UNIVERSITE DE TOULOUSE LE MIRAIL

LICENCE 2 - MIH015X

Statistique pour les Sciences Humaines II (parcours Histoire)

Polycopié de statistique

à l'usage des étudiants inscrits au S.E.D.

———

Agnès Lagnoux **lagnoux@univ-tlse2.fr**

Les étudiants inscrits en première année de Licence dans toute filière ou en deuxième année de Licence d'Histoire peuvent suivre des cours de statistique descriptive, dans le cadre des U.E. optionnelles MIS2OP1X (1er semestre) et uniquement pour les Historiens MIHO15X (2ème semestre) intitulées "Statistique pour les Sciences Humaines I" et "Statistique pour les Sciences Humaines II (parcours Histoire)". L'U.E. optionnelle MIS2OP2X "Statistique pour les Sciences Humaines II" destinée aux étudiants de première année de Licence dans toute filière traitera de Probabilités et Statistique.

L'objectif de ce cours est de donner les outils nécessaires à la compréhension et à l'analyse de documents comportant des données numériques, en liaison avec les Sciences Humaines. Il s'agit d'une initiation à la statistique descriptive qui ne nécessite pas de connaissances spécifiques préalables ; cependant, ce sera l'occasion de revoir avec un peu de recul des notions mathématiques élémentaires qui font partie de la culture générale (calculer des taux de variation, résoudre une équation, utiliser un repère cartésien,...). Ces enseignements sont assurés par des professeurs de mathématiques dépendant du département de Mathématiques et Informatique de l'UFR SES.

Les exercices proposés (situés en fin de polycopié) s'appuient le plus souvent sur des données réelles mais la quantité de données est parfois réduite pour permettre de faire les calculs en un temps raisonnable. Une calculatrice est nécessaire, tous les modèles sont autorisés mais une calculatrice scientifique pour le collège suffit. Une des premières difficultés est de se familiariser avec le vocabulaire spécifique de la statistique. Les mots ayant un sens mathématique précis définis dans ce cours sont en gras. Les premiers exercices fournissent une liste d'exemples permettant d'assimiler ce vocabulaire.

Nous présentons au premier semestre des généralités sur la statistique descriptive concernant une seule varible puis les couples de variables. Nous étudierons aussi l'existence de liaison entre deux variables quelconques. Le deuxième semestre est consacré aux taux de variation, à un type de liaison particulier entre deux variables qui est la corrélation linéaire et enfin à l'étude des séries temporelles. Il est nécessaire d'avoir suivi les cours du premier semestre pour aborder le deuxième. A la fin de chaque semestre, nous illustrerons toutes les notions abordées dans le cours en utilisant le tableur Excel.

N'hésitez pas à me contacter par courrier électronique ou téléphone si vous avez une question

précise concernant le cours ou bien l'organisation de l'U.E.. Il y aura un regroupement en fin de semestre (la date vous sera communiquée plus tard), je vous conseille vivement d'y particper. Enfin, les remarques, critiques et suggestions concernant ce nouveau polycopié sont les bienvenues.

Bon courage !

Responsable des U.E. MIS2OP1X, MIHO11X et MIHO15X :
Agnès Lagnoux
U.F.R. S.E.S.
Département de Mathématiques et Informatique
Bureau 1039, bâtiment 13
Tél : 05-61-50-46-11, e-mail : lagnoux@univ-tlse2.fr.

Organisation : Pour les étudiants inscrits en contrôle continu, il y a 12 séances de Cours/TD (2 heures par semaine pendant 12 semaines) au deuxième semestre. Si vous souhaitez (et pouvez) assister à quelques séances, n'hésitez pas à vous renseigner sur l'horaire du cours : il n'est pas facile de comprendre seul certaines notions.
Vous **devez** renvoyer le devoir qui se trouve à la fin du polycopié avant le 15 mars 2014 (il sera corrigé et noté à titre indicatif).

Evaluation : Une épreuve écrite de statistique aura lieu en avril. La calculatrice ainsi qu'une feuille manuscrite recto-verso sont autorisées à l'examen (vous pouvez vous inspirer des fiches proposées à la fin du polycopié).

Table des matières

Chapitre 1

Taux de variation et courbes semi-logarithmiques

Ce chapitre ne nécessite aucun pré-requis. Il est quand-même très important puisqu'y sont abordées des notions fondamentales utilisées très fréquemment dans la vie de tous les jours : ils s'agit des pourcentages (TVA, soldes, intérêts de placements) et des taux de variation. En plus de leur utilité pratique, ces notions sont indispensables dans la plupart des concours administratifs ainsi que dans la suite de vos études.

1.1 Pourcentage

Un **pourcentage** est une fraction dont le dénominateur est 100.
Notation : $a\% = \dfrac{a}{100}$ $(a = 100 \times a\%)$.

Exemple 1.1. $12\% = \dfrac{12}{100} = 0,12$; $12 = 100 \times 12\%$.

Un pourcentage permet d'écrire un rapport entre deux grandeurs de même nature, exprimées dans la même unité : il n'a donc pas lui-même d'unité.

Remarque 1.1. *Il faut bien comprendre que les pourcentages n'ont pas d'intérêt mathématique puisqu'on ne fait que multiplier et diviser un nombre par 100 en introduisant une nouvelle notation. Ils sont simplement utiles pour exprimer un rapport dans le langage courant.*

Nous allons utiliser les pourcentages dans la suite pour exprimer des taux de variation, mais n'oublions pas que les pourcentages ne sont pas toujours des taux de variation, ils peuvent être des fréquences.

1.2 Taux de variation

1.2.1 Généralités

Soit x une valeur donnée et a un nombre (positif ou négatif, supérieur à -100). Si x subit une variation de $a\%$, la nouvelle valeur sera :

$$x + \frac{a}{100}\,x = \left(1 + \frac{a}{100}\right)\,x$$

Autrement dit, si x varie de $a\%$, x est multiplié par $\left(1 + \frac{a}{100}\right)$.
On appellera **multiplicateur** le nombre $\boxed{M = 1 + \frac{a}{100}}$.

Attention : C'est le multiplicateur qui sera utile pour bien comprendre les variations successives et le taux de variation moyen mais il faut savoir passer rapidement du multiplicateur au taux de variation correspondant et inversement.

Méthode :

Lorsqu'on connaît la valeur de départ V_1 et la valeur d'arrivée V_2 :

$$\boxed{\text{Multiplicateur} = \frac{V_2}{V_1}.}$$

Puis on soustrait 1 pour avoir le taux de variation :

$$\boxed{\text{Taux de variation} = \frac{V_2}{V_1} - 1 = \text{Multiplicateur} - 1}$$

Et donc :

$$\boxed{\text{Multiplicateur} = \text{Taux de variation} + 1}.$$

Ces relations sont très simples mais il faut faire attention lorsque le taux de variation est exprimé en pourcentage, on parle alors aussi de **pourcentage de variation :**

$$\boxed{\text{Taux de variation} = 100 \times \left(\frac{V_2}{V_1} - 1\right)\%}$$

Propriétés :

- Si le taux de variation est positif, c'est-à-dire si le multiplicateur est supérieur à 1, il s'agit d'une augmentation.
- Si le taux de variation est négatif, c'est-à-dire si le multiplicateur est compris entre 0 et 1, il s'agit d'une diminution.

Remarque 1.2. *Une autre façon équivalente de définir le taux de variation (mais le multiplicateur apparaît de façon moins évidente) est :*

$$\boxed{\textit{Taux de variation} = \frac{V_2 - V_1}{V_1}.}$$

Exemple 1.2. *Considérons le tableau suivant concernant l'évolution de la population active en France.*

Années	*Salariés (en milliers)*	*Population active (en milliers)*	*%de salariés dans la population active*
1851	11954	21877	
1954		19219	

a) Remplir les cases vides sachant que le nombre de salariés a augmenté de $5,07\%$.

b) Quel est le pourcentage de diminution de la population active ?

c) Quel est le taux de variation du pourcentage de salariés dans la population active ?

(Réponses : 12560 *;* $54,64\%$ *;* $65,35\%$ *;* $12,15\%$ *;* $19,60\%$ *)*

1.2.2 Variations successives

Soit x une valeur donnée qui subit deux variations successives de $a\%$ puis $b\%$. Soient y la nouvelle valeur après la première variation et z la nouvelle valeur après la seconde variation. On a donc le schéma suivant :

$$x \overset{a\%}{\Longrightarrow} y \overset{b\%}{\Longrightarrow} z$$

Par le calcul,

$$y = \left(1 + \frac{a}{100}\right) x$$

puis

$$z = \left(1 + \frac{b}{100}\right) y = \left(1 + \frac{b}{100}\right)\left(1 + \frac{a}{100}\right) x$$

Le multiplicateur permettant de passer de x à z est donc $\left(1 + \frac{b}{100}\right)\left(1 + \frac{a}{100}\right)$.

Remarque 1.3.

$$\left(1 + \frac{b}{100}\right)\left(1 + \frac{a}{100}\right) = \left(1 + \frac{a + b + \frac{ab}{100}}{100}\right)$$

et donc le pourcentage de variation globale est $a + b + \dfrac{ab}{100}$ *et non* $a + b$*. Ainsi une variation de* $a\%$ *suivie d'une variation de* $b\%$ *n'est pas égale à une variation de* $(a + b)\%$.

De façon générale, si l'on a n variations successives, le multiplicateur permettant de passer de la première à la dernière valeur est le produit des n multiplicateurs. Une fois qu'on a trouvé le multiplicateur global, on a plus qu'à calculer le taux de variation correspondant en retranchant 1.

Cas particulier :
Lorsqu'il y a n variations successives et identiques de $a\%$, le multiplicateur est $\left(1+\frac{a}{100}\right)^n$.

Exemple 1.3. *Le gouvernement annonce trois augmentations du tabac : 6% par an pendant 3 ans. De combien est l'augmentation globale ?*

$\left(1+\frac{6}{100}\right)^3 = 1,06^3 = 1,191016$, *soit une augmentation globale de* $19,1\%$.

1.2.3 Taux de variation moyen

On suppose que x a subi une variation globale de $a\%$ sur une durée totale de n périodes. On cherche à savoir quel est le taux de variation moyen par période.

Exemple 1.4. *Le prix du litre de SP95 a augmenté de 40% en 10 ans. On souhaite déterminer l'augmentation moyenne annuelle. (Réponse 3,82%.)*

Le **taux de variation moyen** $\alpha\%$ **sur** n **périodes** est le pourcentage tel que n variations successives de $\alpha\%$ donnent une variation de $a\%$.
En termes de multiplicateurs, on a :

$$\left(1+\frac{\alpha}{100}\right)^n = 1+\frac{a}{100}$$

$$1+\frac{\alpha}{100} = \sqrt[n]{1+\frac{a}{100}} \left(\text{ou } \left(1+\frac{a}{100}\right)^{\frac{1}{n}}\right) \text{ (lire : "racine } n\text{ième de } 1+\frac{a}{100}\text{")}$$

$$\frac{\alpha}{100} = \sqrt[n]{1+\frac{a}{100}} - 1 \text{ soit } \boxed{\alpha = 100\left(\sqrt[n]{1+\frac{a}{100}} - 1\right)}.$$

Il est indispensable pour faire les exercices de savoir calculer une racine $n$ième avec la calculatrice. Les exemples suivants permettent de s'entraîner.

Exemple 1.5. *1. Une quantité x augmente de 20% la première année et de 40% la deuxième. Quel est le taux de variation global ?*

($1,2 \times 1,4 = 1,68$ *, soit un taux de variation de* 68%*).*

Quel est le taux de variation annuel moyen ?

($\sqrt{1,68} = 1,2961$ *, soit un taux annuel moyen d'augmentation de* $29,61\%$ *).*

Remarque 1.4. *Le taux moyen n'est ni la moyenne des taux, ni le taux global divisé par 2.*

2. Un paquet de cigarettes coûtait 10F en 1988 et 20F en 1998.

Quel a été le taux annuel moyen d'augmentation ?

($2^{0,1} = 1,07177$; taux annuel moyen : $7,18\%$).

Quel a été le taux trimestriel moyen d'augmentation ?

($2^{0,025} = 1,01747$; taux trimestriel moyen : $1,75\%$).

Remarque 1.5. *Ce n'est pas le quart du taux annuel moyen.*

1.3 Graphiques semi-logarithmiques

1.3.1 Introduction

Rassurez-vous : malgré leur nom, l'utilisation des graphiques semi-logarithmiques ne nécessite en aucune façon de connaître ou savoir calculer la fonction logarithme. Les curieux pourront quand même se reporter à la section 1.3.4 pour en savoir plus.

Ici il s'agit d'un **problème d'échelle** : au lieu d'utiliser les échelles arithmétiques que l'on connaît depuis l'école maternelle et où la distance entre la graduation 1 et la graduation 2 est la même qu'entre la graduation 2 et la graduation 3 et ainsi de suite, on introduit un autre système de graduation.

A la place de la proportionnalité précédente (**proportionnalité arithmétique**), on repère des graduations proportionnelles aux logarithmes des valeurs (**proportionnalité logarithmique**). Dans le premier cas, la distance entre les graduations 4 et 3, par exemple, donnera 4-3=1cm sur une feuille de papier où l'échelle est arithmétique. Dans le deuxième cas, cette distance sera $\log 4 - \log 3 = 1cm$. Pour une lecture agréable du graphique, on écrira la valeur donnée et non son logarithme.

Mais pourquoi recourir à des graphiques semi-logarithmiques plus complexes que ceux que nous connaissons déjà ? Quels en sont les avantages ?

Problème 1 : Le cas de la saturation des axes

Par exemple, voici les effectifs des agents de l'Etat dans la ville V et dans la région R sur trois années :

Années	Nb d'agents en V	Nb d'agents en R
2005	100	90000
2006	300	920000
2007	250	85000

Il est intéressant de se demander si ces effectifs augmentent ou diminuent de la même manière (proportionnellement ou pas) en même temps. Pour cela, nous avons envie de faire un graphique afin d'en juger visuellement. Mais si on utilise une échelle habituelle (arithmique), on aura beau tenter toutes les combinaisons possibles de graduation, les deux séries ne rentreront jamais dans le même graphique. Même si on fabriquait une échelle verticale de 10 mètres de long, le message visuel serait inopérant montrant une courbe non fluctuante tout en bas et une autre tout en haut à plusieurs mètres. **Les valeurs de la première série de l'ordre de quelques centaines sont bien trop éloignées de celles de la deuxième série.**

La solution : le diagramme semi-logarithmique !

Problème 2 : Le cas des parallèles non proportionnelles

Un autre problème peut survenir lorsqu'on visualise des comparaisons graphiques en échelle arithmétique.

Par exemple, nous nous intéressons à la production de deux entreprises sur deux années. L'entreprise A produit 200 la première année puis 400 la deuxième alors que l'entreprise B produit 300 la première année puis 500 la deuxième.

Calculons maintenant les taux de variation correspondants :

$400/200 - 1 = 1$ soit 100% pour l'entreprise A ;

$500/300 - 1 = 0,6667$ soit $66,67\%$ pour l'entreprise B.

Visuellement, les deux évolutions apparaissent parallèles. Pourtant, les productions ont augmenté dans des proportions bien différentes. Le message visuel est donc ici trompeur : on dirait à première vue que les productions évoluent à la même vitesse. Or il n'en est rien. **Les échelles arithmétiques traduisent mal les variations relatives.**

La solution : le diagramme semi-logarithmique !

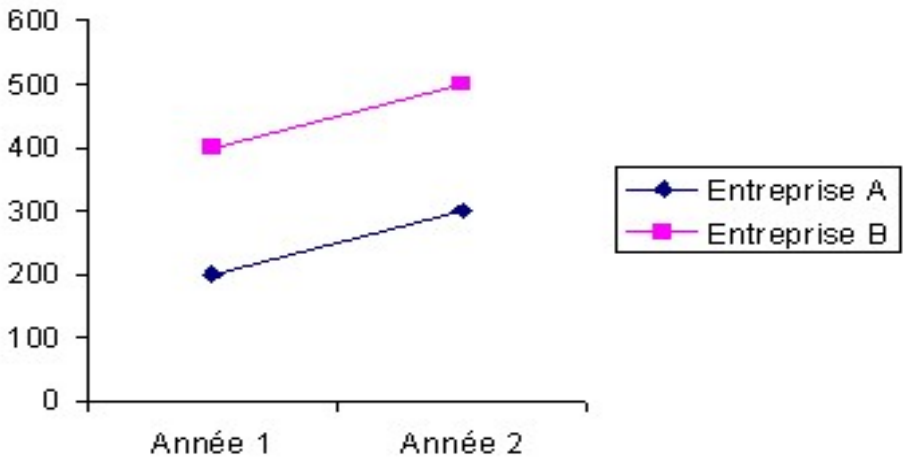


FIGURE 1.1 – *Production des entreprises A et B.*

1.3.2 L'utilisation des supports et du papier semi-logarithmique

Pour bien comprendre

ÉCHELLES ARITHMÉTIQUES	**ÉCHELLES LOGARITHMIQUES**
C'est le domaine de l'addition $(+)$	C'est le domaine de la multiplication $(\times)$
Elles décrivent des partitions.	Elles décrivent des évolutions.
Ex : Victor mange 20% de la tarte aux pommes. S'il en reprend 5%, il aura pris deux parts de tarte $20\% + 5\% = 25\%$. Il aura mangé un quart de tarte aux pommes.	Ex : Elise a grandi de 10% par an sur trois ans. Son taux de croissance annuel est de 0.10 et l'évolution aboutit à $(1 + 0.1)^3$. Sa taille de départ est multipliée par $(1 + 0.1)^3 = 1.33$.
Les échelles arithmétiques traduisent bien les **écarts absolus** : si 1cm vaut 10, 2cm valent 20.	Les échelles logarithmiques traduisent bien les **variations relatives** : 1/10=10/100=100/1000=10%.

Un repère semi-logarithmique se construit en graduant l'axe des abscisses (axe horizontal) comme d'habitude, régulièrement (par exemple, 1 cm correspond à 10 ans), mais on doit reporter sur l'axe des ordonnées (axe vertical) non pas les valeurs données mais les logarithmes décimaux de ces valeurs. Cependant, pour une lecture agréable du graphique, on écrira la valeur donnée et non son logarithme. Dans la pratique, on gradue à l'avance l'axe des ordonnées (il existe du papier déjà gradué). Comme ce sont des logarithmes décimaux, les distances physiques (sur le

papier) entre les multiples de 10 sont égales.

On appelle **module** d'un repère semi-logarithmique, la distance sur l'axe des ordonnées entre les graduations 1 et 10 (c'est-à-dire aussi entre 10 et 100 ou entre 100 et 1000,...). On voit donc déjà que l'on pourra placer sur la même feuille, par exemple à 4 modules, les nombres d'agents de V et de R de l'exemple ci-dessus. Ce qui résoud notre problème de saturation des axes.

Par ailleurs, les propriétés mathématiques des logarithmes (cf section 1.3.4) font que deux droites ou deux courbes qui se présentent de façon parallèle sur le papier semi-logarithmique correspondent bien à des évolutions proportionnelles. Ainsi le problème des parallèles non proportionnelles est lui aussi résolu.
En effet, à une variation de $a\%$ entre deux observations successives correspond toujours la même hauteur sur le graphique :

$$\log\left(x + \frac{a}{100}\,x\right) - \log(x) = \log\left(\frac{100+a}{100}\right),$$

cette valeur ne dépend pas de x mais uniquement de a. Ainsi, un graphique semi-logarithmique fait apparaître les variations en pourcentage et un petit graphique annexe représentant les pentes correspondant à des variations types pour une période donnée permet une meilleure lecture.

Remarque 1.6. *Un détail sur les distances : sur un papier arithmétique, si 4-3 donnent un centimètre, 5-4 et 179-178 aussi. Par contre, les différences entre les logarithmes ne donnent pas une proportionnalité constante :*
- $\log 2 - \log 1 = 0,30103$*, la distance pourra être de 3,01 millimètres sur l'échelle ;*
- $\log 3 - \log 2 = 0,176$*, la distance pourra être de 1,76 millimètres sur l'échelle ;*
- $\log 4 - \log 3 = 0,1249$*, la distance pourra être de 1,25 millimètres sur l'échelle ;*

mais
- $\log 20 - \log 10 = 0,30103$*, la distance pourra être de 3,01 millimètres sur l'échelle ;*
- $\log 30 - \log 20 = 0,176$*, la distance pourra être de 1,76 millimètres sur l'échelle ;*
- $\log 40 - \log 30 = 0,1249$*, la distance pourra être de 1,25 millimètres sur l'échelle ;*

et aussi
- $\log 200 - \log 100 = 0,30103$*, la distance pourra être de 3,01 millimètres sur l'échelle ;*
- $\log 300 - \log 200 = 0,176$*, la distance pourra être de 1,76 millimètres sur l'échelle ;*
- $\log 400 - \log 300 = 0,1249$*, la distance pourra être de 1,25 millimètres sur l'échelle.*

On prend donc conscience que les graduations de chaque module se resserrent de plus en plus de bas en haut sur l'axe vertical.

1.3.3 Application

Exemple 1.6. *Voici un tableau donnant l'évolution de la population de New York, en milliers d'habitants, de 1800 à 1910 :*

Année	*1800*	*1810*	*1820*	*1830*	*1840*	*1850*	*1860*	*1870*	*1880*	*1890*	*1900*	*1910*
Population	*75*	*110*	*130*	*230*	*400*	*700*	*1100*	*1400*	*1900*	*2300*	*3200*	*4500*

a) Faisons un graphique arithmétique (habituel) représentant cette évolution.

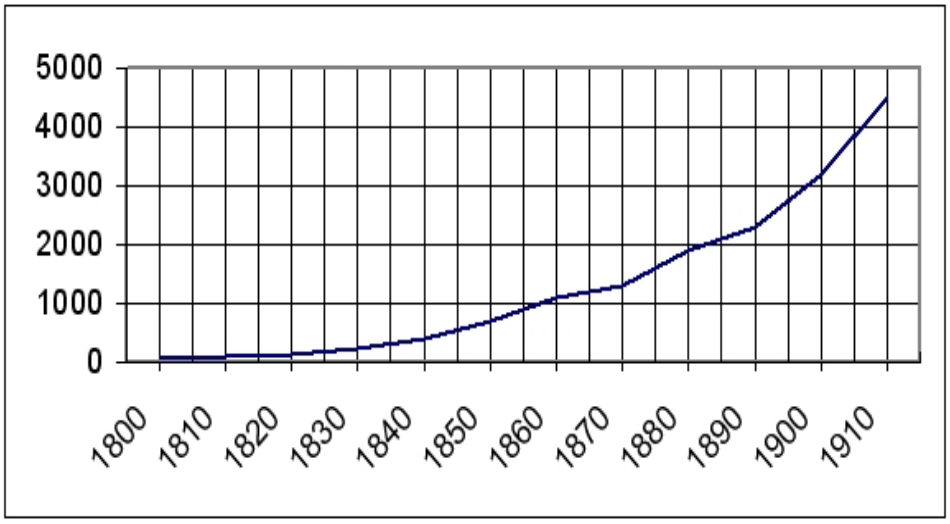


FIGURE 1.2 – *Graphique arithmétique de l'évolution.*

Nous voyons bien que, comme dans l'exemple du nombre d'agents, une représentation graphique classique (en échelle arithmétique) n'est pas adaptée vu l'écart de valeur entre 75 et 4500 : le début de la courbe est complètement écrasé. Nous allons donc faire un graphique semi-logarithmique et utiliser du papier semi-logarithmique.

b) Faisons maintenant un graphique semi-logarithmique. On prendra 1 cm pour 10 ans sur l'axe des abscisses et on gradue le premier module de 10 à 100 sur l'axe des ordonnées. Combien de modules sont nécessaires ?

La plus petite valeur étant 75 et la plus grande 4500, il nous faut 3 modules : le premier allant de 10 à 100, le second de 100 à 1000 et le troisième de 1000 à 10000.
Le 1 du premier module va donc représenter 10. Par conséquent, le 1 du second module représentera 100, le 1 du troisième module 1000... La dernière valeur du troisième module sera 9999.

Les données allant de 75 à 4500, trois modules seront donc suffisants. Ainsi toutes les données seront sur le graphe.

On reporte ensuite point par point les valeurs du phénomène étudié ; ici la population de New-York. Il faut prendre garde au deuxième module de ne pas confondre la valeur 120 avec la valeur 200.

On écrit les valeurs en face de chaque graduation choisie et on relie les points par des segments pour obtenir l'allure finale du phénomène.

Remarque 1.7. *Notons que sur ce graphique un module est représenté par 1,3 cm.*

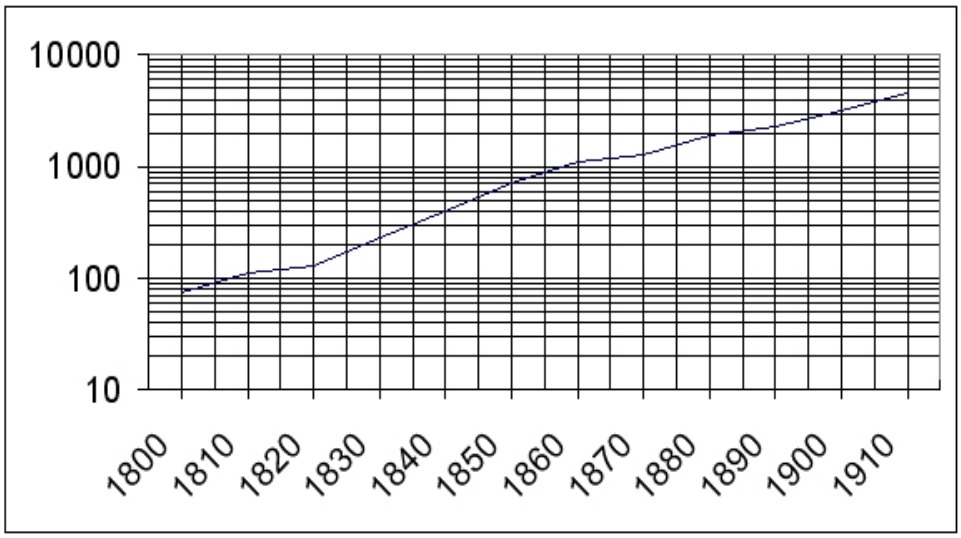


FIGURE 1.3 – *Graphique semi-logarithmique de l'évolution.*

c) Faire un graphique annexe représentant des augmentations sur 10 ans de 10%, 20%, 30%, 50% *et* 100%.

D'après les propriétés des graphiques semi-logarithmiques sur les variations relatives, il suffit de faire des flèches de même origine donnant la pente correspondant à ces augmentations. Plus précisément, comme $m = 1,8$ *cm, une augmentation de*

- 10% *sera représentée par une flèche de hauteur* $m * \log\left(1 + \frac{10}{100}\right) = 0,0538$ *cm pour 10 ans ;*
- 20% *sera représentée par une flèche de hauteur* $m * \log\left(1 + \frac{20}{100}\right) = 0,1029$ *cm pour 10 ans ;*
- 30% *sera représentée par une flèche de hauteur* $m * \log\left(1 + \frac{30}{100}\right) = 0,1481$ *cm pour 10 ans ;*
- 50% *sera représentée par une flèche de hauteur* $m * \log\left(1 + \frac{50}{100}\right) = 0,2289$ *cm pour 10 ans ;*
- 100% *sera représentée par une flèche de hauteur* $m * \log\left(1 + \frac{100}{100}\right) = 0,3913$ *cm pour 10 ans.*

Etant donné qu'il est difficile de réprésenter ces quantités de l'ordre de quelques millimètres, on les double et on obtient la hauteuir correspondante pour 20 ans. On allongera donc la flèche.

Quelles remarques peut-on faire en regardant la courbe semi-logarithmique sur l'évolution de la population de New York ?

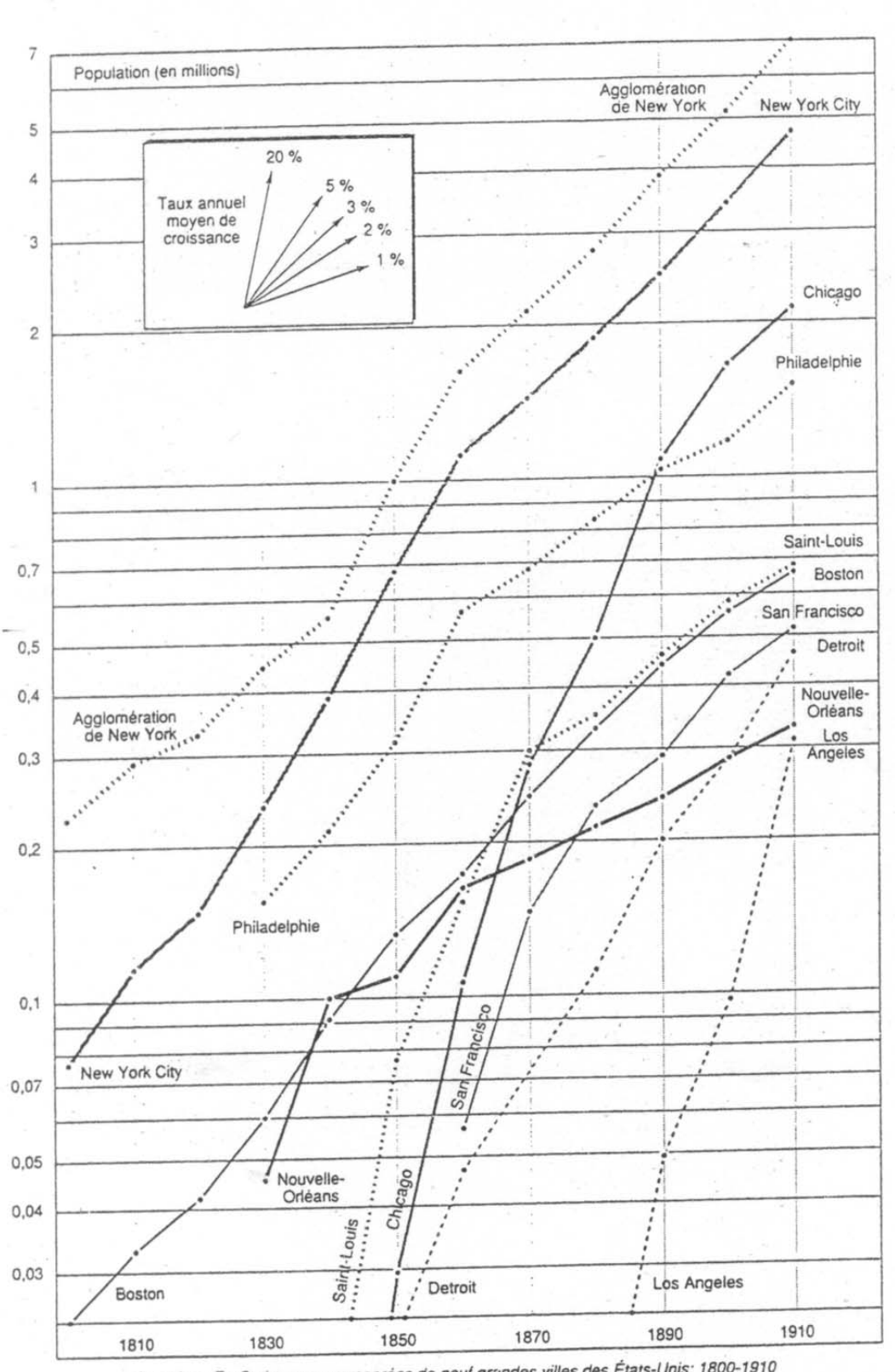
d) Faire un graphique annexe représentant les taux annuels moyens de croissance de 1%, 2%, 3%, 5% *et* 20%.
On fait la même chose et on multiplie par 20 les résultats de façon à les représenter plus aisément.

Le document de la page suivante permet de comparer les croissances de 9 grandes villes des Etats-Unis. La lecture d'un tel graphique peut prêter à confusion si l'on n'a pas compris le principe de construction et l'intérêt des courbes semi-logarithmiques. Pour bien comprendre le graphique, on peut par exemple reconstituer le tableau de données pour une des villes, mesurer le module du repère semi-logarithmique, mesurer la hauteur qui correspond à un doublement de la population... Par exemple essayons de refaire les calculs qui ont permis de construire le graphique annexe. Ici $m = 7,7$ *cm, une augmentation annuelle moyenne de*

- 1% *sera représentée par une flèche de hauteur* $m * \log\left(1 + \frac{1}{100}\right) = 0,0333$ *cm pour 1 an ;*
- 2% *sera représentée par une flèche de hauteur* $m * \log\left(1 + \frac{2}{100}\right) = 0,0662$ *cm pour 1 an ;*
- 3% *sera représentée par une flèche de hauteur* $m * \log\left(1 + \frac{3}{100}\right) = 0,0988$ *cm pour 1 an ;*
- 5% *sera représentée par une flèche de hauteur* $m * \log\left(1 + \frac{5}{100}\right) = 0,1632$ *cm pour 1 an ;*
- 20% *sera représentée par une flèche de hauteur* $m * \log\left(1 + \frac{20}{100}\right) = 0,6097$ *cm pour 1 an.*

Les flèches correspondant à la hauteur sur 20 ans, on multiplie par 20 ces quantités : une augmentation annuelle moyenne de

- 1% *sera représentée par une flèche de hauteur* $0,67$ *cm pour 20 ans ;*
- 2% *sera représentée par une flèche de hauteur* $1,32$ *cm pour 20 ans ;*
- 3% *sera représentée par une flèche de hauteur* $1,98$ *cm pour 20 ans ;*
- 5% *sera représentée par une flèche de hauteur* $3,26$ *cm pour 20 ans ;*
- 20% *sera représentée par une flèche de hauteur* $12,19$ *cm pour 20 ans.*



Graphique E - Croissances comparées de neuf grandes villes des États-Unis; 1800-1910

FIGURE 1.4 – *Graphique semi-logarithmique.*

Exemple 1.7. *Soient les données ci-après représentant le nombre de visiteurs ayant fréquenté un complexe touristique.*

Années	*Nb de visiteurs*	*Années*	*Nb de visiteurs*
1998	*350*	*2003*	*18500*
1999	*850*	*2004*	*22000*
2000	*1200*	*2005*	*95500*
2001	*2000*	*2006*	*74000*
2002	*20000*	*2007*	*60000*

Représentez ces données sur un diagramme semi-logarithmique.

1.3.4 Pour les curieux

Cette section n'est pas nécessaire à la compréhension des courbes semi-logarithmiques et peut être ignorée en première lecture.

Un peu d'Histoire

À la fin du XVI^e siècle, le développement de l'astronomie, de la navigation, du commerce,... oblige les savants à faire de longs et pénibles calculs. Or, il est clair qu'il est plus facile d'additionner que de multiplier. D'où l'idée de remplacer des nombres positifs a et b par des logarithmes tels que :

$$\log(ab) = \log(a) + \log(b).$$

Les premières tables de logarithmes sont dues à l'écossais John Neper (1550-1617). À sa suite, l'anglais Henri BRIGGS (1561-1631) eut l'idée d'utiliser le nombre 10 comme base et d'établir ainsi des tables de logarithmes décimaux, plus adaptés aux calculs numériques. Ces tables eurent un grand succès, tant elles répondaient aux besoins contemporains ; l'astronome allemand KEPLER (1571-1630) dédia ses Tabulae rudolphinae à Neper, ses calculs ayant été largement facilités par l'emploi des logarithmes. Aujourd'hui les calculatrices ont remplacé les tables de logarithmes.

Rappels sur la fonction logarithme népérien (ln)

La fonction ln associe à tout nombre strictement positif x le nombre $\ln x$ appelé **logarithme népérien** de x. Cette fonction ne sera pas étudiée en détail dans ce cours : c'est la bijection réciproque de la fonction exponentielle.

Cette fonction est croissante et tend vers l'infini quand x tend vers l'infini mais le "taux de variation" diminue (la dérivée de $\ln x$ est $1/x$ donc diminue quand x augmente) ; géométriquement,

la tangente à la courbe tend vers l'horizontale quand x augmente. Voici sa représentation :

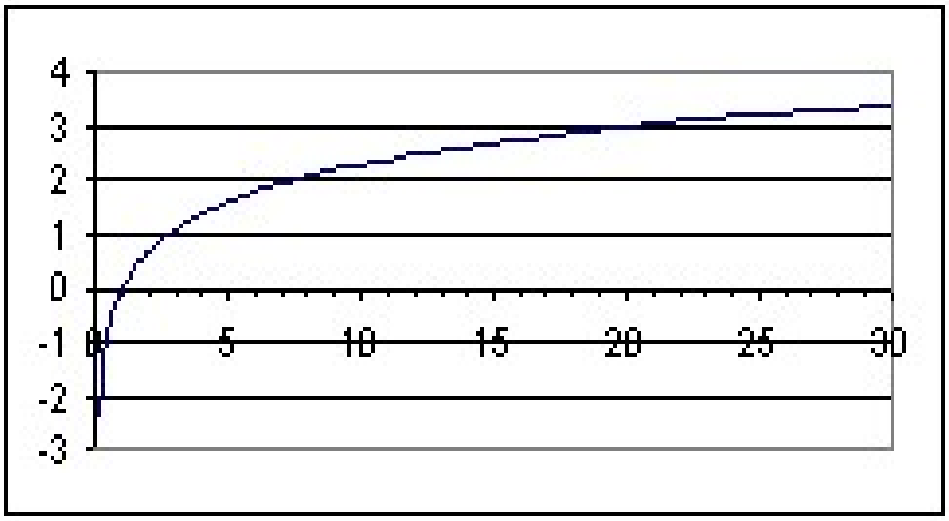


FIGURE 1.5 – *Courbe représentative de ln.*

Les deux propriétés fondamentales qui serviront dans la suite sont les suivantes :

$$\ln(xy) = \ln(x) + \ln(y) \tag{1}$$

$$\ln(x/y) = \ln(x) - \ln(y) \tag{2}$$

Le logarithme décimal

Pour tracer des courbes semi-logarithmiques, on s'intéressera à la fonction **logarithme décimal**, noté $\log(x)$, définie à partir du logarithme népérien par :

$$\log(x) = \ln(x)/\ln(10).$$

Remarquons qu'on a encore les propriétés (exercice) :

$$\log(xy) = \log(x) + \log(y) \tag{1}$$

$$\log(x/y) = \log(x) - \log(y) \tag{2}$$

Remarquons aussi que pour tout entier naturel n, on a $\log(10^n) = n$.

Construction d'un diagramme semi-logarithmique Dans ce paragraphe, nous allons construire à la main et à l'aide de la calculatrice un repère semi-logarithmique. Rappelons que le terme semi-logarithmique signifie que seul l'axe vertical est gradué en logarithmes, l'axe horizontal restant en graduations arithmétiques.

Grâce à la calculatrice, on calcule

- $\log 1 = 0$;
- $\log 2 = 0,30103$, que nous pouvons approcher par 0,30 ;
- $\log 3 = 0,47712$, que nous pouvons approcher par 0,48 ;
- $\log 4 = 0,60206$, que nous pouvons approcher par 0,60 ;
- et ainsi de suite jusqu'au log de 10 qui est tout simplement égal à 1.

Il suffit de reporter les chiffres après la virgule des ces logarithmes jusqu'à 10 sur une échelle arithmétique graduée régulièrement de 0 à 10 pour obtenir une conversion en échelle logarithmique comme ci-dessous :

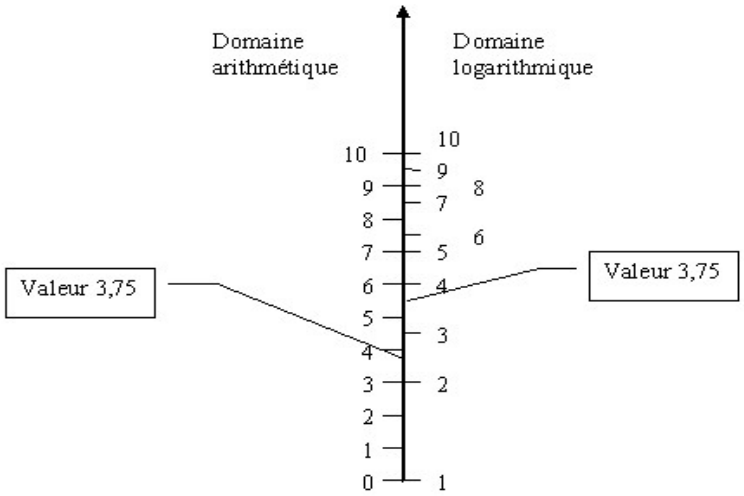


FIGURE 1.6 – *Un module.*

Nous venons de construire le premier module d'un graphique semi-logarithmique : il est gradué de 1 à 10. On voit que les graduations sont de plus en plus resserées à mesure qu'elles augmentent. Cela vient du fait que contrairement à l'échelle arithmétique où la différence entre deux graduations est constamment égale à l'unité, en échelle logarithmique cette différence est variable.

Si nous passons de 10 à 20 puis à 30 et ainsi de suite jusqu'à 100, nous obtiendrons exactement les mêmes chiffres après la virgule que ceux obtenus précédemment. Ainsi par exemple $\log 4 = 0,60206$, $\log 40 = 1,60206$ et aussi $\log 400 = 2,60206$. Dès lors nous pouvons continuer la graduation de l'échelle logarithmique pour les valeurs allant de 10 à 99 en reportant les chiffres après la virgule exactement comme pour le premier module. On obtient ainsi le second module. Et ainsi de suite pour les modules suivants.

Selon les ordres de grandeur des valeurs observées, on choisira le nombre de modules adéquat et la valeur du premier module.

Chapitre 2

Liaison linéaire entre deux variables quantitatives

Soient X et Y deux variables quantitatives définies sur une même population Ω. On considère le couple de variables (X, Y) dont les modalités sont les couples $(x_i \ ; \ y_i)$ où $x_i = X(\omega_i)$ et $y_i = Y(\omega_i)$ pour l'individu ω_i.

L'étude du couple (X, Y) a pour but, entre autres, de mettre en évidence l'existence d'une causalité, d'un lien entre les variables X et Y.

Nous avons vu au premier semestre comment déterminer s'il y avait un lien entre deux variables (avec le coefficient χ^2) et le cas échéant, de quantifier l'importance de ce lien (avec le coefficient ϕ).
Nous allons nous intéresser ici à un type de liaison particulier. On se propose d'examiner le lien **linéaire** entre X et Y et de répondre donc aux questions suivantes :

- Y a-t-il une liaison linéaire entre X et Y du type $Y = aX + b$?
- Si oui, quelles sont les valeurs de a et de b ?

Remarque 2.1. *Notons tout de suite qu'une liaison linéaire parfaite entre X et Y sera rarement vérifiée. On cherche plus exactement une relation du type :*

$$Y \ \simeq \ aX + b,$$

ou encore :

$$Y \ = \ aX + b + erreur,$$

où le terme d'erreur représente la "variabilité" autour du modèle.

Tout au long de ce chapitre, nous travaillerons sur les exemples suivants.

Exemple 2.1. *Dans une étude de docimologie, on demande à deux professeurs de corriger 5 copies (les mêmes).*

Population Ω : les 5 copies.

Variable X : Note du premier professeur ;

Variable Y : Note du deuxième professeur.

On obtient les résultats suivants :

N° de copie	*1*	*2*	*3*	*4*	*5*
X	*13*	*10*	*11*	*7*	*16*
Y	*12*	*9*	*11*	*8*	*14*

On veut ici étudier l'existence d'un lien particulier entre les notes des deux correcteurs : un lien linéaire qui s'exprime par le fait que plus une note est élevée avec le premier professeur et plus elle le sera avec le second professeur et vice versa pour les notes faibles. On espère bien sûr que ce lien linéaire existe et est fort ce qui traduira l'impartialité et la même exigence entre les deux professeurs.

Exemple 2.2. *Considérons un échantillon de 80 enfants sur lequel on a étudié la taille en cm et le poids en kg. Pour ces deux variables on a regroupé les données en classes :*

- pour la variable poids notée X on a utilisé les classes $]10;12]$, $]12;14]$, $]14;16]$, $]16;18]$.

- pour la variable taille notée Y on a utilisé les classes $]80;90]$, $]90;100]$, $]100;110]$.

Les résultats de l'enquête sont les suivants

Y \ X	$]10;12]$	$]12;14]$	$]14;16]$	$]16;18]$
$]80;90]$	*20*	*2*	*1*	*0*
$]90;100]$	*3*	*31*	*3*	*0*
$]100;110]$	*0*	*4*	*12*	*4*

On veut ici étudier l'existence d'un lien particulier (lien linéaire) entre la taille et le poids sur cet échantillon.

Exemple 2.3. *Dans une population Ω, on a prélevé un échantillon de $N = 8$ individus pour lesquels on a considéré l'âge du décès du père (variable X) et l'âge de leur décès (variable Y). Les résultats sont consignés dans le tableau ci-dessous.*

x_i	*77*	*50*	*54*	*62*	*83*	*62*	*34*	*66*
y_i	*74*	*42*	*68*	*66*	*81*	*79*	*44*	*62*

On veut ici étudier l'existence d'un lien particulier (lien linéaire) entre l'âge du décès du père et celui de son enfant sur cet échantillon.

Remarque 2.2. *Avant de continuer, faisons quelques remarques.*

1. *Tout d'abord, notons que pour le premier et le troisième exemples, nous avons dressé le tableau des données résultant de l'enquête tandis que pour le deuxième exemple, nous disposons du tableau des effectifs. Ce dernier tableau est la conséquence d'un premier traitement des données. Il permet une meilleure lecture mais conduit à une perte d'information : nous ne connaissons plus les réponses de chacun des individus mais seulement le nombre d'individus par modalités.*
2. *Le tableau d'effectifs de l'exemple 2 est aussi appelé* ***table de contingence*** *car nous avons mis ensemble les deux variables (contingence signifie mis ensemble).*
3. *Nous pouvons à partir de là calculer les marge de X et de Y en ajoutant une colonne et une ligne au tableau précédent. Ces marges consituent les effectifs par modalité de chacune des variables X et Y. Elles sont aussi appelées* ***distributions marginales****. Ainsi*

Y \ X	]10; 12]	]12; 14]	]14; 16]	]16; 18]	*Marge de X*
]80; 90]	*20*	*2*	*1*	*0*	*23*
]90; 100]	*3*	*31*	*3*	*0*	*37*
]100; 110]	*0*	*4*	*12*	*4*	*20*
Marge de Y	*23*	*37*	*16*	*4*	$N = 80$

De manière plus générale, X et Y étant 2 variables quantitatives sur un échantillon E, on veut étudier l'existence d'une liaison particulière (appelée corrélation linéaire) entre X et Y.

2.1 Le nuage de points

Un moyen "rapide" de vérifier (avant tout calcul) si une liaison linéaire entre X et Y est plausible est de représenter le **nuage de points**. Il s'agit de représenter graphiquement les données concernant les deux variables quantitatives X et Y. On considère le plan muni d'un repère orthogonal (les axes sont perpendiculaires). Pour cela on gradue deux axes, l'un horizontal gradué suivant les modalités de X et l'autre vertical suivant les modalités de Y. On place alors un point A_i correspondant à chaque couple (x_i, y_i) de modalités observé. La grosseur du point permet de représenter de façon figurative l'effectif n_i de chaque couple : concrètement, plus le point est gros, plus l'effectif l'est également.

• L'ensemble des points A_i de coordonnées (x_i, y_i) s'appelle le **nuage de points associé au couple** (X, Y) .

• Notons $\overline{X}$ la moyenne des valeurs x_i et $\overline{Y}$ la moyenne des valeurs y_i. Le point M de coordonnées $(\overline{X}, \overline{Y})$ s'appelle le **point moyen** du nuage.

Exemple 2.1 (suite) : *on obtient les points* $A_1(13;12)$, $A_2(10;9)$, $A_3(11;11)$, $A_4(7;8)$ *et* $A_5(16;14)$, *que l'on place dans un repère orthogonal.*

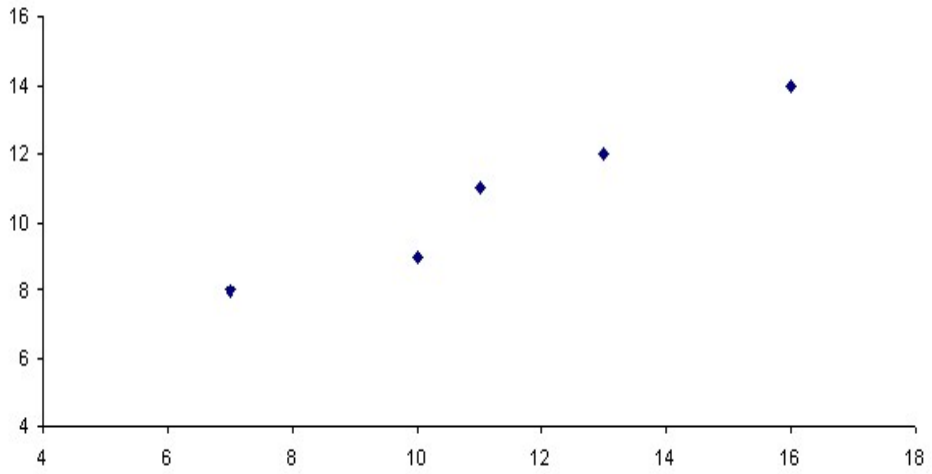


FIGURE 2.1 – *Le nuage de points pour la docimologie*

Les points ne sont pas alignés mais le nuage a une forme allongée. On va essayer de tracer une droite qui passe "aussi près que possible" de ces points. On peut également calculer les coordonnées du point moyen : le point moyen est ici $M(11,4;10,8)$ puisque

$$\overline{X} = \frac{13+10+11+7+16}{5} = 11,4$$

et

$$\overline{Y} = \frac{12+9+11+8+14}{5} = 10,8.$$

Exemple 2.2 (suite) : *Le nuage de points est le suivant :*

En observant ce nuage de points, on remarque une tendance : plus la taille augmente, plus le poids augmente également. On a donc un nuage "croissant" dont la tendance est matérialisée par

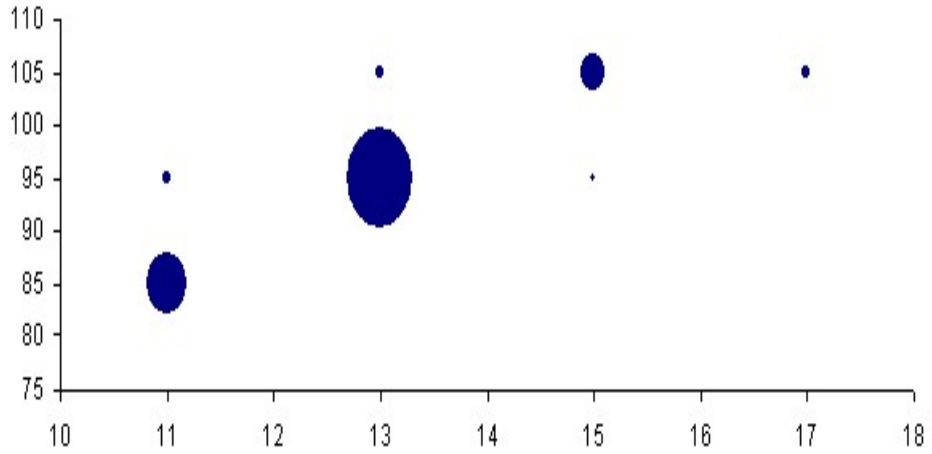


FIGURE 2.2 – *Le nuage de points pour la relation poids/taille*

une droite croissante passant parmi les points du nuage.

Exemple 2.3 (suite) : *Le nuage de points est représenté dans la figure ci-dessous.*

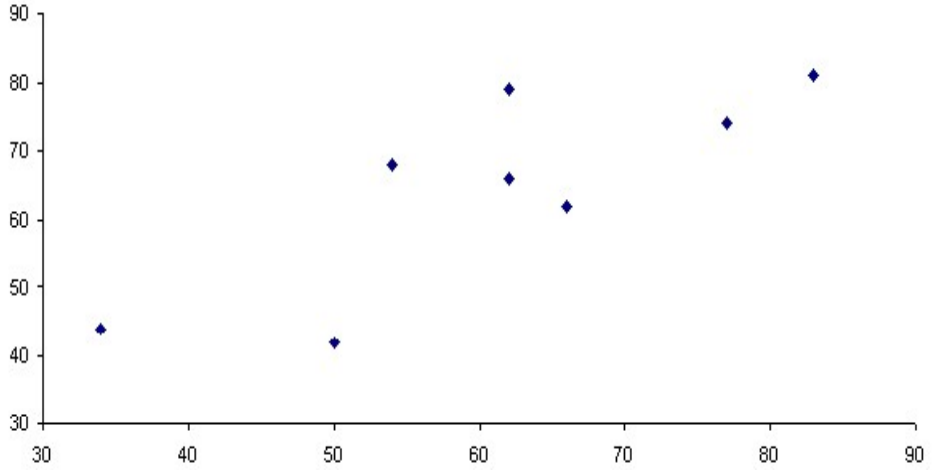


FIGURE 2.3 – *Le nuage de points pour les décès*

On voit que le nuage de points ci-dessus est relativement "étiré" ce qui laisse supposer l'existence d'une liaison linéaire forte entre l'âge du décès du père et l'âge du décès du fils.

On veut maintenant déterminer de façon quantitative s'il y a un lien linéaire entre X et Y. On va donc définir un indice qui permet de "mesurer" l'alignement des points du nuage. On verra ensuite dans le cas où les points du nuage sont "presque" alignés comment déterminer la droite qui passe au plus près des points du nuage.

2.2 Le coefficient de corrélation linéaire

On commence par définir une nouvelle quantité : la covariance qui va intervenir dans les calculs du coefficient de corrélation linéaire.

2.2.1 La covariance

On appelle **covariance** de X et Y et on note $\text{Cov}(X,Y)$ le nombre défini par :

$$\text{Cov}(X,Y) = \frac{1}{N}\sum n_i(x_i - \overline{X})(y_i - \overline{Y}).$$

De même que pour la variance, l'expression qui sert à définir la covariance n'est pas la plus pratique à utiliser pour les calculs. En développant l'expression ci-dessus de $\text{Cov}(X,Y)$, on obtient le résultat suivant :

$$\boxed{\text{Cov}(X,Y) = \left(\frac{1}{N}\sum n_i x_i y_i\right) - \overline{XY}}$$

C'est ce résultat qui sera utilisé pour les calculs.

Cette expression ressemble à celle de la variance mais il est ici tenu compte des écarts entre les données et leur moyenne, à la fois pour X et pour Y.

Exemple 2.1 (suite) : *Calculons* $\text{Cov}(X,Y)$ *en utilisant les deux formules.*

$$\begin{aligned}
\text{Cov}(X,Y) &= \frac{1}{5}\sum_{i=1}^{5}(x_i - 11,4)(y_i - 10,8) \\
&= \frac{1}{5}\left[(13-11,4)(12-10,8) + \cdots + (16-11,4)(14-10,8)\right] \\
&= \frac{1}{5} \times 31,4 = 6,28
\end{aligned}$$

$$\begin{aligned}
\text{Cov}(X,Y) &= \frac{1}{5}\sum_{i=1}^{5} x_i y_i - 11,4 \times 10,8 \\
&= \frac{1}{5}(13 \times 12 + \cdots + 14 \times 16) - 11,4 \times 10,8 \\
&= 129,4 - 123,12 = 6,28
\end{aligned}$$

Exemple 2.2 (suite) : *Déterminons* $\mathrm{Cov}(X,Y)$. *Pour cela il faut d'abord effectuer un certain nombre de calculs. Pour calculer moyennes et écart-types, on doit d'abord remplacer chaque classe par son centre. Ensuite, plutôt que de construire le tableau d'effectifs de* X *puis celui de* Y*, on va construire le tableau d'effectifs du couple* (X,Y) *qui permettra de faire tous les calculs avec un seul tableau.* ***Chaque ligne du tableau ci-dessous correspond donc à une case de la table de contingence, mais il est inutile de représenter les couples de modalités dont l'effectif est nul.***

Ainsi, le couple $(85,11)$ *a pour effectif 20, car 20 individus ont une taille dans la classe* $]80;90]$ *et un poids dans la classe* $]10;12]$. *Le couple* $(85;17)$ *a pour effectif 0 et ne figure donc pas dans le tableau ci-dessous.*

y_i	x_i	n_i	$n_i y_i$	$n_i(y_i^2)$	$n_i x_i$	$n_i(x_i^2)$	$n_i x_i y_i$
85	11	20	1700	144500	220	2420	18700
85	13	2	170	14450	26	338	2210
85	15	1	85	7225	15	225	1275
95	11	3	285	27075	33	363	3135
95	13	31	2945	279775	403	5239	38285
95	15	3	285	27075	45	675	4275
105	13	4	420	44100	52	676	5460
105	15	12	1260	132300	180	2700	18900
105	17	4	420	44100	68	1156	7140
		80	7570	720600	1042	13792	99380

On a rajouté au tableau les 3 colonnes permettant de calculer les moyennes et la covariance de X *et* Y *(et deux colonnes supplémentaires qui sera utilisée au paragraphe suivant pour le calcul des variances). On obtient donc les résultats suivants :*

$$\overline{Y} = \frac{7570}{80} = 94,63 \qquad \mathrm{Var}(Y) = \frac{720600}{80} - \left(\frac{7570}{80}\right)^2 = 53,61 \qquad \sigma_Y = 7,32$$

$$\overline{X} = \frac{1042}{80} = 13,03 \qquad \mathrm{Var}(X) = \frac{13792}{80} - \left(\frac{1042}{80}\right)^2 = 2,75 \qquad \sigma_X = 1,66$$

$$\mathrm{Cov}(X,Y) = \frac{99380}{80} - \frac{1042}{80}\frac{7570}{80} = 9,76$$

Exemple 2.3 (suite) : *Déterminons* $\mathrm{Cov}(X,Y)$ *en nous aidant d'un tableau*

x_i	*77*	*50*	*54*	*62*	*83*	*62*	*34*	*66*	*488*
y_i	*74*	*42*	*68*	*66*	*81*	*79*	*44*	*62*	*516*
x_i^2	*5929*	*2500*	*2916*	*3844*	*6889*	*3844*	*1156*	*4356*	*31434*
y_i^2	*5476*	*1764*	*4624*	*4356*	*6561*	*6241*	*1936*	*3844*	*34802*
$x_i \times y_i$	*5698*	*2100*	*3672*	*4092*	*6723*	*4898*	*1496*	*4092*	*32771*

On a ainsi

$$\overline{X} \;=\; \frac{488}{8} \;=\; 61, \qquad \overline{Y} \;=\; \frac{516}{8} \;=\; 64,5$$

et on en déduit :

$$\mathrm{Cov}(X,Y) \;=\; \frac{32771}{8} - 61 \times 64,5 \;=\; 161,88$$

Propriétés

a) On vérifie facilement que $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

b) La covariance peut prendre des valeurs positives ou négatives contrairement à la variance qui est toujours positive ou nulle.

c) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$; on dit que la covariance est **symétrique**.

d) **Transformation affine des données**. Soient a, b, c et d quatre nombres réels quelconques. Posons $Z = aX + b$ et $T = cY + d$. On a alors :

$$\mathrm{Cov}(Z,T) \;=\; a \times c \times \mathrm{Cov}(X,Y).$$

En particulier, si on prend $a = 1$ et $c = 1$, on voit que la covariance est invariante par translation :

$$\mathrm{Cov}(X+b, Y+d) \;= \mathrm{Cov}(X,Y).$$

Comme dans le cas de la moyenne et de la variance ces relations peuvent s'avérer utiles pour simplifier les calculs de la covariance par changement de variables.

e) **Inégalité de Cauchy-Schwarz**. On a l'inégalité suivante liant la covariance et les variances de X et Y (démonstration admise) :

$$\mathrm{Cov}^2(X,Y) \;\leq \sigma_X^2 \sigma_Y^2,$$

ou encore :

$$|\mathrm{Cov}(X,Y)| \;\leq \sigma_X \sigma_Y.$$

2.2.2 Le coefficient de corrélation linéaire

On appelle **coefficient de corrélation linéaire** ou **coefficient de Bravais Pearson**, le nombre noté $r(X,Y)$ ou $r_{X,Y}$ tel que :

$$\boxed{r(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y}.}$$

Propriétés :

a) Le coefficient de corrélation linéaire est symétrique :

$$r(X,Y) \ = r(Y,X).$$

b) L'inégalité de Cauchy-Schwarz donne :

$$-1 \ \leq \ r(X,Y) \ \leq 1.$$

c) **Transformation affine des données**. Soient a, b, c et d quatre nombres réels quelconques ($a \neq 0$ et $c \neq 0$). Posons $Z = aX + b$ et $T = cY + d$. On a alors :

$$r(Z,T) \ = \ \begin{cases} r(X,Y), \text{ si } a \text{ et } c \text{ sont de même signe,} \\ -r(X,Y), \text{ si } a \text{ et } c \text{ sont de signes opposés.} \end{cases}.$$

En particulier, pour $a = c = 0$, on voit que le coefficient de corrélation linéaire est invariant par translations et pour $b = d = 0$, il est invariant au signe près par homothéties.
On peut à nouveau utiliser ces relations pour simplifier les calculs du coefficient de corrélation linéaire.

Exemple 2.1 (suite) : *on a*

$$\begin{aligned}
\mathrm{Var}(X) &= \frac{1}{5}\sum_{i=1}^{5} y_i^2 - \bar{Y}^2 = \frac{1}{5}(13^2 + 10^2 + 11^2 + 7^2 + 16^2) - 11,4^2 = 9,04 \\
\mathrm{Var}(Y) &= \frac{1}{5}\sum_{i=1}^{5} y_i^2 - \bar{Y}^2 = \frac{1}{5}(12^2 + 9^2 + 11^2 + 8^2 + 14^2) - 10,8^2 = 4,56 \\
\sigma_X &= \sqrt{9,04} = 3,0067 \\
\sigma_Y &= \sqrt{4,56} = 2,1354 \\
r(X,Y) &= \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{6,28}{3,0067 \times 2,1354} = 0,978
\end{aligned}$$

Attention : en arrondissant grossièrement à 3 et 2, on obtient $r(X,Y) > 1$ ce qui est impossible.

Il y a donc une forte corrélation linéaire positive entre X et Y (heureusement) et donc dans notre exemple, il est intéressant de faire un ajustement affine.

Exemple 2.2 (suite) : *Grâce au tableau, les variances et écarts-type sont*

$$\mathrm{Var}(Y) = \frac{720600}{80} - \left(\frac{7570}{80}\right)^2 = 53,61 \qquad \sigma_Y = 7,32$$

$$\mathrm{Var}(X) = \frac{13792}{80} - \left(\frac{1042}{80}\right)^2 = 2,75 \qquad \sigma_X = 1,66$$

Déterminons $r(X,Y)$: on trouve $r(X,Y) = \dfrac{9,76}{7,32 \times 1,66} = 0,80.$

Exemple 2.3 (suite) : *Déterminons* $\mathrm{Cov}(X,Y)$ *: on a*

$$\begin{aligned} \sigma_X^2 &= \frac{31434}{8} - 61^2 = 208,25 \\ \sigma_X &= 14,43 \\ \sigma_Y^2 &= \frac{34802}{8} - 64,5^2 = 190 \\ \sigma_Y &= 13,78. \end{aligned}$$

On en déduit :

$$r(X,Y) = \frac{161,88}{14,43 \times 13,78} = 0,81$$

On avait vu que le nuage de points ci-dessus est relativement "étiré" ce qui confirme, avec la valeur de $r(X,Y) = 0,81$ trouvée ci-dessus, qu'il existe une liaison linéaire forte entre l'âge du décès du père et l'âge du décès du fils.

Signification

(i) Examinons ce qu'il se passe si $r(X,Y) = \pm 1$. On a alors :

$$\mathrm{Cov}(X,Y)^2 = \sigma_X^2 \sigma_Y^2.$$

On pourrait alors montrer que les variables X et Y sont liées par une relation du type :

$$Y = aX + b,$$

où a et b sont deux réels quelconques. Cela signifie qu'il existe une **liaison linéaire parfaite** entre X et Y. De plus si $r(X,Y) = 1$ alors a est strictement positif et si $r(X,Y) = -1$ alors a

est strictement négatif.

(ii) Par ailleurs, si $r(X,Y) = 0$ (ce qui signifie que $\mathrm{Cov}(X,Y) = 0$), il n'existe aucune forme de liaison linéaire entre X et Y.

(iii) Enfin, en dehors de ces valeurs, $|r(X,Y)|$ est d'autant plus proche de 1 que la liaison linéaire entre X et Y est grande.

Remarque 2.3. *1) Il est important de noter que $r(X,Y)$ mesure la liaison* **linéaire** *entre X et Y : on peut avoir $r(X,Y) = 0$ et pourtant avoir une forte liaison entre X et Y (ces variables peuvent être liées par un autre type de liaison, par exemple quadratique,...).*
2) Notons que dans certains cas on peut avoir un coefficient de corrélation linéaire relativement proche de 1 (par exemple de l'ordre de 0,75) et un nuage de points assez éloigné d'une droite (par exemple en forme de "T"). Dans ces cas, on cherche une autre forme de liaison entre X et Y ou on sépare la population en deux sous-populations, ... On voit ainsi l'importance de l'observation du nuage de points : on supposera l'existence d'une liaison linéaire entre X et Y lorsque celui-ci sera étiré (proche d'une droite) et que le coefficient de corrélation linéaire sera proche de 1.

2.3 Relation entre $r(X,Y)$ et le coefficient φ

2.3.1 Rappels sur le coefficient φ

Afin de déterminer l'existence d'un lien (quelconque) entre deux variables X et Y, on calcule le coefficient χ^2

$$\chi^2 = \Sigma \frac{(O_{ij} - T_{ij})^2}{T_{ij}},$$

où

- les O_{ij} sont les effectifs observés

- les T_{ij} sont les effectifs théoriques que l'on doit obtenir s'il y a indépendance entre les deux variables. L'expression de T_{ij} est donnée par

$$T_{ij} = \frac{L_i \times C_j}{N}$$

avec N la taille de la population, L_i l'effectif marginal de la modalité i de X et C_j l'effectif marginal de la modalité Y.

Lorsque χ^2 est proche de zéro, on conclut à l'indépendance des deux variables. Dans le cas contraire, on peut déterminer l'importance du lien en calculant le coefficient φ donné par

$$\varphi = \sqrt{\frac{\chi^2}{N\,[\min(p,q)-1]}},$$

où

- p est le nombre de lignes de la table de contingence de X et Y (ou nombre de modalités de X) ;

- q est le nombre de colonnes de la table de contingence de X et Y (ou nombre de modalités de Y) ;

- $\min(p,q)$ est simplement le plus petit des deux nombres p et q.

Convention : *On considère en général que le lien est*

- *faible lorsque φ est inférieur à 0,3 ;*

- *moyen lorsque φ est compris entre 0,3 et 0,5 ;*

- *fort lorsque φ est supérieur à 0,5.*

Exemple : *On étudie la variable X qui associe à chaque individu son intérêt pour la lecture, avec les modalités : "Fort", "Moyen", "Faible" et Y celle qui associe à chaque individu sa spécialité scolaire, sachant que seules les modalités " L ", " ES " et " S " ont été observées.*
Voici le tableau des effectifs conjoints observés

X \ Y	*L*	*ES*	*S*	*Marge de X*
Fort	*10*	*1*	*1*	*12*
Moyen	*0*	*4*	*0*	*4*
Faible	*0*	*0*	*4*	*4*
Marge de Y	*10*	*5*	*5*	*N=20*

Après quelques calculs, on obtient le tableau des effectifs conjoints théoriques suivant :

X \ Y	*L*	*ES*	*S*	*Marge de X*
Fort	*6*	*3*	*3*	*12*
Moyen	*2*	*1*	*1*	*4*
Faible	*2*	*1*	*1*	*4*
Marge de Y	*10*	*5*	*5*	*N=20*

On calcule maintenant χ^2 en s'appyant sur le tableau suivant :

O_{ij}	T_{ij}	$(O_{ij}-T_{ij})^2$	$\frac{(O_{ij}-T_{ij})^2}{T_{ij}}$
10	*6*	*16*	*2,67*
0	*2*	*4*	*2*
0	*2*	*4*	*2*
1	*3*	*4*	*1,33*
4	*1*	*9*	*9*
0	*1*	*1*	*1*
1	*3*	*4*	*1,33*
0	*1*	*1*	*1*
4	*1*	*9*	*9*
$N = 20$	$N = 20$	-	$\chi^2 = 29,33$

La somme de la dernière colonne nous donne donc la valeur du $\chi^2 = 29,33 \neq 0$ *; et enfin,*

$$\varphi = \sqrt{\frac{\chi^2}{N\,[\min(p,q)-1]}} = \sqrt{\frac{29,33}{40}} = 0,86.$$

0,86 étant proche de 1, on peut conclure à l'existence d'un lien fort entre X *et* Y *sur* Ω *c'est-à-dire entre l'intérêt pour la lecture et la spécialité scolaire pour les 20 individus de l'enquête.*

2.3.2 Relation entre $r(X,Y)$ et φ sur des exemples

Examinons l'exemple suivant : sur un échantillon de 100 personnes, on considère la variable âge (avec pour regroupement en classes]5 ;35],]35 ;65],]65 ;95]) et la variable nombre de frères et sœurs. Les observations sont rassemblées dans le tableau ci-dessous.

Nb. frères/sœurs \ Classe de X	]5 ;35]	]35 ;65]	]65 ;95]
0	0	30	0
1	60	0	0
2	0	0	10

On se pose toujours la même question : existe-t-il un lien entre l'âge et le nombre de frères et sœurs pour ces 100 personnes ?

On remarque aisément qu'il y a un lien total en X et Y puisque à chaque classe d'âge correspond une et une seule possibilité pour le nombre de frères et sœurs. Ceci se traduit par $\varphi = 1$.

Remarque 2.4. *Dans cet exemple précis, puisqu'il y a beaucoup d'effectifs conjoints nuls, les données auraient pu être représentées par le tableau suivant (qui ne contient que les effectifs conjoints non nuls)*

Classe de X	*]5 ;35]*	*]35 ;65]*	*]65 ;95]*
y_i	*1*	*0*	*2*
n_i	*60*	*30*	*10*

plus concis mais aussi plus difficile à interpréter.

Étudions maintenant l'existence d'une corrélation linéaire entre X et Y. Commençons par représenter le nuage de points.

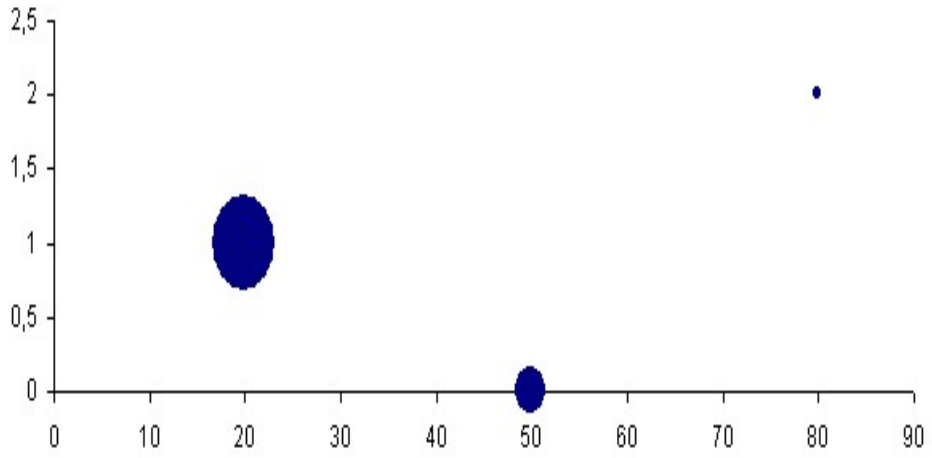


FIGURE 2.4 – *Le nuage de points pour la relation entre l'âge et le nombre de frères et sœurs*

x_i	y_i	n_i	$n_i x_i$	$n_i y_i$	$n_i x_i y_i$	$n_i x_i^2$
20	1	60	1200	60	1200	24000
50	0	30	1500	0	0	75000
80	2	10	800	20	1600	64000
		100	3500	80	2800	163000

$$\bar{X} = \frac{3500}{100} = 35 \quad \bar{Y} = \frac{80}{100} = 0,8 \quad \mathrm{Cov}(X,Y) = \frac{2800}{100} - 35 \times 0,8 = 0$$

d'où $r(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = 0$. Il est donc inutile ici de calculer les écart-types.

On a donc indépendance linéaire entre X et Y pour ces 100 personnes.

En résumé, on a donc un lien total entre X et Y ($\varphi = 1$) mais ce lien n'est pas du tout linéaire puisqu'il y a indépendance linéaire ($r(X,Y) = 0$).

Autre situation

Si dans une autre situation, on trouve $r(X,Y)$ valant -1 ou 1, on a donc une corrélation linéaire totale entre X et Y, il y a donc un lien total entre X et Y et donc $\varphi = 1$.

Conclusion

En fait, φ et $r(X,Y)$ apportent des informations différentes sur X et Y : φ nous dit qu'il y a un lien entre X et Y et $r(X,Y)$ nous renseigne sur la nature de ce lien : c'est une corrélation linéaire.

2.4 La régression linéaire

Lorsqu'une liaison linéaire forte entre deux variables X et Y a été mise à jour, on a alors une relation du type :

$$Y \simeq aX + b,$$

où les coefficients a et b sont inconnus. Le problème est donc de proposer des valeurs pour ces coefficients ou en d'autres termes de les **estimer** au vu des résultats obtenus sur l'échantillon. Si les points du nuage sont parfaitement alignés (sur une même droite), il serait facile de donner des valeurs à a et b : il suffirait en effet de prendre pour a la pente de la droite sur laquelle se trouvent les points du nuage et pour b la valeur en $x = 0$ (la solution se trouve en résolvant un système de deux équations à deux inconnues à partir de deux points du nuage). Le problème est que les points du nuage sont rarement (parfaitement) alignés : ils sont "proches" d'une droite. Une méthode pour estimer les valeurs de a et b est la méthode des **moindres carrés** décrite à la section suivante.

2.4.1 La méthode des moindres carrés

Reprenons l'exemple 2.2 concernant le poids et la taille et son nuage de points.

Nous allons mesurer l'éloignement des points du nuage par rapport à une droite Δ.

Rappel :

1) On rappelle qu'une équation de droite donne la relation entre l'abscisse (lue horizontalement) et l'ordonnée (lue verticalement) d'un point de la droite. Ainsi pour une droite d'équation

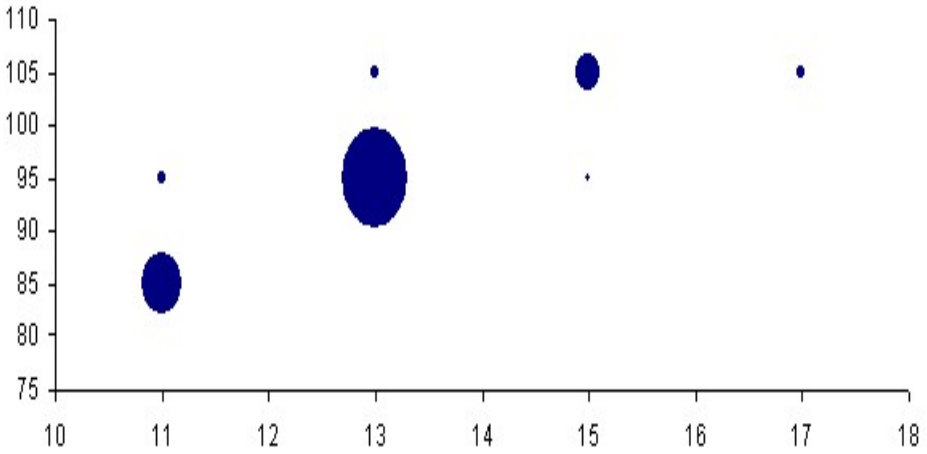


FIGURE 2.5 – *Le nuage de points pour la relation poids/taille*

$y = ax + b$, le point de la droite d'abscisse x_i aura pour ordonnée $ax_i + b$.

2) Le nombre a est appelé la **pente de la droite** ou encore **coefficient directeur** de la droite car il détermine la direction (la pente) de la droite. Lorsque a est positif la droite est croissante, lorsque a est négatif la droite est décroissante.

Le nombre b s'appelle l'**ordonnée à l'origine** car c'est l'ordonnée du point de la droite d'abscisse 0 (intersection de la droite avec l'axe des ordonnées).

3) Deux points (et une règle) suffisent pour tracer une droite. Pour représenter une droite lorsqu'on connaît son équation, il suffit de placer deux points (par exemple les points de coordonnées $(0; b)$ et $(1; a + b)$) puis tracer la droite passant par ces 2 points.

Détermination de la droite de régression :

Pour la droite tracée sur le graphique ci-dessus, d'équation $y = ax + b$, on désire mesurer son éloignement par rapport aux points du nuage.

Pour cela, pour chaque point A_i du nuage d'abscisse x_i et d'ordonnée y_i on mesure l'écart vertical entre le point et la droite. Le point situé sur la droite à la même verticale que A_i a pour abscisse x_i, il aura donc pour ordonnée $ax_i + b$. Ainsi l'écart vertical entre le point A_i et la droite sera $[y_i - (ax_i + b)]$. Afin de n'avoir que des quantités positives, on va élever au carré chacun de ces écarts, ce qui donne $[y_i - (ax_i + b)]^2$. Afin d'obtenir un seul nombre représentant globalement l'éloignement du nuage par rapport à la droite Δ, on prend comme d'habitude la moyenne de

ces carrés d'écarts. Cela donne une quantité qu'on appellera A. On a donc

$$A = \frac{1}{N}\sum_{i=1}^{N} n_i \left[y_i - (ax_i + b)\right]^2 .$$

On cherche maintenant la droite qui passe au plus près des points du nuage : celle pour laquelle la quantité A (c'est-à-dire la moyenne des carrés des écarts verticaux par rapport aux points du nuage) sera la plus petite ; c'est pourquoi lorsqu'on souhaite faire une approximation d'un nuage de points par une droite on parle d'**ajustement affine par la méthode des moindres carrés**. Cette droite est appelée **droite de régression de Y en X** et est notée : $\Delta_{Y/X}$.
Un calcul assez compliqué (et qui ne sera pas détaillé ici) montre que cette droite a pour équation

$$y = ax + b \quad \text{avec} \quad a = \frac{\text{Cov}(X,Y)}{\sigma_X^2} \quad \text{et} \quad b = \bar{Y} - a\bar{X}$$

Toujours par un calcul non détaillé ici, on obtient la valeur de A pour cette droite :

$$A = \sigma_Y^2 \left[1 - r(X,Y)^2\right] .$$

Pour notre exemple on obtient :

$$a = \frac{\text{Cov}(X,Y)}{\sigma_X^2} = \frac{9,76}{2,75} = 3,55 \quad \text{et} \quad b = \bar{Y} - a\bar{X} = 94,63 - 3,55 \times 13,03 = 48,39$$

et $\Delta_{Y/X}$ a donc pour équation $y = 3,55x + 48,39$.

Rappelons ici le résultat obtenu concernant $\Delta_{Y/X}$; ce résultat est à retenir : il sera utilisé au paragraphe suivant :

> La droite de régression de Y en X notée $\Delta_{Y/X}$ a pour équation
> $y = ax + b \quad \text{avec} \quad a = \frac{\text{Cov}(X,Y)}{\sigma_X^2} \quad \text{et} \quad b = \bar{Y} - a\bar{X}$

Propriétés :

1) Cette droite passe par le point moyen $M(\bar{X};\bar{Y})$. Puisqu'il suffit de deux points pour tracer une droite, on pourra, pour tracer $\Delta_{Y/X}$, placer les points $B(0;b)$ et $M(\bar{X};\bar{Y})$

2) Le coefficient directeur a de $\Delta_{Y/X}$, $\text{Cov}(X,Y)$ et $r(X,Y)$ sont de même signe :

- Lorsqu'ils sont positifs, on parle de **corrélation positive** (y augmente quand x augmente).
- Lorsqu'ils sont négatifs, on parle de **corrélation négative** (y diminue quand x augmente).

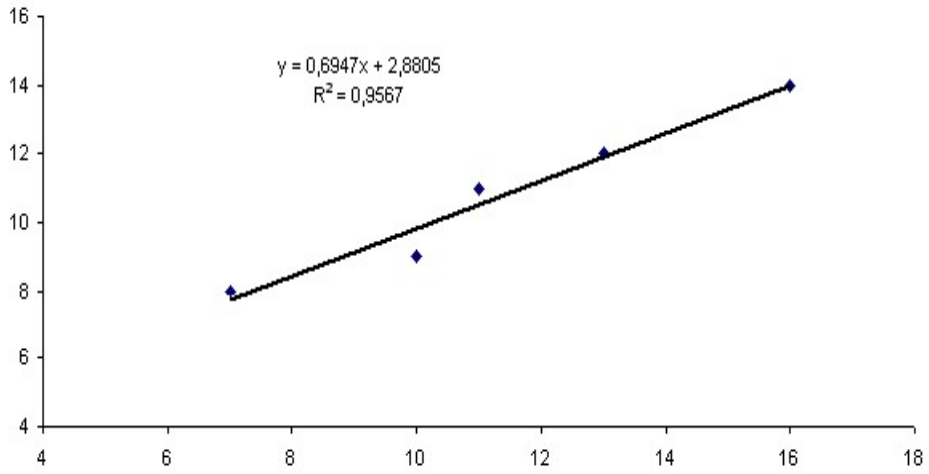
3) En pratique, il faut commencer par tracer le nuage de points puis calculer $r(X,Y)$ et ce n'est que si la corrélation linéaire est assez forte que l'on cherchera la droite de régression de Y en X.

Exemple 2.1 (suite) : *calculons l'équation de la droite de régression de Y en X et traçons cette*

droite. Nous avons déjà calculé $\bar{X}$, $\bar{Y}$, $\mathrm{Cov}(X,Y)$ *et* $\mathrm{Var}(X)$. *Calculons* a *et* b *:*

$$a = \frac{6,28}{9,04} = 0,69 \qquad b = 10,8 - 0,69 \times 11,4 = 2,88.$$

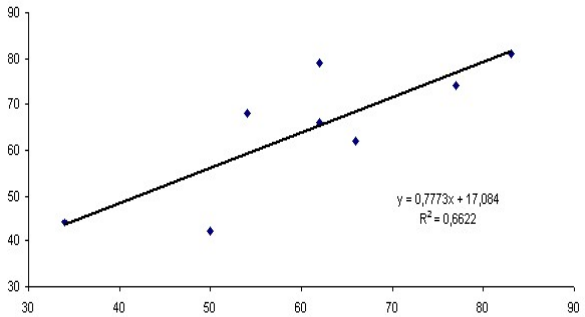
L'équation de $\Delta_{Y/X}$ *est :* $y = 0,69x + 2,88$.



Exemple 2.3 (suite) : *calculons l'équation de la droite de régression de* Y *en* X *et traçons cette droite. Nous avons déjà calculé* $\bar{X}$, $\bar{Y}$, $\mathrm{Cov}(X,Y)$ *et* $\mathrm{Var}(X)$. *Calculons* a *et* b *:*

$$a = \frac{161,88}{208,25} = 0,78 \qquad b = 64,5 - 0,78 \times 61 = 17,08.$$

L'équation de $\Delta_{Y/X}$ *est :* $y = 0,78x + 17,08$.



2.4.2 Propriétés et interprétation du coefficient de corrélation linéaire

De par sa construction A est toujours positif, or $A = \sigma_Y^2 \left[1 - r(X,Y)^2\right]$. On retrouve que

$$-1 \leq r(X,Y) \leq 1.$$

Ce résultat peut être utile pour détecter une erreur de calcul.

La valeur de A sera nulle lorsque $r(X,Y)$ vaudra -1 ou 1 et sera maximale lorsque $r(X,Y)$ vaudra 0. D'où les interprétations de $r(X,Y)$.

Si $r(X,Y) = 0$.
Dans ce cas l'éloignement des points du nuage avec la droite de régression de Y en X est maximal. On dira alors que X et Y sont **linéairement indépendants**.

Si $r(X,Y) > 0$.
Dans ce cas la droite de régression de Y en X est croissante ; on parle alors de **corrélation linéaire croissante** entre X et Y. Lorsque $r(X,Y)$ est proche de 1, A est proche de 0, les points du nuage sont donc presque alignés, on a donc une forte corrélation linéaire croissante (ou positive) entre X et Y.

Dans le cas extrême $r(X,Y) = 1$, les points du nuage sont alors parfaitement alignés, on peut donc parler de **corrélation linéaire croissante totale** : pour un individu, sa donnée suivant X détermine entièrement sa donnée suivant Y.
Arbitrairement, on considèrera la corrélation linéaire croissante **faible** lorsque $0 < r(X,Y) < 0,3$, **moyenne** lorsque $0,3 \leq r(X,Y) \leq 0,7$ et **forte** lorsque $r > 0,7$.

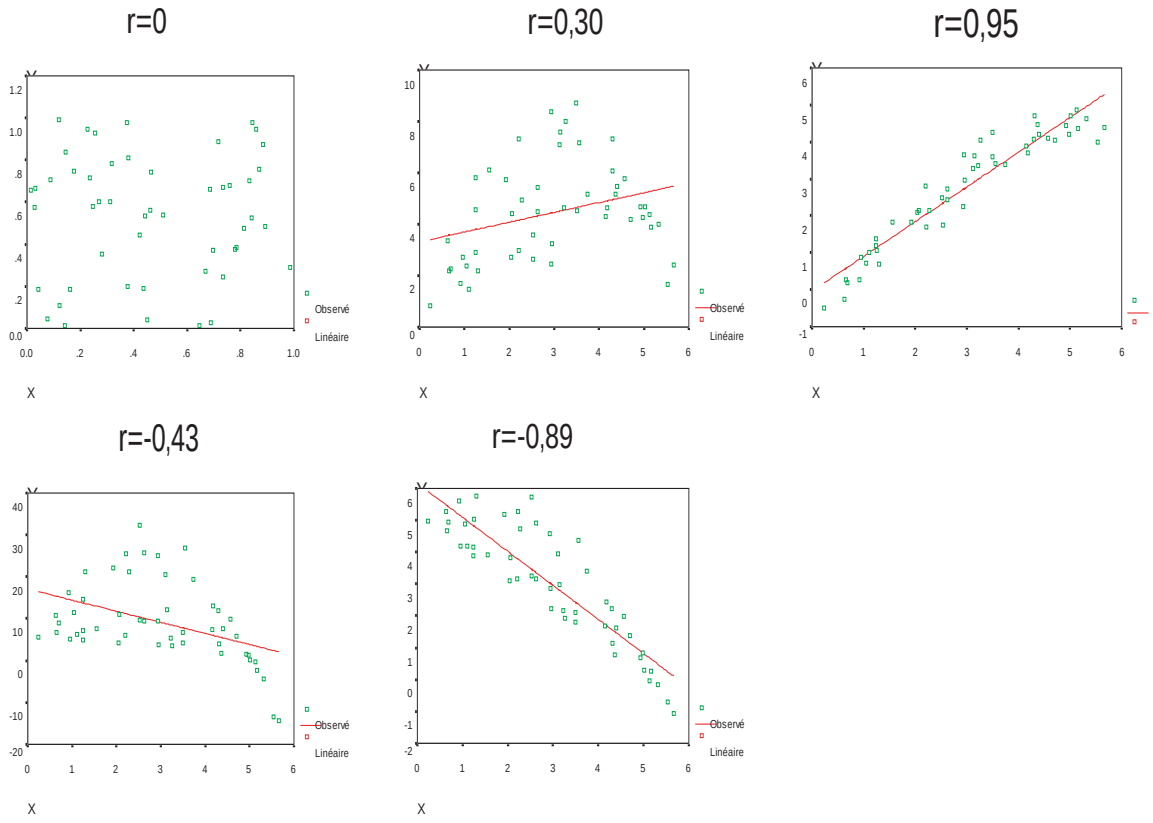
Si $r(X,Y) < 0$.
Dans ce cas la droite de régression de Y en X est décroissante ; on parle alors de corrélation linéaire décroissante entre X et Y. Lorsque $r(X,Y)$ est proche de -1, A est proche de 0, les points du nuage sont donc presque alignés, on a donc une forte corrélation linéaire décroissante (ou négative) entre X et Y.

Dans le cas extrême $r(X,Y) = -1$, les points du nuage sont alors parfaitement alignés, on peut donc parler de **corrélation linéaire décroissante totale** : pour un individu, sa donnée suivant

X détermine entièrement sa donnée suivant Y.

Arbitrairement, on considèrera la corrélation linéaire décroissante **faible** lorsque $-0,3 < r(X,Y) < 0$, **moyenne** lorsque $-0,7 \leq r(X,Y) \leq -0,3$ et **forte** lorsque $r < -0,7$.

Voici quelques exemples de nuages de points avec la valeur du coefficient de Bravais-Pearson qui permettent de mieux comprendre ce que traduit la valeur de $r(X,Y)$ au niveau de nuage de points :



2.4.3 Utilisation de la droite de régression $\Delta_{Y/X}$ pour faire des prévisions

Lorsqu'on observe un forte corrélation linéaire ($r(X,Y)$ proche de -1 ou de 1), les points du nuage sont presque alignés. Il est donc légitime, pour une observation dont on ne connaît que la donnée suivant X, d'utiliser l'équation de la droite de régression de Y en X afin de prévoir (ou estimer) la donnée correspondante suivant Y. Pour cela on remplace simplement x dans l'équation par la donnée suivant X, y donne alors une estimation de la donnée de Y correspondante. Bien sûr on peut échanger les rôles et remplacer y afin d'obtenir x, on aura alors une petite équation à résoudre. Graphiquement cela revient à placer le point du nuage permettant la prévision sur la droite de régression.

Exemple 2.1 (suite) : *Nous souhaitons estimer la note donnée par le second professeur d'un élève qui aurait eu 12 avec le premier.*
Nous avions l'équation de $\Delta_{Y/X}$ *est :* $y = 0,69x + 2,88$. *Donc*

$$y = 0,69 \times 12 + 2,88 = 11,22.$$

De même, si nous souhaitons estimer la note donnée par le premier professeur d'un élève qui aurait eu 15 avec le second,

$$x = \frac{15 - 2,88}{0,69} = 17,45.$$

Exemple 2.2 (suite) : *Nous souhaitons estimer la taille d'un enfant de 15kg.*
Nous avions l'équation de $\Delta_{Y/X}$ *est :* $y = 3,55x + 48,39$. *Donc*

$$y = 3,55 \times 15 + 48,39 = 101,64.$$

De même, si nous souhaitons estimer le poids d'un enfant de 115cm,

$$x = \frac{115 - 48,39}{3,55} = 18,76.$$

Exemple 2.3 (suite) : *Nous souhaitons estimer l'âge du décès d'un fils à partir de celui de son père de 60 ans.*
Nous avions l'équation de $\Delta_{Y/X}$ *est :* $y = 0,78x + 17,08$. *Donc*

$$y = 0,78 \times 60 + 17,08 = 63,72.$$

De même, si nous souhaitons estimer l'âge du décès d'un père à partir de celui de son fils de 60 ans,

$$x = \frac{60 - 17,08}{0,78} = 55,21.$$

Exemple 2.4. *On a demandé à une personne sachant taper à la machine de dactylographier un texte de 30 mots, puis un de 40 mots, etc.*
Voici les résultats obtenus :

Nombre de mots	*30*	*40*	*50*	*60*	*70*	*80*
Temps mis en secondes	*50*	*63*	*85*	*102*	*115*	*133*

1) Étude de la situation proposée

Un individu est ici une copie de mots, il y a 6 essais, notre échantillon se compose donc de N=6 copies. A chaque individu on associe le nombre de mots copiés (variable X) et le temps mis en secondes (variable Y).

Le tableau de données ci dessus nous donne directement les couples $(x_i; y_i)$ en relation, chaque couple étant observé une seule fois on a $n_i = 1$.

Pour rapprocher cet exemple de la présentation utilisée pour l'exemple précédent, on peut construire la table de contingence de X et Y qui sera :

Y \ X	30	40	50	60	70	80
50	1	0	0	0	0	0
63	0	1	0	0	0	0
85	0	0	1	0	0	0
102	0	0	0	1	0	0
115	0	0	0	0	1	0
133	0	0	0	0	0	1

Il est ici évident que X et Y sont totalement liés. En effet à une modalité de X correspond une unique modalité de Y (et réciproquement). On aura donc $\varphi = 1$. Bien que cela soit inutile on pourrait calculer χ^2 et en déduire φ. Attention car on aurait alors 36 effectifs théoriques à calculer : bien que la plupart des effectifs conjoints observés soient nuls, aucun effectif théorique n'est nul par contre.

2) On va d'abord étudier l'existence d'une corrélation linéaire entre X et Y sur notre échantillon. *Pour cela, traçons le nuage de points puis calculons $r(X, Y)$.*

Le tableau d'effectifs du couple (X,Y) sera :

x_i	y_i	n_i	$n_i x_i$	$n_i(x_i^2)$	$n_i y_i$	$n_i(y_i^2)$	$n_i x_i y_i$
30	50	1	30	900	50	2500	1500
40	63	1	40	1600	63	3969	2520
50	85	1	50	2500	85	7225	4250
60	102	1	60	3600	102	10404	6120
70	115	1	70	4900	115	13225	8050
80	133	1	80	6400	133	17689	10640
		6	330	19900	548	55012	33080

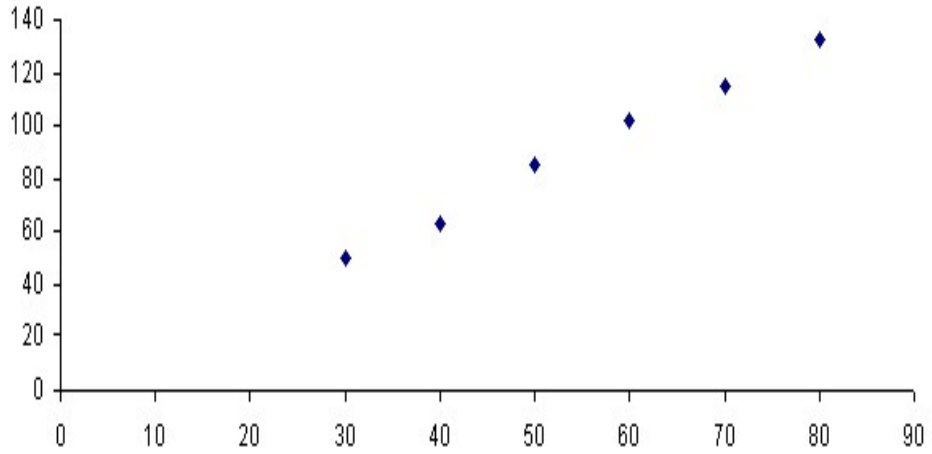


FIGURE 2.6 – *Le nuage de points pour la relation temps de copie/nombre de mots*

$$\bar{X} = \frac{330}{6} = 55 \qquad \sigma_X^2 = \frac{19900}{6} - 55^2 = 291,67 \qquad \sigma_X = 17,08$$
$$\bar{Y} = \frac{548}{6} = 91,33 \quad \sigma_Y^2 = \frac{55012}{6} - 91,33^2 = 826,89 \quad \sigma_Y = 28,76$$

$$\mathrm{Cov}(X,Y) = \frac{33080}{6} - 55 \times 91,33 = 490 \qquad r(X,Y) = \frac{490}{17,08 \times 826,89} = 0,998$$

On a donc une très forte corrélation linéaire positive entre le nombre de mots et le temps mis, pour ces 6 copies. Cela se traduit graphiquement par un nuage de points presque alignés. Il est donc légitime de se servir de la droite de régression de Y en X pour faire des prévisions.

3) Détermination de la droite de régression de Y en X

La droite $\Delta_{Y/X}$ a pour équation $y = ax + b$ avec $a = \dfrac{\mathrm{Cov}(X,Y)}{\sigma_X^2} = \dfrac{490}{291,67} = 1,68$ et $b = \bar{Y} - a\bar{X} = -1,07$. Cela donne : $y = 1,68x - 1,07$.

On peut maintenant faire une estimation du temps qui serait mis pour une dictée de 100 mots. Il suffit de remplacer x par 100, d'où $y = 1,68 \times 100 - 1,07 = 166,93$. On estime donc à 167 secondes le temps nécessaire pour copier 100 mots.

On peut inversement estimer le nombre de mots qui peuvent être copiés en 200 secondes. Il faut ici remplacer y par 200, d'où $200 = 1.68x - 1.07$, ce qui donne $x = \dfrac{200 + 1,07}{1,68} = 119,68$. On estime donc que 120 mots peuvent être copiés en 200 secondes.

Exemple 2.5. *Considérons la population constituée des dix pays en passe d'adhérer à l'U.E. en*

2002

	Population	*PNB global*	*PNB/habitant*
Chypre	*740*	*9*	*12162*
Slovénie	*1991*	*19*	*9543*
Malte	*373*	*3*	*8043*
République Tchèque	*10315*	*54*	*5235*
Hongrie	*10193*	*43*	*4219*
Slovaquie	*5343*	*19*	*3556*
Pologne	*38618*	*134*	*3470*
Estonie	*1466*	*4*	*2729*
Lituanie	*3709*	*8*	*2157*
Lettonie	*2490*	*5*	*2008*

Notons X la variable "population" (en milliers d'habitants), Y la variable "PNB global" (en milliards de dollars) et Z la variable "PNB par habitant" (en dollars).

1. *Pour les dix pays en passe d'adhérer à l'UE, calculer le coefficient de corrélation linéaire entre la population et le PNB global. Interpréter le résultat.*
2. *Que pensez-vous a priori du coefficient de corrélation linéaire entre la population et le PNB/habitant (On pourra justifier sa réponse en raisonnant au niveau mondial) ? Calculer ce coefficient et conclure.*

Commençons par faire un tableau de calcul qui sera utile pour répondre aux deux questions.

Population	*PNB*	*PNB/habitant*	x_i^2	y_i^2	z_i^2	$x_i y_i$	$x_i z_i$
740	*9*	*12162*	*547600*	*81*	*147918188*	*6660*	*9000000*
1991	*19*	*9543*	*3964081*	*361*	*91067766*	*37829*	*19000000*
373	*3*	*8043*	*139129*	*9*	*64688167*	*1119*	*3000000*
10315	*54*	*5235*	*106399225*	*2916*	*27406215*	*557010*	*54000000*
10193	*43*	*4219*	*103897249*	*1849*	*17796429*	*438299*	*43000000*
5343	*19*	*3556*	*28547649*	*361*	*12645525*	*101517*	*19000000*
38618	*134*	*3470*	*1491349924*	*17956*	*12040099*	*5174812*	*134000000*
1466	*4*	*2729*	*2149156*	*16*	*7444783*	*5864*	*4000000*
3709	*8*	*2157*	*13756681*	*64*	*4652285*	*29672*	*8000000*
2490	*5*	*2008*	*6200100*	*25*	*4032193*	*12450*	*5000000*
75238	*298*	*53121,0766*	*1756950794*	*23638*	*389691649*	*6365232*	*298000000*

Il n'est pas étonnant de retrouver la deuxième colonne dans la dernière.

Calculons $r(X, Y)$:

$$\bar{X} = \frac{75238}{10} = 7523,8$$
$$\bar{Y} = \frac{298}{10} = 29,8$$
$$\sigma_X = \sqrt{\frac{1756950794}{10} - 7523,8^2} = 10912,7225$$
$$\sigma_Y = \sqrt{\frac{23638}{10} - 29,8^2} = 10912,7225$$
$$\mathrm{Cov}(X,Y) = \frac{6365232}{10} - 7523,8 \times 29,8 = 412313,96$$
$$r(X,Y) = \frac{412313,96}{10912,7225 \times 38,4156} = 0,9835$$

Le coefficient de corrélation linéaire est très proche de 1, on peut donc considérer que, pour ces dix pays, il y a une forte corrélation linéaire positive entre la population et le PNB global. Il ne faut pas déduire de ce résultat qu'il y a toujours corrélation linéaire entre la population et le PNB global ; en effet, même si on peut penser que de façon générale, le PNB global est lié à la population, cette corrélation n'est linéaire que si on considère des pays dont les niveaux de développement sont voisins (par exemple, on peut constater que le PNB global des Pays-Bas est deux fois plus grand que celui de la Turquie alors que les Turcs sont quatre fois plus nombreux). Nous faisons de la statistique descriptive : les conclusions que l'on tire ne portent que sur la population étudiée.

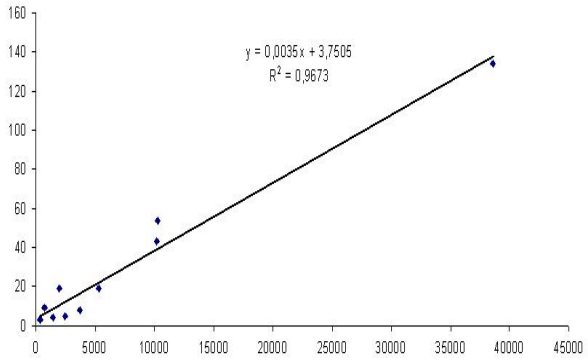
On peut penser a priori qu'il n'y a pas de corrélation linéaire entre la population d'un pays et le PNB par habitant. En effet, il existe des petits pays riches (le Luxembourg,...), des grands pays riches (les Etats-Unis,...), des petits pays pauvres (la Sierra Leone,...) et des grands pays pauvres (l'Inde,...) : dans chaque catégorie, beaucoup d'exemples viennent à l'esprit permettant de penser qu'il n'y a pas de lien. Calculons $r(X,Z)$:

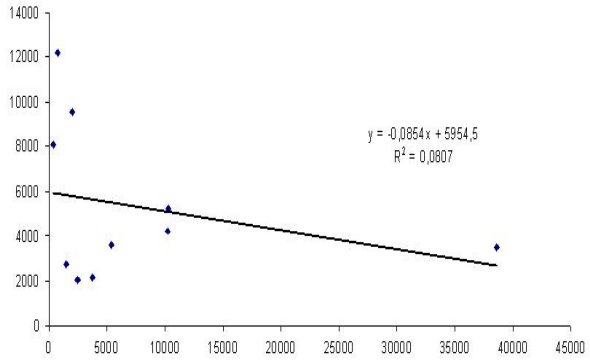
$$\bar{Z} = \frac{53121,0766}{10} = 5312,10766$$
$$\sigma_Z = \sqrt{\frac{389691649}{10} - 5312,10766^2} = 3278,8226$$
$$\mathrm{Cov}(X,Z) = \frac{298000000}{10} - 7523,8 \times 5312,10766 = -10167235,61$$
$$r(X,Z) = \frac{-10167235,61}{10912,7225 \times 3278,8226} = -0,2842$$

Ce résultat confirme une faible corrélation linéaire.

Voici les nuages de points correspondants aux deux situations avec les droites de régression. Dans le premier cas, la droite est un bon modèle pour représenter le lien entre X et Y ; dans le

deuxième cas, il est clair que beaucoup de points s'écartent de cette droite.





Chapitre 3

Les séries temporelles

3.1 Introduction

Une **série chronologique** ou **temporelle** est une suite de valeurs échelonnées dans le temps correspondant à l'évolution d'un phénomène. Nous avons déjà rencontré des séries temporelles :

• lorsqu'on étudie la population d'une ville (ou d'un pays) tous les dix ans entre 1800 et 1910 (les grandeurs mesurées sont des niveaux) ;

• lorsqu'on mesure le taux d'inflation mensuel pendant un an (les grandeurs mesurées sont des flux : taux de variation durant une période).

Le temps pouvant être considéré comme une variable, les séries temporelles sont des cas particuliers de séries statistiques à deux variables. Nous noterons T la variable temps et Y la variable dont on étudie l'évolution.

L'étude d'une série chronologique a pour but de créer un modèle mathématique décrivant au mieux la série des données afin de faire des prévisions à court terme. En particulier, le modèle devra tenir compte

- d'une tendance notée (V_t) : comportement moyen à moyen ou long terme ;
- de variations saisonnières (W_t) : comportement cyclique se répétant à intervalles de temps réguliers.

Nous allons développer dans ce cours le modèle additif défini de la façon suivante pour tout t :

$$Y_t = V_t + W_t + \epsilon_t;$$

où ϵ_t est le bruit du modèle, dû aux imprécisions de mesures, à l'aléa de la réalité et à l'erreur du modèle.

Remarque 3.1. *Il existe d'autres modèles :*

- le modèle multiplicatif : $Y_t = V_t W_t \epsilon_t$.

- le modèle mixte : $Y_t = V_t W_t + \epsilon_t$.

Rappelons que le but est de pouvoir faire des préictions à court terme. Nous allons donc traiter chaque composante séparément, l'estimer et la prédire. Puis en utilisant le fait que le modèle est additif est s'écrit selon

$$Y_t = V_t + W_t + \epsilon_t,$$

proposer une prédiction de Y.

Méthodologie :

1. Représentation graphique des données.
2. Lissage de la courbe par la méthode des moyennes mobiles afin d'isoler la tendance et estimation de cette dernière.
3. Isolement et estimation des variations saisonnières.
4. Prédiction à court terme.

3.2 Exemple

Nous allons considérer l'exemple suivant. Voici un tableau qui donne les consommations trimestrielles en électricité d'une entreprise de vente par Internet pendant les trois premières années :

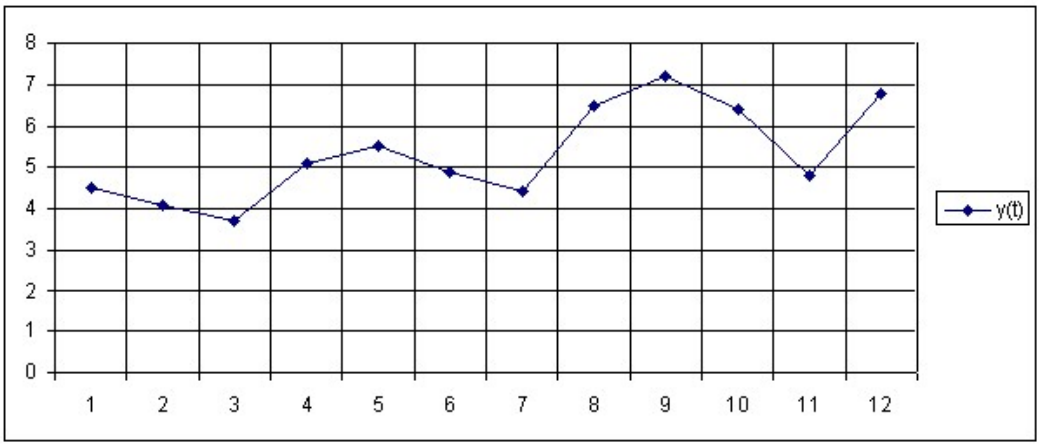
Trimestres \ Années	1997	1998	1999
1	4,5	5,5	7,2
2	4,1	4,9	6,4
3	3,7	4,4	4,8
4	5,1	6,5	6,8

Notons T la variable temps (3 périodes avec 4 observations par période) et Y la variable "consommation en électricité (en milliers de kWh)".

Pour la variable T, on commence par numéroter les valeurs de 1 à 12 (taille de la série statistique, nombre d'observations) dans l'ordre chronologique. Cette numérotation correspond à un changement de variable analogue à ceux que nous avons rencontrés en exercices dans le cours du premier semestre et va grandement simplifier les calculs.

3.3 Représentations graphiques

3.3.1 Représentation cartésienne

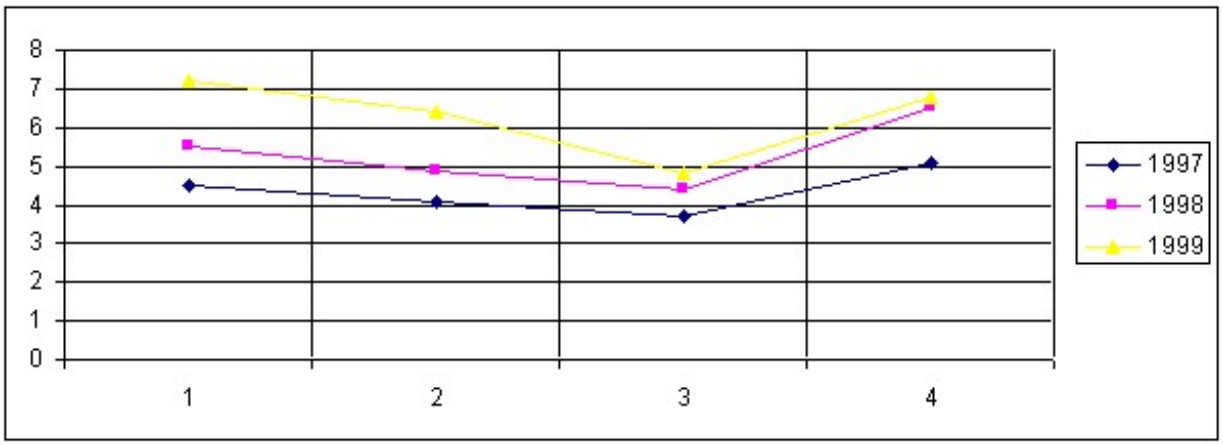


La courbe n'est pas strictement croissante mais il apparaît qu'il y a globalement une hausse de la consommation avec des pics ou des creux selon les trimestres (des variations saisonnières). Ainsi la tendance V_t a l'air d'être croissante tandis que la composante siasonnière semble être liée aux trimestres.

3.3.2 Représentation cartésienne avec périodes superposées

Pour confirmer que les variations sont bien liées aux trimestres, on peut s'aider de graphes superposés (un pour chaque période) et voir si les courbes obtenues

- sont effectivement empilées : mettant en évidence la tendance croissante ou décroissante ;
- ont la même allure : les pics et les creux sont bien liés aux " saisons ".



Ici la croissance s'explique facilement par la hausse de l'activité et les investissements, les variations pouvant être dues aux saisons.

Remarque 3.2. *Le terme "saison" doit être compris au sens large, il s'agit des subdivisions de la période, on parlera de façon générale de variations saisonnières pour des variations mensuelles, hebdomadaires,... Pour des relevés journaliers, les saisons seront les jours de la semaine ; pour des relevés mensuels, les saisons seront les mois de l'année ; pour des relevés trimestriels, les saisons seront ici les saisons au sens usuel : printemps, été, automne, hiver...*

3.4 Lissage d'une courbe : série des moyennes mobiles

On se demande si, en faisant abstraction des variations saisonnières, la série statistique suit grossièrement une droite c'est-à-dire la tendance est-elle linéaire ? Pour répondre à cette question, nous allons désaisonnaliser la série à l'aide d'un outil appelé moyenne mobile.

Soit $y(t)$ une série chronologique de taille N (pour chaque valeur entière de t comprise entre 1 et N, on observe une valeur de Y que l'on note $y(t)$). On suppose que cette série est découpée en périodes avec p observations par période (c'est à dire p "saisons").
On appelle **série des moyennes mobiles d'ordre** p la série des valeurs $z(t)$ calculées de la façon suivante :

- Si p est pair (c'est le cas lorsqu'on découpe l'année en 4 trimestres), alors $p = 2k$ et

$$z(t) = \frac{\frac{y(t-k)}{2} + y(t-k+1) + \ldots + y(t) + \ldots + y(t+k-1) + \frac{y(t+k)}{2}}{p}$$

  (il y a p+1 termes au numérateur avec $y(t)$ au milieu).
- Si p est impair (c'est le cas lorsqu'on découpe la semaine en 7 jours), alors $p = 2k+1$ et

$$z(t) = \frac{y(t-k) + y(t-k+1) + \ldots + y(t) + \ldots + y(t+k-1) + y(t+k)}{p}$$

  (il y a p termes au numérateur avec $y(t)$ au milieu).

Ici, la période est l'année qui est découpée en 4 trimestres (4 observations par période) ; on va donc calculer les moyennes mobiles d'ordre 4 :

	t	$y(t)$	$z(t)$
1997	1	4,5	-
	2	4,1	-
	3	3,7	4,475
	4	5,1	4,7
1998	1	5,5	4,8875
	2	4,9	5,15
	3	4,4	5,5375
	4	6,5	5,9375
1999	1	7,2	6,175
	2	6,4	6,2625
	3	4,8	-
	4	6,8	-

Pour calculer la moyenne mobile z au temps t, nous avons besoin des valeurs de $y(t-2)$, $y(t-1)$, $y(t)$, $y(t+1)$ et $y(t+2)$. La première moyenne mobile que l'on peut calculer est donc $z(3)$, et on a :

$$z(3) = \frac{\frac{4,5}{2} + 4,1 + 3,7 + 5,1 + \frac{5,5}{2}}{4} = 4,475$$

puis

$$z(4) = \frac{\frac{4,1}{2} + 3,7 + 5,1 + 5,5 + \frac{4,9}{2}}{4} = 4,7$$

et ainsi de suite jusqu'à $z(10)$.

Remarque 3.3. *1) La série des moyennes mobiles n'est définie que pour*

$$k + 1 \leq t \leq N - k.$$

2) Les propriétés de la moyenne mobile d'ordre p entraînent que son application

- conserve une tendance linéaire : si Y a une tendance linéaire, Z aura la même tendance ;

- supprime toute saisonnalité de période p : si Y a une composante saisonnière, Z n'en aura pas ;

- atténue de façon optimale le bruit : le bruit de Z sera plus faible que celui de Z.

D'après la représentation graphique (3.4), nous voyons bien que la série des moyennes mobiles a pour effet de lisser la courbe des données brutes en gommant les variations saisonnières et en atténuant les irrégularités dues au bruit. On fait ainsi mieux apparaître la tendance.
La série des moyennes mobiles fait clairement apparaître ici une tendance linéaire croissante, nous allons donc calculer la droite de régression à partir de la série des moyennes mobiles.

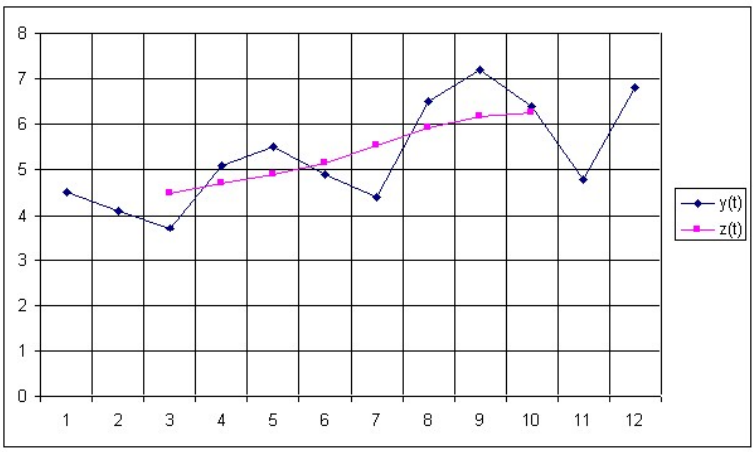


FIGURE 3.1 – Représentation de la consommation en électricité et transformation par moyenne mobile

3.5 La droite de tendance (ou trend)

Puisque nous avons mis en évidence une tendance linéaire croissante, nous allons déterminer la droite de régression associée.

Attention : dans le modèle que nous allons construire, on calcule l'équation de la droite de tendance avec les valeurs $z(t)$ (la série sans variation saisonnière) et non avec les données brutes $y(t)$.

Notons

1) Z la variable "moyenne mobile".

2) t_Z la variable temps en prenant uniquement les valeurs t pour lesquelles on peut calculer $z(t)$.

3) N_Z le nombre de données observées pour Z.

Dans l'exemple, on prend les valeurs t_Z de 3 à 10, les valeurs de $z(t)$ correspondantes et N_Z vaut 8.

La droite de tendance est la droite de régression de Z en t_Z, on notera son équation sous la forme : $x(t) = at + b$.

Il suffit donc de trouver le coefficient directeur a et l'ordonnée à l'origine b, comme on l'a vu au

chapitre précédent. Grâce à la numérotation de la variable temps, les calculs sont grandement simplifiés. En effet,

$$\bar{t}_Z = \frac{N+1}{2} \quad \text{et} \quad \mathrm{Var}(t_Z) = \frac{N_Z^2 - 1}{12}.$$

Dans la deuxième formule, le dénominateur est toujours 12, il ne dépend pas de N (la taille de la série statistique de départ) ni de N_Z.

Revenons à notre exemple et calculons l'équation de la droite de tendance. On a

$$\bar{t}_Z = \frac{12+1}{2} = 6,5 \quad \text{et} \quad \mathrm{Var}(t_Z) = \frac{8^2 - 1}{12} = 5,25.$$

On sait que $a = \dfrac{\mathrm{Cov}(Z, t_Z)}{\mathrm{Var}(t_Z)}$ et que $b = \overline{Z} - a\bar{t}_Z$; il reste donc à calculer $\overline{Z}$ et $\mathrm{Cov}(Z, t_Z)$:

t	$z(t)$	$tz(t)$
3	4,475	13,425
4	4,7	18,8
5	4,8875	24,4375
6	5,15	30,9
7	5,5375	38,7625
8	5,9375	47,5
9	6,175	55,575
10	6,2625	62,625
	43,125	292,025

Ainsi

$$\overline{Z} = \frac{43,125}{8} = 5,390625 \quad \text{et} \quad \mathrm{Cov}(Z, T_Z) = \frac{292,025}{8} - 6,5 \times 5,390625 = 1,4640625.$$

et

$$a = \frac{1,4640625}{5,25} = 0,28 \quad \text{et} \quad b = 5,390625 - 0,278869 \times 6,5 = 3,58.$$

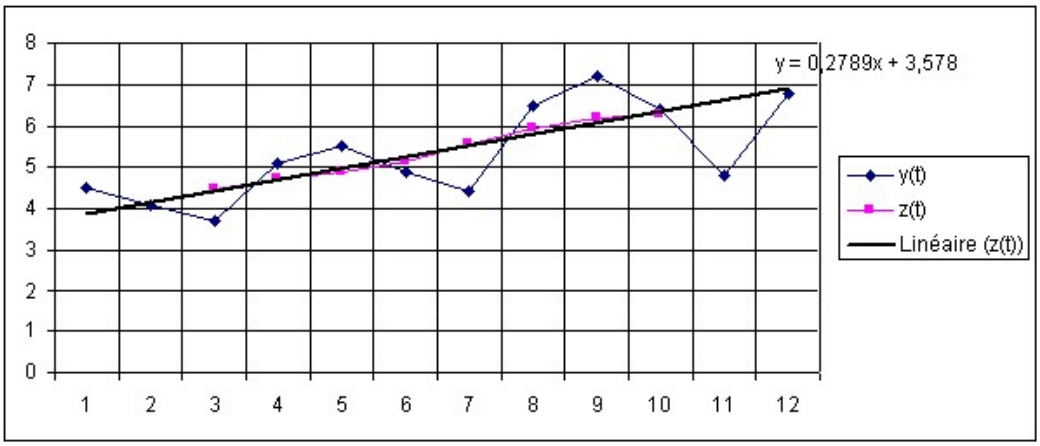
On se contentera pour donner les résultats de deux chiffres après la virgule mais pour faire les calculs, on conservera davantage de chiffres lorsqu'on réutilise un résultat.

L'équation de la droite de tendance est donc $x(t) = 0,28t + 3,58$.

Remarque 3.4. *Si on calcule le coefficient de corrélation linéaire, on trouve $r(Z, T_Z) = 0,9915$, ce qui confirme une forte corrélation linéaire positive. Par contre, en considérant les données brutes $y(t)$, le coefficient de corrélation linéaire est $r(Y, T) = 0,6986$, il y a donc une faible corrélation linéaire à cause des variations saisonnières. Cependant, l'équation de la droite de*

régression de Y en T est $y = 0,22t + 3,88$. Ainsi, si l'on considérait comme droite de tendance la droite $\Delta_{Y/T}$ (au lieu de Δ_{Z/T_Z}), on obtiendrait un modèle assez proche de celui que nous allons maintenant construire.

Traçons la droite de tendance sur le graphique précédent (il suffit de placer les points $B(0; 3,58)$ et le point moyen $M(6,5; 5,39)$).



Il est clair que le modèle linéaire n'est pas satisfaisant pour décrire l'évolution de Y puisque beaucoup de valeurs s'écartent de la droite mais on va s'en servir comme première approximation pour construire un modèle plus précis tenant compte des variations saisonnières.

3.6 Les coefficients saisonniers

Pour réaliser une meilleure approximation, on va calculer des coefficients saisonniers, notés $s(t)$, qui prennent la même valeur sur chaque période. Dans l'exemple, il y a 4 saisons par période donc 4 coefficients à calculer.

Méthode d'estimation des coefficients saisonniers :

1) On calcule d'abord les écarts algébriques entre les composantes linéaires et les valeurs observées $\delta(t) = y(t) - x(t)$.

2) On calcule ensuite des composantes saisonnières (une pour chaque saison), que l'on notera $s'(t)$, en faisant la moyenne de ces écarts algébriques pour chaque saison considérée.

Exemple 3.1. *Calculons la composante saisonnière $s'(1)$ pour le premier trimestre. Les valeurs observées aux premiers trimestres de chacune des 3 périodes d'observation sont : $y(1) = 4,5$,*

$y(5) = 5,5$ *et* $y(9) = 7,2$. *Les composantes linéaires sont :*

$$x(1) = 0,28\times 1+3,58 = 3,86, \quad x(5) = 0,28\times 5+3,58 = 4,97, \quad x(9) = 0,28\times 9+3,58 = 6,09.$$

Les écarts algébriques sont :

$$\delta(1) = 4,5 - 3,86 = 0,64, \quad \delta(5) = 5,5 - 4,98 = 0,53, \quad \delta(9) = 7,2 - 6,1 = 1,11.$$

(On remarque que ces écarts algébriques sont tous positifs : graphiquement, cela signifie qu'au premier trimestre, les valeurs observées sont toutes au-dessus du trend.)

La moyenne de ces écarts algébriques est

$$s'(1) = \frac{\delta(1) + \delta(5) + \delta(9)}{3} = \frac{0,64 + 0,53 + 1,11}{3} = 0,761.$$

3) On souhaite que dans le modèle construit, les variations saisonnières autour du trend se compensent sur une période, c'est-à-dire que la somme des coefficients saisonniers soit nulle. Autrement dit, on veut avoir : $\sum_{t=1}^{p} s(t) = 0$ où p est, comme précédemment, le nombre d'observations par période, c'est-à-dire le nombre de coefficients saisonniers. Pour cela, on calcule :

$$S_0 = \sum_{t=1}^{p} s'(t) = s'(1) + \ldots + s'(p)$$

Puis on pose :

$$s(t) = s'(t) - \frac{S_0}{p}$$

Ainsi

$$\sum_{t=1}^{p} s(t) = \sum_{t=1}^{p} \left(s'(t) - \frac{S_0}{p} \right) = \sum_{t=1}^{p} s'(t) - \sum_{t=1}^{p} \frac{S_0}{p} = S_0 - p \times \frac{S_0}{p} = S_0 - S_0 = 0.$$

Exemple 3.2. *Terminons les calculs à l'aide d'un tableau.*

t	$y(t)$	$x(t) = 0,28t + 3,58$	$y(t) - x(t)$	$s'(t)$	$s(t) = s'(t) - S0/4$
1	*4,5*	*3,86*	*0,64*	*0,761*	*0,827*
2	*4,1*	*4,14*	*-0,04*	*-0,118*	*-0,052*
3	*3,7*	*4,41*	*-0,71*	*-1,23*	*-1,164*
4	*5,1*	*4,69*	*0,41*	*0,324*	*0,390*
5	*5,5*	*4,97*	*0,53*	*0,761*	*0,827*
6	*4,9*	*5,25*	*-0,35*	*-0,118*	*-0,052*
7	*4,4*	*5,53*	*-1,13*	*-1,23*	*-1,164*
8	*6,5*	*5,81*	*0,69*	*0,324*	*0,390*
9	*7,2*	*6,09*	*1,11*	*0,761*	*0,827*
10	*6,4*	*6,37*	*0,03*	*-0,118*	*-0,052*
11	*4,8*	*6,65*	*-1,85*	*-1,23*	*-1,164*
12	*6,8*	*6,92*	*-0,12*	*0,324*	*0,390*

Pour remplir la dernière colonne, on calcule d'abord la somme :

$$S_0 = 0,761 - 0,118 - 1,23 + 0,324 = -0,2625$$

Puis $\frac{S_0}{4} = -0,066$ *et on a*

$$s(1) = s'(1) - (-0,066) = 0,761 + 0,066 = 0,827,\ldots$$

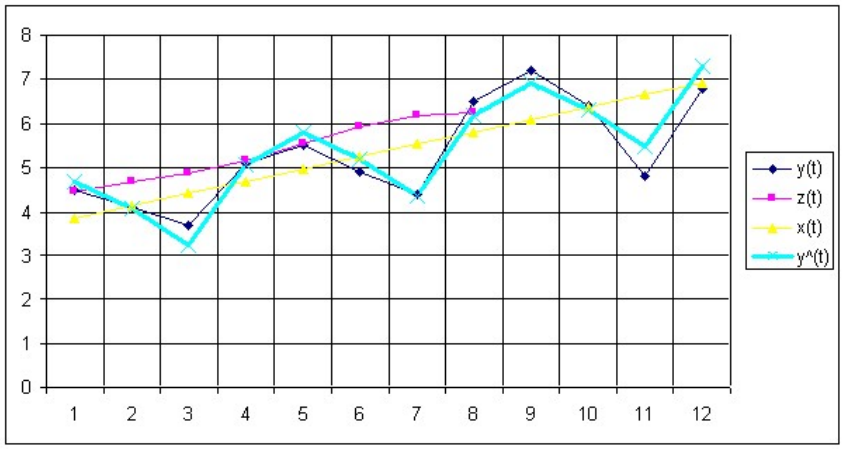
3.7 Prévisions à court terme

Nous travaillons avec le modèle additif comme expliqué en introduction. Dans ce modèle, le coefficient saisonnier est ajouté à la composante linéaire :

$$\hat{y}(t) = x(t) + s(t).$$

Le chapeau sur y permet de bien différencier les données brutes (y) du modèle théorique que l'on construit ($\hat{y}$) et qui nous servira pour faire des prévisions à court terme. $x(t)$ représente l'estimation de la tendance par régression linéaire et $s(t)$ celle de la composante saisonnière.

Voici (en trait épais) la représentation graphique du modèle que l'on a construit pour décrire la série chronologique en tenant compte des variations saisonnières :



Cette courbe oscille régulièrement de part et d'autre de la droite de tendance.

En considérant ce modèle, calculons les prévisions pour la période 4 (année 2000) : La période 4 correspond aux valeurs 13, 14 ,15 et 16 de la variable T. Et donc les consommations prévues

pour 2000 sont :

$$\begin{aligned}
\hat{y}(13) &= x(13) + s(13) = x(13) + s(1) = (0,28 \times 13 + 3,58) + 0,827 = 7,595 \\
\hat{y}(14) &= x(14) + s(14) = x(14) + s(2) = (0,28 \times 14 + 3,58) - 0,052 = 6,938 \\
\hat{y}(15) &= x(15) + s(15) = x(15) + s(3) = (0,28 \times 15 + 3,58) - 1,164 = 6,048 \\
\hat{y}(16) &= x(16) + s(16) = x(16) + s(4) = (0,28 \times 16 + 3,58) + 0,390 = 7,824
\end{aligned}$$

- 7,595 au premier trimestre (en milliers de kWh),
- 6,938 au deuxième trimestre,
- 6,048 au troisième trimestre,
- 7,824 au quatrième trimestre.

Chapitre 4

Utilisation d'un tableur

EXCEL est un logiciel qui permet d'effectuer des calculs, particulièrement des calculs répétitifs, et d'avoir instantanément la mise à jour des résultats lors de changements des données.

Pour les fonctions de base d'Excel, se reporter au TP du premier semestre.

4.1 Révisions des notions du semestre précédent

On considère le tableau de données ci-dessous concernant l'ensemble des pays de l'Union Européenne ainsi que les pays candidats en 2002 à l'entrée dans l'U.E. (Les chiffres sont ceux de 1996, source : Atlaséco 1999).
La population est donnée en milliers d'habitants.
Le PNB (Produit National Brut) global est exprimé en milliards de dollars.
Le PNB par habitant est en dollars.
La colonne **F** indique le nombre de frontières communes avec les 28 pays (le tunnel sous la Manche ne compte pas comme frontière entre la France et le Royaume-Uni, Gibraltar n'induit pas une frontière entre l'Espagne et le Royaume-Uni... Par exemple, le Luxembourg touche la France, la Belgique et l'Allemagne donc F=4).
La colonne " Rang " donne le rang mondial pour le PNB global.

	Membre de l'UE	Population	PNB global	PNB/habitant	F	Rang
Luxembourg	Oui	416	19	45673	4	69
Danemark	Oui	5262	170	32307	2	26
Allemagne	Oui	81912	2342	28592	9	3
Autriche	Oui	8059	226	28043	7	22
Suède	Oui	8843	240	27140	2	21
France	Oui	58375	1535	26296	6	4
Belgique	Oui	10159	267	26282	5	20
Pays-Bas	Oui	15517	399	25714	3	13
Finlande	Oui	5125	120	23415	2	35
Italie	Oui	57380	1193	20791	4	5
Royaume-Uni	Oui	58782	1148	19530	2	6
Irlande	Oui	3626	62	17099	2	46
Espagne	Oui	39260	574	14620	3	9
Chypre	Non*	740	9	12162	1	83
Grèce	Oui	10475	125	11933	3	33
Portugal	Oui	9930	103	10373	2	36
Slovénie	Non*	1991	19	9543	4	70
Malte	Non*	373	3	8043	1	131
République Tchèque	Non*	10315	54	5235	5	49
Hongrie	Non*	10193	43	4219	5	51
Slovaquie	Non*	5343	19	3556	5	68
Pologne	Non*	38618	134	3470	5	32
Turquie	Non	62697	184	2935	3	24
Estonie	Non*	1466	4	2729	2	116
Lituanie	Non*	3709	8	2157	3	91
Lettonie	Non*	2490	5	2008	3	107
Roumanie	Non	22608	35	1548	3	56
Bulgarie	Non	8356	9	1077	4	84

*Pays en passe d'adhérer à l'Union Européenne

Ouvrir le fichier Doc UE.xls

4.1.1 Trier

Lorsqu'on dispose d'un tableau de données avec plusieurs variables, on peut souhaiter trier les individus en fonction de chaque variable.

Exemple : *Pour l'instant, les pays de l'exemple sont classés par PNB/habitant décroissant. Trions-les du plus peuplé au moins peuplé.*

- *Sélectionner la plage* ***A1 :G29*** *(c'est-à-dire tout le tableau).*
- *Dans le menu* ***Données****, cliquer sur* ***Trier****.*
- *Trier par* ***Population****,* ***Décroissant****.*
- *Cliquer sur* ***OK****.*

On peut utiliser le TRI pour dresser facilement un tableau d'effectifs ou bien réaliser un regroupement en classes. A partir du tableau d'effectifs on pourra ensuite calculer la moyenne, la variance et l'écart-type de la variable. Nous verrons plus loin que EXCEL permet aussi d'avoir directement un tableau d'effectifs à partir d'un tableau de données.

Exercice 1 : Nous allons d'abord étudier la variable ***F*** *qui est de type quantitatif discret ; voici un tableau d'effectifs et de fréquences, avec des colonnes de calculs que nous allons utiliser pour déterminer la moyenne et la variance de* ***F*** *.*

$F(x_i)$	*Effectifs* n_i	*Fréquences en %*	$n_i x_i$	$n_i(x_i)^2$
1				
2				
3				
4				
5				
6				
7				
8				
9				
Somme				

- Après avoir trié les pays selon la variable ***F****, compléter les effectifs dans le tableau ci-dessus puis recopier ce tableau sur la feuille de calcul sous le tableau de données (cellule* ***A40****).*

- Calculer la taille de la population (utiliser Σ*).*

- Compléter la colonne "Fréquences" en calculant la fréquence de la première modalité puis en utilisant la fonction de remplissage (pour que la cellule ***B49*** *contenant la taille de la population soit conservée comme diviseur lors du remplissage, taper* ***B$49*** *au lieu de* ***B49*** *: le $ fixe la ligne 49).*

- Sur le même principe, saisir les formules $n_i x_i$ *et* $n_i(x_i)^2$ *pour la première ligne puis utiliser le remplissage vers le bas pour compléter le tableau. Utiliser le remplissage, vers la droite pour obtenir la somme de chaque colonne puis terminer les calculs de moyenne et de variance dans les cellules* ***G44*** *et* ***G46*** *(on écrira " moyenne = " dans la cellule* ***F44*** *et " variance= " dans la*

*cellule **F46**).*

4.1.2 Utilisation des fonctions de calcul

On peut calculer la moyenne, la variance, l'écart-type mais aussi la médiane en utilisant directement les fonctions d'EXCEL.

***Exemple** : Retrouvons les résultats de l'exercice 1.*

- *Sous la série statistique correspondant à la variable **F**, taper =.*
- *Parmi les fonctions proposées (à gauche de la barre de formules) choisir **MOYENNE** (puis **VAR.P**, puis **ECARTYPEP**, puis **MEDIANE**, on rentrera les résultats dans les 4 cases sous la série statistique (on peut faire de même pour chacune des variables quantitatives du tableau de données) .*
- *Sélectionner avec la souris la série statistique correspondant à la variable **F**.*
- *Cliquer sur **OK**.*

Exercice 2 : Nous allons voir comment des calculs de moyennes et variances permettent de mettre en évidence le lien entre les variables "Membre de l'UE" et "PNB/habitant".
On partage l'ensemble W des 28 pays en 3 groupes :

- Ω_1 *: les pays membres de l'UE.*
- Ω_2 *: les pays en passe d'adhérer à l'UE.*
- Ω_3 *: Turquie, Bulgarie et Roumanie.*

On s'intéresse à la variable "PNB/habitant", notée X.

- Calculer la moyenne de X, notée m, et la variance de X, notée v, sur toute la population O (utiliser les fonctions d'EXCEL). Cette moyenne est-elle le PNB/habitant de l'ensemble des 28 pays ?

- Sur chaque groupe Ω_i, calculer la moyenne de X, notée m_i et la variance de X, notée v_i (utiliser le tri et les fonctions d'EXCEL). Compléter les phrases suivantes :

$$\begin{aligned} &\text{Sur } \Omega_1, \text{ la moyenne est } m_1 = \quad \text{et la variance est } v_1 = \\ &\text{Sur } \Omega_2, \text{ la moyenne est } m_2 = \quad \text{et la variance est } v_2 = \\ &\text{Sur } \Omega_2, \text{ la moyenne est } m_3 = \quad \text{et la variance est } v_3 = \end{aligned}$$

Une simple comparaison des moyennes permet de penser qu'un tel écart n'est pas fortuit mais que le PNB par habitant est bien un des critères qui distingue les trois groupes. On peut également constater que la variance de X sur Ω est bien plus grande que la variance de X sur Ω_1 c'est-à-dire

que l'hétérogénéité de l'U.E. pour le PNB/h serait bien plus grande avec ces 28 pays.
On peut aussi mesurer le lien entre X et la variable nominale Y dont les modalités sont "membre de l'U.E.", "en passe d'adhérer à l'U.E.", "autre candidat" en calculant le rapport entre la variance intergroupe (variance des moyennes de chaque groupe Ω_i affectées des effectifs des groupes) et la variance totale. Ce rapport, toujours compris entre 0 et 1, indique un lien fort lorsqu'il est proche de 1 et faible lorsqu'il est proche de 0. On trouve une variance intergroupe égale à $\left(var\ inter = \dfrac{15m_1^2 + 10m_2^2 + 3m_3^2}{28} - m^2 \right)$ *soit un rapport de 0,84 ; on peut donc considérer que le lien entre X et Y est fort.*

4.1.3 Couple de variables

Une autre façon de mettre en évidence un lien entre 2 variables est d'étudier le couple de variables à partir d'une table de contingence (tableau de distribution des effectifs conjoints). EXCEL permet de dresser des tables de contingence à partir d'un tableau de données mais lorsqu'on traite des variables quantitatives continues, il est nécessaire de commencer par faire un regroupement en classes.

***Exemple** : Les variables "PNB/habitant" et "Population" sont quantitatives continues ; nous allons faire un regroupement en classes avec les intervalles suivants :*

PNB/habitant (en milliers de dollars) :	*]0 ;10]*	*]10 ;20]*	*]20 ;30]*	*]30 ;40]*	*]40 ;50]*
Population (en millions d'habitants) :	*]0 ;1]*	*]1 ;10]*	*]10 ;30]*	*]30 ;50]*	*]50 ;100]*

Pour effectuer ce regroupement en classes :

- *Rajouter 2 colonnes au tableau de données intitulées "PNB/hab (classes)" et "Population (classes)".*
- ***Trier** par "PNB/habitant" puis utiliser la fonction **remplissage** (menu **Edition**) pour compléter rapidement la nouvelle colonne "PNB/hab (classes)" .*
- ***Trier** par "Population" puis utiliser la fonction **remplissage** (menu **Edition**) pour compléter rapidement la nouvelle colonne "Population (classes)".*

Nous allons maintenant regarder des tableaux croisés (tables de contingence) pour les couples de variables (Membre de l'UE , PNB/hab (classes)) et (Membre de l'UE , Population (Classes)). On peut se douter que le lien est fort dans le premier cas et faible dans le deuxième, nous allons voir des graphiques qui mettent cela en évidence.

***Exemple** :*

• *Sélectionner tout le tableau de données (plage* ***A1 :I29****).*

• *Dans le menu* ***Données****, choisir* ***Rapport de tableau croisé dynamique****.*

• *Cliquer sur* ***Suivant*** *(la nouvelle fenêtre précise alors que vous avez bien sélectionné la plage* ***A1 :I29****)*

• *Cliquer de nouveau sur* ***Suivant****, la dernière fenêtre, étape 3/3, propose de placer le tableau croisé sur une nouvelle feuille, cliquer sur* ***Terminer*** *pour confirmer ce choix.*

• *Apparaît alors un tableau constitué de 3 parties :* ***Colonne****,* ***Ligne*** *et* ***Données*** *: avec la souris, placer le mot "****pays****" dans la partie* ***Données****, la variable "****Membre de l'UE****" dans la partie****Colonne*** *(on obtient alors le tableau d'effectifs de la variable "Membre de l'UE"), la variable "* ***PNB/hab (classes)*** *" dans la partie* ***Ligne*** *(on obtient alors la table de contingence de ce couple de variable, avec les distributions marginales des effectifs).*

• *Cliquer sur l'icône* ***Assistant Graphique*** *du* ***Tableau croisé dynamique****.*

• *Cliquer sur le bouton droit de la souris sur le graphique obtenu pour sélectionner le* ***Type de graphique*** *souhaité et dans* ***Histogramme****, choisir* ***Histogramme empilé 100****%. On obtient pour chaque classe la proportion de pays appartenant à chacune des trois catégories. Ici, il n'y a que des pays candidats dans la première classe alors qu'il n'y a plus de pays candidats à partir de la troisième classe, le lien entre les 2 variables est évident.*

• *Revenir sur la feuille où se trouve le* ***Tableau croisé dynamique*** *et remplacer à l'aide de la souris la variable "PNB/hab (classes)" par "Population (classes)". Cela modifie automatiquement le graphique qui cette fois ne permet pas de parler de lien entre les deux variables (les 5 rectangles "se ressemblent").*

4.1.4 Représentations graphiques

1) Etude d'une variable qualitative

Il est possible de représenter graphiquement

- *une variable qualitative nominale par un* ***diagramme en colonnes*** *ou* ***en bâtons*** *ou un* ***diagramme en secteurs****.*
- *une variable qualitative ordinale par une* ***boîte à moustaches****.*

Les variables quantitatives discrètes pouvant être considérées comme des variables qualitatives nomibales ou ordinales, nous illustrerons cette section en étudiant la variable F.

A) Le diagramme en colonnes ou en bâtons

Exemple *: Nous allons représenter la variable "F".*

• *Sélectionner le tableau d'effectifs de la variable F.*

• *Cliquer sur l'icône* ***Assistant Graphique*** → ***Histogramme****. En appuyant sur l'onglet* ***Maintenir appuyé pour visualiser****, on peut voir le graphique.*

• *Cliquer sur* ***Suivant*** *(la nouvelle fenêtre précise alors la plage de données sélectionnée). Cliquer sur l'onglet* ***Série****.*

Dans le cadre ***Nom****, vous pouvez rentrer* ***Variable F****.*

Dans le cadre ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de F en sélectionnant la bonne plage de données dans le fichier Excel.*

• *Cliquer sur* ***Suivant****. Vous pouvez alors modifier les options de légende, quadrillage, titre... On peut entre autres faire figurer les valeurs.*

• *Cliquer sur* ***Suivant****. Vous pouvez alors choisir de faire apparaître le graphique sur la même feuille de calcul ou dans une autre fenêtre. Cliquer sur* ***Terminer****.*

B) Le diagramme en secteurs

Exemple *: Nous allons représenter la variable "F" cette fois par un diagramme en secteurs.*

• *Sélectionner le tableau d'effectifs de la variable F.*

• *Cliquer sur l'icône* ***Assistant Graphique*** → ***Secteurs*** *puis sur* ***Suivant****.*

• *Cliquer sur l'onglet* ***Série****.*

Dans le cadre ***Nom****, vous pouvez rentrer* ***Variable F****.*

Dans le cadre ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de F en sélectionnant la bonne plage de données dans le fichier Excel. Cliquer sur* ***Terminer****.*

C) La boîte à moustaches (Box-Plot en anglais)

	F
Minimum	*1*
D_1	*2*
Q_1	*2*
Médiane	*3*
Q_3	*5*
D_9	*6*
Maximum	*9*

• *Sélectionner les deux colonnes du tableau ci-dessus sans les deux premières et dernière lignes. Cliquer sur l'icône* ***Assistant Graphique****.*

1. *sélectionner un graphique en courbes avec marques puis cliquer sur* ***Suivant****.*
2. *sélectionner l'option* ***Lignes*** *→ cliquer sur l'onglet* ***Série*** *→ cliquer dans la partie* ***Étiquettes des abscisses (X)*** *puis sélectionner la cellule contenant F → cliquer sur* ***Suivant****.*
3. *sélectionner* ***Légende****; ne pas cocher* ***Afficher la légende****, cliquer sur* ***Terminer*** *pour créer le graphique.*

• *Activer la deuxième série de données "1er Quartile", ouvrir la boîte de dialogue* ***Format de série de données*** *:*

1. *Cliquer sur l'onglet* ***Motifs****, sélectionner Aucun pour l'option* ***Trait****; Aucune pour l'option* ***Marque****.*
2. *Cliquer sur l'onglet* ***Options****, sélectionner Lignes haut/bas et Barres hausse/baisse.*
3. *Cliquer sur l'onglet* ***Ordre des séries*** *: la série "1 er Quartile" est sélectionnée. Cliquer sur Déplacer vers le haut, sélectionner la série "3e Quartile", cliquer sur déplacer vers le bas. L'ordre final des séries doit être : 1er Quartile, 1er Décile, Médiane, 9e Décile, 3e Quartile.*

• *Activer la série de données "1er Décile", ouvrir la boîte de dialogue* ***Format de série de données****, cliquer sur l'onglet* ***Motifs****, sélectionner Aucun pour l'option* ***Trait****; Barre horizontale pour l'option* ***Marque****, taille 10 pts . On peut modifier les couleurs à cet endroit.*

• *recommencer la même manipulation avec les séries "Médiane" et "9e Décile", ouvrir la boîte de dialogue Format de série de données, cliquer sur l'onglet Motifs. Sélectionnez Aucun pour l'option Trait; Barre horizontale pour l'option Marque, taille 10 pts.*

• *Activer la série de données "3e Quartile", ouvrir la boîte de dialogue* ***Format de série de données****, cliquer sur l'onglet* ***Motifs****. Sélectionnez Aucun pour l'option* ***Trait****; Aucune pour l'option* ***Marque****.*

• *Activer le graphique Dans la barre de menu Graphique → cliquer sur* ***Ajouter des données****. Quand la boîte de dialogue* ***Ajouter des données*** *apparaît sélectionner la cellule contenant le minimum → cliquer sur OK.*

1. *Activer la série de données "Minimum". Ouvrir la boîte de dialogue* ***Type de graphique*** *et choisir* ***Nuages de points*** *→ cliquer sur OK.*
2. *Activer la série de données "Minimum", ouvrir la boîte de dialogue* ***Format de série de données*** *→ cliquer sur l'onglet* ***Motifs****. Sélectionnez Aucun pour l'option* ***Trait****, et un cercle pour l'option* ***Marque****.*

• *Recommencer avec la série de données "Maximum".*

2) Etude d'une variable quantitative

Il est possible de représenter graphiquement

- *une variable quantitative discrète par un* ***diagramme en bâtons****.*
- *une variable quantitative continue par un* ***histogramme****. Dans le cas où les classes ont même amplitude, on peut de plus tracer le polygone des effectifs. Dans le cas contraire, on ne représente pas les effectifs mais les densités d'effectifs.*

A) L'histogramme avec des classes de même amplitude

Exemple *: Nous allons représenter la variable "PNB/hab" avec le regroupement en classes suivant* $]0;10]$*,* $]10;20]$*,* $]20;30]$*,* $]30;40]$ *et* $]40;50]$*. On dispose du tableau suivant*

Classes	*Effectifs*
$]0;10]$	*12*
$]10;20]$	*6*
$]20;30]$	*8*
$]30;40]$	*1*
$]40;50]$	*1*

• *Sélectionner le tableau d'effectifs de la variable PNB/hab.*

• *Cliquer sur l'icône* ***Assistant Graphique*** *et choisir* ***Histogramme*** *puis sur* ***Suivant****.*

• *Cliquer sur l'onglet* ***Série****.*

Dans le cadre ***Nom****, vous pouvez rentrer* ***PNB/hab****.*

Dans le cadre ***Valeurs****, sélectionner les différents effectifs de PNB/hab en sélectionnant la bonne plage de données dans le fichier Excel (deuxième colonne du tableau ci-dessus).*

Dans le cadre ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de PNB/hab en sélectionnant la bonne plage de données dans le fichier Excel (première colonne du tableau). Cliquer sur* ***Terminer****.*

• *Cliquer sur une des colonnes de l'histogramme. Cliquer sur l'onglet* ***Options****, régler la largeur de l'intervalle sur 0. Ainsi les différentes barres de l'histogramme sont collées.*

B) Le polygone des effectifs *On va maintenant ajouter sur le même graphique le polygone des effectifs. On modifie le tableau des données en ajoutant des lignes fictives :*

Classes	*Effectifs*
	0
]0; 10]	*12*
]10; 20]	*6*
]20; 30]	*8*
]30; 40]	*1*
]40; 50]	*1*
	0

• *Sélectionner le tableau d'effectifs de la variable PNB/hab (deuxième colonne du tableau).*

• *Cliquer sur l'icône* ***Assistant Graphique*** *puis sur l'onglet* ***Types personnalisés*** *et choisir* ***Courbes-Histogramme*** *puis sur* ***Suivant****.*

• *Cliquer sur l'onglet* ***Série****.*

Dans le cadre ***Nom****, vous pouvez rentrer* ***Histogramme des effectifs****.*

Dans le cadre ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de PNB/hab en sélectionnant la bonne plage de données dans le fichier Excel (première colonne du tableau ci-dessus).*

• *Cliquer sur* ***Ajouter****.*

Dans le cadre ***Nom****, vous pouvez rentrer* ***Polygone des effectifs****.*

Dans les cadres ***Valeurs*** *et* ***Etiquettes des abscisses****, sélectionner les mêmes plages de données que pour l'histogramme (première et deuxième colonnes du tableau ci-dessus)* → ***Terminer****.*

• *Cliquer sur une des colonnes de l'histogramme. Cliquer sur l'onglet* ***Options****, régler la largeur de l'intervalle sur 0. Ainsi les différentes barres de l'histogramme sont collées.*

C) L'histogramme avec des classes d'amplitudes différentes

Exemple *: Nous allons représenter la variable "PNB/hab" avec un regroupement en classes différents*]0; 5],]5; 10],]10; 20],]20; 30] *et*]30; 50]. *On dispose donc du tableau suivant*

Classes	*Effectifs*
]0; 5]	*9*
]5; 10]	*3*
]10; 20]	*6*
]20; 30]	*8*
]30; 50]	*2*

On transforme ce tableau en un tableau un peu artificiel avec les densités d'effectifs

Début des classes	*Densités d'effectifs*	*Début des classes*	*Densités d'effectifs*
0	*0*	*20*	*0,6*
0	*1,8*	*20*	*0*
5	*1,8*	*20*	*0,8*
5	*0*	*30*	*0,8*
5	*0,6*	*30*	*0*
10	*0,6*	*30*	*0,1*
10	*0*	*50*	*0,1*
10	*0,6*	*50*	*0*

• *Sélectionner les deux colonnes du tableau ci-dessus.*

• *Cliquer sur l'icône* ***Assistant Graphique*** *et choisir* ***Nuage de points reliés par une courbe sans marquage des données*** *puis sur* ***Suivant****.*

• *Cliquer sur l'onglet* ***Série****.*

Dans le cadre ***Nom****, vous pouvez rentrer* ***PNB/hab****.*

Dans le cadre ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de PNB/hab en sélectionnant la bonne plage de données dans le fichier Excel (première colonne du tableau ci-dessus)* → ***Terminer****.*

• *Cliquer sur une des colonnes de l'histogramme. Cliquer sur l'onglet* ***Options****, régler la largeur de l'intervalle sur 0. Ainsi les différentes barres de l'histogramme sont collées.*

3) Etude de deux variables

Dans le cas de deux variables, il est possible de représenter la distribution conjointe et les distributions conditionnelles.

A) Distribution conjointe

Exemple *: Nous allons représenter les variables "Membre" et "PNB/hab" avec le regroupement en classes suivant* $]0;10]$, $]10;20]$, $]20;30]$, $]30;40]$ *et* $]40;50]$. *On dispose du tableau suivant*

		PNB par habitant en milliers de dollars				
		]0 ;10]	*]10 ;20]*	*]20 ;30]*	*]30 ;40]*	*]40 ;50]*
Membre de l'U.E.	*OUI*	*0*	*5*	*8*	*1*	*1*
	NON	*12*	*1*	*0*	*0*	*0*

• *Sélectionner le tableau d'effectifs.*

• *Cliquer sur l'icône* ***Assistant Graphique*** *et choisir* ***Histogramme*** *puis sur* ***Suivant****.*

- *Dans l'onglet* ***Plage de données****, choisir* ***Série en lignes*** *plutôt que* ***colonnes****.*
- *Cliquer sur l'onglet* ***Série****.*

Pour la série 1 :

Dans le cadre ***Nom****, vous pouvez rentrer* ***Oui****.*

Dans le cadre ***Valeurs****, sélectionner les différents effectifs de PNB/hab pour les membres (première ligne de données du tableau ci-dessus).*

Dans le cadre ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de PNB/hab.*

Pour la série 2 :

Dans le cadre ***Nom****, vous pouvez rentrer* ***Non****.*

Dans le cadre ***Valeurs****, sélectionner les différents effectifs de PNB/hab pour les non membres (deuxième ligne de données du tableau ci-dessus).*

Dans le cadre ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de PNB/hab.*

- *Cliquer sur* ***Terminer****.*

Remarque 4.1. *On peut aussi inverser les séries.*

B) Distributions conditionnelles

Exemple *: Nous allons représenter la distribution de la variable "PNB/hab" conditionnellement à "Membre". On travaille sur le même tableau.*

- *Sélectionner le tableau d'effectifs.*
- *Cliquer sur l'icône* ***Assistant Graphique*** *et choisir* ***Histogramme*** *puis* ***Histogramme empilé 100%*** *puis sur* ***Suivant****.*
- *Dans l'onglet* ***Plage de données****, choisir* ***Série en lignes*** *plutôt que* ***colonnes****.*
- *Cliquer sur l'onglet* ***Série****. Pour chacune des séries, rentrer dans nom la classe correspondante du PNB/hab.*

Dans le cadre ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de Membre : oui et non.*

- *Cliquer sur* ***Terminer****.*

4.2 La corrélation linéaire

L'objectif de ce paragraphe est d'utiliser le logiciel EXCEL pour automatiser les calculs statistiques étudiés dans le chapitre sur la corrélation linéaire (moyenne, variance, écart-type, covariance, coefficient de corrélation linéaire) mais aussi de voir comment obtenir facilement le nuage de

points associé à un couple de variables ainsi que la droite de régression. Comme au paragraphe précédent, nous allons regarder un cas où la corrélation linéaire est forte et un cas où elle est faible. Considérons uniquement la population des 10 pays en passe d'adhérer à l'UE (qui ont effectivement adhéré à l'UE le 1/5/2004).

Notons :

- *X la variable " Population " (non regroupée en classes)*
- *Y la variable " PNB global "*
- *Z la variable " PNB/habitant " (non regroupée en classes)*

On s'intéresse aux couples (X, Y) (on peut penser qu'il y a un lien mais est-il linéaire ?) et (X, Z) (on peut penser qu'il n'y a pas de lien puisque de façon générale, il existe des petits pays riches, des grands pays riches, des petits pays pauvres et des grands pays pauvres : beaucoup d'exemples viennent à l'esprit). A l'aide du ***TRI*** *et des fonctions* ***COPIER/COLLER****, dresser le tableau suivant sur une nouvelle feuille (compléter par des colonnes et des lignes de calcul).*

Pays	*Population*	*PNB global*	*PNB/habitant*	x_i^2	y_i^2	z_i^2	$x_i y_i$	$x_i z_i$
Chypre	*740*	*9*	*12162*					
Slovénie	*1991*	*19*	*9543*					
Malte	*373*	*3*	*8043*					
Rép. Tchèque	*10315*	*54*	*5235*					
Hongrie	*10193*	*43*	*4219*					
Slovaquie	*5343*	*19*	*3556*					
Pologne	*38618*	*134*	*3470*					
Estonie	*1466*	*4*	*2729*					
Lituanie	*3709*	*8*	*2157*					
Lettonie	*2490*	*5*	*2008*					
Somme								
Moyenne								
Variance								
Ecart-type								

D'où

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \\ \mathrm{Cov}(X,Z) &= \\ r(X,Y) &= \\ r(X,Z) &= \end{aligned}$$

Remplir toutes les cases en utilisant en particulier la fonction ***Remplissage*** *c'est-à-dire saisir la formule sur la première ligne (respectivement colonne) puis* ***Remplissage en bas*** *(respecti-*

vement ***à droite****).*

Attention : les résultats doivent s'adapter automatiquement si on modifie les valeurs de départ.

A l'issue des calculs, on trouve $r(X,Y) = 0,9835$ *: le coefficient de corrélation linéaire est très proche de 1, on peut donc considérer que, pour ces dix pays, il y a une forte corrélation linéaire positive entre la population et le PNB global. Il ne faut pas déduire de ce résultat qu'il y a toujours corrélation linéaire entre la population et le PNB global; en effet, même si on peut penser que de façon générale, le PNB global est lié à la population, cette corrélation n'est linéaire que si on considère des pays dont les niveaux de développement sont voisins. Nous faisons de la statistique descriptive : les conclusions que l'on tire ne portent que sur la population étudiée. Par contre,* $r(X,Z) = -0,2842$ *: le coefficient de corrélation linéaire est proche de 0, ce qui confirme une faible corrélation linéaire négative.*

Pour terminer, nous allons voir comment obtenir directement ces résultats avec EXCEL en utilisant un graphique :

1. *Sélectionner avec la souris les colonnes Population, PNB global et PNB/habitant (plage* ***B1 :D11****).*
2. *Cliquer sur l'icône* ***Graphique*** *de la barre d'outils.*
3. *Choisir le graphique* ***Nuage de points****.*
4. *Cliquer 3 fois sur* ***Suivant*** *puis sur* ***Terminer****, on obtient un graphique qui donne la population en abscisse et le PNB en ordonnée mais ce graphique n'a pas de sens puisque le PNB global et le PNB par habitant ne sont pas exprimés dans la même unité, nous allons donc d'abord supprimer les points qui correspondent au nuage* (X,Z) *en cliquant sur un des points du nuage (tous les points du nuage correspondant au PNB/habitant sont simultanément sélectionnés) puis en appuyant sur la touche* ***Suppr*** *du clavier.*
5. *Cliquer maintenant* ***avec le bouton droit de la souris*** *sur un des points du nuage correspondant au PNB global; dans la fenêtre qui s'ouvre, choisir* ***Ajouter une courbe de tendance*** *dans* ***Type*** *choisir* ***Linéaire*** *et dans* ***Options****, cocher* ***Afficher l'équation sur le graphique et Afficher le coefficient de détermination*** *(*R^2*) sur le graphique puis cliquer sur* ***OK*** *: on obtient la droite de régression de* Y *en* X *traversant le nuage de points ainsi que son équation et le carré du coefficient de corrélation linéaire que nous avons*

calculé (on peut calculer la racine pour vérifier).

6. *Recommencer les 5 étapes en conservant cette fois le nuage de points associé au couple* (X, Z).

4.3 Taux de variation et courbe semi-logarithmique

Le tableau suivant donne l'évolution de la population de New York, en milliers d'habitants, de 1800 à 1910.

Année	*1800*	*1810*	*1820*	*1830*	*1840*	*1850*	*1860*	*1870*	*1880*	*1890*	*1900*	*1910*
Population	*75*	*110*	*130*	*230*	*400*	*700*	*1100*	*1300*	*1900*	*2300*	*3200*	*4500*

Reproduire ce tableau sur une feuille de calcul EXCEL sur la plage ***A1 :M2*** *(on prendra soin de mettre en forme le tableau en ajoutant les bordures et en réduisant la largeur des colonnes B,...,M de sorte à avoir tout le tableau sur l'écran :* ***sélectionner avec la souris les colonnes B,...,M*** *puis dans* ***Format****, choisir* ***Colonne****, cliquer sur* ***Largeur*** *et taper* ***6)***

Pour chaque décennie, on souhaite calculer le taux de variation annuel moyen d'augmentation de la population (notons " TAM sur 10 ans ", la variable qui donne le taux de variation annuel moyen entre les années n-10 et n) et rajouter une ligne au tableau ci-dessus : Dans la cellule ***C3****, faire le calcul du taux de variation annuel moyen de la population de New York entre 1800 et 1810 :* ***taper*** *=, parmi les fonctions proposées, choisir* ***puissance*** *(lire le mode d'emploi de cette fonction pour continuer) et terminer le calcul. Afin de ne pas recommencer 10 fois ce calcul, utiliser la fonction de* ***Remplissage à droite*** *(dans le menu* ***Edition****). Pour avoir les résultats en pourcentage, sélectionner la plage* ***C3 :M3*** *avec la souris et cliquer sur l'icône % de la barre d'outils (on gardera deux chiffres après la virgule dans l'expression en pourcentage).*

On veut maintenant mettre en évidence la croissance exponentielle de la population de New York au XIXème siècle à l'aide d'une représentation graphique cartésienne (courbe classique) puis, grâce à une courbe semi-logarithmique, mettre en évidence le ralentissement de l'augmentation à partir de 1860 :

1. *Sélectionner avec la souris la plage* ***A1 :M2****.*
2. *Dans le menu* ***Insertion****, cliquer sur* ***graphique*** *(ou cliquer directement sur* ***l'icône Graphique*** *de la barre d'outils), choisir* ***Nuage de points reliés par une courbe****, cliquer sur* ***Suivant*** *plusieurs fois puis sur* ***Terminer*** *: on obtient une courbe qui rappelle celle de*

la fonction exponentielle (on peut essayer, en cliquant sur le graphique de modifier certain paramètres ou bien d'agrandir le graphique de sorte à le rendre plus lisible et agréable).

3. *Pour obtenir la courbe semi-logarithmique, reprendre les étapes 1 et 2, placer le nouveau graphique (pour l'instant identique au premier) sous le précédent,* ***cliquer sur la courbe avec le bouton droit de la souris****, cliquer sur* ***Type de graphique*** *et dans* ***Types personnalisés****, choisir* ***Logarithmique*** *cliquer sur* ***OK****. On obtient une courbe semi-logarithmique représentant l'évolution de la population de New York et la encore on peut modifier à sa convenance quelques paramètres pour affiner le résultat.*

4.4 Les séries temporelles

4.4.1 Traitement des tableaux

Nous considérons l'emple du cours concernant les consommations trimestrielles en électricité d'une entreprise de vente par Internet pendant les trois premières années :

Années / *Trimestres*	*1997*	*1998*	*1999*
1	*4,5*	*5,5*	*7,2*
2	*4,1*	*4,9*	*6,4*
3	*3,7*	*4,4*	*4,8*
4	*5,1*	*6,5*	*6,8*

On étudie donc un phénomène dont les saisons sont trimestrielles.

Pour traiter les séries temporelles, nous avons besoin des données rangées de deux façons différentes selon les deux tableaux suivants :

années	*trimestres*	*rang t*	*variable* $y(t)$
1	T_1	*1*	$y(1)$
1	T_2	*2*	$y(2)$
1	T_3	*3*	$y(3)$
1	T_4	*4*	$y(4)$
2	T_1	*5*	$y(5)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
10	T_3	*39*	$y(39)$
10	T_4	*40*	$y(40)$

trimestres	*année 1*	*année 2*	...	*année 10*
T_1	$y(1)$	$y(5)$	...	...
T_2	$y(2)$	...	...	...
T_3	$y(3)$	...	...	$y(39)$
T_4	$y(4)$	...	...	$y(40)$

S'il n'y a pas trop de données, il est possible de construire l'un à partir de l'autre en utilisant les fonctions ***Copier*** *et* ***Coller****.*
Construire à la main les deux tableaux.

S'il y a trop de données, on peut utiliser la fonction tableau croisé dynamique (prendre soin de donner des libellés par ordre alphanumérique aux années et aux trimestres car les résultats seront renvoyés dans cet ordre).

a) Transformation du tableau 1 en tableau 2
Supposons que nous disposions des données sous la forme du tableau 1. Construire donc à la main le tableau 1.

- *Sélectionner le tableau entier (y compris la première ligne).*
- *Sélectionner* ***Rapport de tableau croisé dynamique*** *dans le menu* ***Données****.*
- *Cliquer deux fois sur* ***suivant****.*
- *Dans l'onglet* ***Disposition****, déplacer avec la souris l'icône de la variable* ***trimestres*** *dans le cadre ligne puis l'icône de la variable* ***années*** *dans le cadre colonne et enfin l'icône de la variable* ***variable*** *y dans le cadre données. Cliquer sur* ***suivant****.*
- *Choisir de placer le tableau dans la feuille existante et sélectionner la plage de données qui recevra le second tableau. Cliquer sur* ***terminer****.*

b) Transformation du tableau 2 en tableau 1 *Avant d'utiliser la fonction tableau croisé dynamique, on commence par transposer le tableau 2 situé par exemple en F1 :P5.*

- *Le tableau 2 comportant 5 lignes et 11 colonnes, sélectionner une plage de 55 cellules sur 5 colonnes et 11 lignes où l'on désire faire apparaître les résultats.*
- *Taper =* ***TRANSPOSE(F1 :P5)****. Appuyer ensemble les touches* ***CTRL*** *,* ***MAJ****,* ***Entrée*** *(la touche MAJ est la touche majuscule) pour valider la saisie.*
- *Sélectionner le nouveau tableau, y compris la première ligne.*
- *Sélectionner* ***Rapport de tableau croisé dynamique*** *dans le menu* ***Données****.*
- *Cliquer deux fois sur* ***suivant****.*
- *Dans l'onglet* ***Disposition****, déplacer avec la souris l'icône de la variable* ***trimestres*** *dans le cadre ligne. Déplacer successivement avec la souris les icônes des variables* ***T1,T2,T3,T4****, dans*

le cadre données. Cliquer ***suivant****.*

• *Choisir de placer le tableau dans la feuille existante et sélectionner la plage de données qui recevra le second tableau. Cliquer sur* ***terminer****.*

4.4.2 Analyse de la série

a) Détermination de la tendance *On se sert du tableau 1, où les valeurs des rangs et de la variable se situent par exemple en C2 :D13.*

- par régression linéaire :

• *Sélectionner une plage de 12 cellules sur une colonne (autant que d'observations Y) où l'on désire faire apparaître les résultats.*

• *Taper =* ***TENDANCE(D2 :D13 ;C2 :C13 ;C2 :C13)****. Appuyer ensemble les touches* ***CTRL****,* ***MAJ****,* ***Entrée*** *(la touche MAJ est la touche majuscule) pour valider la saisie.*

- par moyennes mobiles :

On calcule à la main la première valeur puis on étire la formule pour obtenir les autres valeurs de la série transformée $z(t)$ par moyenne mobile.

Remarque 4.2. *On pourrait penser utiliser les menus* ***Outils****,* ***Utilitaire d'analyse****,* ***Moyenne mobile****, mais seul le cas d'une moyenne mobile d'ordre impair est presque exact (les valeurs sont décalées par rapport aux rangs concernés) et cette fonction n'est pas disponible dans toutes les versions Excel. Il est donc préférable et aussi rapide d'écrire la formule d'obtention de la moyenne mobile et d'utiliser la fonction Recopier.*

On calcule ensuite la droite de régression avec Excel. On obtient le a et le b et on calcule la série obtenue par régression linéaire.

b) Détermination des coefficients saisonniers *On retranche à la série de départ la série obtenue par régression.*
Les résultats étant obtenus dans un tableau du style 1, il faut le transformer en tableau 2, à partir duquel seront facilement calculés les coefficients saisonniers. ***Ne pas oublier de retirer la moyenne des coefficients saisonniers !***

c) Série désaisonnalisée *Elle sera facilement obtenue sous forme de tableau 2 à partir des coefficients saisonniers puis transformée sous forme de tableau 1 pour les représentations graphiques.*

d) Représentations graphiques *Les données à représenter se présentent sous la forme suivante*

rang t	*variable* y_t	*tendance* x_t	$\hat{y}_t$.

• *Sélectionner le tableau entier (y compris la première ligne).*

• *Sélectionner l'icône* ***assistant graphique****.*

• *Dans type de graphiques sélectionner* ***nuage de points*** *; dans sous-types de graphiques sélectionner celui où les point sont reliés par des segments de droite. Cliquer* ***suivant****.*

• *Cliquer deux fois sur* ***suivant****.*

• *Cliquer sur* ***terminer*** *pour obtenir le graphique sur la même feuille ou préciser sur nouvelle feuille.*

e) Prévisions *On peut maintenant faire des prévisions. Par exemple, prédire les valeurs des stocks pour l'année 2000. Retrouve-t-on bien les résultats donnés en cours ?*

Chapitre 5

Fiches récapitulatives

Voici quatre fiches récapitulatives synthétisant les différentes notions vues dans ce cours. Elles vous aideront dans vos révisions et vous pourrez vous en inspirer pour réaliser la feuille recto-verso manuscrite A4 autorisée à l'examen.

Fiche résumé chapitre 1 : Taux de variation et courbes semi-logarithmiques

Pourcentage *: C'est une fraction dont le dénominateur est 100 :* $a\% = \dfrac{a}{100}$ *(*$a = 100 \times a\%$*).*

Multiplicateur et taux de variation *:*

Si une valeur x *subit une variation de* $a\%$ *(*a *positif ou négatif, supérieur à -100), sa nouvelle valeur est :* $\left(1 + \dfrac{a}{100}\right) x$ *et* $1 + \dfrac{a}{100}$ *est le multiplicateur.*

Lorsqu'on connaît la valeur V_1 *de départ et la valeur* V_2 *d'arrivée :*

$$\boxed{\text{Multiplicateur} = \frac{V_2}{V_1} = \text{Taux de variation} + 1 \qquad \text{Taux de variation} = \frac{V_2}{V_1} - 1 = \frac{V_2 - V_1}{V_1} = \text{Multiplicateur} -}$$

Variations successives

Soit x *une valeur donnée qui subit deux variations successives de* $a\%$ *puis* $b\%$*. Le multiplicateur permettant de passer à la nouvelle valeur est* $\boxed{\left(1 + \dfrac{a}{100}\right)\left(1 + \dfrac{b}{100}\right)}$.

De façon générale, si l'on a n *variations successives, le multiplicateur permettant de passer de la première à la dernière valeur est le produit des* n *multiplicateurs.*

Lorsqu'il y a n *variations successives et identiques de* $a\%$*, le multiplicateur est* $\left(1 + \dfrac{a}{100}\right)^n$.

Taux de variation moyen

Si x *a subi une variation globale de* $a\%$*, le taux de variation moyen* $\alpha\%$ *sur* n *périodes est le pourcentage tel que* n *variations successives de* $\alpha\%$ *donnent une variation de* $a\%$*.*

Le multiplicateur moyen est $\boxed{1 + \dfrac{\alpha}{100} = \left(1 + \dfrac{a}{100}\right)^{1/n}}$.

Courbes semi-logarithmiques

Pour tracer des courbes semi-logarithmiques, on s'intéressera à la fonction logarithme décimal, notée $\log(x)$*, définie à partir du logarithme népérien par :*

$$\log(x) = \ln(x)/\ln(10)$$

Le logarithme décimal est une fonction strictement croissante qui " tasse " les valeurs :

$$\log(1) = 0, \log(10) = 1, \log(100) = 2, \log(1000) = 3, \ldots$$

Un repère semi-logarithmique se construit en graduant l'axe des abscisses (axe horizontal) comme d'habitude, régulièrement (par exemple, 1 cm correspond à 10 ans) mais on doit reporter sur l'axe

des ordonnées (axe vertical) non pas les valeurs données mais les logarithmes décimaux de ces valeurs. Cependant, pour une lecture agréable du graphique, on écrira la valeur donnée et non son logarithme. Cela permet de représenter sur un graphique lisible de très grandes variations.

On appelle ***module*** *d'un repère semi-logarithmique la distance sur l'axe des ordonnées entre les graduations 1 et 10 (c'est-à-dire aussi entre 10 et 100 ou entre 100 et 1000 ou ...). C'est en fait l'unité sur l'axe des ordonnées puisque* $\log(10) - \log(1) = 1$.

Remarque importante : A une variation de $a\%$ *entre deux valeurs correspond toujours la même hauteur sur le graphique. Ainsi, un graphique semi-logarithmique fait apparaître les variations en pourcentage ; plus précisément, une pente régulière sur un graphique semi-logarithmique signifie que le taux de variation a été régulier sur la période.*

Fiche résumé chapitre 2 : La corrélation linéaire

Définitions *: Soient X et Y deux variables définies sur une même population Ω de taille N. Les modalités de (X,Y) sont notées (x_i, y_i), $\overline{X}$ est la moyenne des valeurs x_i et $\overline{Y}$ la moyenne des valeurs y_i.*

- *L'ensemble des points A_i de coordonnées $(x_i; y_i)$ s'appelle le* ***nuage de points*** *associé au couple (X,Y).*
- *Le point M de coordonnées $(\overline{X}; \overline{Y})$ s'appelle le* ***point moyen*** *du nuage.*

On souhaite définir un coefficient qui permette de mesurer l'allongement du nuage de points ; plus précisément, on cherche à savoir si l'on peut considérer que les points du nuage sont globalement proches d'une certaine droite traversant le nuage de points. Si la réponse est positive, on dira qu'il y a une ***corrélation linéaire*** *entre les variables X et Y.*

Covariance

$$\boxed{\mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{k} n_i x_i y_i}{N} - \overline{XY}}$$

Droite de régression de Y en X (notée $D_{Y/X}$)

Chaque fois que l'on trace une droite D d'équation $y = ax + b$ qui traverse le nuage de points, on peut projeter sur cette droite les points A_i parallèlement à l'axe des ordonnées (c'est-à-dire verticalement) et noter P_i les projetés.

Notons A_iP_i la distance entre A_i et son projeté vertical P_i. On veut minimiser la somme des carrés de ces distances. Plus précisément, le point P_i a pour coordonnées $(x_i; ax_i + b)$ et on veut minimiser la somme des $(y - (ax_i + b))^2$.

*C'est en ce sens que la droite que l'on cherche passe le plus près possible de l'ensemble des points du nuage ; c'est pourquoi lorsqu'on souhaite faire une approximation d'un nuage de points par une droite on parle d'****ajustement affine par la méthode des moindres carrés****.*

Il existe une droite unique associée au nuage de points $A_i(x_i; y_i)$ telle que la somme des $A_iP_i^2$ soit minimale. Cette droite passe par le point moyen $(\overline{X}; \overline{Y})$ et a pour équation

$$\boxed{y = ax + b \text{ où } a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \text{ et } b = \overline{Y} - a\overline{X}}$$

Cette droite s'appelle la ***droite de régression de*** Y ***en*** X *et on la note* $D_{Y/X}$.

Coefficient de corrélation linéaire

$$\boxed{r(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}}$$

Le coefficient de corrélation linéaire est toujours compris entre -1 et 1.

- *Si* $r(X,Y)$ *est proche de 1, il y a une forte corrélation linéaire positive (concrètement, cela traduit le fait que quand* X *augmente,* Y *augmente et ceci de façon linéaire).*
- *Si* $r(X,Y)$ *est proche de 0, la corrélation linéaire est faible.*
- *Si* $r(X,Y)$ *est proche de -1, il y a une forte corrélation linéaire négative (concrètement, cela traduit le fait que quand* X *augmente,* Y *diminue et ceci de façon linéaire).*

Fiche résumé chapitre 3 : Les séries temporelles

1-Définition et objectifs

Une série ***temporelle*** *ou* ***chronologique*** *est l'observation à certains instants d'un phénomène qui évolue en fonction du temps. Soit Y la variable statistique étudiée, on note $Y(t)$ la valeur de cette variable à l'instant t. Soient t_1, t_2,... , t_N les N instants de mesures, la série chronologique correspondante est $(Y(t_1), \ldots, Y(t_N))$.*

Objectif *: MODÉLISER et PRÉDIRE*

La modélisation se fait en considérant deux composantes :

- *la tendance qui caractérise l'allure générale de la série*
- *la saisonnalité qui permet de prendre en compte le caractère saisonnier éventuel de la série.*

2-Représentation graphique

Il est important de représenter la série pour effectuer une modélisation adéquate :

- *une représentation cartésienne de la série permet d'en voir l'allure générale. Elle permet d'avoir des indications sur la tendance et sur l'existence d'une composante saisonnière.*
- *une représentation cartésienne avec périodes superposées permet de préciser la valeur de la période choisie.*

3-Modélisation

Lissage d'une courbe : la série des moyennes mobiles

On se demande si en faisant abstraction des variations saisonnières, la série statistique suit grossièrement une droite. Pour répondre à cette question, nous allons désaisonnaliser la série à l'aide des moyennes mobiles. On suppose que la série est découpée en périodes avec p observations par période (c'est à dire p " saisons ").

On appelle série des moyennes mobiles d'ordre p la série des valeurs $z(t)$ suivante :

- *Si p est pair (c'est le cas lorsqu'on découpe l'année en 4 trimestres),*

$$z(t) = \frac{\frac{y(t-k)}{2} + y(t-k+1) + \ldots + y(t) + \ldots + y(t+k-1) + \frac{y(t+k)}{2}}{p}$$

(il y a p+1 termes au numérateur avec $y(t)$ au milieu).

• *Si p est impair (c'est le cas lorsqu'on découpe la semaine en 7 jours),*

$$z(t) = \frac{y(t-k) + y(t-k+1) + \ldots + y(t) + \ldots + y(t+k-1) + y(t+k)}{p}$$

(il y a p termes au numérateur avec $y(t)$ au milieu).

Remarque : La série des moyennes mobiles comporte $N - 2k$ termes.

La droite de tendance (ou trend)

On suppose que l'évolution de la série désaisonnalisée se fait selon une droite appelée droite de tendance : c'est la droite de régression de Z en T_Z, où T_Z est la série des instants prenant en compte seulement ceux pour lesquels $z(t)$ est calculé. On notera son équation sous la forme :

$$x(t) = at + b$$

et comme au chapitre 2, on a : $a = \dfrac{\mathrm{Cov}(Z, T_Z)}{\mathrm{Var}(T_Z)}$ et $b = \overline{Z} - a\overline{T}_Z$, en prenant

$$\boxed{\overline{T}_Z = \frac{N+1}{2}} \text{ et } \boxed{\mathrm{Var}(T_Z) = \frac{N_Z^2 - 1}{12}}$$

Les coefficients saisonniers

Nous allons développer le modèle additif (il existe aussi un modèle multiplicatif que nous ne verrons pas ici). Dans ce modèle, le coefficient saisonnier est ajouté à la composante linéaire :

$$\boxed{\hat{y}(t) = x(t) + s(t)}$$

Le chapeau sur y permet de bien différencier les données brutes y du modèle théorique que l'on construit $\hat{y}$ et qui nous servira pour faire des prévisions à court terme.

• *On calcule d'abord les composantes saisonnières (une pour chaque saison), que l'on notera $s'(t)$, en faisant la moyenne des écarts algébriques entre les composantes linéaires et les valeurs observées ($\delta(t) = y(t) - x(t)$) pour chaque saison considérée.*

• *On souhaite que dans le modèle construit, les variations saisonnières autour du trend se compensent sur une période i.e. $\sum_{t=1}^{p} s(t) = 0$. Pour cela, on calcule* $\boxed{S_0 = \sum_{t=1}^{p} s'(t) = s'(1) + \ldots + s'(p)}$ *et on pose* $\boxed{s(t) = s'(t) - \frac{S_0}{p}}$.

4-Prévisions à court terme

On se sert du modèle $\boxed{\hat{y}(t) = x(t) + s(t)}$ *construit pour effectuer les prévisions à court terme.*

Chapitre 6

Devoir à rendre

Exercice 1 Le tribut de la ligue de Délos *(Source : Patrice Brun, Impérialisme et démocratie à Athènes, Ed. Armand Colin, 2005, p. 35-36.)*

Le tableau ci-dessous donne le montant des sommes exigées chaque année de quelques communautés membres de la ligue de Délos (sorte d'OTAN en Grèce au Vème siècle ***AVANT*** *J.C.).*
La première colonne concerne le montant annuel (stable) de la décénie 450-440. La deuxième colonne concerne le montant de l'année 425, qui a subi une révision exceptionnelle. L'unité de monnaie est le millier de drachme (dr.).

Lieux	*450-440*	*425*	*Lieux*	*450-440*	*425*
ANAPHE	*1*	*1*	*IMBROS*	*6*	*6*
ANDROS	*12*	*90*	*IOS*	*3*	*6*
ATHENA DIADES	*4*	*6*	*KEOS*	*24*	*60*
CARYSTOS	*18*	*30*	*KYTHNOS*	*18*	*36*
ERETRIE	*18*	*90*	*LEMNOS*	*24.3*	*24*
GRYNCHE	*1*	*2*	*MYCONOS*	*6*	*12*

Nous nous intéressons à l'existence éventuelle d'une corrélation linéaire entre le montant annuel (stable) de la décenie 450-440 et celui de l'année 425.

1. *Quelle est la population considérée ? Quelle est sa taille ?*
2. *Quelles sont les variables X et Y étudiées ? Précisez leur type ?*
3. *Représenter le nuage de points de coordonnées $M_i=(x_i,\ y_i)$ correspondant à cette série statistique. Commentez.*

4. *Calculer le coefficient de corrélation linéaire r.*
5. *Un ajustement linéaire est-il justifié ? Si oui, déterminer l'équation de la droite de régression D de y en x par la méthode des moindres carrés. Tracer la droite D sur le graphique.*

   ***On indiquera les moyennes de** X **et** Y**, les valeurs de** $\mathrm{Var}(X)$**,** $\mathrm{Var}(Y)$**,** $s(X) = \sigma(X)$**,** $s(X) = \sigma(X)$ **et** $\mathrm{Cov}(X, Y)$ ***intervenant dans les calculs**.*

Exercice 2 Évolution de la population française entre 1950 et 2000

Le tableau suivant donne l'évolution de la population française entre 1950 et 2000 (en milliers d'habitants).

	Effectifs des 0-19 ans	*Population totale*	*Pourcentage*
1950	*12494*		*30%*
2000	*14686*	*58744*	

1. *Compléter le tableau en indiquant les calculs faits.*
2. *Calculer l'évolution en pourcentage de la population totale française entre 1950 et 2000.*
3. *Calculer le taux de croissance annuel moyen de la population totale française entre 1950 et 2000.*
4. *Entre 2000 et 2050, on prévoit une augmentation de 9% de la population totale alors que la population des 0-19 ans devrait baisser de 13%. Quelle sera la part des 0-19 ans en pourcentage de la population totale, en 2050, si ces prévisions sont exactes ?*

Exercice 3 Absence journalière

Dans une grande entreprise, on a mesuré l'absence journalière pendant 4 semaines : chaque semaine comporte 5 jours de travail. Voici les résultats (on donne ici le nombre d'employés absents) :

	Semaine 1	*Semaine 2*	*Semaine 3*	*Semaine 4*
Lundi	*1*	*2*	*4*	*5*
Mardi	*0*	*3*	*4*	*6*
Mercredi	*5*	*7*	*10*	*11*
Jeudi	*2*	*4*	*2*	*3*
Vendredi	*0*	*1*	*2*	*4*

1. Calculer la première, la deuxième et la dernière moyenne mobile. Expliquer l'utilité de la série des moyennes mobiles.

2. On donne l'équation de la droite de tendance calculée à partir de la série des moyennes mobiles : $x(t) = 0,26t + 1,15$ *ainsi que la somme des composantes saisonnières :* $S_0 = -3$.

Calculer les trois premiers coefficients saisonniers dans le modèle additif.

3. Prévoir le nombre d'absents pour les 3 premiers jours de la cinquième semaine.

Chapitre 7

Enoncé des exercices

7.1 Exercices du chapitre 1 - Taux de variation

Exercice 1. *Le prix du litre d'essence est de 5 F le 1er janvier 1993. Il augmente de 25% dans les six premiers mois, puis diminue de 25% dans les six derniers mois de l'année 1993. Quel est le prix du litre d'essence au 1er janvier 1994 ?*
Lorsqu'une hausse de a% est suivie d'une baisse de a%, y a-t-il toujours globalement une baisse ? Qu'en est-il si la baisse a lieu avant la hausse ?

Exercice 2. *Certains disent "les femmes gagnent en moyenne 33% de moins que les hommes" alors que d'autres disent "les hommes gagnent en moyenne 50% de plus que les femmes". Qu'en pensez-vous ?*

Exercice 3. *Le tableau suivant indique l'évolution de la démographie française entre 1920 et 1987. Les effectifs sont donnés en millions de personnes.*

Années / Sexe	1920	1946	1970	1977	1987
Hommes	18,8	20,1	22,6	25,3	26,6
Femmes	20,2	23,9	27,4	27,7	28,4
Total	39	44	50	53	55

Source : INSEE, 1989.

1. Définir la population étudiée. Quelle est la variable (ou le caractère) dont on étudie l'évolution ? Quel est l'ensemble des modalités et le type de cette variable ?

2. Calculer le taux de croissance de la population française féminine entre 1920 et 1977.

3. Calculer le taux annuel moyen de croissance de la population française pour la période allant de 1920 à 1987.

Exercice 4. *Voici un tableau donnant les salaires annuels nets moyens (en francs) des hommes et des femmes en 1969 et 1972.*

	Hommes		*Femmes*	
	1969	*1972*	*1969*	*1972*
Région parisienne	*21729*	*28916*	*13970*	*18686*
Midi-Pyrénées	*13680*	*18409*	*8933*	*12392*
France	*16224*	*21841*	*10812*	*13133*

Source : INSEE.

1. Calculer le taux de variation du salaire des hommes en France entre 1969 et 1972. Faire le même calcul pour les femmes.

2. Calculer le taux de variation annuel moyen du salaire des hommes entre 1969 et 1972 pour la Région parisienne puis pour Midi-Pyrénées.

3. Vérifier que, en 1972, dans la Région parisienne, les hommes gagnent en moyenne 54,75% de plus que les femmes. Calculer ce pourcentage pour Midi-Pyrénées.

4. En 1969, sur l'ensemble des salariés français (hommes et femmes), le salaire moyen est de 14669 F. En déduire le pourcentage de femmes parmi les salariés en 1969.

Exercice 5. *Voici un tableau donnant le nombre d'étudiants inscrits dans l'enseignement supérieur public en France en 1900, 1929 et 1939.*

	Total étudiants	*dont étudiantes*
1900	*29377*	*3,50%*
1929	*69961*	*22,90%*
1939	*78972*	*30%*

1. Calculer le taux de variation du nombre total d'étudiants entre 1900 et 1939 puis, sur cette même période, le taux de variation du nombre d'étudiantes.

2. Calculer le taux annuel moyen de croissance du nombre d'étudiantes entre 1929 et 1939.

Exercice 6. *1. La population de l'Allemagne est passée de 36 millions en 1861 à 65 millions en 1911.*

a) Calculer le taux de variation sur cette période.

b) Calculer le taux annuel moyen de croissance.

2. La construction de voies ferrées en Europe augmente de 343% entre 1850 et 1870 puis augmente de 171% entre 1870 et 1900 . Il y avait 23500 km de voies ferrées en 1850.

a) Combien y a-t-il de kilomètres de voies ferrées en 1900 ?

b) Quel est le taux trimestriel moyen d'augmentation entre 1850 et 1870 ?

Exercice 7. *On étudie la répartition de la population active des Etats Unis entre 1870 et 1920 (en milliers) suivant les secteurs d'activité : primaire, secondaire et tertiaire.*

	1870	*1880*	*1890*	*1900*	*1910*	*1920*
Primaire	*7097*	*8980*	*10568*	*11816*	*12799*	*12809*
Secondaire	*2643*	*3842*	*5526*	*7199*	*10657*	*12861*
Tertiaire	*3185*	*4571*	*7225*	*10058*	*13916*	*16763*

Source : Statistiques internationales rétrospectives. La population active et sa structure, 1968.

1. Calculer pour chaque année les pourcentages correspondants.

2. Quel est le taux de variation entre 1870 et 1920 de la part du secteur primaire dans la population active ? Faire les calculs pour les deux autres secteurs.

3. Quel est le taux de variation moyen sur 10 ans de la part pour chacun des secteurs ?

7.2 Exercices du chapitre 1 - Courbes semi-logarithmiques

Exercice 8. *Voici (document de la page suivante) des courbes semi-logarithmiques donnant l'évolution de quelques grandes productions entre 1800 et 1914. Pour chacun des graphiques, préciser le module du repère semi-logarithmique et la hauteur correspondant à un doublement de la production. Commenter les courbes concernant l'acier en précisant pour chaque pays et chaque période le taux de variation annuel moyen.*

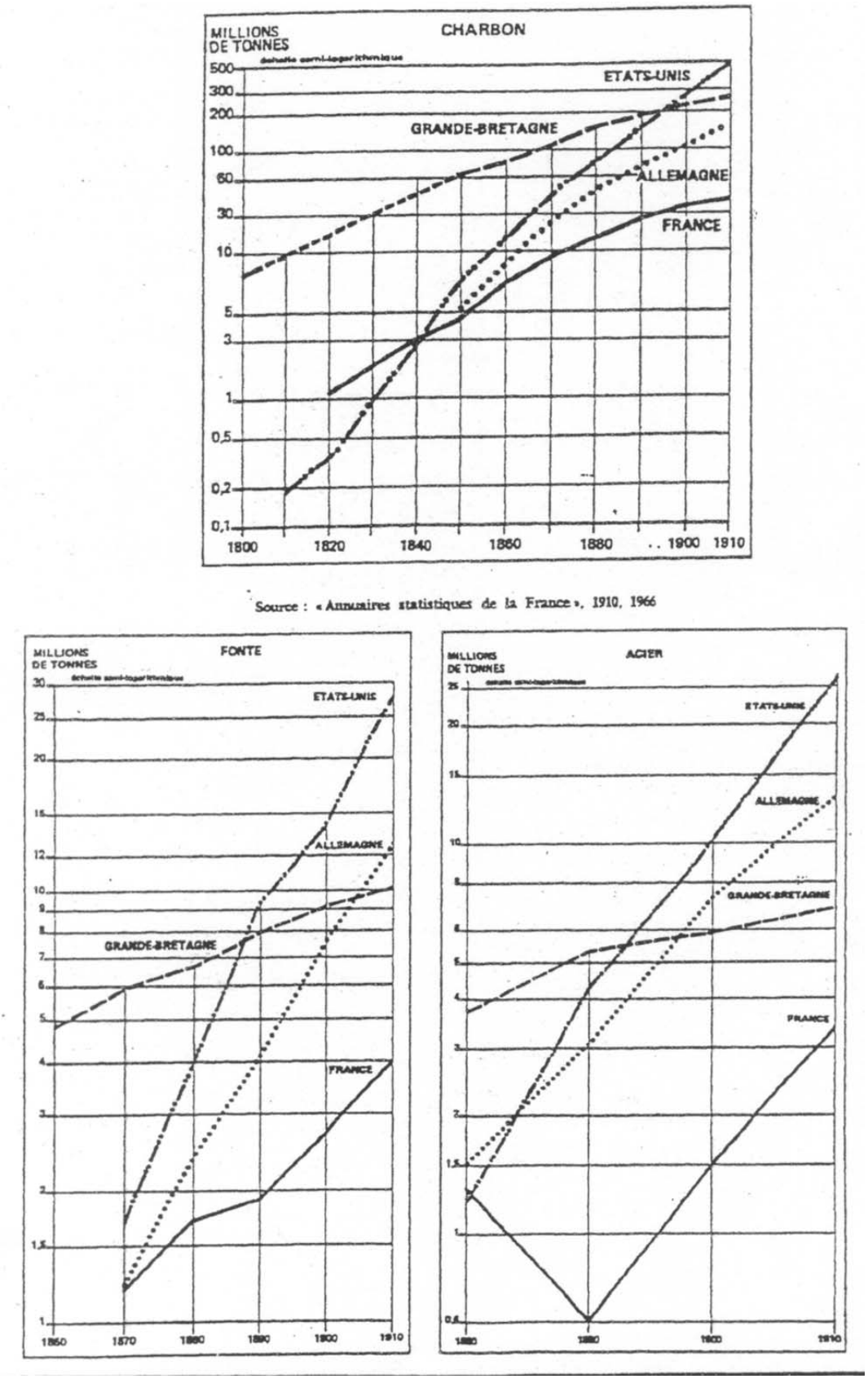


FIGURE 7.1 – *Graphique semi-logarithmique de quelques grandes productions entre 1800 et 1914.*

Exercice 9. *Voici l'évolution sur les six dernières années des taux (en pourcentages) de départ en vacances d'hiver des Gersois selon la catégorie sociale du chef de ménage :*

Années	*Cadres supérieurs*	*Cadres moyens*	*Employés*	*Ouvriers*
2005	*50%*	*39%*	*21%*	*9%*
2006	*69%*	*54%*	*30%*	*13%*
2007	*48%*	*37%*	*21%*	*8%*
2008	*75%*	*58%*	*33%*	*14%*
2009	*56%*	*43%*	*24%*	*10%*
2010	*73%*	*56%*	*31%*	*14%*

1. Pourquoi le type de diagramme semi-logarithmique est-il adapté dans ce cas particulier ?

2. Combien doit-on utiliser de modules ?

3. Construisez-le sur le papier semi-logarithmique du graphique suivant. Conclusion ?

4. Construisez-le cette fois-ci en échelle arithmétique. Conclusion ?

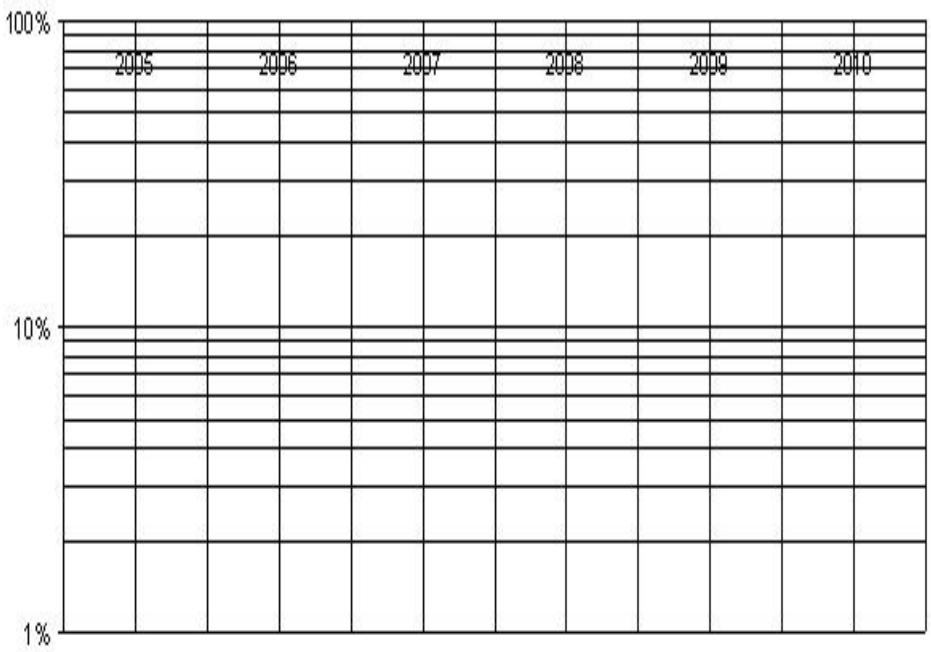


FIGURE 7.2 – *Graphique semi-logarithmique des vacanciers.*

7.3 Exercices du chapitre 2

Exercice 10. *Le tableau suivant donne la consommation de graisse par an et par personne en Norvège, ainsi que le taux de mortalité par athérosclérose pour 100 000 habitants, pendant une période qui couvre la seconde guerre mondiale.*

Année	*Consommation de graisse en kg par an et par personne, X*	*Taux de mortalité par athérosclérose pour 100 000 habitants, Y*
1938	*14,4*	*29,1*
1939	*16*	*29,7*
1940	*11,6*	*29,2*
1941	*11*	*26*
1942	*10*	*24*
1943	*9,6*	*23,1*
1944	*9,2*	*23*
1945	*10,4*	*23,1*
1946	*11,4*	*25,2*
1947	*12,5*	*26,1*

1. Construire le nuage de points.

2. Calculer les caractéristiques numériques des variables X et Y (moyenne, variance et écart-type).

3. Calculer la covariance entre X et Y.

4. Calculer le coefficient de corrélation linéaire.

5. Utiliser les questions 1. et 4. pour conclure.

6. Calculer les coefficients de régression linéaire de Y sur X. Tracer sur le graphique de la question 1. la droite de régression.

7. En 1948, la consommation de graisse était de 13 kg par personne. Le taux de mortalité par athérosclérose n'a pas été relevé cette année là. Donner une estimation de ce taux en utilisant la droite de régression obtenue à la question précédente.

Exercice 11. *Le tableau suivant indique les différentes valeurs obtenues pour deux variables quantitatives X et Y dans un échantillon de taille 8.*

X	1,3	1,5	2,5	2,5	2,7	3	4	5
Y	2,3	3,5	3,5	4,5	4,7	3	2	1

1. Construire le nuage de points.

2. Calculer le coefficient de corrélation linéaire.

3. Peut-on conclure dans ce cas à l'existence d'un lien linéaire entre X et Y ? Que suggère plutôt le nuage de points ?

4. À l'aide du nuage de points, extraire de l'échantillon initial deux sous-échantillons de taille 4 chacun et calculer le coefficient de corrélation linéaire pour ces deux échantillons. Que constate-t-on ?

Exercice 12. *Sur 12 ouvriers d'une entreprise, on a observé en 1995 l'ancienneté (X en années) et le salaire mensuel (Y en Francs).*

X	7	15	15	16	5	12	2	20	14	9	15	8
Y	8100	10200	8400	11400	6900	9600	6300	10500	10800	8100	9300	7500

1. Construire le nuage de points. Commenter.

2. Déterminer les moyennes et écarts-type de X et de Y.

3. Calculer la covariance et le coefficient de corrélation linéaire. Commenter.

4. Réaliser la régression linéaire de Y sur X. Tracer la droite de régression sur le graphique de la question 1.

5. Déterminer les valeurs ajustées, c'est-à-dire les valeurs $\widehat{y}_i = ax_i + b, i = 1, \ldots, 12$, où a et b sont les coefficients de la droite de régression (respectivement la pente et la valeur à l'origine).

6. Déterminer les résidus $y_i - \widehat{y}_i$. Calculer la moyenne des résidus : que constate-t-on ?

7. Deux ouvriers de l'entreprise ont respectivement 4 et 18 ans d'ancienneté ; donner une estimation de leur salaire à l'aide de la droite de régression.

Exercice 13. *On a demandé à une personne de mettre une note X entre 0 et 10 à 8 vins en fonction de l'étiquette figurant sur la bouteille, puis de les noter de nouveau en les dégustant en aveugle. La variable Y correspond à la deuxième note.*

Note sur l'étiquette	10	5	7	7	7	9	9	8
Note sur le goût	3	8	7	9	6	4	2	5

1. Calculer le coefficient de corrélation linéaire.

2. En fonction du résultat obtenu à la question précédente, tracer le nuage de points et conclure.

3. Réaliser s'il y a lieu la régression linéaire de Y sur X.

Exercice 14. *L'étude de la durée hebdommadaire du travail dans l'ensemble de l'industrie sauf bâtiment, génie civil et agricole conduit pour les trimestres des années 1979, 1980 et 1981 au tableau suivant.*

trimestre	*durée hebdomadaire (heures)*
1	*40,7*
2	*40,67*
3	*40,68*
4	*40,68*
5	*40,6*
6	*40,55*
7	*40,49*
8	*40,45*
9	*40,33*
10	*40,29*
11	*40,24*
12	*40,16*

1. On note X la variable "temps" (numéro du trimestre) et Y la variable "durée hebdomadaire du travail". Représenter le nuage de points.

2. On note X' la variable définie par $X' = X - 6$ et Y' la variable définie par

$$Y' = 100(Y - 40,45).$$

a) Calculer le coefficient de corrélation linéaire entre X' et Y'. En déduire le coefficient de corrélation linéaire entre X et Y.

b) Calculer l'équation de la droite de régression de Y' sur X'. En déduire l'équation de la droite de régression de Y sur X.

3. Faire une prévision de la durée hebdomadaire du travail pour le premier trimestre de l'année 1982 en utilisant la droite de régression.

Exercice 15. *À l'oral d'un examen, chaque candidat est interrogé en première et seconde langue. On note X et Y les notes obtenues dans chaque matière. Les résultats obtenus par 100 candidats sont regroupés dans le tableau ci-dessous.*

X \ Y	2	6	10	14	18
2	2	1	0	0	0
6	5	12	3	1	0
10	5	10	25	5	0
12	0	3	12	10	1
16	0	0	1	2	2

1. Écrire les séries statistiques relatives à chaque variable X et Y.

2. Calculer les moyennes et écarts-type de X et Y.

3. Calculer la covariance et le coefficient de corrélation linéaire. Représenter le nuage de points. Commenter.

Exercice 16. *On considère la variable X (temps passé chaque jour devant la télé) et la variable âge. Sur un échantillon de 25 personnes on a obtenu la table d'effectifs suivante :*

Temps Âge	*20-30 ans*	*30-40 ans*	*40-50 ans*
1 h-2 h	5	0	0
2 h-3 h	2	7	0
3 h-4 h	0	4	2
4 h-5 h	0	0	5

1. Calculer la covariance.

2. Calculer le coefficient de Bravais-Pearson.

3. Déterminer le droite de régression de la variable Âge par rapport à la variable Temps.

4. Peut-on estimer le temps passé devant la télévision d'une personne de 50 à 60 ans ?

Exercice 17. *Dans une grosse boîte de publicité, on a voulu, pour des raisons économiques, faire une étude comparative entre le temps T passé en jour de travail à tenter d'améliorer un projet déjà existant et l'évolution V en pourcentage de son impact sur les clients potentiels. On a obtenu sur 7 projets les résultats suivants au cours des 6 derniers mois :*

Temps passé : ti	12	15	10	4	20	30	8
Évolution : vi	25	27	21	10	28	12	20

1. Déterminer la covariance entre T et V.

2. Existe-t-il un lien linéaire entre le temps passé et l'évolution en pourcentage concernant ces 7 projets ?

Exercice 18. *Pour comparer les facultés de mémorisation visuelle et auditive chez les enfants de dix ans, on a demandé à 30 enfants de lire une liste de 15 mots, puis quelques minutes après, d'écrire ceux dont ils se souviennent. On a compté alors le nombre de mots mémorisés (variable X).*

On a ensuite demandé aux mêmes enfants de mémoriser les mots d'une liste en comportant 15 (de même difficulté que la liste précédente), mais la liste étant cette fois lue aux enfants. Y étant la variable "nombre de mots mémorisés ", pour ce deuxième test, voici les résultats obtenus à partir des deux expériences, après regroupement en classes :

X \ Y	$]0;5]$	$]5;10]$	$]10;15]$
$]0;5]$	3	4	0
$]5;10]$	4	5	1
$]10;15]$	0	9	4

Déterminer le coefficient de corrélation linéaire de X et Y. De la valeur obtenue de ce coefficient peut-on conclure concernant ces 30 enfants, que lorsque l'un d'entre eux mémorise moins bien qu'un des autres par le visuel, il a tendance à mémoriser également moins bien par l'audition ?

Exercice 19. *Une entreprise veut effectuer une prévision sur le volume de ses ventes pour les années 2000 et 2001. Pour cela elle étudie le volume de ses ventes pour la période de 1990 à 1999. Les résultats sont donnés dans le tableau ci-après.*

Année	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Ventes réalisées	3400	4000	3200	3700	3600	3100	3300	3500	4200	4100

On notera T la variable "Année - 1990", et V la variable "Ventes réalisées".

1. Déterminer la droite de régression de V en T.

2. En calculant le paramètre approprié, dire s'il est légitime d'utiliser le droite de régression de V en T aux fins de prévisions.

3. En cas de réponse positive à la question b), déterminer les volumes présumés de

ventes pour les années 2000 et 2001.

Exercice 20. *La conscription a permis très tôt grâce au nombre de jeunes gens concernés de faire des études statistiques très intéressantes car portant sur un échantillon important. Le tableau suivant donne une idée de l'évolution de l'illettrisme chez les conscrits au XIX-ième siècle :*

X : variable "année" (1 correspond à l'année 1801, 5 correspond à l'année 1805,...)

Y : variable "pourcentage d'illettrés parmi les conscrits".

x_i	1	5	9	13	16	21	25
y_i	53	49	47	43	40	38	36

1. Calculer la covariance puis le coefficient de corrélation linéaire de (X, Y). Expliquer pourquoi la simplification pour la variable "année" ne modifie pas ces nombres.

2. Si la corrélation linéaire est forte utiliser la droite de régression de Y en X pour prévoir le pourcentage d'illettrés parmi les conscrits en 1830.

Exercice 21. ***Répartition des sexes dans l'enseignement féminin entre 1845 et 1860.***

Extrait d'un article sur la présence des hommes dans l'enseignement secondaire féminin entre 1845 et 1860, de Rébecca ROGERS, paru dans Clio (Revue d'histoire des femmes).

- Instituteur a un féminin, professeur n'en a pas. (Déclaration d'un professeur homme en 1845) Les années 1840 voient l'apparition d'un courant réformateur et critique qui se penche sur l'éducation féminine. C'est en 1845 que la question de la présence d'hommes dans l'enseignement des jeunes filles devient une affaire publique et polémique : la femme de lettres Louise Dauriat dénonce dans un mémoire remis à la Chambre des pairs le scandale qu'introduisent les hommes dans les pensionnats féminins. " il n'est que trop vrai que parmi les maîtres, soit qu'ils enseignent les sciences, soit qu'ils enseignent les arts, il en est qui ne craignent pas d'envelopper dans leurs infâmes séductions de jeunes pensionnaires..."

Le tableau suivant donne, pour quelques années sur la période, la répartition des professeurs selon leur sexe, pour l'ensemble des pensionnats parisiens.

Années	*Professeurs femmes*	*Professeurs hommes*
1845	*341*	*898*
1846	*327*	*928*
1847	*332*	*949*
1849	*431*	*758*
1850	*418*	*688*
1852	*431*	*675*
1854	*516*	*679*

On définit la variable X comme étant le nombre de femmes professeurs et la variable Y comme étant le nombre d'hommes professeurs.

1. *Calculer la moyenne, la variance et l'écart-type de la variable X.*
2. *On donne les informations suivantes :*

La moyenne de Y est de 796,4 avec un écart-type de 115,3.

Calculer la covariance et le coefficient de corrélation linéaire entre les variables X et Y. Que peut-on en conclure ?

3. *Déterminer l'équation de la droite de régression linéaire de Y en X.*

Exercice 22. *Au cours de la décennie 1970-1980, les effectifs employés au fond dans les houillères françaises et la production nette de charbon ont évolué de la façon suivante :*

Année	*Effectifs du fond (milliers de personnes)*	*Production nette de charbon (millions de tonnes)*
1970	*71,3*	*40,1*
1971	*65,3*	*35,8*
1972	*57,6*	*32,7*
1973	*50,4*	*28,4*
1974	*47,1*	*25,7*
1975	*45,8*	*25,6*
1976	*42,2*	*25,1*
1977	*38,6*	*24,4*
1978	*35,9*	*22,4*
1979	*32,7*	*21,1*
1980	*30,8*	*20,7*

Étudier la liaison linéaire entre les effectifs employés au fond et la production de charbon.

Exercice 23. *Voici un tableau donnant la natalité en France au XIX-ième siècle (taux pour 1000).*

Année	*1800*	*1810*	*1820*	*1830*	*1840*	*1850*	*1860*	*1870*	*1880*	*1890*	*1900*	*1910*
Taux pour 1000	*32,9*	*31,6*	*31,7*	*30,3*	*28,5*	*27,1*	*26,9*	*23,1*	*25,2*	*22,9*	*22,4*	*19*

Avant d'étudier la corrélation linéaire, nous allons simplifier la variable année de la façon suivante :

$$1800 \leftarrow 1, \quad 1810 \leftarrow 2, \quad \ldots \quad , 1910 \leftarrow 12$$

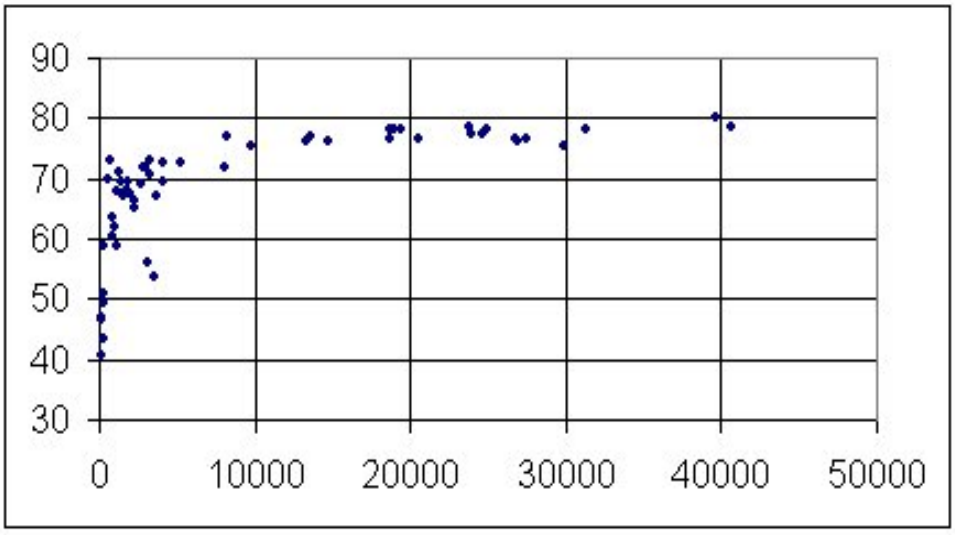
1. Avant de faire les calculs, expliquer quelles vont être les conséquences de cette simplification (de ce changement de variable) sur la covariance, le coefficient de corrélation linéaire et l'équation de la droite de régression.

2. Faire les calculs puis utiliser la droite de régression pour faire des prévisions pour l'an 2000. Qu'en pensez vous ?

Exercice 24. *Nous avons dit en cours que la corrélation linéaire n'est pas le seul type de lien qui peut exister entre deux variables. Nous allons voir ici un exemple de corrélation log-linéaire, qui fait intervenir nos connaissances sur la corrélation linéaire mais aussi la fonction "logarithme népérien" que nous avons rencontrée au chapitre 1. On souhaite mettre en évidence un lien éventuel entre le PNB par habitant et l'espérance de vie, le tableau suivant fournit les données pour l'année 1997 dans 56 pays répartis sur les 5 continents.*

PAYS	PNB/habitants (en $)	Espérance de vie à la naissance (années)	PAYS	PNB/habitants (en $)	Espérance de vie à la naissance (années)
Afghanistan	276	43,5	Iran	1957	67,5
Afrique du sud	3160	56	Irlande	14710	76
Albanie	670	73	Italie	19020	78
Algérie	1600	67	Japon	39640	80
Allemagne	27510	76,5	Maroc	1110	68
Argentine	8030	72	Mexique	3320	73
Australie	18720	78	Norvège	31250	78
Belgique	24710	77,5	Nouvelle-Zélande	13340	76
Bolivie	800	60,5	Pays-Bas	24000	77,5
Brésil	3640	67	Pérou	2310	66,5
Bulgarie	1330	71	Pologne	2790	72
Cambodge	270	49,5	Portugal	9740	75,5
Canada	19380	78	Roumanie	1480	69,5
Chili	4160	72,5	Royaume-Uni	18700	76,5
Chine	620	70	Russie	2240	65
Colombie	1910	69,5	Rwanda	180	40,5
Croatie	3250	70,5	Singapour	26730	76,5
Danemark	29890	75,5	Soudan	227	51
Egypte	790	63,5	Suède	23750	78,5
Espagne	13580	77	Suisse	40630	78,5
Etats Unis	26980	76	Tchad	180	46,5
Ethiopie	100	47	Thaïlande	2740	69
Finlande	20580	76,5	Tunisie	1820	68
France	24990	78	Ukraine	1630	68
Gabon	3490	53,5	Uruguay	5170	72,5
Grèce	8210	77	Vénézuela	3020	72
Hongrie	4120	69,5			
Inde	340	59			
Indonésie	980	62			
Irak	1165	59			

Notons X la variable PNB/h et Y la variable Espérance de vie à la naissance.
Le tableau confirme déjà que globalement, plus le PNB/h est fort plus l'espérance de vie est élevée. Le nuage de point associé au couple (X, Y) va permettre d'affiner cette première observation.



La forme du nuage de point rappelle la courbe de la fonction logarithme népérien, nous allons donc étudier la corrélation linéaire entre $\ln X$ *et* Y *et trouver ainsi une courbe dont l'équation est de la forme*

$$y = a \ln x + b$$

qui passe aussi près que possible de l'ensemble des points.

Si l'on procède comme dans les exercices précédents, on trouve $r(X, Y) = 0,6535$*, ce qui indique une faible corrélation linéaire positive.*

1. *Montrer que* $r(\ln X, Y) = 0,862$*. Commenter ce résultat.*
2. *Montrer que l'équation de la courbe cherchée est*

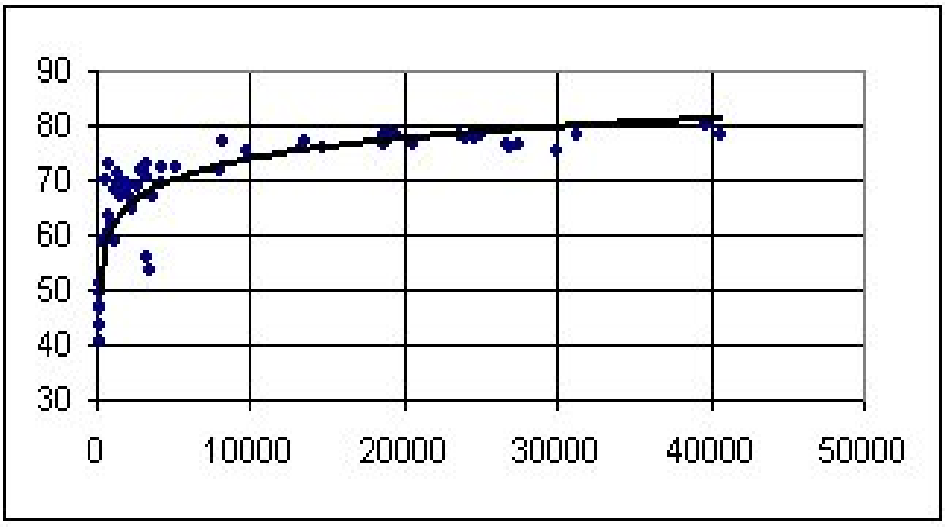
$$y = 5,27 \ln x + 25,43.$$

Pour réduire les calculs, voici quelques résultats partiels :

$$\sum y_i = 3845,5, \sum y_i^2 = 269543,25;$$

$$\sum \ln x_i = 459,77, \sum (\ln x_i)^2 = 3921,37, \sum y_i \ln x_i = 32344,7$$

Voici la courbe d'équation $y = 5,27 \ln x + 25,43$ *traversant le nuage de points.*



7.4 Exercices du chapitre 3

Exercice 25. *On a relevé le nombre de mariages dans une ville du sud-ouest de la France chaque trimestre pendant 3 ans :*

Année / Trimestre	1998	1999	2000
1	10	11	12
2	12	14	15
3	13	15	17
4	11	12	12

On notera Y la variable dont on étudie l'évolution.

1. Représenter graphiquement cette série chronologique (avec périodes superposées puis avec périodes successives). Commenter.

2. Calculer la série des moyennes mobiles, lisser la courbe.

3. Calculer l'équation de la droite de tendance et tracer cette droite sur le graphique précédent.

4. Calculer les quatre coefficients saisonniers (pour le modèle additif).

5. Utiliser le modèle construit pour prévoir le nombre de mariages dans cette ville en 2002.

Exercice 26. *Le tableau ci-dessous donne les indices trimestriels de la production industrielle (base 100 en 1980) en ce qui concerne les produits de la parachimie et de la pharmacie :*

Année / Trimestre	1982	1983	1984
1	106	106,1	109,6
2	108,8	107,1	108,5
3	97,9	98,5	103
4	108,5	111,2	115,5

Source : INSEE - Division indicateurs conjoncturels d'activité.

Faire une étude complète de cette série chronologique (reprendre les 4 premières questions de l'exercice 1). On note toujours Y la variable dont on étudie l'évolution et afin de réduire les calculs, voici un tableau (à compléter) qui fournit quelques résultats :

t	$y(t)$	$z(t)$	$t * z(t)$	$x(t) = 0,49t + 103,04$	$y(t) - x(t)$	$s'(t)$	$s(t)$
1	106			103,53	2,47	1,743	
2	108,8			104,02	4,78	2,153	
3	97,9	105,31	315,94	104,51	-6,61		
4	108,5	105,11	420,45	105	3,5		
5	106,1	104,98	524,88	105,49	0,61		
6	107,1	105,39	632,33	105,98	1,12		
7	98,5	106,16	743,14	106,47	-7,97		
8	111,2	106,78	854,20	106,96	4,24		
9	109,6						
10	108,5						
11	103						
12	115,5						

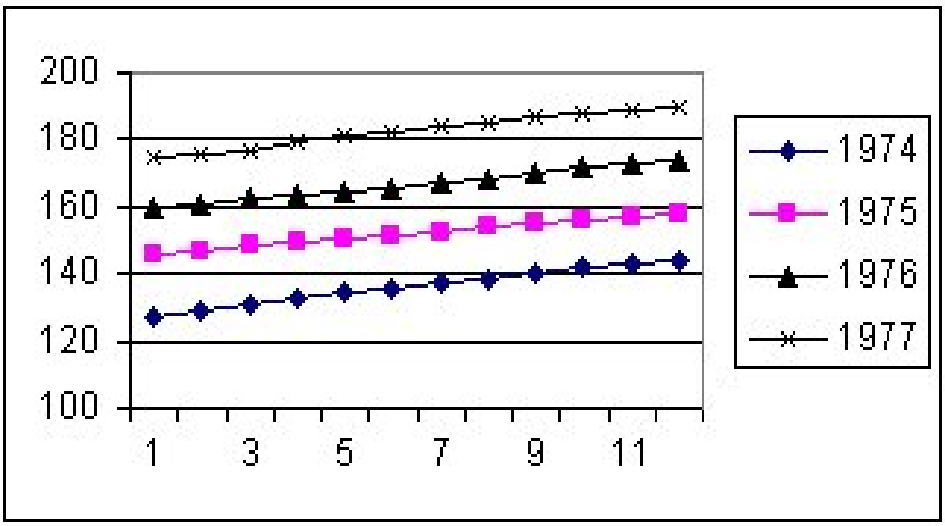
Quel est l'indice prévisible pour le 3-ième trimestre 1985 ?

Exercice 27. *On se propose d'étudier l'évolution de l'indice mensuel des prix à la consommation de 1974 à 1977. Les observations mensuelles sont consignées dans le tableau ci-dessous.*

	1974	1975	1976	1977
Janvier	127,4	145,9	159,9	174,3
Février	129,1	147	161	175,5
Mars	130,6	148,2	162,4	177,1
Avril	132,7	149,5	163,8	179,4
Mai	134,3	150,6	164,9	181,1
Juin	135,8	151,7	165,6	182,5
Juillet	137,5	152,8	167,2	184,1
Août	138,6	153,8	168,4	185,1
Septembre	140,1	155,1	170,2	186,7
Octobre	141,8	156,3	171,8	188,2
Novembre	143,1	157,3	173,2	188,9
Décembre	144,3	158,2	173,8	189,4

1. Calculer la première et la dernière moyenne mobile.

2. Voici la représentation graphique de cette série chronologique avec périodes superposées :



Est-il intéressant ici de calculer la série des moyennes mobiles ?

3. Qu'apporterait le calcul des variations saisonnières ?

4. Quelle méthode vous paraît la mieux adaptée pour faire des prévisions ?

Exercice 28. *On étudie la fréquentation des hôtels américains de Disneyland-Paris entre 1995 et 1997.*
Le nombre de clients (en milliers) est donné dans le tableau ci-dessous :

Année	*Trimestre 1*	*Trimestre 2*	*Trimestre 3*	*Trimestre 4*
1995	*8,2*	*12,3*	*32,7*	*8,3*
1996	*10*	*14,5*	*37,3*	*11,5*
1997	*11*	*17,3*	*41,3*	*13,3*

Source : Disneyland-Paris, Marne la Vallée, 1998.

1. Compléter la série des moyennes mobiles dont les 6 dernières valeurs sont : 16,95 ; 17,925 ; 18,45 ; 18,925 ; 19,775 ; 20,5. Représenter sur un même graphique la série des données et la série des moyennes mobiles. Commenter.

2. Déterminer l'équation de la droite de tendance et la tracer sur le même graphique.

3. On donne la somme des composantes saisonnières : $S_0 = 0,567$. Calculer les coefficients saisonniers du 1-er et du 3-ième trimestre.

4. En utilisant le modèle que l'on a construit, quelle fréquentation peut on prévoir pour le troisième trimestre 2000 ?

Exercice 29. *On étudie l'évolution du nombre de billets vendus (en milliers) dans un complexe cinématographique lors des trois premières années :*

	Janv-Fév	*Mars-Avril*	*Mai-Juin*	*Juillet-Août*	*Sep-Oct*	*Nov-Déc*
1997	*100*	*82*	*70*	*40*	*62*	*91*
1998	*105*	*94*	*73*	*43*	*72*	*106*
1999	*111*	*99*	*84*	*52*	*77*	*118*

1. Représenter graphiquement ces données dans un repère cartésien.

2. Calculer la série des moyennes mobiles.

3. Montrer que l'équation de la droite de tendance est : $x(t) = 1,34t + 68,93$.
Tracer cette droite sur le graphique précédent.

4. On donne la somme des composantes saisonnières : $S_0 = 3,04$. Calculer le coefficient saisonnier pour Juillet-Août. Utiliser ce coefficient pour faire des prévisions pour Juillet-Août 2000.

Exercice 30. *Voici les résultats du chiffre d'affaire d'un magasin d'insecticides d'Egletons (département de la Corrèze) au cours des douze derniers trimestres (en milliers de francs) :*

	1998	*1999*	*2000*
1er trimestre	*27*	*29*	*30*
2ème trimestre	*10*	*12*	*13*
3ème trimestre	*15*	*16*	*17*
4ème trimestre	*26*	*27*	*28*

1. Sur le graphique de la page suivante, faire une représentation cartésienne avec périodes successives de la série chronologique précédente. Quelles conjectures peut-on tirer de ce graphique ?

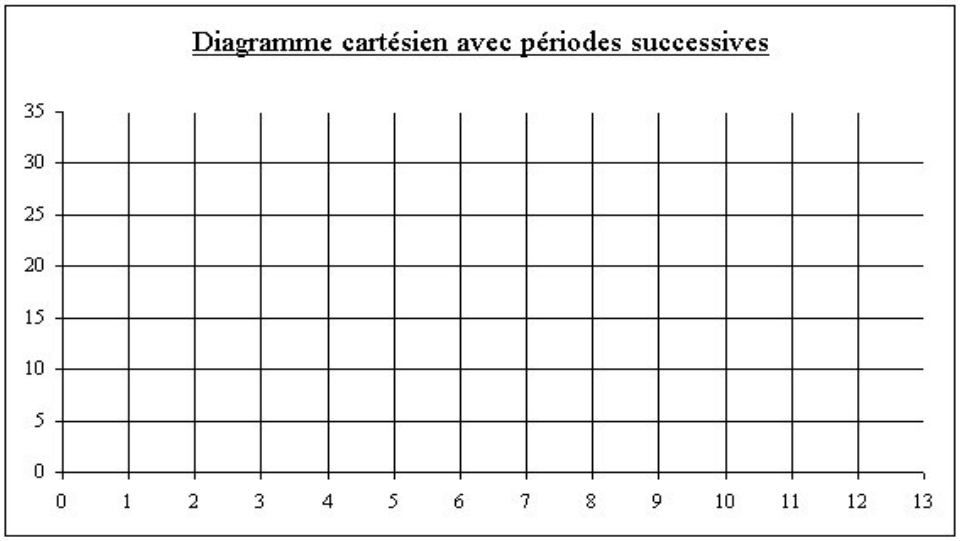


FIGURE 7.3 – Diagramme cartésien pour les insecticides

2. Afin de dégager la tendance, le propriétaire du magasin décide d'utiliser la méthode des moyennes mobiles ; compléter le tableau avec les moyennes mobiles, $z(t)$, non calculées.

3. Pour avancer les calculs, on donne $\overline{Z} = 20,938$. Montrer que l'équation de la droite de tendance est $x(t) = 0,289t + 19,061$. Tracer sur le graphique précédent cette droite.

4. Compléter sur le tableau les différents calculs relatifs aux coefficients saisonniers ($\delta(t)$, $s'(t)$ et $s(t)$).

5. Tracer sur le graphique les estimations (les valeurs de $\hat{y}(t)$) que donne le modèle pour l'année 2000. Au vu des résultats obtenus, le modèle vous paraît-il pertinent ?

6. Déduire des calculs précédents, la prévision du chiffre d'affaire du magasin pour le troisième trimestre 2001.

t	$z(t)$	$\delta(t)$	$s'(t)$	$s(t)$
1		7,650		
2		-9,639		
3	19,750		-5,084	-4,978
4			5,627	5,733
5	20,625	8,494		
6	20,875	-8,795		
7	21,125	-5,084	-5,084	-4,978
8	21,375	5,627	5,627	5,733
9	21,625	8,338		
10		-8,951		
11		-5,240	-5,084	-4,978
12		5,471	5,627	5,733

Exercice 31. *Le tableau suivant donne l'évolution de l'indice du coût de la construction (base 100 au 4ème trimestre 1953). Cet indice est utilisé par exemple pour la révision annuelle des loyers. On notera Y la variable "indice du coût de la construction" et T le temps.*

Trimestres \ Années	1998	1999	2000
1	1058	1071	1083
2	1058	1074	1089
3	1057	1080	1093
4	1074	1065	1127

Source : INSEE.

On pourra utiliser le tableau et le graphique ci-dessous.

1. Représenter graphiquement cette série chronologique avec périodes successives. Peut-on parler ici de variations saisonnières ?

2. Compléter la série des moyennes mobiles (expliciter le calcul des $z(t)$) et représenter cette série sur le graphique ci-dessus. Quel est l'intérêt des moyennes mobiles ? Peut-on utiliser le modèle additif pour décrire cette série chronologique ?

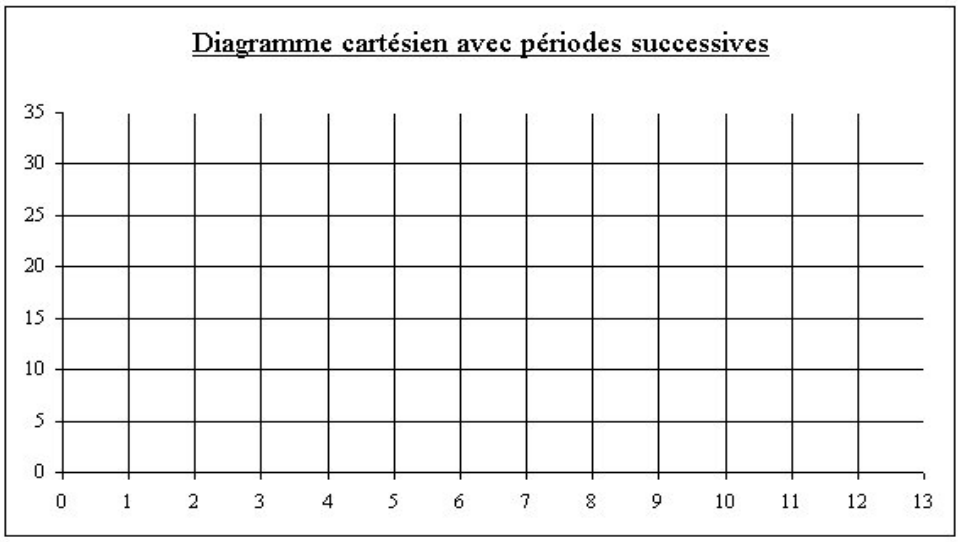


FIGURE 7.4 – Diagramme cartésien pour la construction

t	$y(t)$	$z(t)$
1	*1058*	
2	*1058*	
3	*1057*	*1063,375*
4	*1074*	*1067*
5	*1071*	*1071,875*
6	*1074*	*1073,625*
7	*1080*	*1074*
8	*1065*	
9	*1083*	
10	*1089*	
11	*1093*	
12	*1127*	

Chapitre 8

Corrigé des exercices

8.1 Correction des exercices du Chapitre 1

Exercice 1.

$$5 \times \left(1 + \frac{25}{100}\right) \times \left(1 - \frac{25}{100}\right) = 4,6875F$$

Remarque *:* $1,25 \times 0,75 = 0,9375$*, il y a donc une baisse de 6,25%.*

Lorsqu'il y a une hausse de a% suivie d'une baisse de a%, le multiplicateur global est :

$$\left(1 + \frac{a}{100}\right)\left(1 - \frac{a}{100}\right) = 1 - \left(\frac{a}{100}\right)^2 \leq 1,$$

il y a donc globalement une baisse.

Si la baisse a lieu avant la hausse, la conclusion est la même.

Exercice 2. *Soit* F *le salaire moyen des femmes et* H *celui des hommes :*

$$F = 0,67 \times H \Longleftrightarrow H = 1, \times F.$$

Les deux phrases sont équivalentes.

Exercice 3. *1.* Ω *: population française en 1920, 1946, 1970, 1977 et 1987.*
On peut aussi considérer qu'il y a 5 populations : Ω_1*=population française en 1920, ...,* Ω *=population française en 1987. On étudie la variable " sexe " dont les modalités sont hommes et femmes, cette variable est nominale.*

2. Le multiplicateur est : $\dfrac{27,7}{20,2} = 1,3713$ *et 1,3713 - 1 = 0,3713 = 37,13%.*

Le taux de croissance de la population française féminine entre 1920 et 1977 est $\boxed{37,13\%}$.

3. Le multiplicateur global est : $\dfrac{55}{39} = 1,410$ *et* $1,41^{\frac{1}{67}} = 1,0051$ *et* $1,0051 - 1 = 0,0051 = 0,51\%$.

Le taux annuel moyen de croissance de la population française entre 1920 et 1987 est $\boxed{0,51\%}$.

Exercice 4. *1. Hommes :* $\dfrac{21841}{16224} - 1 = 34,62\%$. *Femmes :* $\dfrac{13133}{10812} - 1 = 21,47\%$.
2. Région parisienne : $\left(\dfrac{28916}{21729}\right)^{1/3} - 1 = 9,99\%$. *Midi-Pyrénées :* $\left(\dfrac{18409}{13680}\right)^{1/3} - 1 = 10,40\%$.
3. $\dfrac{28916}{18686} - 1 = 54,75\%$ *et* $\dfrac{18409}{12392} - 1 = 48,56\%$.
4. Notons f la fréquence de femmes :

$$16224(1-f) + 10812f = 14669 \iff f = 0,2873 = 28,73\%.$$

Exercice 5. *1. Taux de variation du nombre total d'étudiants entre 1900 et 1939 :* $\dfrac{78972}{29377} = 2,6882$. *Soit une augmentation de* $\boxed{168,82\%}$.

Taux de variation du nombre d'étudiantes entre 1900 et 1939 :

$$29377 \times \frac{3,5}{100} = 1028,195\ ; \quad 78972 \times \frac{30}{100} = 23691,6\ ; \quad \frac{23691,6}{1028,195} = 22,0419.$$

Soit une augmentation de $\boxed{2204,19\%}$ *(Dans un tel contexte, le multiplicateur est plus parlant que le taux de variation).*

2. $69961 \times \dfrac{22,9}{100} = 16021,069\ ; \quad \dfrac{23691,6}{16021,069} = 1,4788\ ; \quad 1,4788^{\frac{1}{10}} = 1,0399.$
Soit un taux annuel moyen de croissance de $\boxed{3,99\%}$.

Exercice 6. *1. a)* $\dfrac{65}{36} = 1,8056$. *Soit un taux d'augmentation de* $\boxed{80,56\%}$.
b) $1911 - 1861 = 50$ *;* $1,8056^{\frac{1}{50}} = 1,0119$. *Taux annuel moyen de croissance :* $\boxed{1,19\%}$.

2. a) $23500 \times \left(1 + \dfrac{343}{100}\right) \times \left(1 + \dfrac{171}{100}\right) = 282124,55$. *Il y a* $\boxed{282124,55\ km}$ *de voies ferrées en Europe en 1900.*

b) $4 \times (1870 - 1850) = 80$; $\left(1 + \frac{343}{100}\right)^{\frac{1}{80}} = 1,0188.$

Le taux trimestriel moyen d'augmentation est $\boxed{1,88\%}$.

Exercice 7. *1.*

	1870	*1880*	*1890*	*1900*	*1910*	*1920*
Primaire	*7097*	*8980*	*10568*	*11816*	*12799*	*12809*
Secondaire	*2643*	*3842*	*5526*	*7199*	*10657*	*12861*
Tertiaire	*3185*	*4571*	*7225*	*10058*	*13916*	*16763*
	12925	*17393*	*23319*	*29073*	*37372*	*42433*

	1870	*1880*	*1890*	*1900*	*1910*	*1920*
Primaire	*54,9%*	*51,6%*	*45,3%*	*40,6%*	*34,2%*	*30,2%*
Secondaire	*20,4%*	*22,1%*	*23,7%*	*24,8%*	*28,5%*	*30,3%*
Tertiaire	*24,6%*	*26,3%*	*31,0%*	*34,6%*	*37,2%*	*39,5%*

2. et 3.

	Taux de variation entre 1870 et 1920	*Tx de variation moyen sur 10 ans entre 1870 et 1920*
Primaire	*-45,0%*	*-11,3%*
Secondaire	*48,2%*	*8,2%*
Tertiaire	*60,3%*	*9,9%*

Exercice 8.

Charbon :

Distance entre les graduations 0,1 et 1 : 2 cm.

Distance entre les graduations 1 et 10 : 2,3 cm.

Distance entre les graduations 10 et 100 : 1,6 cm.

Il apparaît donc que ce repère probablement gradué à la main n'est pas tout à fait correct.

Fonte :

Le module est 6,7 cm et la hauteur correspondant à un doublement est $\boxed{2,02\ cm}$. *Cette distance se retrouve simplement de la façon suivante :* $log(10) - log(1) = 1$ *est représenté par 6,7cm,* $log(2) - log(1) = 0,30$ *donc on fait le produit en croix* $6,7 * 0,30/1 = 2,02$ *pour savoir à quelle distance correspond un doublement.*

Acier :

Le module est 6,1 cm et la hauteur correspondant à un doublement est $\boxed{1,84\ cm}$.

Allemagne : sur toute la période de 1880 à 1910 le taux de variation est relativement stable et le multiplicateur global est 9 (on passe de 1,5 à 13,5). $9^{1/30} = 1,0760$ *soit un taux annuel moyen de* $\boxed{7,60\%}$.

Pour les 3 autres pays, on distingue nettement les périodes 1880/1890 et 1890/1910.

États-Unis :

- *Période 1880/1890 : hauteur = 3,3 cm, notons m le multiplicateur global, on a :*

$$6,1 \times \log m = 3,3 \iff \log m = \frac{3,3}{6,1} \iff m = 10^{3,3/6,1} = 3,4752;$$

$3,4752^{1/10} = 1,1327$. *Soit un taux annuel moyen d'augmentation de* $\boxed{13,27\%}$.

- *Période 1890/1910 : hauteur = 5 cm, notons m le multiplicateur global, on a :*

$$\log m = \frac{5}{6,1} \iff m = 10^{5/6,1} = 6,6019;$$

$6,6019^{1/20} = 1,099$. *Soit un taux annuel moyen d'augmentation de* $\boxed{9,90\ \%}$.

Grande-Bretagne :

- *Période 1880/1890 : hauteur = 0,9 cm, taux annuel moyen d'augmentation de* $\boxed{3,46\%}$.
- *Période 1890/1910 : hauteur = 0,5 cm, taux annuel moyen d'augmentation de* $\boxed{0,95\%}$.

France :

- *Période 1880/1890 : hauteur = 2,1 cm, il s'agit ici d'une baisse de la production :*

$$6,1 \times \log m = -2,1 \iff \log m = \frac{-2,1}{6,1} \iff m = 10^{-2,1/6,1} = 0,4526;$$

$0,4526^{1/10} = 0,9238$. *Soit un taux annuel moyen de variation de* $\boxed{-7,62\%}$.

- *Période 1890/1910 : hauteur = 4,6 cm , taux annuel moyen d'augmentation de* $\boxed{9,28\%}$.

***Remarque** : on peut aussi trouver ces taux de variation en lisant les valeurs effectives des productions et en se ramenant ainsi au même type d'exercices que les précédents, la méthode utilisée ici permet de voir que la hauteur sur le graphique détermine le taux de variation.*

Exercice 9. *1. On sait que si l'on obtient des courbes ou des droites "sensiblement parallèles" sur un diagramme semi-logarithmique, cela signifie que les évolutions dans le temps des phénomènes sont "relativement identiques". C'est dans ce but de comparaison que le diagramme semi-log paraît bien adapté.*

2. La valeur la plus basse 8 nous impose de choisir l'échelle du premier module de 1 à 10. La valeur la plus haute ne dépassant évidemment pas 100%, nous travaillerons sur un papier à deux modules, le second module étant gradué de 10 à 100.

3. cf. figure. Les évolutions sont sensiblement parallèles. Les Gersois ont donc adopté des comportements de départ en vacances d'hiver identiques : quand le taux de départ baisse pour une des catégories, il baisse aussi dans la même proportion pour les autres et réciproquement.

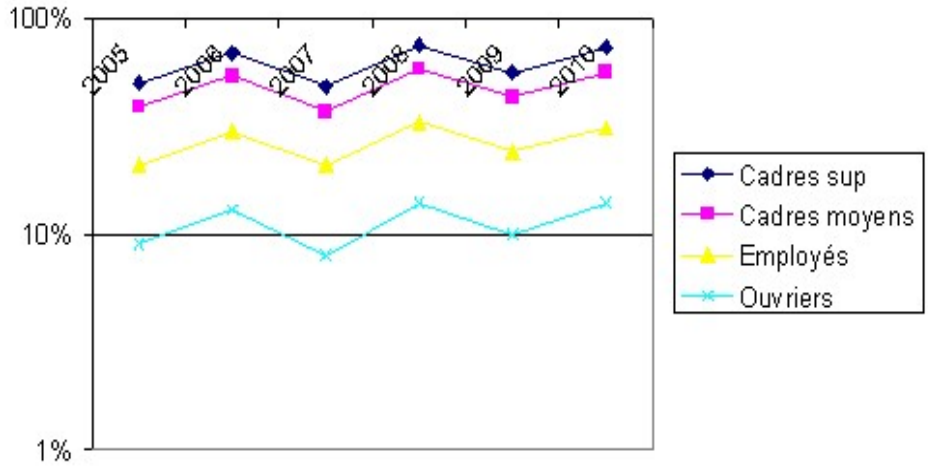


FIGURE 8.1 – *Graphique semi-logarithmique des départs en vacances d'hiver.*

Un graphique en coordonnées arithmétiques n'aurait pas donné le même message visuel comme on peut en juger ci-dessous surtout en comparant la courbe la plus élevée et celle la moins élevée.

8.2 Correction des exercices du Chapitre 2

Exercice 10. *1. Cf. Figure.*

2. En notant X la variable consommation de graisse et Y la variable taux de mortalité, on a :

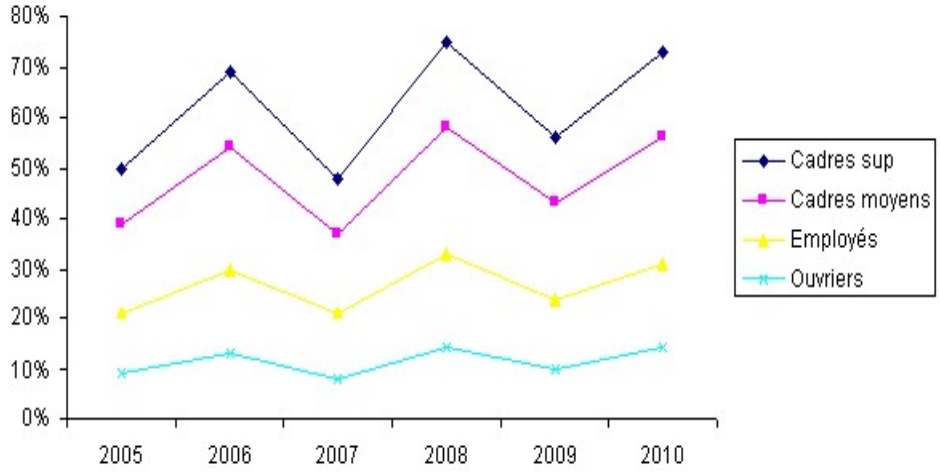


FIGURE 8.2 – *Graphique arithmétique des départs en vacances d'hiver.*

x_i	y_i	x_i^2	y_i^2	$x_i \times y_i$
14,4	*29,1*	*207,36*	*846,81*	*419,04*
16	*29,7*	*256*	*882,09*	*475,2*
11,6	*29,2*	*134,56*	*852,64*	*338,72*
11	*26*	*121*	*676*	*286*
10	*24*	*100*	*576*	*240*
9,6	*23,1*	*92,16*	*533,61*	*221,76*
9,2	*23*	*84,64*	*529*	*211,6*
10,4	*23,1*	*108,16*	*533,61*	*240,24*
11,4	*25,2*	*129,96*	*635,04*	*287,28*
12,5	*26,1*	*156,25*	*681,21*	*326,25*
116,1	*258,5*	*1390,09*	*6746,01*	*3046,09*

On en déduit $\overline{X} = 11,61$, $\sigma_X^2 = 4,22$, $\sigma_X = 2,05$, $\overline{Y} = 25,85$, $\sigma_Y^2 = 6,38$ *et* $\sigma_Y = 2,53$.

3. Calcul de la covariance :

$$\mathrm{Cov}(X,Y) \;=\; \frac{3046,09}{10} - (11,61 \times 25,85) \;=\; 4,4905.$$

4. Calcul du coefficient de corrélation linéaire :

$$r(X,Y) \;=\; \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \;=\; \frac{4,4905}{2,05 \times 2,53} \;=\; 0,87.$$

5. La forme du nuage de points "étirée" (les points sont proches d'une droite) et la valeur du coefficient de corrélation linéaire proche de 1 rendent plausible l'existence d'une liaison linéaire forte entre X *et* Y.

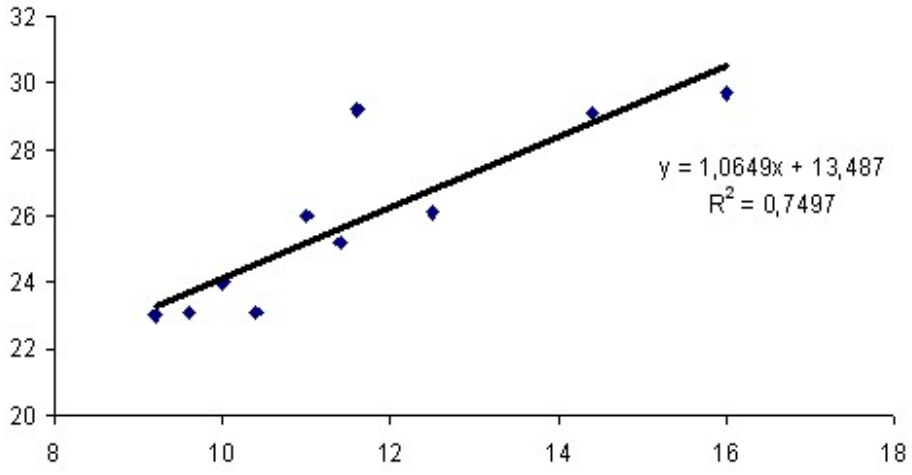


FIGURE 8.3 – *Nuage de points et droite de régression pour les variables graisse/taux de mortalité.*

6. La pente a de la droite de régression de Y sur X est donnée par :

$$a \;=\; \frac{\mathrm{Cov}(X,Y)}{\sigma_X^2} \;=\; 1,06,$$

et la valeur à l'origine b est donnée par :

$$b \;=\; \overline{Y} - a\overline{X} \;=\; 13,49.$$

L'équation de la droite de régression de Y sur X est donc :

$$y \;=\; 1,06x + 13,49.$$

7. Une estimation du taux de mortalité par athérosclérose en 1948 est :

$$\widehat{y}(1948) \;=\; 1,06 \times 13 + 13,49 \;=\; 27,27.$$

Exercice 11. *1.*

2. Après calculs, on obtient : $r(X,Y) \;=\; -0,55$.

3. La valeur (moyenne) du coefficient de corrélation linéaire et surtout la forme du nuage de points (en forme de "T") suggère qu'il n'y a pas de liaison linéaire entre X et Y mais qu'il y a éventuellement deux populations pour lesquelles la liaison linéaire entre X et Y est forte.

4. Considérons un premier sous-échantillon constitué par les points 1, 2, 4 et 5. On obtient le tableau :

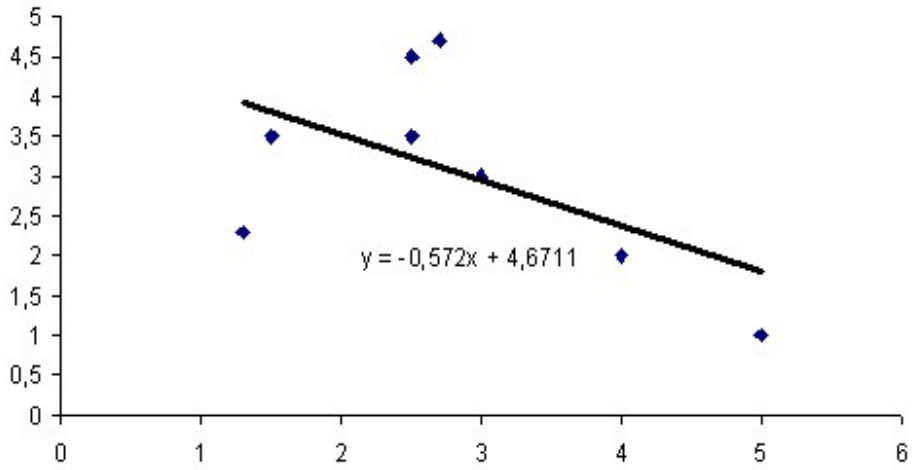


FIGURE 8.4 – *Nuage de points et droite de régression pour les variables X et Y.*

X	1,3	1,5	2,5	2,7
Y	2,3	3,5	4,5	4,7

Le coefficient de corrélation linéaire pour ce sous-échantillon est :

$$r_1(X,Y) \;=\; 0,94.$$

Considérons un second sous-échantillon formé par les points 3, 6, 7 et 8. On obtient le tableau :

X	2,5	3	4	5
Y	3,5	3	2	1

Le coefficient de corrélation linéaire pour ce sous-échantillon est :

$$r_2(X,Y) \;=\; -1.$$

On constate donc que les deux coefficients de corrélation linéaire sur chaque sous-échantillon sont proches de 1 en valeur absolue (égal à -1 pour le second), ce qui confirme l'existence d'une liaison linéaire forte sur chaque sous-échantillon.

Exercice 12. *1.*

2. On a $\overline{X} = 11,5$, $\sigma_X = 5,06$, $\overline{Y} = 8925$ *et* $\sigma_Y = 1555$.

3. La covariance entre X et Y est :

$$\mathrm{Cov}(X,Y) \;=\; \frac{1314000}{12} - (11,5 \times 8925) \;=\; 6862,5,$$

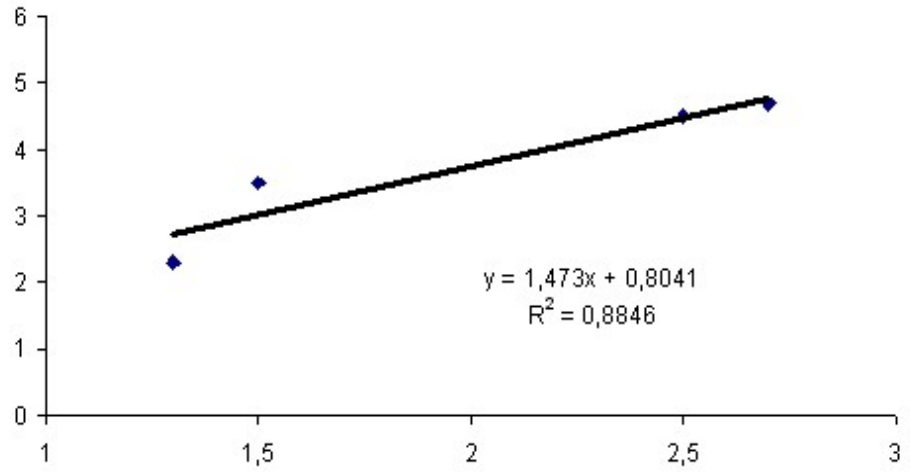


FIGURE 8.5 – *Nuage de points et droite de régression pour le premier sous-échantillon.*

et le coefficient de corrélation linéaire vaut :

$$r(X,Y) \;=\; \frac{6862,5}{5,06 \times 1555} \;=\; 0,87.$$

Le coefficient de corrélation est proche de 1 et le nuage de points est "étiré" : on peut donc faire l'hypothèse d'une liaison linéaire forte entre l'ancienneté et le salaire.

4. On a $a \;=\; 268$, *et* $b \;=\; 5840$.
L'équation de la droite de régression de Y *sur* X *est donc :* $y \;=\; 268x + 5840$.

5. Le tableau des valeurs observées, des valeurs ajustées et des résidus est le suivant :

y_i	*8100*	*10200*	*8400*	*11400*	*6900*	*9600*	*6300*	*10500*	*10800*	*8100*	*9300*	*7500*
$\widehat{y}_i$	*7718*	*9864*	*9864*	*10132*	*7181*	*9059*	*6377*	*11205*	*9596*	*8254*	*9864*	*7986*
$y_i - \widehat{y}_i$	*382*	*336*	*-1464*	*1268*	*-281*	*541*	*-77*	*-705*	*1204*	*-154*	*-564*	*-486*

6. La moyenne des résidus est nulle.

7. Une estimation du salaire d'un ouvrier ayant 4 ans d'ancienneté est :

$$\widehat{Y}_4 \;=\; 268 \times 4 + 5840 \;=\; 6912F,$$

et une estimation du salaire d'un ouvrier ayant 18 ans d'ancienneté est :

$$\widehat{Y}_{18} \;=\; 268 \times 18 + 5840 \;=\; 10664F.$$

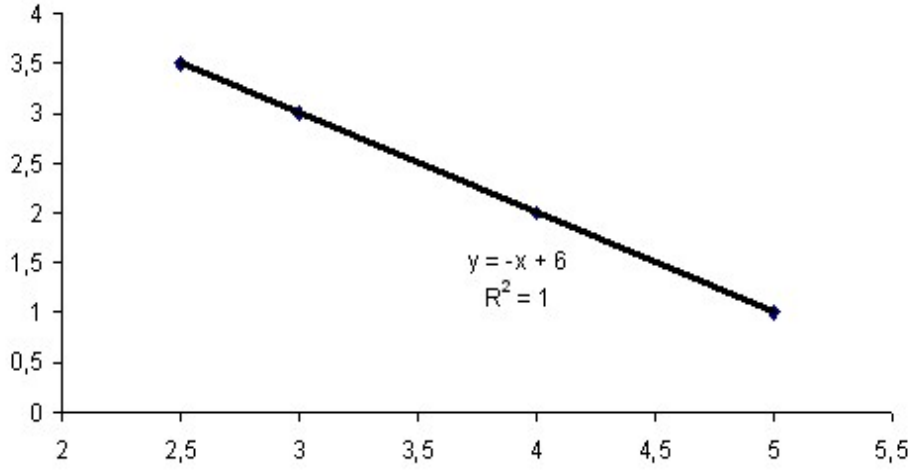


FIGURE 8.6 – *Nuage de points et droite de régression pour le premier sous-échantillon.*

Exercice 13. *On doit ici calculer le coefficient de Bravais-Pearson sur un échantillon de 8 vins, entre la variable "note sur l'étiquette" notée X et la variable "note sur le goût" notée Y.*

x_i	y_i	n_i	$n_i x_i$	$n_i (x_i)^2$	$n_i y_i$	$n_i (y_i^2)$	$n_i x_i y_i$
10	3	1	10	100	3	9	30
5	8	1	5	25	8	64	40
7	7	1	7	49	7	49	49
7	9	1	7	49	9	81	63
7	6	1	7	49	6	36	42
9	4	1	9	81	4	16	36
9	2	1	9	81	2	4	18
8	5	1	8	64	5	25	40
		8	62	498	44	284	318

moyennes : $\overline{X} = \dfrac{62}{8} = 7,75$ *;* $\overline{Y} = \dfrac{44}{8} = 5,5$

variances : $\mathrm{Var}(X) = \dfrac{498}{8} - (7,75)^2 = 2,19$ *;* $\mathrm{Var}(Y) = \dfrac{284}{8} - (5,5)^2 = 5,25$

écarts-type : $\sigma_X = 1,48$ *;* $\sigma_Y = 2,29$ *;*

covariance : $\mathrm{Cov}(X,Y) = \dfrac{318}{8} - 7,75 \times 5,5 = -2,88$ *;*

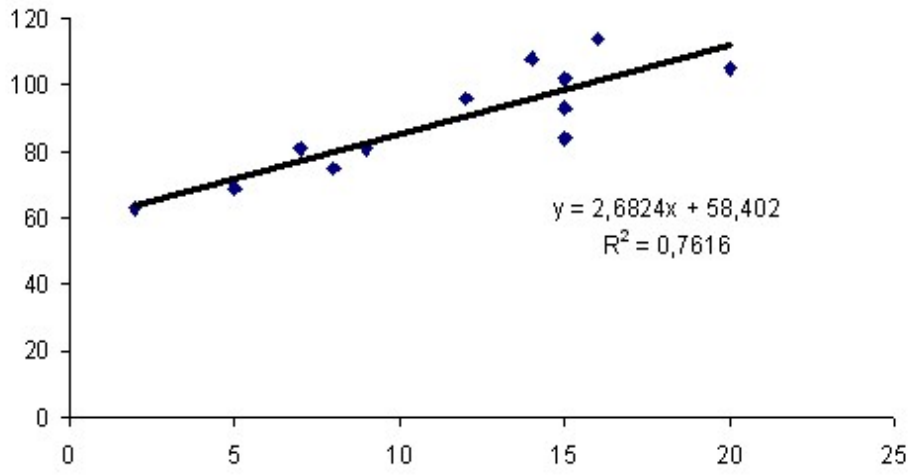


FIGURE 8.7 – *Nuage de points et droite de régression pour les variables ancienneté et salaire moyen.*

$$\textit{corrélation : } r(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y} = \frac{-2,88}{1,48 \times 2,29} = -0,85.$$

$r(X,Y)$ étant proche de -1, on a une forte corrélation linéaire décroissante entre les deux notations pour ces 8 vins.

Le coefficient de corrélation linéaire étant proche de 1, on peut tracer le nuage de points qui est "étiré" ainsi que la droite de régression.

3. Les coefficients de la droite de régression de Y sur X sont $a = -1,31$ et $b = 15,7$.

Exercice 14. *1.*

2.

x'_i	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
y'_i	25	22	23	23	15	10	4	0	-12	-16	-21	-29

a) On a $r(X',Y') = -0,97$. On en déduit que $r(X,Y) = -0,97$ puisque les coefficients des transformations affines de X en X' et de Y en Y' sont de même signe (1 et 100).

b) Les coefficients de la droite de régression de Y' sur X' sont $a' = -5,17$ et $b' = 6,25$. On en déduit les coefficients de la droite de régression de Y sur X :

$$a = \frac{a'}{100} = -0,0517,$$

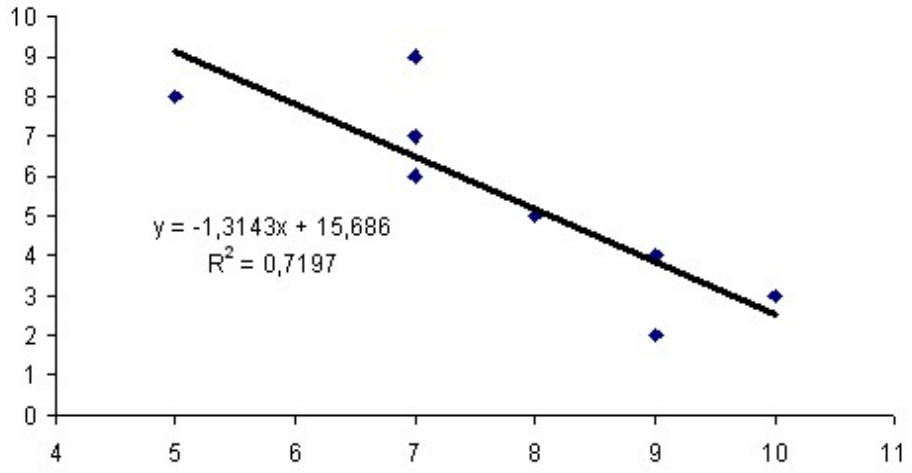


FIGURE 8.8 – *Nuage de points et droite de régression pour les variables note et vin.*

et :

$$b \;=\; \overline{Y} - a\overline{X} \;=\; \frac{\overline{Y}'}{100} + 40,45 - \frac{a'(\overline{X}' + 6)}{100} \;=\; \frac{b'}{100} + 40,45 - \frac{6a'}{100} \;=\; 40,82.$$

3. Pour le trimestre 13, une prévision de la durée hebdomadaire du travail est :

$$-0,0517 \times 13 + 40,82 \;=\; 40,15h.$$

Exercice 15. *1. Le tableau statistique relatif à la variable X est le suivant :*

x_i	2	6	10	12	16	Totaux
n_i	3	21	45	26	5	100
$x_i \times n_i$	6	126	450	312	80	974
$x_i^2 \times n_i$	12	756	4500	3744	1280	10292

Le tableau statistique relatif à la variable Y est le suivant :

y_i	2	6	10	14	18	Totaux
n_i	12	26	41	18	3	100
$x_i \times n_i$	24	156	410	252	54	896
$x_i^2 \times n_i$	48	936	4100	3528	972	9584

2. On a :

$$\overline{X} \;=\; \frac{974}{100} \;=\; 9,74,$$

$$\sigma_X^2 \;=\; \frac{10292}{100} - (9,74)^2 \;=\; 8,0524,\; \sigma_X \;=\; \sqrt{8,0524} \;=\; 2,84,$$

$$\overline{Y} \;=\; \frac{896}{100} \;=\; 8,96,$$

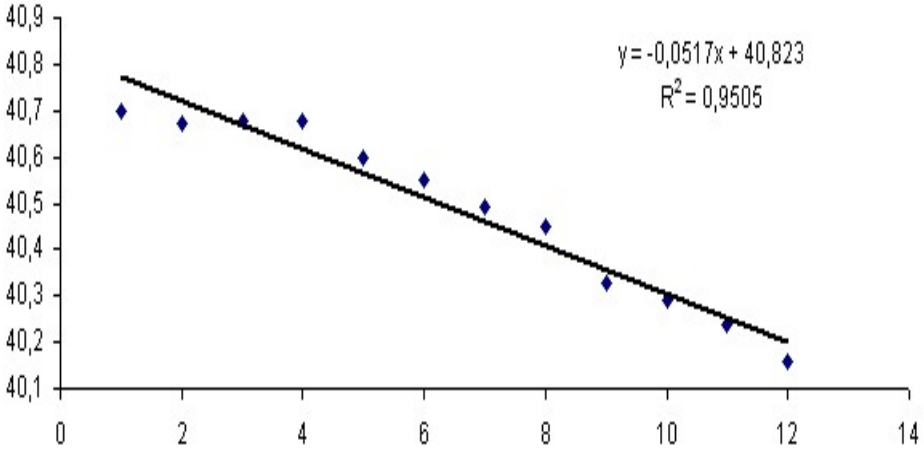


FIGURE 8.9 – *Nuage de points et droite de régression pour les variables trimestre et durée hebdomadaire sans changement de variables.*

$$\sigma_Y^2 \;=\; \frac{9584}{100} - (8,96)^2 \;=\; 15,5584,\;\; \sigma_Y \;=\; \sqrt{15,5584} \;=\; 3,94.$$

3. On a :

$$\mathrm{Cov}(X,Y) \;=\; \frac{1}{n}\sum_{i=1}^{5}\sum_{j=1}^{5} x_i y_j n_{ij} \;-\; \overline{X}\,\overline{Y},$$

où x_i et y_i sont respectivement les modalités des variables X et Y et n est l'effectif cumulé du couple de modalités (x_i, y_j). On obtient :

$$\begin{aligned}\mathrm{Cov}(X,Y) \;&=\; \frac{1}{100}(8+12+60+432+180+84+100+600+2500+700+216\\ &\quad +1440+1680+216+160+448+576) - (9,74\times 8,96) = \; 6,85.\end{aligned}$$

$$r(X,Y) \;=\; \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \;=\; 0,6.$$

Le coefficient de corrélation linéaire et le nuage de points indiquent que la liaison linéaire entre X et Y est assez faible.

Exercice 16. *1. On doit ici calculer le coefficient de Bravais-Pearson sur un échantillon de 25 individus, entre la variable "temps journalier passé devant la télévision" notée X et la variable "âge" notée Y. Les modalités de ces 2 variables étant regroupées en classes, on travaillera avec les centres des classes.*

Tout d'abord pour calculer la covariance entre X et Y, nous avons besoin des moyennes de

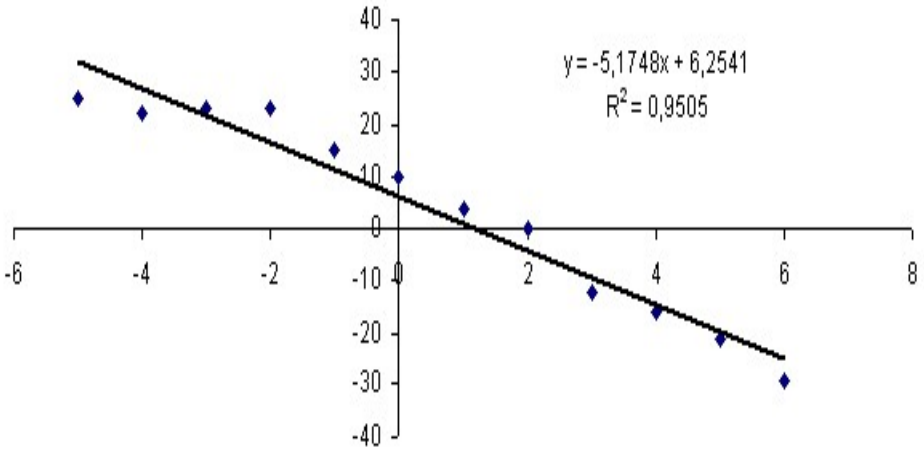


FIGURE 8.10 – *Nuage de points et droite de régression pour les variables trimestre et durée hebdomadaire avec changement de variables.*

X et Y. Pour le coefficient de corrélation linéaire, nous aurons aussi besoin des variances et écarts-type de X et Y. Pour nous guider dans les calculs, nous nous aiderons du tableau suivant

x_i	y_i	n_i	$n_i x_i$	$n_i (x_i)^2$	$n_i y_i$	$n_i (y_i^2)$	$n_i x_i y_i$
1,5	25	5	7,5	11,25	125	3125	187,5
2,5	25	2	5,0	12,50	50	1250	125,0
2,5	35	7	17,5	43,75	245	8575	612,5
3,5	35	4	14,0	49,00	140	4900	490,0
3,5	45	2	7,0	24,50	90	4050	315,0
4,5	45	5	22,5	101,25	225	10125	1012,5
		25	73,5	242,25	875	32025	2742,5

moyennes : $\overline{X} = \dfrac{73,5}{25} = 2,94$ *;* $\overline{Y} = \dfrac{875}{25} = 35$

variances : $\mathrm{var}(X) = \dfrac{242,25}{25} - (2,94)^2 = 1,05$ *;* $\mathrm{var}(Y) = \dfrac{32025}{25} - 35^2 = 56$

écarts-type : $\sigma_X = 1,02$ *;* $\sigma_Y = 7,48$

covariance : $\mathrm{cov}(X,Y) = \dfrac{2742,5}{25} - 2,94 \times 35 = 6,8.$

2. $r(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sigma_X\, \sigma_Y} = \dfrac{6,8}{1,02 \times 7,48} = 0,89.$

$r(X,Y)$ étant proche de 1, on a une forte corrélation linéaire croissante entre le temps journalier passé devant la télé et l'âge, pour les 25 personnes de notre échantillon.

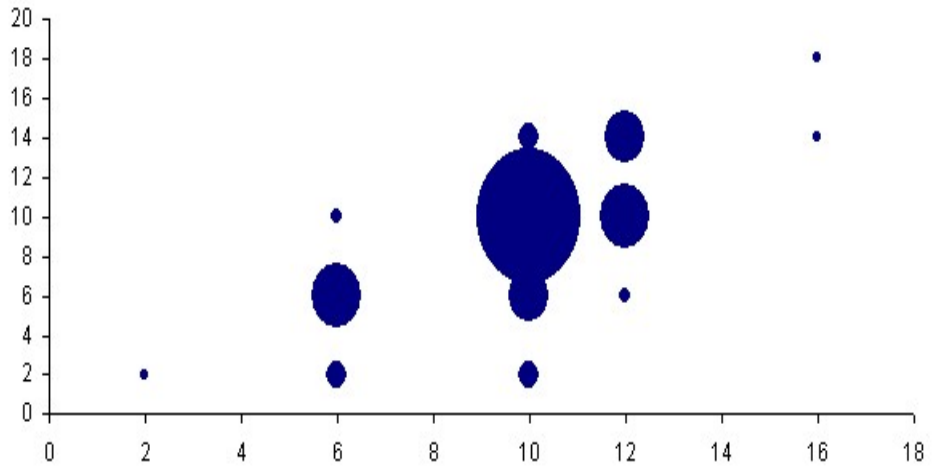


FIGURE 8.11 – *Nuage de points pour les variables notes premère et seconde langue.*

3. Avec les notations choisies, il faut déterminer la droite de régression de Y *en* X*. Elle a pour équation* $y = ax + b$ *avec* $a = \dfrac{\text{Cov}(X,Y)}{\text{var}(X)} = \dfrac{6,8}{1,05} = 6,5$ *et* $b = \overline{Y} - a\overline{X} = 35 - 6,50 \times 2,94 = 15,89$.

La droite de régression de Y *en* X *a pour équation :* $y = 6,5x + 15,89$.

4. La corrélation linéaire étant forte (car $r(X,Y)$ *est proche de 1), il est légitime d'utiliser la droite de régression de* Y *en* X *pour faire une estimation.*

Le centre de la classe $50 - 60$ *étant 55, il suffit de remplacer* y *par 55 dans l'équation de la droite de régression de* Y *en* X*. La valeur de* x *correspondante sera donc l'estimation du temps journaliel passé devant la télévision.*

$x = \dfrac{y - 15,89}{6,50} = \dfrac{55 - 15,89}{6,50} = 6,02$. *On estime donc à 6 heures le temps journalier passé devant la télévision pour une personne de 50 à 60 ans.*

Exercice 17. *1. On doit ici calculer le coefficient de Bravais-Pearson sur un échantillon de 7 projets, entre la variable "temps passé" notée* T *et la variable "évolution" notée* V*.*

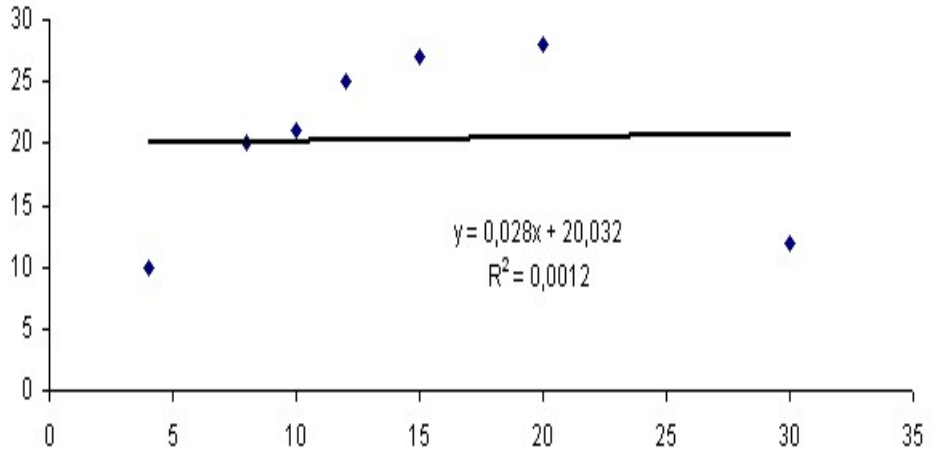


FIGURE 8.12 – *Nuage de points pour les variables temps passé devant la télévision et âge.*

t_i	v_i	n_i	$n_i t_i$	$n_i(t_i)^2$	$n_i v_i$	$n_i(v_i^2)$	$n_i t_i v_i$
12	25	1	12	144	25	625	300
15	27	1	15	225	27	729	405
10	21	1	10	100	21	441	210
4	10	1	4	16	10	100	40
20	28	1	20	400	28	784	560
30	12	1	30	900	12	144	360
8	20	1	8	64	20	400	160
		7	99	1849	143	3223	2035

moyennes : $\overline{T} = \dfrac{99}{7} = 14,14$ *;* $\overline{V} = \dfrac{143}{7} = 20,43$

variances : $\mathrm{Var}(T) = \dfrac{1849}{7} - (14,14)^2 = 64,12$ *;* $\mathrm{Var}(V) = \dfrac{3223}{7} - (20,43)^2 = 43,10$

écarts-type : $\sigma_T = 8,01$ *;* $\sigma_V = 6,57$

covariance : $\mathrm{Cov}(T,V) = \dfrac{2035}{7} - 14,14 \times 20,43 = 1,80.$

2. $r(T,V) = \dfrac{\mathrm{Cov}(T,V)}{\sigma_T\,\sigma_V} = \dfrac{1,80}{8,01 \times 6,57} = 0,03.$

$r(T,V)$ *étant proche de 0, on a presque indépendace linéaire entre le temps passé et l'évolution en pourcentage concernant les 7 projets de notre échantillon.*

Exercice 18. *On doit ici calculer le coefficient de Bravais-Pearson sur un échantillon de 30 enfants, entre la variable "mots mémorisés en lisant" notée X et la variable "mots mémorisés en écoutant" notée Y. Les modalités de ces 2 variables étant regroupées en*

classes, on travaillera avec les centres des classes.

x_i	y_i	n_i	$n_i x_i$	$n_i (x_i)^2$	$n_i y_i$	$n_i (y_i^2)$	$n_i x_i y_i$
2,5	2,5	3	7,5	18,75	7,5	18,75	18,75
2,5	7,5	4	10	25	30	225	75
7,5	2,5	4	30	225	10	25	75
7,5	7,5	5	37,5	281,25	37,5	281,25	281,25
7,5	12,5	1	7,5	56,25	12,5	156,25	93,75
12,5	7,5	9	112,5	1406,25	67,5	506,25	843,75
12,5	12,5	4	50	625	50	625	625
		30	255	2637,5	215	1837,5	2012,5

moyennes : $\overline{X} = \dfrac{255}{30} = 8,5$ *;* $\overline{Y} = \dfrac{215}{30} = 7,17$

variances : $\mathrm{Var}(X) = \dfrac{2637,5}{30} - (8,5)^2 = 15,67$ *;* $\mathrm{Var}(Y) = \dfrac{1837,5}{30} - (7,17)^2 = 9,89$

écarts-type : $\sigma_X = 3,96$ *;* $\sigma_Y = 3,15$

covariance : $\mathrm{Cov}(X,Y) = \dfrac{2012,5}{30} - 8,5 \times 7,17 = 6,17$ *;*

corrélation : $r(X,Y) = \dfrac{\mathrm{cov}(X,Y)}{\sigma_X\, \sigma_Y} = \dfrac{6,17}{3,96 \times 3,15} = 0,50.$

$r(X,Y)$ *étant positif, on a une corrélation linéaire croissante moyenne entre les deux types de mémorisation pour ces 30 enfants.*

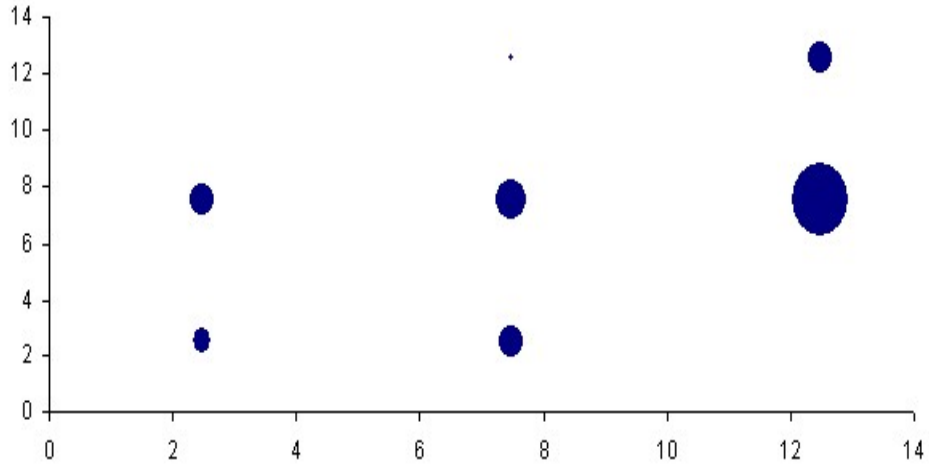


FIGURE 8.13 – *Nuage de points pour le nombre de mots mémorisés.*

Exercice 19. *1.*

t_i	v_i	n_i	$n_i t_i$	$n_i (t_i)^2$	$n_i v_i$	$n_i (v_i^2)$	$n_i t_i v_i$
0	3400	1	0	0	3400	11560000	0
1	4000	1	1	1	4000	16000000	4000
2	3200	1	2	4	3200	10240000	6400
3	3700	1	3	9	3700	13690000	11100
4	3600	1	4	16	3600	12960000	14400
5	3100	1	5	25	3100	9610000	15500
6	3300	1	6	36	3300	10890000	19800
7	3500	1	7	49	3500	12250000	24500
8	4200	1	8	64	4200	17640000	33600
9	4100	1	9	81	4100	16810000	36900
		10	45	285	36100	131650000	166200

moyennes : $\overline{T} = \dfrac{45}{10} = 4,5$ *;* $\overline{V} = \dfrac{36100}{10} = 3610$

variances : $\mathrm{var}(T) = \dfrac{285}{10} - (4,5)^2 = 8,25$ *;* $\mathrm{var}(V) = \dfrac{131650000}{10} - (3610)^2 = 132900$

écarts-type : $\sigma_T = 2,87$ *;* $\sigma_V = 364,56$

covariance : $\mathrm{Cov}(T,V) = \dfrac{166200}{10} - 4,5 \times 3610 = 375$ *;*

corrélation : $r(T,V) = \dfrac{\mathrm{Cov}(T,V)}{\sigma_T\, \sigma_V} = \dfrac{375}{2,87 \times 364,55} = 0,36.$

La droite de régression de V *en* T *a pour équation* $v = at + b$ *avec* $a = \dfrac{\mathrm{Cov}(T,V)}{\mathrm{Var}(T)} = \dfrac{375}{8,25} = 45,46$ *et* $b = \overline{V} - a\overline{T} = 33610 - 45,46 \times 4,5 = 3405,46.$

La droite de régression de V *en* T *a donc pour équation* $v = 45,46t + 3405,46.$

2. $r(T,V)$ *étant faible, on ne doit pas utiliser la droite de régression de* V *en* T *pour faire des estimations et la question 3. est sans objet.*

Exercice 20. *1. On sait que* $\mathrm{Cov}(X+1800, Y) = \mathrm{Cov}(X,Y)$ *et* $r(X+1800, Y) = r(X,Y)$. *Ensuite*

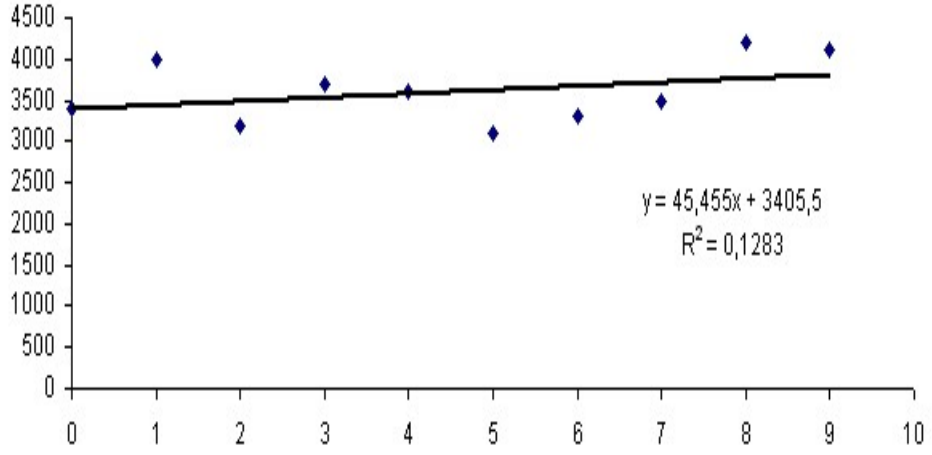


FIGURE 8.14 – *Nuage de points et droite de régression pour les variables année et ventes.*

$$\sum_{i=1}^{7} x_i = 90; \quad \bar{X} = \frac{90}{7} = 12,857$$

$$\sum_{i=1}^{7} y_i = 306; \quad \bar{Y} = \frac{306}{7} = 43,7142$$

$$\sum_{i=1}^{7} x_i y_i = 3618; \quad \mathrm{Cov}(X,Y) = \frac{3618}{7} - 12,857 \times 43,7142 = -45,176$$

$$\sum_{i=1}^{7} x_i^2 = 1598; \quad \mathrm{Var}(X) = \frac{1598}{7} - 12,857^2 = 62,98; \quad \sigma(X) = 7,936$$

$$\sum_{i=1}^{7} y_i^2 = 13608; \quad \mathrm{Var}(Y) = \frac{13608}{7} - 43,7142^2 = 33,07; \quad \sigma(Y) = 5,750$$

$$r(X,Y) = \frac{-45,176}{7,936 \times 5,75} = -0,99.$$

2. Le coefficient de corrélation linéaire est proche de -1*, il y a donc une forte corrélation linéaire négative, on peut utiliser la droite de régression pour faire des prévisions. L'équation de* $\Delta_{Y/X}$ *est* $y = ax + b$ *avec*

$$a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \frac{-45,176}{62,98} = -0,7173$$

et

$$b = \overline{Y} - a\overline{X} = 43,7142 + 0,7173 \times 12,857 = 52,9365.$$

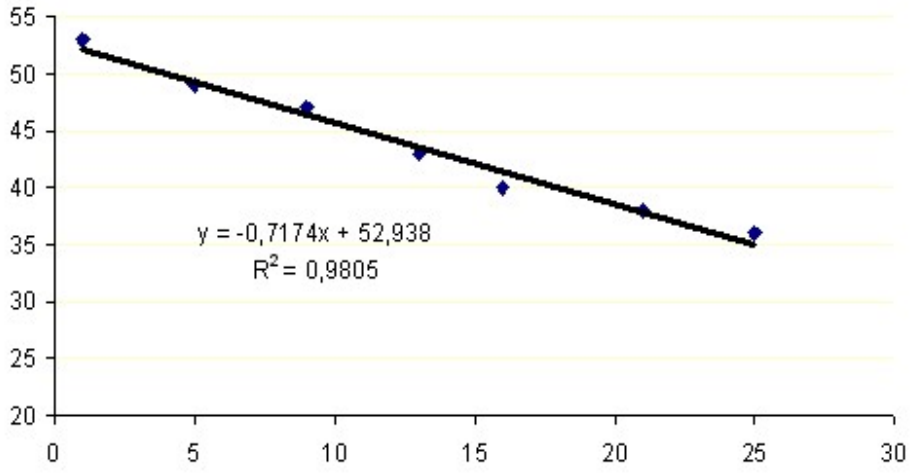


FIGURE 8.15 – *Nuage de points et droite de régression pour les variables année et pourcentage d'illettrés.*

L'équation de $\Delta_{Y/X}$ *est* $y = -0,7173x + 52,9365$.
On prend $x = 30$, $-0,7173 \times 30 + 52,9365 = 31,42$. *Avec ce modèle, on peut prévoir 31,42% d'illettrés parmi les conscrits en 1830.*

Exercice 21. *1. La moyenne de* X *est de 2796 / 7 = 399, 43 professeurs femmes.*
$\mathrm{Var}(X) = 1145936/7 - 399,43^2 = 4160,1$; $\sigma(X) = 64,51$.

2. $\mathrm{Cov}(X,Y) = 1/7 \times 2180313 - 399,43 \times 796,4 = -6632,75$.
$r(X,Y) = -6632,75/(64,51 \times 115,3) = -0,89$.
Il y a donc corrélation linéaire négative entre les deux variables.

3. On a $a = -6632,75/4504 = -1,47$; $b = 1384,5$. *Droite de régression linaire :*

$$y = -1,47x + 1384.$$

Exercice 22. *Notons* X *la variable "effectifs au fond" et* Y *la variable "production de*

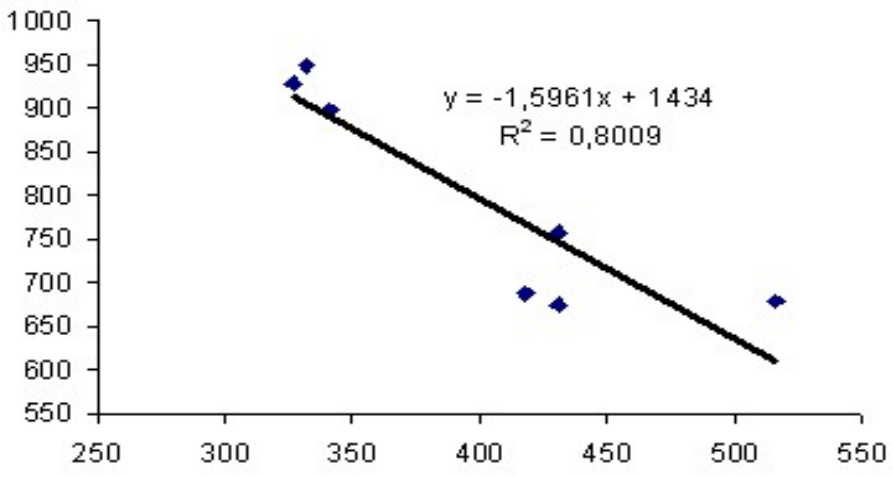


FIGURE 8.16 – *Nuage de points et droite de régression pour les variables année et professeurs.*

charbon".

$$\sum_{i=1}^{11} x_i = 517,7; \quad \bar{X} = 47,06$$
$$\sum_{i=1}^{11} y_i = 302; \quad \bar{Y} = 27,45$$
$$\sum_{i=1}^{11} x_i y_i = 15027,45; \quad \mathrm{Cov}(X,Y) = 74,33$$
$$\sum_{i=1}^{11} x_i^2 = 26099,29; \quad \mathrm{Var}(X) = 157,75; \quad \sigma(X) = 12,56$$
$$\sum_{i=1}^{11} y_i^2 = 8682,18; \quad \mathrm{Var}(Y) = 35,52; \quad \sigma(Y) = 5,96$$
$$r(X,Y) = 0,99.$$

Il y a une forte corrélation linéaire positive entre la production de charbon et les effectifs employés au fond.

Exercice 23. *1. Notons T' la variable temps dont les valeurs sont 1800 , 1810 , 1820 , 1830 , ... , 1910 et notons T la variable temps dont les valeurs sont 1 , 2 , 3 , 4 , ... , 12. On remarque que*

$$T = \frac{T' - 1790}{10} \Longleftrightarrow T' = 10T + 1790.$$

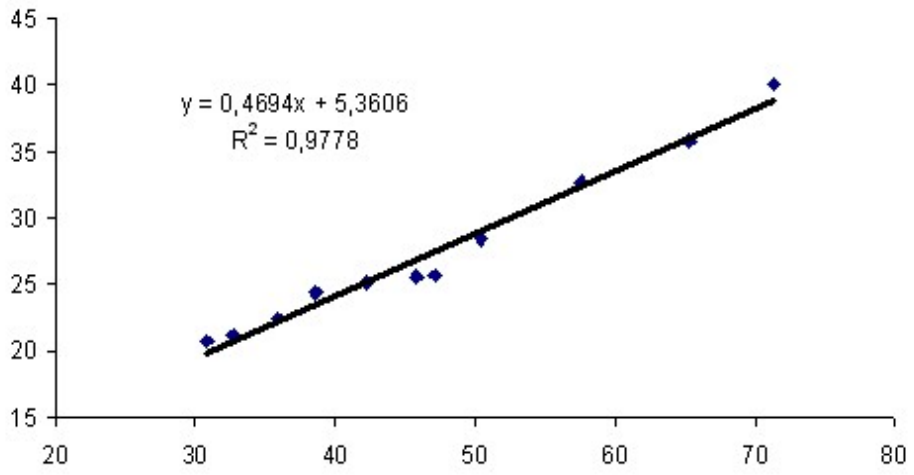


FIGURE 8.17 – *Nuage de points et droite de régression pour les variables effectifs employés au fond et production de charbon.*

Notons Y la variable " Taux de natalité ".
En utilisant les propriétés de la covariance, on a :

$$\mathrm{Cov}(T', Y) = \mathrm{Cov}(10T + 1790, Y) = \mathrm{Cov}(10T, Y) = 10\mathrm{Cov}(T, Y).$$

En utilisant les propriétés de $r(X, Y)$ vues dans l'exercice 8.2 on a :

$$r(T', Y) = r(10T + 1790, Y) = r(10T, Y) = r(T, Y).$$

Ce type de changement de variable ne modifie pas la corrélation linéaire, nous l'utiliserons systématiquement dans le chapitre sur les séries chronologiques.
Notons $y = a't' + b'$ l'équation de $\Delta_{Y/T'}$ et $y = at + b$ l'équation de $\Delta_{Y/T}$.
D'après les propriétés de la variance,

$$\mathrm{Var}(T') = \mathrm{Var}(10T + 1790) = \mathrm{Var}(10T) = 100\mathrm{Var}(T).$$

On en déduit que

$$a' = \frac{\mathrm{Cov}(T', Y)}{\mathrm{Var}(T')} = \frac{10\mathrm{Cov}(T, Y)}{100\mathrm{Var}(T)} = \frac{1}{10}a.$$

D'après les propriétés de la moyenne,

$$\overline{T'} = \overline{10T + 1790} = \overline{10T} + 1790 = 10\overline{T} + 1790.$$

On en déduit que

$$b' = \overline{Y} - a\overline{T'} = \overline{Y} - \frac{1}{10}a\left(10\overline{T} + 1790\right) = \overline{Y} - a\overline{T} - a179 = b - 179a.$$

En posant $t' = 10t + 1790$, *on constate que* $y = a't' + b' \iff y = at + b$. *Nous pouvons donc utiliser l'équation de* $\Delta_{Y/T}$ *pour faire des prévisions.*

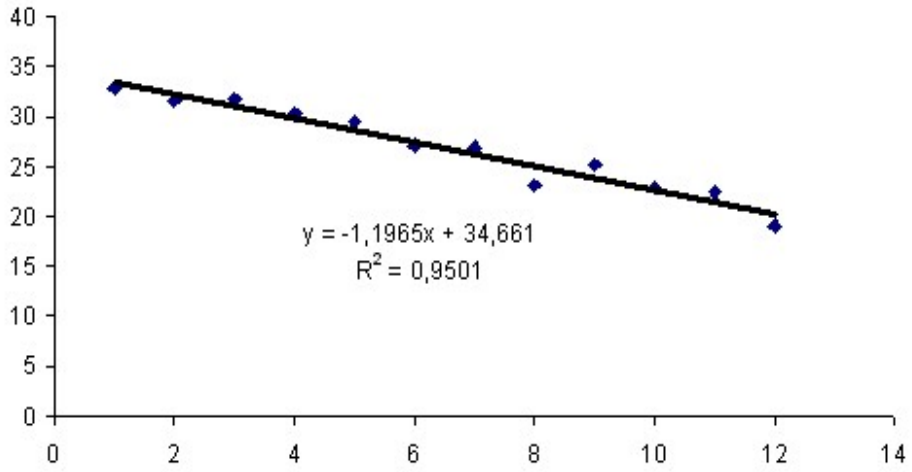


FIGURE 8.18 – *Nuage de points et droite de régression pour les variables année et natalité.*

2.

$$\sum t_i = 78; \quad \overline{T} = 6,5$$
$$\sum y_i = 321,6; \quad \overline{Y} = 26,8$$
$$\sum t_i y_i = 1920,8; \quad \mathrm{Cov}(X,Y) = -14,13$$
$$\sum t_i^2 = 650; \quad \mathrm{Var}T = 11,90; \quad \sigma(T) = 3,45$$
$$\sum y_i^2 = 8830,04; \quad \mathrm{Var}(Y) = 17,56; \quad \sigma(Y) = 4,19$$
$$r(X,Y) = -0,99.$$

Il y a une forte corrélation linéaire décroissante et il est donc intéressant de tracer la droite de régression. L'équation de $\Delta_{Y/X}$ *est* $y = ax + b$ *avec*

$$a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \frac{-14,13}{11,9} = -1,19 \quad \textit{et} \quad b = \overline{Y} - a\overline{X} = 26,86 + 1,19 \times 6,5 = 34,53.$$

L'équation de $\Delta_{Y/X}$ *est* $y = -1,19x + 34,53$.

On prend $t = 21$, $-1,19 \times 21 + 34,53 = 9,54$, *soit un taux de natalité pour 1000 de 9,54. En réalité, ce taux est d'environ 13, la prévision est donc relativement loin de la réalité et si on utilise encore* $\Delta_{Y/T}$ *pour des prévisions futures on trouverait par exemple pour 2100*

un taux de natalité négatif, ce qui est absurde. Même lorsque la corrélation linéaire est forte, il ne faut pas abuser de ce modèle, on ne peut faire que des prévisions à court terme.

Exercice 24. *Les résultats numériques sont donnés dans l'énoncé. Cet exercice dépasse le cadre du programme de cette UE, il donne une ouverture sur d'autres types de corrélation que l'on peut étudier entre deux variables quantitatives en se basant sur la corrélation linéaire.*

8.3 Correction des exercices du Chapitre 3

Exercice 25. *1. Les représentations graphiques font apparaître des variations saisonnières.*

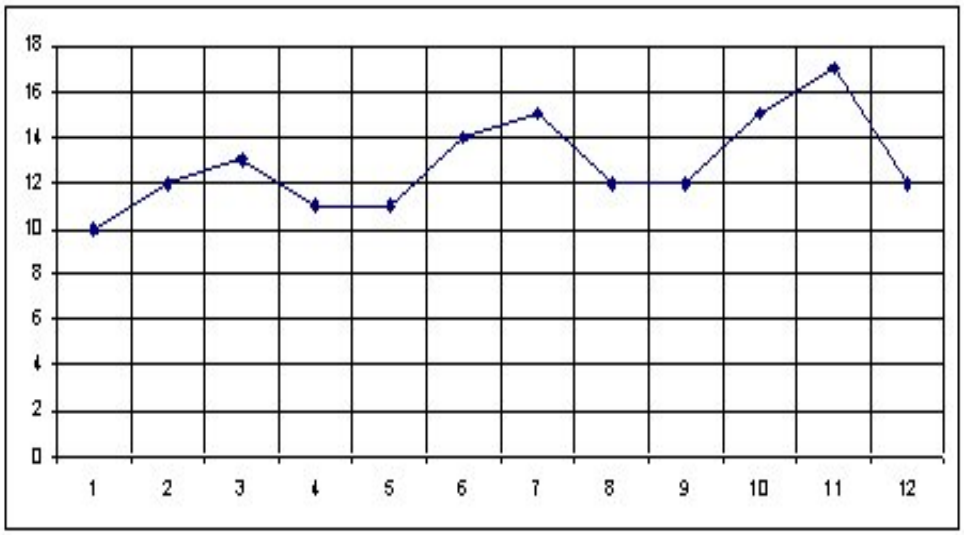


FIGURE 8.19 – Représentation graphique des mariages avec périodes successives

2.3.4. Les relevés étant trimestriels, on considère les moyennes mobiles d'ordre 4 :

$$z(t) = \frac{y(t-2)/2 + y(t-1) + y(t) + y(t+1) + y(t+2)}{4}.$$

Ensuite on réalise la régression de z sur t_z. On a d'après le cours

$$\bar{t}_z = \frac{N+1}{2} = 6,5 \qquad \mathrm{Var}(t_z) = \frac{N_z^2 - 1}{12} = 5,25$$

où $N = 12$ et $N_z = 8$. On a aussi $\overline{z} = \dfrac{103,25}{8} \approx 12,91$ puis

$$\mathrm{Cov}(z, t_z) = \frac{685,25}{8} - \frac{103,25}{8} 6,5 \approx 1,77$$

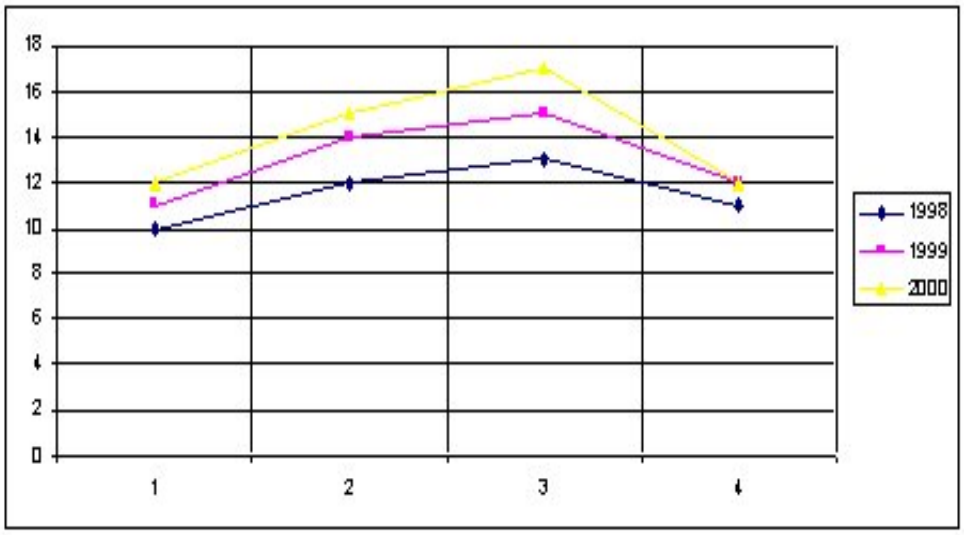


FIGURE 8.20 – Représentation graphique des mariages avec périodes superposées

et $a = \dfrac{\text{Cov}(z, t_z)}{\text{Var}(t_z)} \approx 0,34$ *et* $b = \overline{z} - a\overline{t}_z \approx 12,72$.

On peut présenter l'ensemble des calculs dans le tableau suivant :

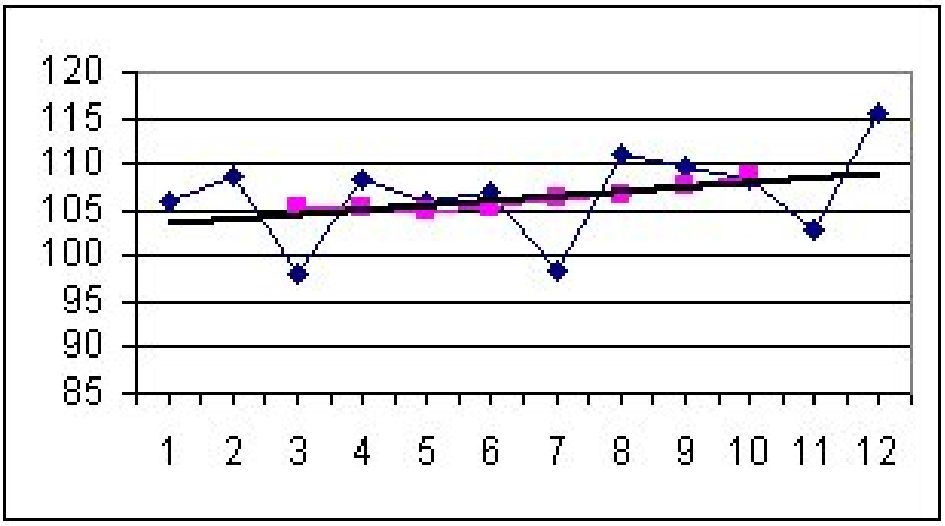
t	$y(t)$	$z(t)$	$t.z(t)$	$x(t) = 0,336t + 10,724$	$y(t) - x(t)$	$s'(t)$	$s(t)$
1	*10*			*11,06*	*-1,06*	*-1,404*	*-1,329*
2	*12*			*11,396*	*0,604*	*0,927*	*1,002*
3	*13*	*11,63*	*34,88*	*11,732*	*1,268*	*1,924*	*1,999*
4	*11*	*12,00*	*48,00*	*12,068*	*-1,068*	*-1,745*	*-1,670*
5	*11*	*12,50*	*62,50*	*12,404*	*-1,404*		
6	*14*	*12,88*	*77,25*	*12,74*	*1,26*		
7	*15*	*13,13*	*91,88*	*13,076*	*1,924*		
8	*12*	*13,38*	*107,00*	*13,412*	*-1,412*		
9	*12*	*13,75*	*123,75*	*13,748*	*-1,748*		
10	*15*	*14,00*	*140,00*	*14,084*	*0,916*		
11	*17*			*14,42*	*2,58*		
12	*12*			*14,756*	*-2,756*		
		103,25	*685,25*	-	-	$S0 = -0,3$	

5.

t	$x(t) = 0,336t + 10,724$	$s(t)$	$x(t) + s(t)$
17	*16,436*	*-1,329*	*15,107*
18	*16,772*	*1,002*	*17,774*
19	*17,108*	*1,999*	*19,107*
20	*17,444*	*-1,670*	*15,774*

Exercice 26.

t	$y(t)$	$z(t)$	$t \times z(t)$	$x(t) = 0,49t + 103,04$	$y(t) - x(t)$	$s'(t)$	$s(t)$
1	*106*			*103,53*	*2,47*	*1,743*	*1,243*
2	*108,8*			*104,02*	*4,78*	*2,153*	*1,653*
3	*97,9*	*105,31*	*315,94*	*104,51*	*-6,61*	*-6,670*	*-7,170*
4	*108,5*	*105,11*	*420,45*	*105*	*3,5*	*4,773*	*4,273*
5	*106,1*	*104,98*	*524,88*	*105,49*	*0,61*		
6	*107,1*	*105,39*	*632,33*	*105,98*	*1,12*		
7	*98,5*	*106,16*	*743,14*	*106,47*	*-7,97*		
8	*111,2*	*106,78*	*854,20*	*106,96*	*4,24*		
9	*109,6*	*107,51*	*967,61*	*107,45*	*2,15*		
10	*108,5*	*108,61*	*1086,13*	*107,94*	*0,56*		
11	*103*			*108,43*	*-5,43*		
12	*115,5*			*108,92*	*6,58*		
		849,85	*5544,66*	-	-	$S_0 = 2$	



3ième trimestre 1985 signifie $t = 15$ *d'où* $\hat{y}(15) = 0,49 \times 15 - 7,170 = 103,22$.

Exercice 27. *1. On veut calculer des moyennes mobiles d'ordre 12. La première que l'on peut calculer correspond au mois de juillet 1974* $(t = 7)$ *:*

$$z(7) = \frac{\frac{127,4}{2} + 129,1 + \ldots + 144,3 + \frac{145,9}{2}}{12} = 137,046.$$

La dernière que l'on peut calculer correspond au mois de juin 1977 $(t = 42)$ *:*

$$z(42) = \frac{\frac{173,8}{2} + 174,3 + \ldots + 188,9 + \frac{189,4}{2}}{12} = 182,042.$$

2. La courbe est déjà lisse, il est inutile de calculer la série des moyennes mobiles : les valeurs seraient très proches des valeurs observées.

3. Il n'y a pas de variations saisonnières.

4. Pour faire des prévisions, on pourrait étudier simplement la corrélation linéaire.

Exercice 28.

t	$y(t)$	$z(t)$	$x(t) = 0,26t + 1,15$	$y(t) - x(t)$	$s'(t)$	$s(t) = s'(t) + 0,6$
1	1		1,41	-0,41	-0,36	0,24
2	0		1,67	-1,67	-0,37	0,23
3	5	1,6	1,93	3,07	4,37	4,97
4	2	1,8	2,19	-0,19	-1,39	-0,79
5	0	2,4	2,45	-2,45	-2,65	-2,05
6	2	2,8	2,71	-0,71		
7	3	3,2	2,97	0,03		
8	7	3,4	3,23	3,77		
9	4	3,8	3,49	0,51		
10	1	4	3,75	-2,75		
11	4	4,6	4,01	-0,01		
12	4	4,2	4,27	-0,27		
13	10	4,4	4,53	5,47		
14	2	4,6	4,79	-2,79		
15	2	5	5,05	-3,05		
16	5	5,2	5,31	-0,31		
17	6	5,4	5,57	0,43		
18	11	5,8	5,83	5,17		
19	3		6,09	-3,09		
20	4		6,35	-2,35		
					$S_0 = -3$	

1. On calcule ici des moyennes mobiles d'ordre 4.

2. On est obligé de calculer tous les coefficients $s'(t)$ pour faire la somme puis calculer les coefficients saisonniers $s(t)$.

3.

t	$x(t) = 0,26t + 1,15$	$s(t) = s'(t) + 0,6$	$x(t) + s(t)$
21	6,61	0,24	6,85
22	6,87	0,23	7,1
23	7,13	4,97	12,1

Exercice 29.

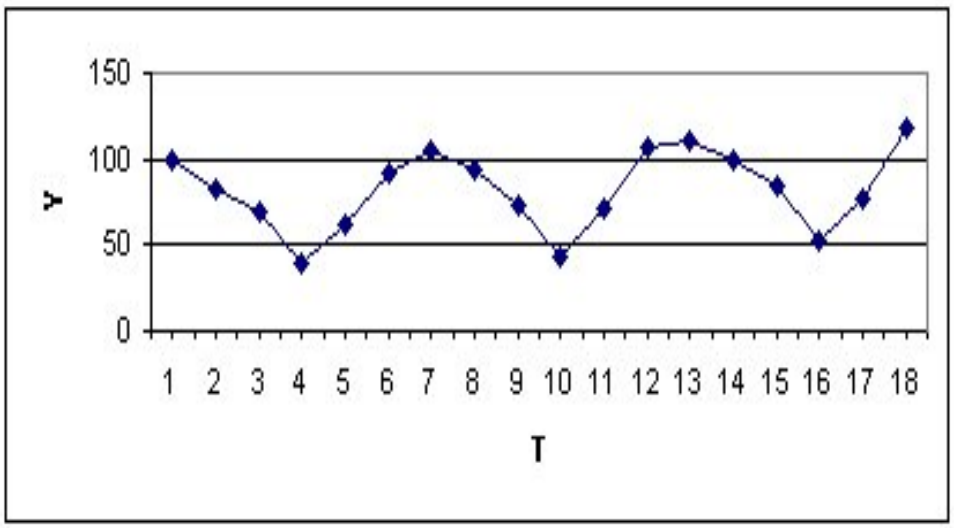
t	$y(t)$	$z(t)$	$t.z(t)$	$x(t) = 0,70t + 13,45$	$y(t) - x(t)$	$s'(t)$	$s(t)$
1	*8,2*			*14,15*	*-5,95*	*-7,217*	*-7,359*
2	*12,3*			*14,85*	*-2,55*	*-2,950*	*-3,092*
3	*32,7*	*15,6*	*46,8*	*15,55*	*17,15*	*18,750*	*18,608*
4	*8,3*	*16,1*	*64,4*	*16,25*	*-7,95*	*-8,017*	*-8,159*
5	*10*	*16,95*	*84,75*	*16,95*	*-6,95*		
6	*14,5*	*17,925*	*107,55*	*17,65*	*-3,15*		
7	*37,3*	*18,45*	*129,15*	*18,35*	*18,95*		
8	*11,5*	*18,925*	*151,4*	*19,05*	*-7,55*		
9	*11*	*19,775*	*177,975*	*19,75*	*-8,75*		
10	*17,3*	*20,5*	*205*	*20,45*	*-3,15*		
11	*41,3*			*21,15*	*20,15*		
12	*13,3*			*21,85*	*-8,55*		
		144,23	*6967,03*	-	-	$S0 = 0,57$	

4. Prévisions pour le 3ième trimestre 2000 :

$$\hat{y}(23) = x(23) + s(3) = 0,7 \times 23 + 13,45 + 18,608 = 48,158.$$

On prévoit donc avec ce modèle $\boxed{48158 \text{ visiteurs}}$ *pour l'été 2000.*

Exercice 30. *1. Notons T la variable temps numérotée dans l'ordre chronologique. Notons Y la variable "nombre de billets vendus (en milliers)".*



2. Il faut calculer des moyennes mobiles d'ordre 6.

t	$y(t)$	$z(t)$	$t.z(t)$	$x(t) = 1,34t + 68,93$	$y(t) - x(t)$	$s'(t)$	$s(t) = s'(t) - 0,51$
1	*100*			*70,27*	*29,73*	*27,02*	*26,51*
2	*82*			*71,61*	*10,39*	*12,02*	*11,51*
3	*70*			*72,95*	*-2,95*	*-5,32*	*-5,83*
4	*40*	*74,58*	*298,33*	*74,29*	*-34,29*	*-37,33*	*-37,84*
5	*62*	*76,00*	*380,00*	*75,63*	*-13,63*	*-13,34*	*-13,85*
6	*91*	*77,25*	*463,50*	*76,97*	*14,03*	*19,99*	*19,48*
7	*105*	*77,75*	*544,25*	*78,31*	*26,69*		
8	*94*	*78,83*	*630,67*	*79,65*	*14,35*		
9	*73*	*80,92*	*728,25*	*80,99*	*-7,99*		
10	*43*	*82,67*	*826,67*	*82,33*	*-39,33*		
11	*72*	*83,58*	*919,42*	*83,67*	*-11,67*		
12	*106*	*84,92*	*1019,00*	*85,01*	*20,99*		
13	*111*	*86,58*	*1125,58*	*86,35*	*24,65*		
14	*99*	*87,75*	*1228,50*	*87,69*	*11,31*		
15	*84*	*89,17*	*1337,50*	*89,03*	*-5,03*		
16	*52*			*90,37*	*-38,37*		
17	*77*			*91,71*	*-14,71*		
18	*118*			*93,05*	*24,95*		
		980	*9501,67*	-	-	$S0 = 3,04$	

3. Pour tracer la droite de tendance, on peut placer le point $B(0; 68,93)$ *et le point moyen :*

$$\overline{T} = \quad = 9,5 \quad \overline{Z} = \frac{980}{12} = 81,67$$

i.e. le point $M(9,5; 81,67)$ *appartient à la droite.*

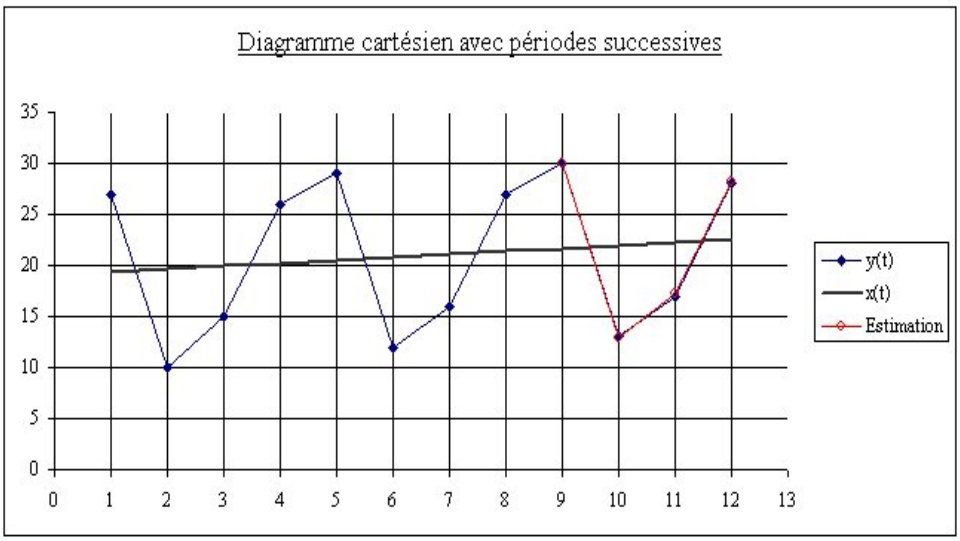
4. Prévisions pour Juillet-Août 2000 :

$$\hat{y}(22) = x(22) + s(4) = 1,34 \times 22 + 68,93 - 37,84 = 60,57,$$

soit $\boxed{\textit{60570 billets vendus}}$.

Exercice 31. *1. On commence par faire le diagramme cartésien :*

2. 3. et 4. On fait ensuite le tableau :



t	$y(t)$	$z(t)$	$tz(t)$	$x(t)$	$\delta(t)$	$s'(t)$	$s(t)$	*estimation*
1	27			19,35	7,65	8,16066667	8,26683333	27,61683
2	10			19,639	-9,639	-9,12833333	-9,02216667	10,61683
3	15	19,75	59,25	19,928	-4,928	-5,084	-4,97783333	14,95017
4	26	20,25	81	20,217	5,783	5,627	5,73316667	25,95017
5	29	20,625	103,125	20,506	8,494		8,26683333	28,77283
6	12	20,875	125,25	20,795	-8,795		-9,02216667	11,77283
7	16	21,125	147,875	21,084	-5,084		-4,97783333	16,10617
8	27	21,375	171	21,373	5,627		5,73316667	27,10617
9	30	21,625	194,625	21,662	8,338		8,26683333	29,92883
10	13	21,875	218,75	21,951	-8,951		-9,02216667	12,92883
11	17			22,24	-5,24		-4,97783333	17,26217
12	28			22,529	5,471		5,73316667	28,26217

D'où l'on déduit :

$$\overline{T'} = 6,5 \ ; \qquad \mathrm{Var}(T') = 0,25 \ ; \qquad \overline{Z} = 20,9375$$

$$\mathrm{Cov}(T', Z) = 1,515625 \ ; \qquad S_0/4 = -0,10616667$$

$$a = 0,28869048 \ ; \qquad b = 19,0610119.$$

6. Prévision pour le 3ième trimestre 2001 : $\boxed{18,418}$ *milliers de francs (soit 18418 francs).*

Exercice 32. *1. Les variations observées ne sont pas des variations saisonnières.*

2. $z(8) = 1077,375$, $z(9) = 1080,875$ *et* $z(10) = 1090,25$. *Les moyennes mobiles lissent la courbe et font apparaître une tendance linéaire croissante. Le modèle additif est inadapté.*

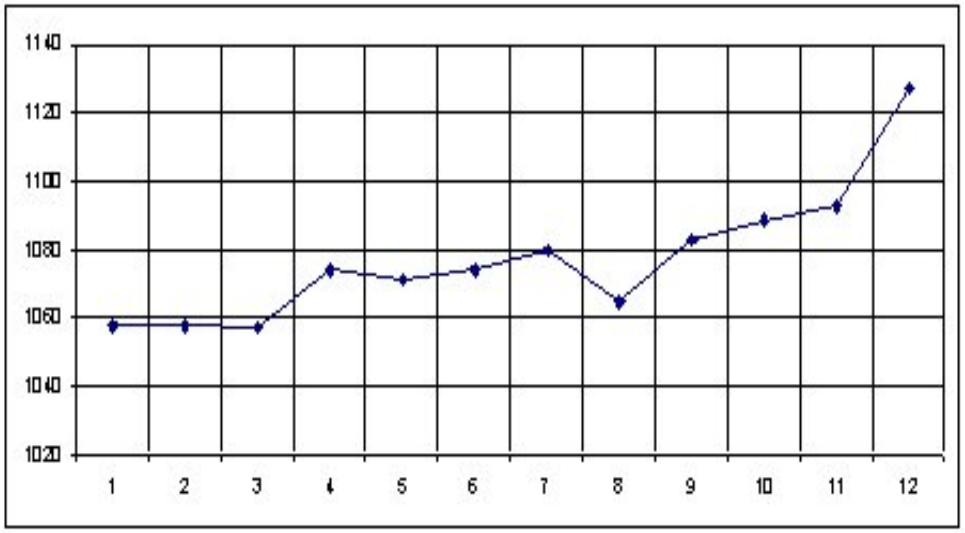


FIGURE 8.21 – Représentation graphique de la construction avec périodes successives

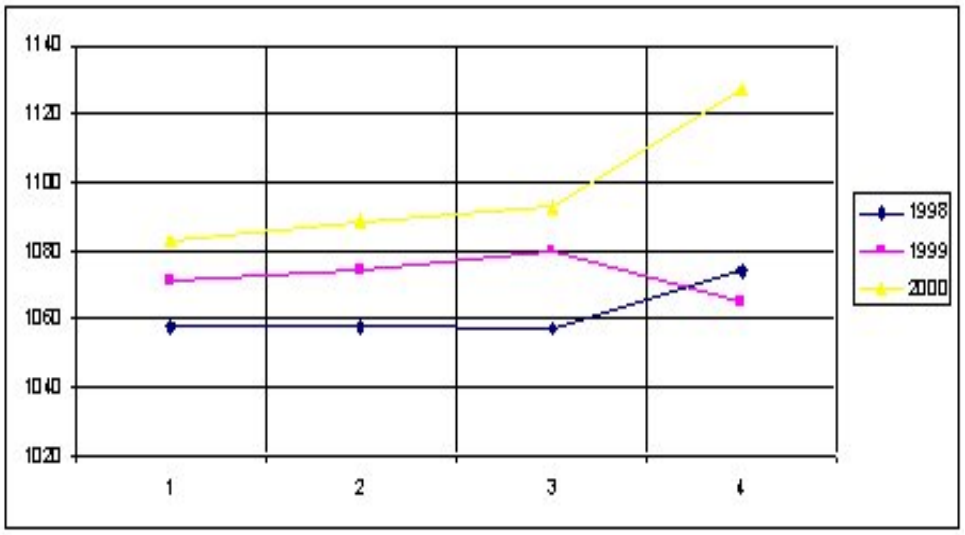


FIGURE 8.22 – Représentation graphique de la construction avec périodes superposées