le nombre d'électeurs favorables à A soit strictement inférieur à 25.

Puisque Y représente le nombre d'électeurs favorables à A, la variable aléatoire $(50 - Y)$ représente le nombre d'électeurs favorables à B. Par un raisonnement identique à celui fait pour Y, la variable aléatoire $(50 - Y)$ suit une loi binomiale $\mathcal{B}(50 ; 0{,}48)$. Lorsque B a la majorité, on a $\{50 - Y > 25\}$ et la probabilité cherchée vaut :

$$P(Y < 25) = P(50 - Y > 25) = 1 - P(50 - Y \leq 25)$$

Le calcul (programme ou table) donne $P(50 - Y \leq 25) = 0{,}6648$ pour la loi $\mathcal{B}(50 ; 0{,}48)$, et il y a donc près de 33,5 % de chances qu'un échantillon de 50 personnes donne un résultat contraire à la réalité ! Ceci est dû à la conjonction de deux éléments :

• le résultat final est assez « serré » car les deux termes de l'alternative sont très peu séparés en probabilité (0,52 contre 0,48) ce qui signifie que si l'échantillon était l'exact reflet de la population, on aurait 26 contre 24, soit 2 voix de différence seulement ;

• on n'interroge que 50 personnes, et cela est bien peu, compte tenu des scores réels, pour discriminer les 2 candidats de manière fiable (donc crédible).

On notera enfin sur cet exemple que si le nombre de votants favorable à A suit une loi binomiale $\mathcal{B}(50 ; 0{,}52)$, la proportion de votants favorables à A suit une loi binomiale en proportion.

B. Le schéma hypergéométrique

Dans le schéma binomial, on répète une épreuve de Bernoulli n fois, mais de telle façon que les épreuves soient indépendantes.

Cette condition peut paraître peu réaliste. En pratique lorsqu'on tire un échantillon de taille n dans une population de taille $N (n < N)$, le bon sens veut qu'on ne prenne pas 2 fois le même individu, ce qui équivaut à tirer l'échantillon **sans remise** (on parle encore de **tirage exhaustif**). Les variables aléatoires de Bernoulli associées aux différents éléments de l'échantillon, et indicatrices de la présence ou de l'absence d'un caractère donné, sont, du fait du tirage sans remise, mutuellement dépendantes.

La variable aléatoire Y égale au nombre d'individus de l'échantillon possédant le caractère considéré est dans ce cas somme de n v.a. de Bernoulli dépendantes, et de même paramètre.

Notons p, la proportion d'individus dans la population (dont on désigne la taille par N) possédant le caractère étudié, et étudions la loi de cette variable aléatoire Y d'abord en ce qui concerne les valeurs possibles, puis pour ce qui est des probabilités associées.

Le nombre d'individus de la population possédant le caractère étudié est égal à Np, et le nombre de ceux qui ne le possède pas est égal à Nq. Le nombre maximum d'individus de l'échantillon possédant le caractère étudié ne peut être supérieur ni à la taille de l'échantillon, ni à Np. Par conséquent, la valeur maximum de Y est égale à $\min(n, Np)$. Le nombre minimum d'individus de l'échantillon possédant le caractère étudié est, bien entendu, au moins égal à0, mais aussi au moins égal à $(n - Nq)$. En effet, si le nombre d'individus ne possédant pas le caractère étudié, soit Nq, est plus petit que la taille n de l'échantillon, on aura au moins $(n - Nq)$ individus qui posséderont le caractère étudié dans l'échantillon. Il s'ensuit que le nombre minimum d'individus de l'échantillon possédant le caractère étudié est égal à $\max(0, n - Nq)$.

La variable aléatoire Y peut prendre toutes les valeurs entières comprises entre :

$$\max(0, n - Nq) \quad \text{et} \quad \min(n, Np)$$

Pour le calcul de $P(Y = k)$, k étant l'une des valeurs possibles entre $\max(0, n - Nq)$ et $\min(n, Np)$, on peut utiliser la méthode combinatoire classique et calculer le rapport du nombre des occurences favorables au nombre des occurences possibles.

Les occurences possibles sont représentées par le nombre d'échantillons de taille n qu'on peut extraire sans remise d'une population de taille N, c'est-à-dire $\binom{N}{n}$.

Les occurrences favorables sont représentées par les échantillons de taille n pour lesquels k individus possèdent le caractère étudié et ($n - k$) individus ne le possèdent pas.

Ces cas favorables sont ceux où l'on a tiré k unités parmi les Np ayant le caractère étudié, en nombre $\binom{Np}{k}$, et ($n - k$) unités parmi les Nq ne le possédant pas, en nombre $\binom{Nq}{n-k}$. On a donc :

$$P(Y = k) = \frac{\binom{Np}{k}\binom{Nq}{n-k}}{\binom{N}{n}} \quad \text{pour } \max(0, n - Nq) \leq k \leq \min(n, Np)$$

On dit que la variable aléatoire Y suit une **loi hypergéométrique de paramètres N, n et p**, ce qu'on note $\mathcal{H}(N ; n ; p)$.

L'espérance d'une telle variable aléatoire est $E(Y) = np$ puisque Y est la somme de n variables de Bernoulli de paramètre p. La variance est égale à :

$$\text{var}(Y) = np(1 - p) \cdot \frac{N - n}{N - 1}$$

mais, le calcul est plus délicat en raison de la non indépendance des variables de Bernoulli. Le terme correctif $\frac{N-n}{N-1} = 1 - \frac{n-1}{N-1}$ est appelé **facteur d'exhaustivité**. On remarque tout de suite que si le taux de sondage n/N est très petit, ce facteur d'exhaustivité est très proche de 1, et donc que l'expression de la variance d'une loi hypergéométrique est très voisine de celle d'une loi binomiale.

Plus généralement on peut montrer que la loi $\mathcal{H}(N ; n ; p)$ peut être approximée par une loi $\mathcal{B}(n ; p)$ dès que le taux de sondage n/N est inférieur à 10 %. Cette conclusion justifie l'utilisation des calculs sous l'hypothèse d'indépendance dès que le taux de sondage est assez petit, même si le tirage est exhaustif.

C'est la raison pour laquelle en pratique, malgré des tirages d'échantillons le plus souvent exhaustifs, on se réfère à la loi binomiale, les probabilités calculées à l'aide de la loi binomiale donnant une bonne approximation des probabilités de la loi hypergéométrique dès que le taux de sondage est assez petit (c'est-à-dire inférieur à 10 %).

En conclusion, la loi hypergéométrique $\mathcal{H}(N ; n ; p)$ est la distribution d'une somme de n variables aléatoires de Bernoulli non indépendantes. Une variable aléatoire hypergéométrique représente, dans un contexte de tirage exhaustif – c'est-à-dire de variables dépendantes – le nombre de réalisations parmi n épreuves de Bernoulli de l'un des termes d'une alternative. Elle prend des valeurs comprises entre $\max(0, n - Nq)$ et $\min(n, Np)$.

Cette variable aléatoire a la même espérance np que la variable binomiale qui serait obtenue dans un contexte d'indépendance, mais sa variance est plus petite, diminuée dans un rapport

$$\frac{N-n}{N-1}$$

appelé facteur d'exhaustivité. Dans le cas d'une très grande population ou plus généralement d'un taux de sondage faible (inférieur à 0,1), on peut supposer les tirages indépendants et remplacer la loi hypergéométrique $\mathcal{H}(N\,;\,n\,;\,p)$ par la loi binomiale $\mathcal{B}(n\,;\,p)$

C. La loi géométrique et la loi de Pascal

On se place dans une optique totalement différente, les conditions de base restant inchangées, c'est-à-dire qu'il y a toujours une succession d'épreuves de Bernoulli de même paramètre p, mais *dont on ne connaît pas le nombre de répétitions* : on ne s'arrête que lorsque le résultat auquel on s'intéresse est obtenu pour la 1re fois (cas de la *loi géométrique*) ou pour la K^{e} fois (*loi de Pascal*).

À chaque épreuve élémentaire, est associée une variable de Bernoulli X_i qui prend la valeur 1 si le résultat auquel on s'intéresse s'est réalisé, et la valeur 0 sinon. On pose :

$$P(X_i = 1) = p \quad \text{et} \quad P(X_i = 0) = 1 - p = q$$

On suppose que les épreuves sont répétées indépendamment les unes des autres. On désigne par Y le **nombre total d'épreuves réalisées jusqu'à l'obtention du premier résultat élémentaire de probabilité p.** Il est clair que Y peut prendre toute valeur entière au moins égale à 1 (c'est-à-dire strictement positive), et que ces valeurs peuvent être aussi grandes que l'on veut. Nous rencontrons ici pour la première fois une variable aléatoire dont le nombre de valeurs possibles est infini.

Cette définition doit être bien comprise, car dans certains cas on s'intéresse au nombre Z d'épreuves précédant la première réalisation du résultat de probabilité p, et on a bien sûr : $Z = Y - 1$

Pour ce qui concerne la variable aléatoire Y, si le résultat codé 1 se produit pour la 1re fois à la k^{e} épreuve, cela signifie que les $(k-1)$ premières épreuves ont produit le résultat complémentaire codé 0 de probabilité q. En raison de l'indépendance des épreuves on a :

$$P(Y = k) = \underbrace{q \cdot q \cdot \ldots \cdot q}_{(k-1)\text{ fois}} \cdot p = q^{k-1} \cdot p$$

On en déduit la fonction de répartition :

$$P(Y \leq n) = \sum_{k=1}^{n} q^{k-1} \cdot p = p \cdot \sum_{k=0}^{n-1} q^k = p \cdot \frac{1-q^n}{1-q} = 1-q^n$$

On calcule aussi :

$$E(Y) = \sum_{k=1}^{\infty} k \cdot q^{k-1} \cdot p = p \cdot \sum_{k=1}^{\infty} k \cdot q^{k-1} = p \cdot \sum_{k=1}^{\infty} \frac{d}{dq}(q^k)$$

La série de terme général q^k étant absolument convergente, la série des dérivées est égale à la dérivée de la série :

$$E(Y) = p \cdot \sum_{k=1}^{\infty} \frac{d}{dq}(q^k) = p \cdot \frac{d}{dq}\left(\sum_{k=1}^{\infty} q^k\right)$$

$$= p \cdot \frac{d}{dq}\left(\frac{q}{1-q}\right) = p \cdot \frac{1}{(1-q)^2} = \frac{1}{p}$$

On calcule de même le moment d'ordre 2 :

$$E(Y^2) = \sum_{k=1}^{\infty} k^2 \cdot q^{k-1} \cdot p = p \cdot \sum_{k=1}^{\infty} k^2 \cdot q^{k-1}$$

$$= p \cdot \sum_{k=1}^{\infty} \Big(k(k-1)+k\Big) q^{k-1} = p \cdot \sum_{k=1}^{\infty} k(k-1) q^{k-2} q + p \sum_{k=1}^{\infty} k q^{k-1}$$

Le second terme de l'expression obtenue n'est autre que $E(Y)$. Pour le premier terme, on remarque que :

$$p \cdot \sum_{k=1}^{\infty} k(k-1) q^{k-2} q = p \cdot q \cdot \sum_{k=1}^{\infty} k(k-1) q^{k-2}$$

$$= p \cdot q \cdot \sum_{k=2}^{\infty} k(k-1) q^{k-2} = p \cdot q \cdot \sum_{k=2}^{\infty} \frac{d^2}{dq^2}(q^k)$$

car encore une fois la double dérivation sous le signe somme est licite en raison de la convergence absolue de la série.

On obtient :

$$p \cdot \sum_{k=1}^{\infty} k(k-1) q^{k-2} q = p \cdot q \frac{d^2}{dq^2}\left(\sum_{k=2}^{\infty} q^k\right)$$

$$= p \cdot q \cdot \frac{d^2}{dq^2}\left(\frac{q^2}{1-q}\right) = p \cdot q \cdot \frac{2}{(1-q)^3} = \frac{2q}{p^2}$$

et par conséquent :

$$E(Y^2) = \frac{2q}{p^2} + \frac{1}{p} = \frac{2q+p}{p^2} = \frac{q+p+q}{p^2} = \frac{q+1}{p^2}$$

On déduit la variance d'une variable aléatoire de loi géométrique :

$$\text{var}(Y) = \frac{q+1}{p^2} - \frac{1}{p^2} = \frac{q}{p^2}$$

Dans l'étude de la modélisation des situations concrètes de ce type, on doit faire très attention de préciser si on s'intéresse au nombre total Y d'épreuves alternatives réalisées jusqu'à l'obtention du premier résultat élémentaire de probabilité p (cas étudié), ou si on s'intéresse au nombre Z d'épreuves élémentaires de probabilité $(1 - p)$ réalisées jusqu'à l'obtention du premier résultat de probabilité p.

Comme nous l'avons déjà mentionné $Z = Y - 1$. Les valeurs possibles de Z sont toutes les valeurs entières positives ou nulle, alors que les valeurs possibles de Y sont toutes les valeurs entières strictement positives. La relation entre Y et Z implique qu'on peut calculer les probabilités associées à Z à partir de celles de Y :

$$P(Z = k) = P(Y = k+1) = q^k \cdot p$$

$$P(Z \leq n) = P(Y \leq n+1) = 1 - q^{n+1}$$

L'espérance mathématique de Z est égale à celle de Y diminuée d'une unité :

$$E(Z) = E(Y-1) = E(Y) - 1 = \frac{1}{p} - 1 = \frac{1-p}{p} = \frac{q}{p}$$

alors que les variances de Y et Z sont égales :

$$\text{var}(Z) = \text{var}(Y-1) = \text{var}(Y) = \frac{q}{p^2}$$

En résumé

La loi géométrique de paramètre p caractérise le nombre d'épreuves de Bernoulli indépendantes qu'il faut réaliser pour obtenir pour la 1 re fois le résultat (de l'épreuve de Bernoulli) auquel on s'intéresse (codé 1). L'espérance est égale à $\frac{1}{p}$ et la variance à $\frac{1-p}{p^2}$

La loi de Pascal est la généralisation de la loi géométrique lorsqu'on recherche l'obtention pour la K^e fois du résultat considéré. Une variable aléatoire de Pascal Y dépend de deux paramètres p et K et peut prendre toutes valeurs entières au moins égales à K.

Pour calculer $P(Y = j)$ pour $j \geq K$, on remarque que si le j^e essai a donné le résultat de probabilité p, c'est qu'au cours des $(j - 1)$ essais précédents, on aura obtenu $(K - 1)$ fois ce résultat et $(j - K)$ fois le résultat contraire. On applique la combinatoire du schéma binomial, et la probabilité d'observer $(j - K)$ fois le résultat de probabilité $\{q = 1 - p\}$ et $(K - 1)$ fois le résultat de probabilité p au cours de $(j - 1)$ essais est donnée par :

$$\binom{j-1}{K-1} p^{K-1}(1-p)^{j-K}$$

Pour obtenir l'événement $\{Y = j\}$, il faut et il suffit que dans les $(j - 1)$ premiers essais, on ait obtenu $(K - 1)$ fois le résultat de probabilité p et $(j - K)$ fois le résultat contraire, et que le j^e essai donne le résultat de probabilité p. En raison de l'indépendance des épreuves :

$$P(Y = j) = \binom{j-1}{K-1} p^{K}(1-p)^{j-K} \quad \text{pour} \quad j \geq K$$

On peut montrer que moyenne et variance de la loi de Pascal de paramètres p et K sont donnés par :

$$E(Y) = \frac{K}{p} \quad \text{et} \quad \text{var}(Y) = \frac{K \cdot (1-p)}{p^2}$$

On doit bien porter attention au fait que la ressemblance avec les probabilités d'une loi binomiale n'est qu'apparente. En effet, non seulement la somme des exposants des termes p et $(1 - p)$ n'est pas égale au nombre $(j - 1)$, mais ces probabilités sont définies pour toutes les valeurs de j au moins égales à K, et donc pour un ensemble de valeurs non borné. Pour une loi de Pascal, généralisant la loi géométrique, *c'est le nombre total d'épreuves, et non pas le nombre d'épreuves conduisant au résultat de probabilité p qui est aléatoire*.

Ces deux lois présentent une différence très importante avec la loi binomiale : le nombre de répétitions de l'épreuve élémentaire de Bernoulli n'est pas connu, et c'est lui qui représente l'aléatoire du problème. En particulier, une variable géométrique peut prendre toute valeur entière positive, sans limite supérieure.

L'exemple suivant montre l'application de ces modèles et l'interprétation de leurs caractéristiques.

▶ **Exemple**

Supposons qu'on observe en moyenne 5 % de pièces défectueuses en sortie d'une chaîne de production lorsqu'elle est optimisée. Si on souhaite connaître la probabilité qu'un échantillon de 20 pièces issu de cette chaîne ne contienne aucune pièce défectueuse, on associe à chaque pièce un caractère à deux modalités, et cette modélisation de base amène à définir des variables de Bernoulli.

Le paramètre de ces variables de Bernoulli étant égal à 0,05 puisque si 5 % des pièces en moyennes sont défectueuses, cela revient à dire que la

probabilité qu'une pièce prise au hasard soit défectueuse est égale à 0,05 (chapitre 5). On peut supposer les tirages indépendants en raison de la grande taille de la population (ici la production).

Le schéma binomial est ici adapté puisqu'on recherche la probabilité d'un nombre donné de défectueux sur un échantillon de taille fixée. Pour cette loi $\mathcal{B}(20 ; 0{,}05)$, on a $P(X = 0) = (0{,}95)^{20} = 0{,}3585$

Si d'autre part, on cherche à calculer la probabilité que le premier défectueux ne soit pas l'une des 20 premières pièces, on gardera la modélisation des unités statistiques par les aléas de Bernoulli de paramètre 0,05 toujours supposés indépendants pour les mêmes raisons. Mais le nombre de pièces étudiées n'étant plus donné, ce nombre devient l'aléa dont on a besoin de déterminer la loi de probabilité.

Soit Y le nombre de pièces observées jusqu'à l'obtention de la première pièce défectueuse. La variable aléatoire Y est une variable aléatoire distribuée selon une loi géométrique de paramètre 0,05 ; par conséquent :

$$P(Y \geq 21) = \sum_{k \geq 21} 0{,}95^{k-1} \cdot 0{,}05 = 0{,}05 \cdot \sum_{k \geq 21} 0{,}95^{k-1} = 0{,}05 \cdot \sum_{j \geq 20} 0{,}95^{j}$$

$$\Rightarrow \quad P(Y \geq 21) = 0{,}05 \cdot 0{,}95^{20} \cdot \sum_{j \geq 0} 0{,}95^{j}$$

$$= 0{,}05 \cdot 0{,}95^{20} \cdot \frac{1}{1 - 0{,}95} = 0{,}95^{20} = 0{,}3585$$

L'espérance mathématique de cette variable aléatoire Y étant égale à 20, on doit tirer en moyenne 20 pièces pour en observer une défectueuse, c'est-à-dire qu'avant de tirer une pièce défectueuse, on tire, en moyenne, 19 pièces qui ne le sont pas.

La relation entre tous ces résultats est laissée au lecteur.

Si on s'était intéressé au nombre de pièces à examiner pour en tirer deux défectueuses, on aurait une loi de Pascal d'espérance mathématique égale à 40. Ici encore, on laisse au lecteur le soin de comparer les deux derniers résultats.

Ces deux lois, loi géométrique et loi de Pascal, interviennent particulièrement en contrôle de qualité, mais aussi dans la surveillance des événements dont une certaine fréquence de survenue est interprétée en terme de signal d'alarme.

Les formules de la loi géométrique sont suffisamment simples pour que les calculs ne posent aucune difficulté avec une petite calculatrice, et pour la loi de Pascal, on peut recourir à quelques pas de programme comme pour la loi binomiale.

→ ***Remarque***

Les lois binomiale, hypergéométrique, géométrique et de Pascal sont donc toutes construites sur la base de la répétition d'épreuves à deux

issues (ou épreuves de Bernoulli). À l'exception de la loi hypergéométrique, elles se placent toutes dans un contexte d'épreuves indépendantes dont la caractéristique p (probabilité de l'issue de l'alternative qu'on cherche à observer) est constante au cours du temps.

Ceci correspond à une notion très développée dans la modélisation des phénomènes dépendant du temps, à savoir la *stationnarité*. Cette propriété n'est pas systématiquement rencontrée, et il faut apporter la plus grande attention à l'analyse de cette hypothèse dans toutes les situations qu'on cherche à représenter.

Beaucoup de cas ne correspondent pas en effet à une succession stationnaire d'épreuves de Bernoulli indépendantes. Nous avons vu que l'indépendance stricto sensu pouvait quelquefois servir de représentation approchée à des tirages exhaustifs (pour un taux de sondage suffisamment faible), mais on devra soigneusement analyser le contexte pour reconnaître s'il est celui d'une parfaite stationnarité (c'est-à-dire de constance dans le temps du paramètre p des épreuves de Bernoulli successives), s'il est celui d'une stationnarité approximative, ou si cette condition ne peut être supposée (auquel cas les outils à mettre en œuvre sont plus complexes et débordent du propos de cet ouvrage).

Le tableau suivant résume de façon synthétique les principaux modèles construits à partir de l'itération du schéma de Bernoulli.

Loi	Nombre d'itérations	Valeur minimale	Valeur maximale	Type de tirage	Espérance	Variance
Bernoulli	fixé	0	1	sans	p	$p(1-p)$
Binomiale	fixé	0	n	indépendant	np	$np(1-p)$
Hypergéométrique	fixé	$\max(0, n-Nq)$	$\min(n, Np)$	exhaustif	np	$np(1-p)\frac{N-n}{N-1}$
Géométrique	aléatoire	1	sans	indépendant	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Pascal	aléatoire	K	sans	indépendant	$\frac{K}{p}$	$\frac{K(1-p)}{p^2}$

III. La loi de Poisson

Cette loi peut être envisagée dans un contexte empirique (statistique), ou d'analyse (probabiliste).

A. Définitions et propriétés

Définition

Une variable aléatoire X ayant pour valeur possible tout nombre entier positif ou nul, et telle que :

$$P(X = k) = e^{-m} \cdot \frac{m^k}{k!}$$

pour tout $k \geq 0$ entier, est dite distribuée selon une *loi de Poisson*[1] *de paramètre m, m* étant un nombre réel strictement positif.

On remarque tout de suite qu'une telle variable aléatoire présente une différence essentielle avec les variables de Bernoulli ou binomiales, car elle est discrète, mais non finie (c'est-à-dire ici que les valeurs possibles ne sont pas limitées supérieurement). Nous avons déjà rencontré cette situation avec la loi géométrique et la loi de Pascal.

Il s'agit bien d'une distribution de probabilité car, il est facile de le constater que :

– toutes les probabilités sont positives ;

– la somme des probabilités est égale à 1, compte tenu de l'expression de la série exponentielle :

$$\sum_{k \geq 0} P(X = k) = \sum_{k \geq 0} e^{-m} \cdot \frac{m^k}{k!} = e^{-m} \cdot \sum_{k \geq 0} \frac{m^k}{k!} = e^{-m} \cdot e^m = e^0 = 1$$

Le calcul de la moyenne est assez simple :

$$E(X) = \sum_{k \geq 0} k \cdot P(X = k) = \sum_{k \geq 0} k \cdot e^{-m} \cdot \frac{m^k}{k!}$$

$$= e^{-m} \cdot \sum_{k \geq 0} k \cdot \frac{m^k}{k!} = e^{-m} \cdot \sum_{k \geq 1} k \cdot \frac{m^k}{k!}$$

car le premier terme de la somme est nul. Par conséquent,

$$E(X) = e^{-m} \cdot \sum_{k \geq 1} \frac{m \cdot m^{k-1}}{(k-1)!} = m \cdot e^{-m} \cdot \sum_{k \geq 1} \frac{m^{k-1}}{(k-1)!}$$

1. Siméon-Denis Poisson (1781-1840), mathématicien, probabiliste et physicien français, à qui on doit d'important développements sur la loi des grands nombres, sur les suites d'épreuves de Bernouilli, sur la loi de… Poisson, mais aussi sur les applications des probabilités dans les domaines du droit.

Le changement de variable $j = k - 1$ donne :

$$E(X) = m \cdot e^{-m} \cdot \sum_{j \geq 0} \frac{m^j}{j!} = m \cdot e^{-m} \cdot e^m = m$$

Ce résultat justifie la notation m adoptée pour le paramètre de la loi de Poisson puisque ce paramètre n'est autre que la valeur moyenne.

En ce qui concerne le moment d'ordre 2, le calcul du même type donne :

$$E(X^2) = \sum_{k \geq 0} k^2 \cdot P(X = k) = \sum_{k \geq 0} k^2 \cdot e^{-m} \cdot \frac{m^k}{k!}$$
$$= e^{-m} \cdot \sum_{k \geq 0} k^2 \cdot \frac{m^k}{k!} = e^{-m} \cdot \sum_{k \geq 1} k^2 \cdot \frac{m^k}{k!}$$

car le 1 er terme de la série étant nul, on peut commencer cette somme pour $k = 1$

On obtient ensuite :

$$E(X^2) = e^{-m} \sum_{k \geq 1} k^2 \cdot \frac{m^k}{k!} = e^{-m} \sum_{k \geq 1} (k(k-1) + k) \cdot \frac{m^k}{k!}$$
$$= e^{-m} \left(\sum_{k \geq 1} \{k(k-1)\} \cdot \frac{m^k}{k!} + \sum_{k \geq 1} k \cdot \frac{m^k}{k!} \right)$$

Le deuxième terme de la dernière parenthèse n'est autre que $E(X)$ soit m. Pour le calcul du premier terme, on remarque que le terme initial pour $k = 1$ est nul. On débute la somme à $k = 2$, et on simplifie par $k(k–1)$:

$$E(X^2) = e^{-m} \cdot \sum_{k \geq 1} k(k-1) \frac{m^k}{k!} + m = e^{-m} \cdot \sum_{k \geq 2} k(k-1) \frac{m^k}{k!} + m$$
$$= e^{-m} \cdot \sum_{k \geq 2} \frac{m^2 \cdot m^{k-2}}{(k-2)!} + m$$

soit :

$$E(X^2) = m^2 \cdot e^{-m} \cdot \sum_{k \geq 2} \frac{m^{k-2}}{(k-2)!} + m$$

En faisant le changement de variable $\{ j = k - 2\}$ dans la dernière somme, on retrouve encore le développement de la série exponentielle, d'où :

$$E(X^2) = m^2 \cdot e^{-m} \cdot e^m + m = m^2 + m$$

Et on déduit la variance :

$$\text{var}(X) = E(X^2) - \left(E(X) \right)^2 = m^2 + m - m^2 = m$$

Ce résultat a un intérêt considérable, comme on le verra plus loin :

Pour une distribution de Poisson, moyenne et variance sont égales (et égales à la valeur du paramètre).

On peut aussi calculer la fonction génératrice :

$$g_X(u) = E(u^X) = \sum_{k \geq 0} u^k e^{-m} \frac{m^k}{k!} = e^{-m} \sum_{k \geq 0} \frac{(um)^k}{k!} = e^{-m} \cdot e^{um} = e^{-m(1-u)}$$

ce qui permet d'obtenir le moment factoriel d'ordre $r (r \in \mathbb{N}^*)$: $\mu_{[r]}(X) = m^r$

La propriété suivante est très utile dans la construction des modèles régis par des lois de Poisson.

Propriété 1

Si X_1 et X_2 sont deux variables aléatoires indépendantes qui suivent des lois de Poisson respectivement de paramètres m_1 et m_2, alors $Y = X_1 + X_2$ suit une loi de Poisson de paramètre $m_1 + m_2$

En effet, la variable Y peut prendre toutes les valeurs entières, positives ou nulle. Calculons la probabilité qu'elle prenne l'une quelconque de ces valeurs.

$$P(Y = k) = P\left(\bigcup_{i=0}^{i=k} (\{X_1 = i\} \cap \{X_2 = k - i\})\right)$$

$$= \sum_{i=0}^{i=k} P(\{X_1 = i\} \cap \{X_2 = k - i\})$$

$$= \sum_{i=0}^{i=k} P(\{X_1 = i\}) \cdot P(\{X_2 = k - i\})$$

donc :

$$P(Y = k) = \sum_{i=0}^{i=k} P(X_1 = i) \cdot P(X_2 = k - i) = \sum_{i=0}^{i=k} e^{-m_1} \cdot \frac{m_1^i}{i!} \cdot e^{-m_2} \cdot \frac{m_2^{k-i}}{(k-i)!}$$

soit :

$$P(Y = k) = e^{-(m_1 + m_2)} \sum_{i=0}^{i=k} \frac{m_1^i \cdot m_2^{k-i}}{i!(k-i)!} = \frac{e^{-(m_1 + m_2)}}{k!} \sum_{i=0}^{i=k} \frac{k!}{i!(k-i)!} m_1^i \cdot m_2^{k-i}$$

et on reconnaît dans la dernière somme le développement du binôme de Newton appliqué à la somme $(m_1 + m_2)^k$. Ce qui permet d'écrire :

$$P(Y = k) = e^{-(m_1 + m_2)} \cdot \frac{(m_1 + m_2)^k}{k!}$$

ce résultat prouvant le résultat annoncé.

➞ ***Remarque***

Ce résultat peut s'étendre à une somme finie de variables aléatoires indépendantes distribuées toutes selon des *lois de Poisson*.

Il existe une forme réciproque de cette propriété.

Propriété 2

Si les variables aléatoires indépendantes X et Y sont telles que la somme $(X + Y)$ est distribuée selon une loi de Poisson, alors les variables X et Y sont elles-mêmes distribuées selon des lois de Poisson.

On ne démontrera pas cette propriété très utile. Il faut remarquer ici qu'on n'a pas le moyen direct de décomposer (pour cette propriété 2) le paramètre de $(X + Y)$ en deux paramètres, l'un pour X et l'autre pour Y.

Une propriété, elle aussi caractéristique de la loi de Poisson, est celle qui suit, obtenue aisément en écrivant le rapport des probabilités et en simplifiant :

Propriété 3

Si X suit une loi de Poisson de paramètre m, on a :

$$\frac{P(X = k)}{P(X = k - 1)} = \frac{m}{k}$$

Cette propriété implique la croissance des probabilités ponctuelles $P(X = k)$ tant que $k \leq m$, et la décroissance (rapide puisqu'inversement proportionnelle à k) dès que $k > m$.

D'autre part si m est un entier, le rapport $\frac{P(X = m)}{P(X = m - 1)}$ est égal à 1. Ceci signifie qu'il existe deux valeurs, m et $m - 1$, qui ont même probabilité. Cette probabilité commune est la plus élevée d'après ce qu'on vient de voir. Par conséquent, *la loi de Poisson possède deux valeurs modales lorsque son paramètre est un nombre entier*.

B. Abord statistique

D'après les propriétés qui viennent d'être montrées, on remarque qu'il est justifié d'envisager une loi de Poisson comme **un modèle représentatif de données statistiques discrètes pour lesquelles la variable ne prend que des valeurs entières, positives ou nulle, et pour lesquelles :**

- la moyenne et la variance sont sensiblement égales ;
- les rapports $\frac{f_k}{f_{k-1}}$ de 2 fréquences consécutives sont inversement proportionnels à k

Il est fréquent que cette dernière condition ne soit vérifiée que pour les faibles valeurs de k. Dans la pratique, on accorde moins d'importance aux entorses à cette propriété pour les queues de distribution.

Enfin, on prendra garde de bien noter qu'il ne s'agit là que d'une indication. Il est indispensable de justifier le choix d'un modèle par un jugement d'adéquation.

C. Abord probabiliste

Il s'agit maintenant de poser la loi de Poisson comme modèle d'une épreuve aléatoire avec l'aide d'une analyse raisonnée de cette épreuve. Un résultat est nécessaire à cette démarche.

Propriété 4

Les probabilités d'une loi binomiale $\mathcal{B}(n\,;p)$ peuvent être approximées par les probabilités d'une loi de Poisson de paramètre np si les conditions suivantes sont réalisées :

$$n > 50 \quad \text{et} \quad p < 0{,}1$$

Ceci implique que la loi de Poisson peut être considérée comme l'approximation d'une loi binomiale qui représente la somme d'un grand nombre ($n > 50$) d'aléas de Bernoulli de faible paramètre ($p < 0{,}1$).

On remarque à ce sujet que si une variable aléatoire est distribuée selon une loi binomiale $\mathcal{B}(n\,;p)$ pour laquelle $n > 50$ et $p < 0{,}1$, on aura $q \approx 1$ et par conséquent $np \approx npq$. Si on approxime cette loi binomiale par une loi de Poisson $\mathcal{P}(np)$, on imagine que les deux lois doivent être assez proches pour que les espérances mathématiques, d'une part, et les variances, d'autre part, soient voisines, sinon même égales. Or, les espérances sont égales toutes

deux à np, mais les variances respectivement égales à npq pour la loi binomiale et à np pour la loi de Poisson sont proches puisque $q \approx 1$. La valeur de l'approximation apparaît bien liée à la faible valeur de p

Bien évidemment, par symétrie, et en particulier en tenant compte de la propriété 2 vue au § II.A, pour une variable X suivant une loi $\mathcal{B}(n\ ;\ p)$ où $n > 50$ et $p > 0{,}9$, cette approximation sera appliquée à la variable $(n - X)$ qui suit une loi $\mathcal{B}(n\ ;\ 1 - p)$

Lorsqu'un événement a une faible probabilité ($p < 0{,}1$) d'apparition lors d'une épreuve élémentaire, et si on répète cette épreuve un grand nombre de fois ($n > 50$), le nombre total de réalisations de l'événement considéré suit à peu près une loi de Poisson de paramètre np. Les graphiques de la figure 6.3a et 6.3b montrent la comparaison entre les diagrammes en bâtons de plusieurs lois binomiales et des lois de Poisson qui sont proposées comme approximation. On retrouve en examinant ces graphiques que plus p est petit, meilleure est la qualité de l'approximation.

Pour cette raison, la loi de Poisson a été appelée la loi des petites probabilités, ou loi des faibles occurences, ou *loi des phénomènes rares*.

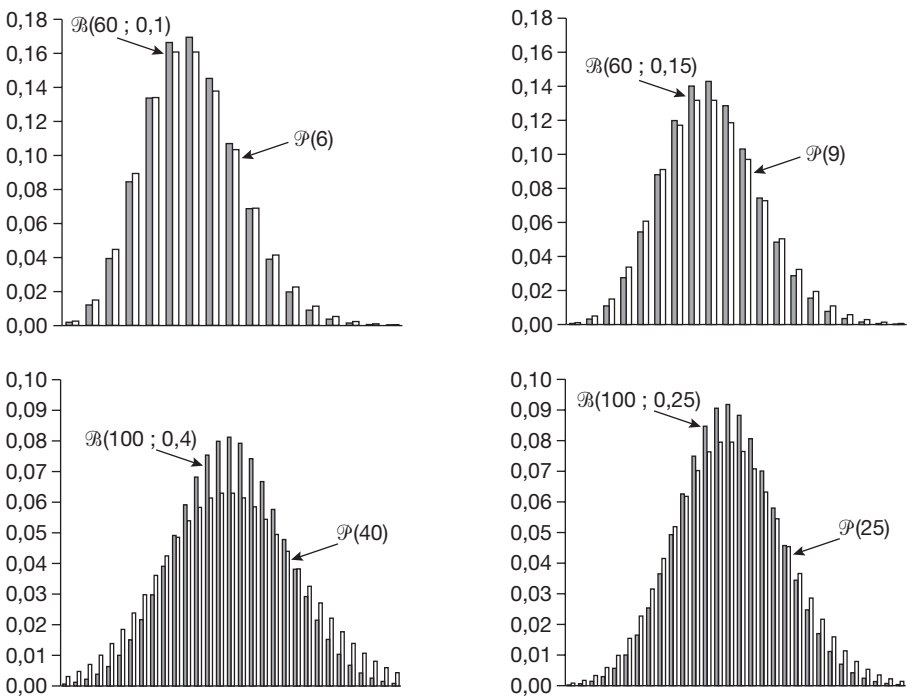


Figure 6.3a – Approximations de mauvaise qualité par la loi de Poisson

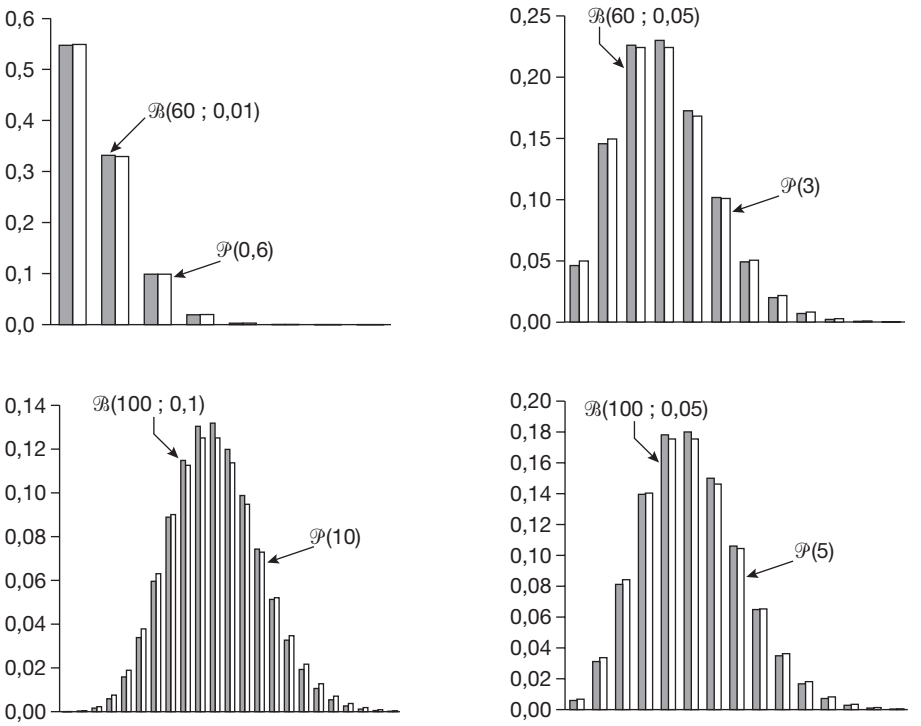


Figure 6.3b – Approximations correctes par la loi de Poisson

Voici quelques exemples où cette loi est évoquée :

- nombre de pièces défectueuses dans un échantillon de grande taille prélevé dans une production où la proportion des pièces défectueuses est faible ;
- nombre de naissances de quadruplés, de quintuplés, par an dans un pays fixé ;
- nombre d'appels intercontinentaux sur une ligne pendant une période donnée.

Les formules des probabilités de lois binomiale et de Poisson montrent bien l'intérêt de la seconde. Même avec une bonne calculatrice, il n'est pas aisé (et parfois pas possible directement) de calculer les probabilités d'une loi binomiale. Mais cependant, on ne recourra à l'approximation par la loi de Poisson que lorsqu'on ne peut aisément obtenir le résultat exact, c'est-à-dire qu'on ne cherchera pas à approximer la loi binomiale tant que le calcul est simple.

Les tables de la loi de Poisson (*cf.* annexe IV) donnent les valeurs de la fonction de répartition pour les valeurs du paramètre :

– entre 0 et 1 par pas de 0,1

– entre 1 et 10 par pas de 0,5

– entre 10 et 25 par pas de 1

Comme pour la loi binomiale, l'usage des tables est limité à certaines valeurs du paramètre. On peut utiliser quelques pas de programme pour tous les calculs (annexe II).

Toutefois, lorsque m dépasse 25, on utilise l'approximation par la loi de Gauss et la correction de continuité (chapitre 7).

La loi de Poisson de paramètre m est une loi caractérisée par l'égalité de sa moyenne et de sa variance, et dont les probabilités p_k croissent tant que $k < m$, puis décroissent. Elle peut être envisagée comme une approximation de la loi binomiale, donc comme la loi approchée de la somme d'un grand nombre ($n > 50$) d'aléas de Bernoulli indépendants représentant une alternative dont l'une des issues est de probabilité très faible ($p < 0,1$). Elle est ainsi directement postulée comme modèle représentatif du nombre d'occurences au cours du temps pour un événement particulièrement peu probable, mais dont la probabilité de survenue est stable.

On ajoutera encore qu'il s'agit d'un modèle particulièrement utilisé en pratique dans la gestion des files d'attente notamment, et qui est à l'origine de développements très importants dans l'analyse des séries d'événements (processus poissoniens).

On n'oubliera pas :

1. Que la loi de Bernoulli représente toute situation d'alternative dans laquelle une issue est codée 0 et l'autre codée 1

2. Que la loi binomiale correspond à la somme d'un nombre fixé de variables aléatoires de Bernoulli de même paramètre et indépendantes (donc à un tirage avec remise)

3. Que la loi hypergéométrique correspond à la somme d'un nombre fixé de variables aléatoires de Bernoulli de même paramètre, mais dans un tirage sans remise

4. Que la loi géométrique correspond à la somme de variables aléatoires de Bernoulli de même paramètre et indépendantes, mais en nombre aléatoire

5. Que la moyenne et la variance d'une variable aléatoire de Poisson sont égales

Exercices *(corrigés page 319)*

Exercice 6.1

Des sondages permettent de constater que 10 % de la population est constituée de gauchers. On considère donc, dans cet exercice, que la probabilité qu'un individu pris au hasard soit gaucher est égale à 0,1 et celle qu'il soit droitier est égale à 0,9.

1. Calculez la probabilité qu'un groupe de 10 individus contienne :
– au moins un gaucher ;
– au plus trois gauchers.

2. Un atelier de couture est équipé de 9 paires de ciseaux pour droitiers et de 3 paires de ciseaux pour gauchers. Quelle est la probabilité que chacun des 10 membres du personnel trouve une paire de ciseaux à sa convenance ?

3. Soit Z la variable aléatoire égale au nombre de personnes ayant trouvé une paire de ciseaux à sa convenance. Établir un tableau donnant Z en fonction du nombre Y de gauchers dans les 10 membres du personnel. En déduire la loi de probabilité de Z.

Exercice 6.2

Une compagnie d'assurances envisage de créer des polices d'assurances individuelles contre un certain type d'accidents. Une enquête préalable du service statistique a permis d'estimer qu'au cours d'une année, chaque personne a une chance sur 5000 environ d'être victime d'un accident couvert par ce type de police, et que la compagnie pourra vendre en moyenne 10 000 polices d'assurance de ce type par an.

Déterminez la probabilité que le nombre d'accidentés ne dépasse pas trois par an (on supposera que chaque personne assurée a au plus un accident par an).

Exercice 6.3

La société Alpha a vendu deux machines de pesage à la société Beta qui est une société de prestations de services. La Société Beta loue ces machines à la journée. Le prix de location lui laisse, par jour et par machine, une marge brute de 20 €. Chaque machine est immobilisée 1 jour sur 10 au hasard, pour réglage et contrôle.

1. Donnez, en la justifiant, la loi de la variable aléatoire Y égale au nombre de machines disponibles un jour quelconque.

2. Par ailleurs, on admet que le nombre d'entreprises désirant louer une machine pour une journée est une variable aléatoire Z ainsi définie :

Valeurs de Z	0	1	2	3
Probabilité	0,1	0,2	0,4	0,3

Cette loi de demande reste invariable au cours du temps et n'a aucune incidence sur le planning des immobilisations pour vérification, car aucune régularité temporelle n'a été décelée. Une entreprise est satisfaite si elle repart avec une machine. Soit N, la variable aléatoire égale au nombre d'entreprises satisfaites au cours d'une journée.

2.1. Quelles sont les valeurs possibles de N ?

2.2. En déduire la loi de N, ainsi que son espérance mathématique. Donnez la marge brute moyenne réalisée au cours d'une journée.

3. La société Alpha constate qu'une machine sur 20 tombe en panne la première année (on suppose qu'une machine ne peut pas tomber plus d'une fois en panne au cours d'une année). Le coût de réparation est de 200 € par machine.

Au lieu de garantir les machines pendant un an, la société Alpha propose de faire un discount de 20 € par machine aux acheteurs éventuels qui se chargeront eux-mêmes des réparations.

Vous êtes président-directeur général d'une société qui achète 60 machines. Quel choix ferez-vous si votre décision est prise en fonction de chacun des deux critères suivants :
– vouloir ne pas y perdre en moyenne ;
– refuser de courir un risque supérieur à 1 % de voir la remise consentie être inférieure au coût de réparation.

Exercice 6.4

Lors de tests d'accès à un ordinateur central par réseau télématique, on a constaté que 95 % des essais permettaient une connexion correcte. Une entreprise doit se connecter 5 fois dans la journée pour la mise à jour de ses fichiers. Soit Y le nombre d'essais nécessaires pour se connecter 5 fois.

1. Déterminez la loi de probabilité de la v.a. Y, ainsi que son espérance et sa variance.

2. Calculez $P(Y = 5)$ et $P(Y > 6)$.

Exercice 6.5

Soit X_t le nombre de demandes d'accès à une ressource informatique en partage, pendant un intervalle de temps de durée t fixée. On suppose X_t distribuée selon une loi de Poisson de paramètre $\mu = \lambda t$. Lorsque cette ressource est saturée, ce qui se produit avec une probabilité π, la demande ne peut être satisfaite. Soit Y_t le nombre de demandes satisfaites durant l'intervalle de temps de longueur t.

1. Montrez que la v.a. conditionnelle $\{Y_t | X_t = x\}$ suit une loi binomiale dont on déterminera les paramètres.

2. Montrez que Y_t suit une loi de Poisson de paramètre $\mu \cdot \pi$

3. On se donne $\mu = 10$ et $\pi = 0{,}2$, l'unité de temps étant la seconde.
Calculez : $P(Y_t < 8)$ et $P(3 < Y_t < 10)$

Exercice 6.6

Dans une étude sur le comportement d'achat de consommateurs, on suppose qu'à chaque minute, une unité (au maximum) d'un certain produit a 1 % de chances d'être vendue. On suppose les achats de ce produit effectués à des temps différents, indépendants les uns des autres.

1. Quelle est la loi de probabilité exacte du nombre d'unités de ce produit vendues en 30 min ?
Calculez la probabilité de vendre au moins 3 unités en 30 min.

2. Le magasin est ouvert 7 h 30 par jour. Quel est le nombre moyen d'unités vendues par jour ?

Par quelle loi peut-on approcher la loi de probabilité du nombre d'unités de ce produit vendues en un jour ?

3. Chaque matin, le stock est reconstitué à 8 unités pour le premier produit, et à 220 unités pour le second. Quelle est la probabilité de rupture de stock pour chacun des deux produits ?

Exercice 6.7

Dans un grand magasin, des observations sur un grand nombre de jours ouvrables au rayon des magnétoscopes ont amené à faire l'hypothèse selon laquelle le nombre de magnétoscopes X vendus au cours d'un jour ouvrable quelconque suit une loi de Poisson de paramètre 5. Les ventes sont supposées indépendantes.

1. Calculez la probabilité de chacun des événements suivants :
- la vente journalière de magnétoscopes est au plus égale à 2 ;
- la vente journalière de magnétoscopes est au plus égale à 2 ou au moins égale à 6 ;
- la vente journalière de magnétoscopes est au plus égale à 6 sachant qu'elle est au moins égale à 2.

2. Donnez, en la justifiant, la loi de la somme des ventes de deux jours consécutifs.

Calculez la probabilité que la somme des ventes de deux jours consécutifs soit égale à 10.

3. Le directeur du magasin décide de faire pendant une semaine une campagne publicitaire sur les magnétoscopes.

Il estime que, pendant cette semaine, la vente journalière suivra toujours une loi de Poisson et que son paramètre λ sera égal à 6 avec une probabilité égale à 2/3 ou à 8 avec une probabilité égale à 1/3.

Quelle est alors la probabilité que, pendant cette campagne publicitaire, la vente journalière de magnétoscopes soit au moins égale à 3 ?

Exercice 6.8

Dans une grande ville, la régie des transports urbains dispose de 1 000 autobus. Des observations antérieures ont montré que la probabilité qu'un autobus tombe en panne un jour donné est égale à 0,0025. Soit Y le nombre d'autobus en panne un jour donné.

1. Déterminez, en la justifiant, la loi de la variable aléatoire Y. Calculez son espérance et sa variance.

2. Donnez, en la justifiant, une loi approximative de la loi de la variable aléatoire Y.

3. Calculez $P(3 < Y < 7)$.

4. Quelle doit être la capacité minimum du service de maintenance des autobus pour que la probabilité que toutes les pannes soient traitées dans la journée, soit au moins égale à 0,998 ?

5. Le service de maintenance peut, en fait, réparer 6 pannes par jour. Calculez la probabilité que, un jour quelconque, ce service soit dans l'incapacité de réparer tous les autobus tombés en panne.

6. Soit Z, le nombre de jours de l'année (année de 365 jours) pour lesquels la maintenance est insuffisante. Déterminez, en la justifiant, la loi de la variable aléatoire Z, ainsi qu'une loi approchée. En déduire la probabilité que la maintenance soit suffisante tous les jours de l'année.

D'après examen de juin 2000, GEA 1re année Paris IX-Dauphine

Exercice 6.9

Le nombre de véhicules se présentant au péage *A* de l'autoroute du « Soleil » pendant un certain intervalle de temps *T* est supposé suivre une loi de Poisson de paramètre 3, et le nombre de véhicules se présentant au péage *B* de la même autoroute pendant le même intervalle de temps est supposé suivre une loi de Poisson de paramètre 2.

1. Déterminez la loi du nombre de véhicules se présentant à l'un ou l'autre des deux péages pendant un laps de temps *T* en précisant l'hypothèse nécessaire.

2. Calculez la probabilité que 8 véhicules se présentent à l'un ou l'autre des deux péages pendant un intervalle de temps *T*.

3. Sachant que 8 véhicules se sont présentés à l'un ou l'autre des deux péages, quelle est la probabilité qu'il y en ait eu 5 au péage *A* (toujours pendant un même laps de temps *T*) ?

D'après examen de juin 2006, DUGEAD 1re année Paris-Dauphine

Exercice 6.10

Les données suivantes, basées sur les annales de dix corps de l'armée prussienne sur une période de vingt ans à la fin du XIXe siècle, rendent compte du nombre de cavaliers tués par une ruade de cheval au cours d'une année.

Nombre de décès x_k	0	1	2	3	4	5 ou +
Cumul des années n_k	109	65	22	3	1	0

Données recueillies par le statisticien L. Bortkiewicz

1. Calculez la moyenne et la variance de cette distribution.

2. Proposez, en la justifiant, une loi de probabilité pour ajuster cette distribution.

D'après examen de septembre 2005, DUGEAD 1re année Paris-Dauphine

7. Les principaux modèles statistiques continus

I. Modèles continus simples

A. La loi uniforme continue

Nous avons déjà abordé la notion d'équiprobabilité dans les distributions statistiques discrètes au § I.B du chapitre 6 avec la loi uniforme discrète. Nous allons l'adapter au cas d'une variable aléatoire continue. Pour une telle variable, on ne peut pas parler de probabilité pour des valeurs isolées, et on imaginera la probabilité comme une masse répartie de façon diffuse. Il est clair alors que l'équiprobabilité se traduira par une probabilité d'intervalle *proportionnelle* à la longueur de l'intervalle. La probabilité cumulée sur tout $\mathbb{R}$ étant limitée à l'unité, on ne pourra avoir de probabilité non nulle que sur un sous-ensemble borné de $\mathbb{R}$.

Définition

Une variable aléatoire X, absolument continue, suit une *loi uniforme continue* sur l'intervalle $[a, b] \subset \mathbb{R}$ si sa densité de probabilité est donnée par :

$$f_X(x) = \begin{cases} k & \text{si} \quad x \in [a, b] \\ 0 & \text{si} \quad x \notin [a, b] \end{cases}$$

Compte tenu des propriétés d'une densité de probabilité, il résulte que $k > 0$ et que :

$$1 = \int_R f(x)dx = \int_a^b k dx = k(b-a) \quad \Rightarrow \quad k = \frac{1}{b-a}$$

L'intervalle sur lequel la densité n'est pas nulle est nécessairement fini. Cette contrainte apparaît tout à fait naturelle si on interprète la probabilité comme une masse.

D'autre part,

$$F_X(x) = \int_{-\infty}^{x} f(t)dt$$

par conséquent $F_X(x) = 0$ si $x \leq a$, alors que pour $x \in]a, b[$, on aura :

$$F_X(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{a} 0 \cdot dt + \int_{a}^{x} \frac{1}{b-a}dt = \frac{x-a}{b-a}$$

et enfin si

$$x \geq b,\ F_X(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{a} 0 \cdot dt + \int_{a}^{b} \frac{1}{b-a}dt + \int_{b}^{x} 0 \cdot dt = \frac{b-a}{b-a} = 1$$

La densité de probabilité d'une loi uniforme continue est donc constante par morceaux, tandis que sa fonction de répartition est linéaire croissante par morceaux (*cf.* figure 7.1, i et ii).

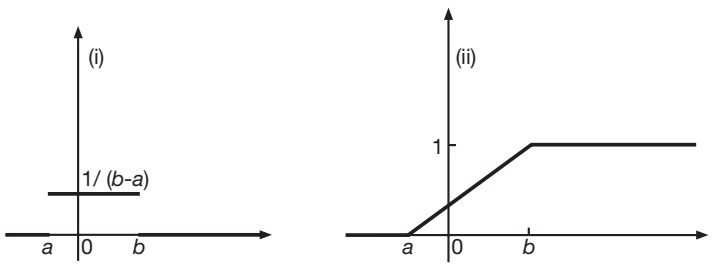


Figure 7.1 – Densité (i) et fonction de répartition (ii) de la loi uniforme continue

Pour ce qui concerne les moments de X :

$$E(X) = \int_{R} xf(x)dx = \int_{a}^{b} \frac{x}{b-a}dx = \frac{1}{b-a}\frac{b^2-a^2}{2} = \frac{b+a}{2}$$

et plus généralement :

$$E(X^k) = \frac{1}{b-a}\int_{a}^{b} x^k dx = \frac{1}{k+1} \cdot \frac{b^{k+1}-a^{k+1}}{b-a}$$

ce qui donne en particulier :

$$E(X^2) = \frac{1}{3}(b^2 + ab + a^2)$$

et par conséquent :

$$\sigma^2(X) = E(X^2) - (E(X))^2 = \frac{(b-a)^2}{12}$$

On note que la loi uniforme continue est symétrique, et que par conséquent, sa médiane et sa moyenne sont confondues au milieu de l'intervalle [a, b]. Comme pour la loi uniforme discrète, ce résultat est naturel compte tenu de l'équiprobabilité. Le calcul du coefficient d'asymétrie γ_1 de Fisher retrouve cette propriété puisque $\gamma_1 = 0$. On notera que cette distribution n'a pas de mode au sens strict.

Les fractiles de la loi uniforme continue sont très aisément calculables, comme pour toute distribution continue dont la fonction de répartition s'exprime analytiquement, et ici le calcul est particulièrement simple. En effet le fractile x_α d'ordre α est défini par $F_X(x_\alpha) = \alpha$. Il correspond à :

$$\frac{x_\alpha - a}{b-a} = \alpha, \text{ soit } x_\alpha = a + \alpha \cdot (b-a)$$

On retrouve la valeur de la médiane, égale à la moyenne, soit :

$$x_{0,5} = a + \frac{1}{2}(b-a) = \frac{b+a}{2}$$

Il faut remarquer que pour cette loi, la probabilité de voir une réalisation appartenir à un intervalle donné ne dépend que de la longueur de cet intervalle, et qu'elle ne dépend pas de la position de cet intervalle. Deux intervalles de même longueur auront la même probabilité, à condition qu'ils soient tous deux inclus dans [a, b], domaine de définition de X.

Compte tenu de la symétrie de cette loi, on peut aussi adopter la définition suivante.

Une variable aléatoire X est *uniforme continue* sur [$a - h, a + h$] si sa densité est définie par :

$$f(x) = \begin{cases} \dfrac{1}{2h} & \text{si } x \in [a-h, a+h] \\ 0 & \text{si } x \notin [a-h, a+h] \end{cases}$$

L'équivalence des deux définitions est laissée au lecteur. C'est, historiquement, cette seconde définition qui est à l'origine d'une grande utilisation de la loi uniforme continue dans le domaine de la prise en compte des erreurs d'arrondi [1].

1. On a ainsi pu étudier que la répartition des erreurs d'arrondi suit une loi uniforme continue dans de nombreux cas.

Cette distribution uniforme recouvre très naturellement la notion d'équiprobabilité dans le contexte de répartitions continues, et elle a été probablement utilisée comme telle avant le XVIIIe siècle, date des premiers écrits la concernant.

La loi uniforme continue est, en raison de son lien avec l'équiprobabilité, à l'origine de multiples modélisations (en sociologie, économie, gestion de flux, gestion de stocks, biologie, physique). On doit également mentionner son intérêt pour la simulation[1] des modèles, quel que soit le domaine d'application. Le résultat fondamental sur lequel reposent ces simulations est le suivant.

Théorème

Soit X une variable aléatoire continue dont la fonction de répartition $F(x)$ est supposée bijective. Alors la variable aléatoire $Y = F(X)$ suit une loi uniforme continue sur [0 ; 1].

En effet, la fonction F étant bijective et monotone, elle admet une réciproque qu'on note F^{-1}. Si on écrit la fonction de répartition de Y, on obtient :

$$P(Y < y) = P(F(X) < y) = P(X < F^{-1}(y)) = F[F^{-1}(y)] = y$$

ce qui prouve le résultat annoncé.

Partant donc d'une réalisation y de variable aléatoire distribuée selon une loi uniforme sur [0 ; 1], on peut reconstruire une réalisation x de variable aléatoire X de fonction de répartition F donnée en calculant $x = F^{-1}(y)$. Connaissant des réalisations de loi uniforme discrète, il est possible d'obtenir des réalisations pour une loi quelconque à partir du moment où on connaît sa fonction de répartition (analytiquement ou avec ses valeurs point par point).

Cette méthode extrêmement simple dans son principe (et basée sur un résultat élémentaire) permet une très grande quantité d'applications, que ce soit entre autres pour la simulation de modèles réels (flux, stocks…) ou pour l'étude de phénomènes aléatoires dont la distribution n'est pas connue *a priori*.

B. La loi exponentielle

Dans ce paragraphe, on présente la loi exponentielle sous son aspect le plus simple, sans tenir compte de la famille de lois dans laquelle elle se place. Afin de ne pas donner un contexte trop abstrait et mathématique, on ne parlera pas des *lois gamma* et des propriétés qui en découlent pour la loi exponentielle. Toutefois, en conclusion de ce paragraphe, on mentionnera les relations de ce modèle exponentiel avec des modèles correspondant à des schémas précis d'identification. Le but de cette présentation étant essentiellement de comprendre la nature des phénomènes aléatoires pour lesquels on envisage une représentation de type loi exponentielle ou dérivée de ce type.

1. Une présentation simple de la simulation, avec des exemples, est donnée à l'annexe III.

Définition

On dit qu'une variable continue X suit une *loi exponentielle de paramètres* $\lambda > 0$ et θ lorsque sa densité est :

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-(x-\theta)/\lambda} & \text{si} \quad x \geq \theta \\ 0 & \text{si} \quad x < \theta \end{cases}$$

La figure 7.2 montre l'allure des courbes représentatives de la densité de la loi exponentielle de paramètres θ et λ.

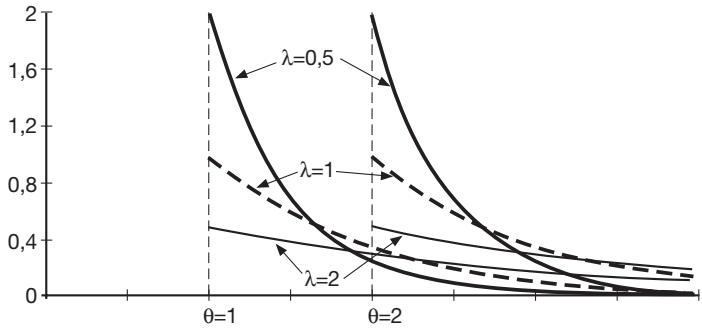


Figure 7.2 – Densités de lois exponentielles pour différentes valeurs de θ (1 *et* 2) *et de* λ (0,5, 1 *et* 2)

Les situations usuelles correspondent au choix de $\theta = 0$, ce que nous garderons pour la suite, en prenant pour densité la fonction :

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} & \text{si} \quad x \geq 0 \\ 0 & \text{si} \quad x < 0 \end{cases}$$

qui est ainsi une distribution à un seul paramètre. Le cas particulier où $\lambda = 1$ est dit *loi exponentielle standard*.

Un calcul élémentaire montre que la fonction de répartition (si $\theta = 0$) est donnée par :

$$F(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ 1 - e^{-x/\lambda} & \text{si} \quad x \geq 0 \end{cases}$$

Calculons les moments de cette distribution :

$$E(X^k) = \int_0^\infty \frac{1}{\lambda} x^k \cdot e^{-x/\lambda} \cdot dx$$

qui devient par le changement de variable $t = \frac{x}{\lambda}$:

$$E(X^k) = \int_0^\infty \frac{1}{\lambda} \cdot (\lambda t)^k \cdot e^{-t} \cdot \lambda \cdot dt = \lambda^k \cdot \int_0^\infty t^k \cdot e^{-t} \cdot dt$$

dont le calcul est très simple [1] en appliquant une intégration par parties à la dernière intégrale :

$$I_k = \int_0^\infty t^k \cdot e^{-t} \cdot dt = [-t^k \cdot e^{-t}]_0^\infty + \int_0^\infty k \cdot t^{k-1} \cdot e^{-t} \cdot dt = k \cdot I_{k-1}$$

ce qui permet d'écrire :

$I_k = kI_{k-1} = k \cdot (k-1) \cdot I_{k-2} = \ldots = k \cdot (k-1) \cdot (k-2) \cdot \ldots \cdot 2 \cdot 1 \cdot I_0$

et puisque I_0 est égal à 1 (intégrale de la densité), on obtient finalement $I_k = k!$ et :

$$E(X^k) = k! \cdot \lambda^k$$

Pour cette distribution exponentielle à un paramètre λ (θ étant nul), *l'espérance et l'écart-type sont tous deux égaux à* λ*, alors que la variance est égale à* λ^2

$$E(X) = \lambda \quad \text{et} \quad \text{var}(X) = \lambda^2$$

Dans le cas particulier de la loi exponentielle standard, il y a égalité de l'espérance, de la variance et de l'écart-type, tous égaux à l'unité. On peut établir un parallèle avec la loi discrète qui possède une propriété analogue, c'est-à-dire la loi de Poisson pour laquelle l'espérance est égale à la variance, mais non à l'écart-type (chapitre 6, § III.A).

Ce modèle correspond à des phénomènes aléatoires dont les « valeurs »[2] positives sont d'autant moins probables qu'elles sont grandes, la décroissance étant de type exponentiel. En pratique, on a souvent recours à un modèle exponentiel lorsqu'on a une variable continue positive, dont la moyenne et l'écart-type sont sensiblement égaux, et dont les probabilités d'intervalles de longueur fixe décroissent rapidement au fur et à mesure qu'ils concernent des valeurs élevées.

1. On notera que cette intégrale I_k n'est autre que la valeur de la fonction eulérienne classique gamma pour l'entier $(k + 1)$, soit $\Gamma(k + 1)$.

2. Le terme « valeur » ici ne doit pas prêter à confusion et concerne en toute rigueur un intervalle infiniment petit entourant une valeur ponctuelle ; ce terme est utilisé ici pour ne pas alourdir inutilement la présentation.

Une situation très classique aussi où on envisage un modèle exponentiel est celle où on s'intéresse au délai de survenue d'événements aléatoires dans le temps (souvent appelé *durée de vie*), et où on admet que le devenir X d'un individu (au sens statistique du terme) ne dépend pas de son âge :

$$P(X \leq x_0 + x | X > x_0) = P(X \leq x) \qquad \forall x > 0\,, \quad \forall x_0 > 0$$

On peut montrer que cette condition implique que X suit une loi de type exponentiel.

Ces modèles de durée de vie sont particulièrement utilisés en économie du travail et dans l'étude de l'amortissement des investissements, mais aussi bien entendu en fiabilité des matériels et en médecine.

Parmi les autres domaines d'application de la loi exponentielle, on citera la démographie et les files d'attente.

Les deux propriétés suivantes (données sans justification ni démonstration) peuvent être utiles pour l'identification d'une distribution exponentielle :

1. Si X_1 et X_2 sont deux variables indépendantes absolument continues telles que $V = \min(X_1, X_2)$ et $W = (X_1 - X_2)$ soient indépendantes, alors X_1 et X_2 sont des variables aléatoires exponentielles de même paramètre θ, mais pouvant avoir des écarts-type λ_1 et λ_2 différents.

2. Si X_1 et X_2 sont deux variables de même distribution absolument continue, et si $T = \dfrac{X_1}{X_1 + X_2}$ suit une loi uniforme continue sur [0 ; 1], indépendante de $(X_1 + X_2)$, alors X_1 et X_2 sont distribuées selon une loi exponentielle de mêmes paramètres $\theta = 0$ et λ.

L'écriture très simple de la fonction de répartition :

$$F(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ 1 - e^{-x/\lambda} & \text{si} \quad x \geq 0 \end{cases}$$

rend tous les calculs très simples avec une petite calculatrice.

Ainsi pour la loi exponentielle de paramètre $\lambda = 2$, on peut calculer la valeur du premier décile D_1 tel que $F(D_1) = 0{,}1$, d'où $\exp(-D_1/2) = 0{,}9$ et $D_1 = 0{,}21072$

De même, la valeur du premier quartile Q_1 est telle que $F(Q_1) = 0{,}25$, d'où :

$$\exp(-Q_1/2) = 0{,}75 \text{ ou encore } Q_1 = -2\ln(0{,}75) = 0{,}57536$$

De même encore, la valeur de la médiane Q_2 est telle que $F(Q_2) = 0{,}5$, d'où :

$$Q_2 = -2\ln(0{,}5) = 1{,}386$$

Ces trois valeurs particulières ont été reportées sur la figure 7.3.

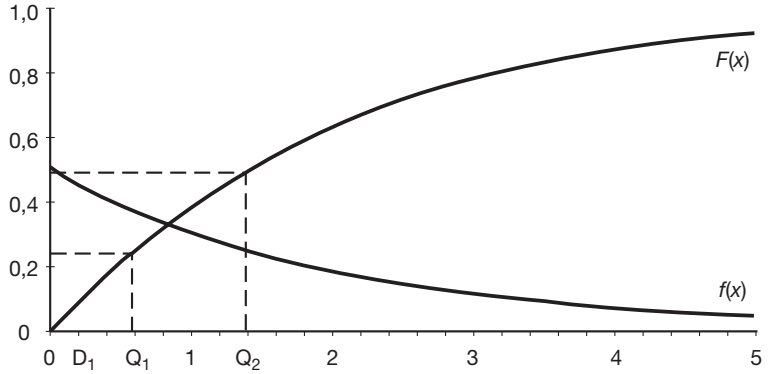


Figure 7.3 – Loi exponentielle de paramètres $\theta = 0$ et $\lambda = 2$: premier décile, premier quartile, médiane

Parmi les distributions de probabilité étroitement liées à la loi exponentielle, on citera :

• la *loi de Weibull* (étudiée au § IV.C) très utilisée dans les modèles de durée de vie ; elle correspond à une variable aléatoire X telle que $Y = (X - \lambda)^c$ suive une distribution exponentielle de paramètres $\theta = 0$ et λ ;

• la *première loi de Laplace* utile en statistique dès lors qu'on dispose de données dans lesquelles certaines valeurs sont « extrêmes » ; elle correspond à une double distribution exponentielle « en miroir » autour de la valeur θ ;

• la *loi dite du χ^2 à 2 degrés de liberté* (reprise au § III.A) qui n'est autre qu'une loi exponentielle pour laquelle $\theta = 0$ et $\lambda = 2$;

• la *loi d'Erlang*, extrêmement utilisée en gestion des files d'attente et fiabilité, est la loi suivie par la variable

$$Y = \sum_{j=1}^{n} \alpha_j X_j$$

dans laquelle les X_j ($j = 1, 2, \ldots, n$) sont indépendantes, toutes de loi exponentielle standard, et où les α_j sont tous distincts.

II. La loi normale ou loi de Laplace-Gauss

On dit encore *loi de Gauss* ou *loi gaussienne*, ou plus simplement *une gaussienne* (au lieu de variable aléatoire distribuée selon une loi de Gauss).

A. La loi normale centrée réduite

Définition

Une variable aléatoire suit une loi normale centrée réduite si elle peut prendre toute valeur réelle et si sa densité de probabilité est donnée par :

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

Cette fonction de densité est une fonction paire, et son graphique admet l'axe des ordonnées comme axe de symétrie. Il y a un maximum pour $x = 0$ qui correspond au mode de cette distribution. Compte tenu de deux points d'inflexion, le graphique est simple à tracer et présente l'allure caractéristique connue sous le nom de *courbe en cloche* (*cf.* figure 7.4).

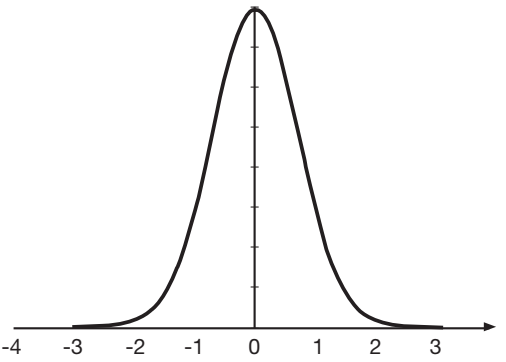


Figure 7.4 – Densité de la loi normale centrée réduite

Il n'existe pas de fonction analytiquement exprimable qui corresponde à une primitive de la fonction de densité *f*. La fonction de répartition d'une loi normale centrée réduite s'écrit :

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)dt$$

Le produit $t \cdot f(t)$ étant une fonction impaire : $\int_{-\infty}^{+\infty} t \cdot \exp(-t^2/2) \cdot dt = 0$

et il s'ensuit que $E(X) = 0$. Ce résultat justifie le nom de *variable centrée*.

Plus généralement la fonction $t^{2k+1} \cdot f(t)$ étant impaire, on a : $E(X^{2k+1}) = 0$

Pour le calcul de la variance, on calcule d'abord $E(X^2)$ par une intégration par parties et on obtient :

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} t^2 \cdot \exp(-t^2/2) \cdot dt$$

$$= \frac{1}{\sqrt{2\pi}} \left\{ [-t \cdot \exp(-t^2/2)]_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} \exp(-t^2/2) \cdot dt \right\}$$

Le premier terme de l'accolade étant nul, il s'ensuit :

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(-t^2/2) \cdot dt = 1$$

puisque le second membre n'est autre que l'intégrale de la densité sur l'ensemble des réels.

Ce résultat justifie le nom de *variable réduite*.

Cette distribution de probabilité possède une moyenne égale à 0. Le graphique étant symétrique par rapport à l'axe des ordonnées (parité de la densité), on a une surface totale (égale à 1) comprise entre la courbe et l'axe des abscisses, partagée en deux parties égales par l'axe vertical (soit 0,5 à gauche et 0,5 à droite). La médiane de cette distribution est aussi égale à 0. Enfin, le sommet de la « cloche » est au point $x = 0$

Propriété 1

Pour la *loi normale centrée réduite* , la valeur 0 représente à la fois la moyenne, la médiane et le mode.

On verra au § II.C, et au-delà, l'importance de cette propriété pour l'ensemble des applications du calcul des probabilités liées à la loi normale, centrée réduite ou non.

Par la suite cette v.a. normale centrée réduite sera toujours notée U pour bien l'identifier.

B. La loi normale $\mathcal{N}(m ; \sigma)$

Une variable normale centrée réduite U a pour moyenne 0 et pour variance 1. Prenons alors une variable X telle que $X = aU + b$ (avec $a \in \mathbb{R}$ et $b \in \mathbb{R}$). Il est clair que :

$$E(X) = aE(U) + b = b \quad \text{et} \quad \text{var}(X) = a^2 \cdot \text{var}(U) = a^2$$

Mais on peut aller plus loin encore et déterminer la fonction de répartition de X puis sa densité :

$$F_X(x) = P(X < x) = P(aU + b < x) = \begin{cases} P\left(U < \dfrac{x-b}{a}\right) & \text{si} \quad a > 0 \\ 1 - P\left(U < \dfrac{x-b}{a}\right) & \text{si} \quad a < 0 \end{cases}$$

$$\Rightarrow F_X(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{\frac{x-b}{a}} \exp(-t^2/2) \cdot dt & \text{si} \quad a > 0 \\ 1 - \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{\frac{x-b}{a}} \exp(-t^2/2) \cdot dt & \text{si} \quad a < 0 \end{cases}$$

On dérive cette fonction par rapport à x pour obtenir la densité de X :

$$f_X(x) = \frac{1}{|a|\sqrt{2\pi}} \exp\{-(x-b)^2/2a^2\}$$

Sur cette expression, on remarque que pour $a = 1$ et $b = 0$, on retrouve la densité de la loi normale centrée réduite. Dans le cas général, b représente la moyenne et $|a|$ l'écart-type.

Ceci conduit à poser que X suit une *loi normale de moyenne m et d'écart-type* σ lorsque X prend toute valeur réelle avec la densité :

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-(x-m)^2/2\sigma^2\}$$

et la loi normale centrée réduite en est un cas particulier. Ce résultat d'une importance pratique considérable, peut se présenter sous la forme générale suivante.

Propriété 2

Si X est une variable aléatoire normale, alors toute fonction du 1er degré (fonction affine) de X suit aussi une loi normale.

Ainsi que nous l'avons vu, la densité d'une variable normale, donc sa loi, dépend de deux paramètres. On a montré que le premier paramètre n'est autre que sa moyenne tandis que le second correspond à son écart-type. **Une variable aléatoire normale est entièrement déterminée par sa moyenne et son écart-type**.

Dans le cas d'une variable aléatoire X distribuée selon une loi normale $\mathcal{N}(m\,;\,\sigma)$, la variable aléatoire $Y = aX + b$ suivant aussi une loi normale avec $E(Y) = am + b$ et $\text{var}(Y) = a^2\sigma^2$ (l'écart-type de Y valant $\cdot|a|\,\sigma$), Y est distribuée selon une loi normale $\mathcal{N}(am + b\,;\,|a|\sigma)$.

En particulier, on peut construire la variable $\dfrac{X - m}{\sigma}$ de moyenne nulle et d'écart-type égal à 1 ; on dit alors que $\dfrac{X - m}{\sigma}$ est la variable normale centrée réduite déduite de X. C'est elle qui permet de faire aisément tous les calculs relatifs à X (§ II.C).

Réciproquement, toute v.a. X distribuée selon une loi de Gauss $\mathcal{N}(m\,;\,\sigma)$ peut s'écrire $X = \sigma U + m$ où U est une variable aléatoire distribuée selon une loi de Gauss centrée réduite.

L'étude de la densité

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}$$

d'une variable aléatoire normale $\mathcal{N}(m\,;\,\sigma)$ montre une courbe « en cloche » avec un axe de symétrie vertical en $x = m$.

La valeur de la moyenne détermine l'axe de la courbe de densité. Pour deux densités correspondant à un même écart-type, mais à deux moyennes différentes, on obtient deux courbes décalées (translatées) l'une par rapport à l'autre.

Comme on le constate aisément, le sommet de la courbe en cloche a pour ordonnée :

$$\frac{1}{\sigma\sqrt{2\pi}}$$

Cette valeur inversement proportionnelle à σ signifie que plus grand est l'écart-type d'une loi de Gauss, plus petit est le maximum de sa densité. Étant donné que la surface totale sous la courbe est constante (et égale à l'unité), on en déduit que la courbe est d'autant plus « aplatie » que l'écart-type est grand.

Au total, pour une valeur moyenne constante (sinon, il suffit de raisonner en translatant la courbe), plus l'écart-type est grand (c'est-à-dire plus la dispersion de la distribution est élevée), plus la densité aura des extrémités (aussi appelées *queues* de distribution) épaisses, compensant ainsi un sommet peu marqué. Inversement, plus l'écart-type est petit (c'est-à-dire plus la distribution est concentrée autour de sa moyenne) et plus le sommet de sa densité sera élevé, diminuant d'autant l'épaisseur aux extrémités (*cf.* figure 7.5).

Un écart-type petit correspond à une distribution resserrée autour de la moyenne, c'est-à-dire montrant par la finesse des queues de distribution que la probabilité de s'écarter « beaucoup » de la moyenne diminue très fortement en sens inverse de l'écart-type.

Nous avons déjà dit qu'une loi de Gauss était totalement caractérisée par sa moyenne et son écart-type. Nous voyons maintenant que, pour une telle distribution, l'écart-type donne une excellente appréciation de la dispersion.

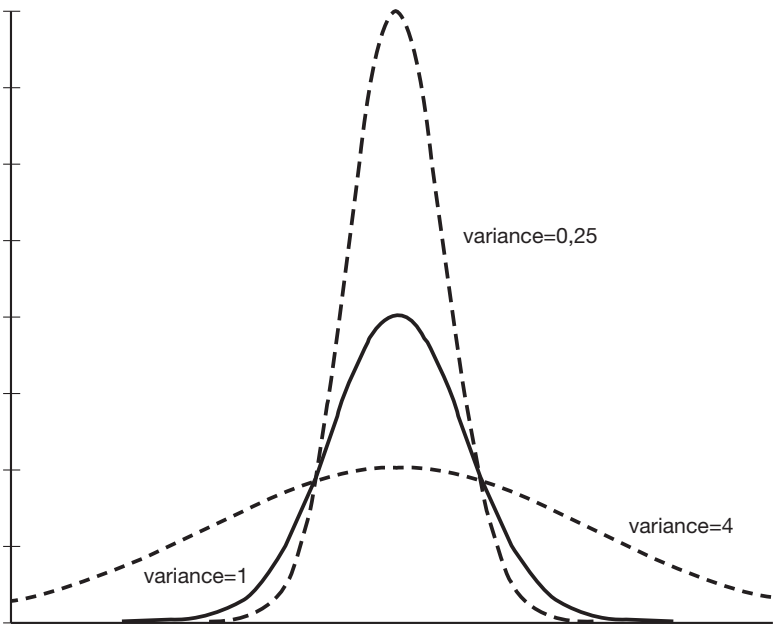


Figure 7.5 – Effet concentrateur de la diminution de la variance d'une loi normale

Pour finir, précisons ces notions de dispersion autour de la moyenne à l'aide de quelques résultats exprimés en terme de probabilité d'observer une valeur s'écartant de la moyenne de plus de k écarts-types (dispersion relative), pour plusieurs valeurs de k

En effet, soit X une v.a. distribuée selon une loi $\mathcal{N}(m\,;\,\sigma)$. Une valeur x qui s'écarte de la moyenne m (dans un sens ou dans l'autre, c'est-à-dire vers les valeurs inférieures ou vers les valeurs supérieures) de plus de k fois l'écart-type σ est caractérisée par l'inégalité :

$$|x - m| > k\sigma$$

La probabilité cherchée est :

$$P(|X-m| > k\cdot\sigma) = P\left(\frac{|X-m|}{\sigma} > k\right) = P\left(\left(\frac{X-m}{\sigma} < -k\right) \cup \left(\frac{X-m}{\sigma} > k\right)\right)$$

soit, $$P(|X-m| > k\cdot\sigma) = P\left(\frac{X-m}{\sigma} < -k\right) + P\left(\frac{X-m}{\sigma} > k\right)$$

La variable $\frac{X-m}{\sigma}$ étant centrée réduite, sa fonction de répartition est désignée par F_U. D'autre part, la densité de la loi normale centrée réduite étant symétrique par rapport à l'axe vertical, il en résulte que (*cf.* figure 7.6) :

$$F_U(-k) = P(U < -k) = P(U > k) = 1 - P(U < k) = 1 - F_U(k)$$

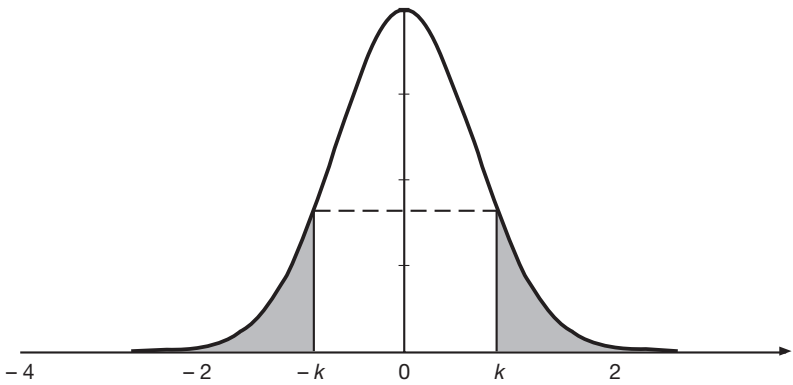


Figure 7.6 – Utilisation de la symétrie d'une loi de Gauss

Ce résultat implique que :

$$P(|X-m| > k\sigma) = P\left(\frac{X-m}{\sigma} < -k\right) + P\left(\frac{X-m}{\sigma} > k\right)$$

$$= 1 - F_U(k) + 1 - F_U(k) = 2(1 - F_U(k))$$

Nous verrons au § II.C la pratique de la lecture des tables. Nous admettrons pour le moment les différentes valeurs de $F_U(k)$ suivantes :

k	0,5	1	1,5	2	2,5	3
$F_U(k)$	0,6915	0,8413	0,9332	0,9772	0,9938	0,9987

Ceci permet d'obtenir le tableau des probabilités cherchées :

k	0,5	1	1,5	2	2,5	3
$P(\|X - m\| > k\sigma)$	0,6170	0,3174	0,1336	0,0456	0,0124	0,0026

Ces calculs donnent des résultats indépendants de m et de σ qui peuvent être convertis en dispersions absolues pour des valeurs données de σ comme on le verra également.

On note ainsi qu'il y a plus de 60 % des observations issues d'une loi $\mathcal{N}(m\,;\sigma)$ qui s'écartent de la moyenne de plus d'un demi écart-type, mais qu'il n'y a que 0,26 % (environ un quart de pour cent) qui s'écarte de la moyenne de plus de 3 écarts-types. En particulier pour une loi normale centrée réduite, cela signifie que 99,74 % des observations sont comprises entre – 3 et + 3. On remarquera encore que plus des deux tiers des observations issues d'une loi $\mathcal{N}(m\,;\sigma)$ sont comprises dans l'intervalle $[m - \sigma\,;\, m + \sigma]$, alors qu'il n'y en a plus que 4,56 % (moins de 5 %) à sortir de l'intervalle $[m - 2\sigma\,;\, m + 2\sigma]$. Tous ces éléments montrent bien la signification de l'écart-type d'une loi normale en termes de concentration des valeurs autour de la moyenne. Ce point est tout à fait fondamental pour la pratique des applications (estimations et tests) des calcu ls fondés sur la loi normale.

Pour finir de caractériser la loi de Laplace-Gauss, examinons les deux coefficients γ_1 et γ_2, caractérisant respectivement l'asymétrie et l'aplatissement.

Puisque $\gamma_1 = \dfrac{\mu_3}{\mu_2^{3/2}}$ où les μ_k désignent les moments centrés, il s'ensuit que $\gamma_1 = 0$. Ceci est tout à fait naturel et cohérent puisque ce coefficient vise à rechercher les entorses à la symétrie de la distribution.

Pour le coefficient d'aplatissement, $\gamma_2 = \dfrac{\mu_4}{\mu_2^2} - 3$ qui caractérise un degré de décroissance aux extrémités de la distribution, le calcul (pour μ_4 on procède par intégration par parties successives et on obtient $\mu_4 = 3\sigma^2$) donne $\gamma_2 = 0$. Ce coefficient d'aplatissement a été choisi afin d'avoir, par la loi normale, un échelon d'aplatissement relatif à la valeur 0.

Parmi les propriétés essentielles de la distribution de Gauss, on doit retenir :

La distribution normale est caractérisée par sa symétrie par rapport à la moyenne et, moyenne, médiane et mode sont confondus. Les coefficients d'asymétrie γ_1 et d'aplatissement γ_2 sont nuls.

Une autre propriété de la loi normale est essentielle dans la pratique.

Propriété 3

Si on a *n variables aléatoires normales* $\mathcal{N}(m_i ; \sigma_i)$ *et indépendantes*, alors leur somme suit une loi normale $\mathcal{N}(m ; \sigma)$.

On admettra ce résultat qui nécessite le calcul de la densité de la somme de v.a. continues indépendantes (produit de convolution, présenté au § III.A).

D'après ce qui précède, la loi de la somme admet pour paramètres respectivement :

– la somme des moyennes : $m = \sum_{i=1}^{n} m_i$

– la racine carrée de la somme des variances : $\sigma = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$

En effet, la moyenne d'une somme de variables aléatoires est toujours égale à la somme des moyennes, d'où la valeur de m. D'autre part, les variables aléatoires étant supposées indépendantes, la variance de la somme est égale à la somme des variances, ce qui montre le second résultat.

On notera bien que le second paramètre de la somme est la racine carrée de la somme des variances. Les variances s'additionnent lorsque les variables sont indépendantes[1], mais il n'en est jamais de même pour les écarts-types.

C. Usage des tables

Deux tables relatives à la loi de Gauss sont utilisées : la table de la *fonction de répartition* et la table des *fractiles* (annexes IV). Nous allons les examiner dans cet ordre.

Pour bien comprendre leur généralité, rappelons tout d'abord le point le plus fondamental des calculs de probabilités liés à une loi de Gauss :

$$X \rightsquigarrow \mathcal{N}(m ; \sigma) \quad \Leftrightarrow \quad U = \frac{X - m}{\sigma} \rightsquigarrow \mathcal{N}(0 ; 1)$$

1. On dit aussi parfois variables orthogonales en référence au théorème de Pythagore de la géométrie classique.

Par conséquent la probabilité d'un événement lié à X peut toujours s'exprimer par la probabilité d'un événement lié à U. Les tables de la loi de Gauss centrée et réduite permettent ainsi de calculer les probabilités associées à une loi de Gauss de moyenne et d'écart-type quelconques.

Tout d'abord examinons les utilisations de la table de la *fonction de répartition* de la loi $\mathcal{N}(0 ; 1)$. C'est une table à double entrée par laquelle on détermine la valeur de $P(U < u)$ pour $u \in [0 ; 3{,}5]$ donné.

On cherche :

i) la ligne correspondant à la partie entière et au 1 er chiffre décimal de u ;

ii) la colonne correspondant au 2 e chiffre décimal de u ;

puis à l'intersection de cette ligne et de cette colonne, on lit la probabilité cherchée.

Exemple

Par exemple si U suit une loi $\mathcal{N}(0 ; 1)$, on lit directement à l'intersection de la ligne correspondant à 0,3 et de la colonne correspondant à 0,08 :

$$P(U < 0{,}38) = 0{,}6480$$

et de même à l'intersection de la ligne portant 1,9 et de la colonne portant 0,06 :

$$P(U < 1{,}96) = 0{,}9750$$

Pour une loi normale quelconque X, la procédure est presque identique. Il faut simplement se ramener à une loi normale centrée réduite U, selon

$$U = \frac{X - m}{\sigma}$$

L'écart-type σ étant strictement positif :

si $\{X < a\}$, alors $\{X - m < a - m\}$ et $\left\{U = \frac{X - m}{\sigma} < \frac{a - m}{\sigma}\right\}$

mais aussi réciproquement si $\left\{U < \frac{a - m}{\sigma}\right\}$, alors $\{X < a\}$

Les événements $\{X < a\}$ et $\left\{U < \frac{a - m}{\sigma}\right\}$ étant identiques, ils ont la même probabilité. On s'est ramené à une lecture de table de loi normale centrée réduite.

▶ **Exemples**

Si X suit une loi $\mathcal{N}(3\ ;\ 2)$, $U = \dfrac{X-3}{2}$ suit une loi $\mathcal{N}(0\ ;\ 1)$

$$P(X < 6{,}24) = P\left(\frac{X-3}{2} < \frac{6{,}24-3}{2}\right) = P(U < 1{,}62) = F_U(1{,}62) = 0{,}9474$$

Si X suit une loi $\mathcal{N}(-4\ ;\ 5)$, $U = \dfrac{X+4}{5}$ suit une loi $\mathcal{N}(0\ ;\ 1)$

$$P(X < 1{,}65) = P\left(\frac{X+4}{5} < \frac{1{,}65+4}{5}\right) = P(U < 1{,}13) = F_U(1{,}13) = 0{,}8708$$

On remarque que la table n'est donnée que pour des valeurs de u (le seuil) comprises entre 0 et 3,49. Les extensions sont très simples :

• pour une valeur $u < 0$, on utilise la symétrie de la loi normale centrée réduite (*cf.* figure 7.7) :

$$F_U(-u) = P(U < -u) = P(U > +u) = 1 - \Pr(U < +u) = 1 - F_U(+u)$$

Par exemple, pour une variable X distribuée selon une loi $\mathcal{N}(4\ ;\ 2)$ calculons $P(X < 2)$:

$$P(X < 2) = P\left(\frac{X-4}{2} < -1\right) = F_U(-1) = 1 - F_U(1) = 0{,}1587$$

• pour les « grandes » valeurs de $|u|$ (c'est-à-dire au moins égales à 3,5) on dispose d'une ligne supplémentaire en bas de table, moins détaillée et s'arrêtant à $u = 4{,}5$. Ceci s'explique par le fait (étudié plus haut au § précédent) qu'une loi de Gauss n'a que moins de 0,30 % de chances de s'écarter de plus de 3 écarts-types de sa moyenne. Les probabilités cumulées (c'est-à-dire les valeurs de la fonction de répartition) pour des seuils supérieurs à 3 sont très proches de 1, et varient extrêmement peu. On le constate à la lecture de la ligne des grandes valeurs puisque lorsque le seuil passe de 4 à 4,5, la fonction de répartition n'augmente que de 0,00003 environ (alors qu'elle augmente de 0,017 entre 2 et 2,5, soit environ 550 fois plus !) ;

• enfin pour toute valeur de u contenant plus de 2 décimales, on fait l'habituelle interpolation linéaire. Si U suit une loi $\mathcal{N}(0\ ;\ 1)$, la valeur de $P(U < 1{,}645)$ se calcule en remarquant que 1,645 est exactement au milieu entre 1,64 et 1,65, valeurs pour lesquelles les probabilités cumulées sont respectivement de 0,9495 et 0,9505. On prend donc $P(U < 1{,}645) = 0{,}95$ (la valeur plus précise est en réalité de 0,94449).

Prenons un autre exemple, en calculant $P(X > 4{,}94)$ pour une v.a. X distribuée selon une loi $\mathcal{N}(-2\ ;\ 4)$. On écrira :

$$P(X > 4{,}94) = 1 - P(X < 4,\ 94) = 1 - P\left(\frac{X+2}{4} < \frac{4{,}94+2}{4}\right)$$

$$= 1 - P\left(\frac{X+2}{4} < 1{,}735\right) = 1 - F_U(1{,}735)$$

Pour obtenir le résultat, on relève dans la table :

$$F_U(1{,}73) = 0{,}9582 \text{ et } F_U(1{,}74) = 0{,}9591$$

ce qui permet d'obtenir $F_U(1{,}735) = 0{,}95865$ par interpolation, et $P(X > 4{,}94) = 0{,}04135$

Pour le calcul de la fonction de répartition F_U, il peut être souvent nécessaire d'utiliser une formule approchée, comme la formule de Hastings présentée à l'annexe II. De nombreux calculs automatiques, par exemple ont recours à ce procédé.

Parmi les autres utilisations de la table de la fonction de répartition de la loi de Gauss centrée réduite, on retrouve souvent le calcul de probabilités d'intervalles. Ce calcul repose sur la formule établie au chapitre 5, $P(a < X < b) = F(b) - F(a)$ dans laquelle F désigne la fonction de répartition de la v.a. continue X.

Prenons l'exemple de la v.a. X distribuée selon une loi $\mathcal{N}(-3\ ;\ 2)$ et pour laquelle on souhaite connaître $P(-4 < X < 0)$. On centre et on réduit au niveau des deux inégalités pour obtenir :

$$P(-4 < X < 0) = P(-0{,}5 < U < 1{,}5)$$

La probabilité cherchée est égale à :

$$\begin{aligned} F_U(1{,}5) - F_U(-0{,}5) &= F_U(1{,}5) - (1 - F_U(0{,}5)) = F_U(1{,}5) + F_U(0{,}5) - 1 \\ &= 0{,}9332 + 0{,}6915 - 1 = 0{,}6247 \end{aligned}$$

Tous les autres calculs, comme par exemple ceux de probabilités conditionnelles reposent de même sur le passage à une v.a. centrée réduite, et éventuellement sur les formules usuelles du calcul de base des probabilités.

Ainsi pour une variable aléatoire X distribuée selon une loi de Gauss $\mathcal{N}(1\ ;\ 3)$, on écrira pour calculer $P(0 < X < 2 \mid X > -2)$:

$$P(0 < X < 2 \mid X > -2) = \frac{P\Big((0 < X < 2) \cap (X > -2)\Big)}{P(X > -2)} = \frac{P(0 < X < 2)}{P(X > -2)}$$

$$= \frac{P(X < 2) - P(X < 0)}{1 - P(X < -2)} = \frac{F_U\left(\frac{2-1}{3}\right) - F_U\left(\frac{0-1}{3}\right)}{1 - F_U\left(\frac{-2-1}{3}\right)}$$

$$= \frac{F_U(0{,}33) - F_U(-0{,}33)}{1 - F_U(-1)} = \frac{2F_U(0{,}33) - 1}{F_U(1)} \approx 0{,}31$$

On remarquera à ce stade que pour les probabilités d'intervalles, il est indifférent de considérer des intervalles fermés, ouverts ou mixtes puisque la probabilité d'un point pour une v.a. continue est nulle (comme on l'a vu au chapitre 5, § II.B).

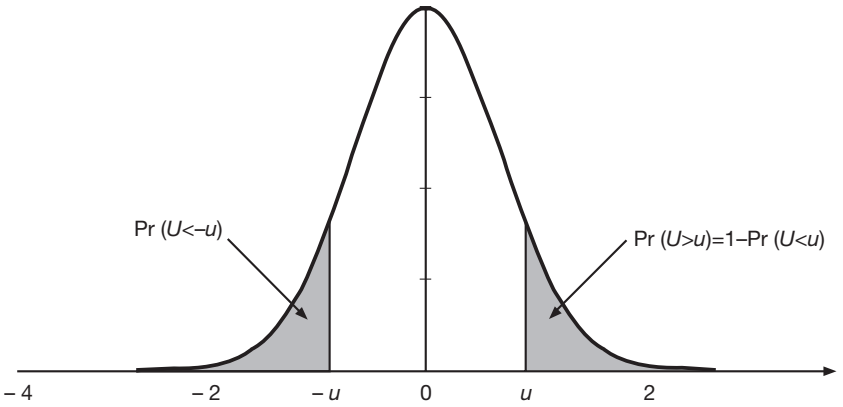


Figure 7.7 – Calcul de probabilité d'intervalles symétriques pour une loi de Gauss centrée réduite

Un calcul très fréquent est celui de $P(|U| < u)$, avec $u > 0$. Nous sommes dans le cas d'intervalles symétriques par rapport à la moyenne (*cf.* figure 7.7), ce qui revient à chercher :

$P(-u < U < +u) = P(U < +u) - P(U < -u)$, or $P(U < -u) = 1 - P(U < u)$

On a le résultat : $P(|U| < u) = 2P(U < u) - 1 = 2F_U(u) - 1$

Par complémentarité, on obtient également :

$$P(|U| > u) = 1 - P(|U| < u) = 2[1 - F_U(u)]$$

Par exemple : $P(|U| < 1{,}96) = 0{,}95$
$P(|U| < 1{,}645) = 0{,}90$

On peut retrouver ainsi quelques caractéristiques utiles de toute distribution normale.

Si X suit une loi normale $\mathcal{N}(m\ ;\ \sigma)$, en donnant à u successivement les valeurs 1, 2 et 3, on trouve que la probabilité que :

- X s'écarte de sa moyenne d'au plus 1 écart-type est $2 \cdot 0{,}8413 - 1 = 0{,}6826$
- X s'écarte de sa moyenne d'au plus 2 écarts-types est $2 \cdot 0{,}9772 - 1 = 0{,}9544$
- X s'écarte de sa moyenne d'au plus 3 écarts-types est $2 \cdot 0{,}9987 - 1 = 0{,}9974$

On pourrait aussi présenter ces résultats sous la forme (*cf.* figure 7.8) :

– il n'y a que 31,74 % des observations d'une loi normale qui s'écartent de la moyenne de plus de 1 écart-type ;

– il n'y a que 4,56 % des observations d'une loi normale qui s'écartent de la moyenne de plus de 2 écarts-types ;

– il n'y a que 0,26 % des observations d'une loi normale qui s'écartent de la moyenne de plus de 3 écarts-types.

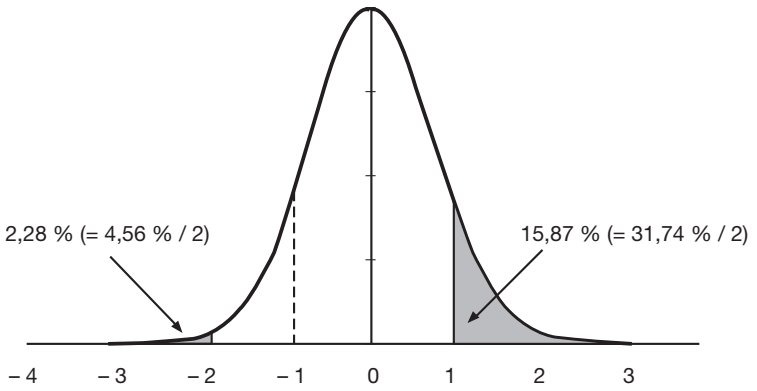


Figure 7.8 – Dispersion de la loi normale

Dans ce domaine gaussien, la valeur 1,96 est à retenir, car elle correspond à 95 % des observations, ou par complémentarité,

Il n'y a que 5 % des observations qui s'écartent de la moyenne de plus de 1,96 fois l'écart-type.

La troisième table relative à la loi de Gauss est celle des *fractiles*.

Rappelons tout d'abord qu'on appelle fractile d'ordre α $(0 \leq \alpha \leq 1)$ pour une distribution de fonction cumulative F, la valeur x_a telle que $F(x_\alpha) = \alpha$

C'est donc la valeur seuil x_α telle que sur l'ensemble $]-\infty ; x[$, on accumule une probabilité α.

Il est conventionnel, nous l'avons déjà mentionné, de désigner par U la loi normale centrée réduite, et par conséquent, on note u_α le fractile d'ordre α $(0 \leq \alpha \leq 1)$ pour cette répartition $\mathcal{N}(0 ; 1)$.

L'utilisation de cette table des fractiles présente une particularité : on peut avoir à la lire de deux façons différentes. En effet, la probabilité α pour laquelle on cherche le fractile se lit soit :

– sur la colonne de gauche (2 premiers chiffres décimaux de α) et la ligne supérieure (3e chiffre décimal de α) si $\alpha < 0,5$

– sur la colonne de droite (2 premiers chiffres décimaux de α) et la ligne inférieure (3e chiffre décimal de α) si $\alpha > 0,5$

Il suffit de prendre alors la valeur à l'intersection de la ligne et de colonne déterminées, en l'affectant :

• du signe + si $\alpha > 0{,}5$

• du signe – si $\alpha < 0{,}5$

Donnons quelques exemples d'application pour une loi de Gauss centrée réduite :

• le fractile d'ordre 0,010 est égal à $u_{0{,}01} = -2{,}3263$

• le fractile d'ordre 0,950 est égal à $u_{0{,}95} = +1{,}6449$ (on peut comparer ce résultat à celui obtenu plus haut dans ce même paragraphe, dans un exemple d'interpolation linéaire)

• le fractile d'ordre 0,250 (premier quartile) est égal à $u_{0{,}25} = -0{,}6745$

• le fractile d'ordre 0,750 (troisième quartile) est égal à $u_{0{,}75} = +0{,}6745$

La comparaison de ces deux derniers fractiles fait bien sûr ressortir la symétrie de la loi.

Notons encore deux fractiles qui jouent un grand rôle en statistique décisionnelle :

$$u_{0{,}975} = +1{,}96 \quad \text{et} \quad u_{0{,}995} = +2{,}5758$$

Si on s'intéresse à une v.a. X de loi normale quelconque $\mathcal{N}(m\ ;\ \sigma)$, on obtiendra le fractile x_α d'ordre α, par :

$$\alpha = P(X < x_\alpha) = P\left(\frac{X-m}{\sigma} < \frac{x_a - m}{\sigma}\right) = F_U\left(\frac{x_\alpha - m}{\sigma}\right)$$

la quantité $\dfrac{x_a - m}{\sigma}$ étant le fractile u_α d'ordre α d'une loi de Gauss centrée réduite.

On peut écrire $u_\alpha = \dfrac{x_a - m}{\sigma}$, soit $x_\alpha = m + u_\alpha \cdot \sigma$

Ainsi par exemple, le fractile d'ordre 0,675 pour une loi $\mathcal{N}(2\ ;\ 0{,}5)$ s'obtient à partir du fractile d'ordre 0,675 de la loi de Gauss centrée réduite lu sur la table, soit $u_{0{,}675} = 0{,}4538$, et vaut :

$$2 + 0{,}5 \cdot 0{,}4538 = 2 + 0{,}2269 = 2{,}2269$$

Par la même méthode, le fractile d'ordre 0,333 pour une loi $\mathcal{N}(-1\ ;\ 2)$ vaut :

$$-1 + 2 \cdot (-0{,}4316) = -1{,}8632$$

Dans le premier cas, 32,5 % des observations issues d'une population distribuée selon une loi $\mathcal{N}(2\ ;\ 0{,}5)$ sont supérieures à 2,2269. Dans le second cas, on peut constater qu'il y a une chance sur trois d'obtenir une observation tirée dans une population distribuée selon une loi $\mathcal{N}(-1\ ;\ 2)$ qui soit inférieure à – 1,8632

Ces calculs de fractiles sont particulièrement utiles pour l'obtention d'intervalles de confiance et la réalisation de tests.

Enfin, nous mentionnons l'existence (et l'utilisation) de la table dite de l'*écart-réduit*. Elle permet de lire les quantités $P(|U| > u)$, autrement dit les probabilités pour qu'une gaussienne s'écarte de sa moyenne de plus de u fois son écart-type.

Cette table présente un intérêt limité, car elle nécessite quelques calculs complémentaires dès lors que l'intervalle auquel on s'intéresse n'est pas symétrique autour de la moyenne. De plus on sait que

$$P(|U| < u) = 2(1 - F_U(u))$$

et ces probabilités sont en fait immédiates à obtenir à partir de la table de la fonction de répartition sans justifier de table supplémentaire.

D. Abord statistique de la loi normale

On a vu dans le paragraphe précédent que pour une variable normale, moyenne, médiane et mode sont confondus, et que la répartition est totalement symétrique par rapport à cette valeur.

Cette propriété essentielle est souvent utilisée devant des données observées pour poser une *hypothèse de normalité* c'est-à-dire pour envisager un modèle fondé sur une loi normale.

C'est donc à partir de l'examen de l'histogramme des données recueillies et de ses caractères de symétrie que l'on peut rechercher un modèle gaussien. Dans ce cas, on prendra tout de même bien soin de vérifier quelques caractéristiques de la dispersion des données pour avoir une appréciation plus complète. Pour cela, on examinera le pourcentage des observations qui s'écartent de la moyenne de moins d'un écart-type et de moins de deux écarts-types, et on comparera ces pourcentages aux valeurs théoriques qui seraient obtenues si les observations étaient purement gaussiennes, c'est-à-dire respectivement 68,26 % et 95,44 %. Très grossièrement, on recherche des pourcentages pas trop éloignés de 70 % et de 95 %.

Toutefois, même lorsque des données vérifient à la fois la symétrie et cette propriété de dispersion, on ne peut conclure qu'à une apparence de normalité (plus ou moins approximative selon les entorses à la symétrie et aux pourcentages de dispersion). Il s'agit là, comme ce qui a été présenté pour la loi de Poisson ou la loi exponentielle, d'une méthode pragmatique visant à reconnaître grossièrement un contexte possible de loi de Gauss.

Le diagramme *quantile-quantile*, présent dans la plupart des logiciels statistiques et très facile à construire avec un tableur muni de possibilités de représentations graphiques, permet une appréciation graphique de la concordance entre une distribution observée et un modèle théorique. Dans ce graphe, l'axe des ordonnées porte les fractiles de la distribution observée, tandis que l'axe des abscisses porte les fractiles correspondants de la loi théorique.

Le nuage des points s'aligne sur la première bissectrice lorsque la distribution théorique proposée est une bonne représentation des observations. On doit remarquer que l'appréciation de l'alignement des points le long de la bissectrice peut être considérée comme subjective. Toutes les déviations par rapport à l'alignement (extrémités présentant une courbure, points éloignés…) peuvent être repérées et analysées. En cas d'alignement, le type de modèle est alors retenu, et il reste à apprécier ses paramètres par une éventuelle translation et/ou inclinaison par rapport à la première bissectrice :

– un alignement sur une parallèle à la première bissectrice fera évoquer une erreur sur le choix de la caractéristique de position (moyenne…) de la distribution théorique ;

– un alignement sur une droite passant par l'origine mais inclinée par rapport à la première bissectrice évoquera une erreur sur la caractéristique de dispersion (écart-type…) ;

– un alignement sur une droite ne passant pas par l'origine et inclinée par rapport à la première bissectrice évoquera une erreur sur le choix des caractéristiques de position et de dispersion.

Prenons comme illustration l'exemple suivant. Un magasin désire adapter ses produits à sa clientèle, et pour cela, étudie le nombre de clients selon l'âge, à partir d'un échantillon de 100 clients. On a obtenu la répartition suivante :

Âge	< 20	[20-25[	[25-30[	[30-35[	[35-40[	[40-45[	[45-50[	≥ 50
Nombre de clients	8	10	13	17	22	11	12	7

Peut-on accepter une hypothèse de normalité pour l'âge des clients avec une moyenne 35 et un écart-type 10,5 ?

On calcule pour la borne supérieure de chaque classe le quantile théorique d'une loi normale centrée réduite correspondant à la fréquence cumulée observée, et on déduit le quantile théorique correspondant à la loi normale $\mathcal{N}(35 ; 10{,}5)$.

Quantile observé x_i	20	25	30	35	40	45	50
Fréquence cumulée F_i	0,08	0,18	0,31	0,48	0,70	0,81	0,93
Quantile théorique $\mathcal{N}(0 ; 1)$	– 1,4051	– 0,9154	– 0,4959	– 0,0502	0,5244	0,8779	1,4758
Quantile théorique $\mathcal{N}(35 ; 10{,}5)$	20,25	25,39	29,79	34,47	40,51	44,22	50,50

Appelons u_i^* et x_i^*, les quantiles théoriques $\mathcal{N}(0 ; 1)$ et $\mathcal{N}(35 ; 10{,}5)$ correspondant au i^e quantile observé x_i

À partir du quantile u_i^* tel que $F_U(u_i^*) = F_i$, on calcule le quantile x_i^* :

$$x_i^* = 10{,}5 \cdot u_i^* + 35$$

Puisque les points (x_i^* , x_i) sont à peu près alignés le long de la première bissectrice (*cf.* figure 7.9), l'ajustement par la loi normale $\mathcal{N}(35 ; 10{,}5)$ est retenu.

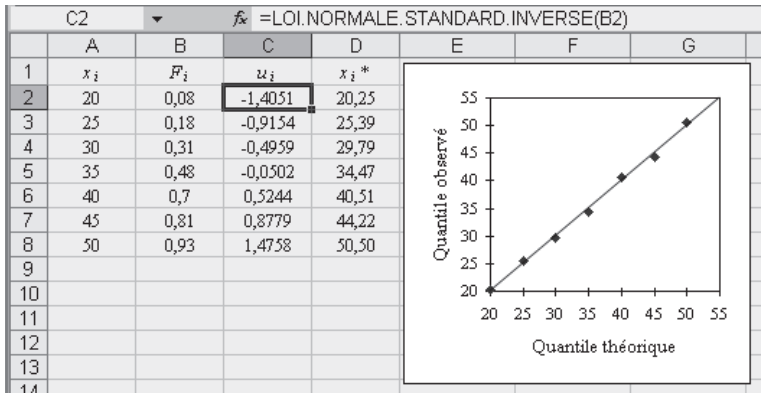


Figure 7.9 – Diagramme Quantile-Quantile d'une répartition observée sensiblement normale

Notons que le diagramme Quantile-Quantile ne s'applique pas seulement pour un modèle gaussien. On peut tracer un *diagramme Quantile-Quantile* pour un ajustement par une loi continue dont la fonction de répartition est strictement croissante, c'est-à-dire une loi dont la fonction de répartition est bijective sur l'intervalle correspondant à des valeurs non nulles de la fonction de densité et ne présentant pas de « trous »[1].

Nous en montrerons l'application pour la loi log-normale (§ IV.A) et pour la loi de Pareto (§ IV.B).

Le diagramme Quantile-Quantile s'utilise aussi pour comparer deux séries d'observations.

E. Abord probabiliste de la loi normale

C'est comme loi approchée qu'il est fait l'usage le plus habituel de la loi normale :

- soit on la retient pour des arguments statistiques tels que ceux évoqués au paragraphe précédent ;
- soit on remplace la distribution réelle (qui peut être connue ou inconnue) par une distribution gaussienne lorsqu'elle est une bonne approximation.

Cette recherche du recours à une loi de Gauss est guidée par deux considérations :

– d'une part, les calculs relatifs à des gaussiennes présentent une grande simplicité (notamment en raison de la symétrie) ;

1. Goldfarb B., Pardoux C., « Comment faire les diagrammes Quantile – Quantile ($Q - Q$) et Probabilité – Probabilité ($P - P$) ? », *La Revue de Modulad*, n° 33, juillet 2005 (www.modulad.fr, « Excel'Ense »).

– d'autre part, des résultats (dits asymptotiques) qui seront abordés ultérieurement donnent à de nombreuses v.a. particulières des lois sensiblement gaussiennes dès lors que leur application est fondée sur de très grandes séries d'observations.

Très concrètement, on peut être conduit à poser une hypothèse de normalité dans trois grandes situations.

i) Lorsqu'on a construit un schéma binomial où $n > 50$ et *p compris entre* 0,4 et 0,6 , on approxime (pour les calculs) par une loi normale $\mathcal{N}(np\ ;\ \sqrt{npq})$. La condition $npq > 18$ est quelquefois utilisée dans un souci de simplicité.

ii) Lorsqu'on a un modèle de Poisson dont le paramètre m est supérieur à 25, on approxime (pour les calculs) par une loi normale $\mathcal{N}(m\ ;\ \sqrt{m}\,)$.

Dans ces deux cas, la justification rigoureuse de l'approximation nécessite la connaissance de certaines formules d'analyse mathématique, telles que la formule de Stirling. Nous les admettrons sans démonstration. Quelques illustrations de ces approximations et de leurs limites sont données aux figures 7.10a et 7.10b.

iii) La somme de n variables aléatoires indépendantes, *suivant la même loi,* de moyennes m et d'écarts-types σ suit approximativement une loi normale dont la moyenne est la somme des moyennes, et l'écart-type est la racine carrée de la somme des variances, et ce dès que n est assez grand, soit en pratique $n > 30$

L'approximation de la loi binomiale par la loi normale en est un cas particulier de cette dernière situation puisqu'une v.a. binomiale est la somme de v.a. de Bernoulli.

Ce résultat joue un rôle essentiel dans toute la statistique classique. Sa démonstration est en dehors du cadre de ce livre.

Théorème central-limite (ou de la limite centrale, ou encore de la limite centrée)

X_i , $i = 1, 2, \ldots , n,$ v.a. indépendantes, de même loi,
de moyenne m, d'écart-type σ

$$\Rightarrow \sum_{i=1}^{n} \frac{X_i - m}{\sigma} \quad \xrightarrow{L} \quad \mathcal{N}(0\ ;\ \sqrt{n})$$

Insistons sur la nécessaire existence de m et de σ pour l'utilisation de cette forme du théorème central-limite, inapplicable sinon (loi de Cauchy, § 3.2).

Ce théorème a été étendu à la convergence en loi de variables aléatoires n'ayant pas la même distribution (donc avec des moyennes et des variances différentes, mais cependant toutes finies), sous la condition dite de Lindeberg, exprimant que les variables

$$\frac{X_i - m_i}{\sqrt{\sum \sigma_i^2}} \text{ sont « très petites » en probabilité}$$

Sous cette dernière forme, on peut alors interpréter la loi de Gauss comme la loi approximative des phénomènes résultant d'un grand nombre de « petites » causes indépendantes, et qui s'additionnent, sans qu'aucune de ces causes ne soit prédominante. Cette apparente généralité fait postuler trop souvent une hypothèse de normalité, en fait par défaut d'information (et/ou d'analyse) sur les causes d'un phénomène que l'on cherche à étudier.

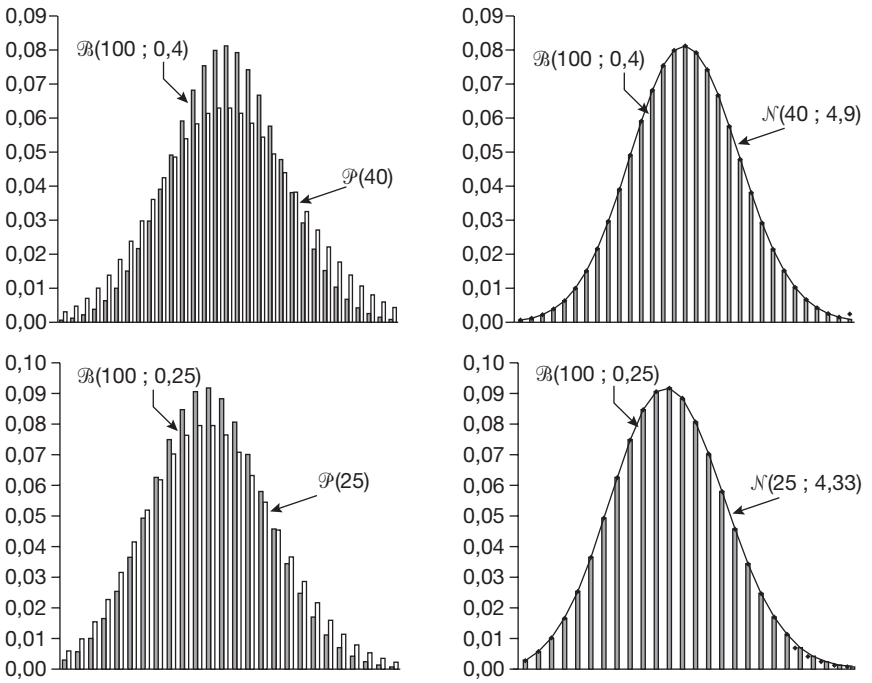


Figure 7.10a – Approximations correctes par la loi de Gauss, incorrectes par la loi de Poisson

Notons encore qu'un domaine d'application particulier de la loi normale par cette dernière approche est l'étude de la variable aléatoire

$$\overline{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

qui, lorsque les variables X_i sont toutes de même loi et indépendantes [1], porte le nom de *moyenne empirique*.

1. On parle alors de variables indépendantes et identiquement distribuées (soit i.i.d. en abrégé).

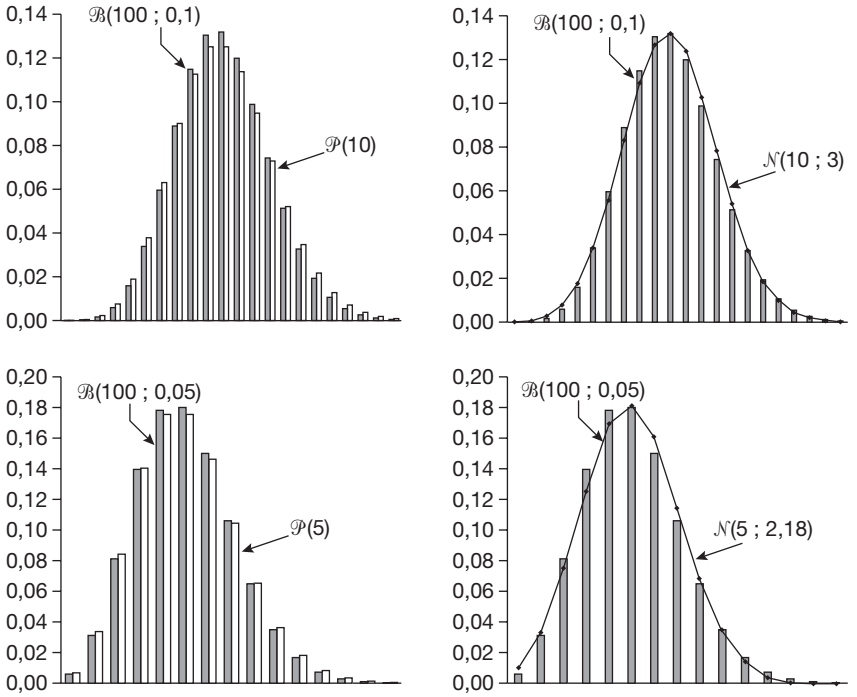


Figure 7.10b – Approximations incorrectes par la loi de Gauss, correctes par la loi de Poisson

L'ensemble des variables X_i caractérise n individus extraits d'une même population ; on l'appelle *n-échantillon* extrait de cette population. Dans ce contexte, $\overline{X_n}$ désigne la moyenne cet échantillon, mais au sens aléatoire, c'est-à-dire compte tenu des *fluctuations d'échantillonnage* . La moyenne empirique d'un échantillon dès lors que sa taille est assez grande (en pratique $n > 30$) suit à peu près une loi normale.

Ce résultat est d'une application extrêmement courante et montre déjà que le recours à la loi de Gauss est très classique pour des raisons qui mêlent, tout à la fois, les arguments statistiques et les arguments probabilistes. Les lois présentées au paragraphe suivant sont ainsi les lois fondamentales de l'univers gaussien.

De tout ce qui précède, on peut dresser le diagramme de la figure 7.11 qui résume les diverses approximations envisagées et montre bien la position clé de la loi de Gauss dans la modélisation de l'aléatoire, justifiant le terme souvent utilisé de *statistique gaussienne*.

La suite de ce chapitre (et notamment les lois du khi-deux, de Student, et de Fisher-Snedecor) relativisera légèrement cette apparence. On ne doit pas conclure à tort qu'un phénomène est gaussien en raison des multiples approximations. Nous avons déjà vu par exemple que la loi binomiale $\mathcal{B}(n\,;\,p)$ où $n = 100$ et $p = 0{,}05$ peut être approximée par une loi de Poisson de paramètre 5, et non pas par une loi de Gauss. Le dernier paragraphe de ce chapitre montrera plusieurs distributions de probabilité correspondant à des situations typées non gaussiennes. L'une d'elles, la loi de Pareto, définit un contexte probabiliste (univers parétien) différent de celui de la loi de Gauss.

On fera enfin particulièrement attention à ne pas donner le sens commun du mot « normal » pour une population distribuée selon une loi de ce type, cette interprétation étant le plus souvent admise en même temps que la généralisation abusive citée ci-dessus.

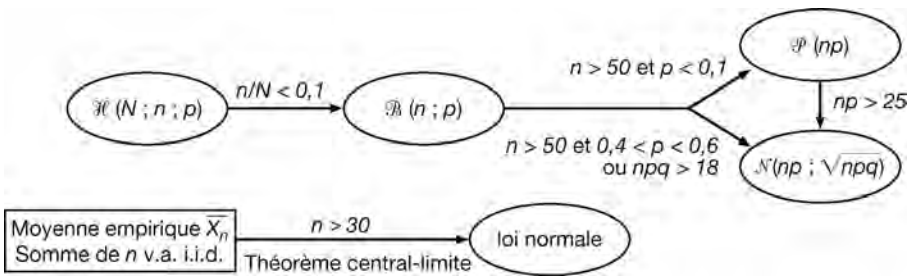


Figure 7.11 – Synthèse des approximations par la loi de Gauss

F. Correction de continuité

Une difficulté se pose lors de l'approximation d'une loi discrète (binomiale ou Poisson par exemple) par la loi normale qui est continue. En effet, les probabilités sont concentrées en des points pour la loi discrète, alors que la loi normale affecte la probabilité 0 à tout point. Si on a une loi $\mathcal{B}(100\,;\,0{,}4)$, comment calculer $P(X = 50)$ à partir de la loi normale ?

L'approximation normale est ici totalement justifiée (c'est même un des meilleurs cas !). Cette probabilité $P(X = 50)$ a une valeur exacte, difficile à calculer (dépassement de capacité), égale à 0,0103 ; la loi binomiale, la loi continue, lui affecte une valeur nulle !

On pallie cette difficulté par une *correction dite « de continuité »*, de la façon suivante.

Si X est une variable discrète qu'on approche par une loi normale $\mathcal{N}(m\,;\,\sigma)$, place la valeur k dans un intervalle symétrique et de largeur unité, et on pose :

$$P(X = k) = P(k - 0{,}5 < X < k + 0{,}5) \approx F_U\left(\frac{k + 0{,}5 - m}{\sigma}\right) - F_U\left(\frac{k - 0{,}5 - m}{\sigma}\right)$$

Ainsi par exemple, dans le cas de la loi $\mathcal{B}(100\,;\,0{,}4)$, approximée par la loi $\mathcal{N}(40\,;\,4{,}9)$

$$P(X = 50) \approx F_U(+\,2{,}14) - F_U(+\,1{,}94) = 0{,}01$$

alors que la valeur exacte calculée par la formule des probabilités binomiales est 0,0103

La formule présentée pour ce calcul n'est à utiliser que si le calcul par la loi exacte est trop délicat. Enfin, on doit noter que ce mode de calcul n'a de sens que pour obtenir des probabilités de loi discrète par des calculs approchés utilisant la loi normale.

III. Les lois dérivées de la loi normale

A. La loi du khi-deux

Si on dispose de n v.a. $\{X_i,\ i = 1,\ 2\ldots,\ n\}$ indépendantes et de même loi de Gauss $\mathcal{N}(m\,;\,\sigma)$, alors la variable aléatoire appelée moyenne empirique

$$\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ suit une loi } \mathcal{N}\left(m\,;\,\frac{\sigma}{\sqrt{n}}\right)$$

et caractérise « la moyenne » des X_i. Cette moyenne empirique permet de justifier plus encore l'utilisation de la loi de Gauss.

Dans le même contexte de gaussiennes indépendantes et de même loi, la v.a. Q définie par

$$Q = \frac{\sum_{i=1}^{n} (X_i - m)^2}{\sigma^2}$$

joue aussi un rôle essentiel. Pour des réalisations $\{ x_i , i = 1, 2,\dots , n\}$ de ces variables X_i, la variable aléatoire Q prend la valeur

$$\frac{\sum_{i=1}^{n} (x_i - m)^2}{\sigma^2} = \frac{n}{\sigma^2} \cdot \frac{\sum_{i=1}^{n} (x_i - m)^2}{n} = \frac{n \cdot s_n^2}{\sigma^2}$$

dans laquelle on reconnaît la variance s_n^2 de la série des réalisations. Au facteur multiplicatif près $\frac{n}{\sigma^2}$, la variable Q va décrire les réalisations de la variance des observations.

Ce rôle de caractéristique de la variance des observations est historiquement[1] celui qui a conduit à son étude détaillée. Cette v.a. Q peut aussi s'écrire :

$$Q = \frac{\sum_{i=1}^{n} (X_i - m)^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - m}{\sigma} \right)^2$$

c'est-à-dire comme somme de n carrés de v.a. gaussiennes centrées réduites indépendantes. C'est cette distribution qu'on étudie sous le nom de loi du *khi-deux à n degrés de liberté*, notée $\chi^2(n)$. Elle apparaît comme étroitement liée à l'étude de la variance. Dans ce paragraphe, nous présentons seulement la distribution du khi-deux, ses propriétés essentielles, la pratique des calculs, et sa place fondamentale dans l'ensemble des méthodes statistiques.

Définition

Si X est une v.a. distribuée selon une loi $\mathcal{N}(m ; \sigma)$, alors la loi de la v.a. $\left(\frac{X - m}{\sigma} \right)^2$ est dite *loi du khi-deux à 1 degré de liberté* , notée $\chi^2(1)$

Propriété

La densité de probabilité d'une loi $\chi^2(1)$ est donnée par :

$$\begin{cases} \frac{1}{\sqrt{2\pi}} e^{-x/2} x^{-1/2} & \text{si} \quad x > 0 \\ 0 & \text{si} \quad x \leq 0 \end{cases}$$

1. Par le mathématicien, probabiliste et démographe français I. J. Bienaymé (1796-1878) entre 1838 et 1852.

En effet, si Q suit une loi $\chi^2(1)$, on peut écrire $Q = U^2$ où U est une v.a. normale centrée réduite. On écrit la fonction de répartition de Q comme suit :

$$F_Q(x) = P(U^2 < x) = P(|U| < \sqrt{x}) = \int_{-\sqrt{x}}^{+\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

$$= \int_{-\infty}^{+\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt - \int_{-\infty}^{-\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

La dérivée de cette dernière expression par rapport à x donne la densité cherchée :

$$f_Q(x) = \frac{1}{\sqrt{2\pi}} e^{-x/2} \cdot \frac{1}{2\sqrt{x}} + \frac{1}{\sqrt{2\pi}} e^{-x/2} \cdot \frac{1}{2\sqrt{x}}$$

$$= \frac{1}{\sqrt{2\pi x}} e^{-x/2} = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2}$$

Bien entendu cette dernière expression est valable si $x > 0$. Compte tenu du fait que Q est un carré, l'événement $Q < x$ est impossible si $x \leq 0$, ce qui implique $f_Q(x) = 0$ si $x \leq 0$

Nous avons vu au chapitre 6 comment obtenir la loi d'une somme de deux v.a. discrètes indépendantes. Dans le contexte des v.a. continues, on admettra le résultat suivant.

Théorème

Si X et Y sont deux v.a. absolument continues indépendantes, de densités respectives $f_X(x)$ et $g_Y(y)$, alors la densité de probabilité de la somme $Z = X + Y$ est donnée par :

$$h_Z(z) = \int_{-\infty}^{+\infty} f_X(x) \cdot g_Y(z-x) dx = \int_{-\infty}^{+\infty} g_Y(y) \cdot f_X(z-y) dy$$

Cette expression qui lie les densités $f_X(x)$ et $g_Y(y)$ des v.a. X et Y est appelée *produit de convolution* de f_X et g_Y

Appliquons ce résultat à la somme des v.a. X et Y indépendantes et suivant chacune une loi $\chi^2(1)$:

$$f_X(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-x/2} x^{-1/2} & \text{si} \quad x > 0 \\ 0 & \text{si} \quad x \leq 0 \end{cases}$$

$$g_Y(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} e^{-y/2} y^{-1/2} & \text{si} \quad y > 0 \\ 0 & \text{si} \quad y \leq 0 \end{cases}$$

Ces résultats permettent d'obtenir la densité de la somme Z de deux v.a. distribuées selon des lois :

$$h_Z(z) = \frac{1}{2\pi} \int_0^z x^{-1/2} e^{-x/2} (z-x)^{-1/2} e^{-(z-x)/2} dx$$

$$= \frac{1}{2\pi} e^{-z/2} \int_0^z \Big(x(z-x) \Big)^{-1/2} dx$$

$$= \frac{1}{2\pi} e^{-z/2} \int_0^1 \Big(t(1-t) \Big)^{-1/2} dt$$

La dernière intégrale a été obtenue avec le changement de variable $x = z \cdot t$, et le calcul usuel donne la valeur π. La première intégration se fait entre 0 et z puisque $f_X(x) = 0$ si $x < 0$ et $g_Y(z-x) = 0$, si $z - x < 0$, soit si $x > z$. La densité de la somme Z est donnée par :

$$h_Z(z) = \begin{cases} \dfrac{1}{2} e^{-z/2} & \text{si} \quad z > 0 \\ 0 & \text{si} \quad z \leq 0 \end{cases}$$

Cette dernière expression n'est autre que la fonction densité de la loi exponentielle de paramètres $\theta = 0$ et $\lambda = 2$.

En tenant compte du fait que $\Gamma(1) = 1$ et que $z^0 = 1$, on peut écrire la densité de Z pour $z > 0$ sous la forme [1] :

$$\frac{1}{2^\alpha \Gamma(\alpha)} z^{\alpha - 1} e^{-z/2} \text{ pour } \alpha = 1$$

Cette formule :

$$f(x) = \begin{cases} \dfrac{1}{2^\alpha \Gamma(\alpha)} x^{\alpha - 1} e^{-x/2} & \text{si} \quad x > 0 \\ 0 & \text{si} \quad x \leq 0 \end{cases}$$

1. Il est particulièrement utile de se servir dans ce chapitre, de la fonction gamma, définie en tout point $x > 0$ par $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ avec $\Gamma\left(\frac{1}{2}\right) = \sqrt{x}$, $\Gamma(x) = (x-1)\Gamma(x-1)$ et $\Gamma(n) = (n-1)$

représente d'une part, lorsque $\alpha = 1/2$, la densité de la loi $\chi^2(1)$, et d'autre part, lorsque $\alpha = 1$, la densité d'une somme de deux lois $\chi^2(1)$

On peut montrer plus généralement pour toute valeur de α que cette formule donne la loi d'une somme de $2\,\alpha$ v.a. distribuées selon des lois $\chi^2(1)$

Ceci explique que, par extension, on appellera $\chi^2(n)$ la loi suivie par la somme de n carrés de variables aléatoires gaussiennes centrées réduites indépendantes. La densité de cette loi $\chi^2(n)$, dite loi du *khi-deux à n degrés de liberté* (*ddl* en abrégé), est donnée par la formule ci-dessus avec $\alpha = n/2$

Pour une v.a. X suivant une loi $\chi^2(1)$, il est très aisé d'obtenir les deux premiers moments, puisque $X = U^2$ où U est une v.a. gaussienne centrée réduite :

$$E(X) = E(U^2) = \text{var}(U) + E(U)^2 = 1$$

De même, on peut écrire :

$$\text{var}(X) = E(X^2) - E(X)^2 = E(U^4) - 1 = 3 - 1 = 2$$

Ces résultats se généralisent tout de suite au cas d'une loi $\chi^2(n)$ puisqu'une telle distribution est la somme de n v.a. i.i.d. de loi $\chi^2(1)$. Par conséquent, pour une v.a. Y de loi $\chi^2(n)$, on a :

$$E(Y) = n \quad \text{et} \quad \text{var}(Y) = 2n$$

Cette loi est une loi asymétrique, qui coïncide avec la loi exponentielle pour $\alpha = 1$, c'est-à-dire pour 2 ddl.

On peut calculer les caractéristiques de forme :

$$\gamma_1 = \sqrt{\frac{8}{n}} \quad \text{et} \quad \gamma_2 = \sqrt{\frac{12}{n}}$$

montrant bien l'asymétrie, mais aussi la tendance (*cf. infra*) vers une loi symétrique (la loi normale) lorsque le nombre de degrés de liberté augmente. De façon tout à fait évidente, ces deux coefficients tendent vers la valeur 0 qu'ils prennent pour une loi de Gauss.

La figure 7.12 donne la forme des distributions $\chi^2(n)$ pour quelques valeurs de n.

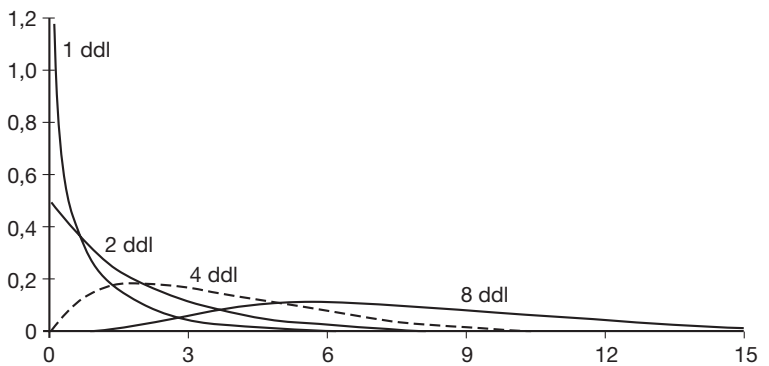


Figure 7.12 – Densités de quelques lois de khi-deux (à 1, 2, 4 et 8 ddl)

On remarque sur cette figure que, pour 1 ddl, la densité se comporte en $1/\sqrt{x}$ au voisinage de $x = 0$, et qu'elle décroît exponentiellement vers 0. Pour 2 ddl, il s'agit de la densité de la loi exponentielle de paramètres $\theta = 0$ et $\lambda = 2$. En dehors de ces deux cas particuliers, toutes les courbes partent de l'origine (d'autant moins rapidement que le nombre de ddl est élevé), elles présentent un maximum et décroissent lentement vers 0. Il faut noter que plus le nombre de ddl est élevé, plus la dissymétrie est atténuée. La loi du $\chi^2(n)$ étant la loi d'une somme de n v.a. i.i.d. (de loi $\chi^2(1)$ commune), le résultat évoqué au § II.D, indiquant que la loi d'une somme de v.a. i.i.d. peut être approximée par une loi de Gauss, montre bien que la loi du khi-deux à n ddl tend à se comporter comme une loi de Gauss lorsque n devient grand.

Cette relation entre la loi du khi-deux et la loi normale est traduite numériquement par la *formule de Wilson-Hilferty* :

$$P(\chi^2(n) < x) \approx F_U \left\{ \left(\left(\frac{x}{n} \right)^{1/3} - 1 + \frac{2}{9n} \right) \sqrt{\frac{9n}{2}} \right\}$$

reliant la fonction de répartition de la loi $\chi^2(n)$ à celle de la loi de Gauss centrée réduite qui est tabulée. Cette formule est une excellente approximation dès lors que le nombre n de ddl n'est pas trop petit (en pratique dès que $n > 10$).

Une autre possibilité de calculs approchés pour la fonction de répartition de la loi $\chi^2(n)$ est donnée par la formule de Fisher :

$$P(\chi^2(n) < x) \approx F_U \left(\sqrt{2x} - \sqrt{2n - 1} \right)$$

plus simple, mais donnant une moins bonne approximation. On ne l'utilise que pour $n > 30$

Enfin la loi du khi-deux à $2n$ ddl présente une relation très intéressante pour les calculs avec la loi de Poisson. Si X suit une loi $\chi^2(2n)$, et si Y suit une loi de Poisson de paramètre x, alors :

$$P(X > 2x) = P(Y > n - 1)$$

La figure 7.13 met bien en évidence l'allure dissymétrique de la courbe générale (c'est-à-dire pour $n > 2$), tant que le nombre de ddl n'est pas trop élevé.

Pour les calculs relatifs à la loi du khi-deux on dispose essentiellement de la table de *fractiles* (annexe IV). Cette table est à double entrée. Dans la colonne de gauche, on recherche la ligne correspondant aux degrés de liberté de la loi étudiée et dans la ligne supérieure, on recherche la probabilité cumulée α qui définira le fractile, noté $\chi^2_\alpha(2n)$. Donnons quelques exemples :

$\chi^2_{0,5}(5) = 4,351$	$\chi^2_{0,95}(10) = 18,307$	$\chi^2_{0,01}(8) = 1,64$	$\chi^2_{0,99}(6) = 16,812$
$\chi^2_{0,5}(30) = 29,336$	$\chi^2_{0,95}(40) = 55,76$	$\chi^2_{0,01}(40) = 22,16$	$\chi^2_{0,99}(30) = 50,892$

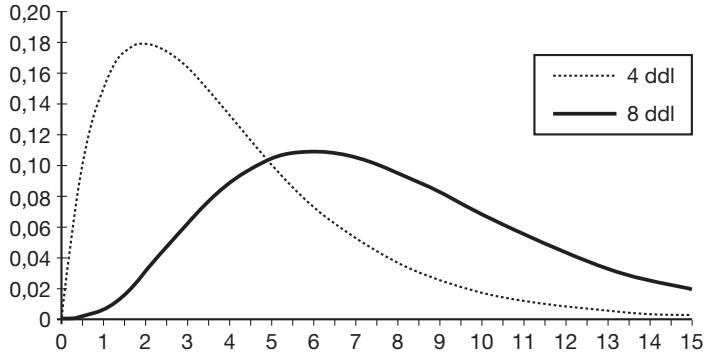


Figure 7.13 – Dissymétrie de la loi du khi-deux

La formule classique $P(a < X < b) = F(b) - F(a)$ permet de déterminer des probabilités d'intervalles à condition de disposer d'une table des fractiles assez détaillée. Dans le cas contraire, on peut recourir à des interpolations linéaires. Cependant, il vaut mieux éviter autant que possible cette dernière méthode et privilégier par exemple la formule de Wilson-Hilferty.

Donnons-en un exemple. Soit X une v.a. distribuée selon une loi de khi-deux à 25 ddl, et soit à calculer $P(14 < X < 39) = P(X < 39) - P(X < 14)$ Aucun des deux seuils ne se trouvant dans la table des fractiles, utilisons la formule de Wilson-Hilferty :

$$\begin{cases} P(\chi^2(25) < 14) \approx F_U \left\{ \left(\left(\frac{14}{25} \right)^{1/3} - 1 + \frac{2}{225} \right) \sqrt{\frac{225}{2}} \right\} = F_U(-1{,}770) \\ P(\chi^2(25) < 39) \approx F_U \left\{ \left(\left(\frac{39}{25} \right)^{1/3} - 1 + \frac{2}{225} \right) \sqrt{\frac{225}{2}} \right\} = F_U(1{,}789) \end{cases}$$

On obtient 0,925 comme valeur approximative par lecture de la table de la loi de Gauss.

Mentionnons pour finir, une propriété qui découle (comme dans le cas de la loi binomiale) directement de la définition concrète (somme de carrés de gaussiennes centrées réduites indépendantes) de la loi du khi-deux.

Si X et Y sont deux v.a. indépendantes distribuées selon des lois de khi-deux respectivement à n_1 et n_2 ddl, alors la v.a. $Z = X + Y$ est distribuée selon une loi $\chi^2(n_1 + n_2)$

B. La loi de Student

Définition

Si U et Y sont deux v.a. indépendantes suivant respectivement une loi $\mathcal{N}(0\,;\,1)$ et une loi $\chi^2(\nu)$, on dit que le quotient

$$\frac{U}{\sqrt{\frac{Y}{\nu}}} = \sqrt{\nu}\frac{U}{\sqrt{Y}}$$

suit une *loi de Student*[1] à ν degrés de liberté (ddl). On la note T_ν

On peut montrer que la densité de la v.a. T_ν est donnée par :

$$f_{T_\nu}(t) = \frac{1}{\sqrt{\nu}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Il s'agit donc d'une densité symétrique par rapport à l'axe des ordonnées. On en déduit que T_ν est une v.a. centrée :

$$E(T_\nu) = 0$$

Le calcul de la variance, ici égale à $E(T_\nu^2)$, donne pour $\nu > 2$:

$$\text{var}(T_\nu) = \frac{\nu}{\nu-2}$$

Lorsque $\nu = 1$ ou $\nu = 2$, la loi de Student n'admet pas de variance finie.

La loi de Student à 1 ddl est assez particulière. En effet, elle correspond au quotient de deux gaussiennes centrées réduites indépendantes puisque le dénominateur est la racine carrée d'une loi $\chi^2(1)$, c'est-à-dire la racine du carré d'une gaussienne centrée réduite. Cette loi T_1 porte aussi le nom de *loi de Cauchy*.

Sa densité s'écrit : $\frac{1}{\pi}\cdot\frac{1}{1+t^2}$. Elle est représentée sur la figure 7.14, en comparaison avec la loi de Gauss centrée réduite, par rapport à laquelle elle présente des queues de distributions très épaisses.

1. Student était le pseudonyme choisi par le statisticien William Sealy Gosset (1876-1937). Il fut l'un des premiers statisticiens du monde de l'entreprise, consacrant sa carrière à l'industrie agro-alimentaire (brasseries) au sein de laquelle il a toujours été reconnu à la fois comme industriel et comme scientifique. Très associé aussi au monde universitaire, il a largement contribué au développement scientifique de cette période.

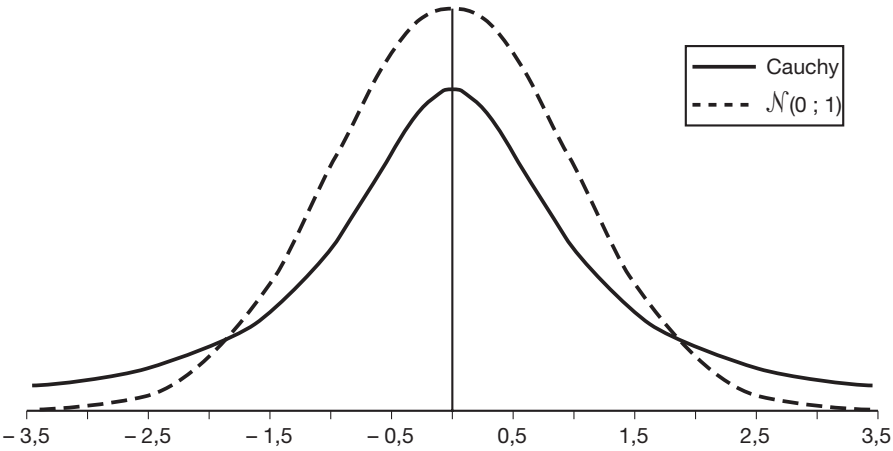


Figure 7.14 – Loi de Cauchy et loi de Gauss centrée réduite

Cela est encore plus évident si on compare la loi de Cauchy à la loi de Gauss centrée qui possède le même maximum, soit 0,3183, ce qui correspond à un écart-type égal à 1,2533. Dans cette comparaison, représentée figure 7.15, on voit que la probabilité qu'une v.a. de Cauchy dépasse la valeur 1 est égale à 0,25 (1 est ainsi le troisième quartile, et par symétrie –1 est le premier quartile de la loi de Cauchy), alors que la probabilité qu'une v.a. de Gauss $\mathcal{N}(0 ; 1{,}2533)$ dépasse la valeur 1 est égale à 0,213. De même, la probabilité qu'une v.a. de Cauchy dépasse la valeur 2 est égale à 0,1476, alors que pour la v.a. $\mathcal{N}(0 ; 1{,}2533)$, cette probabilité est égale à 0,055

Cette loi de Student à 1 ddl, ou loi de Cauchy, présente la particularité de n'avoir aucun moment fini autre que son espérance mathématique (qui est nulle). On retiendra que cette situation n'est pas du seul domaine de la théorie, mais qu'elle correspond au rapport de deux gaussiennes centrées réduites indépendantes.

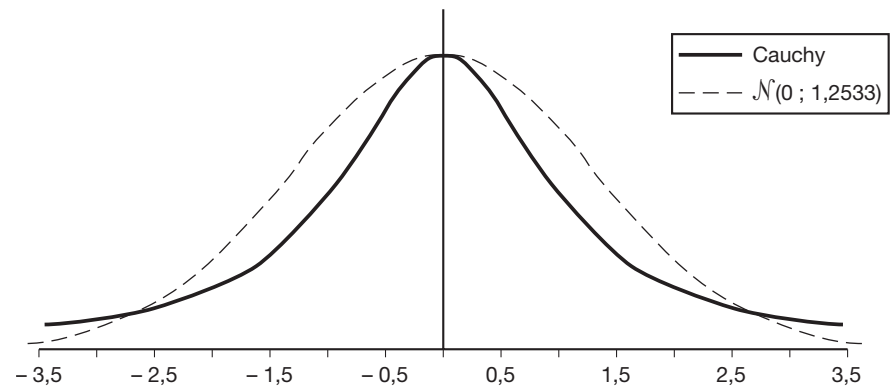


Figure 7.15 – Loi de Cauchy et loi $\mathcal{N}(0 ; 1{,}2533)$: comparaison des aplatissements

Dans le cas général de la loi de Student à ν ddl, nous avons vu que la moyenne est nulle, que la variance est supérieure à 1 et se rapproche de cette valeur au fur et à mesure que ν augmente. On a représenté sur la figure 7.16, les densités des lois de Student à 1, 2, 5 et 15 ddl ainsi que la densité de la loi de Gauss centrée réduite. Il apparaît clairement sur cette figure que la loi de Student devient très proche de la loi $\mathcal{N}(0 ; 1)$ lorsque son nombre de ddl augmente. En pratique, cette approximation est de bonne qualité dès lors que $\nu > 40$

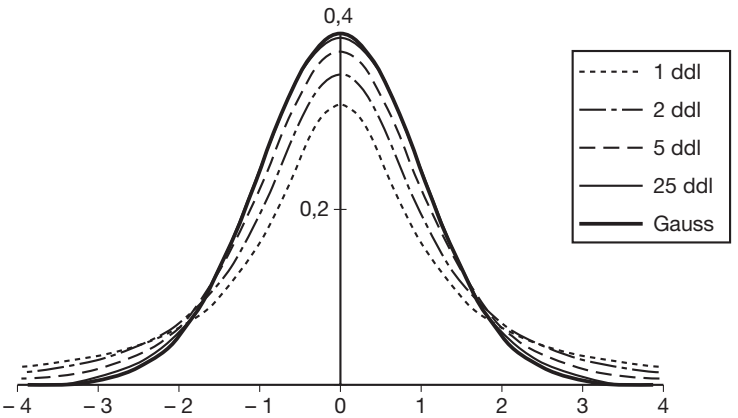


Figure 7.16 – Convergence de la loi de Student vers la loi de Gauss centrée réduite

Sur cette figure, on constatera aussi la relation existant entre les fractiles $t_\alpha(\nu)$ et u_ν de même ordre a respectivement pour la loi de Student T_k et pour la loi $\mathcal{N}(0 ; 1)$:

$$\begin{cases} t_\alpha(\nu) < u_\alpha < 0 & \text{si} \quad \alpha < 0{,}5 \\ t_\alpha(\nu) > u_\alpha > 0 & \text{si} \quad \alpha > 0{,}5 \end{cases}$$

De même, on remarquera que la symétrie de la loi de Student lie les fractiles $t_\alpha(\nu)$ et $t_{1-\alpha}(\nu)$ par la relation : $t_{1-\alpha}(\nu) = - t_\alpha(\nu)$

Le coefficient d'asymétrie γ_1 est bien entendu nul puisque la densité étant une fonction paire, tous les moments impairs sont nuls. Le calcul du coefficient d'aplatissement γ_2 est long, et nous admettrons le résultat :

$$\gamma_2 = \frac{6}{\nu - 4} \text{ valable si } \nu > 4$$

L'ensemble de ces deux résultats montre bien l'affinité de la loi de Student avec la loi de Laplace-Gauss, mais aussi la limite de cette similitude. En effet, on retrouve la symétrie de la loi et l'allure « en cloche » de la densité, mais c'est l'aplatissement aux extrémités qui fait la différence (et nous l'avons explicité pour la loi de Cauchy). Cependant, la valeur du coefficient

d'aplatissement de Fisher montre que cette différence s'atténue au fur et à mesure que le nombre de ddl augmente.

Les calculs relatifs à la loi de Student utilisent la table des *fractiles* (Annexe IV). Cette table se lit en recherchant :

i) dans la 1re colonne, la ligne correspondant aux ddl de la loi considérée ;

ii) dans la 1re ligne, la colonne correspondant à l'ordre α du fractile ;

et le fractile $t_\alpha(\nu)$ se lit à l'intersection de la ligne et de la colonne déterminés.

Ces fractiles sont donnés pour des valeurs de α supérieures à 0,5. Si $\alpha < 0{,}5$ on utilise la symétrie de la loi de Student et la formule $t_\alpha(\nu) = t_{1-\alpha}(\nu)$

Exemples

- le fractile d'ordre 0,9 d'une loi de Student à 5 ddl est égal à 1,476
- le fractile d'ordre 0,3 d'une loi de Student à 10 ddl est égal à – 0,542
- le fractile d'ordre 0,975 d'une loi de Student à 15 ddl est égal à 2,131
- le fractile d'ordre 0,975 d'une loi de Student à 30 ddl est égal à 2,042
- le fractile d'ordre 0,025 d'une loi de Student à 100 ddl est égal à – 1,984

Dans les deux derniers cas, les fractiles correspondant de la loi de Gauss centrée réduite sont égaux à 1,96 et à – 1,96. On retrouve bien que l'approximation de la loi de Student par la loi de Gauss est d'autant plus valable lorsque le nombre de ddl est élevé (en particulier s'il dépasse 40).

Comme pour la loi de Gauss et la loi du khi-deux, on dispose de formules approchées, pour la fonction de répartition ainsi que pour les fractiles, utilisées notamment pour les calculs répétés sur ordinateur.

Pour les fractiles, on utilisera l'approximation dite de Fisher et Cornish :

$$t_\alpha(\nu) \approx u_\alpha + \frac{1}{4\nu}u_\alpha(u_\alpha^2 + 1) + \frac{1}{96\nu^2}u_\alpha(5u_\alpha^4 + 16u_\alpha^2 + 3)$$

$$+ \frac{1}{384\nu^3}u_\alpha(3u_\alpha^6 + 19u_\alpha^4 + 17u_\alpha^2 - 15)$$

qui donne de bons résultats même pour de faibles ddl. Dès que $\nu > 30$, on pourra utiliser la formule beaucoup plus simple :

$$t_\alpha(\nu) \approx u_\alpha + \frac{u_\alpha}{4\nu}(1 + u_\alpha^2)$$

Pour la fonction de répartition dans le cas général, c'est-à-dire pour $\nu > 2$, les formules sont fastidieuses à écrire. Elles n'ont d'intérêt que pour des programmes de calculs et de simulation. Dans l'annexe II, nous indiquons les formules valables pour 1 ddl (l'erreur commise dans l'approximation par ces formules n'excède pas 0,001).

On notera que pour 2 ddl, la fonction de densité s'intègre sans difficultés et qu'on a la formule exacte pour la fonction de répartition :

$$P(0 < T_2 < t) = \frac{t}{2\sqrt{2 + t^2}} \text{ valable pour toute valeur de } t > 0$$

La loi de Student est utilisée principalement pour l'estimation et les tests. Dans ce qui suit, on justifie brièvement ce rôle.

Au paragraphe III.A, nous avons envisagé les quantités aléatoires $\overline{X_n}$ et Q dans le contexte de n v.a. X_i ($i = 1, 2, \ldots, n$) de loi $\mathcal{N}(m\,;\sigma)$. Ces deux v.a. suivent respectivement des lois $\mathcal{N}(m\,;\frac{\sigma}{\sqrt{n}})$ et $\chi^2(n)$. Nous pouvons écrire :

$$Q = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 \cdot \frac{n}{\sigma^2} = \frac{n}{\sigma^2} \cdot V^2 \quad \text{où} \quad V^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$$

la variable aléatoire V^2 prenant la valeur correspondant aux observations $\{x_i\}$, et pouvant être calculée quand on connaît la moyenne m.

De nombreuses situations ne correspondent pas à ce cas. Il faut souvent utiliser la v.a. :

$$S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2$$

pour représenter la variance, à partir de séries obtenues par échantillonnage, et remplacer la v.a. Q par :

$$Q' = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2}{\sigma^2} = \frac{(n-1)\cdot S^2_{n-1}}{\sigma^2}$$

Pour cette v.a. Q', on peut écrire :

$$Q' = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right)^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X_n}}{\sigma}\right)^2 = \sum_{i=1}^{n} Y_i^2 \quad \text{où} \quad Y = \frac{X_i - \overline{X_n}}{\sigma}$$

Ces v.a. Y_i sont gaussiennes, mais ne sont pas indépendantes puisque l'une d'entre elles s'exprime en fonction des autres :

$\sum_{i=1}^{n}\left(X_i - \overline{X_n}\right) = \sum_{i=1}^{n} X_i - n\overline{X_n} = 0$. On ne peut donc pas dire que leur somme, c'est-à-dire Q', est distribuée selon une loi $\chi^2(n)$. On montre qu'elle est en fait distribuée selon une loi $\chi^2(n-1)$, et qu'elle est indépendante de $\overline{X_n}$

Considérons maintenant la variable aléatoire :

$$T_{n-1} = \frac{\left(\frac{\overline{X_n} - m}{\sigma}\sqrt{n}\right)}{\sqrt{\frac{Q'}{n-1}}} = \frac{\overline{X_n} - m}{S_{n-1}} \cdot \sqrt{n}$$

D'après la définition donnée en tête de ce paragraphe, T_{n-1} suit une loi de Student à ($n - 1$) ddl. C'est la v.a. qui est utilisée en lieu et place de

$U = \frac{\overline{X_n} - m}{\sigma} \cdot \sqrt{n}$ lorsqu'on cherche à étudier la moyenne d'une répartition gaussienne dont la variance n'est pas connue au départ.

Toutefois, nous avons vu que lorsque le nombre de ddl augmente, la loi de Student peut être approximée par la loi de Gauss (sur la figure 7.16, on voit que pour 25 ddl, les deux densités sont déjà voisines). **On retiendra que le modèle de la loi de Student s'applique à l'étude de la moyenne d'une loi de Gauss dont la variance n'est pas connue, lorsqu'on ne dispose que d'un petit échantillon.**

C. La loi de Fisher-Snedecor

Définition

Si X et Y sont deux v.a. indépendantes distribuées selon des lois de khi-deux respectivement à ν_1 et ν_2 ddl, la v.a. :

$$F = \frac{\left(X/\nu_1\right)}{\left(Y/\nu_2\right)} = \frac{\nu_2}{\nu_1} \cdot \frac{X}{Y}$$

est dite suivre une loi de Fisher à ν_1 et ν_2 degrés de liberté. On la note $F(\nu_1, \nu_2)$

On fera très attention dans la pratique de cette distribution à l'ordre des degrés de liberté. La loi de Fisher-Snedecor [1] à ν_1 et ν_2 ddl n'est pas la même que la loi de Fisher-Snedecor à ν_2 et ν_1 ddl.

En effet, on peut écrire :

$$P(F(\nu_1, \nu_2) < c) = P\left(\frac{\nu_2}{\nu_1} \cdot \frac{X}{Y} < c\right) = P\left(\frac{\nu_1}{\nu_2} \cdot \frac{Y}{X} > \frac{1}{c}\right)$$

$$= 1 - P\left(F\left((\nu_2, \nu_1) < \frac{1}{c}\right)\right)$$

Si cette probabilité est égale à α, alors c n'est autre que le fractile $f_\alpha(\nu_1, \nu_2)$. Par conséquent, $1/c$ correspond au fractile $f_{1-\alpha}(\nu_2, \nu_1)$. On obtient ainsi la relation très utile, notamment dans la lecture des tables :

$$f_{1-\alpha}(\nu_2, \nu_1) = \frac{1}{f_\alpha(\nu_1, \nu_2)}$$

On peut montrer, par un calcul d'intégrales assez long, que la densité de la loi de Fisher-Snedecor $F(\nu_1, \nu_2)$ est donnée par la formule :

$$g_F(x)\begin{cases} \dfrac{\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \dfrac{x^{(\nu_1/2)-1}}{(\nu_2+\nu_1 x)^{(\nu_1+\nu_2)/2}} & \text{si} \quad x > 0 \\ 0 & \text{si} \quad x < 0 \end{cases}$$

expression dans laquelle $\Gamma(\nu)$ représente la fonction eulérienne gamma (§ III.A).

Lorsque $\nu_1 = 1$, la densité, comportant un terme en

$$x^{-1/2}(\nu_2 + \nu_1 x)^{-(\nu_2+1)/2} = \frac{1}{\sqrt{x(\nu_2 + \nu_1 x)^{\nu_2+1}}}$$

admet l'axe des ordonnées comme asymptote.

Lorsque $\nu_1 = 2$, la densité décroît régulièrement.

En dehors de ces cas, comme on le voit sur la figure 7.17, la densité de la loi de Fisher-Snedecor présente un maximum après une croissance d'autant plus rapide que les degrés de liberté du numérateur sont peu élevés, puis une décroissance lente. C'est une densité très dissymétrique.

1. L'étude de cette loi en tant que rapport de deux lois de khi-deux rapportées à leurs degrés de liberté est due au statisticien anglais Ronald Aymler Fisher (1890-1962), tandis que les développements numériques, et notamment l'établissement des tables ont été réalisés par le statisticien américain George Waddel Snedecor (1881-1974).

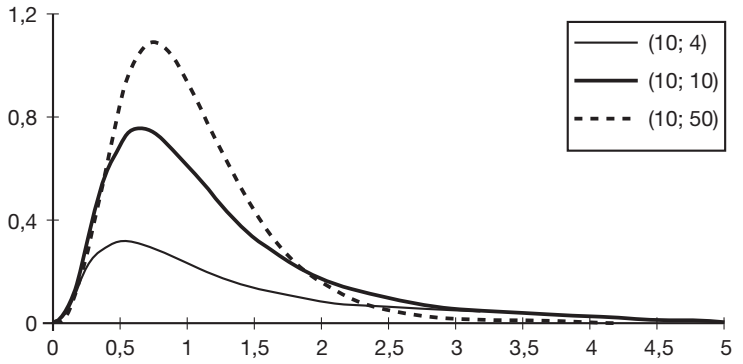


Figure 7.17 – Densités de lois de Fisher pour $\nu_1 = 10$ *et quelques valeurs de* ν_2

On remarquera que la loi de Fisher-Snedecor où $\nu_1 = 1$ se ramène à la loi de Student. En effet, le numérateur de la définition étant le carré d'une gaussienne centrée réduite, il s'ensuit qu'une telle loi $F(1, \nu_2)$ est le carré d'une v.a. distribuée selon une loi de Student à ν_2 ddl.

On peut déduire une relation entre les fractiles de la loi $F(1, \nu_2)$ et ceux de la loi $T(\nu_2)$, puisque :

$$1 - \alpha = P\Big(F(1, \nu_2) < f_{1-\alpha}(1, \nu_2)\Big) = P\Big(T^2(\nu_2) < f_{1-\alpha}(1, \nu_2)\Big)$$

soit encore :

$$1 - \alpha = P\Big(-\sqrt{f_{1-\alpha}(1, \nu_2)} < T(\nu_2) < +\sqrt{f_{1-\alpha}(1, \nu_2)}\Big)$$

$$= 2P\Big(T(\nu_2) < +\sqrt{f_{1-\alpha}(1, \nu_2)}\Big) - 1$$

Ceci revient à écrire :

$$P\Big(T(\nu_2) < +\sqrt{f_{1-\alpha}(1, \nu_2)}\Big) = 1 - \alpha/2$$

et on obtient la relation entre fractiles :

$$\sqrt{f_{1-\alpha}(1, \nu_2)} = t_{1-\alpha/2}(\nu_2)$$

qui est une traduction numérique de la propriété 1 ci-dessous.

Propriété 1

Si X est une v.a. distribuée selon une loi de Fisher à 1 et ν_2 ddl, alors X est le carré d'une v.a distribuée selon une loi de Student à ν_2 ddl.

Le calcul des caractéristiques de la loi de Fisher-Snedecor montre que :

i) $E\Big(F(\nu_1, \nu_2)\Big) = \dfrac{\nu_2}{\nu_2 - 2}$, l'espérance de cette loi ne dépend pas de ν_1, et n'est définie que pour $\nu_2 > 2$

ii) $\text{var}\Big(F(\nu_1, \nu_2)\Big) = \dfrac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$, la variance de cette loi n'est définie que pour $\nu_2 > 4$

On notera une relation, très utile au niveau des calculs, entre la loi de Fisher-Snedecor dont les degrés de liberté sont pairs et la loi binomiale (du type de celle qui est mentionnée au § III.A entre la loi de Poisson et la loi du khi-deux).

Propriété 2

Si X est une v.a. distribuée selon une loi de Fisher-Snedecor $F(2(n-r-1), 2r)$, et Y une v.a. distribuée selon une loi binomiale $\mathcal{B}(n\,;\,p)$, on a la relation :

$$P\left(X > \frac{1-p}{p} \cdot \frac{n}{n-r-1}\right) = P(Y \geq r)$$

Enfin, toujours pour la pratique au niveau des calculs, la formule suivante est une approximation correcte des fractiles de la loi de Fisher-Snedecor lorsque les degrés de liberté sont tous les deux suffisamment grands (au moins égaux à 50) :

$$f_\alpha(\nu_1, \nu_2) \approx \frac{1}{2}(\nu_2^{-1} - \nu_1^{-1}) + u_\alpha\sqrt{\frac{1}{2}(\nu_1^{-1} + \nu_2^{-1})}$$

où u_α désigne le fractile d'ordre α de la loi de Gauss centrée réduite.

La loi de Fisher-Snedecor joue un rôle essentiel dans tous les problèmes posés par l'étude de la variance. Elle est le support des méthodes d'analyse de la variance et d'analyse discriminante.

Les trois lois que nous avons présentées sous cet aspect de lois dérivées de la loi normale correspondent toujours à des variables aléatoires reliées à des variances d'échantillonnage dans des populations supposées gaussiennes. Ce ne sont donc pas des distributions qu'on peut identifier par une démarche concrète. Elles sont aussi souvent utilisées en tant que lois approchées, correspondant à des résultats asymptotiques.

Elles sont toutes trois caractérisées par un ou deux paramètres appelés *degrés de liberté* . Cette notion se justifie mathématiquement (géométriquement) pour la loi du khi-deux, et se déduit pour les deux autres distributions de leur construction à partir de lois du khi-deux.

On caractérise la complexité d'une structure par le nombre de ses paramètres **indépendants** qui la définissent (nombre de degrés de liberté).

Nous avons vu qu'une v.a. du khi-deux à n degrés de liberté est la somme de n v.a., carrés de v.a. gaussiennes centrées réduites indépendantes, et nous avons évoqué que la non-indépendance des v.a. dont on additionne les carrés diminue les degrés de liberté de la loi du khi-deux correspondante. On peut ainsi regarder l'ensemble des v.a. gaussiennes de départ comme engendrant linéairement un ensemble (espace) dont la dimension est alors le nombre de ces v.a. linéairement indépendantes. Chaque relation linéaire qui relie certaines de ces variables aléatoires diminue d'une unité la dimension de l'ensemble considéré, c'est-à-dire l'ensemble sur lequel les éléments statistiques sont définis.

IV. Quelques autres modèles continus courants

A. La loi log-normale

Soit une variable aléatoire continue X prenant des valeurs supérieures à un nombre donné x_0, la densité $f_X(x)$ étant nulle si $x \leq x_0$. Si la variable $Z = \ln(X - x_0)$ est distribuée selon une loi de Gauss $\mathcal{N}(m\,;\,\sigma)$, on dit que X est distribuée selon une loi log-normale notée $\mathcal{LN}(m, \sigma, x_0)$:

$$Z = \ln(X - x_0) \rightarrow \mathcal{N}(m\,;\,\sigma) \quad \Rightarrow Z = \sigma U + m \quad \text{avec} \quad U \rightarrow \mathcal{N}(0\,;\,1)$$

$$\Rightarrow \ln(X - x_0) = \sigma U + m \Rightarrow U = \frac{1}{\sigma}(\ln(X - x_0) - m)$$

Pour obtenir la densité de probabilité de X, on détermine d'abord sa fonction de répartition :

$$\begin{aligned} P(X \leq x) &= P(X - x_0 \leq x - x_0) \\ &= P[\ln(X - x_0) - m \leq \ln(x - x_0) - m] \\ &= P\left(U \leq \frac{\ln(x - x_0) - m}{\sigma}\right) \end{aligned}$$

car le logarithme est une fonction croissante. Par conséquent :

$$P(X \leq x) = \int_{-\infty}^{A} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \qquad \text{pour} \quad A = \frac{\ln(x - x_0) - m}{\sigma}$$

et en dérivant par rapport à x, on obtient la densité d'une loi log-normale :

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{A^2}{2}\right) \cdot \frac{1}{\sigma} \cdot \frac{1}{(x - x_0)}$$

$$= \frac{1}{\sigma\sqrt{2\pi}(x - x_0)} \cdot \exp\left\{-\frac{1}{2}\left(\frac{\ln(x - x_0)}{\sigma}\right)^2\right\}$$

La loi log-normale dépend de 3 paramètres, m, σ et x_0. Les deux premiers sont les moments de la loi normale associée, tandis que le dernier est une caractéristique de position puisque c'est la valeur minimale de cet aléa.

On peut montrer que l'espérance mathématique et la variance de X sont données par :

$$E(X) = x_0 + \exp\left(m + \frac{1}{2}\sigma^2\right)$$

$$\text{var}(X) = e^{2m} \cdot e^{\sigma^2} \cdot (e^{\sigma^2} - 1)$$

Déterminons maintenant la médiane Me :

$$F(Me) = 0{,}5 \Rightarrow A = \frac{\ln(Me - x_0) - m}{\sigma} = 0 \quad \Rightarrow \quad Me = x_0 + e^m$$

Plus généralement, le fractile d'ordre α d'une loi log-normale $\mathcal{LN}(m, \sigma, x_0)$, noté x_α, s'obtient de la façon suivante :

$$P(X \leq x_\alpha) = P\left(\ln(X - x_0) \leq \ln(x - x_0)\right) = P\left(\sigma U + m \leq \ln(x_\alpha - x_0)\right)$$

$$= P\left(U \leq \frac{\ln(x_\alpha - x_0) - m}{\sigma}\right) = \alpha$$

$$\Rightarrow u_\alpha = \frac{\ln(x_\alpha - x_0) - m}{\sigma} \Rightarrow x_\alpha = x_0 + e^{m + \sigma u_\alpha}$$

Le mode Mo de la distribution log-normale, correspondant au maximum de la densité, est :

$$Mo = x_0 + \exp(m - \sigma^2)$$

Puisque $\sigma^2 > 0$, on a $\exp(-\sigma^2) < 1$ et $\exp(\sigma^2/2) > 1$, ce qui implique que le mode Mo, la médiane Me et l'espérance mathématique $E(X)$ vérifient :

$$Mo < Me < E(X)$$

La figure 7.18 donne l'allure de la densité pour quelques valeurs de m et de σ, avec $x_0 = 0$

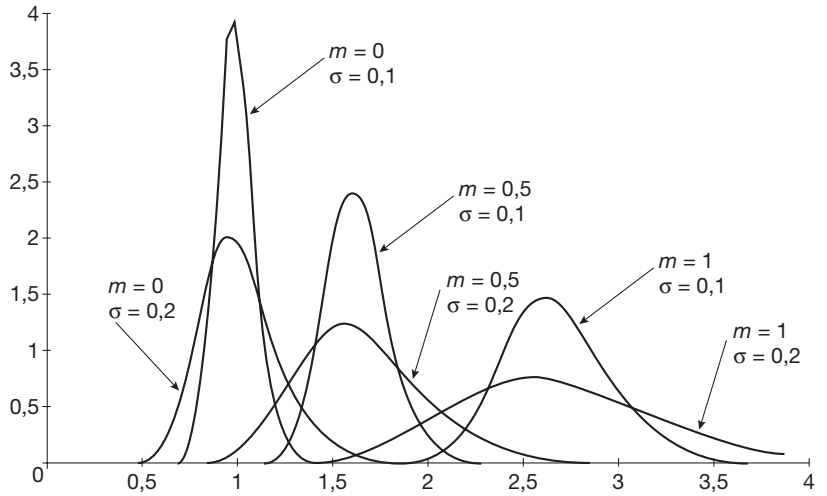


Figure 7.18 – Influence des paramètres m (0 / 0,5 / 1) *et* σ (0,1 / 0,2) *sur la densité de la loi log-normale*

L'expression du coefficient d'asymétrie $\gamma_1 = (e^{\sigma^2} + 2)\sqrt{e^{\sigma^2} - 1}$ montre que l'asymétrie dépend de σ (*cf.* figure 7.19).

En particulier, lorsque σ devient très petit, on obtient des distributions log-normales ressemblant fortement à des distributions normales, tout en ne prenant que des valeurs strictement positives.

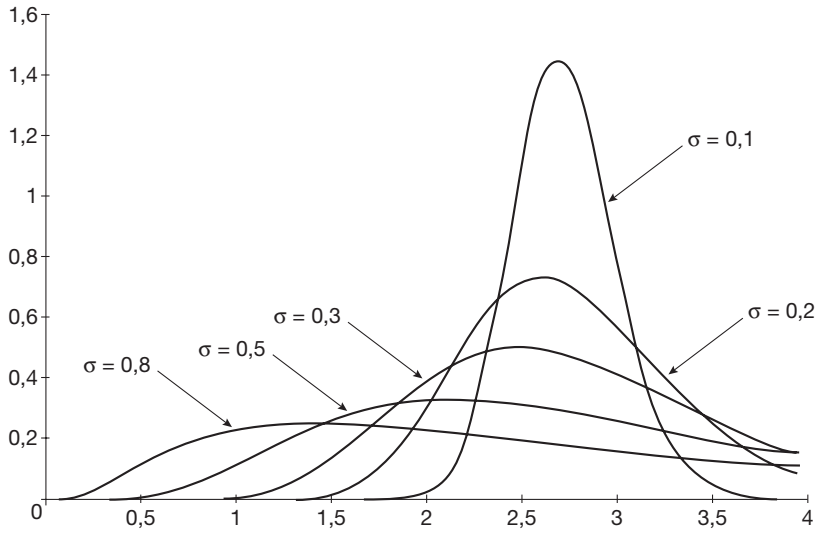


Figure 7.19 – Asymétrie de la loi log-normale $\mathcal{LN}(0, 1, \sigma)$ *en fonction de* σ

Cette dernière propriété fait de la loi log-normale un modèle très intéressant pour de nombreux phénomènes à valeurs positives, et pour lesquels la loi de Gauss donnerait des probabilités non nulles à des domaines de valeurs négatives. Par conséquent, lorsqu'une distribution gaussienne (respectant donc les caractéristiques de symétrie et d'aplatissement) apparaît adaptée à la représentation d'un phénomène aléatoire qui, toutefois, ne prend que des valeurs positives, on recherchera une loi log-normale dont les paramètres seront adaptés.

Dans ce contexte, on trouve notamment des répartitions de poids, de taille…

La durée des communications téléphoniques est très souvent un bon exemple de modélisation par la loi log-normale. Dans tous les cas, lorsque l'hypothèse de répartition log-normale est adoptée, il suffit de prendre le logarithme de toutes les observations pour se ramener à un contexte de loi normale, et d'appliquer alors toutes les propriétés vues au § II.

Les distributions de revenus sont souvent modélisées par la loi log-normale en raison de leur asymétrie, de leurs valeurs toujours bornées à gauche, et de l'effet atténuateur de la transformation logarithmique.

▶ **Exemple**

Illustrons ce propos en ajustant les distributions des salaires « Ensemble » en 2000 (chapitre 1, tableau 1.5) par une loi log-normale.

Si X suit une loi log-normale $\mathcal{LN}(m\ ,\ \sigma\ ,\ x_0)$,

alors $U\ =\ (\ln((X - x_0) - m))/\sigma$ suit une loi normale centrée réduite.

Dans ce cas, les points de coordonnées $\{\ u_i\ ,\ \ln(x_i - x_0)\}$ sont alignés, u_i étant le fractile d'ordre $i/10$ de la loi $\mathcal{N}(0\ ;\ 1)$ si x_i est le i^e décile observé, et x_0 le salaire minimum (égal à 6 200 € en 2000).

Le résultat de l'ajustement est présenté graphiquement (*cf.* figure 7.20). Les 9 points étant proches de l'alignement sur le graphique, l'ajustement par une loi log-normale peut être retenu.

C2 : =LOI.NORMALE.INVERSE(B2;0;1)

	A	B	C	D	E
1	*Déciles*	F_i	u_i	x_i	$\ln(x_i - x_0)$
2	D_1	0,1	-1,28	10 790	8,43
3	D_2	0,2	-0,84	12 220	8,70
4	D_3	0,3	-0,52	13 520	8,90
5	D_4	0,4	-0,25	14 910	9,07
6	*Médiane*	0,5	0,00	16 500	9,24
7	D_6	0,6	0,25	18 410	9,41
8	D_7	0,7	0,52	20 890	9,59
9	D_8	0,8	0,84	24 780	9,83
10	D_9	0,9	1,28	32 810	10,19
11					
12					

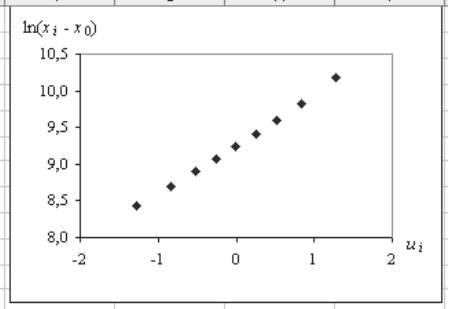


Figure 7.20 – Distribution des salaires « Ensemble » en 2000 ajustée par une loi log-normale

B. La loi de Pareto

Définition

On dit que la v.a. continue X est distribuée selon une loi de Pareto de paramètres α et $x_0 > 0$ si sa densité est donnée par :

$$f_X(x) = \begin{cases} \dfrac{\alpha}{x_0}\left(\dfrac{x_0}{x}\right)^{\alpha+1} & \text{si} \quad x \geq x_0 \\ 0 & \text{si} \quad x < x_0 \end{cases}$$

On notera d'abord que cette fonction $f_X(x)$ ne définit une densité que si $\alpha > 0$

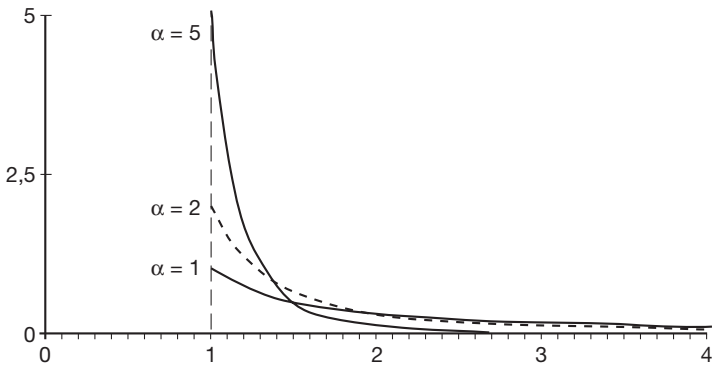


Figure 7.21 – Densités de lois de Pareto, pour $x_0 = 1$, et plusieurs valeurs de α

La fonction de répartition de X (*cf.* figure 7.22) est donnée par un calcul élémentaire. Elle est, bien entendu, nulle pour $x < x_0$, et si $x \geq x_0$, on a :

$$F_X(x) = \int_{x_0}^{x} \left(\frac{\alpha}{x_0}\right)\left(\frac{x_0}{t}\right)^{\alpha+1} dt = \left(\frac{\alpha}{x_0}\right) x_0^{\alpha+1} \int_{x_0}^{x} \frac{dt}{t^{\alpha+1}} = 1 - \left(\frac{x_0}{x}\right)^{\alpha}$$

On écrira donc :

$$f_X(x) = \begin{cases} 1 - \left(\dfrac{x_0}{x}\right)^{\alpha} & \text{si} \quad x \geq x_0 \\ 0 & \text{si} \quad x < x_0 \end{cases}$$

La probabilité d'une valeur supérieure à un seuil fixé x, tel que $\{x > x_0 > 0\}$, est égale à $\left(\frac{x_0}{x}\right)^{\alpha}$

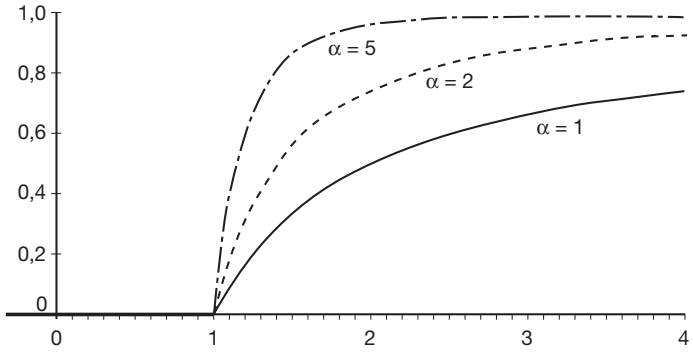


Figure 7.22 – Fonctions de répartition de lois de Pareto pour $x_0 = 1$ et plusieurs valeurs de α

Pareto[1] a introduit sa loi pour décrire des unités économiques par des caractères de taille (revenu, chiffre d'affaires, budget d'investissement…). Pour de telles grandeurs, on postule le plus souvent que le nombre d'individus dont le caractère étudié dépasse un seuil x est donné par $N_x = \frac{C}{x^{\alpha}}$ où C et α sont des constantes. L'application de la loi de Pareto aux distributions de revenus est une des plus usuelles, le paramètre α étant en général voisin de 2

Le calcul des moments est simple. On a d'abord :

$$E(X) = \int_{x_0}^{\infty} x\left(\frac{\alpha}{x_0}\right)\left(\frac{x_0}{x}\right)^{\alpha+1} dx = \alpha x_0^{\alpha} \int_{x_0}^{\infty} \frac{dx}{x^{\alpha}} = \frac{\alpha x_0}{\alpha - 1}$$

mais ce résultat n'est valable (convergence de l'intégrale) que si $\alpha > 1$ Remarquons qu'une v.a. distribuée selon la loi de Pareto ne prend que des valeurs positives ($x_0 > 0$), ce qui implique que son espérance mathématique est positive. Un résultat correspondant à $\alpha < 1$, soit $(\alpha - 1) < 0$, serait absurde.

Plus généralement, on a :

$$E(X^k) = \int_{x_0}^{\infty} x^k\left(\frac{\alpha}{x_0}\right)\left(\frac{x_0}{x}\right)^{\alpha+1} dx = \alpha x_0^{\alpha} \int_{x_0}^{\infty} \frac{dx}{x^{\alpha+1-k}} = \frac{\alpha x_0^k}{\alpha - k}$$

1. La loi étudiée dans ce paragraphe porte le nom de l'économiste italo-suisse Wilfrid Pareto (1848-1923). C'est à lui qu'on doit l'hypothèse, qu'il a supposée « universelle », de la décroissance en $x^{-\alpha}$ de la proportion des individus dont le revenu dépasse la valeur x.

mais ce calcul n'a de sens que si l'intégrale est convergente, c'est-à-dire si $(\alpha - k + 1) > 1$, soit si $\alpha > k$. Le moment d'ordre k n'est donc défini que lorsque $\alpha > k$. En particulier, la variance n'est définie que si $\alpha > 2$. Son calcul est simple :

$$\text{var}(X) = E(X^2) - (E(X))^2 = \frac{\alpha x_0^2}{\alpha - 2} - \frac{\alpha^2 x_0^2}{(\alpha - 1)^2} = \frac{\alpha x_0^2}{(\alpha - 2)(\alpha - 1)^2}$$

Les lois conditionnelles dérivées d'une loi de Pareto ont la propriété essentielle d'être encore des lois de Pareto. Si X suit une loi de Pareto de paramètres α et x_0, et si $x_1 \geq x_0$

$$P(X \leq x | X \geq x_1) = \frac{P(x_1 \leq X \leq x)}{1 - P(X < x_1)} = \frac{(x_0/x_1)^\alpha - (x_0/x)^\alpha}{(x_0/x_1)^\alpha} = 1 - (x_1/x)^\alpha$$

ce qui montre bien que la v.a. $(X | X \geq x_1)$ suit une loi de Pareto de paramètres α et x_1

On a ainsi ramené l'« origine » de la distribution en x_1, sans rien changer à sa forme. On notera aussi qu'on a :

$$E(X | X \geq x_1) = \frac{\alpha x_1}{\alpha - 1}$$

L'écriture de la fonction de répartition pour $x \geq x_0$, $F(x) = 1 - \left(\frac{x_0}{x}\right)^\alpha$ permet de voir que :

$$\ln(1 - F(x)) = \alpha \ln(x_0) - \alpha \ln(x)$$

Les points de coordonnées $\{\ln(x) ; \ln(1 - F(x))\}$ sont donc situés sur une droite de pente $-\alpha$, coupant l'axe vertical en un point d'ordonnée $\alpha \ln(x_0)$

Réciproquement, si les points de coordonnées $\{\ln(x) ; \ln(1 - F(x))\}$ pour une distribution de fonction cumulative F sont situés sur une même droite de pente $-\alpha$, cette distribution correspond à une loi de Pareto puisque si $\ln(1 - F(x)) = -\alpha \ln(x) + K$, on a :

$$1 - F(x) = x^{-\alpha} \cdot e^K = (A/x)^\alpha \quad \text{avec } A^\alpha = e^K$$

Cette représentation graphique fournit une méthode très simple (et efficace) pour apprécier si une distribution observée, soit $(x_i, i = 1, 2, \ldots, n)$ peut être représentée par un modèle de Pareto, en traçant le nuage des points de coordonnées ,$\{\ln(x_i) ; \ln(1 - F(x_i))\}$ où $F(x_i)$ est la valeur de la fonction cumulative en x_i (chapitre 1). L'utilisation d'échelles logarithmiques sur les deux axes est très appréciable. Elle est très courante sur les logiciels.

On peut ainsi valider l'hypothèse de représentation par une loi de Pareto, mais aussi identifier les paramètres puisque la constante α est la valeur abso-

lue de la pente de la droite (elle peut être calculée par la méthode des moindres carrés, chapitre 3).

Il faut remarquer que dans cette représentation graphique, on trouve en ordonnée les valeurs (sur une échelle logarithmique) de $\{1 - F(x)\}$, c'est-à-dire de la proportion résiduelle au-delà de la valeur x. Ceci nous reporte à l'introduction de cette loi par Pareto, et à l'utilisation très fréquente des échelles logarithmiques dans l'étude des variables de type taille.

▶ **Exemple**

L'exemple suivant illustre une modélisation par une loi de Pareto du chiffre d'affaire des 25 premiers groupes français de l'industrie et des services en 2001. Les données sont extraites des *Tableaux de l'Économie Française 2003-2004*.

G2 fx =16774*PUISSANCE(E2;-1/1,646)

	A	B	C	D	E	F	G
1	**Société**	*CAHT* (millions d'€)	ln(*CA*)	F_i	1-F_i	ln(1-F_i)	x_i*
2	TotalFinaElf	105 318	11,565	0,960	0,040	-3,219	118 558
3	Carrefour	69 486	11,149	0,920	0,080	-2,526	77 812
4	Vivendi Universal	57 360	10,957	0,880	0,120	-2,120	60 823
5	PSA Peugeot Citroën	51 663	10,852	0,840	0,160	-1,833	51 069
6	France Telecom	43 026	10,670	0,800	0,200	-1,609	44 595
7	Suez	42 359	10,654	0,760	0,240	-1,427	39 919
8	EDF	40 716	10,614	0,720	0,280	-1,273	36 350
9	Les Mousquetaires	37 200	10,524	0,680	0,320	-1,139	33 518
10	Renault	36 351	10,501	0,640	0,360	-1,022	31 203
11	Saint-Gobain	30 390	10,322	0,600	0,400	-0,916	29 268
12	Pinault-Printemps-La Redoute	27 799	10,233	0,560	0,440	-0,821	27 622
13	Groupe Auchan	26 200	10,174	0,520	0,480	-0,734	26 200
14	Alcatel Alsthom	25 353	10,141	0,480	0,520	-0,654	24 956
15	Galec (Leclerc)	25 000	10,127	0,440	0,560	-0,580	23 857
16	Alstom	23 453	10,063	0,400	0,600	-0,511	22 878
17	Aventis	22 941	10,041	0,360	0,640	-0,446	21 998
18	Groupe Casino (Rallye)	21 984	9,998	0,320	0,680	-0,386	21 203
19	Bouygues	20 473	9,927	0,280	0,720	-0,329	20 479
20	Airbus (EADS)	20 427	9,925	0,240	0,760	-0,274	19 817
21	SNCF	20 129	9,910	0,200	0,800	-0,223	19 209
22	Vonci	17 172	9,751	0,160	0,840	-0,174	18 648
23	La poste	17 028	9,743	0,120	0,880	-0,128	18 129
24	Publicis Groupe	16 667	9,721	0,080	0,920	-0,083	17 646
25	Michelin	15 775	9,666	0,040	0,960	-0,041	17 195
26	Havas	14 950	9,612	0,000	1,000	0,000	16 774
27	Moyenne	33 169					
28	Médiane	25 353					

Tableau 2.1 –Tableau 7.1

Les points $\{\ln(x_i)\ ;\ \ln(1 - F_i)\}$ sont à peu près alignés (*cf.* figure 7.23). Le calcul de la droite des m oindres carrés donne pour estimation des paramètres du modèle de Pareto $\alpha = 1{,}646$ et $x_0 = 16\,774$.

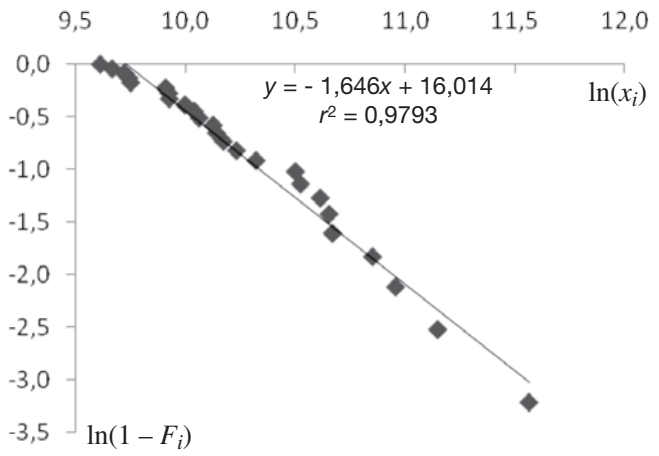


Figure 7.23 – Série des valeurs des 25 premiers chiffres d'affaires français ajustée par une loi de Pareto

Un diagramme quantile-quantile obtenu à partir de la loi de Pareto associée à ces paramètres montre lui aussi que le modèle n'est pas inadapté, mais que l'ajustement n'est pas parfait en raison la première valeur particulièrement élevée.

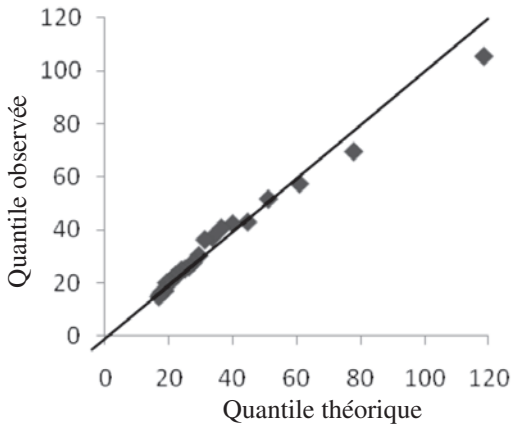


Figure 7.24 – Diagramme Quantile-Quantile (unité : milliards d'euros) Loi de Pareto de paramètres $\alpha = 1{,}646$ et $x_0 = 16\,774$

Pour terminer, on notera que la moyenne d'une loi de Pareto de paramètres α et x_0 est égale à $\alpha \cdot x_0/(\alpha - 1)$, et que sa médiane est égale à $x_0 \cdot 2^{1/\alpha}$

Pour l'exemple, on peut comparer la moyenne observée égale à 33 169 millions d'euros à la moyenne du modèle théorique égale à 42 740 millions d'euros, et la médiane observée égale à 25 353 millions d'euros à la médiane du modèle théorique égale à 25 558 millions d'euros.

Les paramètres du modèle de Pareto ont été évalués à l'aide de toutes les observations avec une première valeur particulièrement élevée. Ceci explique la surévaluation non négligeable de la moyenne par le modèle.

D'autre part, on notera que les médianes (théorique et observée), plus petites que les moyennes correspondantes, indiquent une distribution dissymétrique, étalée vers la droite (chapitre 1).

C. La loi de Weibull

Définition

Une v.a. continue X suit une *loi de Weibull* de paramètres a, $b > 0$ et $c > 0$, si sa densité est donnée par :

$$f(x) = \begin{cases} \dfrac{c}{b}\left(\dfrac{x-a}{b}\right)^{c-1} e^{-\left(\frac{x-a}{b}\right)^c} & \text{si} \quad x > a \\ 0 & \text{si} \quad x \leq a \end{cases}$$

Si X suit une loi de Weibull de paramètres a, b et c, alors $Y = \dfrac{X-a}{b}$ suit une loi de Weibull de paramètres 0, 1 et c. En effet :

$$P(Y < y) = P\left(\frac{X-a}{b} < y\right) = P(X < a + by) = \int_a^{a+by} \frac{c}{b}\left(\frac{x-a}{b}\right)^{c-1} e^{-\left(\frac{x-a}{b}\right)^c} dx$$

et la densité de Y est égale à la dérivée de cette dernière expression. Pour $a + by > a$, soit si $y > 0$:

$$f(y) = b\frac{c}{b}\left(\frac{a+by-a}{b}\right)^{c-1} e^{-\left(\frac{a+by-a}{b}\right)^c} = cy^{(c-1)}e^{-y^c} \quad \text{et}$$

$$f(y) = 0 \quad \text{si} \quad (y < 0)$$

On appelle *loi de Weibull standard* de paramètre c, notée $W(c)$, la loi de $Y = \dfrac{X-a}{b}$ lorsque X suit une loi de Weibull de paramètres a, b et c. On

remarque que pour $c = 1$, la loi de Weibull standard correspond à la loi exponentielle. La figure 7.25 représente les densités de la loi de Weibull standard pour les valeurs $c = 1, 2, 3$ et 5. On voit que cette densité est asymétrique, et présente un maximum si $c > 1$ pour :

$$x = \left(\frac{c-1}{c}\right)^{1/c}$$

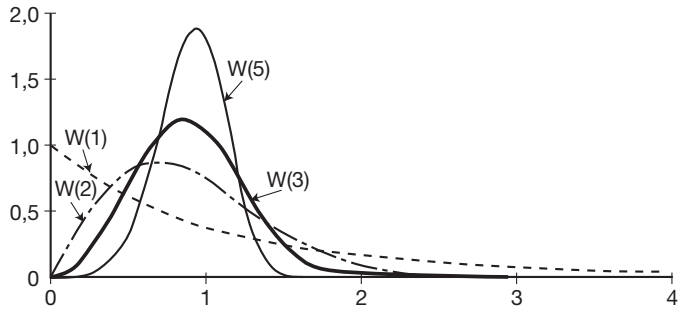


Figure 7.25 – Densités de la loi de Weibull standard

Lorsque $c > 1$, le mode de la distribution de Weibull standard se rapproche de 1 lorsque c tend vers l'infini. Dans le cas général de la loi de Weibull de paramètres a, b et c, le mode converge rapidement vers $(a + b)$ lorsque c augmente.

Par contre si $0 < c < 1$, le mode est en 0, car la densité est décroissante.

La fonction de répartition de la loi de Weibull standard est donnée par :

$$F(x) = \int_0^x c t^{c-1} e^{-t^c} dt$$

soit en posant $u = t^c$: $F(x) = \int_0^{x^c} e^{-u} du = [-e^{-u}]_0^{x^c} = 1 - e^{-x^c}$ si $x > 0$

et naturellement $F(x) = 0$ si $x \leq 0$

Dans le cas général, posant $X = a + bY$, où Y suit une loi de Weibull standard, on a :

$$P(X < x) = P(a + bY < x) = P\left(Y < \frac{x-a}{b}\right) = 1 - e^{-\left(\frac{x-a}{b}\right)^c} \quad \text{si } x > a$$

$$\text{et } P(X < x) = 0 \quad \text{si } x \leq a$$

Ceci nous montre que la médiane est le point $w_{0,5}$ tel que :

$$w_{0,5} = b \cdot (\ln 2)^{1/c} + a$$

Le calcul de la moyenne de la loi de Weibull standard donne :

$$E(Y) = \int_0^\infty c x^c e^{-x^c} dx = \int_0^\infty t^{1/c} e^{-t} dt = \Gamma\left(1 + \frac{1}{c}\right)$$

obtenu en posant $t = x^c$ dans la première intégrale, et en exprimant la seconde intégrale à l'aide de la fonction gamma (§ III.A).

De même, on peut calculer $E(Y^2)$:

$$E(Y^2) = \int_0^\infty c x^{c+1} e^{-x^c} dx = \int_0^\infty t^{2/c} e^{-t} dt = \Gamma\left(1 + \frac{2}{c}\right)$$

ce qui permet donc d'écrire la variance :

$$\text{var}(Y) = \Gamma\left(1 + \frac{2}{c}\right) - \Gamma\left(1 + \frac{1}{c}\right)^2$$

L'expression des moments dans le cas général de la loi de Weibull de paramètres a, b et c provient de la relation $X = a + bY$:

$$E(X) = a + b \cdot \Gamma\left(1 + \frac{1}{c}\right)$$

$$\text{var}(X) = b^2 \left\{ \Gamma\left(1 + \frac{2}{c}\right) - \left(\Gamma\left(1 + \frac{1}{c}\right)\right) \right\}^2$$

La dissymétrie de la loi standard de Weibull, observée sur la figure 7.26, varie avec la valeur du paramètre c. La moyenne tend vers 1 au fur et à mesure que c augmente, tandis que la variance décroît. Les coefficients d'asymétrie et d'aplatissement de Fisher montrent que pour c à peu près égal à 3,6 on obtient une courbe presque symétrique ($\gamma_1 \approx 0$), mais dont l'aplatissement est légèrement moindre que celui de la loi de Gauss ($\gamma_2 < 0$).

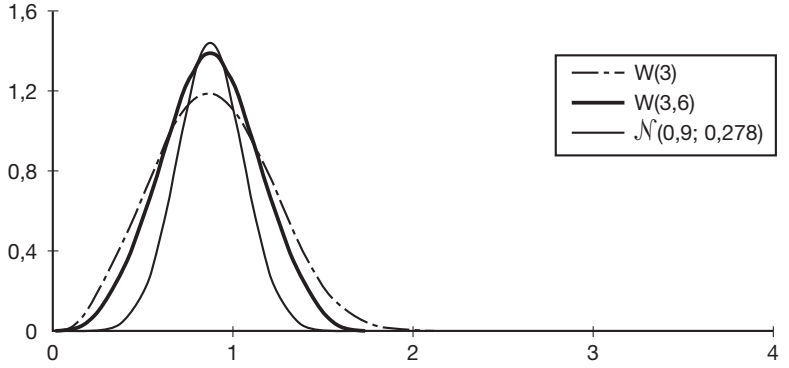


Figure 7.26 – Lois de Weibull standard et loi de Gauss

Il s'ensuit (*cf.* figure 7.26) que pour des valeurs de c proches de 3,6, la distribution standard de Weibull et la loi de Gauss ont des formes assez proches. On notera que pour cette valeur $c = 3{,}6$, la moyenne de la loi standard de Weibull est égale à 0,9011 et son écart-type à 0,278

Il est intéressant de noter que pour une loi de Weibull standard de paramètre c, l'expression de la fonction de répartition permet d'écrire :

$$\ln(1 - F_W(y)) = -y^c$$

soit encore :

$$\ln\{-\ln(1 - F_W(y))\} = c\ln(y)$$

ce qui montre qu'on peut, comme pour la loi de Pareto, mais avec une échelle « log-log » sur l'axe vertical, évaluer graphiquement si une distribution observée peut être représentée par une loi de Weibull standard.

L'intérêt particulier de la loi de Weibull réside dans la propriété suivante.

Propriété

Si Y suit une loi de Weibull standard $W(c)$, alors Y^c suit une loi exponentielle de paramètre 1

En effet, soit $Z = Y^c$. Si $z < 0$, on aura $P(Z < z) = 0$ et si $z > 0$, on peut écrire :

$$P(Z < z) = P(Y^c < z) = P(Y < z^{1/c}) = \int_0^{z^{1/c}} c t^{c-1} e^{-t^c} dt$$

et la densité, nulle pour $z < 0$, s'obtenant par dérivation pour $z > 0$, est bien égale à e^{-z}

La loi de Weibull [1] doit ainsi s'envisager comme une généralisation du modèle exponentiel, tout particulièrement dans les contextes où l'étude porte sur le temps écoulé jusqu'à la réalisation d'un certain événement.

D. La loi logistique

Définition

Une v.a. X continue suit une *loi logistique* si sa fonction densité est donnée par :

$$f(x) = \frac{\exp\left(-\frac{x-\alpha}{\beta}\right)}{\beta\left(1 + \exp\left(-\frac{x-\alpha}{\beta}\right)\right)^2} \text{ avec } \beta > 0$$

1. Nommée ainsi en raison des travaux du statisticien suédois Waloddi Weibull qui en a montré l'intérêt (entre 1939 et 1954) pour une très large gamme d'applications, industrielles notamment.

Un calcul très simple montre alors que la fonction de répartition est donnée par :

$$F(x) = \left(1 + \exp\left(-\frac{x-\alpha}{\beta}\right)\right)^{-1}$$

Si on pose $Y = \frac{X-\alpha}{\beta}$, on obtient la forme réduite de la distribution logistique, définie par la densité $\frac{e^{-y}}{(1+e^{-y})^2}$, et pour laquelle la fonction de répartition est $\frac{1}{1+e^{-y}}$. La densité de Y est symétrique puisque :

$$\frac{e^{-y}}{(1+e^{-y})^2} = \frac{\left(1/e^{y}\right)}{\left(1+1/e^{y}\right)^2} = \frac{\left(1/e^{y}\right)}{\left((e^{y}+1)/e^{y}\right)^2} = \frac{e^{-y}(e^{y})^2}{(e^{y}+1)^2} = \frac{e^{y}}{(1+e^{y})^2}$$

Ceci conduit à constater que la variable Y est centrée, $E(Y) = 0$, et que $E(X) = \alpha$. Le paramètre α de la définition de la distribution logistique est donc égal à sa moyenne. Le calcul de la variance donne le résultat [1] :

$$\text{var}(Y) = \frac{\pi^2}{3}, \text{ ce qui donne : } \text{var}(X) = \frac{\beta^2\pi^2}{3}$$

La courbe représentative de la densité est encore une courbe en cloche, en raison d'une part, de la symétrie et d'autre part, de l'existence d'une asymptote horizontale, d'un maximum et de deux points d'inflexion. La comparaison de cette densité avec celle de la loi normale est justifiée si l'on choisit des paramètres qui assurent l'égalité des moyennes et l'égalité des variances. Compte tenu de ce qui précède, il faut choisir la loi logistique de paramètres :

$$\alpha = 0 \quad \text{et} \quad \beta = \frac{\sqrt{3}}{\pi} \approx 0{,}5513$$

On observe alors (*cf.* figure 7.27) que les deux densités sont assez proches. En raison de la symétrie, le coefficient d'asymétrie γ_1 est nul, et la comparaison entre le coefficient d'aplatissement γ_2, égal à 0 pour la loi de Gauss centrée réduite et à 1,2 pour la loi logistique, traduit bien la limite de leur ressemblance.

1. Le calcul passe par le développement en série de $\frac{1}{1+e^{-y}}$ et l'utilisation des fonctions eulériennes.

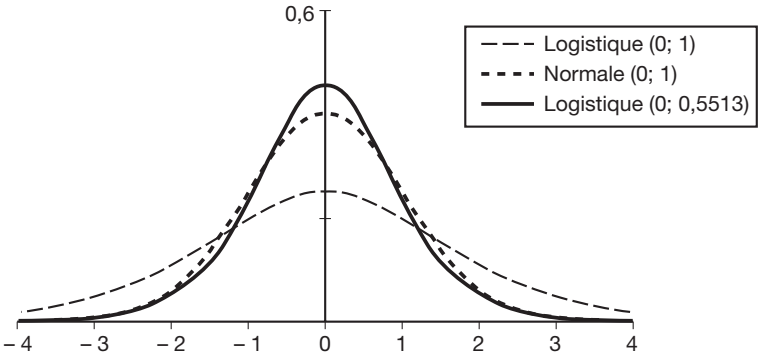


Figure 7.27 – Densités comparées de lois logistiques et de la loi normale centrée réduite

Deux relations concernant la loi logistique standard ($\alpha = 0$ et $\beta = 1$) sont particulièrement utiles :

$$f(y) = F(y) \cdot (1 - F(y))$$

$$y = \ln\left\{\frac{F(y)}{1 - F(y)}\right\}$$

Elles font ressortir l'intérêt de la loi logistique dans les situations binaires. Cette loi est particulièrement utile pour modéliser des systèmes où la réponse (aléatoire) à une intervention est du type « tout ou rien » (ou encore positive ou négative). Pour ces situations, on suppose souvent que la proportion de réponses positives suit une loi logistique et on parle alors d'analyse de type « logit ». On choisit de préférence la loi logistique dans ces problèmes, mais certains modélisent par la loi de Gauss et on parle alors d'analyse de type « probit ». Mais l'interprétation « mécaniste » n'est alors pas aussi fine que celle découlant de la loi logistique.

La fonction définissant la fonction de répartition de la loi logistique standard est utilisée intensivement pour représenter les phénomènes de croissance (entre deux limites) avec le temps. Dans ce contexte, elle est obtenue en partant d'une équation différentielle traduisant la proportionnalité en tout point entre d'une part le taux de croissance, et d'autre part le produit des différences avec les valeurs extrêmes, soit :

$$\frac{dF}{dx} = c(F(x) - l)(L - F(x))$$

V. Bilan

Les modèles présentés dans ce chapitre et dans le précédent sont ceux qui sont le plus souvent utilisés dans l'étude de situations aléatoires concrètes. On n'oubliera pas cependant que d'autres modèles présentent un intérêt certain pour la modélisation. Ils ne peuvent être présentés ici, mais on se doit de citer :

– la loi binomiale négative ;

– la loi log-normale, la loi de Wald, les lois gamma et bêta ;

– les lois de khi-deux, de Student et de Fisher-Snedecor décentrées.

D'autre part, nous n'avons envisagé que les lois de variables aléatoires à valeurs réelles. Les distributions de v.a. à valeurs dans $\mathbb{R}^n$ n'ont pas été abordées. Leur utilisation est cependant fréquente.

L'ensemble de ce chapitre montre l'intérêt particulier de la loi de Gauss. Approximation de lois discrètes ou de lois continues, mais aussi loi approchée des moyennes d'échantillonnage, la distribution gaussienne est au centre de l'édifice statistique dès qu'on cherche à modéliser des situations aléatoires.

Ses propriétés (symétrie, stabilité après transformation linéaire) et son utilisation particulièrement simple pour les calculs expliquent le recours très fréquent à cette loi. Cependant, on ne doit pas se laisser abuser, et nous avons vu que plusieurs problèmes spécifiques conduisent à d'autres lois, soit par construction (cas des lois du khi-deux, de Student, de Fisher-Snedecor), soit par ajustement (lois de Pareto, exponentielle, logistique, de Weibull).

Le sens de l'hypothèse de normalité émerge de cette position centrale de la loi de Gauss, comme aussi de son apparence. La loi de Gauss est celle qui s'impose lorsque le phénomène qu'on étudie ne présente pas de déterminant prédominant. En ce sens, la distribution normale correspond à une répartition sans caractéristique ou individualisation particulière. Elle serait donc presque le modèle à utiliser si aucun autre ne s'imposait. D'ailleurs, c'est historiquement ainsi qu'elle a émergé pour représenter de façon « universelle » les phénomènes au sens de la moyenne[1] (c'est-à-dire observés comme moyennes arithmétiques).

On se gardera bien de vouloir à tout prix poser une hypothèse de normalité dans une attitude descriptive. Réservant la loi de Gauss pour des phénomènes agrégeant réellement de multiples causes indépendantes les unes des autres sans cause prédominante, on recherchera toujours le modèle (moins passe-partout, mais donc plus « savoureux » et surtout plus riche) décrivant vraiment au mieux les observations. La controverse entre univers gaussien et univers parétien en est une illustration.

1. D'abord par Laplace à la fin du 18e siècle, puis par Gauss en 1809, et enfin par Galton en 1889.

On n'oubliera pas :

1. Pour la loi de Gauss

• La symétrie par rapport à la moyenne.

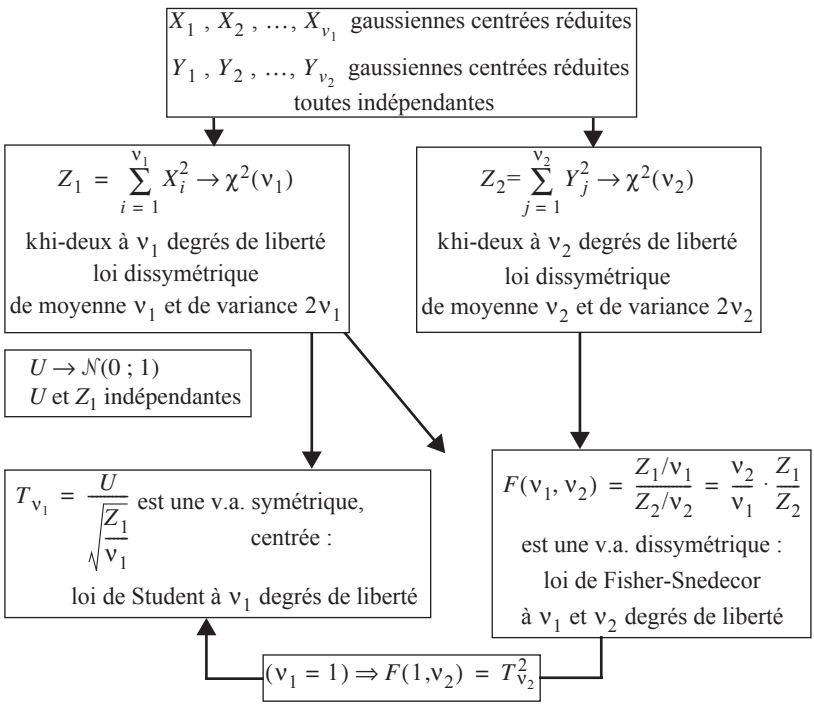
• La pratique des calculs par centrage et réduction

$$X \to \mathcal{N}(m\ ;\ \sigma) \Leftrightarrow U = \frac{X - m}{\sigma} \to \mathcal{N}(0\ ;1)$$

• Son intérêt comme modèle approché (limite centrale, lois binomiale et de Poisson).

• Une somme de gaussiennes indépendantes est une gaussienne.

2. Pour les lois dérivées de la loi de Gauss



3. La loi de Pareto de paramètres x_0 et α peut décrire des grandeurs au moins égales à x_0 et pour lesquelles le nombre d'observations qui dépassent un seuil x est proportionnel à $x^{-\alpha}$. Son moment d'ordre k n'existe que pour $k < \alpha$. C'est une loi dissymétrique.

4. La loi de Weibull standard de paramètre c est une loi dissymétrique souvent utilisée pour décrire le temps écoulé jusqu'à la réalisation d'un événement donné (décès, faillite, panne, reprise d'activité…). Elle généralise la loi exponentielle.

Testez-vous (*les réponses sont données page 287*)

1. Une variable aléatoire de Bernoulli :

a) a une loi de probabilité ne dépendant que d'un seul paramètre

b) a une espérance égale à son écart-type

c) a une variance maximum lorsque son paramètre est égal à 0,5

d) est une variable indicatrice

2. Si X suit une loi binomiale $\mathcal{B}(n ; p)$:

a) X est la somme de n v.a. de Bernoulli de même espérance mathématique

b) X est la somme de n v.a. de Bernoulli indépendantes

c) l'écart-type de X est égal à $np(1 - p)$

d) $Y = X + 1$ suit une loi binomiale $\mathcal{B}(n ; p)$

3. Soit X une variable aléatoire binomiale $\mathcal{B}(n ; p)$ et $Y = n - X$:

a) pour toute valeur entière k comprise entre 0 et n, $P(X = k) = P(Y = n - k)$

b) Y suit une loi binomiale $\mathcal{B}(n ; 1 - p)$

c) X et Y ont même espérance mathématique

d) X et Y ont même variance

4. Si X suit une loi de Poisson de paramètre 10 :

a) $\text{var}(X) = 10$

b) $P(X = 10) = P(X = 9)$

c) $Y = 2X + 1$ suit une loi de Poisson de paramètre 21

d) la fonction de répartition de X est une fonction continue

5. Si X suit une loi géométrique de paramètre p :

a) on peut aussi dire que X suit une loi de Pascal

b) X est une somme de v.a. de Bernoulli indépendantes de paramètre p

c) l'espérance de X dépend du nombre de tirages

d) $\text{E}(X) > \text{var}(X)$

6. Si X est distribuée selon une loi $\mathcal{B}(n ; p)$:

a) si $n = 10$ et $p = 0,1$, alors $P(X = 4) = P(X = 5)$

b) si $n = 60$ et $p = 0,05$, alors $P(X \geq 3) = 0,5768$

c) si $n = 4$ et $p = 0,01$, alors $P(X = 0) \approx 0,96$

d) si $n = 50$ et $p = 0,08$, alors $P(3 < X \leq 10) = 0,573$

7. Pour une population distribuée selon une loi de Gauss $\mathcal{N}(1 ; 1)$:

a) la médiane est égale à 1

b) la moyenne est égale à 0

c) le quart des individus est caractérisé par une valeur supérieure à 1

d) la moitié des individus est caractérisée par une valeur inférieure à 0

8. Pour une population distribuée selon une loi $\mathcal{N}(0 ; 1)$, le pourcentage d'individus qui sont caractérisés par une valeur comprise entre – 1,96 et + 1,96 est égal à :

a) 75%

b) 90%

c) 95%

d) 99%

9. Si X suit une loi de Gauss $\mathcal{N}(2 ; 5)$:

a) $P(X = 2) = F_U(0) = 1/3\sqrt{2\pi}$

b) $P(X < 1) = P(X > + 3) = 0{,}4207$

c) $P(– 3 < X < + 3) = 0{,}4206$

d) $P(1 < X < 3) = 0{,}1586$

10. Si X suit une loi de Gauss $\mathcal{N}(m ; \sigma)$:

a) $P(X > m) = 0{,}5$

b) sa moyenne est égale à sa médiane

c) $F_x(m + x) = F_x(m – x)$

d) le graphe de la densité est symétrique par rapport à la droite d'équation $x = 0$

11. Si X est une variable aléatoire $\mathcal{N}(1 ; 1)$ et Y une variable aléatoire $\mathcal{B}(1000 ; 0{,}01)$ et si X et Y sont indépendantes :

a) $E(X + Y) = 11$

b) $\text{var}(X + Y) = 10{,}9$

c) Y suit approximativement une loi de Poisson de paramètre 10

d) le coefficient de corrélation linéaire entre X et Y est égal à + 1

12. Si une population est représentée par une variable aléatoire X de loi $\mathcal{N}(m ; \sigma)$:

a) 25% des individus s'écartent de la moyenne de plus d'un écart-type

b) 50 % des individus sont au-dessus de la moyenne

c) à peu près 5 % des individus s'écartent de la moyenne de plus de 2 fois l'écart-type

d) $F_x(m + x) = 1 – F_x(m – x)$

13. Si X est une variable aléatoire $\mathcal{N}(1 ; 1)$ et Y une variable aléatoire $\mathcal{B}(100 ; 0{,}5)$ et si X et Y sont indépendantes :

a) $E(X + Y) = 51$

b) X suit approximativement une loi de Poisson de paramètre 50

c) $X + Y$ suit approximativement une loi de Gauss

d) $Z = 2X + 3$ suit une loi de Gauss

14. Soit X une variable aléatoire $\mathcal{N}(– 3 ; 1)$:

a) X est une variable aléatoire réduite

b) $X + 3$ est une variable aléatoire centrée réduite

c) $P(X + 3 > 0) = P(X + 3 < 0)$

d) $E(X^2) = 1$

15. Si X suit une loi exponentielle de paramètres $\theta = 0$ et $\lambda = 2$

a) la fonction de répartition de X est une fonction continue

b) l'espérance de X est égale à sa variance

c) la fonction de densité de X est une fonction qui admet un axe de symétrie

d) X ne prend que des valeurs supérieures à 2

16. Soit Y une somme de n variables aléatoires indépendantes $X_1, \ldots X_2, \ldots X_n$:

a) si les X_i sont des v.a. binomiales, alors Y suit une loi binomiale

b) si les X_i suivent des lois de Gauss, alors Y suit une loi de Gauss

c) si les X_i suivent des lois exponentielles, alors Y suit une loi exponentielle

d) si les X_i suivent des lois de Pareto, alors Y suit une loi de Pareto

17. Si la demande hebdomadaire d'un produit dans un magasin suit une loi binomiale $\mathcal{B}(30 ; 0,45)$, alors si on suppose les demandes hebdomadaires indépendantes entre elles, la demande annuelle de ce même produit (1 an = 52 semaines) :

a) suit une loi binomiale $\mathcal{B}(1560 ; 0,45)$

b) suit approximativement une loi de Gauss $\mathcal{N}(702 ; 19,65)$

c) est une somme de v.a. de Bernoulli indépendantes

d) ne peut pas être égale à 1 600

18. Si la demande quotidienne d'un produit dans un magasin suit une loi binomiale $\mathcal{B}(40 ; 0,05)$, alors si on suppose les demandes quotidiennes indépendantes entre elles, la demande de ce même produit pour 25 jours de fonctionnement de ce magasin suit :

a) à peu près une loi de Poisson $\mathcal{P}(50)$

b) à peu près une loi normale $\mathcal{N}(50 ; 6,9)$

c) une loi binomiale $\mathcal{B}(40 ; 0,2)$

d) une loi de Poisson $\mathcal{P}(0,2)$

Exercices *(corrigés page 324)*

Exercice 7.1

Le prix X d'un certain article est supposé distribué selon une loi de Gauss de paramètres $m = 45$ € et $\sigma = 4$ €

1. Calculez $P(X < 39)$, $P(X \geq 48)$ et $P(35 < X < 48)$

2. Calculez $P(|X - m|) \leq \sigma$

3. Calculez $P(41 \leq X \geq 49 | X \geq 39)$

Exercice 7.2

Les gains mensuels en euros d'un représentant sont supposés suivre une loi normale. Il a pu constater, sur un grand nombre de mois, la répartition suivante de ses gains :

Gain > 3 000 : 4,46 %

2 400 < Gain ≤ 3 000 : 93,26%

Gain ≤ 2 400 : 2,28%

1. Calculez la moyenne et l'écart-type de la loi normale envisagée.

2. Si on suppose les gains du représentant indépendants d'un mois à l'autre, quelle est la loi de probabilité de la variable aléatoire égale au gain du représentant pendant 3 mois ?

3. Quelle est la probabilité que le représentant gagne plus de 8 700 € en 3 mois ?

Exercice 7.3

Une usine fabrique des imprimantes laser dont la durée de vie X (exprimée en millions de pages) est une variable aléatoire normale $\mathcal{N}(2 ; 0,3)$.

1. Calculez la probabilité p que la durée de vie d'une imprimante tirée au hasard dans la production soit supérieure à 2,5 millions de pages. Dans la suite de l'exercice, on arrondira cette probabilité p pour ne conserver que 2 chiffres après la virgule.

2. On teste 60 imprimantes tirées au hasard dans la production. Déterminez, en la justifiant, la loi de la variable Y égale au nombre d'imprimantes dont la durée de vie est supérieure à 2,5 millions de pages. Donnez, en la justifiant, une loi approchée de Y.

3. Calculez la probabilité que parmi les 60 imprimantes testées :

– exactement 10 % des imprimantes aient une durée de vie supérieure à 2,5 millions de pages,

– au moins 5 % des imprimantes aient une durée de vie supérieure à 2,5 millions de pages.

Exercice 7.4

Un groupe de presse décide de lancer un nouveau quotidien. Une enquête permet de conclure que, pendant les 30 jours consécutifs à la date du lancement, la demande journalière (exprimée en milliers d'exemplaires) est une v.a. X dont la loi de probabilité peut être considérée comme une loi normale de paramètres $m = 100$ et $\sigma = 25$.

1. Calculez la probabilité d'avoir pendant la période considérée :

– une demande journalière comprise entre 75 et 125 ;

– une demande journalière de plus de 150.

2. Calculez les valeurs n_1 et n_2 telles que :

– la demande journalière soit supérieure à n_1, 9 jours sur 10 ;

– la demande journalière soit inférieure à n_2, 4 jours sur 10.

3. Déterminez un intervalle symétrique autour de m ayant 90 % de chances de contenir la demande journalière.

Exercice 7.5

Un vigneron commercialise des vins de qualité différentes qu'il répartit en deux classes : la classe des vins courants dits « du terroir », la classe des vins de qualité, appelés « grand cru », et vendus 6 € la bouteille. Malgré le soin apporté à l'embouteillage, il subsiste des erreurs d'étiquetage, et on admet qu'un acheteur de vin « grand cru » aura une probabilité $p = 0{,}12$ d'avoir en fait une bouteille de vin ordinaire.

1. Un restaurateur achète 200 bouteilles « grand cru » au vigneron. Soit Y la v.a. égale au nombre de bouteilles de vin courant parmi les 200 bouteilles achetées.

Déterminez la loi de probabilité de la v.a. Y, ainsi que son espérance et sa variance. Donnez, en la justifiant, une approximation de la loi de Y.

2. Calculez : $P(Y > 20)$ et $P(Y < 30 \mid Y > 20)$.

3. Au fur et à mesure de la consommation des 200bouteilles, le restaurateur a pu détecter chacune des bouteilles de type courant. Il décide alors de ne payer que les bouteilles de qualité effectivement livrées et de refuser tout paiement pour les bouteilles de vin ordinaire.

Calculez, dans cette hypothèse, la probabilité d'un bénéfice néanmoins positif pour le vigneron sachant que chaque bouteille de vin courant lui revient à 1,5 € et que chaque bouteille de vin de qualité lui revient à 3,5 €.

Exercice 7.6

Lors la naissance de jumeaux, on note λ la probabilité qu'il s'agisse de vrais jumeaux et on fait les deux hypothèses suivantes :

i) deux vrais jumeaux sont toujours de même sexe, et la probabilité qu'ils soient des garçons est égale à 1/2 ;

ii) deux faux jumeaux ont des sexes indépendants et chacun des deux enfants est un garçon avec une probabilité égale à 1/2.

Soit A, B et C les événements suivants relatifs à la naissance de deux jumeaux :

A = { 2 garçons }

B = { 2 filles }

C = { 1 garçon et une fille }

1. Calculez en fonction de λ les probabilités des événements A, B et C.

2. Soit Y la variable aléatoire égale au nombre de fois où on a eu un garçon et une fille sur 1 000 naissances de jumeaux.

Donnez en fonction de λ, et en la justifiant, la loi de probabilité de la variable aléatoire Y.

Donnez l'espérance et la variance de Y en fonction de λ.

3. On suppose $\lambda = 0{,}35$; par quelle loi peut-on approximer la loi de Y ? (justifiez votre réponse).

Déterminez les probabilités des événements $\{Y > 300\}$ et $\{310 \leq Y \leq 350 \mid Y > 300\}$.

Exercice 7.7

Après avoir fait remplir un long questionnaire portant sur l'audience de la presse magazine à 200 individus, un institut de sondage a établi la distribution suivante pour la durée d'interview (en minutes) concernant ces 200 individus :

Durée (min)	< 25	[25 ; 30[	[30 ; 35[	[35 ; 40[	[40 ; 45[	[45 ; 50[	≥ 50
Effectif	18	32	36	40	30	24	20

1. Calculez la médiane de cette distribution.

2. On ajuste cette distribution par une loi normale $\mathcal{N}(37 ; 10)$. Représentez le diagramme quantile-quantile. Quel jugement permet-il de porter sur la qualité de l'adéquation de la distribution observée à ce modèle théorique ?

3. On suppose pour la suite de ce problème que la durée X d'une interview suit une loi normale $\mathcal{N}(37 ; 10)$.

3.1. Soient $X_1, X_2, \ldots, X_n$ les variables aléatoires associées aux durées de n interviews, on suppose les v.a. X_i indépendantes et identiquement distribuées à X.

Que représente la variable aléatoire $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$? Donnez, en la justifiant, sa loi en fonction de n.

3.2. Calculez la probabilité que la durée moyenne de 6 interviews n'excède pas 35 minutes.

3.3. Pour quelles valeurs de n, la durée moyenne de n interviews n'excède pas 45 minutes avec une probabilité au moins égale à 0,99 ?

Exercice 7.8

Durant une année, on a observé que 70 % des actions enregistrées sur le marché boursier de Londres ont augmenté de valeur, tandis que 30 % sont restées stationnaires ou ont baissé.

1. Au début de l'année, un service de conseils financiers a choisi 10 actions pour les placer dans la rubrique des actions « spécialement recommandées ».

1.1. Pour un non-initié qui considérait ces 10 actions choisies aléatoirement (au hasard et avec remise), donnez, en la justifiant, la loi de probabilité du nombre d'actions à la hausse.

1.2. Calculez les probabilités des événements suivants :

– accroissement de valeur pour au moins 8 actions ;
– accroissement de valeur pour moins de 4 actions.

2. On considère maintenant que le service de conseils financiers a choisi 100 actions. Toujours pour un non-initié qui considérait ces 100 actions choisies aléatoirement (au hasard et avec remise), calculez, en justifiant le recours à une loi de probabilité approchée, les probabilités des événements suivants :

– accroissement de valeur pour au moins 80 % des actions ;
– accroissement de valeur pour moins de 40 % des actions.

D'après examen de juin 2002, GEA 1 re année Paris IX-Dauphine

Exercice 7.9

Un appareil électronique est soumis à des impulsions séparées par des intervalles de temps variables, indépendants les uns des autres. On suppose que la durée Y (exprimée en secondes) séparant deux impulsions successives est une v.a. définie ainsi :

$$Y = 2 + \lambda X$$

où λ est un paramètre strictement positif et X une v.a. exponentielle de paramètre 1.

1. Déterminez en fonction de λ :

– $E(Y)$ et $\text{var}(Y)$;

– la fonction de répartition de la variable aléatoire Y.

2. On suppose dans cette question le paramètre λ égal à 5.

2.1. Calculez $P(Y < 2)$ et $P(2 \leq Y \leq 5)$.

2.2. Afin d'étudier si la loi théorique retenue pour Y représente bien le phénomène étudié, on a mesuré de façon indépendante 10 durées séparant 11 impulsions successives et on a obtenu les résultats suivants en secondes :

2,3 3,5 3,9 4,7 5,1 6,8 7,9 9,6 13,1 15,5

Tracez le diagramme quantile-quantile pour juger la qualité de l'ajustement de cette distribution observée par la loi retenue. Conclusion.

Exercice 7.10

Afin de mieux connaître sa clientèle, le gérant du cinéma Paradiso fait procéder à un sondage. Il obtient pour un échantillon aléatoire de taille 100 la répartition par âge suivante :

Âge	< 15 ans	[15 ; 20[	[20 ; 25[	[25 ; 30[	[30 ; 35[	[35 ; 40[	[40 ; 50[	≥ 50 ans
Effectif	4	13	22	28	15	10	5	3

1. Calculez la moyenne et l'écart-type de cette distribution ; on supposera l'âge minimum égal à 10 ans et l'âge maximum égal à 70 ans.

2. Calculez la médiane de cette distribution.

3. On suppose que la distribution de l'âge des clients du cinéma Paradiso suit une loi normale de paramètres $m = 28$ ans et $\sigma = 9{,}5$ ans.

3.1. Quel est le pourcentage théorique des clients qui ont entre 18 et 35 ans ?

Calculez le pourcentage observé, à partir de l'échantillon de taille 100, pour la même classe d'âge.

3.2. Calculez l'âge théorique A tel que 75 % des clients ait un âge supérieur à A.

Calculez, à partir de l'échantillon, l'âge A' tel que 75 % des individus de l'échantillon ait un âge supérieur à A'. Comment s'appelle cette valeur A' ?

4. Tracez le diagramme quantile-quantile pour juger la qualité de l'ajustement de la distribution observée par une loi normale de paramètres $m = 28$ ans et $\sigma = 9{,}5$ ans.

L'hypothèse précédente vous semble-t-elle justifiée ?

Exercice 7.11

On considère que la durée du temps d'attente T (mesuré en minutes) du bus que doit prendre Valérie pour se rendre à l'Université, est distribuée selon une loi exponen-

tielle de moyenne 5 mn, c'est-à-dire que la variable aléatoire T admet la densité de probabilité suivante :

$$f_T(T) = \begin{cases} \frac{1}{5} \cdot e^{-t/5} & \text{si} \quad t \geq 0 \\ 0 & \text{sinon} \end{cases}$$

1. Quelle est la fonction de répartition de la variable aléatoire T ?

2. Quelle est la probabilité que le temps d'attente T dépasse 8 minutes ? Dans la suite de l'exercice, on arrondira cette probabilité à sa première décimale.

3. Valérie utilise le métro avec un seul ticket si elle attend le bus plus de 8 mn. Il faut deux tickets pour le bus, mais Valérie a une nette préférence pour le bus qu'elle utilise si le temps d'attente ne dépasse pas 8 mn. Soit Y, le nombre de trajets « allers » effectués en bus en n jours.

3.1. Donnez, en la justifiant, la loi de Y.

3.2. Calculez son espérance et sa variance.

4. Soit Z, la variable aléatoire égale au nombre de tickets utilisés par Valérie pour ses trajets « allers » en n jours.

4.1. Déterminez la loi de Z.

4.2. Calculez son espérance et sa variance.

D'après examen de septembre 2002, GEA 1re année Paris IX-Dauphine

Exercice 7.12

Le tableau ci-dessous donne les dix meilleurs résultats nets des grandes entreprises françaises en 2001 (Source : *Tableaux de l'Économie Française 2003-2004,* INSEE) :

Société	Résultats nets (millions d'€)
TotalFinaElf	$7\,658 = x_{10}$
Suez	$2\,087 = x_9$
PSA Peugeot Citroën	$1\,691 = x_8$
Sanofi-Synthélabo	$1\,585 = x_7$
Aventis	$1\,505 = x_6$
L'Oréal	$1\,291 = x_5$
Carrefour	$1\,266 = x_4$
Saint Gobain	$1\,134 = x_3$
Renault	$953 = x_2$
Gaz de France	$891 = x_1$

Soit F_i la proportion d'entreprises dont les résultats nets X sont inférieurs à x_i.

1. Peut-on considérer les 10 points de coordonnées $\{\{\ln(x_i), \ln(1 - F_i)\}, i = 1 \text{ à } 10\}$ approximativement alignés ? (On calculera le coefficient de corrélation linéaire et les coefficients de la droite des moindres des carrés).

Représentez graphiquement le nuage de ces 10 points, ainsi que la droite des moindres carrés.

2. En déduire que l'on peut ajuster la distribution de X par une loi de Pareto de paramètres α et x_0 qu'on évaluera à l'aide des résultats précédents.

Exercice 7.13 (suite de l'exercice 3.9)

On choisit le modèle quadratique puisque la part de variation de Y non expliquée par ce modèle est plus faible qu'avec le modèle linéaire.

On envisage un ajustement de la distribution des résidus du modèle quadratique par une loi de Gauss de paramètres $m = 0$ et $\sigma = 22$.

1. Calculez la série des 12 résidus de ce modèle, et rangez-les par ordre croissant.

2. Représentez le diagramme Quantile – Quantile.

Quel jugement permet-il de porter sur la qualité de l'adéquation de cette distribution par la loi de Gauss envisagée ?

D'après examen de juin 2006, DUGEAD 1re année Paris – Dauphine

Exercice 7.14

Une société de fabrication de boissons décide de lancer une nouvelle boisson à faible teneur en sucre. Les études effectuées montrent que la teneur X d'une bouteille d'un litre de cette boisson suit une loi normale de moyenne 70 g et d'écart-type 25 g.

1. Calculez la probabilité que la teneur en sucre d'une bouteille d'un litre diffère de la teneur moyenne d'au plus 10 g.

2. On choisit au hasard 25 bouteilles. Soient $X_1, X_2, \ldots, X_{25}$ les variables aléatoires associées. On les suppose indépendantes et identiquement distribuées à X.

Que représente la variable aléatoire $\overline{X} = \frac{1}{25}\sum_{i=1}^{25} X_i$? Donnez sa loi (justifier).

Calculez la probabilité que la teneur moyenne en sucre des 25 bouteilles diffère de la moyenne $m = 70$ g d'au plus 10 g.

3. La société décide de modifier légèrement à la fois la composition et le procédé de fabrication. La variable X suit maintenant une loi normale de moyenne m et d'écart-type σ avec m et σ inconnus. Les essais conduisent aux observations suivantes :

Teneur x_i	40	50	60	70	80	90	100
F_i	0,11	0,23	0,40	0,60	0,77	0,89	0,96

Déterminez les fractiles u_i définis par $F_U(u_i) = F_i$ où F_U désigne la fonction de répartition de la loi normale centrée réduite.
Représentez le nuage des points (u_i, x_i).

En déduire une évaluation de chacun des nouveaux paramètres m et σ en utilisant la méthode des moindres carrés.

Exercice 7.15

On donne la série statistique ordonnée des niveaux de vente d'une revue hebdomadaire dans 100 points de distribution pour une semaine donnée :

32	40	53	54	59	65	66	72	75	80
84	85	89	93	95	95	95	101	104	105
105	105	105	106	107	107	108	110	111	111
112	113	113	114	114	115	116	119	119	122
122	122	123	124	124	124	124	126	127	127
127	127	127	129	129	130	130	130	130	130
131	132	132	135	135	138	138	140	141	141
142	143	143	143	144	146	147	150	152	152
153	156	156	158	158	158	158	158	159	160
160	162	166	166	168	170	176	192	195	196

La moyenne de cette série d'observations est égale à 124,6 et l'écart-type à 32

1. Déterminez les trois quartiles de cette distribution. Proposez trois indicateurs de tendance centrale, deux indicateurs de dispersion, et donnez leurs valeurs.

2. On note X la variable aléatoire suivant une loi $\mathcal{N}(125 ; 30)$

2.1 Calculez la probabilité que la v.a. X appartienne à l'intervalle centré en sa moyenne et de longueur égale à :

a) 2 fois l'écart-type

b) 4 fois l'écart-type

2.2 Déterminez les déciles de la distribution de la v.a. X, c'est-à-dire les valeurs x_i $(1 \leq i \leq 9)$ telles que :

$$P(X < x_i) = i/10$$

3. Classez la distribution observée en 10 classes déterminées par les déciles x_i calculés précédemment (question 2.2).

4. Comparez à l'aide d'un diagramme les fréquences cumulées observées avec les fréquences cumulées théoriques (probabilités cumulées en pourcentage) pour apprécier la qualité de l'ajustement de la distribution observée par la loi $\mathcal{N}(125 ; 30)$. Conclusion.

D'après examen de septembre 2009, DEGEAD 1 re année Paris – Dauphine

Réponses aux questionnaires « Testez-vous »

Chapitre 1

1. Bonnes réponses : b) et c)

Trois quartiles séparent l'intervalle de variation [min(X), max(X)] en quatre intervalles contenant chacun le *quart* de l'effectif, donc 75 % des observations sont supérieures au 1er quartile.

2. Bonnes réponses : a), b) et d)

La boîte de distribution contient 50 % des observations et la médiane, égale à la moyenne, est le milieu de l'intervalle interquartile.

3. Bonnes réponses : c) et d)

On ne peut utiliser dans ce cas que des caractéristiques indépendantes des unités.

4. Bonnes réponses : a) et b)

La médiane de la série des écarts absolus à la moyenne est une mesure de la dispersion ; les quartiles Q_1 et Q_3 sont des indicateurs de position, mais non de tendance centrale.

5. Bonnes réponses : c) et d)

Pour calculer la note moyenne et l'écart-type des deux classes réunies, il faut utiliser des formules avec des pondérations (§ III.B.1 pour la propriété 4 de la moyenne et § III.C.4 pour la propriété 4 de la variance).
L'écart absolu moyen à la médiane est le plus petit écart absolu moyen (§ III.C.3), d'autre part, l'écart-type est toujours supérieur à l'écart absolu moyen à la moyenne arithmétique (§ III.C.4), donc, l'écart absolu moyen des notes à la médiane est au plus égal à 4 pour la classe 1.

6. Bonnes réponses : a) et d)

7. Bonne réponse : c)

Se référer à la moyenne géométrique (§ III.B.2).

8. Bonnes réponses : b) et c)

9. Bonnes réponses : a) et c)

Chapitre 2

1. Bonne réponse : b)

L'indice des prix actuellement calculé par l'INSEE est un indice-chaîne de Laspeyres.

2. Bonne réponse : c)

Un indice de Paasche est souvent un indice des quantités, mais il peut aussi être un indice des prix (§ II.A).

L'indice de Paasche est souvent inférieur à l'indice de Laspeyres, mais les pondérations de ces deux indices n'étant pas les mêmes, il est possible que cette propriété ne soit pas vérifiée (§ II.C).

3. Bonnes réponses : a), b) et c)

Un indice des *dépenses* ou du *coût de la vie* sert à mesurer l'évolution du niveau des dépenses de consommation entre deux périodes distinctes.
S'agissant des sommes à débourser par un ménage compte tenu des prix, mais aussi des quantités achetées, c'est un indice de valeur qui est donc réversible et transitif.

4. Bonnes réponses : a), b) et d)

Le taux annuel moyen de variation peut être déterminé graphiquement en utilisant une représentation graphique avec une ordonnée logarithmique (§ IV.A).

5. Bonnes réponses : c) et d)

Le taux de croissance pour période 1999-2001 est égal à :

$$1{,}029 \cdot 1{,}038 \cdot 1{,}021 - 1 \approx 9\ \%$$

Chapitre 3

1. Bonnes réponses : a) et d)

2. Bonnes réponses : a), b) et c)

On a deux variables liées par une relation linéaire inverse, leur coefficient de corrélation linéaire est égal à – 1 et les pentes des deux droites des moindres sont négatives.

3. Bonnes réponses : a) et b)

Le nuage de points est formé de 2 points, et par 2 points, on peut faire passer une droite, la représentation graphique nous montre que la droite est de pente positive, on a donc $r = + 1$

4. Bonne réponse : c)

Cette question a pour objectif de sensibiliser à l'attention qui doit être nécessairement portée à la signification des variables (§ II.E).

5. Bonnes réponses : b), c) et d)

On a un tableau de profils en colonne.

6. Bonnes réponses : a), b) et d)

La moyenne $\bar{x}$ est une moyenne pondérée des moyennes conditionnelles (§ I.C) ; en ce qui concerne les moyennes conditionnelles, elles s'obtiennent à partir du tableau de contingence, ou à partir du tableau des profils en lignes pour les moyennes conditionnelles de Y à X fixé et du tableau des profils en colonnes pour les moyennes conditionnelles de X à Y fixé.

7. Bonnes réponses : a), b) et d)

Si $a = 20$ et $b = 5$, il y a indépendance puisque les profils en colonnes sont identiques ($r = 0$).
Si $a = 0$ et $b = 0$, alors on a seulement deux observations différentes sur deux variables et dans ce cas, $r = \pm 1$ puisque le nuage de points est réduit à deux points (ici $r = -1$).
Si $a = 0$ et $b = 10$, il n'y a ni indépendance, ni liaison linéaire, donc $r \neq \pm 1$ et $r \neq 0$ (en fait $r = -0{,}5$).
Si $a = 10$ et $b = 10$, il y a indépendance puisque les profils en lignes sont identiques.

8. Bonne réponse : d)

Lorsqu'on connaît les marges, il suffit de connaître $(k - 1) \cdot (l - 1)$ effectifs du tableau de contingencedu fait des liaisons entre les effectifs marginaux et les effectifs du tableau ; le nombre $(k - 1) \cdot (l - 1)$ est appelée nombre de *degrés de liberté*.

9. Bonnes réponses : a), c) et d)

26 % des malades ont pris un somnifère *et* ont bien dormi.

10. Bonnes réponses : a) et d)

On a un tableau de profils en colonnes. Si les deux variables nominales étaient indépendantes, les deux profils-colonnes seraient identiques.

Les pourcentages de deux lignes ne s'aditionnent pas. Pour calculer le pourcentage total des salariés (hommes et femmes réunis), il faut utiliser les effectifs (total des emplois).

On obtient pour les non-salariés :

$$(13{,}4 \cdot 13\,670 + 7{,}3 \cdot 12\,243)/(13\,670 + 12\,243) = 10{,}5\ \%$$

Et pour les salariés :

$$(86{,}6 \cdot 13\,670 + 92{,}7 \cdot 12\,243)/(13\,670 + 12\,243) = 89{,}5\ \%$$

Chapitre 4

1. Bonnes réponses : a), c) et d)

Il y a $(T - p + 1)$ moyennes mobiles centrées de longueur impaire p et $(T - p)$ moyennes mobiles centrées de longueur paire p.

2. Bonne réponse : b) et d)

Si le facteur saisonnier est proportionnel à la tendance, on choisit le modèle multiplicatif, et dans ce cas, la courbe joignant les maxima est à peu près parallèle à celle qui joint les minima sur un graphique à ordonnée logarithmique (§ II).
Des maxima distants de 5 dates peuvent indiquer une composante saisonnière de période 5, mais cette seule information n'est pas suffisante pour choisir le modèle adapté.

3. Bonnes réponses : a) et b)

La moyenne mobile centrée de longueur p rend constante les séries périodiques de période p et de période sous-multiple de p.
La moyenne mobile centrée de longueur 2 p éliminent la composante saisonnière de période p puisque la somme des coefficients saisonniers sur une période est nulle.
La somme de p termes successifs divisée par p donne une évaluation de la tendance pour la date correspondant à celle du terme du milieu des p termes.
On peut calculer $(T - p)$ moyennes mobiles centrées de longueur p si p est pair, et $(T - p + 1)$ moyennes mobiles centrées de longueur p si p est impair, on a donc toujours au moins $(T - p)$ moyennes mobiles centrées.

4. Bonnes réponses : c) et d)

Le lissage exponentiel simple ne peut s'envisager que pour une chronique sans saisonnalité et sans évolution tendancielle ; la prévision tient d'autant plus compte des valeurs récentes de la série que la constante de lissage α est élevée.

Chapitre 5

1. Bonnes réponses : a) et d)

2. Bonnes réponses : a) et b)

3. Bonne réponse : c)

4. Bonnes réponses : b) et d)

Car $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ et A et B étant indépendants, $P(A \cap B) = P(A) \cdot P(B)$

5. Bonnes réponses : c) et d)

Soit A l'événement *le lièvre est touché par au moins un des chasseurs* , alors $P(A) = 1 - P(\bar{A})$

6. Bonnes réponses : c) et d)

7. Bonnes réponses : a), b) et d)

8. Bonnes réponses : b) et c)

9. Bonnes réponses : a) et d)

Seule la loi de probabilité d'une variable aléatoire *continue* est définie par une fonction de densité.

10. Bonnes réponses : b) et c)

11. Bonne réponse : a)

La réponse d) est vraie si a est positif.

12. Bonnes réponses : a), b) et d)

13. Bonnes réponses : a) et c)

14. Bonnes réponses : a) et c)

15. Bonnes réponses : a), b) et c)

Puisque $P\{(X = x_i) \cap (Y = y_i)\} = P(X = x_i \,|\, Y = y_i) \cdot P(Y = y_i)$ on peut calculer la loi jointe ;

$$P(X = 1) = P(X = 1 \,|\, Y = 1) \cdot P(Y = 1) + P(X = 1 \,|\, Y = 2) \cdot P(Y = 2) + P(X = 1 \,|\, Y = 3) \cdot P(Y = 3) = 0{,}2$$

16. Bonnes réponses : a), b) et c)

17. Bonne réponse : a)

Pour spécifier la loi jointe, il faut connaître en sus des k lois conditionnelles $\{Y \,|\, X = x_i\}$ la loi marginale de X, ou en sus des l lois conditionnelles $\{X \,|\, Y = y_j\}$ la loi marginale de Y.

Chapitres 6 et 7

1. Bonnes réponses : a), c) et d)

2. Bonnes réponses : a) et b)

3. Bonnes réponses : a), b) et d)

4. Bonnes réponses : a) et b)

5. Bonnes réponses : a) et b)

6. Bonnes réponses : b), c) et d)

La loi $\mathcal{B}(60 ; 0{,}05)$ peut être approchée par une loi $\mathcal{P}(3)$

$\Rightarrow P(X \geq 3) = 1 - P(X \leq 2) = 0{,}5768$

si $n = 4$ et $p = 0{,}01$: $P(X = 0) = (0{,}99)^4 \approx 0{,}96$

si $n = 50$ et $p = 0{,}08$: $P(3 < X \leq 10) = P(X \leq 10) - P(X \leq 3) = 0{,}573$

7. Bonne réponse : a)

8. Bonne réponse : c)

9. Bonnes réponses : b), c) et d)

$P(X = 2) = 0$, car pour une variable aléatoire continue, la probabilité d'un point est nulle.

10. Bonnes réponses : a) et b)

11. Bonnes réponses : a), b) et c)

Les v.a. X et Y étant indépendantes, leur coeficient de corrélation linéaire est nul.

12. Bonnes réponses : b), c), et d)

13. Bonnes réponses : a), c) et d)

On peut approcher la loi de Y par une loi de Gauss, X et Y étant indépendantes, la v.a. $(X + Y)$ suit approximativement une loi de Gauss puisque la somme de deux variables aléatoires gaussiennes indépendantes est gaussienne .

14. Bonnes réponses : a), b) et c)

$E(X^2) = \text{var}(X) + (E(X))^2 = 10$

15. Bonne réponse : a)

Pour une v.a. exponentielle, l'espérance est égale à l'écart-type et ses valeurs possibles sont supérieures à θ, donc à 0.

16. Bonne réponse : b)

Une somme de variables aléatoires binomiales indépendantes suit une loi binomiale si tous les paramètres p_i sont égaux.

17. Bonnes réponses : a), b), c) et d)

18. Bonnes réponses : a) et b)

La demande du produit pour 25 jours de fonctionnement suit une loi $\mathcal{B}(1\,000 ; 0{,}05)$, et on est dans les conditions d'approximation par la loi normale et aussi par la loi de Poisson.

Corrigés des exercices*

Chapitre 1

Exercice 1.1

1. *Population* : formée des 5 sous-populations (non disjointes) de l'ensemble des familles en France en 1968, 1975, 1982, 1990 et 1999 .
Unité statistique : une famille parmi cet ensemble de familles .
Caractère étudié : nombre d'enfants de 0 à 18 ans en milliers, caractère quantitatif discret.

2. Le nombre total de familles, le nombre de familles sans enfant, le nombre de familles avec 1 enfant, le nombre de familles avec 2 enfants augmentent au cours de la période 1968-1999, tandis que le nombre de familles de 4 enfants et le nombre de familles de 5 enfants ou plus diminuent. Le nombre total d'enfants augmente de 1968 à 1975, et diminue ensuite.

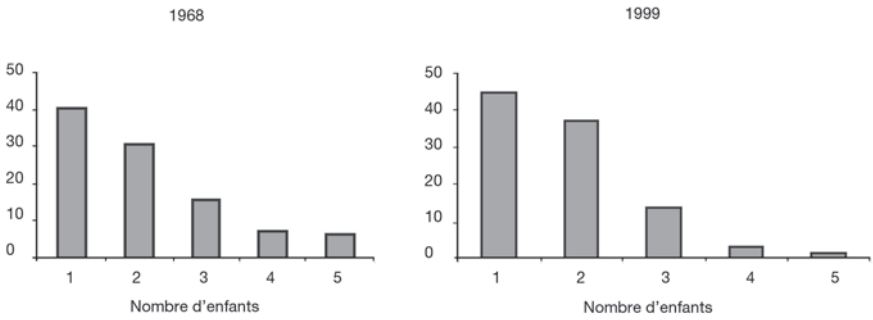
3. Pour tracer les diagrammes en bâtons, il est préférable d'utiliser les distributions de fréquences (en %).

	1968	1975	1982	1990	1999
Familles avec enfants	6 760	7 340	7 610	7 491	7 418
1 enfant	40,3	42,4	43,4	43,8	44,7
2 enfants	30,4	32,3	35,9	36,8	37,4
3 enfants	15,7	14,8	14,2	14,2	13,6
4 enfants	7,1	5,8	4,1	3,5	3,1
5 enfants ou plus	6,5	4,7	2,4	1,8	1,2
Total fréquences	100	100	100	100	100
Nombre total d'enfants	14 569	14 826	14 294	13 748	13 308
Moyenne	2,16	2,02	1,88	1,84	1,79
Écart-type	1,88	1,58	1,14	1,01	0,89

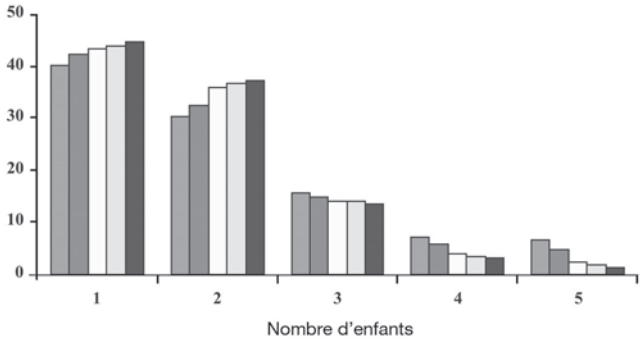
Mode de chaque distribution : « 1 enfant »

Le nombre de familles nombreuses diminuant, la dispersion autour de la valeur moyenne diminue.

* Les onglets renvoient au chapitre du cours correspondant.



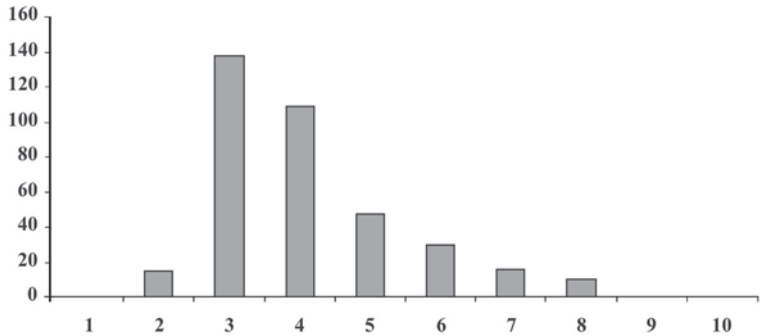
Diagrammes en bâtons juxtaposés



Exercice 1.2

1. *Population* : formée de 7 sous-populations associées chacune à une année (2000 à 2006) ; l'ensemble des jours d'une année constitue la population de l'année.
Unité statistique : une journée d'une année .
Caractère étudié : niveau de l'indice de la qualité de l'air, caractère qualitatif, mais aussi ordinal (les modalités du caractère sont ordonnées).

2. Diagramme en bâtons de la distribution 2006 : mode = « niveau 3 ».



3.

Année	2000	2001	2002	2003	2004	2005	2006	2000-2006
Niveau moyen	3,6	3,7	3,7	4,1	3,6	3,6	3,7	3,7

Le niveau moyen a été particulièrement élevé en 2003.

Exercice 1.3

1. $\overline{x_G} = 9{,}75 \quad s_G = 3{,}945 \quad \overline{x_P} = 11{,}1 \quad s_P = 3{,}727$

$Me_G = 10 \quad Me_P = 10 + 2 \cdot \dfrac{25 - 10}{25} = 11{,}2$

Pour une distribution symétrique, la moyenne est égale à la médiane.

2. $\bar{x} = 0{,}3 \cdot \overline{x_G} + 0{,}7 \cdot \overline{x_P} = 10{,}695 \approx 10{,}7$

3. Tous les indicateurs sont multipliés par 10

Exercice 1.4

1. *Population* : le personnel ouvrier d'un secteur industriel.
Unité statistique : un homme ou une femme de ce personnel.
Caractère étudié : le salaire annuel net en euros, variable statistique continue.

2.
a) On peut en fait proposer 4 indicateurs de tendance centrale :

$\overline{x_H} = 15\,400$ € $\quad Me = 14\,800$ €

$(Q_1 + Q_3)/2 = 15\,205$ € $\quad (D_1 + D_9)/2 = 15\,585$ €

b) On peut en fait proposer 3 indicateurs de dispersion :

$s_H = 3\,620$ € $\quad (Q_3 - Q_1) = 4\,910$ € $\quad (D_9 - D_1) = 9\,270$ €

c) On peut en fait proposer 3 indicateurs de dispersion relative :

$s_H / \overline{x_H} \approx 0{,}235 \quad (Q_3 - Q_1) / Me \approx 0{,}332 \quad (D_9 - D_1) / Me \approx 0{,}626$

3.

$$\frac{11 \cdot 82 + 13 \cdot 34 + 15 \cdot 12 + 18 \cdot n_4}{1\,524} = \frac{12 \cdot (128 + n_4)}{1\,536}$$

$$\Rightarrow \quad n_4 = 2 \quad \Rightarrow \quad N = 130$$

4. $s_F \approx 1\,509$ € $\quad s_F / \overline{x_F} \approx 0{,}125$

5. $\bar{x} = \dfrac{180 \cdot \overline{x_H} + 130 \cdot \overline{x_F}}{310} \approx 13\,974$ €

Exercice 1.5

Appelons x le coût total de la main d'œuvre :

$$\text{coût horaire moyen} = \frac{\text{coût total}}{\text{nombre total d'heures}} = \frac{x}{\frac{0{,}7x}{8} + \frac{0{,}3x}{10}} = \frac{1}{\frac{0{,}7}{8} + \frac{0{,}3}{10}} \approx 8,51 \text{ €}$$

$\Rightarrow$ moyenne harmonique pondérée

Exercice 1.6

1.

B_1 : $\sqrt[10]{(1{,}12)^2(1{,}08)^4(1{,}06)^4} = 1{,}0798 \approx 1{,}08$

$\Rightarrow$ taux de croissance moyen = 8 %

B_2 : $\sqrt[10]{(1{,}1)^3(1{,}08)^3(1{,}07)^4} = \sqrt[10]{2{,}2} = 1{,}0819 \approx 1{,}082$

$\Rightarrow$ taux de croissance moyen = 8,2 %

2. La banque B_1 est la moins performante. Soit x son taux durant la 3e période. On peut calculer x pour que le taux moyen de croissance égale celui de la banque B_2 :

$$(1{,}12)^2\,(1{,}08)^4\,(1 + x)^4 = 2{,}2 \quad \Rightarrow \quad x \approx 6{,}5\ \%$$

Exercice 1.7

1. *Population* : les 30 premiers groupes français de l'industrie et des services selon leur *CAHT* en 2001.

Unité statistique : un groupe parmi les 30 premiers groupes français de l'industrie et des services selon leur *CAHT* en 2001.

Caractères étudiés : deux caractères quantitatifs, le *CAHT* en millions d'€ et l'effectif.

2. *CA* : $n = 30$

$\bar{x} = 30\,000$ millions d'€ $\quad s_X = 19\,729$ millions d'€

Effectif : $\quad n = 30$

$\bar{y} \approx 134\,448 \quad s_Y \approx 87\,248$

3.1.

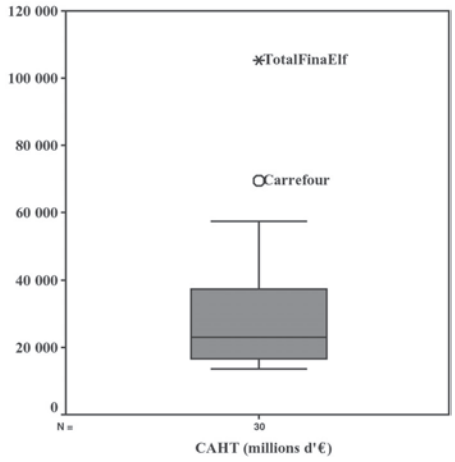
$n = 30 \quad \Rightarrow \quad P(Me) = 15{,}5 \quad \Rightarrow \quad Me = 23\,197$ millions d'€

$P(Q) = 8 \quad \Rightarrow \quad Q_1 = 16\,667$ millions d'€ et $Q_3 = 37\,200$ millions d'€

3.2. et 3.3. Graphiques *SPSS*

Frequency	Stem &	Leaf
6,00	1 .	334444
4,00	1 .	5677
6,00	2 .	000123
4,00	2 .	5567
1,00	3 .	0
2,00	3 .	67
3,00	4 .	023
,00	4 .	
1,00	5 .	1
1,00	5 .	7
2,00	Extremes (> = 69	486)

Stem width : 10 000
Each leaf : 1 case(s)



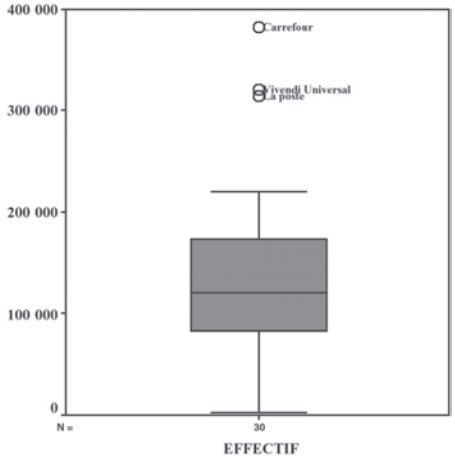
4. Le diagramme « branche et feuille » ne peut s'envisager que pour des distributions de population de taille peu élevée, contrairement à l'histogramme où l'hypothèse d'équirépartition à l'intérieur des classes n'est réaliste qu'avec un effectif suffisant dans chaque classe. Cette représentation permet de plus de ne pas perdre l'information valeur par valeur et aussi d'étiqueter éventuellement les observations.
La boîte de distribution met en évidence une valeur éloignée (Carrefour) et une valeur extrême (TotalFinaElf). Cette distribution asymétrique étalée vers les valeurs élevées sera modélisée par la loi de Pareto au chapitre 7, § IV.B.

5.1. La série étant ordonnée selon le *CA*, il faut maintenant l'ordonner selon l'effectif
$n = 30 \quad \Rightarrow \quad P(Me) = 15{,}5 \quad \Rightarrow \quad Me = 120\,510$
$P(Q) = 8 \quad \Rightarrow \quad Q_1 = 82\,892$ et $Q_3 = 173\,329$

5.2. et **5.3.** Graphiques *SPSS*
La boîte de distribution met en évidence trois valeurs éloignées : Carrefour, Vivendi Universal et La Poste.

Frequency	Stem &	Leaf
5,00	0 .	02234
5,00	0 .	57899
11,00	1 .	00111222234
4,00	1 .	6789
2,00	2 .	02
3,00	Extremes (> = 313854)	

Stem width : 100 000
Each leaf : 1 case(s)



Exercice 1.8

1. *Population* : ensemble des ménages en France en 1999 .
Unité statistique : un ménage en France en 1999 .
Caractère étudié : le revenu annuel en euros, hors revenus du patrimoine quantitatif continu.

2. $\bar{x} = \frac{1}{10} \sum_{i=1}^{10} \overline{x}_i \approx 25\,255$

3. Les moyennes n'étant pas égales aux centres des classes, l'hypothèse d'équirépartition n'est pas justifiée.

4. Indicateurs de tendance centrale : $\bar{x} \approx 25\,255$, $Me = D_5 = 20\,631$, $(D_9 + D_1)/2 = 26\,973$
Indicateur de dispersion : $D_9 - D_1 = 39\,338$
Indicateur de dispersion relative : $(D_9 - D_1)D_5 \approx 1{,}9$

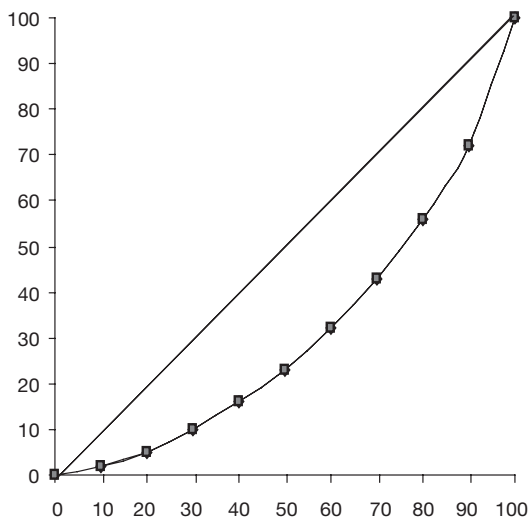
5. Distribution asymétrique étalée vers les valeurs élevées, car la médiane est inférieure à la moyenne (et on a aussi $Me < (D_9 + D_1)/2$). Les distributions de revenus (ou de salaires) sont toujours asymétriques et étalées vers les valeurs élevées.

6. Indicateur de disparité des revenus : $D_9/D_1 = 6{,}4$
$\Rightarrow$ les 10 % les mieux lotis perçoivent plus de 6 fois plus que les 10 % les moins bien lotis.

7. 16 % des revenus sont perçus par les 4 dixièmes des ménages aux revenus les plus faibles.

8. Courbe de concentration ou courbe de Lorenz

F_i (%)	10	20	30	40	50	60	70	80	90	100
R_i (%)	2	5	10	16	23	32	43	56	72	100



L'indice de Gini mesure ici la concentration des revenus des ménages. Il est égal au double de l'aire comprise entre la courbe de concentration et la bissectrice.
Cet indice est compris entre 0 et 1.
La valeur minimum 0 correspond au cas où la courbe est confondue avec la bissectrice et au cas de l'équirépartition : tous les individus ont une part égale du revenu.
La courbe s'éloigne de la bissectrice lorsque l'inégalité s'accroît.
A contrario, si un seul ménage détient la totalité du revenu, tous les autres ayant un revenu nul, l'indice de Gini vaut 1. Dans cette situation, la courbe est confondue avec les côtés du carré : axe des abscisses et segment vertical reliant le point {100 ; 0} au point {100 ; 100}.

Exercice 1.9

1. *Population* : les exploitations agricoles de France métropolitaine en 1979, 1988, 2000 et 2005
Unité statistique : une exploitation agricole de France métropolitaine en 1979, 1988, 2000 et 2005
Caractère étudié : la taille de la *SAU*, variable statistique continue

2. Soit c_1, c_2 et c_3 les taux annuels moyens de variation au cours de chacune des 3 périodes :

$(1 + c_1)^9 = \dfrac{1\,017}{1\,263} = (0{,}80522)^9 \Rightarrow c_1 \approx -2{,}4\ \%$

$(1 + c_2)^{12} = \dfrac{664}{1\,017} = (0{,}65290)^{12} \Rightarrow c_2 \approx -3{,}5\ \%$

$(1 + c_3)^3 = \dfrac{545}{664} = (0{,}96127)^5 \Rightarrow c_3 \approx -3{,}9\ \%$

Le taux annuel moyen de variation c de 1979 à 2005 est une moyenne géométrique pondérée des 3 taux c_1, c_2 et c_3 : $1 + c = \sqrt[26]{(1 + c_1)^9 \cdot (1 + c_2)^{12} \cdot (1 + c_3)^5}$

$\Rightarrow 1 + c = \sqrt[26]{\dfrac{545}{1\,263}} \approx 0{,}96819 \Rightarrow c \approx -3{,}2\ \%$

3.

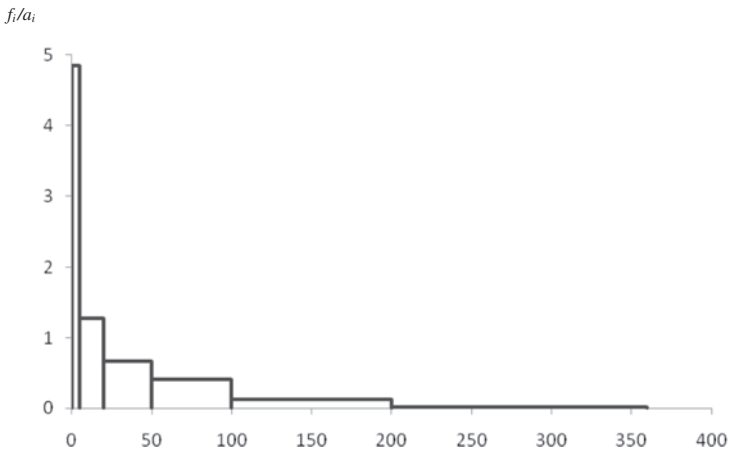
	1979	1988	2000	2005
SAU moyenne	23	28	42	50
SAU moyenne des exploitations de 50 ha ou plus	88	90	106	112

Le nombre des exploitations agricoles diminue, la taille moyenne des *SAU* augmente, ainsi que la taille moyenne des exploitations de 50 ha ou plus.

4. Le centre de la dernière classe étant par hypothèse la *SAU* moyenne des exploitations de 200 ha ou plus est égale en 2005 à 280 (= 4 762/17). On évalue ainsi la *SAU* maximum approximativement à 360 ha.

L'histogramme comporte 6 classes : 6 rectangles de hauteur f_i/a_i.

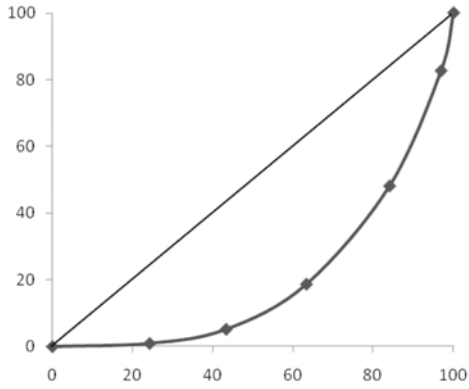
Taille de *SAU*	[0, 5[	[5, 20[	[20, 50[	[50, 100[	[100, 200[	[200, 360[
f_i	24,2	19,1	20,0	20,7	12,8	3,1
f_i/a_i	4,844	1,272	0,667	0,415	0,128	0,019



Pour construire un histogramme à classes inégales, se reporter à la page EXCEL'ENSE du n° 34 de la revue *Modulad*, « Réaliser un histogramme » (www.modulad.fr).

Courbe de concentration

p_i (%)	0	24,2	43,3	63,3	84,0	96,9	100
q_i (%)	0	1,0	5,2	18,7	48,1	82,7	100



Chapitre 2

Exercice 2.1

1. $I^A_{04/00}(p) = 112{,}5 \quad I^B_{04/00}(p) = 140 \quad I^C_{04/00}(p) = 100$

2. La moyenne arithmétique des indices élémentaires pondérée par la part des dépenses engagées par l'entreprise pour chacune des matières premières en 2000, est l'indice de Laspeyres des prix en 2004, base 2000.
Dépense 2000 = 8 000 + 2 000 + 3 000 = 13 000

$$L_{04/00}(p) = \frac{8\,000}{13\,000} \cdot 112{,}5 + \frac{2\,000}{13\,000} \cdot 140 + \frac{3\,000}{13\,000} \cdot 100 \approx 113{,}8$$

3. $I^A_{04/00}(q) = 60 \quad I^B_{04/00}(q) = 100 \quad I^C_{04/00}(q) = 160$

$$L_{04/00}(q) = \frac{8\,000}{13\,000} \cdot 60 + \frac{2\,000}{13\,000} \cdot 100 + \frac{3\,000}{13\,000} \cdot 160 \approx 89{,}2$$

4. Dépense 2004 = 5 400 + 2 800 + 4 800 = 13 000 $\Rightarrow$ $I_{04/00}(v) = 100$

5. Taux de variation des prix : 13,8 %
Taux de variation des quantités : – 10,8 %
Taux de variation des dépenses : 0 %
La non-variation de la dépense totale s'explique par une compensation entre les évolutions des prix et des quantités consommées : diminution de la quantité de matière première *A* dont le prix a augmenté, stagnation de la quantité de matière première *B* dont le prix a augmenté, et augmentation de la quantité de matière première *C* dont le prix a stagné.

Exercice 2.2

1. Soit c le taux annuel moyen de variation pendant entre 1980 et 2000 :

$$(1+c)^{20} = 1{,}5 \cdot \frac{168}{130} \cdot (1+0{,}06)^9 = 3{,}275 = (1{,}061)^{20} \quad \Rightarrow \quad c = 6{,}1\ \%$$

2. $CA_{2000} = CA_{1980} \cdot (0{,}95)^{20} \cdot 3{,}275 = CA_{1980} \cdot 1{,}174 \quad \Rightarrow \quad I_{2000/1980}(CA) = 117{,}4$

Exercice 2.3

Soit c_1 le taux trimestriel moyen de croissance entre le 31 décembre 1998 et le 30 septembre 2009 :
$(1 + c_1)^{43} = 1{,}4145 \approx (1{,}0081)^{43} \Rightarrow$ taux trimestriel moyen de croissance = 0,81 %
Soit c_2 le taux annuel moyen de croissance entre le 31 décembre 1998 et le 30 septembre 2008 :
$(1 + c_2)^{10} = 1{,}3908 = (1{,}0335)^{10} \Rightarrow$ taux annuel moyen de croissance = 3,35 %

Exercice 2.4

1. Indice de valeur de la production des « Produits végétaux » $_{2008/2007} = 99{,}7$

Indice de valeur de la production des « Oléagineux, protéagineux » $_{2008/2007} = 103{,}2$

Indice de valeur de la production des « Vins » $_{2008/2007} = 97{,}9$

2. Évolution 2008/2007 des prix à la production des « Produits végétaux »

$$= 99{,}7/103{,}6 - 1 \approx 0{,}038 = -3{,}8\ \%$$

Évolution 2008/2007 des prix à la production des « Oléagineux, protéagineux »

$$= 103{,}2/104{,}8 - 1 \approx 0{,}015 = -1{,}5\ \%$$

3. Évolution 2008/2007 du volume de la production des « Vins »

$$= 97{,}9/103{,}7 - 1 \approx 0{,}056 = -5{,}6\ \%$$

4. La diminution en valeur de la production de « Produits végétaux » entre 2007 et 2008 est due à une diminution des prix qui n'a pas été totalement compensée par l'augmentation du volume de la production.

L'augmentation en valeur de la production des « Oléagineux, protéagineux » entre 2007 et 2008 est due à l'augmentation du volume de la production qui a plus que compensé la diminution des prix.

La diminution en valeur de la production de « Vins » entre 2007 et 2008 est due à une diminution du volume de la production qui n'a pas été totalement compensée par l'augmentation des prix.

Exercice 2.5

1. La variation relative de la consommation médicale entre 1970 et 2000 peut s'écrire sous forme d'indice, elle est alors égale au rapport (multiplié par 100) de la consommation en 2000 par la consommation en 1970 : (123,545/6,494)·100-1 902,4

2.

Année	Indice 1970 = 100	Année	Indice 1970 = 100	Année	Indice 1970 = 100
1970	100				
1971	115,7	1981	545,1	1991	1 346,2
1972	131,9	1982	633,6	1992	1 439,4
1973	151,4	1983	721,4	1993	1 519,2
1974	178,4	1984	800,7	1994	1 568,5
1975	222,5	1985	878,4	1995	1 636,2
1976	258,9	1986	950,2	1996	1 682,2
1977	289,7	1987	997,4	1997	1 710,1
1978	347,2	1988	1 084,7	1998	1 735,8
1979	401,6	1989	1 176,1	1999	1 803,0
1980	465,3	1990	1 261,3	2000	1 902,4

3. et **4.**

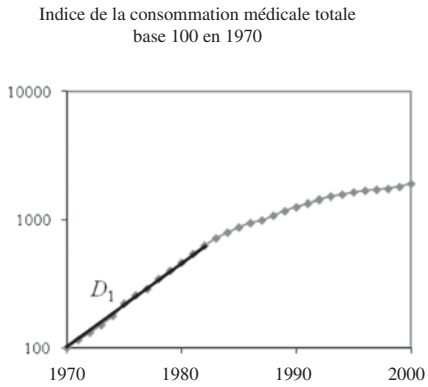


Figure 1 – Ordonnée logarithme

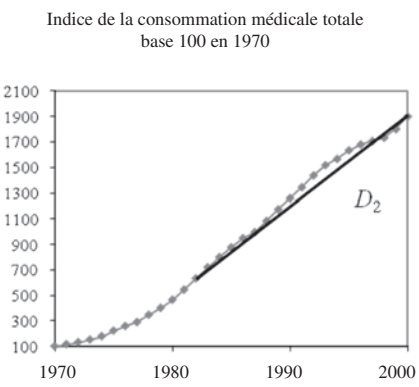


Figure 2 – Ordonnée arithmétique

Pendant la période 1970-1982, les points sont quasi alignés, le taux annuel de croissance c de l'indice est donc quasi constant et égal à la pente de la droite joignant les deux points extrêmes (*cf.* figure 1) :

$$(1+c)^{12} = 6{,}336 \quad \Rightarrow \quad c = \sqrt[12]{6{,}336-1} \quad \Rightarrow \quad c = 16{,}6\,\%$$

Entre 1982 et 2000, l'évolution n'est plus exponentielle, mais quasi-linéaire avec une augmentation annuelle moyenne de l'indice égale à la pente de la droite joignant les deux points extrêmes (*cf.* figure 2) :

$$a = ((CM_{2000} - CM_{1982})/CM_{1970}) \cdot 100/18 = (1902{,}4 - 633{,}6)/18 \approx 70{,}5\,\%$$

$\Rightarrow$ L'indice de la consommation médicale est passé d'une évolution exponentielle avec un taux annuel moyen de croissance de 16,6 % à une évolution quasi-linéaire avec une augmentation annuelle moyenne égale à 70,5 %.

5. Tenir compte de la variation des prix permet de passer d'un *indice de valeur* à un *indice de volume* et d'évaluer le taux annuel moyen de croissance du volume de la consommation entre 1970 et 1982 en s'affranchissant de l'illusion monétaire *:*

$$IndiceCM_{82/70} \cdot 100/3{,}187 = 198{,}8 \quad \Rightarrow \quad c' = \sqrt[12]{1{,}988} - 1 = 5{,}9\,\%$$

Entre 1982 et 2000, l'évolution n'est plus exponentielle, mais linéaire et on peut calculer l'augmentation annuelle moyenne de l'indice de volume :

$$\begin{aligned} a' &= \frac{1}{18} \cdot \left(\left(\frac{CM_{2000}}{5{,}584} - \frac{CM_{1982}}{I_{82/70}} \right) \cdot 100 \Big/ CM_{1970} \right) \\ &= \frac{1}{18} \cdot \left(\frac{Indice\,CM_{00/70}}{5{,}584} - \frac{Indice\,CM_{82/70}}{I_{82/70}} \right) \\ &= \frac{1}{18} \cdot \left(\frac{1902{,}4}{5{,}584} - \frac{633{,}6}{3{,}187} \right) \\ &\approx \left(340{,}7 - 198{,}8\right)/18 \approx 7{,}9\,\% \end{aligned}$$

6. En déflatant, on est passé, sur la période 1970-1982, d'un taux annuel moyen de croissance de 16,6 % pour l'indice de valeur à un taux annuel moyen de croissance de 5,9 % pour l'indice de volume, et sur la période 1982-2000, d'une augmentation annuelle moyenne de 70,5 % pour l'indice de valeur à une augmentation annuelle moyenne de 7,9 % pour l'indice de volume.

Chapitre 3

Exercice 3.1

1. *Population* : 30 000 assurés pour le risque « véhicules à moteur ».
Caractères étudiés : puissance fiscale, variable statistique discrète,
kilométrage parcouru la dernière année, variable statistique continue.

2. Distribution marginale de Y

Y milliers de km	< 10	$[10 ; 20[$	$[20 ; 30[$	$[30 ; 40[$	≥ 40
$f_{\bullet j}$	14 %	17 %	20 %	33 %	16 %

D'après l'énoncé : 1re classe $\rightarrow [2 ; 10[$ dernière classe $\rightarrow [40 ; 50[$
$\bar{y} = 27\,140$ km $\quad s_Y = 12\,648$ km

$Me = 20 + 10 \cdot \dfrac{50 - 31}{20} = 29{,}5$ milliers de km $= 29\ 500$ km

3. La distribution du kilométrage parcouru par les possesseurs d'une voiture d'une puissance fiscale d'au plus 6 CV est une distribution conditionnelle :

$Y \mid X \leq 6$	< 10	$[10 ; 20[$	$[20 ; 30[$	$[30 ; 40[$	≥ 40
Fréquence (%)	41,4 $\left(\dfrac{11{,}6}{28} \cdot 100\right)$	35,0 $\left(\dfrac{9{,}8}{28} \cdot 100\right)$	14,3 $\left(\dfrac{4{,}0}{28} \cdot 100\right)$	9,3 $\left(\dfrac{2{,}6}{28} \cdot 100\right)$	0

$\bar{y}\,/X \leq 6 = 14\,564$ km $\quad s_{Y/X \leq 6} = 9\,211$ km

Exercice 3.2

1. *Population* : les 100 salariées femmes et les 140 salariés homme d'une entreprise.
Unité statistique : un homme ou une femme parmi les 240 salariés.
Caractères étudiés : le salaire mensuel en euros, variable statistique continue,
l'ancienneté exprimée en années, variable statistique continue.

2. Parmi les 50 femmes ayant moins de 8 ans d'ancienneté, 44 gagnent moins de 2200 € :

$$44 \cdot 100 / 50 = 88 \ \%$$

3. Femmes : $\overline{x_1} = 2\,000$ € $\quad s_1 \approx 548$ €

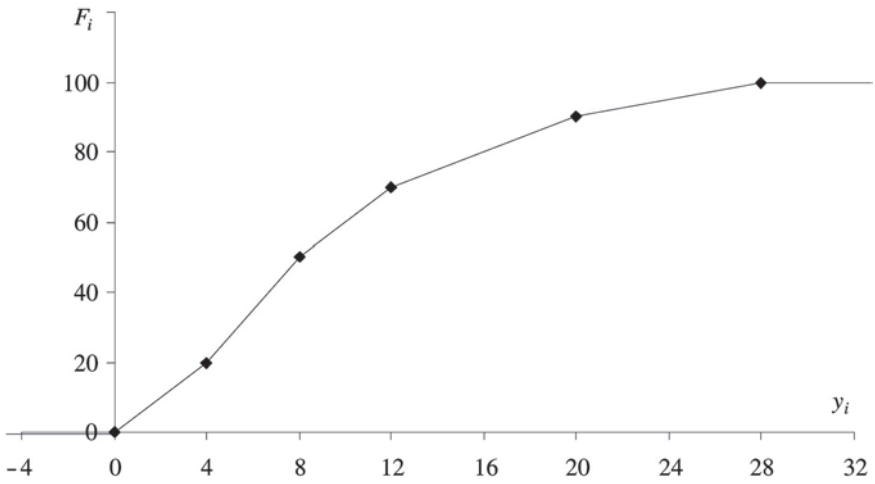
Hommes : $\overline{x_2} = 2\,508$ € $\quad s_2 \approx 697{,}50$ €

Ensemble : $\bar{x} = \dfrac{n_1 \cdot \overline{x_1} + n_2 \cdot \overline{x_2}}{n_1 + n_2} \approx 2\,296$ €

4. Moyenne et écart-type de l'ancienneté des femmes (en années) : $\bar{y} = 9{,}8 \quad s_Y \approx 6{,}7$

5. Graphe des fréquences cumulées : ligne brisée qui joint les points (y_i , F_i)

y_i	≤ 0	4	8	12	20	≥ 28
F_i (%)	0	20	50	70	90	100



6.

$Y \mid X \geq 1\,800$	[0 ; 4[	[4 ; 8[	[8 ; 12[	[12 ; 20[	[20 ; 28[
Fréquence (%)	13,3 (8/60)	33,3 (20/60)	16,7 (10/60)	20 (12/60)	16,7 (10/60)

7. $\hat{a} = r \cdot \dfrac{s_Y}{s_1} = 0{,}45 \cdot \dfrac{6{,}7}{548} \approx 0{,}0055 \qquad \hat{b} = \bar{y} - \hat{a} \cdot \bar{x} = 9{,}8 - 0{,}0055 \cdot 2000 \approx -1{,}2$

Point d'intersection : ($\bar{x}$, $\bar{y}$) = (2 000 ; 9,8)

Exercice 3.3

1. Indice *PIB* en 1991 = Indice 1990 (1 + variation 1991/100) = 102,6 (1 + 1/100) = 103,6

D4 f_x =D3*(1+B4/100)

	A	B	C	D	E
1	Année	Variation *PIB*	Variation *Consommation*	Indice *PIB*	Indice *Consommation*
2	1989			100	100
3	1990	2,6	2,5	102,6	102,5
4	1991	1	0,6	103,6	103,1
5	1992	1,4	1	105,1	104,1
6	1993	-0,9	-0,4	104,1	103,7
7	1994	2,2	1,4	106,4	105,2
8	1995	2,1	1,7	108,7	107,0
9	1996	1,1	1,6	109,9	108,7
10	1997	2,2	0,4	112,3	109,1
11	1998	3,5	3,9	116,2	113,4
12	1999	3,3	3,5	120,0	117,3
13	2000	3,9	3,6	124,7	121,6
14	2001	1,9	2,6	127,1	124,7
15	2002	1	2,4	128,4	127,7
16	2003	1,1	2	129,8	130,3
17	2004	2,5	2,5	133,0	133,5
18	2005	1,9	2,6	135,5	137,0
19	2006	2,2	2,4	138,5	140,3
20	2007	2,3	2,5	141,7	143,8
21	2008	0,4	1	142,3	145,2
22	2009	-2,2	0,8	139,1	146,4

$r = 0{,}992 \approx 1 \quad \Rightarrow$ liaison approximativement linéaire

2. et **3.**

$\hat{a} = 1{,}09 \quad \hat{b} = -11{,}40$ (résultat à obtenir avec une calculatrice ou avec Excel®)

Calcul du coefficient de corrélation linéaire et des coefficients de la droite des moindres carrés avec Excel :

r = COEFFICIENT.CORRELATION(D2:D22;E2:E22)

$\hat{a}$ = INDEX(DROITEREG(E2:E22;D2:D22);1)

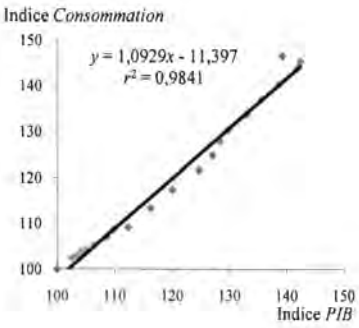
$\hat{b}$ = INDEX(DROITEREG(E2:E22;D2:D22);2)

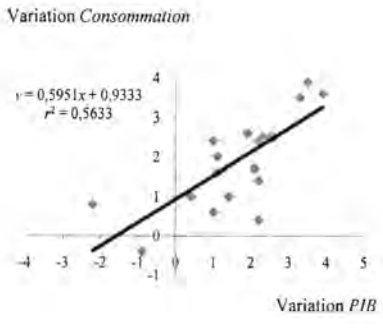
La part de variation de Y expliquée par la relation linéaire est égale à r^2, soit 98 %.

X' = variation en volume du PIB

Y' = variation en volume de la consommation privée

$r^2(X',Y') \approx 0{,}563 = (0{,}75)^2$





Pour obtenir et tracer avec Excel la droite des moindres carrés qui ajuste le nuage de points :

Onglet « Graphique », « Ajouter une courbe de tendance », type « Linéaire », option « Afficher l'équation sur le graphique », et « Afficher le coefficient de détermination (R^2) sur le graphique »

4. Les variables indice de volume du PIB et indice de volume de la consommation privée sont liées linéairement au temps :

$$r^2 \text{ (}\textit{indice PIB}\text{, }\textit{temps}\text{)} \approx 0{,}974$$

$$r^2 \text{ (}\textit{indice Consommation}\text{, }\textit{temps}\text{)} \approx 0{,}964$$

La liaison linéaire entre ces deux indices est due à leur liaison linéaire avec une même 3e variable qui est le temps.

L'étude des variations relatives permet d'éliminer la tendance. On dit alors qu'on a « stationnarisé » les séries.

Exercice 3.4

1. 810 / 1 000 = 81 %

2. 440 / 810 = 54,3 %

3.

	Pas du tout satisfait	Plutôt pas satisfait	Plutôt satisfait	Très satisfait	***Total***
> 2 ans d'ancienneté	10	50	245	195	500
≤ 2 ans d'ancienneté	40	90	205	165	500

Profils en ligne en pourcentage :

> 2 ans d'ancienneté	2	10	49	39	100
≤ 2 ans d'ancienneté	8	18	41	33	100

4.

	Pas du tout ou plutôt satisfait	Plutôt ou très satisfait
> 2 ans d'ancienneté	60	440
≤ 2 ans d'ancienneté	130	370

5. $n_{12} = \dfrac{500 \cdot 810}{1\,000} = 405$

Exercice 3.5

1. $r = -0{,}928 \quad \hat{a} = -0{,}018 \quad \hat{b} = 3{,}513$ (utilisation d'une calculatrice)

2. $\bar{y} = -0{,}018 \cdot x + 3{,}513$
Les deux droites des moindres carrés ont des pentes de même signe.

3. Recette globale $= 1\,000 \cdot x \cdot y \approx -18 \cdot x^2 + 3\,513 \cdot x$

$\Rightarrow \quad \dfrac{\partial R}{\partial x} = -18 \cdot 2x + 3\,513 \quad \Rightarrow \quad \dfrac{\partial R}{\partial x} = 0$ pour $x_c = 96{,}3$ tonnes

La recette globale est une fonction croissante de x entre $[0\ ,\ x_c]$ et décroissante pour $x > x_c$
$\Rightarrow$ la valeur critique que les producteurs ont intérêt à ne pas dépasser est la valeur : $x_c = 96{,}3$ tonnes

Exercice 3.6

1. *Population* : les n ménages (hors étudiants) interrogés pour l'enquête logement 1996 de l'INSEE.
Unité statistique : un ménage parmi les n ménages interrogés.
Caractères : niveau de vie en F/uc/mois (quantitatif continu), type socio-économique (qualitatif).

2. Profils en ligne et profils en colonne, ou distributions conditionnelles selon le niveau de vie et selon le type socio-économique.

Profils en colonne

Niveau de vie (en F/uc/mois) / Type socio-économique	Inférieur au 1er décile < 3 700	du 1er décile au 3e quartile [3 700 ; 9 933[	du 3e quartile au 9e décile [9 933 ; 13 900[	Au moins égal au 9e décile ≥ 13 900	Ensemble
Communes agricoles	22	13	7	5	12
Communes et quartiers ouvriers	41	41	31	22	37
Communes et quartiers des classes moyennes tertiaires	28	34	39	32	34
Communes et quartiers techniques très qualifiés	3	5	11	13	7
Quartiers huppés	6	6	12	28	9
Ensemble	100	100	100	100	100

Profils en ligne

Niveau de vie (en F/uc/mois) / Type socio-économique	Inférieur au 1er décile < 3 700	du 1er décile au 3e quartile [3 700 ; 9 933[	du 3e quartile au 9e décile [9 933 ; 13 900[	Au moins égal au 9e décile ≥ 13 900	Ensemble
Communes agricoles	18	70	8	4	100
Communes et quartiers ouvriers	11	71	12	6	100
Communes et quartiers des classes moyennes tertiaires	8	65	17	9	100
Communes et quartiers techniques très qualifiés	5	51	25	19	100
Quartiers huppés	6	44	20	30	100
Ensemble	10	65	15	10	100

3. Quartiers « huppés » : Me = 9 933 F/uc/mois

4. On peut proposer comme indicateur de disparité : $D_9/D_1 \approx 3{,}76$
Nombre sans dimension qui indique que le 9 e décile est 3,76 fois plus élevé que le 1er décile.

5. $((11 + 12) \cdot 0{,}15 + (13 + 28) \cdot 0{,}10)/0{,}25 = 30{,}2\ \%$

6. $\dfrac{0{,}05 \cdot 0{,}07n + 0{,}06 \cdot 0{,}09n}{0{,}16n} = 0{,}0556 = 5{,}56\ \%$

$\dfrac{0{,}51 \cdot 0{,}07n + 0{,}44 \cdot 0{,}09n}{0{,}16n} = 47{,}06\ \%$

$\dfrac{0{,}25 \cdot 0{,}07n + 0{,}20 \cdot 0{,}09n}{0{,}16n} = 22{,}19\ \%$

$\dfrac{0{,}19 \cdot 0{,}07n + 0{,}30 \cdot 0{,}09n}{0{,}16n} = 25{,}19\ \%$

Niveau de vie (F/uc/mois)	< 3 700	[3 700 ; 9 933[	[9 933 ; 13 900[	≥ 13 900	Total
Fréquence (%)	5,56	47,06	22,19	25,19	100

Exercice 3.7

1.

X	$n_{i\bullet}$	$\bar{y}_i$	$\sum_{j=1}^{n_{i\bullet}} (y_{ij} - \bar{y}_i)^2$
[5 ; 7[	12	7,17	27,67
[7 ; 9[	28	9	60
[9 ; 11[	39	10	120
[11 ; 13[	54	11,37	206,6
[13 ; 15[	41	12,58	161,95
[15 ; 17[	22	14,36	69,1
[17 ; 19]	4	16,5	11

$\bar{y} = 11{,}2$
$Sc_{intra} = 656{,}32$
$Sc_{inter} = 782{,}6$
$SC_{tot} = 1438{,}92$

$$\eta^2_{X/Y} = \frac{782{,}6}{1\,438{,}92} = 0{,}5843$$

Y	$n_{\bullet j}$	$\bar{x}_j$	$\sum_{i=1}^{n_{\bullet j}} (x_{ji} - \bar{x}_j)^2$
[5 ; 7[	10	6,8	17,6
[7 ; 9[	32	9,3	108,9
[9 ; 11[	59	10,8	312,95
[11 ; 13[	46	12,26	156,87
[13 ; 15[	36	13,9	127,55
[15 ; 17[	14	15,14	29,7
[17 ; 19]	3	17,33	2,67

$\bar{x} = 11{,}65$
$Sc_{intra} = 756{,}24$
$Sc_{inter} = 921{,}24$
$SC_{tot} = 1677{,}48$

$$\eta^2_{X/Y} = \frac{921{,}24}{1\,677{,}48} = 0{,}549$$

2. La courbe de régression de Y en x est une ligne brisée qui joint les points $(x_i, \bar{y}_i)$ x_i étant le centre de la i^e classe de X.

3. Les variables X et Y étant quantitatives, on peut mesurer leur liaison à l'aide du coefficient de corrélation linéaire : $r^2 = 0{,}542 = (0{,}736)^2$
On retrouve : $0 \leq r^2 \leq \min(\eta^2_{X/Y} ; \eta^2_{X/Y}) \leq \max(\eta^2_{X/Y} ; \eta^2_{X/Y}) \leq 1$

Exercice 3.8

1.

Eau minérale	X^C	Y^C
Arcens	2	3
Arvie	1	4
Badoit	2	2
Beckerich	1	1
Châteauneuf	3	4
Eau de Perrier	1	1
Faustine	3	2
La Salvetat	1	1
Perrier	1	1
Puits St-Georges	1	3
Pyrénées	1	1
Quézac	3	2
San Pellegrino	1	1
St-Diéry	1	3
St-Jean	2	2
St-Pierre	2	3
St-Yorre	3	4
Vernet	2	2
Vernière	1	2
Vichy-Célestins	3	4
Wattwiller	2	1

2. Distribution conjointe

X^C \ Y^C	1	2	3	4	
1	6	1	2	1	10
2	1	3	2	0	6
3	0	2	0	3	5
Total	7	6	4	4	21

Tableau des profils en ligne

X^C \ Y^C	1	2	3	4	
1	60	10	20	10	100
2	16,7	50	33,3	0	100
3	0	40	0	60	100
Profil moyen	33,3	28,6	19,1	19	100

3. Les profils en ligne n'étant pas identiques, les deux variables ne sont pas indépendantes.

4. Distribution conditionnelle de X^C sachant $\{Y > 300 \text{ mg/l}\}$:

X^C	Effectif
$C1_X$	3
$C2_X$	2
$C3_X$	3

Exercice 3.9

1. Taux trimestriel moyen : $(1 + c)^{11} = 339/117 \approx 2{,}90 = (1{,}10)^{11} \Rightarrow c \approx 10\ \%$

2.

2.1. $\hat{a} = 16{,}52 \qquad \hat{b} = 98{,}50$

2.2. $r = 0{,}90$

$\Rightarrow$ Part de variation de Y non expliquée par le modèle $= 1 - r^2 = 1 - 0{,}81 = 19\ \%$

3.

t^2	Nombre de contrats souscrits
1	117
4	178
9	149
16	189
25	145
36	173
49	170
64	223
81	223
100	281
121	285
144	339

3.1. $\hat{a} = 1{,}3$ $\hat{b} = 135{,}45$

3.2. $r = 0{,}94$

$\Rightarrow$ Part de variation de Y non expliquée par le modèle $= 1 - r^2 = 1 - 0{,}88 = 12\ \%$

4. Le modèle quadratique est préférable au modèle linéaire puisque le coefficient de corrélation linéaire est plus élevé pour ce modèle. On peut aussi dire que la part de variation de Y non expliquée est plus faible avec ce modèle.

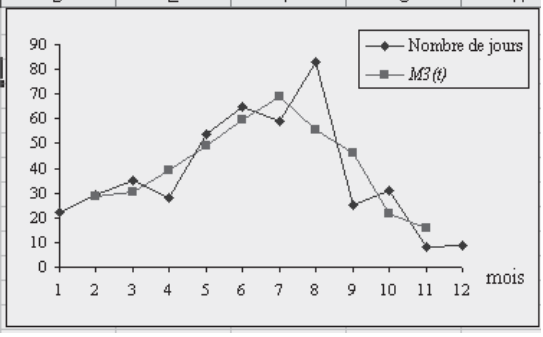
Chapitre 4

Exercice 4.1

1 et 2.

C3 : =(B2+B3+B4)/3

	A	B	C
1	Mois	Nombre de jours	$M_3(t)$
2	1	22	
3	2	29	28,67
4	3	35	30,67
5	4	28	39,00
6	5	54	49,00
7	6	65	59,33
8	7	59	69,00
9	8	83	55,67
10	9	25	46,33
11	10	31	21,33
12	11	8	16,00
13	12	9	



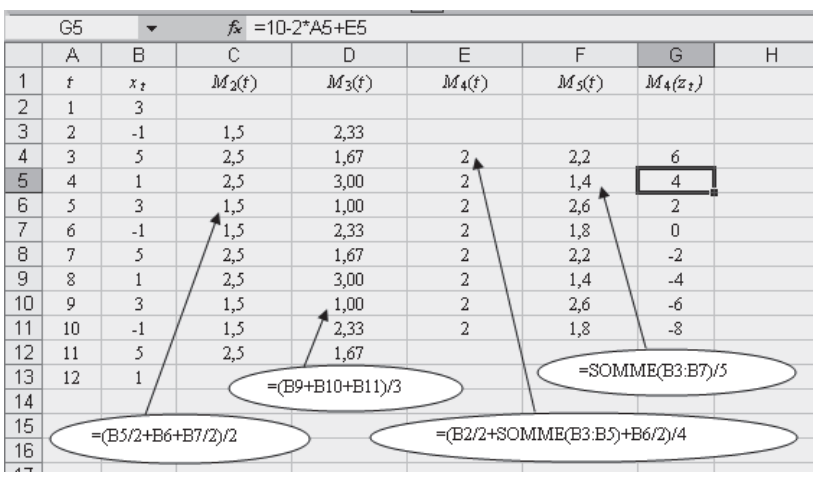
La moyenne mobile lisse la série chronologique, et permet d'évaluer la tendance.

Exercice 4.2

1.

G5 : =10-2*A5+E5

	A	B	C	D	E	F	G
1	t	x_t	$M_2(t)$	$M_3(t)$	$M_4(t)$	$M_5(t)$	$M_4(z_t)$
2	1	3					
3	2	-1	1,5	2,33			
4	3	5	2,5	1,67	2	2,2	6
5	4	1	2,5	3,00	2	1,4	4
6	5	3	1,5	1,00	2	2,6	2
7	6	-1	1,5	2,33	2	1,8	0
8	7	5	2,5	1,67	2	2,2	-2
9	8	1	2,5	3,00	2	1,4	-4
10	9	3	1,5	1,00	2	2,6	-6
11	10	-1	1,5	2,33	2	1,8	-8
12	11	5	2,5	1,67			
13	12	1					



La série x_t est périodique de période 4.
Toutes les suites de moyennes mobiles sont aussi périodiques de période 4.
La suite des moyennes mobiles de longueur 4 est constituée de termes constants égaux à la moyenne des termes sur une période.

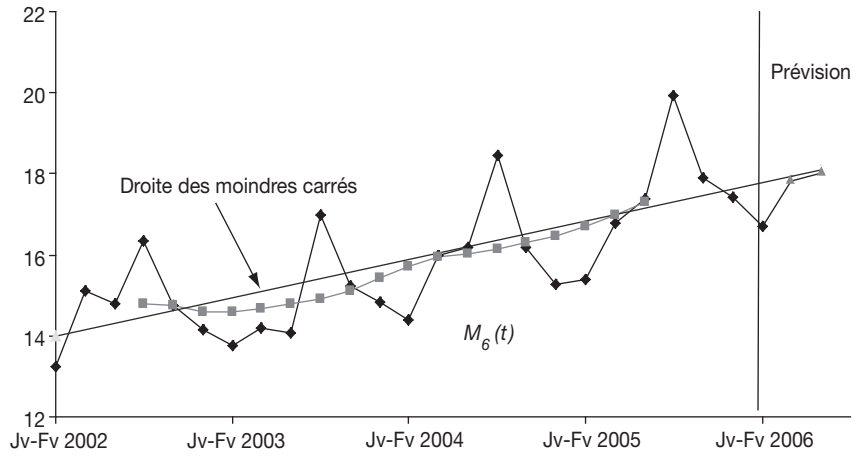
2. La moyenne mobile tranforme une série alignée en elle-même, donc la série $y_t = 10 - 2t$ est transformée en elle-même, et la suite des moyennes mobiles de longueur 4 de la série z_t est égale à : $10 - 2t + 2 = 12 - 2t$ ($t = 3$ à 10).

Exercice 4.3

1. Le schéma additif est adapté puisque les lignes brisées qui joignent d'une part, les maxima (distants de 6 dates), et d'autre part, les minima (distants de 6 dates) sont à peu près parallèles.

2. à 6. La période du mouvement saisonnier étant égale à 6, on évalue la tendance par la suite des moyennes mobiles de longueur 6.

t	x_t	$M_6(t)$	$x_t - M_6(t)$	s_t	s^*_t	*CVS*	Tendance	Prévision	Erreur de prévision
1	13,3				– 1,1	14,4	14,0		
2	15,1				– 0,2	15,3	14,1		
3	14,8				– 0,1	14,9	14,3		
4	16,3	14,8	1,6	2,0	2,0	14,3	14,5		
5	14,8	14,7	0,0	0,0	0,1	14,7	14,6		
6	14,2	14,6	– 0,5	– 0,8	– 0,7	14,9	14,8		
7	13,8	14,6	– 0,8	– 1,2	– 1,1	14,9	14,9		
8	14,2	14,7	-0,5	– 0,2	– 0,2	14,4	15,1		
9	14,1	14,8	– 0,7	– 0,2	– 0,1	14,2	15,3		
10	17,0	14,9	2,1	2,0	2,0	14,9	15,4		
11	15,2	15,1	0,1	0,0	0,1	15,2	15,6		
12	14,8	15,4	– 0,6	– 0,8	– 0,7	15,5	15,7		
13	14,4	15,7	– 1,3	– 1,2	– 1,1	15,5	15,9		
14	16,0	15,9	0,1	– 0,2	– 0,2	16,1	16,1		
15	16,2	16,0	0,1	– 0,2	– 0,1	16,3	16,2		
16	18,5	16,2	2,3	2,0	2,0	16,4	16,4		
17	16,2	16,3	– 0,1	0,0	0,1	16,1	16,5		
18	15,3	16,5	– 1,2	– 0,8	– 0,7	16,0	16,7		
19	15,4	16,7	– 1,3	– 1,2	– 1,1	16,5	16,9		
20	16,8	17,0	– 0,2	– 0,2	– 0,2	16,9	17,0		
21	17,4	17,3	0,1	– 0,2	– 0,1	17,5	17,2		
22	19,9				2,0	17,9	17,3		
23	17,9				0,1	17,8	17,5		
24	17,4				– 0,7	18,1	17,7		
25	**17,2**				– 1,1		**17,8**	**16,7**	**0,5**
26	**18,5**				– 0,2		**18,0**	**17,8**	**0,7**
27	**18,6**				– 0,1		**18,1**	**18,0**	**0,6**



Coefficients de la droite des moindres carrés ajustant la série *CVS* :

$\hat{a} = 0{,}16 \quad \hat{b} = 13{,}82$

$\Rightarrow \hat{y}_t = 0{,}16 \cdot t + 13{,}82$

$\Rightarrow \hat{x}_{25} = \hat{y}_{25} - 1{,}1 = 16{,}7 \quad \hat{x}_{26} = \hat{y}_{26} - 0{,}2 = 17{,}8 \quad \hat{x}_{27} = \hat{y}_{27} - 0{,}1 = 18$

$\Rightarrow$ Les erreurs de prévision étant toujours positives, l'erreur absolue moyenne est égale à l'erreur moyenne de prévision : $= (0{,}5 + 0{,}7 + 0{,}6)/3 = 0{,}6$

Pour tracer avec Excel la droite des moindres carrés qui ajuste la série CVS : onglet « Graphique », « Ajouter une courbe de tendance », type « Linéaire ». On peut utiliser ensuite l'onglet « Options » pour « Afficher l'équation sur le graphique » et pour « Afficher le coefficient de détermination (R^2) sur le graphique ».

Exercice 4.4

1. Sur la représentation graphique, on remarque une composante saisonnière de période 4.

2. à 6.

I2 : =21,6*A2+3270,8

	A	B	C	D	E	F	G	H	I	J
1	t	x_t	$M_4(t)$	$x_t / M_4(t)$	$1 + s_t$	*CVS*	$1 + e_t$	e_t	*tendance*	*Prévision*
2	1	3428			1,062	3228,7			3292,4	
3	2	3295			0,979	3365,8			3314,0	
4	3	3376	3341,6	1,010	1,006	3355,1	1,004	0,004	3335,6	
5	4	3195	3364,6	0,950	0,953	3353,5	0,997	-0,003	3357,2	
6	5	3573	3376,8	1,058	1,062	3365,3	0,997	-0,003	3378,8	
7	6	3334	3397,1	0,981	0,979	3405,7	1,003	0,003	3400,4	
8	7	3434	3426,5	1,002	1,006	3412,7	0,996	-0,004	3422,0	
9	8	3300	3452,4	0,956	0,953	3463,7	1,003	0,003	3443,6	
10	9	3703	3475,9	1,065	1,062	3487,7	1,003	0,003	3465,2	
11	10	3411	3493,1	0,976	0,979	3484,3	0,997	-0,003	3486,8	
12	11	3545			1,006	3523,0			3508,4	
13	12	3327			0,953	3492,1			3530,0	
14	13								3551,6	3771
15	14								3573,2	3498

Calcul des coefficients de la droite des moindres carrés avec Excel ® :

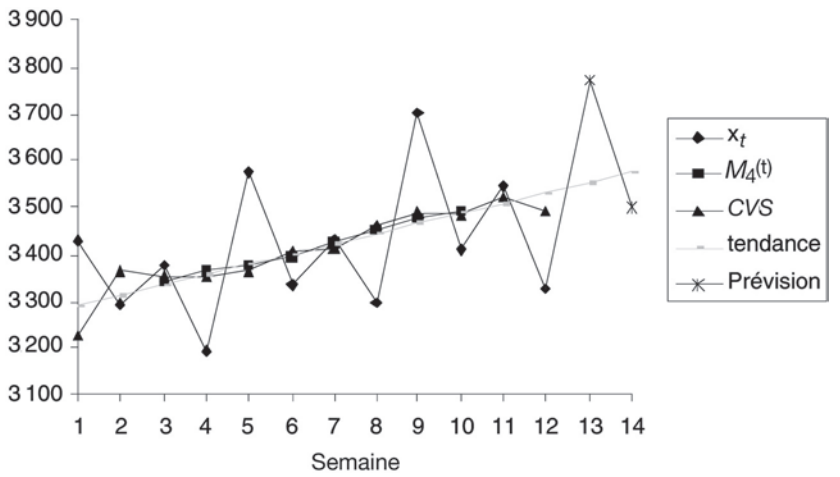
$\hat{a}$ = Index(Droitereg(F2:F13;A2:A13);1) = 21,6

$\hat{b}$ = Index(Droitereg(F2:F13;A2:A13);2) = 3 270,8

$\Rightarrow \quad \bar{y}_t = 21{,}6 \cdot \ t + 3\ 270{,}8$

$\Rightarrow \quad \hat{x}_{13} = (21{,}6 \cdot 13 + 3\ \ 271) \cdot 1{,}062 = 3\ \ 771$

$\hat{x}_{14} = (21{,}6 \cdot 14 + 3\ \ 271) \cdot 0{,}979 = 3\ \ 499$

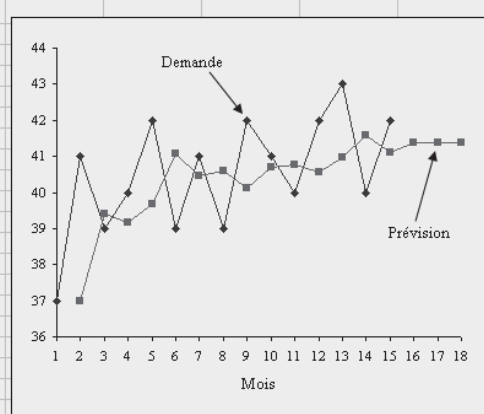


Exercice 4.5

1.

C8 fx =0,3*B7+0,7*C7

	A	B	C	D	E	F
1	Mois	Demande	*LES*	e_t	$ABS(e_t)$	$(e_t)^2$
2	1	37				
3	2	41	37	4,00	4,00	16,00
4	3	39	39,40	-0,40	0,40	0,16
5	4	40	39,16	0,84	0,84	0,71
6	5	42	39,66	2,34	2,34	5,46
7	6	39	41,07	-2,07	2,07	4,27
8	7	41	40,45	0,55	0,55	0,31
9	8	39	40,61	-1,61	1,61	2,60
10	9	42	40,13	1,87	1,87	3,50
11	10	41	40,69	0,31	0,31	0,10
12	11	40	40,78	-0,78	0,78	0,61
13	12	42	40,55	1,45	1,45	2,11
14	13	43	40,98	2,02	2,02	4,07
15	14	40	41,59	-1,59	1,59	2,52
16	15	42	41,11	0,89	0,89	0,79
17	16		41,38			
18	17		41,38			
19	18		41,38			
20				*EM*	*EAM*	*EQM*
21				0,56	1,48	3,09



Jusqu'au 6e mois inclus : $\hat{x}_t = 0{,}6 \cdot x_t + 0{,}4 \cdot \hat{x}_{t-1}$

À partir du 7e mois : $\hat{x}_t = 0{,}3 \cdot x_t + 0{,}7 \cdot \hat{x}_{t-1}$

2. Une constante de lissage élevée jusqu'au 6 e mois inclus permet un taux de réponse plus rapide au changement de niveau du début de la chronique.

3. Erreur moyenne = 0,56

Erreur absolue moyenne = 1,48

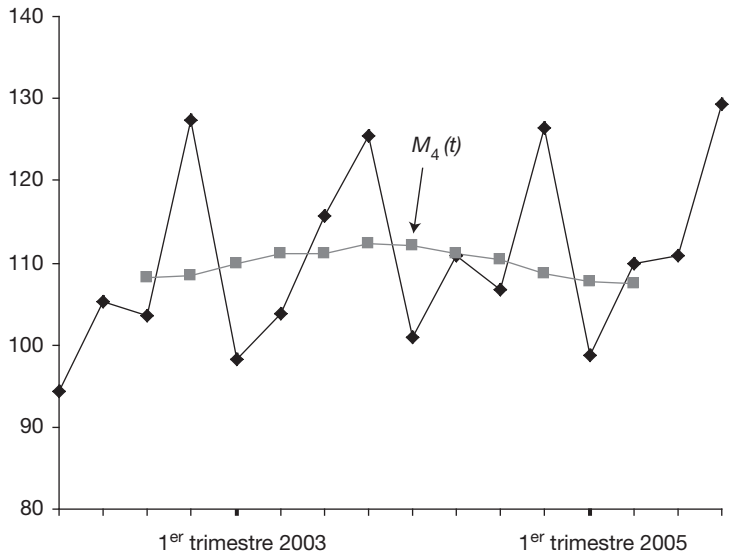
Erreur quadratique moyenne = 3,09

4. Sous l'hypothèse d'une série sans tendance, la demande est supposée constante. Les prévisions de la demande pour les mois 16, 17 et 18 sont égales à 41,4 unités.

Exercice 4.6

1. Cette chronique a une composante saisonnière de période 4 et une tendance approximativement constante sur la période 2002-2005. Les deux schémas de composition peuvent être envisagés. Nous choisissons le schéma additif.

2. Pour une chronique avec une composante saisonnière de période 4, la moyenne mobile de longueur 4 élimine la saisonnalité et permet d'évaluer la tendance.

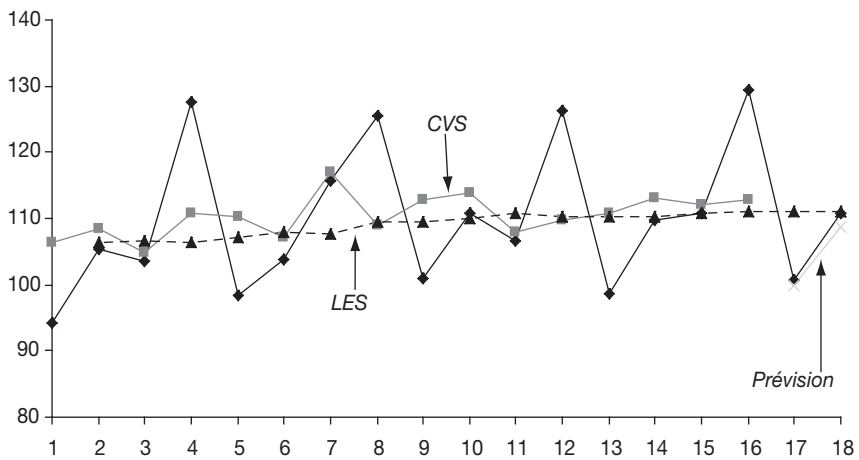


3. à 5.

t	x_t	$M_4(t)$	$x_t - M_4(t)$	s_t	CVS	LES ($\alpha = 0{,}3$)	Prévision	Erreur
1	94,2	108,12	– 4,65	– 12,04	106,3			
2	105,3	108,44	19,03	– 3,26	108,5	106,3		
3	103,5	109,78	– 11,48	– 1,26	104,7	107,0		
4	127,5	111,06	– 7,30	16,57	110,9	106,3		
5	98,3	111,14	4,56	– 12,04	110,3	107,7		
6	103,8	112,34	13,16	– 3,26	107,0	108,5		
7	115,7	112,07	– 11,17	– 1,26	117,0	108,0		
8	125,5	111,03	– 0,30	16,57	108,9	110,7		
9	100,9	110,30	– 3,70	– 12,04	112,9	110,2		
10	110,7	108,78	17,52	– 3,26	114,0	111,0		
11	106,6	107,71	– 13,48	– 1,26	107,9	111,9		
12	126,3	107,46	– 2,20	16,57	109,7	110,7		
13	98,7			– 12,04	110,8	110,4		
14	109,8			– 3,26	113,0	110,5		
15	110,8			– 1,26	112,1	111,3		
16	129,4			16,57	112,8	111,5		
17	**100,8**					**111,5**	**99,9**	**0,9**
18	**110,8**					**111,5**	**108,6**	**2,2**

La série *CVS* pouvant être considérée sans tendance, on peut utiliser le lissage exponentiel simple.
Après avoir resaisonnalisé les prévisions obtenues avec le *LES*, on obtient les prévisions de l'indice trimestriel pour les deux premiers trimestres 2006 qu'on peut comparer aux observations.
On obtient :

Erreur moyenne de prévision = Erreur absolue moyenne de prévision = 1,55



Chapitre 5

Exercice 5.1

a) 75/120 = 0,625
b) 50/120 ≈ 0,417
c) 45/120 = 0,375
d) 100/120 ≈ 0,833

Exercice 5.2

A = « Être allé en Espagne » B = « Avoir pris l'avion »
$P(A) = 0{,}6$
$P(B) = 0{,}45$
$P(A \cap B) = 0{,}25$
$P(\overline{A} \cap \overline{B})$ $P(\overline{A \cup B})$ $1 - P(A \cup B)$
$= 1 - (P(A) + P(B) - P(A \cap B)) = 0{,}2$

Exercice 5.3

1. A = « Lire Notre Campus » B = « Lire la Vie Étudiante »
$P(A) = 23\,522/32\,564 \approx 0{,}722$
$P(B) = 18\,859/32\,564 \approx 0{,}579$
$P(A \cap B) = 11\,422/32\,564 \approx 0{,}351$

a) $P(\overline{A} \cap \overline{B})$ $P(\overline{A \cup B})$ $1 - P(A \cup B)$ $1 - (P(A) + P(B) - P(A \cap B)) \approx 0{,}05$

b) $P(A \cap \overline{B})$ $P(A) - P(A \cap B)) \approx 0{,}371$

2. a) $0{,}05^2 = 0{,}0025$
b) $0{,}351 \cdot 0{,}05 \approx 0{,}018$

Exercice 5.4

$n = 2$: 4 événements élémentaires $\{P,P\}, \{P,F\}, \{F,P\}, \{F,F\}$
$P(A) = 3/4$ $P(B) = 1/2$ $P(A \cap B) = 1/2 \neq P(A) \cdot P(B)$
$\Rightarrow$ A et B non indépendants
$n = 3$: 8 événements élémentaires
$\{P,P,P\}, \{P,P,F\}, \{P,F,P\}, \{P,F,F\}, \{F,P,P\}, \{F,P,F\}, \{F,F,P\}, \{F,F,F\}$
$P(A) = 1/2$ $P(B) = 3/4$ $P(A \cap B) = 3/8 = P(A) \cdot P(B)$
$\Rightarrow$ A et B indépendants
et si on continue…, on peut montrer que A et B ne sont indépendants que pour $n = 3$

Exercice 5.5

1. Il y a 2^5 familles différentes de 5 enfants et $\binom{5}{3}$ familles de 5 enfants avec 3 filles et 2 garçons. Par hypothèse toutes les familles sont équiprobables :

$$P(\text{trois filles et deux garçons}) = \frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}} = \frac{\binom{5}{3}}{2^5} = \frac{10}{32} = 0{,}3125$$

2. Toutes les familles ne sont plus équiprobables. On a une probabilité égale à $(0{,}48)^3 \cdot (0{,}52)^2$ d'avoir une famille de 5 enfants avec 3 filles et 2 garçons, et toutes les familles étant incompatibles, on a :

$$\Rightarrow \quad P(\text{trois filles et deux garçons}) = \binom{5}{3} (0{,}48)^3 \cdot (0{,}52)^2 = 0{,}299$$

Exercice 5.6

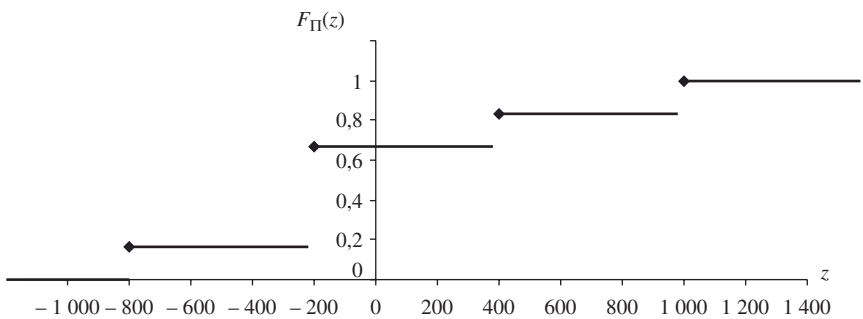
1. $E(X) = 5/3 = 1{,}667 \quad \sigma_X = 0{,}943$ (utiliser les fonctions statistiques d'une calculatrice)
$\Rightarrow \quad E(\Pi(X)) = 600 \cdot E(X) - 800 = 200$ € $\quad \sigma_{\Pi(X)} = 600 \cdot \sigma_X = 565{,}68$ €
Signification de l'espérance du profit : sur un très grand nombre de jours, la moyenne du profit sera voisine de 200 €.

2. Loi de probabilité du profit :

valeur de X	0	1	2	3
valeur de Π	– 800	– 200	400	1 000
Probabilité	1/6	1/6	1/2	1/6

Fonction de répartition du profit : fonction en escalier, continue à droite, les points de discontinuité correspondant aux valeurs possibles du profit.

z	< -800	$[-800, -200[$	$[-200, 400[$	$[400, 1\,000[$	$\geq 1\,000$
$F_\Pi(z) = P(\Pi \leq z)$	0	1/6	2/6	5/6	1



Exercice 5.7

1. $E(B) = 0 \cdot 0{,}6 + 500 \cdot 0{,}25 + 1 \quad 500 \cdot 0{,}1 + 2 \quad 500 \cdot 0{,}05 = 400$ €
$\Rightarrow$ Pour espérer un bénéfice moyen de 50 € par assuré, le montant de la cotisation doit être fixé à 450 €

2. Le montant encaissé de la part de deux assurés est égal à 900 €.

⇒ Il faut qu'au plus un seul des deux assurés ait un sinistre, et le règlement de ce sinistre ne peut pas dépasser 500 €

⇒ La probabilité demandée est égale à : $0{,}6^2 + 2 \cdot 0{,}6 \cdot 0{,}25 = 0{,}66$

Exercice 5.8

1. $E(X) = 1{,}1$ $\quad \text{var}(X) = 0{,}69$ (utiliser les fonctions statistiques d'une calculatrice)

2. $Y = \sum_{i=1}^{12} X_i \quad \Rightarrow$ valeurs possibles de Y : $\{0, 1, 2, \ldots, 24\}$

$$E(Y) = \sum_{i=1}^{12} E(X_i) = 12 \cdot 1{,}1 = 13{,}2$$

$$\text{var}(Y) = \sum_{i=1}^{12} \text{var}(X_i) = 12 \cdot 0{,}69 = 8{,}28$$ car les v.a. X_i sont indépendantes

$$P(Y = 0) = P\left(\bigcap_{i=1}^{12} \{X_i = 0\}\right) = \prod_{i=1}^{12} P(X_i = 0) = 0{,}3^{12} = 5{,}3 \cdot 10^{-7}$$

car les X_i sont indépendantes

3.1.

X \ Z	0	1	2	3	4
0	1	0	0	0	0
1	0,1	0,6	0,3	0	0
2	$\underbrace{0{,}1^2}_{0{,}01}$	$\underbrace{2 \cdot 0{,}6 \cdot 0{,}1}_{0{,}12}$	$\underbrace{2 \cdot 0{,}3 \cdot 0{,}1 + 0{,}6^2}_{0{,}42}$	$\underbrace{2 \cdot 0{,}6 \cdot 0{,}3}_{0{,}36}$	$\underbrace{0{,}3^2}_{0{,}09}$

3.2.

$$P(Z = 0) = P\big(\{(Z = 0) \cap (X = 0)\} \cup \{(Z = 0) \cap (X = 1)\} \cup \{(Z = 0) \cap (X = 2)\}\big)$$
$$= 1 \cdot 0{,}3 + 0{,}1 \cdot 0{,}3 + 0{,}01 \cdot 0{,}4 = 0{,}334$$
$$P(Z = 1) = 0{,}6 \cdot 0{,}3 + 0{,}12 \cdot 0{,}4 = 0{,}228$$
$$P(Z = 2) = 0{,}3 \cdot 0{,}3 + 0{,}42 \cdot 0{,}4 = 0{,}258$$
$$P(Z = 3) = 0{,}36 \cdot 0{,}4 = 0{,}144$$
$$P(Z = 4) = 0{,}09 \cdot 0{,}4 = 0{,}036$$
$E(Z) = 1{,}32$ $\quad \text{var}(Z) \approx 1{,}39$ (utilisation d'une calculatrice)

Exercice 5.9

Appelons D l'événement :
{une pellicule tirée au hasard dans la production est défectueuse}
$P(D) = P(\{D \text{ et machine } A\} \cup \{D \text{ et machine } B\} \cup \{D \text{ et machine } C\})$

Les 3 événements étant incompatibles :
$P(D) = P(D \cap A) + P(D \cap B) + P(D \cap C) = 0{,}2 \cdot 0{,}06 + 0{,}5 \cdot 0{,}05 + 0{,}3 \cdot 0{,}03 = 0{,}046$
$P(A|D)$ $P(B|A)$ $P(A)/P(D) = 0{,}261$
$P(A|D)$ $P(B|A)$ $P(A)/P(\overline{D}) = 0{,}94 \cdot 0{,}2/0{,}954 = 0{,}197$

Exercice 5.10

1.

X \ Y	u	0	1	Loi de Y
0	1/4	a	1/8	15/40 + a
1	1/5	b	1/10	12/40 + b
Loi de X	18/40	$a + b$	9/40	27/40 + a + b

$$\Rightarrow \quad a + b = \frac{13}{40} \qquad \frac{18}{40} \cdot \left(\frac{15}{40} + a\right) = \frac{1}{4} \quad \Rightarrow \quad a = \frac{10}{18} - \frac{15}{40} = \frac{13}{72} \quad \Rightarrow \quad b = \frac{13}{90}\, a$$

2.

X	0	1
P	$\frac{15}{40} + a = \frac{40}{72} = \frac{5}{9}$	$\frac{12}{40} + b = \frac{40}{90} = \frac{4}{9}$

Puisque X et Y sont indépendantes, les lois conditionnelles de X pour les différentes valeurs de Y sont identiques à la loi marginale de X.

3.

X \ Y	u	0	1	Loi de Y
0	1/4	1/5	1/8	23/40
1	1/5	1/8	1/10	17/40
Loi de X	18/40	13/40	9/40	1

$$E(X) = \frac{17}{40} \qquad E(Y) = \frac{18}{40} \cdot u + \frac{9}{40} \qquad E(X \cdot Y) = \frac{u}{5} + \frac{1}{10}$$

Si $\rho = 0$, alors $\text{cov}(X, Y) = 0$:

$$\Rightarrow \quad E(X \cdot Y) = \frac{u}{5} + \frac{1}{10} = E(X) \cdot E(Y) = \frac{17}{40} \cdot \left(\frac{18}{40} \cdot u + \frac{9}{40}\right)$$

$$\Rightarrow \quad 2u - \frac{17 \cdot 18}{160}\, u = \frac{17 \cdot 9}{160} - 1 \quad \Rightarrow \quad u = -0{,}5$$

Exercice 5.11

1.

X \ Y	0	1	Loi de X
0	p	$1/2 - p$	1/2
1	$1/3 - p$	$1/6 + p$	1/2
Loi de Y	1/3	2/3	1

Toutes les probabilités devant être comprises entre 0 et 1, on doit avoir :

$$\begin{cases} 0 \le p \le 1 \\ p \le 1/2 \\ p \le 1/3 \\ p \le 5/6 \end{cases} \quad \Rightarrow \quad 0 \le p \le 1/3$$

2. $E(X) = 1/2 \quad E(Y) = 4/3 \quad E(X \cdot Y) = 2 \cdot (p + 1/6)$
$\Rightarrow \quad \text{cov}(X, Y) = 2 \cdot (p + 1/6) - 2/3 = 2p - 1/3$

$E(X^2) = 1/2 \quad E(Y^2) = 8/3 \quad \Rightarrow \quad \text{var}(X) = 1/4 \quad \text{var}(Y) = 8/9$

$$\Rightarrow \quad \rho(X, Y) = \frac{2p - 1/3}{\frac{1}{2} - \frac{2\sqrt{2}}{3}} = \frac{6p - 1}{\sqrt{2}}$$

Chapitre 6

Exercice 6.1

1. Au i^e individu ($i = 1$ à 10), on associe une variable de Bernoulli :

$$X_i = \begin{cases} 1 & \text{si } i^e \text{ individu gaucher} \quad p = 0{,}1 \\ 0 & \text{sinon} \quad q = 0{,}9 \end{cases}$$

Soit Y, le nombre de gauchers parmi les 10 individus :
Y est une somme de 10 v.a. indépendantes de Bernoulli de même paramètre $p = 0{,}1$

$$\Rightarrow \quad Y = \sum_{i=1}^{10} X_i \rightarrow \mathcal{B}(10 ; 0{,}1)$$

On utilise les tables de la loi Binomiale pour calculer les probabilités :
$P(Y \ge 1) = 1 - P(Y = 0) = 1 - 0{,}3486 = 0{,}6514 \quad P(Y \le 3) = 0{,}9872$

2. Pour que chacun des 10 membres du personnel trouve une paire de ciseaux à sa convenance, il faut que le nombre de gauchers soit au plus égal à 3 et au moins égal à 1 :
$P(1 \le Y \le 3) = P(Y \le 3) - P(Y < 1) = P(Y \le 3) - P(Y = 0) = 0{,}9872 - 0{,}3486 = 0{,}6386$

3.

Y	0	1	2	3	4	5	6	7	8	9	10
Z	9	10	10	10	9	8	7	6	5	4	3

En utilisant la table d'une loi binomiale de paramètres 10 et 0,1, on déduit la loi de Z :

Z	7	8	9	10
P	0,0001	0,0015	0,3599	0,6385

Les probabilités associées aux valeurs 3, 4, 5 et 6 sont négligeables.

Exercice 6.2

1. Au i^e assuré (i = 1 à 10 000), on associe une variable de Bernoulli :

$$X_i = \begin{cases} 1 & \text{si } i^e \text{ assuré a un accident de ce type} \quad p = 1/5\,000 \\ 0 & \text{sinon} \quad q = 1 - p \end{cases}$$

Soit Y, le nombre d'accidents par an parmi les 10 000 assurés.
On suppose les X_i indépendantes, Y est alors une somme de 10 000 v.a. indépendantes de Bernoulli de même paramètre $p = 1/5\,000$:

$$Y = \sum_{i=1}^{10\,000} X_i \to \mathcal{B}(10\,000 ; 1/5\,000) \xrightarrow[n > 50 \quad \text{et} \quad p < 0{,}1]{} \mathcal{P}(2)$$

$$\Rightarrow \quad P(Y \leq 3) = 0{,}8571 \quad \text{(lecture de table)}$$

Exercice 6.3

1. À la $i^{ème}$ machine (i = 1,2), on associe :

$$X_i = \begin{cases} 1 & \text{si } i^{ème} \text{ machine disponible} \quad p = 0{,}9 \\ 0 & \text{sinon} \end{cases}$$

Le nombre Y de machines disponibles un jour quelconque est une somme de 2 v.a. indépendantes de Bernoulli de même paramètre $p = 0{,}9 \quad \Rightarrow \quad Y \to \mathcal{B}(2 ; 0{,}9)$

Valeurs de Y	0	1	2
Probabilité	0,01	0,18	0,81

2.1.

Valeurs de N

Y \ Z	0	1	2	3
0	0	0	0	0
1	0	1	1	1
2	0	1	2	2

2.2.

Valeurs de N	0	1	2
Probabilité	0,109	0,324	0,567

$P(N = 2) = P\big(\{Y = 2 \cap Z = 2\} \cup \{Y = 2 \cap Z = 3\}\big)$
$= P\big(\{Y = 2 \cap Z = 2\}\big) + P\big(\{Y = 2 \cap Z = 3\big)$
$= P(Y = 2) \cdot (P(Z = 2) + P(Z = 3)) = 0{,}81 \cdot 0{,}7 = 0{,}567$
$P'(N = 1) = 0{,}18 \cdot (0{,}2 + 0{,}4 + 0{,}3) + 0{,}81 \cdot 0{,}2 = 0{,}324$
$P'(N = 0) = 1 - 0{,}324 - 0{,}567 = 0{,}109$
$E(N) = 1{,}458 \quad \Rightarrow$ Marge brute moyenne = 29,16 €

3. Y désigne maintenant le nombre de machines tombant en panne au cours de la 1re année.

$$X_i = \begin{cases} 1 & \text{si } i^e \text{ machine en panne au cours de la 1 re année} \quad p = 0{,}05 \\ 0 & \text{sinon} \end{cases}$$

$(i = 1, \ldots, 60)$

Hypothèse : les 60 v.a. de Bernoulli sont indépendantes
$\Rightarrow$ Y est une somme de 60 v.a. indépendantes de Bernoulli de même paramètre $p = 0{,}05$

$$Y = \sum_{i=1}^{60} X_i \to \mathcal{B}(60\,;\,0{,}05) \xrightarrow[\text{n} > 50 \quad \text{et} \quad p < 0{,}1]{} \mathcal{P}(3)$$

i) On a en moyenne 3 pannes puisque $E(Y) = 3$. En cas de machines non garanties, le coût moyen est donc égal à : $3 \cdot 200 = 600$ €.

Le discount étant de 1 200 €, on choisit le discount.
ii) La remise est inférieure au coût de réparation si plus de 6 machines tombent en panne :
$P(Y > 6) = 1 - P(Y \leq 6) = 1 - 0{,}9665 = 0{,}0335 > 1\ \%$
$\Rightarrow$ on choisit la garantie pour chaque machine.

Exercice 6.4

1. À chaque essai, on peut associer une v.a. de Bernoulli de paramètre 0,95. Le nombre d'essais Y nécessaires pour se connecter 5 fois suit une loi de Pascal de paramètres $K = 5$ et $p = 0{,}95$. Appliquons les résultats du cours :

$$E(Y) = \frac{K}{p} = \frac{5}{0{,}95} = 5{,}26 \qquad \text{var}(Y) = \frac{K \cdot (1 - p)}{p^2} = 0{,}277$$

2. Pour avoir $\{Y = 5\}$, il faut s'être connecté les 5 fois avec succès :
$P(Y = 5) = (0{,}95)^5 \approx 0{,}774$
La v.a. Y peut prendre toutes les valeurs entières au moins égales à 5 :
$P(Y > 6) = 1 - P(Y = 5) - P(Y = 6) = 1 - 0{,}774 - 5 \cdot (0{,}95)^5 \cdot 0{,}05 = 0{,}0325$

Exercice 6.5

1. Lorsque $X_t = x$, le nombre de demandes satisfaites Y_t peut prendre toutes les valeurs entières de 0 à x, et chaque demande est satisfaite avec une probabilité $(1 - \pi)$.
$\{Y_t | X_t \quad x\} \implies \mathcal{B}(x\,;\,1 - \pi)$

2. $P(Y_t = k) = P(\{Y_t = k\} \cap \{X_t \geq k\}) = \sum_{x \geq k} P(Y_t = k \,|\, X_t = x) \cdot P(X_t = x)$

$$= \sum_{x \geq k} \binom{x}{k} \cdot (1-\pi)^k \cdot \pi^{x-k} \cdot e^{-\mu} \cdot \frac{(\mu)^x}{x!} = e^{-\mu} \sum_{x \geq k} \frac{x!}{k!(x-k)!} \cdot (1-\pi)^k \cdot \pi^{x-k} \cdot \frac{(\mu)^x}{x!}$$

$$= e^{-\mu} \sum_{x \geq k} \frac{1}{k!(x-k)!} \cdot \left(\frac{1-\pi}{\pi}\right)^k \cdot (\mu \cdot \pi)^x = \frac{e^{-\mu}}{k!} \cdot \left(\frac{1-\pi}{\pi}\right)^k \sum_{x \geq k} \frac{1}{(x-k)!} \cdot (\mu \cdot \pi)^k$$

$$= \frac{e^{-\mu}}{k!} \cdot \left(\frac{1-\pi}{\pi}\right)^k \cdot (\mu \cdot \pi)^k \sum_{x \geq 0} \frac{(\mu \cdot \pi)^x}{x!} = e^{-\mu} \cdot \frac{((1-\pi) \cdot \mu)^k}{k!} \cdot e^{\mu \cdot \pi}$$

$$= \left(e^{-(1-\pi) \cdot \mu} \cdot \frac{((1-\pi) \cdot \mu)^k}{k!} \right)$$

$$\Rightarrow \quad Y_t \to \mathcal{P}((1-\pi) \cdot \mu)$$

3. $\mu = 10$ et $\pi = 0{,}2 \quad \Rightarrow \quad Y_t \to \mathcal{P}(8) \quad \Rightarrow \quad P(Y_t < 8) = 0{,}4530$
$P(3 < Y_t \leq 10) = P(Y_t \leq 10) - P(Y_t \leq 3) = 0{,}8159 - 0{,}0424 = 0{,}7735$

Exercice 6.6

1. À la i^e minute ($i = 1$ à 30), on associe :

$$X_i = \begin{cases} 1 & \text{si vente } i^{\text{e}} \text{ minute} \quad p = 0{,}01 \\ 0 & \text{sinon} \quad q = 0{,}99 \end{cases}$$

Soit Y, le nombre d'unités vendues en 30 min.
$\Rightarrow$ Y est une somme de 30 v.a. indépendantes de Bernoulli de même paramètre $p = 0{,}01$

$$\Rightarrow \quad Y = \sum_{i=1}^{30} X_i \to \mathcal{B}(30\,;\,0{,}01) \quad \Rightarrow \quad P(Y \geq 3) = 1 - P(Y \leq 2) = 1 - 0{,}9967 = 0{,}0033$$

2. Le magasin est ouvert 7 h 30 par jour, donc 450 min par jour.

$$Z = \text{nombre d'unités vendues par jour} = \sum_{i=1}^{450} X_i \to \mathcal{B}(450\,;\,0{,}01) \quad \Rightarrow \quad E(Z) = 4{,}5$$

$$Z \to \mathcal{B}(450\,;\,0{,}01) \xrightarrow[n > 50 \quad \text{et} \quad p < 0{,}1]{} \mathcal{P}(4{,}5)$$

3. $P(Z > 8) = 1 - P(Z \leq 8) = 1 - 0{,}9597 = 0{,}0403$

Exercice 6.7

1. $X \to \mathcal{P}(5) \quad P(X \leq 2) = 0{,}1247$
$P(\{X \leq 2\} \cup \{X \geq 6\}) = 0{,}1247 + 1 - 0{,}6160 = 0{,}5087$

$$P\langle X \leq 6 \,|\, X \geq 2 \rangle = \frac{P(2 \leq X \leq 6)}{P(X \geq 2)} = \frac{P(X \leq 6) - P(X < 2)}{1 - P(X < 2)} = \frac{0{,}7622 - 0{,}0404}{1 - 0{,}0404} = 0{,}7522$$

2. $Y = X_1 + X_2 \to \mathcal{P}(10)$, car Y est somme de deux v.a. de Poisson indépendantes
$P(Y = 10) = 0{,}5831 - 0{,}4580 = 0{,}1251$

3. $P(X \geq 3) = P(\{X \geq 3 \cap \lambda = 6\} \cup \{X \geq 3 \cap \lambda = 8\} = P(\{X \geq 3 \cap \lambda = 6\} + P(\{X \geq 3 \cap \lambda = 8\})$
$= P(\{X \geq 3 \mid \lambda = 6\} \cdot P(\lambda = 6) + P(X \geq 3 \mid \lambda = 8) \cdot P(\lambda = 8) = 0{,}938 \cdot 2/3 + 0{,}9862/3 \approx 0{,}954$

Exercice 6.8

1. $X_i = \begin{cases} 1 \text{ si } i^{\text{e}} \text{ autobus en panne} & p = 0{,}0025 \\ 0 \text{ sinon} & q = 0{,}9975 \end{cases}$

$\Rightarrow Y$ est somme de 1 000 v.a. de Bernoulli de même paramètre $p = 0{,}0025$ qu'on suppose indépendantes.

$$\Rightarrow Y = \sum_{i=1}^{1\,000} X_i \rightarrow \mathcal{B}(1\,000 ; 0{,}0025) \quad \Rightarrow \quad E(Y) = 2{,}5 \quad \text{var}(Y) \approx 2{,}5$$

2. $$\rightarrow Y = \sum_{i=1}^{1\,000} X_i \rightarrow \mathcal{B}(1\,000 ; 0{,}0025) \longrightarrow \mathcal{P}(2{,}5)$$

$n > 50 \quad p < 0.1$

$P(3 < Y < 7) = P(Y \leq 6) - P(Y \leq 3) = 0{,}9858 - 0{,}7576 = 0{,}2282$

3. $P(Y \leq 8) = 0{,}9989 \quad \Rightarrow$ La capacité du service de maintenance doit au moins égale à 8 autobus pour que la probabilité que toutes les pannes soient traitées dans la journée, soit au moins égale à 0,998.

4. $P(Y > 6) = 1 - P(Y \leq 6) = 1 - 0{,}9858 = 0{,}0142$

5. $$Z = \sum_{i=1}^{365} X_i \rightarrow \mathcal{B}(365 ; 0{,}0142) \longrightarrow \mathcal{P}(5{,}183)$$

$n > 50 \quad p < 0.1$

$P(Z = 0) = e^{-5{,}183} \approx 0{,}0056$

Exercice 6.9

1. $X_1 \rightarrow \mathcal{P}(3) \quad X_2 \rightarrow \mathcal{P}(2)$
X_1 et X_2 indépendantes $\Rightarrow Y = X_1 + X_2 \rightarrow \mathcal{P}(5)$

2. $P(Y = 8) = 0{,}0653$

3. $$P(X_1 = 5/Y = 8) = \frac{P(\{X_1 = 5\} \cap \{x_2 = 3\})}{P(Y = 8)}$$

$$= \frac{P(\{X_1 = 5\} \cdot \{X_2 = 3\})}{P(Y = 8)} = \frac{0{,}1008 \cdot 0{,}1804}{0{,}0653} = 0{,}2785$$

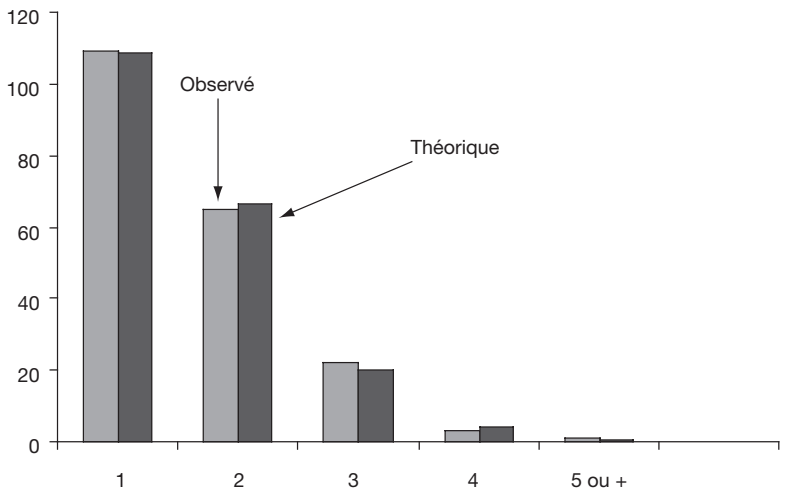
Exercice 6.10

1. $\bar{x} = 0{,}61 \quad \text{Var}(X) = 0{,}6079$

2. Variable discrète :
– les valeurs possibles sont les nombres entiers positifs ou nuls ;
– la moyenne est peu différente de la variance.

On a une variable discrète à valeurs positives ou nulles avec une moyenne quasi-égale à la variance. On peut envisager une loi de Poisson $\mathcal{P}$ (0,61). La comparaison des diagrammes en bâtons des distributions observée et théorique montre une excellente adéquation.

Nombre de décès x_k	0	1	2	3	4	5 ou +	Total
Cumul des années n_k	109	65	22	3	1	0	200
Probabilité $\mathcal{P}$ (0,61)	0,5434	0,3314	0,1011	0,0206	0,0031	0,0004	1
Nombre théorique n^*_k	108,7	66,3	20,2	4,1	0,6	0,1	200



Chapitre 7

Exercice 7.1

1. $X \to \mathcal{N}(45\ ;\ 4)$
$P(X < 39) = F_U((39 - 45)/4) = F_U(-1{,}5) = 1 - F_U(1{,}5) = 1 - 0{,}9332 = 0{,}0668$
$P(X \geq 48) = 1 - P(X < 48) = 1 - F_U(0{,}75) = 1 - 0{,}7734 = 0{,}2266$
$P(39 < X < 48) = P(X \leq 48) - P(X \leq 39) = 0{,}7734 - 0{,}0668 = 0{,}7066$

2. $P(|X - m| \leq \sigma) = P(m - \sigma \leq X \leq m + \sigma) = F_U(1) - F_U(-1) = 2F_U(1) - 1 = 0{,}6826$
(en fait, résultat de cours : § II.C)

3. Puisque $m = 45$ et $\sigma = 4$, on a : $\{41 \leq X \leq 49\} \Leftrightarrow \{m - \sigma \leq X \leq m + \sigma\}$
$P(41 \leq X \leq 49 \,|\, X \geq 39) = P(\{41 \leq X \leq 49\} \cap \{X \geq 39\})/P(X \geq 39)$
$= P(41 \leq X \leq 49)/P(X \geq 39)$
$= 0{,}6826 / 0{,}9332 = 0{,}7315$

Exercice 7.2

1. $P(X \leq 2\,400) = F_U((2\,400 - m)/\sigma) = 0{,}0228 \quad \Rightarrow \quad (2\,400 - m)/\sigma = -2$
$P(X > 3\,000) = 0{,}0446 \quad \Rightarrow \quad P(X \leq 3\,000) = 0{,}9554 \quad \Rightarrow \quad (3\,000 - m)/\sigma = 1{,}7$
On résout un système de 2 équations à 2 inconnues : $m \approx 2\,724$ € $\quad \sigma \approx 162$ €
$\Rightarrow \quad X \rightarrow \mathcal{N}(2\,724 ; 162)$

2. Soit X_i le gain du i^e mois, par hypothèse, les X_i sont *iid* à X (*iid* pour « indépendants et identiquement distribués »).
La v.a. Y égale au gain pendant trois mois est une somme de 3 v.a. normales indépendantes et par conséquent, suit une loi normale :

$$Y = \sum_{i=1}^{3} X_i \rightarrow \mathcal{N}(3m ; \sigma\sqrt{3}), \text{ soit : } \mathcal{N}(8\,172 ; 280{,}6)$$

3. $P(Y > 8\,700) = 1 - F_U(528/280{,}6) = 1 - F_U(1{,}88) = 1 - 0{,}9699 = 0{,}301$

Exercice 7.3

1. $p = P(X > 2{,}5) = 1 - P(X \leq 2{,}5) = 1 - F_U(1{,}67) = 1 - 0{,}9525 = 0{,}0475 \approx 0{,}05$

2. À la i^e imprimante tirée, on associe une v.a. de Bernoulli X_i de paramètre 0,05 :

$$X_i = \begin{cases} 1 & \text{si durée de vie} > 2{,}5 \text{ millions de pages} \\ 0 & \text{sinon} \end{cases}$$

$\Rightarrow \quad Y$ est une somme de 60 v.a. indépendantes de Bernoulli de même paramètre p (les X_i sont considérées indépendantes puisque les imprimantes ont été tirées au hasard dans une production supposée suffisamment importante pour avoir un taux de sondage n/N inférieur à 10 %)

$$\Rightarrow \quad Y = \sum_{i=1}^{60} X_i \rightarrow \mathcal{B}(60 ; 0{,}05) \xrightarrow[n > 50,\ p < 0{,}1]{} \mathcal{P}(3)$$

3. En utilisant les tables de la loi de Poisson, on obtient :
$P(Y = 6) = 0{,}9665 - 0{,}9161 = 0{,}0504$
$P(Y \geq 3) = 1 - P(Y \leq 2) = 1 - 0{,}4232 = 0{,}5768$

Exercice 7.4

1. $P(75 \leq X \leq 125) = F_U(1) - F_U(-1) = 0{,}6826$
en fait : $P(75 \leq X \leq 125) = P(|X - m| \leq \sigma)$
$P(X > 150) = 1 - F_U(2) = 1 - 0{,}9772 = 0{,}0228$
en fait : $P(X > 150) = P(X > 2\sigma)$

2. $P(X > n_1) = 0{,}9 \Rightarrow F_U\left(\dfrac{n_1 - 100}{25}\right) = 0{,}1 \Rightarrow \dfrac{n_1 - 100}{25} = -1{,}2816 \Rightarrow n_1 \approx 68$

$P(X < n_2) = 0{,}4 \Rightarrow F_U\left(\dfrac{n_2 - 100}{25}\right) = 0{,}4 \Rightarrow \dfrac{n_2 - 100}{25} = -0{,}2533 \Rightarrow n_2 \approx 94$

3. On cherche a tel que : $P(|X - m| < a) = 0{,}9$

$$\Rightarrow \quad P\left(\frac{|X-m|}{\sigma} < \frac{a}{\sigma}\right) = 0{,}9 \quad \Rightarrow \quad 2F_U\left(\frac{a}{\sigma}\right) - 1 = 0{,}9$$

$$\Rightarrow \quad F_U\left(\frac{a}{\sigma}\right) = 0{,}95 \quad \Rightarrow \quad \frac{a}{\sigma} = 1{,}645 \quad \Rightarrow \quad a = 41{,}125$$

$$\Rightarrow \quad [m - a \,;\, m + a] = [58{,}87 \,;\, 141{,}13]$$

Exercice 7.5

1. Soit X_i la variable aléatoire de Bernoulli associée la i^e bouteille « grand cru » achetée :

$$X_i = \begin{cases} 1 & \text{si } i^e \text{ bouteille « vin courant »} \quad p = 0{,}12 \\ 0 & q = 0{,}88 \end{cases} \qquad i = 1, \ldots, 200$$

Le nombre Y de bouteilles de vin courant parmi les 200 bouteilles achetées est égal à la somme des 200 variables de Bernoulli X_i : $Y = \sum_{i=1}^{200} X_i$

Les 200 bouteilles étant supposées tirées au hasard dans l'ensemble des bouteilles « grand cru » avec un taux de sondage inférieur à 10%, la v.a. Y suit une loi binomiale $B(200 \,;\, 0{,}12)$.

$$E(Y) = np = 24 \qquad \text{var}(Y) = npq = 21{,}12$$

Puisque $npq = 21{,}12 > 18$, la loi de Y peut être approchée par la loi normale $\mathcal{N}(24 \,;\, 4{,}6)$.

2. Comme on approxime une loi discrète par la loi normale, on fait la correction de continuité :

$P(Y > 20) = 1 - P(Y \leq 20) = 1 - F_U\{(20 + 0{,}5 - 24)/4{,}6\} = 1 - F_U(-0{,}76) = 0{,}7764$

$$P(Y < 30 \,|\, Y > 20) = \frac{P(20 < Y < 30)}{P(Y > 20)} = \frac{P(Y < 30) - P(Y \leq 20)}{0{,}7764}$$

$$P(Y<30 \,|\, Y>20) = \frac{F_U(1{,}2) - F_U(-0{,}76)}{0{,}7764} = \frac{0{,}8849 - 0{,}2236}{0{,}7764} = 0{,}8517$$

3. Les bouteilles de type courant, en nombre Y, occasionnent une perte unitaire de 1,5 €. Les bouteilles réellement « grand cru », en nombre $(200 - Y)$, créent un bénéfice unitaire de 2,50 € (= 6 – 3,5). Donc, au total :

$Bénéfice = 2{,}5 \cdot (200 - Y) - 1{,}5Y = 500 - 4Y \quad \Rightarrow \quad P(Bénéfice > 0) = P(Y < 125) \approx 1$

Exercice 7.6

1. Sachant que la probabilité d'une réunion de 2 événements incompatibles est égale à la somme des probabilités de ces événements, on a :

$P(A) = P(\{\text{vrais jumeaux et 2 garçons}\} \cup \{\text{faux jumeaux et 2 garçons}\})$
$= P(\text{vrais jumeaux et 2 garçons}) + P(\text{faux jumeaux et 2 garçons})$
$P(A) = P(2\,\text{G} \,|\, \text{vrais jumeaux}) \cdot P(\text{vrais jumeaux}) + P(2\,\text{G} \,|\, \text{faux jumeaux}) \cdot P(\text{faux jumeaux})$
$= \lambda/2 + (1 - \lambda)/4 = (\lambda + 1)/4 = P(B)$
$P(C) = P(\{\text{faux jumeaux}\} \cap \{\text{1 garçon et 1 fille}\})$
$= P(\text{1 garçon et 1 fille} \,|\, \text{faux jumeaux}) \cdot P(\text{faux jumeaux}) = (1 - \lambda)/2$

2. À la i^e naissance, on associe :

$$X_i = \begin{cases} 1 & \text{si } i^e \text{ naissance avec 1 } G \text{ et 1 } F \quad p = (1 - \lambda)/2 \\ 0 & \text{sinon} \end{cases}$$

$\Rightarrow$ Y est une somme de 1 000 v.a. indépendantes de Bernoulli de même paramètre p

$\Rightarrow$ $Y = \sum_{i=1}^{1\,000} X_i \rightarrow \mathcal{B}(1\,000 ; (1 - \lambda)/2)$

$E(Y) = 500 \cdot (1 - \lambda)$ $\quad$ $\text{var}(Y) = 250 \cdot (1 - \lambda^2)$

3. Si $\lambda = 0{,}35$: $\quad Y \rightarrow \mathcal{B}(1\,000 ; 0{,}325) \xrightarrow[npq > 18]{} \mathcal{N}(325 ; 14{,}8)$

$$P(Y > 300) = 1 - P(Y \leq 300) \simeq 1 - F_U\left(\frac{300 - 325}{14{,}8}\right) = F_U(1{,}69) \approx 0{,}9545$$

$$P(310 \leq Y \leq 350) = P(Y \leq 350) - P(Y < 310)$$
$$= F_U(1{,}69) - F_U(-1{,}01) = 0{,}9545 - (1 - 0{,}8438) = 0{,}7983$$

$$\Rightarrow \quad P(310 \leq Y \leq 350 \,|\, Y > 300) = P(\{310 \leq Y 350\} \cap \{Y > 300\})/P(Y > 300)$$
$$= 0{,}7983/0{,}9545 \approx 0{,}836$$

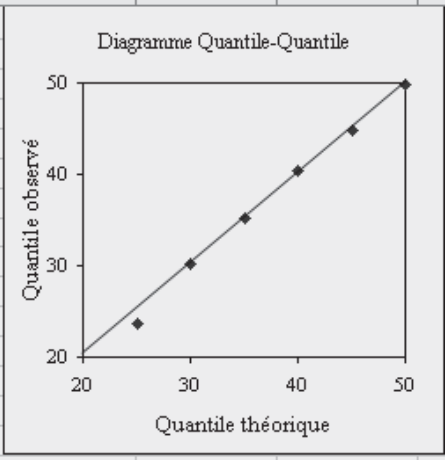
Exercice 7.7

1. Pour une variable statistique continue, on calcule une valeur approchée de la médiane par interpolation linéaire : $Me \in [35 ; 40[\quad \Rightarrow \quad Me = 35 + 5 \cdot \frac{100 - 86}{40} = 36{,}75$

2.

E2 : =37+10*D2

	A	B	C	D	E
1	x_i	f_i	F_i	u_i	x_i*
2	25	9	0,09	-1,34	23,59
3	30	16	0,25	-0,67	30,26
4	35	18	0,43	-0,18	35,24
5	40	20	0,63	0,33	40,32
6	45	15	0,78	0,77	44,72
7	50	12	0,9	1,28	49,82



La loi normale $\mathcal{N}(37\ ;\ 10)$ est adaptée puisque le nuage des points (x_i^*, x_i) est approximativement aligné le long de la première bissectrice.

3.1. $\overline{X_n}$ représente la durée moyenne des n interviews

$\sum_{i=1}^{n} X_i \ \rightarrow\ \mathcal{N}(n \cdot m,\ \sigma\sqrt{n})$ puisque les v.a. X_i sont *iid* à X

(*iid* pour « indépendantes et identiquement distribuées »)

$$\Rightarrow \quad \overline{X_n} \ \rightarrow\ \mathcal{N}\left(37\ ;\ \frac{10}{\sqrt{n}}\right)$$

3.2. $P(\overline{X_6} \leq 35) = F_U(-0{,}2\sqrt{6}) = F_U(-0{,}49) = 1 - 0{,}6879 = 0{,}3121$

3.3. $P(\overline{X_n} \leq 45) = F_U(8\sqrt{n}/10) \geq 0{,}99 \quad \Rightarrow \quad 0{,}8\sqrt{n} \geq 2{,}3263$

$$\Rightarrow \quad n \geq (2{,}3263/0{,}8)^2 = 8{,}46 \quad \Rightarrow \quad n \geq 9$$

Exercice 7.8

1.1. Soit Y le nombre d'actions en hausse parmi les 10 actions

$$X_i = \begin{cases} 1 & \text{si } i^e \text{ action en hausee} \quad p = 0{,}7 \\ 0 & \text{sinon} \quad q = 0{,}3 \end{cases}$$

$\Rightarrow$ Y est somme de 10 v.a. de Bernoulli indépendantes de même paramètre $p = 0{,}7$

$$\Rightarrow \quad Y = \sum_{i=1}^{10} X_i \rightarrow \mathcal{B}(10\ ;\ 0{,}7) \quad \Rightarrow \quad Z = 10 - Y \ \rightarrow\ \mathcal{B}(10\ ;\ 0{,}3)$$

1.2. $P(Y \geq 8) = P(Z \leq 2) = 0{,}3828 \qquad P(Y < 4) = P(Z > 6) = 1 - 0{,}9894 = 0{,}0106$

2. $Y = \sum_{i=1}^{100} X_i \rightarrow \mathcal{B}(100\ ;\ 0{,}7) \xrightarrow[npq = 21 > 18]{} \mathcal{N}(70\ ;\ \sqrt{21})$

$$P(Y \geq 80) = 1 - P(Y < 80) = 1 - F_U\left(\frac{80 - 0{,}5 - 70}{\sqrt{21}}\right) = 1 - F_U(2{,}07) = 1 - 0{,}9808 = 0{,}192$$

$$P(Y < 40) = F_U\left(\frac{40 - 0{,}5 - 70}{\sqrt{21}}\right) = F_U(-6{,}65) \approx 0$$

Exercice 7.9

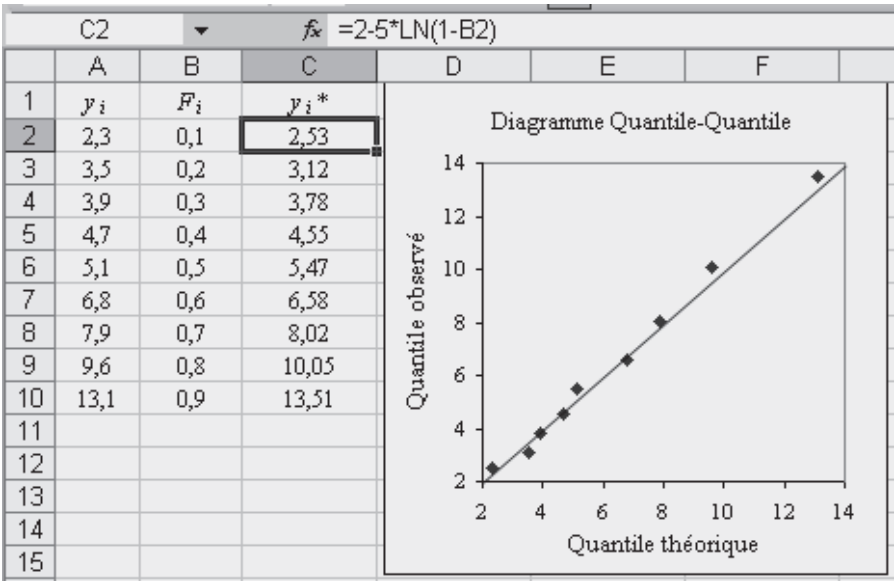
1. X suit une loi exponentielle de paramètre 1 : $E(X) = 1$ et $\text{var}(X) = 1$ (*cf.* cours)
$E(Y) = 2 + \lambda \cdot E(X) = 2 + \lambda \qquad \text{var}(Y) = \lambda^2 \cdot \text{var}(X) = \lambda^2$

$$F_Y(y) = P(Y \leq y) = P\left(X \leq \frac{y-2}{\lambda}\right) = \begin{cases} 1 - e^{-(y-2)/\lambda} & \text{si } y \geq 2 \\ 0 & \text{si } y < 2 \end{cases}$$

La v.a. Y suit en fait une loi exponentielle de paramètres λ et $\theta = 2$.

2.1. $P(Y < 2) = F_Y(2) = 0$
$P(2 \leq Y \leq 5) = F_Y(5) - F_Y(2) = 1 - e^{-3/5} \approx 0{,}45$

2.2. $y_i^* = 2 - 5 \ln(1 - F_i)$



	A	B	C
1	y_i	F_i	y_i^*
2	2,3	0,1	2,53
3	3,5	0,2	3,12
4	3,9	0,3	3,78
5	4,7	0,4	4,55
6	5,1	0,5	5,47
7	6,8	0,6	6,58
8	7,9	0,7	8,02
9	9,6	0,8	10,05
10	13,1	0,9	13,51
11			
12			
13			
14			
15			

La loi exponentielle de paramètres $\lambda = 5$ et $\theta = 2$ est adaptée puisque le nuage des points $(y_i^*,\ y_i)$ est approximativement aligné le long de la première bissectrice.

Exercice 7.10

1. $\bar{x} = 28{,}1 \qquad s_X = 9{,}375$

2. $Me = 25 + 5 \cdot \dfrac{50 - 39}{28} = 26{,}96$

3.1. $P(18 < X < 35) = F_U\left(\dfrac{35 - 28}{9{,}5}\right) - F_U\left(\dfrac{18 - 28}{9{,}5}\right)$

$= F_U(0{,}74) - F_U(-1{,}05) = 0{,}7704 - 1 + 0{,}8531 = 62{,}35\ \%$

À partir de l'échantillon : $2 \cdot 0{,}13/5 + 0{,}22 + 0{,}28 + 0{,}15 = 70{,}2\ \%$

3.2. $P(X > A) = 0{,}75 \quad \Rightarrow \quad P(X \leq A) = 0{,}25 \quad \Rightarrow \quad \dfrac{A - 28}{9{,}5} = u_{25\,\%} = -0{,}675$

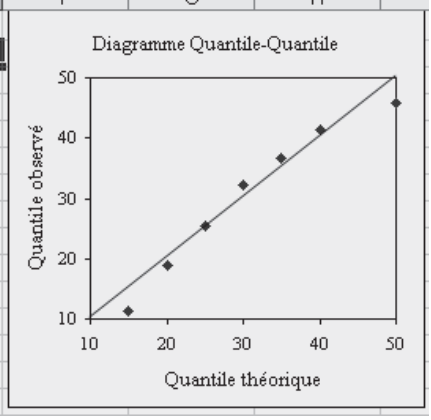
$A = 28 - 9{,}5 \cdot 0{,}675 = 28 - 6{,}4125 \approx 21{,}6$

$A' = Q_1 = 20 + 5 \cdot \dfrac{25 - 17}{22} = 21{,}82$ A' est le premier quartile

4.

E2 : =28+9,5*D2

	A	B	C	D	E
1	x_i	f_i	F_i	u_i	x_i *
2	15	0,04	0,04	-1,75	11,37
3	20	0,13	0,17	-0,95	18,94
4	25	0,22	0,39	-0,28	25,35
5	30	0,28	0,67	0,44	32,18
6	35	0,15	0,82	0,92	36,70
7	40	0,1	0,92	1,41	41,35
8	50	0,05	0,97	1,88	45,87



L'ensemble n'est pas trop éloigné de la bissectrice
⇒ il n'y a pas lieu de remettre l'ajustement en cause.

Exercice 7.11

1. $F_T(t) = \int_{-\infty}^{t} f(u)du = \begin{cases} 0 & \text{si } t < 0 \\ \frac{1}{5}\int_0^t e^{-u/5}du = -[e^{-u/5}]_0^t = 1 - e^{-t/5} & \text{si } t \geq 0 \end{cases}$

2. $P(T > 8) = 1 - F_T(8) = e^{-8/5} = e^{-1,6} \approx 0,202$

3.1. $X_i = \begin{cases} 1 & \text{si temps d'attente} \leq 8 \text{ min} \quad p \approx 0,8 \\ 0 & \text{sinon} \quad q \approx 0,2 \end{cases}$

⇒ Y est une somme de n v.a. indépendantes de Bernoulli de même paramètre $p = 0,8$

$\Rightarrow Y = \sum_{i=1}^{n} X_i \rightarrow \mathcal{B}(n\,;\,0,8)$

3.2. $E(Y) = 0,8 \cdot n \qquad \text{var}(Y) = 0,16 \cdot n$

4.1.

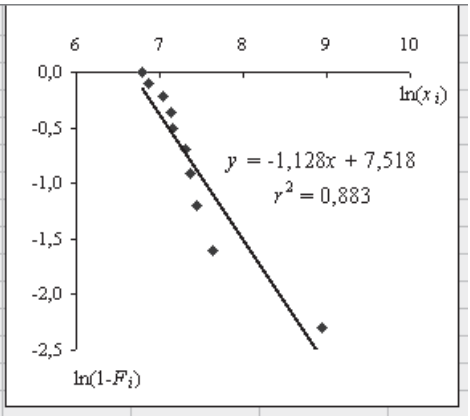
Y	0	…	k	…	n
$Z = Y + n$	n	…	$n + k$	…	$2n$
Probabilité	$0{,}2^n$		$\binom{n}{k} 0{,}8^n\, 0{,}2^{n-k}$		$0{,}8^n$

4.2. $E(Z) = E(Y + n) = 1{,}8 \cdot n$ $\quad \text{var}(Z) = \text{var}(Y + n) = \text{var}(Y) = 0{,}16 \cdot n$

Exercice 7.12

1.

x_i	F_i	$\ln(x_i)$	$1-F_i$	$\ln(1-F_i)$
7 658	0,9	8,944	0,1	-2,303
2 087	0,8	7,643	0,2	-1,609
1 691	0,7	7,433	0,3	-1,204
1 585	0,6	7,368	0,4	-0,916
1 505	0,5	7,317	0,5	-0,693
1 291	0,4	7,163	0,6	-0,511
1 266	0,3	7,144	0,7	-0,357
1 134	0,2	7,034	0,8	-0,223
953	0,1	6,860	0,9	-0,105
891	0,0	6,792	1	0,000



Calcul du coefficient de corrélation linéaire et des coefficients de la droite des moindres carrés avec Excel® :
r = Coefficient corrélation (C2:C11;E2:E1)
$\hat{a}$ = Index(Droitereg(E2:E11; C2:C11);1)
$\hat{b}$ = Index(Droitereg(E2:E11; C2:C11);2)

$r = -0{,}94 \quad \Rightarrow \quad r$ étant voisin de 1, on peut considérer les 10 points approximativement alignés : $\ln(1 - F_i) \approx -1{,}128 \cdot \ln(x_i) + 7{,}518$

Pour tracer avec Excel la droite des moindres carrés : onglet « Graphique », « Ajouter une courbe de tendance », type « Linéaire ». On peut utiliser ensuite l'onglet « Options » pour « Afficher l'équation sur le graphique » et pour « Afficher le coefficient de détermination (R^2) sur le graphique ».

2. La fonction de répartition d'une loi de Pareto est fonction de 2 paramètres α et x_0 :

$$F_X(x) = 1 - \left(\frac{x_0}{x}\right)^{\alpha} \quad \Rightarrow \quad 1 - F_X(x) = \left(\frac{x_0}{x}\right)^{\alpha} \quad \Rightarrow \quad \ln(1 - F_X(x)) = \alpha \cdot (\ln(x_0) - \ln(x))$$

Les points $\{\ln(x_i) ; \ln(1 - F_i)\}$ étant quasi-alignés ($r = -0{,}94$), l'ajustement de la distribution par une loi de Pareto est justifié, et on peut évaluer ses paramètres :
$\alpha = 1{,}128 \quad \ln(x_0) \approx 6{,}667 \quad \Rightarrow \quad x_0 = e^{6{,}667} \approx 786$

Exercice 7.13

1. $\hat{y}_t = 1{,}3 \cdot t^2 + 135{,}5$

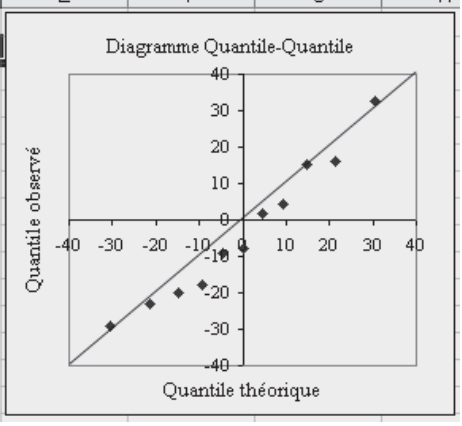
t^2	Nombre de contrats souscrits	$\hat{y}_t$	e_t
1	117	136,8	– 19,8
4	178	140,7	37,3
9	149	147,2	1,8
16	189	156,3	32,7
25	145	168,0	– 23,0
36	173	182,3	– 9,3
49	170	199,3	– 29,3
64	223	218,8	4,2
81	223	240,9	– 17,9
100	281	265,7	15,3
121	285	293,0	– 8,0
144	339	323,0	16,0

Résidus croissants
– 29,3
– 23,0
– 19,8
– 17,9
– 9,3
– 8,0
1,8
4,2
15,3
16,0
32,7
37,3

2.

D2 : =22*C2

	A	B	C	D
1	x_i	F_i (%)	u_i	x_i *
2	-29,3	8,3	-1,3832	-30,4
3	-23,0	16,7	-0,9673	-21,3
4	-19,8	25,0	-0,6745	-14,8
5	-17,9	33,3	-0,4308	-9,5
6	-9,3	41,7	-0,2103	-4,6
7	-8,0	50,0	0,0000	0,0
8	1,8	58,3	0,2103	4,6
9	4,2	66,7	0,4308	9,5
10	15,3	75,0	0,6745	14,8
11	16,0	83,3	0,9673	21,3
12	32,7	91,7	1,3832	30,4
13				
14				
15				



Les points étant peu éloignés de la bissectrice, on ne rejette pas l'ajustement par la loi normale $\mathcal{N}$ (0 ; 22).

Exercice 7.14

1. $P(|X - m| < 10) = P(-10 < X - m < 10) = F_U(10/25) - F_U(-10/25)$

$= (2F_U(0{,}4) - 1) = 2 \cdot 0{,}6554 - 1 = 0{,}3108$

2. $\bar{X}_{25}$ = teneur moyenne en sucre des 25 bouteilles

$\sum_{i=1}^{25} X_i \to \mathcal{N}(1\,750 ; 125)$ puisque les X_i sont *iid* à X de distribution $\mathcal{N}(70 ; 25)$

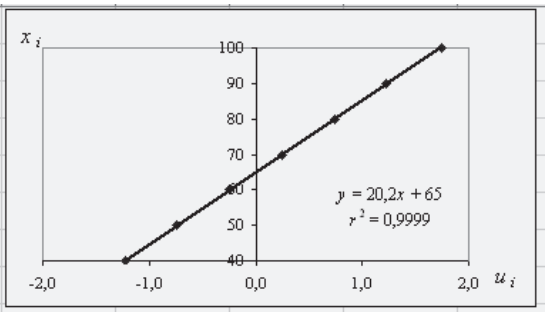
(*iid* pour « indépendantes et identiquement distribuées »)

$$E(\overline{X_{25}}) = E(X) \qquad \text{var}(\overline{X_{25}}) = \text{var}(X)/25 = 25$$

(*cf.* loi de la v.a. appelée moyenne empirique, chapitre 7, § III.A).

$\Rightarrow \overline{X}_{25} \to \mathcal{N}(70 ; 5)$ $\quad P(|\overline{X} - m| < 10) = P(|\overline{X} - m| < 2\,\sigma_{\overline{X}}) = 0{,}9544$

Teneur (g) : x_i	*F_i (%)*	*u_i*
40	11	-1,2265
50	23	-0,7388
60	40	-0,2533
70	60	0,2533
80	77	0,7388
90	89	1,2265
100	96	1,7507



Les points (u_i , x_i) étant alignés, on a : $x_i = au_i + b$

Les u_i étant les fractiles d'une loi normale centrée-réduite, on a : $u_i = x_i - m/\sigma$

Les paramètres de la droite des moindres carrés sont donc des évaluations de la moyenne et de l'écart-type : $m^* \approx 65$ g et $\sigma^* \approx 20$ g

Exercice 7.15

1. Profondeur $(Me) = 50{,}5 \Rightarrow Me = 127$

Profondeur $(Q) = 25{,}5 \Rightarrow Q_1 = 107 \quad Q_3 = (144 + 146)/2 = 145$

Trois indicateurs de tendance centrale :

Moyenne = 124,6 $\quad Me = 127 \quad (Q_1 + Q_3)/2 = 126$

Deux indicateurs de dispersion :

$s_X = 32 \qquad EIQ = 38$

2.1 $P(m - \sigma \le X \le m + \sigma) = F_U(1) - F_U(-1) = 0{,}6826$

$P(m - 2\sigma \le X \le m + 2\sigma) = F_U(2) - F_U(-2) = 0{,}9544$

2.2

$$P(X < x_1) = 0{,}1 \Rightarrow F_U\left(\frac{x_1 - 125}{30}\right) = 0{,}1 \Rightarrow x_1 - 125 = -30 \cdot 1{,}2816 \Rightarrow x_1 = 86{,}552$$

$F_U(u_i)$	0,1	0,2	0,3	0,4	0,5
u_i	– 1,2816	– 0,8416	– 0,5244	– 0,2533	0
$x_i = 30u_i + 125$	86,6	99,8	109,3	117,4	125,0

$F_U(u_i)$	0,6	0,7	0,8	0,9
u_i	2,2533	0,5244	0,8416	1,2816
$x_i = 30u_i + 125$	192,6	140,7	150,2	163,4

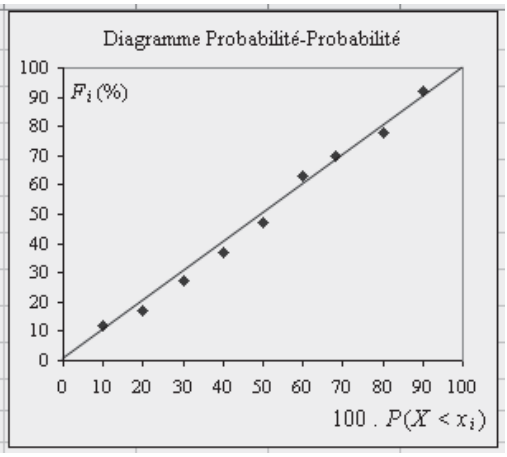
3.

Valeur	$< x_1$	$[x_1 ; x_2[$	$[x_2 ; x_3[$	$[x_3 ; x_4[$	$[x_4 ; x_5[$
Effectif	12	5	10	10	10

Valeur	$[x_5 ; x_6[$	$[x_6 ; x_7[$	$[x_7 ; x_8[$	$[x_8 ; x_9[$	$\geq x_9$
Effectif	16	5	10	14	8

4.

Valeur	$100 \cdot P(X < x_i)$	F_i (%)
x_1	10	12
x_2	20	17
x_3	30	27
x_4	40	37
x_5	50	47
x_6	60	63
x_7	68	70
x_8	80	78
x_9	90	92



On a construit un diagramme Probabilité-Probabilité qui permet de comparer les probabilités cumulées théoriques aux fréquences cumulées.

Les points sont quasi alignés sur la bissectrice, les pourcentages cumulés théoriques et observés sont très proches, l'ajustement de la distribution observée par la loi normale $\mathcal{N}$ (125 ; 30) est retenu.

Annexes

I. Formulaire élémentaire de combinatoire

Sous le nom de combinatoire, on regroupe ici les résultats essentiels de dénombrement sur les ensembles.

A. Ensemble des parties d'un ensemble

Soit $\mathbb{A}$ un ensemble de N éléments. L'ensemble $\mathcal{P}(\mathbb{A})$ des parties de $\mathbb{A}$, comporte 2^N éléments.

B. Arrangements avec répétition

On s'intéresse à un ensemble $\mathbb{A}$ de N éléments, dans lequel on sélectionne k individus, chacun pouvant être choisi plusieurs fois (tirages avec répétition).

Le nombre de sélections possibles de k individus de $\mathbb{A}$, par un tel procédé (ou encore d'arrangements avec répétition) est de :

$$N^k$$

En effet, pour le premier individu on a N choix possibles. Chacun de ces choix est associé à n'importe lequel des N choix possibles pour le second. On continue ainsi jusqu'au choix du dernier (N possibilités également).

C'est par exemple le cas, pour le nombre de résultats possibles pour une suite de N épreuves identiques ayant chacune les mêmes k résultats élémentaires possibles.

C. Permutations

Soit $\mathbb{A}$ un ensemble de N éléments, on appelle permutation sur $\mathbb{A}$ une suite de N éléments de $\mathbb{A}$. Ceci revient à dire que l'on a disposé N objets de $\mathbb{A}$ dans un ordre déterminé. Il faut remarquer que dans cette définition générale, les objets peuvent ne pas être distincts. Pour cette raison, on introduit la notion de permutation sans répétition, dans laquelle les éléments de $\mathbb{A}$ sont distincts. Cette dernière définition revient donc à dire qu'une permutation (sans répétition) de $\mathbb{A}$ est un rangement particulier de ses éléments.

Pour un ensemble $\mathbb{A}$ à N éléments il existe $N \cdot (N-1) \cdot (N-2) \cdot \ldots \cdot 2 \cdot 1$ permutations sans répétitions distinctes. Ce résultat est simple à montrer par récurrence. La valeur ainsi obtenue est appelée factorielle N, et elle est notée

$$N!$$

Plus généralement, si N_1, N_2, N_k sont les nombres de résultats distincts qui peuvent être obtenus sur k épreuves successives, alors le nombre de résultats distincts possibles à l'issue de la série des k épreuves est donné par :

$$N_1 \cdot N_2 \cdot \ldots \cdot N_k$$

On peut montrer aisément qu'il existe N^N permutations avec répétitions de N éléments.

Dans ce qui suit on ne parlera que de permutation sans répétition.

D. Arrangements sans répétition

On se place donc encore dans le cas d'un ensemble $\mathbb{A}$ à N éléments distincts. Le nombre de choix possibles ordonnés de k objets de $\mathbb{A}$ est appelé nombre d'arrangements sans répétition de k objets parmi N, et est désigné par A_N^k. On a :

$$A_N^k = \frac{N!}{(N-k)!} = N \cdot (N-1)(N-2) \cdot \ldots \cdot (N-k+1)$$

En effet, après avoir choisi le premier élément (N choix possibles), il ne reste plus que $(N-1)$ choix possibles pour le second. Chaque choix du premier peut être associé à n'importe lequel des choix du second, et on a $N(N-1)$ choix possibles pour les 2 premiers éléments sélectionnés. En poursuivant le raisonnement, on obtient le résultat annoncé.

Soit ainsi une tombola dotée de 4 prix, pour laquelle 20 billets ont été émis et tous vendus. Le nombre de résultats posssibles correspond alors au nombre de choix possible de 4 individus (les 4 gagnants) parmi 20 (les 20 possesseurs de billets), et l'ordre des gagnants est ici important puisque les prix sont distincts et de valeurs en général très différentes. On a

$$A_{20}^{4} = \frac{20!}{16!} = 17 \cdot 18 \cdot 19 \cdot 20 = 116\,280$$

situations différentes observables à l'issue du tirage de la tombola.

E. Combinaisons sans répétition

Dans le cas précédent, l'ordre dans lequel se trouvent les k individus sélectionnés dans l'ensemble $\mathbb{A}$ est important, et il convenait de distinguer deux sélections dans lesquelles les individus tirés seraient les mêmes mais ne seraient pas affectés aux mêmes positions (ou rangs de tirage).

Nous considérons souvent aussi des cas où cet ordre n'a pas de signification précise. Pour un ensemble $\mathbb{A}$ de N éléments dans lequel on sélectionne k individus sans répétition sans tenir compte de l'ordre, on désigne alors le nombre de choix possibles par $\binom{N}{k}$, qu'on appelle nombre de combinaisons de N individus pris k à k.

On sait que tous les choix résultant aux mêmes k individus donneront donc une seule combinaison de k éléments pris parmi les N de $\mathbb{A}$. Tous ces choix sont les permutations des k éléments, et il en existe $k!$

Il en résulte que le nombre de combinaisons $\binom{N}{k}$ est égal au nombre d'arrangements A_N^k divisé par $k!$:

$$\binom{N}{k} = \frac{A_N^k}{k!} = \frac{N!}{k!(N-k)!}$$

Dans l'exemple précédent de la tombola à 20 billets vendus et 4 prix, si les prix étaient identiques, on parlerait de

$$\binom{20}{4} = \frac{20!}{4! \cdot 16!} = 4\,845$$

résultats possibles à l'issue du tirage (soit 4 ! = 24 fois moins que si les prix sont différents).

Mentionnons quelques formules très utilisées relatives à ce nombre de combinaisons :

$$\binom{N}{0} = 1 \qquad \binom{N}{1} = N \qquad \binom{N}{k} = \binom{N}{N-k}$$

ainsi que :

$$\binom{N}{k} = \binom{N-1}{k} + \binom{N-1}{k-1}$$

dont l'application très connue est le triangle de Pascal qui est un tableau triangulaire dont chaque ligne correspond à une valeur de N et chaque colonne à une valeur de k ($k = 0, 1, 2, \ldots, N$). Sa construction se déduit de la formule ci-dessus. Ainsi pour les 8 premières lignes :

1

1 1

1 2 = 1 + 1 1

1 3 = 2 + 1 3 = 1 + 2 1

1 4 = 3 + 1 6 = 3 + 3 4 = 1 + 3 1

1 5 = 4 + 1 10 = 6 + 4 10 = 4 + 6 5 = 1 + 4 1

1 6 = 5 + 1 15 = 10 + 5 20 = 10 + 10 15 = 5 + 10 6 = 1 + 5 1

1 7 = 6 + 1 21 = 15 + 6 35 = 20 + 15 35 = 15 + 20 21 = 6 + 15 7 = 1 + 6

en se rappelant que les extrêmes de chaque ligne sont toujours égaux à 1.

Citons aussi la formule bien connue du binôme de Newton :

$$(a+b)^N = \sum_{k=1}^{N} \binom{N}{k} \cdot a^k \cdot b^{N-k}$$

dont un cas particulier est :

$$2^N = \sum_{k=1}^{N} \binom{N}{k}$$

obtenu en prenant $a = 1$ et $b = 1$.

Enfin, on citera une formule de récurrence permettant de calculer les nombres de combinaisons par quelques pas de programme :

$$\binom{N}{k} = \frac{N-k+1}{k} \cdot \binom{N}{k-1}$$

On retrouve son application à l'annexe II.

F. Coefficients multinomiaux

Le nombre total de différents partages d'un ensemble à N éléments en k sous-ensembles disjoints, contenant respectivements $n_1, n_2, \ldots, n_k$ éléments est donné par le coefficient multinomial :

$$\frac{N!}{n_1! \cdot n_2! \cdot \ldots \cdot n_k!}$$

C'est une généralisation du nombre de combinaisons, $\binom{N}{k}$, où l'on partageait l'ensemble $\mathbb{A}$ à N éléments en deux sous-ensembles, l'un de k éléments et l'autre des $(N - k)$ éléments restants.

II. Principaux modèles de probabilités : méthodes de calculs

A. Loi binomiale

La formule donnant les probabilités individuelles d'une loi $\mathcal{B}(n, p)$ permet de construire une procédure itérative. En effet, si X suit une telle loi :

$$\begin{aligned}
P(X = k+1) &= \binom{n}{k+1} p^{k+1}(1-p)^{n-k-1} \\
&= \frac{n!}{(k+1)(n-k-1)!} p^{k+1}(1-p)^{n-k-1} \\
&= \frac{n!(n-k)}{k!(k+1)(n-k)!} p^k p \frac{(1-p)^{n-k}}{(1-p)} \\
&= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \frac{(n-k)p}{(k+1)(1-p)} \\
&= \frac{(n-k)}{(k+1)} \frac{p}{(1-p)} P(X = k)
\end{aligned}$$

On écrira donc très facilement, à partir de la valeur $P(X = 0) = (1 - p)^n$, toutes les probabilités individuelles d'une loi binomiale, en multipliant la précédente par

$$\frac{(n-k)p}{(k+1)(1-p)}$$

Partant d'une somme égale à $P(X = 0)$, on obtiendra les probabilités cumulées en ajoutant à chaque fois la nouvelle probabilité individuelle à la somme de l'étape précédente.

B. Loi de Poisson

Ici encore, la formule des probabilités individuelles permet d'écrire une formule de calcul itératif (formule déjà présentée et utilisée au chapitre 6, § III). Si Y suit une loi $\mathcal{P}(m)$, on a la formule :

$$P(Y = k + 1) = \frac{m}{(k+1)} P(Y = k)$$

qui permet de programmer le calcul des probabilités individuelles, partant de $P(Y = 0) = \mathrm{e}^{-m}$. La programmation du calcul des probabilités cumulées se fait comme pour celles de la loi binomiale, en additionnant la nouvelle probabilité individuelle à la somme de l'étape précédente, après avoir débuté la somme par $P(Y = 0)$.

C. Loi de Gauss centrée réduite

Le calcul de valeurs de la fonction de répartition peut s'effectuer à l'aide de formules approchées. Les plus utiles sont les formules de Hastings. Simples à programmer, elles permettent d'obtenir la valeur de la fonction cumulative en un point $u > 0$. Pour la valeur de $F_U(u)$ en un point $u < 0$, on se sert de l'égalité $F_U(u) = 1 - F_U(u)$. Nous donnons deux formules. La seconde est plus simple, mais un peu moins précise que la première.

a) $$F_U(u) \approx 1 - (a_1 z + a_2 z^2 + a_3 z^3) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{u^2}{2}}$$ dans laquelle :

$$a_1 = 0{,}4361836 \qquad a_2 = 0{,}1201676 \qquad a_3 = 0{,}9372980$$

$$z = \frac{1}{1 + 0{,}33267u}$$

L'erreur commise est au plus égale à 0,00001.

b)

$$F_U(u) \approx 1 - \frac{1}{2}(1 + 0{,}196854u + 0{,}115194u^2 + 0{,}000344u^3 + 0{,}019527u^4)^{-4}$$

L'erreur commise est au plus égale à 0,00025.

D. Loi du khi-deux

On utilise la formule, vue au chapitre 7, § III.A, de Wilson-Hilferty :

$$P(\chi^2(n) < x) \approx F_U\left(\left(\left(\frac{x}{n}\right)^{1/3} - 1 + \frac{2}{9n}\right)\sqrt{\frac{9n}{2}}\right)$$

la détermination de la valeur de la fonction de répartition de la loi de Gauss centrée réduite se faisant par l'une des formules données ci-dessus.

E. Loi de Student

On a vu au chapitre 7, III.B, que la loi de Student à 2 ddl possède une fonction de répartition simple permettant des calculs exacts. Dans le cas général, on utilise deux formules d'approximation ; l'une pour la fonction de répartition, l'autre pour les fractiles. La première formule est due à Fisher et s'écrit, pour $t > 0$ (pour $t < 0$, on utilise la symétrie) :

$$P(T_\nu < t) = F_U(t) - \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\left(\frac{1}{4}t(t^2+1)\frac{1}{\nu} + \frac{1}{96}t(3t^6 - 7t^4 - 5t^2 - 3)\frac{1}{\nu^2}\right.$$

$$\left. + \frac{1}{384}t(t^{10} - 11t^8 + 14t^6 + 6t^4 - 3t^2 - 15)\frac{1}{\nu^3}\right)$$

L'erreur commise est au plus égale à 0,000005. C'est donc une excellente approximation.

Dans le cas particulier de la loi de Student à 1 ddl, on peut utiliser la formule suivante qui donne un résultat entaché d'une erreur au plus égale à 0,001, mais qui ne nécessite pas de calcul de fonction de répartition de la loi normale.

$$\begin{cases} P(0 < T_1 < t) \approx \dfrac{1}{\pi}\left(t - \dfrac{t^3}{3} + \dfrac{t^5}{5} - \dfrac{t^7}{7}\right) & \text{si } 0 \le t \le 0{,}5 \\ P(0 < T_1 < t) \approx \dfrac{1}{4} + \dfrac{1}{\pi}\left(\dfrac{1}{2}(t-1)^2 - \dfrac{1}{4}(t-1)^2 + \dfrac{1}{12}(t-1)^3 - \left(\dfrac{1}{40}(t-1)^5\right)\right. & \\ & \text{si } 0{,}5 \le t \le 1 \\ P(0 < T_1 < t) \approx \dfrac{1}{2} + \dfrac{1}{\pi}\left(\dfrac{1}{t} - \dfrac{1}{3t^3} + \dfrac{1}{5t^5} - \dfrac{1}{7t^7}\right) & \text{si } t \ge 1{,}5 \end{cases}$$

La formule suivante permet une approximation des fractiles de la loi de Student à partir de ceux de la loi de Gauss centrée réduite. Elle est due à Fisher et Cornish :

$$t_\alpha(\nu) \approx u_\alpha + \frac{1}{4}u_\alpha(u_\alpha^2 + 1)\frac{1}{\nu} + \frac{1}{96}u_\alpha(5u_\alpha^4 + 16u_\alpha^2 + 3)\frac{1}{\nu^2} + \frac{1}{384}u_\alpha(3u_\alpha^6 + 19u_\alpha^4 + 17u_\alpha^2 - 15)\frac{1}{\nu^3}$$

F. Loi de Fisher-Snedecor

Pour cette loi, on peut utiliser la formule suivante, due à G.W. Cochran, et qui donne les fractiles d'une loi $F(\nu_1, \nu_2)$ en fonction des fractiles de la loi normale centrée réduite :

$$f_\alpha(\nu_1, \nu_2) \approx d\left(1 + \frac{1}{3}(u_\alpha^2 - 1)\right) + u_\alpha c\left(1 - \frac{c^2}{6}(u_\alpha^2 + 3)\right)^{-1/2}$$

où et $\quad d = \dfrac{1}{2}\left(\dfrac{1}{\nu_2} - \dfrac{1}{\nu_1}\right) \qquad c = \dfrac{1}{2}\left(\dfrac{1}{\nu_1} + \dfrac{1}{\nu_2}\right)$

Cette formule est en fait dérivée de la formule de Wilson-Hilferty pour chaque loi de khi-deux au numérateur et au dénominateur de la loi de Fisher-Snedecor. On ne connait pas précisément de borne supérieure de l'erreur commise avec cette approximation, même si on peut considérer que cette approximation est de bonne qualité.

III. Introduction à la simulation des lois de probabilité

A. La place des méthodes de simulation

Les activités économiques sont tributaires de contraintes et d'influences complexes, sources de variation importantes sur la ou les grandeurs étudiées. Dans certains cas, on peut obtenir une solution analytique au fonctionnement d'un système complexe, mais le plus souvent il est nécessaire de recourir à l'étude de scénarii sous la forme d'une analyse de sensibilité, ou mieux encore à la simulation lorsque la partie aléatoire peut être décrite par des distributions de probabilités. Ainsi, dans une modélisation de flux, la prise en compte des interactions de toutes sortes génère des modèles mathématiques délicats, de même que pour l'établissement de valeurs (*pricing*), les calculs sont basés sur des modèles aléatoires dont la résolution complète n'est pas toujours nécessaire et/ou possible.

Toutes les méthodes scientifiques de gestion ont bénéficié de l'explosion des ressources de calcul des ordinateurs, qui ont donné l'occasion d'une large diffusion des méthodes dites « intensives » comme la simulation. Les tableurs actuels sont tous munis d'un générateur de nombres « pseudo-aléatoires » de qualité suffisante pour la plupart des besoins courants. Avec la mise à disposition d'une bibliothèque de fonctions (mathématiques, statistiques, logiques, etc.), les méthodes de simulation sont devenues un ensemble d'outils d'aide à la décision très largement accessible et répandu. Quelques bibliothèques de programmes (comme le logiciel **R**) organisées autour d'un langage très simple complètent les instruments de base.

B. Les principes de la simulation sur tableur

Simuler est une façon d'imiter. Simuler le comportement d'un système complexe consiste à en reconstituer fictivement des réalisations. On parle de simulation aléatoire lorsque celles-ci sont obtenues à l'aide de réalisations fictives de variables aléatoires de distributions connues.

Pour toute simulation, on part de réalisations « artificielles » de la distribution uniforme continue sur l'intervalle]0 ; 1[qui sont « fabriquées » (simulées) au travers de la fonction ALEA(). L'appel de cette fonction dans K cellules d'une feuille de tableur permet d'obtenir $\{x_i, i = 1,\ldots,K\}$, K réalisations indépendantes de la distribution uniforme continue sur]0 ; 1[

Pour obtenir des valeurs simulées de la distribution de Bernoulli de paramètre p on utilise la fonction logique :

SI ALEA() < p

en affectant la valeur 1 comme résultat lorsque la condition est réalisée, et la valeur 0 sinon.

En effet, la probabilité d'avoir un résultat de loi uniforme continue sur]0 ; 1[inférieur à p est égale à p (chapitre 7, § I.A).

C. Simulation de lois discrètes

On peut obtenir une réalisation simulée d'une loi binomiale $\mathcal{B}(n ; p)$ en simulant n réalisations de lois de Bernoulli de paramètre p comme on vient de voir, et en faisant la somme des résultats puisqu'une variable binomiale $\mathcal{B}(n ; p)$ est une somme de n variables de Bernoulli indépendantes et de même paramètre p. On peut aussi simuler une distribution géométrique $\mathcal{G}(p)$ en simulant des réalisations de lois de Bernoulli de paramètre p jusqu'à l'obtention de la première valeur 1

La simulation de valeurs issues d'une distribution de Poisson demande une assez bonne pratique de l'utilisation d'un tableur (avec macros). Elle peut aussi être obtenue à partir des propriétés de certains modèles simples de files d'attente[1] ; la méthode est évoquée à propos de la simulation de la loi exponentielle dans le paragraphe suivant.

D. Simulations de lois continues

Pour obtenir des réalisations simulées d'une distribution continue lorsque sa fonction de répartition est inversible on utilise le résultat donné à la fin du paragraphe sur la distribution uniforme (chapitre 7, § I.A) : si X est une variable aléatoire continue dont la fonction de répartition F est bijective (donc inversible), alors la variable aléatoire $Y = F(X)$ a une distribution uniforme continue sur]0 ; 1[

En effet, X peut s'obtenir par $X = F^{-1}(Y)$ où Y est uniforme continue sur]0 ; 1[, donc pouvant être obtenue par l'appel à la fonction ALEA().

L'exemple de la distribution exponentielle est l'un des plus utilisés. La fonction de répartition de la distribution exponentielle de paramètres λ et θ est donnée par :

$$y = F(x) = \begin{cases} 0 & \text{si } x < \theta \\ 1 - \exp(-(x-\theta)/\lambda) & \text{si } x \geq \theta \end{cases}$$

Pour $x \geq \theta$ on a $x = \theta - \lambda \ln(1 - y)$

1. *Cf.* par exemple l'ouvrage de Dodge et Melfi en bibliographie.

Pour toute valeur de y, nombre pseudo-aléatoire généré par la fonction ALEA(), on obtient ainsi une valeur x d'une loi exponentielle par un calcul élémentaire.

La simulation de valeurs issues de distributions exponentielles permet de simuler des systèmes de files d'attente générés par l'accès aléatoire d'utilisateurs à une ressource partagée (un guichet par exemple) où les intervalles entre deux arrivées successives sont distribués selon une loi exponentielle et où les temps d'utilisation de la ressource sont aussi distribués selon une loi exponentielle (files notées M/M/C)[1]. L'étude de ces files d'attente (ici M/M/1) permet de montrer que le nombre d'arrivées par intervalle de temps fixe est aléatoire et distribué selon une loi de Poisson. On retrouve donc ici une possibilité de simuler des valeurs issues d'une loi de Poisson à partir de la simulation d'une file d'attente reposant sur des lois exponentielles.

Pour obtenir des réalisations simulées d'une distribution continue avec une fonction de répartition non inversible, on doit recourir à des méthodes plus élaborées, telles que la méthode d'acceptation-rejet. Elles ne sont pas présentées ici, mais nous donnerons simplement deux méthodes très utilisées pour simuler des valeurs de lois de Gauss.

Méthode 1 : faire la somme de 12 valeurs simulées de loi uniforme continue sur]0 ; 1[obtenues avec la fonction ALEA(). Par application du théorème central limite (chapitre 7, § II.E), la distribution de la somme de 12 variables uniformes continues sur]0 ; 1[peut être approximée par une loi de Gauss $\mathcal{N}(6 ; 1)$, et on obtient une loi de Gauss centrée réduite en retranchant 6 au résultat de cette somme de 12 valeurs de loi uniforme (il est conseillé de montrer ce résultat en exercice).

Méthode 2 (Box et Müller) : simuler deux valeurs indépendantes x_1 et x_2 de loi uniforme continue sur]0 ; 1[avec la fonction ALEA(). On obtient ensuite deux valeurs indépendantes de loi de Gauss centrée réduite en calculant :

$$u_1 = \sqrt{-2\ln x_1} \cdot \cos(2\pi x_2)$$

$$u_2 = \sqrt{-2\ln x_1} \cdot \sin(2\pi x_2)$$

Pour obtenir une valeur simulée y d'une loi de Gauss $\mathcal{N}(m ; \sigma)$, il suffit d'avoir une valeur simulée u d'une loi de Gauss centrée réduite (par une des méthodes précédentes par exemple) et calculer $y = \sigma \cdot u + m$

Les liens établis entre les différentes distributions continues montrent par exemple encore que pour obtenir une valeur simulée d'une distribution du khi-deux à 2 degrés de libertés, il suffit de se donner deux valeurs simulées indépendantes de loi de Gauss centrée réduite et de faire la somme de leurs carrés. Or, en appliquant la méthode de Box et Müller en partant des valeurs x_1 et x_2

1. La lettre M fait référence au caractère sans mémoire (*memoryless*) de cette distribution.

issues d'une loi uniforme continue sur]0 ; 1[, on obtient deux valeurs indépendantes u_1 et u_2 d'une loi de Gauss centrée réduite telles que $u_1^2 + u_2^2 = -2\ln x_1$

La fonction ALEA() génère donc une valeur d'une distribution du khi-deux à 2 degrés de liberté en calculant –2ln(ALEA())

Le lecteur peut trouver ensuite comment simuler facilement des valeurs d'une loi de khi-deux à nombre pair de degrés de liberté.

E. Quelques exemples et applications

1) Simulation d'une loi binomiale

La simulation de la loi binomiale est illustrée à partir du nombre de filles dans des classes de CP (ayant toutes 25 élèves), en supposant que la répartition des enfants à la naissance est de 48 % de filles et de 52 % de garçons (estimations démographiques classiques).

Dans ce cadre nous avons simulé (*cf.* tableau 1) 12 classes de 25 élèves avec le tableur Excel.

Tableau 1 – Simulation d'une loi $\mathcal{B}$ (25 ; 0,48)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	0,9242	0	0,2283	1	0,4891	0	0,0058	1	0,2202	1	0,4473	1	0,0810	1	0,9259	0	0,2284	1	0,6307	0	0,3345	1	0,6787	0
2	0,2875	1	0,4583	1	0,0993	1	0,5404	0	0,0881	1	0,8812	0	0,9741	0	0,9779	0	0,9222	0	0,0041	1	0,7683	0	0,7698	0
3	0,8741	0	0,3507	1	0,6390	0	0,1078	1	0,8462	0	0,5870	0	0,1550	1	0,2586	1	0,4659	1	0,7793	0	0,1172	1	0,3193	1
4	0,1213	1	0,4892	0	0,3855	1	0,2055	1	0,4115	1	0,8387	0	0,8322	0	0,2903	1	0,2292	1	0,4420	1	0,5185	0	0,2394	1
5	0,7902	0	0,5560	0	0,4861	0	0,2341	1	0,2967	1	0,7908	0	0,0474	1	0,1421	1	0,5657	0	0,7957	0	0,7424	0	0,8386	0
6	0,8602	0	0,8987	0	0,0078	1	0,1949	1	0,1418	1	0,0302	1	0,5001	0	0,2854	1	0,7903	0	0,5457	0	0,9242	0	0,8081	0
7	0,5260	0	0,8509	0	0,5675	0	0,2391	1	0,2530	1	0,4430	1	0,0304	1	0,9534	0	0,8289	0	0,0324	1	0,8888	0	0,3591	1
8	0,3723	1	0,6123	0	0,0455	1	0,1817	1	0,8110	0	0,3121	1	0,8116	0	0,0713	1	0,1215	1	0,6514	0	0,3452	1	0,8414	0
9	0,3513	1	0,8553	0	0,2781	1	0,6549	0	0,8854	0	0,3162	1	0,5301	0	0,8401	0	0,1977	1	0,0557	1	0,5319	0	0,1163	1
10	0,9045	0	0,3976	1	0,9561	0	0,1590	1	0,4950	0	0,6177	0	0,3638	1	0,0665	1	0,8854	0	0,5385	0	0,1413	1	0,6692	0
11	0,6732	0	0,8036	0	0,7759	0	0,6735	0	0,2284	1	0,0945	1	0,1499	1	0,4837	0	0,8576	0	0,0330	1	0,7825	0	0,5491	0
12	0,5125	0	0,6815	0	0,1597	1	0,3583	1	0,0843	1	0,5808	0	0,0626	1	0,7515	0	0,5961	0	0,0884	1	0,1943	1	0,1864	1
13	0,5035	0	0,6952	0	0,0102	1	0,2199	1	0,0486	1	0,4476	1	0,5840	0	0,1206	1	0,8853	0	0,9912	0	0,1401	1	0,6358	0
14	0,9403	0	0,5895	0	0,3775	1	0,4505	1	0,4045	1	0,1940	1	0,9394	0	0,0988	1	0,7774	0	0,2774	1	0,6775	0	0,0640	1
15	0,1236	1	0,7828	0	0,9459	0	0,7377	0	0,4477	1	0,6461	0	0,7694	0	0,8165	0	0,2045	1	0,2712	1	0,9518	0	0,4242	1
16	0,2934	1	0,9184	0	0,5787	0	0,6284	0	0,0896	1	0,6001	0	0,1020	1	0,1825	1	0,9937	0	0,8767	0	0,0048	1	0,7613	0
17	0,9486	0	0,1081	1	0,0515	1	0,6817	0	0,6019	0	0,5898	0	0,3254	1	0,2995	1	0,1440	1	0,6095	0	0,1056	1	0,9443	0
18	0,8300	0	0,7535	0	0,3809	1	0,9407	0	0,3988	1	0,5970	0	0,3167	0	0,0489	1	0,4532	1	0,4516	1	0,7852	0	0,8344	0
19	0,3592	1	0,9668	0	0,1325	1	0,9317	0	0,2403	1	0,2411	1	0,5923	0	0,8691	0	0,6210	0	0,2659	1	0,7387	0	0,3535	1
20	0,2783	1	0,3474	1	0,6083	0	0,5905	0	0,2725	1	0,6106	0	0,8066	0	0,6332	0	0,8509	0	0,7018	0	0,3174	1	0,5950	0
21	0,7101	0	0,3165	1	0,3756	1	0,6270	0	0,7297	0	0,5420	0	0,9278	0	0,9204	0	0,1143	1	0,4484	1	0,8613	0	0,9174	0
22	0,8166	0	0,6675	0	0,4445	1	0,3020	1	0,2901	1	0,6357	0	0,3826	1	0,2337	1	0,5382	0	0,7904	0	0,2369	1	0,0742	1
23	0,6671	0	0,1158	1	0,4151	1	0,8382	0	0,8867	0	0,4045	1	0,1575	1	0,1739	1	0,2974	1	0,9938	0	0,7118	0	0,2313	1
24	0,6032	0	0,6911	0	0,4749	1	0,4743	1	0,7193	0	0,4193	1	0,5022	0	0,7423	0	0,9551	0	0,5365	0	0,8884	0	0,1938	1
25	0,2138	1	0,4871	0	0,5489	0	0,0584	1	0,0695	1	0,0942	1	0,2580	1	0,2333	1	0,8252	0	0,5874	0	0,0381	1	0,2248	1
27		**9**		**8**		**15**		**14**		**17**		**12**		**12**		**14**		**10**		**11**		**11**		**12**
28																								
29			théorique		observé																			
30	moyenne		12		12,08																			
31	variance		6,24		6,08																			
32																								
33																								

=SOMME(H3:H27)

=ALEA()

=SI(O20<0,48;1;0)

Colonne A : valeurs de la fonction ALEA()

Colonne B : simulation de valeurs de loi de Bernoulli de paramètre 0,48 par la fonction SI(A1 < 0,48 ; 1 ; 0) dans la cellule B1, puis tirée vers le bas sur 25 lignes (croix en bas à droite de la cellule).

Colonnes C et D, E et F, G et H, I et J, K et L, M et N, O et P, Q et R, S et T, U et V, W et X remplies de manière similaire aux colonnes A et B.

Les colonnes B, D, F, H, J, L, N, P, R, T, V et X contiennent chacune une suite de 25 valeurs (0 ou 1) réalisations de naissances simulées, et modélisent chacune une classe de CP de 25 enfants dont on obtient le nombre de filles en faisant la somme de la colonne.

Ce nombre de filles est en théorie la somme de 25 aléas de Bernoulli indépendants de même paramètre 0,48 ; il est distribué selon une loi binomiale $\mathcal{B}$ (25 ; 0,48)

La moyenne théorique (25 · 0,48 =12) et la variance théorique (25 · 0,48 · 0,52 = 6,24) sont comparées à la moyenne et à la variance des valeurs simulées (lignes 30 et 31 du tableau1).

Il est aussi possible de simuler presque instantanément un jeu de pile ou face répété 5 000, 10 000 ou même 100 000 fois pour une pièce pipée ou pour une pièce non pipée, et d'observer la convergence des fréquences de pile vers la probabilité théorique imposée dans la simulation, ce qui illustre la loi des grands nombres.

2) Simulation d'une loi exponentielle

Le second exemple est celui des lois exponentielles. Dans les systèmes à file d'attente, une ressource en quantité disponible limitée (guichet, serveur informatique, imprimante, etc.) est soumise à des demandes qui peuvent excéder ses capacités de réponse instantanées. C'est bien entendu ce que chacun a déjà vécu et observé à la caisse d'un magasin, dans une station-service, aux guichets d'un service public, par exemple.

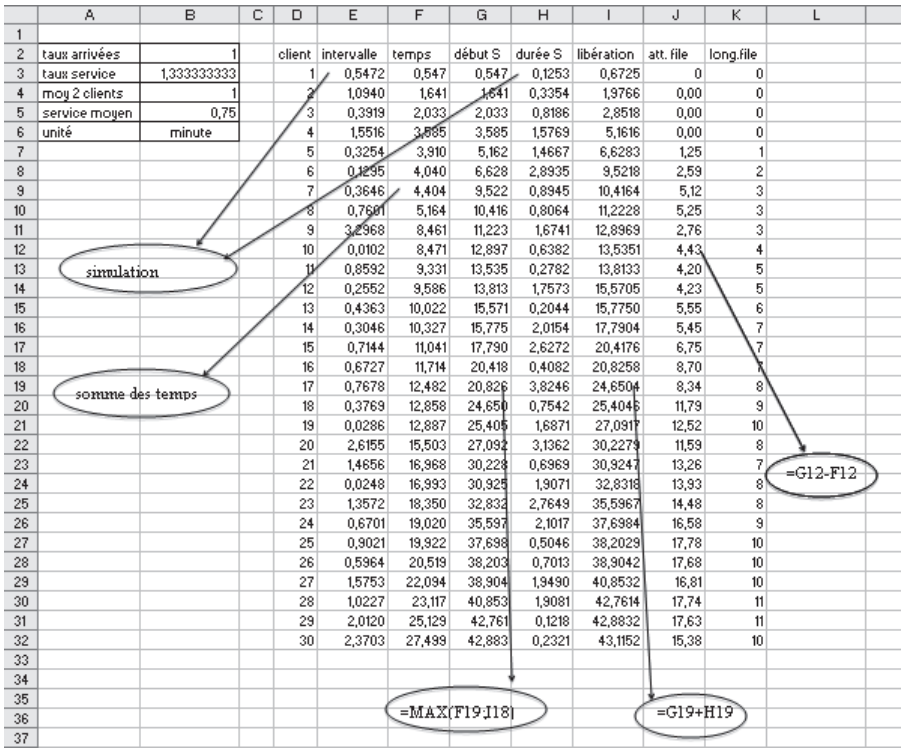
Le modèle simple de file d'attente à une seule ressource en partage est celui où les demandes (ou arrivées) sont aléatoires, indépendantes, et arrivent séparées par des intervalles de temps distribués selon une loi exponentielle, les temps de service (réponses aux demandes) étant eux aussi distribués selon une loi exponentielle.

Ces deux variables aléatoires (temps séparant deux demandes successives, temps de service) ont été simulées à l'aide des nombres pseudo-aléatoires d'Excel ; quelques calculs expliqués ci-dessous permettent de construire arrivées et départs (par libération de la ressource), ainsi que les temps d'attente et le nombre de demandes en attente (longueur de la file).

L'exemple choisi (*cf.* tableau 2), avec la minute pour unité de temps, est celui où les temps séparant les arrivées sont répartis selon une loi exponentielle de paramètre 1, et les temps de service sont répartis selon une loi exponentielle de paramètre 4/3. Autrement dit il y a en moyenne une demande par minute, et le temps de service moyen est de 0,75 minute, soit 45 secondes.

Ce modèle de file d'attente est noté M/M/1 (chaque M caractérisant la distribution exponentielle *memoryless*), des délais inter-arrivées puis des temps de service).

Tableau 2 – Simulation d'une file d'attente de type M/M/1 pour 30 arrivées



	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	taux arrivées	1		client	intervalle	temps	début S	durée S	libération	att. file	long.file	
3	taux service	1,333333333		1	0,5472	0,547	0,547	0,1253	0,6725	0	0	
4	moy 2 clients	1		2	1,0940	1,641	1,641	0,3354	1,9766	0,00	0	
5	service moyen	0,75		3	0,3919	2,033	2,033	0,8186	2,8518	0,00	0	
6	unité	minute		4	1,5516	3,585	3,585	1,5769	5,1616	0,00	0	
7				5	0,3254	3,910	5,162	1,4667	6,6283	1,25	1	
8				6	0,1295	4,040	6,628	2,8935	9,5218	2,59	2	
9				7	0,3646	4,404	9,522	0,8945	10,4164	5,12	3	
10				8	0,7691	5,164	10,416	0,8064	11,2228	5,25	3	
11				9	3,2968	8,461	11,223	1,6741	12,8969	2,76	3	
12				10	0,0102	8,471	12,897	0,6382	13,5351	4,43	4	
13				11	0,8592	9,331	13,535	0,2782	13,8133	4,20	5	
14				12	0,2552	9,586	13,813	1,7573	15,5705	4,23	5	
15				13	0,4363	10,022	15,571	0,2044	15,7750	5,55	6	
16				14	0,3046	10,327	15,775	2,0154	17,7904	5,45	7	
17				15	0,7144	11,041	17,790	2,6272	20,4176	6,75	7	
18				16	0,6727	11,714	20,418	0,4082	20,8258	8,70	7	
19				17	0,7678	12,482	20,826	3,8246	24,6504	8,34	8	
20				18	0,3769	12,858	24,650	0,7542	25,4046	11,79	9	
21				19	0,0286	12,887	25,405	1,6871	27,0917	12,52	10	
22				20	2,6155	15,503	27,092	3,1362	30,2279	11,59	8	
23				21	1,4656	16,968	30,228	0,6969	30,9247	13,26	7	
24				22	0,0248	16,993	30,925	1,9071	32,8318	13,93	8	
25				23	1,3572	18,350	32,832	2,7649	35,5967	14,48	8	
26				24	0,6701	19,020	35,597	2,1017	37,6984	16,58	9	
27				25	0,9021	19,922	37,698	0,5046	38,2029	17,78	10	
28				26	0,5964	20,519	38,203	0,7013	38,9042	17,68	10	
29				27	1,5753	22,094	38,904	1,9490	40,8532	16,81	10	
30				28	1,0227	23,117	40,853	1,9081	42,7614	17,74	11	
31				29	2,0120	25,129	42,761	0,1218	42,8832	17,63	11	
32				30	2,3703	27,499	42,883	0,2321	43,1152	15,38	10	
33												
34												
35												
36												
37												

Colonne D : numéro d'ordre de la demande (du client) qui arrive.

Colonne E : temps séparant deux arrivées successives simulées à l'aide de la fonction :

$$= - LN(1-ALEA())$$

Colonne : F dates réelles d'arrivées des demandes (des clients).

Dans la première cellule de la colonne (F3), on place la valeur contenue en E3

Dans la seconde cellule (F4), on additionne la date d'arrivée du précédent au temps écoulé jusqu'à l'arrivée du suivant : F4 = F3 + E4

Puis on réitère sur la suite de la colonne en tirant la cellule F4 vers le bas (croix en bas à droite de la cellule).

Colonne H : durées des services simulés à l'aide de la fonction :

$$= (4/3)*LN(1-ALEA())$$

Colonne G : date de début du service du client, qui est :

– sa date d'arrivée si le guichet est libre à ce moment ;

– ou bien égale à la date où le guichet est libéré par le client en cours de service

G3 = F3, puis G4 = MAX(F4 ;I3)

et le reste de la colonne par progression selon les valeurs de la colonne I.

Colonne I : date de libération du guichet (fin de service au client en cours). C'est la somme de la date de début de service et de la durée de service du client considéré

I3 = G3 + H3 et le reste de la colonne par progression

Colonne J : temps d'attente égal à la date de début de service diminuée de la date d'arrivée

J3 = G3 – F3 et le reste de la colonne par progression

Colonne K : longueur de la file d'attente lorsqu'un client arrive, qui est égale au numéro d'ordre de ce client diminué du numéro d'ordre du client qui sera servi :

– si le client qui arrive est servi tout de suite, cette longueur est nulle ;

– si le client qui est en cours de service est le précédent, la file sera de longueur 1, etc.

K5 = D5 – EQUIV(F5;G3:G5;1)

Cette procédure par simulation permet de comprendre l'influence des paramètres des lois exponentielles des temps entre demandes et des temps de service.

En gardant constant le premier, on peut voir que l'augmentation du second (qui signifie que le temps de service moyen est diminué donc que le guichet se libère plus vite) entraîne des attentes moins longues et une file moins fournie, alors que sa diminution (qui implique que le temps de service moyen est augmenté) allonge le temps d'attente moyen et la longueur de la file. Ces résultats peuvent être démontrés rigoureusement pour la file M/M/1, mais le recours à la simulation est parfois nécessaire pour évaluer le temps d'attente moyen, la longueur moyenne de la file, et comprendre le mécanisme et les conséquences d'un système impliquant une (ou plusieurs) file(s) d'attente.

3) Simulation d'une loi de Gauss

Nous avons déjà indiqué au § D de cette annexe (*cf.* méthode 1) que le théorème central limite (chapitre 7, § II.E) justifiait l'utilisation de la somme de 12 valeurs de la fonction ALEA() à laquelle on retranche 6 pour obtenir une valeur simulée de loi de Gauss centrée réduite.

L'exemple du tableau 3 porte sur 100 valeurs simulées (dont nous montrons les 25 premières) et illustre la qualité de cette méthode de simulation à l'aide d'un diagramme Quantile-Quantile.

Pour obtenir les quantiles théoriques, les valeurs simulées ont été triées par ordre croissant, et dans la colonne située à gauche on a porté le numéro

d'ordre de ce rangement (de 1 à 100) ; l'observation portant le numéro *i* est donc le quantile observé d'ordre *i*/100

La colonne de droite donne le quantile théorique de ce même ordre par application de la fonction :

= LOI.NORMALE.STANDARD.INVERSE(D*i*/100)

Tableau 3 – Simulation d'une loi de Gauss centrée réduite et calcul des quantiles théorique

C	D	E	F	G	H	I	J
	1	-2,339171976	-2,326347874				
	2	-1,935647642	-2,053748911				
	3	-1,904843002	-1,880793608				
	4	-1,894734582	-1,750686071				
	5	-1,807661612	-1,644853627				
	6	-1,68907906	-1,554773595				
	7	-1,60564648	-1,475791028				
	8	-1,600562944	-1,40507156				
	9	-1,499071175	-1,340755034				
	10	-1,470268583	-1,281551566				
	11	-1,347357655	-1,22652812				
	12	-1,345083913	-1,174986792				
	13	-1,222651196	-1,126391129				
	14	-1,209901163	-1,080319341				
	15	-1,146316619	-1,036433389				
	16	-1,108756007	-0,994457883				
	17	-1,106803406	-0,954165253				
	18	-1,046916254	-0,915365088				
	19	-1,031896302	-0,877896295				
	20	-0,945761521	-0,841621234				
somme de 12 ALEA() - 6	21	-0,917504826	-0,806421247	=LOI.NORMALE.STANDARD.INVERSE(D15/100)			
	22	-0,851327269	-0,772193214				
	23	-0,803950429	-0,738846849				
	24	-0,774842297	-0,706302563				
	25	-0,645920486	-0,67448975				

Pour la dernière valeur, on ne calcule pas LOI.NORMALE.STANDARD.INVERSE(1) qui n'est pas défini, mais on choisit une valeur arbitraire proche de 1, comme 0,995 ou 0,999

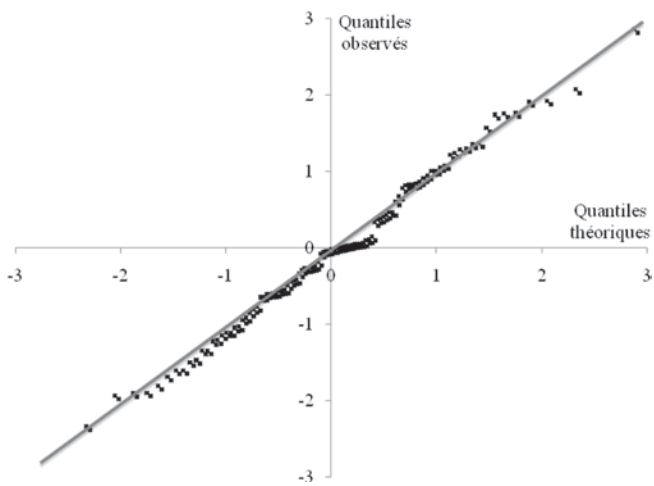


Figure 1 – Diagramme Quantile-Quantile obtenu pour les 100 valeurs

Le diagramme Quantile-Quantile obtenu sur pour 100 valeurs (*cf.* figure 1) justifie parfaitement l'utilisation de cette méthode de simulation pour obtenir des valeurs d'une loi de Gauss centrée réduite. Les valeurs peuvent servir ensuite pour toute situation concrète où un phénomène aléatoire est gouverné par une loi de Gauss.

En effet, avec des valeurs simulées u_i d'une loi de Gauss centrée réduite, on obtient des valeurs simulées y_i d'une loi de Gauss quelconque $\mathcal{N}(m ; \sigma)$ en calculant $y_i = \sigma \cdot u_i + m$

Il en est de même pour la simulation d'une loi log-normale en partant d'une loi de Gauss.

IV. Tables

À l'exception de la table des fractiles de la loi du khi-deux, les tables suivantes sont extraites de l' *Aide-mémoire statistique* (CISIA•CERESTA, 1999).

PROBABILITÉS CUMULÉES DE LA LOI BINOMIALE

Probabilités cumulées $\Pr(X \leq c) = \sum_{k=0}^{k=c} \binom{n}{k} p^k (1-p)^{n-k}$

n	c	p=1%	p=2%	p=3%	p=4%	p=5%	p=6%	p=7%	p=8%	p=9%	p=10%
	0	0,9044	0,8171	0,7374	0,6648	0,5987	0,5386	0,4840	0,4344	0,3894	0,3486
	1	0,9957	0,9838	0,9655	0,9418	0,9139	0,8824	0,8483	0,8121	0,7746	0,7361
	2	0,9999	0,9991	0,9972	0,9938	0,9885	0,9812	0,9717	0,9599	0,9460	0,9298
	3	1	1	0,9999	0,9996	0,9990	0,9980	0,9964	0,9942	0,9912	0,9872
10	4			1	1	0,9999	0,9998	0,9997	0,9994	0,9990	0,9984
	5					1	1	1	1	0,9999	0,9999
	6									1	1
	0	0,8179	0,6676	0,5438	0,4420	0,3585	0,2901	0,2342	0,1887	0,1516	0,1216
	1	0,9831	0,9401	0,8802	0,8103	0,7358	0,6605	0,5869	0,5169	0,4516	0,3917
	2	0,9990	0,9929	0,9790	0,9561	0,9245	0,8850	0,8390	0,7879	0,7334	0,6769
	3		0,9994	0,9973	0,9926	0,9841	0,9710	0,9529	0,9294	0,9007	0,8670
20	4		1	0,9997	0,9990	0,9974	0,9944	0,9893	0,9817	0,9710	0,9568
	5			1	0,9999	0,9997	0,9991	0,9981	0,9962	0,9932	0,9887
	6				1	1	0,9999	0,9997	0,9994	0,9987	0,9976
	7						1	1	0,9999	0,9998	0,9996
	8								1	1	0,9999
	9										1
	0	0,7397	0,5455	0,4010	0,2939	0,2146	0,1563	0,1134	0,0820	0,0591	0,0424
	1	0,9639	0,8794	0,7731	0,6612	0,5535	0,4555	0,3694	0,2958	0,2343	0,1837
	2	0,9967	0,9783	0,9399	0,8831	0,8122	0,7324	0,6488	0,5654	0,4855	0,4114
	3	0,9998	0,9971	0,9881	0,9694	0,9392	0,8974	0,8450	0,7842	0,7175	0,6474
30	4	0,9999	0,9996	0,9982	0,9937	0,9844	0,9685	0,9447	0,9126	0,8723	0,8245
	5	1	1	0,9997	0,9989	0,9967	0,9921	0,9838	0,9707	0,9519	0,9268
	6			1	0,9999	0,9994	0,9983	0,9960	0,9918	0,9848	0,9742
	7				1	0,9999	0,9997	0,9992	0,9980	0,9959	0,9922
	8					1	0,9999	0,9999	0,9996	0,9910	0,9980
	9						1	1	0,9999	0,9998	0,9995
	10								1	1	0,9999
	11										1
	0	0,6050	0,3642	0,2181	0,1299	0,0769	0,0453	0,0266	0,0155	0,0090	0,0052
	1	0,9106	0,7358	0,5553	0,4005	0,2794	0,1900	0,1265	0,0827	0,0532	0,0338
	2	0,9862	0,9216	0,8106	0,6767	0,5405	0,4162	0,3108	0,2260	0,1605	0,1117
	3	0,9984	0,9822	0,9372	0,8609	0,7604	0,6473	0,5327	0,4253	0,3303	0,2503
50	4	0,9999	0,9968	0,9832	0,9510	0,8964	0,8206	0,7290	0,6290	0,5277	0,4312
	5	1	0,9995	0,9963	0,9856	0,9622	0,9224	0,8650	0,7919	0,7072	0,6161
	6		0,9999	0,9993	0,9964	0,9882	0,9711	0,9417	0,8981	0,8404	0,7702
	7		1	0,9999	0,9992	0,9968	0,9906	0,9780	0,9562	0,9232	0,8779
	8			1	0,9999	0,9992	0,9973	0,9927	0,9834	0,9672	0,9421
	9				1	0,9998	0,9993	0,9978	0,9944	0,9875	0,9755
	10					1	0,9998	0,9994	0,9983	0,9957	0,9906
	11						1	0,9999	0,9995	0,9987	0,9968
	12							1	0,9999	0,9996	0,9990
	13								1	0,9999	0,9997
	14									1	0,9999
	15										1

PROBABILITÉS CUMULÉES DE LA LOI DE POISSON

Probabilités cumulées $\Pr(X \leq c) = \sum_{k=0}^{k=c} e^{-m} \frac{m^k}{k!}$

c	m=0,1	m=0,2	m=0,3	m=0,4	m=0,5	m=0,6	m=0,7	m=0,8	m=0,9
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066
1	0,9953	0,9825	0,9631	0,9384	0,9098	0,8781	0,8442	0,8088	0,7725
2	0,9998	0,9988	0,9964	0,9920	0,9856	0,9769	0,9659	0,9526	0,9372
3	1	0,9999	0,9997	0,9992	0,9982	0,9966	0,9942	0,9909	0,9866
4		1	1	0,9999	0,9998	0,9996	0,9992	0,9986	0,9977
5				1	1	1	0,9999	0,9998	0,9997
6							1	1	1

Probabilités cumulées $\Pr(X \leq c) = \sum_{k=0}^{k=c} e^{-m} \frac{m^k}{k!}$

c	m=1,0	m=1,5	m=2,0	m=2,5	m=3,0	m=3,5	m=4,0	m=4,5	m=5,0
0	0,3679	0,2231	0,1353	0,0821	0,0498	0,0302	0,0183	0,0111	0,0067
1	0,7358	0,5578	0,4060	0,2873	0,1991	0,1359	0,0916	0,0611	0,0404
2	0,9197	0,8088	0,6767	0,5438	0,4232	0,3208	0,2381	0,1736	0,1247
3	0,9810	0,9344	0,8571	0,7576	0,6472	0,5366	0,4335	0,3423	0,2650
4	0,9963	0,9814	0,9473	0,8912	0,8153	0,7254	0,6288	0,5321	0,4405
5	0,9994	0,9955	0,9834	0,9579	0,9161	0,8576	0,7851	0,7029	0,6160
6	0,9999	0,9991	0,9955	0,9858	0,9665	0,9347	0,8893	0,8311	0,7622
7	1	0,9998	0,9989	0,9958	0,9881	0,9733	0,9489	0,9134	0,8666
8		1	0,9998	0,9989	0,9962	0,9901	0,9786	0,9597	0,9319
9			1	0,9997	0,9989	0,9967	0,9919	0,9829	0,9682
10				0,9999	0,9997	0,9990	0,9972	0,9933	0,9863
11				1	0,9999	0,9997	0,9991	0,9976	0,9945
12					1	0,9999	0,9997	0,9992	0,9980
13						1	0,9999	0,9997	0,9993
14							1	0,9999	0,9998
15								1	0,9999
16									1

PROBABILITÉS CUMULÉES DE LA LOI DE POISSON (suite)

Probabilités cumulées $\Pr(X \leq c) = \sum_{k=0}^{k=c} e^{-m} \frac{m^k}{k!}$

c	m=5,5	m=6,0	m=6,5	m=7,0	m=7,5	m=8,0	m=8,5	m=9,0	m=9,5
0	0,0041	0,0025	0,0015	0,0009	0,0006	0,0003	0,0002	0,0001	0,0001
1	0,0266	0,0174	0,0113	0,0073	0,0047	0,0030	0,0019	0,0012	0,0008
2	0,0884	0,0620	0,0430	0,0296	0,0203	0,0138	0,0093	0,0062	0,0042
3	0,2017	0,1512	0,1118	0,0818	0,0591	0,0424	0,0301	0,0212	0,0149
4	0,3575	0,2851	0,2237	0,1730	0,1321	0,0996	0,0746	0,0550	0,0403
5	0,5289	0,4457	0,3690	0,3007	0,2414	0,1912	0,1496	0,1157	0,0885
6	0,6860	0,6063	0,5265	0,4497	0,3782	0,3134	0,2562	0,2068	0,1649
7	0,8095	0,7440	0,6728	0,5987	0,5246	0,4530	0,3856	0,3239	0,2687
8	0,9044	0,8472	0,7916	0,7291	0,6620	0,5925	0,5231	0,4557	0,3918
9	0,9462	0,9161	0,8774	0,8305	0,7764	0,7166	0,6530	0,5874	0,5218
10	0,9747	0,9574	0,9332	0,9015	0,8622	0,8159	0,7634	0,7060	0,6453
11	0,9890	0,9799	0,9661	0,9466	0,9208	0,8881	0,8487	0,8030	0,7520
12	0,9955	0,9912	0,9840	0,9730	0,9573	0,9362	0,9091	0,8758	0,8364
13	0,9983	0,9964	0,9929	0,9872	0,9784	0,9658	0,9486	0,9261	0,8981
14	0,9994	0,9986	0,9970	0,9943	0,9897	0,9827	0,9726	0,9585	0,9400
15	0,9998	0,9995	0,9988	0,9976	0,9954	0,9918	0,9862	0,9780	0,9665
16	0,9999	0,9998	0,9996	0,9990	0,9980	0,9963	0,9934	0,9889	0,9823
17	1	1	0,9998	0,9996	0,9992	0,9984	0,9970	0,9947	0,9911
18			1	0,9999	0,9997	0,9993	0,9987	0,9976	0,9957
19				1	0,9999	0,9997	0,9995	0,9989	0,9980
20					1	0,9999	0,9998	0,9996	0,9991
21						1	0,9999	0,9998	0,9996
22							1	0,9999	0,9998
23								1	0,9999
24									1

PROBABILITÉS CUMULÉES DE LA LOI DE POISSON (suite)

Probabilités cumulées $\Pr(X \leq c) = \sum_{k=0}^{k=c} e^{-m} \frac{m^k}{k!}$

c	m=10	m=11	m=12	m=13	m=14	m=15	m=16	m=17	m=18
0									
1	0,0005	0,0002	0,0001						
2	0,0028	0,0012	0,0005	0,0002	0,0001				
3	0,0104	0,0049	0,0023	0,0010	0,0005	0,0002	0,0001		
4	0,0293	0,0151	0,0076	0,0037	0,0018	0,0009	0,0004	0,0002	0,0001
5	0,0671	0,0375	0,0203	0,0107	0,0055	0,0028	0,0014	0,0007	0,0003
6	0,1302	0,0786	0,0458	0,0259	0,0142	0,0076	0,0040	0,0021	0,0010
7	0,2203	0,1432	0,0895	0,0540	0,0316	0,0180	0,0100	0,0054	0,0029
8	0,3329	0,2320	0,1550	0,0997	0,0620	0,0374	0,0220	0,0126	0,0071
9	0,4580	0,3405	0,2424	0,1658	0,1093	0,0698	0,0433	0,0261	0,0154
10	0,5831	0,4599	0,3472	0,2517	0,1756	0,1184	0,0774	0,0491	0,0304
11	0,6968	0,5793	0,4616	0,3532	0,2600	0,1847	0,1270	0,0847	0,0549
12	0,7916	0,6887	0,5760	0,4631	0,3584	0,2676	0,1931	0,1350	0,0917
13	0,8645	0,7813	0,6816	0,5730	0,4644	0,3622	0,2745	0,2009	0,1426
14	0,9166	0,8541	0,7721	0,6751	0,5704	0,4656	0,3675	0,2808	0,2081
15	0,9513	0,9075	0,8445	0,7636	0,6693	0,5680	0,4667	0,3714	0,2867
16	0,9730	0,9442	0,8988	0,8355	0,7559	0,6640	0,5659	0,4677	0,3750
17	0,9857	0,9679	0,9371	0,8905	0,8272	0,7487	0,6593	0,5440	0,4686
18	0,9928	0,9824	0,9626	0,9302	0,8826	0,8193	0,7423	0,6550	0,5622
19	0,9965	0,9908	0,9787	0,9574	0,9235	0,8751	0,8122	0,7363	0,6509
20	0,9984	0,9954	0,9884	0,9751	0,9521	0,9169	0,8681	0,8055	0,7307
21	0,9993	0,9978	0,9939	0,9860	0,9712	0,9468	0,9107	0,8615	0,7991
22	0,9997	0,9990	0,9969	0,9925	0,9833	0,9672	0,9617	0,9048	0,8551
23	0,9999	0,9996	0,9985	0,9962	0,9907	0,9805	0,9633	0,9367	0,8989
24	1	0,9999	0,9993	0,9982	0,9950	0,9888	0,9777	0,9593	0,9313
25		1	0,9997	0,9992	0,9974	0,9938	0,9869	0,9748	0,9554
26			0,9999	0,9997	0,9987	0,9967	0,9926	0,9848	0,9718
27			1	0,9999	0,9994	0,9983	0,9960	0,9912	0,9827
28				1	0,9997	0,9992	0,9979	0,9950	0,9897
29					0,9999	0,9996	0,9989	0,9973	0,9941
30					1	0,9998	0,9995	0,9986	0,9967
31						0,9999	0,9998	0,9993	0,9982
32						1	0,9999	0,9996	0,9990
33							1	0,9998	0,9995
34								0,9999	0,9998
35								1	0,9999
36									1

FRACTILES DE LA LOI NORMALE RÉDUITE

Cette table donne les valeurs absolues des fractiles, u_p de la loi normale réduite tels que :

$$F(u_p) = \int_{-\infty}^{u_p} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = P$$

Pour $P < 0{,}5$ (colonne de gauche et ligne supérieure) les fractiles u_p sont négatifs.

Pour $P > 0{,}5$ (colonne de droite et ligne inférieure) les fractiles u_p sont positifs.

P	0,000	0,001	0,002	0,003	0,004	0,005	0,006	0,007	0,008	0,009	0,010	
0,00	∞	3,0902	2,8782	2,7478	2,6521	2,5758	2,5121	2,4573	2,4089	2,3656	2,3263	0,99
0,01	2,3263	2,2904	2,2571	2,2262	2,1973	2,1701	2,1444	2,1201	2,0969	2,0749	2,0537	0,98
0,02	2,0537	2,0335	2,0141	1,9954	1,9774	1,9600	1,9431	1,9268	1,9110	1,8957	1,8808	0,97
0,03	1,8808	1,8663	1,8522	1,8384	1,8250	1,8119	1,7991	1,7866	1,7744	1,7624	1,7507	0,96
0,04	1,7507	1,7392	1,7279	1,7169	1,7060	1,6954	1,6849	1,6747	1,6646	1,6546	1,6449	0,95
0,05	1,6449	1,6352	1,6258	1,6164	1,6072	1,5982	1,5893	1,5805	1,5718	1,5632	1,5548	0,94
0,06	1,5548	1,5464	1,5382	1,5301	1,5220	1,5141	1,5063	1,4985	1,4909	1,4833	1,4758	0,93
0,07	1,4758	1,4684	1,4611	1,4538	1,4466	1,4395	1,4325	1,4255	1,4187	1,4118	1,4051	0,92
0,08	1,4051	1,3984	1,3917	1,3852	1,3787	1,3722	1,3658	1,3595	1,3532	1,3469	1,3408	0,91
0,09	1,3408	1,3346	1,3285	1,3225	1,3165	1,3106	1,3047	1,2988	1,2930	1,2873	1,2816	0,90
0,10	1,2816	1,2759	1,2702	1,2646	1,2591	1,2536	1,2481	1,2426	1,2372	1,2319	1,2265	0,89
0,11	1,2265	1,2212	1,2160	1,2107	1,2055	1,2004	1,1952	1,1901	1,1850	1,1800	1,1750	0,88
0,12	1,1750	1,1700	1,1650	1,1601	1,1552	1,1503	1,1455	1,1407	1,1359	1,1311	1,1264	0,87
0,13	1,1264	1,1217	1,1170	1,1123	1,1077	1,1031	1,0985	1,0939	1,0893	1,0848	1,0803	0,86
0,14	1,0803	1,0758	1,0714	1,0669	1,0625	1,0581	1,0537	1,0494	1,0450	1,0407	1,0364	0,85
0,15	1,0364	1,0322	1,0279	1,0237	1,0194	1,0152	1,0110	1,0069	1,0027	0,9986	0,9945	0,84
0,16	0,9945	0,9904	0,9863	0,9822	0,9782	0,9741	0,9701	0,9661	0,9621	0,9581	0,9542	0,83
0,17	0,9542	0,9502	0,9463	0,9424	0,9385	0,9346	0,9307	0,9269	0,9230	0,9192	0,9154	0,82
0,18	0,9154	0,9116	0,9078	0,9040	0,9002	0,8965	0,8927	0,8890	0,8853	0,8816	0,8779	0,81
0,19	0,8779	0,8742	0,8705	0,8669	0,8633	0,8596	0,8560	0,8524	0,8488	0,8452	0,8416	0,80
0,20	0,8416	0,8381	0,8345	0,8310	0,8274	0,8239	0,8204	0,8169	0,8134	0,8099	0,8064	0,79
0,21	0,8064	0,8030	0,7995	0,7961	0,7926	0,7892	0,7858	0,7824	0,7790	0,7756	0,7722	0,78
0,22	0,7722	0,7688	0,7655	0,7621	0,7588	0,7554	0,7521	0,7488	0,7454	0,7421	0,7388	0,77
0,23	0,7388	0,7356	0,7323	0,7290	0,7257	0,7225	0,7192	0,7160	0,7128	0,7095	0,7063	0,76
0,24	0,7063	0,7031	0,6999	0,6967	0,6935	0,6903	0,6871	0,6840	0,6808	0,6776	0,6745	0,75
0,25	0,6745	0,6713	0,6682	0,6651	0,6620	0,6588	0,6557	0,6526	0,6495	0,6464	0,6433	0,74
0,26	0,6433	0,6403	0,6372	0,6341	0,6311	0,6280	0,6250	0,6219	0,6189	0,6158	0,6128	0,73
0,27	0,6128	0,6098	0,6068	0,6038	0,6008	0,5978	0,5948	0,5918	0,5888	0,5858	0,5828	0,72
0,28	0,5828	0,5799	0,5769	0,5740	0,5710	0,5681	0,5651	0,5622	0,5592	0,5563	0,5534	0,71
0,29	0,5534	0,5505	0,5476	0,5446	0,5417	0,5388	0,5359	0,5330	0,5302	0,5273	0,5244	0,70
	0,010	0,009	0,008	0,007	0,006	0,005	0,004	0,003	0,002	0,001	0,000	P

FRACTILES DE LA LOI NORMALE RÉDUITE (suite)

(suite)

P	0,000	0,001	0,002	0,003	0,004	0,005	0,006	0,007	0,008	0,009	0,010	
0,30	0,5244	0,5215	0,5187	0,5158	0,5129	0,5101	0,5072	0,5044	0,5015	0,4987	0,4959	0,69
0,31	0,4959	0,4930	0,4902	0,4874	0,4845	0,4817	0,4789	0,4761	0,4733	0,4705	0,4677	0,68
0,32	0,4677	0,4649	0,4621	0,4593	0,4565	0,4538	0,4510	0,4482	0,4454	0,4427	0,4399	0,67
0,33	0,4399	0,4372	0,4344	0,4316	0,4289	0,4261	0,4234	0,4207	0,4179	0,4152	0,4125	0,66
0,34	0,4125	0,4097	0,4070	0,4043	0,4016	0,3989	0,3961	0,3934	0,3907	0,3880	0,3853	0,65
0,35	0,3853	0,3826	0,3799	0,3772	0,3745	0,3719	0,3692	0,3665	0,3638	0,3611	0,3585	0,64
0,36	0,3585	0,3558	0,3531	0,3505	0,3478	0,3451	0,3425	0,3398	0,3372	0,3345	0,3319	0,63
0,37	0,3319	0,3292	0,3266	0,3239	0,3213	0,3186	0,3160	0,3134	0,3107	0,3081	0,3055	0,62
0,38	0,3055	0,3029	0,3002	0,2976	0,2950	0,2924	0,2898	0,2871	0,2845	0,2819	0,2793	0,61
0,39	0,2793	0,2767	0,2741	0,2715	0,2689	0,2663	0,2637	0,2611	0,2585	0,2559	0,2533	0,60
0,40	0,2533	0,2508	0,2482	0,2456	0,2430	0,2404	0,2378	0,2353	0,2327	0,2301	0,2275	0,59
0,41	0,2275	0,2250	0,2224	0,2198	0,2173	0,2147	0,2121	0,2096	0,2070	0,2045	0,2019	0,58
0,42	0,2019	0,1993	0,1968	0,1942	0,1917	0,1891	0,1866	0,1840	0,1815	0,1789	0,1764	0,57
0,43	0,1764	0,1738	0,1713	0,1687	0,1662	0,1637	0,1611	0,1586	0,1560	0,1535	0,1510	0,56
0,44	0,1510	0,1484	0,1459	0,1434	0,1408	0,1383	0,1358	0,1332	0,1307	0,1282	0,1257	0,55
0,45	0,1257	0,1231	0,1206	0,1181	0,1156	0,1130	0,1105	0,1080	0,1055	0,1030	0,1004	0,54
0,46	0,1004	0,0979	0,0954	0,0929	0,0904	0,0878	0,0853	0,0828	0,0803	0,0778	0,0753	0,53
0,47	0,0753	0,0728	0,0702	0,0677	0,0652	0,0627	0,0602	0,0577	0,0552	0,0527	0,0502	0,52
0,48	0,0502	0,0476	0,0451	0,0426	0,0401	0,0376	0,0351	0,0326	0,0301	0,0276	0,0251	0,51
0,49	0,0251	0,0226	0,0201	0,0175	0,0150	0,0125	0,0100	0,0075	0,0050	0,0025	0,0000	0,50
	0,010	0,009	0,008	0,007	0,006	0,005	0,004	0,003	0,002	0,001	0,000	P

Grandes valeurs de u

P	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}
u_p	3,7190	4,2649	4,7534	5,1993	5,6120	5,9978

FONCTION DE RÉPARTITION DE LA LOI NORMALE RÉDUITE

Cette table donne pour $u \geq 0$, la valeur $P = F(u)$ de la fonction de répartition de la loi normale réduite telle que :

$$P = F(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Pour $u < 0$: $P = F(u) = 1 - F(-u)$

u	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	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	$0,9^303$	$0,9^306$	$0,9^310$	$0,9^313$	$0,9^316$	$0,9^318$	$0,9^321$	$0,9^324$	$0,9^326$	$0,9^329$
3,2	$0,9^331$	$0,9^334$	$0,9^336$	$0,9^338$	$0,9^340$	$0,9^342$	$0,9^344$	$0,9^346$	$0,9^348$	$0,9^350$
3,3	$0,9^352$	$0,9^353$	$0,9^355$	$0,9^357$	$0,9^358$	$0,9^360$	$0,9^361$	$0,9^362$	$0,9^364$	$0,9^365$
3,4	$0,9^366$	$0,9^368$	$0,9^369$	$0,9^370$	$0,9^371$	$0,9^372$	$0,9^373$	$0,9^374$	$0,9^375$	$0,9^376$
3,5	$0,9^377$	$0,9^378$	$0,9^378$	$0,9^379$	$0,9^380$	$0,9^381$	$0,9^381$	$0,9^382$	$0,9^383$	$0,9^383$
3,6	$0,9^384$	$0,9^385$	$0,9^385$	$0,9^386$	$0,9^386$	$0,9^387$	$0,9^387$	$0,9^388$	$0,9^388$	$0,9^389$
3,7	$0,9^389$	$0,9^390$	$0,9^400$	$0,9^404$	$0,9^408$	$0,9^412$	$0,9^415$	$0,9^418$	$0,9^422$	$0,9^425$
3,8	$0,9^428$	$0,9^431$	$0,9^433$	$0,9^436$	$0,9^438$	$0,9^441$	$0,9^443$	$0,9^446$	$0,9^448$	$0,9^450$
3,9	$0,9^452$	$0,9^454$	$0,9^456$	$0,9^458$	$0,9^459$	$0,9^461$	$0,9^463$	$0,9^464$	$0,9^466$	$0,9^467$

N.B. La notation $0,9^303$, par exemple, équivaut à 0,99903

FRACTILES DE LA LOI DE STUDENT

Cette table donne les valeurs des fractiles $t_P(\nu)$ de la loi de Student pour $P \geq 0{,}60$

Pour les valeurs $P \leq 0{,}40$, on a $t_P(\nu) = t_{1-P}(\nu)$

ν \ P	0,60	0,70	0,80	0,90	0,95	0,975	0,990	0,995	0,999	0,9995
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
32	0,256	0,530	0,853	1,309	1,694	2,037	2,449	2,738	3,365	3,622
34	0,255	0,529	0,852	1,307	1,691	2,032	2,441	2,728	3,348	3,601
36	0,255	0,529	0,852	1,306	1,688	2,028	2,434	2,719	3,333	3,582
38	0,255	0,529	0,851	1,304	1,686	2,024	2,429	2,712	3,319	3,566
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,298	1,676	2,009	2,403	2,678	3,261	3,496
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
70	0,254	0,527	0,847	1,294	1,667	1,994	2,381	2,648	3,211	3,435
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
90	0,254	0,526	0,846	1,291	1,662	1,987	2,368	2,632	3,183	3,402
100	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

N.B. Pour $\nu > 40$, on pourra utiliser l'approximation suivante :

$$t_P(\nu) = u_P + \frac{(u_P^3 + u_P)}{4\nu}$$

où u_P est le fractile d'ordre P de la loi normale réduite.

FRACTILES DE LA LOI DU KHI-DEUX

ν \ P	0,005	0,01	0,025	0,05	0,1	0,25	0,5	0,7	0,75	0,9	0,95	0,975	0,99	0,995	0,999
1	0,000	0,0002	0,001	0,004	0,016	0,102	0,455	1,074	1,32	2,706	3,841	5,02	6,635	7,88	10,827
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,408	2,77	4,605	5,991	7,38	9,210	10,60	13,815
3	0,072	0,115	0,216	0,352	0,584	1,21	2,366	3,665	4,11	6,251	7,815	9,35	11,345	13,00	16,266
4	0,207	0,297	0,484	0,711	1,064	1,92	3,357	4,878	5,39	7,779	9,488	11,14	13,277	15,00	18,467
5	0,412	0,554	0,831	1,15	1,610	2,67	4,351	6,064	6,63	9,236	11,070	12,83	15,086	16,86	20,515
6	0,676	0,872	1,24	1,64	2,204	3,45	5,348	7,231	7,84	10,645	12,592	14,45	16,812	18,65	22,457
7	0,989	1,24	1,69	2,17	2,833	4,25	6,346	8,383	9,04	12,017	14,067	16,01	18,475	20,37	24,322
8	1,34	1,64	2,18	2,73	3,490	5,07	7,344	9,524	10,22	13,362	15,507	17,53	20,090	22,03	26,125
9	1,73	2,09	2,70	3,33	4,168	5,90	8,343	10,656	11,39	14,684	16,919	19,02	21,666	23,66	27,877
10	2,16	2,56	3,25	3,94	4,865	6,73	9,342	11,781	12,55	15,987	18,307	20,48	23,209	25,25	29,588
11	2,60	3,05	3,82	4,57	5,578	7,58	10,341	12,899	13,70	17,275	19,675	21,92	24,725	26,82	31,264
12	3,07	3,57	4,40	5,23	6,304	8,43	11,340	14,011	14,85	18,549	21,026	23,34	26,217	28,35	32,909
13	3,57	4,11	5,01	5,89	7,042	9,30	12,340	15,119	15,98	19,812	22,362	24,74	27,688	29,87	34,528
14	4,07	4,66	5,63	6,57	7,790	10,16	13,339	16,222	17,12	21,064	23,685	26,12	29,141	31,37	36,123
15	4,60	5,23	6,26	7,26	8,547	11,03	14,339	17,322	18,25	22,307	24,996	27,49	30,578	32,85	37,697
16	5,14	5,81	6,91	7,96	9,312	11,91	15,338	18,418	19,37	23,542	26,296	28,85	32,000	34,31	39,252
17	5,70	6,41	7,56	8,67	10,085	12,79	16,338	19,511	20,49	24,769	27,587	30,19	33,409	35,76	40,790
18	6,26	7,01	8,23	9,39	10,865	13,67	17,338	20,601	21,60	25,989	28,869	31,53	34,805	37,19	42,312
19	6,84	7,63	8,91	10,12	11,651	14,56	18,338	21,689	22,72	27,204	30,144	32,85	36,191	38,62	43,820
20	7,43	8,26	9,59	10,85	12,443	15,45	19,337	22,775	23,83	28,412	31,410	34,17	37,566	40,03	45,315
21	8,03	8,90	10,28	11,56	13,240	16,34	20,337	23,858	24,93	29,615	32,671	35,48	38,932	41,43	46,797
22	8,64	9,54	10,98	12,34	14,041	17,24	21,337	24,939	26,04	30,813	33,924	36,78	40,289	42,83	48,268
23	9,26	10,20	11,69	13,09	14,848	18,13	22,337	26,018	27,14	32,007	35,172	38,08	41,638	44,21	49,728
24	9,89	10,86	12,40	13,85	15,659	19,03	23,337	27,096	28,24	33,196	36,415	39,36	42,980	45,59	51,179
25	10,56	11,52	13,12	14,61	16,47	19,940	24,34	28,172	29,340	34,38	37,652	40,65	44,314	46,96	52,620
26	11,20	12,20	13,84	15,38	17,29	20,840	25,34	29,246	30,430	35,56	38,885	41,92	45,642	48,32	54,052
27	11,84	12,88	14,57	16,15	18,114	21,75	26,336	30,319	31,53	36,741	40,113	43,19	46,963	49,67	55,476
28	12,49	13,56	15,31	16,93	18,939	22,66	27,336	31,391	32,62	37,916	41,337	44,46	48,278	51,02	56,893
29	13,15	14,26	16,05	17,71	19,768	23,56	28,236	32,461	33,71	39,087	42,557	45,72	49,588	52,36	58,302
30	13,82	14,95	16,79	18,49	20,599	24,48	29,336	33,530	34,80	40,256	43,773	46,98	50,892	53,70	59,703
40	20,73	22,16	24,43	26,51	29,05	33,66	39,34	44,16	45,62	51,80	55,76	59,34	63,69	66,78	73,44
50	28,01	29,71	32,36	34,76	37,69	42,94	49,33	54,72	56,33	63,17	67,50	71,42	76,15	79,50	86,69
60	35,55	37,48	40,48	43,19	46,46	52,29	59,33	65,23	66,98	74,40	79,08	83,30	88,38	91,96	99,63
70	43,29	45,44	48,76	51,74	55,33	61,70	69,33	75,69	77,58	85,53	90,53	95,02	100,42	104,22	112,34
80	51,18	53,54	57,15	60,39	64,28	71,14	79,33	86,12	88,13	96,58	101,88	106,63	112,33	116,33	124,86
90	59,21	61,75	65,65	69,13	73,29	80,62	89,33	96,52	98,64	107,56	113,14	118,14	124,12	128,31	137,22
100	67,34	70,06	74,22	77,93	82,36	90,13	99,33	106,91	109,14	118,50	124,34	129,56	135,81	140,18	149,46

Bibliographie

Ouvrages de base

ESCOFIER B., PAGES J., *Initiation aux traitements statistiques, Méthodes, méthodologie*, Presses universitaires de Rennes, 1997.
GIARD V., *Statistique appliquée à la gestion* , 8e éd., Économica, 2003.
GRAIS B., *Statistique descriptive*, coll. « Éco sup », 3e éd., Dunod, 2004.
GRAIS B., *Méthodes statistiques*, coll. « Éco sup », 4e éd., Dunod, 2006.
MORINEAU A., C HATELIN Y.-M. *et al.* , *L'analyse statistique des données : apprendre, comprendre et réaliser avec Excel* , Ellipses, 2005.

Ouvrages généraux

DODGE Y., *Statistique, Dictionnaire encyclopédique*, Springer Verlag, 2004.
DROESBEKE J.-J., T ASSI P., *Histoire de la statistique* , « Que sais-je » n° 2527, 2e éd., PUF, 1997.
INSEE, « Pour comprendre l'indice des prix », *Insee-Méthodes*, n° 81-82, 1998.

Ouvrages d'approfondissement

ANTOINE Ch., *Les Moyennes*, coll. « Que sais-je ? », n° 3383, PUF, 1998.
DODGE Y., MELFI G., *Premiers pas en simulation* , Springer Verlag, 2008.
SAPORTA G., *Probabilités, analyse des données et statistique*, 2e éd., Éditions Technip, 2006.
TENENHAUS M., *Statistique, méthodes pour décrire, expliquer et prévoir* , 2e éd., Dunod, 2007.

Sites Internet

Cours de statistique en ligne : www.agro-montpellier.fr/cnam-fr/statnet/
Module et méthodes de traitement : www.modulad.fr, onglet « Excel'Ense »

Logiciels

Excel 2007®, Microsoft.
PASW Statistics (2010), nouveau nom de SPSS version 18 pour Windows, Mac OS ; SPSS, Chicago, Illinois, 2004.
JMP® (2009) version 8pour Mac OS, pour Windows ou pour Linux, produit par SAS, SAS Institute Inc., Cary, NC, USA, 2004.
R version 2.11.1 (2010). Logiciel libre multi plates-formes (CNU General Public Licence), The R Foundation, www.r-project.org

Lexique anglais/français

A

Arithmetic mean – Moyenne arithmétique

B

Bernoulli trial – Épreuve de Bernoulli
Binomial distribution – Loi binomiale
Box plot – Boîte de distribution, boîte à moustache, boîte à pattes
Box-and-whisker plot – Boîte de distribution, boîte à moustaches, boîte à pattes

C

Categorical variable – Variable qualitative, variable nominale
Centered random variable – Variable aléatoire centrée
Central limit theorem – Théorème central-limite
Coefficient of kurtosis – Coefficient d'aplatissement
Coefficient of skewness – Coefficient d'asymétrie
Coefficient of variation – Coefficient de variation
Composite index number – Indice synthétique
Conditional frequency – Fréquence conditionnelle
Correlation coefficient– Coefficient de corrélation
Concentration index – Indice de concentration
Conditional distribution – Distribution conditionnelle
Conditional probability – Probabilité conditionnelle
Contingency table – Tableau de contingence
Continuous random variable – Variable aléatoire continue
Convergence in distribution – Convergence en loi
Convergence in second-order mean – Convergence en moyenne quadratique
Convergence in probability – Convergence en probabilité
Correlation ratio – Rapport de corrélation
Covariance – Covariance
Cumulative function – Fonction cumulative
Cumulative frequency – Effectif cumulé
Cumulative frequency curve – Courbe cumulative
Cumulative distribution function – Fonction de répartition

D

Decile – Décile
Degree of freedom – Degré de liberté
Depth – Profondeur
Discrete random variable – Variable aléatoire discrète
Dummy variable – Variable indicatrice

E

Equally probable – Équiprobabilité
Equiprobability – Équiprobabilité
Exhaustive sampling – Tirage exhaustif
Expected value – Espérance mathématique
Exponential smoothing – Lissage exponentiel

F

Forecasting – Prévision
Frequency – Effectif
Frequency distribution – Distribution observée
Frequency table – Tableau de fréquence

G

Gaussian distribution – Loi de Gauss
Geometric distribution – Loi géométrique
Geometric mean – Moyenne géométrique

H

Harmonic mean – Moyenne harmonique
Histogram – Histogramme

I

Independence – Indépendance
Index number – Indice élémentaire
Individual – Individu
Interquartile range – Intervalle interquartile

L

Law of large numbers – Loi des grands nombres
Least-squares regression line – Droite des moindres carrés
Line chart – Diagramme en bâtons

M

Marginal distribution – Distribution marginale
Median – Médiane
Mean – Moyenne
Mean absolute error of prediction – Erreur absolue moyenne de prévision
Mean deviation – Écart absolu moyen
Mean square error of prediction – Erreur quadratique moyenne de prévision
Measure of location – Indicateur de position
Measure of shape – Indicateur de forme
Measure of skewness – Indicateur d'asymétrie
Measure of variability – Indicateur de dispersion
Modality – Modalité
Mode – Mode
Moving average – Moyenne mobile
Moving median – Médiane mobile

N

Normal distribution – Loi normale

O

Observation – Observation
Outlier – Valeur éloignée, valeur extrême

P

Pair of random variables – Couple de variables aléatoires
Percentile – Centile
Pie chart – Diagramme circulaire
Population – Population
Probability – Probabilité
Probability density function– Fonction de densité de probabilité
Probability distribution – Loi de probabilité

Q

Quantile – Quantile, fractile
Quantile-Quantile plot – Diagramme Quantile-Quantile
Quartile – Quartile
Quantitative variable – Variable quantitative

R

Random component – Composante aléatoire
Random experiment – Expérience aléatoire
Random variable – Variable aléatoire
Range – Étendue
Regression curve – Courbe de régression
Relative frequency – Fréquence
Response category – Modalité

S

Sample – Échantillon
Sample space – Ensemble fondamental
Sampling without replacement– Tirage exhaustif
Scatter plot – Graphique de dispersion
Seasonal component – Composante saisonnière
Seasonally adjusted data – Données corrigées des variations saisonnières
Skewness – Asymétrie
Standard deviation – Écart-type
Standard normal distribution – Loi normale centrée réduite
Standardized normal distribution– Loi normale centrée réduite
Standardized random variable – Variable aléatoire centrée-réduite
Statistical unit – Unité statistique
Stem and leaf diagram – Diagramme « branche et feuille »

T

Time series – Chronique, série chronologique
Trend – Tendance à long terme

U

Uniform distribution – Loi uniforme

V

Variance – Variance

Lexique français/anglais

A

Asymétrie – *Skewness*

B

Boîte de distribution – *Box plot, box-and-whisker plot*
Boîte à moustaches – *Box plot, box-and-whisker plot*
Boîte à pattes – *Box plot, box-and-whisker plot*

C

Centile – *Percentile*
Chronique – *Time series*
Coefficient d'aplatissement – *Coefficient of kurtosis*
Coefficient d'asymétrie – *Coefficient of skewness*
Coefficient de corrélation – *Correlation coefficient*
Coefficient de variation – *Coefficient of variation*
Composante saisonnière – *Seasonal component*
Composante aléatoire – *Random component*
Convergence en loi – *Convergence in distribution*
Convergence en moyenne quadratique – *Convergence in second-order mean*
Convergence en probabilité – *Convergence in probability*
Couple de variables aléatoires –*Pair of random variables*
Courbe cumulative – *Cumulative frequency curve*
Courbe de régression – *Regression curve*
Covariance – *Covariance*

D

Décile – *Decile*
Degré de liberté – *Degree of freedom*
Diagramme « branche et feuille » – *Stem and leaf diagram*
Diagramme circulaire – *Pie chart*
Diagramme en bâtons – *Line chart*
Diagramme Quantile-Quantile – *Quantile-Quantile plot*
Distribution conditionnelle – *Conditional distribution*
Distribution marginale – *Marginal distribution*
Distribution observée – *Frequency distribution*
Droite des moindres carrés – *Least-squares regression line*

E

Écart absolu moyen – *Mean deviation*
Écart-type – *Standard deviation*
Échantillon – *Sample*
Effectif – *Frequency*
Effectif cumulé – *Cumulative frequency*
Ensemble fondamental – *Sample space*

Épreuve de Bernoulli – *Bernoulli trial*
Équiprobabilité – *Equiprobability, equally probable*
Erreur absolue moyenne de prévision – *Mean absolute error of prediction*
Erreur quadratique moyenne de prévision – *Mean square error of prediction*
Espérance mathématique – *Expected value*
Étendue – *Range*
Expérience aléatoire – Random experiment

F

Fonction cumulative – *Cumulative function*
Fonction de densité de probabilité – *Probability density function*
Fonction de répartition – *Cumulative distribution function*
Fractile – *Quantile*
Fréquence – *Relative frequency*
Fréquence conditionnelle – *Conditional frequency*

G

Graphique de dispersion – *Scatter plot*

H

Histogramme – *Histogram*

I

Indépendance – *Independence*
Indicateur d'asymétrie – *Measure of skewness*
Indicateur de dispersion – *Measure of variability*
Indicateur de forme – *Measure of shape*
Indicateur de position – *Measure of location*
Indice de concentration – *Concentration Index*
Indice élémentaire – *Index number*
Indice synthétique – *Composite index number*
Individu – *Individual, observation*
Intervalle interquartile – *Interquartile range*

L

Lissage exponentiel – *Exponential smoothing*
Loi binomiale – *Binomial distribution*
Loi de Gauss – *Gaussian distribution*
Loi de probabilité – *Probability distribution*
Loi des grands nombres –*Law of large numbers*
Loi géométrique – *Geometric distribution*
Loi normale – *Normal distribution*
Loi normale centrée réduite – *Standardized normal distribution, standard normal distribution*
Loi uniforme – *Uniform distribution*

M

Médiane – *Median*
Médiane mobile – *Moving median*
Modalité – *Modality, response category*
Mode – *Mode*
Moyenne – *Mean*
Moyenne arithmétique – *Arithmetic mean*
Moyenne géométrique – *Geometric mean*

Moyenne harmonique – *Harmonic mean*
Moyenne mobile – *Moving average*

P

Population – *Population*
Prévision – *Forecasting*
Probabilité – *Probability*
Probabilité conditionnelle – *Conditional probability*
Profondeur – *Depth*

Q

Quantile – *Quantile*
Quartile – *Quartile*

R

Rapport de corrélation – *Correlation ratio*

S

Série chronologique – *Time series*
Série corrigée des variations saisonnières – *Seasonally adjusted series*

T

Tableau de contingence – *Contingency table*
Tableau de fréquence – *Frequency table*
Tendance à long terme – *Trend*
Théorème central-limite – *Central limit theorem*
Tirage exhaustif – *Exhaustive sampling, sampling without replacement*

U

Unité statistique – *Statistical unit*

V

Valeur éloignée – *Outlier*
Valeur extrême – *Outlier*
Variable aléatoire – *Random variable*
Variable aléatoire centrée – *Centered random variable*
Variable aléatoire centrée-réduite – *Standardized random variable*
Variable aléatoire continue – *Continuous random variable*
Variable aléatoire discrète – *Discrete random variable*
Variable indicatrice – *Dummy variable*
Variable nominale – *Categorical variable*
Variable quantitative – *Quantitative variable*
Variable qualitative – *Categorical variable*
Variance – *Variance*

Index

F

H

I

K

L

M

P

Q

R

S

T

U

V

W