# Éléments de statistique descriptive (Christian AMBROSINI)

**Ce cours a été dispensé entre 1998 et 2011 auprès des étudiants de Licence de la Faculté de Sciences Economiques et de Gestion et de la Faculté de Droit et Science Politique (AES) de l'Université Lumère-Lyon 2.**

## PLAN DU COURS

**Chapitre 2 : mise en relation de deux caractères qualitatifs**

1. Les représentations graphiques appropriées
    11. Les tuyaux d'orgue à base variable
    12. Les secteurs semi-circulaires
2. La notion d'indépendance entre deux caractères
    21. Principe général
    22. Les tableaux théoriques d'indépendance

**Chapitre 3 : mise en relation d'un caractère qualitatif et d'un caractère quantitatif**

1. Les représentations graphiques appropriées
2. Dépendance et indépendance entre les deux caractères
3. Calcul des paramètres marginaux et conditionnels du caractère quantitatif
4. Notion et calcul du rapport d'explication
    41. Principe général (décomposition de la variance)
    42. Application à l'exemple

**Chapitre 4 : mise en relation de deux caractères quantitatifs**

1. Recherche de la dépendance ou de l'indépendance linéaire de deux caractères quantitatifs : cas des tableaux individus-variables
    11. Données générales
    12. La droite des moindres carrés, expression de la dépendance linéaire entre deux caractères quantitatifs
    13. Calcul de l'intensité de la dépendance linéaire entre deux caractères quantitatifs : le coefficient de corrélation linéaire
    14. Droite d'ajustement des moindres carrés et coefficient de corrélation linéaire
    15. Exemple
2. Recherche de la dépendance ou de l'indépendance linéaire de deux caractères quantitatifs : cas des tableaux à double entrée
    21. Cadre général et formules de calculs
    22. Relations entre paramètres marginaux et conditionnels
    23. Les rapports de corrélation et le coefficient de corrélation linéaire
    24. Les courbes de régression
    25. Indépendance totale entre deux caractères quantitatifs
    26. Exemple

**Partie 3 : les indices statistiques**

**Chapitre 1 : les indices élémentaires**

1. Définition
2. Propriétés des indices élémentaires
    21. La circularité
    22. La réversibilité
    23. La multiplication des indices élémentaires

**Chapitre 2 : les indices synthétiques**

1. Les indices de Laspeyres
    11. L'indice des prix de Laspeyres
    12. L'indice des quantités (ou de volume) de Laspeyres
2. Les indices de Paasche
    21. L'indice des prix de Paasche
    22. L'indice des quantités (ou de volume) de Paasche
3. L'indice de valeur (ou de dépense) globale

**Chapitre 3 : l'indice des prix à la consommation de l'Insee**

**Partie 4 : les séries chronologiques**

**Chapitre 1 : représentation graphique des séries chronologiques**

1. Une distinction fondamentale : données en termes de stocks et en termes de flux
2. La représentation graphique de séries d'indices en coordonnées arithmétiques

**Chapitre 2 : lissage des séries chronologiques**

1. Les composantes d'une série chronologique
2. Le lissage d'une série chronologique
    21. Méthodes empiriques de lissage d'une série chronologique
    22. Méthode analytique de lissage d'une série chronologique

**Chapitre 3 : désaisonnalisation des séries chronologiques**

1. Les hypothèses qui sous-tendent une désaisonnalisation
    11. Hypothèses générales
    12. Hypothèses relatives aux différentes composantes de la série
    13. Hypothèses relatives à la composition des éléments constitutifs de la série chronologique
2. Les méthodes de traitement des séries en vue d'une désaisonnalisation
3. Les étapes successives d'une désaisonnalisation
    Étape 1 : repérage et traitement des valeurs "aberrantes"
    Étape 2 : choix d'un schéma de composition
    Étape 3 : estimation du mouvement extra saisonnier
    Étape 4 : calcul des variations saisonnières
    Étape 5 : calcul des coefficients (ou des rapports) saisonniers
    Étape 6 : calcul (éventuel) des coefficients (ou des rapports) saisonniers corrigés
    Étape 7 : calcul de la série corrigée des variations saisonnières (cvs)
4. Exemple de désaisonnalisation
5. Quelques compléments
    51. Évolution d'une grandeur en moyenne annuelle et en glissement
    52. Exemple d'utilisation de la méthode analytique de lissage d'une série chronologique
    53. Filtrage d'une série chronologique

## INTRODUCTION

### 1. Objet et définition de la statistique descriptive

L'étude rationnelle d'un phénomène économique nécessite :

- premier temps : **collecter de l'information**.

À cet effet, on organise généralement des enquêtes auprès des agents économiques concernés (entreprises, ménages, etc.), sous forme d'entretiens téléphoniques ou en face à face ou encore au moyen de questionnaires à remplir et à renvoyer par courrier ou par mail.

On peut aussi travailler éventuellement sur des données statistiques déjà existantes.

L'ensemble des informations dont on dispose à ce stade représente ce que l'on appelle les **données brutes**.

- deuxième temps : **classement (mise en ordre) des données**.

On réalise une première manipulation sur les données brutes en en **classant** les valeurs, généralement dans l'ordre croissant. À cette étape, on peut **représenter graphiquement la série de données**, afin de faire apparaître la physionomie de celle-ci, en première approximation.

Mais on remarquera que la totalité de l'information disponible présente, en même temps, le désavantage d'être difficilement manipulable, si l'on dispose d'un grand nombre de données. En effet, on ne perçoit pas aisément les caractéristiques principales de la série des valeurs envisagées. D'où :

- troisième temps : **traitement des données**.

On cherche à **résumer** la série de valeurs au moyen d'un petit nombre d'**indicateurs statistiques appropriés**, qui refléteront au mieux les caractéristiques principales de la série.

C'est dans ce but que nous étudierons dans la suite ces indicateurs, tels la moyenne, la médiane ou encore l'écart-type.

**Remarque importante** :: dès maintenant, il faut souligner que le fait de résumer par un seul (ou quelques) nombre une série de valeurs représente un **compromis** entre une manipulation facilitée des données (et leur interprétation) d'une part et l'exhaustivité des données brutes d'autre part. Il résulte de cet écart une perte d'information qu'il convient de limiter au mieux.

**Exemple de perte d'information** : considérons deux étudiants dont les profils de notes (résultant de trois interrogations successives dans une même matière) sont les suivants :

étudiant A : 7, 10 et 13 étudiant B : 13, 10 et 7

S'il s'agit de faire passer les étudiants dans l'année supérieure, le critère de la moyenne (égale à 10 dans les deux cas) permet de résumer chacune de ces deux séries et allège les manipulations de données, sachant que la valeur représentée par une moyenne est "parlante". À l'inverse, l'indicateur moyenne ne permet plus ici de percevoir la **qualité** de la progression de l'étudiant sur l'ensemble d'un semestre. Ce "résumé" de la série des trois valeurs représente, du même coup, une **perte qualitative d'information** qui peut s'avérer par ailleurs déterminante.

Ce que nous venons de décrire brièvement ici constitue les trois étapes principales de ce que l'on appelle la **statistique descriptive** :

1ère étape : collecte de données → données brutes ("en vrac").

2ème étape : on ordonne, on classe les données. En général aussi, à ce stade, on réalise des **représentations graphiques**.

3ème étape : on résume les données au moyen d'indicateurs numériques, en vue de caractériser l'ensemble des valeurs initiales, selon un éclairage particulier.

A partir de ce qui précède, on peut proposer la **définition** suivante **de la statistique descriptive** :

**La statistique descriptive est un ensemble de méthodes qui permettent d'ordonner des observations statistiques, de les représenter graphiquement et de les résumer par des caractéristiques diverses.**

## 2. Concepts de base de la statistique descriptive

### 21. Population statistique et individu statistique

Historiquement, la démographie représente le premier domaine d'application de la statistique. C'est la raison pour laquelle apparaît ce terme de population.

**Définition d'une population statistique** : la population statistique est l'ensemble fini des éléments qu'on se propose d'étudier.

Le terme population doit être pris au sens large, c'est-à-dire pas seulement des êtres humains ou des êtres vivants, mais aussi des ensembles d'objets ou d'évènements factuels.

Exemples :

- ensemble d'êtres humains : la population de la France ou le personnel d'une entreprise à une date donnée ;

- un **stock** d'objets à une date donnée : un ensemble de pots de peinture en stock dans l'entreprise Machin au 31 janvier de l'année t ;

- un **flux** d'objets sur une période donnée : les demandes d'emploi déposées à Pôle Emploi au cours du mois de septembre de l'année t ;

- autres ensembles : le nombre d'accidents de la route ayant eu lieu en Rhône-Alpes durant une année.

**Définition de l'individu statistique** : un individu statistique (ou unité statistique) représente un élément de la population considérée.

Exemples : si l'on reprend les différents exemples de populations vues ci-dessus, on aura : une personne résidant en France ; un employé de l'entreprise ; un pot de peinture du stock de l'entreprise Machin ; - une demande d'emploi déposée à Pôle Emploi ; - un accident de la route ayant eu lieu en Rhône-Alpes durant l'année considérée.

Remarque importante : la définition d'une population statistique n'est pas toujours évidente.

Par exemple, la population d'une agglomération donnée, mesurée en nombre d'habitants (individus ou unités statistiques) peut sembler a priori quelque chose de facile à déterminer.

En réalité, on peut légitimement se poser la question de savoir si l'on doit ou non inclure dans cette population, à un moment donné, les élèves internes des établissements scolaires ou les étudiants en cité universitaire ou ayant un logement sur Lyon pour l'année universitaire.

En effet, il s'agit-là de populations transitoires, dont les individus retournent chez eux (hors de l'agglomération) les fins de semaine ou lors des périodes de vacances.

Or, si l'on s'intéresse, par exemple, aux comportements migratoires (déplacements) de la population envisagée, durant la semaine ou en fin de semaine, il est clair que la définition de la population initiale sur laquelle on va travailler n'est pas neutre, par rapport au sujet d'étude.

### 22. Les caractères statistiques

**Définition :** le **caractère** désigne la façon dont l'observation des individus d'une population est réalisée.

Chaque individu d'une population peut, en général, être décrit selon un ou plusieurs caractères.

Exemples : considérons la population d'un amphi ou d'un groupe de TD. On peut l'étudier selon : le sexe, l'âge, la taille, le poids, la nationalité, la série de baccalauréat, la commune de résidence des étudiants, etc.

Remarque : si l'on envisage l'étude de cette population en considérant un seul caractère à la fois, on aura affaire à des **séries statistiques à un caractère (à une dimension)**. Si l'on envisage l'étude de cette population en considérant 2 caractères simultanément, on aura affaire à des **séries statistiques à 2 caractères (à 2 dimensions)**. Enfin, si l'on envisage d'étudier cette population en considérant **n** caractères à la fois, alors on aura affaire à des **séries statistiques à n dimensions**. Dans ce dernier cas, les analyses renvoient aux **statistiques multi-dimensionnelles**, qui forment le fond des **analyses de données**.

Autre exemple : considérons la production mensuelle d'un constructeur automobile. Cette production représente la population étudiée. Dès lors, on peut envisager cette population selon divers caractères : - la couleur du véhicule produit, - la puissance du moteur, - le nombre de places des véhicules produits, - le type de véhicules produits (poids lourd, véhicule utilitaire léger, voiture / véhicule à essence, véhicule diesel), etc.

**Remarque importante : pour caractériser une population de manière appropriée, il faut s'attacher à considérer les caractères les plus pertinents et les plus discriminants.**

On peut retenir que "trop d'information tue l'information" : si l'on prend en compte trop de caractères, on reste près des données brutes, en disposant d'une information maximale, mais il est alors difficile de synthétiser et de mettre en évidence les grandes tendances de la série étudiée (éléments permanents et invariants caractéristiques d'une série donnée).

### 23. Les modalités

**Définition** : chaque caractère peut présenter plusieurs états, qu'on appelle des **modalités.**

Par exemple, si l'on s'intéresse à l'état matrimonial d'une personne, on peut retenir, par exemple, les deux classifications suivantes : - un schéma à 2 modalités : marié / non marié ; - un schéma à 5 modalités : célibataire, marié, veuf, divorcé, non déclaré (non réponse). Mais, dans le cas du premier schéma, le classement des veufs est ambigu. Cet exemple montre que l'on doit choisir soigneusement les modalités d'un caractère.

**De manière générale, les modalités relatives à un même caractère doivent être :**

**- incompatibles : un même individu ne peut présenter qu'une seule modalité, parmi toutes celles qui sont proposées,**

**- exhaustives : toutes les situations possibles doivent être prévus (un individu possède toujours l'une (et une seule) des modalités proposées),**

**- non ambiguës : pour éviter d'entraîner des erreurs de classement, les libellés des questions (qualitatives) dans les enquêtes doivent être très précis et ne pas faire l'objet de recouvrements possibles** (cf. schéma à deux modalités ci-dessus).

### 24. Classification des caractères statistiques

Si l'on reprend les exemples précédents, on s'aperçoit que les différents caractères évoqués ne peuvent pas tous être traités de la même façon. Il convient en effet de distinguer trois types de caractères :

- les caractères qualitatifs,
- les caractères quantitatifs discrets,
- les caractères quantitatifs continus.

#### 241. Les caractères qualitatifs

**Les modalités des caractères qualitatifs ne sont pas mesurables.**

Exemples : le sexe, la couleur des yeux, une série de baccalauréat, un état matrimonial, la commune de résidence, etc.

On peut distinguer :

- les caractères qualitatifs **ordinaux**, pour lesquels les états correspondent à un rang (hiérarchie). Ex. : un degré de satisfaction (faible, moyen, fort), un niveau de diplôme (bepc, bac, licence, master, doctorat), etc.

- les caractères qualitatifs **nominaux**, pour lesquels l'ordre des modalités est conventionnel (exemple les nomenclatures INSEE)..

Dans ce dernier cas, les modalités sont ordonnées, **par conventions particulières**, à l'intérieur de listes (tableaux), qu'on appelle des **nomenclatures**. Ces nomenclatures doivent présenter des modalités exhaustives, incompatibles et non ambiguës.

Exemple des nomenclatures INSEE : la nomenclature des indices mensuels de prix à la consommation, la nomenclature des activités économiques et des produits (NAP), la nomenclature des professions et catégories socioprofessionnelles (PCS). Ces différentes nomenclatures sont hiérarchisées selon un degré plus ou moins fort d'agrégation.

En ce qui concerne la nomenclature des professions et catégories socioprofessionnelles par exemple, les présentations les plus agrégées sont en 8, 24 ou 42 postes. La ventilation maximale est de 455 postes (voir page suivante).

Remarque importante :

souvent, on repère les modalités d'un caractère qualitatif par des valeurs numériques.

Par exemple, dans le tableau de la page suivante, si l'on considère les seules rubriques indiquées en caractères gras, on peut leur attribuer un numéro (1 = agriculteurs exploitants ; 2 = artisans, commerçants, chefs d'entreprise ; etc.).

Mais il faut clairement comprendre que ces valeurs numériques sont ici appliquées, **par commodité et de manière purement conventionnelle**, à des modalités qui restent néanmoins qualitatives.

France Tef 2002 / 03

## Répartition de la population de 15 ans et plus selon la catégorie sociale

| Catégorie socioprofessionnelle (PCS) | 1990 | | 1999 | | Évolution 90/82 | Évolution 99/90 |
|---|---|---|---|---|---|---|
| | Total milliers | Part des femmes % | Total milliers | Part des femmes % | % | % |
| Agriculteurs sur petite exploitation | 338 | 41,3 | 106 | 40,4 | – 51,0 | – 68,5 |
| Agriculteurs sur moyenne exploitation | 310 | 35,5 | 165 | 29,8 | – 31,3 | – 46,9 |
| Agriculteurs sur grande exploitation | 365 | 34,3 | 371 | 30,2 | 10,7 | 1,7 |
| **Agriculteurs exploitants** | **1 013** | **37,0** | **642** | **31,8** | **– 31,1** | **– 36,6** |
| Artisans | 850 | 23,5 | 761 | 23,1 | – 5,3 | – 10,5 |
| Commerçants et assimilés | 796 | 45,1 | 725 | 38,9 | – 0,2 | – 8,9 |
| Chefs d'entreprise de 10 salariés ou plus | 177 | 15,8 | 173 | 15,7 | 31,5 | – 2,1 |
| **Artisans, commerçants, chefs d'entreprise** | **1 823** | **32,2** | **1 659** | **29,2** | **– 0,4** | **– 9,0** |
| Professions libérales | 311 | 31,9 | 355 | 34,3 | 30,6 | 14,2 |
| Cadres de la Fonction publique | 288 | 28,1 | 373 | 35,7 | 19,6 | 29,6 |
| Professeurs, professions scientifiques | 564 | 50,4 | 670 | 52,6 | 58,8 | 18,6 |
| Profession de l'information, des arts et spectacles | 170 | 42,7 | 234 | 44,0 | 49,2 | 37,9 |
| Cadres admin. et commerciaux d'entreprise | 759 | 30,2 | 806 | 35,0 | 33,3 | 6,2 |
| Ingénieurs, cadres techniques d'entreprise | 601 | 11,2 | 727 | 15,1 | 58,0 | 21,0 |
| **Cadres, professions intellectuelles supérieures** | **2 693** | **30,9** | **3 165** | **34,8** | **41,9** | **17,5** |
| Instituteurs et assimilés | 757 | 65,1 | 903 | 65,2 | – 2,7 | 19,2 |
| Profess. interm. santé et travail social | 784 | 76,6 | 1 032 | 76,4 | 27,6 | 31,7 |
| Clergé, religieux | 48 | 41,8 | 21 | 22,1 | – 19,3 | – 57,2 |
| Profess. interm. admin. Fonction publique | 396 | 50,3 | 444 | 57,1 | 42,2 | 12,2 |
| Profess. interm. administratives des entreprises | 1 392 | 46,3 | 1 843 | 51,5 | 43,8 | 32,4 |
| Techniciens | 762 | 12,8 | 944 | 14,4 | 12,9 | 23,8 |
| Contremaîtres, agents de maîtrise | 574 | 7,5 | 576 | 8,9 | 1,7 | 0,3 |
| **Professions intermédiaires** | **4 714** | **44,5** | **5 763** | **48,1** | **19,7** | **22,3** |
| Employés civils, agents de service de la Fonction publique | 1 998 | 79,5 | 2 343 | 77,7 | 18,3 | 17,3 |
| Policiers et militaires | 414 | 6,7 | 523 | 9,1 | 7,4 | 26,3 |
| Employés administratifs d'entreprise | 2 344 | 84,7 | 2 178 | 84,6 | – 8,3 | – 7,1 |
| Employés de commerce | 969 | 81,0 | 1 115 | 77,8 | 31,4 | 15,1 |
| Personnels des services directs aux particuliers | 1 189 | 83,6 | 1 649 | 84,7 | 34,3 | 38,7 |
| **Employés** | **6 913** | **77,8** | **7 809** | **76,5** | **10,6** | **13,0** |
| Ouvriers qualifiés de type industriel | 1 640 | 13,6 | 1 540 | 14,6 | 2,8 | – 6,1 |
| Ouvriers qualifiés de type artisanal | 1 603 | 8,3 | 1 574 | 9,6 | 7,8 | – 1,8 |
| Chauffeurs | 622 | 3,0 | 640 | 5,0 | 9,8 | 2,8 |
| Ouvriers qualifiés manutention,magasinage, transport | 409 | 8,3 | 400 | 8,2 | 4,8 | – 2,2 |
| Ouvriers non qualifiés de type industriel | 2 136 | 39,6 | 1 724 | 36,0 | – 10,9 | – 19,3 |
| Ouvriers non qualifiés de type artisanal | 932 | 30,6 | 887 | 32,2 | – 12,6 | – 4,7 |
| Ouvriers agricoles | 282 | 22,9 | 297 | 27,8 | – 4,3 | 5,4 |
| **Ouvriers (y compris agricoles)** | **7 623** | **21,0** | **7 062** | **20,2** | **– 2,2** | **– 7,4** |
| Anciens agriculteurs exploitants | 1 273 | 53,7 | 1 080 | 52,3 | – 4,4 | – 15,2 |
| Anciens artis. commerç., chefs d'entrep. | 773 | 50,9 | 946 | 45,6 | 4,4 | 22,3 |
| Anciens cadres | 578 | 22,6 | 743 | 21,4 | 70,0 | 28,5 |
| Anciennes professions intermédiaires | 1 036 | 45,8 | 1 507 | 48,9 | 42,2 | 45,6 |
| Anciens employés | 2 675 | 70,2 | 3 204 | 73,3 | 41,8 | 19,8 |
| Anciens ouvriers (y compris agricoles) | 2 886 | 35,7 | 3 155 | 37,7 | 21,5 | 9,3 |
| **Retraités** | **9 221** | **49,8** | **10 634** | **51,1** | **24,6** | **15,3** |
| Chômeurs n'ayant jamais travaillé | 276 | 65,1 | 352 | 58,0 | – 24,0 | 27,2 |
| Militaires du contingent | 231 | 0,6 | 86 | 1,9 | – 8,7 | – 63,0 |
| Élèves ou étudiants de 15 ans ou plus | 5 261 | 50,4 | 5 433 | 51,0 | 23,1 | 3,3 |
| Autres inactifs de moins de 60 ans | 4 124 | 87,6 | 3 518 | 80,4 | – 25,9 | – 14,7 |
| Autres inactifs de 60 ans ou plus | 1 982 | 96,1 | 1 945 | 92,2 | – 1,5 | – 1,9 |
| **Autres sans activité professionnelle** | **11 875** | **70,4** | **11 334** | **67,0** | **– 4,8** | **– 4,6** |
| **Effectif total** | **45 875** | **51,9** | **48 068** | **52,0** | **6,5** | **4,8** |

### 242. Les caractères quantitatifs

**Leurs modalités sont mesurables. À chaque modalité est attaché un nombre**. Les caractères quantitatifs prennent aussi le nom de **variables statistiques**. Leurs différentes modalités sont les valeurs possibles de la variable statistique.

Il convient de distinguer à nouveau, d'une part les **caractères quantitatifs discrets**, d'autre part les **caractères quantitatifs continus**.

#### 2421. Les caractères quantitatifs discrets

**Ceux-ci ne peuvent prendre que des valeurs isolées (ponctuelles), qui sont, le plus souvent, des nombres entiers.**

Exemples : - nombre d'enfants d'un ménage, - nombre de pièces d'un logement, - nombre de salariés d'une entreprise.

**Familles selon le nombre d'enfants***

France Tef 2002 / 03 — milliers

| | 0 | 1 | 2 | 3 | 4 ou plus | Ensemble |
|---|---|---|---|---|---|---|
| 1962 | 4 229 | 2 798 | 2 085 | 1 141 | 1 069 | **11 322** |
| 1968 | 4 346 | 2 967 | 2 333 | 1 256 | 1 161 | **12 063** |
| 1975 | 4 876 | 3 333 | 2 665 | 1 293 | 1 009 | **13 176** |
| 1982 | 5 420 | 3 548 | 3 118 | 1 325 | 708 | **14 119** |
| 1990 | 6 484 | 3 667 | 3 345 | 1 349 | 546 | **15 391** |
| 1999 | 7 492 | 3 617 | 3 255 | 1 268 | 465 | **16 097** |

* Enfants célibataires de moins de 25 ans. Source : recensements de 1962 à 1999.

Remarque : on peut traiter des variables discrètes en continu, par regroupement de leurs valeurs en classes (voir point suivant 2422.).

## 2422. Les caractères quantitatifs continus

**Un caractère quantitatif est continu s'il peut prendre toutes les** (un ensemble de) **valeurs possibles à l'intérieur d'un intervalle donné, appelé classe.**

Exemples : taille, poids, âge, chiffre d'affaires, salaire, etc.

Tera 1997 / 98

**Population active totale selon le sexe et l'âge en 1990**

(en milliers)

| | Ain | Ardèche | Drôme | Isère | Loire | Rhône | Savoie | Haute-Savoie | Rhône-Alpes | France |
|---|---|---|---|---|---|---|---|---|---|---|
| **Hommes actifs** | **127,3** | **67,1** | **103,6** | **262,8** | **179,1** | **388,7** | **92,1** | **160,8** | **1 381,5** | **14 235,9** |
| de 15 à 24 ans | 16,6 | 8,4 | 13,1 | 32,5 | 23,5 | 47,7 | 11,6 | 20,4 | 173,8 | **1 784,9** |
| de 25 à 34 ans | 35,2 | 18,3 | 28,7 | 74,8 | 48,7 | 112,2 | 26,9 | 46,9 | 391,6 | **4 077,6** |
| de 35 à 49 ans | 51,9 | 27,2 | 41,5 | 106,6 | 73,2 | 150,5 | 36,8 | 64,3 | 552,1 | **5 656,1** |
| de 50 à 59 ans | 20,2 | 11,6 | 17,5 | 42,9 | 30,1 | 67,4 | 14,5 | 24,6 | 228,7 | **2 331,9** |
| 60 ans ou plus | 3,4 | 1,6 | 2,7 | 6,0 | 3,6 | 10,9 | 2,4 | 4,7 | 35,3 | **385,5** |
| **Femmes actives** | **93,3** | **49,6** | **77,7** | **198,2** | **140,2** | **315,6** | **68,9** | **121,8** | **1 065,4** | **11 106,0** |
| de 15 à 24 ans | 12,4 | 6,3 | 9,9 | 24,9 | 18,2 | 40,8 | 9,0 | 16,4 | 138,0 | **1 461,6** |
| de 25 à 34 ans | 27,3 | 14,4 | 22,8 | 60,5 | 40,3 | 97,3 | 21,1 | 37,7 | 321,3 | **3 382,6** |
| de 35 à 49 ans | 37,7 | 19,8 | 31,0 | 80,1 | 55,4 | 120,4 | 27,0 | 48,2 | 419,6 | **4 296,1** |
| de 50 à 59 ans | 13,4 | 7,4 | 11,7 | 27,6 | 22,6 | 47,8 | 9,7 | 16,2 | 156,4 | **1 629,9** |
| 60 ans ou plus | 2,5 | 1,5 | 2,4 | 5,1 | 3,7 | 9,3 | 2,1 | 3,4 | 30,1 | **335,8** |

Remarque : la frontière entre discret et continu est cependant quelque peu conventionnelle dans certains cas. Exemple : **par nature, une durée est une variable quantitative continue**. En effet, à l'intérieur de l'intervalle 10 h - 10 h 15, le temps peut prendre toutes les valeurs possibles (infinité de valeurs). Toutefois, on peut poser l'hypothèse que, compte tenu d'une mesure ponctuelle, grossière ou **arrondie** du temps, on raisonne en minutes révolues, ou en heures révolues, ou encore en années révolues (pour l'âge par exemple). Dans un tel cas, à la limite et de manière conventionnelle, on peut considérer la variable quantitative correspondante comme étant du type discret et la traiter en tant que telle. Il en est de même pour un salaire ou un chiffre d'affaires, si l'on considère des valeurs arrondies à l'euro près.

**De manière générale cependant, on retiendra que lorsqu'on peut utiliser des décimales, cela signifie que la variable quantitative est continue.**

**De même, on considérera comme continue une variable qui peut prendre un si grand nombre de valeurs, qu'il devient nécessaire de les regrouper en classes (cas de variables discrètes traitées en continu).**

Exemple : supposons une épreuve d'examen, notée sur 100 points, au quart de point près. Nous avons affaire ici à un caractère quantitatif discret, dont le nombre de modalités est égal à 401 (valeurs entières des notes, plus celles qui correspondent aux quarts, aux demis et aux trois-quarts de points). Un traitement statistique s'avère difficile en l'état. De plus, il est peu intéressant car l'information est trop abondante et colle trop aux données brutes. Cela empêche de faire clairement apparaître les grandes tendances de la série étudiée. C'est pourquoi, dans un tel cas, on effectue généralement des regroupements en classes et, in fine, on traite la variable statistique en question en continu.

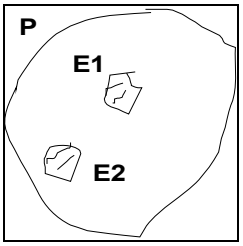
## 3. Place de la statistique descriptive dans l'ensemble de la démarche statistique

En général, pour des raisons de temps (pour rassembler l'information) et / ou de coût économique de l'opération de collecte et / ou de traitement des données, on ne travaille que sur des échantillons, et non pas sur la population envisagée tout entière.

Par suite, les résultats trouvés ne concernent qu'un (ou quelques) échantillon, et non pas la population tout entière. Or, ce qui est intéressant, dans une analyse économique (notamment en matière de prévision et de modélisation), c'est de pouvoir disposer de paramètres valant pour l'ensemble d'une population.

Le problème qui se pose alors est le suivant : comment fournir une approximation valable des paramètres recherchés, au niveau de l'ensemble de la population, à partir des résultats apportés par l'étude de l'échantillon ?

La réponse à cette question nécessite de mettre en oeuvre le calcul des probabilités.




L'analyse combinatoire permet de savoir de combien de façons on peut tirer d'échantillons (de même taille ou non) $E_i$ dans une même population P.

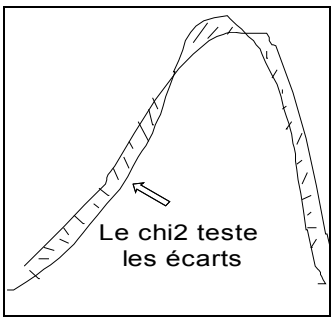
Sur chacun de ces échantillons, on peut calculer des paramètres (notamment la moyenne et l'écart-type), dont on comparera ensuite les valeurs avec celles obtenues pour l'ensemble de la population P. Ensuite, on en déduit les similitudes et / ou un biais entre les valeurs obtenues pour les $E_i$ et P.

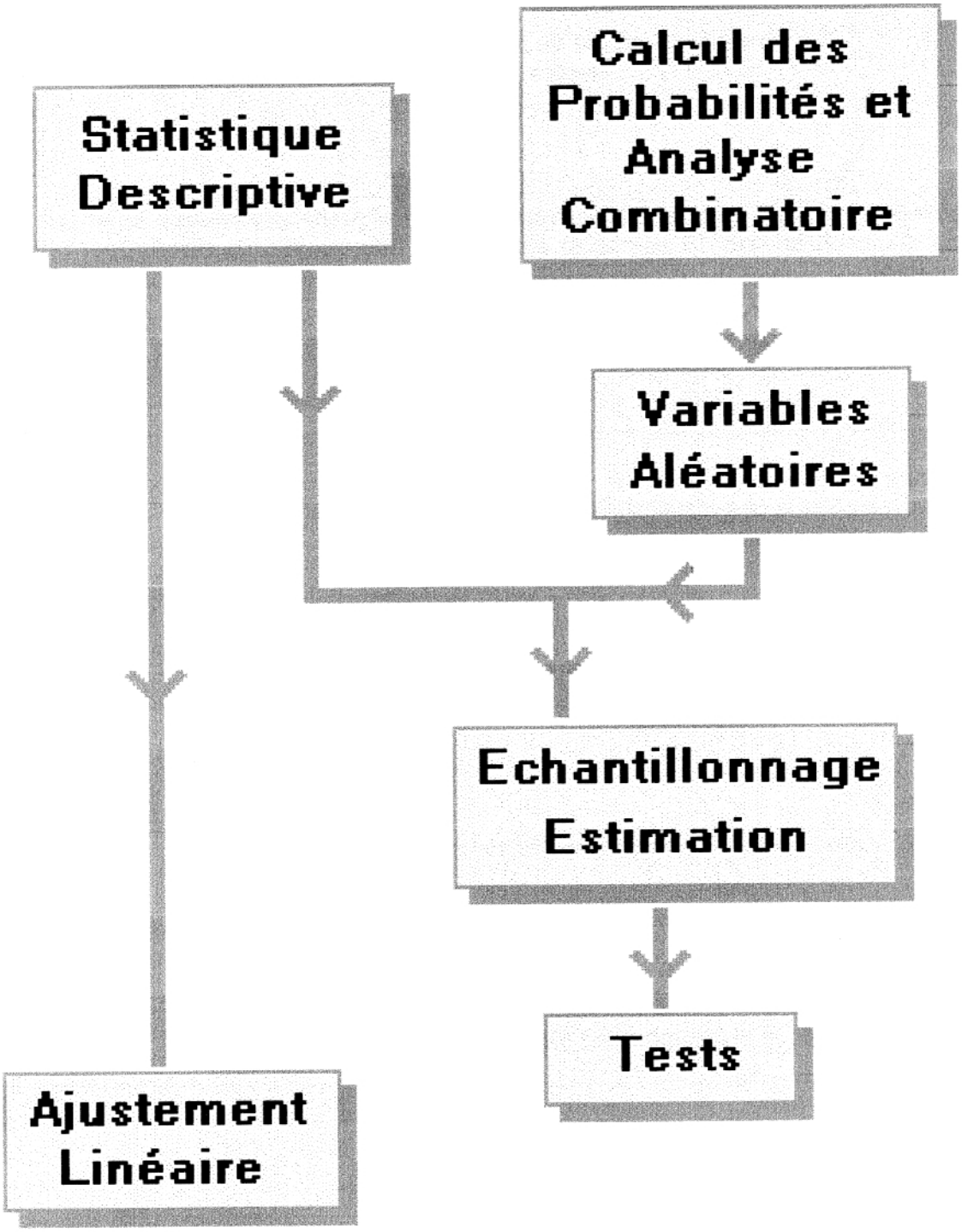
Plus généralement, le calcul de des probabilités, enrichi de la notion de variable aléatoire, permet aux statisticiens de déterminer des distributions "connues" (= distributions théoriques mathématisées).

On compare alors (test du $\chi^2$) les distributions des échantillons (avec leur propres paramètres) avec les distributions théoriques qui utilisent les mêmes paramètres et les mêmes valeurs.

Des tests statistiques permettent alors de savoir si les écarts observés entre les distributions empiriques et théoriques sont ou non imputables au hasard.

Cela permet de généraliser (ou non) les résultats obtenus sur un échantillon à l'ensemble d'une population.




Statistique Descriptive
Calcul des Probabilités et Analyse Combinatoire
Variables Aléatoires
Echantillonnage Estimation
Tests
Ajustement Linéaire

# PARTIE 1 : ÉTUDE DES SÉRIES STATISTIQUES À UN CARACTÈRE

## CHAPITRE 1 : PRÉSENTATION DES DONNÉES ET REPRÉSENTATIONS GRAPHIQUES

Remarque liminaire :

On peut envisager deux types d'approches, en vue de présenter des données statistiques :

a) Si l'on réalise une enquête, on dispose initialement de données brutes (c'est-à-dire "en vrac", selon l'ordre dans lequel on aura traité les questionnaires, effectué les mesures, etc.).

Dans ce cas, il est nécessaire, en tout premier lieu, d'ordonner les différentes valeurs de la série (de la plus petite à la plus grande, en général), puis de les classer dans un tableau récapitulatif. Il s'agit d'un premier résumé de la série, pour lequel on regroupe les effectifs partiels correspondant à chacune des modalités du caractère.

b) si l'on réalise une étude statistique sur des données existantes, celle-ci sont déjà ordonnées et classées dans des tableaux.

Dans ce cas (et dans le cas précédent, à partir de cette étape), il va s'agir de déterminer les éléments principaux qui caractérisent les séries étudiées, à savoir : la population statistique, l'individus statistique, le caractère, et le type du caractère (qualitatif, quantitatif discret ou continu).

La forme générale des tableaux statistiques (séries à un caractère) est la suivante :

| **Modalités du caractère ($x_i$)** | **Effectifs par modalité ($n_i$)** |
|---|---|
| $x_1$ | $n_1$ |
| $x_2$ | $n_2$ |
| … | … |
| $x_i$ | $n_i$ |
| … | … |
| $x_k$ | $n_k$ |
| | $\mathbf{n} = \sum_{i=1}^{k} n_i$ |

x représente le caractère ; $x_i$ représente de la modalité i du caractère x ; **n** représente **l'effectif total** de la population statistique étudiée (si l'étude est exhaustive) ou de l'échantillon statistique étudié (si l'étude n'est pas exhaustive) ; $n_i$ représente l'effectif partiel correspondant à la seule modalité i du caractère (on parle quelquefois de fréquences absolues).

Remarques :

- l'intitulé de la première colonne (modalités) reprend la donnée du terme général $x_i$.

- on suppose ici que le caractère x prend k modalités distinctes.

- la sommation est effectuée sur le nombre total de modalités :

$$n = n_1 + \ldots + n_i + \ldots + n_k = \sum_{i=1}^{k} n_i$$

On complète souvent le tableau précédent par une colonne supplémentaire, dans laquelle on fait apparaître les **fréquences relatives** ou, plus généralement en Économie, les pourcentages correspondant à chaque modalité du caractère x.

Cette procédure facilite les comparaisons entre tableaux relatifs au même caractère, mais portant sur des échantillons ou des populations différentes en taille.

| **Modalités du caractère ($x_i$)** | **Effectifs par modalité ($n_i$)** | **Fréquences relatives par modalité ($f_i$)** | **Pourcentages relatifs par modalité ($f_i$ x 100)** |
|---|---|---|---|
| $x_1$ | $n_1$ | $f_1$ | $f_1$ x 100 |
| $x_2$ | $n_2$ | $f_2$ | $f_2$ x 100 |
| … | … | … | … |
| $x_i$ | $n_i$ | $f_i$ | $f_i$ x 100 |
| … | … | … | … |
| $x_k$ | $n_k$ | $f_k$ | $f_k$ x 100 |
| | $n = \sum_{i=1}^{k} n_i$ | $1 = \sum_{i=1}^{k} f_i$ | $100 = \sum_{i=1}^{k} (f_i \times 100)$ |

Remarque : que l'on ait affaire à des caractères qualitatifs ou quantitatifs, la structure des tableaux de données correspond à celle qui est indiquée ici. Dans le cas d'un caractère quantitatif continu, on fait apparaître des classes de valeurs, dans la première colonne, afin de matérialiser les modalités du caractère.

Par contre, des différences importantes interviennent, au moment de construire des représentations graphiques adéquates, selon chaque type de caractère statistique. C'est ce que nous allons aborder dans les trois points suivants.

## 1. Les caractères qualitatifs

### 11. Présentation des données

La forme générale des tableaux est la suivante :

| **Modalités du caractère ($x_i$)** | **Effectifs par modalité ($n_i$)** | **Fréquences relatives par modalité ($f_i$)** | **Pourcentages relatifs par modalité ($f_i$ x 100)** |
|---|---|---|---|
| $x_1$ | $n_1$ | $f_1$ | $f_1$ x 100 |
| $x_2$ | $n_2$ | $f_2$ | $f_2$ x 100 |
| … | … | … | … |
| $x_i$ | $n_i$ | $f_i$ | $f_i$ x 100 |
| … | … | … | … |
| $x_k$ | $n_k$ | $f_k$ | $f_k$ x 100 |
| | $n = \sum_{i=1}^{k} n_i$ | $1 = \sum_{i=1}^{k} f_i$ | $100 = \sum_{i=1}^{k} (f_i \times 100)$ |

avec : $x_i$ = modalité i du caractère x ; k = nombre de modalités prises par le caractère x ; $n_i$ = effectif partiel qui correspond à $x_i$ ; n = effectif total ; $f_i = n_i / n$

Remarque : on a vu dans l'introduction que les modalités de ces caractères ne sont pas mesurables. Il en résulte que l'ordre de ces modalités est complètement conventionnel dans le cas des caractères qualitatifs nominaux. Dans le cas des caractères qualitatifs ordinaux, il existe une hiérarchie (par exemple : faible, moyen, fort).

En pratique, dans le cas des caractères qualitatifs nominaux, on classe les modalités selon un ordre croissant ou décroissant des effectifs (ou des fréquences relatives), de façon à ce qu'il n'y ait pas de confusion avec les représentations graphiques des caractères quantitatifs discrets (un diagramme en tuyaux d'orgue ne correspond pas du tout à la même réalité qu'un diagramme en bâtons).

Exemple :

**La population des départements de la région Rhône-Alpes au recensement de 1999.**

Le classement des modalités est établi selon l'ordre alphabétique (par convention).

| Départements | Population 1999 (en milliers d'hab.) |
|---|---|
| **Ain** | 515 |
| **Ardèche** | 286 |
| **Drôme** | 438 |
| **Isère** | 1 094 |
| **Loire** | 728 |
| **Rhône** | 1 579 |
| **Savoie** | 373 |
| **Hte-Savoie** | 632 |
| **Total** | 5 645 |

Source : TEF Insee.

Population statistique : les 5 645 000 habitants de la région Rhône-Alpes en 1999.
Individu statistique : l'un des 5 645 000 habitants.
Caractère statistique : les départements de la région Rhône-Alpes.
Type du caractère : qualitatif nominal.

Il est possible de ramener les effectifs de chaque département à des pourcentages relatifs par rapport à la population totale de la région Rhône-Alpes :

| Départements | Population 1999 (en milliers d'hab.) | Poids relatif de chaque département (en %) |
|---|---|---|
| **Ain** | 515 | 9,1 |
| **Ardèche** | 286 | 5,1 |
| **Drôme** | 438 | 7,8 |
| **Isère** | 1 094 | 19,4 |
| **Loire** | 728 | 12,9 |
| **Rhône** | 1 579 | 28,0 |
| **Savoie** | 373 | 6,6 |
| **Hte-Savoie** | 632 | 11,2 |
| **Total** | 5 645 | 100,0 |

On a : $f_i \times 100 = \frac{n_i}{n} \times 100$. Par exemple : $0{,}091 \times 100 = \frac{515}{5\,645} \times 100 = 9{,}1\%$

Remarques importantes :

a) dans un souci de cohérence, on homogénéise l'écriture des pourcentages. Si l'on arrondit les résultats à un chiffre après la virgule, il est nécessaire de faire éventuellement apparaître des zéros.

b) **pas de précision illusoire** : si une valeur décimale n'est pas significative, on arrondit à la décimale précédente, voire même à la valeur entière la plus proche.

Que l'on cherche à représenter les effectifs ou les fréquences relatives, trois types de représentations graphiques sont disponibles pour illustrer un **caractère qualitatif** de manière appropriée, c'est-à-dire **sans déformer la réalité des données** :

- les diagrammes en tuyaux d'orgue,
-les diagrammes en secteurs (ou camemberts),
-les cartogrammes.

Quel que soit le type de diagramme retenu, on utilise la première colonne du tableau précédent et l'une des deux autres colonnes.

Pour mémoire, signalons les diagrammes figuratifs, très utilisés dans la presse.

Si l'on veut ici représenter les effectifs relatifs à chaque département de manière non déformante, au moyen d'un diagramme figuratif, on doit poser une hypothèse.

Par exemple, on convient de dessiner une figurine dont la hauteur est proportionnelle à l'effectif d'un département donné. On peut retenir par exemple : 1 cm = 100 000 habitants. Selon cette hypothèse, on aurait ainsi :

Ardèche = 2,86 cm, Savoie = 3,73 cm, ..., jusqu'à Rhône = 15,79 cm.

On peut ainsi créer un graphique en abscisse duquel on trouve chacun des départements (classés dans un ordre croissant de leur population) et en ordonnée duquel on place, pour chaque département, le nombre de centimètres correspondant à la population de ce département.

En fait, ce type de représentation peut donner lieu à des interprétations erronées. En effet :

1) lorsque ce type de graphique est utilisé, il est rarement indiqué explicitement s'il s'agit de considérer la hauteur ou la surface (voire même un pseudo volume 3D) d'une figurine, à rapporter à l'effectif qu'elle est censée représenter.

2) même si l'on sait par exemple que la hauteur doit être proportionnelle à l'effectif correspondant, se pose un problème de proportions "esthétiques" de la figure. Celle-ci va également tendre à gagner en largeur (donc en surface !), au fur et à mesure où l'effectif (donc la hauteur) croît.

Il en résulte donc le plus souvent une déformation de la réalité que ne reflète pas la lecture directe du tableau de données correspondant.

Conclusion : retenir qu'il vaut mieux ne pas utiliser ce type de représentation graphique (ou, en tout cas, l'utiliser avec précaution).

## 12. Représentations graphiques appropriées

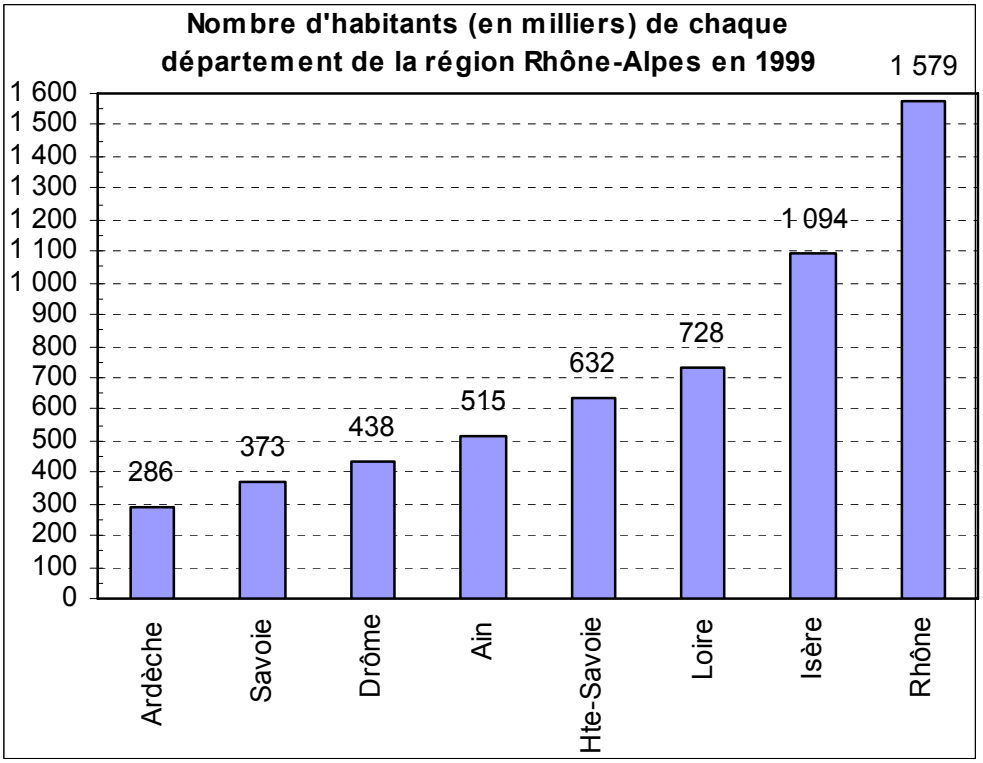
### 121. Les diagrammes en tuyaux d'orgue

Il s'agit d'une représentation non ambiguë des caractères qualitatifs. En effet, ce type de représentation est neutre (pas d'effet de déformation des données), car les bases de chaque rectangle (tuyau d'orgue) sont toutes identiques.

Il est conseillé de placer les modalités du caractère dans un ordre croissant ou décroissant, afin qu'il n'y ait pas de confusion avec les représentations graphiques des caractères quantitatifs discrets (un diagramme en tuyaux d'orgue ne correspond pas du tout à la même réalité qu'un diagramme en bâtons).

Sur notre exemple, on a (selon les effectifs) :

| | Population 1999 (en milliers) |
|---|---|
| **Ardèche** | 286 |
| **Savoie** | 373 |
| **Drôme** | 438 |
| **Ain** | 515 |
| **Hte-Savoie** | 632 |
| **Loire** | 728 |
| **Isère** | 1 094 |
| **Rhône** | 1 579 |



## 122. Les diagrammes en secteurs ("camemberts")

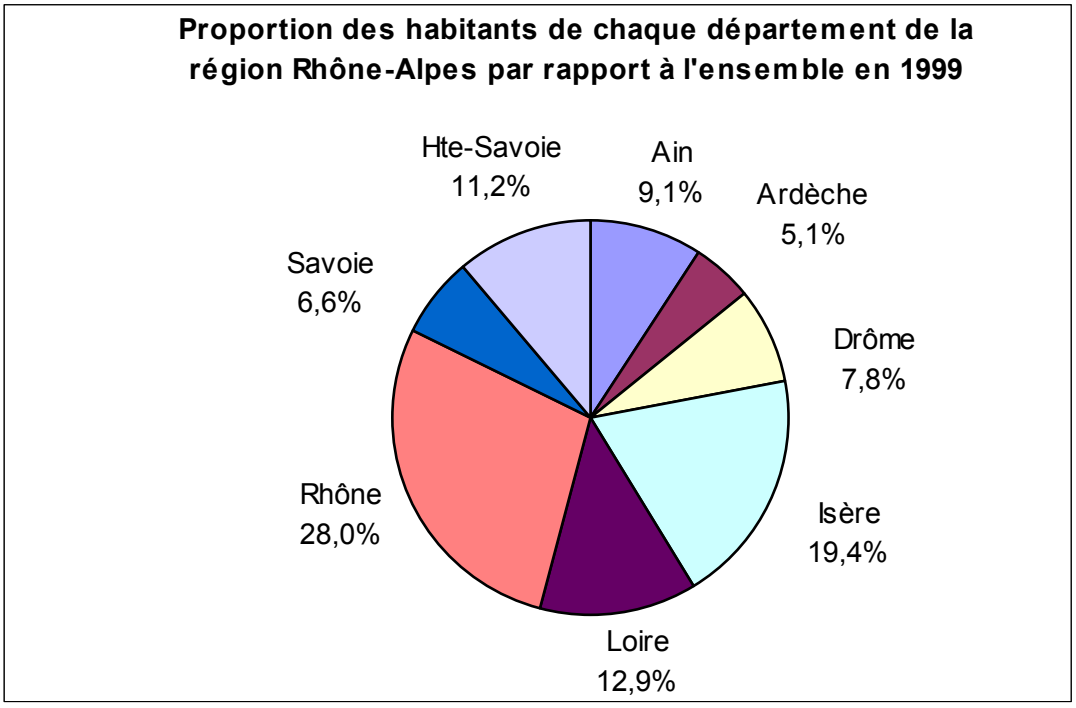
De même que les diagrammes en tuyaux d'orgue, les diagrammes en secteurs ne déforment pas la réalité des données. Ces diagrammes sont particulièrement bien adaptés à la représentation graphique des données exprimées en fréquences relatives ou en pourcentages.

Lorsqu'on travaille sur le papier, il est nécessaire, pour chaque modalité, de calculer l'angle au centre qui correspondant à un pourcentage donné. En effet, dans la colonne relative aux pourcentages, on raisonne sur un total de 100. Par ailleurs, une circonférence correspond à 360°. Il faut donc transformer chaque pourcentage en degrés d'angle, de façon à pouvoir tracer chaque secteur circulaire sur le graphique. **Si l'on travaille en pourcentages, il suffit de multiplier chaque pourcentage par 3,6 pour obtenir l'angle au centre correspondant, exprimé en degrés** (ici, on arrondit la valeur des angles au centre à l'entier le plus proche).

Remarque : pour un tracé rigoureux, il est nécessaire d'utiliser un compas et un rapporteur.

On obtient ainsi :

| Départements | Population 1999 (en milliers d'hab.) | Poids relatif de chaque département (en %) | Angle au centre (en degrés) |
|---|---|---|---|
| **Ain** | 515 | 9,1 | 33 |
| **Ardèche** | 286 | 5,1 | 18 |
| **Drôme** | 438 | 7,8 | 28 |
| **Isère** | 1 094 | 19,4 | 70 |
| **Loire** | 728 | 12,9 | 46 |
| **Rhône** | 1 579 | 28,0 | 101 |
| **Savoie** | 373 | 6,6 | 24 |
| **Hte-Savoie** | 632 | 11,2 | 40 |
| **Rh.-Alpes** | 5 645 | 100,0 | 360 |



### 123. Les cartogrammes

Ces représentations graphiques peuvent être utilisées quel que soit le type du caractère statistique envisagé (qualitatif, quantitatif discret ou quantitatif continu).

**Avantage** : un cartogramme permet d'introduire la **dimension spatiale des données**, qu'on ne peut rendre d'une autre façon.

**Inconvénient** : dans un certain nombre de cas, la réalité des données est faussée par ce type de représentation.

Par exemple si l'on travaille sur les départements français, ces derniers ont des superficies différentes, de même que des densités de population différentes. Selon le caractère considéré, ces différences déforment les données statistiques de manière plus ou moins prononcée (sur ou sous-évaluation visuelle de l'importance d'une modalité par rapport à une autre).

En ce qui concerne les caractères qualitatifs, on peut prendre l'exemple classique d'un référendum. Le lendemain du scrutin, la presse édite des cartes des départements français avec une couleur différente selon le bord politique majoritaire dans un département donné.

Cela permet de repérer, d'un seul coup d'oeil, les zones géographiques où la majorité des électeurs ont voté pour telle ou telle personne (ou telle ou telle couleur politique).

Remarque :

La France métropolitaine compte 96 départements : 89 départements + 2A, 2B (Corse) + les cinq départements de la région parisienne (91,92,93,94,95).

Certaines cartes font apparaître en plus les quatre départements d'outre-mer : Guadeloupe, Guyane, Martinique et Réunion.

D'autres ajoutent également les cinq territoires d'outre-mer : Mayotte, Nouvelle-Calédonie, Polynésie, Saint-Pierre-et-Miquelon, terres Australes et antarctiques.

Soit au total 105 circonscriptions administratives.

**ATTENTION !**

Dans l'exemple précédent, si l'on se réfère à la France métropolitaine par exemple, on a :

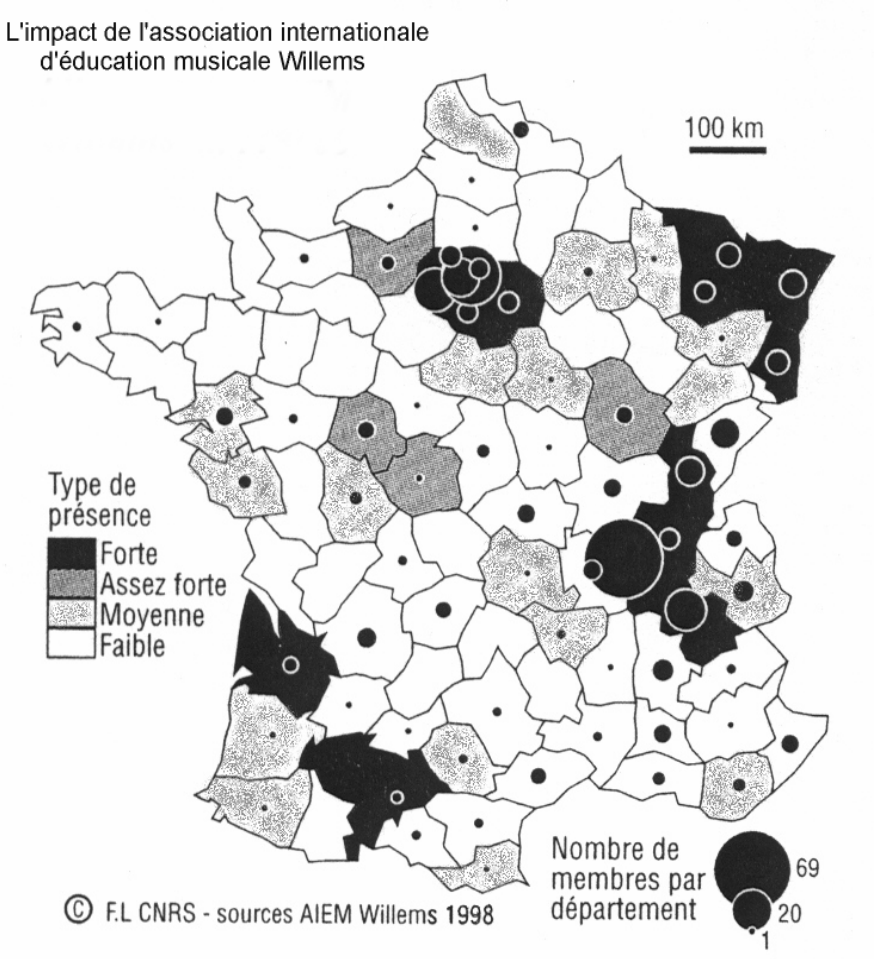
Population statistique : les 96 départements français métropolitains.

Individu statistique : l'un des 96 départements.

Caractère statistique : la couleur politique majoritaire d'un département.

Type du caractère : qualitatif nominal.

Autre exemple de cartogramme relatif à un caractère qualitatif :




Remarques terminales :

1) un **tableau de données** doit toujours faire apparaître des valeurs homogènes (même nombre de décimales) et être accompagné :

- d'une source (aussi détaillée que possible),
- d'une date (au moins l'année),
- des unités qui affectent les caractères et les populations statistiques,
- d'un titre.

2) un **graphique** doit, de plus, faire apparaître clairement les intitulés des axes en abscisse et en ordonnée, de même que des éléments complémentaires appropriés (par exemple, une unité d'aire, dans le cas d'un histogramme).

3) dans le cas des caractères qualitatifs, du fait que les modalités ne sont pas mesurables, il n'est pas possible de mener des calculs supplémentaires en vue de "résumer" les séries correspondantes.

De manière générale, le calcul d'une moyenne n'a pas de sens pratique concret. Si l'on reprend notre exemple, il ne serait pas impossible de calculer un effectif de population moyenne par département pour l'ensemble de la région Rhône-Alpes. Cependant, ce résultat n'aurait guère de signification de portée générale.

**Deuxième exemple relatif au traitement des caractères qualitatifs**

Considérons à nouveau les départements de la région Rhône-Alpes et envisageons cette fois-ci les superficies de ces départements. On peut établir le tableau suivant :

| Départements | Superficie (en km²) | Poids relatif de chaque département (en %) | Angle au centre (en degrés) |
|---|---|---|---|
| **Ain** | 5 762 | 13,2 | 47 |
| **Ardèche** | 5 529 | 12,7 | 46 |
| **Drôme** | 6 530 | 14,9 | 54 |
| **Isère** | 7 431 | 17,0 | 61 |
| **Loire** | 4 781 | 10,9 | 39 |
| **Rhône** | 3 249 | 7,4 | 27 |
| **Savoie** | 6 028 | 13,8 | 50 |
| **Hte-Savoie** | 4 388 | 10,0 | 36 |
| **Rh.-Alpes** | 43 698 | 100,0 | 360 |

**Diagramme en tuyaux d'orgue**

| | Superficie (en km²) |
|---|---|
| **Isère** | 7 431 |
| **Drôme** | 6 530 |
| **Savoie** | 6 028 |
| **Ain** | 5 762 |
| **Ardèche** | 5 529 |
| **Loire** | 4 781 |
| **Hte-Savoie** | 4 388 |
| **Rhône** | 3 249 |

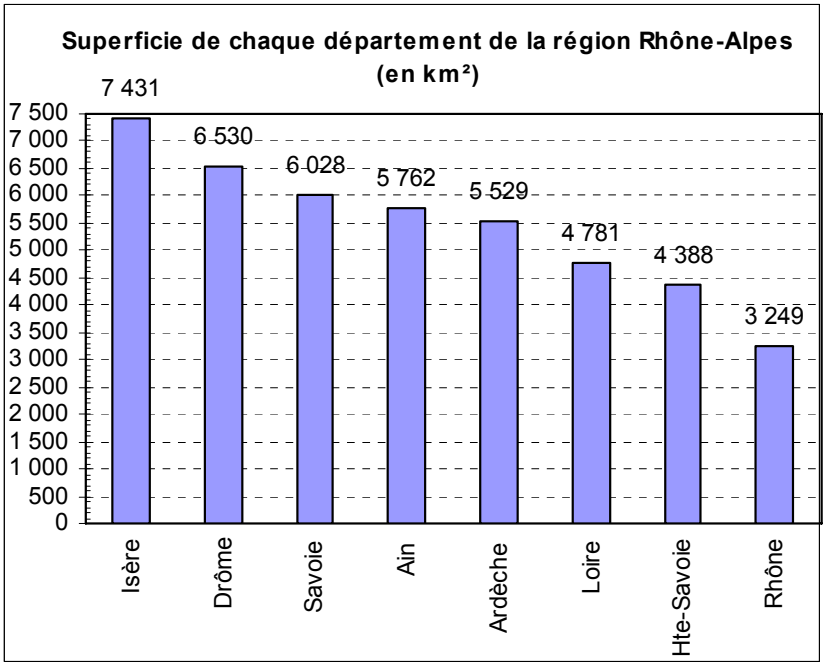
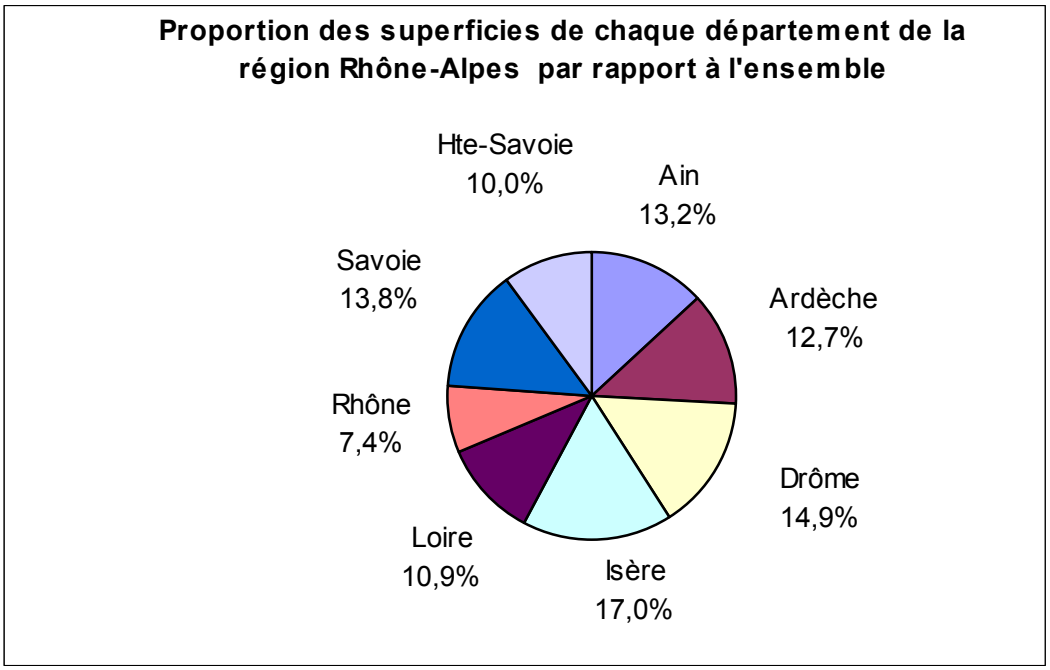


**Diagramme en secteurs**

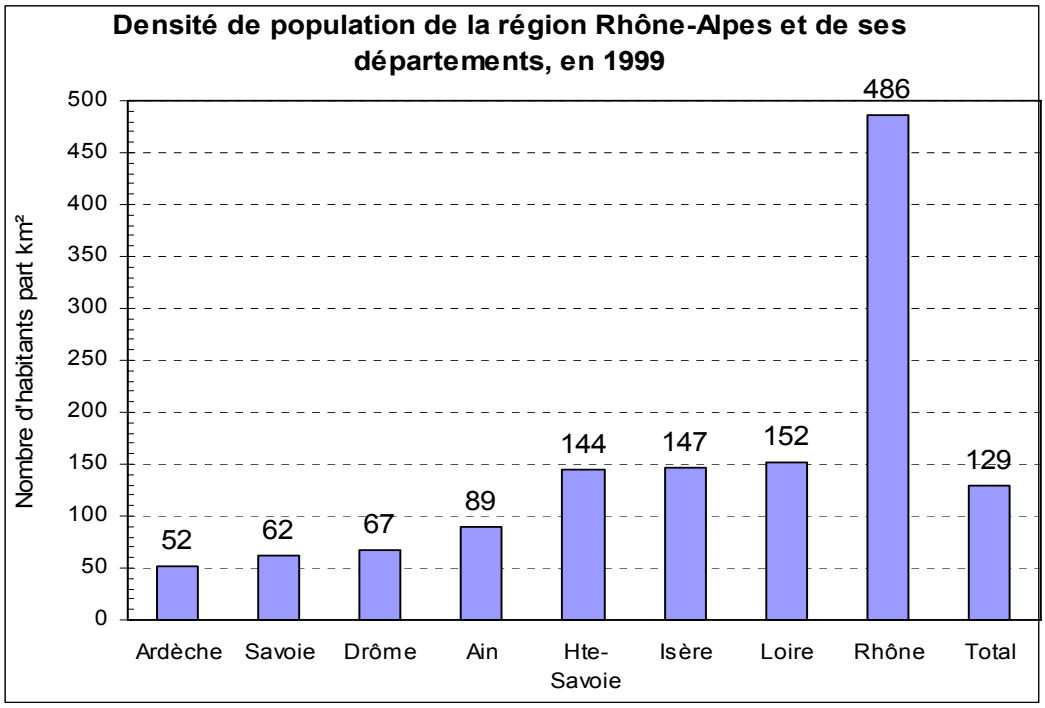



Compte tenu du fait de la connaissance simultanée de la population et de la superficie de chaque département, on peut en tirer des enseignements sur les **densités** (interprétation statistique et économique, en termes de dynamisme économique).

$$\text{Densité (en hab. / km²)} = \frac{\text{population (en nb d' hab.)}}{\text{superficie (en km²)}}$$

On constate que le plus petit des 8 départements de la région Rhône-Alpes (département du Rhône) est aussi le plus peuplé. Sa densité et de loin la plus élevée de la région Rhône-Alpes : 486 hab. / km².

| Départements | Population 1999 (en milliers d'hab.) | Superficie (en km²) | Densité (en hab. / km²) |
|---|---|---|---|
| **Ain** | 515 | 5 762 | 89 |
| **Ardèche** | 286 | 5 529 | 52 |
| **Drôme** | 438 | 6 530 | 67 |
| **Isère** | 1 094 | 7 431 | 147 |
| **Loire** | 728 | 4 781 | 152 |
| **Rhône** | 1 579 | 3 249 | 486 |
| **Savoie** | 373 | 6 028 | 62 |
| **Hte-Savoie** | 632 | 4 388 | 144 |
| **Rh.-Alpes** | 5 645 | 43 698 | 129 |




## 2. Les caractères quantitatifs discrets

Les modalités des caractères quantitatifs discrets sont mesurables (c'est-à-dire qu'elles s'expriment sous forme numérique). Les valeurs ne sont pas le résultat de conventions comme, par exemple, les numéros qu'on peut affecter aux rubriques des nomenclatures Insee.

Les valeurs prises par les modalités sont discrètes, c'est-à-dire discontinues, isolées. En général, il s'agit de valeurs entières.

### 21. Présentation des données

La forme générale des tableaux est la suivante :

| **Modalités caractère ($x_i$)** | **Effectifs ($n_i$)** | **Fréquences relatives ($f_i$)** | **Pourcentages relatifs ($f_i$ x 100)** |
|---|---|---|---|
| $x_1$ | $n_1$ | $f_1$ | $f_1$ x 100 |
| $x_2$ | $n_2$ | $f_2$ | $f_2$ x 100 |
| … | … | … | … |
| $x_i$ | $n_i$ | $f_i$ | $f_i$ x 100 |
| … | … | … | … |
| $x_k$ | $n_k$ | $f_k$ | $f_k$ x 100 |
| | $n = \sum_{i=1}^{k} n_i$ | $1 = \sum_{i=1}^{k} f_i$ | $100 = \sum_{i=1}^{k} (f_i \times 100)$ |

avec : $x_i$ = modalité i du caractère x ; k = nombre de modalités prises par le caractère x ; $n_i$ = effectif partiel qui correspond à $x_i$ ; n = effectif total ; $f_i = n_i / n$.

Remarque : pour réaliser un diagramme en bâtons (parfois appelé **diagramme différentiel**), on utilise la première colonne du tableau et l'une ou l'autre des trois colonnes suivantes.

Exemple

Une entreprise dispose de 153 machines. Un mois durant, le service entretien de l'entreprise note, tous les jours et pour chaque machine, le nombre de pannes (six modalités) :

| Nombre de pannes par machine, sur un mois (xi) | Nombre de machines concernées (ni) | Fréquences relatives (fi) (en %) |
|---|---|---|
| 0 | 69 | 45,1 |
| 1 | 41 | 26,8 |
| 2 | 19 | 12,4 |
| 3 | 13 | 8,5 |
| 4 | 8 | 5,2 |
| 5 | 3 | 2,0 |
| Totaux | 153 | 100,0 |

**Population statistique :** les 153 machines composant le parc machine de l'entreprise.
**Individu statistique :** l'une des 153 machines de ce parc.
**Caractère statistique :** le nombre de pannes, sur un mois.
**Type du caractère :** quantitatif discret.

### 22. Représentations graphiques appropriées

Nous présentons ici deux types de graphiques appropriés à la représentation des caractères quantitatifs discrets : les diagrammes en bâtons et les cartogrammes.

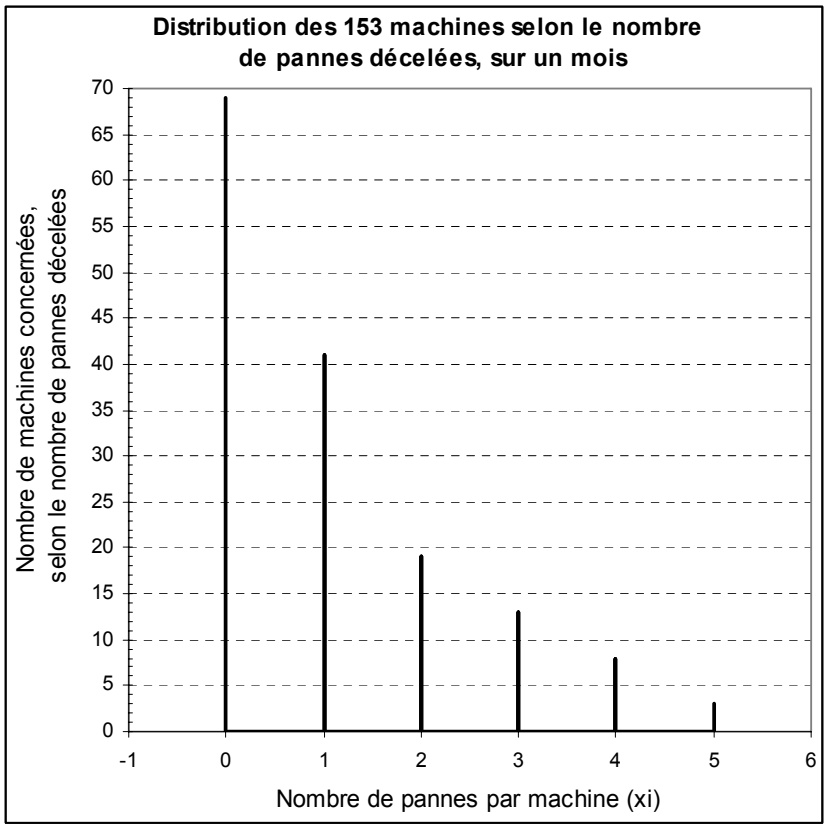
Remarques :

Il est également possible de représenter graphiquement, sans déformation, une série quantitative discrète au moyen d'un diagramme en secteurs. Toutefois, le diagramme en bâtons apporte des informations supplémentaires (dans le cas d'échantillons, l'allure du polygone des effectifs associé au diagramme en bâtons est utile, pour déterminer l'utilisation d'une loi de probabilité discrète adéquate, lorsqu'on cherche à extrapoler les résultats de calculs effectués sur un échantillon à la population correspondante tout entière).

De même, un diagramme cumulé des effectifs (ou des fréquences), que nous ne traiterons pas ici, peut être associé à une série quantitative discrète (fonction en escalier).

## 221. Les diagrammes en bâtons (diagrammes différentiels)

En reprenant notre exemple, nous obtenons (selon les effectifs) :




Le **diagramme différentiel** d'une série quantitative discrète s'appelle le **diagramme en bâtons**.

Ce diagramme retrace la **distribution** des effectifs (ou des fréquences relatives), selon les modalités du caractère étudié (ici, le nombre de pannes par machine).

Remarques :

1) si l'on raisonne selon les fréquences relatives ou les pourcentages (colonne de droite du tableau), l'allure du diagramme en bâtons est évidemment la même que lorsqu'on raisonne selon les effectifs.

2) ici, on constate que la hauteur des bâtons décroît régulièrement, selon les modalités croissantes du caractère x.

Contrairement aux caractères qualitatifs nominaux, pour lesquels l'ordre des tuyaux et conventionnels, il n'est évidemment pas question ici de permuter les bâtons, car **l'ordre du nombre de pannes 0,1, 2, 3, 4, 5 est signifiant**, en termes de mesure.

Dans le cas présent, cela signifie que pour la majorité des machines il y a heureusement pas ou peu de pannes.

3) lorsque le nombre de modalités d'une variable quantitative discrète devient trop grand, on regroupe les modalités (initialement discrètes) en classes et l'on traite les données correspondantes en continu.

### 222. Les cartogrammes

Nous avons déjà vu plus haut que ces représentations graphiques peuvent être utilisées quel que soit le type du caractère statistique envisagé (qualitatif, quantitatif discret, quantitatif continu).

**Avantage** : un cartogramme permet d'introduire la **dimension spatiale des données**, qu'on ne peut rendre d'une autre façon.

**Inconvénient** : dans un certain nombre de cas, la réalité des données est faussée par ce type de représentation.

Exemples :

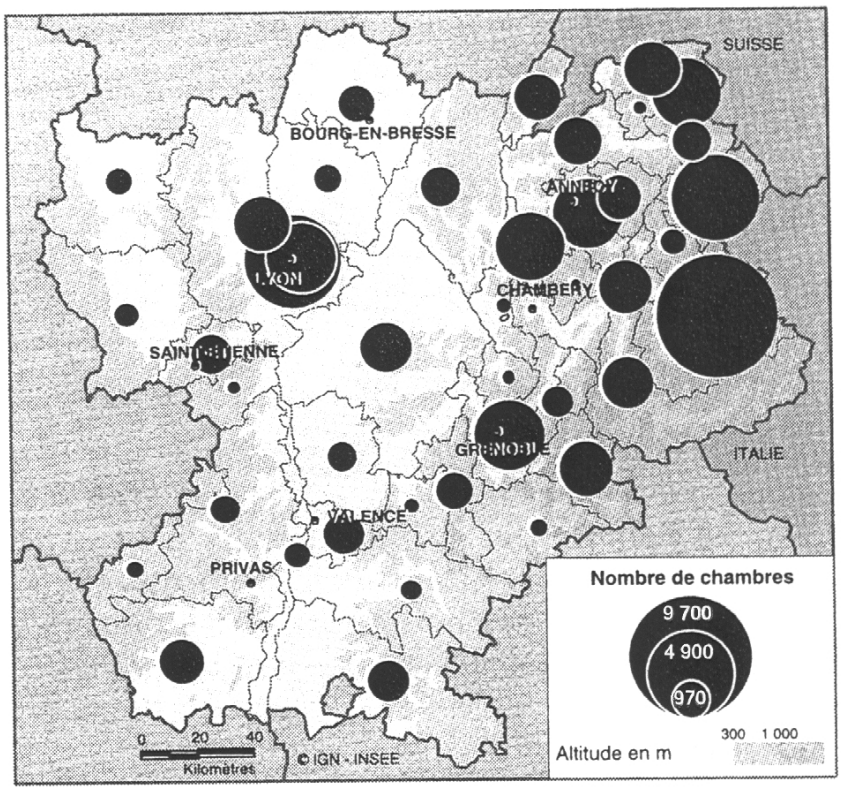
1) On peut envisager une étude portant, pour une année donnée, sur l'étude de la population des moins de 20 ans d'une agglomération donnée.

Considérons par exemple le Grand Lyon, qui est composé de 55 communes.

Population statistique : les 55 communes du Grand Lyon.
Individu statistique : l'une des 55 communes du Grand Lyon.
Caractère statistique : le nombre de jeunes de moins de 20 ans.
Type du caractère quantitatif discret (puisqu'il s'agit d'un nombre de personnes).

Sur un fond de carte des communes du Grand Lyon, on disposera, au niveau de chaque commune, des **cercles dont la surface sera rigoureusement proportionnelle aux effectifs**.

2) Parc hôtelier dans la région Rhône-Alpes, au 31 décembre 1996.




Parc hôtelier en région Rhône-Alpes, au 31 / 12 / 1996

Tera 1997 / 98

Ce cartogramme ne déforme pas la réalité des données si les cercles sont bien proportionnels aux effectifs.

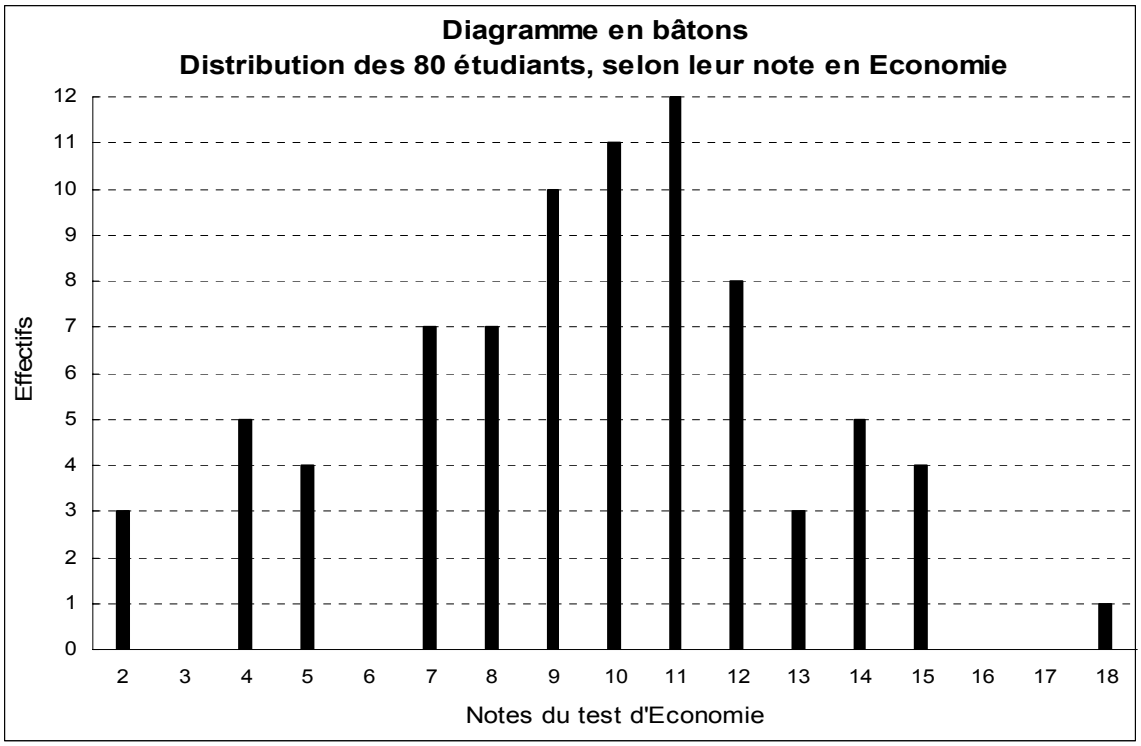
Population statistique : les arrondissements des départements de la Région Rhône-Alpes.
Individu statistique : l'un des arrondissements des départements de la Région Rhône-Alpes.
Caractère statistique : le nombre total de chambres des hôtels de chaque arrondissement.
Type du caractère quantitatif discret (puisqu'il s'agit d'un nombre de chambres).

**Deuxième exemple relatif au traitement des caractères quantitatifs discrets**

Les notes (valeurs entières) de 80 étudiants à un test d'Économie :

| **Notes obtenues (sur 20, val. entières) (xi)** | **Nombre d'étudiants (ni)** | **Fréquences relatives (fi) (en %)** |
|---|---|---|
| 2 | 3 | 3,8 |
| 3 | **0** | 0,0 |
| 4 | 5 | 6,3 |
| 5 | 4 | 5,0 |
| 6 | **0** | 0,0 |
| 7 | 7 | 8,8 |
| 8 | 7 | 8,8 |
| 9 | 10 | 12,5 |
| 10 | 11 | 13,8 |
| 11 | 12 | 15,0 |
| 12 | 8 | 10,0 |
| 13 | 3 | 3,8 |
| 14 | 5 | 6,3 |
| 15 | 4 | 5,0 |
| 16 | **0** | 0,0 |
| 17 | **0** | 0,0 |
| 18 | 1 | 1,3 |
| **Totaux** | 80 | 100,0 |

Population statistique : les 80 étudiants ayant composé en Économie
Individu statistique : l'un de ces 80 étudiants.
Caractère statistique : la note du test d'économie.
Type du caractère quantitatif discret (par hypothèse, valeurs entières).




**Bien mentionner les valeurs de toutes les modalités, même si l'effectif est nul !**

### 3. Les caractères quantitatifs continus

Les modalités, mesurables, peuvent prendre toute valeur à l'intérieur d'une plage de valeurs (classe).

En général, lorsqu'on peut utiliser des décimales, on a affaire à un caractère (= une variable) continue.

Comme pour les caractères quantitatifs discrets, on peut représenter un caractère quantitatif continu au moyen d'un diagramme différentiel et d'un diagramme intégral.

a) Le **diagramme différentiel** d'une série statistique retrace la **distribution** des effectifs (ou des fréquences), selon les modalités du caractère étudié.

Les modalités apparaissent sous la forme de plages de valeurs possibles, qu'on appelle des **classes**, à l'intérieur de chacune desquelles sont rangés les individus, dont la valeur du caractère est comprise entre les **extrémités** de l'une des classes confectionnées.

Deux principes sont à respecter pour la construction correcte des classes :

- principe **d'exhaustivité** : tout individu de la population étudiée est rangé (classé) dans une classe et une seule.

- Principe **d'incompatibilité** : les valeurs de deux classes successives ne peuvent pas se chevaucher.

b) Le **diagramme intégral** d'une série statistique retrace la **répartition** des effectifs (ou des fréquences) cumulés (on parle de fonction cumulative), selon les modalités du caractère étudié (i.e. ici, selon les classes).

Remarque : pour chacun de ces diagrammes, les représentations graphiques correspondantes sont radicalement différentes, selon qu'on a affaire à un caractère discret ou à un caractère continu.

**De manière générale, il convient de bien se rappeler que la première étape relative au traitement de données statistiques, consiste à repérer la nature et le type du caractère. En effet, c'est en fonction du caractère et de son type qu'en découlent les représentations graphiques possibles, ainsi que les traitements statistiques possibles.**

### 31. Présentation des données

La forme générale des tableaux est la suivante :

| Modalités caractère Classes ($x_i$) | Effectifs ($n_i$) | Fréquences relatives ($f_i$) | Pourcentages relatifs ($f_i$ x 100) | Effectifs cumulés N (x) | Fréquences cumulées F (x) | Pourcentages cumulés F (x) x 100 |
|---|---|---|---|---|---|---|
| $x_1$ | $n_1$ | $f_1$ | $f_1$ x 100 | N (1) | F (1) | F (1) x 100 |
| $x_2$ | $n_2$ | $f_2$ | $f_2$ x 100 | N (2) | F (2) | F (2) x 100 |
| … | … | … | … | … | … | … |
| $x_i$ | $n_i$ | $f_i$ | $f_i$ x 100 | N (i) | F (i) | F (i) x 100 |
| … | … | … | … | … | … | … |
| $x_k$ | $n_k$ | $f_k$ | $f_k$ x 100 | N (k) = n | F (k) = 1 | F (k) x 100 = 100 |
| | $n = \sum_{i=1}^{k} n_i$ | $1 = \sum_{i=1}^{k} f_i$ | $100 = \sum_{i=1}^{k} (f_i \times 100)$ | | | |

$x_i$ = **classe** correspondant à la modalité i du caractère x ;

k = nombre de modalités prises par le caractère x ;

$n_i$ = effectif partiel qui correspond à $x_i$ ;

n = effectif total ; $f_i = n_i / n$ ;

N (i) = effectif cumulé des individus de la population étudiée, dont la valeur du caractère est inférieure ou égale à i (c-à-d $n_1 + n_2 + \ldots + n_i$) ;

F (i) = fréquence cumulée (proportion) des individus de la population étudiée, dont la valeur du caractère est inférieure ou égale à i (c-à-d $f_1 + f_2 + \ldots + f_i$).

## 32. Représentations graphiques appropriées

Remarque : nous considérons ici des limites de classes complètement définies (valeurs précisées). Il n'en est pas toujours ainsi : voir le point 4., dans le cas contraire.

### 321. Amplitudes de classes identiques

**Exemple**

On considère une population (échantillon) de 200 personnes. On étudie cette population selon le caractère "Poids" de ces personnes.

La première chose à faire, est de **caractériser correctement les données**, dans le but de représenter correctement ces dernières.

Population statistique : un échantillon de 200 personnes.
Individu statistique : l'une des 200 personnes de l'échantillon.
Caractère statistique : le poids des personnes.
Type du caractère : quantitatif continu (décimales possibles).

Pour traiter ce type de variable, on recourt à des **classes de poids**. Chaque classe de poids correspond à **une** modalité de la variable statistique "Poids" des individus.

On suppose qu'à partir des données, le tableau suivant a pu être établi :

| Extrémités de classes ei | Classes xi | Effectifs ni | Effectifs cumulés N (x) | Fréquences relatives fi x 100 | Fréquences cumulées F (x) x 100 |
|---|---|---|---|---|---|
| 40 | | | 0 | | 0,0 |
| | [40-50[ | 32 | | 16,0 | |
| 50 | | | 32 | | 16,0 |
| | [50-60[ | 47 | | 23,5 | |
| 60 | | | 79 | | 39,5 |
| | [60-70[ | 51 | | 25,5 | |
| 70 | | | 130 | | 65,0 |
| | [70-80[ | 36 | | 18,0 | |
| 80 | | | 166 | | 83,0 |
| | [80-90[ | 19 | | 9,5 | |
| 90 | | | 185 | | 92,5 |
| | [90-100[ | 15 | | 7,5 | |
| 100 | | | 200 | | 100,0 |
| **Total** | | 200 | | 100,0 | |

Remarques relatives à la construction du tableau :

1°) **On décale l'écriture des extrémités de classes $e_i$**, de manière à "encadrer" les classes : ces extrémités représentent en effet les limites de chacune des classes. La construction de la courbe des effectifs (ou des fréquences) cumulés étant établie sur les valeurs qui correspondent aux extrémités de classes, les valeurs des colonnes correspondantes du tableau sont donc, elles aussi, décalées par rapport à l'écriture des effectifs (ou des fréquences) partiels.

Remarque : un tableur ne permet pas une mise en œuvre commode de ce décalage.

2°) La confection des classes doit satisfaire aux conditions d'exhaustivité et d'incompatibilité.

Par ailleurs, les crochets peuvent être conventionnellement fixés dans l'autre sens, pourvu qu'on respecte les deux conditions précédentes.

Remarque : contrairement au cas des caractères discrets, le décalage réalisé sur les extrémités de classes présente une signification concrète dans le cas des caractères continus. En effet, dans une même classe, toutes les valeurs intermédiaires peuvent être prises effectivement.

En toute rigueur, par analogie aux variables discrètes, lorsqu'on raisonne sur la base de la définition anglo-saxonne pour cumuler des effectifs ou des fréquences, on devrait utiliser les crochets dans le sens ouvrant à gauche et fermant à droite ( ] --- ] ).

### 3211. Le diagramme différentiel ou histogramme

Il s'agit d'une représentation graphique spécifique aux caractères continus. Cette représentation, qui retrace la **distribution** des effectifs (ou des fréquences) selon les modalités du caractère étudié, est composée d'un ensemble de rectangles accolés les uns aux autres.

Pour un tracé rigoureux de ces rectangles, **de façon à ce que la réalité des données numériques ne soit pas déformée par la représentation graphique**, on pose 2 hypothèses :

- hypothèse **d'équirépartition des effectifs (ou des fréquences) dans chaque classe**.

On suppose qu'à l'intérieur de chaque classe, l'effectif partiel correspondant $n_i$ est **uniformément réparti** à l'intérieur de la classe.

- on pose hypothèse que **la surface de chaque rectangle est proportionnelle à l'effectif (ou à la fréquence) partiel correspondant**.

De chacune de ces hypothèses, résulte un principe de construction particulier :

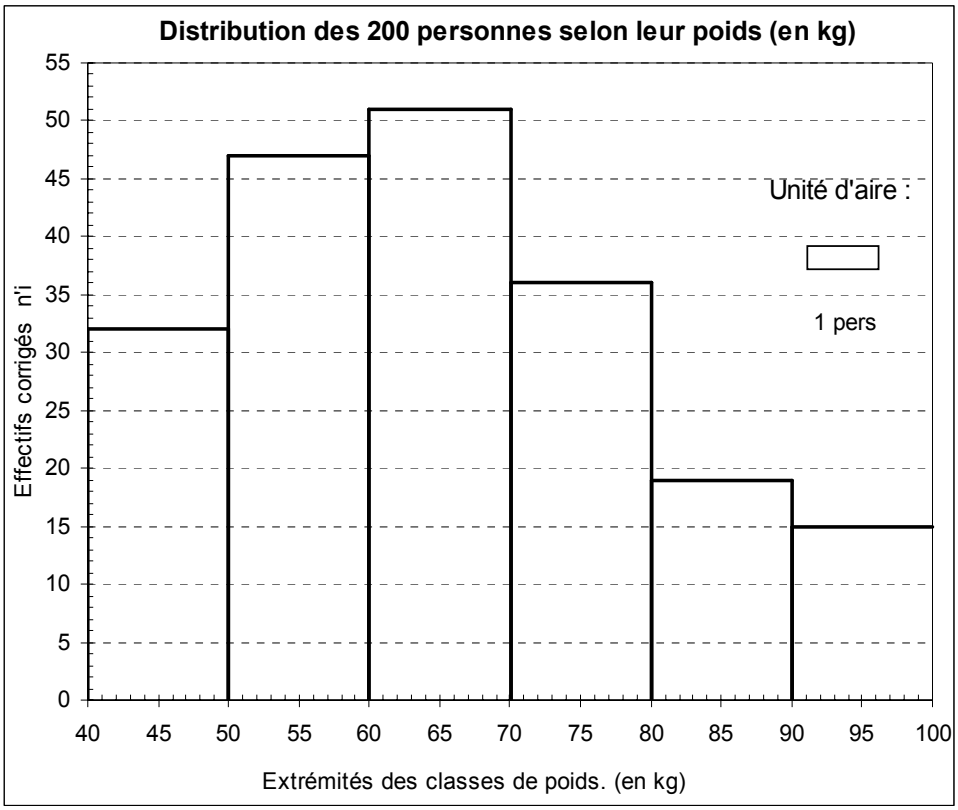
- suite à l'hypothèse d'équirépartition, le sommet de chaque rectangle, par construction, est fermé par un segment de droite horizontal.

- lorsque les amplitudes de classes sont inégales, afin de vérifier la $2^{ème}$ hypothèse, il est nécessaire de **corriger les effectifs (ou les fréquences) de certaines classes, ce qui va avoir pour effet de modifier les hauteurs des rectangles correspondants**.

Remarque : c'est seulement au point 322. que nous aurons besoin de mettre en œuvre la deuxième hypothèse est donc que nous devrons corriger les effectifs de certaines classes avant de pouvoir tracer correctement l'histogramme.

Pour le présent exemple, nous n'avons pas besoin d'ajouter des colonnes supplémentaires au tableau pour réaliser ce tracé.

**Remarque : prévoir le tracé ultérieur du deuxième diagramme, sur la même feuille, en dessous de l'histogramme, avec la même échelle en abscisse. Cette disposition facilitera les comparaisons.**




Au-delà du titre du graphique, des intitulés précis de chacun des deux axes, il est absolument indispensable de faire apparaître une autre information : **l'unité d'aire**.

La deuxième hypothèse vue plus haut, précise que **c'est la surface des rectangles qui est proportionnelle aux effectifs**. Cela signifie donc que, pour une échelle donnée du graphique, la surface comprise entre l'axe des abscisses et les sommets des rectangles correspond très exactement à 200 personnes, ou bien à 100 % si l'on avait raisonné en fréquences (en %).

**Ainsi, l'unité d'aire, que l'on fixe généralement égale à 1 cm², permet de savoir immédiatement à combien de personnes ou à quel pourcentage de personnes correspond 1 cm² de la surface située sous l'histogramme.**

Remarque : notamment dans le cas où l'on raisonne selon les effectifs, et afin d'être certain de disposer d'une unité d'aire renvoyant une valeur entière, on peut considérer directement l'un des rectangles (pour des raisons d'encombrement, en général le plus petit) comme unité d'aire. Toutefois, cette solution présente un inconvénient majeur si le graphique est destiné à être utilisé à titre de comparaison avec un ou plusieurs autres graphiques qui ne disposeront peut-être pas de la même unité d'aire.

Si l'on retient la première solution, la plus recommandée, on détermine le nombre de personnes qui correspondent à une surface de 1 cm² de la façon suivante :

on mesure la hauteur et la largeur de l'un quelconque des rectangles et l'on en déduit la surface (en cm²). Le tableau de données nous indique l'effectif réel correspondant au rectangle dont on a calculé la surface. Pour avoir le nombre de personnes par cm², il suffit donc de calculer le rapport de l'effectif réel correspondant au rectangle utilisé, à la surface calculée :

Effectif / cm² = effectif réel du rectangle choisi / surface du rectangle (en cm²)

### 3212. Le diagramme intégral ou fonction cumulée des effectifs (ou des fréquences)

À travers un histogramme, on cherche à faire apparaître le poids relatif de chacune des classes, en considérant leurs effectifs partiels.

Les fonctions (ou courbes) cumulées (ou cumulatives) mettent en évidence la façon dont les effectifs partiels se cumulent depuis la première jusqu'à la dernière classe.

Cette représentation graphique retrace la **répartition** des effectifs (ou des fréquences) cumulés, selon les modalités du caractère étudié

Afin de tracer les courbes correspondantes, il est nécessaire d'ajouter des colonnes au tableau de données.

**Définitions**

**N (x) représente l'effectif cumulé des individus de la population, dont la valeur du caractère est inférieure ou égale à x**.

**F (x) représente la proportion (fréquence) cumulée des individus de la population, dont la valeur du caractère est inférieure ou égale à x**.

Remarque : tout comme le diagramme différentiel des caractères discrets et continus est très différent (diagramme en bâtons dans le premier cas et histogramme dans le second), il en est de même pour les diagrammes intégraux respectifs (à la courbe cumulée discontinue, en escalier, des variables discrètes, on substitue ici une courbe cumulée continue pour les variables continues).

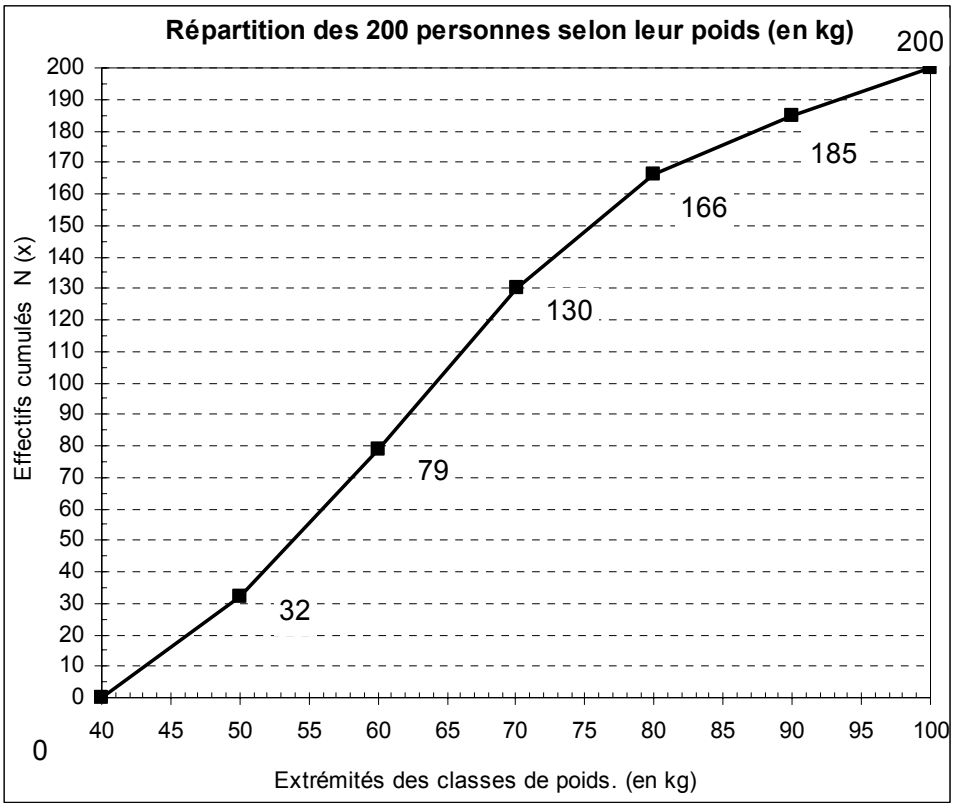
**Hypothèse** : comme pour l'histogramme, le tracé de la fonction cumulée est sous-tendu par l'hypothèse **d'équirépartition des effectifs (ou des fréquences) à l'intérieur de chaque classe**.

Il en découle le principe de construction suivant : la courbe est composée de segments de droite successifs.

Remarque importante pour la construction de la courbe cumulée : le point de départ d'un segment, de même que son point d'arrivée, correspondent à une **extrémité de classe**.

**Par convention**, on considère que : $N(-\infty) = 0$ et que : $F(-\infty) = 0$.

De même : $N(+\infty) = n$ et que : $F(+\infty) = 1$.




Sur le graphique, on observe que **le segment de droite dont la pente est la plus élevée correspond toujours au rectangle le plus haut de l'histogramme, c'est-à-dire à l'effectif (corrigé) le plus nombreux**.

Ici, il s'agit de la classe [ 60- 70 [, d'effectif 51 (= 130 -79).

Nous verrons au chapitre 2 que cette classe spéciale prend le nom de classe modale.

### 3213. Relation entre les deux diagrammes

Il existe une relation directe entre les deux représentations graphiques d'une variable statistique continue (c'est vrai également pour une variable discrète, mais plus délicat à interpréter).

Pour faire apparaître cette relation, supposons que l'on cherche à connaître, par exemple, le **nombre de personnes de la population dont le poids est inférieur ou égal à 65 kilos**.

Nous disposons de trois façons pour répondre à la question de :

**Première méthode** : **graphiquement, sur le diagramme de la courbe cumulée**, on monte une verticale au point d'abscisse 65, jusqu'à la rencontre de la courbe cumulée, puis on tire une horizontale jusque sur l'axe des ordonnées. La valeur indiquée, ici environ 105 personnes, représente précisément le nombre de personnes de la population dont le poids est inférieur ou égal à 65 kilos.

**Deuxième méthode** : **graphiquement, sur l'histogramme**, au point d'abscisse 65, on monte une verticale de façon à séparer cet histogramme en deux parties. On peut hachurer la partie gauche. Cette surface hachurée correspond au nombre de personnes pesant 65 kilos ou moins, puisque la surface des rectangles est proportionnelle aux effectifs **réels**.

Ainsi, on peut déterminer ce nombre de deux façons :

- La classe [40-50[ contient 32 personnes, la classe [50-60[ en contient 47 et la classe [60-70[ en contient 51. L'application de l'hypothèse d'équirépartition nous permet de déduire qu'entre 60 et 65, nous avons la moitié des personnes de la classe correspondante, soit 25,5. Au total (surface hachurée), on trouve 104,5 personnes (32 + 47 + 25,5).

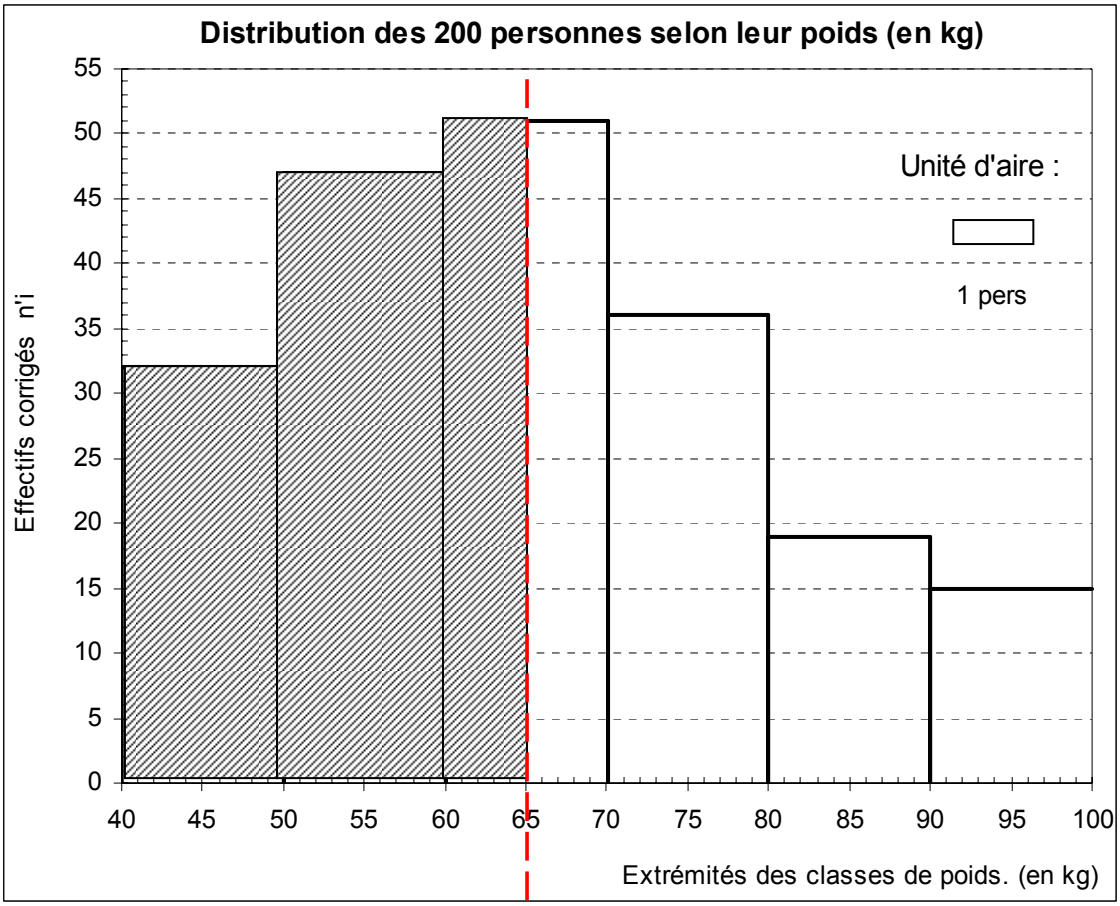
- Compte tenu de l'unité d'aire affectée à l'histogramme (ici = 1 personne), il suffit de mesurer la surface de chacun des 3 rectangles qui composent l'aire hachurée. On retrouve 104,5 personnes (à l'erreur de lecture graphique près).

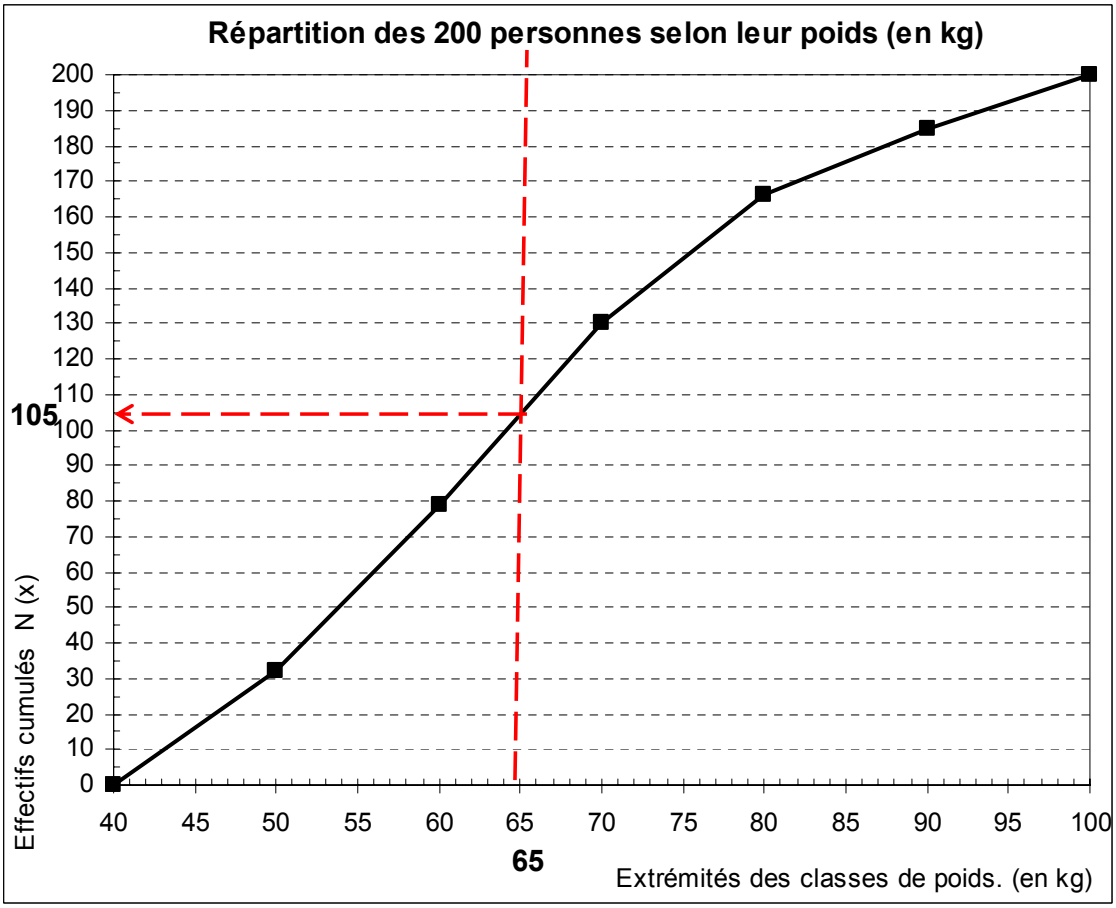
Remarque : dans la mesure où l'on raisonne sur des individus, on est tenu d'arrondir la valeur trouvée à l'entier le plus proche.

**Troisième méthode** : on peut calculer **algébriquement** la valeur N (x) qui correspond à x = 65 kg, à partir du tableau de données et des valeurs qui nous permettent de tracer la courbe cumulée.

Dans la mesure où le tracé de N (x) est sous-tendu par une hypothèse d'équirépartition, on parle **d'interpolation linéaire** (on retrouvera cette méthode au chapitre 2, pour le calcul des quartiles).

Sur le graphique de la fonction cumulée, considérons la classe à l'intérieur de laquelle se trouve la valeur d'abscisse x = 65.


Distribution des 200 personnes selon leur poids (en kg)
Effectifs corrigés n'i
Unité d'aire :
1 pers
0
5
10
15
20
25
30
35
40
45
50
55
40
45
50
55
60
65
70
75
80
85
90
95
100
Extrémités des classes de poids. (en kg)


Répartition des 200 personnes selon leur poids (en kg)
Effectifs cumulés N (x)
105
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
40
45
50
55
60
65
70
75
80
85
90
95
100
65
Extrémités des classes de poids. (en kg)

Lorsqu'on a affaire à une droite, les règles de proportionnalité permettent d'écrire :

$$\frac{N(65) - 79}{130 - 79} = \frac{65 - 60}{70 - 60}$$

D'où l'on tire :

$$N(65) = 79 + (130 - 79) \times \frac{65 - 60}{70 - 60} = 79 + 51 \times \frac{5}{10} = 79 + 25{,}5 = 104{,}5$$

51 représente l'effectif relatif de la classe dans laquelle on se trouve, soit la classe [60 - 70[.

10 représente l'amplitude de la classe dans laquelle on se trouve.

Au total donc, on dénombre 104,5 personnes (on arrondit ici ce résultat à **105**) dont le poids est inférieur ou égal à 65 kilos.

Remarque : on pourrait raisonner en pourcentage, en vue de déterminer la proportion de personnes dont le poids est inférieur ou égal à 65 kilos. Il suffit ici de poser le rapport 105 / 200 et de remultiplier le résultat par 100, pour obtenir **52,5 %**.

Remarque : dans le cas général, on peut écrire la formule d'interpolation linéaire suivante, pour un calcul portant sur les effectifs :

$$N(x) = N(ei) + [N(ei+1) - N(ei)] \times (x - ei) / (ei+1 - ei)$$

avec : x = valeur critique ; $[N(ei+1) - N(ei)] = ni$ ; $(ei+1 - ei) = ai$

Dans le cas d'une proportion, on remplace simplement les valeurs des effectifs cumulés par celle des fréquences cumulées :

$$F(x) = F(ei) + [F(ei+1) - F(ei)] \times (x - ei) / (ei+1 - ei)$$

avec : x = valeur critique ; $[F(ei+1) - F(ei)] = fi$ ; $(ei+1 - ei) = ai$

**Remarques de prolongement**

La valeur N(x) = 104,5 (courbe cumulée) qui correspond à l'abscisse x = 65 kg est égale à l'aire de l'histogramme comprise entre l'abscisse 40 et l'abscisse 65 (hachures).

Par conséquent, une valeur numérique N(x) correspond à une surface (hachurée). Cf. cours de Math sur les intégrales.

La courbe cumulée est la primitive (d'où le nom de diagramme intégral) de la courbe enveloppe de l'histogramme (d'où le nom de diagramme différentiel). Cette courbe enveloppe est la dérivée de la courbe cumulée.

Remarque : ce raisonnement devient rigoureux lorsque le nombre de classes devient infini (cf. la définition de l'intégrale de Riemann, à partir de la surface de rectangles d'amplitude tendant vers zéro). Dans ce cas, le contour de l'histogramme tend vers une courbe lissée, de même que celui de la courbe cumulée.

Lorsque le nombre de classes tend vers l'infini, l'histogramme tend vers la fonction de distribution de la variable x. La courbe cumulée tend vers la fonction de répartition de la variable x. D'où les intitulés donnés à chacun de ces graphiques.

Il existe un lien avec les lois de probabilité des variables aléatoires. La fonction de densité de probabilité f(x) de la variable aléatoire X et la dérivée de la fonction de répartition F(x) de cette même variable aléatoire.

**Attention !**

Certains praticiens de la statistique font comme si la condition du nombre de classes (tendant vers l'infini) était vérifiée. Pour cette raison, ils adjoignent ou substituent une courbe de distribution à l'histogramme, afin d'avoir une idée de l'allure de cette distribution (pour renvoyer plus facilement à telle ou telle loi de probabilité théorique). Le même raisonnement est également étendu à la courbe cumulée (ou encore à la courbe de concentration de Lorenz).

**Mais** :

1) la construction méthodique précise de la courbe de distribution associée à un histogramme est un exercice périlleux (sauf à avoir de bonnes raisons de penser que telle ou telle loi s'applique au phénomène observé). Il en résulte une précision illusoire car on ne connaît pas l'équation de la courbe f(x).

2) dans un trop grand nombre de manuels, on rencontre une incohérence fautive, par l'utilisation de deux hypothèses différentes, appliquées l'une à la distribution, l'autre à la répartition de la variable envisagée. Ainsi, posant l'hypothèse d'équirépartition dans le cas de la distribution, on trace un histogramme et posant une hypothèse de normalité dans le cas de la répartition, on trace une courbe cumulée lissée, ce qui occasionne une incohérence totale lors de la comparaison des deux représentations graphiques.

En conclusion, si l'on veut avoir une idée de l'allure de la distribution de la série (asymétrie, aplatissement,...), il suffit d'effectuer un tracé rapide à main levée ou bien de mettre en œuvre un certain nombre de calculs (coefficient de Yule, de Pearson, de Fisher, …, basés sur l'hypothèse … d'équirépartition !). L'objectif est de suggérer la loi de probabilité théorique à appliquer à un échantillon, de façon à généraliser les résultats obtenus sur une population plus large.

Remarque : si l'on opte pour une hypothèse différente de l'équirépartition, c'est souvent au profit d'une hypothèse de normalité de la distribution.

Dans ce cas, avant le mode de la série, les effectifs tendent à être disposés dans les classes de façon croissante. Après le mode de la série, les effectifs tendent à être disposés dans les classes de façon décroissante.

C'est ce raisonnement qui permet d'ajuster une fonction de distribution à l'histogramme correspondant.

En tout état de cause, si l'on ne dépasse pas le stade de la statistique descriptive, ce tracé empirique est quelque peu inutile, dans la mesure où les calculs relatifs aux paramètres tels que la moyenne ou l'écart-type sont, eux, toujours sous-tendus par l'hypothèse d'équirépartition des effectifs dans les classes !

### 322. Amplitudes de classes différentes

La démarche mise en oeuvre dans l'exercice précédent, va se retrouver ici, mais va inclure une étape supplémentaire, du fait que les amplitudes de classe ne sont plus toutes identiques.

**Conséquence fondamentale : le tracé de l'histogramme est modifié**.

En effet, comme on l'a vu dans le point 321., les surfaces des rectangles sont, par hypothèse, proportionnelle aux effectifs.

**Pour que cette hypothèse soit respectée, lorsque les amplitudes de classes sont différentes, il est nécessaire d'effectuer une correction qui va modifier la hauteur de certains rectangles, de façon à ce que la représentation graphique de la série reste fidèle aux données numériques de départ**.

Nous avons déjà vu que le tracé d'un histogramme correct requiert les hypothèses suivantes :

- équirépartition des effectifs (ou des fréquences) dans chaque classe : en conséquence, chaque classe est représentée par un rectangle dont le sommet est parallèle à l'axe des abscisses.

- surface des rectangles proportionnelle aux effectifs : en conséquence, il est nécessaire de corriger les effectifs n'i (ou les fréquences f'i), lorsque les amplitudes de classes sont différentes.

Lorsque les amplitudes de classes sont différentes, on complète le tableau précédent de la manière suivante :

| **Modalités du caractère Classes ($x_i$)** | **Effectifs ($n_i$)** | **Amplitudes des classes ($a_i$)** | **Effectifs corrigés ($n'_i$)** |
|---|---|---|---|
| $x_1$ | $n_1$ | $a_1 = e_2 - e_1$ | $n'_1 = n_1 / a_1 \times a_0$ |
| $x_2$ | $n_2$ | $a_2 = e_3 - e_2$ | $n'_2 = n_2 / a_2 \times a_0$ |
| … | … | … | … |
| $x_i$ | $n_i$ | $a_i = e_{i+1} - e_i$ | $n'_i = n_i / a_i \times a_0$ |
| … | … | … | … |
| $x_k$ | $n_k$ | $a_k = e_{k+1} - e_k$ | $n'_k = n_k / a_k \times a_0$ |

Par exemple :

$a_0$ représente la valeur de l'amplitude de classe qui revient le plus souvent dans la colonne des $a_i$.

$a_1$, amplitude de la 1ère classe $x_1$, est obtenue en retranchant à la valeur de la borne supérieure ($e_2$) de cette classe, la valeur de la borne inférieure ($e_1$) de cette même classe.

Par suite, l'effectif corrigé $n'_1$ de la 1ère classe est obtenu en divisant l'effectif réel $n_1$ par l'amplitude $a_1$ de la classe correspondante et en multipliant le résultat par $a_0$.

**Ce sont les valeurs de la dernière colonne qu'on reporte en ordonnée de l'histogramme.**

Remarque : on procède évidemment de manière similaire si l'on travaille, non pas sur les effectifs, mais sur les fréquences ou les pourcentages. Dans le cas des fréquences, par exemple, on aura la formule générale suivante : $f'_i = f_i / a_i \times a_0$ .

**Exemple**

Dans le tableau suivant, nous considérons la distribution des chiffres d'affaires (en millions d'euros annuels) de 200 entreprises.

Comme précédemment, la première chose à faire, est de caractériser correctement les données, dans le but de représenter correctement ces dernières.

Population statistique : un échantillon de 200 entreprises.

Individu statistique : l'une des 200 entreprises de l'échantillon.

Caractère statistique : le CA (en M€ annuels) des entreprises.

Type du caractère : quantitatif continu (décimales possibles).

Pour traiter ce type de variable, on recourt ici à des classes de chiffres d'affaires. Chaque classe de CA correspond à **une** modalité de la variable statistique "Chiffre d'affaires" des entreprises.

On suppose qu'à partir des données, le tableau suivant a pu être établi :

| **Classes xi** | **Effectifs ni** | **Extrémités de classes ei** | **Amplitudes de classes ai** | **Effectifs corrigés n'i** | **Effectifs cumulés N (x)** | **Fréquences relatives fi x 100** | **Fréquences cumulées F (x) x 100** |
|---|---|---|---|---|---|---|---|
| | | 3,3 | | | 0 | | 0,0 |
| [3,3-3,5[ | 11 | | 0,2 | **27,5** | | 5,5 | |
| | | 3,5 | | | 11 | | 5,5 |
| [3,5-4,0[ | 31 | | **0,5** | 31,0 | | 15,5 | |
| | | 4,0 | | | 42 | | 21,0 |
| [4,0-4,5[ | 41 | | **0,5** | 41,0 | | 20,5 | |
| | | 4,5 | | | 83 | | 41,5 |
| [4,5-5,0[ | 60 | | **0,5** | 60,0 | | 30,0 | |
| | | 5,0 | | | 143 | | 71,5 |
| [5,0-5,3[ | 43 | | 0,3 | **71,7** | | 21,5 | |
| | | 5,3 | | | 186 | | 93,0 |
| [5,3-5,5[ | 14 | | 0,2 | **35,0** | | 7,0 | |
| | | 5,5 | | | 200 | | 100,0 |
| **Total** | 200 | | | | | 100,0 | |

On réalise le même décalage que précédemment pour les colonnes ei, N(x) et F(x).

Ici, nous retenons : $a_0 = 0,5$, car il s'agit de la valeur d'amplitude de classe qui revient le plus souvent.

### 3221. Construction de l'histogramme (diagramme différentiel)

Les deux hypothèses que nous avons posées dans l'exemple précédent sont toujours valables ici. L'histogramme respecte donc toujours l'hypothèse d'équirépartition des effectifs dans les classes. Mais, de façon à toujours respecter l'hypothèse indiquant que les surfaces des rectangles sont proportionnelles aux effectifs des classes correspondantes, **il est nécessaire ici de corriger les effectifs réels des classes selon le principe suivant** :

$$n_i' = n_{corr} = \frac{n_i}{a_i} \ (x\ a_0)$$

**L'effectif corrigé** d'une classe $n_i'$ ou $n_{corr}$ est égal au rapport de l'effectif $n_i$ de cette classe à l'amplitude $a_i$ de cette classe. Éventuellement, on multiplie la valeur de ce rapport par celle d'une **amplitude de référence** $a_0$. Dans ce cas, pour limiter les calculs de, on a intérêt à prendre comme amplitude de référence celle qui revient le plus souvent.

Dans notre exemple, les amplitudes de classes prennent trois valeurs différentes : $a_i$ = 0,2 ; 0,3 ; 0,5. Ici, nous prenons **$a_0$ = 0,5** , dans la mesure où cette valeur d'amplitude revient 3 fois sur 6 classes. L'intérêt de ce choix provient du fait que les effectifs des 3 classes correspondantes n'ont pas besoin d'être modifiés, ce qui est appréciable lorsqu'on fait les calculs à la main.

Remarque : si l'on raisonne en fréquences relatives (ou en pourcentage), on procède de la même façon, en utilisant la formule suivante :

$$f_i' = f_{corr} = \frac{f_i}{a_i} \ (x\ a_0)$$

Dans le tableau de calcul, il convient d'ajouter 2 colonnes :

- l'une pour indiquer l'amplitude de chaque classe (qu'on détermine simplement par différence entre la valeur de l'extrémité supérieure de la classe et la valeur de l'extrémité inférieure de cette même classe) ;

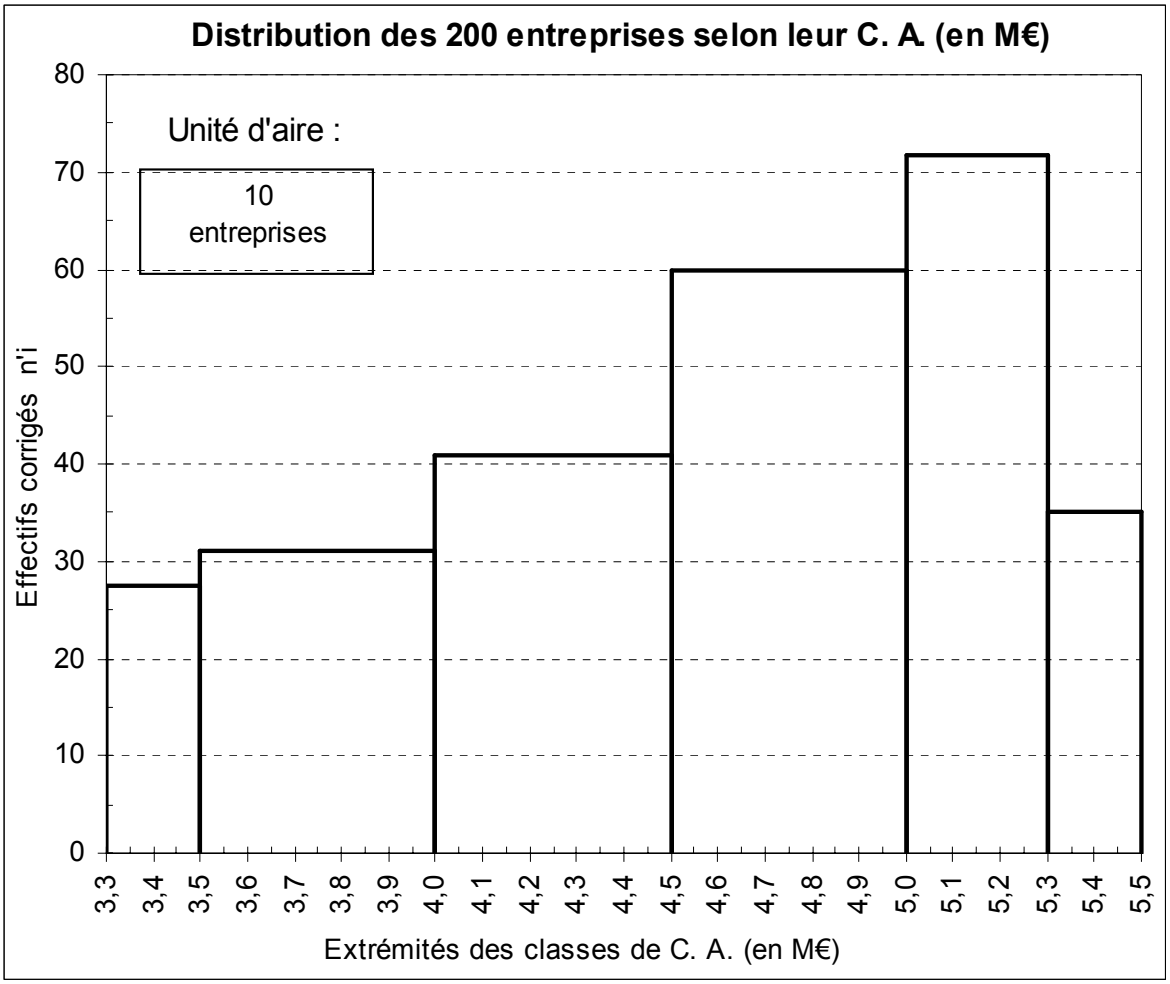
- l'autre pour calculer les effectifs corrigés sur la base de l'une ou l'autre des formules ci-dessus.

**Comme dans l'exercice précédent, on tracera les deux diagrammes sur la même feuille de papier, en faisant correspondre les abscisses de chaque graphique.**

Bien penser à renseigner l'intitulé de chaque axe et de donner un titre au graphique.

Enfin ici, **l'unité d'aire** est fondamentalement requise, puisque les effectifs réels sont modifiés. **On n'a donc plus que ce moyen pour connecter la réalité des données à la représentation graphique**.

Sur le graphique ci-après, nous avons arbitrairement considéré le tiers inférieur du deuxième rectangle, dont nous savons que l'effectif réel est de 10 entreprises.


Distribution des 200 entreprises selon leur C. A. (en M€)
Unité d'aire :
10
entreprises
Effectifs corrigés n'i
0
10
20
30
40
50
60
70
80
3,3
3,4
3,5
3,6
3,7
3,8
3,9
4,0
4,1
4,2
4,3
4,4
4,5
4,6
4,7
4,8
4,9
5,0
5,1
5,2
5,3
5,4
5,5
Extrémités des classes de C. A. (en M€)

### 3222. La fonction cumulé des effectifs (ou des fréquences) = diagramme intégral de la série statistique

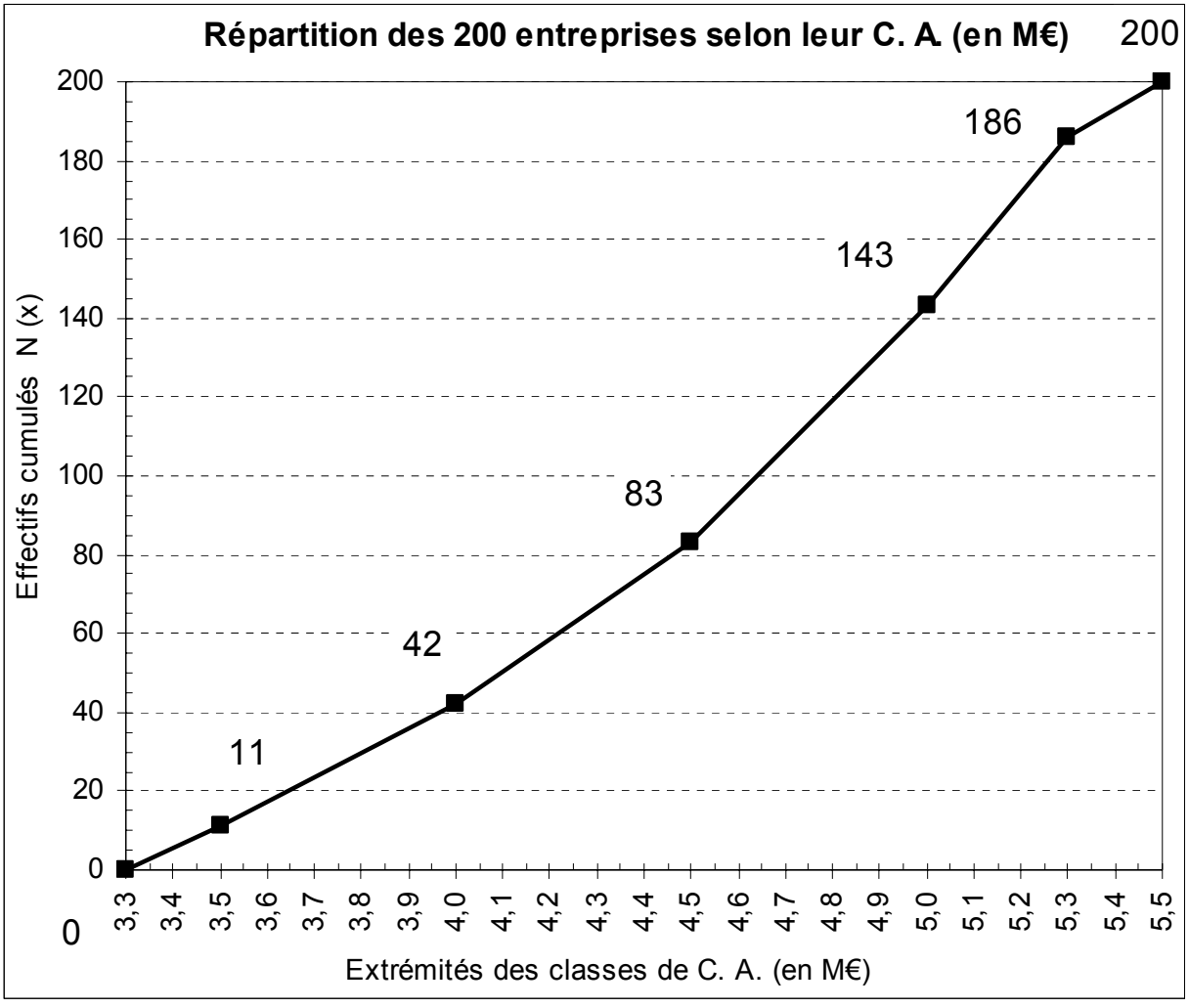
Rappels :

a) il s'agit d'un diagramme de **répartition** des effectifs (ou des fréquences) cumulés d'une série statistique.

b) N (x) représente l'effectif cumulé des individus de la population, dont la valeur du caractère (ici le CA) est inférieure ou égale à x.

c) F (x) représente la proportion (fréquence) cumulée des individus de la population, dont la valeur du caractère est inférieure ou égale à x.

On complète le tableau en cumulant les effectifs (ou les fréquences, ou les pourcentages). On respecte l'hypothèse de tracé, déjà utilisée dans l'exemple précédent, à savoir l'équirépartition des effectifs dans chaque classe (segments de droite).



### 3223. La relation histogramme - courbe cumulée

Pour la mettre en évidence ici, cherchons à déterminer le nombre d'entreprises dont le chiffre d'affaires annuel est inférieur ou égal à **4,3 M€**.

Nous avons vu précédemment que nous disposerons de trois méthodes (2 graphiques et une algébrique) pour trouver la solution :

a) lecture directe à partir de la courbe cumulée.

b) lecture à partir de l'histogramme ou utilisation de l'unité d'aire.

c) mise en œuvre d'une interpolation linéaire.

Ici, dans l'exemple, la fonction de répartition permet de renvoyer immédiatement le nombre d'entreprises ayant un chiffre d'affaires compris entre 3,3 et 4,3 M€, soit environ 67 entreprises (à l'erreur de lecture graphique près).

La surface hachurée de l'histogramme correspond à ce même nombre d'entreprises. On peut déterminer ce nombre de deux façons :

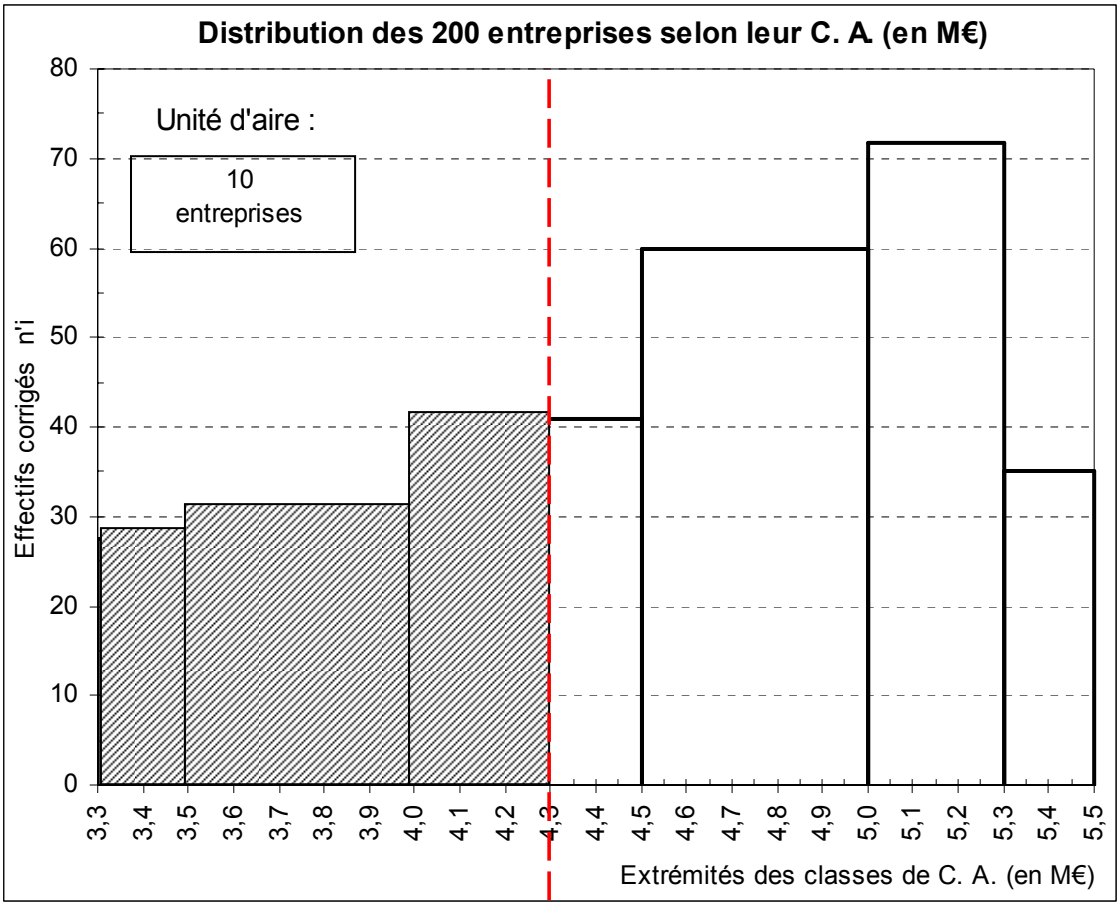
- La classe [3,3-3,5[ contient 11 entreprises, la classe [3,5-4,0[ en contient 31 et la classe [4,0-4,5[ en contient 41.

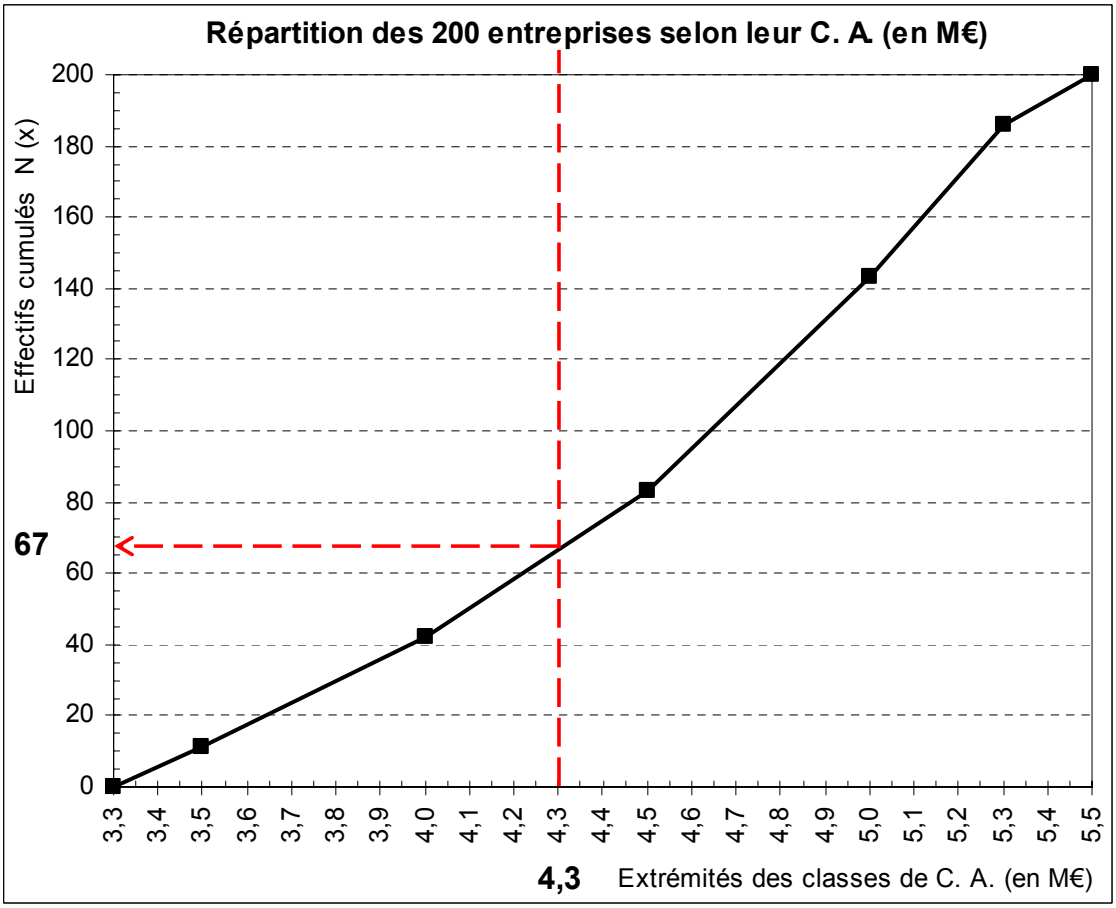
L'application de l'hypothèse d'équirépartition nous permet de déduire qu'entre 4,0 et 4,3 nous avons 3/5 des entreprises de la classe correspondante, soit environ 25.

Au total (surface hachurée), on retrouve bien 67 entreprises (11 + 31 + 25) .

- Compte tenu de l'unité d'aire affectée à l'histogramme (10 entreprises, sur l'histogramme réalisé sur tableur), il suffit de mesurer la surface de chacun des 3 rectangles qui composent l'aire hachurée.

On retrouve environ 67 entreprises (à l'erreur de lecture graphique près).


Distribution des 200 entreprises selon leur C. A. (en M€)
Unité d'aire :
10
entreprises
Effectifs corrigés n'i
0
10
20
30
40
50
60
70
80
3,3
3,4
3,5
3,6
3,7
3,8
3,9
4,0
4,1
4,2
4,3
4,4
4,5
4,6
4,7
4,8
4,9
5,0
5,1
5,2
5,3
5,4
5,5
Extrémités des classes de C. A. (en M€)


Répartition des 200 entreprises selon leur C. A. (en M€)
Effectifs cumulés N (x)
67
0
20
40
60
80
100
120
140
160
180
200
3,3
3,4
3,5
3,6
3,7
3,8
3,9
4,0
4,1
4,2
4,3
4,4
4,5
4,6
4,7
4,8
4,9
5,0
5,1
5,2
5,3
5,4
5,5
4,3
Extrémités des classes de C. A. (en M€)

Sur le graphique de la fonction cumulée, considérons la classe à l'intérieur de laquelle se trouve la valeur d'abscisse x = 4,3.

Lorsqu'on a affaire à une droite, les règles de proportionnalité permettent d'écrire :

$$\frac{N(4,3) - 42}{83 - 42} = \frac{4,3 - 4}{4,5 - 4}$$

D'où l'on tire :

$$N(4,3) = 42 + (83 - 42) \times \frac{4,3 - 4}{4,5 - 4} = 42 + 41 \times \frac{0,3}{0,5} = 42 + 24,6 = 66,6$$

41 représente l'effectif relatif de la classe dans laquelle on se trouve, soit la classe [4,0 - 4,5[.

0,5 représente l'amplitude de la classe dans laquelle on se trouve.

Au total donc, on dénombre 66,6 entreprises (on arrondit ici ce résultat à **67**) dont le CA est inférieur ou égal à 4,3 M€.

Remarque : on pourrait raisonner en pourcentage (c'est généralement plus pertinent, comme on l'a déjà vu, à partir du moment où l'on se sert des résultats à titre de comparaison avec d'autres populations ou échantillons), en vue de déterminer la proportion de personnes dont le CA est inférieur ou égal à 4,3 M€.

Il suffit ici de poser le rapport 67 / 200 et de remultiplier le résultat par 100, pour obtenir **33,5 %**.

Remarque : dans le cas général, on peut écrire la formule d'interpolation linéaire suivante, pour un calcul portant sur les effectifs :

$$N(x) = N(ei) + [N(ei+1) - N(ei)] \times (x - ei) / (ei+1 - ei)$$

avec : x = valeur critique ; $[N(ei+1) - N(ei)] = ni$ ; $(ei+1 - ei) = ai$

Dans le cas d'une proportion, on remplace simplement les valeurs des effectifs cumulés par celle des fréquences cumulées :

$$F(x) = F(ei) + [F(ei+1) - F(ei)] \times (x - ei) / (ei+1 - ei)$$

avec : x = valeur critique ; $[F(ei+1) - F(ei)] = fi$ ; $(ei+1 - ei) = ai$

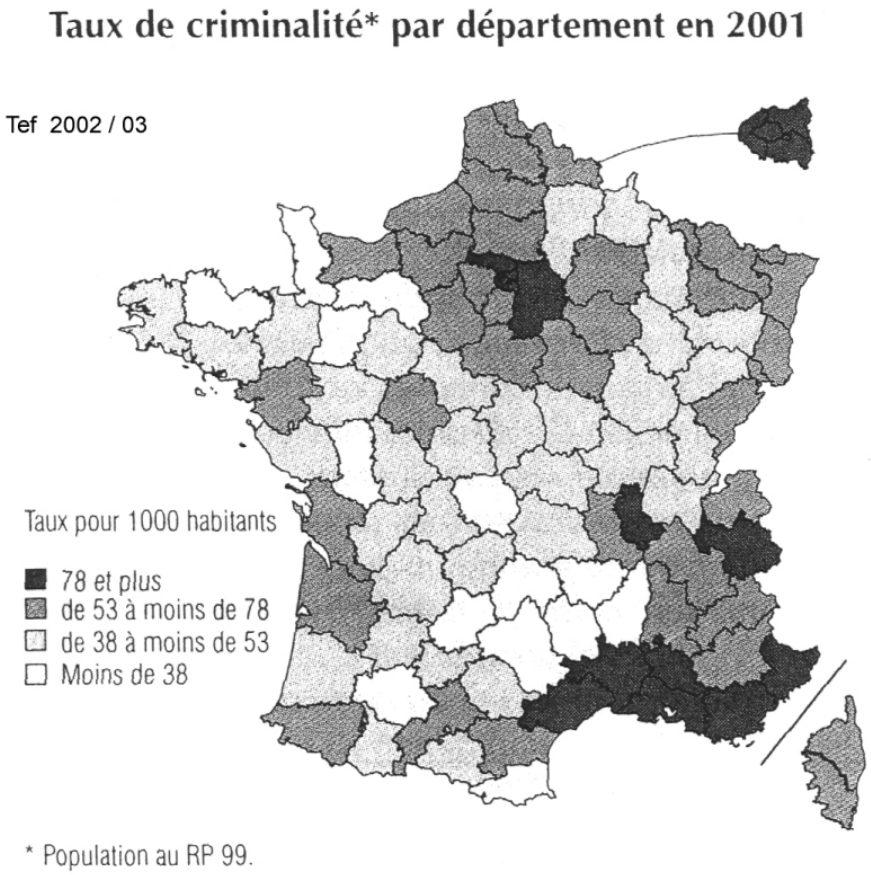
Deux remarques terminales :

a) une **pyramide des âges** est un histogramme double qu'on a fait pivoter de 90°. Le caractère de la série est l'âge des personnes. Il s'agit d'un caractère quantitatif continu, dont les modalités ont été regroupées en classes d'âge. Les deux histogrammes (hommes et femmes) sont accolés par leur base et placés verticalement.

b) attention à la dénomination "Histogramme" des tableurs qui renvoie en fait à un diagramme en tuyaux d'orgue. On peut cependant l'utiliser pour tracer des histogrammes, mais uniquement dans le cas où les amplitudes de classes sont identiques (on joue sur la largeur des rectangles, de façon à ce qu'ils apparaissent accolés les uns aux autres sur le graphique). Si l'on a affaire à des classes d'amplitude variable, il est nécessaire d'utiliser un graphique "Nuage de points" et l'on doit faire subir un traitement particulier aux extrémités de classes et aux effectifs corrigés pour tracer correctement un véritable histogramme.

### 323. Les cartogrammes

Exemple :




**Attention !** les cartogrammes du type précédent, relatifs aux caractères quantitatifs continus, génèrent des erreurs visuelles d'interprétation, dans la mesure où la superficie des différents départements n'est pas la même. Selon le caractère envisagé, le défaut de ce type de représentation graphique est plus ou moins gênant, eu égard au degré de précision recherché.

Remarque : de tels: cartogrammes peuvent s'appliquer à des **caractères discrets traités en continu**.

**Remarque terminale sur les caractères quantitatifs discrets et continus**

Dans les tableaux de données, il arrive parfois qu'on trouve des modalités pour lesquelles des regroupements ont été effectués.

Cas d'une variable discrète : par exemple, on peut avoir une modalité "3 enfants ou plus". Ce type de modalité ne pose pas de problème particulier pour tracer les diagrammes associés aux caractères discrets.

Mais si l'on veut calculer une moyenne ou un écart-type, il devient indispensable de poser une hypothèse sur le contenu de cette modalité.

Cas d'une variable continue : par exemple, on peut avoir les modalités "Moins de 20 ans" ou "Plus de 1 000 €".

Pour tracer les diagrammes associés aux caractères continus, tout autant que pour calculer une moyenne ou un écart-type sur la série correspondante, il est nécessaire de déterminer une borne inférieure (s'il s'agit de la 1ère classe) ou une borne supérieure (s'il s'agit de la dernière classe).

Dans un tel cas si, au-delà du tableau de données étudié, on ne dispose d'aucune information complémentaire, on se conformera au principe suivant :

- borne supérieure de la dernière classe : on retiendra comme amplitude de la dernière classe celle de l'avant-dernière classe.

Par exemple, si celle-ci est [900 ; 1 000[ (amplitude = 100), la borne supérieure de la dernière classe sera égale à 1 000 + 100 = 1 100 €, soit [1 000 ; 1 100[

- borne inférieure de la 1ère classe : on retiendra comme amplitude de la 1ère classe, soit celle de la 2ème classe, soit éventuellement zéro si cette valeur apparaît plus pertinente.

Par exemple, si la 2ème classe est [20 ; 25[ (amplitude = 5), la borne inférieure de la 1ère classe peut être 20 - 5 = 15, soit [15 ; 20[ , ou [0 ; 20[.

Si, au-delà du tableau de données, on dispose d'informations complémentaires, on doit alors justifier la valeur retenue pour borner la classe, sur la base de ces informations complémentaires.

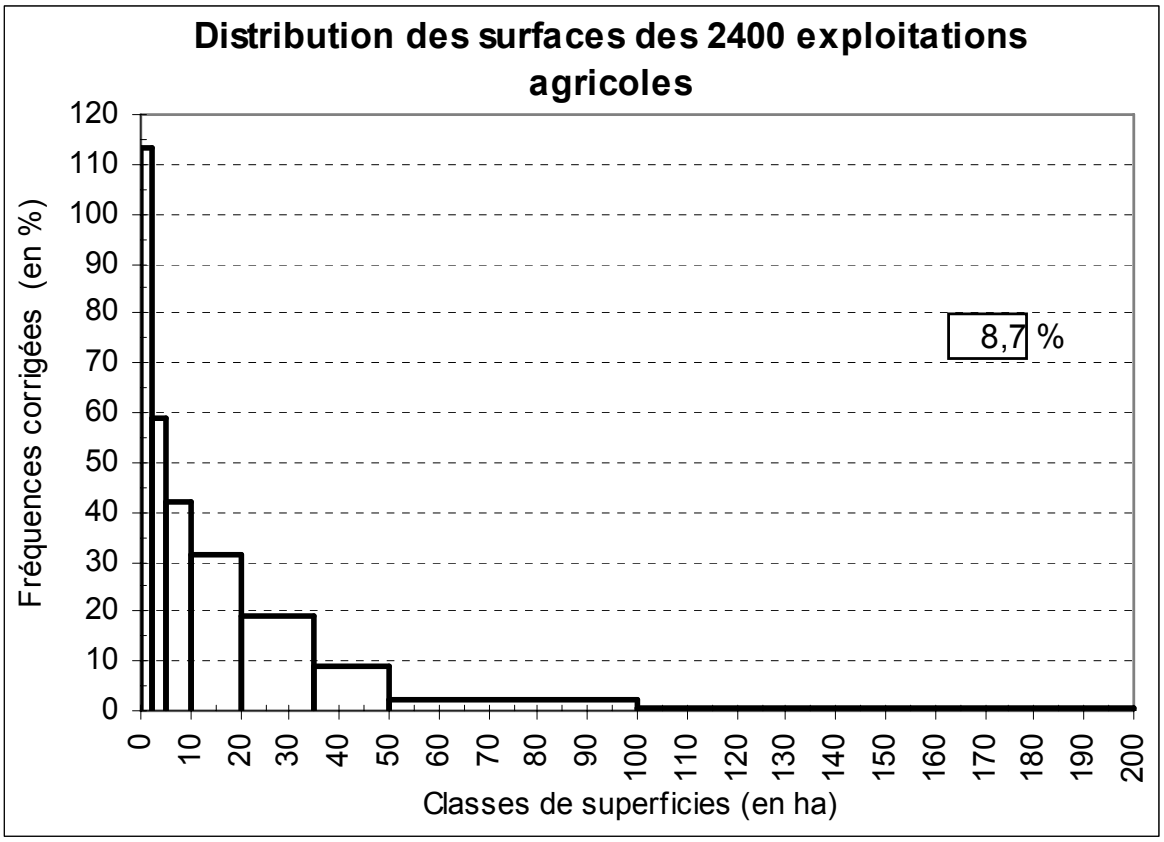
**Caractères quantitatifs continus : second exemple avec amplitudes de classes différentes**

Population statistique : 2 400 exploitations agricoles.
Individu statistique : l'une de ces 2 400 exploitations.
Caractère statistique : la surface (en ha) des exploitations.
Type du caractère : quantitatif continu.

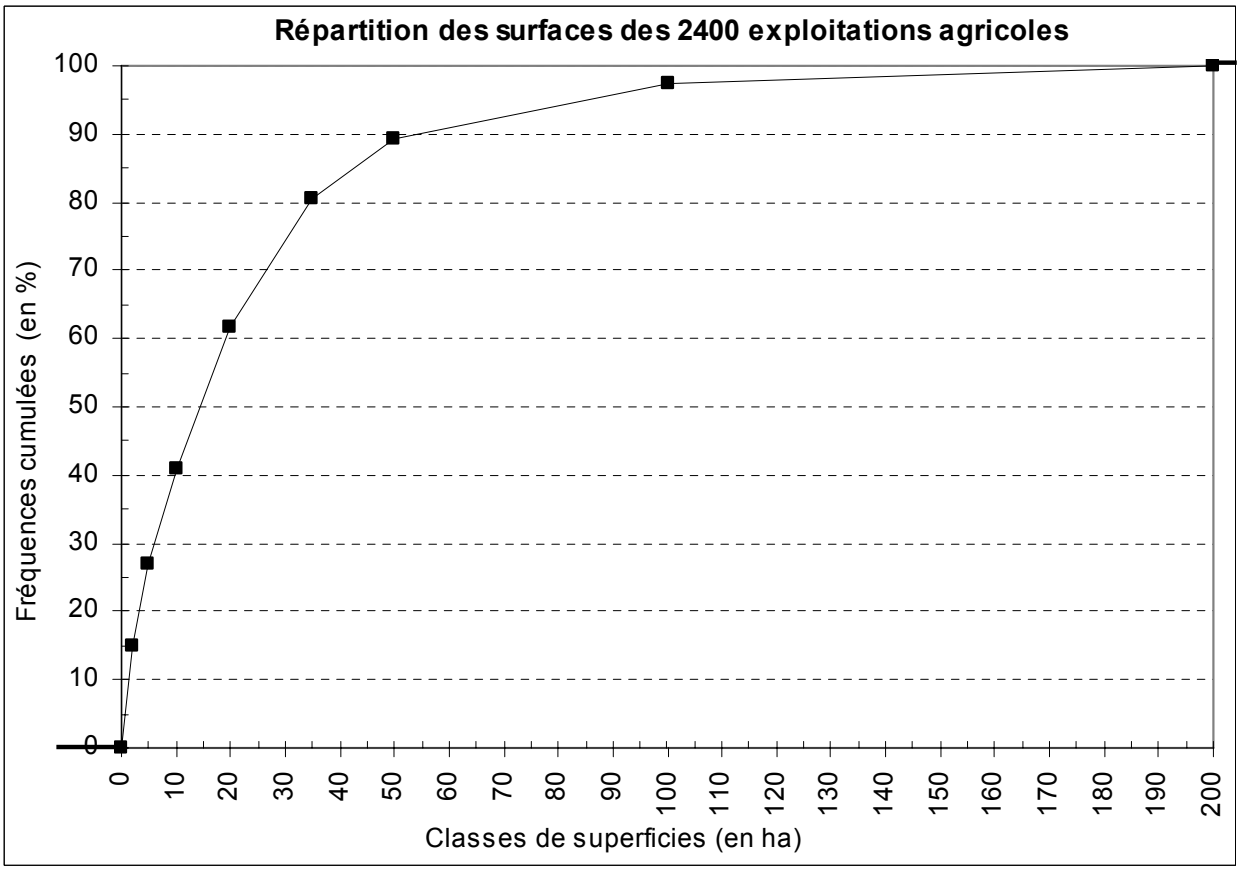
Dans cet exemple, le raisonnement est effectué sur les fréquences.

Ici, nous retenons : $f'_i = f_i / a_i \times 15$

| Classes de superficies xi | Effectifs ni | Fréquences fi (en %) | Fréquences cumulées Fi (en %) | ei | ai | Fréquences corrigées f'i |
|---|---|---|---|---|---|---|
| | | | 0 | 0 | | 0 |
| **[ 0 - 2 [** | 362 | 15,1 | 15,1 | 2 | 2 | 113,1 |
| **[ 2 - 5 [** | 283 | 11,8 | 26,9 | 5 | 3 | 59,0 |
| **[ 5 - 10 [** | 336 | 14,0 | 40,9 | 10 | 5 | 42,0 |
| **[ 10 - 20 [** | 502 | 20,9 | 61,8 | 20 | 10 | 31,4 |
| **[ 20 - 35 [** | 451 | 18,8 | 80,6 | 35 | 15 | 18,8 |
| **[ 35 - 50 [** | 209 | 8,7 | 89,3 | 50 | 15 | 8,7 |
| **[ 50 - 100 [** | 197 | 8,2 | 97,5 | 100 | 50 | 2,5 |
| **[ 100 - 200 [** | 60 | 2,5 | 100,0 | 200 | 100 | 0,4 |
| **Total** | **2 400** | 100,0 | | | | |




L'unité d'aire coïncide avec le rectangle [35-50[, qui correspond à une proportion de 8,7 % d'exploitations (valeur réelle de la fréquence).



Si l'on retient la valeur **40 ha** en abscisse (surfaces des exploitations), la courbe de répartition permet de renvoyer immédiatement la proportion de logements dont la superficie est inférieure ou égale à 40 ha, soit environ **84 %** (à l'erreur près de lecture graphique).

La surface hachurée de l'histogramme correspond à cette même proportion de logements On peut déterminer celle-ci de deux manières :

- la classe [0-2[ contient 15,1 % de logements, la classe [2-5[ en contient 11,8, la classe [5-10[ en contient 14,0, la classe [10-20[ en contient 20,9, la classe [20-35[ en contient 18,8. Enfin, l'application de l'hypothèse d'équirépartition nous permet de déduire qu'entre 35 et 40 nous en avons 8,7 / 3 = 2,9. Au total (surface hachurée), on obtient **83,5 %** d'exploitations.

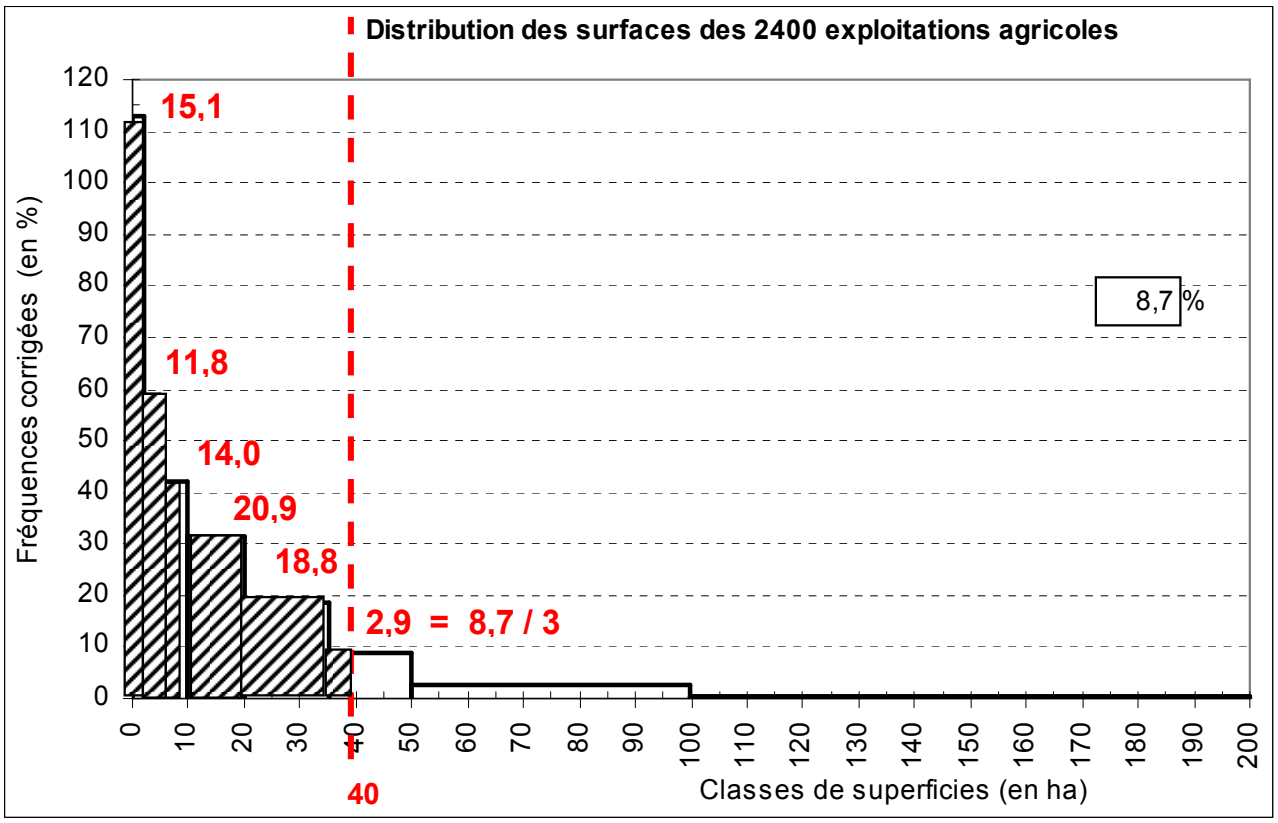
- compte tenu de l'unité d'aire affectée à l'histogramme (8,7 % d'exploitations correspondent à l'aire du dernier rectangle), il suffit de mesurer la surface de chacun des rectangles qui composent l'aire hachurée. On retrouve **83,5 %** d'exploitations (à l'erreur près de lecture graphique).

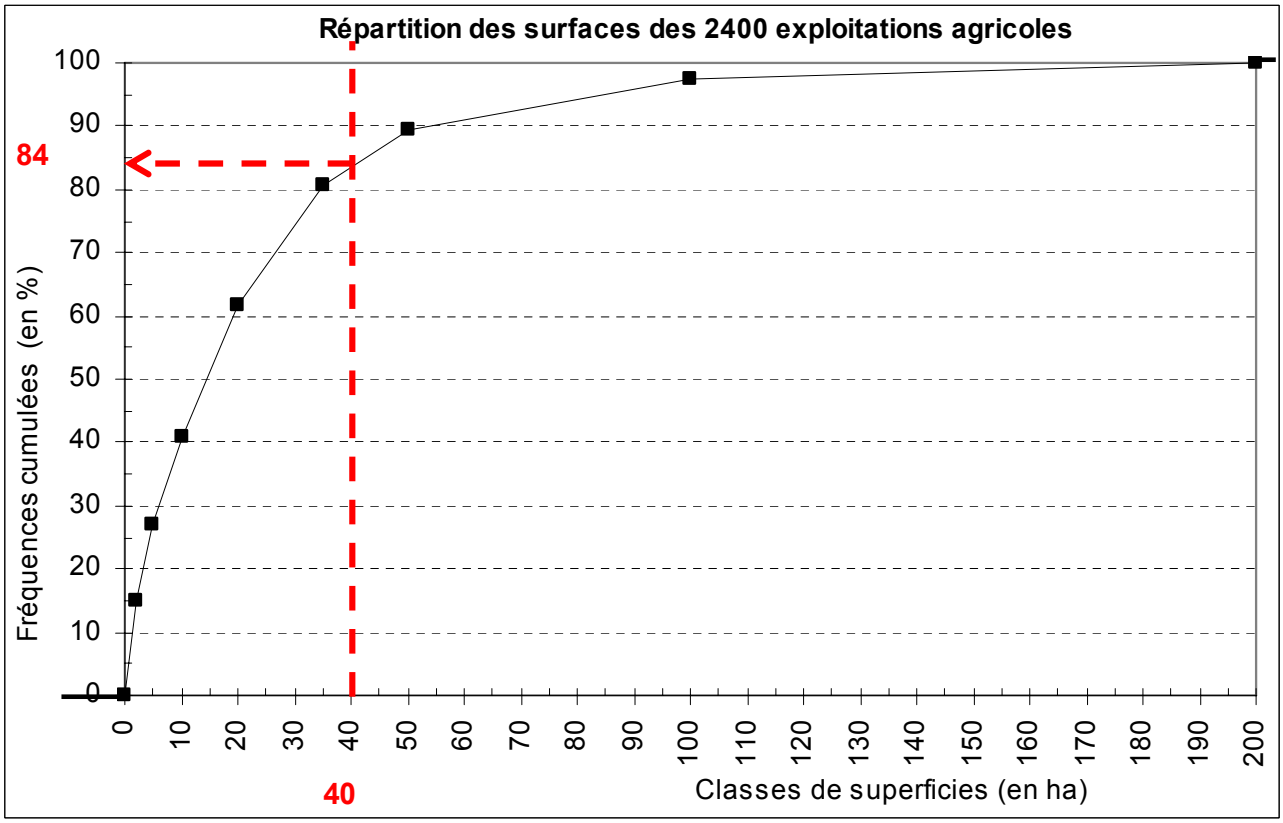
Dans le cas d'une proportion, on peut réaliser un calcul algébrique direct, au moyen d'une interpolation linéaire. On détermine la valeur recherchée F(x) de la fréquence cumulée, de la façon suivante :

$$F(x) = F(ei) + [F(ei+1) - F(ei)] \times (x - ei) / (ei+1 - ei)$$

$$\text{Ici :} \quad F(40) = 80,6 + (89,3 - 80,6) \times (40 - 35) / (50 - 35) = 83,5\ \%$$

Remarque : dans l'histogramme ci-dessous, les fréquences mentionnées au-dessus des rectangles hachurés correspondent aux fréquences **réelles** et non aux fréquences corrigées.



## 4. Points particuliers

### 41. La question des classes non bornées

Exemple : le tableau suivant présente la ventilation des 537 entreprises d'un secteur industriel donné, selon leur taille (exprimée en nombre de personnes employées) :

| Taille des entreprises xi | Classes de tailles xi | Nombre d'entreprises ni | Amplitude des classes ai | Effectif employé par classe |
|---|---|---|---|---|
| **20 - 49** | [ 20 - 50 [ | 265 | 30 | 8 630 |
| **50 - 99** | [ 50 - 100 [ | 72 | 50 | 4 910 |
| **100 - 199** | [ 100 - 200 [ | 72 | 100 | 9 389 |
| **200 - 499** | [ 200 - 500 [ | 63 | 300 | 19 879 |
| **500 et plus** | 500 et plus | 65 | | 252 854 |
| **Total** | | **537** | | **295 662** |

Population statistique : 537 entreprises d'un secteur industriel donné.
Individu statistique : l'une de ces 537 entreprises.
Caractère statistique : la taille des entreprises selon le nombre de personnes employées.
Type du caractère : quantitatif discret traité en continu.

**Étape 1** : en vue de tracer un histogramme et / ou de réaliser des calculs ultérieurs sur cette série, on est amené à créer des classes (2ème colonne du tableau), pour respecter les principes d'exhaustivité et d'incompatibilité associés aux modalités du caractère.

**Étape 2** : quelle valeur va-t-on donner à la borne supérieure de la dernière classe ?

**De manière générale, en l'absence d'informations complémentaires** sur le secteur industriel considéré, on retient le principe suivant :

- **borne supérieure de la dernière classe : on retient comme amplitude de la dernière classe celle de l'avant-dernière classe**.

Ici, l'amplitude de l'avant-dernière classe [200 ; 500[ est de 300. Par conséquent, on fixe la borne supérieure de la dernière classe à : 500 + 300 = 800 personnes employées, soit [500 ; 800[ . Tout se passe comme si, dans la 1ère colonne du tableau de données, on avait : 500-799 , à la place de 500 et plus.

- **borne inférieure de la 1ère classe** : on retient comme amplitude de la 1ère classe :

- soit **éventuellement zéro** (si cette valeur apparaît pertinente) ;
- soit **l'amplitude de la 2ème classe**.

Supposons ici qu'on ait eu "Moins de 50", à la place de "20 - 49". On aurait alors retenu zéro (classe [0 - 50[), cette amplitude coïncidant, dans le cas présent, avec l'amplitude de la 2ème classe, soit 50.

Si, au-delà du tableau de données, on dispose d'**informations complémentaires**, on doit en tenir compte et justifier la valeur retenue pour borner la classe, sur la base de ces informations complémentaires (sinon les représentations graphiques s'en trouveraient faussées, de même que les calculs de moyenne et d'écart-type, qui sont sensibles aux bornes des classes).

Ces précautions méthodologiques sont indispensables si l'on veut éviter, autant que faire se peut, au-delà des seuls calculs mathématiques, des interprétations économiques plus ou moins erronées, selon le niveau de l'erreur commise.

Exemples concrets :

- les entreprises de TCU (Ratp 40 000 employés ; Tcl 4 000 Marseille 3 000 ; Lille 2 600 ; puis une centaine de réseaux, de 10 à 500 employés).

- les entreprises de TRM (38 000 environ, dont les deux tiers de moins de 3 employés).

Pour mémoire, étape 3 :

La colonne de droite du tableau correspond à une information complémentaire importante. Elle indique l'effectif total employé par l'ensemble des entreprises d'une classe de taille donnée.

C'est évidemment l'information portant sur la dernière classe qui nous importe ici.

En divisant l'effectif employé de cette dernière classe par le nombre d'entreprises de cette classe, on obtient le nombre **moyen** d'employés de chaque entreprise de la classe :

$$252\,854 / 65 = 3\,890(,1)$$

Nous justifierons au chapitre 2 (calcul de la moyenne arithmétique, dans le cas d'un caractère quantitatif continu), l'affirmation suivante : **la valeur 3 890(,1) représente le centre $c_i$ de la dernière classe** (hypothèse d'équirépartition des effectifs dans les classes).

Soit a l'amplitude de la dernière classe [ 500 - x [ .

On a alors : $a / 2 = 3890 - 500 = 3\,390$

et donc : $a = 3\,390 \times 2 = 6\,780$ (amplitude de la dernière classe).

Enfin, pour obtenir la valeur de la borne supérieur x de la dernière classe, il suffit d'ajouter cette amplitude (6 780) à la valeur de la borne inférieure de la classe (500) :

$x = 500 + 6\,780 = 7\,280$

Toute précision étant ici quelque peu illusoire, on arrondit la valeur trouvée à 7 300.

D'où finalement, on a l'écriture de la **dernière classe bornée : [ 500 - 7 300 [** .

Conclusion : il y a donc loin de 800 (valeur fixée par défaut, en l'absence d'information complémentaire) à 7 300.

Bien noter que, dans chaque cas, les représentations graphiques (histogramme et courbe cumulée) sont différentes, tout autant que le résultat des calculs de paramètres tels que la moyenne ou l'écart-type. Par ailleurs, l'interprétation socio-économique des résultats peut s'en trouver plus ou moins notablement infléchie.

## 42. La question du regroupement des données en classes (agrégation des données)

Supposons qu'après avoir mené une enquête, nous disposions des données brutes. Celles-ci apparaissent sous forme non classée ("en vrac", ou plus exactement dans l'ordre où elles ont été saisies).

Dans un premier temps, ces données peuvent être ordonnées, sans que l'on cherche à les regrouper en classes. Mais si les données sont trop nombreuses, on doit se résoudre à effectuer un regroupement.

A ce moment-là, deux questions se posent :

a) Quel nombre de classes faut-il créer ?

b) Quelles hypothèses faut-il poser pour créer ces classes ? Quelles bornes va-t-on affecter à ces classes ?

### 421. Nombre optimal de classes à créer

Y a-t-il un nombre optimal de classes, lorsqu'on effectue des regroupements ?

Position du problème :

a) Plus on crée de classes, plus on reste proche de la structure exacte (diversifiée) de la population étudiée. En d'autres termes, on dispose d'une information quasi exhaustive par rapport aux données brutes.

b) Mais les tableaux de traitement seront de grandes dimensions, les graphiques peu lisibles (car l'information, "trop riche", ne permettra pas de dégager les tendances lourdes de la série) et les interprétations pourront devenir difficiles.

Conclusion : la baisse du nombre de classes pallie donc les éléments négatifs précédents, mais au prix d'une perte d'information plus ou moins importante.

Certains auteurs (Monjallon, Aïvazian, …) montrent qu'un compromis acceptable peut être trouvé entre la perte d'information d'une part et une plus grande facilité de manipulation et de traitement des données, ainsi que d'interprétation d'une série d'autre part.

Des regroupements en classes sont recommandés, dès lors que la taille de la population (ou de l'échantillon) sur lequel on travaille dépasse 50 observations : **$n > 50$** et simultanément si le nombre de modalités (variable discrète) est supérieur à 10 : **$k > 10$** .

De manière générale, retenir que le nombre de classes doit osciller entre **8 à 10 au minimum** (car en deçà, on gomme littéralement les tendances générales caractéristiques de la série) et **20 à 25 au maximum** (car au-delà, on perd en clarté et en lisibilité des tendances générales, de même que le traitement des données peut devenir plus complexe du fait de tableaux lourds à manipuler).

Remarque importante : pour des raisons pédagogiques, en td, les exercices portent sur des tableaux à dimensions plus faibles, qui ne correspondent donc généralement pas à des traitements et des analyses réelles pertinentes.

## 422. La structure du découpage en classes

On cherche à répondre à la question de savoir quelles valeurs donner aux extrémités de classes.

Deux méthodes principales sont généralement mises en œuvre :

### 4221. Création de classes d'amplitude identique

L'avantage de cette méthode est de simplifier les calculs, puisqu'il n'est pas nécessaire de corriger les effectifs (ou les fréquences) lorsqu'on trace un histogramme.

La mise en œuvre de cette méthode présente un inconvénient majeur, lorsqu'on a affaire à des séries très dissymétriques (à gauche ou à droite). En effet, on risque alors d'avoir des effectifs très différents d'une classe à l'autre.

Graphiquement, cela empêche de visualiser correctement les grandes tendances de la série. On risque de passer à côté d'une caractéristique importante d'une population, relativement à un caractère donné, parce que l'analyse ne sera pas assez fine pour certaines modalités de la variable.

Exemple

Si l'on s'intéresse à des maladies infantiles, ordonner l'ensemble d'une population (enfants + adultes) selon des classe d'amplitude égale, aura pour conséquence négative de très mal étudier (et donc de visualiser) le phénomène envisagé.

Si l'on considère par exemple des classes d'amplitude de 5 ans, on aura [0-5[, [5-10[, [10-15[, …, alors même que la quasi-totalité des effectifs sera située dans les deux premières, voire la seule première classe.

On a le même problème avec les entreprises de certains secteurs d'activité, quand elles sont étudiées selon leur nombre d'employés (par exemple les TRM), ou encore lorsqu'on traite des données relatives aux salaires ou aux revenus. Dans tous ces cas, les séries sont très dissymétriques (fortement étalées à droite).

### 4222. Création de classes d'effectifs similaires

Le plus souvent possible, pour pallier l'inconvénient précédent, on a intérêt à créer des classes dont l'effectif de chacune sera du même ordre de grandeur.

Le principal avantage de cette méthode est de percevoir aisément les classes "critiques", selon le phénomène étudié : là où les amplitudes de classes sont faibles (à effectifs à peu près égaux), cela signifie qu'une part importante de la population (ou de l'échantillon) concernée est sensible aux valeurs correspondantes des modalités du caractère étudié.

Contrairement à la méthode précédente, les amplitudes de classes peuvent être très différentes. Par exemple, s'il s'agit d'une maladie infantile, nous aurons :

[0-1[, [1-2[, [2-3[, [3-4[, [4-6[, [6-8[, [8-10[, [10-15[, [15-20[, [20-30[, [30-50[, 50 et +.

Remarques générales sur les deux méthodes :

a) si l'on crée trop de classes, on dispose paradoxalement de "trop" d'information. Il en résulte un histogramme à l'allure très accidentée, qui peut empêcher de déduire correctement la tendance générale du phénomène étudié.

b) si l'on crée trop peu de classes, on perd trop d'information essentielle sur la nature du phénomène étudié. Dans ce cas, on risque de gommer plus ou moins complètement les spécificités de la série étudiée.

Conclusion générale sur les regroupements en classes

Méthode et rigueur sont de mise pour réaliser cette opération. Il faut un "doigté", qu'on acquiert avec l'expérience ("art" statistique).

Se souvenir que les résultats de calculs de moyenne ou d'écart-type sont très sensibles aux limites de classes qui vont être retenues. Notamment dans le cas de séries très dissymétriques, selon le nombre de classes et la détermination des extrémités de classes, la valeur de ces paramètres peut varier du simple au double ! (notamment si, en plus, l'hypothèse d'équirépartition des effectifs dans les classes, que l'on pose pour réaliser les calculs, est mal vérifiée à l'intérieur des classes).

### 43. Le traitement des variables discrètes en continu

Exemple

Considérons un échantillon de 60 piles électriques, testées selon leur durée de fonctionnement (en heures révolues) :

| | | | | | |
|---|---|---|---|---|---|
| 147 | 116 | 76 | 149 | 184 | 162 |
| 137 | 144 | 146 | 128 | 136 | 137 |
| 132 | 92 | 186 | 153 | 141 | 130 |
| 143 | 183 | 142 | 134 | 137 | 100 |
| 134 | 150 | 138 | 89 | 135 | 131 |
| 141 | 157 | 122 | 98 | 49 | 110 |
| 149 | 143 | 145 | 146 | 136 | 115 |
| 77 | 153 | 144 | 112 | 65 | 118 |
| 139 | 117 | 146 | 170 | 167 | 140 |
| 135 | 141 | 154 | 150 | 145 | 115 |

Population statistique : 60 piles électriques testées.
Individu statistique : l'une de ces 60 piles.
Caractère statistique : la durée de vie de chaque pile (en heures révolues).
Type du caractère : quantitatif discret traité en continu.

Nous effectuons trois types de regroupements en considérant des effectifs à peu près similaires dans chacune des classes, en confectionnant successivement :

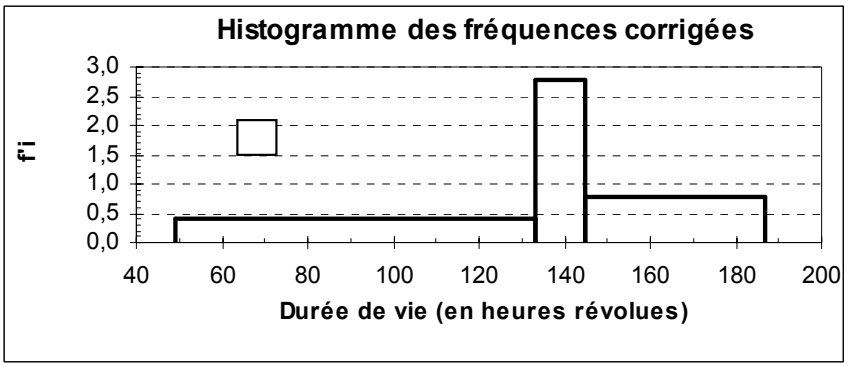
- 3 classes (effectifs de 20 individus par classe) ;

- 10 classes (effectifs de 6 individus par classe) ;

- 6 classes (effectifs de 10 individus par classe).

**Regroupement en classes d'effectifs constants (20 individus par classe)**

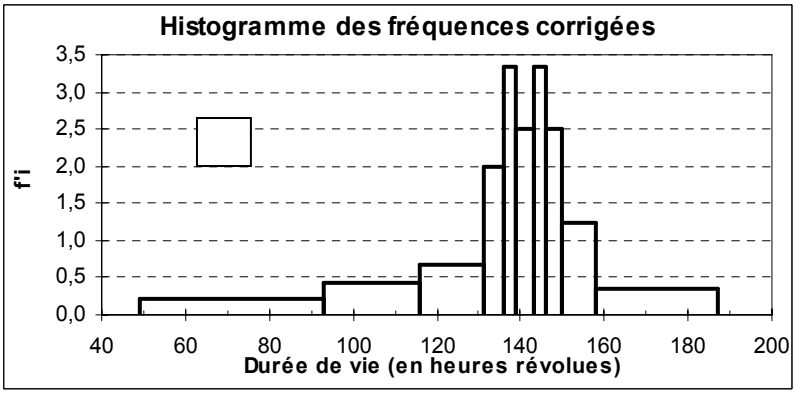
| Durée | Effectifs | Fréq. | ai | f'i corr. |
|---|---|---|---|---|
| **[49-133[** | 20 | 33,3% | 84 | 0,40 |
| **[133-145[** | 20 | 33,3% | 12 | 2,78 |
| **[145-187[** | 20 | 33,3% | 42 | 0,79 |
| | 60 | 100,0% | | |




Manifestement, le nombre de classes retenu ici est trop faible. On gomme à l'excès les caractéristiques de la série. On supprime trop d'information et l'interprétation en est mise à mal.

**Regroupement en classes d'effectifs constants (6 individus par classe)**

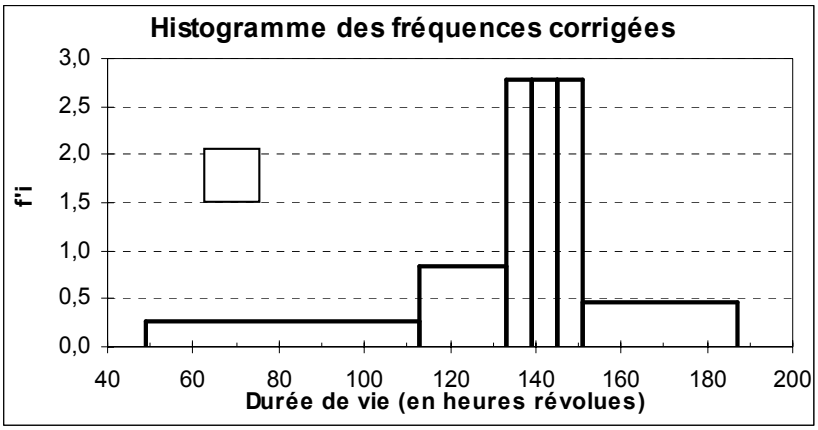
| Durée | Effectifs | Fréq. | Ai | f'i corr. |
|---|---|---|---|---|
| **[49-93[** | 6 | 10,0% | 44 | 0,23 |
| **[93-116[** | 6 | 10,0% | 23 | 0,43 |
| **[116-131[** | 6 | 10,0% | 15 | 0,67 |
| **[131-136[** | 6 | 10,0% | 5 | 2,00 |
| **[136-139[** | 6 | 10,0% | 3 | 3,33 |
| **[139-143[** | 6 | 10,0% | 4 | 2,50 |
| **[143-146[** | 6 | 10,0% | 3 | 3,33 |
| **[146-150[** | 6 | 10,0% | 4 | 2,50 |
| **[150-158[** | 6 | 10,0% | 8 | 1,25 |
| **[158-187[** | 6 | 10,0% | 29 | 0,34 |
| | 60 | 100,0% | | |




On se rapproche de l'allure accidentée du diagramme en bâtons que l'on peut tracer pour cette série. Il y a ici "trop" d'information pour bien cerner l'allure générale du phénomène, notamment dans sa partie centrale. La série est insuffisamment regroupée.

**Regroupement en classes d'effectifs constants (10 individus par classe)**

| Durée | Effectifs | Fréq. | ai | f'i corr. |
|---|---|---|---|---|
| **[49-113[** | 10 | 16,7% | 64 | 0,26 |
| **[113-133[** | 10 | 16,7% | 20 | 0,83 |
| **[133-139[** | 10 | 16,7% | 6 | 2,78 |
| **[139-145[** | 10 | 16,7% | 6 | 2,78 |
| **[145-151[** | 10 | 16,7% | 6 | 2,78 |
| **[151-187[** | 10 | 16,7% | 36 | 0,46 |
| | 60 | 100,0% | | |




Ici, on discerne mieux la tendance principale de la série, sans perdre autant d'information que dans le premier cas de figure, un peu caricatural.

Remarques terminales sur le traitement des variables discrètes en continu :

1) Le tracé d'une courbe cumulée n'est pas différent de celui d'une variable continue par nature. Son tracé est établi sur les limites de classes que l'on retient.

2) Les calculs de moyenne et d'écart-type sont effectués sur une base un peu différente de ceux effectués sur une variable continue par nature. Concrètement, on diminue de une unité la valeur de la borne supérieure de chaque classe avant de calculer le centre de classe, pour tenir compte du fait qu'au-delà de la dernière valeur entière d'une classe donnée, il n'y a aucun individu puisqu'on est parti de valeurs entières (variable discrète). Cette procédure permet de rapprocher les valeurs de la moyenne et de l'écart-type de la réalité (l'hypothèse d'équirépartition des effectifs à l'intérieur des classes ayant tendance à éloigner la valeur de ces paramètres des valeurs obtenues sans regroupement des données en classes).

## CHAPITRE 2 : ÉTUDE DES VARIABLES QUANTITATIVES A UN CARACTÈRE

Nous avons déjà signalé plus haut que, contrairement aux caractères qualitatifs, on peut mener sur les caractères quantitatifs discrets ou continus un certain nombre de calculs qui vont permettre de caractériser précisément les séries statistiques correspondantes. Dans la mesure où ces caractères possèdent des modalités mesurables, on parle indifféremment de caractères ou de variables statistiques discrètes ou continues.

Après avoir ordonné (classé) les données brutes et après avoir représenté graphiquement ces données, on cherche ici à **caractériser les séries quantitatives par des valeurs numériques**. Il s'agit d'une **étape de synthèse, de résumé d'une série**, sous la forme d'un petit nombre de paramètres signifiants.

En effet, les différents graphiques étudiés dans le chapitre 1, si leur représentation est effectuée correctement, permettent une première analyse rapide des séries. Leurs allures générales diverses donnent déjà une idée au lecteur de quelques grandes tendances du phénomène étudié. Toutefois, cette lecture est subjective, restant à la seule appréciation de l'analyste.

Afin de renforcer l'objectivité du jugement de l'opérateur, on réalise un certain nombre de calculs numériques sur les données, qui permettent une analyse plus pertinente. Cette objectivité est nécessaire car, en général, on est amené à comparer la série étudiée avec d'autres séries.

Notamment, si les représentations graphiques à comparer sont relativement proches les unes des autres, il est plus facile de comparer des valeurs numériques, synthétisant chacune des séries étudiées. Globalement, cela permet de simplifier la comparaison et de l'envisager sur une base plus objective.

Si l'on opte pour cette procédure, une question importante se pose immédiatement :

"**Quelle nombre unique doit-on (ou peut-on) utiliser pour résumer la série statistique étudiée, de façon à rendre compte le plus fidèlement possible de celle-ci ?**"

Le statisticien anglais Yule a précisé un certain nombre de conditions que doivent remplir les différents indicateurs, permettant de caractériser une série statistique. Cela de façon à ce que ces indicateurs soient acceptés par tous les opérateurs et interprétés de la même façon partout dans le monde, dans un souci de plus grande objectivité.

Il est notamment important que les définitions ne soient pas ambiguës, sinon des interprétations arbitraires et subjectives pourraient être générées par les analystes.

Pour être statistiquement acceptable, un indicateur particulier (c'est-à-dire un nombre qui résume la série statistique étudiée d'un point de vue déterminé), doit, si possible, vérifier les **six conditions suivantes, appelées conditions (critères) de Yule** :

1) l'indicateur est défini de manière objective.

Cela veut dire qu'un utilisateur de cet indicateur ne peut lui adjoindre un jugement de valeur personnel. Deux utilisateurs différents d'un même indicateur doivent parvenir au même résultat.

2) l'indicateur tient compte de toutes les observations de la série étudiée.

Remarque : parfois (cf. étude des séries chronologiques) on est amené à éliminer certaines valeurs d'une série, qualifiées "d'aberrantes", en vue d'éviter de fausser des calculs de portée générale ayant trait au phénomène étudié (calculs de coefficients ou de rapports saisonniers). Noter qu'il s'agit déjà là de l'introduction d'une certaine subjectivité (jugement de valeur).

3) l'indicateur a une signification concrète parlante.

Même un non spécialiste doit pouvoir saisir aisément l'utilité de tel ou tel indicateur (a contrario l'utilisation des indices de Fisher).

4) l'indicateur doit être facile à calculer.

5) l'indicateur doit être peu sensible aux fluctuations d'échantillonnage.

Par exemple, lorsque dans une même population statistique, on tire plusieurs échantillons différents, la moyenne arithmétique portant sur le caractère étudié ne sera jamais exactement la même d'un échantillon à l'autre. La moyenne est sensible aux fluctuations d'échantillonnage au contraire par exemple de la médiane.

6) l'indicateur doit se prêter facilement au calcul algébrique.

En vue d'assurer une manipulation aisée en statistique mathématique (utilisation de lois de probabilité). De façon également à agréger facilement les résultats proposés par un tel indicateur (passage de plusieurs sous-populations à une population plus large).

La moyenne arithmétique se prête très bien au calcul algébrique, au contraire de la médiane.

Dans un premier temps, si l'on cherche à résumer une série statistique par un seul nombre, il importe de saisir le trait dominant de cette série, l'élément qui revient en force. Ce nombre va donner une indication sur la tendance lourde de la série. On parle de **tendance centrale** de la série étudiée.

Mais nous verrons ensuite que, lors de la comparaison effectuée entre plusieurs populations ou échantillons, on est amené à caractériser les séries par un deuxième nombre (voire au-delà). En effet, deux séries peuvent présenter une caractéristique de tendance centrale très proche et il sera donc nécessaire de les distinguer selon d'autres paramètres.

De manière générale, on caractérise une série quantitative par des indicateurs (paramètres) de tendance centrale, de dispersion, de forme et de concentration.

### 1. Les caractéristiques de tendance centrale (ou de position) d'une série

Ici l'on cherche à résumer une série statistique par un seul nombre. On s'efforce de saisir le trait dominant, la tendance lourde de la série. C'est pourquoi l'on parle d'indicateurs (ou de caractéristiques) de tendance centrale.

Usuellement, on détermine le mode (Mo), la médiane (Me), la moyenne arithmétique ($\bar{x}$), la moyenne géométrique (G) et parfois la moyenne harmonique (H) d'une série.

Chacun de ces indicateurs ne vérifie pas simultanément l'ensemble des six conditions de Yule précédentes. C'est la raison pour laquelle, souvent, certains d'entre eux sont utilisés de manière conjointe, afin de pallier les faiblesses relatives de l'un ou de l'autre.

## 11. Le mode

### 111. Cas des variables discrètes

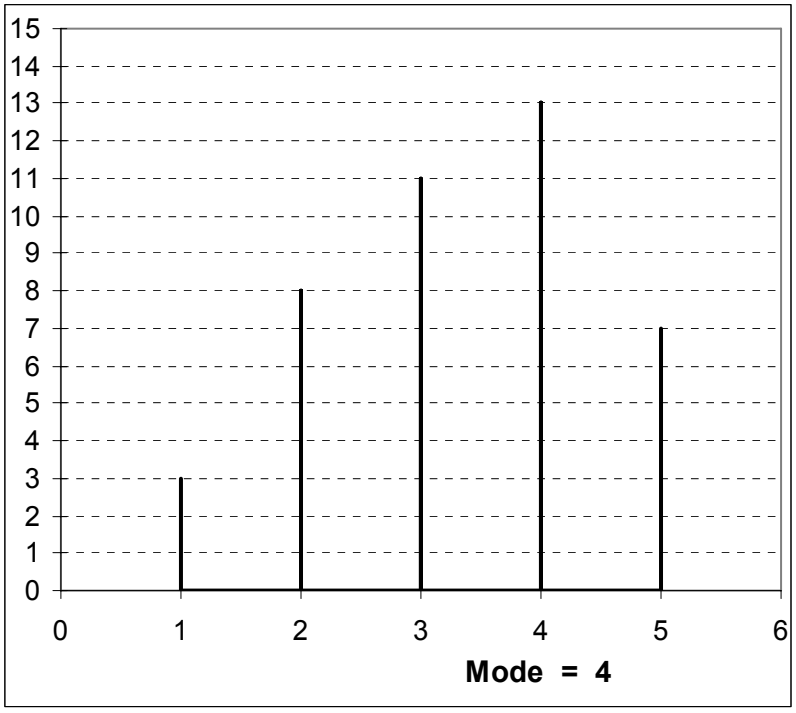
Définition : le mode est la valeur de la modalité ($x_i$) qui correspond à l'effectif $n_i$ (ou la fréquence $f_i$) le plus nombreux.

On détermine la valeur modale directement dans le tableau de données ou bien à partir du diagramme en bâtons (diagramme différentiel).

Exemple

| xi | ni |
|---|---|
| 1 | 3 |
| 2 | 8 |
| 3 | 11 |
| **4** | **13** |
| 5 | 7 |
| | 42 |

Le mode est égal à la valeur de xi correspondant à l'effectif partiel ni le plus grand : **Mo = 4** .




Le mode est égal à la valeur de xi correspondant au bâton le plus grand : **Mo = 4** .

Remarque : il peut exister des distributions plurimodales. En général, cela dénote la présence de deux ou plusieurs sous-populations aux caractéristiques spécifiques, au sein de la population générale étudiée.

### 112. Cas des variables continues

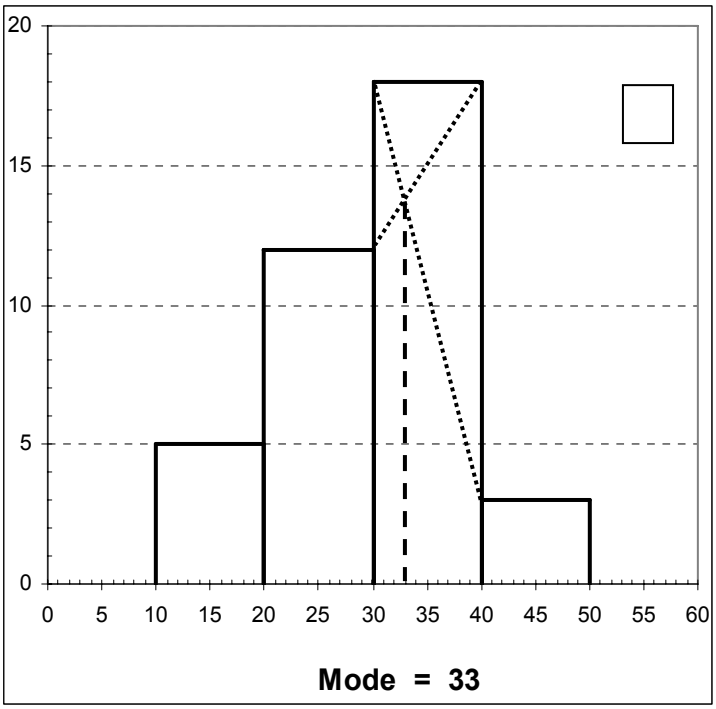
Définition : la classe modale est la valeur de la modalité ($x_i$) qui correspond à l'effectif **corrigé** $n'_i$ (ou la fréquence corrigée $f'_i$) le plus nombreux.

On détermine la classe modale directement dans le tableau de données ou bien à partir de l'histogramme (diagramme différentiel).

Exemple

| xi | ni |
|---|---|
| [ 10 - 20 [ | 5 |
| [ 20 - 30 [ | 12 |
| **[ 30 - 40 [** | **18** |
| [ 40 - 50 [ | 3 |
| | 38 |

Ici, les classes sont d'amplitude égale. Il n'est donc pas nécessaire de corriger les effectifs. À l'effectif le plus élevé (18) correspond la **classe [ 30 - 40 [** , qui est donc la classe modale.




On repère facilement cette même classe modale sur l'histogramme.

Dans le cas d'une variable continue, on peut déterminer une **valeur ponctuelle du mode**, de plusieurs manières :

a) sur la base de l'hypothèse d'équirépartition des effectifs dans les classes, on peut retenir la valeur qui correspond au **centre de la classe modale** (approximation qui sera d'autant meilleure que la série vérifie mieux l'hypothèse posée, et inversement). D'où ici : **Mo = 35** .

b) si l'on a tracé l'histogramme de la série étudiée, la classe modale correspond à l'effectif corrigé $n'_i$ (ou la fréquence corrigée $f'_i$) le plus nombreux. Dans ce cas, on peut obtenir une valeur ponctuelle du mode qui constitue une meilleure approximation, en utilisant la méthode graphique des diagonales, illustrée dans l'exemple ci-dessus.

L'histogramme ci-dessus fait apparaître une classe modale [ 30 - 40 [. Pour obtenir une valeur ponctuelle du mode, on peut considérer le centre de la classe modale, soit 35 (c'est la valeur dont on doit se contenter si l'on ne trace pas l'histogramme et si l'on ne veut pas mettre en œuvre une méthode algébrique, basé sur une double interpolation linéaire). Mais la méthode des diagonales permet ici d'obtenir une valeur plus proche de la réalité de la série, car le rectangle situé à gauche de la classe modale (donc l'effectif de cette classe) est plus important que celui de droite.

Méthode : le 1$^{er}$ segment (la 1$^{ère}$ diagonale) est obtenu en reliant le sommet gauche du rectangle modal et le sommet gauche du rectangle situé immédiatement à sa droite. Le 2$^{ème}$ segment (la 2$^{ème}$ diagonale) est obtenu en reliant le sommet droit du rectangle modal et le sommet droit du rectangle situé immédiatement à sa gauche. La valeur ponctuelle du mode est donnée par l'abscisse du point d'intersection des deux segments, soit ici environ : **Mo = 33** .

c) on peut aussi mettre en œuvre une méthode algébrique, basée sur une double interpolation linéaire (pour mémoire).

Quelle est la meilleure solution ?

En général, cela a peu d'importance sur le fond, car le mode vérifie seulement trois des conditions de Yule (objectivité, détermination concrète, calcul aisé) et, dans la pratique le plus souvent, cela est insuffisant pour un bon résumé de la tendance centrale d'une série statistique.

Le principal inconvénient du mode provient du fait que cet indicateur ne tient pas compte de la totalité des modalités du caractère.

On améliore donc un peu l'efficacité de cet indicateur lorsqu'on utilise la méthode des diagonales (ou celle de la double interpolation linéaire) puisque, dans ces cas, la valeur du mode repose sur la prise en compte de trois modalités du caractère au lieu d'une seule lorsqu'on utilise la méthode du centre de classe.

Souvent cependant, on conserve la classe modale, comme ordre de grandeur qu'on peut utiliser à titre de comparaison avec d'autres caractéristiques de tendance centrale, comme la médiane ou la moyenne arithmétique.

Remarque terminale : dans le cas de séries très dissymétriques à droite ou à gauche, le mode peut largement suffire pour caractériser la tendance centrale de ces séries. Cf. le nombre d'entreprises du secteur des TRM, selon le nombre de salariés.

## 12. La médiane

Définition : la médiane correspond à la modalité ($x_i$) du caractère x qui partage l'effectif total en deux sous-ensembles égaux.

Plus rigoureusement, dans le cas d'une variable discrète, on peut donner la définition suivante :

la médiane (ou valeur médiane) d'une série statistique est la valeur de la modalité ($x_i$) du caractère x qui vérifie simultanément les deux propriétés suivantes :

- pour la moitié au moins des individus de la population, la valeur du caractère est inférieure ou égale à $x_i$ ;

- pour la moitié au moins des individus de la population, la valeur du caractère est supérieure ou égale à $x_i$.

### 121. Cas des variables discrètes

#### 1211. Séries de données classées et non regroupées selon les modalités du caractère étudié

La détermination de la médiane dépend de la parité de la valeur de l'effectif total :

- lorsque l'effectif total n est une valeur impaire, la médiane est égale à la valeur de la modalité qui correspond à la $\frac{n+1}{2}^{\text{ème}}$ observation.

- lorsque l'effectif total n est une valeur paire, l'intervalle médian est égal aux valeurs de la modalité qui correspondent aux $\left(\frac{n}{2};\frac{n}{2}+1\right)^{\text{èmes}}$ observations.

**C'est le rang de la donnée et non sa valeur qui importe !**

Exemples

**n impair** : soit la série : 3 7 10 13 14 15 15

Ici, n = 7 observations.

Par conséquent, conformément au principe énoncé, la médiane correspond à la $\frac{7+1}{2} = 4^{\text{ème}}$ observation, soit la valeur 13 dans la suite des modalités. Donc : **Me = 13**.

**n pair** : soit la série : 3 7 10 13 14 15

Ici, n = 6 observations.

Par conséquent, conformément au principe précédent, l'intervalle médian correspond aux $\left(\frac{6}{2};\frac{6}{2}+1\right)^{\text{èmes}}$, soit les $(3 ; 4)^{\text{èmes}}$ observations, soit l'intervalle (10 ; 13) des modalités.

Donc : **intervalle médian = (10 ; 13)**.

Cas particulier :

Supposons que la série ait été : : 3 7 13 13 14 15

L'intervalle médian devient (13 ; 13) . Dans ce cas, la valeur de la médiane ne correspond plus à un intervalle, mais est égale à une valeur entière unique : **Me = 13**.

### 1212. Séries de données classées et regroupées selon les modalités du caractère étudié

Lorsque les données sont regroupées selon les différentes modalités du caractère, on raisonne de la même façon que précédemment sur les effectifs (ou les fréquences) cumulés.

On peut aussi déterminer graphiquement la médiane ou l'intervalle médian sur le diagramme intégral représentatif de la série discrète (fonction cumulée discontinue en escalier).

Exemple

Nous reprenons ici l'exemple utilisé pour la détermination du mode, mais nous ajoutons une colonne qui correspond aux effectifs cumulés :

| xi | ni | N(x) |
|---|---|---|
| 1 | 3 | 3 |
| 2 | 8 | 11 |
| **3** | 11 | **22** |
| 4 | 13 | 35 |
| 5 | 7 | 42 |
| | **42** | |

La connaissance des effectifs cumulés est nécessaire pour situer le milieu de l'effectif ordonné. Nous avons ici un effectif paire (42). On recherche donc un intervalle médian.

Cet intervalle est compris entre les 21$^{ème}$ et 22$^{ème}$ observations. D'après la colonne de droite du tableau, ces observations correspondent à la modalité 3 du caractère.

On se trouve ici dans le cas particulier signalé précédemment : **Me = 3**.

Pour aider à la compréhension, on peut dégrouper chaque modalité de la série selon ses effectifs (cf. point 1211.) :

1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, …

La médiane, qui est située entre les 21$^{ème}$ et 22$^{ème}$ observations, tombe entre les deux derniers 3 de la série non regroupée.

### 122. Cas des variables continues

Comme pour le mode, on peut déterminer, dans un premier temps, une classe médiane à partir des effectifs (ou des fréquences) cumulés. **On recherche la valeur de la modalité qui correspond à la (n / 2)$^{ème}$ observation** (sauf à raisonner très rigoureusement, on n'a plus à tenir compte de la parité de la valeur de l'effectif total, dans la mesure où la variable est continue).

Exemple

Nous reprenons ici l'exemple utilisé pour la détermination du mode, mais nous ajoutons une colonne qui correspond aux effectifs cumulés :

| xi | ni | ei | N(x) |
|---|---|---|---|
| | | 10 | 0 |
| [ 10 - 20 [ | 5 | | |
| | | 20 | 5 |
| [ 20 - 30 [ | 12 | | |
| | | 30 | 17 |
| [ 30 - 40 [ | 18 | | |
| | | 40 | 35 |
| [ 40 - 50 [ | 3 | | |
| | | 50 | 38 |
| | 38 | | |

Nous avons ici : n / 2 = 38 / 2 = 19. Cela signifie, d'après la colonne de droite du tableau, que la 19$^{ème}$ observation est située dans la classe [ 30 - 40 [ . Par conséquent, la classe médiane est la classe [ 30 - 40 [ . **Classe médiane = classe [ 30 - 40 [** .

Comme pour le mode, on peut déterminer une valeur ponctuelle de la médiane :

- graphiquement, en utilisant le diagramme intégral représentatif de la série continue (fonction cumulée continue). Il suffit pour cela de repérer la valeur n / 2 (ou 50 %, si l'on raisonne en pourcentages) et de tirer un segment horizontal, depuis l'axe des ordonnées jusqu'à rencontrer la courbe cumulée. De là, on tire un segment vertical en direction de l'axe des abscisses. La valeur ponctuelle de la médiane est alors donnée par le point d'intersection du segment vertical et de l'axe des abscisses.

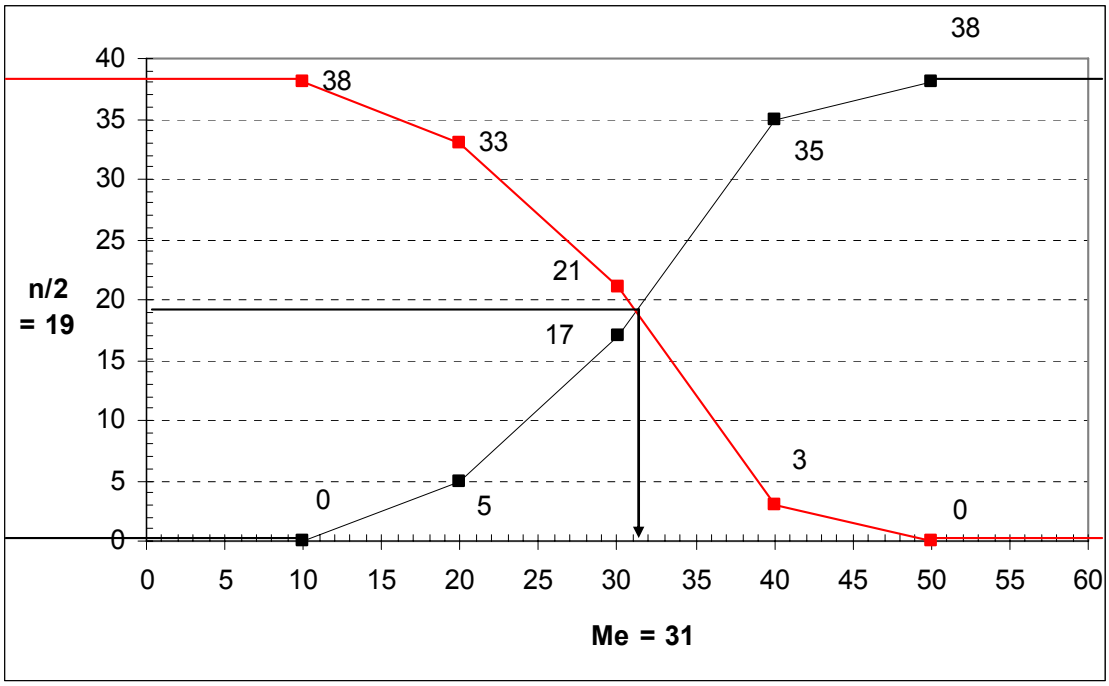
- algébriquement, par interpolation linéaire, par application de la formule :

$$Me = e_i + a_i \times \frac{n/2 - N(e_i)}{n_i}$$

$e_i$ = extrémité inférieure de la classe dans laquelle se trouve la médiane ;
$a_i$ = amplitude de la classe dans laquelle se trouve la médiane ;
n = effectif total ;
$N(e_i)$ = effectif cumulé qui correspond à l'extrémité $e_i$ ;
$n_i$ = effectif de la classe dans laquelle se trouve la médiane.

Remarque : si l'on raisonne en pourcentages, n / 2 est à remplacer par 50 (%) et $N(e_i)$ par $F(e_i) \times 100$.

Détermination graphique




Détermination algébrique par interpolation linéaire

À l'inverse du calcul effectué au chapitre 1, on détermine ici une valeur en abscisse, connaissant une valeur en ordonnée. Cependant, le principe de l'interpolation linéaire reste le même. On a : Lorsqu'on a affaire à une droite, les règles de proportionnalité permettent d'écrire :

$$\frac{Me - 30}{40 - 30} = \frac{n/2 - N(30)}{N(40) - N(30)}$$

D'où l'on tire :

$$Me = 30 + (40-30) \times \frac{n/2 - N(30)}{N(40) - N(30)} = 30 + 10 \times \frac{19-17}{35-17} = 30 + 10 \times \frac{1}{9} = 31(,1)$$

Remarque : on peut également déterminer la valeur de la médiane graphiquement par l'abscisse du point d'intersection des courbes cumulées ascendante et descendante $N(x)$ et $n - N(x)$, ou $F(x)$ et $1 - F(x)$ si l'on raisonne en fréquences (voir graphique ci-dessus).

Remarque terminale : **la médiane est-elle un bon indicateur de tendance centrale** ?

Seule la condition de Yule, relative au calcul algébrique, n'est pas vérifiée par la médiane.

Cela représente un inconvénient majeur dans le cas où la population est composée de plusieurs sous-populations. Dès lors, il n'y a pas de relation algébrique simple entre les médianes de chacune des sous-populations et la médiane de la population totale, car l'on raisonne sur les rangs des observations et non sur les valeurs de ces dernières. C'est la raison pour laquelle, dans un tel cas, il est nécessaire d'utiliser la moyenne arithmétique plutôt que la médiane.

Au contraire, lorsque certaines valeurs extrêmes d'une série sont très grandes par rapport à l'ensemble des autres valeurs (il peut s'agir de valeurs aberrantes), la médiane devient intéressante, comme caractéristiques de tendance centrale, car sa valeur n'est pas influencée par le niveau absolu atteint par certaines données. La médiane vérifie en effet la condition de Yule relative aux fluctuations d'échantillonnage.

Exemple :

Soit la série : 3, 5, 8, 15, 18, 25, 312

Cette série comporte un nombre impair d'observations (n = 7). Par conséquent, la médiane correspond ici à la quatrième observation, soit la valeur 15 : Me = 15 .

Il apparaît clairement ici que le niveau atteint par la dernières données n'est pas représentatif de la valeur centrale de cette série (encore que la faiblesse du nombre d'observations puisse laisser subsister un doute).

Dans ce cas, la valeur de la médiane (15) est nettement plus significative de cette valeur centrale que la moyenne arithmétique, dont on va voir qu'elle prend en compte le niveau (valeur numérique) de chaque donnée.

En effet, si l'on calcule la moyenne arithmétique de cette série, on a : xbar = 386 / 7 = 55,1

Cette valeur est très supérieure à 15 et, simultanément, très supérieure aux six premières valeurs de la série.

Dans un tel cas, on peut penser que 312 est une valeur aberrante (c'est-à-dire exceptionnelle, accidentel). Il peut s'agir d'une erreur de mesure où deux saisies de la donner en question. Si l'on a de bonnes raisons de penser cela, alors on "supprime" (on ne tient pas compte de) la valeur en question.

Avec les six premières valeurs, on a : 3, 5, 8, 15, 18, 25

On a cette fois-ci un intervalle médian (8,15).

Le calcul de la moyenne donne : xbar2 = 74 / 6 = 12,3

On constate que les valeurs de tendance centrale relatées par chacun des deux indicateurs sont maintenant similaires. Dans la mesure où la médiane se prête mal au calcul algébrique, dans ce deuxième cas, on aurait alors de bonnes raisons de choisir la moyenne arithmétique comme indicateur de tendance centrale pertinent.

Remarquons enfin qu'ici le mode est totalement inopérant, quel que soit l'effectif considéré, car les effectifs sont toujours tous égaux à un. Cet indicateur n'apporte donc aucune information intéressante.

## 13. La moyenne arithmétique

Dans la pratique courante, on est amené à utiliser l'une des 3 moyenne suivantes : - la moyenne arithmétique, - la moyenne géométrique, - la moyenne harmonique. Nous aborderons les deux derniers types de moyenne ultérieurement.

### 131. Cas des variables discrètes

Définition : la moyenne arithmétique est la valeur du rapport de la somme des valeurs observées $x_i$ au nombre n des observations.

Moyenne arithmétique simple : $\overline{x} = \frac{1}{n} \sum_{i=1}^{k} x_i$ où : k = nombre de modalités ; n = effectif total.

Rappel : dans le cas d'une moyenne simple, on a : k = n.

Cette formule est utilisable lorsque, à chaque valeur de xi, ne correspond qu'une seule observation.

| xi | ni |
|---|---|
| 3 | 1 |
| 6 | 1 |
| 7 | 1 |
| 10 | 1 |
| 12 | 1 |
| 14 | 1 |
| 18 | 1 |
| **70** | **7** |

Ici xbar = 70 / 7 = 10 .

Il est rare cependant que les effectifs $n_i$ soient tous égaux à un. Dès lors, il est nécessaire de pondérer chaque modalité par les effectifs correspondants, pour mener un calcul correct de la moyenne arithmétique. Ce calcul tient compte des valeurs numériques représentatives des modalités de x.

Moyenne arithmétique pondérée (par les effectifs) : $\overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i$

avec : $n_i$ = effectif partiel correspondant à la modalité $x_i$.

Exemple

| xi | Ni | ni xi |
|---|---|---|
| 1 | 3 | 3 |
| 2 | 8 | 16 |
| 3 | 11 | 33 |
| 4 | 13 | 52 |
| 5 | 7 | 35 |
| | **42** | **139** |

**xbar = 139 / 42 = 3,3(1)** . Nous avions : Mo = 4 et Me = 3 .

Remarque : si l'on raisonne en fréquences, on écrit : $\bar{x} = \sum_{i=1}^{k} f_i x_i$ , avec : $f_i = n_i / n$

### 132. Cas des variables continues

Le calcul de la moyenne arithmétique d'une variable continue nécessite de poser deux hypothèses :

- hypothèse d'équirépartition des effectifs à l'intérieur de chaque classe ;
- centres de classes = moyennes de classes

En effet, dans la mesure où une classe représente une plage de valeurs, il est nécessaire de considérer une valeur ponctuelle de cette classe, afin de pouvoir calculer la moyenne où l'écart-type d'une série statistique. En posant les deux hypothèses précédentes, on privilégie une valeur qui correspond au centre de chaque classe, valeur à partir à partir de laquelle sera effectué le calcul.

On détermine le centre d'une classe en divisant par deux la somme des valeurs de ses extrémités.

Moyenne arithmétique pondérée : $\bar{c} = \frac{1}{n} \sum_{i=1}^{k} n_i c_i$

avec : k = nombre de classes et : $c_i$ = centre de la $i^{ème}$ classe, d'effectif partiel $n_i$.

Si l'on raisonne en fréquences : $\bar{c} = \sum_{i=1}^{k} f_i c_i$ , avec : $f_i = n_i / n$

Exemple

| xi | ni | ci | ni ci |
|---|---|---|---|
| [ 10 - 20 [ | 5 | 15 | 75 |
| [ 20 - 30 [ | 12 | 25 | 300 |
| [ 30 - 40 [ | 18 | 35 | 630 |
| [ 40 - 50 [ | 3 | 45 | 135 |
| | **38** | | **1 140** |

**cbar = 1 140 / 38 = 30** .

Nous avions : classe modale = [ 30 - 40 [ ; classe médiane = [ 30 - 40 [ .

Mo = 35 (centre de classe) ou 33 (méthode des diagonales). Me = 31 (interpolation linéaire).

Remarquons qu'ici, nous avons : **cbar < Me < Mo** . Cette double inégalité indique que la série est légèrement dissymétrique (étalement à gauche ou oblique à droite) cf. plus loin.

Cas des variables discrètes traitées en continu

Dans un tel cas, on a vu plus haut que le tracé de l'histogramme nécessitait des classes contiguës (afin de satisfaire à la propriété d'exhaustivité), mais on sait qu'il n'existe aucun effectif entre la dernière valeur entière d'une classe et sa limite supérieure, puisque les valeurs réelles intermédiaires n'existent pas.

Par conséquent, il n'est pas judicieux de calculer la moyenne sur la base d'un centre de classes déterminé comme pour une variable continue par nature, car on surévalue systématiquement (donc inutilement) la valeur de la moyenne obtenue.

C'est pourquoi on retranche une unité à la borne supérieure de chaque classe avant de réaliser le calcul du centre de la classe correspondante.

### 133. Une propriété remarquable de la moyenne arithmétique

La somme pondérée des écarts de chaque modalité de la variable x à la moyenne est nulle.

On a : $$\sum_{i=1}^{k} n_i (x_i - \overline{x}) = 0$$

Ce résultat est à l'origine de la mise en œuvre de la variance comme indicateur de dispersion d'une série statistique. En effet, ce dernier est fondé sur la somme pondérée des carrés des écarts de chaque modalité de la variable à la moyenne et l'on sait que lorsqu'on utilise le carré d'une valeur, le problème de son signe ne se pose le plus (voir plus loin).

Exemple

| xi | ni | xi - xbar | ni (xi - xbar) |
|---|---|---|---|
| 1 | 3 | -2,31 | -6,93 |
| 2 | 8 | -1,31 | -10,48 |
| 3 | 11 | -0,31 | -3,40 |
| 4 | 13 | 0,69 | 8,98 |
| 5 | 7 | 1,69 | 11,83 |
| | 42 | | **0,00** |

avec : **xbar = 3,31**

cqfd

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

$$\sum_{i=1}^{k} n_i (x_i - \overline{x}) = \sum_{i=1}^{k} n_i x_i - n_i \overline{x} = \sum_{i=1}^{k} n_i x_i - \sum_{i=1}^{k} n_i \overline{x}$$

$$= n\overline{x} - \overline{x} \sum_{i=1}^{k} n_i = n\overline{x} - n\overline{x} = 0$$

fin cqfd

La moyenne arithmétique vérifie les conditions de Yule, sauf celle qui est relative aux fluctuations d'échantillonnage.

Lorsque c'est possible, s'il existe un risque d'avoir plusieurs valeurs aberrantes (accidentelles) pour une série donnée, alors de préférence, on a intérêt à utiliser la médiane qui est peu sensible aux fluctuations d'échantillonnage.

Conclusion

Le mode permet de fixer un ordre de grandeur plus ou moins grossier de la valeur centrale d'une série, dans la mesure où, dans le cas général, on doit travailler sur la totalité des données pour disposer d'un n'indicateur de tendance central pertinent.

Lorsque les fluctuations d'échantillonnage sont importantes, la médiane est un bon indicateur de tendance centrale. Cependant, l'utilisation de cet indicateur n'est pas pertinent lorsqu'on veut extrapoler des résultats à l'ensemble d'une population, à partir d'un échantillon (inférence statistique).

Si les fluctuations d'échantillonnage sont peu importantes et que les risques de valeurs aberrantes sont réduits, la moyenne arithmétique est sans doute le meilleur indicateur de tendance centrale, ce d'autant plus que sa manipulation est aisée en statistique mathématique.

## 2. Les caractéristiques de dispersion

Dans le point 1., on s'est attaché à résumer une série statistique du point de vue de sa caractéristique principale : la tendance centrale de cette série.

Cependant, à lui seul, l'indicateur utilisé (mode, médiane ou moyenne arithmétique) ne répond qu'imparfaitement à une bonne caractérisation de cette série, notamment (ce qui est le plus souvent le cas) lorsqu'on va comparer la série à d'autres séries étudiant le même phénomène.

Ainsi, la médiane ou la moyenne de plusieurs séries peuvent-elles être du même ordre de grandeur, alors même que la distribution des effectifs (qui peuvent être différents) peut présenter des profils très variés selon les séries.

Dans ce cas, il est impératif de pouvoir disposer d'un autre indicateur qui précise la façon dont l'effectif est distribué autour de la valeur centrale.

Exemple

Considérons les deux séries de cinq notes (sur 20) suivantes :

8 9 10 11 12 et : 2 6 10 14 18

Pour chacune de ces séries, la médiane et la moyenne arithmétique sont égales à 10. Par ailleurs, dans chaque cas, le mode est renvoyé par n'importe laquelle des cinq valeurs de la série.

D'autre part, la taille des échantillons est la même (cinq valeurs dans chaque cas).

Il apparaît donc clairement ici que ce n'est pas le nombre qui caractérise la tendance centrale de chaque série qui permet de les comparer de manière significative. Dans le même temps, il apparaît tout aussi clairement que la 2ème série contient des notes bien plus dispersées que la 1ère autour de la valeur moyenne. C'est précisément cette notion d'étalement de la série de part et d'autre de l'indicateur de tendance centrale que l'on va chercher à résumer par un autre nombre.

L'indicateur en question va rendre compte de la dispersion (étalement) de la série, de part et d'autre de la valeur centrale de la série. C'est la raison pour laquelle on parle d'indicateurs (ou de caractéristiques) de dispersion.

Usuellement, on détermine quelquefois l'étendue, mais on utilise surtout l'intervalle (ou le rapport) interquartile, de même que l'écart-type ($\sigma$) et que le coefficient de variation (CV).

Pour mémoire, signalons l'**étendue** (= intervalle de variation = range) d'une série statistique, qui est égale à la différence entre la plus grande et la plus petite valeur de la série. Dans le cas d'une variable continue, c'est la différence entre l'extrémité supérieure de la dernière classe et l'extrémité inférieure de la première classe.

Cet indicateur est peu utilisé, car il est fondé sur les seules valeurs extrêmes d'une série. Assez souvent, ces dernières sont sujettes à de fortes fluctuations d'échantillonnage. Dans certains cas, ces valeurs extrêmes ne reflètent pas les valeurs du reste de la série (valeurs aberrantes). En d'autres termes, la modification de l'une des valeurs extrêmes peut entraîner un changement important de l'ordre de grandeur de l'étendue, ce qui rend cet indicateur peu significatif.

Sur les séries ci-dessus par exemple, on a respectivement :

étendue = 12 - 8 = 4 (série 1) et : étendue = 18 - 2 = 16 (série 2).

### 21. L'intervalle et le rapport interquartiles

Définition : un **quantile** d'ordre k désigne l'une des (k - 1) valeurs d'une variable statistique, qui permettent de partager l'effectif total d'une population en k sous-ensembles d'effectifs égaux.

Si l'on pose k = 2, il existe k - 1 = 1 valeur qui permet de partager l'effectif total en 2 sous-ensembles d'effectifs égaux : il s'agit de la médiane.

Si l'on pose k = 4, il existe k - 1 = 3 valeurs qui permettent de partager l'effectif total en 4 sous-ensembles d'effectifs égaux : il s'agit des **quartiles**.

Définition : un **quartile** désigne l'une des 3 valeurs d'une variable statistique, qui permettent de partager l'effectif total d'une population en 4 sous-ensembles d'effectifs égaux.

Les 3 quartiles (caractéristiques de position) sont notés **$Q_1$, $Q_2$ et $Q_3$**.

Le **quartile $Q_1$** (ou quartile inférieur) est égal à la valeur de la modalité du caractère x, telle que l'effectif des données inférieures à cette valeur représente au plus le quart de l'effectif total et l'effectif des données supérieures à cette valeur représente au plus les trois-quarts de l'effectif total. Si l'on raisonne en pourcentage, la valeur du quartile $Q_1$ est telle que l'on trouve au plus 25 % de l'effectif total en dessous de cette valeur et au plus 75 % de l'effectif total au-dessus de cette valeur.

Le **quartile $Q_2$** représente tout simplement la médiane.

Le **quartile $Q_3$** (ou quartile supérieur) est égal à la valeur de la modalité du caractère x, telle que l'effectif des données inférieures à cette valeur représente au plus les trois quarts de l'effectif total et l'effectif des données supérieures à cette valeur représente au plus le quart de l'effectif total. Si l'on raisonne en pourcentage, la valeur du quartile $Q_3$ est telle que l'on trouve au plus 75 % de l'effectif total en dessous de cette valeur et au plus 25 % de l'effectif total au-dessus.

**L'intervalle interquartile** est défini par la différence : $\mathbf{Q_3 - Q_1}$.

Par construction même des quartiles, l'intervalle interquartile comprend la moitié (n / 2) la plus centrale de l'effectif total. Si l'on raisonne en pourcentage, l'intervalle interquartile comprend les 50 % les plus centraux de l'effectif total.

Remarque : on rapporte parfois la valeur de cet intervalle à celle de $Q_2$ (la médiane) pour caractériser la dispersion d'une série, au moyen d'un nombre sans dimension :

$$\text{interquartile relatif} = \frac{Q_3 - Q_1}{Q_2}$$

On utilise souvent **le rapport interquartile : $\mathbf{Q_3 / Q_1}$**.

Dans ce cas, le résultat obtenu est un nombre sans dimension, qui facilite les comparaisons inter populations.

**Qu'il s'agisse de l'intervalle ou du rapport interquartiles, plus les valeurs obtenues sont élevées, plus la dispersion de la série autour de la médiane est importante**.

Remarque : dans le cas du rapport interquartile, il convient de nuancer l'affirmation précédente, qui reste cependant vraie dans la majorité des cas. Certains exemples montrent en effet que la dispersion n'augmente pas proportionnellement à la valeur obtenue du rapport interquartile. On observe par exemple que si le quartile $Q_1$ augmente et que, dans le même temps l'intervalle interquartile demeure à peu près constant, alors le rapport interquartile diminue !

On considère parfois que l'effectif pris en compte par l'intervalle interquartile (à savoir n / 2) est insuffisant. Par exemple, lorsque l'Insee mène des analyses approfondies sur des données relatives aux revenus. Dans ces cas-là, on choisit souvent 9 valeurs, qu'on appelle **déciles**, et qui partagent l'effectif total d'une population en 10 sous-ensembles d'effectifs égaux.

On définit les déciles de la même façon que les quartiles. Ainsi, le **décile $\mathbf{D_1}$** est égal à la valeur de la modalité du caractère x, telle que l'effectif des données inférieures à cette valeur représente au plus le dixième de l'effectif total et l'effectif des données supérieures à cette valeur représente au plus les neuf dixièmes de l'effectif total. Si l'on raisonne en pourcentage, la valeur du décile $D_1$ est telle que l'on trouve au plus 10 % de l'effectif total en dessous de cette valeur et au plus 90 % de l'effectif total au-dessus de cette valeur. A noter que le décile 5 représente la médiane de la série.

On définit alors un **intervalle interdécile**, par la différence : $\mathbf{D_9 - D_1}$ , qui comprend les 8 / 10 de l'effectif total (ou 80 % de la population totale). De même, on peut définir un **rapport interdécile : $\mathbf{D_9 / D_1}$**.

Dans le cas d'analyse très fines sur les revenus, l'Insee utilise les **centiles**. On dispose alors de 99 **centiles**. Par exemple, le **centile $\mathbf{C_1}$** est égal à la valeur de la modalité du caractère x, telle que l'effectif des données inférieures à cette valeur représente au plus le centième de l'effectif total et l'effectif des données supérieures à cette valeur représente au plus 99 centièmes de l'effectif total. Si l'on raisonne en pourcentage, la valeur du décile $D_1$ est telle que l'on trouve au plus 1 % de l'effectif total en dessous de cette valeur et au plus 99 % de l'effectif total au-dessus de cette valeur. A noter que le centile 50 représente la médiane de la série.

On définit alors un **intervalle intercentile**, par la différence : $\mathbf{C_{99} - C_1}$ , qui comprend 98 centièmes de l'effectif total (ou 98 % de la population totale). De même, on peut définir un **rapport intercentile : $\mathbf{C_{99} / C_1}$**.

Dans le cas des **variables discrètes**, on utilise rarement les intervalles ou les rapports interquartiles. Du fait que les valeurs des modalités sont ponctuelles, on ne peut évidemment retenir que les valeurs entières les plus proches des résultats des calculs. Il en résulte des approximations souvent non pertinentes dans les comparaisons entre plusieurs séries.

Pour mémoire cependant, si l'on tient à réaliser un calcul pour déterminer les quartiles $Q_1$ et $Q_3$, alors on procède de la même façon que pour la médiane.

Exemple 1 (cf. point 1) :

| xi | ni | N(x) |
|---|---|---|
| 1 | 3 | 3 |
| 2 | 8 | 11 |
| 3 | 11 | 22 |
| 4 | 13 | 35 |
| 5 | 7 | 42 |
| | 42 | |

La colonne des effectifs (ou des fréquences) cumulés permet de déterminer les valeurs entières de $Q_1$, $Q_2$ et $Q_3$. Dans la mesure où,ici, l'effectif est pair (n = 42) :

- la valeur de la modalité qui correspond au quartile $Q_1$ est donnée par les (n / 4 ; [n / 4] +1)$^{èmes}$ observations, soit ici les 10,5$^{ème}$ et 11,5$^{ème}$ observations. Comme les valeurs sans nécessairement entières, $Q_1$ correspond à la 11$^{ème}$ observation, soit à la modalité 2 de x. Donc, on a : $Q_1 = 2$.

- la valeur de la modalité qui correspond au quartile $Q_2$ est donnée par les (n / 2 ; [n / 2] +1)$^{èmes}$ observations (intervalle médian), soit ici les 21$^{ème}$ et 22$^{ème}$ observations. Ces dernières sont inférieures ou égales à la 22$^{ème}$ observation, qui correspond à la modalité 3 de x. Donc, on a : $Q_2 = Me = 3$.

- la valeur de la modalité qui correspond au quartile $Q_3$ est donnée par les (3n / 4 ; [3n / 4] +1)$^{èmes}$ observations. soit ici les 31,5$^{ème}$ et 32,5$^{ème}$ observations. Comme les valeurs sans nécessairement entières, $Q_3$ correspond à la 32$^{ème}$ observation, soit à la modalité 4 de x.. Donc, on a : $Q_3 = 4$.

Exemple 2 : enquête portant sur le nombre de pièces de 120 logements :

| Nombre de pièces (xi) | Nombre de logements (ni) | N(x) |
|---|---|---|
| **1** | 11 | 11 |
| **2** | 22 | 33 |
| **3** | 31 | 64 |
| **4** | 25 | 89 |
| **5** | 17 | 106 |
| **6** | 8 | 114 |
| **7** | 3 | 117 |
| **8** | 2 | 119 |
| **9** | 1 | 120 |

La colonne des effectifs (ou des fréquences) cumulés permet de déterminer les valeurs entières de $Q_1$, $Q_2$ et $Q_3$. Dans la mesure où,ici, l'effectif est pair (n = 120) :

- la valeur de la modalité qui correspond au quartile $Q_1$ est donnée par les (n / 4 ; [n / 4] +1)$^{èmes}$ observations, soit ici les 30$^{ème}$ et 31$^{ème}$ observations. Ces dernières sont inférieures ou égales à la 33$^{ème}$ observation, qui correspond à la modalité 2 de x. Donc, on a : $Q_1 = 2$ pièces / logement.

- la valeur de la modalité qui correspond au quartile $Q_2$ est donnée par les (n / 2 ; [n / 2] +1)$^{èmes}$ observations (intervalle médian), soit ici les 60$^{ème}$ et 61$^{ème}$ observations. Ces dernières sont inférieures ou égales à la 64$^{ème}$ observation, qui correspond à la modalité 3 de x. Donc, on a : $Q_2$ = Me = 3 pièces / logement.

- la valeur de la modalité qui correspond au quartile $Q_3$ est donnée par les (3n / 4 ; [3n / 4] +1)$^{èmes}$ observations. soit ici les 90$^{ème}$ et 91$^{ème}$ observations. Ces dernières sont inférieures ou égales à la 106$^{ème}$ observation, qui correspond à la modalité 5 de x. Donc, on a : $Q_3$ = 5 pièces / logement.

Lorsque l'effectif est impair, on recherche la valeur de la modalité qui correspond à la [(n + 1) / 4]$^{ème}$ observation et qui donne la valeur de $Q_1$. Pour $Q_3$, on recherche la valeur de la modalité qui correspond à la [3(n + 1) / 4]$^{ème}$ observation.

Dans le cas des **variables continues**, on détermine une classe pour le quartile $Q_1$ et une classe pour le quartile $Q_3$ de façon similaire à la classe médiane, à partir des effectifs (ou des fréquences) cumulés.

On recherche respectivement la valeur de la modalité qui correspond à la (n / 4)$^{ème}$ observation ($Q_1$) et celle qui correspond à la (3n / 4)$^{ème}$ observation ($Q_3$)

Et comme pour la médiane, la détermination d'une valeur ponctuelle pour $Q_1$ et $Q_3$ s'effectue soit :

- graphiquement, en utilisant le diagramme intégral représentatif de la série continue (fonction cumulée continue).

**Pour $Q_1$**, on repère la valeur n / 4 (ou 25 %, si l'on raisonne en pourcentages) et on tire un segment horizontal, depuis l'axe des ordonnées jusqu'à ce qu'il rencontre la courbe cumulée. De là, on tire un segment vertical en direction de l'axe des abscisses. La valeur ponctuelle de $Q_1$ est alors donnée par le point d'intersection du segment vertical et de l'axe des abscisses.

**Pour $Q_3$**, on repère la valeur 3n / 4 (ou 75 %) et l'on procède de même.

- algébriquement, par interpolation linéaire, par application de la formule :

$$Q_1 = e_i + a_i \times \frac{n/4 - N(e_i)}{n_i} \qquad Q_3 = e_i + a_i \times \frac{3n/4 - N(e_i)}{n_i}$$

$e_i$ = extrémité inférieure de la classe dans laquelle se trouve $Q_1$ (ou $Q_3$) ;
$a_i$ = amplitude de la classe dans laquelle se trouve $Q_1$ (ou $Q_3$) ;
n = effectif total ;
N ($e_i$) = effectif cumulé qui correspond à l'extrémité $e_i$ ;
$n_i$ = effectif de la classe dans laquelle se trouve $Q_1$ (ou $Q_3$).

Remarque : si l'on raisonne en pourcentages :
- pour $Q_1$, n / 4 est à remplacer par 25 (%) et N ($e_i$) par F ($e_i$) x 100 ;
- pour $Q_3$, 3n / 4 est à remplacer par 75 (%) et N ($e_i$) par F ($e_i$) x 100.

Exemple (cf. point 1) :

| xi | ni | ei | N(x) |
|---|---|---|---|
| | | 10 | 0 |
| [ 10 - 20 [ | 5 | | |
| | | 20 | 5 |
| [ 20 - 30 [ | 12 | | |
| | | 30 | 17 |
| [ 30 - 40 [ | 18 | | |
| | | 40 | 35 |
| [ 40 - 50 [ | 3 | | |
| | | 50 | 38 |
| | 38 | | |

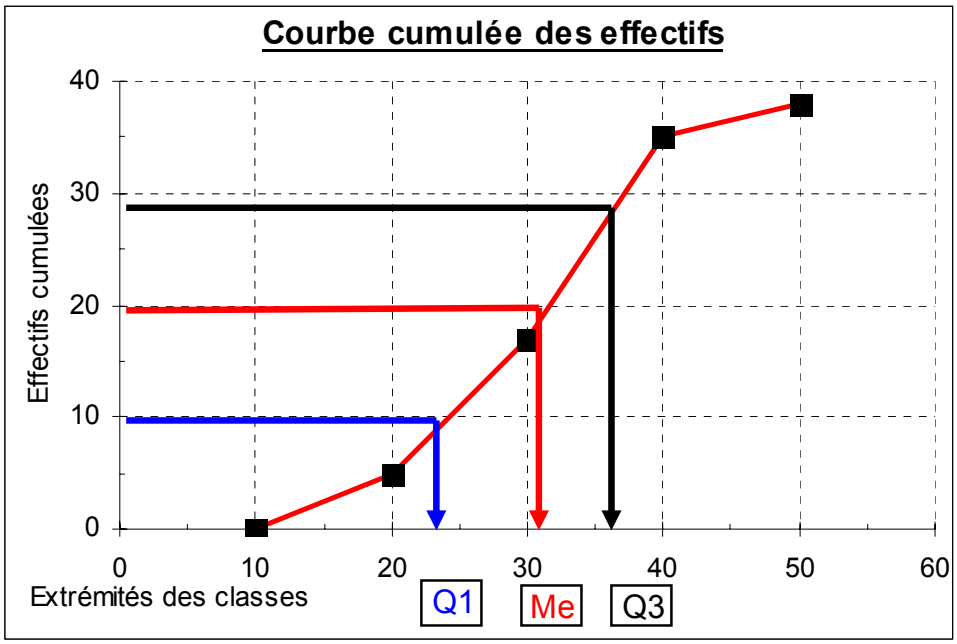
**Pour $Q_1$**, on recherche la valeur de la modalité qui correspond à la $(n / 4)^{ème}$ observation, soit la $9{,}5^{ème}$ observation. La classe $Q_1$ correspond donc à **[20 - 30[** .

**Pour $Q_2$**, on recherche la valeur de la modalité qui correspond à la $(n / 2)^{ème}$ observation, soit la $19^{ème}$ observation. La classe $Q_2$ correspond donc à **[30 - 40[** .

**Pour $Q_3$**, on recherche la valeur de la modalité qui correspond à la $(3n / 4)^{ème}$ observation, soit la $28{,}5^{ème}$ observation. La classe $Q_3$ correspond donc à **[30 - 40[** .

Détermination de la valeur ponctuelle des quartiles

a) Graphiquement (courbe cumulée) :




À l'erreur de lecture graphique près, on obtient environ : **$Q_1$ = 24, Me = 31, $Q_3$ = 36**.

b) algébriquement, par interpolation linéaire :

**Pour $Q_1$ :** $$\frac{Q_1 - 20}{30 - 20} = \frac{n/4 - N(20)}{N(30) - N(20)}$$

$$Q_1 = 20 + (30-20) \times \frac{n/4 - N(20)}{N(30) - N(20)} = 20 + 10 \times \frac{9{,}5-5}{17-5} = 20 + 10 \times \frac{4{,}5}{12} = 23{,}75$$

**Pour $Q_2$ :** $\dfrac{Me - 30}{40 - 30} = \dfrac{n/2 - N(30)}{N(40) - N(30)}$

$$Me = 30 + (40-30) \times \frac{n/2 - N(30)}{N(40) - N(30)} = 30 + 10 \times \frac{19-17}{35-17} = 30 + 10 \times \frac{1}{9} = 31{,}11$$

**Pour $Q_3$ :** $\dfrac{Q_3 - 30}{40 - 30} = \dfrac{3n/4 - N(30)}{N(40) - N(30)}$

$$Q_3 = 30 + (40-30) \times \frac{3n/4 - N(30)}{N(40) - N(30)} = 30 + 10 \times \frac{28{,}5-17}{35-17} = 30 + 10 \times \frac{11{,}5}{18} = 36{,}39$$

L'intervalle interquartile : **$Q_3 - Q_1$** = 36,4 - 23,8 = **12,6** .

Cet intervalle permet de quantifier la dispersion de la partie centrale de l'effectif (la moitié de la population totale est située dans l'intervalle interquartile).

Plus cet intervalle est grand, plus la dispersion autour de la médiane est forte. Mais, de façon générale, cette valeur est à comparer à celle obtenue pour d'autres séries, pour prendre toute sa pertinence.

Le rapport interquartile : **$Q_3 / Q_1$** = 38,4 / 23,8 = **1,61** .

On obtient un nombre sans dimension (indépendant des unités choisies). Cela rend les comparaisons d'échantillons plus pertinente.

**Rappel** : plus la valeur de ce rapport est élevée, plus la dispersion autour de la médiane l'est également. Toutefois, si l'on compare des séries dont les **asymétries** sont très différentes, il se peut que l'affirmation précédente se trouve invalidée.

Remarque terminale sur les quantiles

Que l'on raisonne à partir des quartiles, des déciles ou des centiles, il faut bien avoir présent à l'esprit que seul le rang des observations est pris en compte, et non la valeur des modalités correspondantes.

Si cela représente un avantage, en termes d'insensibilité aux fluctuations d'échantillonnage, assez représente un inconvénient en matière de calcul algébrique effectué sur les écarts ou sur les rapports interquantiles (notamment problème d'agrégation des données d'un ensemble de sous-populations). C'est la raison pour laquelle, en matière de dispersion, on raisonne souvent plus volontiers sur les notions de variance et d'écart-type.

Remarque terminale : les boîtes à moustaches (box and whiskers diagrams)

Lorsqu'on réalise des comparaisons entre diverses populations, on peut compléter le calcul de l'écart interquartile par le tracé d'un graphique appelé boîte à moustaches.

La boîte est limitée par les quartiles $Q_1$ et $Q_3$. Une barre intérieure à la boîte correspond à la médiane.

La moustache supérieure correspond au décile $D_9$ (ou à n) et la moustache inférieure correspond au décile $D_1$ (ou à zéro).

En prenant $D_1$ et $D_9$, au lieu de zéro et n, on cherche à éliminer d'éventuelles valeurs aberrantes. On peut d'ailleurs indiquer sur le diagramme les observations aberrantes entre zéro et $D_1$ d'une part, entre $D_9$ et n d'autre part.

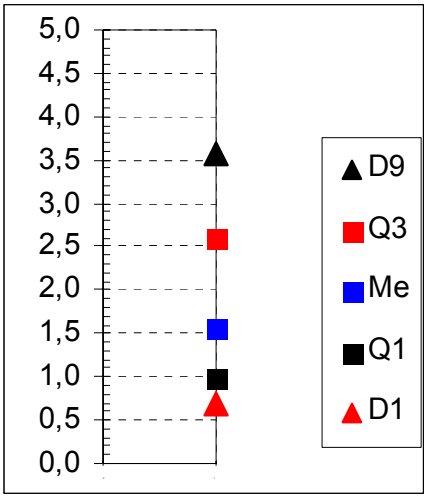
L'intérêt de cette représentation est de pouvoir comparer visuellement des populations différentes, même si les tailles de ces populations ne sont pas les mêmes. Le diagramme donne une idée de la dispersion associée à chacune de ces populations (ou échantillons).

Exemple

On considère ici un échantillon de personnes, interrogées sur leurs revenus mensuels (l'échelle, sur la gauche du diagramme, indique les niveaux de revenus mensuels en K€).

Le traitement statistique des données a fait apparaître les résultats suivants :

**$D_1$ = 0,69 K€ ; $Q_1$ = 0,99 K€ ; $Q_2$ (Me) = 1,56 K€ ; $Q_3$ = 2,60 K€ ; $D_9$ = 3,56 K€ .**




## 22. La variance et l'écart-type

Dans le point 1., on a vu que : $\sum_{i=1}^{k} n_i (x_i - \bar{x}) = 0$. Ce résultat signifie qu'il n'est pas possible d'utiliser directement les écarts à la moyenne arithmétique pour avoir une mesure de la dispersion d'une série autour de cette moyenne.

De quelles solutions possibles dispose-t-on ?

Pour mémoire, signalons l'écart absolu moyen par rapport à la moyenne arithmétique :

$$\bar{e}_{\bar{x}} = \frac{1}{n} \sum_{i=1}^{k} n_i \, | x_i - \bar{x} |$$

et l'écart absolu moyen par rapport à la médiane : $\bar{e}_{Me} = \frac{1}{n} \sum_{i=1}^{k} n_i \mid x_i - Me \mid$

Ces deux expressions sont parfois utilisées dans la pratique, mais la manipulation des valeurs absolues par le calcul algébrique est généralement malaisée.

En pratique, le plus souvent, on retient une méthode qui consiste à élever au carré les écarts $(x_i - \bar{x})^2$, de façon à faire disparaître les signes "moins", qui annulent la somme pondérée. D'où les notions de variance et d'écart-type.

### 221. Définitions

**Définition de la variance**

Soit x un caractère statistique qui prend k modalités $x_i$, auxquels sont associés des effectifs $n_i$ .
La variance $\sigma_x^2$ ou V (x) est la moyenne arithmétique, pondérée par les effectifs, des carrés des écarts à la moyenne arithmétique $\bar{x}$ de la série :

$$\sigma_x^2 = V(x) = \frac{1}{n} \sum_{i=1}^{k} n_i \, (x_i - \bar{x})^2 \qquad \text{avec : } n = \sum_{i=1}^{k} n_i$$

Si l'on raisonne en fréquences, on a :

$$\sigma_x^2 = V(x) = \sum_{i=1}^{k} f_i \, (x_i - \bar{x})^2 \qquad \text{avec : } f_i = \frac{n_i}{n}$$

**Par définition, une variance ne peut jamais être négative, puisque l'on raisonne sur des carrés, associés à des effectifs (ou à des fréquences) qui sont toujours positifs.**

**La variance rend compte des écarts (donc de la dispersion) des observations à la moyenne de celles-ci. Au minimum donc, ces écarts peuvent être égaux à zéro si toutes les observations dont on dispose prennent la même valeur de la modalité du caractère envisagé et, dans ce cas, chacune des observations est égale à la moyenne de celle-ci.**

**Définition de l'écart-type** : c'est la racine carrée de la variance :

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^{k} n_i \, (x_i - \bar{x})^2} = \sqrt{\sum_{i=1}^{k} f_i \, (x_i - \bar{x})^2}$$

**On utilise l'écart-type, et non la variance, comme indicateur de dispersion, de façon à ce que l'unité dans laquelle il est exprimé soit la même que celle de la moyenne arithmétique associée**. Cela permet donc de conserver l'homogénéité des unités.

Interprétation : plus la valeur de l'écart-type est élevée, plus la dispersion de la série autour de la moyenne arithmétique est importante.

## 222. Propriétés de la variance et de l'écart-type

**Propriété 1**

Considérons l'expression générale : $\sum_{i=1}^{k} n_i \, (x_i - a)^2$, dans laquelle **a** est un nombre réel quelconque, différent de zéro.

On démontre que lorsque : $a = \overline{x}$, la somme qui correspond à l'expression ci-dessus est minimale.

Cela signifie que parmi tous les écarts possibles envisageables, on a retenu, pour la définition de la variance et de l'écart-type, l'expression vue au point 221., pour la raison que la somme des carrés des écarts la moyenne arithmétique est **minimale**.

Ce résultat est aussi à l'origine de la notion de droite des moindres carrés (cf. partie 2).

cqfd

Soit : $S = \sum_{i=1}^{k} n_i \, (x_i - a)^2$, c'est-à-dire que l'on considère la somme des carrés des écarts, pondérée par les effectifs partiels, des valeurs d'une série à une valeur réelle quelconque a, différente de zéro.

Le résultat de cette expression dépend de la valeur de a. Comme pour toute fonction, on peut chercher à savoir si elle possède des extrema (minima au maxima). Cela revient à calculer la dérivée de cette expression par rapport à a, considérée comme variable.

On développe tout d'abord le carré, puis on sépare les différents termes selon le signe $\Sigma$ :

$$S = \sum_{i=1}^{k} n_i \, (x_i^2 - 2ax_i + a^2) = \sum_{i=1}^{k} (n_i x_i^2 - 2an_i x_i + a^2 n_i)$$

$$= \sum_{i=1}^{k} n_i x_i^2 - 2a \sum n_i x_i + a^2 \sum n_i = \sum_{i=1}^{k} n_i x_i^2 - 2an\overline{x} + a^2 n$$

$$\text{car : } = \overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i \, x_i \Leftrightarrow n\overline{x} = \sum_{i=1}^{k} n_i \, x_i$$

Dérivons l'expression de S par rapport à a : $\frac{\partial S}{\partial a} = -2n\overline{x} + 2na = 2n\,(a - \overline{x}) = 0$

Il vient : $a = \overline{x}$ et pour savoir si l'on a affaire à un maximum ou un minimum, on calcule la dérivée seconde de S par rapport à a et l'on obtient : $\frac{\partial^2 S}{\partial a^2} = 2n$. La valeur obtenue est toujours positive puisqu'elle ne représente l'effectif de la population (ou de l'échantillon), lui-même toujours positif.

On sait (cf. cours de math) que lorsque la dérivée première s'annule et que, simultanément, la dérivée seconde est positive, on est à un minimum de la fonction.

Par conséquent la somme des carrés des écarts de la moyenne arithmétique est bien minimale.

fin cqfd

**Propriété 2**

A partir du théorème de König, on démontre que l'on peut écrire l'expression de l'écart-type sous la forme suivante :

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2} = \sqrt{\sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2}$$

**Ces formulations représentent des formes opératoires bien adaptées au calcul concret d'un écart-type, une fois qu'on a calculé la moyenne arithmétique.**

Le seul inconvénient de cette nouvelle expression est purement formel, à savoir que la somme des carrés des écarts à la moyenne n'apparaît plus directement.

cqfd

Développons l'expression de la variance, ainsi que nous l'avons fait dans le cas général précédent :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \frac{1}{n}\sum_{i=1}^{k} (n_i x_i^2 - 2\bar{x}n_i x_i + \bar{x}^2 n_i)$$

$$= \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \frac{2\bar{x}}{n}\sum n_i x_i + \frac{\bar{x}^2}{n}\sum n_i = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \frac{2n\bar{x}^2}{n} + \frac{\bar{x}^2 n}{n}$$

$$\text{car : } \sum n_i x_i = n\bar{x} \quad \text{et : } \sum n_i = n$$

$$\text{D'où : } \sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2$$

fin cqfd

Remarque : dans le cas général, le théorème de König se démontre de la manière suivante :

$$\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - a)^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x} + \bar{x} - a)^2 = \frac{1}{n}\sum_{i=1}^{k} n_i [(x_i - \bar{x}) + (\bar{x} - a)]^2$$

$$= \frac{1}{n}\sum_{i=1}^{k} [n_i (x_i - \bar{x})^2 + 2n_i (x_i - \bar{x})(\bar{x} - a) + n_i (\bar{x} - a)^2]$$

$$\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 + \frac{2(\bar{x} - a)}{n}\sum n_i (x_i - \bar{x}) + \frac{(\bar{x} - a)^2}{n}\sum n_i$$

$$\text{D'où : } \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - a)^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 + (\bar{x} - a)^2$$

$$\text{car : } \sum_{i=1}^{k} n_i (x_i - \bar{x}) = 0 \quad \text{et : } \sum n_i = n$$

Si l'on pose : a = 0, il vient :

$$\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 + \bar{x}^2 \text{ et donc : } \sigma^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2$$

Remarque : sous sa forme opérationnelle, l'écart-type comprend un premier terme qui est l'expression d'une moyenne particulière : la moyenne quadratique (quadra = carré), qui est la valeur du rapport de la somme des carrés des valeurs observées $x_i$ , pondérés par les effectifs partiels $x_i$ , au nombre n des observations.

C'est la raison pour laquelle l'écart-type est aussi appelé écart quadratique moyen. En effet, à ce premier terme, on retranche $\bar{x}^2$ , qui est le carré de la moyenne arithmétique.

**Propriété 3**

Dans le cas des variables continues, le traitement des informations regroupées en classe nécessite que l'on pose les deux hypothèses suivantes (cf. moyenne arithmétique) :

- équirépartition des effectifs dans chaque classe ;
- moyenne d'une classe = centre de la classe correspondante.

Comme pour la moyenne, on réalise le calcul de l'écart-type sur la base des centres de classes.

**Remarque terminale importante**

On peut aisément concevoir que la moyenne arithmétique et l'écart type, calculés à partir des valeurs exactes des données contenues à l'intérieur de chaque classe, sont différents de ceux que l'on obtient un à partir des centres de classes.

En pratique, selon que la distribution est à peu près symétrique ou au contraire asymétrique à gauche ou à droite, on peut distinguer trois cas :

Premier cas : la distribution est à peu près symétrique

On peut démontrer que la moyenne arithmétique obtenue avec les centres de classes est assez proche de celle qu'on obtiendrait avec les valeurs exactes des données à l'intérieur de chaque classe.

En effet, on peut observer (par exemple en effectuant des simulations de regroupements en classes de valeurs discrètes) que :

- dans les classes situées avant le mode, les moyennes calculées selon l'hypothèse des centres de classes tendent à surestimer les moyennes qui seraient calculées avec les vraies valeurs des données correspondantes ;

- inversement, dans les classes situées après le mode de la série, les moyennes calculées sur les centres de classes tendent à sous-estimer les moyennes calculées avec les vraies valeurs.

Ainsi, dans la mesure où la distribution des effectifs est à peu près symétrique, il en résulte une certaine compensation et il n'y a donc pas lieu de corriger la moyenne arithmétique calculée sur la base des centres de classes.

Il n'en a pas de même pour la variance ou l'écart-type, car leur formulation est basée sur les carrés des écarts. Donc, les "erreurs" s'ajoutent et ne se compensent pas (pas d'écarts négatifs). On peut donc observer que la variance calculée sur la base des centres de classes est systématiquement surévaluée, dans le cas d'une distribution à peu près symétrique.

C'est la raison pour laquelle le statisticien américain Sheppard a proposé un correctif, qu'on appelle correction de Sheppard :

$$\sigma^2 \text{ corrigée } = \sigma^2 \text{ calculée (sur la base des } c_i) - \frac{a^2}{12}$$

Dans l'expression, a est l'amplitude des classes. Si les amplitudes des classes sont différentes, on retient l'amplitude qui revient le plus souvent ou bien la valeur moyenne des amplitudes.

Deuxième cas : la distribution est étalée à droite

La moyenne, calculée sur la base des centres de classes, est en général sous-évaluée, ce d'autant plus que la dissymétrie de la série est prononcée. Malgré ce constat, il n'existe pas de correctif général utilisable (soit on applique des méthodes au cas par cas, soit on n'effectue aucune correction).

La variance calculée sur la base des centres de classes est ici correctement estimée. Aucune correction n'est donc nécessaire.

Troisième cas : la distribution est étalée à gauche

La moyenne, calculée sur la base des centres de classes, est en général surestimée, ce d'autant plus que la dissymétrie de la série est prononcée. Malgré ce constat, il n'existe pas de correctif général utilisable (soit on applique des méthodes au cas par cas, soit on n'effectue aucune correction).

La variance calculée sur la base des centres de classes est ici correctement estimée. Aucune correction n'est donc nécessaire.

### 223. Le coefficient de variation

Lorsqu'on compare des populations différentes, il est souvent plus commode de disposer d'indicateurs sans dimension (les unités physiques de mesure n'entrent alors plus en ligne de compte).

C'est pourquoi on complète l'étude de la dispersion d'une variable statistique par la donnée du **coefficient de variation qui est égal au rapport de l'écart-type à la moyenne arithmétique** :

$$CV = \frac{\sigma}{\bar{x}}$$

**Plus ce coefficient (qui ne peut être négatif) augmente, plus la dispersion de la série est importante autour de la valeur moyenne**. On retiendra les ordres de grandeur suivants :

$0 \leq CV \leq 0{,}2$ ou $0{,}3$ $\Rightarrow$ dispersion faible
$0{,}3 \leq CV \leq 0{,}5$ ou $0{,}6$ $\Rightarrow$ dispersion moyenne
$CV \geq 0{,}6$ $\Rightarrow$ dispersion forte

CV = 0 signifie que toutes les valeurs de la série sont identiques.
CV > 1 signifie que l'on a $\sigma > \overline{x}$ , ce qui représente une dispersion des données extrêmement forte autour de la moyenne arithmétique.

### 3. Les indicateurs de forme

Deux distributions peuvent avoir la même médiane et le même écart interquartile, la même moyenne arithmétique et le même écart-type. Dans un tel cas (et même lorsqu'il n'en est pas ainsi, car cela de restituer un peu plus d'information sur les séries étudiées), on peut mettre en œuvre une nouvelle batterie d'indicateurs, dans le but de différencier deux ou plusieurs séries.

Ces indicateurs viennent compléter la caractérisation d'une série statistique en donnant un ordre de grandeur de la physionomie de cette série, par le calcul d'un nombre unique, sans avoir besoin de tracer un graphique.

A souligner que la physionomie d'une série peut être pressentie à l'allure du tableau de données. Par exemple, dans de nombreux cas, on peut assez facilement se rendre compte du fait que les effectifs les plus nombreux sont situés plutôt à l'une ou à l'autre extrémité de la série, ou bien au contraire au milieu de cette dernière.

**Le coefficient de Yule**

Il s'agit de l'un des indicateurs de forme les plus usités. Il indique quantitativement l'intensité plus ou moins forte de l'asymétrie d'une distribution vers la droite ou vers la gauche.

Ce coefficient est basé sur l'utilisation des quartiles $Q_1$, $Q_2$ et $Q_3$.

Le principe consiste à comparer les écarts $Q_3 - Q_2$ et $Q_2 - Q_1$ , et de rapporter la différence à l'écart interquartile $Q_3 - Q_1$ , ce qui permet d'obtenir un nombre sans dimension :

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

Interprétation :

$s_Y > 0$ signifie qu'on a un étalement à droite, qui augmente avec la valeur de $s_Y$.

$s_Y < 0$ signifie qu'on a un étalement à gauche, qui augmente avec la valeur (absolue) de $s_Y$.

$s_Y = 0$ signifie que la distribution est symétrique.

Remarque : bien qu'en soi cela ne représente pas une mesure de l'asymétrie d'une série, en l'absence de graphique, les positions respectives des 3 caractéristiques de tendance centrale (Mo, Me et $\overline{x}$) peuvent indiquer, en première approximation, au-delà de l'examen du tableau de données, le caractère plus ou moins asymétrique de la série.

Ainsi, lorsque la double inégalité suivante est vérifiée : **Mo < Me <** $\overline{x}$ , alors on peut en déduire que **la distribution est étalée à droite** (résultat qui peut être confirmé par le tracé du diagramme différentiel ou par le calcul d'un coefficient d'asymétrie).

Inversement, lorsque la double inégalité suivante est vérifiée : **Mo > Me >** $\overline{x}$ , alors on peut en déduire que **la distribution est étalée à gauche**.

**Pour mémoire**, signalons d'autres indicateurs de forme.

### Le coefficient de Pearson

Ce coefficient mesure **l'asymétrie** d'une distribution en comparant les positions de la moyenne arithmétique et du mode, en rapportant la différence à l'écart-type de la distribution, ce qui permet d'obtenir un nombre sans dimension : $s_P = \frac{\bar{x} - Mo}{\sigma}$

$s_p > 0$ signifie qu'on a un étalement à droite, qui augmente avec la valeur de $s_p$.
$s_p < 0$ signifie qu'on a un étalement à gauche, qui augmente avec la valeur (absolue) de $s_p$.
$s_p = 0$ signifie que la distribution est symétrique.

Le calcul de ce coefficient est simple, mais son résultat est sujet à variations, dans le cas d'une variable continue, selon que l'on retient le centre de la classe modale ou que l'on applique la méthode graphique des diagonales, pour déterminer une valeur ponctuelle du mode.

### Moments statistiques et coefficients de Ronald Fisher

Les **moments simples d'ordre r** sont définis par : $m_r = \frac{1}{n} \sum_{i=1}^{k} n_i x_i^r$

Lorsqu'on pose : r = 1, on obtient : $m_1 = \bar{x} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i$

Lorsqu'on pose : r = 2, on obtient : $m_2 = \frac{1}{n} \sum_{i=1}^{k} n_i x_i^2 = \sigma^2 + \bar{x}^2$

Ainsi, la variance de x peut s'écrire : $\sigma^2 = m_2 - m_1^2$

Les **moments centrés (sur la moyenne) d'ordre r** sont définis par :

$$\mu_r = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{x})^r$$

Lorsqu'on pose : r = 2, on obtient : $\mu_2 = \sigma^2 = m_2 - m_1^2$

Lorsqu'on pose : r = 3, on obtient : $\mu_3 = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{x})^3$

Lorsqu'on pose : r = 4, on obtient : $\mu_4 = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \bar{x})^4$

### Coefficient $\gamma_1$ de Fisher

Le coefficient $\gamma_1$ ("gamma 1") de Fisher (nombre sans dimension) est un coefficient qui mesure l'intensité de **l'asymétrie** d'une distribution, en mettant en œuvre la moyenne arithmétique et l'écart-type, dans la logique des moments centrés d'ordre 3 (ou des moments simples d'ordre 3, comme c'est aussi le cas pour la variance à l'ordre 2) :

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}\right)^3} = \frac{m_3 - 3 m_1 m_2 + 2 m_1^3}{\left(\sqrt{m_2 - m_1^2}\right)^3}$$

$\gamma_1 > 0$ signifie qu'on a un étalement à droite, qui augmente avec la valeur de $\gamma_1$ .

$\gamma_1 < 0$ signifie qu'on a un étalement à gauche, qui augmente avec la valeur (absolue) de $\gamma_1$ .

$\gamma_1 = 0$ signifie que la distribution est symétrique.

Cette expression apparemment compliquée est souvent utilisée en statistique mathématique, parce qu'elle est aisément manipulable algébriquement (basée sur l'utilisation de la moyenne est de l'écart type).

### Coefficient $\gamma_2$ de Fisher

Le coefficient $\gamma_2$ de Fisher (nombre sans dimension) est un coefficient qui mesure le degré **d'aplatissement** d'une distribution, par rapport à une distribution normale (dont les paramètres sont déterminés à partir de $\bar{x}$ et $\sigma$).

Ici encore, on met en œuvre la moyenne arithmétique et l'écart-type, dans la logique des moments centrés d'ordre 4 (ou des moments simples d'ordre 4) :

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^4}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}\right)^4} - 3 = \frac{m_4 - 4 m_1 m_3 + 6 m_1^2 m_2 - 3 m_1^4}{\left(\sqrt{m_2 - m_1^2}\right)^4} - 3$$

$\gamma_2 > 0$ signifie que la distribution est leptocurtique, c-à-d d'autant plus pointue par rapport à la loi normale que la valeur de $\gamma_2$ est élevée.

$-2 < \gamma_2 < 0$ signifie que la distribution est platicurtique, c-à-d d'autant plus plate par rapport à la loi normale que la valeur de $\gamma_2$ tend vers - 2 (en - 2, la courbe est plate, dénotant une équirépartition totale des observations).

$\gamma_2 = 0$ signifie que la distribution possède un aplatissement du même ordre que la loi normale.

## 4. Exercices récapitulatifs sur les séries quantitatives continues

**Exercice 1**

### Salaires horaires d'une entreprise (en €)

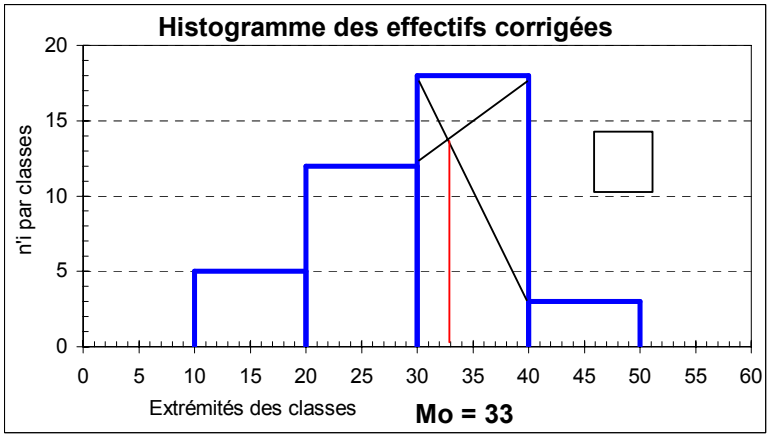
Population statistique : 38 salariés d'une entreprise
Individu statistique : l'un des 38 salariés
Caractère statistique : le salaire horaire (en €)
Type du caractère : quantitatif continu

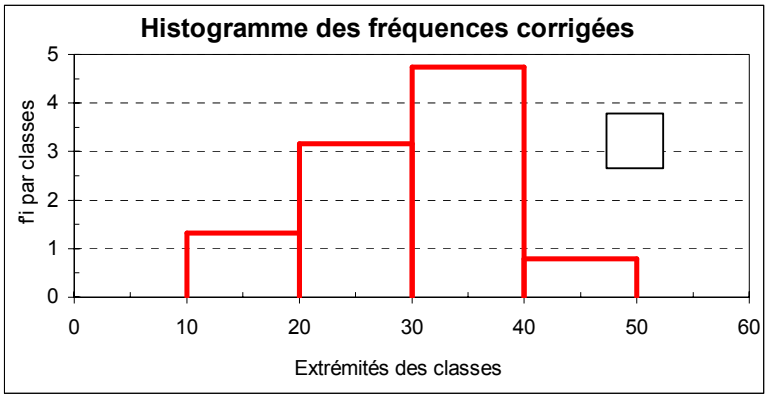
| Salaires horaires | Borne inf | Borne sup | ci | ai | ni | n'i corr. | N(x) | fi.100 | f'i corr. | F(x) | ni ci | ni ci² |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 0 | 0 | | 0,00 | 0 | | |
| [ 10 - 20 [ | 10 | 20 | 15 | 10 | 5 | 5 | 5 | 13 | 1,32 | 13 | 75 | 1 125 |
| [ 20 - 30 [ | 20 | 30 | 25 | 10 | 12 | 12 | 17 | 32 | 3,16 | 45 | 300 | 7 500 |
| [ 30 - 40 [ | 30 | 40 | 35 | 10 | 18 | 18 | 35 | 47 | 4,74 | 92 | 630 | 22 050 |
| [ 40 - 50 [ | 40 | 50 | 45 | 10 | 3 | 3 | 38 | 8 | 0,79 | 100 | 135 | 6 075 |
| **Totaux** | | | | | 38 | | | 100 | | | 1 140 | 36 750 |

**Diagramme différentiel = histogramme**

- Equirépartition des effectifs dans chaque classe (sommet des rectangles parrallèle à l'axe des abscisses)
- Surface des rectangles proportionnelle aux effectifs (calcul des n'i ou des f'i)

| ei | n'i corr. | f'i corr. |
|---|---|---|
| 10 | 0 | 0,00 |
| 10 | 5 | 1,32 |
| 20 | 5 | 1,32 |
| 20 | 0 | 0,00 |
| 20 | 12 | 3,16 |
| 30 | 12 | 3,16 |
| 30 | 0 | 0,00 |
| 30 | 18 | 4,74 |
| 40 | 18 | 4,74 |
| 40 | 0 | 0,00 |
| 40 | 3 | 0,79 |
| 50 | 3 | 0,79 |
| 50 | 0 | 0,00 |




Dans le cas d'une variable continue, on repère la classe modale qui correspond à l'effectif corrigé le plus élevé (ou la fréquence corrigée). Ici, on a donc :

**classe modale = [ 30 - 40 [**

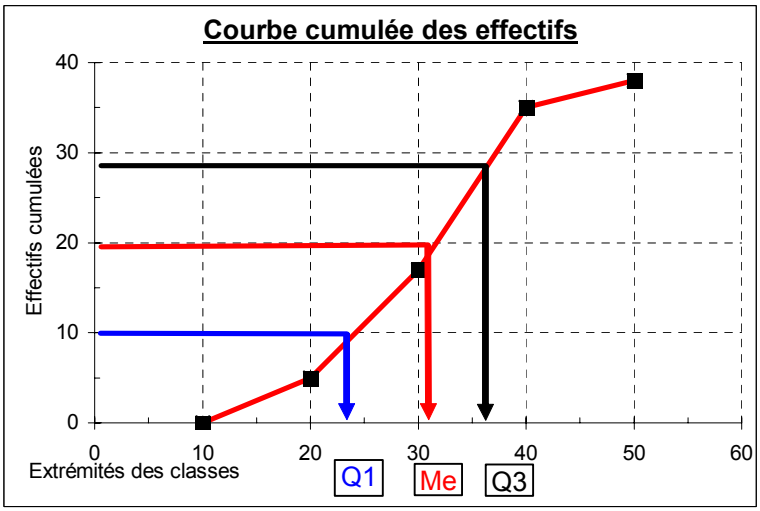
## Diagramme intégral = courbe cumulée continue

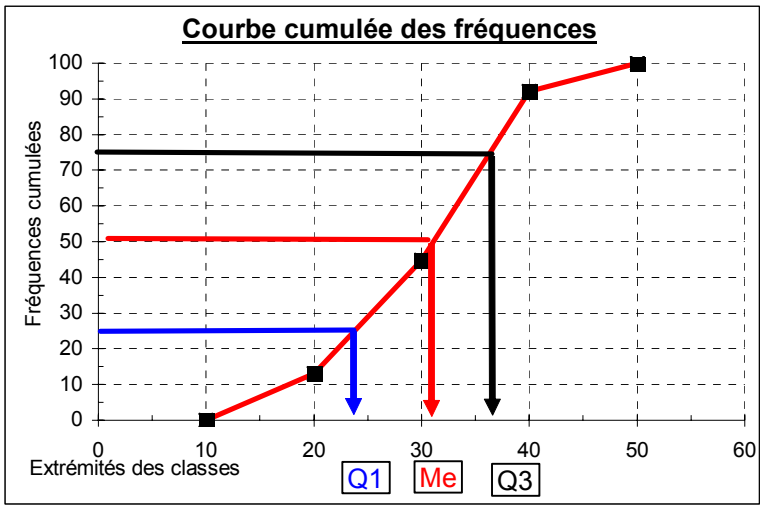
| ei | N(x) | F(x) |
|---|---|---|
| 10 | 0 | 0,00 |
| 20 | 5 | 13,16 |
| 30 | 17 | 44,74 |
| 40 | 35 | 92,11 |
| 50 | 38 | 100,00 |

n / 4 = 9,5

n / 2 = 19

3n / 4 = 28,5




Le quartile Q1 est la valeur de la modalité telle qu'on trouve 25 % de l'effectif au-dessous de cette valeur et 75 % au-dessus.

La médiane (ou quartile Q2) est la valeur de la modalité telle qu'on trouve 50 % de l'effectif au-dessous de cette valeur et 50 % au-dessus (c'est la valeur de la modalité qui partage l'effectif en 2 parties égales).

Le quartile Q3 est la valeur de la modalité telle qu'on trouve 75 % de l'effectif au-dessous de cette valeur et 25 % au-dessus.

**Q1** 25 % des salariés ont moins de 23,75 €/h 75% des salariés ont plus de 23,75 €/h

**Q2** 50 % des salariés ont moins de 31,11 €/h 50 % des salariés ont plus de 31,11

50 % des salariés ont un salaire horaire compris entre 23,75 et 36,39 €/h . **L'écart est de 12,64 €/h**

**Q1 = 23,75 €**
**Me = 31,11 €**
**Q3 - Q1 = 12,64 €** **Q3 = 36,39 €**

- Equirépartition des effectifs dans chaque classe
- Centre de classe = moyenne de la classe

Alors que la **moyenne** est une **caractéristique de tendance centrale**,

l'**écart-type** est une **caractéristique de dispersion** autour de la moyenne.

**Moyenne = 30,00 €/h**

**Var = 67,11**

**Ectyp = 8,19 €/h**

**CV = ectyp / xbar = 0,27**

**La dispersion est faible (< 0,3)**

Le CV permet d'obtenir un nombre sans dimension, qui facilite les comparaisons avec d'autres échantillons.

**coefficient de Yule = ###**

**Etalement à gauche**

Yule est > 0 , il y a étalement à droite
Yule est < 0 , il y a étalement à gauche

**L'étalement à gauche est faible.**

**classe modale = [ 30 - 40 [ ; Me = #### ; Moyenne = 30,00**

ni ci3

16875
187500
771750
273375
1249500

Premier coefficient de Fisher

gamma 1 -0,29
Etalement à gauche

**Exercice 2**

**Les tailles d'une population d'étudiants (en cm)**

Population statistique : 140 étudiants enquêtés  Individu statistique : l'un des 140 étudiants enquêtés

Caractère statistique : la taille (en cm) des étudiants  Type du caractère : quantitatif continu

n'i = ni / ai x 10

| Classes de tailles | Borne inf. | Borne sup. | ci | ai | ni | n'i corr. | N(x) | ni ci | ni ci² |
|---|---|---|---|---|---|---|---|---|---|
| | | 0 | | | | | 0 | | |
| | | 155 | | | | 0 | 0 | | |
| **[155-160[** | 155 | 160 | 157,5 | 5 | 7 | 14 | 7 | 1 102,50 | 173 643,75 |
| **[160-165[** | 160 | 165 | 162,5 | 5 | 12 | 24 | 19 | 1 950,00 | 316 875,00 |
| **[165-170[** | 165 | 170 | 167,5 | 5 | 26 | 52 | 45 | 4 355,00 | 729 462,50 |
| **[170-175[** | 170 | 175 | 172,5 | 5 | 37 | 74 | 82 | 6 382,50 | 1 100 981,25 |
| **[175-180[** | 175 | 180 | 177,5 | 5 | 33 | 66 | 115 | 5 857,50 | 1 039 706,25 |
| **[180-190[** | 180 | 190 | 185,0 | 10 | 21 | 21 | 136 | 3 885,00 | 718 725,00 |
| **[190-200[** | 190 | 200 | 195,0 | 10 | 4 | 4 | 140 | 780,00 | 152 100,00 |
| **Totaux** | | 200 | | | 140 | | 140 | 24 312,50 | 4 231 493,75 |

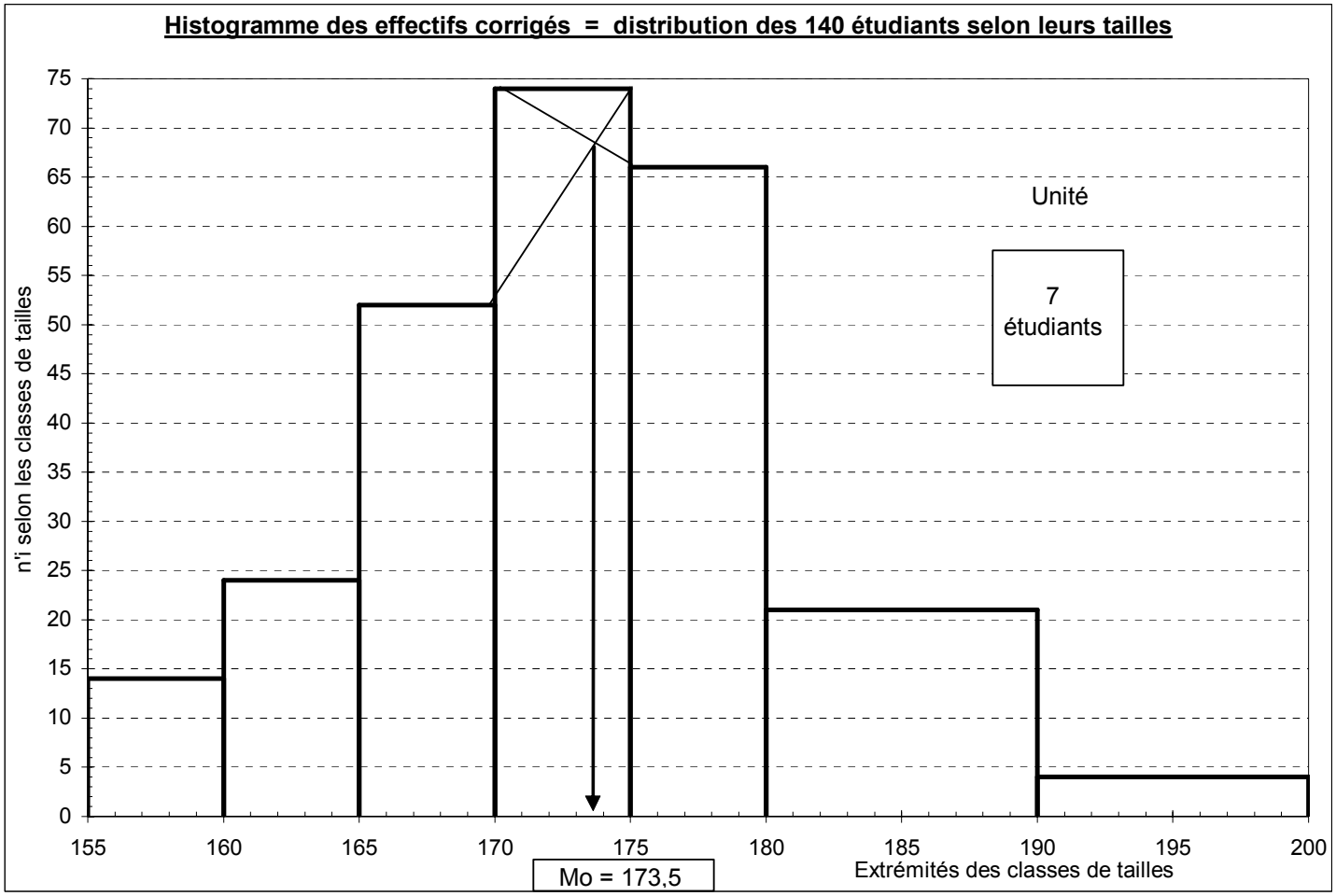
Remarque :

Dans un premier temps, les calculs sont menés sur selon les effectifs. Dans un deuxième temps, les mêmes calculs seront réalisés sur les fréquences relatives. Bien évidemment, les deux méthodes conduisent aux mêmes résultats.

**Diagramme différentiel = histogramme**

- Equirépartition des effectifs dans chaque classe : d'où sommets des rectangles parallèles à l'axe des abscisses.
- Surface des rectangles proportionnelle aux effectifs (calcul des n'i ou des f'i).

| ei | n'i corr. |
|---|---|
| 155 | 0,0 |
| 155 | 14,0 |
| 160 | 14,0 |
| 160 | 0,0 |
| 160 | 24,0 |
| 165 | 24,0 |
| 165 | 0,0 |
| 165 | 52,0 |
| 170 | 52,0 |
| 170 | 0,0 |
| 170 | 74,0 |
| 175 | 74,0 |
| 175 | 0,0 |
| 175 | 66,0 |
| 180 | 66,0 |
| 180 | 0,0 |
| 180 | 21,0 |
| 190 | 21,0 |
| 190 | 0,0 |
| 190 | 4,0 |
| 200 | 4,0 |
| 200 | 0,0 |




L'unité d'aire coïncide avec le 1er rectangle, qui correspond à un effectif de 7 étudiants.

Dans le cas d'une variable continue, on repère la classe modale qui correspond à l'effectif corrigé (ou à la fréquence corrigée) le plus élevé. Ici, on a donc :

**classe modale = [170-175[**

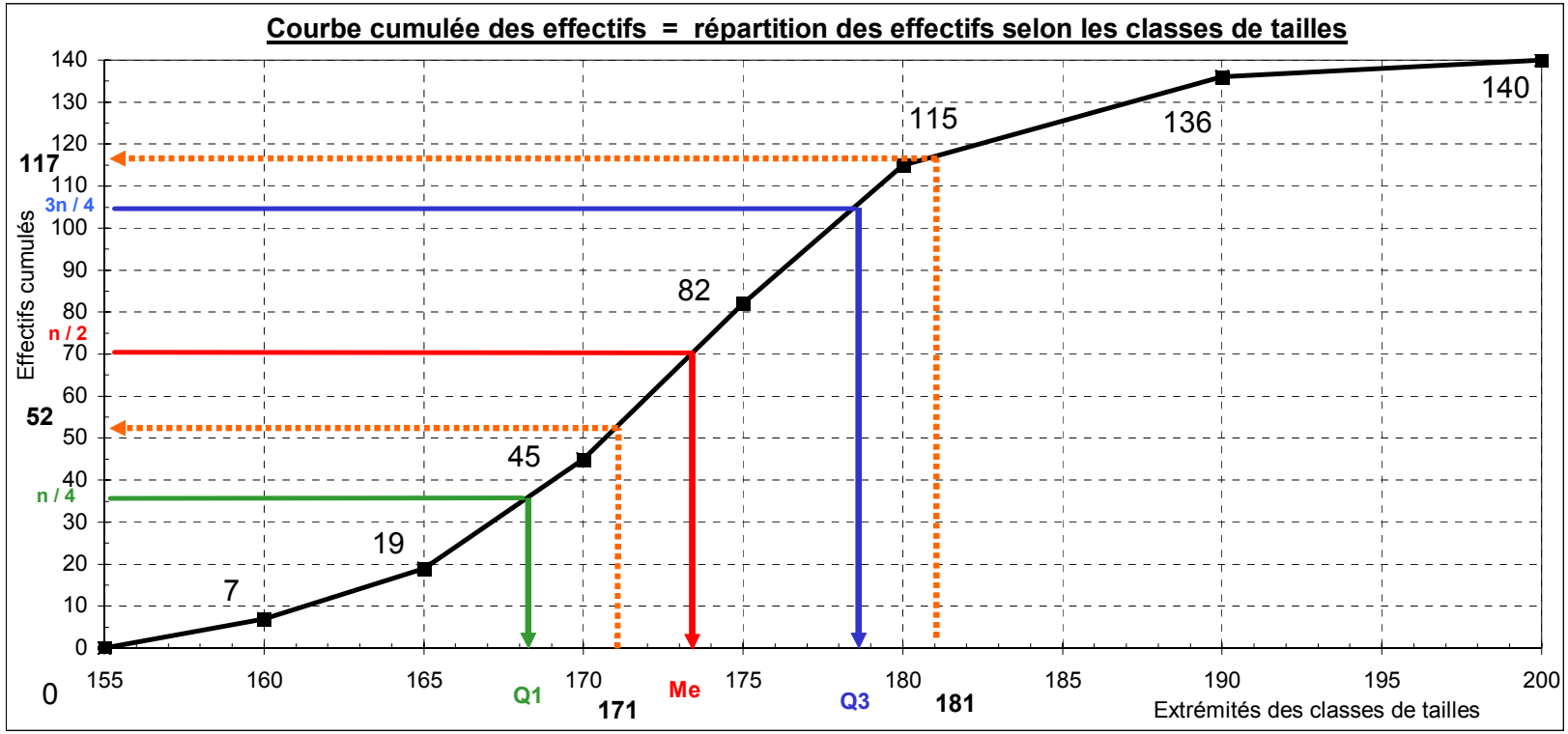
**Mo = 172,5**

On peut envisager deux **valeurs ponctuelles du mode** (caractéristique de tendance centrale souvent peu appropriée) :

- hyp. d'équirépartition des effectifs de la classe modale ---> **centre de la classe** et **Mo = 172,5**

- on tient compte du fait que cette équirépartition est valable pour les classes adjacentes et, graphiquement, par la **méthode des diagonales** (double interpolation linéaire), on a : **Mo = 173,5**

**Diagramme intégral = courbe cumulée continue**




**n = 140** **n/2 = 70,0** **n/4 = 35,0** **3n/4 = 105,0**

Le **quartile Q1** est la valeur de la modalité telle qu'on trouve 25 % de l'effectif au-dessous de cette valeur et 75 % au-dessus.

La **médiane (ou quartile Q2)** (caractéristique de tendance centrale) est la valeur de la modalité telle qu'on trouve 50 % de l'effectif au-dessous de cette valeur et 50 % au-dessus. C'est la valeur de la modalité qui partage l'effectif en 2 sous-ensembles égaux.

Le **quartile Q3** est la valeur de la modalité telle qu'on trouve 75 % de l'effectif au-dessous de cette valeur et 25 % au-dessus.

Ici,on a :

25 % des étudiants mesurent moins de 168,1 cm et 75 % des étudiants mesurent plus de **168,1 cm**
50 % des étudiants mesurent moins de 173,4 cm et 50 % des étudiants mesurent plus de **173,4 cm**
75 % des étudiants mesurent moins de 178,5 cm et 25 % des étudiants mesurent plus de **178,5 cm**

50 % des étudiants ont une taille comprise entre 168,1 cm et 178,5 cm .

**L'écart interquartile Q3 - Q1** (caractéristique de dispersion autour de la médiane) est de **10,4 cm** .

Le **rapport interquartile Q3 / Q1 est de 1,1** (nombre sans dimension).

Ici, **la dispersion autour de la médiane est moyenne**.

| | | |
|---|---|---|
| **Q1 =** | **168,1** | **cm** |
| **Me =** | **173,4** | **cm** |
| **Q3 =** | **178,5** | **cm** |

| | | |
|---|---|---|
| **Q3 - Q1 =** | **10,4** | **cm** |

| | |
|---|---|
| **Q3 / Q1 =** | **1,06** |

**Etendue = range = 45 cm**
(borne sup. dernière classe - borne inf. première classe = 200 - 155)
(caractéristique de dispersion, en général peu appropriée)

- Equirépartition des effectifs dans chaque classe.
- Centres de classes = moyennes de classes

Alors que la moyenne est une caractéristique de tendance centrale, l'écart-type est une caractéristique de dispersion autour de la moyenne.

| | | |
|---|---|---|
| **Moyenne =** | **173,66** | **cm** |
| **Var =** | **66,91** | **cm²** |
| **Ectyp =** | **8,18** | **cm** |

**CV = ectyp / xbar = 0,047**

Le coefficient de variation (CV = ectyp / xbar) permet d'obtenir un nombre sans dimension, qui facilite les comparaisons de dispersion avec d'autres échantillons.

**La dispersion est faible**

**Coefficient de Yule = -0,019**

**Etalement à gauche**

On a : s = (Q3 - 2Me + Q1) / (Q3 - Q1)
(nombre sans dimension / caractéristique d'asymétrie)

Si s de Yule est > 0 , il y a étalement à droite.
Si s de Yule est < 0 , il y a étalement à gauche.

**L'étalement à gauche est très faible** et l'on a :

**classe modale = [170-175[ < Me = 173,4 < Moyenne = 173,7**
**Mo = 172,5**

Compte tenu de la quasi symétrie de la série, les trois valeurs de tendance centrale sont très proche.

**Recherche du nombre d'étudiants dont les tailles sont comprises entre 171 et 181 cm.**

1) Lecture directe de la courbe cumulée

**52** étudiants mesurent au plus 171 cm et **117** étudiants mesurent au plus 181 cm.
Par différence, on a donc **65 étudiants qui mesurent entre 171 cm et 181 cm**, dans la population concernée.

2) Interpolation linéaire

$N(x) = N(ei) + [N(ei+1) - N(ei)] \times (x - ei) / (ei+1 - ei)$

avec : $x$ = valeur critique ; $[N(ei+1) - N(ei)] = ni$ ; $(ei+1 - ei) = ai$

$$\frac{N(171) - 45}{82 - 45} = \frac{171 - 170}{175 - 170}$$

$$N(171) = 45 + (82 - 45) \times \frac{171 - 170}{175 - 170} = 45 + 37 \times \frac{1}{5} = 52{,}4$$

$$\frac{N(181) - 115}{136 - 115} = \frac{181 - 180}{190 - 180}$$

$$N(181) = 115 + (136 - 115) \times \frac{181 - 180}{190 - 180} = 115 + 21 \times \frac{1}{10} = 117{,}1$$

**Les tailles d'une population d'étudiants (en cm)**

Population statistique : 140 étudiants enquêtés  Individu statistique : l'un des 140 étudiants enquêtés

Caractère statistique : la taille (en cm) des étudiants  Type du caractère : quantitatif continu

f'i = fi / ai x 10

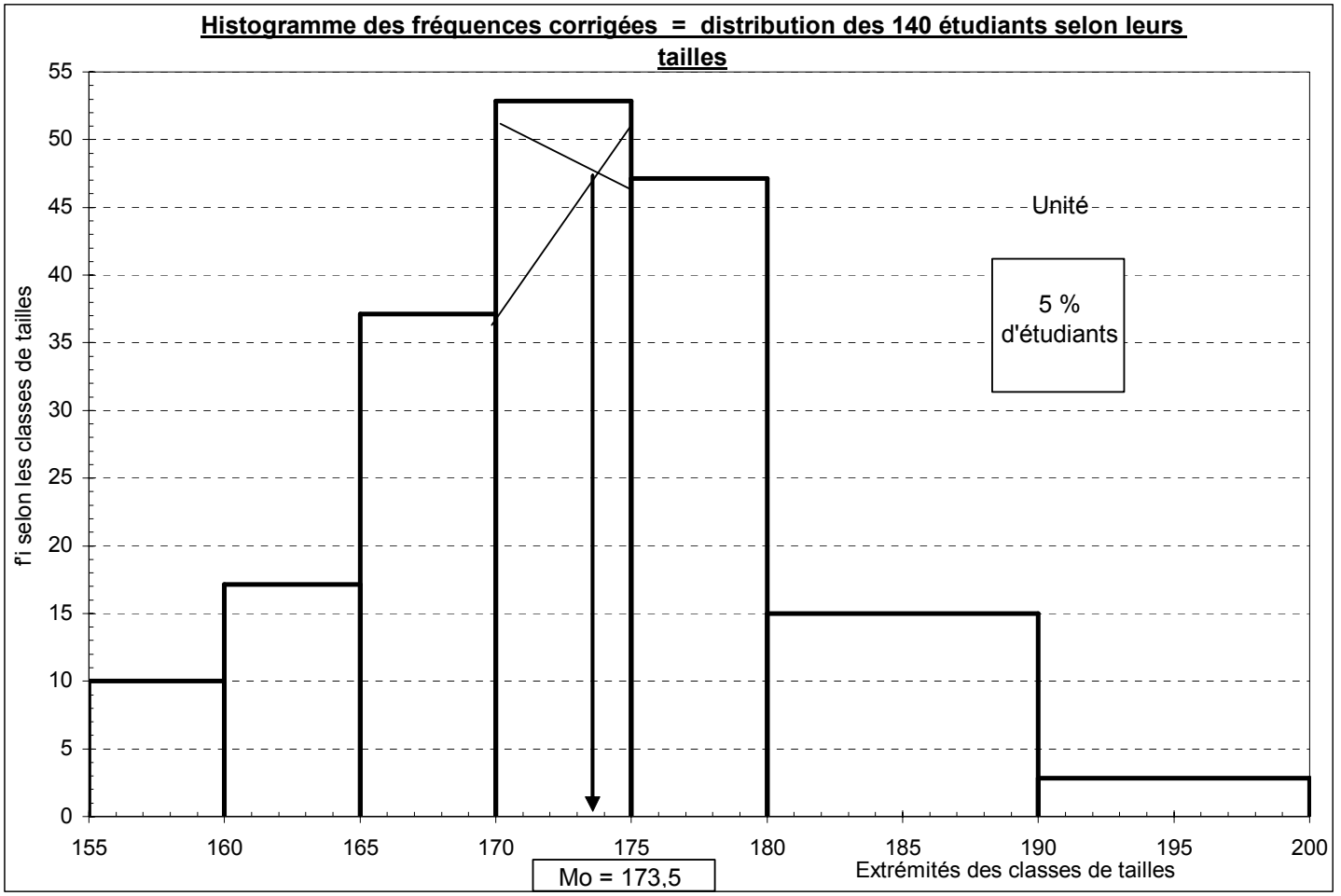
| Classes de tailles | Borne inf. | Borne sup. | ci | ai | fi x 100 | f'i corr x 100 | F(x) x 100 | fi ci x 100 | fi ci² x 100 |
|---|---|---|---|---|---|---|---|---|---|
| | | 0 | | | | | 0 | | |
| | | 155 | | | | 0 | 0 | | |
| **[155-160[** | 155 | 160 | 157,5 | 5 | 5,00 | 10,00 | 5,00 | 787,50 | 124 031,25 |
| **[160-165[** | 160 | 165 | 162,5 | 5 | 8,57 | 17,14 | 13,57 | 1 392,86 | 226 339,29 |
| **[165-170[** | 165 | 170 | 167,5 | 5 | 18,57 | 37,14 | 32,14 | 3 110,71 | 521 044,64 |
| **[170-175[** | 170 | 175 | 172,5 | 5 | 26,43 | 52,86 | 58,57 | 4 558,93 | 786 415,18 |
| **[175-180[** | 175 | 180 | 177,5 | 5 | 23,57 | 47,14 | 82,14 | 4 183,93 | 742 647,32 |
| **[180-190[** | 180 | 190 | 185,0 | 10 | 15,00 | 15,00 | 97,14 | 2 775,00 | 513 375,00 |
| **[190-200[** | 190 | 200 | 195,0 | 10 | 2,86 | 2,86 | 100,00 | 557,14 | 108 642,86 |
| **Totaux** | | 200 | | | 100 | | 100,00 | 17 366,07 | 3 022 495,54 |

Rappel : on raisonne maintenat en utilisant les fréquece relatives.

## Diagramme différentiel = histogramme

- Equirépartition des effectifs dans chaque classe : d'où sommets des rectangles parallèles à l'axe des abscisses.
- Surface des rectangles proportionnelle aux effectifs (calcul des n'i ou des f'i).

| ei | f'i corr. |
|---|---|
| 155 | 0,0 |
| 155 | 10,0 |
| 160 | 10,0 |
| 160 | 0,0 |
| 160 | 17,1 |
| 165 | 17,1 |
| 165 | 0,0 |
| 165 | 37,1 |
| 170 | 37,1 |
| 170 | 0,0 |
| 170 | 52,9 |
| 175 | 52,9 |
| 175 | 0,0 |
| 175 | 47,1 |
| 180 | 47,1 |
| 180 | 0,0 |
| 180 | 15,0 |
| 190 | 15,0 |
| 190 | 0,0 |
| 190 | 2,9 |
| 200 | 2,9 |
| 200 | 0,0 |




L'unité d'aire coïncide avec le 1er rectangle, qui correspond à une proportion de 5 % d'étudiants.

Dans le cas d'une variable continue, on repère la classe modale qui correspond à l'effectif corrigé (ou à la fréquence corrigée) le plus élevé. Ici, on a donc :

**classe modale = [170-175[**

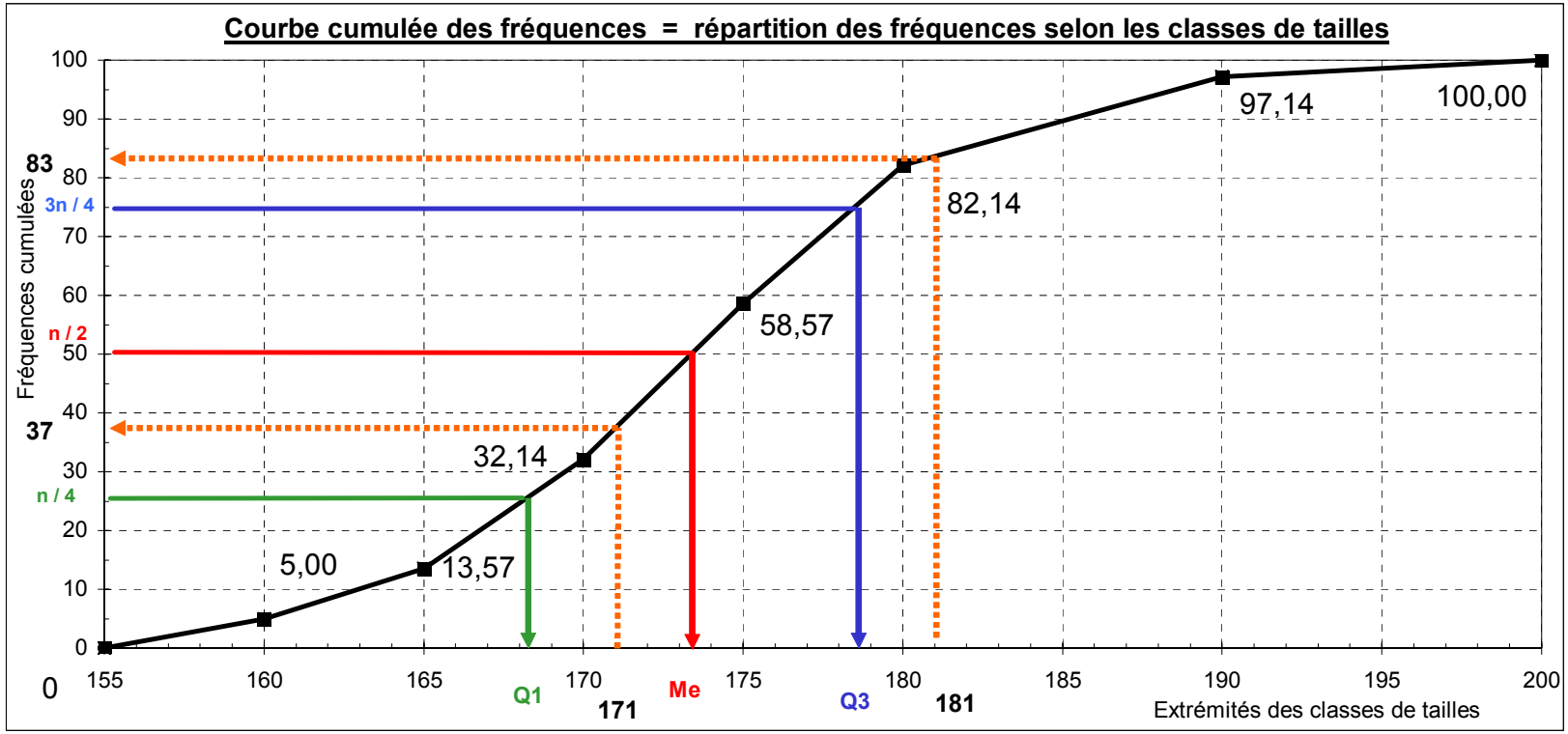
**Mo = 172,5**

On peut envisager deux **valeurs ponctuelles du mode** (caractéristique de tendance centrale souvent peu appropriée) :

- hyp. d'équirépartition des effectifs de la classe modale ---> **centre de la classe** et **Mo = 172,5**

- on tient compte du fait que cette équirépartition est valable pour les classes adjacentes et, graphiquement, par la **méthode des diagonales** (double interpolation linéaire), on a : **Mo = 173,5**

**Diagramme intégral = courbe cumulée continue**




n = **100** **n/2 = 50,0** **n/4 = 25,0** **3n/4 = 75,0**

Le **quartile Q1** est la valeur de la modalité telle qu'on trouve 25 % de l'effectif au-dessous de cette valeur et 75 % au-dessus.

La **médiane (ou quartile Q2)** (caractéristique de tendance centrale) est la valeur de la modalité telle qu'on trouve 50 % de l'effectif au-dessous de cette valeur et 50 % au-dessus. C'est la valeur de la modalité qui partage l'effectif en 2 sous-ensembles égaux.

Le **quartile Q3** est la valeur de la modalité telle qu'on trouve 75 % de l'effectif au-dessous de cette valeur et 25 % au-dessus.

Ici,on a :

25 % des étudiants mesurent moins de 168,1 cm et 75 % des étudiants mesurent plus de **168,1 cm**
50 % des étudiants mesurent moins de 173,4 cm et 50 % des étudiants mesurent plus de **173,4 cm**
75 % des étudiants mesurent moins de 178,5 cm et 25 % des étudiants mesurent plus de **178,5 cm**

50 % des étudiants ont une taille comprise entre 168,1 cm et 178,5 cm .

**L'écart interquartile Q3 - Q1** (caractéristique de dispersion autour de la médiane) est de **10,4 cm** .

Le **rapport interquartile Q3 / Q1 est de 1,1** (nombre sans dimension).

Ici, **la dispersion autour de la médiane est moyenne**.

| | | |
|---|---|---|
| **Q1 =** | **168,1** | **cm** |
| **Me =** | **173,4** | **cm** |
| **Q3 =** | **178,5** | **cm** |

| | | |
|---|---|---|
| **Q3 - Q1 =** | **10,4** | **cm** |

| | |
|---|---|
| **Q3 / Q1 =** | **1,06** |

**Etendue = range = 45 cm**
(borne sup. dernière classe - borne inf. première classe = 200 - 155)
(caractéristique de dispersion, en général peu appropriée)

- Equirépartition des effectifs dans chaque classe.
- Centres de classes = moyennes de classes

Alors que la moyenne est une caractéristique de tendance centrale, l'écart-type est une caractéristique de dispersion autour de la moyenne.

| | | |
|---|---|---|
| **Moyenne =** | **173,66** | **cm** |
| **Var =** | **66,91** | **cm²** |
| **Ectyp =** | **8,18** | **cm** |

**CV = ectyp / xbar = 0,047**

Le coefficient de variation (CV = ectyp / xbar) permet d'obtenir un nombre sans dimension, qui facilite les comparaisons de dispersion avec d'autres échantillons.

**La dispersion est faible**

**Coefficient de Yule = -0,019**

**Etalement à gauche**

On a : s = (Q3 - 2Me + Q1) / (Q3 - Q1)
(nombre sans dimension / caractéristique d'asymétrie)

Si s de Yule est > 0 , il y a étalement à droite.
Si s de Yule est < 0 , il y a étalement à gauche.

L'étalement à gauche est très faible et l'on a :

**classe modale = [170-175[ < Me = 173,4 < Moyenne = 173,7**
**Mo = 172,5**

Compte tenu de la quasi symétrie de la série, les trois valeurs de tendance centrale sont très proche.

**Recherche du nombre d'étudiants dont les tailles sont comprises entre 171 et 181 cm.**

1) Lecture directe de la courbe cumulée

**37 %** des étudiants mesurent au plus 171 cm et **83 %** des étudiants mesurent au plus 181 cm.
Par différence, on a donc **46 % des étudiants qui mesurent entre 171 cm et 181 cm**, dans la population concernée.

2) Interpolation linéaire

$F(x) = F(ei) + [F(ei+1) - F(ei)] \times (x - ei) / (ei+1 - ei)$

avec : x = valeur critique ; $[F(ei+1) - F(ei)] = fi$ ; $(ei+1 - ei) = ai$

$$\frac{F(171) - 32,1}{58,6 - 32,1} = \frac{171 - 170}{175 - 170} \qquad F(171) = 32,1 + (58,6 - 32,1) \times \frac{1}{5} = 37,4$$

$$\frac{F(181) - 82,1}{97,1 - 82,1} = \frac{181 - 180}{190 - 180} \qquad F(181) = 82,1 + (97,1 - 82,1) \times \frac{1}{10} = 83,6$$

**Exercice 3**

**Montants mensuels (en €) des dépôts effectués sur un compte d'épargne par les 500 clients d'un établissement bancaire**

Population statistique : 500 clients d'un établissement bancaire  Individu statistique : l'un des 500 clients

Caractère statistique : le montant mensuel des dépôts (en €)  Type du caractère : quantitatif continu

n'i = ni / ai x 50

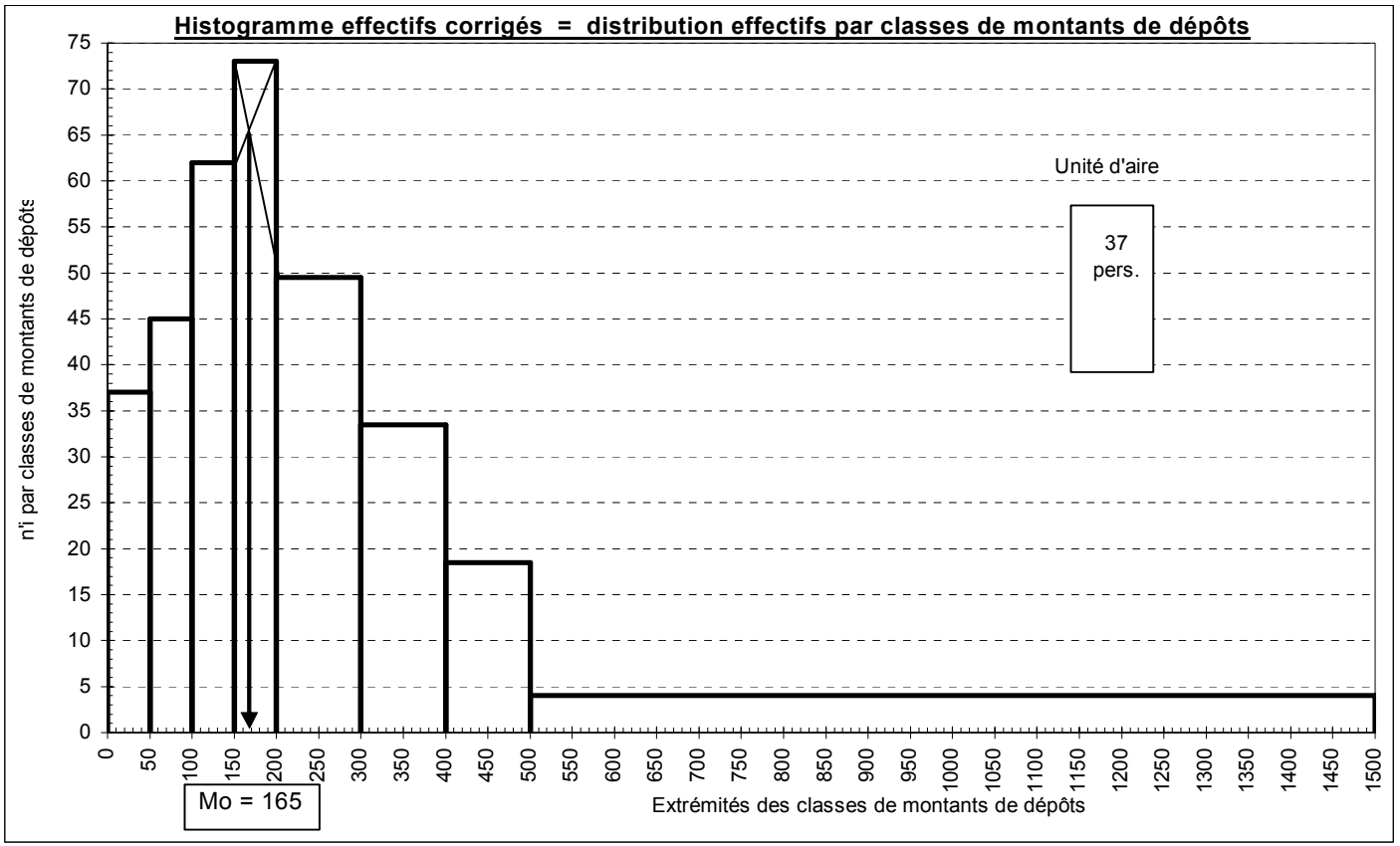
| Classes de dépôts mensuels (€) | Borne inf. | Borne sup. | ci | ai | ni | n'i corr. | N(x) | ni ci | ni ci² | ni ci³ | ni ci⁴ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | | | | | 0 | | | | |
| | | 0 | | | | 0,0 | 0 | | | | |
| **[ 0 - 50 [** | 0 | 50 | 25 | 50 | 37 | 37,0 | 37 | 925 | 23 125 | 578 125 | 14 453 125 |
| **[ 50 - 100 [** | 50 | 100 | 75 | 50 | 45 | 45,0 | 82 | 3 375 | 253 125 | 18 984 375 | 1 423 828 125 |
| **[ 100 - 150 [** | 100 | 150 | 125 | 50 | 62 | 62,0 | 144 | 7 750 | 968 750 | 121 093 750 | 15 136 718 750 |
| **[ 150 - 200 [** | 150 | 200 | 175 | 50 | 73 | 73,0 | 217 | 12 775 | 2 235 625 | 391 234 375 | 68 466 015 625 |
| **[ 200 - 300 [** | 200 | 300 | 250 | 100 | 99 | 49,5 | 316 | 24 750 | 6 187 500 | 1 546 875 000 | 386 718 750 000 |
| **[ 300 - 400 [** | 300 | 400 | 350 | 100 | 67 | 33,5 | 383 | 23 450 | 8 207 500 | 2 872 625 000 | 1 005 418 750 000 |
| **[ 400 - 500 [** | 400 | 500 | 450 | 100 | 37 | 18,5 | 420 | 16 650 | 7 492 500 | 3 371 625 000 | 1 517 231 250 000 |
| **[ 500 - 1 500 [** | 500 | 1 500 | 1 000 | 1 000 | 80 | 4,0 | 500 | 80 000 | 80 000 000 | 80 000 000 000 | 80 000 000 000 000 |
| **Totaux** | | 1 600 | | | 500 | | 500 | 169 675 | 105 368 125 | 88 323 015 625 | 82 994 409 765 625 |

Remarque : comme dans l'exercice précédent, les calculs sont d'abord menés sur les effectifs, avant d'être menés sur les fréquences relatives.

**Diagramme différentiel = histogramme**

Equirépartition des effectifs dans chaque classe : d'où sommets des rectangles parallèles à l'axe des abscisses.
Surface des rectangles proportionnelle aux effectifs (calcul des n'i ou des f'i).

| ei | n'i corr. |
|---|---|
| 0 | 0,0 |
| 0 | 37,0 |
| 50 | 37,0 |
| 50 | 0,0 |
| 50 | 45,0 |
| 100 | 45,0 |
| 100 | 0,0 |
| 100 | 62,0 |
| 150 | 62,0 |
| 150 | 0,0 |
| 150 | 73,0 |
| 200 | 73,0 |
| 200 | 0,0 |
| 200 | 49,5 |
| 300 | 49,5 |
| 300 | 0,0 |
| 300 | 33,5 |
| 400 | 33,5 |
| 400 | 0,0 |
| 400 | 18,5 |
| 500 | 18,5 |
| 500 | 0,0 |
| 500 | 4,0 |
| 1 500 | 4,0 |
| 1 500 | 0,0 |




L'unité d'aire correspond à la surface de l'avant-dernier rectangle. Le tableau nous indique que cette surface équivaut à 37 déposants.

Dans le cas d'une variable continue, on repère la classe modale qui correspond à l'effectif corrigé (ou à la fréquence corrigée) le plus élevé. Ici, on a donc :

**classe modale = [ 150 - 200 [**

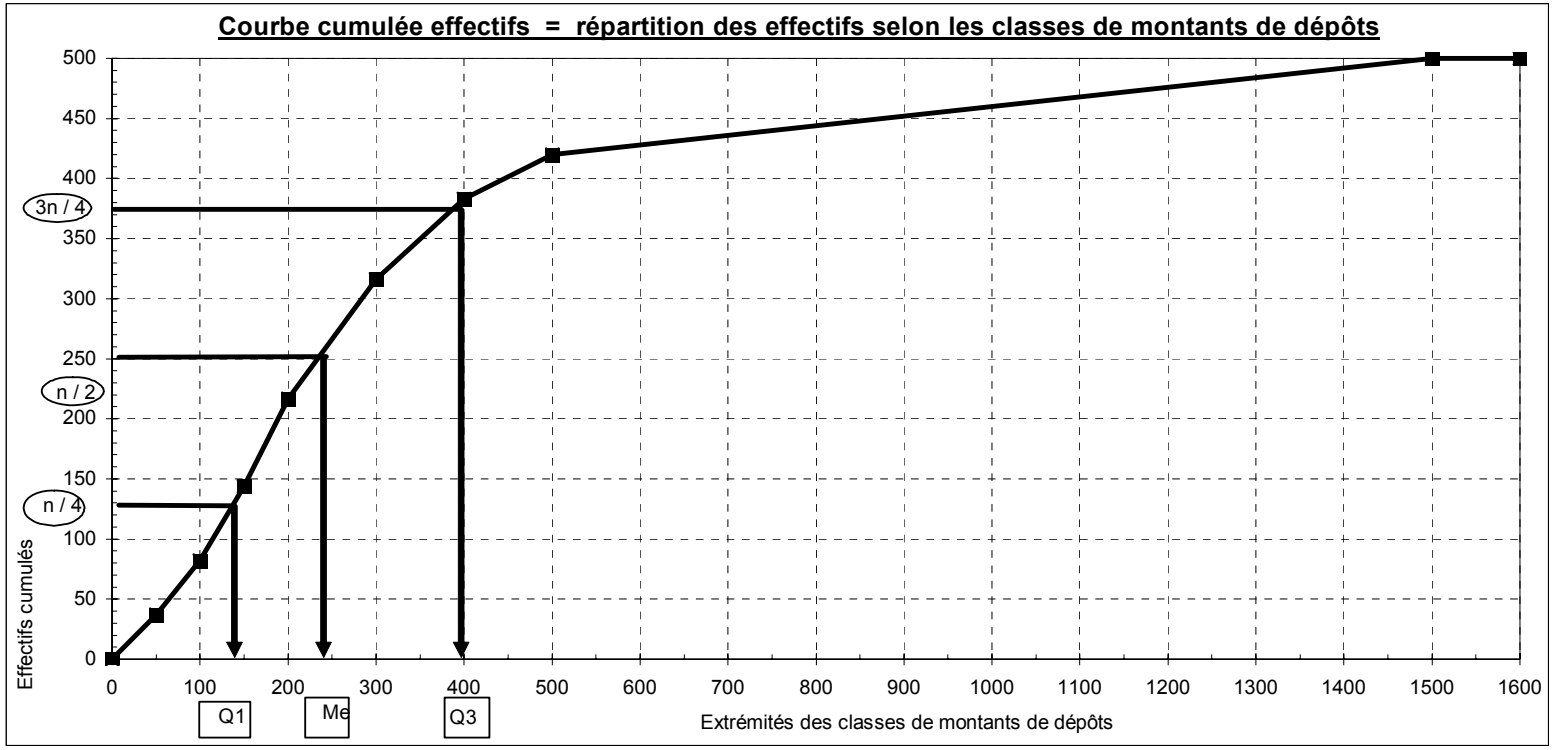
**Mo = 175 €**

On peut envisager deux **valeurs ponctuelles du mode** (caractéristique de tendance centrale souvent peu appropriée) :

- hyp. d'équirépartition des effectifs dans la classe modale ---> centre de la classe et Mo = 175 .

- on tient compte du fait que cette équirépartition est valable pour les classes adjacentes et, graphiquement, par la méthode des diagonales (double interpolation linéaire), on obtient Mo = 165 .

**Diagramme intégral = courbe cumulée continue**



Le **quartile Q1** est la valeur de la modalité telle qu'on trouve 25 % de l'effectif au-dessous de cette valeur et 75 % au-dessus.

La **médiane (ou quartile Q2)** (caractéristique de tendance centrale) est la valeur de la modalité telle qu'on trouve 50 % de l'effectif au-dessous de cette valeur et 50 % au-dessus. C'est la valeur de la modalité qui partage l'effectif en 2 sous-ensembles égaux.

Le **quartile Q3** est la valeur de la modalité telle qu'on trouve 75 % de l'effectif au-dessous de cette valeur et 25 % au-dessus.

Ici,on a :

25 % des déposants ont placé moins de 135 € et 75 % plus de 135 €

50 % des déposants ont placé moins de 230 € et 50 % plus de 230 €

75 % des déposants ont placé moins de 390 € et 25 % plus de 390 €

50 % des déposants ont placé entre 588,06 F et 136,99 F .

L'écart interquartile Q3 - Q1 (caractéristique de dispersion autour de la médiane) est de 251,07 € .

Le rapport interquartile Q3 / Q1 est de 2,83 .
Nombre sans dimension qui indique ici une forte dispersion autour de la médiane.

| | | |
|---|---|---|
| **Q1 =** | **136,99** | **€** |
| **Me =** | **233,33** | **€** |
| **Q3 =** | **388,06** | **€** |

| | | |
|---|---|---|
| **Q3 - Q1 =** | **251,07** | **€** |

| | |
|---|---|
| **Q3 / Q1 =** | **2,83** |

**Etendue = range = 1 500 €**
(borne sup. dernière classe - borne inf. première classe = 1 500 - 0)
(caractéristique de dispersion, en général peu appropriée)

- Equirépartition des effectifs dans chaque classe.
- Centres de classes = moyennes de classes

Alors que la moyenne est une caractéristique de tendance centrale,
l'écart-type est une caractéristique de dispersion autour de la moyenne.

| | | |
|---|---|---|
| **Moyenne =** | **339,35** | **€** |
| **Var =** | **95 577,83** | **€** |
| **Ectyp =** | **309,16** | **€** |

**CV = ectyp / xbar = 0,91**

La dispersion est très forte (> 0,5) .

Le coefficient de variation (CV) permet d'obtenir un nombre sans dimension, qui facilite les comparaisons de dispersion avec d'autres échantillons.

**Coefficient de Yule = 0,23**

**Etalement à droite**

Si s de Yule est > 0 , il y a étalement à droite.
Si s de Yule est < 0 , il y a étalement à gauche.

On a : s = (Q3 - 2Me + Q1) / (Q3 - Q1)
(nombre sans dimension / caractéristique d'asymétrie)

L'étalement à droite est fort et l'on a :

**classe modale = [ 150 - 200 [ < Me = 233,33 < Moyenne = 339,35**

**Mo = 175**

**Pearson 1 = 0,53**

**Etalement à droite (Pearson 1 positif)**

Pearson 1 = (xbar - Mo) / ectyp (nombre sans dimension)

Caractéristique d'asymétrie, souvent peu appropriée.

| | | |
|---|---|---|
| m1 = | xbar = s nici / n = | 339,35 |
| m2 = | s nici² / n = | 210 736,25 |
| m3 = | s nici$^3$ / n = | 176 646 031,25 |
| m4 = | s nici$^4$ / n = | 165 988 819 531,25 |

**Fisher 1 = 1,36**

**Etalement à droite (Fisher 1 positif)**

gamma1 = mu3 / ectyp^3 (nombre sans dimension)

= (m3 - 3m1.m2 + 2m1^3) / ectyp^3

Caractéristique d'asymétrie d'une série statistique.

**Fisher 2 = 0,51**

**Distribution leptocurtique (Fisher 2 positif)**

gamma2 = (mu4 / ectyp^4) - 3 (nombre sans dimension)

= [(m4 - 4m1.m3 + 6m1^2.m2 - 3m1^4) / ectyp^4] - 3

Caractéristique d'aplatissement d'une série statistique.

**Montants mensuels (en €) des dépôts effectués sur un compte d'épargne par les 500 clients d'un établissement bancaire**

Population statistique : 500 clients d'un établissement bancaire  Individu statistique : l'un des 500 clients

Caractère statistique : le montant mensuel des dépôts (en €)  Type du caractère : quantitatif continu

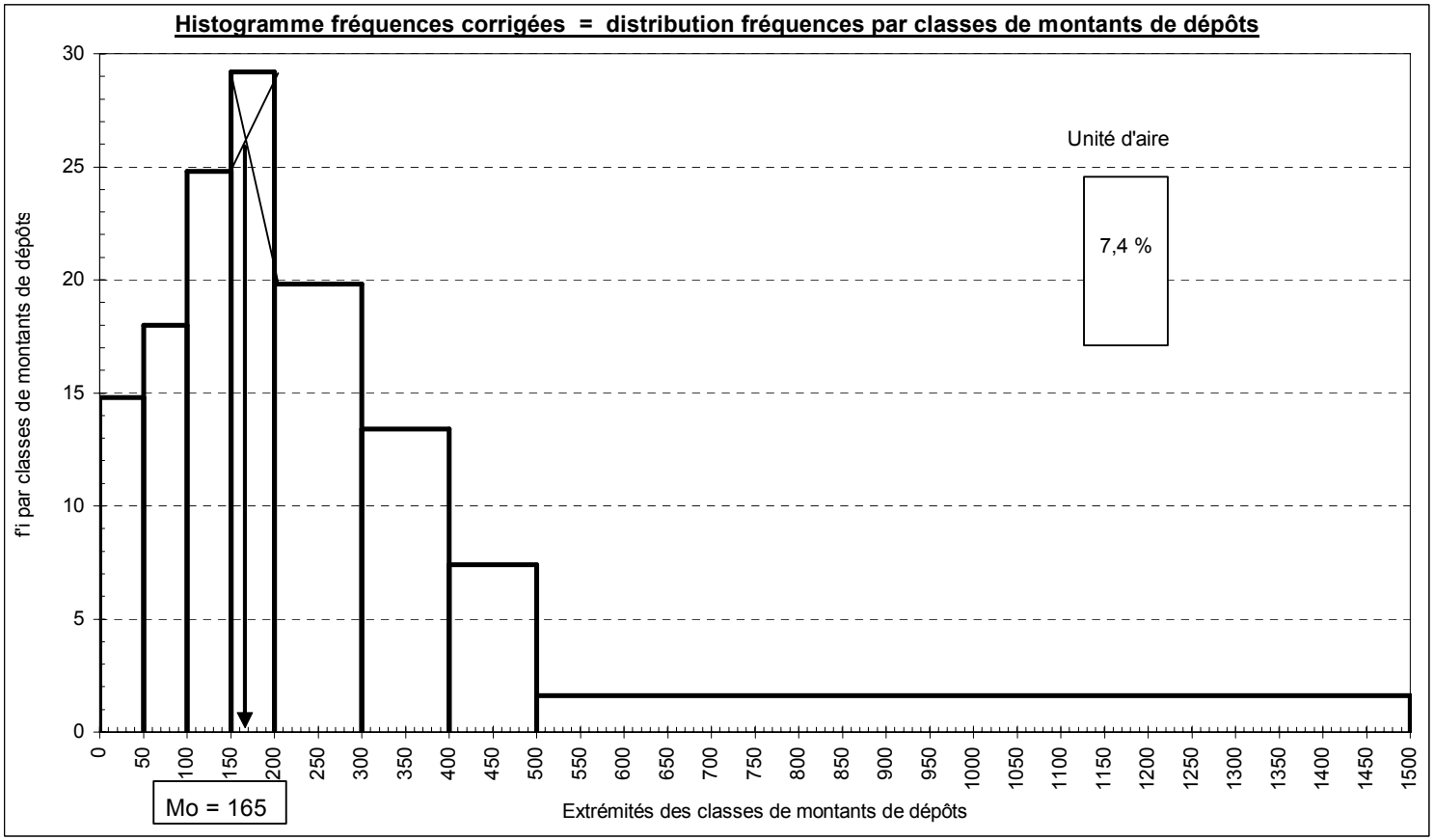
f'i = fi / ai x 10 000

| Classes de dépôts mensuels (€) | Borne inf. | Borne sup. | ci | ai | ni | fi x100 | f'i corr. X10 000 | F(x) x100 | fi ci | fi $ci^2$ | fi $ci^3$ | fi $ci^4$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | | | 0,0% | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | | 0,0 | 0,0% | | | | |
| **[ 0 - 50 [** | 0 | 50 | 25 | 50 | 37 | 7,4% | 14,8 | 7,4% | 1,85 | 46,25 | 1 156,25 | 28 906,25 |
| **[ 50 - 100 [** | 50 | 100 | 75 | 50 | 45 | 9,0% | 18,0 | 16,4% | 6,75 | 506,25 | 37 968,75 | 2 847 656,25 |
| **[ 100 - 150 [** | 100 | 150 | 125 | 50 | 62 | 12,4% | 24,8 | 28,8% | 15,50 | 1 937,50 | 242 187,50 | 30 273 437,50 |
| **[ 150 - 200 [** | 150 | 200 | 175 | 50 | 73 | 14,6% | 29,2 | 43,4% | 25,55 | 4 471,25 | 782 468,75 | 136 932 031,25 |
| **[ 200 - 300 [** | 200 | 300 | 250 | 100 | 99 | 19,8% | 19,8 | 63,2% | 49,50 | 12 375,00 | 3 093 750,00 | 773 437 500,00 |
| **[ 300 - 400 [** | 300 | 400 | 350 | 100 | 67 | 13,4% | 13,4 | 76,6% | 46,90 | 16 415,00 | 5 745 250,00 | 2 010 837 500,00 |
| **[ 400 - 500 [** | 400 | 500 | 450 | 100 | 37 | 7,4% | 7,4 | 84,0% | 33,30 | 14 985,00 | 6 743 250,00 | 3 034 462 500,00 |
| **[ 500 - 1 500 [** | 500 | 1 500 | 1 000 | 1 000 | 80 | 16,0% | 1,6 | 100,0% | 160,00 | 160 000,00 | 160 000 000,00 | 160 000 000 000,00 |
| **Totaux** | 0 | 1 600 | 0 | 0 | 500 | 100,0% | | 100,0% | 339,35 | 210 736,25 | 176 646 031,25 | 165 988 819 531,25 |

**Diagramme différentiel = histogramme**

Equirépartition des effectifs dans chaque classe : d'où sommets des rectangles parallèles à l'axe des abscisses.
Surface des rectangles proportionnelle aux effectifs (calcul des n'i ou des f'i).

| ei | f'i corr. |
|---|---|
| 0 | 0,0 |
| 0 | 14,8 |
| 50 | 14,8 |
| 50 | 0,0 |
| 50 | 18,0 |
| 100 | 18,0 |
| 100 | 0,0 |
| 100 | 24,8 |
| 150 | 24,8 |
| 150 | 0,0 |
| 150 | 29,2 |
| 200 | 29,2 |
| 200 | 0,0 |
| 200 | 19,8 |
| 300 | 19,8 |
| 300 | 0,0 |
| 300 | 13,4 |
| 400 | 13,4 |
| 400 | 0,0 |
| 400 | 7,4 |
| 500 | 7,4 |
| 500 | 0,0 |
| 500 | 1,6 |
| 1 500 | 1,6 |
| 1 500 | 0,0 |




L'unité d'aire correspond à la surface de l'avant-dernier rectangle. Le tableau nous indique que cette surface équivaut à 7,4 % des déposants.

Dans le cas d'une variable continue, on repère la classe modale qui correspond à l'effectif corrigé (ou à la fréquence corrigée) le plus élevé. Ici, on a donc :

**classe modale = [ 150 - 200 [**

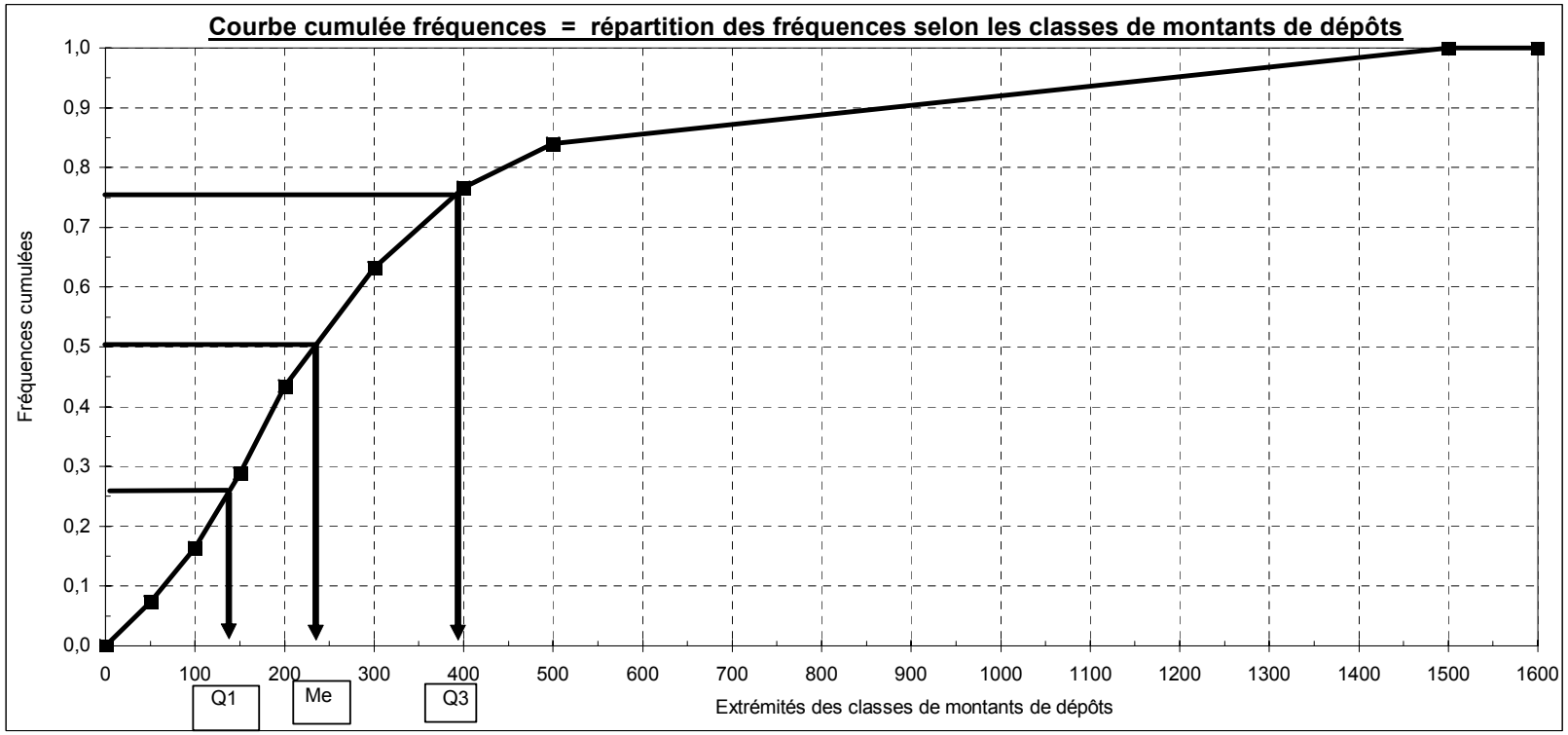
**Mo = 175 €**

On peut envisager deux **valeurs ponctuelles du mode** (caractéristique de tendance centrale souvent peu appropriée) :

- hyp. d'équirépartition des effectifs dans la classe modale ---> centre de la classe et Mo = 175 .

- on tient compte du fait que cette équirépartition est valable pour les classes adjacentes et, graphiquement, par la méthode des diagonales (double interpolation linéaire), on obtient Mo = 165 .

**Diagramme intégral = courbe cumulée continue**



Le **quartile Q1** est la valeur de la modalité telle qu'on trouve 25 % de l'effectif au-dessous de cette valeur et 75 % au-dessus.

La **médiane (ou quartile Q2)** (caractéristique de tendance centrale) est la valeur de la modalité telle qu'on trouve 50 % de l'effectif au-dessous de cette valeur et 50 % au-dessus. C'est la valeur de la modalité qui partage l'effectif en 2 sous-ensembles égaux.

Le **quartile Q3** est la valeur de la modalité telle qu'on trouve 75 % de l'effectif au-dessous de cette valeur et 25 % au-dessus.

Ici,on a :

25 % des déposants ont placé moins de 135 € et 75 % plus de 135 €

50 % des déposants ont placé moins de 230 € et 50 % plus de 230 €

75 % des déposants ont placé moins de 390 € et 25 % plus de 390 €

50 % des déposants ont placé entre 588,06 F et 136,99 F .

L'écart interquartile Q3 - Q1 (caractéristique de dispersion autour de la médiane) est de 251,07 € .

Le rapport interquartile Q3 / Q1 est de 2,83 .
Nombre sans dimension qui indique ici une forte dispersion autour de la médiane.

**Q1 = 136,99 €**

**Me = 233,33 €**

**Q3 = 388,06 €**

**Q3 - Q1 = 251,07 €**

**Q3 / Q1 = 2,83**

**Etendue = range = 1500 €**
(borne sup. dernière classe - borne inf. première classe = 1 500 - 0)
(caractéristique de dispersion, en général peu appropriée)

- Equirépartition des effectifs dans chaque classe.
- Centres de classes = moyennes de classes

Alors que la moyenne est une caractéristique de tendance centrale,
l'écart-type est une caractéristique de dispersion autour de la moyenne.

**Moyenne = 339,35 €**

**Var = 95577,83 €**

**Ectyp = 309,16 €**

**CV = ectyp / xbar = 0,91**

La dispersion est assez forte (> 0,5) .

Le coefficient de variation (CV) permet d'obtenir un nombre sans dimension, qui facilite les comparaisons de dispersion avec d'autres échantillons.

**Coefficient de Yule = 0,23**

**Etalement à droite**

Si s de Yule est > 0 , il y a étalement à droite.
Si s de Yule est < 0 , il y a étalement à gauche.

On a : s = (Q3 - 2Me + Q1) / (Q3 - Q1)
(nombre sans dimension / caractéristique d'asymétrie)

L'étalement à droite est fort et l'on a :

**classe modale = [ 150 - 200 [ < Me = 233,33 < Moyenne = 339,35**
**Mo = 175**

**Pearson 1 = 0,53**

**Etalement à droite (Pearson 1 positif)**

Pearson 1 = (xbar - Mo) / ectyp (nombre sans dimension)

Caractéristique d'asymétrie, souvent peu appropriée.

| | | |
|---|---|---|
| m1 = | xbar = s fici = | 339,35 |
| m2 = | s fici² = | 210 736,25 |
| m3 = | s fici$^3$ = | 176 646 031,25 |
| m4 = | s fici$^4$ = | 165 988 819 531,25 |

**Fisher 1 = 1,36**

**Etalement à droite (Fisher 1 positif)**

gamma1 = mu3 / ectyp^3 (nombre sans dimension)

= (m3 - 3m1.m2 + 2m1^3) / ectyp^3

Caractéristique d'asymétrie d'une série statistique.

**Fisher 2 = 0,51**

**Distribution leptocurtique (Fisher 2 positif)**

gamma2 = (mu4 / ectyp^4) - 3 (nombre sans dimension)

= [(m4 - 4m1.m3 + 6m1^2.m2 - 3m1^4) / ectyp^4] - 3

Caractéristique d'aplatissement d'une série statistique.

**Montants mensuels (en €) des dépôts effectués sur un compte d'épargne par les 500 clients d'un établissement bancaire**

**Population statistique : 500 clients d'un établissement bancaire** — **Caractère statistique : le montant mensuel des dépôts**

**Individu statistique : l'un des 500 clients** — **Type du caractère : quantitatif continu**

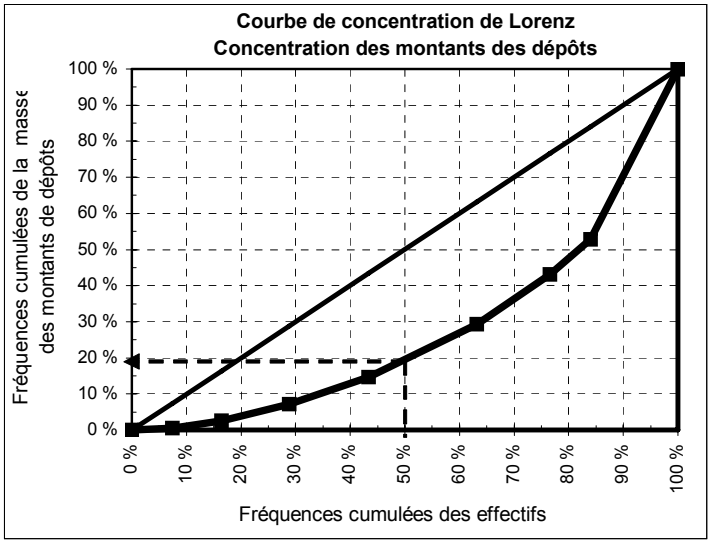
| Classes de dépôts mensuels (€) | Borne inf. | Borne sup. | ci | ai | ni | N(x) | ni ci | nici / Snici x100 (si) | si x100 cumulés S(x) x100 | fi x100 | F(x) x100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 0,0% | | 0,0% |
| [ 0 - 50 [ | 0 | 50 | 25 | 50 | 37 | 37 | 925 | 0,5% | 0,5% | 7,4% | 7,4% |
| [ 50 - 100 [ | 50 | 100 | 75 | 50 | 45 | 82 | 3 375 | 2,0% | 2,5% | 9,0% | 16,4% |
| [ 100 - 150 [ | 100 | 150 | 125 | 50 | 62 | 144 | 7 750 | 4,6% | 7,1% | 12,4% | 28,8% |
| [ 150 - 200 [ | 150 | 200 | 175 | 50 | 73 | 217 | 12 775 | 7,5% | 14,6% | 14,6% | 43,4% |
| [ 200 - 300 [ | 200 | 300 | 250 | 100 | 99 | 316 | 24 750 | 14,6% | 29,2% | 19,8% | 63,2% |
| [ 300 - 400 [ | 300 | 400 | 350 | 100 | 67 | 383 | 23 450 | 13,8% | 43,0% | 13,4% | 76,6% |
| [ 400 - 500 [ | 400 | 500 | 450 | 100 | 37 | 420 | 16 650 | 9,8% | 52,9% | 7,4% | 84,0% |
| [ 500 - 1 500 [ | 500 | 1500 | 1000 | 1000 | 80 | 500 | 80 000 | 47,1% | 100,0% | 16,0% | 100,0% |
| Totaux | | | | | 500 | | 169 675 | 100,0% | 100,0% | 100,0% | 100,0% |

Dans le tableau ci-dessous, la valeur 0,0133 est donnée par :
(0,00 + 11,76) / 2 x 22,7

Le résultat de cette opération est divisé par 10.000, de façon à se ramener à un raisonnement en fréquence (carré de côté égal à 1) .

| F(x) x100 | S(x) x100 | Equi-répart. | Inég. totale |
|---|---|---|---|
| 0,0 % | 0,0 % | 0,0 % | 0 % |
| 7,4 % | 0,5 % | 7,4 % | 0 % |
| 16,4 % | 2,5 % | 16,4 % | 0 % |
| 28,8 % | 7,1 % | 28,8 % | 0 % |
| 43,4 % | 14,6 % | 43,4 % | 0 % |
| 63,2 % | 29,2 % | 63,2 % | 0 % |
| 76,6 % | 43,0 % | 76,6 % | 0 % |
| 84,0 % | 52,9 % | 84,0 % | 0 % |
| 100,0 % | 100,0 % | 100,0 % | 0 % |
| 100,0 % | 100,0 % | 100,0 % | 100 % |

La moitié des clients (qui déposent le moins) détient seulement 20 % du montant total des dépôts.
L'autre moitié (qui dépose le plus) détient 80 % du montant total des dépôts.




| fi | S(x) | S sous courbe |
|---|---|---|
| | 0,0 % | 0,0002 |
| 7,4 % | 0,5 % | 0,0014 |
| 9,0 % | 2,5 % | 0,0060 |
| 12,4 % | 7,1 % | 0,0159 |
| 14,6 % | 14,6 % | 0,0434 |
| 19,8 % | 29,2 % | 0,0484 |
| 13,4 % | 43,0 % | 0,0355 |
| 7,4 % | 52,9 % | 0,1223 |
| 16,0 % | 100,0 % | |

| | | |
|---|---|---|
| **S tot.** | **ss courbe** | **0,273** |
| **S** | **concentr.** | **0,227** |
| **Indice** | **Gini** | **0,454** |

**Concentration moyenne**

**Exercice 4**

**Ancienneté (en années) des 350 employés d'une entreprise**

Population statistique : 350 employés d'une entreprise  Individu statistique : l'un des 350 employés

Caractère statistique : l'ancienneté (en années) des employés  Type du caractère : quantitatif (discret traité en) continu

n'i = ni / ai x 5

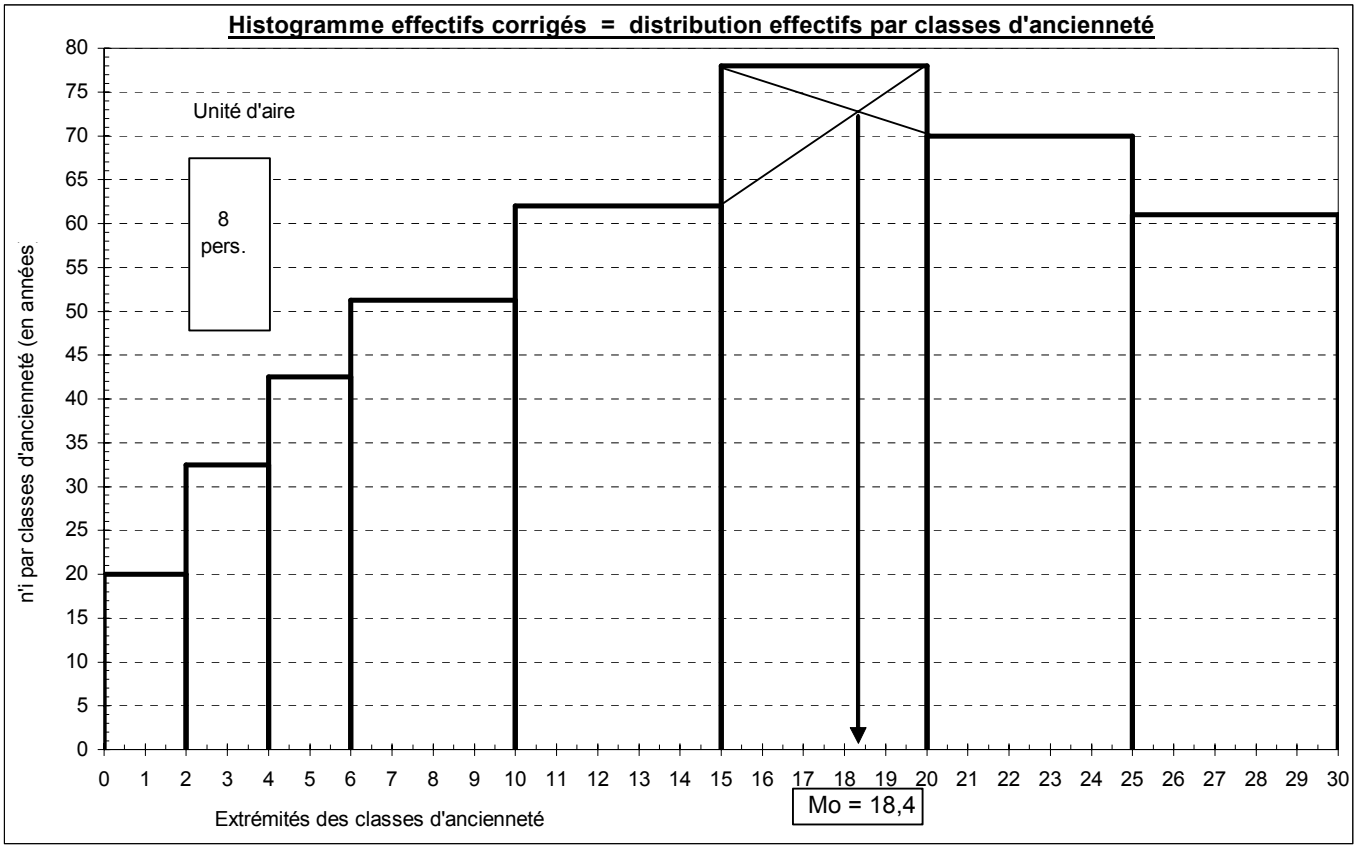
| Classes d'années d'ancienneté | Borne inf. | Borne sup. | ci | ai | ni | n'i corr. | N(x) | ni ci | ni ci² | ni ci³ | ni ci⁴ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | | | | | 0 | | | | |
| | | 0 | | | | 0,0 | 0 | | | | |
| [ 0 - 2 [ | 0 | 2 | 1 | 2 | 8 | 20,0 | 8 | 8 | 8 | 8 | 8 |
| [ 2 - 4 [ | 2 | 4 | 3 | 2 | 13 | 32,5 | 21 | 39 | 117 | 351 | 1 053 |
| [ 4 - 6 [ | 4 | 6 | 5 | 2 | 17 | 42,5 | 38 | 85 | 425 | 2 125 | 10 625 |
| [ 6 - 10 [ | 6 | 10 | 8 | 4 | 41 | 51,3 | 79 | 328 | 2 624 | 20 992 | 167 936 |
| [ 10 - 15 [ | 10 | 15 | 13 | 5 | 62 | 62,0 | 141 | 775 | 9 688 | 121 094 | 1 513 672 |
| [ 15 - 20 [ | 15 | 20 | 18 | 5 | 78 | 78,0 | 219 | 1 365 | 23 888 | 418 031 | 7 315 547 |
| [ 20 - 25 [ | 20 | 25 | 23 | 5 | 70 | 70,0 | 289 | 1 575 | 35 438 | 797 344 | 17 940 234 |
| [ 25 - 30 [ | 25 | 30 | 28 | 5 | 61 | 61,0 | 350 | 1 678 | 46 131 | 1 268 609 | 34 886 758 |
| **Totaux** | | 32 | | | 350 | | 350 | 5 853 | 118 318 | 2 628 554 | 61 835 833 |

Remarque : ici encore, nous traiterons successivement les données sur les effectifs, puis sur les fréquences relatives.

**Diagramme différentiel = histogramme**

Equirépartition des effectifs dans chaque classe : d'où sommets des rectangles parallèles à l'axe des abscisses.
Surface des rectangles proportionnelle aux effectifs (calcul des n'i ou des f'i).

| ei | n'i corr. |
|---|---|
| 0 | 0,0 |
| 0 | 20,0 |
| 2 | 20,0 |
| 2 | 0,0 |
| 2 | 32,5 |
| 4 | 32,5 |
| 4 | 0,0 |
| 4 | 42,5 |
| 6 | 42,5 |
| 6 | 0,0 |
| 6 | 51,3 |
| 10 | 51,3 |
| 10 | 0,0 |
| 10 | 62,0 |
| 15 | 62,0 |
| 15 | 0,0 |
| 15 | 78,0 |
| 20 | 78,0 |
| 20 | 0,0 |
| 20 | 70,0 |
| 25 | 70,0 |
| 25 | 0,0 |
| 25 | 61,0 |
| 30 | 61,0 |
| 30 | 0,0 |




L'unité d'aire correspond à la surface du 1er rectangle. Le tableau nous indique que cette surface équivaut à 8 employés.

Dans le cas d'une variable continue, on repère la classe modale qui correspond à l'effectif corrigé (ou à la fréquence corrigée) le plus élevé. Ici, on a donc :

**classe modale = [ 15 - 20 [**

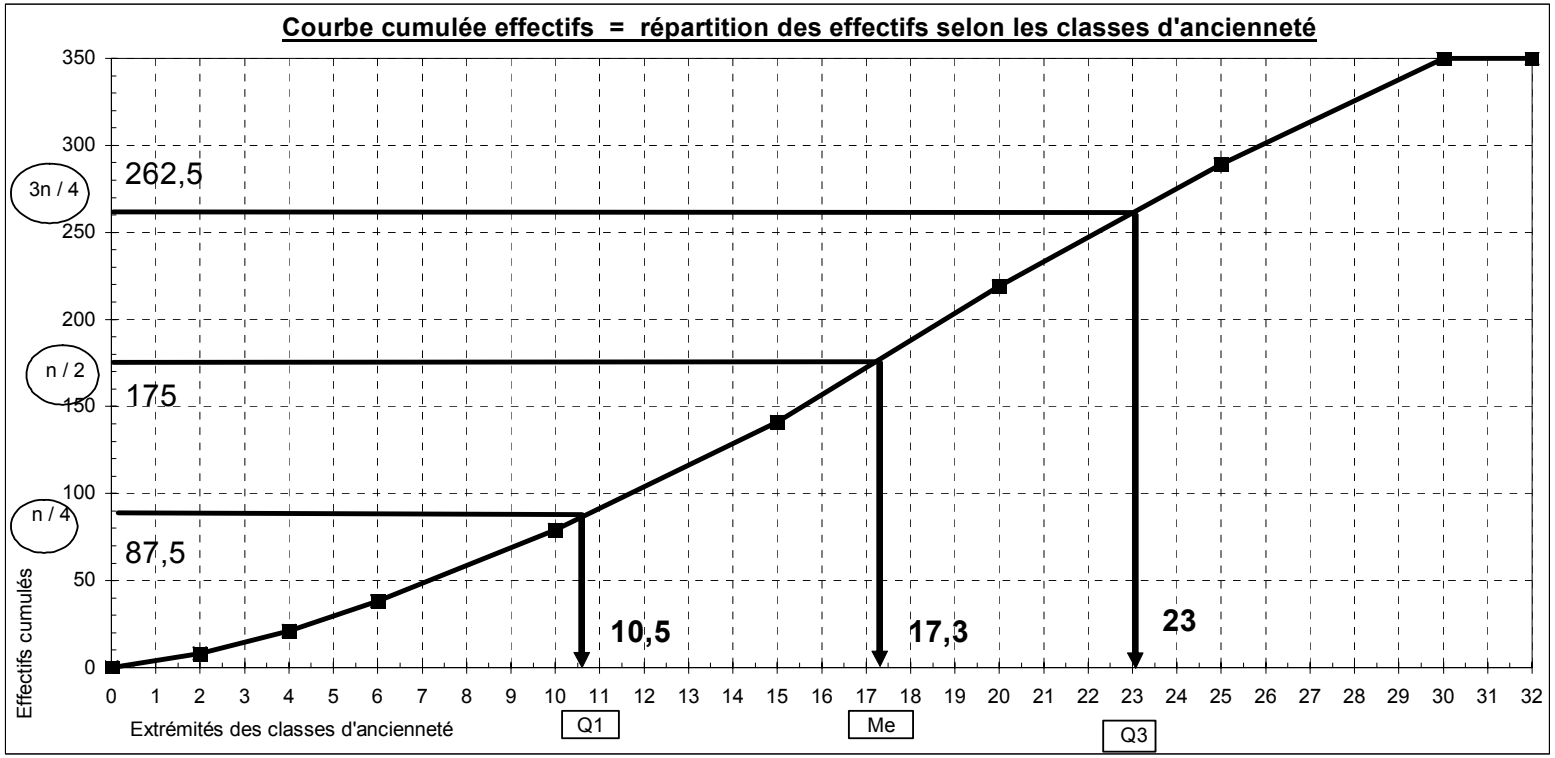
**Mo = 17,5 ans**

On peut envisager deux **valeurs ponctuelles du mode** (caractéristique de tendance centrale souvent peu appropriée) :

- hyp. d'équirépartition des effectifs dans la classe modale ---> centre de la classe et Mo = 17,5 ans d'ancienneté.

- on tient compte du fait que cette équirépartition est valable pour les classes adjacentes et, graphiquement, par la méthode des diagonales (double interpolation linéaire), on obtient Mo = 18,3 ans.

**Diagramme intégral = courbe cumulée continue**



Le **quartile Q1** est la valeur de la modalité telle qu'on trouve 25 % de l'effectif au-dessous de cette valeur et 75 % au-dessus.

La **médiane (ou quartile Q2)** (caractéristique de tendance centrale) est la valeur de la modalité telle qu'on trouve 50 % de l'effectif au-dessous de cette valeur et 50 % au-dessus. C'est la valeur de la modalité qui partage l'effectif en 2 sous-ensembles égaux.

Le **quartile Q3** est la valeur de la modalité telle qu'on trouve 75 % de l'effectif au-dessous de cette valeur et 25 % au-dessus.

Ici,on a :

25 % des employés ont moins de 10,7 ans d'ancienneté et 75 % plus de 10,7 ans d'ancienneté

50 % des employés ont moins de 17,2 ans d'ancienneté et 50 % plus de 17,2 ans d'ancienneté

75 % des employés ont moins de 23,1 ans d'ancienneté et 75 % plus de 23,1 ans d'ancienneté

50 % des employés ont entre 23,1 et 10,7 ans d'ancienneté .

L'écart interquartile Q3 - Q1 (caractéristique de dispersion autour de la médiane) est de 12,4 ans .

Le rapport interquartile Q3 / Q1 est de 2,16 .
Nombre sans dimension qui indique ici une assez forte dispersion autour de la médiane.

| | | |
|---|---|---|
| **Q1 =** | **10,69** | **ans** |
| **Me =** | **17,18** | **ans** |
| **Q3 =** | **23,11** | **ans** |

| | | |
|---|---|---|
| **Q3 - Q1 =** | **12,42** | **ans** |

| | |
|---|---|
| **Q3 / Q1 =** | **2,16** |

**Etendue = range = 30 ans**
(borne sup. dernière classe - borne inf. première classe = 30 - 0)
(caractéristique de dispersion, en général peu appropriée)

- Equirépartition des effectifs dans chaque classe.
- Centres de classes = moyennes de classes

Alors que la moyenne est une caractéristique de tendance centrale, l'écart-type est une caractéristique de dispersion autour de la moyenne.

| | | |
|---|---|---|
| **Moyenne =** | **16,72** | **ans** |
| **Var =** | **58,44** | **ans** |
| **Ectyp =** | **7,64** | **ans** |

**CV = ectyp / xbar = 0,46**

La dispersion est moyenne (< 0,5) .

Le coefficient de variation (CV) permet d'obtenir un nombre sans dimension, qui facilite les comparaisons de dispersion avec d'autres échantillons.

**Coefficient de Yule = -0,05**
**Etalement à gauche**

Si s de Yule est > 0 , il y a étalement à droite.
Si s de Yule est < 0 , il y a étalement à gauche.

On a : s = (Q3 - 2Me + Q1) / (Q3 - Q1)
(nombre sans dimension / caractéristique d'asymétrie)

L'étalement à gauche est faible et l'on a :

**classe modale = [ 15 - 20 [ < Me = 17,18 < Moyenne = 16,72**
**Mo = 17,5**

**Pearson 1 = -0,10**

**Etalement à gauche (Pearson 1 négatif)**

Pearson 1 = (xbar - Mo) / ectyp (nombre sans dimension)

Caractéristique d'asymétrie, souvent peu appropriée.

| | | |
|---|---|---|
| m1 = | xbar = s nici / n = | 16,72 |
| m2 = | s nici² / n = | 338,05 |
| m3 = | s nici$^3$ / n = | 7 510,15 |
| m4 = | s nici$^4$ / n = | 176 673,81 |

**Fisher 1 = -0,22**

**Etalement à gauche (Fisher 1 négatif)**

gamma1 = mu3 / ectyp^3 (nombre sans dimension)

= (m3 - 3m1.m2 + 2m1^3) / ectyp^3

Caractéristique d'asymétrie d'une série statistique.

**Fisher 2 = -0,97**

**Distribution platicurtique (-2 < Fisher 2 < 0)**

gamma2 = (mu4 / ectyp^4) - 3 (nombre sans dimension)

= [(m4 - 4m1.m3 + 6m1^2.m2 - 3m1^4) / ectyp^4] - 3

Caractéristique d'aplatissement d'une série statistique.

## Ancienneté (en années) des 350 employés d'une entreprise

Population statistique : 350 employés d'une entreprise  Individu statistique : l'un des 350 employés

Caractère statistique : l'ancienneté (en années) des employés  Type du caractère : quantitatif (discret traité en) continu

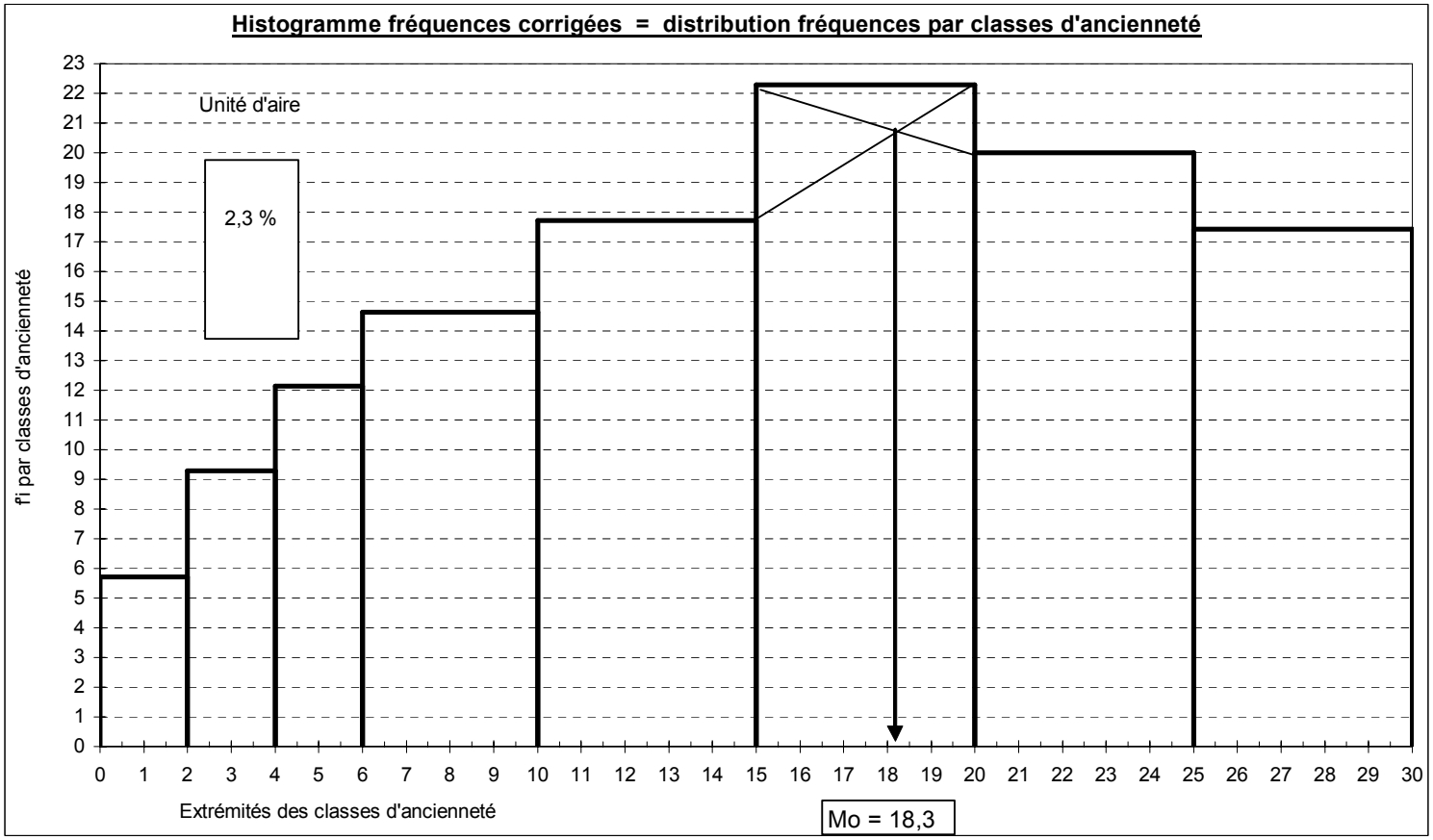
f'i = fi / ai x 500

| Classes d'années d'ancienneté | Borne inf. | Borne sup. | ci | ai | ni | fi x100 | f'i corr. X10000 | F(x) x100 | fi ci | fi ci² | fi ci³ | fi ci⁴ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | | | 0,0% | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | | 0,0 | 0,0% | | | | |
| **[ 0 - 2 [** | 0 | 2 | 1 | 2 | 8 | 2,3% | 5,7 | 2,3% | 0,02 | 0,02 | 0,02 | 0,02 |
| **[ 2 - 4 [** | 2 | 4 | 3 | 2 | 13 | 3,7% | 9,3 | 6,0% | 0,11 | 0,33 | 1,00 | 3,01 |
| **[ 4 - 6 [** | 4 | 6 | 5 | 2 | 17 | 4,9% | 12,1 | 10,9% | 0,24 | 1,21 | 6,07 | 30,36 |
| **[ 6 - 10 [** | 6 | 10 | 8 | 4 | 41 | 11,7% | 14,6 | 22,6% | 0,94 | 7,50 | 59,98 | 479,82 |
| **[ 10 - 15 [** | 10 | 15 | 12,5 | 5 | 62 | 17,7% | 17,7 | 40,3% | 2,21 | 27,68 | 345,98 | 4 324,78 |
| **[ 15 - 20 [** | 15 | 20 | 17,5 | 5 | 78 | 22,3% | 22,3 | 62,6% | 3,90 | 68,25 | 1 194,38 | 20 901,56 |
| **[ 20 - 25 [** | 20 | 25 | 22,5 | 5 | 70 | 20,0% | 20,0 | 82,6% | 4,50 | 101,25 | 2 278,13 | 51 257,81 |
| **[ 25 - 30 [** | 25 | 30 | 27,5 | 5 | 61 | 17,4% | 17,4 | 100,0% | 4,79 | 131,80 | 3 624,60 | 99 676,45 |
| **Totaux** | 0 | 32 | 0 | 0 | 350 | 100,0% | | 100,0% | 16,72 | 338,05 | 7 510,15 | 176 673,81 |

**Diagramme différentiel = histogramme**

Equirépartition des effectifs dans chaque classe : d'où sommets des rectangles parallèles à l'axe des abscisses.
Surface des rectangles proportionnelle aux effectifs (calcul des n'i ou des f'i).

| ei | f'i corr. |
|---|---|
| 0 | 0,0 |
| 0 | 5,7 |
| 2 | 5,7 |
| 2 | 0,0 |
| 2 | 9,3 |
| 4 | 9,3 |
| 4 | 0,0 |
| 4 | 12,1 |
| 6 | 12,1 |
| 6 | 0,0 |
| 6 | 14,6 |
| 10 | 14,6 |
| 10 | 0,0 |
| 10 | 17,7 |
| 15 | 17,7 |
| 15 | 0,0 |
| 15 | 22,3 |
| 20 | 22,3 |
| 20 | 0,0 |
| 20 | 20,0 |
| 25 | 20,0 |
| 25 | 0,0 |
| 25 | 17,4 |
| 30 | 17,4 |
| 30 | 0,0 |




L'unité d'aire correspond à la surface du 1er rectangle. Le tableau nous indique que cette surface équivaut à 2,3 % des employés.

Dans le cas d'une variable continue, on repère la classe modale qui correspond à l'effectif corrigé (ou à la fréquence corrigée) le plus élevé. Ici, on a donc :

**classe modale = [ 15 - 20 [**

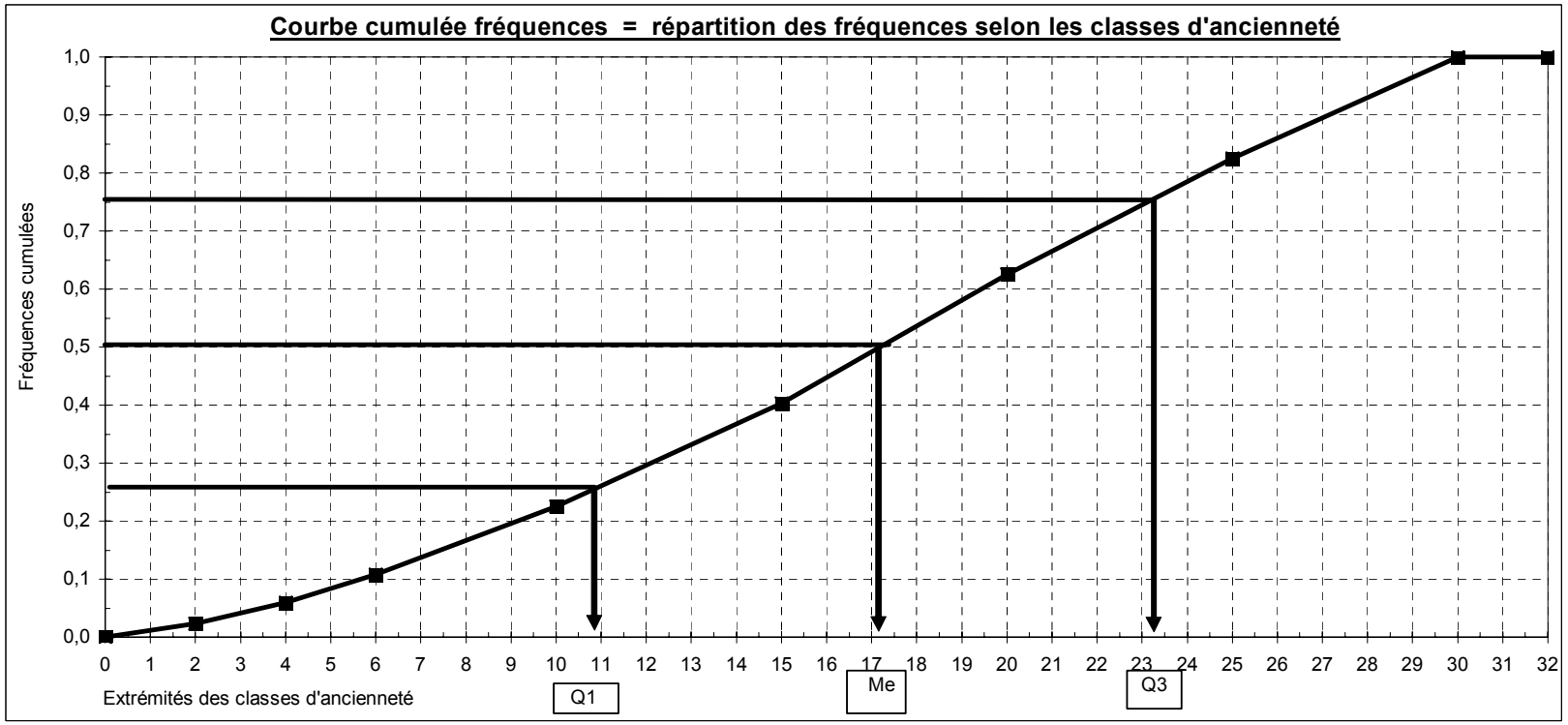
**Mo = 17,5 ans**

On peut envisager deux **valeurs ponctuelles du mode** (caractéristique de tendance centrale souvent peu appropriée) :

- hyp. d'équirépartition des effectifs dans la classe modale ---> centre de la classe et Mo = 17,5 ans d'ancienneté.

- on tient compte du fait que cette équirépartition est valable pour les classes adjacentes et, graphiquement, par la méthode des diagonales (double interpolation linéaire), on obtient Mo = 18,3 ans.

**Diagramme intégral = courbe cumulée continue**



Le **quartile Q1** est la valeur de la modalité telle qu'on trouve 25 % de l'effectif au-dessous de cette valeur et 75 % au-dessus.

La **médiane (ou quartile Q2)** (caractéristique de tendance centrale) est la valeur de la modalité telle qu'on trouve 50 % de l'effectif au-dessous de cette valeur et 50 % au-dessus. C'est la valeur de la modalité qui partage l'effectif en 2 sous-ensembles égaux.

Le **quartile Q3** est la valeur de la modalité telle qu'on trouve 75 % de l'effectif au-dessous de cette valeur et 25 % au-dessus.

Ici,on a :

25 % des employés ont moins de 10,7 ans d'ancienneté et 75 % plus de 10,7 ans d'ancienneté

50 % des employés ont moins de 17,2 ans d'ancienneté et 50 % plus de 17,2 ans d'ancienneté

75 % des employés ont moins de 23,1 ans d'ancienneté et 75 % plus de 23,1 ans d'ancienneté

50 % des employés ont entre 23,1 et 10,7 ans d'ancienneté .

L'écart interquartile Q3 - Q1 (caractéristique de dispersion autour de la médiane) est de 12,4 ans .

Le rapport interquartile Q3 / Q1 est de 2,16 .
Nombre sans dimension qui indique ici une assez forte dispersion autour de la médiane.

**Q1 = 10,69 ans**

**Me = 17,18 ans**

**Q3 = 23,11 ans**

**Q3 - Q1 = 12,42 ans**

**Q3 / Q1 = 2,16**

**Etendue = range = 30 ans**
(borne sup. dernière classe - borne inf. première classe = 30 - 0)
(caractéristique de dispersion, en général peu appropriée)

- Equirépartition des effectifs dans chaque classe.
- Centres de classes = moyennes de classes

Alors que la moyenne est une caractéristique de tendance centrale,
l'écart-type est une caractéristique de dispersion autour de la moyenne.

| | | |
|---|---|---|
| **Moyenne =** | **16,72** | **ans** |
| **Var =** | **58,44** | **ans** |
| **Ectyp =** | **7,64** | **ans** |

**CV = ectyp / xbar = 0,46**

La dispersion est moyenne (< 0,5) .

Le coefficient de variation (CV) permet d'obtenir un nombre sans dimension, qui facilite les comparaisons de dispersion avec d'autres échantillons.

**Coefficient de Yule = -0,05**
**Etalement à gauche**

Si s de Yule est > 0 , il y a étalement à droite.
Si s de Yule est < 0 , il y a étalement à gauche.

On a : s = (Q3 - 2Me + Q1) / (Q3 - Q1)
(nombre sans dimension / caractéristique d'asymétrie)

L'étalement à gauche est faible et l'on a :

**classe modale = [ 15 - 20 [ < Me = 17,18 < Moyenne = 16,72**
**Mo = 17,5**

**Pearson 1 = -0,10**

**Etalement à gauche (Pearson 1 négatif)**

Pearson 1 = (xbar - Mo) / ectyp (nombre sans dimension)

Caractéristique d'asymétrie, souvent peu appropriée.

| | | |
|---|---|---|
| m1 = | xbar = s fici = | 16,72 |
| m2 = | s fici² = | 338,05 |
| m3 = | s fici$^3$ = | 7 510,15 |
| m4 = | s fici$^4$ = | 176 673,81 |

**Fisher 1 = -0,22**

**Etalement à gauche (Fisher 1 négatif)**

gamma1 = mu3 / ectyp^3 (nombre sans dimension)

= (m3 - 3m1.m2 + 2m1^3) / ectyp^3

Caractéristique d'asymétrie d'une série statistique.

**Fisher 2 = -0,97**

**Distribution platicurtique (-2 < Fisher 2 < 0)**

gamma2 = (mu4 / ectyp^4) - 3 (nombre sans dimension)

= [(m4 - 4m1.m3 + 6m1^2.m2 - 3m1^4) / ectyp^4] - 3

Caractéristique d'aplatissement d'une série statistique.

**Ancienneté (en années) des 350 employés d'une entreprise**

**Population statistique : 350 employés d'une entreprise**
**Individu statistique : l'un des 350 employés**

**Caractère statistique : l'ancienneté (en années) des employés**
**Type du caractère : quantitatif (discret traité en) continu**

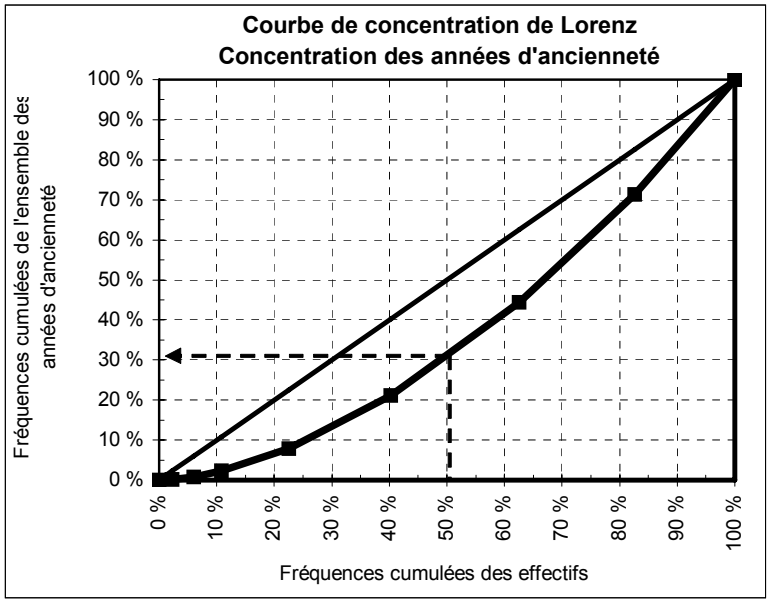
| Classes d'années d'ancienneté | Borne inf. | Borne sup. | ci | ai | ni | N(x) | ni ci | nici / Snici x100 (si) | si x100 cumulés S(x) x100 | fi x100 | F(x) x100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 0,0% | | 0,0% |
| [ 0 - 2 [ | 0 | 2 | 1 | 2 | 8 | 8 | 8 | 0,1% | 0,1% | 2,3% | 2,3% |
| [ 2 - 4 [ | 2 | 4 | 3 | 2 | 13 | 21 | 39 | 0,7% | 0,8% | 3,7% | 6,0% |
| [ 4 - 6 [ | 4 | 6 | 5 | 2 | 17 | 38 | 85 | 1,5% | 2,3% | 4,9% | 10,9% |
| [ 6 - 10 [ | 6 | 10 | 8 | 4 | 41 | 79 | 328 | 5,6% | 7,9% | 11,7% | 22,6% |
| [ 10 - 15 [ | 10 | 15 | 12,5 | 5 | 62 | 141 | 775 | 13,2% | 21,1% | 17,7% | 40,3% |
| [ 15 - 20 [ | 15 | 20 | 17,5 | 5 | 78 | 219 | 1 365 | 23,3% | 44,4% | 22,3% | 62,6% |
| [ 20 - 25 [ | 20 | 25 | 22,5 | 5 | 70 | 289 | 1 575 | 26,9% | 71,3% | 20,0% | 82,6% |
| [ 25 - 30 [ | 25 | 30 | 27,5 | 5 | 61 | 350 | 1 678 | 28,7% | 100,0% | 17,4% | 100,0% |
| Totaux | | | | | 350 | | 5 853 | 100,0% | 100,0% | 100,0% | 100,0% |

Dans le tableau ci-dessous, la valeur 0,0133 est donnée par :
(0,00 + 11,76) / 2 x 22,7

Le résultat de cette opération est divisé par 10.000, de façon à se ramener à un raisonnement en fréquence (carré de côté égal à 1) .

| F(x) x100 | S(x) x100 | Equi-répart. | Inég. totale |
|---|---|---|---|
| 0,0 % | 0,0 % | 0,0 % | 0 % |
| 2,3 % | 0,1 % | 2,3 % | 0 % |
| 6,0 % | 0,8 % | 6,0 % | 0 % |
| 10,9 % | 2,3 % | 10,9 % | 0 % |
| 22,6 % | 7,9 % | 22,6 % | 0 % |
| 40,3 % | 21,1 % | 40,3 % | 0 % |
| 62,6 % | 44,4 % | 62,6 % | 0 % |
| 82,6 % | 71,3 % | 82,6 % | 0 % |
| 100,0 % | 100,0 % | 100,0 % | 0 % |
| 100,0 % | 100,0 % | 100,0 % | 100 % |

La moitié des employés (les + récemment embauchés) représente 30 % de l'ensemble des années d'ancienneté.
L'autre moitié (les plus anciens) représente 70 % de l'ensemble des années d'ancienneté.




| fi | S(x) | S sous courbe |
|---|---|---|
| | 0,0 % | 0,0000 |
| 2,3 % | 0,1 % | 0,0002 |
| 3,7 % | 0,8 % | 0,0007 |
| 4,9 % | 2,3 % | 0,0059 |
| 11,7 % | 7,9 % | 0,0257 |
| 17,7 % | 21,1 % | 0,0730 |
| 22,3 % | 44,4 % | 0,1158 |
| 20,0 % | 71,3 % | 0,1493 |
| 17,4 % | 100,0 % | |

| | | |
|---|---|---|
| **S tot.** | **ss courbe** | **0,371** |
| **S** | **concentr.** | **0,129** |
| **Indice** | **Gini** | **0,259** |

**Concentration faible**

## 5. Autres moyennes

La moyenne arithmétique est un cas particulier de moyenne, qui n'est pas adaptée dans les deux cas suivants :

a) lorsqu'on a affaire à des taux d'accroissement moyens ou bien lorsque des moyennes sont effectuées à partir de coefficients multiplicateurs (cf. maths financières). Dans de tels cas, on utilise une moyenne géométrique G.

b) lorsque les moyennes à calculer sont réalisées à partir de rapports (quotients), comme des vitesses (= distance / temps), alors on utilise une moyenne harmonique H.

Remarque : lorsqu'on a affaire à des moyennes d'écarts au carré par rapport à une valeur (le plus souvent) centrale, on utilise une moyenne quadratique Q (cf. notion de moment centré).

### 51. La moyenne géométrique

On l'utilise obligatoirement (sans quoi l'on obtient des résultats erronés) lorsqu'on raisonne sur des expressions comportant des coefficients multiplicateurs et lorsqu'on cherche à calculer des taux d'accroissement moyens.

Avant de préciser la définition de la moyenne géométrique simple et celle de la moyenne géométrique pondérée, nous revenons sur la **notion de taux de croissance** applicable à l'évolution d'une grandeur (économique) dans le temps.

#### 511. Taux de croissance et propriétés

Hypothèse : soit une grandeur G, qui prend les valeurs suivantes :

$$g_0 \text{ en } t = 0 \quad \text{et} \quad g_1 \text{ en } t = 1$$

Définition : le taux de croissance de G entre les deux dates 0 et 1, est égal à :

$$\text{taux de croissance} = r = (g_1 - g_0) / g_0 = (g_1 / g_0) - 1$$

On exprime généralement le taux de croissance r en pourcentage. On multiplie donc que le résultat obtenu par 100.

Notion de multiplicateur associé au taux de croissance r

On a : $r = (g_1 / g_0) - 1 \quad \Leftrightarrow \quad 1 + r = (g_1 / g_0)$

ou bien : $\mathbf{g_1 = g_0 (1 + r)}$

(1 + r) est le (coefficient) multiplicateur associé au taux de croissance r.

Exemples d'utilisation

Si, entre les dates 0 et 1, on suppose que le niveau atteint par la grandeur G **double**, on aura :

$$g_1 = g_0 (1 + r) = 2 g_0 \quad \text{avec :} \quad 2 = 1 + r. \quad \text{Donc :} \quad r = 1 = 100\ \%$$

Remarque : un accroissement de 100 % correspond à un multiplicateur associé de $2 = 1 + r$.

Si, entre les dates 0 et 1, on suppose que le niveau atteint par la grandeur G **triple**, on aura :

$$g_1 = g_0 (1 + r) = 3 g_0 \quad \text{avec :} \quad 3 = 1 + r. \quad \text{Donc :} \quad r = 2 = 200\ \%$$

Remarque : un accroissement de 200 % correspond à un multiplicateur associé de $3 = 1 + r$.

et ainsi de suite ... Par exemple, un accroissement de 1 500 % correspond à un multiplicateur associé de $16 = 1 + r$, avec : $r = 15 = 1\ 500\ \%$.

Non symétrie des taux de croissance (à la hausse et à la baisse)

Soit les valeurs suivantes de G :

| t | 0 | 1 |
|---|---|---|
| | 100 | 130 |

En appliquant la définition du taux de croissance, on a donc un accroissement de la valeur de G de **30 %** entre les deux dates : $[(130 - 100) / 100) \times 100] = 30\ \%$.

Supposons maintenant que l'on ait eu les valeurs suivantes :

| t | 0 | 1 |
|---|---|---|
| | 130 | 100 |

En appliquant la définition du taux de croissance, on observe donc une diminution de la valeur de G de **23,1 %** entre les deux dates : $[(100 - 130) / 130) \times 100] = - 23{,}1\ \%$.

**Retenir qu'une hausse de r % n'est pas annulée (symétrisée) par une baisse de r %.**

Taux de croissance de deux grandeurs liées entre elles

Hypothèse : on considère trois grandeurs A, B et G, telles que : $\mathbf{g = a \times b}$

En $t = 0$, on a donc ; $g_0 = a_0 \times b_0$ .

En $t = 1$, on a donc ; $g_1 = a_1 \times b_1$ .

À partir de la définition du taux de croissance, on a :

$$g_1 = g_0 (1 + r) \quad \text{et :} \quad a_1 = a_0 (1 + r_1) \quad \text{et :} \quad b_1 = b_0 (1 + r_2)$$

On peut donc écrire : $g_1 = a_1 b_1 = g_0 (1 + r) = a_0 (1 + r_1) \times b_0 (1 + r_2)$

$$g_0 (1 + r) = a_0 b_0 (1 + r_1) (1 + r_2) \Leftrightarrow \underline{\mathbf{(1 + r) = (1 + r_1) . (1 + r_2)}}$$

car : $g_0 = a_0 b_0 \Leftrightarrow g_0 / a_0 b_0 = 1$

Conclusion : le coefficient multiplicateur (1 + r) associé au taux de croissance r d'un produit est égal au produit des coefficients multiplicateurs associés respectivement à chacune des grandeurs du produit.

De même, si **g = a / b**, avec : $g_0 = a_0 / b_0$ en t = 0 et $g_1 = a_1 / b_1$ en t = 1.

Alors, par une démonstration similaire, avec $b, b_0, b_1 \neq 0$ , on obtient :

$$g_1 = g_0 (1 + r) = a_0 (1 + r_1) / b_0 (1 + r_2) \Leftrightarrow \underline{\mathbf{(1 + r) = (1 + r_1) / (1 + r_2)}}$$

car : $g_0 = a_0 / b_0 \Leftrightarrow g_0 / (a_0 / b_0) = 1$

Conclusion : le coefficient multiplicateur (1 + r) associé au taux de croissance r d'un quotient est égal au quotient des coefficients multiplicateurs associés respectivement à chacune des grandeurs du quotient.

Exemples

| | t = 0 | t = 1 | Tx évolution 1 / 0 | Coeff. multiplicateur (1 + r) |
|---|---|---|---|---|
| **a** | 30 | 33 | + 10 % | 1,10 |
| **b** | 25 | 28 | + 12 % | 1,12 |
| **g = a x b** | 750 | 924 | + 23,2 % | 1,232 |

$g_0 = a_0 \times b_0 = 30 \times 25 = 750$ $g_1 = a_1 \times b_1 = 33 \times 28 = 924$

$g_1 = g_0 (1 + r) = a_0 (1 + r_1) \times b_0 (1 + r_2)$
$924 = 750 (1,232) = 30 (1,1) \times 25 (1,12)$

et : $1,232 = (1,1) \times (1,12) \Leftrightarrow (1 + r) = (1 + r_1) . (1 + r_2)$

| | t = 0 | t = 1 | Tx évolution 1 / 0 | Coeff. multiplicateur (1 + r) |
|---|---|---|---|---|
| **a** | 30 | 33 | + 10 % | 1,10 |
| **b** | 25 | 28 | + 12 % | 1,12 |
| **g = a / b** | 1,2 | 1,179 | - 1,8 % | 0,982 |

$g_0 = a_0 / b_0 = 30 / 25 = 1,2$ $g_1 = a_1 \times b_1 = 33 / 28 = 1,179$

$g_1 = g_0 (1 + r) = a_0 (1 + r_1) / b_0 (1 + r_2)$
$1,179 = 1,2 (0,982) = 30 (1,1) / 25 (1,12)$

et : $0,982 = (1,1) / (1,12) \Leftrightarrow (1 + r) = (1 + r_1) / (1 + r_2)$

### 512. Taux de croissance moyen et moyenne géométrique

Supposons qu'on enregistre les valeurs de la grandeur G sur **n périodes** :

| Que | 0 | 1 | 2 | ... | n |
|---|---|---|---|---|---|
| Valeurs de G | $g_0$ | $g_1$ | $g_2$ | ... | $g_n$ |

On peut envisager deux types d'évolution :

1) le taux de croissance r est constant pour chaque période et le coefficient multiplicateur est égal à (1 + r) :

$$g_1 = g_0 (1 + r)$$

$$g_2 = g_1 (1 + r) = g_0 (1 + r)^2$$

…

$$g_n = g_0 (1 + r)^n \text{ (cf. maths financières)} \quad \mathbf{(1)}$$

2) le taux de croissance est différent à chaque période (par exemple, évolution chronologique d'une grandeur économique) :

$$g_1 = g_0 (1 + r_1)$$

$$g_2 = g_1 (1 + r_2) = g_0 (1 + r_1)(1 + r_2)$$

…

$$g_n = g_{n-1} (1 + r_n) = g_0 (1 + r_1)(1 + r_2) \ldots (1 + r_n) \quad (2)$$

Supposons alors que l'on recherche un taux de croissance constant qui, appliqué n fois (i.e. sur les n périodes) à la grandeur G, donne le même niveau $g_n$ (à la date t = n) qu'une suite de taux de croissance différents à chaque période.

Cette hypothèse revient à égaliser les deux résultats précédents :

$$g_n = g_0 (1 + r)^n = g_0 (1 + r_1)(1 + r_2) \ldots (1 + r_n)$$

$$(1 + r)^n = (1 + r_1)(1 + r_2) \ldots (1 + r_n)$$

ou encore : $\boxed{(1+r) = \sqrt[n]{(1+r_1)(1+r_2)\ldots(1+r_n)}}$

Remarque : pour faciliter la compréhension, on peut passer par n = 2 :

$$(1+r)^2 = (1+r_1)(1+r_2) \Leftrightarrow (1+r) = \sqrt{(1+r_1)(1+r_2)}$$

Dans l'expression trouvée, le taux r est appelé **taux de croissance moyen périodique**.

Ce taux r représentant un pourcentage constant qui, appliqué durant la même période de temps, donne la même croissance de la grandeur G que les différents taux successifs $r_1$, $r_2$, ..., $r_n$ .

Notion de moyenne géométrique

Dans l'expression trouvée, posons : $1 + r = G$ .

De même, posons : $1 + r_1 = x_1$ ; $1 + r_2 = x_2$ ; … ; $1 + r_n = x_n$ .

Il vient : $$G = \sqrt[n]{x_1 . x_2 . \ldots . x_n}$$

C'est l'expression de la moyenne géométrique simple. Ici il s'agit de la moyenne géométrique des multiplicateurs associés $1 + r_1$ , $1 + r_2$ , … , $1 + r_n$ .

Définition de la moyenne géométrique

C'est la racine $n^{ième}$ du produit de n valeurs **positives** du caractère x .

Le caractère x est ici le coefficient multiplicateur associé au taux de croissance annuel d'une grandeur économique, telle que le PIB ou un chiffre d'affaires par exemple.

Remarque : on peut résumer l'écriture de la moyenne géométrique, en utilisant le symbole $\prod$, qui s'interprète de la manière suivante : $\prod_{i=1}^{n} x_i = x_1 . x_2 . \ldots . x_n$ ($\prod$ joue ici le même rôle que $\sum$ dans le cas de la moyenne arithmétique).

Par suite, on a : Moyenne géométrique simple : $$G = \sqrt[n]{\prod_{i=1}^{k} x_i}$$

avec : k = nombre de modalités du caractère x et : n = effectif total.

**Remarque : dans la formule, nous distinguons effectif (n) et nombre de modalités (k), même si, dans le cas d'une moyenne simple, on a : n = k.**

Analogie avec la moyenne arithmétique

On a : $G = \sqrt[n]{\prod_{i=1}^{k} x_i} = (\prod_{i=1}^{k} x_i)^{1/n}$ . En passant aux logarithmes décimaux :

$$\log G = \frac{1}{n} \log (\prod_{i=1}^{k} x_i) = \frac{1}{n} \log (x_1 + x_2 + \ldots + x_k)$$

$$\log G = \frac{1}{n} (\log x_1 + \log x_2 + ... + \log x_k) = \frac{1}{n} \sum_{i=1}^{k} \log x_i$$

Ainsi, le logarithme d'une moyenne géométrique est égal à la moyenne arithmétique des logarithmes des observations d'une série statistique.

Moyenne géométrique pondérée

Pour l'introduire, on peut supposer que, parmi les k coefficients multiplicateurs $1 + r_i$ , certains d'entre eux sont identiques. Dès lors, on peut effectuer des regroupements dans un tableau, comme on l'a déjà fait pour la moyenne arithmétique.

Dans l'exemple précédent, on va supposer que l'on a k modalités différentes, c-à-d que k coefficients multiplicateurs $(1 + r_i)$ sont différents.

Par exemple, on peut voir apparaître plusieurs fois le même taux de croissance $r_1$ ; en termes de coefficients multiplicateurs, si $1 + r_1$ revient $n_1$ fois, on aura donc :

$$\boxed{(1+r_1)(1+r_1)\dots(1+r_1) = (1+r)^{n_1}}$$

On peut procéder de même avec chacun des autres taux distincts. Par suite :

$$G = \sqrt[n]{x_1^{n_1} . x_2^{n_2} .... x_k^{n_k} .} = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}}$$

avec : $n_i$ = effectif partiel correspondant à la modalité $x_i$.

Si l'on raisonne en fréquences, l'on obtient : $G = \prod_{i=1}^{k} x_i^{f_i}$ , avec : $f_i = n_i / n$

Remarque : on peut en effet écrire, en posant fi = ni / n :

$$G = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}} = (\prod_{i=1}^{k} x_i^{n_i})^{1/n} = (x_1^{n_1} . x_2^{n_2} .... x_k^{n_k})^{1/n}$$

$$G = x_1^{n_1/n} . x_2^{n_2/n} .... x_k^{n_k/n} = x_1^{f_1} . x_2^{f_2} .... x_k^{f_k}$$

Finalement : $G = \prod_{i=1}^{k} x_i^{f_i}$

Remarque : bien noter que, lorsqu'on raisonne en fréquences, la racine $n^{ième}$ disparaît.

## 513. Exemples

**Exemple 1**

Le tableau suivant donne, sur une année, l'évolution mensuelle du chiffre d'affaires d'une entreprise, exprimée sous la forme de pourcentages d'évolution du chiffre d'affaires d'un mois sur l'autre (détermination au dernier jour de chaque mois) :

| | Janvier | Février | Mars | Avril | Mai | Juin | Juillet | Août | Septembre | Octobre | Novembre | Décembre |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Évolution CA (en %)** | 0,70% | 0,40% | 0,20% | **- 0,30%** | **- 0,20%** | 0,10% | 0,70% | 0,80% | 0,50% | 0,50% | 0,60% | 0,40% |
| **Coeff. multiplic.** | 1,007 | 1,004 | 1,002 | **0,997** | **0,998** | 1,001 | 1,007 | 1,008 | 1,005 | 1,005 | 1,006 | 1,004 |

Le coefficient multiplicateur global pour l'ensemble de l'année est donné par :

$$1 + a = (1 + m_1)(1 + m_2) \ldots (1 + m_{12}) \qquad \text{ou bien :} \qquad 1 + a = \prod_{t=1}^{12} (1 + m_t)$$

$a$ = taux global d'accroissement du chiffre d'affaires sur l'ensemble de la période (année) ;
$m_t$ = taux mensuel réel d'accroissement (mois t) du chiffre d'affaires.

Numériquement : $1 + a = (1{,}007) \times (1{,}004) \times (1{,}002) \times (\mathbf{0{,}997}) \times (\mathbf{0{,}998}) \times 1{,}001) \times \ldots \times (1{,}04) = 1{,}0448$ et : **$a = 4{,}48\ \%$**.

Si l'on recherche le taux mensuel moyen d'accroissement m qui, appliqué chaque mois sur l'ensemble de l'année, donne le même taux annuel d'accroissement que a, alors on pose (moyenne géométrique simple) :

$$1 + m = \sqrt[12]{\prod_{t=1}^{12} (1 + m_t)} = \sqrt[12]{1{,}0448} = 1{,}00366 \qquad \text{et :} \quad \mathbf{m = 0{,}37\ \%}$$

Remarque : ici, il serait faux de calculer une moyenne arithmétique en sommant les pourcentages et en divisant la somme obtenue par 12. Le principe de calcul serait faux, car on change de valeur de référence à la fin de chaque mois. Cependant, lorsque les taux d'évolution sont très petits et que le nombre de périodes n'est pas trop important, on n'observe pas de gros écarts entre la moyenne géométrique et la moyenne arithmétique. À l'inverse, si les pourcentages d'évolution sont élevés et si le nombre de périodes envisagées est grand, les résultats peuvent devenir nettement divergents.

**Exemple 2**

Mise en œuvre d'une moyenne géométrique pondérée

Supposons que le PIB d'un pays s'accroisse de 2 % par an durant 3 ans, puis de 1,7 % par an pendant 2 ans et enfin de 2,2 % par an pendant un an.

Quel est le taux moyen annuel d'accroissement du PIB ?

Soit $x$ la valeur du PIB en $t = 0$. On peut dresser le tableau suivant :

| **t** | **0** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|---|
| **PIB** | x | $(1,02)\,x$ | $(1,02)^2\,x$ | $(1,02)^3\,x$ | $(1,02)^3$ $(1,017)\,x$ | $(1,02)^3$ $(1,017)^2\,x$ | $(1,02)^3$ $(1,017)^2$ $(1,022)\,x$ |

En utilisant la formule de la moyenne géométrique pondérée, on a :

$$1 + a = \sqrt[6]{(1,02)^3 . (1,017)^2 . 1,022} = \sqrt[6]{1,12174} = 1,01933$$

Le taux annuel moyen d'accroissement du PIB sur l'ensemble des 6 années est égal à :

**a = 1,93 %**

### 52. La moyenne harmonique

On l'utilise obligatoirement lorsqu'on cherche à calculer des moyennes qui portent sur des rapports (ratios, quotients).

Définition

La moyenne harmonique H est la valeur de la modalité du caractère x pour laquelle son inverse 1 / H est égal à la moyenne arithmétique de l'inverse des valeurs de la série.

Remarque : la forme 1 / H permet de faire apparaître une analogie de construction avec la moyenne arithmétique :

$$\frac{1}{H} = \frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_k}\right) = \frac{1}{n}\sum_{i=1}^{k}\frac{1}{x_i}$$

Moyenne harmonique simple : $$H = \frac{n}{\sum_{i=1}^{k}\frac{1}{x_i}}$$

Moyenne harmonique pondérée : $$H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$

k = nombre de modalités du caractère x ;
n = effectif total ;
$n_i$ = effectif partiel correspondant à la modalité $x_i$ .

Rappel : dans le cas d'une moyenne simple, on a : n = k.

Si l'on raisonne en fréquences : $$H = \frac{1}{\sum_{i=1}^{k} \frac{f_i}{x_i}}, \quad \text{avec : } f_i = n_i / n$$

**Exemple 1**

Un automobiliste réalise une suite de trajets à des vitesses moyennes différentes. Par ex., on a :

trajet 1 : 60 km à 80 km/h  trajet 2 : 130 km à 125 km/h
trajet 3 : 30 km à 90 km/h  trajet 4 : 20 km à 45 km/h

On veut calculer la vitesse moyenne sur l'ensemble du parcours.

Remarque : il serait erroné ici de calculer une moyenne arithmétique pondérée, en sommant la valeur des trajets partiels (chacun étant multiplié par la vitesse correspondante), puis en divisant cette somme par la distance totale du parcours. Signalons qu'ici, on trouverait : 105,2 km/h.

Nous recherchons ici une vitesse moyenne qui est égale à :

$$\text{vitesse moyenne} = \frac{\text{distance totale parcourue (en km)}}{\text{temps total passé pour l'ensemble des déplacements (en h)}}$$

En ce qui concerne le numérateur de l'expression, il suffit d'ajouter les 4 distances partielles réalisées : 60 + 130 + 30 + 20 = **240 km = D**

Au dénominateur, on a : $$V = \frac{D}{T} \Leftrightarrow T = \frac{D}{V}$$

Cela signifie que le temps total passé pour effectuer l'ensemble des déplacements est égal à :

$$T = \sum_{i=1}^{4} \frac{d_i}{v_i} = \frac{60}{80} + \frac{130}{125} + \frac{30}{90} + \frac{20}{75} = 0,75 + 1,04 + 0,33 + 0,27$$

**T = 2,39 heures**

Remarque : ti = di / vi correspond à km / (km / h) → h

Finalement : $$V = \frac{D}{T} = \frac{240}{2,39} = \mathbf{100,4\ km/h}$$

Il s'agit ici d'une **moyenne harmonique**, portant sur des vitesses horaires (quotients), pondérées par les distances parcourues lors de chaque trajet :

$$V = \frac{D}{T} = \frac{n}{\sum_{i=1}^{4} \frac{n_i}{x_i}}$$

avec : n = distance totale parcourue ; ni = distance parcourue sur le trajet i ; xi = vitesse moyenne sur le trajet i.

$$V = \frac{60 + 130 + 30 + 20}{\frac{60}{80} + \frac{130}{125} + \frac{30}{90} + \frac{20}{75}}$$

$$V = \frac{60 + 130 + 30 + 20}{0{,}75 + 1{,}04 + 0{,}33 + 0{,}27} = \frac{240}{2{,}39} = \mathbf{100{,}4\ km/h}$$

**Exemple 2**

Une personne achète des dollars en échange d'euros, à trois dates différentes et aux cours suivants :

- premier achat : 1 200 €, au cours de 1,05 € / $ ;
- deuxième achat : 1 700 €, au cours de 1,17 € / $ ;
- troisième achat : 900 €, au cours de 1,12 € / $ .

Quel est le cours moyen du dollar en euro (€ / $) pour l'ensemble des trois transactions ?

**Il s'agit d'une moyenne harmonique, pondérée par les sommes échangées, portant sur des taux de change (quotients).**

$$\text{Cours moyen du \$ en €} = \frac{\text{dépense totale en euros}}{\text{achats totaux en dollars}} = \frac{n}{\sum_{i=1}^{3} \frac{n_i}{x_i}}$$

avec : n = dépense totale en euros ; ni = dépense en euros lors de l'opération i ; xi = taux de change appliqué lors de l'opération i .

$$\text{Cours moyen du \$ en €} = \frac{1200 + 1700 + 900}{\frac{1200}{1{,}05} + \frac{1700}{1{,}17} + \frac{900}{1{,}12}} = \frac{1200 + 1700 + 900}{1142{,}86 + 1452{,}99 + 803{,}57}$$

$$\text{Cours moyen du \$ en €} = \frac{3\,800}{3\,399{,}42} = 1{,}118 = \mathbf{1{,}12\ €/\$}$$

Remarque : il serait erroné ici encore de calculer une moyenne arithmétique pondérée, en sommant la valeur des achats partiels (chacun étant multiplié par le taux de change correspondant), puis en divisant cette somme par la somme totale dépensée en euros. Signalons qu'ici, on trouverait : 1,12056 €/$.

**Exemple 3**

Pour quatre communes, on connaît le nombre total d'habitants, ainsi que le nombre d'habitants pour une voiture :

| Communes | Population de la commune | Nombre d'habitants pour une voiture |
|---|---|---|
| **A** | 5 150 | 5,0 |
| **B** | 1 710 | 7,5 |
| **C** | 440 | 8,0 |
| **D** | 420 | 7,0 |
| **Total** | **7 720** | |

Quel est le nombre moyen d'habitants pour une voiture, pour l'ensemble des quatre communes ?

**Ici, on doit utiliser une moyenne harmonique, car le caractère statistique (le nombre d'habitants par voiture) apparaît sous la forme d'un rapport (quotient).**

$$\text{Nb hab. / vp} = \frac{\text{population totale des 4 communes}}{\text{nombre total de voitures dans les 4 communes}} = \frac{n}{\sum_{i=1}^{4} \frac{n_i}{x_i}}$$

avec : n = population totale des 4 communes ; ni = population de la commune i ; xi = nombre d'habitants par voiture dans la commune i .

$$\text{Nb moyen d'hab. / vp} = \frac{5\,150 + 1710 + 440 + 420}{\frac{5\,150}{5} + \frac{1710}{7{,}5} + \frac{440}{8} + \frac{420}{7}} = \frac{5\,150 + 1710 + 440 + 420}{1\,030 + 228 + 55 + 60}$$

$$\text{Nb moyen d'hab. / vp} = \frac{7\,720}{1373} = \textbf{5,6 (2) hab. / voiture}$$

Remarque : il serait erroné ici encore de calculer une moyenne arithmétique pondérée, en sommant le nombre d'habitants de chaque commune (chacun étant multiplié par le taux correspondant "d'occupation" d'une voiture), puis en divisant cette somme par le nombre total d'habitants des communes concernées. Signalons qu'ici, on trouverait : 5,83355 €/$.

Remarque terminale

Pour mémoire, signalons l'existence de la moyenne quadratique Q.

Par définition, il s'agit de la racine carrée de la moyenne arithmétique des carrés des observations d'une série statistique.

On l'utilise lorsqu'on réalise des calculs de moyennes d'écarts à une valeur (notamment à une valeur centrale, telle la moyenne arithmétique). cf. variance et écart-type.

Le fait d'élever au carré les valeurs des observations rend positifs tous les écarts à la valeur centrale. Cela représente un avantage par rapport aux écarts absolus moyens qui utilisent une valeur absolue, d'usage algébrique peu commode.

Moyenne quadratique simple :

$$Q = \sqrt{\frac{1}{n}\,(x_1^2 + x_2^2 + \ldots + x_k^2)} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} x_i^2}$$

avec : k = nombre de modalités du caractère x et : n = effectif total.

Rappel : dans le cas d'une moyenne simple, on a : k = n.

Moyenne quadratique pondérée : $Q = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i\, x_i^2}$

avec : $n_i$ = effectif partiel correspondant à la modalité $x_i$.

Si l'on raisonne en fréquences : $Q = \sqrt{\sum_{i=1}^{k} f_i\, x_i^2}$ , avec : $f_i = n_i / n$

### 53. Comparaisons entre moyennes

On considère la série suivante :

| Caractère x (xi) | Effectifs (ni) | ni xi | ni / xi | xi ni |
|---|---|---|---|---|
| 1 | 20 | 20 | 20,0 | 1 |
| 2 | 30 | 60 | 15,0 | 1 073 741 824 |
| 3 | 15 | 45 | 5,0 | 14 348 907 |
| 4 | 10 | 40 | 2,5 | 1 048 576 |
| 5 | 5 | 25 | 1,0 | 3 125 |
| 6 | 2 | 12 | 0,3 | 36 |
| **Totaux** | **82** | **202** | **43,8** | **1 817 486 218 641 770 000 000 000 000** |

Soit : **1,81749 x 10 27**

Moyenne arithmétique

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i\, x_i = \frac{202}{82} = 2{,}46$$

Moyenne harmonique

$$H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{82}{43{,}8} = 1{,}87$$

Moyenne géométrique

$$G = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}} = \sqrt[82]{1{,}81749 \times 10\text{^}27} = 2{,}15$$

On observe que : $H < G < \bar{x} \Leftrightarrow 1{,}87 < 2{,}15 < 2{,}46$

Cette double inégalité se démontre. Elle est toujours vérifiée, quelle que soit la série statistique envisagée.

Remarque : pour calculer la moyenne géométrique G, on peut passer par les logarithmes selon l'expression :

$$n \log G = 82 \log G = 20 \log 1 + 30 \log 2 + 15 \log 3 + 10 \log 4 + 5 \log 5 + 2 \log 6$$

Il vient : $\log G = 27{,}25947 / 82 = 0{,}33243 \Leftrightarrow G = 10^{0{,}33243} = 2{,}14997 =$ **2,15 G**

### 6. Les caractéristiques de concentration

Remarque préliminaire importante : ce point concerne uniquement les caractères quantitatifs **continus** qui prennent des **valeurs positives**.

La notion de concentration d'une série s'apparente à celle de dispersion.

Prenons l'exemple d'une série relative aux salaires distribués dans une entreprise.

a) Lorsqu'on calcule la **médiane** d'une telle série, on raisonne sur l'effectif total des salariés et l'on peut notamment répondre à la question suivante :

"Pour quel niveau de salaires dans l'entreprise a-t-on 50 % de l'effectif qui gagne en dessous de ce niveau et l'autre moitié qui gagne en dessus de ce niveau" ?

Le raisonnement est fait par rapport aux individus (les salariés) de cette entreprise. La médiane renvoie au nombre de salariés.

De même, lorsqu'on établit la courbe cumulée des effectifs (ou des fréquences), on raisonne également sur le nombre de salariés de l'entreprise. À partir de cette courbe, on peut chercher à savoir quel nombre de personnes gagne entre deux valeurs arbitrairement données.

b) Si maintenant, au lieu de raisonner sur l'ensemble des salariés, on raisonne sur le total des salaires versés par l'entreprise (la masse salariale), on peut déterminer **une autre valeur centrale**, qu'on appelle la **médiale** de la série.

La médiale d'une série statistique partage en deux sous-ensembles égaux la masse salariale (i.e. niveau de salaire moyen d'une classe x nombre d'individus qui touchent ce salaire) versés par une entreprise.

Dans cette optique, on peut répondre à une autre question :

"Quel nombre de salariés se voit attribuer la moitié de la masse salariale distribuée par l'entreprise" ?

Le raisonnement est fait par rapport au total des salaires versés par l'entreprise. La médiale renvoie à la masse salariale.

**Le concept de concentration résulte de ce changement de point de vue.**

On va s'attacher ici à répondre à la question suivante :

"Quelle proportion d'individus dans l'effectif total détient une proportion donnée de l'ensemble de la masse salariale de l'entreprise" ?

Si, graphiquement ou par le calcul, on détermine par exemple que 20 % des salariés qui gagnent le plus se voient attribuer 80 % de la masse salariale de l'entreprise, on dira que la concentration des salaires dans cette entreprise est forte. En effet, simultanément, ce résultat signifie que 80 % des salariés qui gagnent le moins se voient attribuer seulement 20 % de la masse salariale de l'entreprise !

### 61. La courbe de concentration de Max Otto Lorenz

Il s'agit d'une méthode graphique qui permet de rendre compte du phénomène de concentration d'une série statistique.

Nous utilisons les notations suivantes :

ni = effectifs partiels de chaque classe ;
n = effectif total ;
fi = ni / n = fréquences partielles de chaque classe ;
F (x) = fréquences cumulées.

ci = centres de classes ;
nici = effectifs partiels, pondérés par les centres de classes ;
si = nici / $\Sigma$ nici = fréquences partielles construites à partir des nici ;
S (x) = fréquences cumulées des si .

$$s_i = \frac{\text{effectif partiel } n_i \text{ de la classe } i, \text{ pondéré par le centre de classe } c_i \text{ correspondant}}{\text{somme des effectifs, pondérés par le centre de classe correspondant}}$$

$$s_i = \frac{n_i c_i}{\sum_{i=1}^{k} n_i c_i}$$

Si nous reprenons notre exemple :

$$s_i = \frac{\text{masse salariale moyenne versée dans la classe } i}{\text{masse salariale totale versée par l'entreprise}}$$

Le graphique fait apparaître :

- en abscisse les valeurs des fréquences (ou %) cumulées F (x) ;

- en ordonnée, les valeurs des fréquences (ou %) cumulées S (x) .

La courbe de Lorenz permet d'apprécier graphiquement la concentration de la série (i.e. le caractère plus ou moins inégalitaire de la distribution).

Remarque : on ne raisonne pas sur les effectifs, afin d'éviter le problème posé par les unités dans les comparaisons de populations ou d'échantillons. Chaque axe est gradué de 0 à 1 ou de 0 à 100 %. L'ensemble de la représentation graphique apparaît donc sous la forme d'un carré.

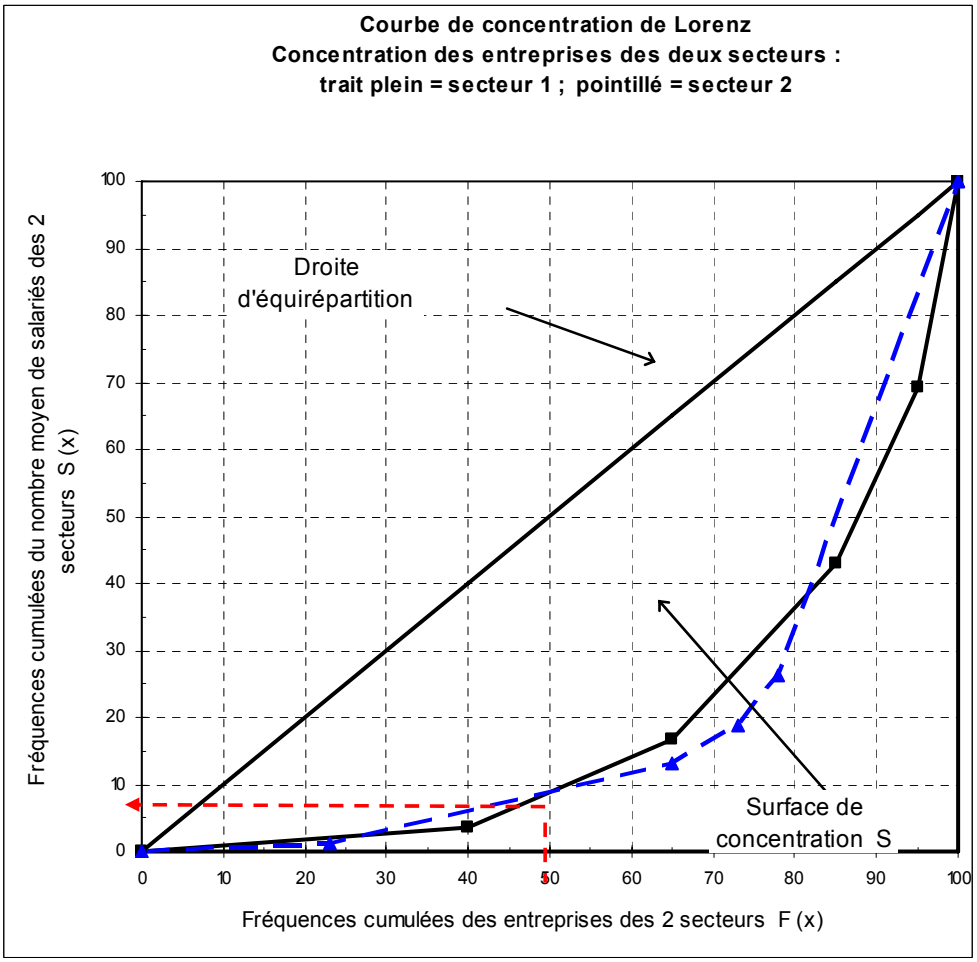
La forme générale du tableau de calcul des F (x) et des S (x) est la suivante :

| **Centres de classes ($c_i$)** | **Effectifs ($n_i$)** | **Fréquences relatives ($f_i = n_i / n$)** | **Fréquences cumulées F (x)** | **$n_i\ c_i$** | $s_i = \frac{n_i\ c_i}{\sum_{i=1}^{k} n_i\ c_i}$ | **Fréquences cumulées S (x)** |
|---|---|---|---|---|---|---|
| $c_1$ | $n_1$ | $f_1$ | F (1) | $n_1\ c_1$ | $s_1$ | S (1) |
| $c_2$ | $n_2$ | $f_2$ | F (2) | $n_2\ c_2$ | $s_2$ | S (2) |
| … | … | … | … | … | … | … |
| $c_i$ | $n_i$ | $f_i$ | F (i) | $n_i\ c_i$ | $s_i$ | S (i) |
| … | … | … | … | … | … | … |
| $c_k$ | $n_k$ | $f_k$ | F (k) = 1 | $n_k\ c_k$ | $s_k$ | S (k) = 1 |
| | n | 1 | | $\Sigma\ n_i\ c_i$ | $\Sigma\ s_i$ | |

Pour F (x) et S (x) : hypothèse d'équirépartition des effectifs dans les classes.

Pour S (x) : centres de classes = moyennes de classes, car on utilise les $c_i$ .

Exemple de tracé



Dans le graphique, la diagonale du carré joue un rôle fondamental. On l'appelle **droite d'équirépartition** :En chacun de ses points, il y a égalité entre la proportion de salariés de l'entreprise et la proportion de masse salariale distribuée par l'entreprise à ces salariés.

L'ensemble des deux côtés droit et bas du carré représente ce que l'on appelle **l'inégalité totale**. On aurait une telle configuration si un seul des salariés détenait la totalité de la masse salariale.

**La courbe de concentration de Lorenz** délimite une zone appelée **surface de concentration**, située entre la droite d'équirépartition et la courbe de concentration.

Concentration nulle : la distribution est totalement égalitaire (chaque employé touche le même salaire). La surface de concentration est égale à zéro et la courbe de concentration est confondue avec la droite d'équirépartition.

Distribution totalement inégalitaire : un seul salarié reçoit l'intégralité de la masse salariale.

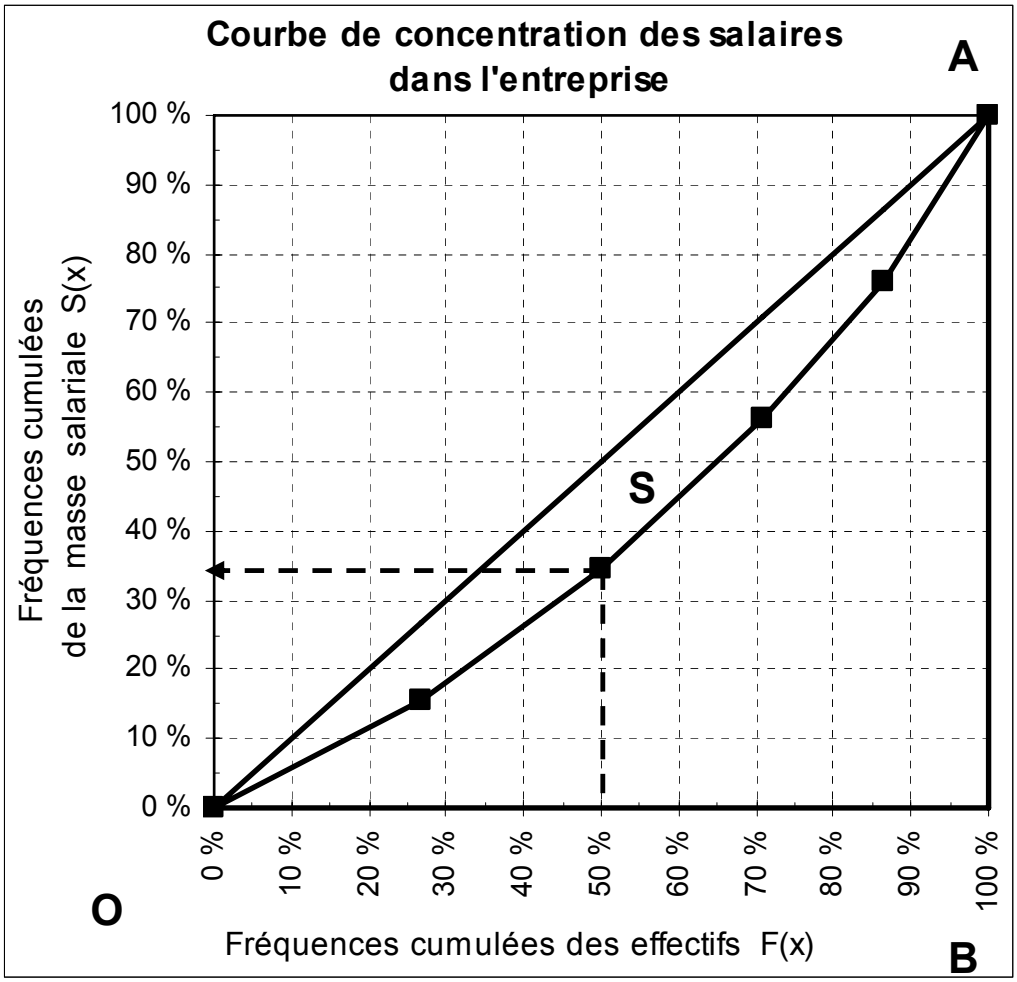
Dans la réalité, la concentration est faible lorsque de la courbe se rapproche de la droite d'équirépartition et elle est forte lorsque la surface de concentration tend à occuper tout le triangle situé en dessous de la diagonale du carré.

Noter que, pour un niveau donné de concentration, on peut avoir affaire à des structures d'inégalité très différentes (cf. graphique ci-dessus).

Exemple

**Les salaires annuels (exprimés en K€) des 212 employés d'une entreprise**

| Salaires annuels (en K€) | ci | ni | fi x100 | F(x) x100 | ni ci | nici / Snici x100 (si) | si x100 cumulés = S(x) x100 |
|---|---|---|---|---|---|---|---|
| **[10-15[** | 12,5 | 57 | 26,9% | 26,9% | 712,5 | 15,6% | 15,6% |
| **[15-20[** | 17,5 | 49 | 23,1% | 50,0% | 857,5 | 18,8% | 34,5% |
| **[20-25[** | 22,5 | 44 | 20,8% | 70,8% | 990,0 | 21,7% | 56,2% |
| **[25-30[** | 27,5 | 33 | 15,6% | 86,3% | 907,5 | 19,9% | 76,1% |
| **[30-45[** | 37,5 | 29 | 13,7% | 100,0% | 1 087,5 | 23,9% | 100,0% |
| **Totaux** | | **212** | 100,0% | 100,0% | **4 555,0** | 100,0% | 100,0% |




Interprétation du graphique

La médiane de la série (Me = 20 K€) partage en deux sous-ensembles égaux la population des salariés de l'entreprise. Le graphique permet de repérer (cf. flèches pointillées) que **la moitié des salariés qui gagnent le moins détient à peu près 35 % de la masse salariale versée par l'entreprise. L'autre moitié, qui gagnent le plus, détient 65 % de cette même masse salariale**.

S'il y avait une stricte égalité de la répartition des salaires, 50 % de l'effectif détiendraient 50 % de la masse salariale. En effet, la droite d'équirépartition établit une proportionnalité entre la proportion de salariés et la proportion de masse salariale correspondante (10 % des salariés détiennent 10 % de la masse salariale ; 20 % des salariés détiennent 20 % de la masse salariale ; etc.).

## 62. L'indice de Gini

### 621. Définition

L'indice de Gini, noté $I_G$ , est une valeur numérique qui résume l'information apportée par la courbe de concentration de Lorenz.

On a : $$\mathbf{I_G = 2 \times S}$$

En effet : $$I_G = \frac{\text{aire (surface) de concentration (S)}}{\text{aire (surface) du triangle OAB}}$$

Si l'on raisonne en fréquences, l'aire du carré est égale à un, puisque les côtés sont eux-mêmes égaux à un.

Par conséquent, l'aire du triangle OAB, située au-dessous de la diagonale est égale à 1 / 2 .

et : $$I_G = \frac{S}{0,5} = \frac{S}{1/2} = 2 \times S$$

Cette définition de l'indice de Gini permet d'obtenir à des valeurs, comprises entre 0 et 1. Il s'agit d'un nombre sans dimension, puisqu'il est établi sur la base du rapport de deux surfaces :

$$0 \leq I_G \leq 1$$

Lorsque la surface de concentration S est nulle, l'indice de Gini est égal à zéro : il y a alors équirépartition totale.

Lorsque la surface de concentration S est égale à 1 / 2 , l'indice de Gini est égal à un : il y a alors inégalité totale.

Plus la valeur de l'indice de Gini est élevée, plus la concentration est forte, c'est-à-dire plus les inégalités sont fortes.

On peut retenir les ordres de grandeur suivants :

$I_G \leq 0,3$ concentration faible ;
$0,3 \leq I_G \leq 0,5$ concentration moyenne ;
$I_G \geq 0,5$ concentration forte.

Remarques :

1) Si l'indice de Gini résume les informations apportées par une courbe de concentration de Lorenz, il en résulte cependant une perte d'information par rapport à cette dernière. En effet, dans certains cas, l'indice de Gini de deux distributions est à peu près le même, alors que l'allure des deux courbes de Lorenz correspondantes révèle des structures d'inégalité très différentes.

2) on peut multiplier par 100 la valeur de l'indice de Gini de pour obtenir un résultat en pourcentage (de concentration).

## 622. Détermination de la valeur de l'indice de Gini

Méthode graphique

Généralement (car l'ordre de grandeur du résultat obtenu est le plus souvent satisfaisant), on détermine graphiquement la valeur de l'indice de Gini, en calculant approximativement le nombre de carreaux contenus dans la surface de concentration S.

Sur le graphique de notre exemple, on dénombre environ 10 carreaux.

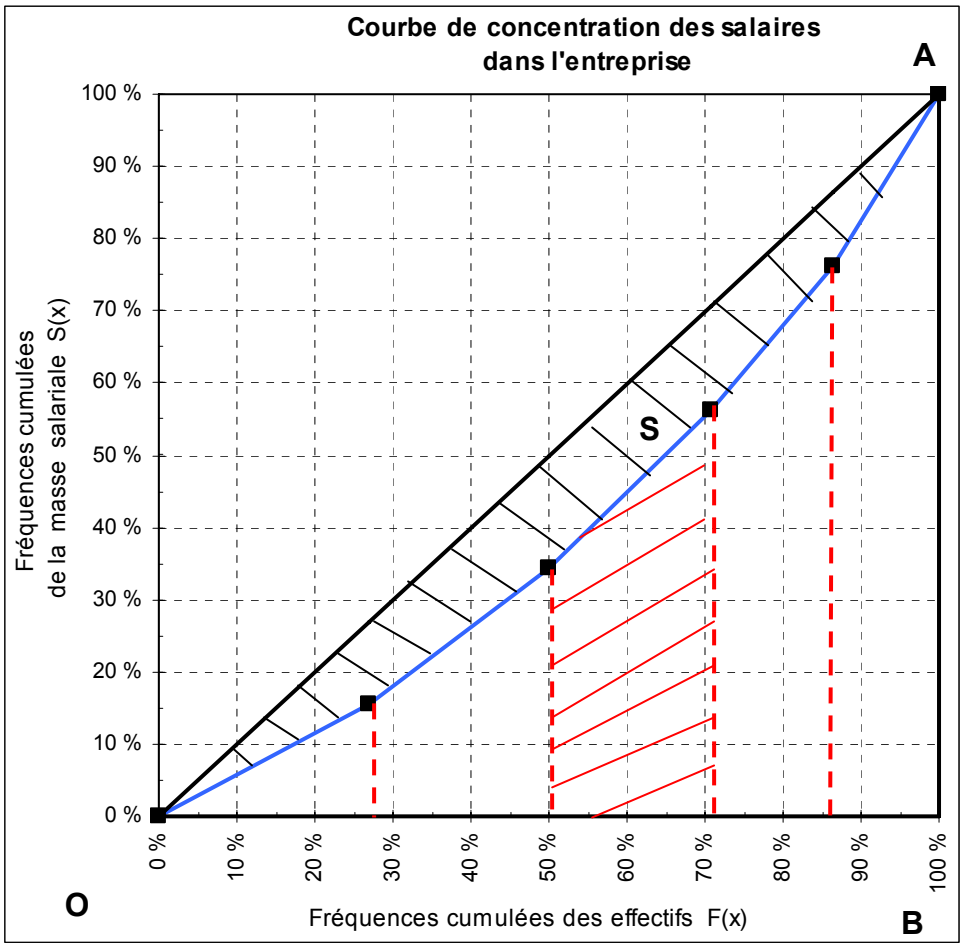
On multiplie cette valeur par 2 et il ne faut pas oublier de rapporter le résultat aux 100 carreaux qui constituent le quadrillage de l'ensemble du carré (ici de 10 % en 10 %), soit : $\mathbf{I_G = 0{,}20}$.

Interprétation : ici, l'indice de Gini est proche de 0 ; la concentration est relativement faible et la répartition de la masse salariale est peu inégalitaire.

Remarque : il faut bien noter ce à quoi correspond un carreau. Ici chaque carreau fait 10 % de côté. Il faut donc rapporter le décompte obtenu à 100 (10 x 10) carreaux. Si l'on raisonne sur des carreaux de 1 % de côté, il faut rapporter le décompte obtenu à 10 000 (100 x 100) carreaux ; etc.

Méthode algébrique (pour mémoire)

Sur le graphique de notre exemple, S représente la surface de concentration, hachurée en noir. Cette surface est comprise entre la diagonale du carré (droite d'équirépartition) et la courbe de concentration de Lorenz (en bleu).



Méthode des trapèzes

On considère la surface située au-dessous de la surface de concentration S pour mettre en œuvre la méthode des trapèzes (dont la surface de l'un d'eux a été matérialisée par des hachures rouges).

a) La formule suivante permet de calculer la surface de chacun des trapèzes rectangles (pour la 1ère classe, on a seulement un triangle), dont les abscisses sont données par les valeurs successives de la fonction cumulée F(x) :

$$\frac{B + b}{2} \times h = \frac{S(e_i) + S(e_{i+1})}{2} \times f_i$$

$e_i$ = extrémité inférieure de la classe i ;

$f_i = F(e_{i+1}) - F(e_i)$

$S(e_{i+1})$ et $S(e_i)$ représentent les valeurs correspondantes des ordonnées pour un trapèze donné.

b) Une fois calculée la surface de chaque trapèze, on somme les valeurs de toutes les surfaces. Cela nous donne la surface comprise entre la courbe de concentration et le bas du triangle OAB.

c) Ensuite, on détermine l'aire de concentration S par différence :

aire S = 0,5 - surface située au-dessous de la courbe de Lorenz

Remarque : 0,5 représente la valeur de la demi-surface du carré de côté 1, située au-dessous de la droite d'équirépartition.

d) Enfin, on multiplie le résultat obtenu par 2 pour obtenir la valeur de l'indice de Gini, conformément à la formule présentée plus haut.

Application à l'exemple :

| fi | S(x) | Surface au-dessous de la courbe de Lorenz |
|---|---|---|
| | **0,000** | **0,021** |
| **0,269** | **0,156** | 0,058 |
| 0,231 | 0,345 | 0,094 |
| 0,208 | 0,562 | 0,103 |
| 0,156 | 0,761 | 0,120 |
| 0,137 | 1,000 | |
| | | Somme = 0,396 |
| | Surface de concentration | = 0,500 - 0,396 = 0,104 |
| | **Indice de Gini** | = 2 x 0,104 = **0,207** |

Par exemple, on a : (0,000 + 0,156) / 2 x 0,269 = 0,021

La valeur trouvée ici de l'indice de Gini (0,207 ou encore 20,7 %) dénote une faible concentration des salaires dans l'entreprise.

Calcul et utilisation de la médiale (Ml)

La médiale (Ml) se détermine de la même façon que la médiane, mais à partir de la colonne S (x).

Dans notre exemple, on a : $\Sigma$ nici / 2 = 4 555 / 2 = 2 277,5 ou encore 50 % dans la colonne S (x). La classe médiale est donc la classe [20 - 25[ . Comme pour la médiane, on peut ici réaliser une interpolation linéaire en vue de calculer une valeur ponctuelle de la médiale :

$$Ml = 20 + (25-20) \times \frac{50 - S(20)}{S(25) - S(20)} = 20 + 5 \times \frac{50-34,5}{56,2-34,5} = 20 + 5 \times \frac{15,5}{21,7} = 23,57$$

Par ailleurs, nous avons : Me = 20 K€

Remarque : en général, on a : Ml $\geq$ Me. C'est en cas d'équirépartition que l'on a : Ml = Me .

**Écart $\Delta$M = Ml - Me = 23,57 - 20 = 3,57 K€**

Cet écart traduit l'intensité de la concentration des salaires dans l'entreprise. Plus il est important, plus la concentration est forte (i.e. plus la répartition des salaires est inégalitaire).

Parfois, afin d'obtenir un nombre sans dimension, on pose le rapport : $\Delta$M / étendue.

Ici, nous avons : $\Delta$M / étendue = 3,57 / (45 - 10) = 3,57 / 35 $\approx$ 0,102 , soit 10,2 %, ce qui dénote une faible concentration des salaires de l'entreprise.

Bien que le cet indicateur soit commode pour réaliser des comparaisons interpopulation, il n'est pas toujours très pertinent du fait que l'on utilise l'étendue dans son calcul.

### 7. Le traitement des effets de structure

Lorsqu'on compare des populations différentes, il est peu fréquent de disposer des mêmes pondérations pour chacune des modalités envisagées. Il en résulte que la comparaison des moyennes de ces populations est entachée d'un biais. Pour éliminer un tel biais, on doit rendre neutres ces pondérations différentes, qui génèrent ce que l'on appelle un **effet de structure**.

Par exemple, on peut vouloir comparer les niveaux de salaires moyens pratiqués dans deux entreprises, sachant que, dans chacune d'entre elles, on distingue plusieurs CSP.

Pour que la comparaison soit pertinente, il convient de veiller à deux choses :

- le nombre total d'employés est différent dans chaque entreprise : on peut remédier à cette difficulté en raisonnant en pourcentage, plutôt que selon les effectifs absolus (au niveau de l'effectif total de chaque entreprise et par CSP).

- le nombre d'employés par CSP est différent dans chaque entreprise. Or, dans chacune d'entre elles, les niveaux moyens des salaires de chaque CSP varient en fonction du nombre d'employés. Il en résulte que la valeur moyenne globale des salaires est influencée par le poids (pondération) de chaque CSP.

Plus la structure des emplois est différente entre les deux entreprises, plus la comparaison entre les niveaux des salaires moyens globaux est biaisée. Il est donc nécessaire de mettre en œuvre une méthode qui évite ce biais, de façon à pouvoir répondre correctement à une question du type : "le niveau des salaires d'une entreprise A est-il plus élevé que celui d'une entreprise B" ?

Exemple

Soit une entreprise qui fabrique trois types différents de moteurs de voiture (A, B et C). Le service de l'entreprise qui contrôle la fabrication a décidé de vérifier la bonne qualité des moteurs produits. Durant un mois donné, on suppose que le service a mis au rebut 532 moteurs de type A, 126 moteurs de type B et 187 moteurs de type C. Par rapport à la production totale du mois, ces trois valeurs correspondent à des taux de rejet respectifs de :

| **Type de moteur** | **A** | **B** | **C** |
|---|---|---|---|
| **Nombre de moteurs mis au rebut** | 532 | 126 | 187 |
| **Taux de rejet = proportion de moteurs mis au rebut** | 8 % | 4 % | 2 % |

Le service de contrôle de la fabrication procède de la façon suivante :

**I.** Il calcule le taux de rejet moyen, constaté durant le mois de contrôle, pour l'ensemble des trois moteurs.

**II.** Non satisfait des résultats, il décide de mettre en place de nouveaux procédés de fabrication, en vue d'améliorer la qualité des moteurs. Durant le mois suivant la mise en œuvre des nouveaux équipements, il procède à une nouvelle vérification, dont les résultats sont consignés ci-après :

| **Type de moteur** | **A** | **B** | **C** |
|---|---|---|---|
| **Nombre de moteurs mis au rebut** | 742 | 82 | 55 |
| **Taux de rejet = proportion de moteurs mis au rebut** | 7 % | 4 % | 1 % |

1°) On calcule à nouveau le taux de rejet moyen pour l'ensemble des trois moteurs.

2°) On compare les taux de rejet globaux avant et après la mise en service des nouveaux équipements, ainsi que l'évolution des taux de rejet par type de moteur.

3°) On recherche des solutions pour éliminer l'effet de structure constatées au 2°).

**I.**

$$\text{On a :} \quad \text{taux de rejet moyen global} = \frac{\text{nombre total de moteurs rejetés}}{\text{nombre total de moteurs produits}}$$

Il s'agit d'une moyenne harmonique : $$H = \frac{n}{\sum_{i=1}^{3} \frac{n_i}{x_i}} = \text{taux de rejet moyen global}$$

avec : $n$ = nombre total de moteurs rejetés, $n_i$ = nombre de moteurs rejetés, pour le moteur i et $x_i$ = taux de rejet du moteur i.

$$\text{taux de rejet moyen global I.} = \frac{532 + 126 + 187}{\frac{532}{0{,}08} + \frac{126}{0{,}04} + \frac{187}{0{,}02}} = \frac{845}{6650 + 3150 + 9350} = \frac{845}{19150}$$

$$= 0{,}0441 \text{ soit } 4{,}41\%$$

**II.**

1°) On reprend le même calcul avec les nouvelles valeurs :

$$\text{taux de rejet moyen global II.} = \frac{742 + 82 + 55}{\frac{742}{0,07} + \frac{82}{0,04} + \frac{55}{0,01}} = \frac{879}{10\,600 + 2\,050 + 5\,500} = \frac{879}{18\,150}$$

$= 0,0484$ soit 4,84 %

2°) Les résultats sont apparemment paradoxaux. En effet, le 2ème taux de rejet moyen global est supérieur au premier, alors même que les taux de rejet par types de moteur sont restés les mêmes ou ont diminué !! On a en effet :

| | **Avant amélioration** | **Après amélioration** |
|---|---|---|
| **Taux de rejet des moteurs de type A** | 8 % | 7 % |
| **Taux de rejet des moteurs de type B** | 4 % | 4 % |
| **Taux de rejet des moteurs de type C** | 2 % | 1 % |

Le paradoxe n'est qu'apparent, car il existe un **effet de structure** qui vient parasiter l'expression des valeurs moyennes et invalide les comparaisons.

En effet, intéressons-nous aux quantités produites de chaque moteur, avant et après la mise en œuvre des nouveaux procédés de fabrication, et comparons l'importance relative de la production de chacun des types de moteurs A, B, C, en calculant des pourcentages "de parts de marché".

Avant la mise en service des nouveaux procédés de fabrication, on a :

| | **A** | **B** | **C** | **Total** |
|---|---|---|---|---|
| **Nombre de moteurs produits** | 6 650 | 3 150 | 9 350 | 19 150 |
| **Poids dans la production totale** | 34,7 % | 16,5 % | 48,8 % | 100,0 % |

Après la mise en service des nouveaux procédés de fabrication, on a :

| | **A** | **B** | **C** | **Total** |
|---|---|---|---|---|
| **Nombre de moteurs produits** | 10 600 | 2 050 | 5 500 | 18 150 |
| **Poids dans la production totale** | 58,4 % | 11,3 % | 30,3 % | 100,0 % |

Que s'est-il passé entre les deux situations ?

La production totale a diminué, passant de 19 150 moteurs produits à 18 150.

La part de la production du moteur de type A, génératrice du plus grand nombre de rejets de fabrication, augmente d'environ 24 points dans la production totale de l'entreprise. Or, dans le même temps, les productions des moteurs de type B et C, moins génératrices de rejets, voient leur part diminuer dans la production totale (respectivement de 5 et de 19 points).

Il en résulte que, même si l'on note une baisse relative d'un point du taux de rejet de A, cette amélioration est plus que contrebalancée par la hausse de la production des moteurs de type A, qui continue de générer beaucoup de rejets relativement aux deux autres types de moteurs.

C'est donc cet effet de structure (c'est-à-dire le changement observé dans les parts respectives de production des trois types de moteurs A, B et C) qu'il faut éliminer, de façon à faire apparaître l'effet résiduel, dû à une variation effective de la qualité des produits (si elle existe et de combien).

3°) La solution du problème consiste à raisonner à quantités constantes de produits fabriqués et à parts de marché identiques, avant (I.) et après (II.) la mise en service des nouveaux procédés de fabrication. Cependant, la solution du problème n'est pas unique. De façon générale, on peut envisager trois types de solutions :

Solution 1

On conserve les quantités produites de la période I. (c'est-à-dire que l'on conserve la structure de fabrication de la période I.) et on leur applique les nouveaux taux de rejet (ceux de la période II.), de façon à recalculer le second taux de rejet moyen.

$$\text{taux de rejet moyen global} = \frac{\text{nombre total de moteurs rejetés}}{\text{nombre total de moteurs produits}}$$

$$\text{taux de rejet moyen global II.} = \frac{6\,650 \times 0{,}07 + 3\,150 \times 0{,}04 + 9\,350 \times 0{,}01}{6\,650 + 3\,150 + 9\,350}$$

$$= \frac{465{,}5 + 126 + 93{,}5}{19\,150} = \frac{685}{19\,150} = 0{,}0358 \quad \text{soit } 3{,}58\,\%$$

Dans ces conditions, on a : **taux de rejet moyen II. 3,58 % < taux de rejet moyen I. 4,41 %**
À quantités produites constantes et à structure de production constante, la qualité de la production globale s'est donc améliorée.

Solution 2

On retient les quantités produites de la période II. (c'est-à-dire que l'on retient la nouvelle structure de fabrication de la période II.) et on leur applique les anciens taux de rejet (ceux de la période I.), de façon à recalculer, cette fois-ci, le premier taux de rejet moyen.

$$\text{taux de rejet moyen global I.} = \frac{10\,600 \times 0{,}08 + 2\,050 \times 0{,}04 + 5\,500 \times 0{,}02}{10\,600 + 2\,050 + 5\,550}$$

$$= \frac{848 + 82 + 110}{18\,150} = \frac{1040}{18\,150} = 0{,}0573 \quad \text{soit } 5{,}73\,\%$$

Dans ces conditions, on a : **taux de rejet moyen II. 4,84 % < taux de rejet moyen I. 5,73 %**
À quantités produites constantes et à structure de production constante, la qualité de la production globale s'est donc améliorée.

Solution 3

On réalise un "compromis", c'est-à-dire qu'on fait la moyenne des quantités produites en I. et en II. pour chaque produit et pour l'ensemble de la production. Ensuite, on applique successivement les taux de rejet I. et II. à ces nouvelles valeurs moyennes.

Les productions moyennes (entre I. et II.) sont égales à :

Moteur A : $\frac{6\,650 + 10\,600}{2} = 8\,625$ Moteur B : $\frac{3\,150 + 2\,050}{2} = 2\,600$

Moteur C : $\frac{9\,350 + 5\,500}{2} = 7\,425$ Production totale : $\frac{19\,150 + 18\,150}{2} = 18\,650$

D'où, pour la période I. :

$$\text{taux de rejet moyen global I.} = \frac{8\,625 \times 0{,}08 + 2\,600 \times 0{,}04 + 7\,425 \times 0{,}02}{18\,650}$$

$$= 0{,}0505 \text{ soit } 5{,}05\ \%$$

Pour la période II. :

$$\text{taux de rejet moyen global II.} = \frac{8\,625 \times 0{,}07 + 2\,600 \times 0{,}04 + 7\,425 \times 0{,}01}{18\,650}$$

$$= 0{,}0419 \text{ soit } 4{,}19\ \%$$

Ici encore, on a : **taux de rejet moyen II. 4,19 % < taux de rejet moyen I. 5,05 %**

À quantités produites constantes et à structure de production constante, la qualité de la production globale s'est donc améliorée.

En conclusion, quelle solution retenir ?

De manière générale, et dépassant le cadre de cet exemple, si l'entreprise estime que le niveau et la structure de la production sont habituellement proches de ceux que l'on observe durant la période I., alors la solution 1 est la plus pertinente.

Si l'entreprise estime que le niveau et la structure de la production sont habituellement proches de ceux que l'on observe durant la période II., alors la solution 2 est la plus pertinente.

Enfin, si l'entreprise ne perçoit pas clairement la tendance de la production dans l'avenir, elle peut retenir la solution 3, qui constitue un compromis entre les deux solutions précédentes.

**Deuxième exemple sur le traitement des effets de structure**

**Entreprise Thermoplastik**

| | Mars 2004 | | Mars 2005 | | |
|---|---|---|---|---|---|
| | **Salaires mensuels (en €)** | **Effectifs** | **Salaires mensuels (en €)** | **Effectifs** | **Variation mars 05 / mars 04** |
| **Ouvriers non qualifiés** | 1 350 | 30 | 1 500 | 70 | **11,11%** |
| **Ouvriers qualifiés** | 1 800 | 70 | 1 950 | 30 | **8,33%** |
| | 1 665,00 | 100 | 1 635,00 | 100 | |

Entre mars 2004 et mars 2005, on note :

a) d'une part que le salaire mensuel de chaque CSP a crû de 150 € ,
b) d'autre part que, si l'effectif total des ouvriers n'a pas changé, les effectifs par CSP sont permutés.

Il ya effet de structure, en ce sens que si l'on calcule le salaire moyen, toutes CSP confondues, **une partie de la variation enregistrée entre les deux dates sur le salaire moyen, provient directement de la variation des pondérations (changement de qualification), observée sur les CSP** (et non du seul impact des variations de salaire de chaque CSP).

Salaire mensuel moyen en mars 2004 :
(1 350 x 30) + (1 800 x 70) / 100

**xbar 04**
**1 665,00 €**

Salaire mensuel moyen en mars 2005 :
(1 500 x 70) + (1 950 x 30) / 100

**xbar 05**
**1 635,00 €**

**-1,80%**

**On constate paradoxalement que le salaire mensuel moyen a diminué de 1,80 % entre les deux dates, alors même que le salaire mensuel de chacune des CSP a augmenté de 150 € !**

1 635 / 1 665 = 0,98198 et : 0,98198 - 1 = - 0,01802 , soit : - 1,80 % .

Afin d'estimer l'influence de ce changement dans la pondération des effectifs de chaque CSP, sur l'évolution réelle du salaire mensuel moyen, il est nécessaire de réaliser le calcul de ce dernier, **sans modifier les effectifs de chaque CSP**.

On observera cependant que l'on a le choix entre :
- solution 1 : conserver les effectifs (ou fréquences) initiaux (mars 2004),
- solution 2 : conserver les effectifs (ou fréquences) finaux (mars 2005).

**Solution 1 :**

$$\frac{\bar{x}_{05}}{\bar{x}_{04}} = \frac{\sum f_{05}\, x_{05}}{\sum f_{04}\, x_{05}} \times \frac{\sum f_{04}\, x_{05}}{\sum f_{04}\, x_{04}}$$

Le 1er terme correspond à l'effet de structure (portant sur les qualif.). Le 2è à l'effet de la variation du revenu, à effectifs 2004.

Salaire mensuel moyen en mars 2004 (S f04 x04) :
(1 350 x 30) + (1 800 x 70) / 100 — **1 665,00 €**

Salaire mensuel moyen en mars 2005 (S f04 x05) :
(1 500 x **30**) + (1 950 x **70**) / 100 — **1 815,00 €**

**-9,92%** **9,01%**

Conclusion : les salaires ont augmenté, en moyenne, de 9,01 % .

1 815 / 1 665 = 1,09009 et : 1,09009 - 1 = 0,09009 , soit : 9,01 % .

Cependant, le salaire mensuel moyen a baissé de 1,80 % , en raison d'un effet de structure qui a entraîné une **baisse de la qualification** des ouvriers :

1 635 / 1 815 = 0,90083 et : 0,90083 - 1 = - 0,09917 , soit : - 9,92 % .

**Solution 2 :**

$$\frac{\bar{x}_{05}}{\bar{x}_{04}} = \frac{\sum f_{05}\, x_{04}}{\sum f_{04}\, x_{04}} \times \frac{\sum f_{05}\, x_{05}}{\sum f_{05}\, x_{04}}$$

Le 1er terme correspond à l'effet de structure (portant sur les qualif.). Le 2è à l'effet de la variation du revenu, à effectifs 2005.

Salaire mensuel moyen en mars 2004 (S f05 x04) :
(1 350 x **70**) + (1 800 x **30**) / 100 — **1 485,00 €**

Salaire mensuel moyen en mars 2005 (S f05 x05) :
(1 500 x 70) + (1 950 x 30) / 100 — **1 635,00 €**

**-10,81%** **10,10%**

Conclusion : les salaires ont augmenté, en moyenne, de 10,10 % .

1 635 / 1 485 = 1,09009 et : 1,10101 - 1 = 0,10101 , soit : 10,10 % .

Cependant, le salaire mensuel moyen a baissé de 1,80 % , en raison d'un effet de structure qui a entraîné une **baisse de la qualification** des ouvriers :

1 485 / 1 665 = 0,89189 et : 0,89189 - 1 = - 0,10811 , soit : - 10,81 % .

## PARTIE 2 : ÉTUDE DES SÉRIES STATISTIQUES À DEUX CARACTÈRES

Jusqu'ici nous avons étudié des populations (ou des échantillons) en ne considérant qu'un seul caractère à la fois (l'âge, le revenu, le nombre de pannes d'une machine, etc.).

Dorénavant, nous allons envisager deux caractères simultanément. Par exemple :

- dans une même population donnée, les poids et taille des personnes qui la constituent (2 caractères quantitatifs).
- dans un pays donné, le nombre de personnes par ménage et la nationalité de ces ménages (un caractère quantitatif et un caractère qualitatif).
- dans une même population donnée, la couleur des yeux et la couleur des cheveux des individus qui la constituent (2 caractères qualitatifs).

Objectifs

1) Étudier deux caractères statistiques simultanément a pour but de déterminer s'il existe une dépendance ou une indépendance entre ces deux caractères.

2) Si une relation existe entre les deux caractères, on cherche à en mesurer l'intensité plus ou moins forte.

Exemples

a) en physique, un grand nombre de phénomènes sont reliés fonctionnellement (relation fonctionnelle). Ainsi, l'allongement d'un ressort est proportionnel à la masse que l'on attache à l'une de ses extrémités.

b) à l'inverse, d'autres phénomènes n'ont aucune incidence l'un sur l'autre. Par exemple, si l'on croise le salaire et la taille (ou le poids) d'un groupe de personnes, on ne décèle aucune relation entre ces deux caractères.

c) dans de nombreux autres cas, la relation de dépendance pouvant exister entre deux caractères sera plus ou moins marquée. Si le caractère 1 représente la série de Bacc. d'une promotion d'étudiants et si le caractère 2 est la note d'Économie (ou de Statistique) obtenue en première année, l'intensité de la liaison décelée entre ces deux caractères est susceptible de varier selon les années (en effet, pour l'ensemble d'une même promotion, la proportion des étudiants d'une série de Bacc. peut changer d'une année sur l'autre ; ou encore les niveaux moyens des étudiants de chaque série peuvent varier chaque année).

Remarques

Comme dans le cas des distributions à un caractère, il n'est pas possible de réaliser les mêmes calculs, selon le type des caractères mis en jeu. De même que précédemment, on ne peut guère réaliser de calculs sur les caractères qualitatifs, contrairement aux caractères quantitatifs.

Cependant, la présentation des données peut être effectuée de la même façon dans tous les cas. C'est pourquoi nous envisagerons tout d'abord un **tableau de base**, utilisable quel que soit le type des caractères envisagés. De nombreux concepts et notations nouveaux vont pouvoir être précisés dans le cadre de cette première approche.

Nous parlerons de **tableaux à double entrée** ou encore de **tableaux de contingence**.

La **contingence** d'une distribution à deux caractères mesure le degré d'interdépendance (ou l'intensité de l'association) entre ces deux caractères.

Sur un tableau de contingence, on peut calculer un coefficient de contingence c², qui permet de dire si deux caractères sont indépendants ou non, en utilisant la loi statistique du $\chi^2$, tel que :

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}} \quad \text{et :} \quad C^2 = \frac{\chi^2}{\chi^2 + n_{..}}$$

Si c² = 0, alors x et y sont indépendants et si c² = 0,5, alors x et y sont fonctionnellement liés.

## CHAPITRE 1 : LES TABLEAUX DE CONTINGENCE

Les tableaux de contingence (= tableaux à double entrée = tableaux croisés) sont des tableaux destinés au traitement de séries statistiques à deux caractères. Dans ces tableaux, les deux caractères sont **croisés** (lecture en ligne et en colonne). Selon le type des caractères étudiés, on effectue le croisement de deux caractères qualitatifs, ou d'un caractère qualitatif et d'un caractère quantitatif, ou encore de deux caractères quantitatifs. Dans chacun de ces cas, les traitements appropriés sont différents.

Principe

Tout comme on avait, pour les séries à un caractère, un tableau composé d'un certain nombre de colonnes, indiquant notamment les modalités du caractère et les effectifs concernés par chaque modalité, on va construire les tableaux à double entrée selon un principe similaire (évidemment avec quelques différences).

Deux points importants

1) Selon qu'on s'intéresse à l'un ou l'autre des caractères que l'on croise, on va trouver les renseignements les concernant, soit en colonne, conventionnellement x (comme pour les séries à un caractère), soit en ligne, conventionnellement y.

2) A l'intersection d'une ligne et d'une colonne du tableau principal, on trouve les effectifs (ou les fréquences relatives), correspondant simultanément à une modalité précise du premier caractère et à une modalité précise du deuxième caractère (effectifs conjoints), sachant que les modalités d'un caractère sont reportées en ligne et les modalités de l'autre caractère sont reportées en colonne.

### 1. Notations conventionnelles et tableau de base

Soit deux caractères x et y.

Objectifs

On veut savoir, pour une population donnée, si ces deux caractères sont liés entre eux :

- l'un des deux caractères influe-t-il unilatéralement sur l'autre et avec quelle intensité ?

- y a-t-il influence réciproque entre les deux caractères et avec quelle intensité ?

- y a-t-il indépendance entre les deux caractères ?

Pour conserver une présentation similaire à celle de l'étude des caractères à une dimension, nous conservons ici **en lignes les modalités du caractère x**. Les **modalités du caractère y** sont donc placées **en colonnes**.

De façon à concrétiser les choses, nous allons travailler sur un petit exemple numérique. Ce n'est que dans un deuxième temps que nous généralisons l'écriture de ce tableau de base général, en vue de l'appliquer à n'importe quel type de caractères (qualitatif, quantitatif discret, quantitatif continu).

Exemple numérique (3 x 3 modalités)

Le **caractère x** (modalités placées dans la première colonne du tableau) représente l'appréciation d'un étudiant en Économie, selon trois modalités A, B et C (avec A > B > C).

Le **caractère y** (modalités placées dans la première ligne du tableau) représente l'appréciation d'un étudiant en Statistiques, selon trois modalités A, B et C (avec A > B > C).

Supposons qu'il s'agisse d'un groupe de TD de 30 étudiants :

| $y_j$ (note Stat) / $x_i$ (notes Eco) | **A** | **B** | **C** |
|---|---|---|---|
| **A** | 6 | 3 | 1 |
| **B** | 4 | 8 | 0 |
| **C** | 2 | 1 | 5 |

À l'intérieur de ce tableau, on trouve, à l'intersection d'une ligne et d'une colonne, des effectifs qu'on appelle **effectifs conjoints**.

Par exemple, à l'intersection de la ligne B et de la colonne B, on dénombre 8 étudiants qui ont obtenu simultanément l'appréciation B en économie et l'appréciation B en statistique.

Afin de repérer commodément une case du tableau, on utilise une **notation matricielle** :

$n_{ij}$ = effectif conjoint, possédant la modalité i du caractère x (en ligne) et la modalité j du caractère y (en colonne).

En reprenant l'exemple précédent, on aura : $n_{22} = 8$ (huit étudiants ont obtenu à la fois B en économie et B en statistique).

Autre exemple : aucun étudiant n'a obtenu simultanément B en économie et C en statistique. C'est pourquoi nous avons : $n_{23} = 0$ .

**Première question** : que peut-on maintenant faire apparaître comme renseignements utiles supplémentaires ?

Nous pouvons faire apparaître les **effectifs obtenus en additionnant les valeurs de chaque ligne et / ou de chaque colonne du tableau**.

| $y_j$ (note Stat) / $x_i$ (notes Eco) | **A** | **B** | **C** | |
|---|---|---|---|---|
| **A** | 6 | 3 | 1 | **10** |
| **B** | 4 | 8 | 0 | **12** |
| **C** | 2 | 1 | 5 | **8** |
| | **12** | **12** | **6** | |

On parle **d'effectifs marginaux** (au sens marges du tableau principal).

**Dans la dernière colonne**, on trouve :

- pour la première ligne, on a l'ensemble des élèves ayant obtenu A en économie (10), quelle que soit leur note de statistique.

- pour la deuxième ligne, on a l'ensemble des élèves ayant obtenu B en économie (12), quelle que soit leur note de statistique.

- pour la troisième ligne, on a l'ensemble des élèves ayant obtenu C en économie (8), quelle que soit leur note de statistique.

Remarque : imaginons un tableau qui comporte seulement la première et la dernière colonne du tableau ci-dessus ; on aurait :

| $x_i$ (notes Eco) | $n_i$ |
|---|---|
| **A** | 10 |
| **B** | 12 |
| **C** | 8 |

Tout se passe comme si l'on étudiait la **distribution du seul caractère x** (i.e. la note d'économie). On parle donc des **effectifs marginaux du caractère x** et l'on parle de **distribution marginale du caractère x**, car on ignore complètement le caractère y (i.e. la note de statistique).

**Dans la dernière ligne**, on trouve :

- pour la première colonne, on a l'ensemble des élèves ayant obtenu A en statistique (12), quelle que soit leur note d'économie.

- pour la deuxième colonne, on a l'ensemble des élèves ayant obtenu B en statistique (12), quelle que soit leur note d'économie.

- pour la troisième colonne, on a l'ensemble des élèves ayant obtenu C en statistique (6), quelle que soit leur note d'économie.

Remarque : imaginons un tableau qui comporte seulement la première et la dernière ligne du tableau initial ; on aurait :

| $y_j$ (note Stat) | A | B | C |
|---|---|---|---|
| $n_j$ | 12 | 12 | 6 |

Tout se passe comme si l'on étudiait la **distribution du seul caractère y** (i.e. la note de statistique). On parle donc des **effectifs marginaux du caractère y** et l'on parle de **distribution marginale du caractère y**, car on ignore complètement le caractère x (i.e. la note d'économie).

**Deuxième question** : comment va-t-on noter littéralement les éléments de ces marges respectives (en colonne (x) et en ligne (y)) ?

Utilisation de la notation matricielle et conventions

| $y_j$ (note Stat)<br>$x_i$ (notes Eco) | A | B | C | $n_{i.}$ |
|---|---|---|---|---|
| **A** | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{1.}$ |
| **B** | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{2.}$ |
| **C** | $n_{31}$ | $n_{32}$ | $n_{33}$ | $n_{3.}$ |
| $n_{.j}$ | $n_{.1}$ | $n_{.2}$ | $n_{.3}$ | |

Si l'on somme les éléments de la première ligne du tableau, on obtient l'effectif marginal correspondant à la première modalité (A) du **caractère x**, c'est-à-dire la note d'économie :

$$n_{1.} = n_{11} + n_{12} + n_{13} = \sum_{j=1}^{3} n_{1j}$$

Le point (.) Indique, lorsqu'on lit le tableau en lignes, que l'on a réalisé une somme sur l'ensemble des modalités de la colonne (deuxième indice).

On procède de la même manière pour les trois lignes du tableau et l'on obtient :

$$n_{2.} = n_{21} + n_{22} + n_{23} = \sum_{j=1}^{3} n_{2j} \quad \text{et :} \quad n_{3.} = n_{31} + n_{32} + n_{33} = \sum_{j=1}^{3} n_{3j}$$

Par convention : $\mathbf{n_{i.}}$ représente l'intitulé de la dernière colonne.

**De même**, si l'on somme les éléments de la première colonne du tableau, on obtient l'effectif marginal correspondant à la première modalité (A) du **caractère y**, c'est-à-dire la note de statistique :

$$n_{.1} = n_{11} + n_{21} + n_{31} = \sum_{i=1}^{3} n_{i1}$$

Le point (▪) Indique, lorsqu'on lit le tableau en colonnes, que l'on a réalisé une somme sur l'ensemble des modalités de la ligne (premier indice).

On procède de la même manière pour les trois colonnes du tableau et l'on obtient :

$$n_{.2} = n_{12} + n_{22} + n_{32} = \sum_{i=1}^{3} n_{i2} \quad \text{et :} \quad n_{.3} = n_{13} + n_{23} + n_{33} = \sum_{i=1}^{3} n_{i3}$$

Par convention : $\mathbf{n_{.j}}$ représente l'intitulé de la dernière ligne.

**Troisième question** : que va-t-on trouver à l'intersection de la dernière colonne et de la dernière ligne du tableau (case située à l'intersection des deux marges) ?

| **$y_j$ (note Stat)**<br>**$x_i$ (notes Eco)** | **A** | **B** | **C** | $\mathbf{n_{i.}}$ |
|---|---|---|---|---|
| **A** | $n_{11}$ | $n_{12}$ | $n_{13}$ | $\mathbf{n_{1.}}$ |
| **B** | $n_{21}$ | $n_{22}$ | $n_{23}$ | $\mathbf{n_{2.}}$ |
| **C** | $n_{31}$ | $n_{32}$ | $n_{33}$ | $\mathbf{n_{3.}}$ |
| $\mathbf{n_{.j}}$ | $\mathbf{n_{.1}}$ | $\mathbf{n_{.2}}$ | $\mathbf{n_{.3}}$ | $\mathbf{n_{..}}$ |

En sommant les effectifs marginaux de la dernière colonne du tableau, on obtient :

$$n_{1.} + n_{2.} + n_{3.} = \mathbf{n_{..}} = 10 + 12 + 8 = 30 \quad \text{(effectif total)}$$

De même, en sommant les effectifs marginaux de la dernière ligne du tableau, on obtient :

$$n_{.1} + n_{.2} + n_{.3} = \mathbf{n_{..}} = 12 + 12 + 6 = 30 \quad \text{(effectif total)}$$

Finalement :

| $y_j$ (note Stat) / $x_i$ (notes Eco) | **A** | **B** | **C** | $\mathbf{n_{i.}}$ |
|---|---|---|---|---|
| **A** | 6 | 3 | 1 | **10** |
| **B** | 4 | 8 | 0 | **12** |
| **C** | 2 | 1 | 5 | **8** |
| $\mathbf{n_{.j}}$ | **12** | **12** | **6** | **30** |

**Généralisation**

Soit **m** le nombre de modalités du **caractère x**.
Soit **p** le nombre de modalités du **caractère y**.

Par convention, les m modalités du caractère x sont disposées dans la 1ère colonne (modalités $x_1$, …, $x_i$, …, $x_m$) et les p modalités du caractère y sont disposées dans la 1ère ligne (modalités $y_1$, …, $y_j$, …, $y_p$).

| $x_i$ \ $y_j$ | $y_1$ | … | $y_j$ | ... | $y_p$ | $\mathbf{n_{i.}}$ |
|---|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | … | $n_{1j}$ | | $n_{1p}$ | $\mathbf{n_{1.}}$ |
| … | … | … | … | … | … | … |
| $x_i$ | $n_{i1}$ | … | $\mathbf{n_{ij}}$ | … | $n_{ip}$ | $\mathbf{n_{i.}}$ |
| ... | … | … | … | … | … | … |
| $x_m$ | $n_{m1}$ | … | $n_{mj}$ | … | $n_{mp}$ | $\mathbf{n_{m.}}$ |
| $\mathbf{n_{.j}}$ | $\mathbf{n_{.1}}$ | … | $\mathbf{n_{.j}}$ | … | $\mathbf{n_{.p}}$ | $\mathbf{n_{..}}$ |

A l'intersection d'une ligne et d'une colonne sont reportés les **effectifs conjoints**. Par exemple, $\mathbf{n_{ij}}$ correspond à l'effectif des individus qui vérifient à la fois la modalité i du caractère x et la modalité j du caractère y.

La dernière ligne indique la somme des effectifs pour chacune des colonnes du tableau. On parle **d'effectifs marginaux de y**. Par exemple, pour la $j^{ème}$ colonne, on a : $n_{.j} = \sum_{i=1}^{m} n_{ij}$ . Par convention, on remarquera que la notation qui correspond à cette somme (terme général) est reprise comme intitulé de la ligne correspondante.

La dernière colonne indique la somme des effectifs pour chacune des lignes du tableau. On parle **d'effectifs marginaux de x**. Par exemple, pour la $i^{ème}$ ligne, on a : $n_{i.} = \sum_{j=1}^{p} n_{ij}$ . Par convention, on remarquera que la notation qui correspond à cette somme (terme général) est reprise comme intitulé de la colonne correspondante.

La valeur $n_{..}$ située dans la case en bas et à droite du tableau représente l'effectif total de la population sur laquelle on travaille. Cette valeur s'obtient de trois façons différentes : - en sommant les effectifs marginaux de x, dans la dernière colonne ; - en sommant les effectifs marginaux de y, dans la dernière ligne ; - en sommant les effectifs conjoints en ligne et en colonne.

$$n_{..} = \sum_{i=1}^{m} n_{i.} = \sum_{j=1}^{p} n_{.j} = \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij}$$

## 2. Tableaux de fréquences

Comme pour les séries à un caractère, on peut définir des fréquences relatives. Mais ici, du fait de la forme de la complexité du tableau à double entrée, on peut définir plusieurs types de fréquences :

- fréquences conjointes (nous venons de définir les effectifs conjoints) ;
- fréquences marginales (nous venons de définir les fréquences marginales en colonne en ligne) ;
- fréquences conditionnelles en colonnes (caractère x) ;
- fréquences conditionnelles en lignes (caractère y) .

### 21. Fréquences conjointes et fréquences marginales

Reprenons tout d'abord notre exemple numérique :

| $y_j$ (note Stat) / $x_i$ (notes Eco) | A | B | C | $n_{i.}$ |
|---|---|---|---|---|
| **A** | 6 | 3 | 1 | **10** |
| **B** | 4 | 8 | 0 | **12** |
| **C** | 2 | 1 | 5 | **8** |
| $n_{.j}$ | **12** | **12** | **6** | **30** |

**Par définition, une fréquence conjointe est égale au rapport d'un effectif conjoint à l'effectif total.**

**Soit :** $f_{ij} = n_{ij} / n_{..}$

Par exemple : $f_{11} = n_{11} / n_{..} = 6 / 30 = 0{,}20 = 20\ \%$ (correspond à la proportion des individus qui vérifient à la fois la modalité 1 (A) du caractère x **et** la modalité 1 (A) du caractère y).

| $y_j$ (note Stat) / $x_i$ (notes Eco) | A | B | C | $f_{i.}$ |
|---|---|---|---|---|
| **A** | 20,0 % | 10,0 % | 3,3 % | 33,3 % |
| **B** | 13,3 % | 26,7 % | 0,0 % | 40,0 % |
| **C** | 6,7 % | 3,3 % | 16,7 % | 26,7 % |
| $f_{.j}$ | 40,0 % | 40,0 % | 20,0 % | 100,0 % |

Dans le tableau ci-dessus, nous avons ajouté les marges du tableau obtenu.

La dernière colonne du tableau correspond aux **fréquences marginales du caractère x**.

De même que pour les effectifs conjoints, si l'on somme les éléments de la première ligne du tableau, on obtient la fréquence marginale correspondant à la première modalité (A) du **caractère x**, c'est-à-dire la note d'économie :

$$f_{1.} = f_{11} + f_{12} + f_{13} = \sum_{j=1}^{3} f_{1j}$$

Par convention : $f_{i.}$ représente l'intitulé de la dernière colonne.

Cependant, on peut obtenir les fréquences marginales du caractère x d'une autre façon, en utilisant les effectifs marginaux et en calculant le rapport de chaque effectif marginal à l'effectif total, selon la formulation générale suivante :

$$f_{i.} = \frac{n_{i.}}{n_{..}} = \sum_{j=1}^{p} f_{ij}$$

La dernière ligne du tableau correspond aux **fréquences marginales du caractère y**.

De même que pour les effectifs conjoints, si l'on somme les éléments de la première colonne du tableau, on obtient la fréquence marginale correspondant à la première modalité (A) du **caractère y**, c'est-à-dire la note de statistiques :

$$f_{.1} = f_{11} + f_{21} + f_{31} = \sum_{i=1}^{3} f_{i1}$$

Par convention : $f_{.j}$ représente l'intitulé de la dernière ligne.

Cependant, on peut obtenir les fréquences marginales du caractère y d'une autre façon, en utilisant les effectifs marginaux et en calculant le rapport de chaque effectif marginal à l'effectif total, selon la formulation générale suivante : $$f_{.j} = \frac{n_{.j}}{n_{..}} = \sum_{i=1}^{m} f_{ij}$$

Dans le **cas général**, on peut construire le tableau suivant, des **fréquences conjointes** :

| $x_i \setminus y_j$ | $y_1$ | ⋯ | $y_j$ | ... | $y_p$ | $f_{i.}$ |
|---|---|---|---|---|---|---|
| $x_1$ | $f_{11}$ | ⋯ | $f_{1j}$ | | $f_{1p}$ | $f_{1.}$ |
| ⋯ | ⋯ | ⋯ | ⋯ | ⋯ | ⋯ | ⋯ |
| $x_i$ | $f_{i1}$ | ⋯ | $f_{ij}$ | ⋯ | $f_{ip}$ | $f_{i.}$ |
| ... | ⋯ | ⋯ | ⋯ | ⋯ | ⋯ | ⋯ |
| $x_m$ | $f_{m1}$ | ⋯ | $f_{mj}$ | ⋯ | $f_{mp}$ | $f_{m.}$ |
| $f_{.j}$ | $f_{.1}$ | ⋯ | $f_{.j}$ | ⋯ | $f_{.p}$ | $f_{..}$ |

A l'intersection d'une ligne et d'une colonne sont reportées les fréquences conjointes. Par exemple, $f_{ij}$ correspond à la proportion des individus qui vérifient à la fois la modalité i du caractère x et la modalité j du caractère y.

La dernière ligne indique la somme des **fréquences marginales de y**. Par exemple, pour la $j^{ème}$ colonne, on a : $f_{.j} = \sum_{i=1}^{m} f_{ij} = \frac{n_{.j}}{n_{..}}$ . Par convention, on remarquera que la notation qui correspond à cette somme (terme général) est reprise comme intitulé de la ligne.

La dernière colonne indique la somme des **fréquences marginales de x**. Par exemple, pour la $i^{ème}$ ligne, on a : $f_{i.} = \sum_{j=1}^{p} f_{ij} = \frac{n_{i.}}{n_{..}}$ . Par convention, on remarquera que la notation qui correspond à cette somme (terme général) est reprise comme intitulé de la colonne.

La valeur $f_{..}$ située dans la case en bas et à droite du tableau représente la proportion totale de la population sur laquelle on travaille. Cette valeur est égale à 1 (ou à 100, si l'on raisonne en pourcentage) et elle s'obtient de trois façons différentes : - en sommant les fréquences marginales de x, dans la dernière colonne ; - en sommant les fréquences marginales de y, dans la dernière ligne ; - en sommant les fréquences conjointes en ligne et en colonne.

$$f_{..} = \sum_{i=1}^{m} f_{i.} = \sum_{j=1}^{p} f_{.j} = \sum_{i=1}^{m} \sum_{j=1}^{p} f_{ij} = 1 = 100\,\%$$

## 22. Fréquences conditionnelles

**Par définition**, les fréquences conditionnelles correspondent au rapport d'un effectif conjoint à un effectif marginal.

Il existe deux types de fréquences conditionnelles :

- les fréquences conditionnelles en colonnes, où fréquences conditionnelles du caractère x.

- les fréquences conditionnelles en lignes où fréquences conditionnelles du caractère y

### 221 Les fréquences conditionnelles en colonnes = fréquences conditionnelles du caractère x

La fréquence de la modalité $x_i$ du caractère x, sous condition que le caractère y prenne la modalité $y_j$, est égale au rapport de l'effectif conjoint $n_{ij}$ à l'effectif marginal $n_{.j}$ :

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

On parle de **fréquences conditionnelles de x**, à j fixé ou à modalité de y fixée, c'est-à-dire à colonne donnée.

Bien noter qu'il existe **p fréquences conditionnelles pour le caractère x** (il y a autant de fréquences conditionnelles de x qu'il y a de modalités de y).

Dans notre exemple, $f_{i/j}$ représente la proportion d'étudiants du groupe de td qui a obtenu la note $x_i$ en économie, parmi tous les étudiants ayant obtenu la note $y_j$ en statistiques. On peut dresser le tableau suivant :

| $y_j$ (note Stat) / $x_i$ (notes Eco) | A | B | C | $f_{i.}$ |
|---|---|---|---|---|
| **A** | 50,0 % | 25,0 % | 16,7 % | 33,3 % |
| **B** | **33,3 %** | 66,7 % | 0,0 % | 40,0 % |
| **C** | 16,7 % | 8,3 % | 83,3 % | **26,7 %** |
| **Total** | 100,0 % | 100,0 % | 100,0 % | 100,0 % |

Par exemple, 33,3 % représente la proportion d'étudiants ayant obtenu un B en économie, parmi tous ceux qui ont eu un A en statistiques.

De même, 26,7 % représente la proportion d'étudiants ayant eu un C en économie, quelle que soit leur note de statistiques (sur la marge, on raisonne sur le seul caractère x, l'économie, tout en ignorant le caractère y, les statistiques).

Dans le cas général, on a p fréquences conditionnelles de x :

| $x_i$ \ $y_j$ | $y_1$ | ⋯ | $y_j$ | ... | $y_p$ | $f_{i.}$ |
|---|---|---|---|---|---|---|
| $x_1$ | $f_{1/1}$ | ⋯ | $f_{1/j}$ | | $f_{1/p}$ | $f_{1.}$ |
| ... | ⋯ | ⋯ | ⋯ | ⋯ | ⋯ | ... |
| $x_i$ | $f_{i/1}$ | ⋯ | $f_{i/j}$ | ⋯ | $f_{i/p}$ | $f_{i.}$ |
| ... | ⋯ | ⋯ | ⋯ | ⋯ | ⋯ | ... |
| $x_m$ | $f_{m/1}$ | ⋯ | $f_{m/j}$ | ⋯ | $f_{m/p}$ | $f_{m.}$ |
| **Total** | **1** | **1** | **1** | **1** | **1** | **1** |

Dans la dernière ligne du tableau, le résultat est égal à 1 (ou 100, si l'on raisonne en pourcentage). Pour chacune des colonnes, on a : $\sum_{i=1}^{m} f_{i/j} = 1$ . On notera que la dernière colonne du tableau contient les fréquences marginales de x.

### 222 Les fréquences conditionnelles en lignes = fréquences conditionnelles du caractère y

La fréquence de la modalité $y_j$ du caractère y, sous condition que le caractère x prenne la modalité $x_i$, est égale au rapport de l'effectif conjoint $n_{ij}$ à l'effectif marginal $n_{i.}$ :

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

**Dans ce cas, il faut être attentif à la permutation des indices de ligne et de colonne, dans la notation et le calcul !!**

On parle de **fréquences conditionnelles de y**, à i fixé ou à modalité de x fixée, c'est-à-dire à ligne donnée.

Bien noter qu'il existe **m fréquences conditionnelles pour le caractère y** (il y a autant de fréquences conditionnelles de y qu'il y a de modalités de x).

Dans notre exemple, $f_{j/i}$ représente la proportion d'étudiants du groupe de td qui a obtenu la note $y_j$ en statistiques, parmi tous les étudiants ayant obtenu la note $x_i$ en économie. On peut dresser le tableau suivant :

| $y_j$ (note Stat) / $x_i$ (notes Eco) | A | B | C | Total |
|---|---|---|---|---|
| **A** | 60,0 % | 30,0 % | 10 % | 100,0 % |
| **B** | 33,3 % | 66,7 % | 0,0 % | 100,0 % |
| **C** | 25,0 % | **12,5 %** | 62,5 % | 100,0 % |
| $\mathbf{f_{.j}}$ | 40,0 % | 40,0 % | **20,0 %** | 100,0 % |

Par exemple, 12,5 % représente la proportion d'étudiants ayant obtenu un B en statistiques, parmi tous ceux qui ont eu un C en économie.

De même, 20,0 % représente la proportion d'étudiants ayant eu un C en statistiques, quelle que soit leur note d'économie (sur la marge, on raisonne sur le seul caractère y, les statistiques, tout en ignorant le caractère x, l'économie).

Dans le cas général, on a m fréquences conditionnelles de y :

| $x_i$ \ $y_j$ | $y_1$ | … | $y_j$ | ... | $y_p$ | Total |
|---|---|---|---|---|---|---|
| $x_1$ | $f_{1/1}$ | … | $f_{j/1}$ | | $f_{p/1}$ | 1 |
| … | … | … | … | … | … | 1 |
| $x_i$ | $f_{1/i}$ | … | $\mathbf{f_{j/i}}$ | … | $f_{p/i}$ | 1 |
| ... | … | … | … | … | … | 1 |
| $x_m$ | $f_{1/m}$ | … | $f_{j/m}$ | … | $f_{p/m}$ | 1 |
| $\mathbf{f_{.j}}$ | $\mathbf{f_{.1}}$ | … | $\mathbf{f_{.j}}$ | … | $\mathbf{f_{.p}}$ | 1 |

Dans la dernière colonne du tableau, le résultat est égal à 1 (ou 100, si l'on raisonne en pourcentage). Pour chacune des lignes, on a : $\sum_{j=1}^{p} f_{j/i} = 1$ . On notera que la dernière ligne du tableau contient toujours les fréquences marginales de y.

**Récapitulation**

**Sommes en lignes** : $n_{i.} = \sum_{j=1}^{p} n_{ij}$ avec : $n_{..} = \sum_{i=1}^{m} n_{i.}$

**Sommes en colonnes** : $n_{.j} = \sum_{i=1}^{m} n_{ij}$ avec : $n_{..} = \sum_{j=1}^{p} n_{.j}$

**Fréquences conjointes** : $f_{ij} = \frac{n_{ij}}{n_{..}}$

**Fréquences marginales de x** : $f_{i.} = \frac{n_{i.}}{n_{..}}$ et : $f_{..} = \sum_{i=1}^{m} f_{i.} = 1$

**Fréquences marginales de y** : $f_{.j} = \frac{n_{.j}}{n_{..}}$ et : $f_{..} = \sum_{j=1}^{p} f_{.j} = 1$

**m fréquences conditionnelles de x (= fréquences conditionnelles en colonnes, à j fixé)** :

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

**p fréquences conditionnelles de y (= fréquences conditionnelles en lignes, à i fixé)** :

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

**Attention à la permutation des indices !!!**

## 23. Relations entre fréquences conjointes, marginales et conditionnelles

Fréquences marginales

De x : $f_{i.} = \frac{n_{i.}}{n_{..}}$ **(1)**

De y : $f_{.j} = \frac{n_{.j}}{n_{..}}$ **(2)**

Fréquences conditionnelles

De y (en lignes) : $f_{j/i} = \frac{n_{ij}}{n_{i.}}$ **(3)**

De x (en colonnes) : $f_{i/j} = \frac{n_{ij}}{n_{.j}}$ **(4)**

A. Multiplions (1) et (3) : $$f_{i.} \times f_{j/i} = \frac{n_{i.}}{n_{..}} \times \frac{n_{ij}}{n_{i.}} = \frac{n_{ij}}{n_{..}} = f_{ij}$$

On a donc finalement : $$f_{ij} = f_{i.} \times f_{j/i}$$

c-à-d que la fréquence conjointe $\mathbf{f_{ij}}$ est égale au produit de la fréquence marginale de x et de la fréquence conditionnelle de y.

B. Multiplions (2) et (4) : $$f_{.j} \times f_{i/j} = \frac{n_{.j}}{n_{..}} \times \frac{n_{ij}}{n_{.j}} = \frac{n_{ij}}{n_{..}} = f_{ij}$$

De même : $$f_{ij} = f_{.j} \times f_{i/j}$$

c-à-d que la fréquence conjointe $\mathbf{f_{ij}}$ est égale au produit de la fréquence marginale de y et de la fréquence conditionnelle de x.

Par conséquent :

$$f_{ij} = f_{i.} \times f_{j/i} = f_{.j} \times f_{i/j}$$

**Application à l'exemple :**

Caractère x : évaluation en économie (A, B, C).

Caractère y : évaluation en statistiques (A, B, C).

n .. = 30 étudiants.

| $y_j$ (note Stat) / $x_i$ (notes Eco) | A | B | C | $n_{i.}$ |
|---|---|---|---|---|
| **A** | 6 | 3 | 1 | 10 |
| **B** | 4 | 8 | 0 | 12 |
| **C** | 2 | 1 | 5 | 8 |
| $n_{.j}$ | 12 | 12 | 6 | 30 |

**Tableau I Effectifs conjoints et marginaux**

| $y_j$ (note Stat) / $x_i$ (notes Eco) | A | B | C | $f_{i.}$ |
|---|---|---|---|---|
| **A** | 20,0 % | 10,0 % | **3,3 %** | **33,3 %** |
| **B** | 13,3 % | 26,7 % | 0,0 % | 40,0 % |
| **C** | 6,7 % | 3,3 % | 16,7 % | 26,7 % |
| $f_{.j}$ | 40,0 % | 40,0 % | **20,0 %** | 100,0 % |

**Tableau II Fréquences conjointes et marginales**

| $y_j$ (note Stat) / $x_i$ (notes Eco) | A | B | C | $f_{i.}$ |
|---|---|---|---|---|
| **A** | 50,0 % | 25,0 % | **16,7 %** | 33,3 % |
| **B** | 33,3 % | 66,7 % | 0,0 % | 40,0 % |
| **C** | 16,7 % | 8,3 % | 83,3 % | 26,7 % |
| **Total** | 100,0 % | 100,0 % | 100,0 % | 100,0 % |

**Tableau III Fréquences conditionnelles de x (en colonnes)**

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

| $y_j$ (note Stat) / $x_i$ (notes Eco) | A | B | C | Total |
|---|---|---|---|---|
| **A** | 60,0 % | 30,0 % | **10 %** | 100,0 % |
| **B** | 33,3 % | 66,7 % | 0,0 % | 100,0 % |
| **C** | 25,0 % | 12,5 % | 62,5 % | 100,0 % |
| $f_{.j}$ | 40,0 % | 40,0 % | 20,0 % | 100,0 % |

**Tableau IV Fréquences conditionnelles de y (en lignes)**

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

Par exemple, vérifions : $f_{ij} = f_{i.} \times f_{j/i} = f_{.j} \times f_{i/j}$

pour : **i = 1** et **: j = 3**

On cherche donc : $f_{13} = f_{1.} \times f_{3/1} = f_{.3} \times f_{1/3}$

D'où : 0,033 = 0,333 x 0,10 = 0,20 x 0,167

Tableau : (II) (II) (IV) (II) (III)

## CHAPITRE 2 : MISE EN RELATION DE DEUX CARACTÈRES QUALITATIFS

Les modalités prises par chacun des 2 caractères x et y ne sont pas mesurables.

Objectifs :

1) quels types de représentation graphique sont-ils adaptés pour rendre compte, sans déformation des données, d'une telle série statistique ?

2) quelles méthodes peut-on mettre en œuvre pour déterminer si les deux caractères sont ou non indépendants ?

On cherche en effet à savoir si x influe sur y, ou bien si y influe sur x, ou enfin si x et y sont deux caractères indépendants.

Remarque importante : dans certains cas, alors qu'une procédure (ou un calcul statistique) est toujours possible à mettre en œuvre, la relation entre deux caractères peut n'avoir aucun sens économiquement (en termes de causalité).

### 1. Les représentations graphiques appropriées

Comme dans le cas de séries à un caractère qualitatif, on peut utiliser des diagrammes en tuyaux d'orgue et des diagrammes en secteurs. Il est cependant nécessaire ici de les adapter, afin que les représentations graphiques ne faussent pas la réalité des données.

#### 11. Les tuyaux d'orgue à base variable

Exemple

On considère la population des 1 500 salariés d'une entreprise, selon deux caractères :

- caractère **x** : la CSP des salariés. Caractère qualitatif.

- caractère **y** : le sexe des salariés. Caractère qualitatif.

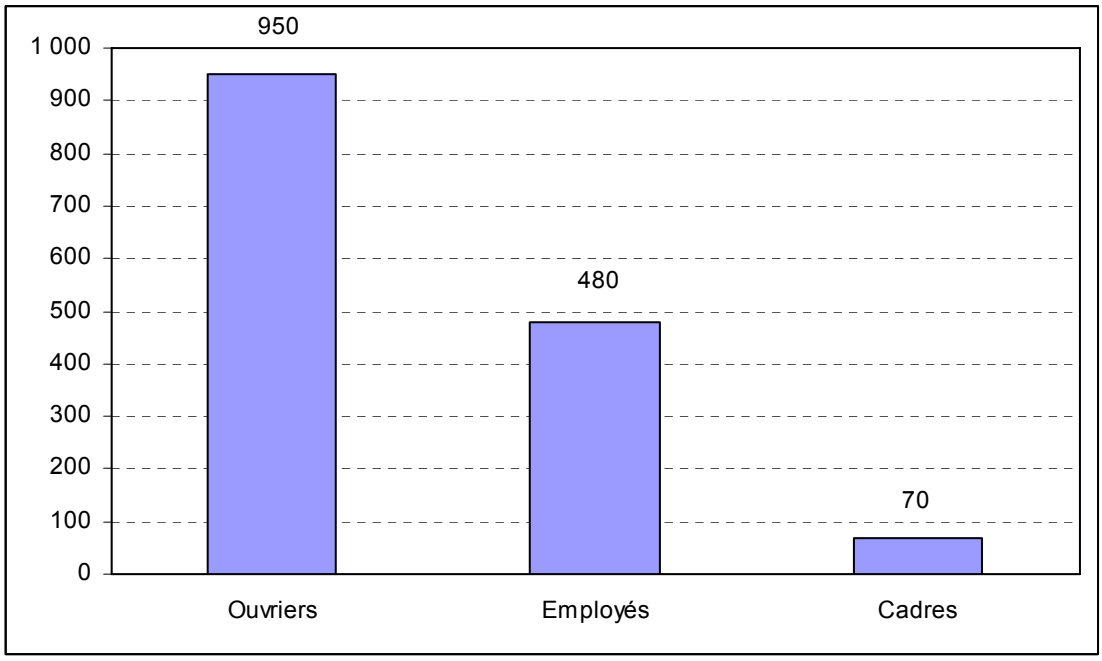
**Tableau des effectifs conjoints**

| Sexe (yj) \ CSP (xi) | H | F | n i. |
|---|---|---|---|
| **Ouvriers** | 800 | 150 | 950 |
| **Employés** | 150 | 330 | 480 |
| **Cadres** | 50 | 20 | 70 |
| **n .j** | 1 000 | 500 | 1 500 |

Remarque :

si l'on raisonnait sur le seul caractère CSP (c'est-à-dire en raisonnant sur la dernière colonne du tableau), on aurait :

| CSP (xi) | Effectifs ni |
|---|---|
| **Ouvriers** | 950 |
| **Employés** | 480 |
| **Cadres** | 70 |
| | 1 500 |




Lorsqu'on prend en compte un deuxième caractère, deux problèmes doivent être résolus correctement, de façon à ce que le graphique ne déforme pas la réalité de l'information donnée dans le tableau initial :

1) pour tenir compte de la présence simultanée du deuxième caractère (ici, le sexe), **on partage chaque tuyau**, de façon à rendre compte du poids (pondération) chacune des deux modalités (H et F) du caractère Sexe.

2) en ce qui concerne l'autre caractère (les CSP), il faut également pondérer chacune de ses modalités (employés, ouvriers, cadres) selon son importance (pondération).

**D'où la notion de tuyau d'orgue à base variable, pour prendre en compte cette pondération**.

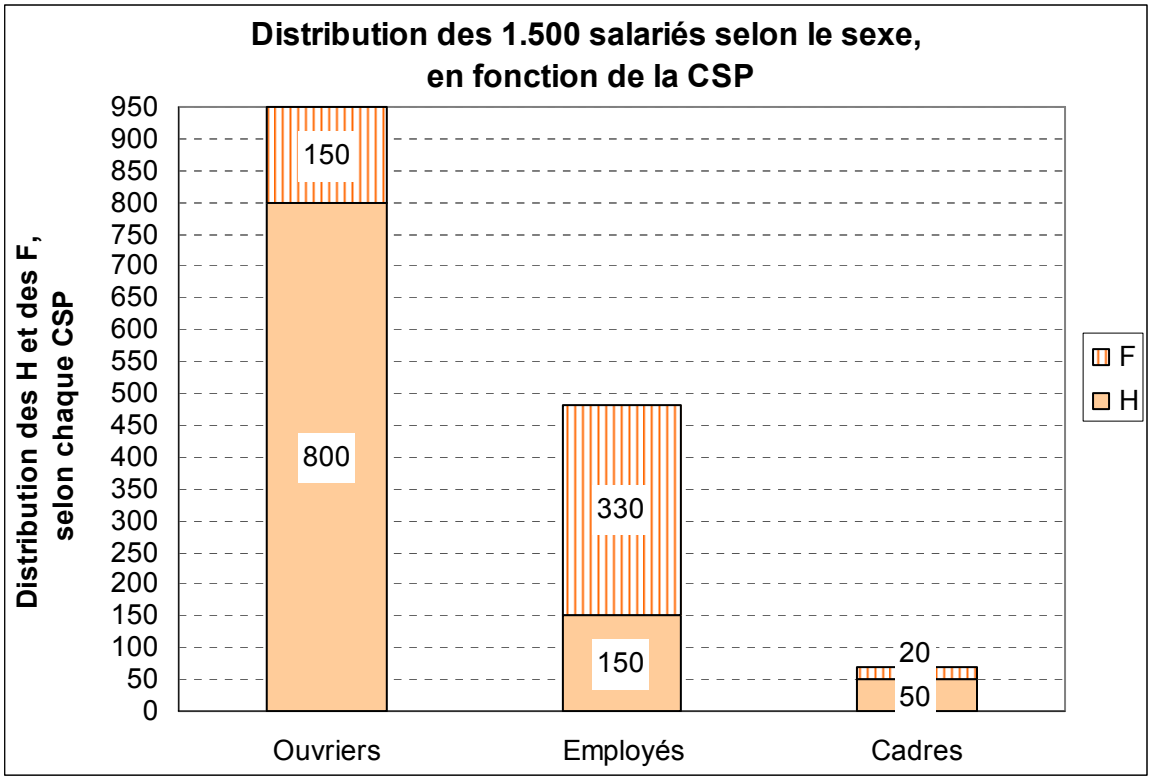
Par ailleurs, il est impératif de raisonner en fréquences (ou en %), de manière à pouvoir comparer correctement la distribution par sexe de chaque CSP.

De fait, on peut raisonner sur les **effectifs** absolus. Dans ce cas, on construit des tuyaux (un par modalité du 1er caractère) qui présentent tous la même amplitude de base, chacun d'entre eux étant segmenté selon le nombre de modalités du 2ème caractère. Cette représentation graphique rend correctement compte des données brutes, mais elle présente deux défauts majeurs :

a) lorsque les effectifs de chaque modalité du 1er caractère sont très différents les uns des autres et si, dans le même temps, le nombre de modalités du 2ème caractère est assez important, ce type de représentation devient rapidement illisible (voir plus loin).

b) lorsqu'on cherche à comparer plusieurs populations (ou échantillons), le raisonnement sur des effectifs (qui peuvent être très différents) n'est pas très efficace.

Nous allons cependant tracer le graphique correspondant (tuyaux d'orgue ordinaires, c'est-à-dire ayant une base de même amplitude), pour des raisons pédagogiques :




Compte tenu des raisons précisées plus haut, on raisonne habituellement sur les **fréquences** (ou les pourcentages). Mais, dans ce cas, pour que le graphique rende correctement compte de la réalité des données, il est nécessaire d'adapter les tuyaux d'orgue :

1) chaque tuyau d'orgue représente une modalité de l'un des deux caractères. Le poids de chacune des modalités de ce caractère est traduit graphiquement par une **base d'amplitude variable** pour chacun des tuyaux (en abscisse).

2) il faut partager chaque tuyau, selon les différentes modalités de l'autre caractère étudié, de telle façon que le découpage réalisé rende compte du poids de chacune des modalités de ce caractère (en ordonnée).

En présence de caractères qualitatifs nominaux, il est préférable de ranger les tuyaux, en abscisse, selon l'importance des fréquences (comme pour les séries à un caractère).

**Première représentation : distribution par sexe des salariés selon les CSP**

Pour cela, on considère le tableau de données en lignes et l'on crée le tableau suivant :

**tableau des fréquences conditionnelles en lignes (en %)**

| Sexe (yj) \ CSP (xi) | H | F | Total |
|---|---|---|---|
| **Ouvriers** | 84,2 | 15,8 | 100 |
| **Employés** | 31,3 | 68,8 | 100 |
| **Cadres** | 71,4 | 28,6 | 100 |
| **f .j** | 66,7 | 33,3 | 100 |

Il s'agit des fréquences conditionnelles de y, c'est-à-dire à CSP donnée : $f_{j/i} = \frac{n_{ij}}{n_{i.}}$

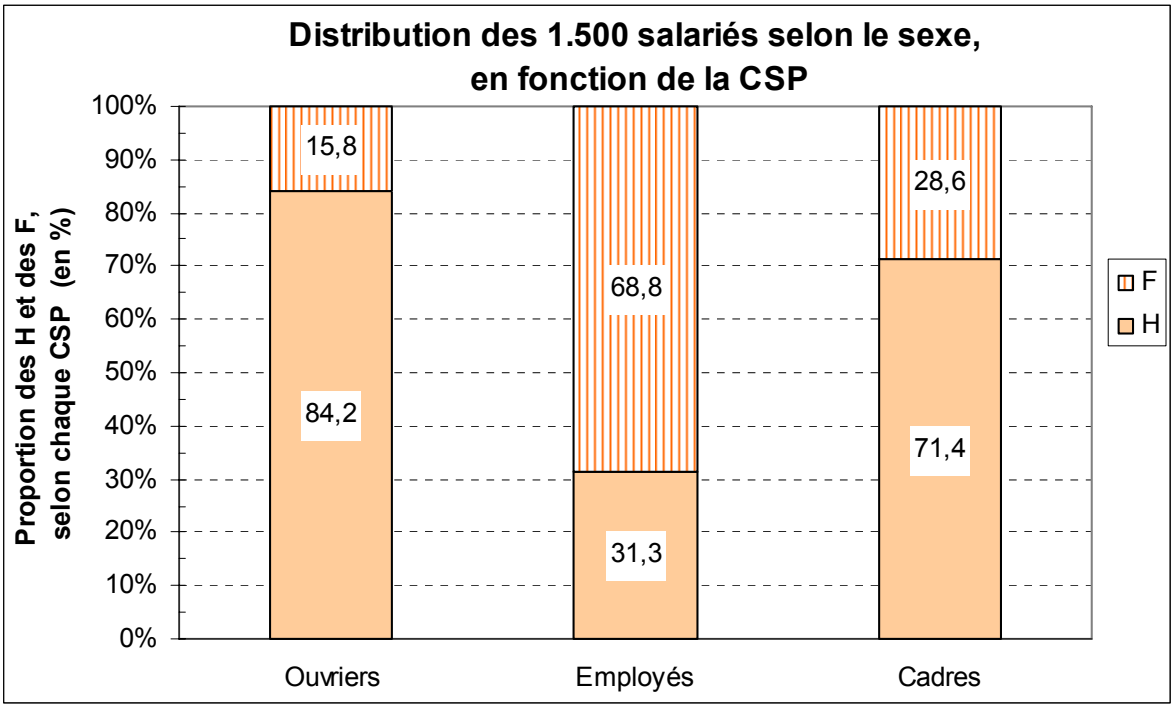
**D'une part, chaque tuyau est partagé en deux, selon les pourcentages calculés dans le tableau précédent**.

**D'autre part, la base de chaque tuyau doit être proportionnelle à l'effectif (ou à la fréquence) de chaque CSP**.

En vue d'obtenir le poids de chacune des CSP, dressons un autre tableau qui nous resservira plus loin :

**tableau des fréquences conjointes et des fréquences marginales (en %)**

| Sexe (yj) \ CSP (xi) | H | F | f i. |
|---|---|---|---|
| **Ouvriers** | 53,3 | 10,0 | **63,3** |
| **Employés** | 10,0 | 22,0 | **32,0** |
| **Cadres** | 3,3 | 1,3 | **4,7** |
| **f .j** | 66,7 | 33,3 | 100,0 |




**<u>Attention</u> : les tuyaux d'orgue ci-dessus (réalisés avec un tableur) ne sont pas corrects en abscisse, car leurs bases ne sont pas variables, en fonction du poids de chaque CSP (<u>63,3%, 32,0%, 4,7%</u>).**

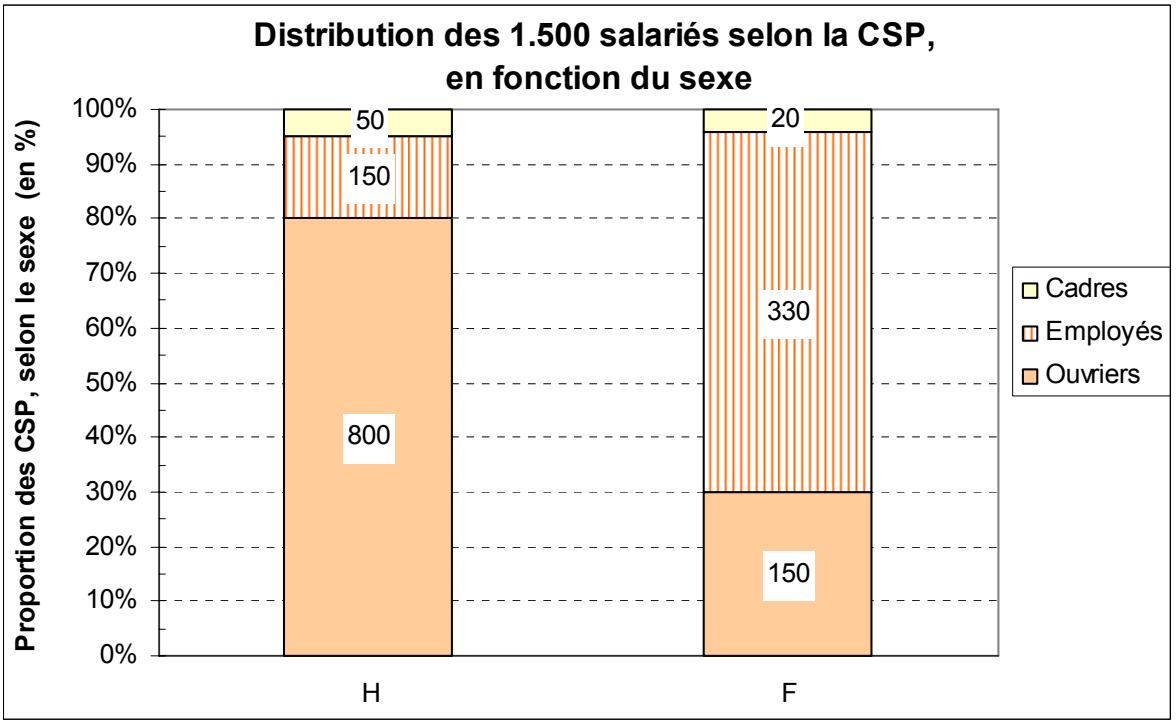
En principe, on classe les tuyaux selon l'importance de leurs effectifs.

**<u>Deuxième représentation : distribution par CSP des salariés selon le sexe</u>**

On peut évidemment envisager le même type de représentation, relativement au sexe, selon les CSP. Dans ce cas, on a besoin des <u>fréquences conditionnelles de x, à colonne fixée, c'est-à-dire à sexe donné</u> : $f_{i/j} = \frac{n_{ij}}{n_{.j}}$

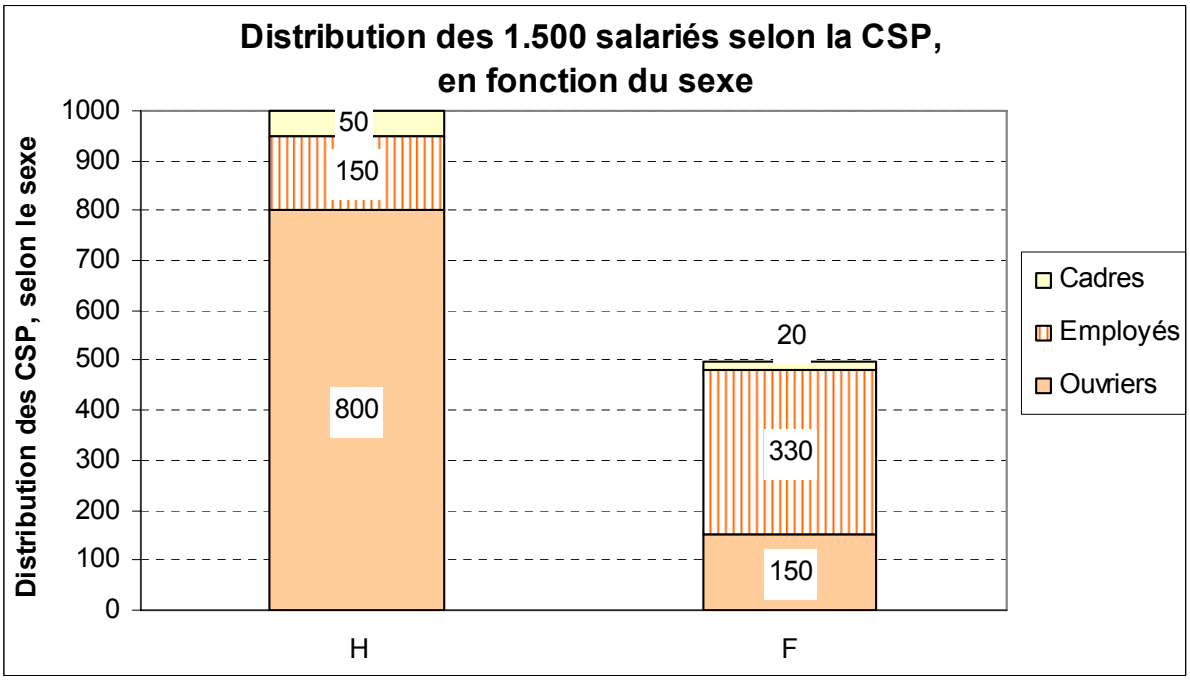
**<u>Tableau des fréquences conditionnelles en colonnes</u> (en %)**

| Sexe (yj) \ CSP (xi) | H | F | f i. |
|---|---|---|---|
| **Ouvriers** | 80,0 | 30,0 | 63,3 |
| **Employés** | 15,0 | 66,0 | 32,0 |
| **Cadres** | 5,0 | 4,0 | 4,7 |
| **Total** | 100,0 | 100,0 | 100,0 |




**<u>Attention</u> : les tuyaux d'orgue ci-dessus (réalisés avec un tableur) ne sont pas corrects en abscisse, car leurs bases ne sont pas variables, en fonction des poids de chaque sexe (<u>66,7%, 33,3%</u>).**

Comme on l'a fait précédemment, on peut tôt aussi raisonner <u>selon les effectifs</u>, en créant des <u>rectangles de même amplitude de base</u> :



### 12. Les secteurs semi-circulaires

Cette autre représentation graphique possible est **bien adaptée lorsque l'un des deux caractères ne présente que deux modalités** (cas de notre deuxième représentation).

Pour construire des secteurs semi-circulaires, deux principes sont à respecter :

1) le poids de chacune des deux modalités du caractère dichotomique est traduit graphiquement par deux demi-cercles, dont la surface est proportionnelle aux effectifs (ou aux fréquences) correspondants.

2) le poids des modalités de l'autre caractère est traduit graphiquement, à l'intérieur de chaque demi-cercle, par des secteurs proportionnels aux effectifs (ou aux fréquences) correspondants. Pour cela, on calcule, à partir des effectifs (ou des fréquences), les angles de chacun des secteurs des demi-cercles.

Considérons à nouveau le tableau initial :

| Sexe (yj) \ CSP (xi) | H | F | n i. |
|---|---|---|---|
| **Ouvriers** | 800 | 150 | 950 |
| **Employés** | 150 | 330 | 480 |
| **Cadres** | 50 | 20 | 70 |
| **n .j** | 1 000 | 500 | 1 500 |

On remarque que **l'effectif H (1 000) est égal au double de l'effectif F (500)**. Selon les fréquences marginales de y, on a respectivement 66,7 % et 33,3 %.

Lorsqu'on place en regard deux secteurs semi-circulaires, on doit tenir compte de cette différence. En effet dans ce type de représentation, c'est **la surface qui est proportionnelle aux effectifs**.

On doit respecter la proportion suivante :

**surface du demi-cercle H = 2 x surface du demi-cercle F** $\Leftrightarrow$ $S_H = 2 \times S_F$

avec : **2 = 1 000 / 500**

Posons :

**$R_H$ = rayon du demi-cercle hommes et : $R_F$ = rayon du demi-cercle femmes**

Par ailleurs, on sait que la surface d'un cercle est égale à : $S = \pi R^2$

Donc, celle d'un demi-cercle est égale à : $S = (\pi R^2) / 2$

Et l'on a : $S_H = (\pi R_H^2) / 2$ De même : $S_F = (\pi R_F^2) / 2$

Comme on a : $S_H = 2 \times S_F \Leftrightarrow (\pi R_H^2) / 2 = 2 (\pi R_F^2) / 2$

Après simplification, il vient : $R_H^2 = 2\ R_F^2$

et : $R_H = \sqrt{2} \cdot R_F$ ou bien : $R_F = R_H / \sqrt{2}$

Cela signifie que si l'on choisit arbitrairement, par ex. :

$R_H = 10$ cm, alors : $R_F = 10 / \sqrt{2} = 10 / 1{,}414 \sim 7{,}1$ cm

Ainsi, les surfaces respectives de chaque demi-cercle sont bien proportionnelles à chacun des deux effectifs (H et F).

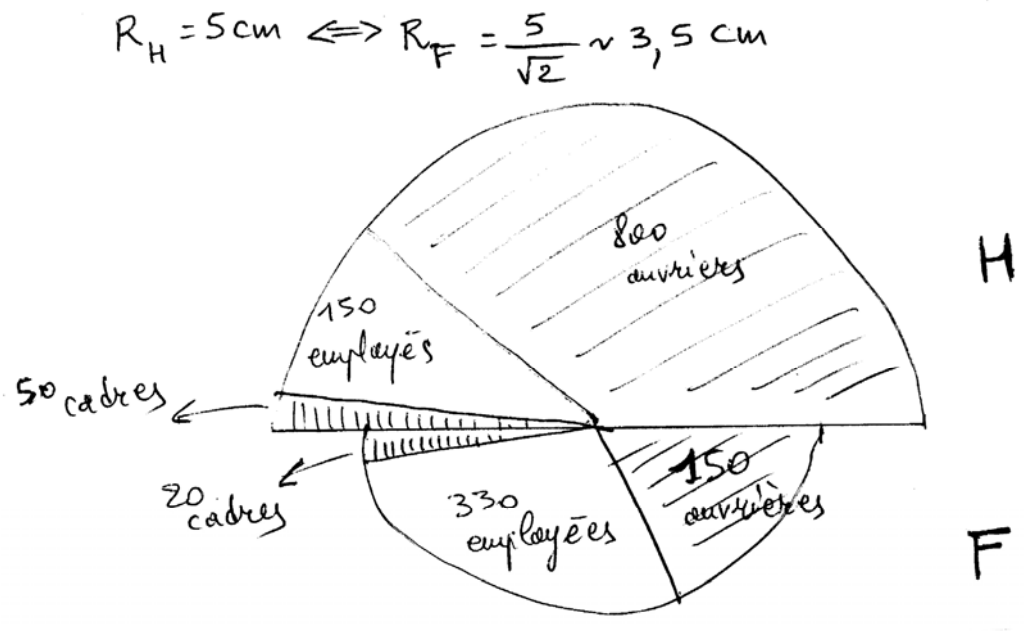
Dans un deuxième temps, il faut **déterminer les secteurs angulaires relatifs à chaque CSP, pour chacun des deux demi-cercles** (H et F).

Comme on raisonne sur un demi-cercle, on calcule les angles sur 180°.

Reprenons le tableau des fréquences conditionnelles en colonnes, présenté plus haut :

| **Sexe (yj) \ CSP (xi)** | **H (en %)** | **F (en %)** | **f i.** | **H (en d°)** | **F (en d°)** |
|---|---|---|---|---|---|
| **Ouvriers** | 80 | 30 | 63,3 | **144** | **54** |
| **Employés** | 15 | 66 | 32,0 | **27** | **119** |
| **Cadres** | 5 | 4 | 4,7 | **9** | **7** |
| **Total** | 100 | 100 | 100,0 | 180 | 180 |

À partir des pourcentages des deux premières colonnes du tableau, il suffit de multiplier les valeurs correspondants par 1,8 pour obtenir les valeurs des angles (en degrés), valeurs qui sont arrondies au degré le plus proche.



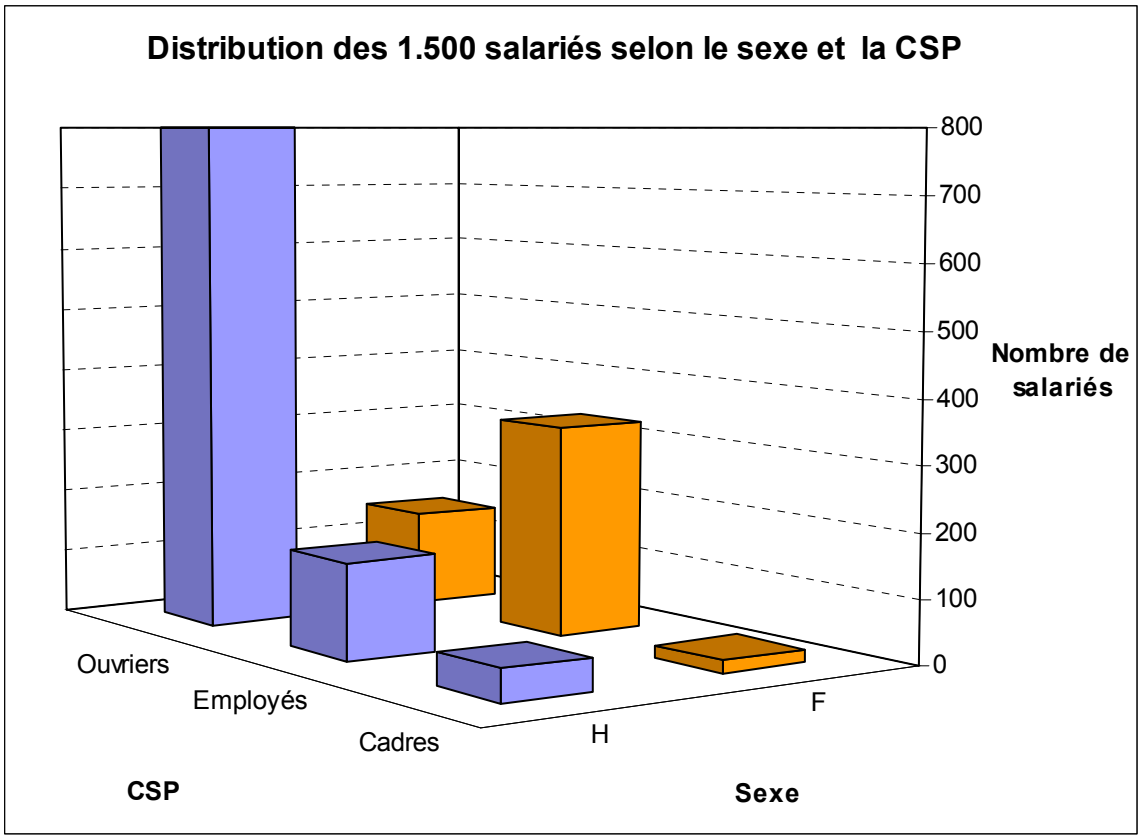
Remarques terminales

1) Lorsque le caractère possède plus de deux modalités, le diagramme semi-circulaire n'est, en général, pas très appropriée, car la construction devient plus complexe.

C'est pourquoi, ici, on ne représentera pas la série selon les CSP, sous cette forme, car cela nécessiterait trois fractions de cercle (de 120° chacun), proportionnelles à l'effectif total par CSP, chacune de ces fractions étant elle-même partagée en deux, selon la proportion H ou F.

2) Enfin, si l'on utilise un tableur, on peut également réaliser des **stéréogrammes**, qui permettent une bonne visualisation si le nombre de modalités de chacun des caractères n'est pas trop élevé.

Ici, par exemple, on obtient :



## 2. La notion d'indépendance entre deux caractères

On conserve les notations et conventions du tableau de base (partie 2, ch. 1, point 1), notamment le fait que les modalités du caractère x sont placées en lignes et celles de y en colonnes.

### 21. Principe général

Revenons sur le tableau de contingence général (m lignes et p colonnes), vu au chapitre 1 :

| $x_i$ \ $y_j$ | $y_1$ | … | $y_j$ | … | $y_p$ | $n_{i.}$ |
|---|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | … | $n_{1j}$ | | $n_{1p}$ | $n_{1.}$ |
| … | … | … | … | … | … | … |
| $x_i$ | $n_{i1}$ | … | $n_{ij}$ | … | $n_{ip}$ | $n_{i.}$ |
| … | … | … | … | … | … | … |
| $x_m$ | $n_{m1}$ | … | $n_{mj}$ | … | $n_{mp}$ | $n_{m.}$ |
| $n_{.j}$ | $n_{.1}$ | … | $n_{.j}$ | … | $n_{.p}$ | $n_{..}$ |

De même :

**Fréquences conjointes** : $f_{ij} = \frac{n_{ij}}{n_{..}}$

**Fréquences marginales de x** : $f_{i.} = \frac{n_{i.}}{n_{..}}$ et : $f_{..} = \sum_{i=1}^{m} f_{i.} = 1$

**Fréquences marginales de y** : $f_{.j} = \frac{n_{.j}}{n_{..}}$ et : $f_{..} = \sum_{j=1}^{p} f_{.j} = 1$

**m fréquences conditionnelles de x (= fréquences conditionnelles en colonnes, à j fixé)** :

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

**p fréquences conditionnelles de y (= fréquences conditionnelles en lignes, à i fixé)** :

$$f_{j/i} = \frac{n_{ij}}{n_{i.}}$$ **<u>Attention à la permutation des indices !!!</u>**

$$f_{ij} = f_{i.} \times f_{j/i} = f_{.j} \times f_{i/j}$$

Dans le tableau précédent, supposons que l'on prenne en compte les seules colonnes suivantes :

| $x_i$ | $n_{i.}$ | $f_{i.}$ |
|---|---|---|
| $x_1$ | $n_{1.}$ | $f_{1.}$ |
| … | … | … |
| $x_i$ | $n_{i.}$ | $f_{i.}$ |
| ... | … | … |
| $x_m$ | $n_{m.}$ | $f_{m.}$ |
| $n_{.j}$ | $n_{..}$ | $f_{..}$ |

On conserve les modalités de x, les effectifs marginaux de x et les fréquences marginales de x.

Tout se passe comme si l'on avait plié le tableau croisé dans le sens de la largeur.

Entre ce nouveau tableau et le tableau initial, les modalités du caractère y ont disparu.

**Quelle que soit la colonne (donc la modalité du caractère y), le caractère x ne dépend pas du caractère y.**

Si l'on considère les **fréquences conditionnelles de x** (c'est-à-dire les fréquences conditionnelles en colonnes, à j fixé ou à colonne fixée) dans le nouveau tableau, elles ne dépendent plus de j (c'est-à-dire d'une quelconque modalité de y).

On peut donc dire que x ne dépend pas de y. Il existe alors une indépendance totale entre x et y.

**Pour cette raison**, on est autorisé à écrire : $f_{i/j} = f_{i.}$

C'est-à-dire que **les fréquences conditionnelles de x (à j fixé ou à colonne fixée) sont égales aux fréquences marginales de x, lorsque les deux caractères x et y sont indépendants**.

Exemple :

| $x_i$ \ $y_j$ | $y_1$ | $y_2$ | $n_{i.}$ |
|---|---|---|---|
| $x_1$ | 3 | 5 | 8 |
| $x_2$ | 6 | 10 | 16 |
| $n_{.j}$ | 9 | 15 | 24 |

| $x_i$ \ $y_j$ | $y_1$ | $y_2$ | $f_{i.}$ |
|---|---|---|---|
| $x_1$ | 33,3 | 33,3 | 33,3 |
| $x_2$ | 66,7 | 66,7 | 66,7 |
| | 100,0 | 100,0 | 100,0 |

Le tableau de gauche est un tableau de contingence des effectifs conjoints, dont les colonnes sont proportionnelles. Le tableau de droite donne, en %, les valeurs des fréquences conditionnelles $f_{i/j}$ de x (en colonnes, à j fixé) et les fréquences marginales de x.

**Lorsque x est indépendant de y, toutes les colonnes du tableau sont identiques**.

On a : **fréquences marginales de x** : $f_{i.} = \frac{n_{i.}}{n_{..}}$

Ici : $f_{1.} = \frac{n_{1.}}{n_{..}} = \frac{8}{24} = \frac{1}{3} = 0{,}33$ et : $f_{2.} = \frac{n_{2.}}{n_{..}} = \frac{16}{24} = \frac{2}{3} = 0{,}67$

Par ailleurs : **fréquences conditionnelles de x (à j fixé)** : $f_{i/j} = \frac{n_{ij}}{n_{.j}}$

Ici : $f_{1/1} = \frac{n_{11}}{n_{.1}} = \frac{3}{9} = \frac{1}{3} = 0{,}33$ et : $f_{1/2} = \frac{n_{12}}{n_{.2}} = \frac{5}{15} = \frac{1}{3} = 0{,}33$

Donc : $f_{1.} = f_{1/1} = f_{1/2}$

De même :

$f_{2/1} = \frac{n_{21}}{n_{.1}} = \frac{6}{9} = \frac{2}{3} = 0{,}67$ et : $f_{2/2} = \frac{n_{22}}{n_{.2}} = \frac{10}{15} = \frac{2}{3} = 0{,}67$

Donc : $f_{2.} = f_{2/1} = f_{2/2}$

**Conséquence** :

Lorsque x est indépendant de y, les colonnes du tableau sont proportionnelles et l'on vérifie que :

$$f_{i/j} = f_{i.} \quad (1)$$

Par ailleurs (chapitre 1, point 23), on a : $f_{ij} = f_{.j} \times f_{i/j}$ (2)

Lorsque x est indépendant de y, on peut remplacer $f_{i/j}$ dans (2) par $f_{i.}$ dans (1).

Il en résulte que **la fréquence conjointe (d'une case du tableau de contingence) est égale au produit des fréquences marginales de la ligne et de la colonne correspondantes** :

$$f^{*}_{ij} = f_{i.} \times f_{.j} \quad (A)$$

Si l'on raisonne en pourcentage, il faut diviser par 100 le produit des fréquences marginales.

Si l'on raisonne selon les **effectifs**, on remplace dans (A) les fréquences par les rapports d'effectifs correspondants :

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

et : $$\frac{n_{ij}}{n_{..}} = \frac{n_{i.}}{n_{..}} \times \frac{n_{.j}}{n_{..}} \quad \Leftrightarrow \quad n^{*}_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$$

C'est-à-dire que lorsque x est indépendant de y, **l'effectif conjoint (d'une case du tableau de contingence) est égale au produit des effectifs marginaux de la ligne et de la colonne correspondantes, divisé par l'effectif total**.

**En conclusion**, à partir des tableaux de contingence des effectifs réels (ou des fréquences réelles), on va pouvoir construire, selon les formules ci-dessus, des **tableaux d'effectifs théoriques (ou de fréquences théoriques) d'indépendance**.

On obtiendra ainsi des tableaux d'indépendance à effectifs théoriques ou à fréquences théoriques, fabriqués de telle façon que le caractère x sera indépendant du caractère y (l'évolution de x ne sera pas influencée par les valeurs prises par le caractère y).

Modalités pratiques

Pour créer à coup sûr un tableau tel que les caractères x et y soient indépendants, il faut tout d'abord remplir les marges du tableau. Puis on calcule les effectifs conjoints (ou les fréquences conjointes), situés à l'intersection d'une ligne et d'une colonne.

On utilise les expressions : $n^{*}_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$ ou : $f^{*}_{ij} = f_{i.} \times f_{.j}$ selon le cas.

Exemple :

| $x_i$ \ $y_j$ | $y_1$ | $y_2$ | $n_{i.}$ |
|---|---|---|---|
| $x_1$ | | | **75** |
| $x_2$ | | | **375** |
| $n_{.j}$ | **150** | **300** | **450** |

| $x_i$ \ $y_j$ | $y_1$ | $y_2$ | $f_{i.}$ |
|---|---|---|---|
| $x_1$ | **25** | **50** | 75 |
| $x_2$ | **125** | **250** | 375 |
| $f_{.j}$ | 150 | 300 | 450 |

On commence par entrer les valeurs (ici arbitraires) qui correspondent aux marges (tableau de gauche). Dans le tableau de droite, on entre les valeurs conjointes des effectifs, selon la 1ère formule ci-dessus. Par ex. : $n^{*}_{11} = \frac{n_{1.} \times n_{.1}}{n_{..}} = \frac{75 \times 150}{450} = 25$

On constate que **les lignes et les colonnes du tableau obtenu sont proportionnelles entre elles**.

Dans un premier temps, nous avons raisonné sur le caractère x, en colonnes. On peut évidemment raisonner sur le caractère y, en lignes. Dans ce cas, on aurait :

| $y_j$ | $y_1$ | ⋯ | $y_j$ | ... | $y_p$ | $n_{i.}$ |
|---|---|---|---|---|---|---|
| $n_{.j}$ | $n_{.1}$ | ⋯ | $n_{.j}$ | ⋯ | $n_{.p}$ | $n_{..}$ |
| $f_{.j}$ | $f_{.1}$ | ⋯ | $f_{.j}$ | ⋯ | $f_{.p}$ | $f_{..}$ |

On conserve les modalités de y, les effectifs marginaux de y et les fréquences marginales de y.

Tout se passe comme si l'on avait plié le tableau croisé dans le sens de la hauteur.

Entre ce nouveau tableau et le tableau initial, les modalités du caractère x ont disparu.

**Quelle que soit la ligne est (donc la modalité du caractère x), le caractère y ne dépend pas du caractère x.**

Si l'on considère les **fréquences conditionnelles de y** (c'est-à-dire les fréquences conditionnelles en lignes, à i fixé ou à ligne fixée) dans le nouveau tableau, elles ne dépendent plus de i (c'est-à-dire d'une quelconque modalité de x).

On peut donc dire que y ne dépend pas de x. Il existe alors une indépendance totale entre x et y.

**Pour cette raison**, on est autorisé à écrire : $f_{j/i} = f_{.j}$

C'est-à-dire que **les fréquences conditionnelles de y (à i fixé ou à ligne fixée) sont égales aux fréquences marginales de y, lorsque les deux caractères x et y sont indépendants**.

Reprenons notre exemple :

| $x_i$ \ $y_j$ | $y_1$ | $y_2$ | $n_{i.}$ |
|---|---|---|---|
| $x_1$ | 3 | 5 | 8 |
| $x_2$ | 6 | 10 | 16 |
| $n_{.j}$ | 9 | 15 | 24 |

| $x_i$ \ $y_j$ | $y_1$ | $y_2$ | |
|---|---|---|---|
| $x_1$ | 37,5 | 62,5 | 100,0 |
| $x_2$ | 37,5 | 62,5 | 100,0 |
| $f_{.j}$ | 37,5 | 62,5 | 100,0 |

Le tableau de gauche est un tableau de contingence des effectifs conjoints, dont les lignes sont proportionnelles. Le tableau de droite donne, en %, les valeurs des fréquences conditionnelles $f_{j/i}$ de y (en lignes, à i fixé) et les fréquences marginales de y.

**Lorsque y est indépendant de x, toutes les lignes du tableau sont identiques**.

On a : **fréquences marginales de y** : $f_{.j} = \frac{n_{.j}}{n_{..}}$

Ici : $f_{.1} = \frac{n_{.1}}{n_{..}} = \frac{9}{24} = \frac{3}{8} = 0{,}375$ et : $f_{.2} = \frac{n_{.2}}{n_{..}} = \frac{15}{24} = \frac{5}{8} = 0{,}625$

Par ailleurs : **fréquences conditionnelles de y (à i fixé)** : $f_{j/i} = \frac{n_{ij}}{n_{i.}}$

Ici : $f_{1/1} = \frac{n_{11}}{n_{1.}} = \frac{3}{8} = 0{,}375$ et : $f_{1/2} = \frac{n_{21}}{n_{2.}} = \frac{6}{16} = 0{,}375$

Donc : $f_{.1} = f_{1/1} = f_{1/2}$

De même :

$f_{2/1} = \frac{n_{12}}{n_{1.}} = \frac{5}{8} = 0{,}625$ et : $f_{2/2} = \frac{n_{22}}{n_{2.}} = \frac{10}{16} = 0{,}625$

Donc : $f_{.2} = f_{2/1} = f_{2/2}$

**Conséquence** :

Lorsque y est indépendant de x, les lignes du tableau sont proportionnelles et l'on vérifie que :

$$f_{j/i} = f_{.j} \quad (1)$$

Par ailleurs (chapitre 1, point 23), on a : $f_{ij} = f_{i.} \times f_{j/i} \quad (2)$

Lorsque y est indépendant de x, on peut remplacer $f_{j/i}$ dans (2) par $f_{.j}$ dans (1).

Il en résulte que **la fréquence conjointe (d'une case du tableau de contingence) est égale au produit des fréquences marginales de la ligne et de la colonne correspondantes** :

$$f^{*}_{ij} = f_{i.} \times f_{.j} \quad (A)$$

Si l'on raisonne en pourcentage, il faut diviser par 100 le produit des fréquences marginales.

Si l'on raisonne selon les **effectifs**, on remplace dans (A) les fréquences par les rapports d'effectifs correspondants :

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

et : $$\frac{n_{ij}}{n_{..}} = \frac{n_{i.}}{n_{..}} \times \frac{n_{.j}}{n_{..}} \quad \Leftrightarrow \quad n^{*}_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$$

C'est-à-dire que lorsque y est indépendant de x, **l'effectif conjoint (d'une case du tableau de contingence) est égale au produit des effectifs marginaux de la ligne et de la colonne correspondantes, divisé par l'effectif total**.

### 22. Les tableaux d'indépendance théorique

Objectif : à partir de l'exemple numérique traité dans le point 1., nous allons construire un tableau d'indépendance théorique selon les effectifs, puis un tableau d'indépendance théorique selon les fréquences, sur la base des résultats du point 21.

Dans un deuxième temps, nous verrons qu'on peut tirer des enseignements sur la dépendance des deux caractères, en comparant, case par case, en lignes et / ou en colonnes :

a) soit les tableaux d'effectifs conjoints réels et d'effectifs théoriques d'indépendance ;
b) soit les tableaux de fréquences conjointes réelles et de fréquences théoriques d'indépendance.

Exemple :

| $x_i$ \ $y_j$ | Hommes | Femmes | $n_{i.}$ |
|---|---|---|---|
| **Ouvriers** | 800 | 150 | 950 |
| **Employés** | 150 | 330 | 480 |
| **Cadres** | 50 | 20 | 70 |
| $n_{.j}$ | 1 000 | 500 | 1 500 |

| $x_i$ \ $y_j$ | H | F | $f_{i.}$ |
|---|---|---|---|
| **O** | 53,3 | 10,0 | 63,3 |
| **E** | 10,0 | 22,0 | 32,0 |
| **C** | 3,3 | 1,3 | 4,7 |
| $f_{.j}$ | 66,7 | 33,3 | 100,0 |

Le tableau de gauche correspond au tableau de contingence des **effectifs conjoints réels** de la série statistique précédente, pour laquelle le caractère x représente la CSP des travailleurs d'une entreprise et le caractère y le sexe de ces travailleurs.

Le tableau de droite donne, en %, les valeurs des **fréquences conjointes réelles** : $f_{ij} = \frac{n_{ij}}{n_{..}}$

On a aussi les **fréquences conditionnelles réelles de x (en colonnes et en %)** : $f_{i/j} = \frac{n_{ij}}{n_{.j}}$

| Sexe (yj) / CSP (xi) | H | F | f i. |
|---|---|---|---|
| **Ouvriers** | 80,0 | 30,0 | 63,3 |
| **Employés** | 15,0 | 66,0 | 32,0 |
| **Cadres** | 5,0 | 4,0 | 4,7 |
| **Total** | 100,0 | 100,0 | 100,0 |

et les **fréquences conditionnelles réelles de y (en lignes et en %)** : $f_{j/i} = \frac{n_{ij}}{n_{i.}}$

| Sexe (yj) / CSP (xi) | H | F | Total |
|---|---|---|---|
| **Ouvriers** | 84,2 | 15,8 | 100 |
| **Employés** | 31,3 | 68,8 | 100 |
| **Cadres** | 71,4 | 28,6 | 100 |
| **f .j** | 66,7 | 33,3 | 100 |

A partir des deux premiers tableaux, on peut construire un **tableau d'indépendance théorique des effectifs** et un **tableau d'indépendance théorique des fréquences** (ici, en %).

Pour cela, on commence par entrer les marges de l'un ou l'autre de ces tableaux, puis on calcule les valeurs conjointes théoriques, comme on a procédé au point 21.

Pour obtenir les effectifs conjoints théoriques d'indépendance, on a : $n^{*}_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$

Pour obtenir les fréquences conjointes théoriques d'indépendance, on a : $f^{*}_{ij} = \frac{f_{i.} \times f_{.j}}{100}$

**Dans le cas des effectifs, les valeurs ont été arrondies à l'entier le plus proche** :

| $x_i$ \ $y_j$ | Hommes | Femmes | $n_{i.}$ |
|---|---|---|---|
| **Ouvriers** | **633** | **317** | 950 |
| **Employés** | **320** | **160** | 480 |
| **Cadres** | **47** | **23** | 70 |
| $n_{.j}$ | 1.000 | 500 | 1.500 |

| $x_i$ \ $y_j$ | H | F | $f_{i.}$ |
|---|---|---|---|
| **O** | **42,2** | **21,1** | 63,3 |
| **E** | **21,3** | **10,7** | 32,0 |
| **C** | **3,1** | **1,6** | 4,7 |
| $f_{.j}$ | 66,7 | 33,3 | 100,0 |

Par exemple : $n^{*}_{11} = \frac{n_{1.} \times n_{.1}}{n_{..}} = \frac{950 \times 1000}{1500} = 633(,3)$

$$\text{et :} \quad f^{*}_{11} = \frac{f_{1.} \times f_{.1}}{f_{..}} = \frac{63,3 \times 66,7}{100} = 42,2\%$$

L'astérisque signifie qu'on calcule des effectifs (ou des fréquences) théoriques, de façon à ce que les caractères x et y soient indépendants, c-à-d qu'ils n'aient aucune influence l'un sur l'autre.

A partir du tableau des fréquences conjointes théoriques d'indépendance, on obtient des tableaux de fréquences conditionnelles théoriques remarquables (on pourrait procéder de même à partir du tableau de les effectifs conjoints théoriques d'indépendance).

Le **tableau des fréquences conditionnelles théoriques en colonnes (caractère x)** :

| Sexe (yj) / CSP (xi) | H | F | f i. |
|---|---|---|---|
| **Ouvriers** | 63,3 | 63,3 | 63,3 |
| **Employés** | 32,0 | 32,0 | 32,0 |
| **Cadres** | 4,7 | 4,7 | 4,7 |
| **Total** | 100,0 | 100,0 | 100,0 |

On constate que la structure de toutes les colonnes est identique. **Le caractère x est donc bien indépendant du caractère y, c'est-à-dire que la CSP des salariés ne dépend pas du sexe de ces derniers**, dans cette entreprise théorique.

Quel que soit le sexe, la distribution des qualifications de travail est la même.

x étant indépendant de y, on a : $f_{i/j} = f_{i.}$

Les fréquences conditionnelles de x sont égales aux fréquences marginales de x.

Le **tableau des fréquences conditionnelles théoriques en lignes (caractère y)** :

| Sexe (yj) / CSP (xi) | H | F | Total |
|---|---|---|---|
| **Ouvriers** | 66,7 | 33,3 | 100,0 |
| **Employés** | 66,7 | 33,3 | 100,0 |
| **Cadres** | 66,7 | 33,3 | 100,0 |
| **f .j** | 66,7 | 33,3 | 100,0 |

On constate que la structure de toutes les lignes est identique. **Le caractère y est donc bien indépendant du caractère x, c'est-à-dire que le sexe des salariés ne dépend pas de la CSP de ces derniers**, dans cette entreprise théorique.

Quelle que soit la CSP envisagée, la proportion hommes / femmes reste la même.

y étant indépendant de x, on a : $f_{j/i} = f_{.j}$

Les fréquences conditionnelles de y sont égales aux fréquences marginales de y.

**Qu'en est-il de la situation réelle dans l'entreprise ?** (par rapport à cette situation théorique d'indépendance entre les deux caractères).

**Première méthode**

Pour le savoir, il faut comparer, case par case, les valeurs du tableau des effectifs conjoints réels avec celle du tableau des effectifs conjoints théoriques d'indépendance. Ou bien on effectue cette comparaison entre le tableau des fréquences conjointes réelles et le tableau des fréquences conjointes théoriques d'indépendance.

Compte tenu d'une part des problèmes d'arrondis qui peuvent exister dans certains cas, et d'autre part du fait qu'il est plus "parlant" de raisonner en pourcentage, il vaut mieux réaliser les comparaisons entre le **tableau des fréquences conjointes réelles et celui des fréquences conjointes théoriques d'indépendance** :

| $x_i$ \ $y_j$ | H | F | $f_{i.}$ |
|---|---|---|---|
| O | 53,3 | 10,0 | 63,3 |
| E | 10,0 | 22,0 | 32,0 |
| C | 3,3 | 1,3 | 4,7 |
| $f_{.j}$ | 66,7 | 33,3 | 100,0 |

| $x_i$ \ $y_j$ | H | F | $f_{i.}$ |
|---|---|---|---|
| O | 42,2 | 21,1 | 63,3 |
| E | 21,3 | 10,7 | 32,0 |
| C | 3,1 | 1,6 | 4,7 |
| $f_{.j}$ | 66,7 | 33,3 | 100,0 |

On peut réaliser l'analyse comparative des deux tableaux, soit en colonnes, soit en lignes.

**A. Analyse comparative des deux tableaux en colonnes**

Dans ce cas, si le caractère x (la CSP) était indépendant du caractère y (le sexe), on aurait, pour chaque sexe, une même répartition des qualifications (cf. ci-dessus le tableau des fréquences conditionnelles théoriques d'indépendance en colonnes).

De fait, en comparant les 2 tableaux de fréquences conjointes, on observe que **les hommes sont surreprésentés dans la réalité chez les ouvriers** (53,3% contre 42,2% dans l'hypothèse d'indépendance) **et sous-représentés chez les employés** (10,0% contre 21,3% dans l'hypothèse d'indépendance). L'indépendance n'est vérifiée que pour les cadres.

En ce qui concerne **les femmes, celles-ci sont sous-représentées dans la réalité chez les ouvrières** (10,0% contre 21,1% dans l'hypothèse d'indépendance) **et surreprésentées chez les employées** (22,0% contre 10,7% dans l'hypothèse d'indépendance). L'indépendance n'est vérifiée que pour les cadres.

Conclusion générale : on peut conclure qu'il existe un **lien entre le fait d'être un homme et simultanément un ouvrier**, dans cette entreprise. De même, il existe un **lien entre le fait d'être une femme et simultanément une employée**, dans cette entreprise.

**B. Analyse comparative des deux tableaux en lignes**

Dans ce cas, si le caractère y (le sexe) était indépendant du caractère x (la CSP), on aurait, pour chaque CSP, une même proportion d'hommes et de femmes (cf. ci-dessus le tableau des fréquences conditionnelles théoriques d'indépendance en lignes).

De fait, en comparant les 2 tableaux de fréquences conjointes, on observe que **pour les ouvriers, les hommes sont surreprésentés dans la réalité** (53,3% contre 42,2% dans l'hypothèse d'indépendance). **A l'inverse, les femmes sont sous-représentées** (10,0% contre 21,1%).

**En ce qui concerne les employés, les hommes sont sous-représentés dans la réalité** (10,0% contre 21,3% dans l'hypothèse d'indépendance). **A l'inverse, les femmes sont surreprésentées** (22,0% contre 10,7%).

Enfin, **pour les cadres, le sexe est à peu près indépendant de la CSP**, puisque les écarts entre les valeurs de chacun des tableaux sont très faibles. Pour les hommes, on a 3,3% dans la réalité contre 3,1% dans l'hypothèse d'indépendance. Pour les femmes, on a 1,3% dans la réalité contre 1,6% dans l'hypothèse d'indépendance.

Conclusion générale : on retrouve ici une conclusion similaire à la précédente, mais en même temps, le point de vue différent : il existe un **lien entre le fait d'être ouvrier et simultanément homme**, dans cette entreprise. De même, il a un **lien entre le fait d'être employée et simultanément femme**, dans cette entreprise.

Sur l'ensemble des deux analyses, on peut avancer qu'en ce qui concerne les cadres, il y a indépendance entre les deux caractères (les écarts observés entre les deux tableaux ne sont pas significatifs).

**Deuxième méthode**

On considère le tableau des effectifs conjoints réels et celui des effectifs conjoints théoriques d'indépendance.

On génère un nouveau tableau. Pour chacune des cases de ce tableau, on pose :

$$\left( \frac{\text{effectif conjoint réel}}{\text{effectif conjoint théorique d'indépendance}} - 1 \right) \times 100$$

Il s'agit d'un pourcentage, tel que :

- s'il est positif, cela signifie que l'effectif réel excède l'effectif théorique correspondant. Cela est donc le signe d'une **surreprésentation réelle**.

- s'il est négatif, cela signifie que l'effectif réel est inférieur à l'effectif théorique correspondant. Cela est donc le signe d'une **sous-représentation réelle**.

Exemple : 1ère case du tableau : $\left( \frac{n_{11}}{n^{*}_{11}} - 1 \right) \times 100 = \left( \frac{800}{633} - 1 \right) \times 100 = +26{,}4\%$

Ce résultat est positif. Il dénote donc une surreprésentation réelle des ouvriers hommes, par rapport à l'hypothèse d'indépendance.

**Cette méthode permet de faire apparaître directement une certaine "intensité" de sur ou sous-représentation de l'effectif correspondant par rapport à l'effectif théorique d'indépendance.**

On obtient le tableau (% des effectifs conjoints réels aux effectifs conjoints théoriques) suivant :

| Sexe (yj) / CSP (xi) | H | F | ([n i./ n i.] - 1) x 100 |
|---|---|---|---|
| **Ouvriers** | **26,4** | -52,7 | 0,0 |
| **Employés** | -53,1 | **106,3** | 0,0 |
| **Cadres** | 6,4 | -13,0 | 0,0 |
| **([n .j / n .j] - 1) x 100** | 0,0 | 0,0 | 0,0 |

Cependant, il est important de relativiser la valeur des pourcentages obtenus (en positif comme en négatif) en remarquant que chacun d'entre eux renvoie à des effectifs absolus réels très différents.

C'est pourquoi il est indispensable de joindre à ce tableau celui des **fréquences conjointes réelles**, de façon à pouvoir apprécier le poids (importance relative) de chacune des cases du tableau précédent.

| $x_i$ \ $y_j$ | H | F | $f_{i.}$ |
|---|---|---|---|
| **O** | **53,3** | 10,0 | 63,3 |
| **E** | 10,0 | **22,0** | 32,0 |
| **C** | 3,3 | 1,3 | 4,7 |
| $f_{.j}$ | 66,7 | 33,3 | 100,0 |

On observe ici que les pourcentages les plus importants du tableau précédent apparaissent sur sa diagonale principale (leur somme représente en effet 75,3 % de l'effectif total). On retrouve cette tendance dans le tableau des fréquences conjointes réelles ci-dessus, pour les femmes employées (+ 106,3 %), mais de manière beaucoup moins nette pour les hommes ouvriers (+ 26,4 %).

Ce constat permet d'avancer que les deux caractères ne sont pas indépendants.

Pour autant, cette analyse ne permette pas de mesurer l'intensité de la liaison qui existe entre les deux caractères x et y. Pour cela, il faudrait réaliser, par ex., un test de "Khi-deux").

Conclusion générale, relative aux deux méthodes :

de manière générale, les deux méthodes montrent que les hommes ouvriers sont surreprésentés dans cette entreprise. Il en est de même pour les femmes employées. Pour les cadres seulement, le sexe est à peu près indépendant de la CSP.

**Remarque terminale**

Au-delà de ce constat de sur ou sous-représentation relative à un caractère donné, peut-on réaliser une mesure plus précise des écarts observés entre les effectifs réels et les effectifs théoriques ? En d'autres termes, peut-on mesurer l'intensité de la liaison qui existe entre CSP et sexe ?

La réponse est oui, mais cela nécessite de mettre en œuvre une procédure utilisant une loi de probabilité, appelée loi de $\chi^2$ ("khi-deux"). Cette méthode consiste, en gros, à tester les carrés des écarts qui existent entre les effectifs théoriques et les effectifs réels (test du $\chi^2$).

Outre ce moyen de mesure quantifiée de l'existence ou non d'une liaison entre les deux caractères, sur les caractères qualitatifs, on peut aussi réaliser des analyses typologiques ou des segmentations (basées sur la notion de distance). Cf. analyses des données et statistiques multidimensionnelles.

Lorsqu'on travaille sur des caractères quantitatifs, il existe d'autres moyens de déterminer la dépendance ou l'indépendance (notamment linéaire) entre deux caractères (coefficient de corrélation linéaire, rapports de corrélation, égalité entre les moyennes marginales et les moyennes conditionnelles).

## CHAPITRE 3 : MISE EN RELATION D'UN CARACTÈRE QUALITATIF ET D'UN CARACTÈRE QUANTITATIF

Les modalités prises par l'un des 2 caractères x et y ne sont pas mesurables. Elles sont mesurables pour l'autre caractère.

Remarque préliminaire aux chapitres 3 et 4

De manière générale, en ce qui concerne les caractères quantitatifs, nous travaillerons uniquement sur des variables discrètes.

En effet, dans le cas des variables continues, si le nombre de classes est faible (c'est-à-dire inférieur à 10 ou 12 classes), il devient hasardeux de poser l'hypothèse des centres de classe, lorsqu'on considère des séries à deux caractères. Il est alors préférable de travailler sur des valeurs moyennes (discrètes), dont on connaît de plus la dispersion.

La raison essentielle de cette disposition tient au fait que la dispersion qu'on observe simultanément sur deux dimensions, s'établit de façon encore plus aléatoire que sur une seule dimension : **l'hypothèse d'équirépartition des effectifs à l'intérieur d'une classe, déjà difficilement soutenable dans certains cas lorsqu'on raisonne sur un seul caractère, ne tient plus du tout lorsqu'on travaille sur deux caractères**. De plus, pour des raisons qui tiennent à la théorie de l'échantillonnage, pour obtenir des résultats valides, les effectifs conjoints doivent être supérieurs à 30 individus (cf. cours de statistiques mathématiques et probabilistes).

Toutefois, dans certains ouvrages ou certains exercices, on trouve des classes (caractères quantitatifs continus). Il faut considérer qu'il s'agit d'hypothèses d'école. Dans la pratique, il vaut mieux éviter de réaliser, sans précautions particulières, de tels types de traitement (cf. traitements d'enquêtes). Ou alors, il faut travailler sur des tableaux de grande taille (de 15 l. x 15 c. à 20 l. x 20 c. au moins), avec des effectifs conjoints d'au moins 30 unités.

De manière générale, les tableaux qui croisent un caractère qualitatif et un caractère quantitatif sont construits de la même façon que ceux qui croisent deux caractères qualitatifs (cf. chapitre 2).

Mais au-delà de cette mise en œuvre, lorsqu'on croise un caractère qualitatif et un caractère quantitatif, on peut mener une analyse quantifiée, à la fois plus précise et plus objective sur le caractère quantitatif. Il est en effet possible de **calculer des paramètres marginaux et des paramètres conditionnels sur le caractère quantitatif**.

On peut mener l'analyse selon les quatre étapes suivantes :

1) des **graphiques spécifiques**, qui vont se démarquer des graphiques vus au chapitre 2.

2) des **tableaux théoriques d'indépendance**, envisagés de la même façon qu'au chapitre 2, en vue de les comparer aux tableaux réels d'effectifs ou de fréquences (cf. plus loin le point 2).

3) des **calculs de moyenne, de variance, d'écart-type et de coefficient de variation**, portant sur le caractère quantitatif (selon chaque modalité du caractère qualitatif).

4) Grâce aux paramètres précédents, on peut calculer un **rapport d'explication** (exprimable en pourcentage), qui mesure quantitativement l'intensité de la liaison existant entre les deux caractères (cf. plus loin le point 4). Il s'agit de mesurer la quantité d'information apportée par le caractère qualitatif. En d'autres termes, on cherche à savoir si les deux caractères sont indépendants ou non.

Remarque : pour calculer le rapport d'explication, on réalise une **décomposition de la variance**.

### 1. Les représentations graphiques appropriées

Si l'on raisonne selon le **caractère qualitatif**, on peut reprendre les éléments décrits au chapitre 2 : diagrammes en tuyau d'orgue à base variable, diagrammes en secteurs semi-circulaires (lorsque le caractère qualitatif ne possède que deux modalités) ou stéréogrammes.

Si l'on raisonne selon le **caractère quantitatif**, on établit un diagramme en bâtons particulier (= diagramme en bâtons segmentés), chaque bâton étant segmenté selon les modalités du caractère qualitatif (en effectifs, ou bien en fréquences ou en pourcentages). Si l'on utilise un tableur, on peut aussi utiliser des stéréogrammes.

**Exemple 1** :

On considère les familles résidant en France (en milliers) au moment du **recensement de 1982** (source : Tef / Insee), selon les deux caractères suivants :

- caractère x (en lignes) : le **nombre d'enfants de 0 à 16 ans** (**caractère quantitatif**).

- caractère y (en colonnes) : la **nationalité** (**caractère qualitatif**).

**Tableaux des effectifs conjoints de réels**

| $x_i$ \ $y_j$ | Nationalité française | Nationalité étrangère | Total ($n_{i.}$) |
|---|---|---|---|
| **0** | 6.815 | 316 | 7.131 |
| **1** | 3.005 | 196 | 3.201 |
| **2** | 2.320 | 179 | 2.499 |
| **3** | 814 | 105 | 919 |
| **4 et plus** | 264 | 107 | 371 |
| **Total ($n_{.j}$)** | 13.218 | 903 | 14.121 |

Population : le nombre de familles résidant en France en 1982 (en milliers).

Caractère x : le nombre d'enfants de 0 à 16 ans des familles. Caractère quantitatif discret.

Caractère y : la nationalité des familles résidant en France en 1982. Caractère qualitatif.

**1°)** A partir du tableau des effectifs conjoints ci-dessus, réaliser un graphique de la distribution des nationalités des familles, **selon le nombre d'enfants de 0 à 16 ans**. Pour quelle raison ce graphique n'est-il pas vraiment approprié, pour rendre compte correctement de cette série à deux caractères ?

**2°)** Pour lever la difficulté mise en évidence à la question 1, construire le tableau des fréquences adéquates.

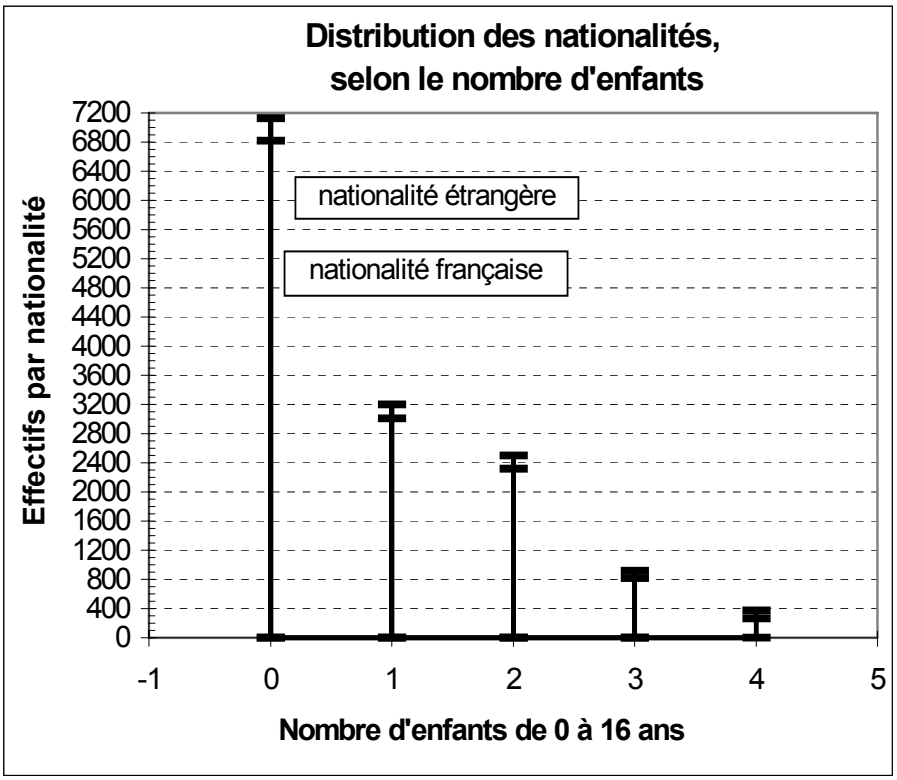
**3°)** De même, toujours à partir du tableau des effectifs conjoints, réaliser un graphique approprié de la distribution du nombre d'enfants de 0 à 16 ans des familles, **selon la nationalité**. On procédera comme aux questions 1 et 2.

**Question 1**

Répondre à la question 1 revient à envisager un **diagramme en bâtons segmentés**.

En abscisse, on trouve les modalités du caractère **quantitatif**, ici le nombre d'enfants de 0 à 16 ans.

En ordonnée, on a les effectifs correspondant à chaque nombre d'enfants. De plus, on ajoute des marques (tirets, segments) qui vont matérialiser les effectifs correspondant à chaque modalité du caractère **qualitatif**.

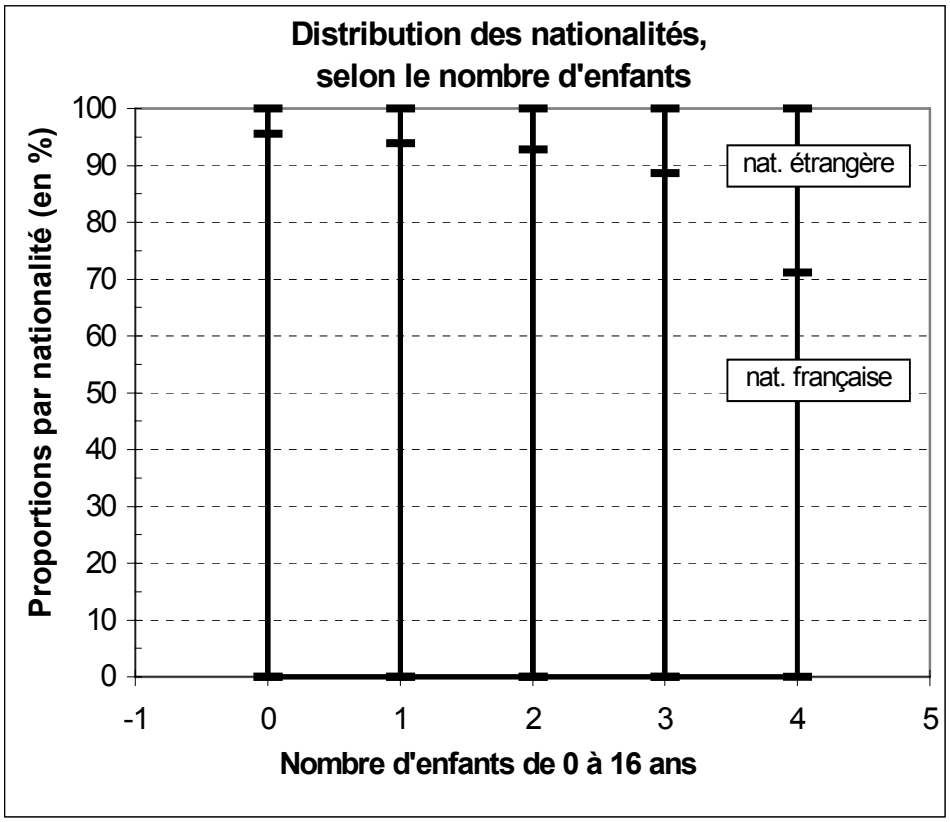



**Question 2**

Bien que correct, le graphique précédent, selon les effectifs, n'est pas approprié, car sa lecture ne permet pas, pour un nombre donné d'enfants, une comparaison précise des **proportions** entre les nationalités.

Il faut donc calculer ici les **fréquences conditionnelles en lignes** (c-à-d les **fréquences conditionnelles de x** : nombre d'enfants de 0 à 16 ans). Les valeurs sont données en % :

| y j / x i \ | Nat. française | Nat. étrangère | Total |
|---|---|---|---|
| **0** | 95,6 | 4,4 | 100,0 |
| **1** | 93,9 | 6,1 | 100,0 |
| **2** | 92,8 | 7,2 | 100,0 |
| **3** | 88,6 | 11,4 | 100,0 |
| **4** | 71,2 | 28,8 | 100,0 |
| **Total (f .j)** | 93,6 | 6,4 | 100,0 |

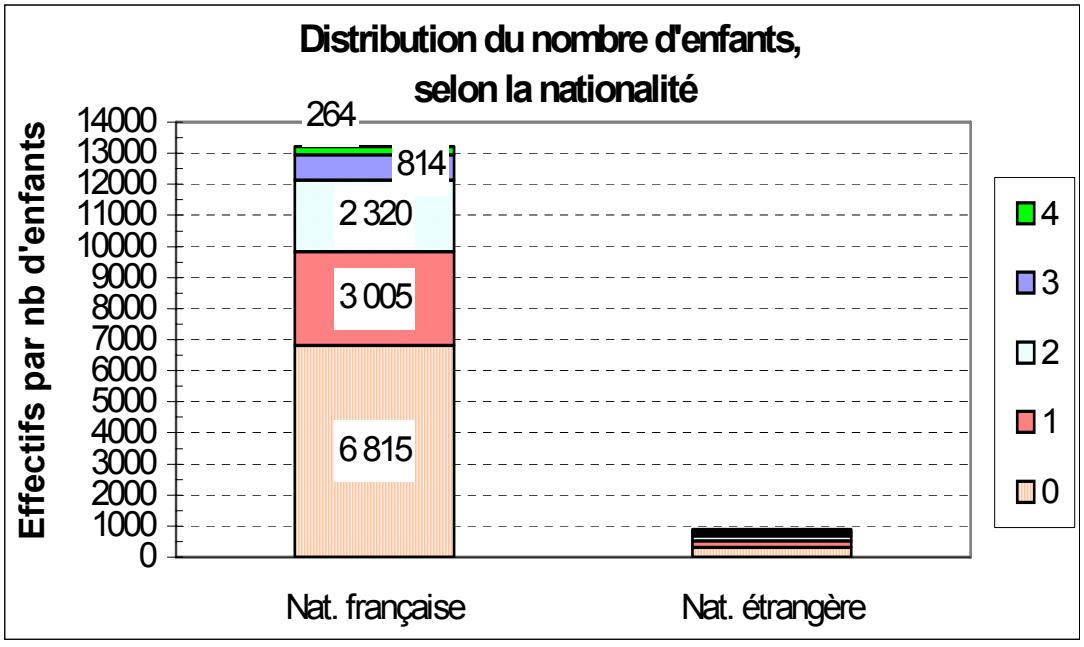



**Question 3**

Répondre à la question 3 revient à envisager un **diagramme en tuyaux d'orgue à base variable**.

En abscisse, on trouve les modalités du caractère **qualitatif**, ici la nationalité des familles.

En ordonnée, on a les effectifs correspondant à chaque nombre d'enfants, c-à-d correspondant à chaque modalité du caractère quantitatif x.



Bien que correct, le graphique précédent, selon les effectifs, n'est pas approprié, car sa lecture ne permet pas, pour une nationalité donnée, une comparaison précise des **proportions** du nombre d'enfants.

Ici, de plus, le déséquilibre numérique entre les 2 modalités du caractère "Nationalité" invalide toute tentative de lecture correcte.

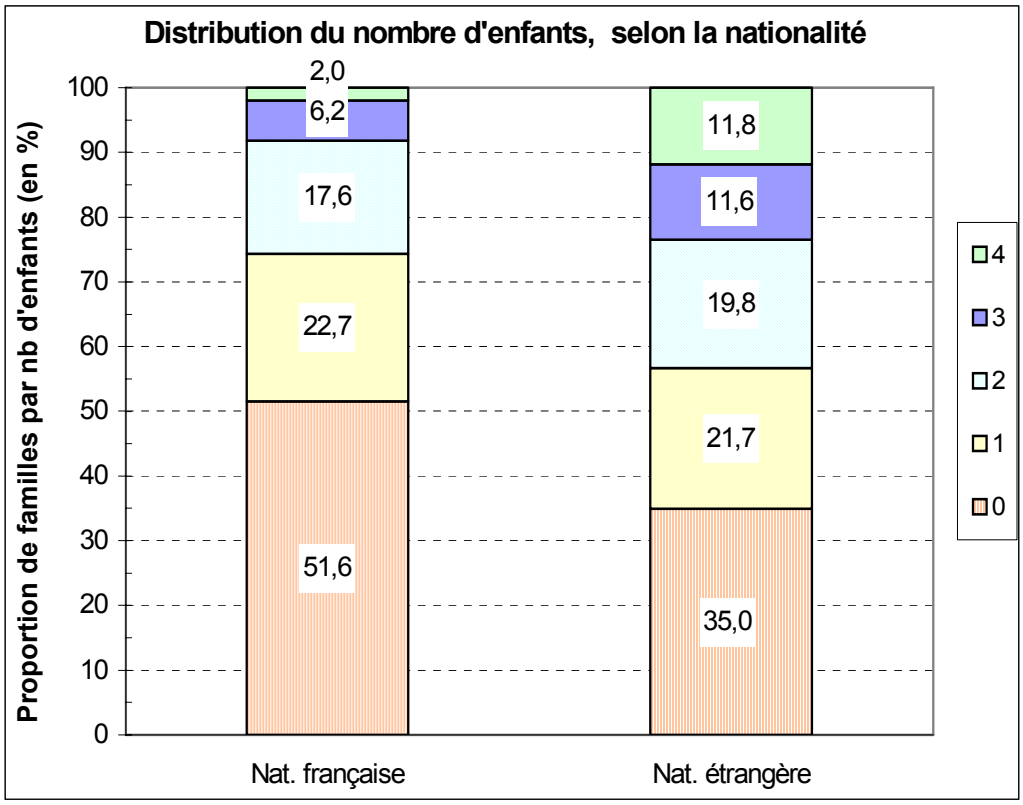
Pour rendre correctement compte des deux distributions simultanément, il est nécessaire de faire apparaître sur le graphique, en abscisse, la proportion relative des deux nationalités, c-à-d les fréquences marginales de y (f .j).

On met en œuvre la technique des tuyaux d'orgue à base variable. C'est pourquoi on construit ici le **tableau des fréquences conjointes réelles** :

| y j / x i \ | Nat. française | Nat. étrangère | Total (f i.) |
|---|---|---|---|
| **0** | 48,3 | 2,2 | 50,5 |
| **1** | 21,3 | 1,4 | 22,7 |
| **2** | 16,4 | 1,3 | 17,7 |
| **3** | 5,8 | 0,7 | 6,5 |
| **4** | 1,9 | 0,8 | 2,6 |
| **Total (f .j)** | 93,6 | 6,4 | 100,0 |

Ici, il faut aussi calculer les **fréquences conditionnelles en colonnes** (c-à-d les **fréquences conditionnelles de y** : nationalité des familles). Les valeurs sont données en % :

| y j / x i \ | Nat. française | Nat. étrangère | Total (f i.) |
|---|---|---|---|
| **0** | 51,6 | 35,0 | 50,5 |
| **1** | 22,7 | 21,7 | 22,7 |
| **2** | 17,6 | 19,8 | 17,7 |
| **3** | 6,2 | 11,6 | 6,5 |
| **4** | 2,0 | 11,8 | 2,6 |
| **Total** | 100,0 | 100,0 | 100,0 |




**Attention** : les tuyaux d'orgue ci-dessus, réalisés avec un tableur ne sont pas corrects en abscisse, car **leurs bases ne sont pas variables, en fonction du poids de chaque nationalité (93,6 %, 6,4 %)**.

**Remarque terminale** : ici, le caractère qualitatif ne présente que deux modalités. On aurait donc également pu utiliser la technique des **secteurs semi-circulaires** (cf. ch. 2). Toutefois, la disproportion des effectifs de chaque modalité est ici préjudiciable à une bonne lecture du demi-cercle "Nationalité étrangère".

### 2. Dépendance et indépendance entre les deux caractères

Deux présentations du tableau sont possibles selon que le caractère quantitatif est placé en ligne ou en colonne.

**Mise en œuvre des tableaux théoriques d'indépendance**

**Cas où le caractère quantitatif est le caractère x**

Rappel : les modalités de x sont placées dans la 1ère colonne du tableau.

On a le tableau général suivant :

| $x_i$ \ $y_j$ | $y_1$ | … | $y_j$ | ... | $y_p$ | $n_{i.}$ |
|---|---|---|---|---|---|---|
| $x_1$ | $N_{11}$ | … | $n_{1j}$ | | $n_{1p}$ | $n_{1.}$ |
| … | … | … | … | … | … | … |
| $x_i$ | $n_{i1}$ | … | $n_{ij}$ | … | $n_{ip}$ | $n_{i.}$ |
| ... | … | … | … | … | … | … |
| $x_m$ | $N_{m1}$ | … | $n_{mj}$ | … | $n_{mp}$ | $n_{m.}$ |
| $n_{.j}$ | $n_{.1}$ | … | $n_{.j}$ | … | $n_{.p}$ | $n_{..}$ |

Exemple : nous poursuivons l'exemple précédent : dans le tableau 1, on considère les familles (en milliers) résidant en France en 1982, selon un caractère quantitatif x, qui indique le nombre d'enfants de 0 à 16 ans des différentes familles, et un caractère qualitatif y qui indique la nationalité (française ou étrangère) de la famille.

On cherche à savoir dans quelle mesure le caractère qualitatif y (la nationalité de la famille) exerce une influence sur le nombre d'enfants (de 0 à 16 ans) de la famille.

**4°)** Après en avoir rappelé le principe de construction, établir le tableau des effectifs conjoints théoriques d'indépendance, de même que celui des fréquences conjointes théoriques d'indépendance.

**5°)** Construire le tableau des fréquences conditionnelles théoriques d'indépendance en lignes, puis celui des fréquences conditionnelles théoriques d'indépendance en colonnes. Constater et expliquer.

**6°)** A partir de la comparaison du tableau des fréquences conjointes réelles et de celui des fréquences conjointes théoriques d'indépendance, analyser, selon deux méthodes différentes, la dépendance ou l'indépendance des deux caractères x et y.

**Tableau 1 : effectifs conjoints réels**

| x i \ y j | Nationalité française | Nationalité étrangère | n i . |
|---|---|---|---|
| **0** | 6 815 | 316 | 7 131 |
| **1** | 3 005 | 196 | 3 201 |
| **2** | 2 320 | 179 | 2 499 |
| **3** | 814 | 105 | 919 |
| **4** | 264 | 107 | 371 |
| **N . j** | 13 218 | 903 | 14 121 |

## Question 4

Rappel : pour construire un tableau des effectifs conjoints théoriques d'indépendance, on utilise la formule : $n_{ij} = (n_{i.} \times n_{.j}) / n_{..}$

Comme il s'agit d'effectifs qui portent sur des nombres entiers, **on arrondit** les résultats obtenus, pour chaque case du tableau, **à l'entier le plus proche**. On obtient le tableau 2 suivant, en utilisant la procédure vue au chapitre 2, point 21

**Tableau 2 : effectifs théoriques d'indépendance**

| x i \ y j | Nationalité française | Nationalité étrangère | n i . |
|---|---|---|---|
| **0** | 6 675 | 456 | 7 131 |
| **1** | 2 996 | 205 | 3 201 |
| **2** | 2 339 | 160 | 2 499 |
| **3** | 860 | 59 | 919 |
| **4** | 347 | 24 | 371 |
| **N . j** | 13 218 | 903 | 14 121 |

Indiquons à nouveau le tableau des fréquences conjointes réelles (en %) :

**Tableau 3 : fréquences conjointes réelles (en %)**

| x i \ y j | Nationalité française | Nationalité étrangère | f i . |
|---|---|---|---|
| **0** | 48,3 | 2,2 | 50,5 |
| **1** | 21,3 | 1,4 | 22,7 |
| **2** | 16,4 | 1,3 | 17,7 |
| **3** | 5,8 | 0,7 | 6,5 |
| **4** | 1,9 | 0,8 | 2,6 |
| **f j** | 93,6 | 6,4 | 100,0 |

Rappel : pour construire un tableau des fréquences conjointes théoriques d'indépendance (exprimées en pourcentage), on utilise la formulation : **f ij = (f i. x f .j) / 100**

**Tableau 4 : fréquences conjointes théoriques d'indépendance (en %)**

| x i \ y j | Nationalité française | Nationalité étrangère | f i . |
|---|---|---|---|
| **0** | 47,3 | 3,2 | 50,5 |
| **1** | 21,2 | 1,4 | 22,7 |
| **2** | 16,6 | 1,1 | 17,7 |
| **3** | 6,1 | 0,4 | 6,5 |
| **4** | 2,5 | 0,2 | 2,6 |
| **f . j** | 93,6 | 6,4 | 100,0 |

## Question 5

**Fréquences conditionnelles théoriques d'indépendance en lignes (en %)** $f_{j/i} = \frac{n_{ij}}{n_{i.}}$

La nationalité (y) est indépendante du nombre d'enfants (x) :

| y j / x i \ | Nat. française | Nat. étrangère | Total |
|---|---|---|---|
| **0** | 93,6 | 6,4 | 100,0 |
| **1** | 93,6 | 6,4 | 100,0 |
| **2** | 93,6 | 6,4 | 100,0 |
| **3** | 93,6 | 6,4 | 100,0 |
| **4** | 93,6 | 6,4 | 100,0 |
| **Total (f .j)** | 93,6 | 6,4 | 100,0 |

**Fréquences conditionnelles théoriques d'indépendance en colonnes (en %)** $f_{i/j} = \frac{n_{ij}}{n_{.j}}$

Le nombre d'enfants (x) est indépendant de la nationalité (y) :

| y j / x i \ | Nat. française | Nat. étrangère | Total (f i.) |
|---|---|---|---|
| **0** | 50,5 | 50,5 | 50,5 |
| **1** | 22,7 | 22,7 | 22,7 |
| **2** | 17,7 | 17,7 | 17,7 |
| **3** | 6,5 | 6,5 | 6,5 |
| **4** | 2,6 | 2,6 | 2,6 |
| **Total** | 100,0 | 100,0 | 100,0 |

**Interprétations** :

Dans le tableau des fréquences conditionnelles théoriques d'indépendance **en lignes**, on constate que les pourcentages sont égaux pour chaque ligne et correspondent à ceux des fréquences marginales f .j .

Cela signifie que la nationalité (y) est indépendante du nombre d'enfants (x). Pour chaque nombre d'enfants, on a la même proportion de familles françaises ou étrangères.

Dans le tableau des fréquences conditionnelles théoriques d'indépendance **en colonnes**, on constate que les pourcentages sont égaux pour chaque colonne et correspondent à ceux des fréquences marginales f i.

Cela signifie que le nombre d'enfants (x) est indépendant de la nationalité (y). Pour chaque nationalité, la distribution des nombres d'enfants est la même.

## Question 6 : analyse de l'indépendance des deux caractères

### 21. Première méthode

On doit comparer deux tableaux (cf. chapitre 2, point 22) :

- soit celui des effectifs conjoints réels et celui des effectifs conjoints théoriques d'indépendance ;

- soit celui des fréquences conjointes réelles et celui des fréquences conjointes théoriques d'indépendance.

En général, la 2ème possibilité est la plus pratique, notamment si les effectifs sont élevés et / ou si les écarts entre les deux tableaux sont importants. Nous faisons ce choix ici pour traiter notre exemple.

Comparons donc les tableaux 3 et 4, afin de déterminer si les deux caractères x et y sont ici indépendants ou non, selon les principes suivants :

a) considérons une case de même ligne et de même colonne dans chacun des tableaux 3 et 4. Si l'on constate que les pourcentages sont identiques ou très proches, cela signifie que, au moins pour cette case, il y a indépendance entre les caractères x et y.

b) à l'inverse, pour une case donnée, si les pourcentages sont plus élevés dans le tableau des fréquences réelles que dans celui des fréquences théoriques d'indépendance, cela signifie que, dans la réalité, il y a plus d'individus possédant les modalités correspondantes de x et y que si les caractères étaient indépendants.

c) de même, toujours pour une case donnée, si les pourcentages sont moins élevés dans le tableau des fréquences réelles que dans celui des fréquences théoriques d'indépendance, cela signifie que, dans la réalité, il y a moins d'individus possédant les modalités correspondantes de x et y que si les caractères étaient indépendants.

d) dans les deux derniers cas, on en déduit que les caractères x et y sont d'autant plus dépendants que l'écart entre les deux pourcentages est important.

**Analyse des tableaux en lignes**

Si la nationalité (y) était totalement indépendante du nombre d'enfants (x), on devrait avoir, pour chaque nombre d'enfants, une même proportion de familles françaises et étrangères (cf. question 5). Un tel cas de figure signifierait que, pour une case donnée, les pourcentages constatés dans chacun des deux tableaux seraient identiques.

De fait, en comparant les deux tableaux, on observe que pour 0 enfant, les familles françaises sont légèrement surreprésentées dans la réalité (48,3 % contre 47,3 % dans l'hypothèse d'indépendance) et donc les familles étrangères sont légèrement sous-représentées (2,2 % contre 3,2 % dans l'hypothèse d'indépendance).

A l'inverse, pour 4 enfants ou plus, les familles françaises sont légèrement sous-représentées dans la réalité (1,9 % contre 2,5 % dans l'hypothèse d'indépendance) et donc les familles étrangères sont légèrement surreprésentées (0,8 % contre 0,2 % dans l'hypothèse d'indépendance).

Dans les autres cas (1, 2 ou 3 enfants), les différences sont insignifiantes : la nationalité est indépendante du nombre d'enfants.

**Remarque importante** : dans un tableau à double entrée mettant en jeu deux caractères, une analyse purement statistique est toujours possible, en lignes comme en colonnes. Cependant, il convient de toujours s'interroger sur la pertinence socio-économique de la relation considérée.

Ici, le fait de chercher à savoir si la nationalité (y) est indépendante du nombre d'enfants (x) n'a pas de sens, en termes de causalité. **Quels que soient les caractères que l'on croise (qualitatifs ou quantitatifs), il faut toujours s'assurer de la signification concrète des relations que l'on analyse**.

**Analyse des tableaux en colonnes**

Si le nombre d'enfants (x) était totalement indépendant de la nationalité des familles (y), quelle que soit la nationalité, on devrait avoir une même proportion dans le nombre d'enfants.

De fait, en comparant les deux tableaux, on observe que les familles françaises sont légèrement surreprésentées dans la réalité (48,3 % contre 47,3 % dans l'hypothèse d'indépendance) pour 0 enfant et légèrement sous-représentées dans la réalité (1,9 % contre 2,5 % dans l'hypothèse d'indépendance) pour 4 enfants. Dans les autres cas (1, 2 ou 3 enfants), les différences ne sont pas significatives.

A l'inverse, les familles étrangères sont légèrement sous-représentées dans la réalité (2,2 % contre 3,2 % dans l'hypothèse d'indépendance) pour 0 enfant et légèrement surreprésentées dans la réalité (0,8 % contre 0,2 % dans l'hypothèse d'indépendance) pour 4 enfants. Dans les autres cas (1, 2 ou 3 enfants), les différences ne sont pas significatives.

**Conclusion de l'analyse** : les constats précédents permettent d'avancer que les deux caractères sont pratiquement indépendants : on peut affirmer que le nombre d'enfants ne dépend pas de la nationalité des familles. La dispersion du nombre d'enfants par famille s'explique par d'autres facteurs que la nationalité, notamment le mode de vie, le niveau de vie, la religion, etc.

## 22. Deuxième méthode

Cette autre méthode possible d'analyse de la dépendance ou de l'indépendance entre deux caractères x et y met en jeu le tableau 1 des effectifs réels, le tableau 2 des effectifs théoriques d'indépendance et le tableau 3 des fréquences conjointes réelles (cf. chapitre 2, point 22).

Dans le tableau 5 ci-après, la valeur de chaque case résulte du calcul suivant :

$$\left( \frac{n_{ij}\ \text{réel}}{n_{ij}\ \text{théorique}} - 1 \right) \times 100 \ .$$

Si la valeur du résultat est positive (+), cette opération permet d'exprimer, en pourcentage, de combien l'effectif réel excède l'effectif théorique d'indépendance. Cela donne un ordre de grandeur de la surreprésentation d'une case du tableau des effectifs réels.

Inversement, si la valeur est négative (-), on a une idée de la sous-représentation d'une case du tableau des effectifs réels par rapport à une hypothèse d'indépendance.

**Tableau 5 : rapport des effectifs réels aux effectifs théoriques d'indépendance (en %)**

| x i \ y j | Nationalité française | Nationalité étrangère |
|---|---|---|
| **0** | **2,1 %** | -30,7 % |
| **1** | **0,3 %** | -4,4 % |
| **2** | **-0,8 %** | 11,9 % |
| **3** | -5,3 % | **78,0 %** |
| **4** | -23,9 % | **345,8 %** |

Remarque : les marges des deux tableaux d'effectifs étant identiques, les écarts sur les marges sont nuls.

Par exemple, la valeur 2,1 % de la 1ère case du tableau 5 résulte du calcul suivant :

$$\left( \frac{6\,815}{6\,675} - 1 \right) \times 100 = + 2{,}1\%$$

Cependant, il est nécessaire de relativiser la valeur des pourcentages obtenus (en positif comme en négatif) en remarquant que chacun d'entre eux renvoie à des effectifs absolus réels très différents. C'est pourquoi il est indispensable de joindre à ce tableau celui des fréquences conjointes réelles (tableau 3, repris ci-après), de façon à pouvoir apprécier le poids (importance relative) de chacune des cases du tableau 5 :

**Tableau 3 : fréquences conjointes réelles (en %)**

| x i \ y j | Nationalité française | Nationalité étrangère |
|---|---|---|
| **0** | **48,3 %** | 2,2 % |
| **1** | **21,3 %** | 1,4 % |
| **2** | **16,4 %** | 1,3 % |
| **3** | 5,8 % | **0,7 %** |
| **4** | 1,9 % | **0,8 %** |

L'analyse simultanée des tableaux 3 et 5 permet d'affirmer que les deux caractères x et y sont pratiquement indépendants.

En effet, on constate notamment un très fort pourcentage dans la dernière case du tableau 5 (345,8 %), mais, dans le même temps, pour la même case du tableau 3, il apparaît que le poids relatif de l'effectif réel correspondant est extrêmement faible (0,8 %). Cela signifie que le caractère y a, de fait, peu d'influence sur le caractère x, à l'échelle de l'ensemble du tableau 1.

On peut généraliser ce constat en comparant une à une les cases des tableaux 3 et 5. On observe que là où les pourcentages du tableau 5 sont importants, ceux du tableau 3 sont négligeables, et inversement.

Ainsi, les familles françaises de 0 à 2 enfants de moins de 16 ans représentent 86 % du total (tableau 3), alors que, simultanément, les pourcentages observés dans le tableau 5 montrent que l'écart entre les effectifs réels et théoriques est très peu important.

En conclusion, on peut donc dire que le caractère y (nationalité) a très peu d'influence sur le caractère x (nombre d'enfants de 0 à 16 ans de la famille).

Cela veut dire que la dispersion du nombre d'enfants par famille s'explique par d'autres facteurs que la nationalité (notamment le mode de vie, le niveau de vie, la religion, etc.).

Remarque : le calcul du rapport d'explication, dont la valeur est de 1,8 %, confirme pleinement cette interprétation (voir plus loin).

### 3. Calcul des paramètres marginaux et conditionnels du caractère quantitatif x

Comme pour une série à un caractère, on peut effectuer des calculs supplémentaires sur le caractère quantitatif x, en le considérant :

a) d'une part indépendamment du caractère qualitatif et l'on calcule les **paramètres marginaux du caractère quantitatif x** ;

b) d'autre part en considérant autant de **sous-populations** qu'il y a de modalités pour le caractère **qualitatif** y et l'on calcule les **paramètres conditionnels du caractère quantitatif x**.

Exemple :

**7°)** Calculer moyenne, écart-type et coefficient de variation du nombre d'enfants par famille.

**8°)** Procéder de même pour chacune des nationalités. Commenter les résultats.

**Question 7 : calcul des paramètres marginaux du caractère quantitatif x**

Ici, on raisonne **globalement sur l'ensemble de la population**, sans plus s'occuper de la nationalité des familles (caractère y). Cependant, on reste dans la logique d'un croisement entre deux caractères. Par conséquent, les notations littérales vont être modifiées, par rapport aux séries à un caractère, en conservant deux indices, de la façon suivante :

**Moyenne marginale de x** : $$\bar{x} = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\, x_i$$

**Variance marginale de x** : $$\sigma_x^2 = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\,(x_i - \bar{x})^2 = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\, x_i^2 - \bar{x}^2$$

**Coefficient de variation de x** : $$CV_x = \frac{\sigma_x}{\bar{x}}$$

**Remarque importante : on doit formuler une hypothèse pour remplacer la modalité "4 et plus" par une modalité mesurable. Ici, on retient la valeur "4".**

| x i \ y j | Nat. française | Nat. étrangère | Total (n i.) | ni. xi | ni. xi ² |
|---|---|---|---|---|---|
| **0** | 6.815 | 316 | 7.131 | 0 | 0 |
| **1** | 3.005 | 196 | 3.201 | 3.201 | 3.201 |
| **2** | 2.320 | 179 | 2.499 | 4.998 | 9.996 |
| **3** | 814 | 105 | 919 | 2.757 | 8.271 |
| **4** | 264 | 107 | 371 | 1.484 | 5.936 |
| **Total (n .j)** | 13.218 | 903 | 14.121 | 12.440 | 27.404 |

Pour l'ensemble de la population, la moyenne, l'écart-type et le coefficient de variation du nombre d'enfants par famille sont de :

**Moyenne = $\bar{x} = \frac{12\,440}{14\,121}$ = 0,88 enfant par famille** (enfants de zéro à 16 ans !).

**Variance = $\sigma_x^2 = \frac{27\,404}{14\,121} - (0,88)^2$ = 1,16 enfant² par famille**

**Ecart-type = 1,08 enfant par famille**

**Coefficient de variation = $CV_x = \frac{1,08}{0,88}$ = 1,22**

Valeur très élevée du CV, qui caractérise une très forte dispersion autour de la valeur moyenne observée (écart-type supérieur à la moyenne !!).

**Question 8 : calcul des paramètres conditionnels du caractère quantitatif x**

Ici, on raisonne successivement sur chaque colonne du tableau, c-à-d sur des **sous-populations, par nationalité**.

Nous allons successivement considérer deux façons de présenter les tableaux de calcul :

- d'abord en se ramenant à une série à un caractère, portant sur une sous-population donnée.

- ensuite, en raisonnant directement sur le tableau à double entrée (forme générale usuelle).

Comme précédemment, on reste dans la logique d'un croisement entre deux caractères. Par conséquent, les notations littérales comportent ici encore deux indices et renvoient à une sous-population particulière, de la façon suivante :

**Moyennes conditionnelles de x** : $\bar{x}_j = \frac{1}{n_{.j}} \sum_{j=1}^{m} n_{ij}\, x_i$

Il y a autant de moyennes conditionnelles de x qu'il y a de modalités de y.

**Variances conditionnelles de x** :

$$\sigma^2_{x_j} = \frac{1}{n_{.j}} \sum_{i=1}^{m} n_{ij} (x_i - \bar{x}_j)^2 = \frac{1}{n_{.j}} \sum_{i=1}^{m} n_{ij} x_i^2 - \bar{x}_j^2$$

Il y a autant de variance conditionnelles de x qu'il y a de modalités de y.

**Coefficient de variation de $x_j$** : $CV_{x_j} = \frac{\sigma_{x_j}}{\bar{x}_j}$

**Première méthode** :

**Nationalité française** (on raisonne sur la seule première colonne du tableau)

| y j / x i \ | Nat. française | ni1 xi | ni1 xi ² |
|---|---|---|---|
| **0** | 6.815 | 0 | 0 |
| **1** | 3.005 | 3.005 | 3.005 |
| **2** | 2.320 | 4.640 | 9.280 |
| **3** | 814 | 2.442 | 7.326 |
| **4** | 264 | 1.056 | 4.224 |
| **Total (n .1)** | 13.218 | 11.143 | 23.835 |

**Moyenne conditionnelle =** $\bar{x}_1 = \frac{11143}{13\,218}$ **= 0,84 enfant par famille** (enfants de 0 à 16 ans !).

**Variance conditionnelle =** $\sigma^2_{x_1} = \frac{23\,835}{13\,218} - (0,84)^2$ **= 1,09 enfant² par famille**

**Ecart-type conditionnel = 1,05 enfant par famille**

**Coefficient de variation =** $CV_{x_1} = \frac{1,05}{0,84}$ **= 1,24**

Valeur très élevée du CV, qui caractérise une très forte dispersion autour de la valeur moyenne observée (écart-type supérieur à la moyenne !!).

**Nationalité étrangère** (on raisonne sur la seule deuxième colonne du tableau)

| x i \ y j | Nat. étrangère | ni2 xi | ni2 xi ² |
|---|---|---|---|
| **0** | 316 | 0 | 0 |
| **1** | 196 | 196 | 196 |
| **2** | 179 | 358 | 716 |
| **3** | 105 | 315 | 945 |
| **4** | 107 | 428 | 1.712 |
| **Total (n .2)** | 903 | 1.297 | 3.569 |

**Moyenne conditionnelle =** $\bar{x}_2 = \frac{1297}{903}$ **= 1,44 enfant par famille** (enfants de 0 à 16 ans !).

**Variance conditionnelle =** $\sigma^2_{x_2} = \frac{3\,569}{903} - (1{,}44)^2$ **= 1,89 enfant² par famille**

**Ecart-type conditionnel = 1,37 enfant par famille**

**Coefficient de variation =** $CV_{x_2} = \frac{1{,}37}{1{,}44}$ **= 0,96**

Valeur très élevée du CV, qui caractérise une très forte dispersion autour de la valeur moyenne observée.

**Remarques terminales** :

Les deux moyennes conditionnelles sont assez sensiblement différentes : le nombre moyen d'enfants (de zéro à 16 ans) par famille étrangère est supérieur à celui des familles françaises.

La dispersion du nombre d'enfants par famille est un peu moins élevée pour les familles étrangères résidant en France en 1982 (cette populations est plus homogène).

De façon générale, on peut noter l'intérêt des informations apportées par ces calculs complémentaires, réalisés sur le caractère quantitatif.

**Deuxième méthode** :

Usuellement, on travaille directement sur le tableau à double entrée, en ajoutant deux lignes supplémentaires.

| $x_i$ \ $y_j$ | Nationalité française | Nationalité étrangère | $n_{i.}$ | $n_{i.} x_i$ | $n_{i.} x_i^2$ |
|---|---|---|---|---|---|
| **0** | 6 815 | 316 | 7 131 | 0 | 0 |
| **1** | 3 005 | 196 | 3 201 | 3 201 | 3 201 |
| **2** | 2 320 | 179 | 2 499 | 4 998 | 9 996 |
| **3** | 814 | 105 | 919 | 2 757 | 8 271 |
| **4** | 264 | 107 | 371 | 1 484 | 5 936 |
| $N_{.j}$ | 13 218 | 903 | 14 121 | 12 440 | 27 404 |
| $\sum_i n_{ij} x_i$ | **11 143** | **1 297** | (12 440) | | |
| $\sum_i n_{ij} x_i^2$ | **23 835** | **3 569** | (27 404) | | |

Dans les deux dernières lignes du tableau, on remarque que l'on enregistre directement les sommes, sans détailler leurs calculs comme dans la première méthode.

## 4. Notion et calcul du rapport d'explication

### 41. Principe général (décomposition de la variance)

Soit **P** une population statistique composée de **m** sous-populations $P_1, P_2, \ldots, P_m$.

a) On démontre que la moyenne $\overline{x}$ de la population **P**, constituée par la réunion des **m** sous-populations, est égale à la moyenne pondérée des moyennes $\overline{x}_j$ des **m** sous-populations :

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{m} n_j \overline{x}_j$$ avec : $n = n_1 + n_2 + \ldots + n_m$ ($n_j$ = effectif de la sous-population $P_j$).

b) On démontre que la variance (totale) $\sigma^2$ de la population **P**, constituée par la réunion des **m** sous-populations, est égale à l'expression suivante :

$$\sigma_x^2 = V(x) = \frac{1}{n} \sum_{j=1}^{m} n_j (\overline{x}_j - \overline{x})^2 + \frac{1}{n} \sum_{j=1}^{m} n_j \sigma_{x_j}^2$$

On constate que la variance de x peut être décomposée en deux termes (on parle de **décomposition de la variance**) :

1) Le premier terme correspond à une variance dans laquelle on considère les carrés des écarts entre la moyenne $\overline{x}_j$ de chaque sous-population et la moyenne $\overline{x}$ de l'ensemble de la population (chaque terme de la somme étant pondéré par les effectifs $n_j$ de la sous-population correspondante). Ce terme rend compte de ce que l'on appelle **variance des moyennes** ou **variance interpopulation**. En d'autres termes, il s'agit d'un calcul de dispersion des moyennes $\overline{x}_j$ des **m** sous-populations, par rapport à la moyenne globale $\overline{x}$ de la population **P**.

Remarque : si les moyennes $\overline{x}_j$ de chacune des sous-populations étaient toutes égales, la variance interpopulation serait nulle.

2) Le deuxième terme correspond à une moyenne calculée sur les variances $\sigma_j^2$ de chacune des **m** sous-populations (chaque terme de la somme étant pondéré par les effectifs $n_j$ de la sous-population correspondante). Ce terme rend compte de ce que l'on appelle **moyenne des variances** ou **variance intrapopulation**. En d'autres termes, il s'agit de calculer la dispersion moyenne qui existe entre les variances $\sigma_j^2$ des m sous-populations.

Remarque : si, à l'intérieur de chaque sous-population, chaque valeur était égale à sa moyenne, alors la variance intrapopulation serait nulle.

**Intérêt de cette décomposition de la variance**

De manière générale, lorsqu'on croise deux caractères, dont l'un au moins est quantitatif, la décomposition de la variance sert à montrer dans quelle mesure la dispersion de l'un des deux caractères est expliquée par l'autre caractère.

Dans le cas où l'on raisonne sur une seule population, composée de plusieurs sous-populations, la décomposition de la variance sert à montrer dans quelle mesure la dispersion du caractère étudié au niveau de la population totale est expliquée par la dispersion qui existe entre les sous-populations.

Pour mesurer l'intensité de cette dispersion, on définit un **rapport d'explication**, qu'on peut exprimer en pourcentage, en multipliant le résultat obtenu par 100, selon la formulation :

$$\text{rapport d'explication} = \text{variance interpopulation} / \text{variance totale}$$

$$\text{soit :} \quad \text{Rapport d'explication} = \frac{\frac{1}{n}\sum_{j=1}^{m} n_j \, (\overline{x}_j - \overline{x})^2}{\sigma_x^2}$$

Remarque : pour le calcul du rapport d'explication, on peut se contenter de déterminer les moyennes conditionnelles, qui permettent ensuite de calculer la variance interpopulation.

**Cas général lorsque que le caractère quantitatif est le caractère x** (cf. exemple précédent).

Rappel : les modalités de x sont placées dans la 1ère colonne du tableau.

Le tableau de calculs en vue de déterminer le rapport d'explication est le suivant :

| $x_i$ \ $y_j$ | $y_1$ | … | $y_j$ | … | $y_p$ | $n_{i.}$ | $n_{i.} x_i$ | $n_{i.} x_i^2$ |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | … | $n_{1j}$ | | $n_{1p}$ | $n_{1.}$ | $n_{1.} x_1$ | $n_{1.} x_1^2$ |
| … | … | … | … | … | … | … | … | … |
| $x_i$ | $n_{i1}$ | … | $n_{ij}$ | … | $n_{ip}$ | $n_{i.}$ | $n_{i.} x_i$ | $n_{i.} x_i^2$ |
| … | … | … | … | … | … | … | … | … |
| $x_m$ | $n_{m1}$ | … | $n_{mj}$ | … | $n_{mp}$ | $n_{m.}$ | $n_{m.} x_m$ | $n_{m.} x_m^2$ |
| $N_{.j}$ | $n_{.1}$ | … | $n_{.j}$ | … | $n_{.p}$ | $n_{..}$ | $\sum_i n_{i.} x_i$ | $\sum_i n_{i.} x_i^2$ |
| $\sum_i n_{ij} x_i$ | $\sum_i n_{i1} x_i$ | … | $\sum_i n_{ij} x_i$ | … | $\sum_i n_{ip} x_i$ | | | |
| $\sum_i n_{ij} x_i^2$ | $\sum_i n_{i1} x_i^2$ | … | $\sum_i n_{ij} x_i^2$ | … | $\sum_i n_{ip} x_i^2$ | | | |

Les deux dernières colonnes du tableau servent à calculer les **paramètres marginaux du caractère x** (moyenne, variance, écart-type, coefficient de variation). On a (bien faire attention aux notations utilisées) :

$$\bar{x} = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} x_i \quad \text{et :} \quad \sigma_x^2 = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} (x_i - \bar{x})^2 = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} x_i^2 - \bar{x}^2$$

Les deux dernières lignes du tableau servent à calculer les **p paramètres conditionnels du caractère x** (p moyennes, p variances, p écarts-types, p coefficients de variation). On a :

**Moyennes conditionnelles de x** : $$\bar{x}_j = \frac{1}{n_{.j}} \sum_{j=1}^{m} n_{ij} x_i$$

Il s'agit de la moyenne calculée sur les seuls éléments de la colonne j du tableau. On peut calculer une moyenne conditionnelle de x pour chacune des p colonnes du tableau. C'est pourquoi le caractère quantitatif x possède p moyennes conditionnelles.

**Variances conditionnelles de x** :

$$\sigma^2_{x_j} = \frac{1}{n_{.j}} \sum_{i=1}^{m} n_{ij} (x_i - \bar{x}_j)^2 = \frac{1}{n_{.j}} \sum_{i=1}^{m} n_{ij} x_i^2 - \bar{x}_j^2$$

Il s'agit de la variance calculée sur les seuls éléments de la colonne j du tableau. On peut calculer une variance conditionnelle de x pour chacune des p colonnes du tableau. C'est pourquoi le caractère quantitatif x possède p variances conditionnelles.

Nous venons de voir que la variance marginale de x peut s'écrire sous la forme d'une somme de deux termes. Les différentes expressions ci-dessus permettent le calcul de chacun de ces deux termes (seul le premier terme importe pour calculer le rapport d'explication). Selon les notations du tableau général, nous avons :

$$\sigma^2_x = V(x) = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} (\bar{x}_j - \bar{x})^2 + \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} \sigma^2_{x_j}$$

expression qu'on peut aussi écrire sous la forme :

$$\sigma^2_x = V(x) = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} \bar{x}_j^2 - \bar{x}^2 + \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} \sigma^2_{x_j}$$

Le 1er terme correspond à la variance interpopulation, ou **variance des moyennes conditionnelles de x**, ou encore variance expliquée de x ($VE_x$). C'est la **part de la variance de x qui est expliquée par le caractère y**.

Le 2ème terme correspond à la variance intrapopulation, ou moyenne des variances conditionnelles de x, ou encore variance résiduelle de x ($VR_x$). C'est la **part de la variance de x qui n'est pas expliquée par le caractère y**.

Pour connaître l'intensité de la liaison qui existe entre le caractère y et le caractère x, c-à-d la **quantité d'information supplémentaire apportée par le caractère y à la connaissance des variations du caractère x**, on calcule le **rapport d'explication** :

$$\text{Rapport d'explication} = \frac{\text{variance des moyennes conditionnelles de x}}{\text{variance marginale de x}}$$

$$\text{Rapport d'explication} = \frac{VE_x}{\sigma^2_x} = \frac{\frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} \bar{x}_j^2 - \bar{x}^2}{\sigma^2_x}$$

### 42. Application à l'exemple

Caractère x : nombre d'enfants de 0 à 16 ans. Caractère y : nationalité.

On se propose de calculer le rapport d'explication relatif au caractère quantitatif x, en vue de savoir dans quelle mesure le caractère qualitatif y (la nationalité de la famille) exerce une influence sur le nombre d'enfants (de 0 à 16 ans) de la famille. On a le tableau suivant :

| $x_i$ \ $y_j$ | Nationalité française | Nationalité étrangère | $n_{i.}$ | $n_{i.} x_i$ | $n_{i.} x_i^2$ |
|---|---|---|---|---|---|
| **0** | 6 815 | 316 | 7 131 | 0 | 0 |
| **1** | 3 005 | 196 | 3 201 | 3 201 | 3 201 |
| **2** | 2 320 | 179 | 2 499 | 4 998 | 9 996 |
| **3** | 814 | 105 | 919 | 2 757 | 8 271 |
| **4** | 264 | 107 | 371 | 1 484 | 5 936 |
| $n_{.j}$ | 13 218 | 903 | 14 121 | 12 440 | 27 404 |
| $\sum_i n_{ij} x_i$ | 11 143 | 1 297 | (12 440) | | |
| $\sum_i n_{ij} x_i^2$ | 23 835 | 3 569 | (27 404) | | |

Les deux dernières colonnes du tableau permettent de calculer les **paramètres marginaux du caractère x**. On obtient des paramètres qui concernent globalement l'ensemble des familles résidant en France :

$$\bar{x} = \frac{12\,440}{14\,121} = 0{,}88 \text{ enfant par famille}$$

$$\sigma_x^2 = \frac{27\,404}{14\,121} - 0{,}88^2 = 1{,}16 \text{ enfant}^2 \text{ par famille}, \quad \sigma_x = 1{,}08 \text{ enfant par famille}$$

$CV_x = 1{,}08 / 0{,}88 = 1{,}22$. La valeur du coefficient de variation, très élevée (supérieure à la valeur de la moyenne), caractérise une très forte dispersion du nombre d'enfants par famille autour de la valeur moyenne.

**On peut aussi raisonner selon chacune des deux colonnes du tableau. Cela revient à considérer indépendamment les deux sous-populations** (les résidents français et les résidents étrangers).

Sur chacune d'entre elles, on peut calculer les mêmes paramètres qui ci-dessus : il s'agit des **paramètres conditionnels de x**, qui sont indicés, chacun, selon le numéro de la colonne concernée.

- colonne 1 : paramètres relatifs aux familles de nationalité française :

$$\bar{x}_1 = \frac{11143}{13\,218} = 0{,}84 \text{ enfant par famille}$$

$\sigma^2_{x1} = \frac{23\,835}{13\,218} - 0{,}84^2 = 1{,}09$ enfant$^2$ par famille , $\sigma_{x1} = 1{,}05$ enfant par famille

$CV_{x1} = 1{,}05 / 0{,}84 = 1{,}37$. La valeur du coefficient de variation, très élevée (supérieure à la valeur de la moyenne), caractérise une très forte dispersion du nombre d'enfants par famille autour de la valeur moyenne.

- colonne 2 : paramètres relatifs aux familles de nationalité étrangère :

$\bar{x}_2 = \frac{1297}{903} = 1{,}44$ enfant par famille

$\sigma^2_{x2} = \frac{3\,569}{903} - 1{,}44^2 = 1{,}89$ enfant$^2$ par famille , $\sigma_{x2} = 1{,}37$ enfant par famille

$CV_{x2} = 1{,}37 / 1{,}44 = 0{,}96$. La valeur du coefficient de variation, très élevée, caractérise une très forte dispersion du nombre d'enfants par famille autour de la valeur moyenne.

Remarque : la dispersion du nombre d'enfants par famille est un peu moins élevée pour les familles étrangères résidant en France en 1982 que pour les familles françaises.

Ces différentes valeurs permettent de calculer la **variance des moyennes conditionnelles de x** (= variance interpopulation), selon l'expression vue plus haut. On obtient ici :

$$\text{variance des moyennes conditionnelles de } x = VE_x$$
$$= \frac{1}{14\,121}\left[13\,218\ (0{,}84 - 0{,}88)^2 + 903\ (1{,}44 - 0{,}88)^2\right]$$
$$= \frac{1}{14\,121}\left(13\,218 \times 0{,}84^2 + 903 \times 1{,}44^2\right) - 0{,}88^2 = 0{,}02$$

Enfin, on calcule le **rapport d'explication** :

$$\text{Rapport d'explication} = VEx / \sigma^2_x = 0{,}02 / 1{,}16 = \mathbf{0{,}018}, \text{ soit } 1{,}8\ \% .$$

Ce résultat signifie que, pour les familles résidant en France en 1982, la nationalité des familles n'explique que 2 % environ de la dispersion relative au nombre d'enfants par famille. 98 % de celle-ci s'explique par d'autres facteurs, tels que le mode de vie, le niveau de vie, la religion, etc.

Remarque : selon la formule vue plus haut, on peut également calculer la moyenne des variances conditionnelles de x (= variance intrapopulation).

On a :

$$\sigma_x^2 = V(x) = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} \bar{x}_j^2 - \bar{x}^2 + \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} \sigma_{x_j}^2$$

et :

$$VR_x = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} \sigma_{x_j}^2 = \frac{1}{14121} (13218 \times 1{,}09 + 903 \times 1{,}89) = 1{,}14$$

En ajoutant cette valeur à VEx = 0,02 , on peut vérifier qu'on retrouve bien la valeur de la variance marginale de x.

**Conclusion générale**

Avec la méthode des tableaux d'indépendance, on doit se contenter d'apprécier plus ou moins subjectivement les sur et sous-représentation observables, case par case, ligne par ligne ou colonne par colonne.

Avec le calcul du rapport d'explication, on peut quantifier et donc rendre plus objectif le diagnostic porté sur le croisement des deux caractères x et y.

Sur notre exemple, avec le rapport d'explication, on se rend mieux compte du caractère négligeable de l'impact de la nationalité (le caractère y) sur le nombre d'enfants par famille (le caractère x).

**Exemple 2** :

**Ici, les modalités du caractère qualitatif sont placées en lignes et celles du caractère quantitatif sont placées en colonnes**.

On considère les familles résidant en France (en milliers) au moment du recensement de 1982, selon les deux caractères suivants :

- caractère x (en lignes) : le **statut de la famille** (**caractère qualitatif**).

- caractère y (en colonnes) : le **nombre d'enfants de 0 à 16 ans** (**caractère quantitatif**).

| x i \ y j | 0 | 1 | 2 | 3 | 4 et + |
|---|---|---|---|---|---|
| **couples** | 6.844 | 2.830 | 2.341 | 869 | 348 |
| **mono parents** | 287 | 371 | 158 | 50 | 23 |

**1°)** A partir du tableau des effectifs conjoints ci-dessus, réaliser un graphique de la **distribution du nombre d'enfants de 0 à 16 ans des familles, selon le statut de la famille**. Pour quelle raison ce graphique n'est-il pas vraiment approprié, pour rendre compte correctement de cette série à deux caractères ?

**2°)** Pour lever la difficulté mise en évidence à la question 1, construire le tableau des fréquences adéquates.

**3°)** De même, toujours à partir du tableau des effectifs conjoints, réaliser un graphique approprié de la **distribution du statut de la famille, selon le nombre d'enfants de 0 à 16 ans**. On procédera comme aux questions 1 et 2.

**4°)** Après en avoir rappelé le principe de construction, établir le tableau des effectifs conjoints théoriques d'indépendance, de même que celui des fréquences conjointes théoriques d'indépendance.

**5°)** Construire le tableau des fréquences conditionnelles théoriques d'indépendance en lignes, puis celui des fréquences conditionnelles théoriques d'indépendance en colonnes. Constater et expliquer.

**6°)** A partir de la comparaison du tableau des fréquences conjointes réelles et de celui des fréquences conjointes théoriques d'indépendance, analyser, selon deux méthodes différentes, la dépendance ou l'indépendance des deux caractères x et y.

**7°)** Calculer moyenne, écart-type et coefficient de variation du nombre d'enfants par famille.

**8°)** Procéder de même pour chacun des statuts familiaux. Commenter les résultats.

Population : le nombre de familles résidant en France en 1982 (en milliers).

Caractère x : le statut des familles. Caractère qualitatif.

Caractère y : le nombre d'enfants de 0 à 16 ans des familles. Caractère quantitatif discret.

**Effectifs conjoints réels**

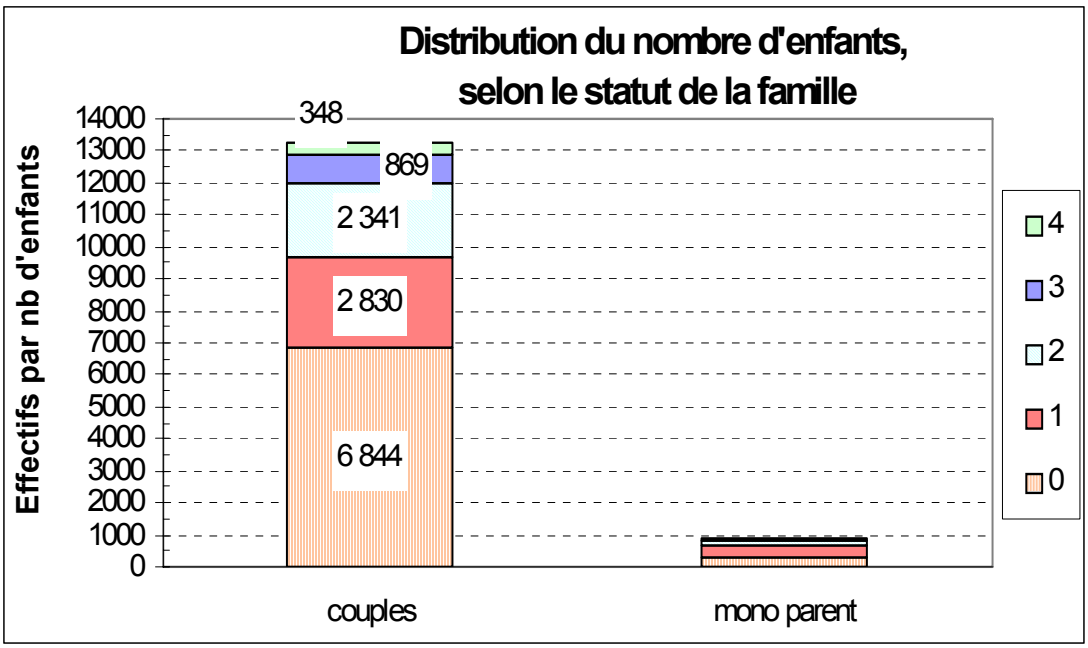
| y j / x i \ | 0 | 1 | 2 | 3 | 4 | Total (n i .) |
|---|---|---|---|---|---|---|
| **couples** | 6.844 | 2.830 | 2.341 | 869 | 348 | 13.232 |
| **mono parent** | 287 | 371 | 158 | 50 | 23 | 889 |
| **Total (n .j)** | 7.131 | 3.201 | 2.499 | 919 | 371 | 14.121 |

**Question 1**

On utilise un **diagramme en tuyaux d'orgue à base variable**.

En abscisse, on trouve les modalités du caractère **qualitatif**, ici le statut de la famille.

En ordonnée, on a les effectifs correspondant à chaque nombre d'enfants, c-à-d correspondant à chaque modalité du caractère quantitatif y.




Bien que correct, le graphique précédent, selon les effectifs, n'est pas approprié, car sa lecture ne permet pas, pour un statut familial donné, une comparaison précise des **proportions** du nombre d'enfants. Ici, de plus, le déséquilibre numérique entre les 2 modalités du caractère "Statut familial" invalide toute tentative de lecture correcte.

**Question 2**

Pour rendre correctement compte des deux distributions simultanément, il est nécessaire de faire apparaître sur le graphique, en abscisse, la proportion relative des deux statuts, c-à-d les fréquences marginales de x (f i.). On met en œuvre la technique des **tuyaux d'orgue à base variable**. C'est pourquoi on construit ici le tableau des fréquences conjointes.

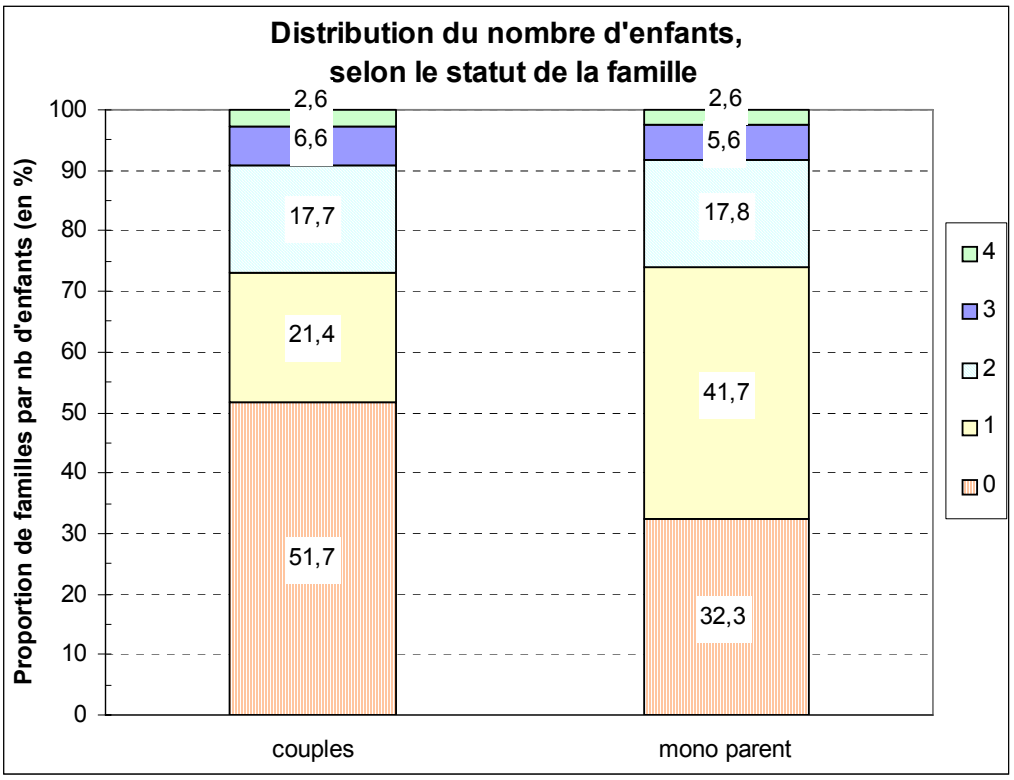
**Fréquences conjointes réelles (en %)**

| y j / x i \ | 0 | 1 | 2 | 3 | 4 | Total (f i .) |
|---|---|---|---|---|---|---|
| **Couples** | 48,5 | 20,0 | 16,6 | 6,2 | 2,5 | 93,7 |
| **mono parent** | 2,0 | 2,6 | 1,1 | 0,4 | 0,2 | 6,3 |
| **Total f .j** | 50,5 | 22,7 | 17,7 | 6,5 | 2,6 | 100,0 |

Pour chaque statut familial, on cherche à faire apparaître les proportions de nombre d'enfants. C'est pourquoi le tableau adéquat est celui des **fréquences conditionnelles en lignes** (c-à-d les **fréquences conditionnelles de y** : statut de la famille). Les valeurs sont données en % :

**Fréquences conditionnelles en lignes (en %)**

| y j / x i \ | 0 | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|---|
| **couples** | 51,7 | 21,4 | 17,7 | 6,6 | 2,6 | 100,0 |
| **mono parent** | 32,3 | 41,7 | 17,8 | 5,6 | 2,6 | 100,0 |
| **Total f .j** | 50,5 | 22,7 | 17,7 | 6,5 | 2,6 | 100,0 |




**Attention** : les tuyaux d'orgue ci-dessus, réalisés avec un tableur, ne sont pas corrects en abscisse, car **leurs bases ne sont pas variables**, **en fonction du poids de chaque statut (93,7 %, 6,3 %)**.

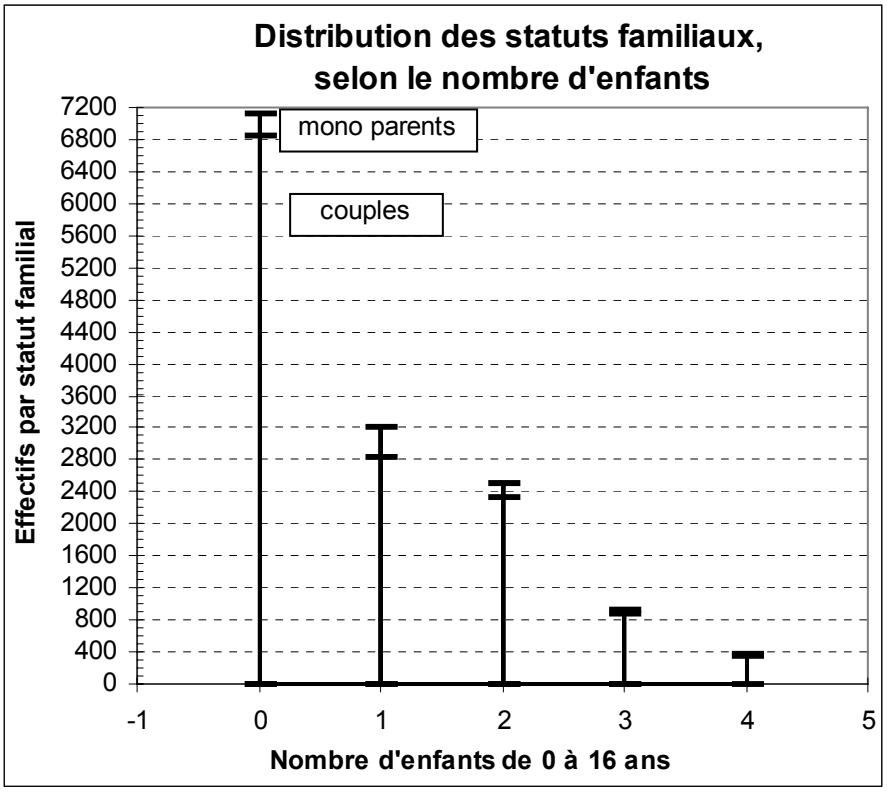
**Remarque terminale** : ici, le caractère qualitatif ne présente que 2 modalités. On aurait également pu utiliser la technique des **secteurs semi-circulaires** (cf. ch. 2). Toutefois, la disproportion des effectifs de chaque modalité est ici préjudiciable à une bonne lecture du demi-cercle "Familles monoparentales".

## Question 3

On utilise un **diagramme en bâtons segmentés**.

En abscisse, on trouve les modalités du caractère **quantitatif**, ici le nombre d'enfants de 0 à 16 ans.

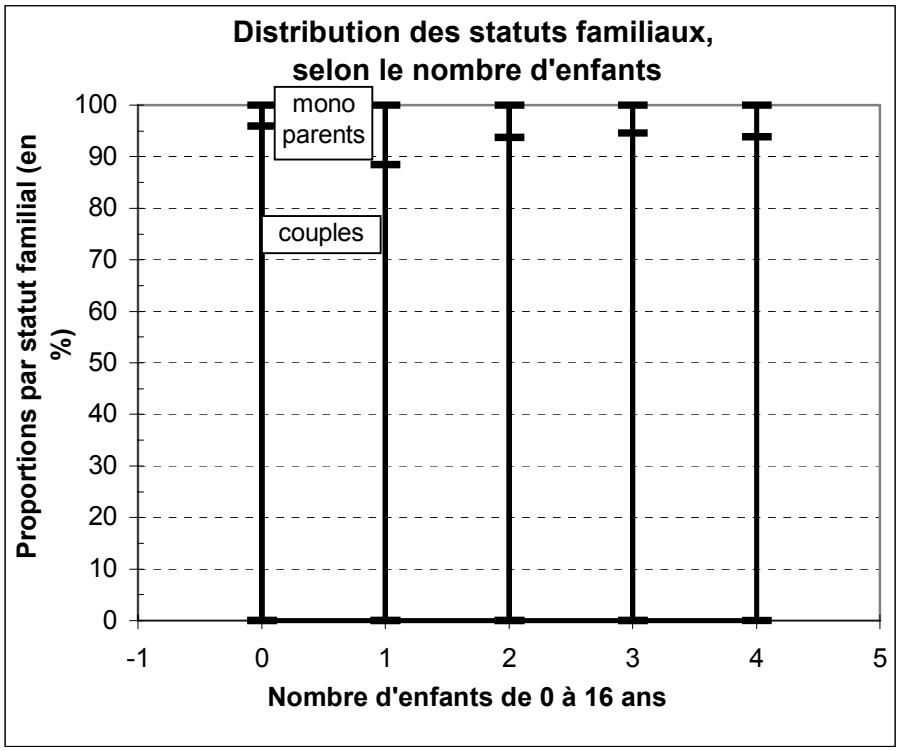
En ordonnée, on a les effectifs correspondant à chaque nombre d'enfants. De plus, on ajoute des marques (tirets, segments) qui vont matérialiser les effectifs correspondant à chaque modalité du caractère **qualitatif**.




Bien que correct, ce graphique, selon les effectifs, n'est pas approprié, car sa lecture ne permet pas, pour un nombre donné d'enfants, une comparaison précise des **proportions** entre les statuts familiaux.

Pour chaque nombre d'enfants, on cherche à faire apparaître les proportions selon le statut familial. C'est pourquoi le tableau adéquat est celui des **fréquences conditionnelles en colonnes** (c-à-d les **fréquences conditionnelles de x** : nombre d'enfants de 0 à 16 ans). Les valeurs sont données en % :

| **y j**<br>**x i ** | **0** | **1** | **2** | **3** | **4** | **Total f i .** |
|---|---|---|---|---|---|---|
| **couples** | 96,0 | 88,4 | 93,7 | 94,6 | 93,8 | 93,7 |
| **mono parent** | 4,0 | 11,6 | 6,3 | 5,4 | 6,2 | 6,3 |
| **Total** | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |




**Question 4**

**Rappel** : pour construire un tableau des **effectifs conjoints théoriques d'indépendance**, on utilise la formule : **n ij = (n i. x n .j) / n ..**

Comme il s'agit d'effectifs qui portent sur des nombres entiers, **on arrondit** les résultats obtenus, pour chaque case du tableau, **à l'entier le plus proche**.

**Effectifs conjoints théoriques d'indépendance (valeurs arrondies)**

| y j / x i \ | 0 | 1 | 2 | 3 | 4 | Total n i . |
|---|---|---|---|---|---|---|
| **couples** | 6.682 | 2.999 | 2.342 | 861 | 348 | 13.232 |
| **mono parent** | 449 | 202 | 157 | 58 | 23 | 889 |
| **Total n .j** | 7.131 | 3.201 | 2.499 | 919 | 371 | 14.121 |

**Rappel** : pour construire un tableau des fréquences conjointes théoriques d'indépendance (exprimées en pourcentage), on utilise la formulation : **f ij = (f i. x f .j) / 100**

**Fréquences conjointes théoriques d'indépendance (en %)**

| y j / x i \ | 0 | 1 | 2 | 3 | 4 | Total f i . |
|---|---|---|---|---|---|---|
| **couples** | 47,3 | 21,2 | 16,6 | 6,1 | 2,5 | 93,7 |
| **mono parent** | 3,2 | 1,4 | 1,1 | 0,4 | 0,2 | 6,3 |
| **Total f .j** | 50,5 | 22,7 | 17,7 | 6,5 | 2,6 | 100,0 |

**Question 5**

**Fréquences conditionnelles théoriques d'indépendance en lignes (en %) (caract. y)**

| y j / x i \ | 0 | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|---|
| couples | 50,5 | 22,7 | 17,7 | 6,5 | 2,6 | 100,0 |
| mono parent | 50,5 | 22,7 | 17,7 | 6,5 | 2,6 | 100,0 |
| Total f .j | 50,5 | 22,7 | 17,7 | 6,5 | 2,6 | 100,0 |

**Fréquences conditionnelles théoriques d'indépendance en colonnes (en %) (caract. x)**

| y j / x i \ | 0 | 1 | 2 | 3 | 4 | Total f i . |
|---|---|---|---|---|---|---|
| couples | 93,7 | 93,7 | 93,7 | 93,7 | 93,7 | 93,7 |
| mono parent | 6,3 | 6,3 | 6,3 | 6,3 | 6,3 | 6,3 |
| Total | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |

**Interprétations** :

Dans le tableau des fréquences conditionnelles théoriques d'indépendance **en lignes**, on constate que les pourcentages sont égaux pour chaque ligne et correspondent à ceux des fréquences marginales f .j . Cela signifie que le nombre d'enfants (y) est indépendant du statut familial (x). Pour chaque statut, la distribution des nombres d'enfants est la même.

Dans le tableau des fréquences conditionnelles théoriques d'indépendance **en colonnes**, on constate que les pourcentages sont égaux pour chaque colonne et correspondent à ceux des fréquences marginales f i. Cela signifie que le statut familial (x) est indépendant du nombre d'enfants (y). Pour chaque nombre d'enfants, on a la même proportion de familles monoparentales ou en couples.

**Question 6**

**Analyse de l'indépendance des deux caractères : méthode 1**

On peut comparer 2 tableaux :

- soit celui des effectifs conjoints réels et celui des effectifs conjoints théoriques d'indépendance ;

- soit celui des fréquences conjointes réelles et celui des fréquences conjointes théoriques d'indépendance.

En général, la 2ème possibilité est la plus pratique, notamment si les effectifs sont élevés et / ou si les écarts entre les 2 tableaux sont importants. Nous faisons ce choix ici pour traiter notre exemple.

**a)** Considérons les tableaux **en lignes** : si le nombre d'enfants (y) était totalement indépendant du statut familial (x), on devrait avoir la même proportion d'enfants pour chaque type de statut familial (cf. 5°), le tableau des fréquences conditionnelles théoriques d'indépendance en lignes).

De fait, en comparant les 2 tableaux de fréquences conjointes, on observe que, pour les couples comme pour les familles monoparentales, les différences entre les fréquences conjointes réelles et les fréquences conjointes d'indépendance sont **minimes et donc peu significatives, quel que soit le nombre d'enfants considéré**.

Légère surreprésentation réelle des couples de 0 enfant (48,5 % contre 47,3 %) et légère sous-représentation réelle des couples de 1 enfant (20,0 % contre 21,2 %). Inversement, légère sous-représentation réelle des monoparents de 0 enfant (2,0 % contre 3,2 %) et légère surreprésentation réelle des monoparents de 1 enfant (2,6 % contre 1,4 %).

**b)** Considérons les tableaux **en colonnes** : si le statut familial (x) était totalement indépendant du nombre d'enfants (y), quel que soit ce nombre, on devrait avoir les mêmes proportions de monoparents et de couples (cf. 5°), le tableau des fréquences conditionnelles théoriques d'indépendance en colonnes).

De fait, en comparant les 2 tableaux de fréquences conjointes, on observe que, **pour chaque nombre d'enfants, les écarts de pourcentages sont minimes, tant pour les couples que pour les familles monoparentales**. Cela induit une quasi indépendance entre les 2 caractères.

Pour 0 enfant, légère surreprésentation réelle des couples (48,5 % contre 47,3 %) et légère sous-représentation réelle des monoparents (2,0 % contre 3,2 %). Pour 1 enfant, légère sous-représentation réelle des couples (20,0 % contre 21,2 %)et légère surreprésentation réelle des monoparents (2,6 % contre 1,4 %).

## Analyse de l'indépendance des deux caractères : méthode 2

On compare deux tableaux :

- le **tableau A**, qui indique, dans chacune de ses cases, le rapport des effectifs réels aux effectifs théoriques d'indépendance (ramené à un pourcentage) ;

- le **tableau B**, qui est celui des fréquences conjointes réelles.

Dans le tableau A, la valeur de chaque case résulte du calcul suivant : $\left( \frac{n_{ij}\ \text{réel}}{n_{ij}\ \text{théorique}} - 1 \right) \times 100$ .

Si la valeur du résultat est positive, cette opération permet d'exprimer, en pourcentage, de combien l'effectif réel excède l'effectif théorique d'indépendance : cela donne un ordre de grandeur de la surreprésentation d'une case du tableau des effectifs réels. Inversement, si la valeur est négative, on a une idée de la sous-représentation d'une case du tableau des effectifs réels par rapport à une hypothèse d'indépendance.

**Tableau A : rapports des effectifs conjoints réels aux effectifs conjoints théoriques (en %)**

| X i \ y j | 0 | 1 | 2 | 3 | 4 | ( [ n i. / n i. ] - 1) x 100 |
|---|---|---|---|---|---|---|
| **couples** | **2,4** | -5,6 | 0,0 | 0,9 | 0,0 | 0,0 |
| **mono parent** | -36,1 | 83,7 | 0,6 | -13,8 | 0,0 | 0,0 |
| **( [ n .j / n .j ] - 1) x 100** | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |

Exemple de calcul, dans la 1ère case du tableau A : $\left( \frac{6\,844}{6\,682} - 1 \right) \times 100 = + 2{,}4\,\%$

Cf. ex. 1 : il faut relativiser la valeur des pourcentages obtenus (positifs ou négatifs) en remarquant que chacun d'entre eux renvoie à des effectifs absolus réels très différents. C'est pourquoi il est indispensable de joindre à ce tableau celui des fréquences conjointes réelles (tableau B), de façon à pouvoir apprécier le poids (importance relative) de chacune des cases du tableau A :

**Tableau B : tableau des fréquences conjointes réelles**

| X i \ y j | 0 | 1 | 2 | 3 | 4 | Total (f i .) |
|---|---|---|---|---|---|---|
| **Couples** | 48,5 | 20,0 | 16,6 | 6,2 | 2,5 | 93,7 |
| **mono parent** | 2,0 | 2,6 | 1,1 | 0,4 | 0,2 | 6,3 |
| **Total f .j** | 50,5 | 22,7 | 17,7 | 6,5 | 2,6 | 100,0 |

Si l'on considère les seules fréquences conjointes réelles qui dépassent 10% (ces **fréquences correspondent aux effectifs les plus significatifs dans la réalité**, soit ici **85,1 %** de l'effectif total), on constate que les pourcentages correspondants du tableau A font apparaître des pourcentages extrêmement faibles. Dans ce dernier, le plus fort pourcentage observé (83,7 %) correspond, pour la même case du tableau B, à l'un des pourcentages les plus faibles (2,6 %).

Il n'y a d'ailleurs pas de pourcentages très élevés dans le tableau A, et ils correspondent à des valeurs tout à fait négligeables dans le tableau B.

**Conclusion générale**

Les constats précédents permettent d'avancer que les deux caractères sont très largement indépendants : **le nombre d'enfants ne dépend pas du statut familial**.

La dispersion du nombre d'enfants par famille s'explique par d'autres facteurs que le statut familial, notamment le mode de vie, le niveau de vie, la religion, etc.

Ce constat est confirmé par la valeur du rapport d'explication, égale à 0,02 % (voir plus loin) .

**Question 7**

Ici, on raisonne **globalement sur l'ensemble de la population**, sans plus s'occuper du statut de la famille (caractère x). Cependant, on reste dans la logique d'un croisement entre deux caractères. Par conséquent, les notations littérales vont être modifiées, par rapport aux séries à un caractère, en conservant deux indices, de la façon suivante :

**Moyenne marginale de y** : $$\bar{y} = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\, y_j$$

**Variance marginale de y** : $$\sigma_y^2 = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\,(y_j - \bar{y})^2 = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\, y_j^2 - \bar{y}^2$$

**Coefficient de variation de y** : $$CV_y = \frac{\sigma_y}{\bar{y}}$$

**Remarque importante : on doit formuler une hypothèse pour remplacer la modalité "4 et plus" par une modalité mesurable. Ici, on retient la valeur "4".**

| y j / x i \ | 0 | 1 | 2 | 3 | 4 | n i . |
|---|---|---|---|---|---|---|
| **couples** | 6.844 | 2.830 | 2.341 | 869 | 348 | 13.232 |
| **mono parent** | 287 | 371 | 158 | 50 | 23 | 889 |
| **n . j** | 7.131 | 3.201 | 2.499 | 919 | 371 | 14.121 |
| **n . j yj** | 0 | 3.201 | 4.998 | 2.757 | 1.484 | 12.440 |
| **n . j yj²** | 0 | 3.201 | 9.996 | 8.271 | 5.936 | 27.404 |

Pour l'ensemble de la population, la moyenne, l'écart-type et le coefficient de variation du nombre d'enfants par famille sont de :

**Moyenne = 0,88 enfant par famille**
**Variance = 1,16 enfant² par famille**
**Ecart-type = 1,02 enfant par famille**
**Coefficient de variation = 1,22**

Valeur très élevée du CV, qui caractérise une très forte dispersion autour de la valeur moyenne observée (écart-type supérieur à la moyenne !!).

**Remarque : les valeurs ci-dessus sont évidemment les mêmes que dans le 1er exemple, puisque l'on travaille sur la même population !**

**Question 8**

Ici, on raisonne successivement sur chaque ligne du tableau, c-à-d sur des **sous-populations, par statut familial**.

Comme précédemment, on reste dans la logique d'un croisement entre deux caractères. Par conséquent, les notations littérales comportent ici encore deux indices et renvoient à une sous-population particulière, de la façon suivante :

**Moyennes conditionnelles de y** : $$\bar{y}_i = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij}\, y_j$$

Il y a autant de moyennes conditionnelles de y qu'il y a de modalités de x.

**Variances conditionnelles de y** :

$$\sigma^2_{y_i} = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij}\,(y_j - \bar{y}_i)^2 = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij}\, y_j^2 - \bar{y}_i^{\,2}$$

Il y a autant de variance conditionnelles de y qu'il y a de modalités de x.

**Coefficient de variation de y i** : $$CV_{y_i} = \frac{\sigma_{y_i}}{\bar{y}_i}$$

**Première méthode** :

**Couples** (on raisonne sur la seule première ligne du tableau)

| $X_i$ \ $y_j$ | 0 | 1 | 2 | 3 | 4 | $n_{1.}$ |
|---|---|---|---|---|---|---|
| **couples** | 6.844 | 2.830 | 2.341 | 869 | 348 | 13.232 |
| $n_{1j}\ y_j$ | 0 | 2.830 | 4.682 | 2.607 | 1.392 | 11.511 |
| $n_{1j}\ y_j^2$ | 0 | 2.830 | 9.364 | 7.821 | 5.568 | 25.583 |

**Moyenne = 0,87 enfant par famille**

**Variance = 1,18 enfant² par famille**

**Ecart-type = 1,08 enfant par famille**

**Coefficient de variation = 1,25**

Valeur très élevée du CV, qui caractérise une très forte dispersion autour de la valeur moyenne observée (écart-type supérieur à la moyenne !!).

**Familles mono-parentales** (on raisonne sur la seule deuxième ligne du tableau)

| $x_i$ \ $y_j$ | 0 | 1 | 2 | 3 | 4 | $n_{2.}$ |
|---|---|---|---|---|---|---|
| **mono parent** | 287 | 371 | 158 | 50 | 23 | 889 |
| $n_{2j}\ y_j$ | 0 | 371 | 316 | 150 | 92 | 929 |
| $n_{2j}\ y_j^2$ | 0 | 371 | 632 | 450 | 368 | 1.821 |

**Moyenne = 1,04 enfant par famille**

**Variance = 0,96 enfant² par famille**

**Ecart-type = 0,98 enfant par famille**

**Coefficient de variation = 0,94**

Valeur très élevée du CV, qui caractérise une très forte dispersion autour de la valeur moyenne observée.

**Remarque terminale** : la dispersion du nombre d'enfants par famille est un peu moins élevée pour les familles monoparentales résidant en France en 1982.

**Deuxième méthode** :

Usuellement, on travaille directement sur le tableau à double entrée, en ajoutant deux colonnes supplémentaires.

| **x i \ y j** | **0** | **1** | **2** | **3** | **4** | **n i .** | $\sum_j n_{ij} y_j$ | $\sum_j n_{ij} y_j^2$ |
|---|---|---|---|---|---|---|---|---|
| **couples** | 6 844 | 2 830 | 2 341 | 869 | 348 | 13 232 | 11 511 | 25 583 |
| **mono parents** | 287 | 371 | 158 | 50 | 23 | 889 | 929 | 1 821 |
| **n . j** | 7 131 | 3 201 | 2 499 | 919 | 371 | 14 121 | (12 440) | (27 404) |
| **n . j yj** | 0 | 3 201 | 4 998 | 2 757 | 1 484 | 12 440 | | |
| **n . j yj²** | 0 | 3 201 | 9 996 | 8 271 | 5 936 | 27 404 | | |

Dans les deux dernières colonnes du tableau, on remarque que l'on enregistre directement les sommes, sans détailler leurs calculs comme dans la première méthode.

**Calcul du rapport d'explication**

**Cas général lorsque que le caractère quantitatif est le caractère y** (cf. exemple précédent).

Rappel : les modalités de y sont placées dans la 1ère ligne du tableau.

| $x_i$ \ $y_j$ | $y_1$ | … | $y_j$ | … | $y_p$ | $n_{i.}$ | $\sum_j n_{ij} y_j$ | $\sum_j n_{ij} y_j^2$ |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | … | $n_{1j}$ | | $n_{1p}$ | $n_{1.}$ | $\sum_j n_{1j} y_j$ | $\sum_j n_{1j} y_j^2$ |
| … | … | … | … | … | … | … | … | … |
| $x_i$ | $n_{i1}$ | … | $n_{ij}$ | … | $n_{ip}$ | $n_{i.}$ | $\sum_j n_{ij} y_j$ | $\sum_j n_{ij} y_j^2$ |
| … | … | … | … | … | … | … | … | … |
| $x_m$ | $n_{m1}$ | … | $n_{mj}$ | … | $n_{mp}$ | $n_{m.}$ | $\sum_j n_{mj} y_j$ | $\sum_j n_{mj} y_j^2$ |
| $n_{.j}$ | $n_{.1}$ | … | $n_{.j}$ | … | $n_{.p}$ | $n_{..}$ | | |
| $n_{.j} y_j$ | $n_{.1} y_1$ | … | $n_{.j} y_j$ | … | $n_{.p} y_p$ | $\sum_j n_{.j} y_j$ | | |
| $n_{.j} y_j^2$ | $n_{.1} y_1^2$ | … | $n_{.j} y_j^2$ | … | $n_{.p} y_p^2$ | $\sum_j n_{.j} y_j^2$ | | |

Les deux dernières lignes du tableau servent à calculer les **paramètres marginaux du caractère y** (moyenne, variance, écart-type, coefficient de variation). On a (bien faire attention aux notations utilisées) :

$$\bar{y} = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\, y_j \quad \text{et :} \quad \sigma_y^2 = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\, (y_j - \bar{y})^2 = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\, y_j^2 - \bar{y}^2$$

Les deux dernières colonnes du tableau servent respectivement à calculer les **m paramètres conditionnels du caractère y** (m moyennes, m variances, m écarts-types, m coefficients de variation). On a :

**Moyennes conditionnelles de y** : $$\bar{y}_i = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij}\, y_j$$

Il s'agit de la moyenne calculée sur les seuls éléments de la ligne i du tableau. On peut calculer une moyenne conditionnelle de y pour chacune des m lignes du tableau. C'est pourquoi le caractère quantitatif y possède m moyennes conditionnelles.

**Variances conditionnelles de y** :

$$\sigma_{y_i}^2 = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij}\, (y_j - \bar{y}_i)^2 = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij}\, y_j^2 - \bar{y}_i^2$$

Il s'agit de la variance calculée sur les seuls éléments de la ligne i du tableau. On peut calculer une variance conditionnelle de y pour chacune des m lignes du tableau. C'est pourquoi le caractère quantitatif y possède m variances conditionnelles.

On peut évidemment transposer ici au caractère y ce que nous avons vu précédemment pour le caractère x : la variance marginale de y peut s'écrire sous la forme d'une somme de deux termes et les différentes expressions ci-dessus permettent le calcul de chacun de ces deux termes (seul le premier terme va importer ici pour le calcul du rapport d'explication). Selon les notations du tableau général, nous avons :

$$\sigma_y^2 = V(y) = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\, (\bar{y}_i - \bar{y})^2 + \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\, \sigma_{y_i}^2$$

expression qu'on peut aussi écrire sous la forme :

$$\sigma_y^2 = V(y) = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\, \bar{y}_i^2 - \bar{y}^2 + \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\, \sigma_{y_i}^2$$

Le 1er terme correspond à la variance interpopulation, ou **variance des moyennes conditionnelles de y**, ou encore variance expliquée de y ($VE_y$). C'est la **part de la variance de y qui est expliquée par le caractère x**.

Le 2ème terme correspond à la variance intrapopulation, ou moyenne des variances conditionnelles de y, ou encore variance résiduelle de y ($VR_y$). C'est la **part de la variance de y qui n'est pas expliquée par le caractère x**.

Pour connaître l'intensité de la liaison qui existe entre le caractère x et le caractère y, c-à-d la quantité d'information supplémentaire apportée par le caractère x à la connaissance des variations du caractère y, on calcule le **rapport d'explication** :

$$\text{Rapport d'explication} = \frac{VE_y}{\sigma_y^2} = \frac{\frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} \bar{y}_i^2 - \bar{y}^2}{\sigma_y^2}$$

**Application à l'exemple**

Caractère x : type de famille Caractère y : nombre d'enfants de 0 à 16 ans.

On se propose de calculer le rapport d'explication relatif au caractère quantitatif y, en vue de savoir dans quelle mesure le caractère qualitatif x (le type de famille) exerce une influence sur le nombre d'enfants (de 0 à 16 ans) de la famille (caractère y). On a le tableau suivant :

| **x i \ y j** | **0** | **1** | **2** | **3** | **4** | **n i .** | $\sum_j n_{ij} y_j$ | $\sum_j n_{ij} y_j^2$ |
|---|---|---|---|---|---|---|---|---|
| **couples** | 6 844 | 2 830 | 2 341 | 869 | 348 | 13 232 | 11 511 | 25 583 |
| **mono parents** | 287 | 371 | 158 | 50 | 23 | 889 | 929 | 1 821 |
| **n . j** | 7 131 | 3 201 | 2 499 | 919 | 371 | 14 121 | (12 440) | (27 404) |
| **n . j yj** | 0 | 3 201 | 4 998 | 2 757 | 1 484 | 12 440 | | |
| **n . j yj²** | 0 | 3 201 | 9 996 | 8 271 | 5 936 | 27 404 | | |

Les 2 dernières lignes du tableau permettent calculer les **paramètres marginaux du caractère y**. On obtient des paramètres qui concernent globalement l'ensemble des familles résidant en France les valeurs sont évidemment les mêmes que dans l'exemple précédent, puisque l'on travaille sur la même population) :

$$\bar{y} = \frac{12\,440}{14\,121} = 0{,}88 \text{ enfant par famille}$$

$$\sigma_y^2 = \frac{27\,404}{14\,121} - 0{,}88^2 = 1{,}16 \text{ enfant}^2 \text{ par famille} \quad , \quad \sigma_y = 1{,}08 \text{ enfant par famille}$$

$CV_y = 1{,}08 / 0{,}88 = 1{,}22$. La valeur du coefficient de variation, très élevée (supérieure à la valeur de la moyenne), caractérise une très forte dispersion du nombre d'enfants par famille autour de la valeur moyenne.

**On peut aussi raisonner selon chacune des deux lignes du tableau. Cela revient à considérer indépendamment les deux sous-populations** (les couples et les familles mono parentales).

Sur chacune d'entre elles, on peut calculer les mêmes paramètres qui ci-dessus : il s'agit des **paramètres conditionnels de y**, qui sont indicés, chacun, selon le numéro de la ligne concernée.

- ligne 1 : paramètres relatifs aux familles constituées de couples :

$$\bar{y}_1 = \frac{11\,511}{13\,232} = 0{,}87 \text{ enfant par famille}$$

$$\sigma^2_{y1} = \frac{25\,583}{13\,232} - 0{,}87^2 = 1{,}18 \text{ enfant}^2 \text{ par famille} \quad , \quad \sigma_{y1} = 1{,}08 \text{ enfant par famille}$$

$CV_{y1} = 1{,}08 / 0{,}87 = 1{,}25$. La valeur du coefficient de variation, très élevée (supérieure à la valeur de la moyenne), caractérise une très forte dispersion du nombre d'enfants par famille autour de la valeur moyenne.

- ligne 2 : paramètres relatifs aux familles monoparentales :

$$\bar{y}_2 = \frac{929}{889} = 1{,}04 \text{ enfant par famille}$$

$$\sigma^2_{y2} = \frac{1821}{889} - 1{,}04^2 = 0{,}96 \text{ enfant}^2 \text{ par famille} \quad , \quad \sigma_{y2} = 0{,}98 \text{ enfant par famille}$$

$CV_{y2} = 0{,}98 / 1{,}04 = 0{,}94$. La valeur du coefficient de variation, très élevée, caractérise une très forte dispersion du nombre d'enfants par famille autour de la valeur moyenne.

Remarque : la dispersion du nombre d'enfants par famille est un peu moins élevée pour les familles monoparentales que pour les familles vivant en couples (résidence en France en 1982).

Ces différentes valeurs permettent de calculer la **variance des moyennes conditionnelles de y** (= variance interpopulation), selon l'expression vue plus haut. On obtient ici :

$$\text{variance des moyennes conditionnelles de } y = VE_y$$
$$= \frac{1}{14\,121}\left[13\,232\,(0{,}87 - 0{,}88)^2 + 889\,(1{,}04 - 0{,}88)^2\right]$$
$$= \frac{1}{14\,121}\,(13\,232 \times 0{,}87^2 + 889 \times 1{,}04^2) - 0{,}88^2 = 0{,}002$$

Enfin, on calcule le **rapport d'explication** :

$$\text{Rapport d'explication} = VEx / \sigma^2_x = 0{,}002 / 1{,}16 = \mathbf{0{,}0016}, \text{ soit } 0{,}2\ \% .$$

Ce résultat signifie que, pour les familles résidant en France en 1982, le type de famille n'explique rien de la dispersion relative au nombre d'enfants par famille. Celle-ci s'explique par d'autres facteurs, tels que le mode de vie, le niveau de vie, la religion, etc.

Remarque : selon la formule vue plus haut, on peut également calculer la moyenne des variances conditionnelles de y (= variance intrapopulation).

On a :

$$\sigma_y^2 = V(y) = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} \bar{y}_i^2 - \bar{y}^2 + \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} \sigma_{y_i}^2$$

et :

$$VR_y = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} \sigma_{y_i}^2 = \frac{1}{14\,121} (13\,232 \times 1{,}18 + 889 \times 0{,}96) = 1{,}163$$

En ajoutant cette valeur à VEy = 0,002 , on vérifie qu'on retrouve bien la valeur de la variance marginale de y.

**Conclusion générale**

Avec la méthode des tableaux d'indépendance, on doit se contenter d'apprécier plus ou moins subjectivement les sur et sous-représentation observables, case par case, ligne par ligne ou colonne par colonne.

Avec le calcul du rapport d'explication, on peut quantifier et donc rendre plus objectif le diagnostic porté sur le croisement des deux caractères x et y.

Sur notre exemple, avec le rapport d'explication, on se rend mieux compte du caractère négligeable de l'impact du statut de la famille (le caractère x) sur le nombre d'enfants par famille (le caractère y).

## CHAPITRE 4 : MISE EN RELATION DE DEUX CARACTÈRES QUANTITATIFS

Remarque (rappel du chapitre 3) :

Un travail effectué sur des caractères dont les modalités sont regroupées en classes (c'est-à-dire des caractères quantitatifs continus) n'est valide que si le nombre de classes est grand (au moins 15 classes pour chacun des deux caractères), sinon l'hypothèse des centres de classes (équirépartition) est hasardeuse. Par ailleurs, l'effectif total doit être conséquent, de façon à ce que les effectifs conjoints soient de taille suffisante ($n_{ij} > 30$). C'est la raison pour laquelle, dans le cadre du cours et des td, nous ne raisonnerons que sur des caractères quantitatifs discrets.

Nous nous attachons ici à déterminer si les caractères que nous croisons sont ou non indépendants. Lorsque deux caractères quantitatifs sont dépendants, on peut donner une mesure quantifiée de l'intensité de leur relation, selon deux logiques :

a) on peut raisonner dans le cadre de la recherche d'une relation **linéaire** entre les deux caractères et mesurer l'intensité de cette relation en mettant en œuvre un coefficient de corrélation **linéaire**.

b) on peut aussi raisonner dans un cadre plus général où la relation qui peut exister entre les deux caractères n'est pas linéaire. Dans ce cas, on mesure cette relation en mettant en œuvre des rapports de corrélation.

Chacune de ces deux approches renvoie à des méthodes de calcul différentes qui sont dévelopée dans la suite.

## 1. Recherche de la dépendance ou de l'indépendance linéaire de deux caractères quantitatifs : cas des tableaux individus-variables

Il s'agit de tableaux non croisés, à partir desquels on cherche à savoir s'il existe ou non une dépendance linéaire entre les deux caractères x et y.

### 11. Données générales

On considère une population de **n** individus et deux caractères **x** et **y**.

Chaque individu **i** de cette population est caractérisé par un couple d'observations **(xi ; yi)**.

Un **tableau individus-variables** se présente sous la forme suivante :

| $x_i$ | $y_i$ |
|---|---|
| $x_1$ | $y_1$ |
| … | … |
| $x_i$ | $y_i$ |
| … | … |
| $x_n$ | $y_n$ |
| $\sum_i x_i$ | $\sum_i y_i$ |

Chaque ligne du tableau renvoie à un individu particulier, caractérisé par l'association d'une certaine valeur de x et d'une certaine valeur de y.

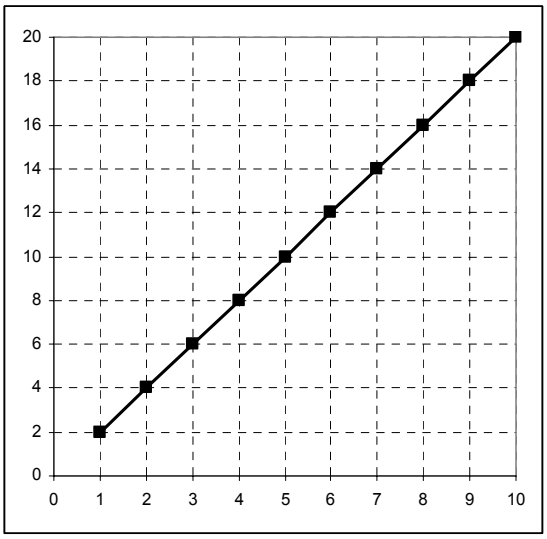
On peut réaliser un graphique en plaçant les valeurs des modalités de x en abscisse et celles de y en ordonnée. On obtient un nuage de points : chaque point matérialise un individu donné de la population.

De manière théorique, en matière de liaison entre les deux caractères, le champ des possibles est limité par les deux cas extrêmes suivants :

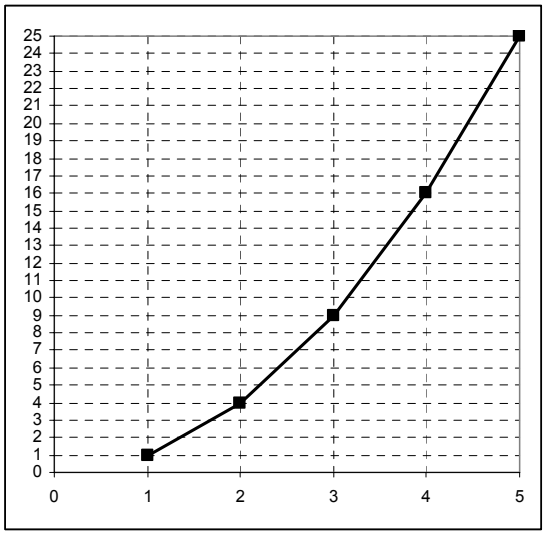
a) il peut exister une **liaison fonctionnelle** entre les deux caractères x et y. Dans ce cas, la connaissance des valeurs de x (la variable explicative) permet de déterminer exactement les valeurs correspondantes de y (la variable expliquée). On peut alors écrire : $y = f(x)$ .

Il s'agit d'une fonction au sens mathématique du terme, qui n'exprime pas nécessairement une relation linéaire entre les deux caractères x et y.

Graphiquement, on aura par exemple :



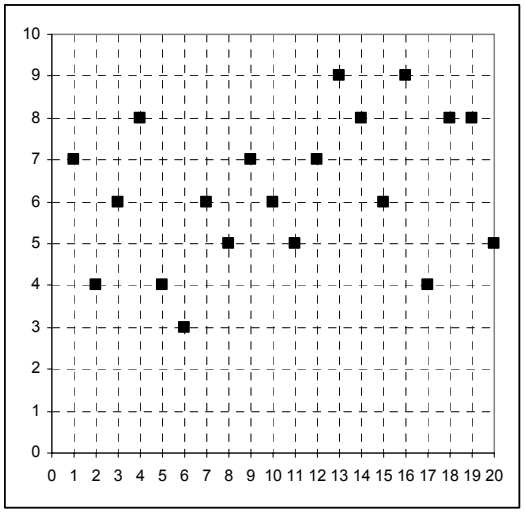

ou :




Une relation fonctionnelle peut ainsi être matérialisée par une relation linéaire ou par une relation non linéaire. **Chaque point du nuage correspond à un individu statistique**. La position de chaque point du nuage est déterminée par la conjonction d'une modalité donnée de x et d'une modalité donnée de y.

b) à l'opposé, il peut exister une **indépendance totale** entre les deux caractères x et y.

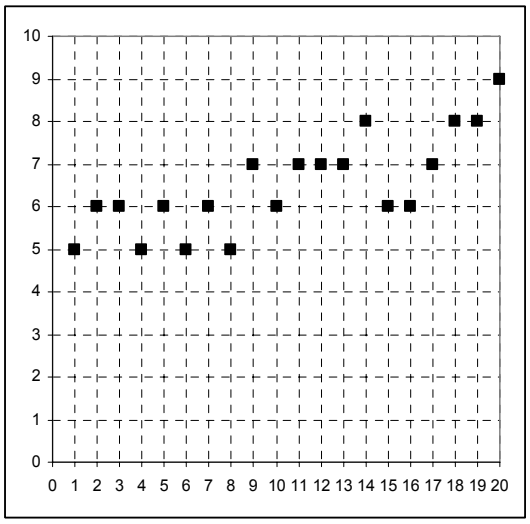
Graphiquement dans ce cas, on obtient généralement un nuage de points de la forme suivante :




Le nuage de points apparaît sous la forme d'une boule.

La réalité économique fait apparaître des situations intermédiaires, dans lesquelles il est possible de mettre en évidence une dépendance plus ou moins forte entre deux caractères x et y.

Dans de tels cas, graphiquement, la forme plus ou moins allongée du nuage de points peut suggérer l'existence d'une relation linéaire entre les deux caractères :




Dans le cas d'un nuage allongé et aplati, il est pertinent de chercher à savoir s'il existe une corrélation (linéaire) entre les deux caractères. Plus le nuage est aplati, plus la liaison entre les deux caractères est forte. À l'inverse, si le nuage tend plus vers la forme d'une boule, la liaison devient faible, voire inexistante, entre les deux caractères.

Mais, on ne peut pas en rester à une appréciation subjective de la lecture des graphiques. Il est nécessaire de pouvoir réaliser un calcul objectif qui permettra de déterminer avec précision l'intensité de la liaison qui peut exister entre les deux caractères.

Algébriquement, un calcul approprié permet une mise en évidence précise de la dépendance (linéaire) qu'on a pu déceler graphiquement. Si elle existe, on peut réaliser une mesure quantifiée de la dépendance entre x et y.

Pour cela, si l'on cherche à mesurer une dépendance linéaire, on utilise le coefficient de corrélation linéaire entre les deux caractères. Si l'on cherche à mesurer une dépendance non linéaire, on utilise les rapports de corrélation entre deux caractères.

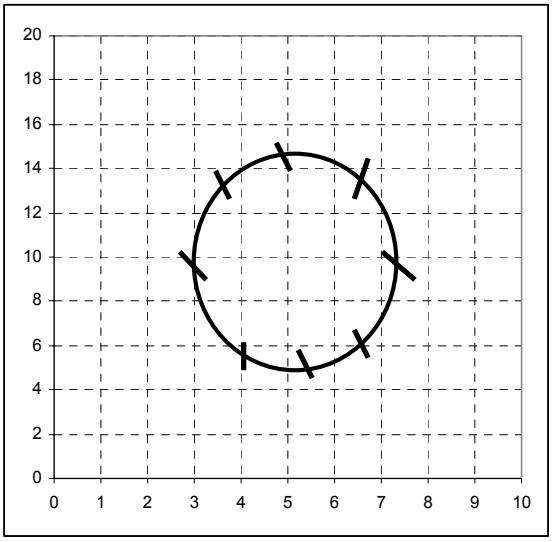
L'avantage de disposer d'une mesure quantifiée de la dépendance linéaire entre deux variables et de faciliter les comparaisons (i.e. qu'on se place dans les mêmes conditions d'analyse) d'échantillons ou de populations différentes.

Remarques :

a) lorsqu'on cherche à mesurer une dépendance linéaire entre deux caractères, en utilisant le coefficient de corrélation linéaire, il se peut que le calcul de ce dernier fasse apparaître une indépendance linéaire totale entre les deux caractères.

Simultanément, il peut exister une relation de type non linéaire entre les 2 caractères en question, qu'il n'est évidemment pas possible de mesurer en utilisant le coefficient de corrélation linéaire.

On peut prendre l'exemple classique suivant :




Dans un tel cas, où les points du nuage sont situés sur un cercle, la valeur du coefficient de corrélation linéaire est égale à zéro, signifiant par là qu'il y a indépendance linéaire totale entre les deux caractères. Mais dans le même temps, si l'on calcule les rapports de corrélation entre ces deux mêmes caractères, ceux-ci seront égaux à 1, car il existe une relation fonctionnelle circulaire entre les deux caractères en question.

b) malgré la limite apportée par la remarque précédente, la méthode qui consiste à faire apparaître une relation linéaire (si elle existe) entre deux caractères x et y est très largement utilisée dans la pratique. Essentiellement pour deux raisons :

α) le traitement mathématique des relations linéaires est plus simple que celui des relations non linéaires, ce qui amène des simplifications dans les calculs.

β) dans certains cas, par exemple lorsqu'on a affaire à des puissances ou des exponentielles (ce qui est fréquemment le cas en économie), on peut se ramener à des relations linéaires entre deux caractères, en posant des changements de variables appropriés.

Exemple :

Si l'on passe l'expression suivante aux logarithmes, on obtient :

$$y = b \cdot e^{a x} \quad \Rightarrow \quad \ln y = \ln b + a x$$

Si alors l'on pose : $\ln y = Y$ et : $\ln b = B$ , il vient : $\mathbf{Y = a x + B}$

Par changements de variables, on se ramène donc à une équation linéaire et il est alors possible d'utiliser la méthode ci-après (appelé méthode des moindres carrés), qui est associée au calcul du coefficient de corrélation linéaire entre deux variables.

### 12. La droite des moindres carrés, expression de la dépendance linéaire entre deux caractères quantitatifs

On parle aussi de droite d'ajustement de y en x (ou de x en y), ou encore de droite de régression (linéaire) de y en x (ou de x en y).

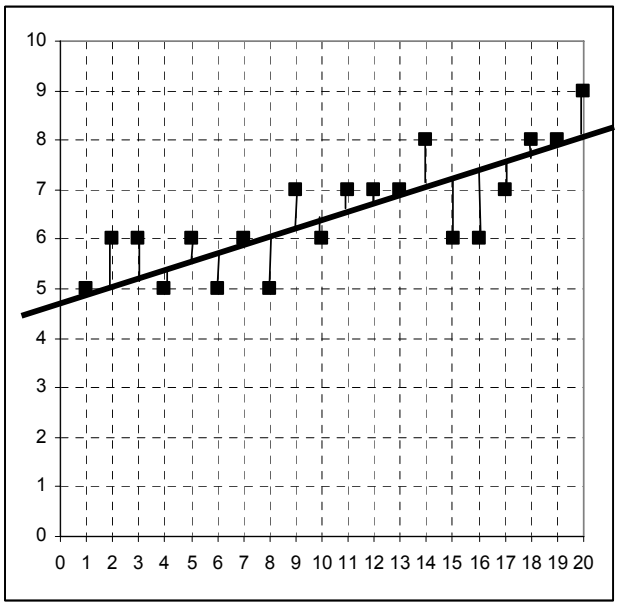
La dépendance linéaire entre deux caractères est exprimée graphiquement par les droites d'ajustement (de y en x, ou de x en y).

Le principe du tracé d'une droite des moindres carrés est le suivant :

on cherche à faire passer une droite à travers le nuage de points des observations, de telle façon que la somme des carrés des distances de chacun des points du nuage à cette droite soit minimale.

D'où le nom de droite des moindres carrés, qui est de la forme classique : $y = ax + b$.

Graphiquement, on a :




Il est à noter qu'habituellement, en mathématique, la distance d'un point à une droite est donnée par la projection orthogonale de ce point sur la droite en question.

Ici, on n'applique pas ce principe là : on appelle distance, au sens des moindres carrés, la longueur du segment qui relie un point du nuage à la droite, la mesure étant effectuée parallèlement à l'axe des ordonnées.

Plus précisément, d'une part on considère les carrés des distances (écarts) de chaque point du nuage à la droite, de façon à éliminer les signes moins pour faciliter les calculs, et d'autre part on cherche à minimiser la somme obtenue : d'où le nom de droite des moindres carrés.

L'équation d'une droite est égale à : $y = ax + b$ .

Le principe est de rechercher les valeurs des coefficients **a** et **b** de cette droite, de façon à ce que la somme des carrés des écarts, entre les points du nuage (c'est-à-dire les valeurs réelles observées) et les valeurs dites "ajustées" (c'est-à-dire les points de la droite d'ajustement de y en x), soit **minimale**.

On montre que le minimum recherché est assuré pour les valeurs suivantes des coefficients a et b de la droite d'ajustement de y en x :

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$b = \bar{y} - a\,\bar{x}$$

avec : n = nombre d'individus statistiques constituant la population (n représente aussi le nombre de points du nuage). $\sigma_{xy}$ = cov (xy) = covariance entre x et y. $\sigma^2_x$ = variance de x.

Les coefficients a et b caractérisent la **droite d'ajustement de y en x (D)**.

## La covariance

Il s'agit d'un indicateur de dispersion, qui tient compte des écarts aux moyennes respectives de chaque variable. Bien noter qu'on n'a pas de carrés sur les écarts (à la différence d'une variance). Ainsi, une covariance peut être **négative**.

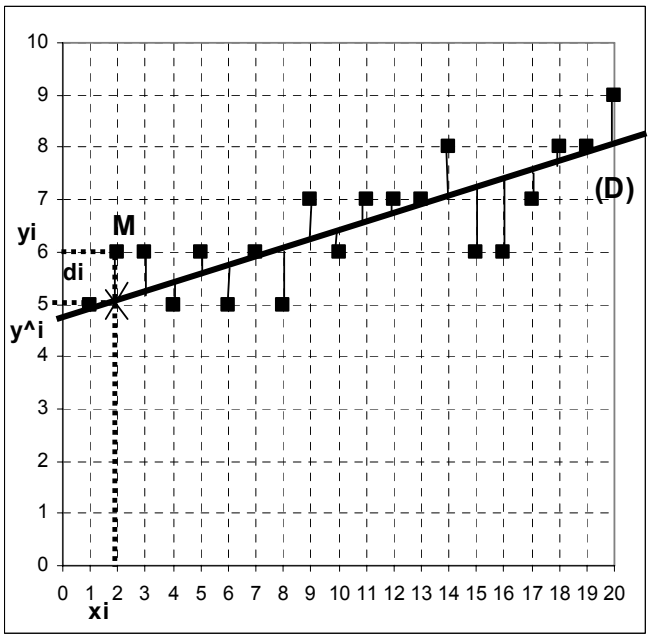
$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}$$

La 2ème expression est obtenue, comme pour la variance, en développant les termes du produit.

**cqfd**

On a : $a = \frac{\sigma_{xy}}{\sigma_x^2}$ et : $b = \bar{y} - a\,\bar{x}$

On cherche à minimiser les carrés des écarts $d_i$ des points M $(x_i; y_i)$ à la droite $y\hat{}_i = ax_i + b$ (**α**).



On a : $\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (\beta)$ avec : n = nb d'individus statistiques = nb de points du nuage.

Reportons (**α**) dans (**β**) : $\min \sum_{i=1}^{n} (y_i - ax_i - b)^2$, fonction de deux variables a et b.

Conditions du 1er ordre :

$$\frac{\partial \sum}{\partial a} = \sum_{i=1}^{n} -2\,x_i\,(y_i - ax_i - b) = -2\sum_{i=1}^{n} x_i\,(y_i - ax_i - b) = 0 \quad (1)$$

$$\frac{\partial \sum}{\partial b} = \sum_{i=1}^{n} -2\,(y_i - ax_i - b) = -2\sum_{i=1}^{n} (y_i - ax_i - b) = 0 \quad (2)$$

Développons (2) : $-2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n} ax_i + 2\sum_{i=1}^{n} b = -\sum_{i=1}^{n} y_i + a\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} b = 0$

$$-n\bar{y} + an\bar{x} + nb = -\bar{y} + a\bar{x} + b = 0 \quad \Rightarrow \quad b = \bar{y} - a\bar{x} \quad (3)$$

Conditions du 2ème ordre :

$$\frac{\partial^2 \sum}{\partial a^2} = 2\sum_{i=1}^{n} x_i^2 > 0 \quad (1') \quad \text{et} : \quad \frac{\partial^2 \sum}{\partial b^2} = 2n > 0 \quad (2')$$

On minimise donc bien la somme des carrés des écarts de chaque point du nuage à la droite d'ajustement de y en x (ce d'autant que n > 10 pour un ajustement linéaire valide).

De (1) et (3), on tire :

$$-2\sum_{i=1}^{n} x_i\,(y_i - ax_i - \bar{y} + a\bar{x}) = 0$$

$$\sum_{i=1}^{n} x_i\,(y_i - ax_i - \bar{y} + a\bar{x}) = \sum_{i=1}^{n} x_i\,[(y_i - \bar{y}) - a\,(x_i - \bar{x})] = 0$$

$$\sum_{i=1}^{n} x_i\,(y_i - \bar{y}) = a\sum_{i=1}^{n} x_i\,(x_i - \bar{x})$$

$$\text{D'où : } a = \frac{\sum_{i=1}^{n} x_i\,(y_i - \bar{y})}{\sum_{i=1}^{n} x_i\,(x_i - \bar{x})} = \frac{\sum_{i=1}^{n} x_i\,y_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i} = \frac{\sum_{i=1}^{n} x_i\,y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}$$

Divisons numérateur et dénominateur par n : $$a = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{\sigma_{xy}}{\sigma_x^2} \quad (4)$$

avec :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \bar{y} + \frac{1}{n}\sum_{i=1}^{n} \bar{x}\bar{y} - \frac{1}{n}\sum_{i=1}^{n} \bar{x} y_i$$

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

Dans : $y = ax + b$ , remplaçons b et a par (3) et (4) :

$$y = \frac{\sigma_{xy}}{\sigma_x^2} x + (\bar{y} - a\bar{x}) \Leftrightarrow y = \frac{\sigma_{xy}}{\sigma_x^2} x + (\bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x})$$

$$y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x}) \Leftrightarrow y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x}) = a(x - \bar{x})$$

C'est l'équation de la droite (D) : $y = ax + b = ax + (y - a\bar{x})$

**fin cqfd**

### 13. Calcul de l'intensité de la dépendance linéaire entre deux caractères quantitatifs : le coefficient de corrélation linéaire

Pour que le tracé d'une droite d'ajustement soit statistiquement justifié, deux paramètres interviennent :

- le premier paramètre concerne la taille n de la population (ou plus généralement de l'échantillon) que l'on étudie, c-à-d le nombre d'observations dont on dispose. **On doit avoir au moins n > 10**.

- le deuxième paramètre est la valeur du **coefficient de corrélation linéaire**, qui précise le degré d'intensité de la dépendance **linéaire** qui peut exister entre deux caractères.

Il s'agit d'un **nombre sans dimension** (donc indépendant des unités choisies pour exprimer les modalités des caractères x et y), que l'on note $\rho_{xy}$ ("rho de xy") :

$$\rho_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\sigma_x \cdot \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

On démontre que les valeurs de $\rho_{xy}$ sont comprises entre -1 et +1 .

**cqfd**

On a : $-1 \leq \rho \leq +1$

Soit les variables centrées : $x' = x - \overline{x}$ et : $y' = y - \overline{y}$

et : $\frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij} (\lambda x'_i + y'_j)^2$ (A) avec $\lambda \in IR$

$$\frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij} (\lambda^2 x'^2_i + 2 \lambda x'_i y'_j + y'^2_j)$$

Développons : $$= \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{p} (n_{ij} \lambda^2 x'^2_i + 2 \lambda n_{ij} x'_i y'_j + n_{ij} y'^2_j)$$

$$= \frac{\lambda^2}{n} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij} x'^2_i + \frac{2\lambda}{n} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij} x'_i y'_j + \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij} y'^2_j$$

On a : $\sum_{j=1}^{p} n_{ij} = n_{i.}$ et : $\sum_{i=1}^{m} n_{ij} = n_{.j}$

Donc :

$$= \frac{\lambda^2}{n} \sum_{i=1}^{m} n_{i.} (x_i - \overline{x})^2 + \frac{2\lambda}{n} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij} (x_i - \overline{x})(y_j - \overline{y}) + \frac{1}{n} \sum_{j=1}^{p} n_{.j} (y_j - \overline{y})^2$$

$$= \lambda^2 \sigma_x^2 + 2\lambda \sigma_{xy} + \sigma_y^2 \quad (A)$$

L'expression (A) est toujours positive ou nulle, car il s'agit d'une somme de carrés, $\forall \lambda$ .

On a une forme trinôme : $a \lambda^2 + b \lambda + c$ , avec a et $\lambda^2$ toujours $\geq 0$ . Par suite, pour que le trinôme soit $\geq 0$ (signe de a), il faut que le discriminant soit $\leq 0$ .

$$\Delta' = \sigma_{xy}^2 - \sigma_x^2 \sigma_y^2 \leq 0 \text{ (inégalité de Schwartz)}$$

On a : $$\rho^2 - 1 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} - 1 \leq 0 \quad \Rightarrow \quad \rho^2 \leq 1 \quad \Rightarrow \quad -1 \leq \rho \leq +1$$

**Fin cqfd**

$\rho_{xy}$ = +1 signifie qu'il existe une relation de dépendance linéaire fonctionnelle **positive** entre les deux caractères y et x.

$\rho_{xy}$ = -1 signifie qu'il existe une relation de dépendance linéaire fonctionnelle **négative** entre les deux caractères y et x.

$\rho_{xy} = 0$ signifie qu'il n'existe pas de relation **linéaire** entre y et x.

**Cela ne signifie pas pour autant qu'il n'existe pas éventuellement une relation non linéaire entre y et x !** (cf. remarque du point 11).

Généralement, on calcule $\rho^2_{xy}$ , appelé **coefficient de détermination**, qui varie entre 0 et + 1.

Multiplié par 100, ce coefficient mesure le **pourcentage d'explication** procurée par un ajustement linéaire (cf. aussi les rapports de corrélation).

Il a la même signification que le rapport d'explication vu au chapitre 3.

Par exemple, si l'on a : $\rho^2_{xy} = 0{,}8$ , cela signifie que l'ajustement linéaire de y (la fonction) en x (la variable) explique 80 % de la dispersion de y (c-à-d de la variance marginale (totale) $\sigma^2_y$ du caractère y).

On peut dire aussi que la connaissance de x explique 80 % de la dispersion des valeurs de y correspondantes.

Concrètement, si x (variable) correspond au revenu et y (fonction) correspond à une dépense de consommation et que, dans le même temps, on a un coefficient de détermination : $\rho^2_{xy} = 0{,}8$ , alors on peut dire que le revenu explique 80 % de la dispersion qui existe sur les dépenses de consommation, dans la population considérée ("le revenu explique 80 % de la consommation").

Remarques :

**Attention ! Ne pas confondre corrélation et causalité !**

Deux caractères peuvent faire apparaître un coefficient de corrélation linéaire très satisfaisant, sans que, pour autant, il existe une **relation de causalité** entre ces 2 caractères.

Exemple : si, lors de la période estivale, on met en relation x = vente de boissons fraîches et y = vente de lunettes de soleil, $\rho_{xy}$ sera très élevé, ce qui n'induit cependant rien sur l'influence directe de l'un de ces deux caractères sur l'autre.

Dans ce cas en effet, c'est une température élevée qui représente la bonne variable explicative de la croissance des ventes de x comme celles de y.

a) $\rho^2_{xy} = 0{,}5$ (donc : $\rho_{xy} = 0{,}71$) signifie que l'ajustement linéaire de y en x explique la moitié de la variance marginale de y.

b) en général (voir exceptions ci-après), $\rho^2_{xy} = 0{,}81$ (donc : $\rho_{xy} = 0{,}9$) est le signe d'un bon ajustement linéaire de y en x (80 % d'explication).

$\rho^2_{xy} = 0{,}9$ (donc : $\rho_{xy} = 0{,}95$) est le signe d'un très bon ajustement linéaire de y en x (90 % d'explication).

c) **lorsque la taille de l'échantillon est trop petite (n < 10), même un coefficient de détermination très élevé n'a pas de signification !**

A l'inverse, si l'échantillon est grand (par ex. n = 1 000 ou 10 000), même un coefficient de détermination médiocre (de l'ordre de 0,4 , par ex.) permet d'avancer l'hypothèse qu'il existe une dépendance linéaire entre y et x (même si, dans ce cas, on ne peut évidemment pas qualifier de bonne la dépendance linéaire en question).

Remarques :

1) On montre que : $V(y) = \sigma_y^2 = VE_y + VR_y = \rho^2 \sigma_y^2 + (1-\rho^2)\sigma_y^2$

C'est-à-dire que la variance marginale (totale) de y est égale à la somme de la part de la variance totale de y qui est expliquée (variance expliquée) par l'ajustement (régression) de y en x, et de la part de la variance totale de y qui n'est pas expliquée (variance résiduelle) par l'ajustement (régression) de y en x

2) Dans le cas général, on a :

Courbe de régression de y en x ($C_{y/x}$)

$$\eta^2_{y/x} = \frac{VE_y}{\sigma^2_y} \Rightarrow VE_y = \eta^2_{y/x}\,\sigma^2_y \quad \text{et} \quad VR_y = (1-\eta^2_{y/x})\,\sigma^2_y$$

Droite d'ajustement de y en x (D)

$$\rho^2 = \frac{VE_y}{\sigma^2_y} \Rightarrow VE_y = \rho^2\,\sigma^2_y \quad \text{et} \quad VR_y\ (1-\rho^2)\,\sigma^2_y$$

## 14. Droite d'ajustement des moindres carrés et coefficient de corrélation linéaire

Comparons les formules du coefficient de corrélation linéaire et de la pente de la droite d'ajustement de y en x :

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \quad \text{et :} \quad a = \frac{\sigma_{xy}}{\sigma^2_x}$$

Il vient : $\sigma_{xy} = \rho_{xy}\,\sigma_x\,\sigma_y = a\,\sigma^2_x \Leftrightarrow a = \rho_{xy}\,\frac{\sigma_y}{\sigma_x}$

Il s'agit de la **formule opératoire** du calcul de la **pente de la droite**, qu'on ne calcule que si la valeur du coefficient de corrélation est suffisamment élevée (en valeur absolue).

De cette formule résulte l'écriture suivante de la **droite (D) d'ajustement de y en x** :

$$(D) \quad y - \overline{y} = \rho_{xy} \frac{\sigma_y}{\sigma_x} (x - \overline{x}) \quad \Leftrightarrow \quad y = \rho_{xy} \frac{\sigma_y}{\sigma_x} x + (\overline{y} - a\,\overline{x})$$

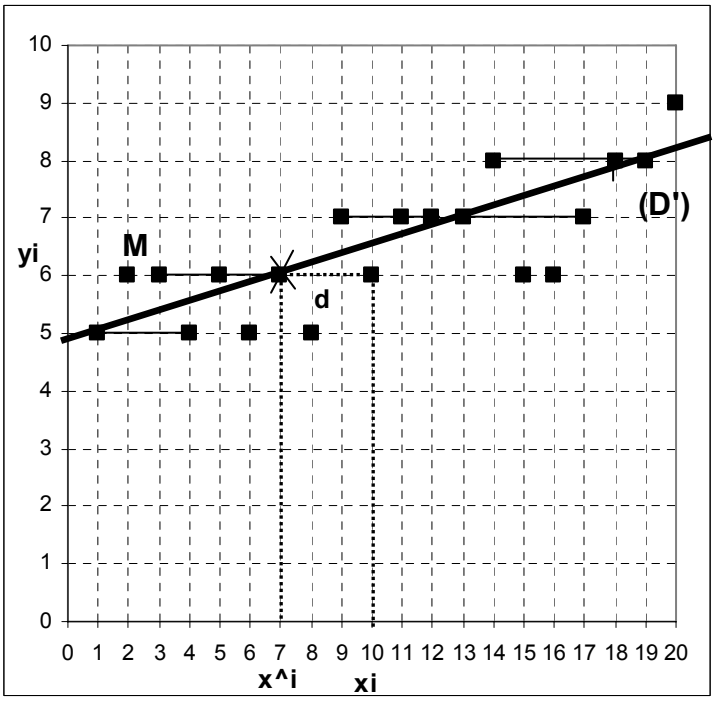
Remarque : de manière générale, une analyse de régression consiste à déterminer la forme concrète de la dépendance qui peut exister entre deux variables.

La formule ci-dessus vaut lorsqu'on considère x comme la variable (explicative) et y comme la fonction (variable expliquée), dans un repère habituel (O, x, y).

Rien n'empêche mathématiquement (**mais pas nécessairement en termes de causalité !!**)de permuter le rôle des deux variables x et y .

**Si y devient la variable et x la fonction**, on peut déterminer **l'équation d'une autre droite des moindres carrés, que l'on désigne par (D')**.

**cqfd**




On appelle **droite d'ajustement linéaire de x en y** (ou droite des moindres carrés de x en y, ou encore droite de régression linéaire de x en y) la droite (D') qui minimise les carrés des écarts des points du nuage à cette droite, les écarts étant mesurés parallèlement à l'axe des abscisses Ox.

**Remarque** : pour déterminer l'équation de la droite (D'), on peut aussi renverser le système d'axe, en permutant x et y.

Dans ce cas, on a : $(D') \Leftrightarrow x = a'' y + b''$

avec : $a'' = \frac{\sigma_{xy}}{\sigma^2_y}$ et : $b'' = \overline{x} - a'' \overline{y}$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \Leftrightarrow \sigma_{xy} = \rho_{xy} \cdot \sigma_x \cdot \sigma_y$$

Or, on a :

$$\text{et : } a'' = \frac{\sigma_{xy}}{\sigma_y^2} = \frac{\rho_{xy} \cdot \sigma_x \cdot \sigma_y}{\sigma_y^2} = \rho_{xy} \frac{\sigma_x}{\sigma_y}$$

On constate que le rapport des écarts-types est inversé par rapport à la droite (D).

Dans un repère inversé, on a donc :

$$(D') \quad x = a''y + b''$$

$$\Leftrightarrow x = \rho_{xy} \frac{\sigma_x}{\sigma_y} y + (\bar{x} - a'' \bar{y})$$

$$x - \bar{x} = a''(y - \bar{y}) = \rho_{xy} \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

alors que pour (D), nous avions : $(D) \quad y - \bar{y} = a(x - \bar{x}) = \rho_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x})$

Si l'on veut représenter les deux droites (D) et (D') dans le même repère habituel (x en abscisses et y en ordonnée), il suffit de permuter le rôle des deux variables x et y dans l'expression de (D').

Remarque : on a : $(D') \quad x = a''y + b''$

avec : $a'' = \frac{\sigma_{xy}}{\sigma_y^2}$ , $b'' = \bar{x} - a'' \bar{y}$ et : $x - \bar{x} = a''(y - \bar{y})$

Si l'on passe en repère habituel, on obtient : $y - \bar{y} = \frac{1}{a''}(x - \bar{x})$

avec : $\frac{1}{a''} = \frac{1}{\rho_{xy}} \frac{\sigma_y}{\sigma_x} = a'$

D'où, finalement : $(D') \quad y = a'x + (\bar{y} - a' \bar{x}) = a'x + b'$

Par conséquent, dans le repère habituel, on a :

$$(D') \quad y = a'x + b' \Leftrightarrow y = \frac{1}{\rho_{xy}} \frac{\sigma_y}{\sigma_x} x + (\bar{y} - a' \bar{x})$$

ou encore : $(D') \quad y - \bar{y} = \frac{1}{\rho_{xy}} \frac{\sigma_y}{\sigma_x} (x - \bar{x})$

Remarque : si l'on calcule (D') dans le référentiel inversé, on a : $x = a'' y + b''$

avec : $b'' = \bar{x} - a'' \bar{y}$ (1)

Par ailleurs : $b' = \bar{y} - a' \bar{x}$ avec : $a' = \frac{1}{a''}$

D'où : $b' = \bar{y} - \frac{\bar{x}}{a''} = \frac{a'' \bar{y} - \bar{x}}{a''} \Leftrightarrow a'' b' = a'' \bar{y} - \bar{x}$

et : $-a'' b' = \bar{x} - a'' \bar{y}$ (2)

De (1) et (2), on tire : $b'' = -a'' b'$ (3) ou : $b'' = -\frac{b'}{a'}$

De (3) : $b' = -\frac{b''}{a''}$ ou : $b' = -a' b''$

Par conséquent, il suffit de remplacer $\rho$ par $1 / \rho$ pour obtenir la valeur des paramètres de (D').

**fin cqfd**

En permutant le rôle des deux caractères (y devenant la variable et x la fonction), **si l'on se place dans le repère géométrique habituel avec x en abscisse et y en ordonnée**, l'équation de la **droite d'ajustement, de x en y (D ')**, est la suivante :

$$(D') \quad y = a'x + b' \quad \Leftrightarrow \quad y = \frac{1}{\rho_{xy}} \frac{\sigma_y}{\sigma_x} x + (\bar{y} - a' \bar{x})$$

Les interprétations relatives à cette droite d'ajustement de x en y sont similaires à celles vues plus haut pour la droite d'ajustement de y en x, en prenant soin de permuter le rôle des deux caractères.

Remarques :

a) on constate que pour obtenir les coefficients a ' et b ' de la droite (D '), il suffit de remplacer $\rho_{xy}$ par $1 / \rho_{xy}$ .

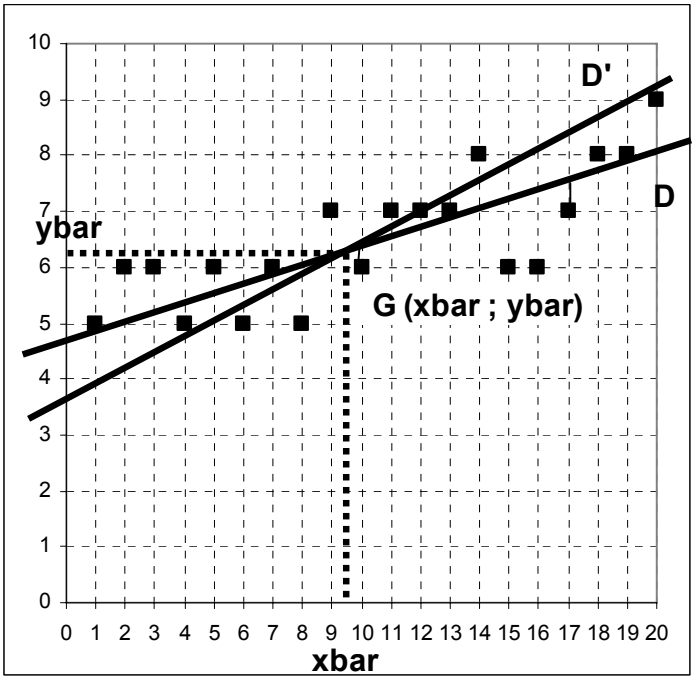
b) les deux droites d'ajustement (D) et (D ') sont confondues si $\rho_{xy} = + 1$ ou si $\rho_{xy} = - 1$ , c-à-d lorsqu'il existe une relation fonctionnelle entre les deux variables.

c) les droites d'ajustement (D) et (D ') sont toutes deux, soit croissantes, soit décroissantes, car $\rho_{xy}$ et $1 / \rho_{xy}$ sont de même signe.

d) **on montre que la pente de la droite (D ') est toujours supérieure ou égale à celle de la droite (D), lorsqu'elles sont représentées dans le même repère.**

e) le point d'intersection G des 2 droites d'ajustement (D) et (D ') correspond au point moyen (ou barycentre, ou centre de gravité) du nuage de points. On le note : $G(\bar{x} ; \bar{y})$

f) **bien qu'il existe deux droites d'ajustement linéaire (de y en x, et de x en y), on ne détermine qu'un seul coefficient de corrélation linéaire.**




## 15. Exemples

Calculer le coefficient de corrélation linéaire (et le cas échéant, déterminer l'équation de la droite de régression de y en x) pour chacune des séries suivantes. Interpréter les résultats obtenus :

**Exo 1**

| x | y |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 4 |

**Exo 2**

| x | y |
|---|---|
| 1 | 1 |
| 1 | 4 |
| 4 | 4 |

**Exo 3**

| x | y |
|---|---|
| 1 | 1 |
| 2 | 8 |
| 4 | 64 |

**Exo 4**

| x \ y | -2 | -1 | 0 | 1 | 2 |
|---|---|---|---|---|---|
| **0** | 1 | | 1 | | 1 |
| **1** | | | 2 | | |
| **2** | 1 | 2 | 3 | 2 | 1 |
| **3** | | | 2 | | |
| **4** | 1 | | 1 | | 1 |

**Exo 5**

| x \ y | 0 | 2 | 6 |
|---|---|---|---|
| **-1** | | | 1 |
| **0** | | 1 | |
| **1** | 1 | | |
| **2** | 1 | | |
| **3** | | 1 | |
| **4** | | | 1 |

**Remarque** : dans la réalité, le nombre de modalités des exercices précédents est insuffisant pour réaliser des ajustements linéaires pertinents. Au minimum, il convient de disposer d'au moins 10 observations.

**Exo 1**

| | xi | yi | xi² | yi² | xi yi |
|---|---|---|---|---|---|
| | 1 | 1 | 1 | 1 | 1 |
| | 2 | 2 | 4 | 4 | 4 |
| | 3 | 4 | 9 | 16 | 12 |
| **Total** | **6** | **7** | **14** | **21** | **17** |

x bar = 2,00 y bar = 2,33 cov xy = 1,00

var x = 0,67 var y = 1,56 **rho xy = 0,98**

ectyp x = 0,82 ectyp y = 1,25 **rho² xy = 0,96**

Rappels :

**La nature des phénomènes observés, la quantité d'informations dont on dispose sur eux, ainsi que le degré de précision ou d'approximation que l'on se fixe, permettent un déplacement de la valeur critique, relative à la pertinence d'un ajustement linéaire.**

Une valeur de rho xy = 0,95, soit : rho² xy ~ 0,90, correspond à un très bon ajustement linéaire.

La valeur de rho² xy (coefficient de détermination) indique le niveau d'explication de la régression de y en x (ou de x en y). En toute rigueur, pour valider un ajustement, il faut n > 10 observations.

Ici, l'ajustement de y en x est valide, car x explique la dispersion de y à 96 %. L'équation de la droite d'ajustement (D) est :

$$\hat{y} = a\,x + b = 1{,}5\,x - 0{,}7$$

avec : a = rho . ectyp y / ectyp x = 1,50 et : b = y bar - a . x bar = - 0,67

**Exo 2**

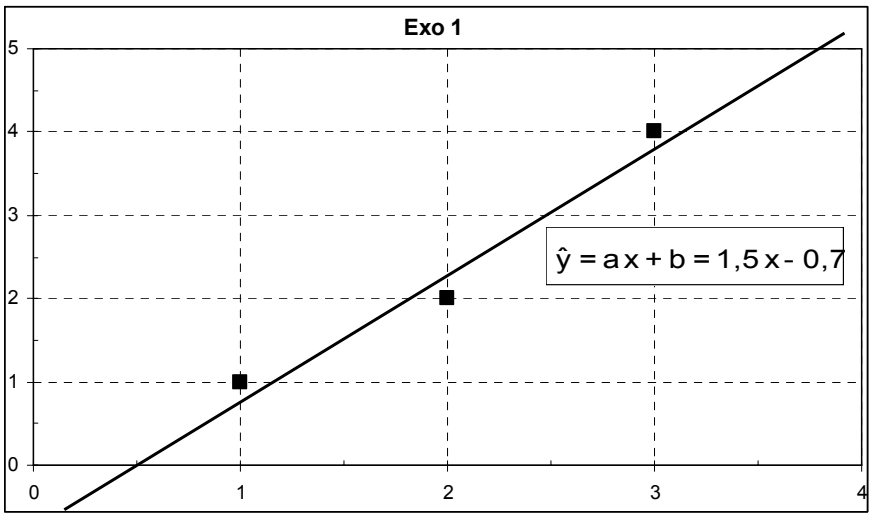
| | xi | yi | xi² | yi² | xi yi |
|---|---|---|---|---|---|
| | 1 | 1 | 1 | 1 | 1 |
| | 1 | 4 | 1 | 16 | 4 |
| | 4 | 4 | 16 | 16 | 16 |
| **Total** | **6** | **9** | **18** | **33** | **21** |

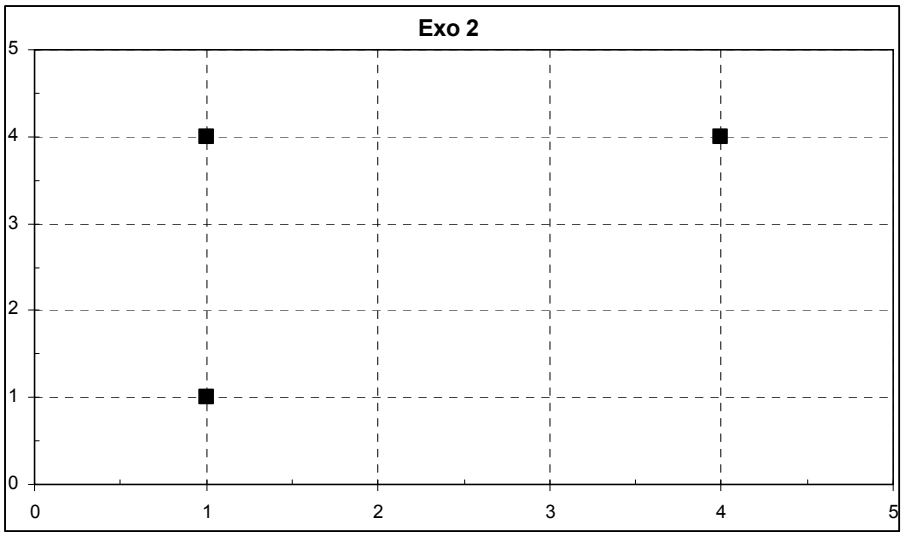
x bar = 2,00 y bar = 3,00 cov xy = 1,00

var x = 2,00 var y = 2,00 **rho xy = 0,50**

ectyp x = 1,41 ectyp y = 1,41 **rho² xy = 0,25**

Ici, le coefficient de corrélation linéaire est tout à fait insuffisant pour justifier la validité d'un ajustement linéaire de y en x (ou de x en y). La connaissance de x n'explique que 25 % de la dispersion de y.




**Exo 3**

| | xi | yi | xi² | yi² | xi yi |
|---|---|---|---|---|---|
| | 1 | 1 | 1 | 1 | 1 |
| | 2 | 8 | 4 | 64 | 16 |
| | 4 | 64 | 16 | 4 096 | 256 |
| **Total** | **7** | **73** | **21** | **4 161** | **273** |

x bar = 2,33 y bar = 24,33 cov xy = 34,22

var x = 1,56 var y = 794,89 **rho xy = 0,97**

ectyp x = 1,25 ectyp y = 28,19 **rho² xy = 0,95**

Ici, l'ajustement de y en x est valide, car les valeurs de x expliquent celle de y à 95 %. L'équation de la droite d'ajustement est :

$$\hat{y} = a\,x + b = 22\,x - 27$$

avec : a = rho . ectyp y / ectyp x = 22,00 et : b = ybar - a . xbar = - 27,00

L'ajustement linéaire, opéré ci-avant, est certes apparemment très correct. Toutefois, si l'on analyse de plus près les données dont nous disposons, nous constatons qu'il existe une relation fonctionnelle particulière entre les variables x et y.

En effet, on a : $x_i^3 = y_i$.

De ce fait, si l'on recherche une corrélation "linéaire" entre $x^3$ et y, on obtient un coefficient de corrélation linéaire égal à + 1, caractéristique d'une relation fonctionnelle entre y et $x^3$.

Posons : $X = x^3$. On a alors le tableau suivant :

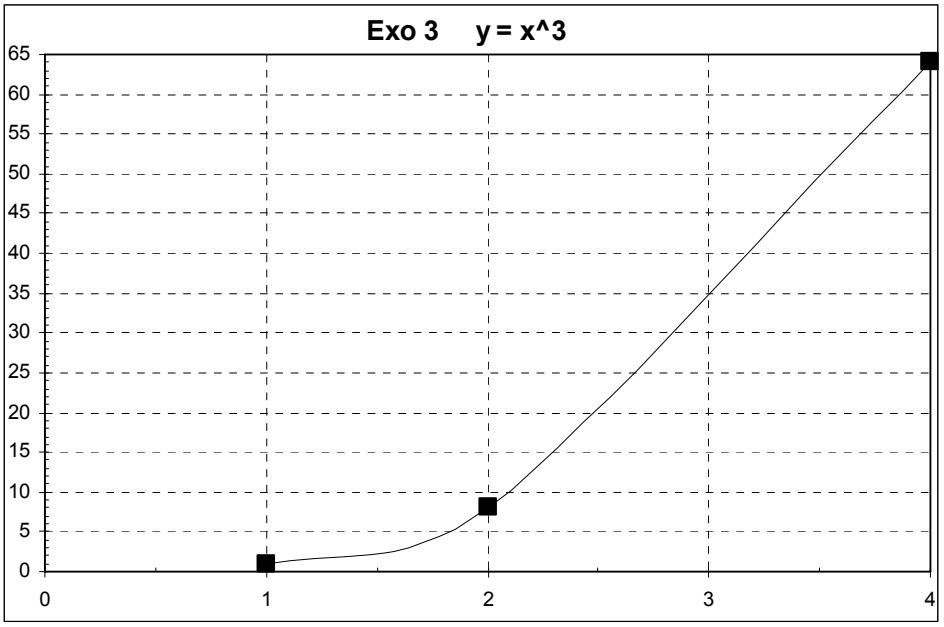
| | $X = x^3$ | yi | $X^2 = yi^2$ | X y |
|---|---|---|---|---|
| | 1 | 1 | 1 | 1 |
| | 8 | 8 | 64 | 64 |
| | 64 | 64 | 4 096 | 4 096 |
| **Total** | **73** | **73** | **4 161** | **4 161** |

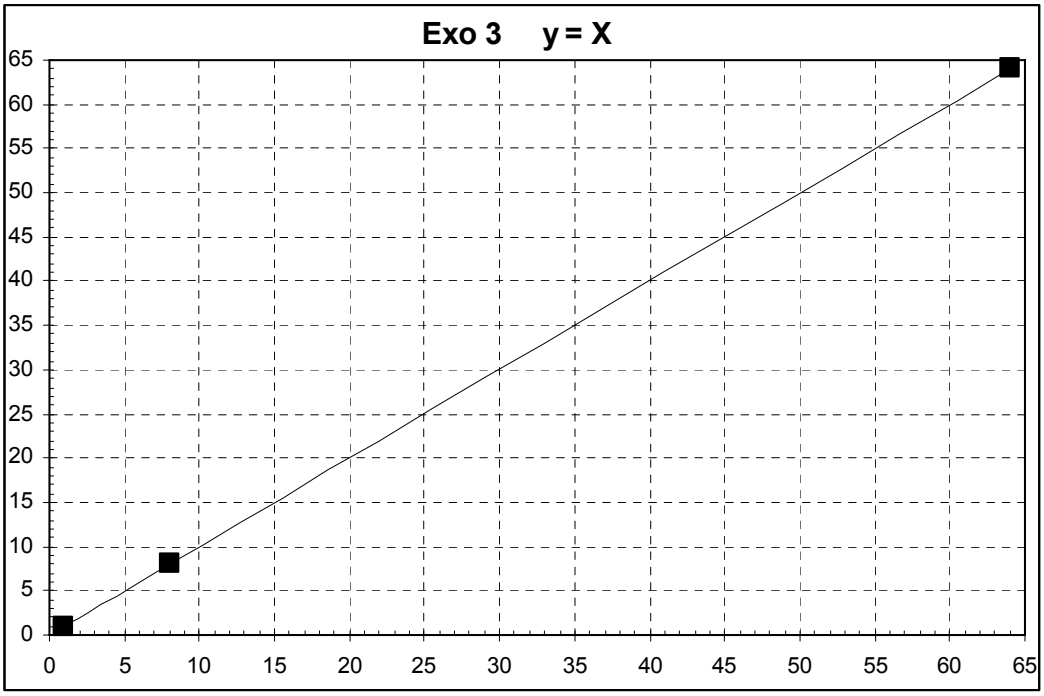
X bar = y bar = 24,33 cov Xy = 794,89

var X = var y = 794,89 **rho Xy = 1,00**

ectyp X = ectyp y = 28,19 **rho² Xy = 1,00**

Il existe une relation fonctionnelle linéaire entre X (= $x^3$) et y.






**Exo 4**

| xi \ yj | - 2 | - 1 | 0 | 1 | 2 | ni. | ni. xi | ni. xi² | nijxiyj |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | | 1 | | 1 | 3 | 0 | 0 | 0 |
| **1** | | | 2 | | | 2 | 2 | 2 | 0 |
| **2** | 1 | 2 | 3 | 2 | 1 | 9 | 18 | 36 | 0 |
| **3** | | | 2 | | | 2 | 6 | 18 | 0 |
| **4** | 1 | | 1 | | 1 | 3 | 12 | 48 | 0 |
| **n.j** | 3 | 2 | 9 | 2 | 3 | 19 | 38 | 104 | 0 |
| **n.j yj** | -6 | -2 | 0 | 2 | 6 | 0 | | | |
| **n.j yj²** | 12 | 2 | 0 | 2 | 12 | 28 | | | |
| **nijyjxi** | -12 | -4 | 0 | 4 | 12 | 0 | | | |

La somme des nij xi yj est nulle. Donc la covariance et le coefficient de corrélation linéaire sont tous deux nuls.

Pour autant ici, les variables x et y ne sont pas indépendantes. Le tableau montre en effet que la distribution est symétrique par rapport au centre de gravité du nuage de points, obtenu en plaçant les nij dans un plan xOy.

**Exo 5**

| xi \ yj | 0 | 2 | 6 | ni. | ni. xi | ni. xi² | nijxiyj |
|---|---|---|---|---|---|---|---|
| **-1** | | | 1 | 1 | -1 | 1 | -6 |
| **0** | | 1 | | 1 | 0 | 0 | 0 |
| **1** | 1 | | | 1 | 1 | 1 | 0 |
| **2** | 1 | | | 1 | 2 | 4 | 0 |
| **3** | | 1 | | 1 | 3 | 9 | 6 |
| **4** | | | 1 | 1 | 4 | 16 | 24 |
| **n.j** | 2 | 2 | 2 | 6 | 9 | 31 | 24 |
| **n.j yj** | 0 | 4 | 12 | 16 | | | |
| **n.j yj²** | 0 | 8 | 72 | 80 | | | |
| **nijyjxi** | 0 | 6 | 18 | 24 | | | |

x bar = 1,50 y bar = 2,67

var x = 2,92 var y = 6,22

ectyp x = 1,71 ectyp y = 2,49

cov xy = 0,00 **rho xy = 0,00** **rho² xy = 0,00**

Ici encore, le coefficient de corrélation linéaire est nul, indiquant qu'il n'existe pas de liaison linéaire entre y et x (ou entre x et y). Mais il existe une relation fonctionnelle non linéaire entre les deux variables y et x.

Pour le vérifier, récrivons le tableau initial sous la forme suivante :

| xi | yi |
|---|---|
| -1 | 6 |
| 0 | 2 |
| 1 | 0 |
| 2 | 0 |
| 3 | 2 |
| 4 | 6 |

Sous cette forme, on remarque mieux que, pour x = 1 et x = 2, la valeur correspondante de y est nulle. Mathématiquement, on a ainsi les racines de l'équation : (x - 1) (x - 2) = 0.

Par ailleurs, le tableau montre que la distribution est symétrique par rapport à une parallèle à l'axe des y, d'équation : x = xbar = 1,5.

On en déduit ainsi qu'il existe une relation de type fonctionnel parabolique entre y et x, de la forme : $y = x^2 - 3x + 2 = (x - 1)(x - 2)$.

**Exo 6**

Dans le tableau ci-après, une entreprise de maintenance informatique a rapproché, pour 14 observations, le nombre d'éléments à changer sur un ordinateur en panne ($x_i$) et le temps correspondant, en minutes révolues, de sa réparation ($y_i$) :

| $x_i$ | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 6 | 7 | 8 | 9 | 9 | 10 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 23 | 29 | 49 | 64 | 74 | 87 | 96 | 97 | 109 | 119 | 149 | 145 | 154 | 166 |

On se propose de calculer le coefficient de corrélation linéaire entre x et y, puis d'en déduire la valeur des coefficients de la droite d'ajustement de y en x : $\hat{y}_i = ax_i + b$ . Ensuite, on représente sur un même graphique la droite d'ajustement de y en x, le nuage de points ($x_i$, $y_i$) et les droites représentatives des moyennes $\bar{x}$ et $\bar{y}$.

* * *

Le tableau de calcul permet de déterminer tous les paramètres utiles. Il permet d'obtenir :

$\bar{x} = 6$ éléments $\quad \sigma_x^2 = 8{,}14$ éléments$^2$ $\quad \sigma_x = 2{,}85$ éléments

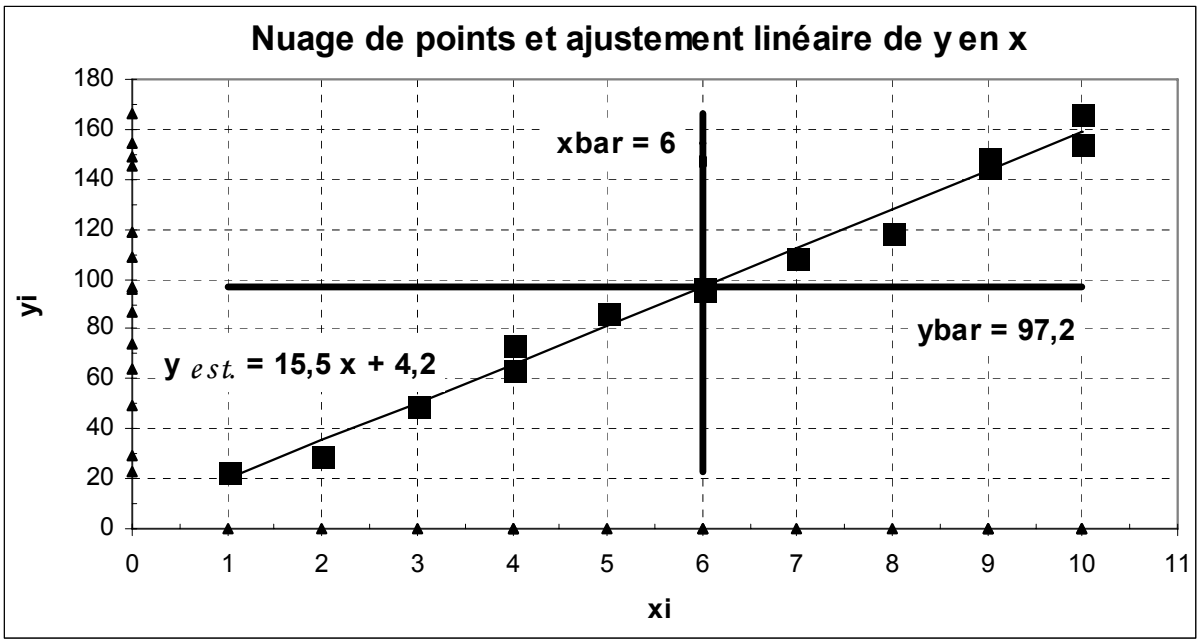
$\bar{y} = 97{,}21$ mn $\quad \sigma_y^2 = 1983{,}45$ mn$^2$ $\quad \sigma_y = 44{,}54$ mn

$\sigma_{xy} = 126{,}29$ éléments$^2$ $\quad \rho_{xy} = 0{,}994$ $\quad \rho_{xy}^2 = 0{,}987$

**Le nombre d'observations est supérieur à 10. Par ailleurs, la valeur du coefficient de corrélation est très élevée (0,994). On peut donc réaliser un ajustement linéaire de y en x satisfaisant.** On peut donc calculer les coefficients a et b de la droite (D) d'ajustement de y en x.

$a = 15{,}51$ et $b = 4{,}16$ On a donc : $\hat{y}_i = 15{,}51\, x_i + 4{,}16$

| $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_i y_i$ | $\hat{y}_i$ |
|---|---|---|---|---|---|
| 1 | 23 | 1 | 529 | 23 | 19,67 |
| 2 | 29 | 4 | 841 | 58 | 35,18 |
| 3 | 49 | 9 | 2 401 | 147 | 50,69 |
| 4 | 64 | 16 | 4 096 | 256 | 66,20 |
| 4 | 74 | 16 | 5 476 | 296 | 66,20 |
| 5 | 87 | 25 | 7 569 | 435 | 81,71 |
| 6 | 96 | 36 | 9 216 | 576 | 97,21 |
| 6 | 97 | 36 | 9 409 | 582 | 97,21 |
| 7 | 109 | 49 | 11 881 | 763 | 112,72 |
| 8 | 119 | 64 | 14 161 | 952 | 128,23 |
| 9 | 149 | 81 | 22 201 | 1 341 | 143,74 |
| 9 | 145 | 81 | 21 025 | 1 305 | 143,74 |
| 10 | 154 | 100 | 23 716 | 1 540 | 159,25 |
| 10 | 166 | 100 | 27 556 | 1 660 | 159,25 |
| **84** | **1 361** | **618** | **160 077** | **9 934** | **1 361,00** |




**La droite d'ajustement de y en x "résume" d'autant mieux le nuage de points que ces derniers sont alignés**. On montre que ce qui est résumé par la droite correspond, pour chaque point du nuage, à la distance entre la droite d'ajustement et la moyenne marginale de y, ici égale à 97,2 (on appelle cela "**variance de y, expliquée par x**"). A l'inverse, ce qui n'est pas résumé par la droite correspond, pour chacun des points du nuage, à la distance entre un point du nuage et la droite d'ajustement (on appelle cela "**variance résiduelle de y, non expliquée par x**").

Remarque : le centre de gravité du nuage des points (xi, yi) correspond au point d'intersection de la droite d'ajustement de y en x (et aussi de x en y), ainsi que des droites correspondant à la moyenne marginale de x et à celle de y.

**Ici, le coefficient de détermination est égal à 0,987. Cela signifie que le nombre d'éléments à changer en cas de panne (x), explique à 98,7 % (soit la quasi totalité) de la dispersion des temps de réparation (y).**

**Remarque fondamentale** :

**ici la recherche d'un ajustement linéaire de x en y n'a pas de sens, car la durée de réparation des machines en panne n'induit pas le nombre d'éléments à changer :**

**il ne pas confondre corrélation et relation de causalité !!**

**Exo 6 bis**

Une entreprise de maintenance informatique a rapproché le nombre d'éléments à changer sur un ordinateur en panne ($x_i$) et le temps correspondant, en minutes révolues, de sa réparation ($y_i$) . Le tableau ci-après retient les 14 observations suivantes :

| $x_i$ | 1 | 2 | 3 | 4 | 4 | 5 | 6 | 6 | 7 | 8 | 9 | 9 | 10 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 23 | 29 | 49 | 64 | 74 | 87 | 96 | 97 | 109 | 119 | 149 | 145 | 154 | 166 |

a) Quels sont la population, les caractères et types de la série statistique envisagée ?

b) Calculer les moyenne, variance et écart-type marginaux des variables x et y. Calculer les coefficients de variation CVx et CVy.

c) Calculer la covariance et le coefficient de corrélation linéaire entre x et y . Interpréter les résultats. En déduire la valeur des coefficients de la droite d'ajustement de y en x :

$$\hat{y}_i = ax_i + b .$$

d) Représenter sur un même graphique la droite d'ajustement de y en x, le nuage de points ($x_i$, $y_i$) et les droites représentatives des moyennes $\bar{x}$ et $\bar{y}$ . Commenter.

e) Déterminer la part de variance expliquée par la régression de y en x , en calculant : $\sigma_E^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ . Matérialiser sur le graphique les <u>écarts expliqués par la régression</u>.

f) Déterminer la part de variance résiduelle, non expliquée par la régression de y en x ,. en calculant : $\sigma_R^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y})^2$ . Matérialiser sur le graphique les <u>écarts résiduels non expliqués par la régression</u>.

g) Vérifier que l'on a : $\sigma_E^2 + \sigma_R^2 = \sigma_y^2$ .

h) A quoi est égal le <u>rapport d'explication</u> $\frac{\sigma_E^2}{\sigma_y^2}$ ?

**a)**

<u>Population</u> : 14 ordinateurs en panne.
<u>Caractères</u> : - nombre d'éléments à changer lors d'une panne (xi).
- nombre de minutes nécessaires à la réparation d'une machine (yi).
<u>Type</u> : quantitatif discret, pour chacun des deux caractères.

| 1 | | | | | 2 | 1 - 2 | (1 - 2)² | 2 - 3 | (2 - 3)² |
|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_i y_i$ | $\hat{y}_i$ | $y_i - \hat{y}_i$ | $(y_i - \hat{y}_i)^2$ | $\hat{y}_i$ - ybar | $(\hat{y}_i$ - ybar$)^2$ |
| 1 | 23 | 1 | 529 | 23 | 19,7 | 3,3 | 11,1 | -77,5 | 6 013,1 |
| 2 | 29 | 4 | 841 | 58 | 35,2 | -6,2 | 38,2 | -62,0 | 3 848,4 |
| 3 | 49 | 9 | 2 401 | 147 | 50,7 | -1,7 | 2,9 | -46,5 | 2 164,7 |
| 4 | 64 | 16 | 4 096 | 256 | 66,2 | -2,2 | 4,8 | -31,0 | 962,1 |
| 4 | 74 | 16 | 5 476 | 296 | 66,2 | 7,8 | 60, 9 | -31,0 | 962,1 |
| 5 | 87 | 25 | 7 569 | 435 | 81,7 | 5,3 | 28,0 | -15,5 | 240,5 |
| 6 | 96 | 36 | 9 216 | 576 | 97,2 | -1,2 | 1,5 | 0,0 | 0,0 |
| 6 | 97 | 36 | 9 409 | 582 | 97,2 | -0,2 | 0,1 | 0,0 | 0,0 |
| 7 | 109 | 49 | 11 881 | 763 | 112,7 | -3,7 | 13,9 | 15,5 | 240,5 |
| 8 | 119 | 64 | 14 161 | 952 | 128,2 | -9,2 | 85,2 | 31,0 | 962,1 |
| 9 | 149 | 81 | 22 201 | 1 341 | 143,7 | 5,3 | 27,7 | 46,5 | 2 164,7 |
| 9 | 145 | 81 | 21 025 | 1 305 | 143,7 | 1,3 | 1,6 | 46,5 | 2 164,7 |
| 10 | 154 | 100 | 23 716 | 1 540 | 159,3 | -5,3 | 27,6 | 62,0 | 3 848,4 |
| 10 | 166 | 100 | 27 556 | 1 660 | 159,3 | 6,8 | 45,6 | 62,0 | 3 848,4 |
| **84** | **1 361** | **618** | **160 077** | **9 934** | **1 361,0** | **0,0** | **348,9** | **0,0** | **27 419,5** |

**b)**

x bar = 6,00 éléments

y bar = 97,21 minutes

var x = 8,14

var y = 1983,45

ectyp x = 2,85 éléments

ectyp y = 44,54 minutes

**CVx = 0,48**

**CVy = 0,46**

La dispersion autour de la moyenne est moyenne pour chacun des deux caractères x et y.

**c)**

cov xy = 126,29

**rho xy = 0,994**

**rho² xy = 0,987**

Coefficients de la droite de régression :

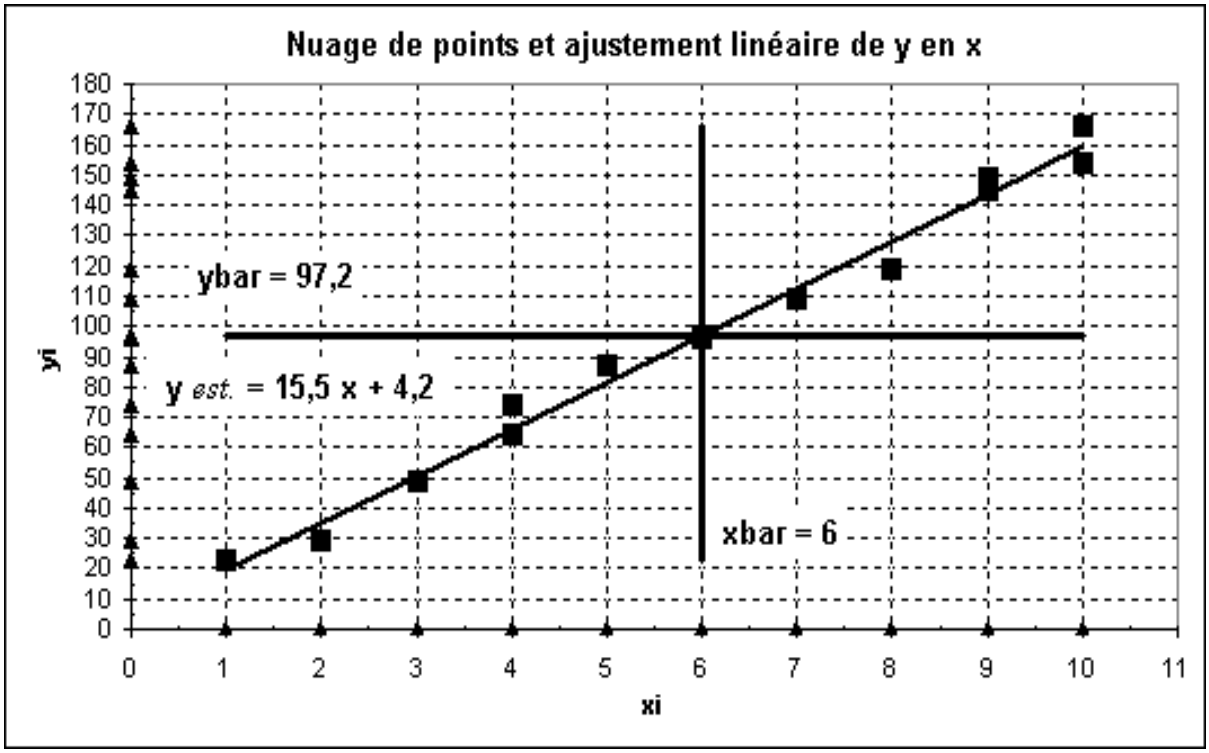
**a = 15,51**

**b = 4,16**

**Le nombre d'éléments à changer explique le temps de réparation à 99 % !**

Ici, le nombre d'observations (n = 14) est suffisant pour valider l'ajustement linéaire. D'où l'équation de la droite d'ajustement de y en x : $\hat{y}_i = 15{,}51\, x_i + 4{,}16$

**d)**




**La droite d'ajustement de y en x "résume" d'autant mieux le nuage de points que ces derniers sont alignés**.

On montre que ce qui est résumé par la droite correspond, pour chaque point du nuage, à la distance entre la droite d'ajustement et la moyenne marginale de y, ici égale à 97,2 (on appelle cela "**variance de y, expliquée par x**").

A l'inverse, ce qui n'est pas résumé par la droite correspond, pour chacun des points du nuage, à la distance entre un point du nuage et la droite d'ajustement (on appelle cela "**variance résiduelle de y, non expliquée par x**").

Cela correspond aux fluctuations aléatoires du temps de réparation, indépendamment du nombre d'éléments à changer.

Remarque : le centre de gravité du nuage des points (xi, yi) correspond au point d'intersection de la droite d'ajustement de y en x (et aussi de x en y), ainsi que des droites correspondant à la moyenne marginale de x et à celle de y.

**e)**

On a : **variance expliquée** = $\sigma_E^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

avec n = nombre d'onservations = nombre de points du nuage.

**VEy** = 27.419,5 / 14 = **1958,5**

C'est la part de la dispersion des temps de réparation (y), expliquée par le nombre d'éléments à changer (x).

**f)**

On a : **variance résiduelle** = $\boxed{\sigma^2_R = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y})^2}$

**VRy** = 348,9 / 14 = **24,9**

C'est la part de la dispersion des temps de réparation (y), non expliquée par le nombre d'éléments à changer (x).

**g)**

On a : **var. expliquée + var. résiduelle = var y =** $\boxed{\sigma^2_E + \sigma^2_R = \sigma^2_y}$ **= 1983,5**

**h)**

On a : $\boxed{\frac{\sigma^2_E}{\sigma^2_y}}$ = 1958,5 / 1983,5 = **0,987** = rho² xy = **coefficient de détermination** = $\boxed{\frac{VEy}{\sigma^2_y}}$

* * *

On a : $\boxed{\sigma^2_y = VEy + VRy = \sigma^2_E + \sigma^2_R = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y})^2}$

**VEy** correspond à la variance des moyennes (conditionnelles) : les ^yi représentent bien des valeurs moyennes (celles qui sont sur la droite d'ajustement de y en x).. Chaque point de ^y correspond à une sous-population (valeur moyenne).

**VRy** de correspond à la moyenne des variances (conditionnelles) : les écarts points-droite correspond à une dispersion entre les temps de réparation, par rapport à un nombre "moyen" d'éléments à changer.

Remarques terminales :

1) un ajustement linéaire nécessite au moins une dizaine d'observations (c'est le cas ici) pour que les coefficients a et b de la droite de régression soient "fiables". En effet, au-delà de 10 observations, les écarts aléatoires (centrés réduits) à la droite de régression commencent à converger vers une loi normale, ce qui n'est pas nécessairement le cas en dessous de 10 observations.

2) ici la recherche d'un ajustement linéaire de x en y n'a pas de sens, car la durée de réparation des machines en panne n'induit pas le nombre d'éléments à changer : **il ne pas confondre corrélation et relation de causalité !!**

## 2. Recherche de la dépendance ou de l'indépendance linéaire ou non de deux caractères quantitatifs (tableaux à double entrée)

### 21. Cadre général et formules de calculs

La logique de construction du tableau ci-après est similaire à celle des deux types de tableaux présentés au chapitre 3.

Ici, dans la mesure où les deux caractères sont quantitatifs, on peut calculer simultanément, à l'aide du même tableau, les paramètres marginaux et conditionnels de chacune des deux variables x et y, en prolongeant le tableau de base en lignes et en colonnes.

Si l'on adjoint le calcul de la covariance (une ligne ou une colonne supplémentaires sont ajoutées au tableau), les paramètres marginaux de x et de y permettent de calculer le coefficient de corrélation linéaire, de même que le coefficient de détermination (linéaire).

Les paramètres conditionnels de x et de y permettent de calculer la variance expliquée de chacun des deux caractères (cf. chap. 3).

Ces variances expliquées permettent alors de calculer les **rapports de corrélation**, qui expriment le degré d'intensité de la relation (**non nécessairement linéaire**), si elle existe, entre y et x, et / ou entre x et y.

Ces rapports jouent le même rôle et ont la même signification que les rapports d'explication vus au chapitre 3. Leur signification s'apparente aussi à celle du coefficient de détermination $\rho^2_{xy}$, calculé lorsqu'on s'intéresse à une éventuelle liaison linéaire entre deux caractères quantitatifs (cf. point 1).

Remarque : on peut également tracer des **courbes de régression** qui résument le nuage de points représentatif de la série étudiée.

A la droite d'ajustement de y en x (D) de l'analyse linéaire correspond la courbe de régression de y en x ($C_{y/x}$) de l'analyse générale.

A la droite d'ajustement de x en y (D ') de l'analyse linéaire correspond la courbe de régression de x en y ($C_{x/y}$) de l'analyse générale.

Ci-après, on trouve le tableau de calculs complet, qui permet de déterminer notamment les paramètres marginaux et les paramètres conditionnels de chacun des deux caractères quantitatifs x et y. De même, on trouve un récapitulatif des formules appropriées :

| | | | | | | | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i \setminus y_j$ | $y_1$ | … | $y_j$ | … | $y_p$ | $n_{i.}$ | $n_{i.}\,x_i$ | $n_{i.}\,x_i^2$ | $\sum_j n_{ij}\,y_j$ | $\sum_j n_{ij}\,y_j^2$ | $\sum_j n_{ij}\,x_i\,y_j$ |
| $x_1$ | $n_{11}$ | … | $n_{1j}$ | | $n_{1p}$ | $n_{1.}$ | $n_{1.}\,x_1$ | $n_{1.}\,x_1^2$ | $\sum_j n_{1j}\,y_j$ | $\sum_j n_{1j}\,y_j^2$ | $\sum_j n_{1j}\,x_i\,y_j$ |
| … | … | … | … | … | … | … | … | … | … | … | … |
| $x_i$ | $n_{i1}$ | … | $n_{ij}$ | … | $n_{ip}$ | $n_{i.}$ | $n_{i.}\,x_i$ | $n_{i.}\,x_i^2$ | $\sum_j n_{ij}\,y_j$ | $\sum_j n_{ij}\,y_j^2$ | $\sum_j n_{ij}\,x_i\,y_j$ |
| … | … | … | … | … | … | … | … | … | … | … | … |
| $x_m$ | $n_{m1}$ | … | $n_{mj}$ | … | $n_{mp}$ | $n_{m.}$ | $n_{m.}\,x_m$ | $n_{m.}\,x_m^2$ | $\sum_j n_{mj}\,y_j$ | $\sum_j n_{mj}\,y_j^2$ | $\sum_j n_{mj}\,x_i\,y_j$ |
| $n_{.j}$ | $n_{.1}$ | … | $n_{.j}$ | … | $n_{.p}$ | $n_{..}$ | $\sum_i n_{i.}\,x_i$ | $\sum_i n_{i.}\,x_i^2$ | | | $\sum_i \sum_j n_{ij}\,x_i\,y_j$ |
| $n_{.j}\,y_j$ | $n_{.1}\,y_1$ | … | $n_{.j}\,y_j$ | … | $n_{.p}\,y_p$ | $\sum_j n_{.j}\,y_j$ | | | | | |
| $n_{.j}\,y_j^2$ | $n_{.1}\,y_1^2$ | … | $n_{.j}\,y_j^2$ | … | $n_{.p}\,y_p^2$ | $\sum_j n_{.j}\,y_j^2$ | | | | | |
| $\sum_i n_{ij}\,x_i$ | $\sum_i n_{i1}\,x_i$ | … | $\sum_i n_{ij}\,x_i$ | … | $\sum_i n_{ip}\,x_i$ | | | | | | |
| $\sum_i n_{ij}\,x_i^2$ | $\sum_i n_{i1}\,x_i^2$ | … | $\sum_i n_{ij}\,x_i^2$ | … | $\sum_i n_{ip}\,x_i^2$ | | | | | | |
| $\sum_i n_{ij}\,x_i\,y_j$ | $\sum_i n_{i1}\,x_i\,y_j$ | … | $\sum_i n_{ij}\,x_i\,y_j$ | … | $\sum_i n_{ip}\,x_i\,y_j$ | $\sum_j \sum_i n_{ij}\,x_i\,y_j$ | | | | | |

**(6)**

**(7)**

**(8)**

**(9)**

**(10)**

**Colonnes (1) et (2) : calcul des <u>paramètres marginaux de x</u>**

Colonne (1) : permet le calcul de la moyenne marginale de x : $\bar{x} = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} x_i$

Colonne (2) : permet le calcul de la variance marginale de x : $\sigma_x^2 = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} x_i^2 - \bar{x}^2$

**Lignes (8) et (9) : calcul des <u>paramètres conditionnels de x</u>**

Ligne (8) : permet le calcul des p moyennes conditionnelles de x : $\bar{x}_j = \frac{1}{n_{.j}} \sum_{j=1}^{m} n_{ij} x_i$

Ligne (9) : permet le calcul des p variances conditionnelles de x : $\sigma_{x_j}^2 = \frac{1}{n_{.j}} \sum_{i=1}^{m} n_{ij} x_i^2 - \bar{x}_j^2$

**Lignes (6) et (7) : calcul des <u>paramètres marginaux de y</u>**

Ligne (6) : permet le calcul de la moyenne marginale de y : $\bar{y} = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} y_j$

Ligne (7) : permet le calcul de la variance marginale de y : $\sigma_y^2 = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} y_j^2 - \bar{y}^2$

**Colonnes (3) et (4) : calcul des <u>paramètres conditionnels de y</u>**

Colonne (3) : permet le calcul des m moyennes conditionnelles de y : $\bar{y}_i = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} y_j$

Colonne (4) : permet le calcul des m variances conditionnelles de y :

$$\sigma_{y_i}^2 = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} y_j^2 - \bar{y}_i^2$$

**Colonne (5) ou ligne (10) : permettent indifféremment le calcul de la <u>covariance</u> $\sigma_{xy}$ :**

$$\sigma_{xy} = \frac{1}{n_{..}} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij} x_i y_j - \bar{x}\,\bar{y}$$

## 22. Relations entre paramètres marginaux et conditionnels

**Relation entre la moyenne marginale et les moyennes conditionnels**

La moyenne de la population totale est égale à la moyenne pondérée des moyennes des différents sous-populations :

$$\bar{x} = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\ \bar{x}_j \qquad \text{et :} \qquad \bar{y} = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\ \bar{y}_i$$

**Relation entre la variance marginale et les variances conditionnelles**

**Caractère x** : $$\sigma_x^2 = V(x) = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\ \bar{x}_j^2 - \bar{x}^2 + \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\ \sigma_{x_j}^2$$

Le premier terme de la somme correspond à la **variance des moyennes conditionnelles de x** ou **variance interpopulation** ou encore **variance expliquée de x** (par y) (= VEx).

Le deuxième terme correspond à la **moyenne des variances** ou **variance intrapopulation** ou encore **variance résiduelle de x** (non expliquée par y) (= VRx).

**Caractère y** : $$\sigma_y^2 = V(y) = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\ \bar{y}_i^2 - \bar{y}^2 + \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\ \sigma_{y_i}^2$$

Le premier terme de la somme correspond à la **variance des moyennes conditionnelles de y** ou **variance interpopulation** ou encore **variance expliquée de y** (par x) (= VEy).

Le deuxième terme correspond à la **moyenne des variances** ou **variance intrapopulation** ou encore **variance résiduelle de y** (non expliquée par x) (= VRy).

## 23. Les rapports de corrélation et le coefficient de corrélation linéaire

Le carré d'un rapport de corrélation, noté $\eta^2$ (on lit "éta deux"), est égal au **rapport de la variance expliquée à la variance marginale**. Il s'agit d'un **rapport d'explication**.

Contrairement au coefficient de détermination, qui est unique (que l'on considère un ajustement linéaire de y en x, ou bien de x en y), il existe deux rapports de corrélation, qui jouent le même rôle que le coefficient de détermination dans le cas linéaire :

Pour le **caractère x**, on a : $$\eta^2_{x/y} = \frac{VE_x}{\sigma_x^2} = \frac{\frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} \bar{x}_j^2 - \bar{x}^2}{\sigma_x^2}$$

Le résultat obtenu est un nombre sans dimension, qui peut être multiplié par 100 pour l'exprimer en pourcentage. Ce rapport précise l'intensité de la liaison qui existe entre x (variable expliquée) et y (variable explicative). Il exprime la **part (pourcentage) d'explication de la dispersion des valeurs de x (variance marginale de x) par rapport à des valeurs données du caractère y**.

Pour le **caractère y**, on a : $$\eta^2_{y/x} = \frac{VE_y}{\sigma_y^2} = \frac{\frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} \bar{y}_i^2 - \bar{y}^2}{\sigma_y^2}$$

Ce rapport précise l'intensité de la liaison qui existe entre y (variable expliquée) et x (variable explicative). Il exprime la **part (pourcentage) d'explication de la dispersion des valeurs de y (variance marginale de y) par rapport à des valeurs données du caractère x**.

On a toujours : $\mathbf{0 \le \rho^2_{xy} \le \eta^2_{y/x} \le \eta^2_{x/y} \le 1}$

On a : $\mathbf{\eta^2_{y/x} = \eta^2_{x/y} = 1}$ , lorsqu'il existe une **liaison fonctionnelle quelconque** (non nécessairement linéaire) entre les deux caractères x et y.

On a : $\mathbf{\eta^2_{y/x} = \eta^2_{x/y} = 0}$ , lorsqu'il y a **indépendance totale** entre les deux caractères x et y. Dans ce cas, on est certain qu'il n'existe de relation de dépendance d'aucune sorte entre les deux variables (voir point 25).

Enfin, on peut avoir : $\rho^2_{xy} = \eta^2_{y/x} = \eta^2_{x/y} = 1$ , lorsqu'il existe une **liaison fonctionnelle linéaire** entre les deux caractères x et y.

**Attention ! Ne pas confondre corrélation et causalité !**

De même que pour le coefficient de corrélation linéaire, deux caractères peuvent faire apparaître une valeur élevée des rapports de corrélation, sans que, pour autant, il existe une relation de causalité directe entre eux. Il peut également arriver que l'un des deux rapports soit pertinent et l'autre pas du tout.

Enfin, dans le cas d'un tableau de contingence, l'expression du coefficient de corrélation linéaire est la suivante :

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n_{..}} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij}\, x_i\, y_j - \bar{x}\,\bar{y}}{\sigma_x \cdot \sigma_y}$$

### 24. Les courbes de régression

Pour représenter graphiquement une série à deux caractères quantitatifs, si le nombre de modalités de chaque caractère n'est pas trop important (car le résultat devient assez vite illisible), on peut utiliser, sur un tableur, un **stéréogramme** 3D, dans lequel l'un des trois axes indique le niveau des effectifs (ou des fréquences) conjoints.

Sur papier ou sur tableur (**graphes "à bulles"**), on peut aussi construire un nuage de points particulier, en associant à chacun des points un cercle dont la surface est proportionnelle aux effectifs (ou aux fréquences) conjoints.

Mais, quelle que soit la forme graphique retenue, il est souvent malaisé d'en déduire l'existence d'une liaison (d'une dépendance) entre les deux caractères étudiés, et surtout l'intensité de cette liaison. C'est pourquoi, généralement, on va plutôt tracer des **courbes de régression**.

Les courbes de régression $C_{y/x}$ (de y en x) et $C_{x/y}$ (de x en y) vont respectivement jouer le même rôle que les droites d'ajustement linéaire (D, de y en x) et (D ', de x en y).

A la droite d'ajustement de y en x (D) de l'analyse linéaire correspond la courbe de régression de y en x ($C_{y/x}$) dans l'analyse générale.

A la droite d'ajustement de x en y (D ') de l'analyse linéaire correspond la courbe de régression de x en y ($C_{x/y}$) dans l'analyse générale.

**Ces courbes vont résumer, dans le plan, les informations fournies par le nuage de points et permettre de mettre éventuellement en évidence des relations de dépendance (non nécessairement linéaires) entre les deux caractères x et y.**

La **courbe de régression $C_{y/x}$ (de y en x)** : au lieu de tenter de représenter le triplet $(x_i ; y_j ; n_{ij})$ dans un espace à trois dimensions, on fait correspondre, à chaque modalité $x_i$, une valeur qui résume (synthétise) les couples $(y_j ; n_{ij})$. Cette valeur est donnée par la moyenne conditionnelle $\overline{y}_i$. Ainsi, pour les m modalités de x, on aura m points de coordonnées $(x_i ; \overline{y}_i)$. De cette façon, on peut tracer un nuage de points "résumé", dans le plan.

On peut dire que **la relation qui lie le caractère y au caractère x est synthétisée par la courbe de variation des moyennes conditionnelles $\overline{y}_i$ de y, en fonction des modalités $x_i$ de x.**

D'où la courbe de régression de y en x, notée $C_{y/x}$, qui porte sur les points $(x_i ; \overline{y}_i)$.

La **courbe de régression $C_{x/y}$ (de x en y)** : on fait correspondre, à chaque modalité $y_j$, une valeur qui résume (synthétise) les couples $(x_i ; n_{ij})$. Cette valeur est donnée par la moyenne conditionnelle $\overline{x}_j$. Ainsi, pour les p modalités de y, on aura p points de coordonnées $(y_j ; \overline{x}_j)$.

On peut dire que **la relation qui lie le caractère x au caractère y est synthétisée par la courbe de variation des moyennes conditionnelles $\overline{x}_j$ de x, en fonction des modalités $y_j$ de y.**

D'où la courbe de régression de x en y, notée $C_{x/y}$, qui porte sur les points $(y_j ; \overline{x}_j)$.

Remarques :

a) La somme des carrés des écarts de chaque point du nuage à chacune des courbes de régression est "minimale" (on fera le parallèle avec les droites d'ajustement linéaire ; voir point 1).

b) Le point d'intersection G des deux courbes de régression correspond au point moyen (ou barycentre, ou centre de gravité) du nuage de points. On le note : $G(\overline{x} ; \overline{y})$.

c) Lorsque : $\eta^2_{y/x} = \eta^2_{x/y} = 1$, il y a **liaison fonctionnelle** entre les deux caractères x et y. Dans ce cas, **les deux courbes de régression sont confondues**.

d) Lorsque : $\eta^2_{y/x} = \eta^2_{x/y} = 0$, il y a **indépendance totale** entre les deux caractères x et y. Dans ce cas, **les deux courbes de régression forment deux droites perpendiculaires** :

- la courbe de régression $C_{y/x}$ (de y en x) est une droite d'ordonnée $\overline{y} = \overline{y}_i$ ;

Dans ce cas, on a également : $f_{i/j} = f_{i.}$ (voir point 25).

- la courbe de régression $C_{x/y}$ (de x en y) est une droite d'abscisse $\overline{x} = \overline{x}_j$.

Dans ce cas, on a également : $f_{j/i} = f_{.j}$ (voir point 25).

e) On a toujours : **pente de $C_{x/y}$ ≥ pente de $C_{y/x}$**.

L'égalité a lieu lorsque les deux courbes sont confondues (relation fonctionnelle).

**Variance expliquée de y** : $$VEy = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} \overline{y}_i^2 - \overline{y}^2 = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} (\overline{y}_i - \overline{y})^2$$

Elle représente la part de la variance marginale de y, qui est **expliquée par la courbe de régression de y en x**.

Graphiquement, pour chaque point du nuage, il s'agit des carrés des écarts pondérés de la courbe de régression $C_{y/x}$ (de y en x) à la moyenne marginale de y (cette dernière apparaît sous la forme d'une droite horizontale sur le graphique).

**Variance résiduelle de y** : $$VRy = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.} \sigma_{y_i}^2$$

Elle représente la part de la variance marginale de y, **non expliquée** par la courbe de régression de y en x.

Graphiquement, pour chaque point du nuage, il s'agit des carrés des écarts pondérés de chacun des points du nuage à la courbe de régression $C_{y/x}$ (de y en x).

De même :

**Variance expliquée de x** : $$VEx = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} \overline{x}_j^2 - \overline{x}^2 = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} (\overline{x}_j - \overline{x})^2$$

Elle représente la part de la variance marginale de x, qui est **expliquée par la courbe de régression de x en y**.

Graphiquement, pour chaque point du nuage, il s'agit des carrés des écarts pondérés de la courbe de régression $C_{x/y}$ (de x en y) à la moyenne marginale de x (cette dernière apparaît sous la forme d'une droite verticale sur le graphique).

**Variance résiduelle de x** : $VRx = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} \sigma^2_{x_j}$

Elle représente la part de la variance marginale de y, **non expliquée** par la courbe de régression de x en y.

Graphiquement, pour chaque point du nuage, il s'agit des carrés des écarts pondérés de chacun des points du nuage à la courbe de régression $C_{x/y}$ (de x en y).

## 25. Indépendance totale entre deux caractères quantitatifs

On montre que l'indépendance entre deux caractères se manifeste quand on a : $f_{i/j} = f_{i.}$ c'est-à-dire lorsque les fréquences conditionnelles de x (à j fixé = à colonne fixée) sont égales aux fréquences marginales de x. En ce cas, les fréquences conditionnelles ne sont plus dépendantes de la colonne envisagée dans le tableau. Cela signifie que x est indépendant de y. Tout se passe comme si l'on raisonnait seulement sur une distribution à un caractère (x), l'information apportée par y étant nulle pour x.

De même, quand on a : $f_{j/i} = f_{.j}$ c-à-d lorsque les fréquences conditionnelles de y (à i fixé = à ligne fixée) sont égales aux fréquences marginales de y. Dans ce cas, les fréquences conditionnelles ne sont plus dépendantes de la ligne envisagée dans le tableau. Cela signifie que y est indépendant de x. Tout se passe comme si l'on raisonnait seulement sur une distribution à un caractère (y), l'information apportée par x étant nulle pour y.

Dans le cas des caractères quantitatifs, on montre de plus que lorsque x est indépendant de y, on a : $\overline{x} = \overline{x}_j \quad \forall j$ c'est-à-dire que les moyennes conditionnelles de x (les $\overline{x}_j$ ) qu'on peut calculer pour chacune des **p** colonnes du tableau (c-à-d pour chacune des modalités j du caractère y) sont égales à la moyenne marginale de x, soit $\overline{x}$ . Les moyennes de chaque colonne du tableau sont alors confondues et égales à la moyenne marginale (colonne $n_{i.}$ du tableau).

De même, quand y est indépendant de x, on a : $\overline{y} = \overline{y}_i \quad \forall i$ c'est-à-dire que les moyennes conditionnelles de y (les $\overline{y}_i$ ) qu'on peut calculer pour chacune des **m** lignes du tableau (c-à-d pour chacune des modalités i du caractère x) sont égales à la moyenne marginale de y, soit $\overline{y}$ . Les moyennes de chaque ligne du tableau sont alors confondues et égales à la moyenne marginale (ligne $n_{.j}$ du tableau).

Exemple :

| $x_i$ \ $y_j$ | 7 | 9 | 11 | 13 | $n_{i.}$ |
|---|---|---|---|---|---|
| **1** | 4 | 12 | 24 | 8 | 48 |
| **2** | 5 | 15 | 30 | 10 | 60 |
| **3** | 2 | 6 | 12 | 4 | 24 |
| **4** | 1 | 3 | 6 | 2 | 12 |
| $n_{.j}$ | 12 | 36 | 72 | 24 | 144 |

Dans ce tableau, les lignes sont proportionnelles entre elles, de même que les colonnes. Il en résulte que les deux caractères x et y sont indépendants.

On peut vérifier que l'on a ici : $\overline{x} = \overline{x}_j = 2$ et que : $\overline{y} = \overline{y}_i = 10{,}5$ .

Par exemple, la moyenne de la 1ère colonne est égale à :

$$\overline{x}_1 = \frac{\sum_{i=1}^{4} x_i \,.\, n_{i1}}{n_{.1}} = \frac{(1 \times 4) + (2 \times 5) + (3 \times 2) + (4 \times 1)}{12} = \frac{24}{12} = 2$$

On trouve également 2 pour les trois colonnes suivantes, ainsi que pour la dernière colonne du tableau qui permet de calculer la moyenne marginale de x :

$$\overline{x} = \frac{\sum_{i=1}^{4} x_i \,.\, n_{i.}}{n_{..}} = \frac{(1 \times 48) + (2 \times 60) + (3 \times 24) + (4 \times 12)}{144} = \frac{288}{144} = 2$$

De même, la moyenne de la 1ère ligne est égale à :

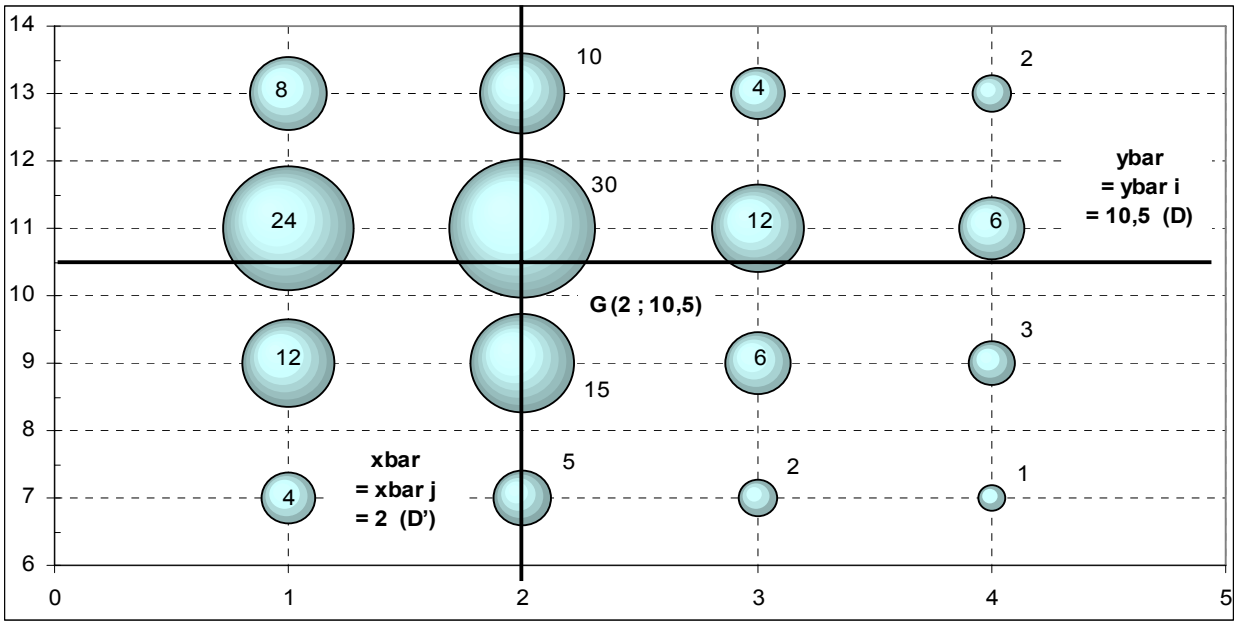
$$\overline{y}_1 = \frac{\sum_{j=1}^{4} y_j \,.\, n_{1j}}{n_{1.}} = \frac{(7 \times 4) + (9 \times 12) + (11 \times 24) + (13 \times 8)}{48} = \frac{504}{48} = 10{,}5$$

On trouve également 10,5 pour les trois lignes suivantes, ainsi que pour la dernière ligne du tableau qui permet de calculer la moyenne marginale de y :

$$\overline{y} = \frac{\sum_{j=1}^{4} y_j \,.\, n_{.j}}{n_{..}} = \frac{(7 \times 12) + (9 \times 36) + (11 \times 72) + (13 \times 24)}{144} = \frac{1\,512}{144} = 10{,}5$$

Dans cet exemple, on peut vérifier qu'on a bien : $\rho_{xy} = \rho^2_{xy} = 0$

Enfin, lorsque x et y sont indépendants, les droites (D) et (D ') sont perpendiculaires.




Lorsque d'une part $\rho = 0$ et d'autre part $f_{i/j} = f_{i.}$ et $\overline{x} = \overline{x}_j$ , de même que $f_{j/i} = f_{.j}$ et $\overline{y} = \overline{y}_i$ , alors on est certain qu'il n'y a aucune liaison (même non linéaire) entre les deux caractères x et y.

### 26. Exemples

**Exemple 1**

Une enquête, réalisée auprès de 350 lycéens, a permis de construire le tableau suivant, dans lequel :

- la **variable x** représente le montant mensuel moyen (exprimé en euros) dont un lycéen dispose sous forme d'argent de poche ;

- la **variable y** représente le montant mensuel moyen (exprimé en euros) qu'il utilise pour consommer :

| $x_i$ \ $y_j$ | **20** | **50** | **80** | **110** |
|---|---|---|---|---|
| **50** | 47 | 8 | | |
| **100** | 11 | 112 | 46 | |
| **150** | 2 | 6 | 58 | 60 |

Objectif :

on se propose d'analyser la dépendance ou l'indépendance des deux caractères x et y, en calculant le coefficient de corrélation linéaire $\rho_{xy}$ et les rapports de corrélation $\eta^2_{y/x}$ et $\eta^2_{x/y}$ .

A cet effet, on détermine les paramètres marginaux de x et de y, le coefficient de corrélation linéaire $\rho_{xy}$ et les paramètres de la droite (D) d'ajustement de y en x.

De même, on détermine les paramètres conditionnels de x et de y, ainsi que les variances expliquées de x et de y.

Un tableau de calculs permet de déterminer tous les paramètres utiles.

Dans un premier temps, nous nous intéressons aux représentations graphiques appropriées à ce type de série statistique.

Dans le cas d'un croisement entre deux caractères quantitatifs et de l'utilisation d'un tableau à double entrée, la représentation graphique la plus appropriée est le **graphe à bulles**. Celui-ci représente le nuage des points (pondérés par les effectifs), sous la forme de **cercles, dont la surface est proportionnelle aux effectifs conjoints de chacune des cases du tableau**.

On peut également représenter sur un tel graphique la droite d'ajustement de y en x, qui "résume" le nuage de points.

Si l'on dispose d'un tableur, on peut également représenter cette série au moyen d'un **stéréogramme**.

**A.**

**1°)** Quels sont la **population**, les **caractères** et **types** de la série statistique envisagée ?

**2°)** Combien de lycéens disposent de 100 € par mois et, simultanément, consomment pour 80 € par mois ?

Combien de lycéens disposent de 50 € par mois et, simultanément, consomment pour 110 € par mois ?

**3°)** A partir du tableau des effectifs conjoints ci-dessus, réaliser un **graphique** approprié pour représenter correctement cette série.

**B.**

**4°)** Établir le tableau des **fréquences conjointes théoriques** d'indépendance.

**5°)** A partir de la comparaison du tableau des **fréquences conjointes réelles** et de celui des fréquences conjointes théoriques d'indépendance, analyser la dépendance ou l'indépendance des deux caractères x et y.

**C.**

**6°)** Calculer les **paramètres marginaux de chacune des variables x et y**. Commenter les résultats.

**7°)** Calculer la **covariance** entre les deux caractères x et y, de deux façons différentes. A partir de cette valeur, calculer le **coefficient de corrélation linéaire** entre x et y, ainsi que le **coefficient de détermination**. Interpréter les résultats.

**8°)** Calculer, le cas échéant, les **coefficients de la droite d'ajustement de y en x** et tracer cette droite sur le graphique précédent.

**9°)** Un ajustement linéaire de x en y a-t-il ici une signification ? Pourquoi ?

**D.**

**10°)** Calculer les **paramètres conditionnels du caractère x**. On s'intéresse plus particulièrement aux lycéens dont les dépenses de consommation mensuelles moyennes sont de 20 €. Quel est le montant moyen du revenu des lycéens concernés ?

**11°)** Calculer les **paramètres conditionnels du caractère y**. On s'intéresse plus particulièrement aux lycéens dont le revenu mensuel moyen est de 150 €. Quel est le montant moyen des dépenses de consommation des lycéens concernés ?

**E.**

**12°)** Calculer le **rapport de corrélation de y en x**, après avoir rappelé les relations qui existent, d'une part entre les moyennes conditionnelles de y et la moyenne marginale de y, d'autre part entre les variances conditionnelles de y et la variance marginale de y.

## A. 1°)

Population : 350 lycéens enquêtés.

Caractères : - "revenu" mensuel moyen d'un lycéen (xi) ;
- dépense mensuelle moyenne de consommation (yj).

Type : quantitatif discret, pour chacun des deux caractères.

**2°)**

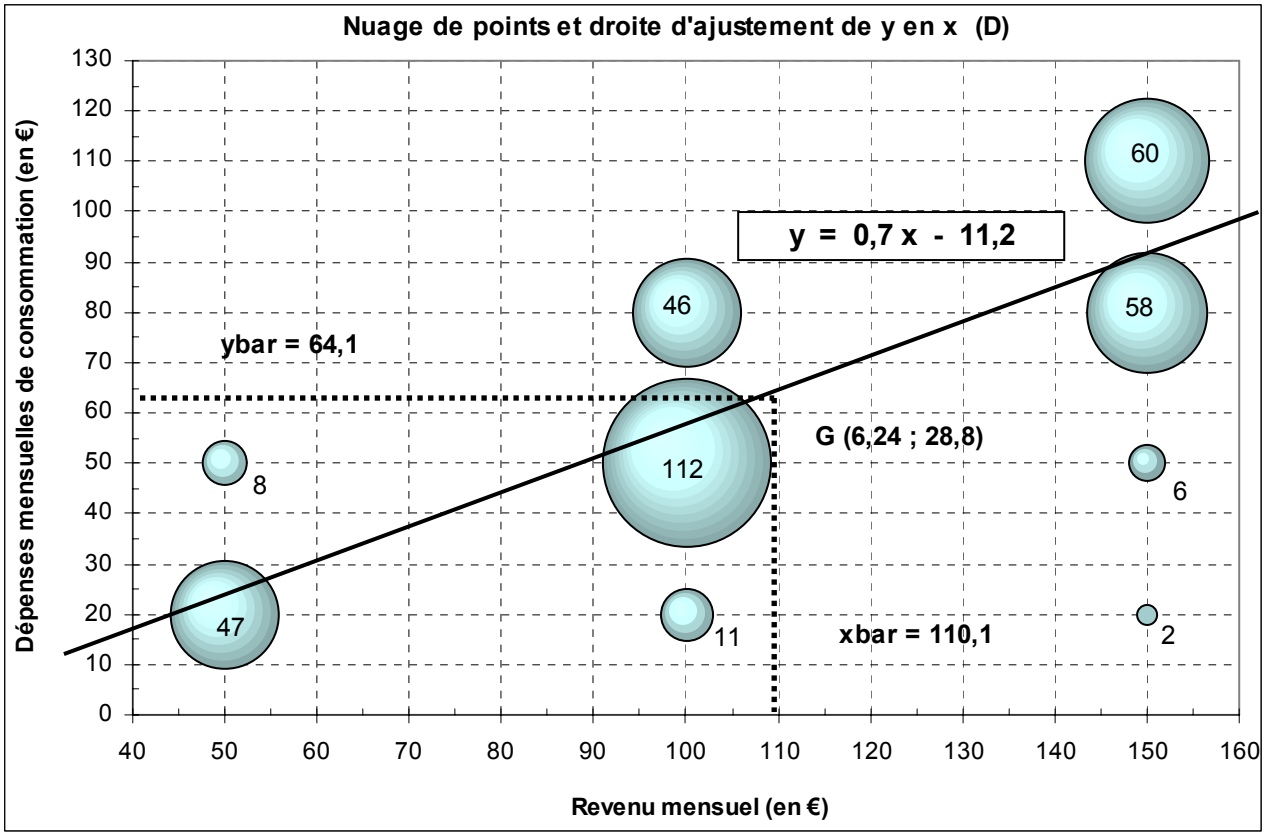
Rappel : afin de repérer aisément une case du tableau, on utilise une notation matricielle.

Si l'on veut connaître le nombre de **lycéens qui disposent de 100 € par mois et qui, simultanément, consomment 80 € par mois**, on considère simultanément la deuxième modalité de x et la troisième modalité de y, soit l'effectif conjoint $\mathbf{n_{23} = 46}$ **lycéens**.

De même, le nombre de **lycéens disposant de 50 € et consommant 110 €** est repéré par l'effectif conjoint $\mathbf{n_{14} = 0}$ **lycéen**.

**3°)**

**Graphe à bulles** : une représentation graphique appropriée de cette série statistique à deux caractères quantitatifs consiste à tracer un **nuage dont les points ont une surface proportionnelle aux effectifs conjoints correspondants** .




Si l'on veut tracer des cercles dont la surface soit rigoureusement proportionnelle aux effectifs, on repère l'effectif le plus important est l'on fixe arbitrairement le rayon du cercle correspondant.

Ensuite, pour chacun des autres cercles du nuage de points, on calcule leur rayon à partir du rapport de l'effectif, relatif à un cercle donné, à l'effectif du cercle de référence.

Exemple

Ici, l'effectif le plus important est égal à **112**. Posons arbitrairement que le rayon du cercle qui correspond à cet effectif est égal à **1 cm**.

Considérons maintenant l'effectif **47** de la première case du tableau.

On a : $\pi R^2_{(47)} = 47$ . Dans le même temps : $\pi R^2_{(112)} = 112$ .

Posons le rapport :

$$\frac{\pi R^2_{(47)}}{\pi R^2_{(112)}} = \frac{47}{112} \Leftrightarrow R^2_{(47)} = \frac{47}{112} R^2_{(112)} \Leftrightarrow R_{(47)} = \sqrt{\frac{47}{112}} R_{(112)} = 0{,}65\ R_{(112)}$$

Comme on a posé : $R_{(112)} = 1$ cm , alors : $R_{(47)} = 0{,}65$ cm

**Tableau de calculs** :

| $x_i$ \ $y_j$ | **20** | **50** | **80** | **110** | $n_{i.}$ | $n_{i.} x_i$ | $n_{i.} x_i^2$ | $\sum_j n_{ij} y_j$ | $\sum_j n_{ij} y_j^2$ | $\sum_j n_{ij} x_i y_j$ |
|---|---|---|---|---|---|---|---|---|---|---|
| **50** | 47 | 8 | | | 55 | 2 750 | 137 500 | 1 340 | 38 800 | 67 000 |
| **100** | 11 | 112 | 46 | | 169 | 16 900 | 1 690 000 | 9 500 | 578 800 | 950 000 |
| **150** | 2 | 6 | 58 | 60 | 126 | 18 900 | 2 835 000 | 11 580 | 1 113 000 | 1 737 000 |
| $n_{.j}$ | 60 | 126 | 104 | 60 | 350 | 38 550 | 4 662 500 | | | 2 754 000 |
| $n_{.j} y_j$ | 1 200 | 6 300 | 8 320 | 6 600 | 22 420 | | | | | |
| $n_{.j} y_j^2$ | 24 000 | 315 000 | 665 600 | 726 000 | 1 730 600 | | | | | |
| $\sum_i n_{ij} x_i$ | 3 750 | 12 500 | 13 300 | 9 000 | | | | | | |
| $\sum_i n_{ij} x_i^2$ | 272 500 | 1 275 000 | 1 765 000 | 1 350 000 | | | | | | |
| $\sum_i n_{ij} x_i y_j$ | 75 000 | 625 000 | 1 064 000 | 990 000 | 2 754 000 | | | | | |

**B. 4°)**

On peut mettre en œuvre des tableaux d'indépendance théorique, comme dans le cas de deux caractères qualitatifs (cf. ch. 2) ou d'un caractère qualitatif et d'un caractère quantitatif (cf. ch. 3).

**Méthode 1**

**Tableau 1 : effectifs réels**

| xi \ yj | 20 | 50 | 80 | 110 | n i. |
|---|---|---|---|---|---|
| **50** | 47 | 8 | | | 55 |
| **100** | 11 | 112 | 46 | | 169 |
| **150** | 2 | 6 | 58 | 60 | 126 |
| **n.j** | 60 | 126 | 104 | 60 | 350 |

Pour construire le tableau des effectifs conjoints théoriques d'indépendance, on utilise la formulation suivante : **n ij = (n i. x n .j) / n ..**

Comme il s'agit d'effectifs portant sur des nombres entiers, on arrondit les résultats obtenus, pour chaque case du tableau, à l'entier le plus proche.

**Tableau 2 : effectifs théoriques d'indépendance (valeurs arrondies)**

| xi \ yj | 20 | 50 | 80 | 110 | n i. |
|---|---|---|---|---|---|
| **50** | 9 | 20 | 16 | 9 | 55 |
| **100** | 29 | 61 | 50 | 29 | 169 |
| **150** | 22 | 45 | 37 | 22 | 126 |
| **n.j** | 60 | 126 | 104 | 60 | 350 |

Comme on l'a vu plus haut, il est plus commode de comparer les tableaux de fréquences conjointes réelles et théoriques d'indépendance. Pour construire le tableau des fréquences conjointes théoriques d'indépendance (exprimées en pourcentage), on utilise la formulation suivante : **f ij = (f i. x f .j) / 100** :

**Tableau 3 : fréquences conjointes réelles (en %)**

| Xi \ yj | 20 | 50 | 80 | 110 | f i. |
|---|---|---|---|---|---|
| **50** | **13,4** | 2,3 | 0,0 | 0,0 | 15,7 |
| **100** | 3,1 | **32,0** | 13,1 | 0,0 | 48,3 |
| **150** | 0,6 | 1,7 | **16,6** | **17,1** | 36,0 |
| **f .j** | 17,1 | 36,0 | 29,7 | 17,1 | 100,0 |

**Tableau 4 : fréquences conjointes théoriques d'indépendance (en %)**

| Xi \ yj | 20 | 50 | 80 | 110 | f i. |
|---|---|---|---|---|---|
| **50** | **2,7** | 5,7 | 4,7 | 2,7 | 15,7 |
| **100** | 8,3 | **17,4** | 14,3 | 8,3 | 48,3 |
| **150** | 6,2 | 13,0 | **10,7** | **6,2** | 36,0 |
| **f .j** | 17,1 | 36,0 | 29,7 | 17,1 | 100,0 |

**5°)**

**a)** Considérons les deux tableaux **en lignes** : si la consommation (y) était totalement indépendante du revenu (x), on devrait retrouver, pour chaque modalité de x (le revenu), les mêmes pourcentages dans chaque ligne du tableau (cf. les f .j réelles : 17,1 %, 36,0 %, 29,7 % et 17,1 %).

De fait, en comparant les deux tableaux de fréquences conjointes, on observe**, pour les lycéens dont le revenu moyen mensuel est de 50 €**, une nette surreprésentation dans le cas où la consommation mensuelle moyenne est de 20 € (13,4 % contre 2,7 % dans l'hypothèse d'indépendance). Pour les autres modalités de y (consommation), on constate une légère sous-représentation.

Pour les **lycéens dont le revenu moyen mensuel est de 100 €**, on observe une nette surreprésentation dans le cas où la consommation mensuelle moyenne est de 50 € (32,0 % contre 17,4 % dans l'hypothèse d'indépendance). Pour les autres modalités de y (consommation), on constate une sous-représentation plus ou moins prononcée.

Pour les **lycéens dont le revenu moyen mensuel est de 150 €**, on observe une forte surreprésentation dans les cas où la consommation mensuelle moyenne est de 80 € et 110 € (respectivement 16,6 % contre 10,7 % dans l'hypothèse d'indépendance, et 17,1 % contre 6,2 %). Pour les deux autres modalités de y (consommation), on constate une nette sous-représentation.

**b)** Considérons les deux tableaux **en colonnes** : si le revenu (x) était totalement indépendant de la consommation (y), on devrait retrouver, pour chaque modalité de y (la consommation), les mêmes pourcentages dans chaque colonne du tableau (cf. les f i. réelles : 15,7 %, 48,3 % et 36,0 %).

**Attention !** Bien noter que **le raisonnement en colonnes, s'il est effectivement possible ici d'un point de vue "mécanique", n'est pas valide dès lors qu'on se place sur le plan de la recherche d'une relation de cause à effet (relation de causalité) entre les deux caractères ! En aucun cas, la consommation n'induit le revenu !**

En **conclusion** : au vu des valeurs obtenues, on peut dire que **la consommation des lycéens est très largement dépendante du niveau de leur revenu**.

**Méthode 2**

**Tableau 5 : rapport des effectifs réels aux effectifs théoriques d'indépendance (en %)**

| xi \ yj | 20 | 50 | 80 | 110 |
|---|---|---|---|---|
| **50** | **422,2%** | -60,0% | -100,0% | -100,0% |
| **100** | -62,1% | **83,6%** | **-8,0%** | -100,0% |
| **150** | -90,9% | -86,7% | **56,8%** | **172,7%** |

**Tableau 3 : fréquences conjointes réelles (en %)**

| xi \ yj | 20 | 50 | 80 | 110 |
|---|---|---|---|---|
| **50** | **13,4** | 2,3 | 0,0 | 0,0 |
| **100** | 3,1 | **32,0** | **13,1** | 0,0 |
| **150** | 0,6 | 1,7 | **16,6** | **17,1** |

Rappel : les pourcentages du tableau 5 sont obtenus de la façon suivante :

$$\left( \frac{\text{effectif conjoint réel}}{\text{effectif conjoint théorique d'indépendance}} - 1 \right) \times 100$$

Par exemple : [(47 / 9) - 1] x 100 = + 422,2 %

L'analyse simultanée des tableaux 3 et 5 est beaucoup moins aisée et précise que lorsqu'on peut disposer d'une valeur quantifiée comme le coefficient de corrélation linéaire ou le rapport de corrélation.

On peut cependant remarquer, en considérant dans le tableau 3 les seules fréquences conjointes réelles qui dépassent 10% (ici, leur somme représente 92,3% de l'effectif total) ou celles qui dépassent 15% (ici, leur somme représente 65,7% de l'effectif total), **que les pourcentages correspondants du tableau 5 sont aussi les plus élevés** (en faisant abstraction des cases qui correspondent à une fréquence réelle nulle dont l'interprétation n'est pas très pertinente).

**Conclusion** : si l'on ne calculait pas le coefficient de corrélation linéaire, le constat précédent permettrait tout de même de dire que les deux caractères x et y ne sont **pas indépendants**, sans toutefois pouvoir donner un ordre de grandeur pertinent de l'intensité de la liaison existant entre ces deux caractères.

On peut simplement dire ici que la consommation des lycéens est très largement dépendante du niveau de leur revenu.

**C. 6°)**

**Paramètres marginaux de x**

$$\bar{x} = \frac{38\,550}{350} = 110{,}14 \text{ euros} \qquad \boxed{\bar{x} = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\, x_i}$$

$$\sigma_x^2 = \frac{4\,662\,500}{350} - 110{,}14^{\,2} = 1189{,}98 \text{ euros}^2 \qquad \boxed{\sigma_x^2 = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\, x_i^2 - \bar{x}^{\,2}}$$

$\sigma_x$ = 34,50 euros

**$CV_x$ = 34,50 / 110,14 = 0,31**. La valeur du coefficient de variation caractérise une **dispersion moyenne des revenus autour du revenu moyen de l'échantillon** des 350 lycéens.

**Paramètres marginaux de y**

$$\bar{y} = \frac{22\,420}{350} = 64{,}06 \text{ euros} \qquad \boxed{\bar{y} = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\, y_j}$$

$$\sigma_y^2 = \frac{1350\,000}{350} - 64{,}06^{\,2} = 841{,}25 \text{ euros}^2 \qquad \boxed{\sigma_y^2 = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\, y_j^2 - \bar{y}^{\,2}}$$

$\sigma_y$ = 29,00 euros

**$CV_y$ = 29,00 / 64,06 = 0,45**. La valeur du coefficient de variation caractérise **une dispersion moyenne (mais significativement plus élevée que la dispersion sur les revenus) des dépenses mensuelles de consommation autour de la dépense mensuelle moyenne de l'échantillon** des 350 lycéens.

**7°) Calculs de la covariance et du coefficient de corrélation linéaire**

La valeur de la **covariance** est déterminée, soit à partir du total de la dernière colonne du tableau, soit à partir du total de la dernière ligne du tableau. On obtient :

$$\boxed{\sigma_{xy} = \frac{1}{n_{..}} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij}\, x_i\, y_j - \bar{x}\,\bar{y}}$$

$$\sigma_{xy} = \frac{2\,754\,000}{350} - 110{,}14 \times 64{,}06 = 813{,}13$$

On calcule ensuite le **coefficient de corrélation linéaire** et le **coefficient de détermination** :

$$\boxed{\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{\dfrac{1}{n_{..}} \displaystyle\sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij}\, x_i\, y_j - \bar{x}\,\bar{y}}{\sigma_x \cdot \sigma_y}}$$

$$\rho_{xy} = \frac{813{,}13}{34{,}50 \times 29{,}00} = 0{,}81 \text{ et : } \rho_{xy}^2 = 0{,}66$$

**8°)**

**La taille de l'échantillon (n = 350) et la valeur du coefficient de corrélation linéaire ($\rho_{xy}$ = 0,81) sont ici suffisantes pour qu'un ajustement linéaire entre les variables y et x soit pertinent.**

Si l'on calcule les coefficients de la **droite d'ajustement de y en x (D)**, on obtient :

$$a = \frac{\frac{1}{n_{..}} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij} (x_i - \bar{x})(y_j - \bar{y})}{\frac{1}{n_{..}} \sum_{i=1}^{m} (x_i - \bar{x})^2} = \frac{\frac{1}{n_{..}} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij}\, x_i\, y_j - \bar{x}\bar{y}}{\frac{1}{n_{..}} \sum_{i=1}^{m} (x_i - \bar{x})^2} = \frac{\sigma_{xy}}{\sigma_x^2} = \rho \frac{\sigma_y}{\sigma_x}$$

$$b = \bar{y} - a\bar{x}$$

$$a = 0{,}81 \times \frac{29{,}00}{34{,}50} = 0{,}68 \quad \text{et :} \quad b = 64{,}06 - 0{,}68 \times 110{,}14 = -11{,}21$$

D'où : (D) **y = 0,68 x - 11,21**

**9°)**

Un ajustement linéaire de x en y, c-à-d lorsque y joue le rôle de variable explicative et x celui de variable expliquée (dépendante), n'a **ici aucune signification** :

**ce n'est pas le montant des dépenses de consommation qui induit le niveau du revenu !!**

**D. 10°)**

On peut raisonner sur une seule colonne ou une seule ligne. Dans ce cas, cela signifie qu'on raisonne sur une **sous-population** particulière.

Nous allons envisager successivement la première colonne et la première ligne du tableau.

**Première colonne du tableau**

Parmi les 350 lycéens, il existe une **sous-population de 60 lycéens qui dépensent 20 € par mois en moyenne**.

Les paramètres relatifs à cette colonne du tableau, que nous allons calculer, sont appelés **paramètres conditionnels de x**. Il y en a autant que de colonnes (modalités de y) dans le tableau.

De manière générale, on a :

**Moyennes conditionnelles de x** : $$\bar{x}_j = \frac{1}{n_{.j}} \sum_{j=1}^{m} n_{ij}\, x_i$$

Il y a autant de moyennes conditionnelles de x qu'il y a de modalités de y.

**Variances conditionnelles de x** :

$$\sigma^2_{x_j} = \frac{1}{n_{.j}} \sum_{i=1}^{m} n_{ij}\,(x_i - \bar{x}_j)^2 = \frac{1}{n_{.j}} \sum_{i=1}^{m} n_{ij}\, x_i^2 - \bar{x}_j^{\,2}$$

Il y a autant de variance conditionnelles de x qu'il y a de modalités de y.

**Coefficient de variation de $x_j$** : $$CV_{x_j} = \frac{\sigma_{x_j}}{\bar{x}_j}$$

Application à l'exemple (on raisonne sur la seule première colonne du tableau) :

| x i | n i1 | ni 1 xi | n i1 xi ² |
|---|---|---|---|
| **50** | 47 | 2 350 | 117 500 |
| **100** | 11 | 1 100 | 110 000 |
| **150** | 2 | 300 | 45 000 |
| **Total (n .1)** | 60 | 3 750 | 272 500 |

**Moyenne conditionnelle = $\bar{x}_1 = \frac{3\,750}{60}$ = 62,50 € / mois**.

Les lycéens qui consomment en moyenne 20 € / mois ont un revenu moyen de 62,50 € / mois.

**Variance conditionnelle = $\sigma^2_{x_1} = \frac{272\,500}{60} - (62,5)^2$ = 635,42 €² / mois**

**Ecart-type conditionnel = 25,21 € / mois**

**Coefficient de variation = $CV_{x_1} = \frac{25,21}{62,50}$ = 0,40**

La valeur du CV caractérise ici une dispersion moyenne autour de la valeur moyenne observée.

**11°)**

**Première ligne du tableau**

Parmi les 350 lycéens, il existe **une sous-population de 55 lycéens qui disposent de 50 € par mois en moyenne**.

Les paramètres relatifs à cette ligne du tableau, que nous allons calculer, sont appelés **paramètres conditionnels de y**. Il y en a autant que de ligne (modalités de x) dans le tableau.

De manière générale, on a :

**Moyennes conditionnelles de y** : $$\bar{y}_i = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij}\, y_j$$

Il y a autant de moyennes conditionnelles de y qu'il y a de modalités de x.

**Variances conditionnelles de y** :

$$\sigma^2_{y_i} = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij}\,(y_j - \bar{y}_i)^2 = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij}\, y_j^2 - \bar{y}_i^{\,2}$$

Il y a autant de variance conditionnelles de y qu'il y a de modalités de x.

**Coefficient de variation de $y_i$** : $$CV_{y_i} = \frac{\sigma_{y_i}}{\bar{y}_i}$$

Application à l'exemple (on raisonne sur la seule première ligne du tableau) :

| y j | n 1j | n 1j yj | n 1j yj² |
|---|---|---|---|
| **20** | 47 | 940 | 18 800 |
| **50** | 8 | 400 | 20 000 |
| **Total (n 1 .)** | 55 | 1 340 | 38 800 |

**Moyenne conditionnelle = $\bar{y}_1 = \frac{1340}{55}$ = 24,36 € / mois**.

Les lycéens qui disposent en moyenne de 50 € / mois dépensent en moyenne 24,36 € / mois.

**Variance conditionnelle = $\sigma^2_{y_1} = \frac{38\,800}{55} - (24{,}36)^2$ = 111,87 €² / mois**

**Ecart-type conditionnel = 10,58 € / mois**

**Coefficient de variation =** $CV_{y_1} = \frac{10,58}{24,36}$ **= 0,43**

La valeur du CV caractérise ici une dispersion moyenne autour de la valeur moyenne observée.

Pour information, les deux pages suivantes présentent l'ensemble des résultats relatifs à chacune des lignes et à chacune des colonnes du tableau.

**10°)** **D.** **Paramètres conditionnels de x**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Moy cond xbar1 = | 62,50 | Var cond x1 = | 635,42 | Ectyp cond x 1 = | 25,21 | CV x1 = | 0,40 |
| Moy cond xbar2 = | 99,21 | Var cond x2 = | 277,15 | Ectyp cond x 2 = | 16,65 | CV x2 = | 0,17 |
| Moy cond xbar3 = | 127,88 | Var cond x3 = | 616,68 | Ectyp cond x 3 = | 24,83 | CV x3 = | 0,19 |
| Moy cond xbar4 = | 150,00 | Var cond x4 = | 0,00 | Ectyp cond x 4 = | 0,00 | CV x4 = | 0,00 |

$$\bar{x}_j = \frac{1}{n_{.j}} \sum_{i=1}^{m} n_{ij}\, x_i \qquad \sigma^2_{x_j} = \frac{1}{n_{.j}} \sum_{i=1}^{m} n_{ij}\, x_i^2 - \bar{x}_j^{\,2}$$

Considérer les seuls lycéens dont les dépenses de consommation mensuelles moyennes s'élèvent à 20 € revient à envisager une **sous-population** particulière parmi les 350 lycéens de l'échantillon.

Tout se passe comme si l'on avait affaire à une série à un seul caractère (x) et que les effectifs correspondant à chacune des modalités de x étaient ceux de la modalité 20 du caractère y. Au total, **60 lycéens** sont concernés (cf. tableau général de calcul).

Pour calculer la moyenne du revenu mensuel moyen de ces 60 lycéens, on procède classiquement en multipliant les valeurs de chaque modalité de x par l'effectif correspondant. En sommant l'ensemble des produits obtenus, on obtient ici :

50 x 47 + 100 x 11 + 150 x2 = **3 750** (cf. tableau général de calcul).

La moyenne recherchée, qu'on appelle **moyenne conditionnelle de x, pour la modalité 20 du caractère y**, est égale à :

**moyenne conditionnelle de x, pour la modalité 20 du caractère y = 3 750 / 60 = 62,50 €.**

En d'autres termes, les 60 lycéens dépensant en moyenne 20 € par mois ont un revenu moyen mensuel de 62,50 €.

On a :

$CV_{x1} = 0,40 \quad CV_{x2} = 0,17 ; \quad CV_{x3} = 0,19 ; \quad CV_{x4} = 0,00.$

Pour les colonnes 2 et 3 du tableau, la valeur des coefficients de variation caractérise une dispersion faible des revenus autour du revenu moyen des lycéens consommant respectivement 50 € et 80 € par mois.

Dans la mesure où les effectifs de la colonne 4 ne sont concernés que par un seul effectif conjoint, il est logique d'avoir une dispersion nulle, ce qui conduit à un coefficient de variation nul lui aussi.

**11°)** **Paramètres conditionnels de y**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Moy cond ybar1 = | 24,36 | Var cond y1 = | 111,87 | Ectyp cond y 1 = | 10,58 | CV y1 = | 0,43 |
| Moy cond ybar2 = | 56,21 | Var cond y2 = | 264,95 | Ectyp cond y 2 = | 16,28 | CV y2 = | 0,29 |
| Moy cond ybar3 = | 91,90 | Var cond y3 = | 386,85 | Ectyp cond y 3 = | 19,67 | CV y3 = | 0,21 |

$$\bar{y}_i = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij}\, y_j \qquad \sigma^2_{y_i} = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij}\, y_j^2 - \bar{y}_i^{\,2}$$

Considérer les seuls lycéens dont les revenus mensuels moyens se montent à 150 € revient à envisager une **sous-population** particulière parmi les 350 lycéens de l'échantillon.

Tout se passe comme si l'on avait affaire à une série à un seul caractère (y) et que les effectifs correspondant à chacune des modalités de y étaient ceux de la modalité 150 du caractère x. Au total, **126 lycéens** sont concernés (cf. tableau général de calcul).

Pour calculer la moyenne des dépenses de consommation mensuelles moyennes de ces 126 lycéens, on procède classiquement en multipliant les valeurs de chaque modalité de y par l'effectif correspondant. En sommant l'ensemble des produits obtenus, on obtient ici : 20 x 2 + 50 x 6 + 80 x 58 + 110 x 60 = **11 580** (cf. tableau général de calcul).

La moyenne recherchée, qu'on appelle **moyenne conditionnelle de y, pour la modalité 150 du caractère x**, est égale à :
**moyenne conditionnelle de y, pour la modalité 150 du caractère x = 11 580 / 126 = 91,90 €.**

C-à-d que les 126 lycéens dont le revenu mensuel moyen est de 150 €, dépensent mensuellement en moyenne 91,90 €.

On a :

$CV_{y1} = 0,43 \quad CV_{y2} = 0,29 ; \quad CV_{y3} = 0,21.$

Pour les lignes 2 et 3 du tableau, la valeur des coefficients de variation caractérise une dispersion faible des dépenses de consommation autour de la dépense mensuelle moyenne des lycéens dont les revenus sont respectivement de 100 € et 150 € par mois.

**E. 12°)** $$\bar{y} = \frac{1}{n..} \sum_{i=1}^{m} n_{i.}\ \bar{y}_i = \frac{1}{350} (55 \times 24{,}36 + 169 \times 56{,}21 + 126 \times 91{,}90) =$$ **64,06**

$$\sigma_y^2 = V(y) = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\ \bar{y}_i^2 - \bar{y}^2 + \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\ \sigma_{y_i}^2 = VE_y + VR_y =$$ **841,25**

$VE_y$

$$= \frac{1}{350} \left[ 55\ (24{,}36 - 64{,}06)^2 + 169\ (56{,}21 - 64{,}06)^2 + 126\ (91{,}90 - 64{,}06)^2 \right]$$ **VEy =**

$$= \frac{1}{350} (55 \times 24{,}36^2 + 169 \times 56.21^2 + 126 \times 91{,}90^2) - 64{,}06^2 =$$ **556,48**

$$\eta_{y/x}^2 = \frac{VE_y}{\sigma_y^2} = \frac{\frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\ \bar{y}_i^2 - \bar{y}^2}{\sigma_y^2} = \frac{556{,}48}{841{,}25} =$$ **$\eta^2$ y / x** **0,6615**

**Le rapport de corrélation de y en x est égal à 0,6615 , soit 66,2 % .**

**L'argent de poche (= le niveau de revenu) dont disposent mensuellement les lycéens de l'échantillon explique, pour 2/3 environ, la variation de leurs achats mensuels de consommation.**

$VR_y$

$$= \frac{55 \times 111{,}87 + 169 \times 264{,}95 + 126 \times 386{,}85}{350} =$$ **VRy =** **284,78**

$$0 \leq \eta^2 \leq +1$$

$$\eta^2_{y/x} \leq \eta^2_{x/y}$$

$$\rho^2 \leq \eta^2_{y/x} \leq \eta^2_{x/y} \text{ si relation linéaire entre x et y}$$

Les inégalités se transforment en égalités lorsqu'il y a relation fonctionnelle entre les variables.

**Les calculs suivants, réalisés à titre pédagogique, n'ont pas de signification socio-économique concrète.**

$$\bar{x} = \frac{1}{n..} \sum_{j=1}^{p} n_{.j}\ \bar{x}_j = \frac{1}{350}\,(60 \times 62{,}50 + 126 \times 99{,}21 + 104 \times 127{,}88 + 60 \times 150{,}00) = \mathbf{110{,}14}$$

$$\sigma_x^2 = V(x) = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\ \bar{x}_j^2 - \bar{x}^2 + \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\ \sigma_{x_j}^2 = VE_x + VR_x = \mathbf{1189{,}98}$$

$$VE_x = \frac{1}{350}\left[60\,(62{,}50 - 110{,}14)^2 + 126\,(99{,}21 - 110{,}14)^2 + 104\,(127{,}88 - 110{,}14)^2 + 60\,(150{,}00 - 110{,}14)^2\right]$$

$$= \frac{1}{350}\,(60 \times 62{,}50^2 + 126 \times 99.21^2 + 104 \times 127{,}88^2 + 60 \times 150{,}00^2) - 110{,}14^2 =$$

**VEx = 798,04**

$$\eta^2_{x/y} = \frac{VE_x}{\sigma_x^2} = \frac{\frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\ \bar{x}_j^2 - \bar{x}^2}{\sigma_x^2} = \frac{798{,}04}{1189{,}98} =$$

**η² x / y**

**0,6706**

$$VR_x = \frac{60 \times 635{,}42 + 126 \times 277{,}15 + 104 \times 616{,}68 + 60 \times 0}{350} =$$

**VRx =**

**391,94**

**Le rapport de corrélation de y en x est égal à 0,6706 , soit 67,1 % .**

**Ici, pas de lien de causalité dans ce sens !!**

S'il avait une signification concrète, ce rapport indiquerait que les achats mensuels de consommation des lycéens expliqueraient, pour 2/3 environ, la variation de leurs "revenus" mensuels.

12°)

$$VE_y = \frac{1}{350}\left[55\ (62{,}50 - 64{,}06)^2 + 169\ (99{,}21 - 64{,}06)^2 + 126\ (127{,}88 - 64{,}06)^2\right]$$
$$= \frac{1}{350}\ (55 \times 62{,}50^2 + 169 \times 99.21^2 + 126 \times 127{,}88^2) - 64{,}06^2 = 556{,}48$$

Le rapport de corrélation : $\eta^2_{y/x}$ = VEy / $\sigma^2_y$ = 556,48 / 841,25 = **0,66**

Ce résultat signifie que l'argent de poche dont disposent mensuellement les lycéens de l'échantillon explique, pour 2/3 environ, la variation de leurs achats mensuels de consommation.

A l'aide des variances conditionnelles de y, on peut calculer la variance intrapopulation (= VRy = variance résiduelle de y) et l'on obtient ici : VRy = 284,78.

On vérifie que l'on a bien : $\sigma^2_y$ = VEy + VRy = 556,48 + 284,78 = 841,25.

$$VE_x = \frac{1}{350}\left[60\ (62{,}50 - 110{,}14)^2 + 126\ (99{,}21 - 110{,}14)^2 + 104\ (127{,}88 - 110{,}14)^2 + 60\ (150{,}00 - 110{,}14)^2\right]$$
$$= \frac{1}{350}\ (60 \times 62{,}50^2 + 126 \times 99.21^2 + 104 \times 127{,}88^2 + 126 \times 150{,}00^2) - 110{,}14^2 = 798{,}04$$

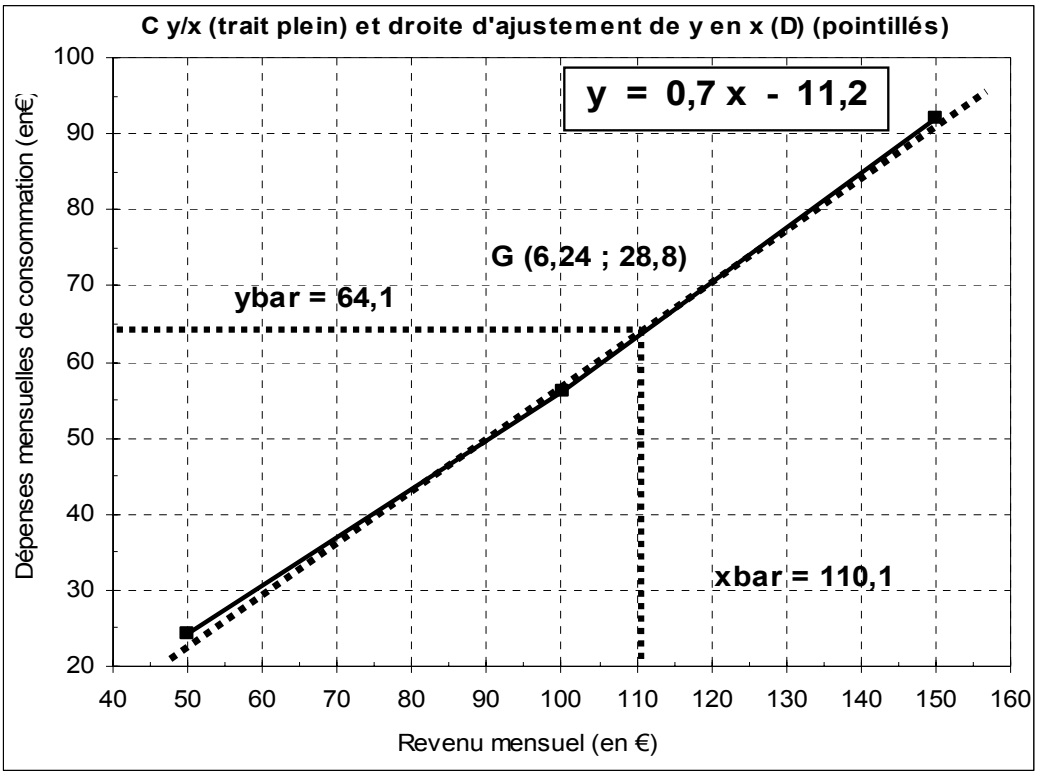
Le rapport de corrélation : $\eta^2_{x/y}$ = VEx / $\sigma^2_x$ = 798,04 / 1 189,98 = **0,67**, déterminé pour des raisons pédagogiques, n'a ici aucune signification concrète, car ce n'est pas la consommation qui influe sur le revenu (cf. même remarque que pour un ajustement linéaire).

A l'aide des variances conditionnelles de x, on peut calculer la variance intrapopulation (= VRx = variance résiduelle de x) et l'on obtient ici : VRx = 391,94.

On vérifie que l'on a bien : $\sigma^2_x$ = VEx + VRx = 798,04 + 391,94 = 1 189,98

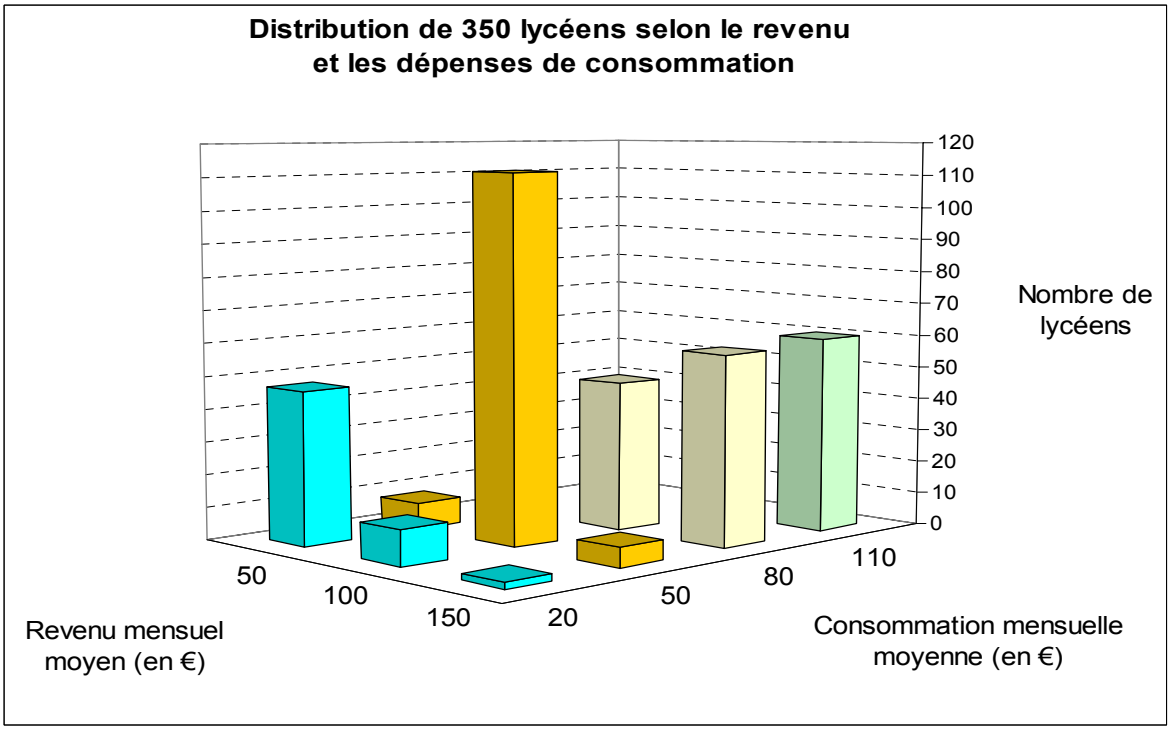
Remarques terminales :

Si l'on porte sur le même graphique la droite d'ajustement de y en x et la courbe de régression de y en x, on obtient les tracés suivants :




Ce graphique confirme bien que l'ajustement linéaire est pertinent, puisque la courbe de régression est presque confondue avec la droite d'ajustement linéaire.

## Stéréogramme de la série



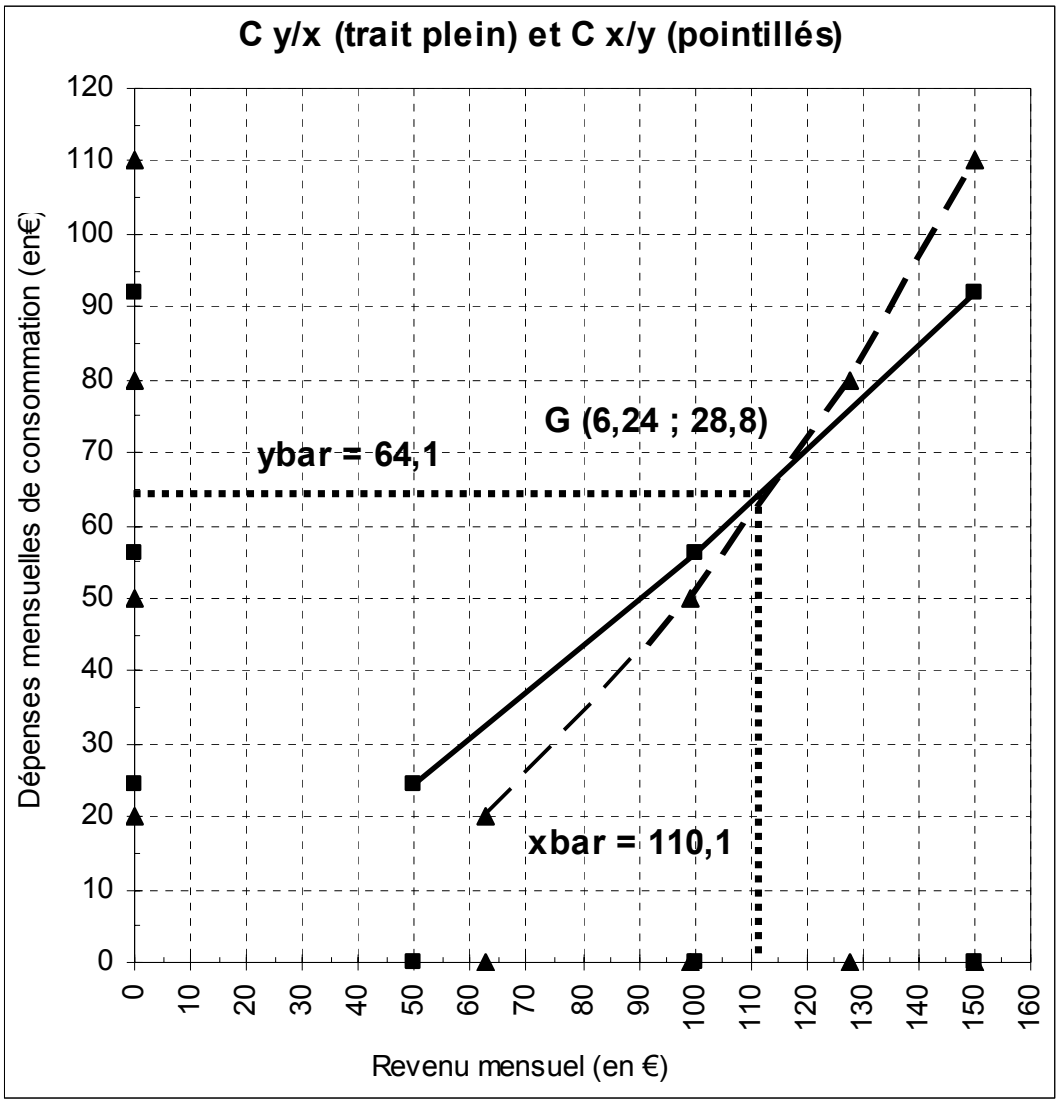
**Tracé des courbes de régression C $_{y/x}$ (de y en x) et C $_{x/y}$ (de x en y)**

| | ybar i |
|---|---|
| 0 | 24,36 |
| 0 | 56,21 |
| 0 | 91,90 |

| xi | ybar i | |
|---|---|---|
| 50 | 24,36 | 0 |
| 100 | 56,21 | 0 |
| 150 | 91,90 | 0 |

| xbar j | yj | |
|---|---|---|
| 62,50 | 20 | 0 |
| 99,21 | 50 | 0 |
| 127,88 | 80 | 0 |
| 150,00 | 110 | 0 |

| | yj |
|---|---|
| 0 | 20 |
| 0 | 50 |
| 0 | 80 |
| 0 | 110 |



**Exemple 2**

Une enquête, réalisée auprès de 100 personnes possédant une automobile, a permis de renseigner le tableau suivant, dans lequel :

- la **variable x** représente la **puissance fiscale des véhicules** (en CV) ;
- la **variable y** représente la **durée de vie moyenne des pneumatiques** (en milliers de kilomètres) :

| $x_i$ \ $y_j$ | **20** | **30** | **35** |
|---|---|---|---|
| **4** | 0 | 6 | 15 |
| **6** | 4 | 35 | 7 |
| **8** | 20 | 11 | 2 |

**A.**

**1°)** Quels sont **la population, les caractères et types** de la série statistique envisagée ?

**2°)** Combien de véhicules de 6 CV ont simultanément une durée de vie de leurs pneumatiques de 35 000 km ? Combien de véhicules ont une durée de vie de leurs pneumatiques de 20 000 km, pour une puissance fiscale de 4 CV ?

**3°)** A partir du tableau des effectifs conjoints ci-dessus, réaliser un **graphique** approprié pour représenter correctement cette série.

**B.**

**4°)** Établir le tableau des **fréquences conjointes théoriques d'indépendance**.

**5°)** A partir de la comparaison du tableau des **fréquences conjointes réelles** et de celui des **fréquences conjointes théoriques d'indépendance**, analyser, selon deux méthodes différentes, la dépendance ou l'indépendance des deux caractères x et y.

**C.**

**6°)** Calculer les **paramètres marginaux de chacune des variables x et y**. Commenter les résultats.

**7°)** Calculer la **covariance** entre les deux caractères x et y, de deux façons différentes. A partir de cette valeur, calculer le **coefficient de corrélation linéaire** entre x et y, ainsi que le **coefficient de détermination**. Interpréter les résultats.

**8°)** Calculer, le cas échéant, les **coefficients de la droite d'ajustement de y en x** et tracer cette droite sur le graphique précédent.

**9°)** Un ajustement linéaire de x en y a-t-il ici une signification ? Pourquoi ?

**D.**

**10°)** On s'intéresse aux véhicules selon la durée de vie de leurs pneumatiques. Calculer les **paramètres conditionnels de x**.

**11°)** On s'intéresse aux véhicules selon leur puissance fiscale. Calculer les **paramètres conditionnels de y**.

**A. 1°)**

**Population :** 100 automobilistes enquêtés.
**Caractères :** - puissance fiscale (en CV) des véhicules (xi) ;
- durée de vie moyenne (en milliers de km) des pneus (yj).
**Type :** quantitatif discret, pour chacun des deux caractères.

| xi \ yj | 20 | 30 | 35 | n i. | n i. xi | n i. xi ² | Σj nij yj | Σj nij xi yj |
|---|---|---|---|---|---|---|---|---|
| **4** | 0 | 6 | 15 | 21 | 84 | 336 | 705 | 2 820 |
| **6** | 4 | 35 | 7 | 46 | 276 | 1 656 | 1 375 | 8 250 |
| **8** | 20 | 11 | 2 | 33 | 264 | 2 112 | 800 | 6 400 |
| **n .j** | 24 | 52 | 24 | 100 | 624 | 4 104 | 2 880 | 17 470 |
| **n.j yj** | 480 | 1 560 | 840 | 2 880 | | | | |
| **n.j yj ²** | 9 600 | 46 800 | 29 400 | 85 800 | | | | |
| **Σi nij xi** | 184 | 322 | 118 | 624 | | | | |
| **Σi nij xi yj** | 3 680 | 9 660 | 4 130 | 17 470 | | | | |

**2°)**

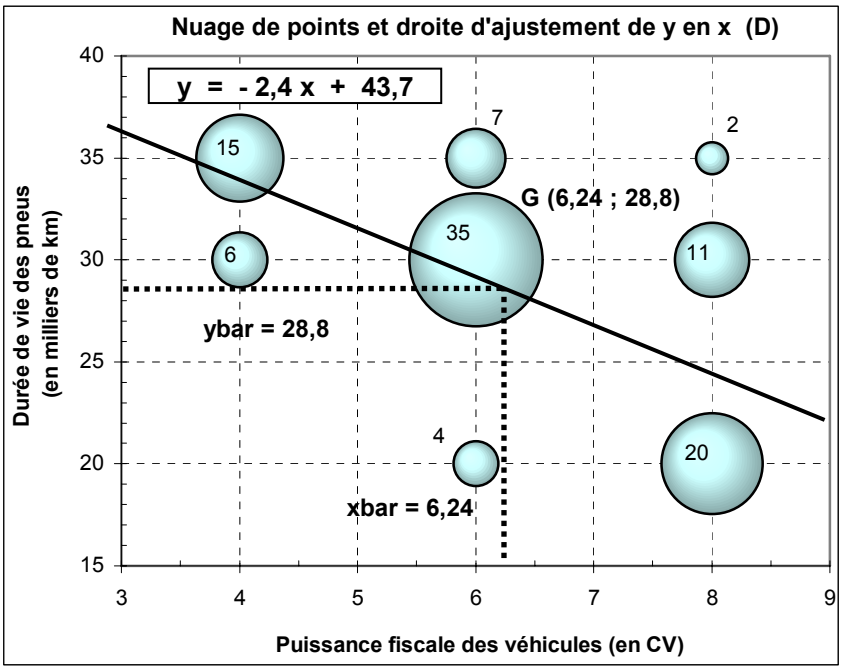
pf = 6 CV et dvp = 35 000 km = **n23 = 7** véhicules.
pf = 4 CV et dvp = 20 000 km = **n11 = 0** véhicule.

**3°)**

Une représentation graphique appropriée de cette série statistique à deux caractères quantitatifs consiste à tracer un **nuage dont les points ont une surface proportionnelle aux effectifs conjoints** correspondants.

| x | y | nij |
|---|---|---|
| 4 | 20 | 0 |
| 4 | 30 | 6 |
| 4 | 35 | 15 |
| 6 | 20 | 4 |
| 6 | 30 | 35 |
| 6 | 35 | 7 |
| 8 | 20 | 20 |
| 8 | 30 | 11 |
| 8 | 35 | 2 |

Remarque :
les cases grises servent uniquement à construire le graphique à l'aide d'un tableur.




Si R(35) = 1 cm par exemple :

alors R(20) = racine (20 / 35) x R(35) = 0,76 cm.
alors R(15) = racine (15 / 35) x R(35) = 0,65 cm.

**B. 4°)**

Il convient tout d'abord de construire un tableau des effectifs conjoints théoriques d'indépendance. Pour cela, on utilise la formulation suivante :

$$n_{ij} = (n_{i.} \times n_{.j}) / n_{..}$$

Comme il s'agit d'effectifs portant sur des nombres entiers, on arrondit les résultats obtenus, pour chaque case du tableau, à l'entier le plus proche.

**Effectifs conjoints théoriques d'indépendance** (valeurs arrondies)

| xi \ yj | 20 | 30 | 35 | n i. |
|---|---|---|---|---|
| **4** | 5 | 11 | 5 | 21 |
| **6** | 11 | 24 | 11 | 46 |
| **8** | 8 | 17 | 8 | 33 |
| **n .j** | 24 | 52 | 24 | 100 |

**Fréquences conjointes théoriques d'indépendance (en %)**

| xi \ yj | 20 | 30 | 35 | f i. |
|---|---|---|---|---|
| **4** | 5,0 | 10,9 | 5,0 | 21,0 |
| **6** | 11,0 | 23,9 | 11,0 | 46,0 |
| **8** | 7,9 | 17,2 | 7,9 | 33,0 |
| **f .j** | 24,0 | 52,0 | 24,0 | 100,0 |

Pour construire un tableau des fréquences conjointes théoriques d'indépendance (exprimées en pourcentage), on utilise la formulation suivante :

$$f_{ij} = (f_{i.} \times f_{.j}) / 100$$

**5°)**

**Fréquences conjointes réelles (en %)**

| xi \ yj | 20 | 30 | 35 | f i. |
|---|---|---|---|---|
| **4** | 0,0 | 6,0 | 15,0 | 21,0 |
| **6** | 4,0 | 35,0 | 7,0 | 46,0 |
| **8** | 20,0 | 11,0 | 2,0 | 33,0 |
| **f .j** | 24,0 | 52,0 | 24,0 | 100,0 |

Remarque : on a un échantillon de taille 100. La donnée de ce tableau est donc immédiate.

**Analyse de l'indépendance des deux caractères : méthode 1**

On peut comparer les 2 tableaux (effectifs conjoints réels et effectifs conjoints théoriques d'indépendance), ou ceux des fréquences conjointes réelles et des fréquences conjointes théoriques d'indépendance, ce que nous faisons ici.

**a)** Considérons les 2 tableaux **en lignes** : si l'usure des pneus (y) était totalement indépendante de la puissance des véhicules (x), on devrait retrouver, pour chaque modalité de x (la puissance fiscale), les mêmes pourcentages dans chaque ligne du tableau (cf. les f .j réelles : 24,0 %, 52,0 %, 24,0 %).

De fait, en comparant les deux tableaux de fréquences conjointes, on observe, pour les **véhicules dont la puissance fiscale est de 4 CV**, une nette surreprésentation des véhicules dont la durée de vie moyenne des pneus est de 35 000 km (15,0 % contre 5,0 % dans l'hypothèse d'indépendance). Pour les autres modalités de y (durées de vie des pneus), on constate une sous-représentation assez importante.

Pour les **véhicules de puissance fiscale 6 CV**, on observe une nette surreprésentation des véhicules à durée de vie moyenne des pneus est de 30 000 km (35,0% contre 23,9% dans l'hypothèse d'indépendance). Pour les autres modalités de y (durées de vie des pneus), on constate une sous-représentation plus ou moins prononcée.

Pour les **véhicules de puissance fiscale 8 CV**, on observe une forte surreprésentation des véhicules à durée de vie moyenne des pneus est de 20 000 km (20,0 % contre 7,9 % dans l'hypothèse d'indépendance). Pour les deux autres modalités de y (durées de vie des pneus), on constate une sous-représentation assez nette.

**b)** Considérons les 2 tableaux **en colonnes** : si la puissance fiscale (x) était totalement indépendante de la durée de vie des pneus (y), on devrait retrouver, pour chaque modalité de y (la durée de vie des pneus), les mêmes pourcentages dans chaque colonne du tableau (cf. les f i. réelles : 21,0 %, 46,0 % et 33,0 %).

**Attention ! Bien noter que le raisonnement en colonnes, effectivement possible ici d'un point de vue "mécanique", n'est pas valide si l'on se place sur le plan de la recherche d'une relation de cause à effet (relation de causalité) entre les deux caractères !**

**En aucun cas, la durée de vie des pneus n'induit (n'a d'influence sur) la puissance fiscale des véhicules !**

**Analyse de l'indépendance des deux caractères : méthode 2**

On peut comparer les deux tableaux ci-après : A (rapport des effectifs réels aux effectifs théoriques d'indépendance) et B (fréquences conjointes réelles).

**Pourcentages des effectifs conjoints réels aux effectifs conjoints théoriques (tableau A)**

| xi \ yj | 20 | 30 | 35 | ([n i./ n i.]-1) x 100 |
|---|---|---|---|---|
| 4 | -100,0 | -45,5 | **200,0** | 0,0 |
| 6 | -63,6 | **45,8** | -36,4 | 0,0 |
| 8 | **150,0** | **-35,3** | -75,0 | 0,0 |
| ([n .j/ n .j]-1) x 100 | 0,0 | 0,0 | 0,0 | 0,0 |

On a par exemple (1ère case du tableau A) : [(0 / 5) - 1] = - 1 soit - 100 %

**Fréquences conjointes réelles (en %) (tableau B)**

| xi \ yj | 20 | 30 | 35 | f i. |
|---|---|---|---|---|
| 4 | 0,0 | 6,0 | **15,0** | 21,0 |
| 6 | 4,0 | **35,0** | 7,0 | 46,0 |
| 8 | **20,0** | **11,0** | 2,0 | 33,0 |
| f .j | 24,0 | 52,0 | 24,0 | 100,0 |

Si l'on considère les seules fréquences conjointes réelles qui dépassent 10 % (ici, leur somme représente 81,0 % de l'effectif total) ou celles qui dépassent 15 % (ici, leur somme représente 70,0 % de l'effectif total), on constate que **les pourcentages correspondants du tableau A sont les plus élevés**.

NB : il faut faire abstraction de la case qui correspond à une fréquence réelle nulle dont l'interprétation n'est pas très pertinente.

**Conclusion générale** : au vu des valeurs obtenues, on peut dire que la durée de vie des pneus des véhicules est largement dépendante de la puissance fiscale des véhicules.

**C. 6°)**

| | | |
|---|---|---|
| moy mgle x = | 6,24 | CV |
| var mgle x = | 2,10 | CV ² |
| ectyp mgl x = | 1,45 | CV |
| CV x = | 0,23 | |

$$\bar{x} = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\, x_i$$

$$\sigma_x^2 = \frac{1}{n_{..}} \sum_{i=1}^{m} n_{i.}\, x_i^2 - \bar{x}^2$$

La dispersion sur les puissances fiscales dans l'échantillon est faible (CVx = 0,23).

| | | |
|---|---|---|
| moy mgle y = | 28,80 | 10^3 km |
| var mgle y = | 28,56 | 10^3 km ² |
| ectyp mgl y = | 5,34 | 10^3 km |
| CV y = | 0,19 | |

$$\bar{y} = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\, y_j$$

$$\sigma_y^2 = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j}\, y_j^2 - \bar{y}^2$$

La dispersion sur les durées de vie des pneus est également faible (CVy = 0,19)

**7°)**

La somme des triples produits nij xi yj , qui sert à calculer la covariance entre x et y, peut être obtenue de 2 façons différentes :
- soit en calculant les sommes intermédiaires ligne par ligne (dernière colonne du tableau),
- soit en calculant les sommes intermédiaires colonne par colonne (dernière ligne du tableau).

cov xy = -5,01

$$\sigma_{xy} = \frac{1}{n_{..}} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij}\, x_i\, y_j - \bar{x}\,\bar{y}$$

**rho xy = -0,65**

**rho ² xy = 0,42**

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n_{..}} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij}\, x_i\, y_j - \bar{x}\,\bar{y}}{\sigma_x \cdot \sigma_y}$$

La valeur du coefficient de détermination permet d'avancer que la régression linéaire de y en x (ou de ... x en y !) explique 42 % du phénomène observé, c-à-d. que la puissance fiscale des véhicules explique à hauteur de 42 % le niveau de la dispersion observée sur les durées de vie des pneus.

**8°)**

Ici, on a $n > 10$ et un coefficient de corrélation linéaire suffisant pour valider un ajustement linéaire.

**Ajustement linéaire de y en x (droite D) : y = - 2,4 x + 43,7**

$$a = \frac{\frac{1}{n_{..}} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij} (x_i - \bar{x}) (y_i - \bar{y})}{\frac{1}{n_{..}} \sum_{i=1}^{m} (x_i - \bar{x})^2} = \frac{\frac{1}{n_{..}} \sum_{i=1}^{m} \sum_{j=1}^{p} n_{ij} x_i y_j - \bar{x}\bar{y}}{\frac{1}{n_{..}} \sum_{i=1}^{m} (x_i - \bar{x})^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

$b = \bar{y} - a\bar{x}$ **a = -2,38** **b = 43,68**

Pour tracer la droite sur le graphique, on peut retenir les deux points :
x = 3 ; y = 36,5 et : x = 8 ; y = 24,5

**9°)**

Un ajustement linéaire de x en y, c-à-d lorsque y joue le rôle de variable explicative et x celui de variable expliquée (dépendante), n'a ici aucune signification :
**ce n'est pas la durée de vie des pneus**
**qui influence la puissance fiscale des véhicules !**

**D. 10°)**

Considérer les véhicules selon la durée de vie de leurs pneus revient à envisager des **sous-populations** particulières parmi les 100 véhicules de l'échantillon.

Tout se passe comme si l'on avait affaire à des séries à un seul caractère (x), dont les effectifs correspondaient sucessivement à chacune des modalités de y.
Pour calculer les **paramètres conditionnels du caractère x** (moyennes, variances, écarts-types et coefficients de variation), on procède classiquement comme dans le cas des séries à un caractère.

Selon chaque durée de vie des pneus, on a successivement :

| xi | n i1 | n i1 xi | n i1 xi ² |
|---|---|---|---|
| **4** | 0 | 0 | 0 |
| **6** | 4 | 24 | 144 |
| **8** | 20 | 160 | 1 280 |
| **n .1** | 24 | 184 | 1 424 |

Durée de vie de 20 000 km.

**xbar 1 =** **7,67** CV
**var =** **0,56** CV ²
**écart-type =** **0,75** CV
**CV =** **0,10**

Valeur qui caractérise une dispersion très faible autour de la valeur moyenne observée.

| xi | n i2 | n i2 xi | n i2 xi ² |
|---|---|---|---|
| **4** | 6 | 24 | 96 |
| **6** | 35 | 210 | 1 260 |
| **8** | 11 | 88 | 704 |
| **n .2** | 52 | 322 | 2 060 |

Durée de vie de 30 000 km.

**xbar 2 =** **6,19** CV
**var =** **1,27** CV ²
**écart-type =** **1,13** CV
**CV =** **0,18**

Valeur qui caractérise une dispersion faible autour de la valeur moyenne observée.

| xi | n i3 | n i3 xi | n i3 xi ² |
|---|---|---|---|
| **4** | 15 | 60 | 240 |
| **6** | 7 | 42 | 252 |
| **8** | 2 | 16 | 128 |
| **n .3** | 24 | 118 | 620 |

Durée de vie de 35 000 km.

**xbar 3 =** **4,92** CV
**var =** **1,66** CV ²
**écart-type =** **1,29** CV
**CV =** **0,26**

Valeur qui caractérise une dispersion faible autour de la valeur moyenne observée.

La première sous-population (durée de vie des pneus de 20 000 km) est ici la plus homogène par rapport à la puissance fiscale des véhicules (CV le plus faible = 0,10).

**11°)**

Considérer les véhicules selon leur puissance fiscale revient à envisager des **sous-populations** particulières parmi les 100 véhicules de l'échantillon.
Tout se passe comme si l'on avait affaire à des séries à un seul caractère (y), dont les effectifs correspondaient sucessivement à chacune des modalités de x.
Pour calculer les paramètres conditionnels du caractère y (moyennes, variances, écarts-types et coefficients de variation), on procède classiquement comme dans le cas de séries à un caractère.

Selon chaque puissance fiscale des véhicules, on a successivement :

Puissance fiscale de 4 CV.

| yj | n 1j | n 1j yj | n 1j yj ² |
|---|---|---|---|
| **20** | 0 | 0 | 0 |
| **30** | 6 | 180 | 5 400 |
| **35** | 15 | 525 | 18 375 |
| **n 1.** | 21 | 705 | 23 775 |

**ybar 1 =** **33,57** 10^3 km
**var =** **5,10** 10^3 km ²
**écart-type =** **2,26** 10^3 km
**CV =** **0,07**

Valeur qui caractérise une dispersion très faible autour de la valeur moyenne observée.

Puissance fiscale de 6 CV.

| yj | n 2j | n 2j yj | n 2j yj ² |
|---|---|---|---|
| **20** | 4 | 80 | 1 600 |
| **30** | 35 | 1 050 | 31 500 |
| **35** | 7 | 245 | 8 575 |
| **n 2.** | 46 | 1 375 | 41 675 |

**ybar 2 =** **29,89** 10^3 km
**var =** **12,49** 10^3 km ²
**écart-type =** **3,53** 10^3 km
**CV =** **0,12**

Valeur qui caractérise une dispersion faible autour de la valeur moyenne observée.

Puissance fiscale de 8 CV.

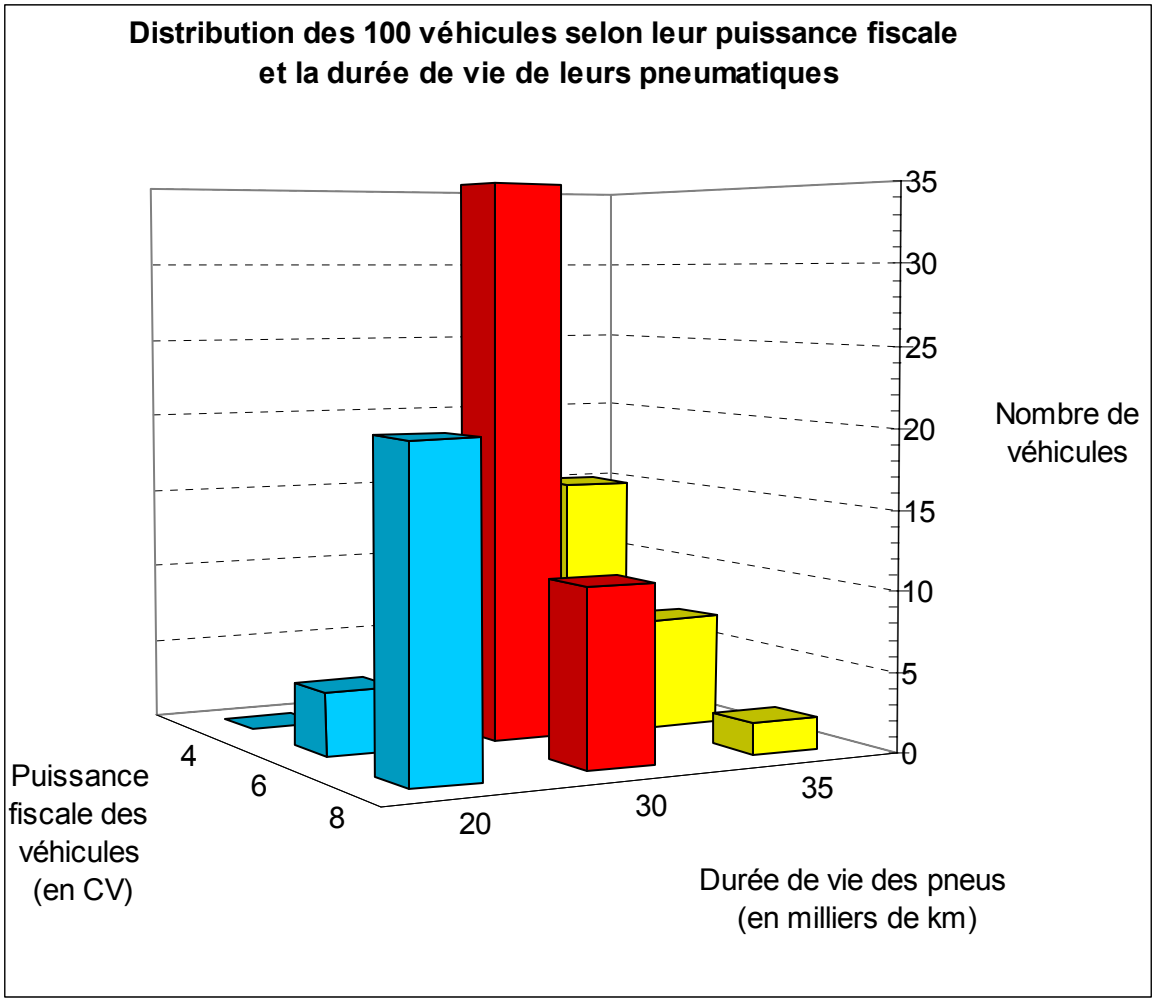
| yj | n 3j | n 3j yj | n 3j yj ² |
|---|---|---|---|
| **20** | 20 | 400 | 8 000 |
| **30** | 11 | 330 | 9 900 |
| **35** | 2 | 70 | 2 450 |
| **n 3.** | 33 | 800 | 20 350 |

**ybar 3 =** **24,24** 10^3 km
**var =** **28,97** 10^3 km ²
**écart-type =** **5,38** 10^3 km
**CV =** **0,22**

Valeur qui caractérise une dispersion faible autour de la valeur moyenne observée.

La première sous-population (puissance fiscale des véhicules = 4 CV) est ici la plus homogène par rapport à la durée de vie des pneumatiques (CV le plus faible = 0,07).

## Stéréogramme de la série



**a)** $n_{2.}$ = nombre total de véhicules de 6 CV = **46** (= effectif marginal correspondant à la 2ème modalité de x)
$n_{.3}$ = nombre total de véhicules dont les pneumatiques ont duré 35.000 km = **24** (= effectif marginal correspondant à la 3ème modalité de y)
$n_{..}$ = taille de l'échantillon = **100**

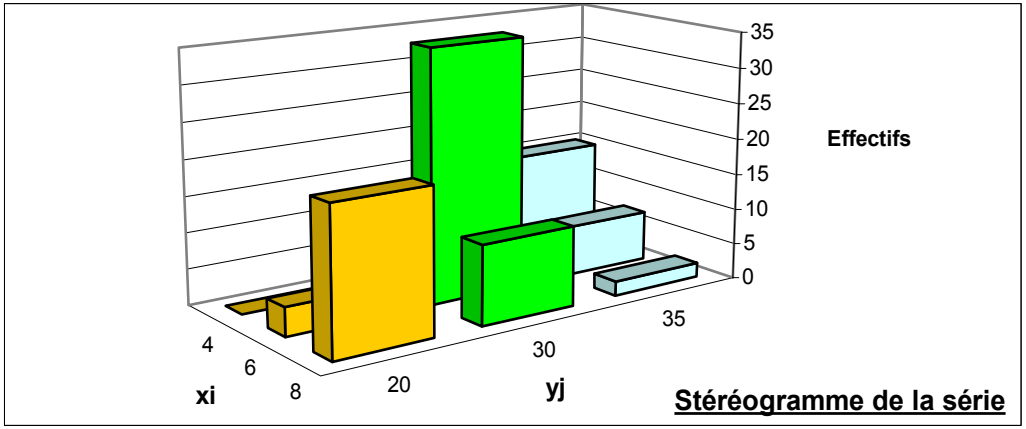
Etant donné que la taille de l'échantillon est égale à 100, les fréquences relatives (x 100) sont égales aux effectifs.

**b)** $f_{.2}$ = proportion des véhicules dont les pneus ont duré 30.000 km = **52%** (= fréquence marginale correspondant à la 2ème modalité de y)
$f_{23}$ = proportion des véhicules de 6 CV dont les pneumatiques ont duré 35.000 km = **7%** (= fréquence partielle relative à l'effectif total)
$f_{3/2}$ **à j fixé** = proportion des véhicules dont les pneumatiques ont duré 30.000 km et qui sont des véhicules de 8 CV = 11 / 52 = **21,15 %**
= fréquence conditionnelle de x selon y (ou à j fixé) : $f_{i/j} = n_{ij} / n_{.j}$

**Remarque :** si l'on avait eu $f_{3/2}$ à **i** fixé, il se serait agi de la proportion des véhicules de 6 CV dont les pneus ont duré 35.000 km = 7 / 46 = 15,22 %
= fréquence conditionnelle de y selon x (ou à i fixé) : $f_{j/i} = n_{ij} / n_{i.}$

**c)**

| xi \ yj | 20 | 30 | 35 | ni. | ni. xi | ni. xi ² | Sj nij yj | y bar i | Sj nij yj ² | var cond yi | Sj nij xi yj |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **4** | ***0*** | ***6*** | ***15*** | 21 | 84 | 336 | 705 | 33,57 | 23 775 | 5,10 | 2 820 |
| **6** | ***4*** | ***35*** | ***7*** | 46 | 276 | 1 656 | 1 375 | 29,89 | 41 675 | 12,49 | 8 250 |
| **8** | ***20*** | ***11*** | ***2*** | 33 | 264 | 2 112 | 800 | 24,24 | 20 350 | 28,97 | 6 400 |
| **n.j** | 24 | 52 | 24 | 100 | 624 | 4 104 | | | | | 17 470 |
| **n.j yj** | 480 | 1 560 | 840 | 2 880 | | | | | | | |
| **n.j yj ²** | 9 600 | 46 800 | 29 400 | 85 800 | | | | | | | |
| **Si nij xi** | 184 | 322 | 118 | | | | | | | | |
| **x bar j** | 7,67 | 6,19 | 4,92 | | | | | | | | |
| **Si nij xi ²** | 1 424 | 2 060 | 620 | | | | | | | | |
| **var cond xj** | 0,56 | 1,27 | 1,66 | | | | | | | | |
| **Si nij xi yj** | 3 680 | 9 660 | 4 130 | 17 470 | | | | | | | |



**Stéréogramme de la série**

**d)**

**xbar 2** = moyenne conditionnelle de x selon y (ou à j fixé) = **7,67 CV**
= valeur moyenne des puissances fiscales des véhicules dont les pneus ont duré 30.000 km

**ybar 1** = moyenne conditionnelle de y selon x (ou à i fixé) = **33 570 km** = valeur moyenne des durées des pneumatiques concernant les véhicules de 4 CV

| | |
|---|---|
| Moyenne marginale de x | 6,24 |
| Variance marginale de x | 2,10 |
| Ecart-type marginal de x | 1,45 |

| | |
|---|---|
| Moyenne marginale de y | 28,80 |
| Variance marginale de y | 28,56 |
| Ecart-type marginal de y | 5,34 |

**f)**

| | |
|---|---|
| COV xy | -5,01 |
| **ρ xy** | **-0,647** |
| **ρ ² xy** | **0,418** |

La valeur du coefficient de corrélation linéaire est ici trop faible pour que l'on puisse admettre un ajustement linéaire valide entre les variables x et y. Par ailleurs, le nombre de couples de points est insuffisant, pour réaliser un ajustement linéaire pertinent.

La valeur du coefficient de détermination permet d'avancer que la régression linéaire de y en x (ou de ... x en y !) n'explique qu'environ 42 % du phénomène observé.

**g)**

| | |
|---|---|
| Variance expliquée de x | 0,91 |
| Variance résiduelle de x | 1,19 |
| Variance totale de x | 2,10 |

| | |
|---|---|
| Variance expliquée de y | 12,18 |
| Variance résiduelle de y | 16,38 |
| Variance totale de y | 28,56 |

Le rapport de corrélation eta² y/x = variance expliquée de y / variance marginale (ou totale) de y est un nombre sans dimension destiné à nous renseigner sur l'intensité de la liaison qui peut exister entre y (fonction) et x (variable).

Le rapport de corrélation eta² x/y = variance expliquée de y / variance marginale (ou totale) de y est un nombre sans dimension destiné à nous renseigner sur l'intensité de la liaison qui peut exister entre x (fonction) et y (variable) .

| | |
|---|---|
| **η² x/y =** | **0,433** |

| | |
|---|---|
| **η² y/x =** | **0,427** |

On a : variance marginale = variance expliquée + variance résiduelle = variance des moyennes conditionnelles + moyenne des variances conditionnelles
(variance totale) (par la régression) (non expliquée par la régression)

Quel que soit le rapport considéré, la variance expliquée ne représente qu'environ 43 % de la variance totale, ce qui est plutôt faible. Il en résulte que les régressions de y en x ou de x en y n'expliquent que 43 % de la variance marginale : les courbes de régression correspondantes résument donc mal une hypothétique liaison entre les variables x et y .

**ATTENTION !**

**Il doit être clair que la recherche d'un ajustement (linéaire ou non) de x en y n'est pas pertinente ici !**
**Seule la relation de y en x a un sens dans cet exemple !**

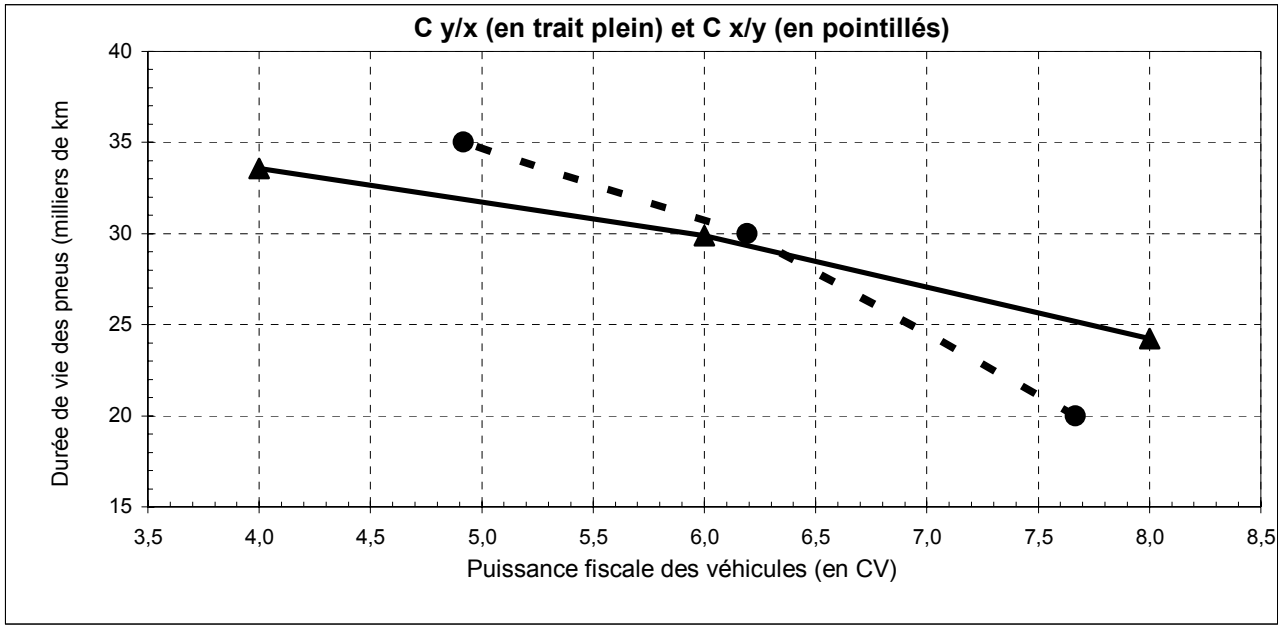
**De manière générale, il convient donc expressément, avant de se lancer dans les calculs (toujours possibles), de vérifier le bien-fondé de ces derniers, en termes de causalité, sous peine d'interprétations farfelues !**

**C'est seulement pour des raisons pédagogiques que les deux courbes de régression ont été tracées, mais seule la courbe rouge (en trait plein) a un sens ici.**

**e)**

| xi | ybar i |
|---|---|
| 4 | 33,57 |
| 6 | 29,89 |
| 8 | 24,24 |

| xbar j | yj |
|---|---|
| 7,67 | 20 |
| 6,19 | 30 |
| 4,92 | 35 |



# PARTIE 3 : LES INDICES STATISTIQUES

Dans les parties précédentes, l'objectif était de caractériser (résumer) des séries à un ou deux caractères au moyen d'un certain nombre de valeurs numériques (tendance centrale, dispersion, asymétrie, concentration).

Dans le cas des séries à deux caractères, on a de plus cherché à résumer l'intensité de la liaison existant entre les deux caractères (rapports de corrélation, coefficient de détermination, courbes de régression, droites d'ajustement linéaire des moindres carrés).

Les indices statistiques cherchent à résumer, par des **nombres sans dimension**, les évolutions d'une grandeur économique dans le temps (par ex. évolution du prix d'un bien sur plusieurs mois ou années) ou dans l'espace (par ex. évolution du prix d'un même dans des régions ou pays différents).

L'utilisation des indices facilite les comparaisons entre variables hétérogènes (unités et / ou valeurs absolues très différentes).

Exemple : comparaison de l'évolution de la production de deux types de bien.

| | **t = 0** | **t = 1** | |
|---|---|---|---|
| **Bien X** | 375.000 unités | 469.625 unités | + 25 % |
| **Bien Y** | 244 unités | 305 unités | + 25 % |

Au vu des valeurs absolues, on a du mal à discerner rapidement que les taux d'accroissement sont les mêmes pour les deux productions entre les deux périodes. Alors qu'en indice base 100, pour chaque bien en t = 0, on aura un indice 125 en t = 1.

Nous allons voir en quoi le rôle des (nombres) indices facilite une lecture directe (donc rapide) des taux d'évolution d'une grandeur (économique) dans le temps ou dans l'espace.

**De manière générale, un indice est formé par le rapport positif ou nul de deux valeurs**.

L'indice permet de lire directement le taux d'évolution d'une grandeur dans le temps ou dans l'espace.

La difficulté réside dans l'établissement du rapport en question, selon qu'il met en jeu des éléments homogènes ou hétérogènes.

Lorsque les éléments sont homogènes (mêmes unités notamment), on élabore des **indices élémentaires**.

Lorsqu'ils sont hétérogènes, on élabore des **indices synthétiques**.

Ces indices sont plus délicats à conceptualiser (notamment en vue de disposer de propriétés intéressantes, comme l'agrégation des données). Les résultats qu'ils fournissent sont parfois plus difficiles à interpréter car, comme pour toute caractéristique statistique qui tend à résumer une distribution, leur est associée une perte d'information par rapport aux données brutes.

## CHAPITRE 1 : LES INDICES ÉLÉMENTAIRES

### 1. Définition

Une grandeur G est **simple** si elle prend une valeur et une seule, pour une situation donnée.

Par ex., à la date 0, la valeur de la grandeur G sera $g_0$ ; à la date 1, la valeur de la grandeur G sera $g_1$ ; etc.

On appelle **indice élémentaire** d'une grandeur simple G, à la date t par rapport à la date 0, le rapport : $I_{t/0}(G) = \frac{g_t}{g_0}$

où : 0 représente la date de référence, ou date de base, et t représente la date courante.

cf. coefficient multiplicateur (1 + t) dans le calcul des taux de croissance.

Remarque : la plupart du temps, les indices sont exprimés en pourcentage, base 100 à la date : t = 0. On écrit alors : $I_{t/0}(G) = \frac{g_t}{g_0} \times 100$ .

**Exemple d'indice temporel** : évolution de la population française sur les trois derniers recensements.

Pop. 1982 : 54,3 M d'habitants

Pop. 1990 : 58,1 M d'habitants

Pop. 1999 : 60,2 M d'habitants

Si l'on retient l'année 1982 comme année de base, on a :

$I_{82/82}(P) = \frac{54,3}{54,3} \times 100 = 100$ soit 1 en base 1 (= coefficient multiplicateur).

$I_{90/82}(P) = \frac{58,1}{54,3} \times 100 = 107,0$ soit 1,07 en base 1.

$I_{99/82}(P) = \frac{60,2}{54,3} \times 100 = 110,9$ soit 1,109 en base 1.

Ainsi, de manière rapide, on peut lire le taux d'évolution de la population :

+ 7,0 % entre 1982 et 1990 ; + 10,9 % entre 1982 et 1999.

**Exemple d'indice spatial** : comparaison d'un salaire horaire selon la zone géographique.

Pour une CSP donnée, on a par ex. une rémunération horaire de :

28,7 € en région parisienne (P),
25,3 € en zone urbaine de province (ZU),
22,5 € en zone rurale (ZR).

Si l'on retient comme salaire de base 100 celui de la région parisienne, on peut construire un indice relatif à ces salaires horaires de la façon suivante :

$$I_{P/P} = \frac{28,7}{28,7} \times 100 = 100$$

$$I_{ZU/P} = \frac{25,3}{28,7} \times 100 = 88,2 \qquad I_{ZR/P} = \frac{22,5}{28,7} \times 100 = 78,4$$

Cela signifie que :

- en zone urbaine de province, le salaire horaire est 100 - 88,2 = 11,8 % moins élevé qu'à Paris,
- celui des zones rurales est 100 - 78,4 = 21,6 % moins élevé qu'à Paris.

## 2. Propriétés des indices élémentaires

### 21. La circularité (ou transitivité, ou transférabilité)

On a : $100 \times I_{t/0}(G) = I_{t/t'}(G) \times I_{t'/0}(G)$ avec : $0 < t' < t$ (1)

Remarque : on remplace l'expression de chaque indice par le rapport des valeurs correspondantes de la grandeur G aux dates considérées : $100 \left(\frac{g_t}{g_0} \times 100\right) = \left(\frac{g_t}{g_{t'}} \times 100\right) \times \left(\frac{g_{t'}}{g_0} \times 100\right)$

Soit aussi : $I_{t/t'}(G) = \frac{I_{t/0}(G)}{I_{t'/0}(G)} \times 100$ (2)

Si l'on connaît les indices aux temps t et t ', par rapport à l'année de base 0, la 2ème égalité permet d'obtenir directement l'indice de t par rapport à t ' .

Si l'on reprend l'exemple de la population française vu plus haut, on obtient :

$$I^{(P)}_{99/90} = \frac{I^{(P)}_{99/82}}{I^{(P)}_{90/82}} \times 100 = \frac{110,9}{107,0} \times 100 = 103,6 = \frac{60,2}{58,1} \times 100 .$$

soit un accroissement de la population de 3,6 % entre 1990 et 1999.

Remarque importante : il ne faut surtout pas ajouter les indices $I_{90/82}$ et $I_{99/90}$ en vue d'obtenir l'indice $I_{99/82}$ . cf. logique des taux de croissance (moyenne géométrique).

Le principe de circularité des indices élémentaires joue un rôle important lors des changements de base (= changement de période (année) de référence).

Si l'on effectue, sur une série, des changements de base successifs, en prenant pour base la situation de la période (mois, année, …) précédente, on aboutit à un **indice-chaîne**, de la forme :

Soit t l'année courante.

Pour comparer la situation de l'année t par rapport à l'année de base t = 0, on peut effectuer le produit des indices (annuels, mensuels, ...) successifs :

$$\frac{I_{t/0}}{100} = \frac{I_{1/0}}{100} \times \frac{I_{2/1}}{100} \times \frac{I_{3/2}}{100} \times \ldots \times \frac{I_{t/t-1}}{100}$$

En remplaçant les indices par leurs expressions en fonction des valeurs absolues des grandeurs correspondantes, on a : $\frac{I_{t/0}}{100} = \frac{\frac{g_1}{g_0} \times 100}{100} \times \frac{\frac{g_2}{g_1} \times 100}{100} \times \frac{\frac{g_3}{g_2} \times 100}{100} \times \ldots \times \frac{\frac{g_t}{g_{t-1}} \times 100}{100}$ .

Après simplifications, il reste : $\frac{I_{t/0}}{100} = \frac{g_t}{g_0}$ .

Remarque : bien que cette propriété ne s'applique pas pour les indices synthétiques, l'Insee est contraint, faute de mieux, de l'utiliser de manière non rigoureuse, pour le calcul de certains indices de prix, lorsque des changements d'année de base sont opérés. Ces derniers sont nécessaires car, sur les indices de prix composites, la composition de la grandeur G envisagée se modifie à chaque période (année).

Exemple de calcul d'indice-chaîne

Une grandeur G évolue dans le temps de la façon suivante :

| t | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Pourcentage d'évolution d'une période sur l'autre | Valeur absolue quelconque $G_0$ Base 100 | +10 % | +20 % | -15 % | -5 % | +10 % |

On a :

$$\frac{I_{5/0}}{100} = \frac{I_{1/0}}{100} \times \frac{I_{2/1}}{100} \times \frac{I_{3/2}}{100} \times \frac{I_{4/3}}{100} \times \frac{I_{5/4}}{100}$$

$$\frac{I_{5/0}}{100} = \frac{110}{100} \times \frac{120}{100} \times \frac{85}{100} \times \frac{95}{100} \times \frac{110}{100} = 1{,}173$$

$$\text{Donc : } I_{5/0} = 117{,}3$$

soit un accroissement de 17,3 % de la grandeur G sur l'ensemble de la période.

Remarque : supposons par exemple qu'on ait eu les données brutes suivantes :

| t | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Valeurs de G (en M€) | 250 | 275 | 330 | 280,5 | 266,48 | 293,12 |

Noter que les valeurs successives de G évoluent selon les pourcentages du tableau précédent.

Il vient : 293,12 / 250 = 1,173 , soit un accroissement de 17,3 % de la grandeur G sur l'ensemble de la période !

### 22. La réversibilité

En base 100, on a : $I_{0/t}(G) = \dfrac{10^4}{I_{t/0}(G)}$ Propriété très utile dans le cas des indices spatiaux.

Remarques :

1) en base 1, on a : $I_{0/t}(G) = \dfrac{1}{I_{t/0}(G)}$

2) on a en effet : $\dfrac{g_0}{g_t} \times 100 = \dfrac{100 \times 100}{\dfrac{g_t}{g_0} \times 100} = \dfrac{g_0}{g_t} = \dfrac{1}{\dfrac{g_t}{g_0}}$

Exemple : si l'on compare 2 régions ou 2 pays, il est évidemment intéressant de pouvoir exprimer un indice par rapport à chacun des deux pays. Supposons que l'indice du prix d'un bien, base 100 en France, soit de 125 en Italie, on a alors :

$$I_{I/F} = 125 \quad \text{et simultanément :} \quad I_{F/I} = \frac{10^4}{125} = 80.$$

Du point de vue français, le bien en question est 25 % plus cher en Italie et du point de vue italien, ce bien est 20 % moins cher en France.

**Proportionnalité des indices élémentaires**

Hypothèse : soit une grandeur G qui prend la valeur $g_0$ en t = 0. En t = t (date courante), on a :
$g_t = a \times g_0$ , expressions dans laquelle a est une constante.

Sous forme d'indice, on peut écrire : $I_{t/0}(G) = \frac{g_t}{g_0} \times 100 = \frac{a.g_0}{g_0} \times 100 = a \times 100$

Conclusion : si une grandeur G est multipliée par une constante a à l'année t, alors son indice $I_{t/0}(G)$ est aussi multiplié par cette constante.

Exemple : en t = 0, on a $g_0 = 350$ et en t = 1, on a $g_1 = 1\,400$, soit a = 4 : $g_1 = 4 . g_0$ .

En termes d'indice, on a : $I_{t/0}(G) = 4 \times 100 = 400$

En effet : $I_{1/0}(G) = \frac{g_1}{g_0} \times 100 = \frac{a.g_0}{g_0} \times 100 = \frac{4.350}{350} \times 100 = 400$

**Homogénéité des indices élémentaires**

Hypothèse : dans une unité donnée (par exemple la tonne), on a : $I_{t/0}(G) = \frac{g_t}{g_0} \times 100$

Supposons que l'unité devienne le kilogramme. La valeur $g_0$ (exprimée en tonne) devient : $1.000 \times g_0$ . De même, la valeur $g_1$ devient : $1.000 \times g_1$ .

Est donc on a, en termes d'indice : $I'_{t/0}(G) = \frac{1\,000 \times g_t}{1\,000 \times g_0} \times 100 = \frac{g_t}{g_0} \times 100 = I_{t/0}(G)$

Conclusion : si l'on change l'unité dans laquelle est exprimée une grandeur G, l'indice ne voit pas sa valeur affectée par ce changement d'unité.

Il s'agit-là d'un avantage à disposer d'un nombre sans dimension.

### 23. La multiplication des indices élémentaires

L'indice élémentaire d'un produit est égal au produit des indices élémentaires correspondants.

Par exemple, si l'on considère trois grandeurs A, B, G, telles que : G = A x B, alors :

$$100 \text{ x } I_{t/0}(G) = I_{t/0}(A) \text{ x } I_{t/0}(B)$$

En effet : $\dfrac{a_t \,.\, b_t}{a_0 \,.\, b_0} \text{ x } 100 = \dfrac{a_t}{a_0} \text{ x } 100 \text{ x } \dfrac{b_t}{b_0} \text{ x } 100$

Exemple

Considérons la recette totale procurée à une entreprise parlant d'un bien : RT = p . q

On a :

| t | 0 | 1 |
|---|---|---|
| p | 20 € | 22 € |
| q | 4 000 unités | 3 500 unités |

Solution 1

$$I_{1/0}(p) = \frac{22}{20} \text{ x } 100 = 110 \quad (+10\,\%)$$

$$I_{1/0}(q) = \frac{3\,500}{4\,000} \text{ x } 100 = 87{,}5 \quad (-12{,}5\,\%)$$

Finalement :

$$100 \text{ x } I_{1/0}(RT) = I_{1/0}(p) \text{ x } I_{1/0}(q) = 110 \text{ x } 87{,}5 = 9625$$

$$\text{D'où : } I_{1/0}(RT) = 96{,}25$$

soit une diminution de 3,75 % de la recette totale entre les deux périodes.

Solution 2

À partir du tableau, on peut aussi calculer $p_0q_0$ et $p_1q_1$, puis réalisé le calcul de l'indice correspondant :

| t | 0 | 1 |
|---|---|---|
| p | 20 € | 22 € |
| q | 4 000 unités | 3 500 unités |
| **RT = p . q** | 80 000 | 77 000 |

Il vient : $I_{1/0}(RT) = \dfrac{77\,000}{80\,000} \text{ x } 100 = 96{,}25$

## CHAPITRE 2 : LES INDICES SYNTHÉTIQUES

Les indices élémentaires correspondent à des grandeurs simples exprimées sous forme numérique.

Les indices synthétiques renvoient à des grandeurs complexes (c'est-à-dire que chaque grandeur complexe est elle-même exprimée par plusieurs nombres). Les grandeurs complexes sont le fruit composite de plusieurs grandeurs simples (cf. l'indice des prix à la consommation de l'INSEE).

En général, l'analyse des phénomènes économiques prend en compte l'évolution simultanée de plusieurs grandeurs simples, souvent hétérogènes entre elles (par ex. un indice général des prix à la consommation). Pour cette raison, on ne peut pas agréger directement les différents éléments étudiés. Il est alors nécessaire d'élaborer des grandeurs complexes, dont l'évolution à travers le temps (ou l'espace) est traduite par des **indices synthétiques**. Ces derniers résument, sous la forme d'un unique nombre sans dimension, l'information apportée par un ensemble d'indices élémentaires.

En Économie, on utilise usuellement des indices de prix (prix de gros, de détail, indices de salaires, évolution de chiffres d'affaires, …), des indices de quantité (ou de volume ; par ex. production industrielle ou agricole, volume du commerce extérieur, …) et des indices de valeur (ou de dépense globale, mettant en jeu simultanément prix et quantités). **En général, ces prix et ces quantités renvoient à des grandeurs économiques complexes, ce qui pose la question de savoir comment construire de façon appropriée les indices synthétiques correspondants**.

### Élaboration d'un indice synthétique

Soit G une grandeur complexe, constituée d'un ensemble de **k** grandeurs simples $g^1, g^2, \ldots, g^k$ .

On note alors : $G = \{ g^1, g^2, \ldots, g^i, \ldots, g^k \}$ .

Chaque grandeur simple peut voir son évolution retracée par un indice élémentaire. Pour la grandeur simple $g^i$ , par exemple : $I_{t/0}(g^i) = \frac{g^i_t}{g^i_0} \times 100$ .

Pour rendre compte de l'évolution de la grandeur complexe G, entre 0 et t, l'objectif est alors de résumer, sous la forme d'un nombre unique, donc par un indice synthétique $I_{t/0}(G)$ , la série G ci-dessus, formée par les k grandeurs $g^i$ (ou bien les indices élémentaires associés à ces grandeurs).

**L'indice synthétique recherché va correspondre à une valeur centrale de la série des k grandeurs ou des k indices élémentaires.**

Historiquement, on a envisagé des indices synthétiques confectionnés à partir de moyennes calculées sur les valeurs des différentes grandeurs simples, constitutives de la grandeur complexe envisagée, sous la forme suivante :

$$I_{t/0}(G) = \frac{\frac{1}{k}\left(g^1_t + g^2_t + \ldots + g^i_t + \ldots + g^k_t\right)}{\frac{1}{k}\left(g^1_0 + g^2_0 + \ldots + g^i_0 + \ldots + g^k_0\right)} \times 100 = \frac{\frac{1}{k}\sum_{i=1}^{k} g^i_t}{\frac{1}{k}\sum_{i=1}^{k} g^i_0} \times 100 = \frac{\bar{g}_t}{\bar{g}_0} \times 100$$

L'indice synthétique précédent a notamment le défaut de varier selon les unités de mesure retenues pour exprimer chaque grandeur simple, ce qui le rend peu opérationnel.

La méthode actuelle d'élaboration des indices synthétiques consiste à calculer des moyennes sur les différents indices élémentaires qui sont associés aux grandeurs simples constitutives de la grandeur complexe étudiée. Ainsi, on a :

$$I_{t/0}(G) = \frac{g_t}{g_0} \times 100 = \frac{1}{k}\left( \frac{g_t^1}{g_0^1} + \frac{g_t^2}{g_0^2} + ... + \frac{g_t^i}{g_0^i} + ... + \frac{g_t^k}{g_0^k} \right) \times 100 = \frac{1}{k} \sum_{i=1}^{k} \frac{g_t^i}{g_0^i} \times 100$$

Exemple : soit trois biens A, B et C, dont on a suivi l'évolution des prix aux dates 0 et 1 :

| | **A** | **B** | **C** |
|---|---|---|---|
| **t = 0** | 50 | 72 | 100 |
| **t = 1** | 50 | 85 | 110 |

On a ici : $I_{1/0}(G) = \frac{1}{3}\left( \frac{50}{50} + \frac{85}{72} + \frac{110}{100} \right) \times 100 = 109{,}35$, soit une hausse de + 9,35 % pour l'ensemble de ces trois biens, entre les dates 0 et 1.

Remarque : la formulation ci-avant ne fait pas apparaître de pondérations. On considère ici que les trois biens ont la même importance, donc le même poids, dans le calcul de l'indice de la grandeur composite G.

Dans les points suivants, nous allons au contraire attribuer des coefficients de pondération différents selon les biens envisagés, au prorata de la dépense relative affectée à l'achat de chacun de ces biens.

**Indices synthétiques usuels en Économie**

L'économiste est souvent amené à raisonner simultanément sur des prix et des quantités. En effet, le producteur s'intéresse à sa recette totale, produit du prix de vente unitaire d'un bien par la quantité produite et vendue. Le consommateur s'intéresse à sa dépense totale, produit du prix d'achat unitaire d'un bien par la quantité achetée et consommée. Dans cette optique, on définit des indices de la valeur globale (= de la dépense totale). Il s'agit d'indices de valeur.

Cependant, lorsque le produit prix x quantité évolue à la hausse ou à la baisse dans le temps, à l'aide du seul indice de la valeur globale, il n'est pas possible de connaître l'influence des seuls prix ou des seules quantités dans la variation du produit global.

Pour cette raison, afin d'isoler l'effet prix ou l'effet quantité, on rend constante l'une des deux variables, afin de repérer l'effet de la seule autre variable (concrètement, cela revient à supprimer un effet de structure !).

Nous allons donc distinguer dans la suite :

- des indices de prix, dont la formulation fait apparaître l'utilisation de quantités constantes, entre deux dates d'observation.

- des indices de quantité (ou de volume), dont la formulation fait apparaître l'utilisation de prix constants, entre deux dates d'observation.

Enfin, selon la date d'observation (o ou t) que l'on retient pour rendre constants les éléments de l'une ou l'autre variable, on construit des indices de Laspeyres (date de base 0) ou des indices de Paasche (date courante t).

## 1. Les indices de Laspeyres (1844)

Il s'agit d'indices de prix ou de quantités (volumes), portant sur des grandeurs complexes, dont les coefficients de pondération sont proportionnels à la dépense relative affectée à chacune des catégories de biens faisant partie de la grandeur complexe considérée.

**C'est le type de pondération envisagé qui distingue les indices de Laspeyres et les indices de Paasche (voir point 2). Pour les premiers, les coefficients de pondération retenus sont associés à la période de base (ou de référence). Pour les seconds, ces coefficients sont associés à la période courante.**

### 11. L'indice des prix de Laspeyres

Formule de définition de l'indice des prix de Laspeyres :

$$L^{p}_{t/0} = \frac{\sum_{i=1}^{n} p^{i}_{t} \cdot q^{i}_{0}}{\sum_{i=1}^{n} p^{i}_{0} \cdot q^{i}_{0}} \times 100$$

où : n = nombre de grandeurs simples (= nombre de biens) intégrées à la grandeur synthétique considérée ;
$p^{i}_{0}$ = prix de la grandeur simple i (en général, un bien économique) à la date 0 (date de base ou date de référence) ;
$q^{i}_{0}$ = quantité de la grandeur simple i (en général, quantité achetée ou consommée d'un bien économique) à la date 0 ;
$p^{i}_{t}$ = prix de la grandeur simple i à la date t (date courante) ;
$L^{p}_{t/0}$ = indice des prix de Laspeyres, entre les dates d'observation 0 et t.

**L'indice de Laspeyres permet de voir comment évoluent les prix, à quantités données de la période de base.**

On démontre qu'on peut aussi écrire :

$$L^{p}_{t/0} = \sum_{i=1}^{n} \frac{p^{i}_{0} \cdot q^{i}_{0}}{\sum_{i=1}^{n} p^{i}_{0} \cdot q^{i}_{0}} \times \frac{p^{i}_{t}}{p^{i}_{0}} \times 100$$

$\frac{p^{i}_{t}}{p^{i}_{0}} = I^{i}_{t/0}(p)$ c-à-d l'indice élémentaire du prix de la grandeur simple i, entre les dates 0 et t,

$\frac{p^{i}_{0} \cdot q^{i}_{0}}{\sum_{i=1}^{n} p^{i}_{0} \cdot q^{i}_{0}}$ = coefficient de pondération (ou coefficient budgétaire) de la grandeur simple i, à la date de base 0.

Sous cette forme, on peut définir l'indice de la façon suivante :

**l'indice des prix de Laspeyres est une moyenne arithmétique des indices élémentaires des prix (des grandeurs simples associées à l'indice synthétique), pondérés par les coefficients de pondération de la date de base.**

**Exemple 1** : un marchand de chaussures possède deux magasins A et B. Dans le tableau suivant, il a consigné, pour les années 1995, 1996 et 1997, le prix de vente moyen p (en €) des paires de chaussures et le nombre total de paires vendues q (en milliers) :

| | Magasin A | | Magasin B | |
|---|---|---|---|---|
| | **p** | **q** | **p** | **q** |
| **1995** | 20 | 10 | 25,5 | 7 |
| **1996** | 21 | 12 | 26,5 | 9,5 |
| **1997** | 21,6 | 11,5 | 30,5 | 14 |

En utilisant la 1ère formule, on a :

$$L^{p}_{96/95} = 104,5 \qquad L^{p}_{97/95} = 113,5 \qquad L^{p}_{97/96} = 109$$

Ainsi, entre 1996 et 1995, on a :

$$L^{p}_{96/95} = \frac{\sum_{i=1}^{n} p^{i}_{96}\ q^{i}_{95}}{\sum_{i=1}^{n} p^{i}_{95}\ q^{i}_{95}} \times 100 = \frac{21 \times 10 + 26,5 \times 7}{20 \times 10 + 25,5 \times 7} \times 100 = 104,49 = 104,5$$

**On constate un accroissement des prix de 4,5 % entre 1995 et 1996, de 13,5 % entre 1997 et 1995 et de 9 % entre 1997 et 1996.**

**Remarque importante : les indices synthétiques de Laspeyres et de Paasche ne vérifient pas les propriétés de circularité et de réversibilité des indices élémentaires.**

Dans l'exemple précédent, il n'y a **pas circularité** et, généralement, la différence observée entre les deux calculs n'est pas toujours aussi faible qu'ici :

$$L^{p}_{97/95} \neq \frac{L^{p}_{97/96} \times L^{p}_{96/95}}{100} \Leftrightarrow 113,5 \neq \frac{109 \times 104,5}{100} \Leftrightarrow 113,5 \neq 113,9$$

Ce résultat est fâcheux dans la mesure où l'INSEE utilise ces indices pour suivre, par exemple, l'évolution des prix à la consommation des ménages. On peut alors se demander pourquoi ils sont tout de même utilisés, moyennant une approximation plus ou moins importante selon les cas. Simplement parce que ces indices permettent d'agréger les données, en calculant des indices globaux qui sont de simples moyennes arithmétiques d'indices pondérés (contrairement à d'autres indices synthétiques qui, en contrepartie sont circulaires et réversibles, comme les indices de Fisher).

De même, il n'y a **pas réversibilité** :

$$L^{p}_{95/97} \neq \frac{10^4}{L^{p}_{97/95}} \Leftrightarrow 86,9 \neq \frac{10^4}{113,5} \Leftrightarrow 86,9 \neq 88,1$$

Cette non réversibilité est très gênante lors des comparaisons spatiales (pays ou régions), car elle nécessite un nouveau calcul systématique lorsqu'on change de zone de référence.

La valeur trouvée **88,1** n'est pas quelconque. On verra plus loin qu'elle correspond à l'indice des prix de Paasche de 1995 par rapport à 1997 !

**Exemple 2** : afin de mettre en jeu les coefficients budgétaires, considérons trois biens A, B et C, dont on connaît les quantités consommées (en unités physiques) en 1995 et 2005, ainsi que les prix unitaires associés (en €) :

| | Quantités consommées | | | Prix unitaires (en €) | | | Coefficients budgétaires | | |
|---|---|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **A** | **B** | **C** | **A** | **B** | **C** |
| **1995** | 40 | 30 | 20 | 20 | 10 | 10 | 8/13 | 3/13 | 2/13 |
| **2005** | 40 | 40 | 10 | 22 | 20 | 12 | 22/45 | 20/45 | 3/45 |

Remarque :

les coefficients budgétaires de 1995 sont calculés à partir du rapport $\frac{p^{i}_{0} \cdot q^{i}_{0}}{\sum_{i=1}^{n} p^{i}_{0} \cdot q^{i}_{0}}$, à la date de base 0. Ceux de 2005 serviront plus loin pour le calcul des indices de Paasche.

En utilisant la 2ème formule, on a : $\boxed{L^{p}_{05/95} = 132{,}3}$, **soit une hausse des prix de 32,3 % entre 1995 et 2005, pour l'ensemble des trois biens A, B et C.**

$$L^{p}_{05/95} = \sum_{i=1}^{n} \frac{p^{i}_{95} \cdot q^{i}_{95}}{\sum_{i=1}^{n} p^{i}_{95} \cdot q^{i}_{95}} \times \frac{p^{i}_{05}}{p^{i}_{95}} \times 100$$

$$= \left(\frac{8}{13} \times \frac{22}{20} + \frac{3}{13} \times \frac{20}{10} + \frac{2}{13} \times \frac{12}{10}\right) \times 100 = 132{,}3$$

On peut ici encore vérifier la non réversibilité de l'indice de Laspeyres :

$$L^{p}_{95/05} \neq \frac{10^4}{L^{p}_{05/95}} \Leftrightarrow 72{,}2 \neq \frac{10^4}{132{,}3} \Leftrightarrow 72{,}2 \neq 75{,}6$$

De manière similaire à l'exemple précédent, la valeur trouvée **75,6** n'est pas quelconque. Elle correspond à l'indice des prix de Paasche de 1995 par rapport à 2005.

### 12. L'indice des quantités de Laspeyres (ou indice de volume de Laspeyres)

Formule de définition de l'indice des quantités de Laspeyres :

$$L^{q}_{t/0} = \frac{\sum_{i=1}^{n} p^{i}_{o} \cdot q^{i}_{t}}{\sum_{i=1}^{n} p^{i}_{0} \cdot q^{i}_{0}} \times 100$$

où : n = nombre de grandeurs simples intégrées à la grandeur synthétique considérée ;
$p^{i}_{0}$ = prix de la grandeur simple i (en général, un bien économique) à la date 0 (date de base ou date de référence) ;
$q^{i}_{0}$ = quantité de la grandeur simple i (en général, quantité achetée ou consommée d'un bien économique) à la date 0 ; $q^{i}_{t}$ = quantité de la grandeur simple i à la date t (date courante) ;
$L^{q}_{t/0}$ = indice de Laspeyres des quantités, entre les dates d'observation 0 et t.

On démontre qu'on peut aussi écrire : $$L^{q}_{t/0} = \sum_{i=1}^{n} \frac{p^{i}_{0} \cdot q^{i}_{0}}{\sum_{i=1}^{n} p^{i}_{0} \cdot q^{i}_{0}} \times \frac{q^{i}_{t}}{q^{i}_{0}} \times 100$$

$\frac{q^{i}_{t}}{q^{i}_{0}} = I^{i}_{t/0}(q)$ c-à-d l'indice élémentaire de la quantité de la grandeur simple i, entre les dates 0 et t,

$\frac{p^{i}_{0} \cdot q^{i}_{0}}{\sum_{i=1}^{n} p^{i}_{0} \cdot q^{i}_{0}}$ = coefficient de pondération (ou coefficient budgétaire) de la grandeur simple i, à la date de base 0.

Sous cette forme, on peut définir l'indice de la façon suivante :

**l'indice des quantités de Laspeyres est une moyenne arithmétique des indices élémentaires des quantités (des grandeurs simples associées à l'indice synthétique), pondérés par les coefficients de pondération de la date de base.**

**Exemple 1** : un marchand de chaussures possède deux magasins A et B. Dans le tableau suivant, il a consigné, pour les années 1995, 1996 et 1997, le prix de vente moyen p (en €) des paires de chaussures et le nombre total de paires vendues q (en milliers) :

| | Magasin A | | Magasin B | |
|---|---|---|---|---|
| | **p** | **q** | **p** | **q** |
| **1995** | 20 | 10 | 25,5 | 7 |
| **1996** | 21 | 12 | 26,5 | 9,5 |
| **1997** | 21,6 | 11,5 | 30,5 | 14 |

En utilisant la 1ère formule, on a :

$$L^{q}_{96/95} = 127,4 \quad L^{q}_{97/95} = 155,1 \quad L^{q}_{97/96} = 121,6$$

Ainsi, entre 1996 et 1995, on a :

$$L^{p}_{96/95} = \frac{\sum_{i=1}^{n} p^{i}_{95}\ q^{i}_{96}}{\sum_{i=1}^{n} p^{i}_{95}\ q^{i}_{95}} \times 100 = \frac{20 \times 12 + 25,5 \times 9,5}{20 \times 10 + 25,5 \times 7} \times 100 = 127,4$$

**On constate un accroissement des quantités vendues de 27,4 % entre 1995 et 1996, de 55,1 % entre 1997 et 1995 et de 21,6 % entre 1997 et 1996.**

Ici encore, on vérifie qu'il n'y a pas circularité :

$$L^{q}_{97/95} \neq \frac{L^{q}_{97/96} \times L^{q}_{96/95}}{100} \Leftrightarrow 155,1 \neq \frac{121,6 \times 127,4}{100} \Leftrightarrow 155,1 \neq 154,9$$

De même, il n'y a pas réversibilité :

$$L^{q}_{95/97} \neq \frac{10^4}{L^{q}_{97/95}} \Leftrightarrow 63{,}6 \neq \frac{10^4}{155{,}1} \Leftrightarrow 63{,}6 \neq 64{,}5$$

**Exemple 2** : afin de mettre en jeu les coefficients budgétaires, considérons trois biens A, B et C, dont on connaît les quantités consommées (en unités physiques) en 1985 et 1995, ainsi que les prix unitaires associés (en €) :

| | Quantités consommées | | | Prix unitaires (en €) | | | Coefficients budgétaires | | |
|---|---|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **A** | **B** | **C** | **A** | **B** | **C** |
| **1995** | 40 | 30 | 20 | 20 | 10 | 10 | 8/13 | 3/13 | 2/13 |
| **2005** | 40 | 40 | 10 | 22 | 20 | 12 | 22/45 | 20/45 | 3/45 |

Remarque : les coefficients budgétaires de 1995 sont calculés à partir du rapport $\frac{p^{i}_{0} \cdot q^{i}_{0}}{\sum_{i=1}^{n} p^{i}_{0} \cdot q^{i}_{0}}$,

.à la date de base 0. Ceux de 2005 serviront plus loin pour les indices de Paasche.

En utilisant la 2ème formule, on a : $\boxed{L^{q}_{05/95} = 100}$, **soit une quantité vendue inchangée entre 1995 et 2005**.

$$L^{q}_{05/95} = \sum_{i=1}^{n} \frac{p^{i}_{95} \cdot q^{i}_{95}}{\sum_{i=1}^{n} p^{i}_{95} \cdot q^{i}_{95}} \times \frac{q^{i}_{05}}{q^{i}_{95}} \times 100$$

$$= \left(\frac{8}{13} \times \frac{40}{40} + \frac{3}{13} \times \frac{40}{30} + \frac{2}{13} \times \frac{10}{20}\right) \times 100 = 100$$

On peut ici encore vérifier la non réversibilité de l'indice de Laspeyres :

$$L^{q}_{95/05} \neq \frac{10^4}{L^{q}_{05/95}} \Leftrightarrow 95{,}6 \neq \frac{10^4}{100} \Leftrightarrow 95{,}6 \neq 100$$

## 2. Les indices de Paasche (1874)

### 21. L'indice des prix de Paasche

Formule de définition de l'indice des prix de Paasche :

$$P^{p}_{t/0} = \frac{\sum_{i=1}^{n} p^{i}_{t} . q^{i}_{t}}{\sum_{i=1}^{n} p^{i}_{0} . q^{i}_{t}} \times 100$$

où : n = nombre de grandeurs simples intégrées à la grandeur synthétique considérée ;
$p^{i}_{t}$ = prix de la grandeur simple i à la date t (date courante) ;
$q^{i}_{t}$ = quantité de la grandeur simple i (en général, quantité achetée ou consommée d'un bien économique) à la date t ;
$p^{i}_{0}$ = prix de la grandeur simple i (en général, un bien économique) à la date 0 (date de base ou date de référence) ;
$P^{p}_{t/0}$ = indice de Paasche des prix, entre les dates d'observation 0 et t.

On démontre qu'on peut aussi écrire :

$$P^{p}_{t/0} = \frac{1}{\sum_{i=1}^{n} \frac{p^{i}_{t} . q^{i}_{t}}{\sum_{i=1}^{n} p^{i}_{t} . q^{i}_{t}} \times \frac{p^{i}_{0}}{p^{i}_{t}}} \times 100$$

$\frac{p^{i}_{t}}{p^{i}_{0}} = I^{i}_{t/0}(p)$ c-à-d l'indice élémentaire du prix de la grandeur simple i, entre les dates 0 et t,

$\frac{p^{i}_{t} . q^{i}_{t}}{\sum_{i=1}^{n} p^{i}_{t} . q^{i}_{t}}$ = coefficient de pondération (ou coefficient budgétaire) de la grandeur simple i, à la date courante t.

Sous cette forme, on peut définir l'indice de la façon suivante :

**l'indice des prix de Paasche est une moyenne harmonique des indices élémentaires des prix (des grandeurs simples associées à l'indice synthétique), pondérés par les coefficients de pondération de la date courante.**

Remarque : la formule ci-dessus n'apparaît pas sous la forme d'une moyenne harmonique classique. Au dénominateur, par commodité d'écriture, on multiplie les coefficients budgétaires par l'inverse des indices élémentaires de prix.

**Exemple 1** : un marchand de chaussures possède deux magasins A et B. Dans le tableau suivant, il a consigné, pour les années 1995, 1996 et 1997, le prix de vente moyen p (en €) des paires de chaussures et le nombre total de paires vendues q (en milliers) :

| | Magasin A | | Magasin B | |
|---|---|---|---|---|
| | **p** | **q** | **p** | **q** |
| **1995** | 20 | 10 | 25,5 | 7 |
| **1996** | 21 | 12 | 26,5 | 9,5 |
| **1997** | 21,6 | 11,5 | 30,5 | 14 |

En utilisant la 1ère formule, on a :

$$P^{p}_{96/95} = 104,5 \quad P^{p}_{97/95} = 115,1 \quad P^{p}_{97/96} = 110,3$$

Ainsi, entre 1996 et 1995, on a :

$$P^{p}_{96/95} = \frac{\sum_{i=1}^{n} p^{i}_{96} \cdot q^{i}_{96}}{\sum_{i=1}^{n} p^{i}_{95} \cdot q^{i}_{96}} \times 100 = \frac{21 \times 12 + 26,5 \times 9,5}{20 \times 12 + 25,5 \times 9,5} \times 100 = 104,46 = 104,5$$

**On constate un accroissement des prix de 4,5 % entre 1995 et 1996, de 15,1 % entre 1997 et 1995 et de 8,5 % entre 1997 et 1996.**

**<u>Remarque</u> : les indices synthétiques de Laspeyres et de Paasche ne vérifient pas les propriétés de circularité et de réversibilité des indices élémentaires.**

Dans l'exemple précédent, il n'y a <u>pas circularité</u> et, généralement, la différence observée entre les deux calculs n'est pas toujours aussi faible qu'ici :

$$P^{p}_{97/95} \neq \frac{P^{p}_{97/96} \times P^{p}_{96/95}}{100} \Leftrightarrow 115,1 \neq \frac{110,3 \times 104,5}{100} \Leftrightarrow 115,1 \neq 115,3$$

De même, il n'y a <u>pas réversibilité</u> : $P^{p}_{95/97} \neq \frac{10^4}{P^{p}_{97/95}} \Leftrightarrow 88,1 \neq \frac{10^4}{115,1} \Leftrightarrow 88,1 \neq 86,9$

**<u>Deux remarques importantes</u>** :

a) dans le cas usuel où les quantités consommées diminuent lorsque les prix augmentent, le plus souvent, on peut vérifier l'inégalité $L^{p} > P^{p}$ (moyenne arithmétique > moyenne harmonique). Cela dépend toutefois de la part relative de chacun des biens qui entrent dans la composition de l'indice, ainsi que des pourcentages d'accroissement des prix et de diminution des quantités consommées correspondantes. On ne vérifie pas ce résultat ici, car les quantités consommées augmentent en même temps que les prix.

b) Il existe une <u>relation de réversibilité particulière entre les indices des prix de Laspeyres et de Paasche, relativement aux dates 0 et t</u>.

On a : $$L^{p}_{0/t} = \frac{10^4}{P^{p}_{t/0}} \quad \text{et :} \quad P^{p}_{0/t} = \frac{10^4}{L^{p}_{t/0}}$$

Dans l'exemple :

$$L^{p}_{95/97} = \frac{10^4}{P^{p}_{97/95}} = \frac{10^4}{115,1} = 86,9 \quad \text{et :} \quad P^{p}_{95/97} = \frac{10^4}{L^{p}_{97/95}} = \frac{10^4}{113,5} = 88,1$$

**Exemple 2** : afin de mettre en jeu les coefficients budgétaires, considérons trois biens A, B et C, dont on connaît les quantités consommées (en unités physiques) en 1995 et 2005, ainsi que les prix unitaires associés (en €) :

| | Quantités consommées | | | Prix unitaires (en €) | | | Coefficients budgétaires | | |
|---|---|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **A** | **B** | **C** | **A** | **B** | **C** |
| **1995** | 40 | 30 | 20 | 20 | 10 | 10 | 8/13 | 3/13 | 2/13 |
| **2005** | 40 | 40 | 10 | 22 | 20 | 12 | 22/45 | 20/45 | 3/45 |

Remarque : les coefficients budgétaires de 2005 sont calculés à partir du rapport $\frac{p_t^i . q_t^i}{\sum_{i=1}^{n} p_t^i . q_t^i}$, à la date courante t.

En utilisant la 2ème formule, on a : $\boxed{P^{p}_{05/95} = 138,5}$, **soit une hausse des prix de 38,5 % entre 1995 et 2005.**

$$P^{p}_{05/95} = \frac{1}{\sum_{i=1}^{n} \frac{p^i_{05}}{\sum_{i=1}^{n} p^i_{05}.q^i_{05}}} x100 = \frac{1}{\sum_{i=1}^{n} \frac{p^i_{05}.q^i_{05}}{\sum_{i=1}^{n} p^i_{05}.q^i_{05}} x \frac{p^i_{95}}{p^i_{05}}} x100 = \left( \frac{1}{\frac{22}{45}x\frac{20}{22} + \frac{20}{45}x\frac{10}{20} + \frac{3}{45}x\frac{10}{12}} \right) x100 = 138,5$$

Ici, on a :

$$P^{p}_{95/05} = \frac{10^4}{L^{p}_{05/95}} = \frac{10^4}{132,3} = 75,6 \quad \text{et :} \quad L^{p}_{95/05} = \frac{10^4}{P^{p}_{05/95}} = \frac{10^4}{138,5} = 72,2$$

### 22. L'indice des quantités de Paasche (ou indice de volume de Paasche)

Formule de définition de l'indice des quantités de Paasche : $\boxed{P^{q}_{t/0} = \frac{\sum_{i=1}^{n} p_t^i . q_t^i}{\sum_{i=1}^{n} p_t^i . q_0^i} \times 100}$

où : n = nombre de grandeurs simples intégrées à la grandeur synthétique considérée ;
$p_t^i$ = prix de la grandeur simple i à la date t (date courante) ;
$q_t^i$ = quantité de la grandeur simple i (en général, quantité achetée ou consommée d'un bien économique) à la date t ;
$q_0^i$ = quantité de la grandeur simple i à la date 0 (date de base ou date de référence) ;
$P^{q}_{t/0}$ = indice de Paasche des quantités, entre les dates d'observation 0 et t.

On démontre qu'on peut aussi écrire : $$P^{q}_{t/0} = \frac{1}{\sum_{i=1}^{n} \frac{p^{i}_{t} . q^{i}_{t}}{\sum_{i=1}^{n} p^{i}_{t} . q^{i}_{t}} \times \frac{q^{i}_{0}}{q^{i}_{t}}} \times 100$$

$\frac{q^{i}_{t}}{q^{i}_{0}} = I^{i}_{t/0}(q)$ c-à-d l'indice élémentaire de la quantité de la grandeur simple i, entre les dates 0 et t,

$\frac{p^{i}_{t} . q^{i}_{t}}{\sum_{i=1}^{n} p^{i}_{t} . q^{i}_{t}}$ = coefficient de pondération (ou coefficient budgétaire) de la grandeur simple i, à la date courante t.

Sous cette forme, on peut définir l'indice de la façon suivante :

**l'indice des quantités de Paasche est une moyenne harmonique des indices élémentaires des quantités (des grandeurs simples associées à l'indice synthétique), pondérés par les coefficients de pondération de la date courante.**

**Exemple 1** : un marchand de chaussures possède deux magasins A et B. Dans le tableau suivant, il a consigné, pour les années 1995, 1996 et 1997, le prix de vente moyen p (en €) des aires de chaussures et le nombre total de paires vendues q (en milliers) :

| | **Magasin A** | | **Magasin B** | |
|---|---|---|---|---|
| | **p** | **q** | **p** | **q** |
| **1995** | 20 | 10 | 25,5 | 7 |
| **1996** | 21 | 12 | 26,5 | 9,5 |
| **1997** | 21,6 | 11,5 | 30,5 | 14 |

En utilisant la 1ère formule, on a :

$$P^{q}_{96/95} = 127,4 \quad P^{q}_{97/95} = 157,3 \quad P^{q}_{97/96} = 123,0$$

Ainsi, entre 1996 et 1995, on a :

$$P^{q}_{96/95} = \frac{\sum_{i=1}^{n} p^{i}_{96} . q^{i}_{96}}{\sum_{i=1}^{n} p^{i}_{96} . q^{i}_{95}} \times 100 = \frac{21 \times 12 + 26,5 \times 9,5}{21 \times 10 + 26,5 \times 7} \times 100 = 127,4$$

**On constate un accroissement des quantités vendues de 27,4 % entre 1995 et 1996, de 57,3 % entre 1997 et 1995 et de 23 % entre 1997 et 1996**.

Ici encore, on vérifie qu'il n'y a pas circularité :

$$P^{q}_{97/95} \neq \frac{P^{q}_{97/96} \times P^{q}_{96/95}}{100} \Leftrightarrow 157,3 \neq \frac{123 \times 127,4}{100} \Leftrightarrow 157,3 \neq 156,7$$

De même, il n'y a pas réversibilité :

$$P^{q}_{95/97} \neq \frac{10^4}{P^{q}_{97/95}} \Leftrightarrow 64,5 \neq \frac{10^4}{157,3} \Leftrightarrow 88,1 \neq 63,6$$

**Deux remarques importantes** :

a) dans le cas usuel où les quantités consommées diminuent lorsque les prix augmentent, le plus souvent, on peut vérifier l'inégalité $L^{q} > P^{q}$. Cela dépend toutefois de la part relative de chacun des biens qui entrent dans la composition de l'indice, ainsi que des pourcentages d'accroissement des prix et de diminution des quantités consommées correspondantes. On ne vérifie pas ce résultat ici, car les quantités consommées augmentent en même que les prix..

b) Il existe une relation de réversibilité particulière entre les indices des quantités de Laspeyres et de Paasche, relativement aux dates 0 et t. On a :

$$L^{q}_{0/t} = \frac{10^4}{P^{q}_{t/0}} \quad \text{et :} \quad P^{q}_{0/t} = \frac{10^4}{L^{q}_{t/0}}$$

Dans l'exemple :

$$L^{q}_{95/97} = \frac{10^4}{P^{q}_{97/95}} = \frac{10^4}{157,3} = 63,6 \quad \text{et :} \quad P^{q}_{95/97} = \frac{10^4}{L^{q}_{97/95}} = \frac{10^4}{155,1} = 64,5$$

**Exemple 2** : afin de mettre en jeu les coefficients budgétaires, considérons trois biens A, B et C, dont on connaît les quantités consommées (en unités physiques) en 1995 et 2005, ainsi que les prix unitaires associés (en €) :

| | Quantités consommées | | | Prix unitaires (en €) | | | Coefficients budgétaires | | |
|---|---|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **A** | **B** | **C** | **A** | **B** | **C** |
| **1995** | 40 | 30 | 20 | 20 | 10 | 10 | 8/13 | 3/13 | 2/13 |
| **2005** | 40 | 40 | 10 | 22 | 20 | 12 | 22/45 | 20/45 | 3/45 |

Remarque : les coefficients budgétaires de 2005, utilisés ici, sont calculés à partir du rapport $\frac{p^{i}_{t} \cdot q^{i}_{t}}{\sum_{i=1}^{n} p^{i}_{t} \cdot q^{i}_{t}}$, à la date courante t.

En utilisant la 2ème formule, on a : $\boxed{P^{q}_{05/95} = 104,7}$, **soit une hausse des quantités vendues de 4,7 % entre 1985 et 1995.**

$$P^{q}_{05/95} = \frac{1}{\dfrac{\dfrac{\sum_{i=1}^{n} p^{i}_{05}.q^{i}_{05}}{\sum_{i=1}^{n} p^{i}_{05}.q^{i}_{05}}}{\dfrac{q^{i}_{05}}{q^{i}_{95}}}} x100 = \frac{1}{\sum_{i=1}^{n} \dfrac{p^{i}_{05}.q^{i}_{05}}{\sum_{i=1}^{n} p^{i}_{05}.q^{i}_{05}} x \dfrac{q^{i}_{95}}{q^{i}_{05}}} x100 = \left( \frac{1}{\frac{22}{45}x\frac{40}{40} + \frac{20}{45}x\frac{30}{40} + \frac{3}{45}x\frac{20}{10}} \right) x100 = 104,7$$

Ici, on a :

$$P^{q}_{95/05} = \frac{10^4}{L^{q}_{05/95}} = \frac{10^4}{100} = 100 \quad \text{et :} \quad L^{q}_{95/05} = \frac{10^4}{P^{q}_{05/95}} = \frac{10^4}{104,7} = 95,6$$

**Les indices de Fisher (1922)**

En combinant les indices de Laspeyres et de Paasche, l'Américain Irving Fisher a proposé les indices suivants en 1922 :

**Indice des prix de Fisher**

L'indice des prix de Fisher est égal à la moyenne géométrique des indices des prix de Laspeyres et des prix de Paasche : $\boxed{F^{p}_{t/0} = \sqrt{L^{p}_{t/0} \times P^{p}_{t/0}} \times 100}$

**Indice des quantités de Fisher**

L'indice des quantités de Fisher est égal à la moyenne géométrique des indices des quantités de Laspeyres et des quantités de Paasche : $\boxed{F^{q}_{t/0} = \sqrt{L^{q}_{t/0} \times P^{q}_{t/0}} \times 100}$

Intérêt des indices de Fisher

Ces deux indices sont **réversibles**, ce qui est particulièrement intéressant dans le cas des indices spatiaux. Comme pour les indices élémentaires, on vérifie donc :

$$\boxed{F^{p}_{0/t} = \frac{10^4}{F^{p}_{t/0}} \quad \text{et :} \quad F^{q}_{0/t} = \frac{10^4}{F^{q}_{t/0}}}$$

Malheureusement, les indices de Fisher ne sont pas circulaires.

Ils présentent par ailleurs d'autres inconvénients :

1) l'interprétation d'une moyenne géométrique n'est pas immédiate (cf. conditions de Yule), alors que les indices de Laspeyres et de Paasche peuvent être écrits sous la forme de moyennes arithmétiques.

2) dans le cas de la moyenne géométrique, la propriété d'agrégation n'est plus vérifiée. Au contraire, les indices de Laspeyres et de Paasche étant utilisables sous forme de moyennes arithmétiques, il est possible de réaliser des calculs d'indices sur diverses sous populations.

Par exemple, si l'on s'intéresse à l'évolution des dépenses des ménages par grandes rubriques (alimentation, logement, produits manufacturés, services, ...), on peut d'abord calculer des indices de Laspeyres et de Paasche pour chacune des grandes rubriques (on commence même par les sous-rubriques). Ensuite, on utilise la propriété de la moyenne de la population totale (par rapport aux sous-populations), tel que :

$$\boxed{\overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i \, \overline{x}_i}$$

où k représente le nombre de sous-populations.

Rappel : la moyenne de la population est la moyenne arithmétique, pondérée par les effectifs des sous-populations, des moyennes de chaque sous-population.

C'est de cette façon on calcule un indice de Laspeyres ou un indice de Paasche global, en agrégeant (regroupant) les données.

Cette propriété d'agrégation des données n'est pas vérifiée pour les indices de Fisher. C'est essentiellement pour cette raison qu'on utilise les indices de Laspeyres et de Paasche, malgré leurs inconvénients (non réversibilité et non circularité). Cf. notamment le calcul de l'indice des prix à la consommation de l'INSEE.

Exemple 1

On obtient :

$$\boxed{F^{p}_{97/95} = \sqrt{L^{p}_{97/95} \times P^{p}_{97/95}} \times 100 = \sqrt{113,5 \times 115,1} = 114,3}$$

et : $$\boxed{F^{p}_{95/97} = \sqrt{L^{p}_{95/97} \times P^{p}_{95/97}} \times 100 = \sqrt{86,9 \times 88,1} = 87,5}$$

On peut alors vérifier que : $$\boxed{F^{p}_{95/97} = \frac{10^4}{F^{p}_{97/95}} \Leftrightarrow 87,5 = \frac{10^4}{114,3}}$$

De même pour les quantités :

$$\boxed{F^{q}_{97/95} = \sqrt{L^{q}_{97/95} \times P^{q}_{97/95}} \times 100 = \sqrt{155,1 \times 157,3} = 156,2}$$

et : $$\boxed{F^{q}_{95/97} = \sqrt{L^{q}_{95/97} \times P^{q}_{95/97}} \times 100 = \sqrt{63,6 \times 64,5} = 64,1}$$

On peut alors vérifier que : $$\boxed{F^{q}_{95/97} = \frac{10^4}{F^{q}_{97/95}} \Leftrightarrow 64,1 = \frac{10^4}{156,2}}$$

Exemple 2

Pour les prix, on a : $F^{p}_{05/95} = \sqrt{132,3 \times 138,5} = 135,4$

et : $F^{p}_{95/05} = \sqrt{72,2 \times 75,6} = 73,9$

On peut alors vérifier que : $F^{p}_{95/05} = \frac{10^4}{F^{p}_{05/95}} \Leftrightarrow 135,4 = \frac{10^4}{73,9}$

De même pour les quantités : $F^{q}_{05/95} = \sqrt{100,0 \times 104,7} = 102,3$

et : $F^{q}_{95/05} = \sqrt{95,6 \times 100,0} = 97,8$

On peut alors vérifier que : $F^{q}_{95/05} = \frac{10^4}{F^{q}_{05/95}} \Leftrightarrow 102,3 = \frac{10^4}{97,8}$

### 3. L'indice de valeur (ou de dépense) globale

En Économie, on ne considère pas toujours indépendamment les variations de prix et les variations de quantités consommées. Il est utile d'envisager la valeur globale (ou dépense totale), telle que : $d = p \cdot q$ .

Il en résulte que, pour le bien i, l'indice de la valeur globale (ou indice de la dépense totale) s'écrit :

$$D^{i}_{t/0} = P^{i}_{t/0} \times Q^{i}_{t/0} \Leftrightarrow D^{i}_{t/0} = \frac{d^{i}_{t}}{d^{i}_{0}} \times 100 = \frac{p^{i}_{t}\, q^{i}_{t}}{p^{i}_{0}\, q^{i}_{0}} \times 100$$

D'où, pour n biens, l'indice de la valeur globale est : $D_{t/0} = \frac{\sum_{i=1}^{n} p^{i}_{t}\, q^{i}_{t}}{\sum_{i=1}^{n} p^{i}_{0}\, q^{i}_{0}} \times 100$

On démontre que l'on a : $D_{t/0} = \frac{L^{p}_{t/0} \cdot P^{q}_{t/0}}{100} = \frac{L^{q}_{t/0} \cdot P^{p}_{t/0}}{100}$

Reprenons l'exemple 2 précédent. Le tableau permet de calculer directement :

$$D_{05/95} = \frac{\sum_{i=1}^{n} p^{i}_{05}\, q^{i}_{05}}{\sum_{i=1}^{n} p^{i}_{95}\, q^{i}_{95}} \times 100 = 138,5$$

On peut aussi utiliser l'une ou l'autre des égalités ci-dessus :

$$D_{05/95} = \frac{132,3 \times 104,7}{100} = \frac{100 \times 138,5}{100} = 138,5$$

## CHAPITRE 3 : L'INDICE DES PRIX A LA CONSOMMATION DE L'INSEE

L'objectif de ce point est de mettre en évidence quelques éléments relatifs aux difficultés liées à la construction de cet indice, largement utilisé dans l'analyse économique.

Notamment, cet indice est utilisé pour les prévisions macro-économiques de la nation (liens avec le budget de l'État et son déficit). Pour réaliser cette prévision, des hypothèses économiques d'évolution sont posées sur le PIB, les exportations, les importations, la consommation des ménages, l'investissement et l'évolution des prix (indices INSEE). Cet ensemble d'éléments permet la mise au point des projets de loi de finances pour une année donnée.

Les premiers indices de prix à la consommation remontent à 1914.

L'indice des prix à la consommation (IPC) actuel constitue, en 1998, la septième génération d'indices depuis 1914.

Remarques :

1) la structure des biens composant l'indice a considérablement évolué dans le temps.

En 1914, on prenait essentiellement en compte les biens alimentaires.

Aujourd'hui, en 2007, l'indice recouvre presque l'ensemble des produits alimentaires, des produits manufacturés et des services.

Les seules exceptions actuelles concernent les jeux de hasard, les services hospitaliers, l'assurance-vie et l'assurance santé.

2) la définition des ménages a changé également. En 1914, il s'agissait de la population ouvrière. Ensuite, durant une longue période, on a considéré les ménages urbains dont le chef était ouvrier ou employé. À l'heure actuelle, on considère tous les ménages des agglomérations de plus de 2 000 habitants.

Aujourd'hui, les relevés de prix sont effectués dans 96 agglomérations de plus de 2 000 habitants, dispersées sur l'ensemble du territoire.

Il s'agit d'un échantillon fixe , de façon à assurer la continuité de l'indice dans le temps.

La base 1998 porte sur 159 groupes de produits. 27 000 points de vente sont concernés. Toutes les formes de commerce sont envisagées (hypermarchés, supermarchés, marchés, magasins traditionnels, ...), au prorata de leur part de marché.

Chaque mois, 180 enquêteurs relèvent environ 160 000 prix (plus de 50 000 séries de tarifs sont collectées au niveau central).

300 personnes travaillent à plein temps à la réalisation de l'indice des prix à la consommation.

D'un point de vue technique, l'indice des prix INSEE est un indice de Laspeyres chaîné annuellement.

Avantages

1) C'est pratique, car les pondérations (= les quantités consommées) restent fixes. Où il suffit donc de relever des prix (si l'on veut construire un indice de volume, on doit relever les quantités consommées).

2) L'indice des prix de Laspeyres étant une moyenne arithmétique, la propriété d'agrégation des moyennes fonctionne. Ainsi, on peut calculer des indices pour diverses sous-rubriques, que l'on agrége ensuite en trois grandes catégories (produits alimentaires, produits manufacturés, services).

Les coefficients de pondération appliqués sont les coefficients budgétaires des ménages (= proportion de bien consommer par les ménages).

On a alors recours à des moyennes pondérées jusqu'à obtenir l'indice général des prix à la consommation (IPC). Le coefficient de pondération total est égal à 10 000.

Inconvénients

1) L'indice de Laspeyres n'est pas circulaire. Il en résulte que le chaînage est de plus en plus faux au fil des années.

2) il est nécessaire de modifier l'année de base assez souvent (à peu près tous les 10 ans), car les habitudes et comportements des consommateurs évoluent plus ou moins rapidement. Il convient donc de modifier la structure du "panier de la ménagère", c'est-à-dire les 159 groupes de produits.

Or à long terme, se pose le problème du suivi de l'évolution d'un indice sur une longue période, durant laquelle sa définition se modifie en partie.

D'où la nécessité d'utiliser des raccords d'indices (coefficients de raccordement).

Coefficients de raccordement = CR = valeur de l'ancien indice / valeur du nouvel indice

On a : $\boxed{CR = \frac{I_{a/0}}{100}}$ . En t = a , on calcule les deux indices (l'ancien et le nouveau) simultanément et l'on pose de, par hypothèse : valeur du nouvel indice = 100.

Exemple de calcul

Considérons la nouvelle base de 1998 et l'ancienne base de 1990.

Supposons qu'un indice base 100 en 1990 soit égal à 137,2 en 1998.

Alors, le coefficient de raccordement est égal à : CR = 137,2 / 100 = 1,372 .

Il s'agit bien évidemment dans le coefficient multiplicateur.

Si, par hypothèse, on veut connaître la valeur de l'indice pour l'année 2003 dans l' ancienne base 1990 (à fin de comparaisons statistiques), on aura par exemple : $I_{03/98} = 118,5$

Grâce au coefficient de raccordement, on pourra écrire :

$$I_{03/90} = I_{03/98} \times CR = 118,5 \times 1,372 = 162,6$$

Remarques :

1) Ce calcul est approximatif, puisque l'indice de Laspeyres n'est pas circulaire.

Rigoureusement en effet : $100 \times I_{03/90} \neq I_{03/98} \times I_{98/90}$

2) Les deux indices (base 1990 et 1998) n'ont plus le même champ : d'une part les produits ne sont plus les mêmes, d'autre part les proportions consommées ne sont plus les mêmes (coefficients de pondération différents).

**Trois problèmes particuliers**

A) le traitement des changements de produits, car d'anciens produits disparaissent et de nouveaux produits apparaissent sur le marché. La solution consiste à rechercher en pratique des produits dont les caractéristiques sont similaires. Par exemple, remplacement progressif des appareils photo argentiques par les appareils photo numériques.

B) choix de l'année de base (année de référence). Il est important de ne pas choisir une année trop bonne ou trop mauvaise. Il eût été inapproprié de retenir 1968 ou 1975 comme année de référence.

C) le traitement des produits frais (fruits et légumes). Leurs pondérations sont variables selon les mois de l'année, de façon à tenir compte des fortes variations de consommation dues au phénomènes saisonniers.

# PARTIE 4 : LES SÉRIES CHRONOLOGIQUES

**Définition** : une série statistique est appelée **série chronologique** (ou série temporelle, ou chronique) lorsque la variable étudiée prend des valeurs ordonnées dans le temps.

Les **principaux objectifs** de l'étude des séries chronologiques sont de :

a) décrire l'évolution d'une grandeur dans le temps, en vue de mieux comprendre et expliquer un phénomène donné ;

b) comparer l'évolution de plusieurs grandeurs dans le temps, de façon à déterminer s'il existe entre elles des relations (notion de covariation) ;

c) faciliter l'élaboration de prévisions conjoncturelles : après avoir posé des hypothèses appropriées, l'analyse de séries chronologiques rétrospectives aide à la prévision des évolutions futures d'un phénomène donné.

Remarque : les séries chronologiques présentent des caractéristiques particulières qui nécessitent des traitements spécifiques. Notamment, on ne peut pas les assimiler à des séries à un caractère, de type variable quantitative discrète.

Nous allons voir que les séries chronologiques brutes sont composées de plusieurs éléments fondamentaux, qu'on essaye d'isoler pour une analyse plus fine et plus pertinente du phénomène observé :

a) le mouvement extra-saisonnier, qui comprend le trend (mouvement de tendance générale à long terme) et le cycle (mouvement cyclique).

b) le mouvement saisonnier (variations saisonnières annuelles).

c) le mouvement résiduel (aléatoire ou accidentel) qui correspond aux variations accidentelles.

Selon que l'on cherche à mettre en évidence l'un ou l'autre de ces mouvements de (dans un but d'analyse de croissance ou de fluctuations à long terme ou à court terme), on met en œuvre des méthodes différentes (lissage, filtrage, désaisonnalisation).

Dans un premier temps, comme pour les séries à un ou deux caractères, on va chercher à rendre compte graphiquement des phénomènes étudiés.

## CHAPITRE 1 : REPRÉSENTATION GRAPHIQUE DES SÉRIES CHRONOLOGIQUES

### 1. Une distinction fondamentale : les données exprimées en termes de stocks et les données exprimées en termes de flux

Avant de représenter une série chronologique brute sur du papier millimétré, il est nécessaire, au préalable, d'avoir répondu à la question suivante : "relativement à la série étudiée, a-t-on affaire à des données exprimant des flux ou à des données exprimant des stocks" ?

a) Les valeurs prises par une grandeur peuvent être mesurées à une date donnée (par ex. le nombre de résidents en France au 31 mars 1999, lors du dernier recensement général de la population ; le stock de capital fixe des entreprises françaises au 31 décembre 2006 ; les avoirs en euros de la Banque de France au 1er janvier 2007 ; le parc des poids lourds immatriculés en France au 31 décembre 2006 ; …). La série chronologique associée à une telle grandeur est composée de données exprimées en termes de **stocks**.

b) Les valeurs prises par une grandeur peuvent être mesurées durant une période de temps comprise entre deux dates données (par ex. l'accroissement d'une population entre deux recensements successifs ; l'évolution des importations ou des exportations françaises durant l'année 2006 ; l'évolution du nombre de demandeurs d'emploi en fin de mois, entre le 31 janvier 2007 et le 28 février 2007 ; …). La série chronologique associée à une telle grandeur est composée de données exprimées en termes de **flux**.

La distinction précédente est fondamentale, en vue de reporter correctement sur un graphique la valeur qui correspond à un événement donné.

Règle applicable aux représentations graphiques des séries chronologiques :

a) si l'on a affaire à des **stocks**, on fait correspondre la valeur atteinte par la grandeur étudiée **à une date donnée** (1er janvier, 31 décembre, …).

b) si l'on a affaire à des **flux**, les points du graphique sont portés à la verticale du **milieu de la période correspondante** (mois, trimestre, année, …).

Dans les deux cas, on place le temps en abscisse et la valeur de la grandeur étudiée en ordonnée. Pour l'ensemble d'une série de données, on obtient une succession de points.

On relie ces points deux à deux par des segments de droite, posant en cela **l'hypothèse d'une linéarité de l'évolution de la grandeur étudiée** entre deux dates successives (stocks) ou entre deux milieux de période successifs (flux).

### 2. La représentation graphique de séries d'indices en coordonnées arithmétiques

Intérêt : au lieu de renvoyer les fluctuations dans le temps des valeurs absolues d'une grandeur, un graphique réalisé à partir de séries d'indices renvoie, pour chaque segment (on pose une hypothèse d'évolution linéaire entre deux dates données d'observation), un taux de croissance (annuel, trimestriel, mensuel, …), qu'on peut mesurer directement sur le graphique. Cette méthode permet aussi de comparer les taux d'évolution de plusieurs grandeurs sur un même graphique. Lorsque les valeurs absolues des grandeurs comparées sont très différentes, c'est d'ailleurs l'une des seules solutions pratiques qui autorise des comparaisons pertinentes.

Exemple :

Le tableau ci-après indique, au 31 décembre de chaque année, l'évolution du nombre de mineurs de fond en France (en milliers) et, pour chaque année, la production de charbon correspondante (en millions de tonnes) :

| Années | Effectif du fond (en milliers) | Production de charbon (en Mt) |
|---|---|---|
| **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** | 42,4 | 25,1 |
| **1977** | 38,6 | 24,4 |
| **1978** | 35,9 | 22,4 |
| **1979** | 32,7 | 21,1 |
| **1980** | 30,8 | 20,7 |

Ces deux séries sont de nature différente :

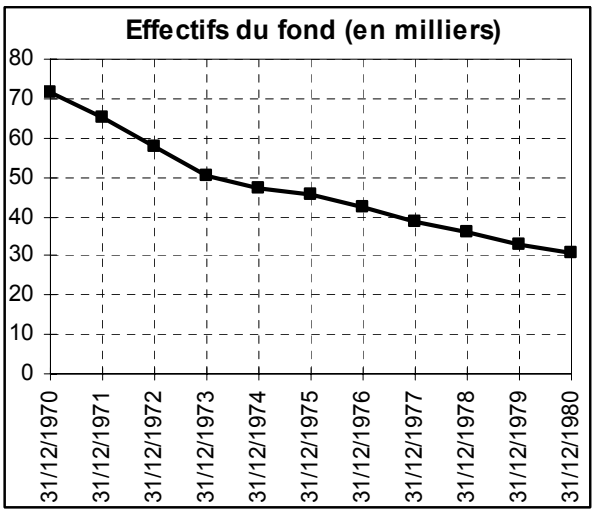
- l'une correspond à un **stock** mesuré à une date déterminée : les **effectifs** au 31-12. Dans ce cas, les points successifs sont portés à l'aplomb de cette date sur un graphique (années en abscisse, effectifs en ordonnée).

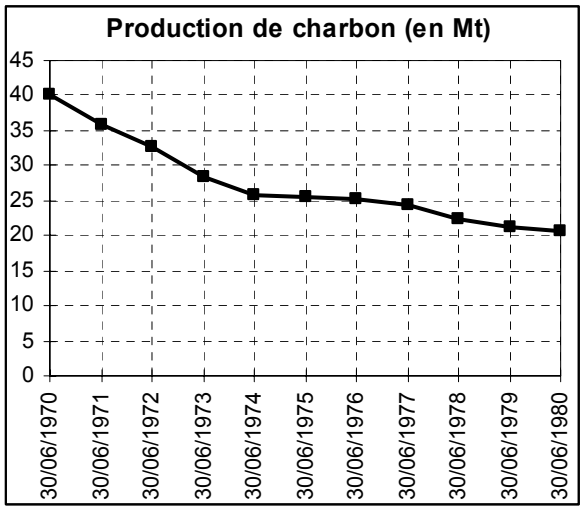
- l'autre correspond à un **flux** qui se rapporte à une période annuelle : la **production** à l'année t. Dans ce cas, les points représentant la production annuelle de charbon sont portés à la verticale du milieu de période sur un graphique (années en abscisse, production de charbon en ordonnée).

On peut représenter séparément chacune de ces séries chronologiques : au 31 décembre pour le nombre de mineurs et au 30 juin pour la production de charbon.

| Années | Effectifs du fond (en milliers) |
|---|---|
| **31/12/1970** | 71,3 |
| **31/12/1971** | 65,3 |
| **31/12/1972** | 57,6 |
| **31/12/1973** | 50,4 |
| **31/12/1974** | 47,1 |
| **31/12/1975** | 45,8 |
| **31/12/1976** | 42,4 |
| **31/12/1977** | 38,6 |
| **31/12/1978** | 35,9 |
| **31/12/1979** | 32,7 |
| **31/12/1980** | 30,8 |

| Années | Production de charbon (en Mt) |
|---|---|
| **30/06/1970** | 40,1 |
| **30/06/1971** | 35,8 |
| **30/06/1972** | 32,7 |
| **30/06/1973** | 28,4 |
| **30/06/1974** | 25,7 |
| **30/06/1975** | 25,6 |
| **30/06/1976** | 25,1 |
| **30/06/1977** | 24,4 |
| **30/06/1978** | 22,4 |
| **30/06/1979** | 21,1 |
| **30/06/1980** | 20,7 |




La comparaison directe des deux graphiques n'est pas pertinente. En effet :

a) on ne compare pas les deux grandeurs aux mêmes dates (31 décembre dans un cas et 30 juin dans l'autre).

b) les échelles utilisées pour représenter chaque grandeur sont très différentes : d'une part on a affaire à des milliers de personnes et d'autre part à des millions de tonnes.

A supposer que le problème a) soit résolu, de manière générale, lorsque les ordres de grandeur des différentes séries brutes sont très dissemblables, leur comparaison sur un même graphique peut s'avérer difficile.

Des difficultés de lecture graphique peuvent apparaître si l'on a affaire à une série chronologique dont les variations périodiques sont importantes (par exemple une évolution exponentielle). De même, si l'on doit représenter plusieurs séries chronologiques dont les valeurs absolues sont très différentes.

Conséquence :

Le plus souvent, on a besoin de pouvoir représenter les séries chronologiques autrement qu'en considérant les valeurs absolues des données brutes. Pour résoudre ce problème, il existe principalement deux solutions :

- on remplace les valeurs absolues par des indices (séries d'indices).

- on utilise un graphique à ordonnée logarithmique, nécessitant un papier spécial appelé papier semi-logarithmique (cf. point 3).

Dans notre exemple, nous allons réaliser successivement deux transformations sur nos séries chronologiques.

**Étape 1**

De manière à ramener la comparaison des deux séries à une même date, nous allons modifier les données relatives aux effectifs des mineurs.

Nous considérons que les effectifs des mineurs sont estimés au 30 juin de chaque année, en calculant la moyenne des effectifs de l'année en cours (au 31 décembre de l'année t) et de l'année précédente (au 31 décembre de l'année t-1).

A noter que pour réaliser cela, on pose une hypothèse de linéarité sur l'évolution de l'effectif des mineurs en cours d'année.

Remarques :

a) on ignore le nombre de mineurs au 31 décembre 1969. Par conséquent, on perd une valeur (au 30 juin 1970). La première valeur disponible, pour la comparaison des deux séries, devient le 30 juin 1971.

À noter que le si l'on possède de l'information par ailleurs sur l'évolution du secteur, il est toujours possible de poser une hypothèse plausible concernant le nombre de mineurs à la date du 31 décembre 1969. Dans ce cas évidemment, on peut calculer une valeur moyenne au 30 juin 1970.

b) on aurait également pu modifier la série des productions annuelles. Mais dans ce cas, on aurait dû poser une hypothèse de linéarité de l'évolution de la production sur deux périodes (deux ans) au lieu d'une seule. L'hypothèse de linéarité pouvant être déjà forte sur une année, on limite donc mieux l'approximation (sur une année), en modifiant la série des effectifs des mineurs (c'est-à-dire la série exprimée en termes de stock).

**Étape 2**

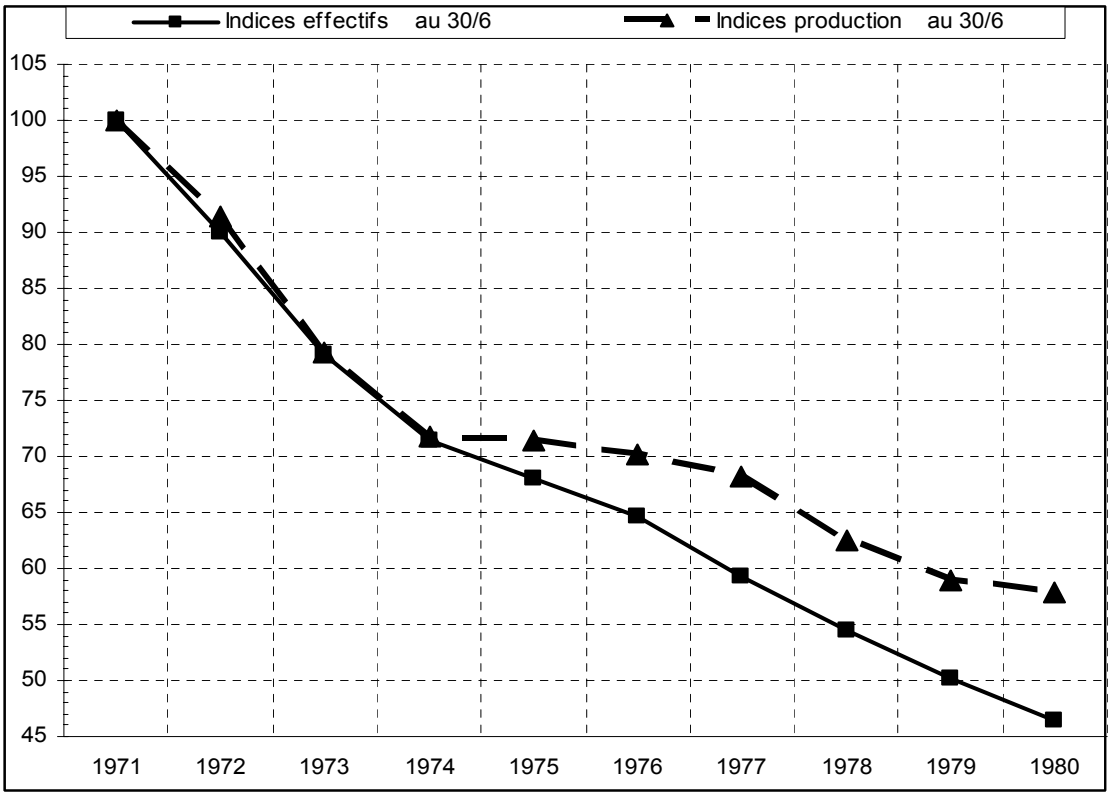
Pour chacune des deux séries, nous considérons un **indice base 100 au 30 juin 1971**. De ce fait, on résout le problème posé par les unités :

| | | | (1) | | (2) |
|---|---|---|---|---|---|
| **Ans** | **Effectif du fond 31/12 (en milliers)** | **Effectif du fond 30/06 (en milliers)** | **Indices effectifs au 30/6** | **Production charbon au 30/6 (en Mt)** | **Indices production au 30/6** |
| 1970 | **71,3** | | | 40,1 | |
| 1971 | **65,3** | **68,3** | 100,0 | **35,8** | 100,0 |
| 1972 | 57,6 | **61,5** | **90,0** | **32,7** | **91,3** |
| 1973 | 50,4 | 54,0 | 79,1 | 28,4 | 79,3 |
| 1974 | 47,1 | 48,8 | 71,4 | 25,7 | 71,8 |
| 1975 | 45,8 | 46,5 | 68,0 | 25,6 | 71,5 |
| 1976 | 42,4 | 44,1 | 64,6 | 25,1 | 70,1 |
| 1977 | 38,6 | 40,5 | 59,3 | 24,4 | 68,2 |
| 1978 | 35,9 | 37,3 | 54,5 | 22,4 | 62,6 |
| 1979 | 32,7 | 34,3 | 50,2 | 21,1 | 58,9 |
| 1980 | 30,8 | 31,8 | 46,5 | 20,7 | 57,8 |

Par exemple : (71,3 + 65,3) / 2 = 68,3

De même : (61,5 / 68,3) x 100 =90,0 et : (32,7 / 35,8) x 100 =91,3

Si l'on reporte sur un même graphique les séries d'indices (1) et (2), on obtient, au 30 juin de chaque année :



Remarques :

a) Les indices étant des nombres sans dimension, la comparaison des deux séries chronologiques devient possible directement (il n'y a plus de problèmes d'unités). Les pentes des segments indiquent les variations relatives, entre la période de référence (1971) et une date donnée.

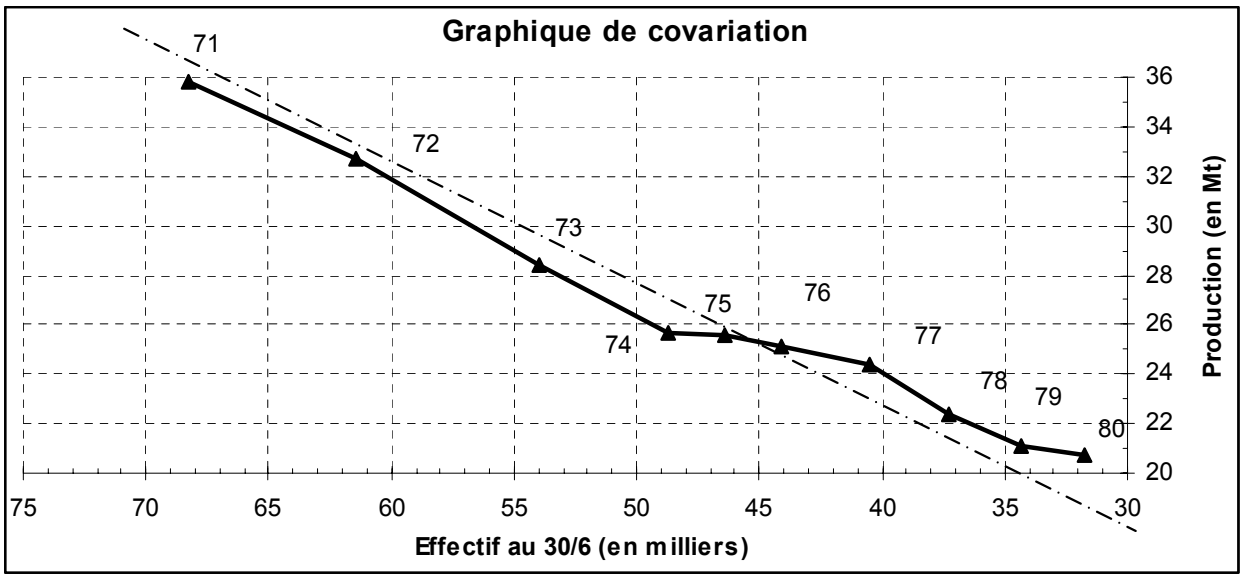
b) Le tableau précédent permet de lire directement la décroissance des effectifs et de la production, simultanée jusqu'en 1974, puis divergente entre 1975 et 1980.

Ainsi., au 30/6/72, on a : Effectifs : 100 - 90 = - 10 %  Production : 100 - 91,3 = - 8,7 %

Au 30/6/80, on a : Effectifs : 100 - 46,5 = - 53,5 %  Production : 100 - 57,8 = - 42,2 %

Le graphique précédent permet de visualiser les deux phénomènes sur une même base. Il apparaît clairement que les deux mouvements sont similaires à la baisse. Toutefois, une différence apparaît entre les deux séries en 1974, quant au rythme de la décroissance.

Cette visualisation peut être complétée par un graphique de covariation (ou de corrélation). Pour réaliser celui-ci, on indique en abscisses les effectifs (en milliers) et en ordonnée la production (en millions de tonnes). Les points du nuage correspondent à chacune des années (au 30 juin) considérées. Ici, on obtient :




Remarque : pour faire apparaître la décroissance des valeurs, l'échelle des abscisses a été inversée.

Ce graphique permet de mettre en évidence la forte covariation de la production et des effectifs dans le temps : chute rapide entre 1970 et 1974, ralentissement des baisses d'effectifs et maintien de la production en 1974-76, puis reprise de la baisse.

On peut interpréter le palier 74-76 relatifs à la production comme un accroissement de la productivité du travail : la production stagne alors que les effectifs continuent de diminuer.

Remarque terminale :

L'analyse graphique précédente permet de mettre en évidence le fait qu'il existe une relation forte de type linéaire entre les deux variables.

Toutefois, on ne parle pas de coefficient de corrélation dans le cas d'une série chronologique. En effet, **chacune de ces séries est liée fonctionnellement au temps**.

La comparaison directe des deux séries risque, le plus souvent, d'être faussée et l'interprétation directe du coefficient de covariation est généralement erronée, car celui-ci ne rend compte que d'une liaison artificielle entre les deux variables.

Rappelons-nous que dans le cas de variables quantitatives non chronologiques, on peut obtenir de très bons coefficients de corrélation linéaire sans pour autant que les deux variables considérées soient directement dépendantes l'une de l'autre.

Exemple : en été, on observe que le nombre de lunettes de soleil vendu s'accroît, de même que le nombre de boissons fraîches vendues. Il en résulte un coefficient de corrélation linéaire élevé, mais la véritable variable explicative de ces hausses correspond à de fortes températures estivales, dues à un taux d'ensoleillement important : c'est la chaleur qui explique la hausse des ventes de lunettes de soleil et de boissons fraîches, sans qu'il existe une relation de cause à effet directe entre ces deux types de vente. Conclusion, ne pas confondre corrélation (linéaire) et (relation de) causalité.

En ce qui concerne les séries chronologiques, plutôt que de corrélation, on parle de **covariation** entre deux séries. On peut calculer un coefficient de covariation linéaire entre deux séries chronologiques. Celui-ci mesure l'intensité d'une éventuelle liaison entre des **variables qui sont, chacune, liées au temps**. L'interprétation de ce coefficient est similaire à celle du coefficient de corrélation linéaire.

$$C = \frac{\frac{1}{k}\sum_{t=1}^{k}(x_t - \bar{x})(y_t - \bar{y})}{\sigma_{x_t} \cdot \sigma_{y_t}} = \frac{\sigma_{x_t, y_t}}{\sigma_{x_t} \cdot \sigma_{y_t}}$$

k représente le nombre d'observations successives.

C (coefficient de covariation) est compris entre -1 et + 1.

Si la valeur de C est proche de +1 ou -1, alors il est intéressant de rechercher les véritables causes économiques qui sont à l'origine de cette forte covariation.

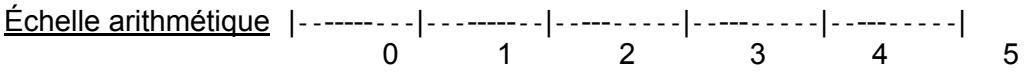
## 3. La représentation graphique des séries chronologiques dans un repère à ordonnée logarithmique

### 31. La logique du papier semi-logarithmique

Une échelle arithmétique renvoie les valeurs absolues des variables concernées. Si les séries que l'on cherche à comparer sur un même graphique présentent des niveaux très différents, il devient difficile de trouver une échelle satisfaisante en ordonnée. Par ailleurs, une échelle arithmétique ne rend pas compte des variations relatives (taux d'évolution) des phénomènes étudiés. Pour ces deux raisons, on a vu dans le point 2 que l'on peut travailler sur des séries d'indices plutôt que sur les valeurs absolues des grandeurs.

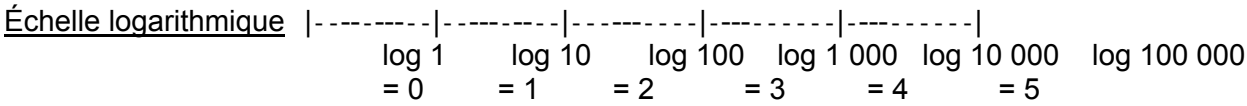
Il existe une autre solution : celle-ci consiste à utiliser une ordonnée logarithmique. Pour cela, on se sert d'un papier millimétré spécial, qui possède une échelle arithmétique en abscisse et une échelle logarithmique (en base 10) en ordonnée. On parle alors de **papier semi-logarithmique**. Ce dernier a pour but d'éviter d'avoir à calculer les logarithmes décimaux de toutes les valeurs que l'on cherche à représenter graphiquement.

Sur une échelle arithmétique, les distances entre 0 et 1, entre 1 et 2, entre 2 et 3, etc. sont caractérisées par des intervalles égaux.




Sur une échelle logarithmique, ce sont les distances entre log 1 (= 0) et log10 (= 1), entre log10 (= 1) et log 100 (= 2), entre log 100 (= 2) et log 1 000 (= 3), etc. qui sont caractérisées par des intervalles égaux. Par suite, la distance log 1 à log 10 est égale à la distance log10 à log 100, elle-même égale à la distance log 100 à log 1 000, etc.

**Pour des longueurs égales, une échelle logarithmique permet de lire des rapports égaux, donc des variations relatives égales (1/10 = 10/100 = 100/1 000 = … = 10 %)**.




En ordonnée logarithmique, l'intervalle entre deux graduations, puissance de 10, s'appelle un **module**. Il existe des feuilles de papier millimétré semi-logarithmique à 2, 3 ou 4 modules (voire plus), selon les besoins. Il est important de noter que, sur de telles feuilles, **l'échelle n'est pas fixée à l'avance**, car selon les phénomènes décrits, on aura par exemple deux modules, allant respectivement de 1 à 10 et de 10 à 100, ou par exemple trois modules, allant respectivement de 100 à 1 000, de 1 000 à 10 000 et de 10 000 à 100 000. C'est pourquoi il devient aisé de représenter sur un même graphique des séries dont les valeurs absolues sont très différentes.

Exemple : pour une grandeur variant de 150 à 150 000, le tracé nécessite un papier semi-logarithmique à quatre modules de (100-1 000 ; 1 000-10 000 ; 1 0000-100 000 ; 100 000-1 000 000).

A l'intérieur de chaque module, on trouve des graduations qui sont, elles aussi, proportionnelles aux logarithmes décimaux des nombres correspondants. Entre le début et la fin d'un même module, on observe un resserrement des graduations. Ce profil se retrouve à l'identique à l'intérieur de chacun des modules.

Ainsi, entre log 1 et log 2, la distance est proportionnelle à :
log 2 - log 1 = 0,30103 - 0 = 0,30103 (soit près du tiers d'un module !)

Ainsi, entre log 2 et log 3, la distance est proportionnelle à :
log 3 - log 2 = 0,47712 - 0,30103 = 0,17609

Ainsi, entre log 3 et log 4, la distance est proportionnelle à :
log 4 - log 3 = 0,60206 - 0,47712 = 0,12494

Ainsi, entre log 4 et log 5, la distance est proportionnelle à :
log 5 - log 4 = 0,69897 - 0,60206 = 0,09691

Ainsi, entre log 5 et log 6, la distance est proportionnelle à :
log 6 - log 5 = 0,77815 - 0,69897 = 0,07918

Ainsi, entre log 6 et log 7, la distance est proportionnelle à :
log 7 - log 6 = 0,84510 - 0,77815 = 0,06695

Ainsi, entre log 7 et log 8, la distance est proportionnelle à :
log 8 - log 7 = 0,90309 - 0,84510 = 0,05799

Ainsi, entre log 8 et log 9, la distance est proportionnelle à :
log 9 - log 8 = 0,95424 - 0,90309 = 0,05115

Ainsi, entre log 9 et log 10, la distance est proportionnelle à :
log 10 - log 9 = 1 - 0,95424 = 0,04576

Remarque : entre log 10 et log 20, la distance est proportionnelle à :
log 20 - log 10 = 1,30103 - 1 = 0,30103 **On retrouve le même écart !**

Entre log 90 et log 100, la distance est proportionnelle à :
log 100 - log 90 = 2 - 1,95424 = 0,04576 **Idem**

**Exemple** : pour un module de 10 cm, on a à peu près : de 1 à 2 = 3,0 cm ;
de 2 à 3 = 1,8 cm ; de 3 à 4 = 1,2 cm ; de 4 à 5 = 1,0 cm ; de 5 à 6 = 0,8 cm ;
de 6 à 7 = 0,7 cm ; de 7 à 8 = 0,6 cm ; de 8 à 9 = 0,5+ cm ; de 9 à 10 = 0,5- cm ;

Par ailleurs, très grossièrement, on a à peu près :

1/3 de 1 à 2- ; 1/3+ de 2 à 5 ; 1/3- de 5 à 10
1/2 de 1 à 3- ; 1/2 de 3 à 10
3/4+ de 1 à 6

En résumé :

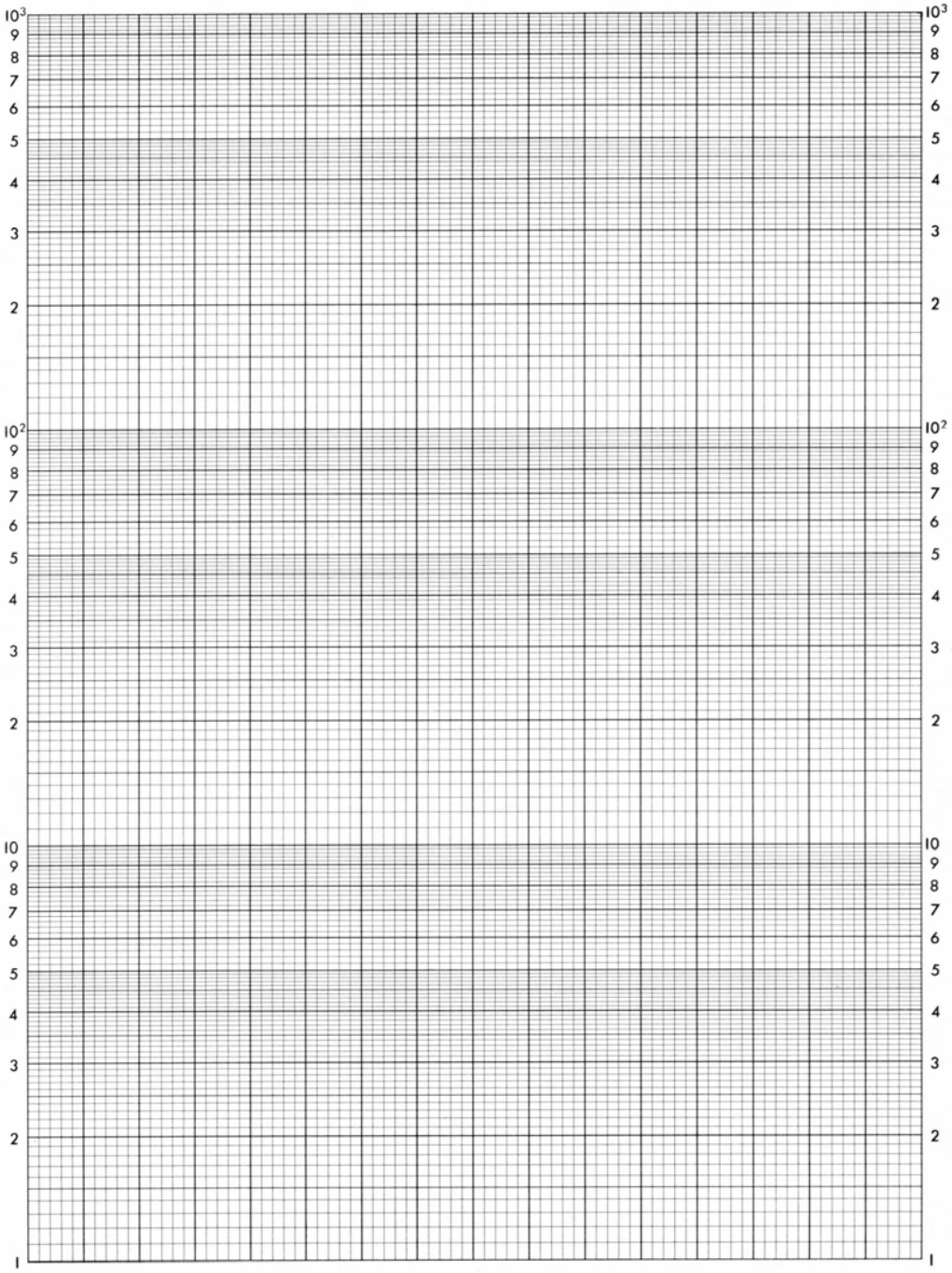
- l'utilisation du papier semi-logarithmique évite tout calcul de logarithme.

- pour des longueurs égales, ce papier permet de lire des rapports égaux, donc des variations relatives égales. En effet : 1/10 = 10/100 = 100/1 000 = … = 10 %

Au contraire, une échelle arithmétique permet de lire des variations absolues égales, pour des longueurs égales : 1 cm = 10 K€ ; 2 cm = 20 K€ ; 3 cm = 30 K€ ; …

Retenir que **si le taux de croissance d'une grandeur est constant**, sur un diagramme semi-logarithmique, la représentation graphique de la série renvoie une **droite**.

En **ordonnée arithmétique**, la représentation graphique d'une telle série renvoie une **exponentielle** (fonction $10^x$).


10³
9
8
7
6
5
4
3
2
10²
9
8
7
6
5
4
3
2
10
9
8
7
6
5
4
3
2
1
B 2270

## 32. Exemple de détermination graphique de taux de croissance

L'utilisation du papier semi-logarithmique permet donc d'une part de représenter facilement une ou plusieurs séries chronologiques dont les variations dans le temps sont importantes, d'autre part de **déterminer graphiquement, sans calcul algébrique, des taux de croissance** (global, moyen, pour une période de temps donnée).

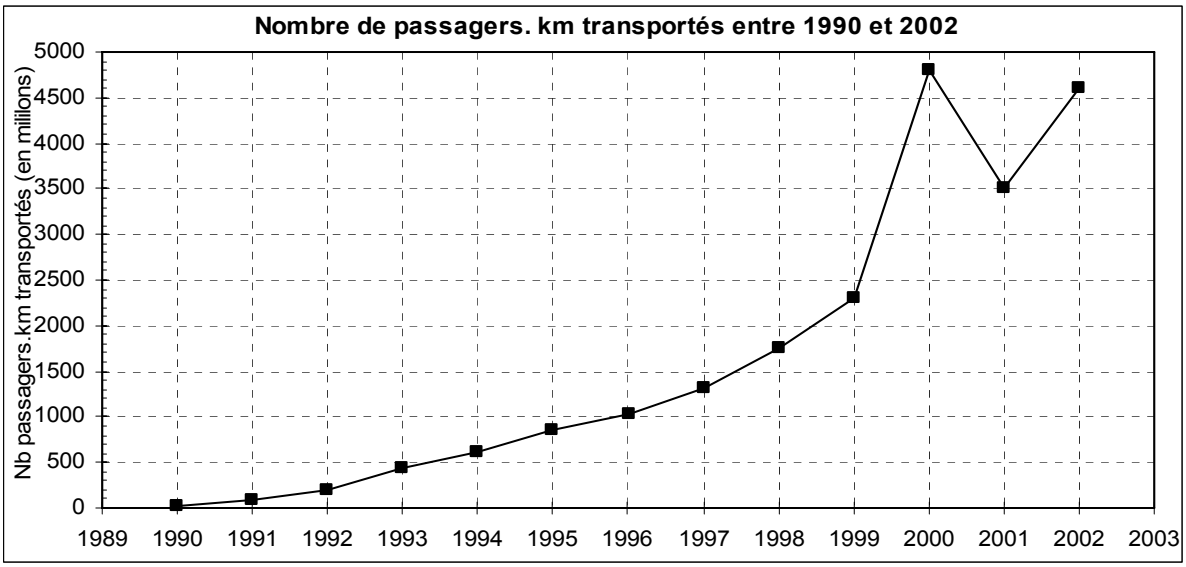
Soulignons que chaque segment de la courbe d'évolution du phénomène étudié indique directement le taux d'évolution de la période (année) correspondante.

Considérons le trafic voyageurs d'une compagnie aérienne, exprimé en passagers.km :

| Années | Nombre de passagers.km transportés (en millions) | Indice base 100 en 1990 |
|---|---|---|
| **1990** | 30 | 100,00 |
| **1991** | 80 | 266,67 |
| **1992** | 200 | 666,67 |
| **1993** | 430 | 1433,33 |
| **1994** | 620 | 2066,67 |
| **1995** | 850 | 2833,33 |
| **1996** | 1 030 | 3433,33 |
| **1997** | 1 320 | 4400,00 |
| **1998** | 1 750 | 5833,33 |
| **1999** | 2 300 | 7666,67 |
| **2000** | 4 800 | 16000,00 |
| **2001** | 3 500 | 11666,67 |
| **2002** | 4 600 | 15333,33 |

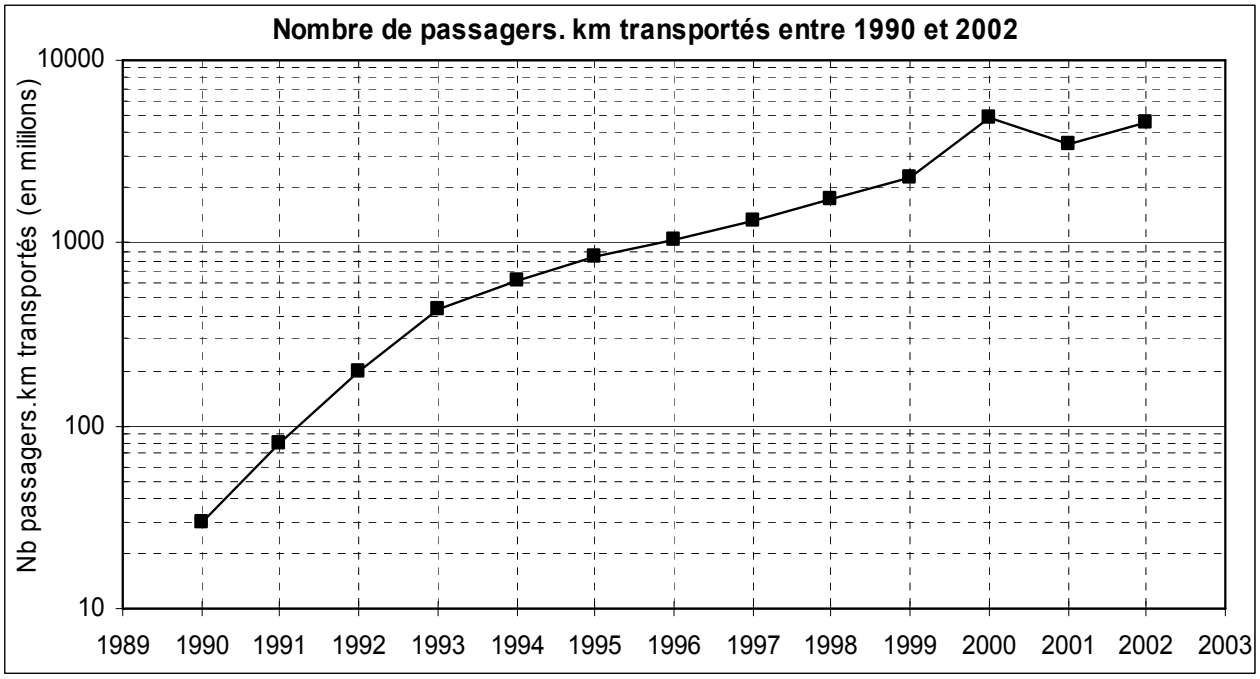
Remarque : les données sont exprimées en termes de **flux** et les valeurs représentées sur un graphique correspondent au milieu de la période (ici l'année) correspondante.

On peut représenter cette série chronologique brute en échelles arithmétiques, mais cela ne permet pas de déterminer graphiquement des taux de croissance.

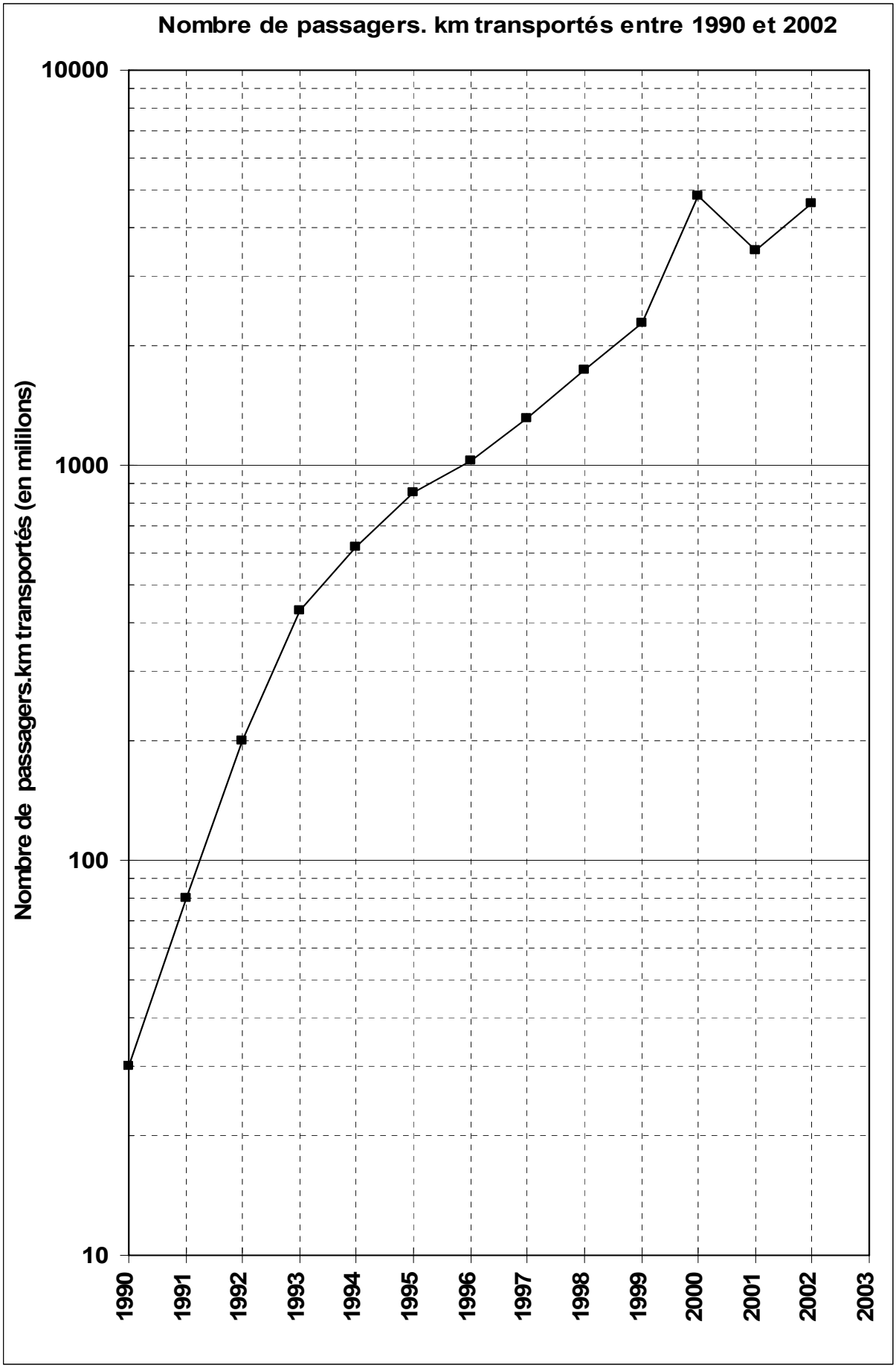


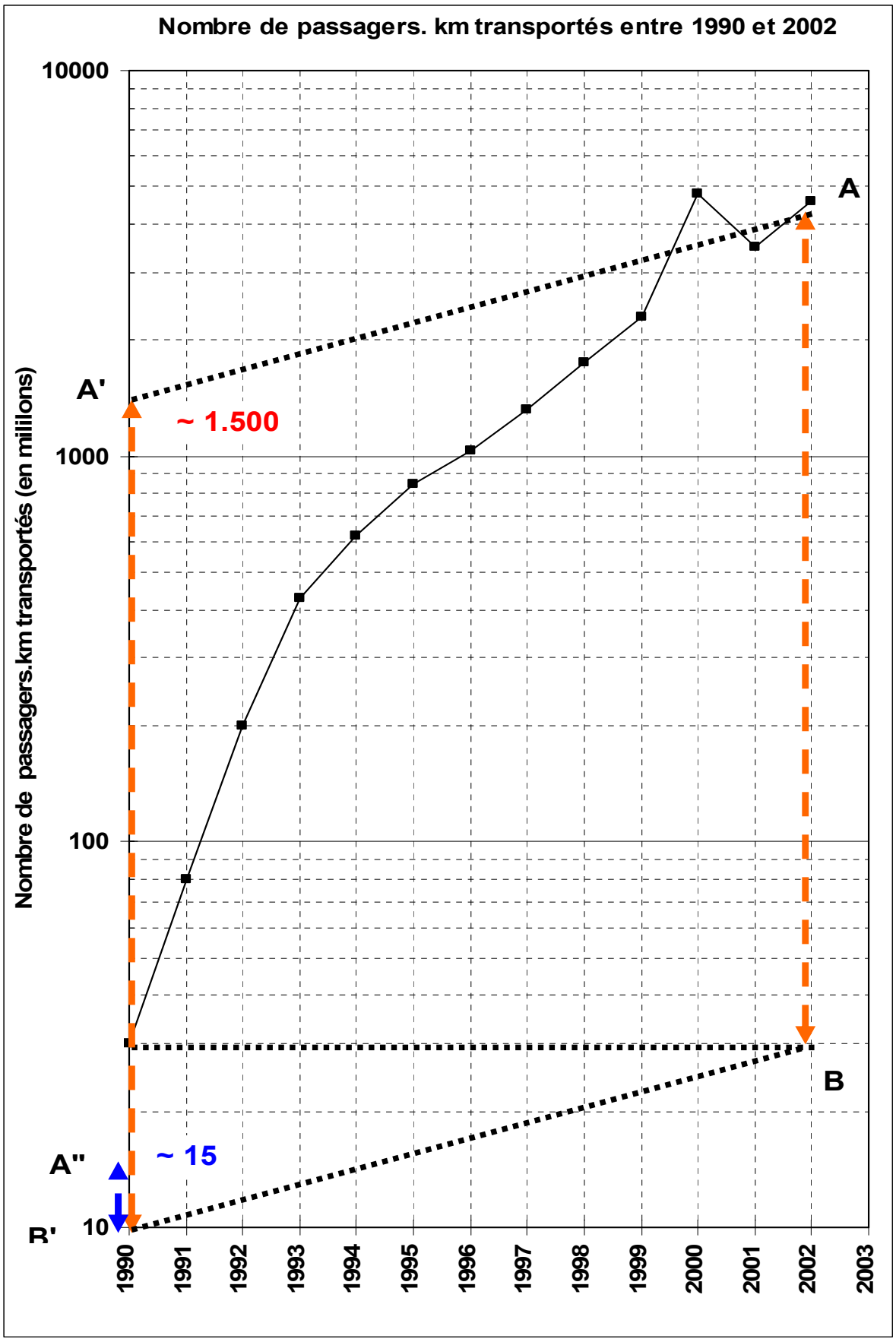
Le graphique suivant, réalisé à l'aide d'un tableur correspond à ce que l'on obtiendrait sur une feuille de papier semi-logarithmique à trois modules, compte tenu de la grande disparité des valeurs de cette série.

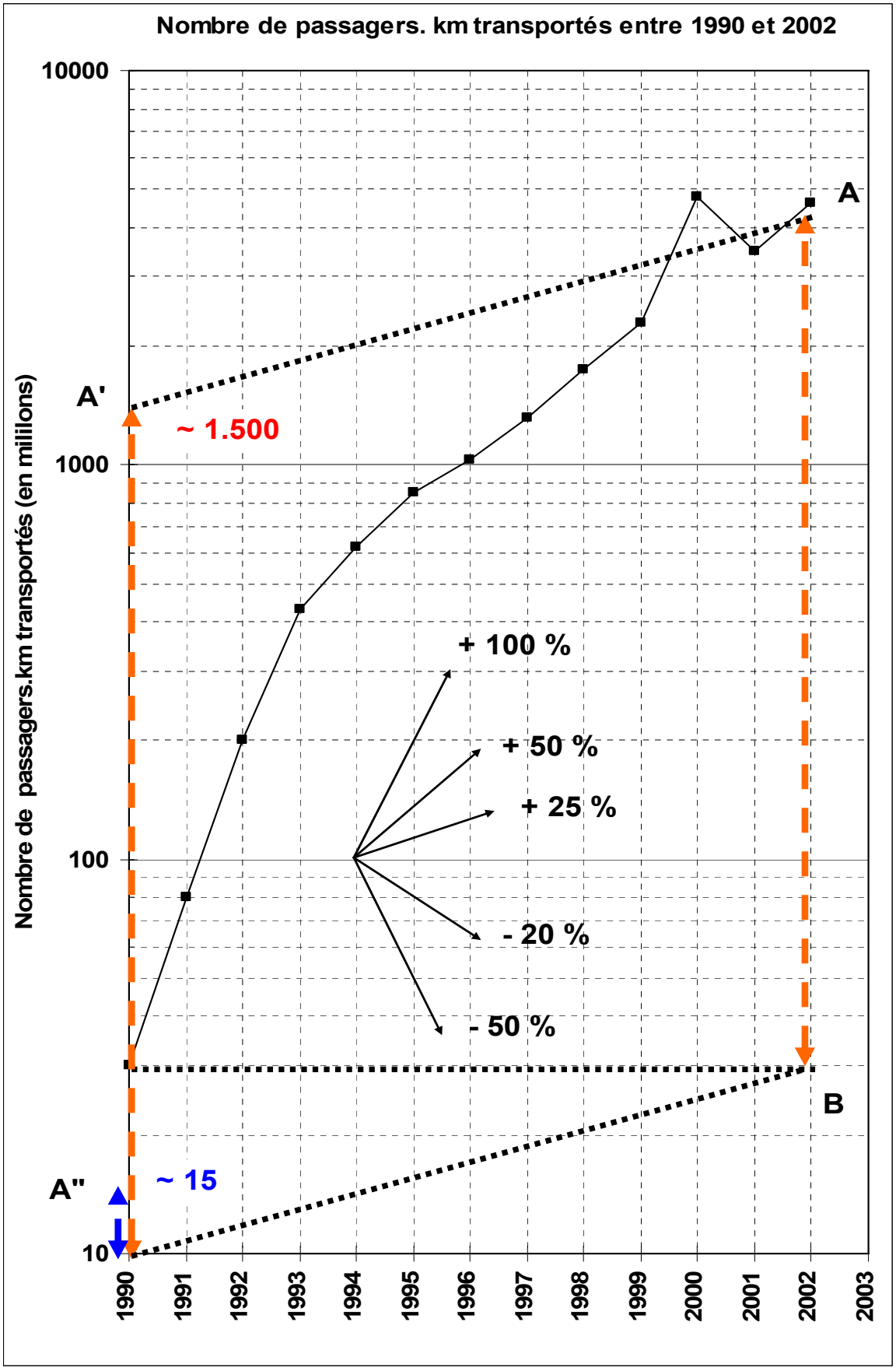
Pour obtenir une ordonnée logarithmique, il suffit de cocher la case correspondante, en même temps que l'on fixe les paramètres relatifs à l'échelle verticale du graphique dans la fenêtre appropriée :




Le graphique de la page suivante reprend le graphique ci-dessus en l'agrandissant, de façon à permettre une meilleure visualisation de la détermination graphique de taux de croissance :


Nombre de passagers. km transportés entre 1990 et 2002
Nombre de passagers.km transportés (en millions)
10000
1000
100
10
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003


Nombre de passagers. km transportés entre 1990 et 2002
Nombre de passagers.km transportés (en millions)
10000
1000
100
10
A
A'
A"
B
B'
~ 1.500
~ 15
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003


Nombre de passagers. km transportés entre 1990 et 2002
Nombre de passagers.km transportés (en millions)
10000
1000
100
10
A
A'
A"
B
~ 1.500
~ 15
+ 100 %
+ 50 %
+ 25 %
- 20 %
- 50 %
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003

A titre de comparaison, l'encadré présente le traitement algébrique associé à cette série.

Taux de croissance global sur l'ensemble de la période 1990-2002 :

$$T = (4.600 - 30) / 30 = 152,33 , \text{ soit : } 15\,233\ \%.$$

Taux de croissance moyen annuel constant (13 dates,12 périodes) :

$$a = \sqrt[12]{\frac{4.600}{30}} - 1 = \sqrt[12]{153,33} - 1 = 1,521 - 1 = 0,521 , \text{ soit : } 52,1\ \% \text{ par an}$$

**Procédure de détermination graphique du taux de croissance global de la variable**

a) On mesure la distance AB (A correspond à la dernière valeur et B à la 1ère).

b) On reporte cette longueur à l'origine d'un module (ici 10), selon A ' B ' .

c) On lit la valeur de l'ordonnée A ' . Ici, celle-ci est à peu près égale à 1 500. Il faut rapporter cette valeur à celle de B ' (origine du module), qui est ici égale à 10. On pose le rapport : valeur de A ' / valeur de B ' = 1 500 / 10 = 150. Ce résultat représente le **coefficient multiplicateur du nombre de passagers.km, de 1990 à 2002 (c-à-d 1 + T)**. Pour obtenir le taux de croissance global du nombre de passagers.km, sur l'ensemble de la période 1990-2002, on pose : T = 150 - 1 = 149, soit : **T = 14 900 %**.

Le nombre de passagers.km de la compagnie aérienne a été multiplié par 150, entre 1990 et 2002, ou encore ce nombre a crû d'environ 15 000 % sur la période 1990-2002.

**Procédure de détermination graphique du taux de croissance moyen annuel constant**

a) On considère la mesure du segment AB (ou du segment A' B').

b) On calcule la valeur AB / 12 = AB / nombre de périodes = A " B ' .
12 représente ici le nombre de périodes considérées, car entre 1990 et 2002, il y a 12 périodes d'accroissement du trafic. Sur le graphique précédent, on a : AB = 14 cm. Donc : A " B ' = 14 / 12 ≈ 1,2 cm.

c) On reporte cette longueur à partir de B ', d'où le segment A ' ' B ' .

d) On lit la valeur de l'ordonnée A ' ' . Ici, celle-ci est à peu près égale à 15. Il faut rapporter cette valeur à celle de B '(origine du module), qui est ici égale à 10. On pose le rapport : valeur de A ' ' / valeur de B ' = 15 / 10 = 1,5. Ce résultat représente le **coefficient multiplicateur du nombre de passagers.km, sur un an, en moyenne (c-à-d 1 + a)**. Le taux de croissance moyen annuel constant du nombre de passagers.km, est donné par : a = 1,5 - 1 = 0,5, soit : **a = 50 %**.

Le nombre de passagers.km de la compagnie aérienne a été multiplié par 1,5 ou encore ce nombre a crû d'environ 50 %, **en moyenne annuelle**, sur l'ensemble de la période 1990-2002.

**Construction d'un abaque (faisceau de droites) pour matérialiser différents taux de croissance (annuels)**

a) On se positionne au niveau de l'origine d'un module (quelconque) et au niveau d'une année (quelconque). Ici, on a retenu la valeur 100 et l'année 1994, par commodité.

b) Compte tenu de l'allure générale de la série, on retient des taux de croissance appropriés. Ici, par exemple, le début de la série montre qu'on a environ un doublement (+ 100 %) du trafic chaque année. Ensuite, on observe un ralentissement entre 1995 et 1999, l'accroissement annuel tournant autour de 50 %. Sur la fin de la série, le contrecoup du 11 septembre 2000 nécessite de tracer des segments décroissants.

c) **Principe du tracé** : pour obtenir un segment représentatif d'une croissance de 100 % sur un an, on positionne l'origine du segment en 100, pour l'année 1994 (choix arbitraire), puis l'on repère le point d'ordonnée 200, pour l'année 1995. Il suffit ensuite de joindre les deux points et, pour une meilleure lisibilité, on prolonge à volonté le vecteur obtenu. Pour obtenir un vecteur décroissant, par ex. - 50 %, on part de la même origine et l'on repère le point d'ordonnée 50, toujours pour l'année 1995 (pour un vecteur décroissant de - 20 %, on repère le point d'ordonnée 80).

## CHAPITRE 2 : LISSAGE DES SÉRIES CHRONOLOGIQUES

Les traitements de ce chapitre sont spécifiques aux séries chronologiques.

Pour analyser correctement une série chronologique, on distingue dans celle-ci plusieurs éléments fondamentaux, appelés composantes de la série chronologique.

### 1. Les composantes d'une série chronologique

4 composantes peuvent être envisagées :

1) la tendance (générale) à long terme (ou trend) : pour un phénomène donné, il s'agit du mouvement qui se manifeste sur une longue période. Ce mouvement est lié à la croissance générale de l'économie.

Par exemple, la diminution de la population active agricole, la croissance de la production industrielle ou encore l'augmentation de la consommation d'électricité, etc.

2) le mouvement cyclique : il s'agit d'un mouvement périodique dont les maximums et les minimums relatifs correspondent à une succession de périodes de prospérité et de dépression. La période moyenne des cycles est variable selon le phénomène étudié, mais supérieure à l'année en général.

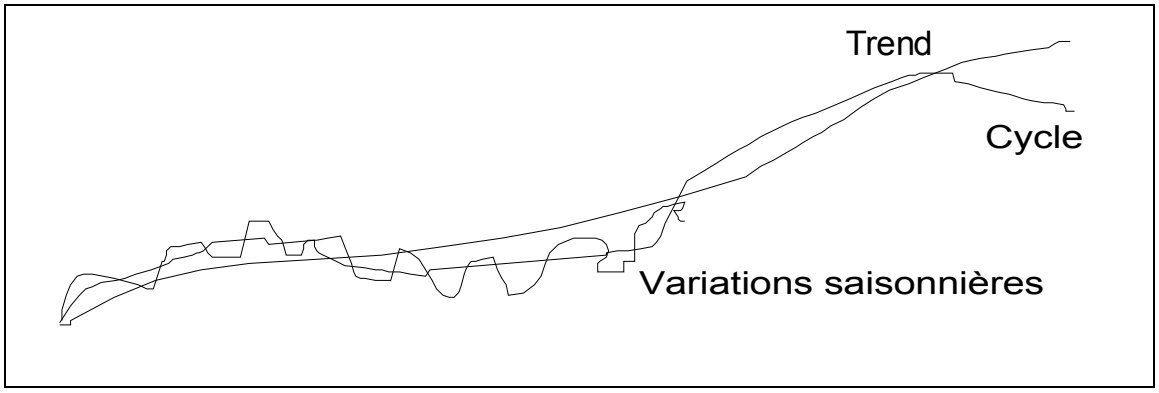
Le cycle s'articule autour de la tendance à long terme en fluctuations, qui sont liées à des variations conjoncturelles. Celles-ci accompagnent les phénomènes de croissance économique. Un cycle présente schématiquement quatre phases : une phase de prospérité (ou d'expansion), à laquelle succède une phase de crise (= retournement de tendance), puis survient une phase de dépression (ou de récession) qui cesse avec une phase de reprise (ou de relance). Exemples : les cycles longs de type Kondratiev (environ 50 ans) ; les cycles de Juglar (environ 8 ans) ; les cycles de Kitchin (environ 4 ans) ; les cycles des affaires, du bâtiment, etc. Il est à noter que les économistes ne sont pas tous d'accord sur l'existence même des cycles. Par ailleurs certains phénomènes économiques semblent n'être pas affectés par des mouvements périodiques dans leur évolution temporelle. En général, on travaille rarement sur des séries chronologiques très longues. De ce fait, en pratique, il est difficile de séparer l'influence des cycles de celle de la tendance à long terme. Pour ces diverses raisons, on regroupera dans la suite ces deux composantes sous la dénomination de mouvement extra-saisonnier (quelquefois dénommé mouvement conjoncturel).

3) les variations saisonnières (ou mouvement saisonnier) : il s'agit de fluctuations périodiques plus ou moins régulières qui viennent se superposer au mouvement extra-saisonnier De manière générale, on parle de mouvement saisonnier lorsque la périodicité du phénomène correspond à une année.

Exemples : la production industrielle est traditionnellement affectée par les périodes de vacances, notamment au mois d'août ; la vente des vêtements, selon la saison, est affectée par des fluctuations périodiques ; la vente des jouets connaît de fortes fluctuations au moment de Noël ; divers articles, comme les lunettes de soleil par exemple, connaissent aussi des fluctuations saisonnières de vente ; etc.

Remarque : de nombreux phénomènes, le plus souvent liés aux activités économiques, connaissent des fluctuations sur des périodes très courtes. Par exemple, les migrations alternantes domicile-travail journalières ; les déplacements hebdomadaires ; les déplacements effectués lors des vacances, etc. (les causes de ces différentes fluctuations sont associées au cycle des saisons, aux modes de vie, aux coutumes, aux dispositions réglementaires, etc.). Dans la suite, nous n'envisagerons pas ce type de phénomènes.

Schématiquement, on peut représenter les trois types de composantes de la manière suivante :




4) les variations accidentelles ou résiduelles : conjointement aux mouvements précédents, qui affectent l'évolution d'une grandeur dans le temps, on peut observer des variations, en général de faible amplitude (en temps et en volume), qu'on distingue en :

a) fluctuations aléatoires, dues à un grand nombre de petites causes qui, individuellement, ont une trop faible influence pour qu'on puisse les prendre en compte, d'où le nom de fluctuations résiduelles.

b) événements occasionnels, d'importance non négligeable, tels les grèves, les krachs financiers, les modifications structurelles de la législation fiscale, sociale ou économique, etc. Dans ce cas, nous verrons qu'il est indispensable de prendre en compte ces accidents de forte amplitude, avant d'effectuer une désaisonnalisation (cf. chapitre 3).

Selon les résultats recherchés (lissage, filtrage, désaisonnalisation), une ou plusieurs des composantes précédentes sont à prendre en considération. Par exemple, si l'on veut faire de la prévision à long terme (à 5 ou 10 ans), il faudra alors traiter la série chronologique envisagée, de façon à éliminer les composants cycliques et saisonnières, pour faire apparaître uniquement le trend.

## 2. Le lissage d'une série chronologique

De manière générale, le graphique des données brutes d'une série chronologique se présente sous la forme d'une courbe plus ou moins accidentée et il est souvent malaisé de repérer les différentes composantes évoquées plus haut (par exemple, des fluctuations boursières). C'est donc pour y voir plus clair, qu'on lisse la série chronologique. Ce faisant, on cherche à atténuer les pics et les creux observés sur le graphique des données brutes, afin de pouvoir mieux discerner l'évolution générale du phénomène sur la longue période. Le lissage permet en effet de mieux faire apparaître le trend, notamment les ruptures de tendance que l'analyse économique du phénomène cherchera à expliquer.

### 21. Méthodes empiriques de lissage d'une série chronologique

On peut tracer sur le graphique de la série chronologique brute une courbe "à main levée" en vue de mieux faire apparaître la tendance générale de la série, mais il s'agit-là d'un procédé très subjectif peu approprié à une analyse approfondie de l'évolution du phénomène observé.

#### 211. La méthode des points médians

On peut améliorer la procédure précédente en mettant en œuvre la méthode des points médians (high low mid points), qui reste néanmoins très subjective.

Le **principe** est le suivant : on trace approximativement les courbes enveloppes (haute et basse) de la série chronologique brute, en considérant d'une part les pics les plus significatifs et d'autre part les creux de la série. Ensuite, au niveau des maximums (les pics), on projette verticalement un segment sur la courbe enveloppe des minimums. On procède de même à l'inverse depuis un creux vers la courbe enveloppe des pics. Enfin, on relie les milieux des segments verticaux obtenus, ce qui donne une approximation grossière de la tendance générale à long terme du phénomène observé.

Cette méthode présente plusieurs inconvénients :

- (trop) forte influence des valeurs "aberrantes" (accidents conjoncturels divers) ;
- trend non défini pour les valeurs extrêmes de la série (perte d'information) ;
- tracé d'autant plus subjectif que les pics et les creux sont peu marqués.

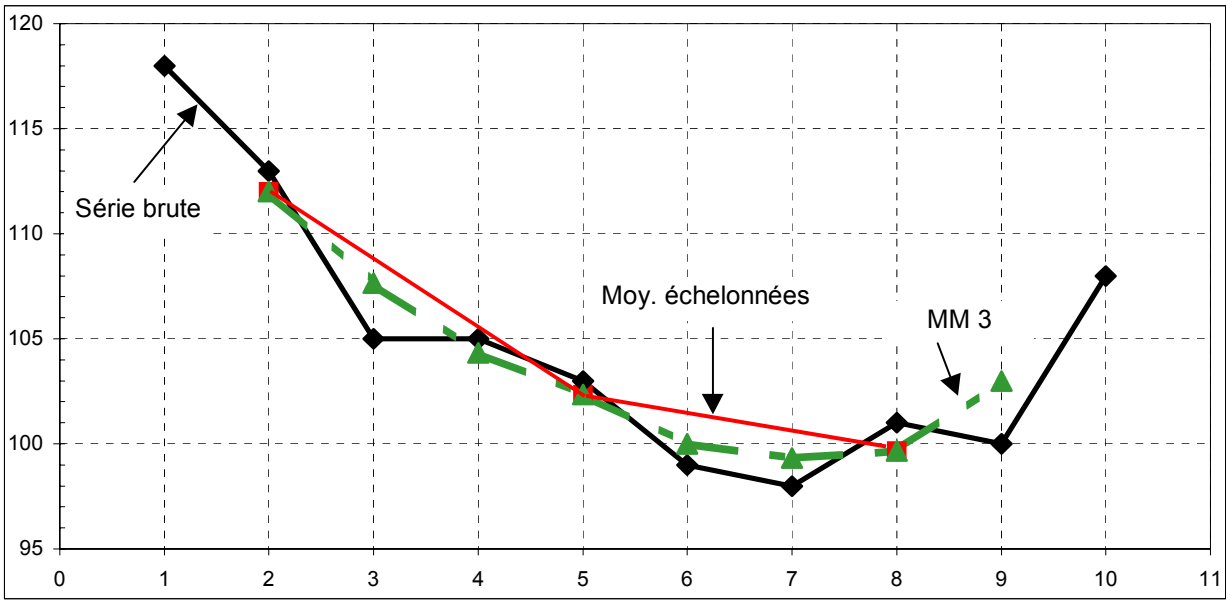
#### 212. Les moyennes échelonnées

À partir des données brutes de la série, on remplace un nombre impair de valeurs (en général 3) par leur moyenne arithmétique. Cette méthode permet donc de lisser la série brute, pour faire apparaître le trend plus clairement, en réduisant les pics et les creux, tout en conservant l'allure générale de la série. Si la méthode des moyennes échelonnées est moins subjective que les deux méthodes évoquées plus haut, elle n'en reste pas moins arbitraire (en effet, on peut retenir n'importe quel nombre de valeurs pour le calcul de la moyenne) et elle génère une grosse perte d'information (si l'on retient trois valeurs, la série obtenue possède trois fois moins de données que la série initiale).

Exemple : dans le tableau suivant, on dispose de 10 données, à partir desquelles on peut calculer 3 valeurs moyennes. Par exemple, la 1ère valeur est obtenue par : (118 + 113 + 105) / 3 = 112,0. La dernière ligne du tableau sera définie et utilisée plus loin :

| Périodes | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Série brute | 118 | 113 | 105 | 105 | 103 | 99 | 98 | 101 | 100 | 108 |
| Moyennes échelonnées | | 112,0 | | | 102,3 | | | 99,7 | | |
| Moyennes mobiles (MM3) | | 112,0 | 107,7 | 104,3 | 102,3 | 100,0 | 99,3 | 99,7 | 103,0 | |

Graphiquement, on obtient :




Remarque : la série brute comportant peu de données, la perte d'information est importante puisque qu'on ne dispose plus que de 3 valeurs pour les moyennes échelonnées. De plus, la dernière valeur de la série brute (108) n'est pas prise en compte par les calculs. Il en résulte qu'on ne repère par le retournement de tendance à la hausse, observé sur la série brute. Ce constat est très gênant si l'on utilise cette méthode pour prendre des décisions à court terme, relativement au phénomène considéré.

### 213. Les moyennes mobiles

Pour éviter les problèmes précédents, la méthode des moyennes mobiles est fondée sur un calcul où les moyennes vont se "chevaucher". On parle de moyennes mobiles, car on va décaler le calcul de période en période, en réutilisant les données du calcul précédent, sauf la première valeur.

Reprenons l'exemple précédent et posons également : n = 3, (nombre de valeurs sur lequel on réalise le calcul d'une moyenne). On parle alors de moyennes mobiles d'ordre 3 que l'on note : MM3.

Dans le tableau précédent, on constate que la 1ère valeur de la dernière ligne (MM3) est identique à celle de la 1ère moyenne échelonnée. Le calcul est en effet le même dans les deux cas. La 2ème valeur (107,7) est alors obtenue par le calcul : (113 + 105 + 105) / 3. Pour cela, nous avons utilisé les 2ème, 3ème et 4ème valeurs de la série brute. De même, la 3ème valeur (104,3) est obtenue par le calcul : (105 + 105 + 103) / 3. Pour cela, nous avons utilisé les 3ème, 4ème et 5ème valeurs de la série brute. On procède ainsi jusqu'à la dernière valeur disponible de la série brute. La dernière moyenne mobile (103,0) est obtenue par : (101 + 100 + 108) / 3.

La méthode des moyennes mobiles est plus efficace que celle des moyennes échelonnées puisqu'on dispose de 8 valeurs au lieu de 3. Sur le graphique, on peut notamment observer que les moyennes mobiles d'ordre 3 rendent compte du retournement de tendance, observé en fin de série.

Toutefois, le choix de l'ordre de la moyenne mobile reste arbitraire. De plus, on constate une perte d'information aux extrémités de la série. Par ailleurs, il existe une difficulté particulière lorsque la moyenne mobile calculée est d'ordre pair (c-à-d si n est pair). En ce cas en effet, la valeur obtenue ne correspond plus à une date donnée, mais tombe entre deux dates successives (cf. exemples suivants).

Remarque : malgré les défauts évoqués, la méthode des moyennes mobiles est très souvent utilisée car elle est facile à mettre en œuvre. Si l'ordre de la moyenne mobile est faible (3 ou 5, par exemple), le suivi du trend de la série sera plus efficace, mais le lissage sera moins important. À l'inverse, si l'ordre de la moyenne mobile est plus élevé, alors le lissage de la série brute sera plus important, mais le suivi du trend sera moins précis, notamment lorsque la série brute fait apparaître des retournements de tendance.

Pour illustrer la remarque précédente, l'exemple suivant présente le calcul de moyennes mobiles de trois ordres différents (MM3, MM4 et MM5) mis en œuvre sur une série chronologique portant sur les soldes d'exécution des lois de finances en France, de 1976 à 1995. Les données sont exprimées en milliards de francs :

| Années | Soldes | MM 3 | MM 4 | MM 5 |
|---|---|---|---|---|
| **1976** | -17,2 | | | |
| **1977** | -19,5 | -24,97 | | |
| **1978** | -38,2 | -32,20 | -29,28 | -27,52 |
| **1979** | -38,9 | -33,63 | -35,70 | -36,94 |
| **1980** | -23,8 | -42,33 | -48,00 | -51,40 |
| **1981** | -64,3 | -59,97 | -67,08 | -71,34 |
| **1982** | -91,8 | -98,00 | -96,16 | -95,06 |
| **1983** | -137,9 | -129,07 | -124,76 | -122,18 |
| **1984** | -157,5 | -151,60 | -143,59 | -138,78 |
| **1985** | -159,4 | -154,73 | -150,53 | -148,00 |
| **1986** | -147,3 | -148,20 | -143,31 | -140,38 |
| **1987** | -137,9 | -128,33 | -128,98 | -129,36 |
| **1988** | -99,8 | -113,37 | -115,65 | -117,02 |
| **1989** | -102,4 | -99,97 | -108,55 | -113,70 |
| **1990** | -97,7 | -110,27 | -122,94 | -130,54 |
| **1991** | -130,7 | -150,17 | -164,90 | -173,74 |
| **1992** | -222,1 | -222,87 | -217,24 | -213,86 |
| **1993** | -315,8 | -280,30 | -266,90 | -258,86 |
| **1994** | -303,0 | -313,83 | | |
| **1995** | -322,7 | | | |

Pour le calcul des moyennes mobiles d'ordre impair (MM3 et MM5), on procède comme dans le premier exemple. Ainsi, la 1ère moyenne mobile d'ordre 3 est égale à : - 24,97 = (-17,2 -19,5 -38,2) / 3, la 2ème moyenne mobile d'ordre 3 est égale à : - 32,20 = (-19,5 -38,2 -38,9) / 3, etc. De même, la 1ère moyenne mobile d'ordre 5 est égale à : - 27,52 = (-17,2 -19,5 -38,2 -38,9 -23,8) / 5, la 2ème moyenne mobile d'ordre 5 est égale à : - 36,94 = (-19,5 -38,2 -38,9 -23,8 -64,3) / 5, etc.

On constate qu'à l'ordre 3, on perd 2 données (la 1ère et la dernière). A l'ordre 5, on en perd 4 (les deux 1ères et les deux dernières), et ainsi de suite.

Le calcul des moyennes mobiles **d'ordre pair**

Si l'on procède comme on vient de le faire pour une moyenne mobile d'ordre pair (ici MM 4), la valeur obtenue ne correspond plus à une année donnée, mais elle tombe entre deux années. On ne peut en rester à une telle approximation, car l'analyse de l'évolution du phénomène étudié en serait par trop faussée.

Pour résoudre cette difficulté, **on calcule la moyenne mobile qui correspond à 2 sommes mobiles successives**.

Ce procédé a pour effet de faire coïncider à nouveau le résultat avec une année donnée. Ici, nous avons n = 4. Une somme mobile comprend donc 4 valeurs et 2 sommes mobiles successives en comprennent alors 8. Par conséquent, on obtient la valeur moyenne en divisant la double somme par 8.

Par exemple, considérons la 1ère valeur de la colonne MM4 du tableau (- 29,28). Pour l'obtenir, on procède de la manière suivante :

- 1ère somme mobile : on ajoute les 4 1ères valeurs de la série brute (colonne des soldes) :
  - 17,2 - 19,5 - 38,2 - 38,9 = - 113,8

Si l'on calculait une moyenne avec ces 4 valeurs, le résultat tomberait entre 1977 et 1978.

- 2ème somme mobile : on ajoute les 2ème, 3ème, 4ème et 5ème valeurs de la série brute :
  - 19,5 - 38,2 - 38,9 - 23,8 = - 120,4

Si l'on calculait une moyenne avec ces 4 valeurs, le résultat tomberait entre 1978 et 1979.

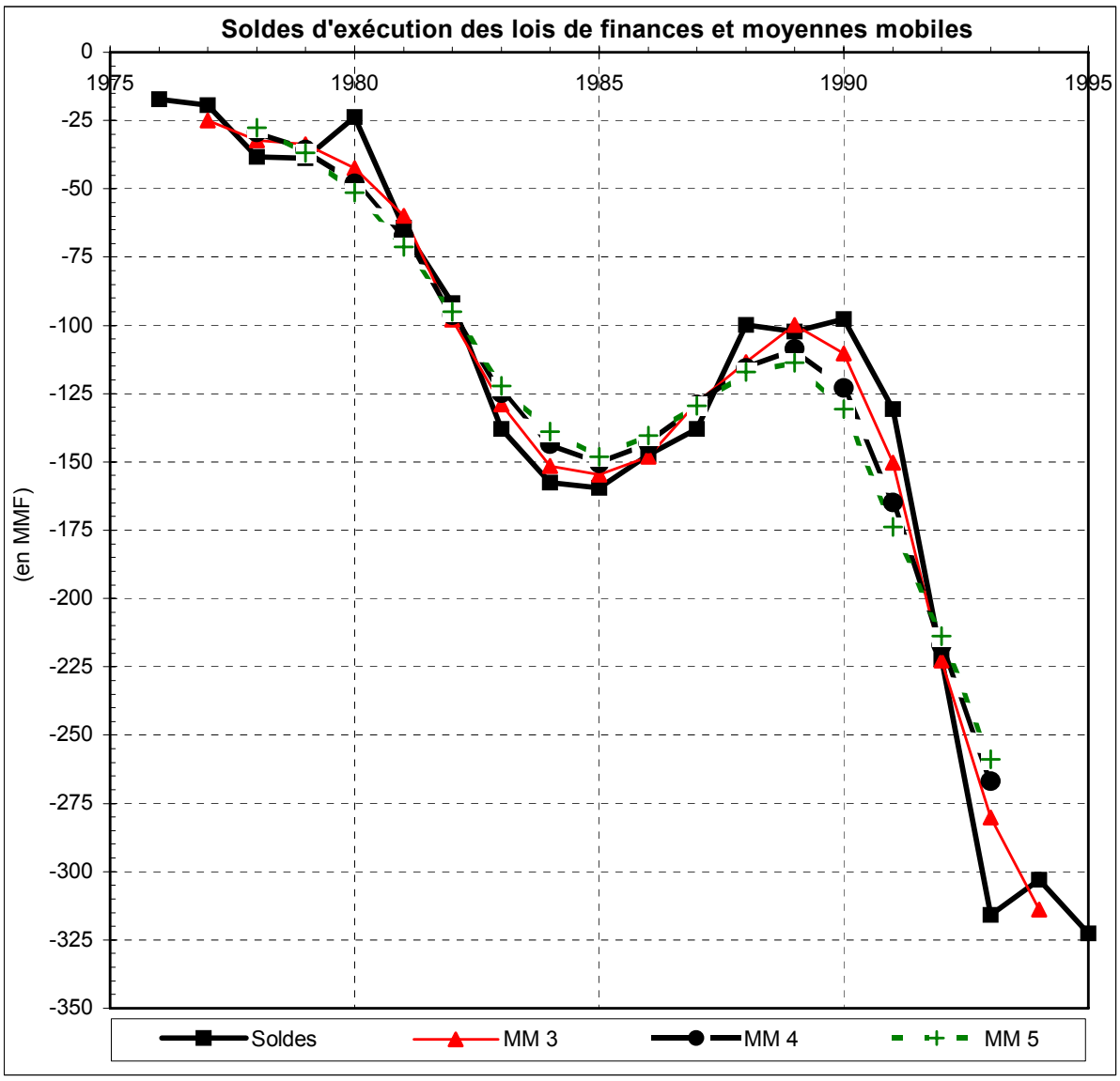
Au lieu de calculer une moyenne pour chacune de ces sommes, on les ajoute (= double somme mobile) et l'on divise le résultat obtenu par 8, puisqu'on a désormais 8 valeurs. La valeur moyenne obtenue coïncide alors bien à une date donnée, à savoir ici l'année 1978 :

(- 113,8 - 120,4) / 8 = - 29,28

On procède de même pour toutes les valeurs de la colonne MM4.

On constate que cette procédure génère une perte d'information identique à celle de l'ordre impair immédiatement supérieur. Pour MM4, comme pour MM5, on perd ici quatre valeurs (les deux 1ères et les deux dernières années). Il en irait par exemple de même pour MM6 et MM7, pour lesquelles on perdrait 6 années (3 au début de la série et 3 à la fin).

On obtient le graphique suivant :




Remarques :

a) plus l'ordre de la moyenne mobile est élevé, plus les pics et les creux de la série brute sont lissés, mais dans le même temps, le suivi du trend de la série devient moins précis.

b) cet exemple montre bien l'inconvénient que représente la perte d'information aux extrémités de la série. Ici, aucune des trois moyennes mobiles calculées ne permet de repérer l'infléchissement de tendance qui apparaît sur les deux dernières années de la série brute.

## 22. Méthode analytique de lissage d'une série chronologique

Contrairement aux méthodes empiriques, qui ne supposent aucune hypothèse sur l'allure du mouvement extra-saisonnier, la méthode analytique, que nous nous contentons de signaler ici sans la décrire, nécessite de poser des hypothèses sur la forme analytique des composantes extra-saisonnier et saisonnière de la série chronologique, pour être mise en œuvre.

Précisons simplement ici que l'on utilise une droite d'ajustement (des moindres carrées), de même type que celle qui tend à relater une liaison de type linéaire entre 2 caractères x et y. Toutefois, dans le cas présent, l'une des deux variables possède un statut particulier : c'est le temps t. Pour le reste (notamment le calcul des coefficients a et b de la droite d'ajustement), le traitement est le même que pour des caractères quantitatifs "classiques".

Avantages de la méthode :

- les fondements théoriques sur lesquels repose l'élaboration de l'ajustement linéaire sont solides, notamment en ce sens qu'on peut évaluer la variance (dispersion) des paramètres estimés, c'est-à-dire qu'on peut calculer la précision avec laquelle on estime les différentes composantes de la série chronologique.

- contrairement à la méthode des moyennes mobiles, on ne perd pas d'information aux extrémités de la série étudiée.

- on n'a pas besoin de poser d'hypothèses arbitraires (par ex. l'ordre de la moyenne mobile).

Inconvénients de la méthode :

- le mouvement extra-saisonnier de la série chronologique doit pouvoir être correctement représenté par une fonction analytique (de type linéaire, exponentiel, polynomial, …). Malheureusement, en économie, ces types d'ajustements (modèles) sont parfois trop simplistes.

- si une série connaît d'importants retournements de tendance, l'ajustement linéaire sera nécessairement mauvais. Dans un tel cas, on utilisera de préférence la méthode des moyennes mobiles. Si l'on tient à utiliser la méthode analytique, il faut pouvoir repérer aussi précisément que possible les retournements de tendance, de façon à utiliser la méthode indépendamment, sur chaque phase constatée (voir exemple au point 52 du chapitre 3).

## CHAPITRE 3 : DESAISONNALISATION DES SÉRIES CHRONOLOGIQUES

Les traitements de ce chapitre sont spécifiques aux séries chronologiques. Nous allons utiliser à nouveau ici les moyennes mobiles et la méthode analytique, évoquées dans le chapitre précédent, en vue de traiter les **fluctuations relatives à l'année**, qui affectent la plupart des séries chronologiques.

Nombre de séries chronologiques, généralement établies en mois ou en trimestres, font apparaître des variations périodiques qu'on appelle des **variations saisonnières**. Ces dernières tendent à se reproduire, sinon à l'identique, du moins de façon plus ou moins similaire d'une année sur l'autre, lorsqu'on étudie un phénomène donné sur plusieurs années.

Par exemple : les ventes de nombreux magasins vont augmenter systématiquement certains mois, selon le type d'activité ; d'autres activités sont assujetties aux variations climatiques saisonnières ; les demandes d'emploi en fin de mois recensées par l'ANPE sont toujours plus élevées en septembre, après les vacances d'été ; l'achat de fuel domestique et la consommation d'électricité tendent à augmenter en hiver pour des raisons de chauffage ; etc.

Ces modulations périodiques doivent être éliminées, si l'on veut rendre comparables entre eux tous les mois ou tous les trimestres de la série chronologique étudiée.

Conséquences :

- les évolutions du mouvement extra-saisonnier (trend + cycle) sont plus clairement perceptibles. C'est l'objectif que se propose la désaisonnalisation d'une série, à savoir étudier et annihiler l'influence des facteurs saisonniers, en vue d'obtenir une série dite corrigée des variations saisonnières (série CVS). Celle-ci fait apparaître le trend de façon plus satisfaisante.

- on peut alors établir des prévisions à court terme et à moyen terme plus fiables, en ce qui concerne l'évolution probable du phénomène étudié dans le futur.

Remarque : la question du long terme renvoie aux changements structurels survenus dans les comportements des individus et des organisations.

Rappel : nous avons vu plus haut qu'on peut envisager quatre composantes dans une série chronologique brute :

- la tendance (générale) à long terme (ou trend) ;
- le mouvement cyclique ;

Nous avons regroupé ces deux composantes en un même mouvement extra-saisonnier

- les variations saisonnières (ou mouvement saisonnier) ;
- les variations accidentelles (ou résiduelles ou aléatoires).

Pour mener à bien une désaisonnalisation, il convient de poser un certain nombre **d'hypothèses, relatives à chacune des composantes de la série chronologique**. Il s'agit de construire un **modèle** qui permette de formaliser (c'est-à-dire de décomposer) systématiquement le mouvement général d'une série chronologique selon les composantes décrites dans le chapitre 2. Bien entendu, il s'agit d'une schématisation de la réalité (cette dernière étant plus complexe), qui permet un traitement "simplifié" de la désaisonnalisation de la série.

Nous désignerons chacun des mouvements précédents par les symboles suivants :

$f_t$ = mouvement extra-saisonnier (trend + cycle) ;

$s_t$ = mouvement saisonnier (annuel) ;

$z_t$ = mouvement résiduel (accidentel).

## 1. Les hypothèses qui sous-tendent une désaisonnalisation

### 11. Hypothèses générales

a) On suppose que le temps se répète à l'identique d'une année sur l'autre, de telle façon que la valeur $y_t$ de la série soit directement comparable à la valeur $y_{t+1}$ , sur l'ensemble de la période d'étude.

En pratique cependant, cette hypothèse n'est pas toujours vérifiée. Par exemple, le nombre de jours du mois de février n'est pas toujours le même (29 jours durant les années bissextiles, ce qui représente une augmentation de 3,6 % (1 / 28) de la durée mensuelle !). Par ailleurs, il existe une fluctuation des fêtes légales : selon que la fête de Pâques ait lieu en mars ou en avril, cela occasionne des différences significatives pour certaines activités économiques).

b) On suppose que les structures de base du phénomène étudié sont stables dans le temps.
En pratique, cela ne sera pas vrai notamment en cas de grèves, de catastrophes naturelles, de conflits divers, etc. cf. plus loin le rôle de $z_t$.

c) On suppose que la grandeur étudiée conserve la même définition sur l'ensemble de la période d'étude.

En pratique cependant, la définition relative à une grandeur peut varier si les méthodes de mesure sont améliorées (par exemple par un champ d'enquête élargi). Dans le cas d'un indice des prix (cf. partie 3), son contenu à long terme (séries longues) tend à changer selon le nombre et le type des produits qu'il inclut.

### 12. Hypothèses relatives aux différentes composantes de la série

a) Notons $f_t$ le mouvement extra-saisonnier On pose a priori qu'il est une fonction quelconque du temps. Toutefois, il faut savoir que l'utilisation de la méthode analytique limite les types de fonctions qu'on peut mettre en œuvre, de façon à valider le modèle.

b) Notons $s_t$ le mouvement saisonnier. Il est supposé de période rigoureusement égale à un an. Cela implique :

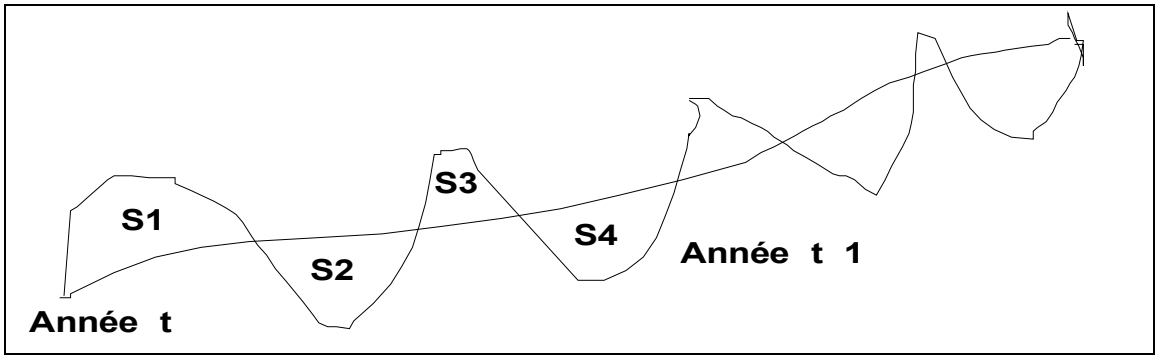
- une répétition à l'identique d'une année sur l'autre, telle que :

$s_t = s_{t+12} = s_{t+24} = \ldots$ (série mensuelle) ; $s_t = s_{t+4} = s_{t+8} = \ldots$ (série trimestrielle).

- pour distinguer $f_t$ et $s_t$ dans les schémas de composition de ces deux composantes (voir plus loin la notion de schéma de composition), on pose l'hypothèse dite de conservation des aires, telle qu'on ait :

$$\sum_{t=1}^{12} s_t = 0 \text{ (série mensuelle)} \quad \text{ou :} \quad \sum_{t=1}^{4} s_t = 0 \text{ (série trimestrielle)}$$

Cela signifie que, **sur une même année, la somme des variations saisonnières est égale à zéro**. La mise en œuvre de cette hypothèse permet le calcul les coefficients (ou des rapports) saisonniers de la série.




Le principe de conservation des aires est tel que : S1 + S3 = S2 + S4.

Remarque : cette hypothèse reflète rarement la réalité, mais est fondamentale dans la logique du calcul des coefficients ou des rapports saisonniers.

c) Notons $z_t$ le mouvement résiduel (dit encore accidentel ou aléatoire). On suppose par hypothèse que ce mouvement est négligeable, c'est-à-dire que l'on considère que la faible amplitude des variations accidentelles n'est pas susceptible d'influencer le trend, sur l'ensemble de la période d'étude. Cela revient à considérer, comme c'est le cas la plupart du temps, que les fluctuations résiduelles de la série sont le fait d'un grand nombre de petites causes qui ne génèrent que de faibles amplitudes de variation et :

- sur une même année : $\sum_{t=1}^{12} z_t = 0$ (série mensuelle) ou : $\sum_{t=1}^{4} z_t = 0$ (série trimestrielle)

- sur n années (période d'étude) : $\sum_{t=1}^{n} z_t = 0$

**Quelles solutions doit-on mettre en œuvre dans les cas où l'on ne peut pas poser l'hypothèse précédente** ? C'est-à-dire lorsque les accidents conjoncturels sont trop importants pour être négligés (grèves majeures, cataclysmes naturels, conflits divers, etc.).

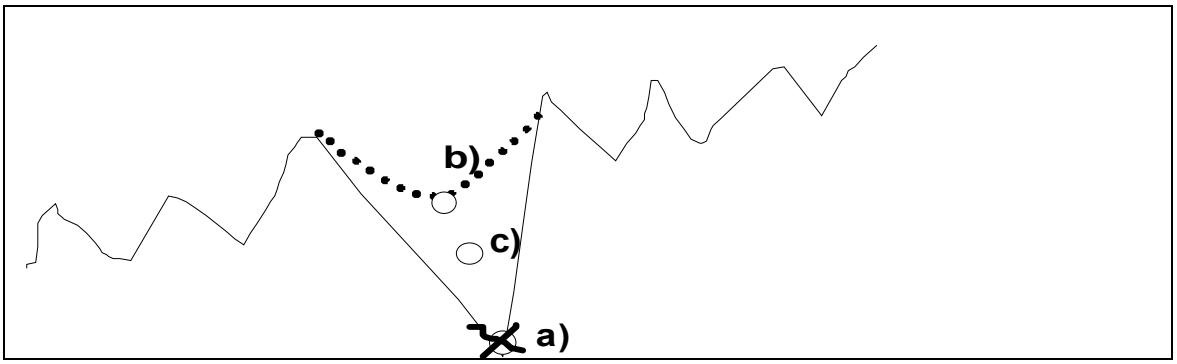
Dans de tels cas, trois types de solutions sont possibles :

a) on écarte la valeur "aberrante" en question de la détermination des coefficients saisonniers (mensuels ou trimestriels), de façon à ne pas fausser le calcul de ces coefficients, qui sont des coefficients moyens ou médians. Mais évidemment, on doit réintégrer l'accident correspondant dans la série corrigée des variations saisonnières, à la fin des calculs. On ne peut en effet gommer le fait que l'accident ait bien eu lieu à une date donnée.

b) à la place de la vraie valeur, on peut aussi retenir une valeur moyenne ou médiane, calculée sur les mois ou trimestres correspondants, sur l'ensemble de la période d'étude ou une partie seulement de celle-ci.

c) si l'on dispose d'informations complémentaires sur le secteur d'activité étudié, on peut aussi choisir une valeur "plausible", dont on peut penser qu'elle aurait pu être obtenue en l'absence de l'accident conjoncturel constaté dans la réalité.

Illustration schématique des solutions précédentes :




Solution a) : on supprime la valeur "aberrante", de façon à ce qu'elle n'intervienne pas dans les calculs relatifs aux coefficients ou aux rapports saisonniers.

Solution b) : valeur moyenne ou médiane obtenue à partir des valeurs des mois ou des trimestres correspondants, sur l'ensemble de la période d'étude ou une partie seulement de celle-ci.

Solution c) : par exemple, une connaissance fine du secteur et de la conjoncture permet d'avancer l'hypothèse d'une valeur plus faible que dans le cas b).

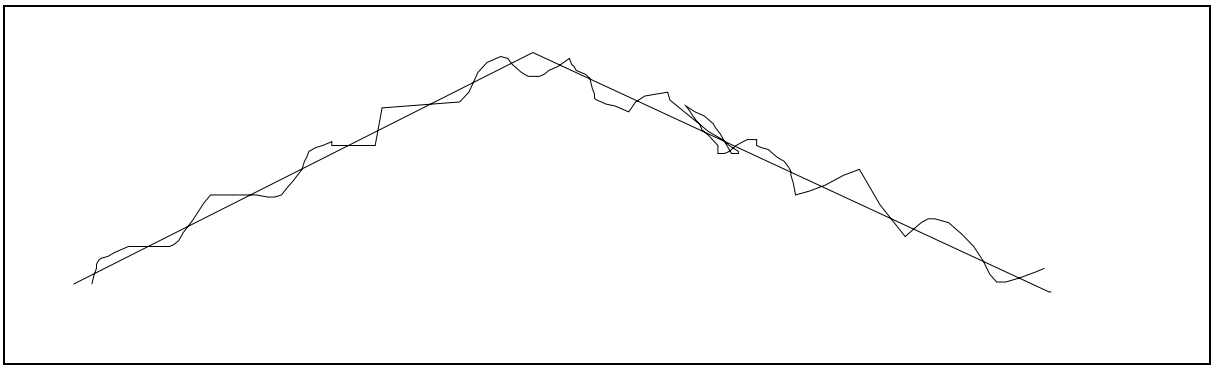
## 13. Hypothèses relatives à la composition des éléments constitutifs de la série chronologique

Ces hypothèses concernent la façon dont sont liés les mouvements extra-saisonnier d'une part et saisonnier d'autre part. Dans la suite en effet, on considérera que le mouvement résiduel est négligeable. On peut envisager trois schémas (modèles) de composition des éléments d'une série chronologique :

### a) le schéma (de composition) additif

On postule que le phénomène étudié en fonction du temps (évolution temporelle du phénomène) se décompose en éléments indépendants les uns des autres. Cela signifie que, quel que soit le niveau atteint (en volume) par le mouvement extra-saisonnier, les variations saisonnières sont indépendantes de ce niveau. On écrit alors :

$$y_t = f_t + s_t + z_t$$

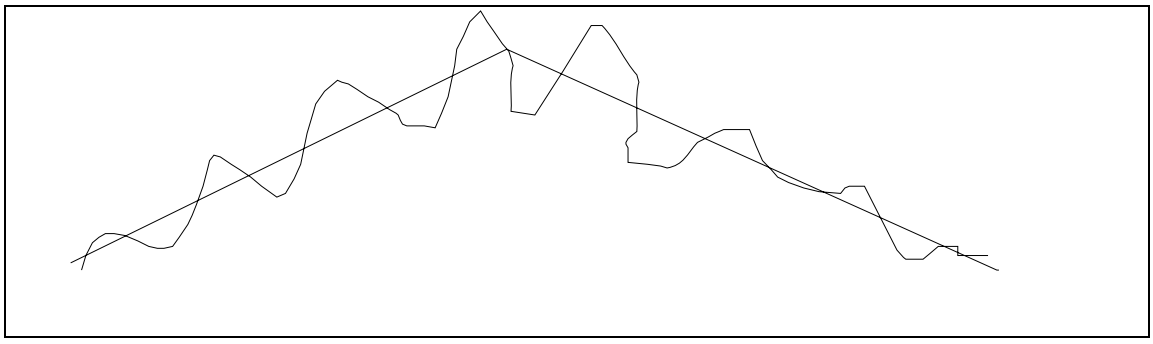


Quels que soient t et le niveau de $f_t$, on constate toujours à peu près une même amplitude des variations saisonnières.

### b) le schéma (de composition) multiplicatif

Dans ce cas, on postule que le phénomène étudié en fonction du temps se décompose en éléments dépendants les uns des autres. Selon le niveau du mouvement extra-saisonnier, les variations saisonnières voient leur amplitude se modifier. On a :

$$y_t = f_t (1 + s_t) + z_t$$



Donc, $s_t$ est fonction du niveau atteint par $f_t$, tel que : $y_t = f_t + f_t . s_t + z_t$

Remarque : parfois, on trouve aussi la forme suivante : $y_t = f_t \cdot s_t \cdot z_t$

c) **le schéma (de composition) mixte**

Dans ce cas, on postule que le mouvement saisonnier se décompose en 2 éléments :

- $s_{t\,(1)}$ qui dépend du niveau de $f_t$ ;
- $s_{t\,(1)}$ qui ne dépend pas du niveau de $f_t$.

On écrit alors : $y_t = f_t\,(1 + s_{t\,(1)}) + s_{t\,(2)} + z_t$

Remarque : le traitement d'une série chronologique, en vue d'une désaisonnalisation, dans le cadre d'un schéma mixte, s'avère plus complexe que dans les deux premiers cas envisagés. Nous ne traiterons pas ce cas ici. On se ramènera à un schéma de composition additif ou à un schéma de composition multiplicatif, en retenant en pratique celui qui est apparemment le "moins mauvais" cf. plus loin la façon de choisir entre un schéma additif et un schéma multiplicatif.

## 2. Les méthodes de traitement des séries en vue d'une désaisonnalisation

Quel que soit le schéma de composition retenu (additif ou multiplicatif), il s'agit d'estimer, pour chaque date d'observation, les valeurs des composantes extra-saisonnier $f_t$ et saisonnière $s_t$ de la série. Pour cela, il existe 2 méthodes principales que nous avons déjà évoquées au chapitre 2 : la méthode des moyennes mobiles et la méthode analytique.

### 21. Estimation de la composante extra-saisonnière par les moyennes mobiles

Nous avons vu le principe d'utilisation des moyennes mobiles au chapitre 2, pour lisser une série chronologique. Dans le cas d'une désaisonnalisation, le problème n'est pas fondamentalement différent de celui d'un lissage.

Il s'agit en effet de **faire apparaître le trend, en éliminant les variations saisonnières**. Mais ici, l'ordre des moyennes mobiles utilisées n'est plus arbitraire comme pour un lissage. Comme on travaille sur une année et qu'on raisonne en mois ou en trimestres (en général), il n'existe qu'une seule possibilité. Si la série est donnée en mois, alors on utilise obligatoirement une moyenne mobile d'ordre 12 (MM12). Si la série est donnée en trimestres, alors on utilise obligatoirement une moyenne mobile d'ordre 4 (MM4).

Dans les 2 cas, il s'agit de moyennes mobiles d'ordre pair. Nous avons vu au chapitre 2 que, dans la mesure où l'on ne tombe plus sur une date d'observation donnée mais entre 2 dates, on calcule les moyennes mobiles sur 2 sommes mobiles consécutives, afin de retomber sur une date donnée. Dans le cas d'une série mensuelle, on perd l'information relative aux 6 premiers mois de la première année et aux 6 derniers mois de la dernière année de l'intervalle d'étude. Au total, on perd un an de données. Dans le cas de données trimestrielles, on perd l'information relative aux 2 premiers trimestres de la première année et aux 2 derniers trimestres de la dernière année de l'intervalle d'étude. Au total, on perd également un an de données. Pour mémoire, il existe des méthodes économétriques appropriées (moyenne d'Henderson, par exemple) destinées à pallier cet inconvénient.

### 22. Estimation de la composante extra-saisonnière par la méthode analytique

Cette méthode, que nous évoquons ici pour mémoire, présente deux avantages importants sur la précédente :

- on peut estimer la valeur de la composante extra-saisonnière en utilisant toute l'information disponible (depuis le premier mois ou le premier trimestre de la période d'étude jusqu'au dernier), à partir de la formulation suivante : $\hat{f}_t = at + b$ , où a et b sont des coefficients à calculer ;

- on peut réaliser des prévisions à court terme, par extrapolation de la tendance générale calculée sur une période de temps donnée.

La méthode analytique présente toutefois deux inconvénients principaux :

- la composante extra-saisonnière doit pouvoir être correctement représentée par une fonction analytique, de type linéaire, exponentiel, polynomial ou logistique (ce qui, en général, ne déforme pas trop la réalité) ;

- si une série chronologique brute fait apparaître de nombreux retournements de tendance assez marqués, la méthode analytique peut devenir non pertinente.

**Quelques compléments sur la méthode analytique**

Généralement, pour simplifier les calculs, on met en œuvre la méthode analytique en utilisant une relation linéaire (modèle linéaire), c-à-d $\hat{f}_t = at + b$. Il s'agit de l'expression du mouvement extra-saisonnier estimé, dans laquelle il s'agit de déterminer les valeurs de a et de b (cf. croisement de 2 caractères quantitatifs).

L'allure de l'évolution du phénomène étudié laisse parfois à penser qu'un ajustement de type exponentiel conviendrait mieux (de nombreux phénomènes économiques connaissent des croissances exponentielles dans le temps). Dans ce cas, la forme du modèle (modèle exponentiel) est la suivante :

$\hat{f}_t = f_0 (1 + r)^t$ (1) avec : $f_0$ = valeur du mouvement extra-saisonnier en t = 0 (valeur initiale) ;
r = taux de croissance constant (caractéristique d'un accroissement exponentiel).

Remarque : ce deuxième cas se ramène en fait au premier si l'on passe l'expression aux logarithmes décimaux (cf. utilisation du papier semi-logarithmique : une exponentielle devient une droite).

En passant aux logarithmes, l'expression (1) s'écrit : $\log \hat{f}_t = \log f_0 + t \log(1 + r)$

Si l'on pose : $\hat{F}_t = \log \hat{f}_t$ ; $B = \log f_0$ ; $A = \log(1 + r)$, alors il vient : $\hat{F}_t = At + B$ (2)

L'expression (1) est représentée par une droite si l'on trace le mouvement extra-saisonnier sur une feuille de papier semi-logarithmique, selon (2). On peut ajuster la droite (2) aux points observés du nuage $(t ; \hat{F}_t) = (t ; \log \hat{f}_t)$, par la méthode des moindres carrés.

En pratique, cela revient à remplacer, dans le calcul des paramètres a et b de la droite d'ajustement, les valeurs $y_t$ par $\log y_t$ (on détermine le mouvement extra-saisonnier à partir de la série chronologique brute ; voir le point 4). On a :

log ft estimé = a*.t + b* D'où : ft estimé = trend estimé = 10^(a*.t + b*) = 10^a*.t . 10^b*

Si l'on hésite entre l'ajustement linéaire et l'ajustement exponentiel, on calcule le coefficient de corrélation linéaire dans chaque cas. Il est important de rappeler que, en soi, la valeur du coefficient de corrélation linéaire n'a pas de signification, puisque le phénomène observé est fonctionnellement relié au temps.

Mais si l'on trouve : $\rho_{t,y_t} \geq \rho_{t,\log y_t}$ , on réalise un ajustement linéaire (modèle linéaire).

Si l'on a : $\rho_{t,y_t} \leq \rho_{t,\log y_t}$ , on réalise un ajustement exponentiel (modèle exponentiel).

Lorsque les courbures de la série brute ne sont pas trop marquées, on a intérêt à choisir la méthode analytique plutôt que celle des moyennes mobiles pour estimer la composante extra-saisonnière de la série. En effet, les propriétés de la droite d'ajustement permettent la mise en œuvre d'une prévision plus robuste à court terme (on peut en effet calculer un intervalle de confiance sur la valeur estimée des paramètres a et b de la droite d'ajustement, ce qui n'est pas possible avec les moyennes mobiles). Par ailleurs, on gagne un an d'informations supplémentaires en termes de données.

Les prévisions à court terme qui découlent de l'utilisation de la méthode analytique sont très utiles aux entreprises (meilleure planification et meilleure gestion de leurs activités) et à l'État (pour les mêmes raisons, ainsi que pour une meilleure préparation du budget de l'État).

Pour réaliser une prévision à court terme, on prolonge la tendance de la droite d'ajustement $\hat{y}_t = at + b$ , en considérant évidemment que cette tendance de l'évolution du phénomène va se maintenir dans le futur, de la façon suivante :

| t | 1 | 2 | … | n | | n+1 | n+2 |
|---|---|---|---|---|---|---|---|
| $y_t$ | $X_1$ | $X_2$ | … | $X_n$ | | $Y_1$ | $Y_2$ |

### 3. Les étapes successives d'une désaisonnalisation

**Étape 1 : repérage et traitement des valeurs "aberrantes"**

On vérifie que les variations accidentelles $z_t$ sont négligeables. En fait, il suffit de parcourir les données (ou de lire le graphique des données brutes) pour repérer des valeurs éventuellement "aberrantes". Si l'on repère des valeurs exceptionnellement basses ou élevées par rapport au reste de la série, on doit rechercher des informations complémentaires (contexte économique de la grandeur étudiée). Puis on applique l'une des solutions évoquées au point 12.

**Étape 2 : choix d'un schéma de composition**

Il s'agit de choisir entre le schéma de composition additif : $y_t = f_t + s_t$ et le schéma de composition multiplicatif : $y_t = f_t (1 + s_t)$.

#### A. Solutions graphiques

a) Représentations graphiques de la série chronologique brute

Pour l'ensemble de la période d'étude, on établit un graphique avec le temps en abscisse et les valeurs (absolues ou relatives) prises par la grandeur étudiée en ordonnée.

- ordonnée arithmétique : si les amplitudes des fluctuations saisonnières restent à peu près les mêmes, quel que soit le niveau atteint par le trend, alors on retient un schéma additif. Si au contraire les amplitudes tendent à se modifier en même temps que le niveau du trend, alors on retient un schéma multiplicatif.

- ordonnée logarithmique : si les amplitudes des fluctuations saisonnières restent à peu près les mêmes, quel que soit le niveau atteint par le trend, alors on retient un schéma multiplicatif.

b) Représentations graphiques en données annuelles superposées

On découpe la série en périodes annuelles (12 mois ou 4 trimestres), que l'on reporte sur le même graphique. Comme précédemment, on place le temps en abscisse et les valeurs (absolues ou relatives) prises par la grandeur étudiée en ordonnée.

- ordonnée arithmétique : si les courbes superposées font apparaître un mouvement saisonnier d'amplitude absolue à peu près constante, alors même que le niveau du trend augmente ou diminue, alors on retient un schéma additif. Si l'on observe un mouvement saisonnier dont l'amplitude croît ou décroît avec le niveau du trend, alors on retient un schéma multiplicatif.

- ordonnée logarithmique : si l'amplitude relative du mouvement saisonnier reste à peu près constante, alors même que le niveau du trend augmente ou diminue, alors on retient un schéma multiplicatif.

c) Graphiques polaires

Il s'agit d'une variante des représentations graphiques en données annuelles superposées. Ici encore, on peut envisager des ordonnées arithmétiques ou logarithmiques comme dans les cas précédents. L'interprétation des graphiques polaires est la même que précédemment.

Si l'on a affaire à une série mensuelle, on utilise un repère d'origine 0, à partir de laquelle on trace douze axes, chacun correspondant à un mois donné. Une valeur (absolue ou relative) prise par la grandeur étudiée est alors placée sur l'axe du mois correspondant. On relie ensuite chronologiquement les différents points successifs par des segments de droite. Comme la méthode précédente, celle-ci permet de repérer facilement les éventuels mouvements saisonniers de la série étudiée.

d) Utilisation des moyennes mobiles

Cette méthode oblige à un calcul supplémentaire, celui des moyennes mobiles de la série étudiée (ce qui ne constitue une perte de temps que si l'on souhaite utiliser ensuite la méthode analytique pour estimer le trend de la série). Sur un graphique, on porte en abscisse la valeur des données de la série brute et en ordonnée les valeurs des moyennes mobiles correspondantes (MM4 si la série est trimestrielle ou MM12 si la série est mensuelle). On relie ensuite les points qui correspondent respectivement à tous les $1^{ers}$ trimestres, tous les $2^{èmes}$ trimestres, etc. (ou à tous les mois de janvier, tous les mois de février, etc.) par des droites (approximativement). Si les 4 droites (trimestres) ou les 12 droites (mois) obtenues sont à peu près parallèles entre elles, on retient le schéma additif. Si les 4 droites (trimestres) ou les 12 droites (mois) convergent à peu près vers l'origine du repère, on retient le schéma multiplicatif.

### B. Solution algébrique

Cette méthode ne nécessite aucun graphique. Sur les 4 trimestres (ou les 12 mois) de chaque année de la période d'étude, on calcule un écart-type. En fonction du niveau atteint par les valeurs de la série brute, on va comparer les évolutions des écarts-types annuels. **Si ces écarts-types annuels restent à peu près constants, quel que soit le niveau atteint par les valeurs de la série, alors on retient un schéma additif. Si au contraire les écarts-types annuels tendent à augmenter ou à diminuer en même temps que le niveau atteint par les valeurs de la série, alors on retient un schéma multiplicatif.**

**Étape 3 : estimation du mouvement extra-saisonnier** (trend)

Cette estimation est réalisée soit en utilisant la méthode des moyennes mobiles (MM4 ou MM12, selon le type de série étudiée) cf. point 21., soit la méthode analytique.

Méthode des moyennes mobiles : $\hat{f}_t = MM4_t$ ou : $\hat{f}_t = MM12_t$

Méthode analytique (pour mémoire) :

$$\hat{f}_t = at + b \text{ (modèle linéaire)} \quad \text{ou :} \quad \hat{f}_t = f_0 (1 + r)^t \text{ (modèle exponentiel)}$$

Remarque : on utilise le modèle linéaire si le coefficient de corrélation $\rho_{t, y_t}$ est supérieur au coefficient de corrélation $\rho_{t, \log y_t}$, et inversement.

**Étape 4 : calcul des variations saisonnières**

À partir de cette étape et pour les suivantes, les calculs sont à distinguer entre le schéma additif et le schéma multiplicatif (rappel : les variations accidentelles sont négligées).

**Schéma additif** : $y_t = f_t + s_t$ ou : $y_{ij} = f_{ij} + s_{ij}$

avec : i = indice des années de la période d'étude ; j = indice du mois (ou du trimestre) d'une année donnée (présentation sous forme de tableau à double entrée).

A l'étape 3, le trend $f_t$ a été estimé par $MM4_t$ (moyennes mobiles).

Il vient donc : $s_t = y_t - MM4_t$. Les valeurs $s_t$ correspondent aux **variations saisonnières** réelles de la série étudiée.

**Schéma multiplicatif** : $y_t = f_t (1 + s_t)$ ou : $y_{ij} = f_{ij} (1 + s_{ij})$

Ici, compte tenu de la forme de l'expression, on a : $(1 + s_t) = \frac{y_t}{MM4_t}$.

**Étape 5 : calcul des coefficients (ou des rapports) saisonniers**

Pour une raison pratique (usage de tableaux à double entrée), on remplace l'indice t associé à chacune des composantes d'une série chronologique par un indice double ij où i représente l'indice des années (i = 1 à n) et j l'indice des trimestres (j = 1 à 4) ou des mois (j = 1 à 12).

**Schéma additif** :

Pour chaque trimestre (ou chaque mois) j, on calcule une première estimation du **coefficient saisonnier** en considérant :

- soit la médiane $s'_j$ des variations saisonnières relatives à ce trimestre (ou à ce mois) sur l'ensemble des valeurs disponibles de la série ;

- soit la moyenne $\bar{s}_j$ ou, le plus souvent, la moyenne tronquée $\bar{s}_j\,tr.$ obtenue en ne prenant pas en compte la plus haute et la plus basse des valeurs de la série (cela permet de limiter l'effet d'éventuelles valeurs "aberrantes").

**Schéma multiplicatif** :

Dans ce cas, on procède la même façon, mais on parle de **rapport saisonnier**, puisque les variations saisonnières apparaissent sous la forme de rapports (cf. étape 4).

Ici, les premières estimations des rapports saisonniers sont donc égales, soit à $(1 + s'j)$, soit à $(1+\bar{s}_j)$ ou à $(1+\bar{s}_j)\,tr.$ .

Principe :

| | **An t** | **An t+1** | **An t+2** | **An t+3** | **…** | **$s'_j$ ou $\bar{s}_j$ (+)**<br>**$(1+s'_j)$ ou $(1+\bar{s}_j)$ (x)** |
|---|---|---|---|---|---|---|
| **Trim. 1** | x | x | x | x | | X |
| **Trim. 2** | | | | | | X |
| **Trim. 3** | | | | | | X |
| **Trim. 4** | | | | | | X |

Dans le cas du schéma additif, on doit vérifier que la somme des 4 valeurs de la dernière colonne est égale à zéro. Dans le cas contraire, il faut corriger les coefficients saisonniers.

Dans le cas du schéma multiplicatif, on doit vérifier que la somme des 4 valeurs de la dernière colonne est égale à 4 (dans le cas de trimestres et à 12 dans le cas de mois). Dans le cas contraire, il faut corriger les rapports saisonniers.

**Étape 6 : calcul (éventuel) des coefficients (ou des rapports) saisonniers corrigés**

Au point 12., nous avons posé une hypothèse de conservation des aires, telle que, sur un an, les variations saisonnières s'annulent.

**Schéma additif** : on vérifie que :

$$\sum_{j=1}^{4\,(ou\,12)} s'_j = 0 \text{ (médiane)} \quad ou : \quad \sum_{j=1}^{4\,(ou\,12)} \bar{s}_j = 0 \text{ (moyenne)} \quad ou : \quad \sum_{j=1}^{4\,(ou\,12)} \bar{s}_j \text{ tr.} = 0 \text{ (moyenne tronquée)}$$

**Schéma multiplicatif** : on vérifie que :

$$\sum_{j=1}^{4\,(ou\,12)} (1+s'_j) = 4\,(ou\,12) \quad ou : \quad \sum_{j=1}^{4\,(ou\,12)} (1+\bar{s}_j) = 4\,(ou\,12) \quad ou : \quad \sum_{j=1}^{4\,(ou\,12)} (1+\bar{s}_j \text{ tr.}) = 4\,(ou\,12)$$

Si les sommes obtenues ne sont pas égales (ou tout au moins proches) aux valeurs requises pour la conservation des aires sur une année, on doit alors réaliser une correction sur les coefficients (ou les rapports) saisonniers de la façon suivante :

**Schéma additif** :

$$s_j = s'_j - \frac{1}{4\,(ou\,12)} \cdot \sum_{j=1}^{4\,(ou\,12)} s'_j \text{ (méd.)} \qquad \text{avec : } s_j = \text{coefficient saisonnier corrigé}$$

$$ou : \; s_j = \bar{s}_j - \frac{1}{4\,(ou\,12)} \cdot \sum_{j=1}^{4\,(ou\,12)} \bar{s}_j \text{ (moy.)} \quad ou : \; s_j = \bar{s}_j \text{ tr.} - \frac{1}{4\,(ou\,12)} \cdot \sum_{j=1}^{4\,(ou\,12)} \bar{s}_j \text{ (moy. tr.)}$$

**Schéma multiplicatif** :

$$(1+s_j) = \frac{1+s'_j}{\frac{1}{4\,(ou\,12)} \cdot \sum_{j=1}^{4\,(ou\,12)} (1+s'_j)} \text{ (méd.)} \qquad \text{avec : } (1+s_j) = \text{rapport saisonnier corrigé}$$

$$ou : \; (1+s_j) = \frac{1+\bar{s}_j}{\frac{1}{4\,(ou\,12)} \cdot \sum_{j=1}^{4\,(ou\,12)} (1+\bar{s}_j)} \text{ (moy.)} \quad ou : \; (1+s_j) = \frac{1+\bar{s}_j \text{ tr.}}{\frac{1}{4\,(ou\,12)} \cdot \sum_{j=1}^{4\,(ou\,12)} (1+\bar{s}_j \text{ tr.})} \text{ (moy. tr.)}$$

Remarque : ici, on choisit d'équirépartir l'écart observé entre les 4 (ou 12) coefficients (ou rapports) saisonniers. Ce n'est pas la seule méthode possible. Si l'on veut réaliser un calcul très précis, par exemple on peut calculer un écart-type sur les variations saisonnières relatives à chaque mois (ou à chaque trimestre). On établit alors une clé de répartition des écarts observés dans la dispersion, selon le poids de chaque mois (ou de chaque trimestre).

**Étape 7 : calcul de la série corrigée des variations saisonnières (CVS)**

**Schéma additif** :

La série corrigée des variations saisonnières est donnée par : $y_t^{CVS} = y_t - s_j$

Sur l'ensemble de la période d'étude, on utilise, pour le même trimestre (ou le même mois) de chaque année, la même valeur du coefficient (ou du rapport) saisonnier corrigé $s_j$, calculé à l'étape 6 (cf. hypothèse d'un mouvement saisonnier qui revient rigoureusement à l'identique chaque année).

**Schéma multiplicatif** :

La série corrigée des variations saisonnières est donnée par : $y_t^{CVS} = \frac{y_t}{1+s_j}$

Remarques terminales :

a) si, lors de l'étape 1, on a soustrait du calcul des valeurs correspondant à des accidents conjoncturels non négligeables, on doit penser à bien les rajouter aux valeurs correspondantes trouvées pour la série CVS, après le calcul de cette dernière.

b) il est possible de lisser encore la série CVS obtenue, en effectuant une MM3, si la tendance générale n'est pas encore suffisamment lisible.

## 4. Exemples

### 41. Exemple de mise en œuvre d'un schéma de composition additif

Le tableau suivant présente l'évolution du chiffre d'affaires d'un magasin (en millions d'euros), sur une période de 5 ans, de 1994 à 1998. L'objectif de l'exercice est de désaisonnaliser la série statistique envisagée. On considère qu'il n'y a aucune valeur "aberrante" dans les données brutes (**étape 1**). Après avoir déterminé le schéma de composition à utiliser, on désaisonnalise au moyen des moyennes mobiles, puis par la méthode analytique :

| | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| **1er trim** | 2,0 | 5,0 | 6,5 | 6,8 | 7,0 |
| **2e trim** | 0,5 | 2,0 | 4,2 | 4,0 | 5,0 |
| **3e trim** | 3,5 | 4,5 | 7,0 | 6,5 | 8,0 |
| **4e trim** | 1,0 | 3,5 | 5,0 | 5,2 | 5,8 |

Détermination **algébrique** du schéma de composition à adopter :

On a :

| **Années** | **1994** | **1995** | **1996** | **1997** | **1998** |
|---|---|---|---|---|---|
| **Ecarts-types** | 1,15 | 1,15 | 1,13 | 1,11 | 1,14 |

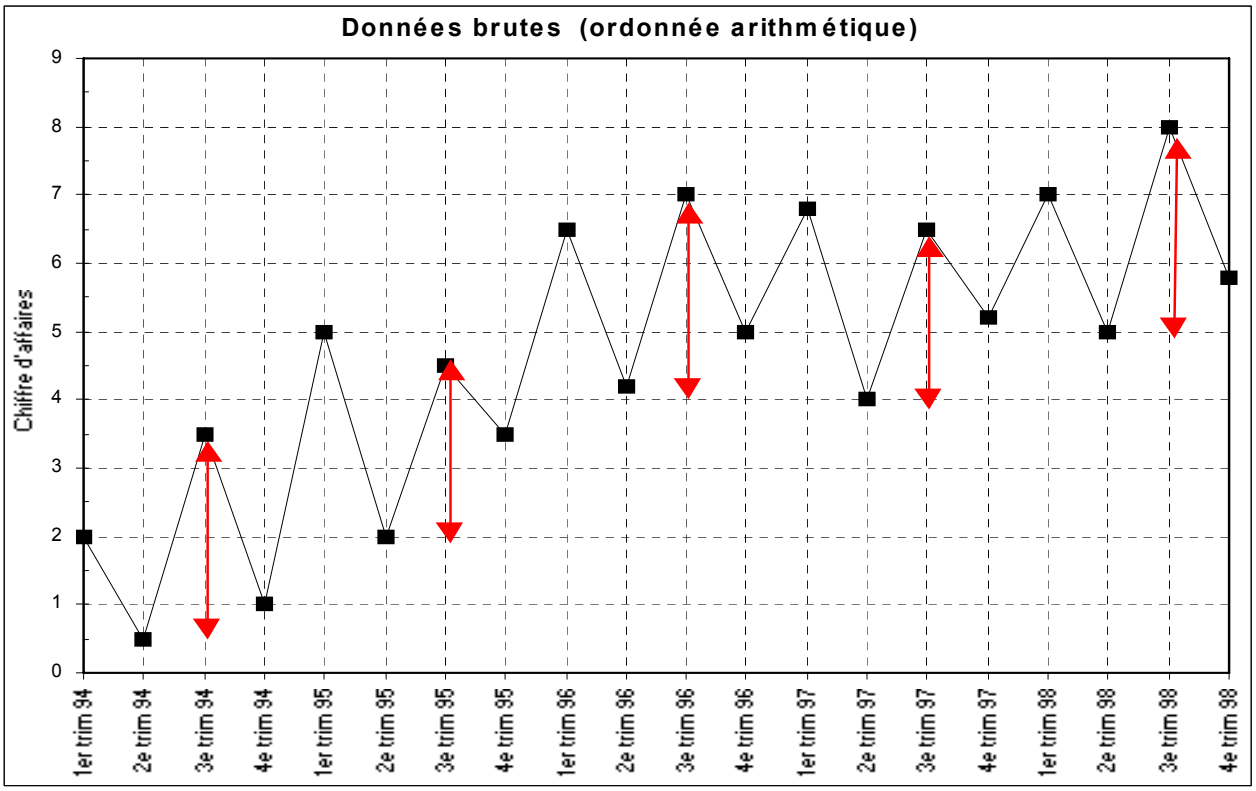
Par exemple, l'écart type pour 1994 se calcule de la façon suivante :

$$\sqrt{\frac{2^2 + 0{,}5^2 + 3{,}5^2 + 1^2}{4} - \left(\frac{2 + 0{,}5 + 3{,}5 + 1}{4}\right)^2}$$

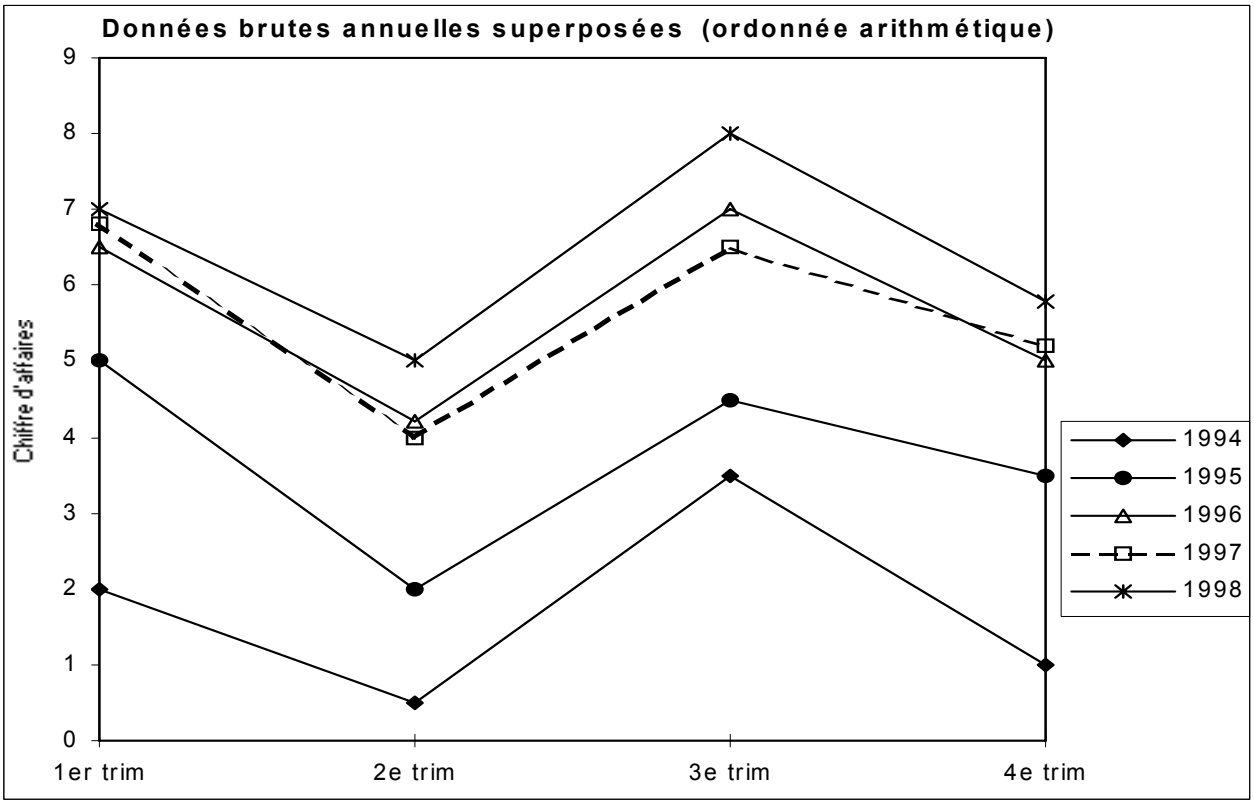
On constate que les écarts-types annuels, calculés sur les 4 trimestres d'une même année, sont à peu près identiques, alors que les valeurs de la série brute tendent à s'accroître fortement sur les 5 années. On retient donc le schéma de composition additif (**étape 2**).

Détermination **graphique** du schéma de composition à adopter :

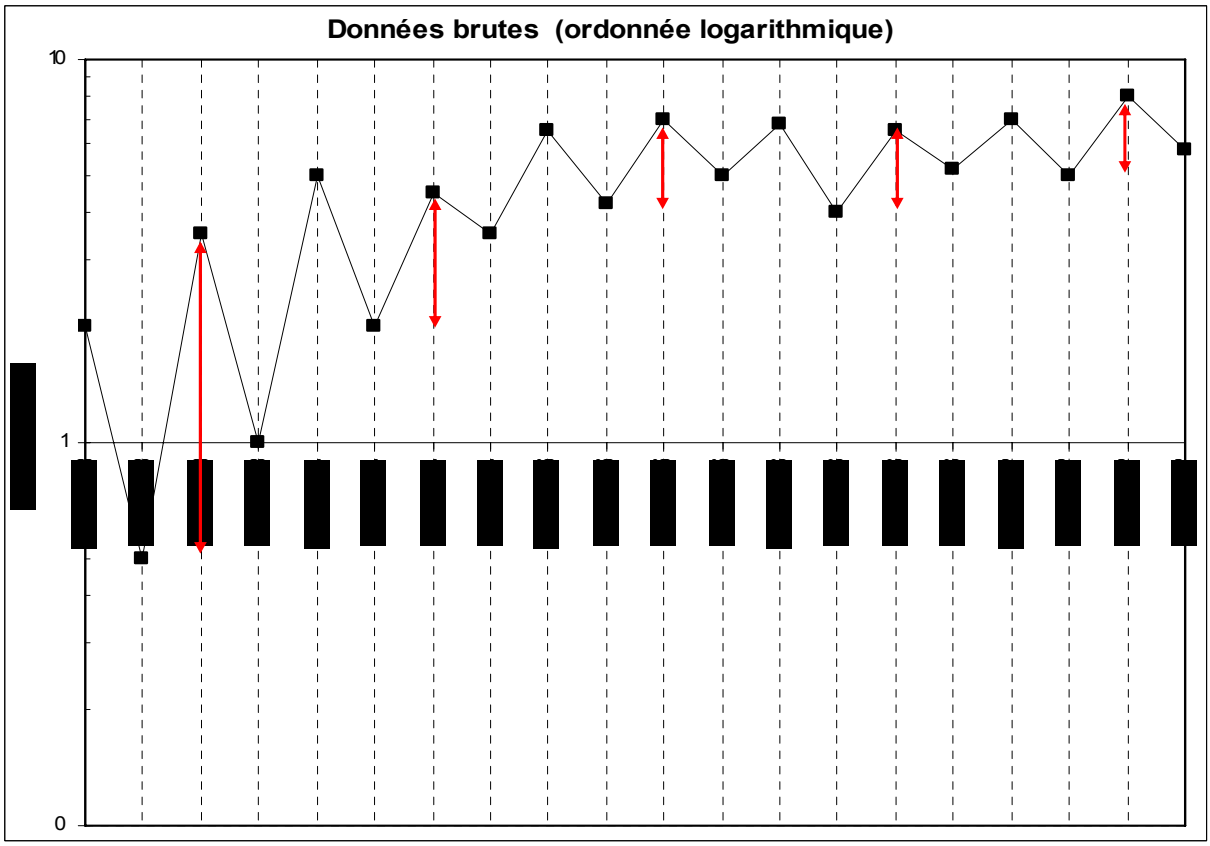
Pour un même trimestre, les amplitudes absolues du mouvement saisonnier sont à peu près les mêmes chaque année (flèches rouges pour les 3èmes trimestres), alors même que le niveau atteint par les valeurs de la série brute s'accroît. Cela conduit à retenir un schéma additif :

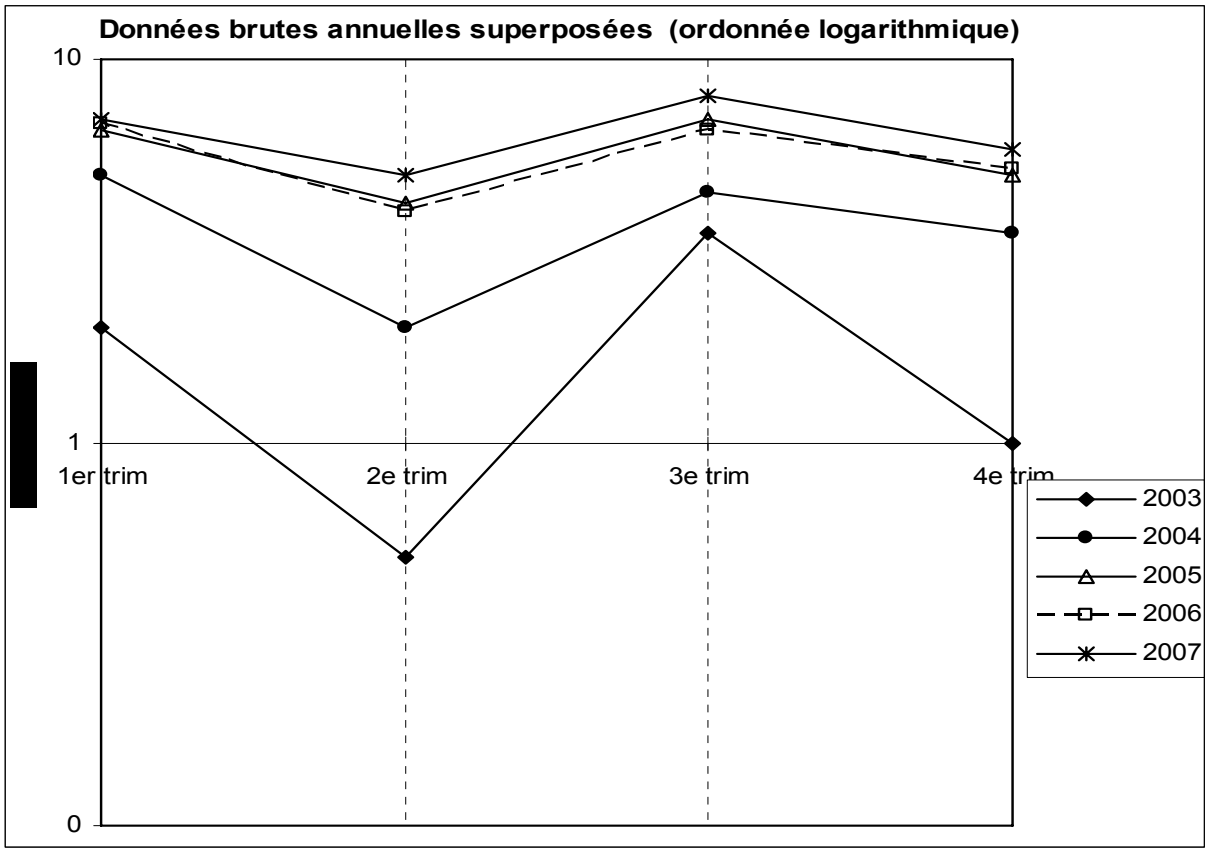



Même si c'est moins net pour deux des 4èmes trimestres, d'une année sur l'autre, les segments relatifs à deux trimestres consécutifs ont tendance à être relativement parallèles entre eux. Cela conduit à choisir un schéma additif :

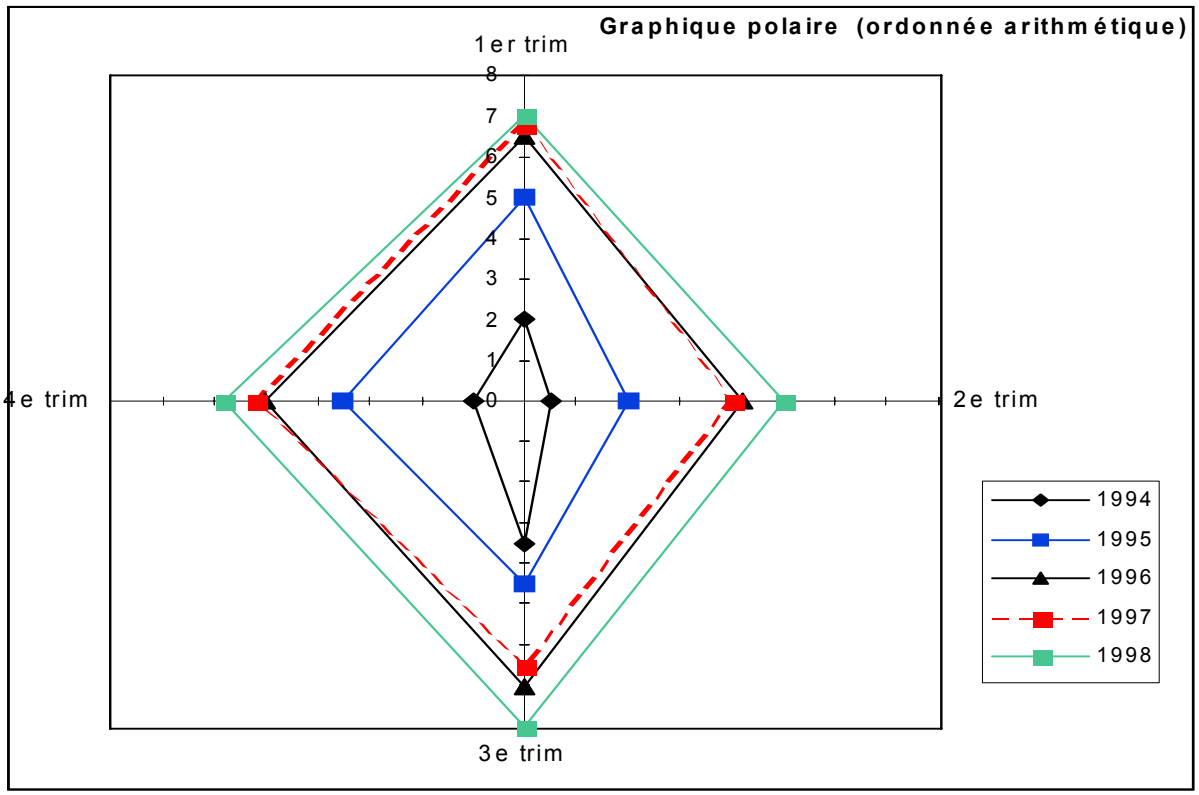



Avec une ordonnée logarithmique, on obtient :

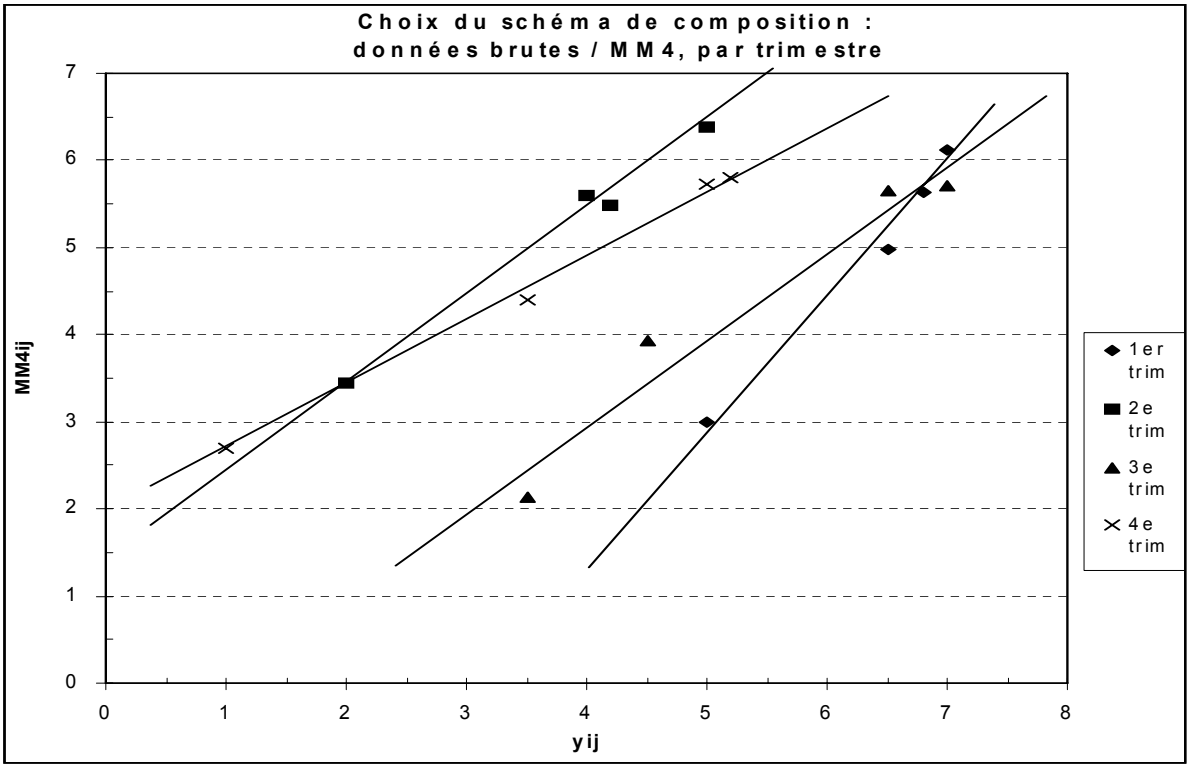





De même, sur un graphique polaire, les segments relatifs à deux trimestres consécutifs ont tendance à être relativement parallèles entre eux. Cela induit le choix d'un schéma additif.
Remarque : le tableur utilisé ici présente un défaut qui ne permet pas de rendre correctement compte de la série brute, chaque courbe annuelle se refermant sur elle-même :



Sur le graphique suivant, qui met en jeu des MM4, les droites ajustées empiriquement et respectivement sur les nuages de points correspondant à chacun des 4 trimestres, seraient parallèles entre elles si l'on avait affaire à un schéma additif pur, ce qui est assez loin d'être le cas ici. Toutefois, à l'inverse, ces droites tendent encore moins à converger toutes les 4 vers l'origine du repère, ce qui signifie qu'on est encore plus loin d'un schéma multiplicatif :




**Désaisonnalisation par la méthode des moyennes mobiles**

On commence par le calcul des moyennes mobiles d'ordre 4 (MM4), à partir du tableau initial des données brutes, afin d'estimer la composante extra-saisonnière de la série (**étape 3**).

On remarque qu'on a une perte d'information de 4 trimestres :

Calcul des moyennes mobiles MMij (MM4)

| | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| **1er trim** | | 3,0 | 5,0 | 5,6 | 6,1 |
| **2e trim** | | 3,4 | 5,5 | 5,6 | 6,4 |
| **3e trim** | 2,1 | 3,9 | 5,7 | 5,7 | |
| **4e trim** | 2,7 | 4,4 | 5,7 | 5,8 | |

Les yij représentent les valeurs du tableau des données brutes. Pour des raisons pédagogiques, on réalise ici le calcul des **coefficients saisonniers**, selon les 2 méthodes (médiane et moyenne).

Pour cela, une fois déterminées les variations saisonnières (**étape 4**), on lit le tableau en ligne et l'on calcule soit la médiane ou la moyenne des 4 valeurs correspondantes (**étape 5**).

Ensuite, on somme les 4 coefficients obtenus pour vérifier le principe de conservation des aires.

Ici, on peut assimiler à zéro les valeurs trouvées, mais pour des raisons pédagogiques, on réalise tout de même une correction des coefficients obtenus :

Calcul des variat saisonnières : $s_{ij} = y_{ij} - MM_{ij}$

| | 1994 | 1995 | 1996 | 1997 | 1998 | Coeff. saisonniers s' j Méd | Coeff. saisonniers sbar j Moy |
|---|---|---|---|---|---|---|---|
| **1er trim** | | 2,0 | 1,5 | 1,2 | 0,9 | 1,338 | 1,391 |
| **2e trim** | | -1,4 | -1,3 | -1,6 | -1,4 | -1,406 | -1,425 |
| **3e trim** | 1,4 | 0,6 | 1,3 | 0,9 | | 1,069 | 1,019 |
| **4e trim** | -1,7 | -0,9 | -0,7 | -0,6 | | -0,813 | -0,978 |
| | | | | Somme | | 0,188 | 0,006 |
| | | | | Somme / 4 | | 0,047 | 0,002 |

Les valeurs corrigées du tableau suivant sont obtenues en retranchant à chaque coefficient saisonnier la valeur 0,047 pour $s'_j$ et la valeur 0,002 pour $sbar_j$. Si l'on effectue les sommes en colonnes sur les coefficients corrigés, on trouve évidemment zéro (**étape 6**) :

| | s'j corr. | sbarj corr. |
|---|---|---|
| **1er trim.** | 1,29 | 1,39 |
| **2ème trim.** | -1,45 | -1,43 |
| **3ème trim.** | 1,02 | 1,02 |
| **4ème trim.** | -0,86 | -0,98 |

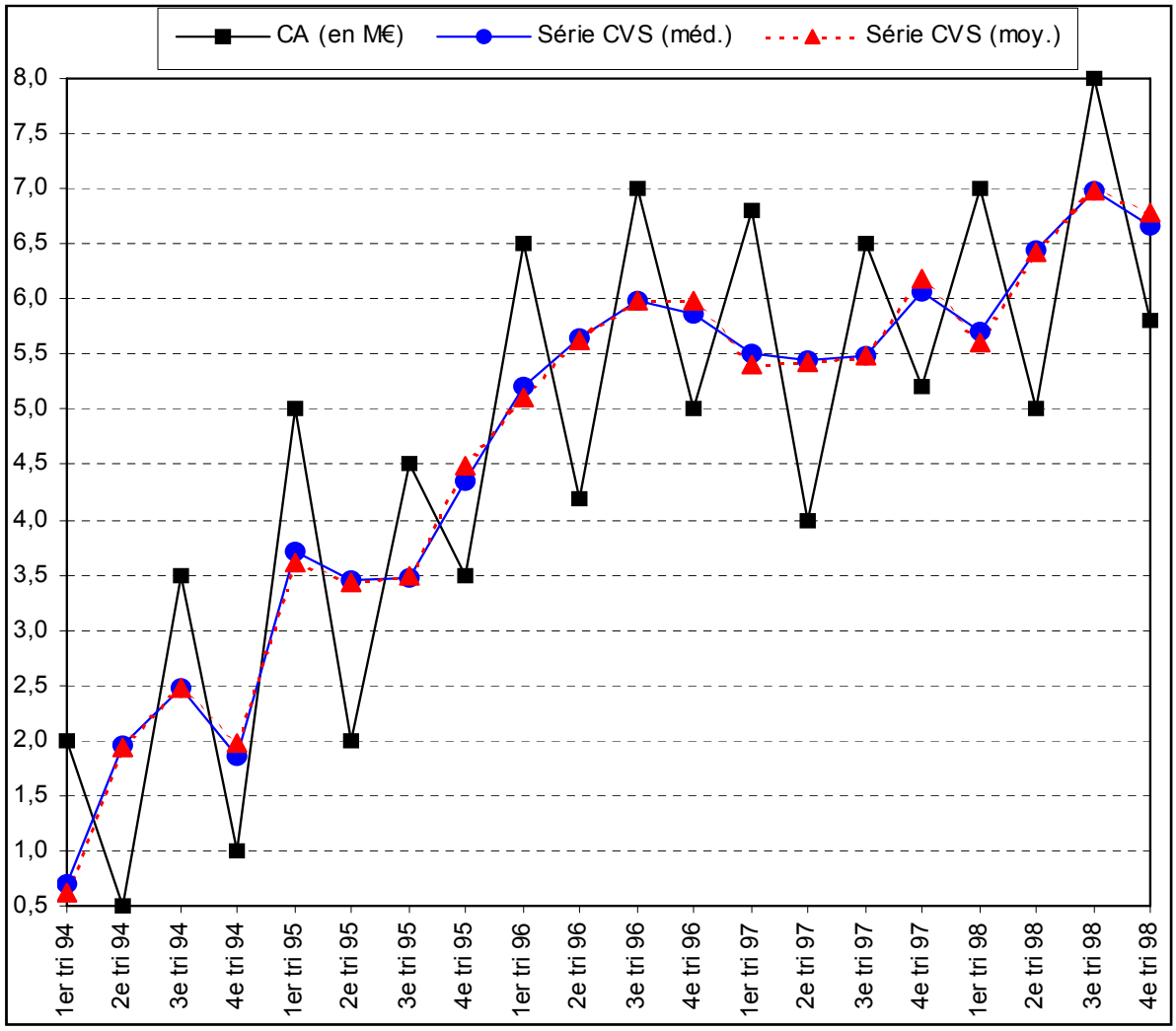
Enfin, en retranchant à chaque donnée brute le coefficient saisonnier corrigé correspondant (selon le trimestre), on obtient les séries CVS (selon la médiane et selon la moyenne) dans les 2 tableaux suivants (**étape 7**) :

| cvs (méd.) | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| **1er trim** | 0,71 | 3,71 | 5,21 | 5,51 | 5,71 |
| **2e trim** | 1,95 | 3,45 | 5,65 | 5,45 | 6,45 |
| **3e trim** | 2,48 | 3,48 | 5,98 | 5,48 | 6,98 |
| **4e trim** | 1,86 | 4,36 | 5,86 | 6,06 | 6,66 |

| cvs (moy.) | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| **1er trim** | 0,61 | 3,61 | 5,11 | 5,41 | 5,61 |
| **2e trim** | 1,93 | 3,43 | 5,63 | 5,43 | 6,43 |
| **3e trim** | 2,48 | 3,48 | 5,98 | 5,48 | 6,98 |
| **4e trim** | 1,98 | 4,48 | 5,98 | 6,18 | 6,78 |

Le graphique suivant indique la série brute et les séries CVS, calculées à partir des médianes et des moyennes.

On constate ici qu'entre les 2 séries CVS les différences sont peu importantes, ce qui n'est pas toujours le cas. Par ailleurs, la tendance générale de la série apparaît nettement mieux, montrant bien que des cassures dans le rythme de la croissance du chiffre d'affaires ont lieu notamment au 1er trimestre 1995 et au 3ème trimestre 1998 :




## Désaisonnalisation par la méthode analytique

Afin d'opter pour le meilleur modèle (linéaire ou exponentiel) d'estimation de la composante extra-saisonnière de la série, on calcule les coefficients de corrélation $\rho_{t,y_t}$ et $\rho_{t,\log y_t}$, à l'aide des 2 tableaux suivants :

| Modèle linéaire | | | | | |
|---|---|---|---|---|---|
| Tri. ($x_t$) | CA ($y_t$) | $x_t^2$ | $y_t^2$ | $x_t\ y_t$ | $y^*_t$ |
| **1** | 2,0 | 1 | 4,0 | 2,0 | 2,1 |
| **2** | 0,5 | 4 | 0,3 | 1,0 | 2,4 |
| **3** | 3,5 | 9 | 12,3 | 10,5 | 2,7 |
| **4** | 1,0 | 16 | 1,0 | 4,0 | 2,9 |
| **5** | 5,0 | 25 | 25,0 | 25,0 | 3,2 |
| **6** | 2,0 | 36 | 4,0 | 12,0 | 3,5 |
| **7** | 4,5 | 49 | 20,3 | 31,5 | 3,7 |
| **8** | 3,5 | 64 | 12,3 | 28,0 | 4,0 |
| **9** | 6,5 | 81 | 42,3 | 58,5 | 4,3 |
| **10** | 4,2 | 100 | 17,6 | 42,0 | 4,5 |
| **11** | 7,0 | 121 | 49,0 | 77,0 | 4,8 |
| **12** | 5,0 | 144 | 25,0 | 60,0 | 5,0 |
| **13** | 6,8 | 169 | 46,2 | 88,4 | 5,3 |
| **14** | 4,0 | 196 | 16,0 | 56,0 | 5,6 |
| **15** | 6,5 | 225 | 42,3 | 97,5 | 5,8 |
| **16** | 5,2 | 256 | 27,0 | 83,2 | 6,1 |
| **17** | 7,0 | 289 | 49,0 | 119,0 | 6,4 |
| **18** | 5,0 | 324 | 25,0 | 90,0 | 6,6 |
| **19** | 8,0 | 361 | 64,0 | 152,0 | 6,9 |
| **20** | 5,8 | 400 | 33,6 | 116,0 | 7,2 |
| **210** | **93,0** | **2870** | **516,1** | **1153,6** | |

xbar = 10,50 ectyp x = 5,77

ybar = 4,65 ectyp y = 2,04

**rho xy = 0,75**

cov xy = 8,85

rho² xy = 0,56

a = 0,27 b = 1,85

$$y^*_t = \hat{f}_t = 0{,}27\,t + 1{,}85$$

| Modèle exponentiel | | |
|---|---|---|
| $\log (y_t)$ | $\log (y_t)^2$ | $x_t \log (y_t)$ |
| 0,30103 | 0,09062 | 0,30103 |
| -0,30103 | 0,09062 | -0,60206 |
| 0,54407 | 0,29601 | 1,63220 |
| 0,00000 | 0,00000 | 0,00000 |
| 0,69897 | 0,48856 | 3,49485 |
| 0,30103 | 0,09062 | 1,80618 |
| 0,65321 | 0,42669 | 4,57249 |
| 0,54407 | 0,29601 | 4,35254 |
| 0,81291 | 0,66083 | 7,31622 |
| 0,62325 | 0,38844 | 6,23249 |
| 0,84510 | 0,71419 | 9,29608 |
| 0,69897 | 0,48856 | 8,38764 |
| 0,83251 | 0,69307 | 10,82262 |
| 0,60206 | 0,36248 | 8,42884 |
| 0,81291 | 0,66083 | 12,19370 |
| 0,71600 | 0,51266 | 11,45605 |
| 0,84510 | 0,71419 | 14,36667 |
| 0,69897 | 0,48856 | 12,58146 |
| 0,90309 | 0,81557 | 17,15871 |
| 0,76343 | 0,58282 | 15,26856 |
| **11,896** | **8,861** | **149,066** |

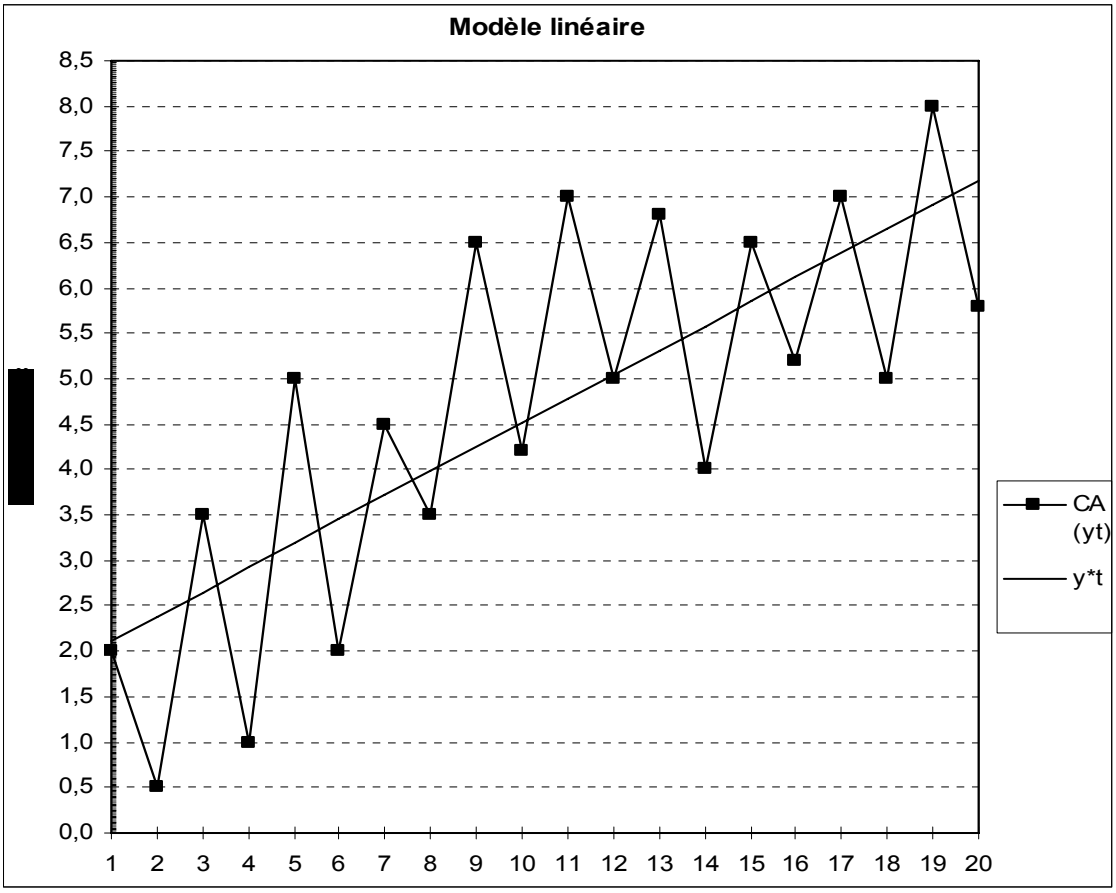
| | | | |
|---|---|---|---|
| xbar = | 10,50 | ectyp x = | 5,77 |
| log (ybar) = | 0,59 | ectyp logy = | 0,30 |
| cov x logy = | 1,21 | **rho x logy =** | **0,70** |
| | | rho² x logy = | 0,49 |
| a* = | 0,036 | b* = | 0,21 |

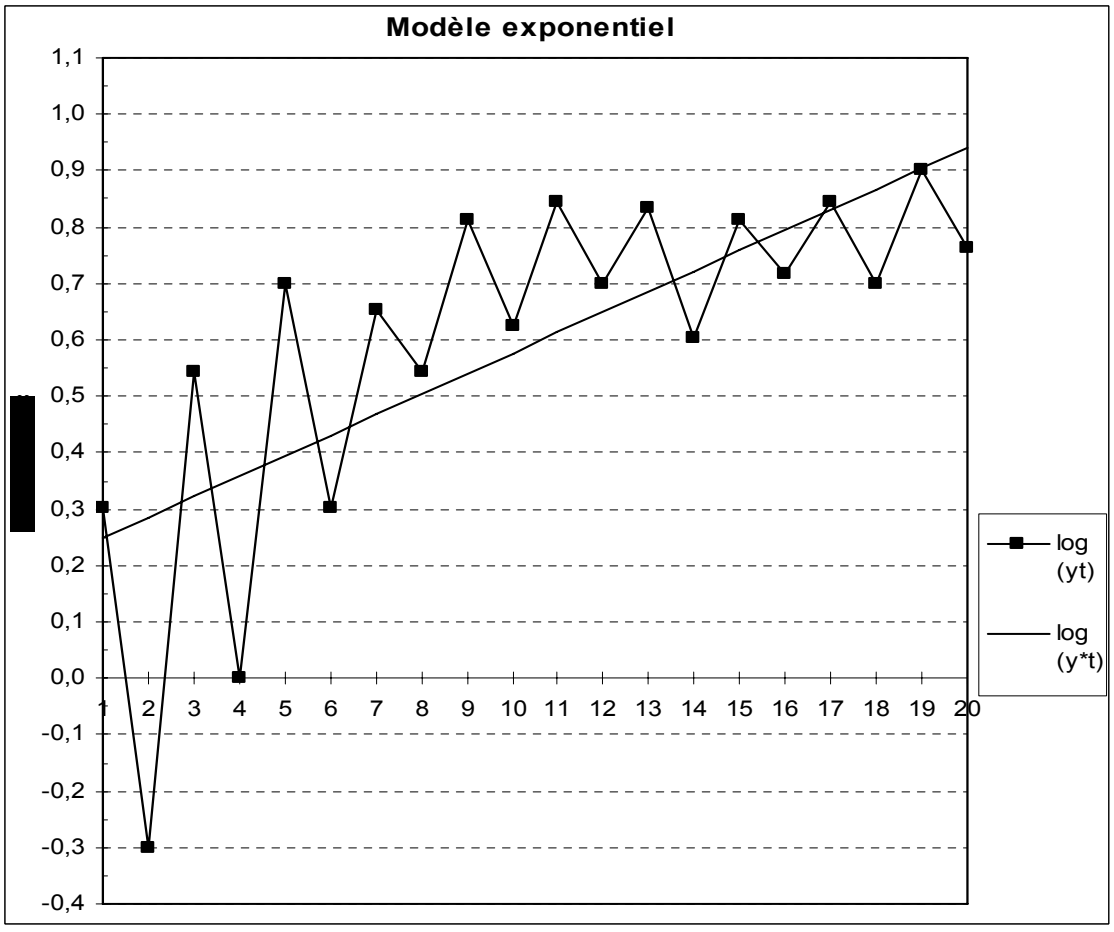
Conclusion :

Le coefficient de corrélation linéaire rho xy (ou le coefficient de détermination rho² xy) du modèle linéaire est ici supérieur à celui du modèle exponentiel. On a :

$$\rho_{t,y_t} \geq \rho_{t,\log y_t} \Leftrightarrow 0,75 > 0,70$$

On retient donc un **ajustement linéaire** pour estimer le trend de la série étudiée.


Modèle linéaire
8,5
8,0
7,5
7,0
6,5
6,0
5,5
5,0
4,5
4,0
3,5
3,0
2,5
2,0
1,5
1,0
0,5
0,0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
CA (yt)
y*t


Modèle exponentiel
1,1
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
-0,1
-0,2
-0,3
-0,4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
log (yt)
log (y*t)

Pour la suite de la désaisonnalisation, on retrouve les mêmes étapes qu'avec la méthode des moyennes mobiles.

Ici cependant, on note que **l'on ne perd pas d'information**. Pour cette raison, on dispose de 5 valeurs par ligne.

C'est pourquoi nous calculons ici une **moyenne tronquée** (sur 3 valeurs, après avoir, dans chaque ligne, soustrait au calcul les valeurs la plus haute et la plus basse), dans la colonne de droite :

**Calcul des variations saisonnières sij = yij - y*ij**

| | 1994 | 1995 | 1996 | 1997 | 1998 | Coeff. saisonniers s' j Méd | sbar j tr Moy tr |
|---|---|---|---|---|---|---|---|
| **1er tri** | -0,120 | 1,815 | 2,249 | 1,484 | 0,619 | 1,484 | 1,306 |
| **2e tri** | -1,886 | -1,452 | -0,317 | -1,582 | -1,647 | -1,582 | -1,560 |
| **3e tri** | 0,847 | 0,782 | 2,217 | 0,652 | 1,086 | 0,847 | 0,905 |
| **4e tri** | -1,919 | -0,484 | -0,049 | -0,915 | -1,380 | -0,915 | -0,926 |
| | | | | **Somme** | | -0,165 | -0,275 |
| | | | | **Somme / 4** | | -0,041 | -0,069 |

Ici encore, on somme les 4 coefficients obtenus pour vérifier le principe de conservation des aires et l'on peut assimiler à zéro les valeurs trouvées, mais pour des raisons pédagogiques, on réalise tout de même une correction des coefficients obtenus :

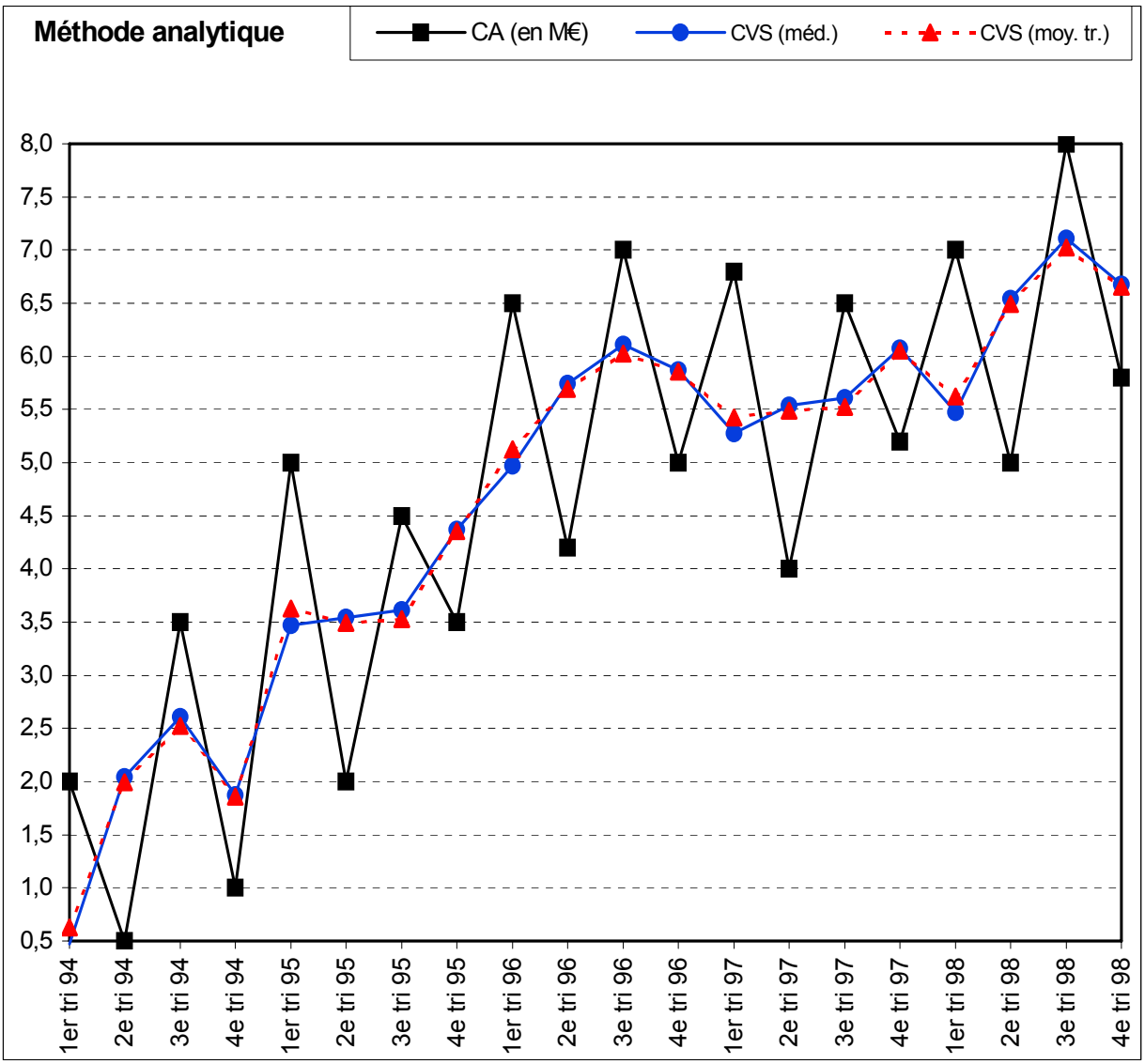
| | s'j corr. | sbarj corr. |
|---|---|---|
| 1er trim. | 1,53 | 1,37 |
| 2ème trim. | -1,54 | -1,49 |
| 3ème trim. | 0,89 | 0,97 |
| 4ème trim. | -0,87 | -0,86 |

En procédant de la même façon que pour les moyennes mobiles, on obtient les séries CVS (selon la médiane et selon la moyenne tronquée) dans les 2 tableaux suivants :

| CVS (méd.) | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| **1er tri** | 0,47 | 3,47 | 4,97 | 5,27 | 5,47 |
| **2e tri** | 2,04 | 3,54 | 5,74 | 5,54 | 6,54 |
| **3e tri** | 2,61 | 3,61 | 6,11 | 5,61 | 7,11 |
| **4e tri** | 1,87 | 4,37 | 5,87 | 6,07 | 6,67 |

| CVS (moy.) | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| **1er tri** | 0,63 | 3,63 | 5,13 | 5,43 | 5,63 |
| **2e tri** | 1,99 | 3,49 | 5,69 | 5,49 | 6,49 |
| **3e tri** | 2,53 | 3,53 | 6,03 | 5,53 | 7,03 |
| **4e tri** | 1,86 | 4,36 | 5,86 | 6,06 | 6,66 |

Dans cet exemple, les courbes CVS calculées selon la méthode des MM et selon la méthode analytique sont peu différentes, car la série ne comporte pas de fortes ruptures de tendance sur l'ensemble de la période d'étude :




### 42. Exemple de mise en œuvre d'un schéma de composition multiplicatif

Nous reprenons ici les valeurs de 1994 de l'exemple précédent, puis nous appliquons à ces valeurs, pour chacune des années suivantes, un coefficient multiplicateur croissant, tel qu'on ait :

- valeurs de 1995 = valeurs de 1994 x 1,3 ; - valeurs de 1996 = valeurs de 1994 x 1,5 ;
- valeurs de 1997 = valeurs de 1994 x 1,8 ; - valeurs de 1998 = valeurs de 1994 x 2,1 .

Nous allons donc étudier ici un mouvement multiplicatif pur, c-à-d que les variations saisonnières sont rigoureusement proportionnelles au niveau absolu de la variable chiffre d'affaires. La démarche générale reste la même que précédemment. Le tableau résultant de ces hypothèses est le suivant :

| | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| **1er trim** | 2,0 | 2,6 | 3,0 | 3,6 | 4,2 |
| **2e trim** | 0,5 | 0,7 | 0,8 | 0,9 | 1,1 |
| **3e trim** | 3,5 | 4,6 | 5,3 | 6,3 | 7,4 |
| **4e trim** | 1,0 | 1,3 | 1,5 | 1,8 | 2,1 |

Détermination **algébrique** du schéma de composition à adopter :

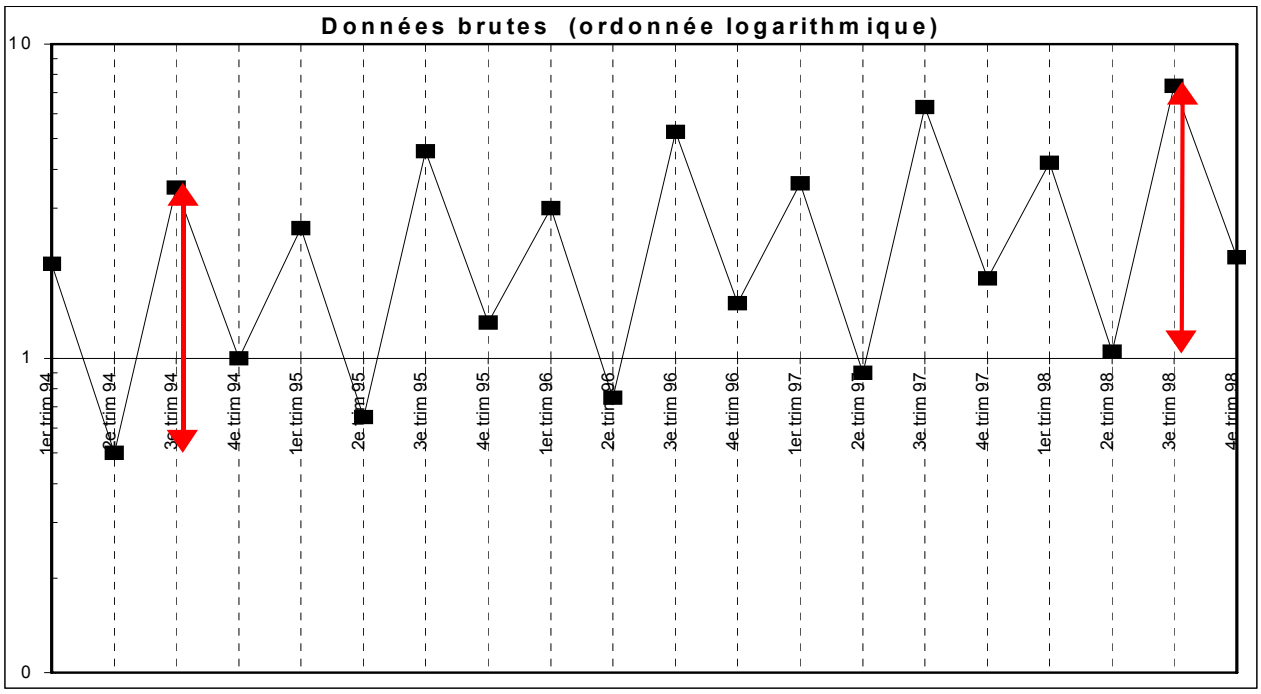
On a :

| **Années** | **1994** | **1995** | **1996** | **1997** | **1998** |
|---|---|---|---|---|---|
| **Ecarts-types** | 1,15 | 1,49 | 1,72 | 2,06 | 2,41 |

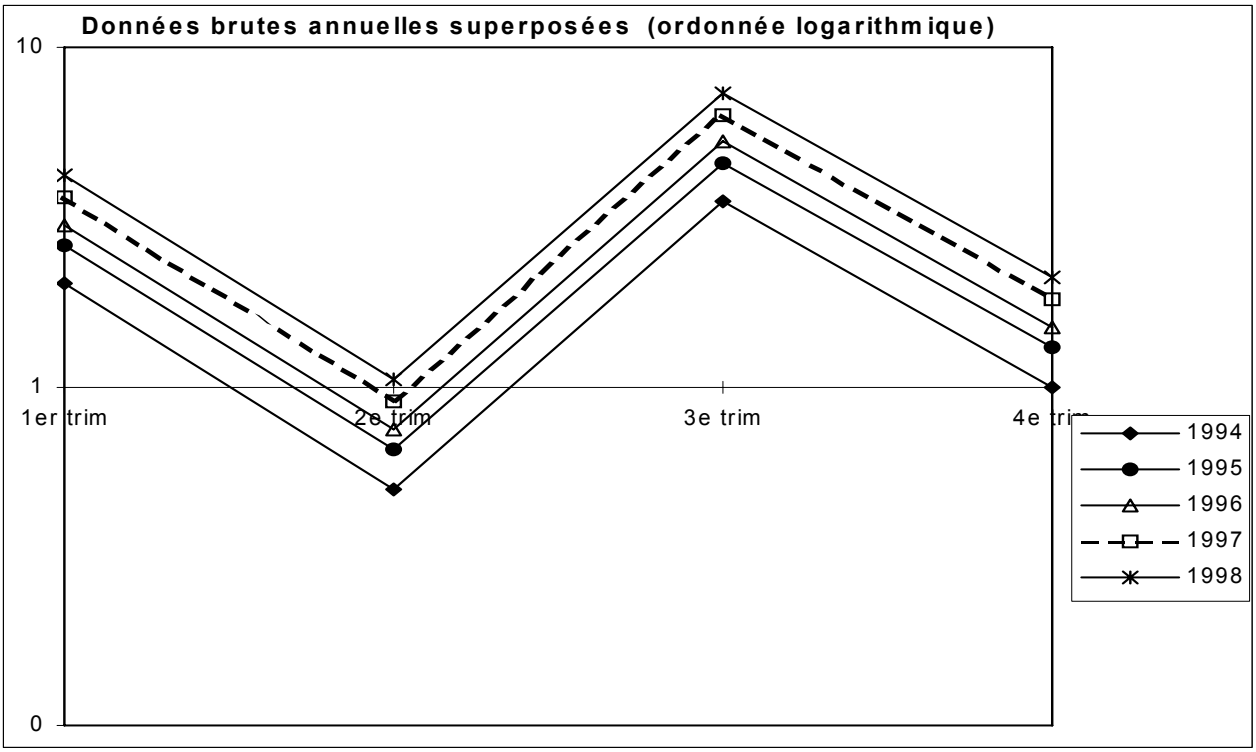
On constate que les écarts-types annuels, s'accroissent en même temps que les valeurs de la série brute sur les 5 années. On retient donc le schéma de composition multiplicatif.

Détermination **graphique** du schéma de composition à adopter :

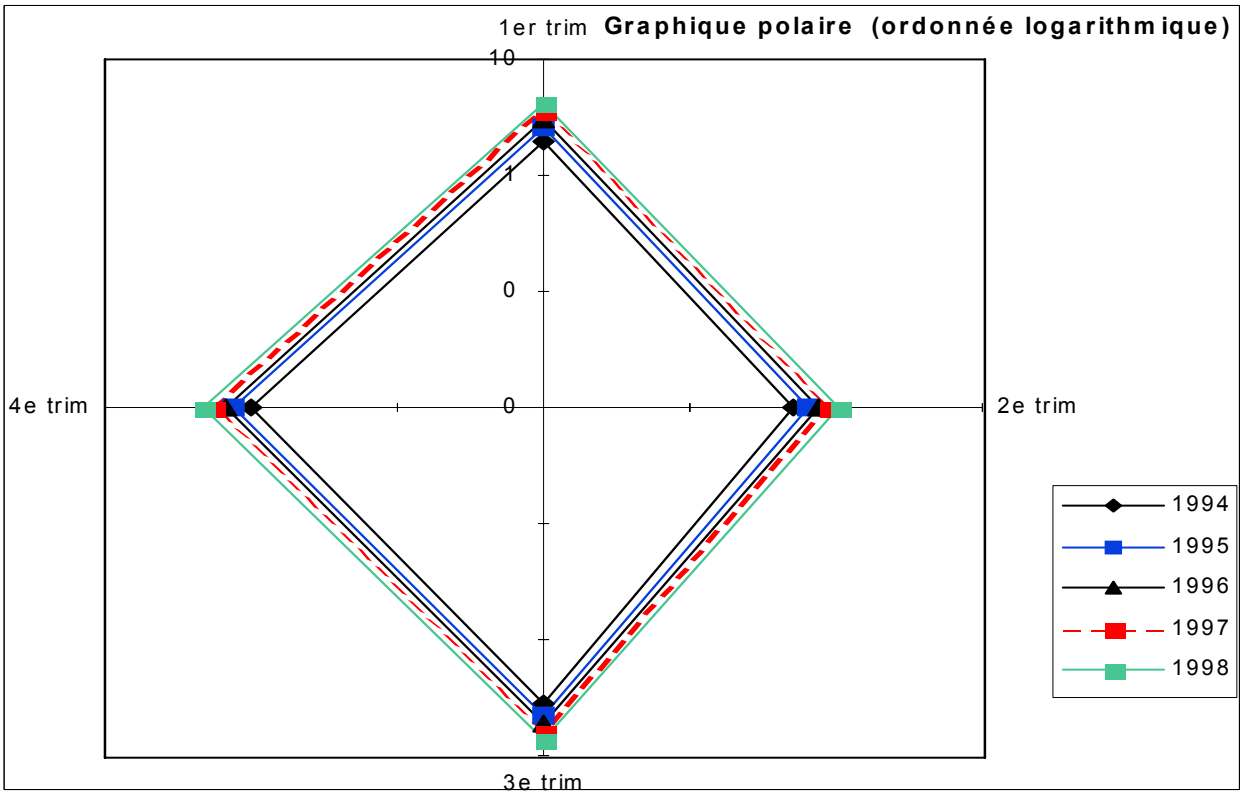
Pour un même trimestre, les amplitudes relatives (ordonnée logarithmique) du mouvement saisonnier sont identiques chaque année (flèches rouges pour les 3$^{èmes}$ trimestres), alors que le niveau atteint par les valeurs de la série brute s'accroît. Cela conduit à retenir un schéma multiplicatif :



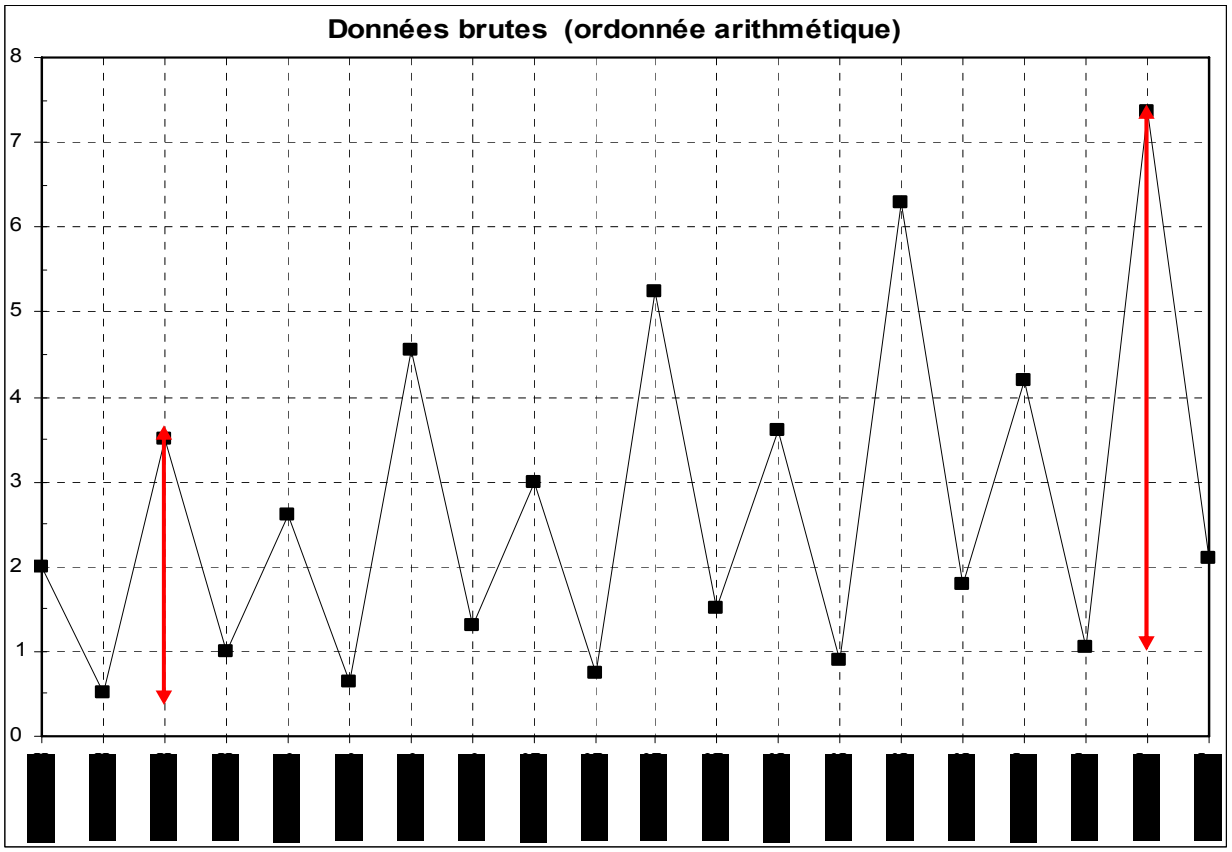
Comme, par construction du tableau de données, on a affaire à un schéma multiplicatif pur, en ordonnée logarithmique les segments relatifs à deux trimestres consécutifs sont rigoureusement parallèles entre eux :

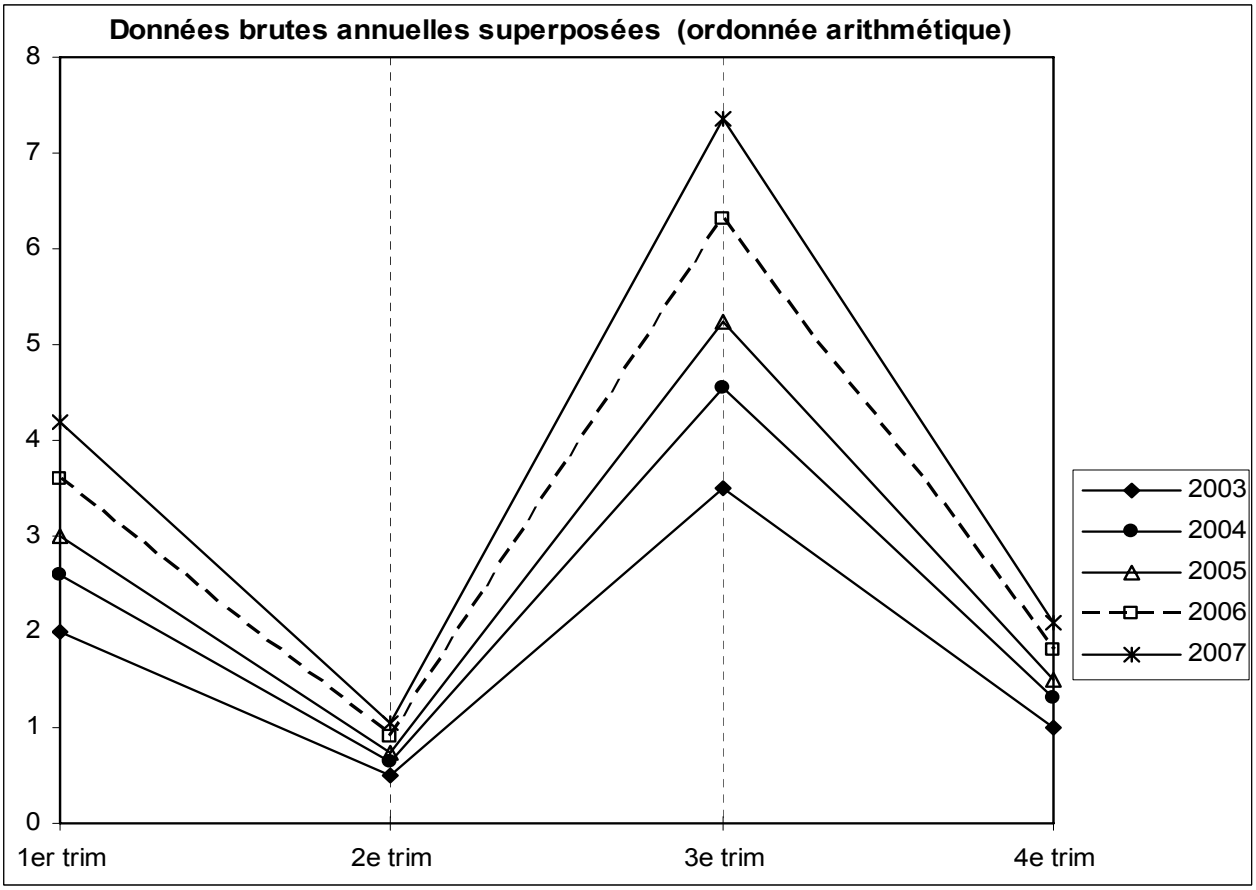



Toujours avec une ordonnée logarithmique, sur un graphique polaire, les segments relatifs à deux trimestres consécutifs sont ici rigoureusement parallèles entre eux :

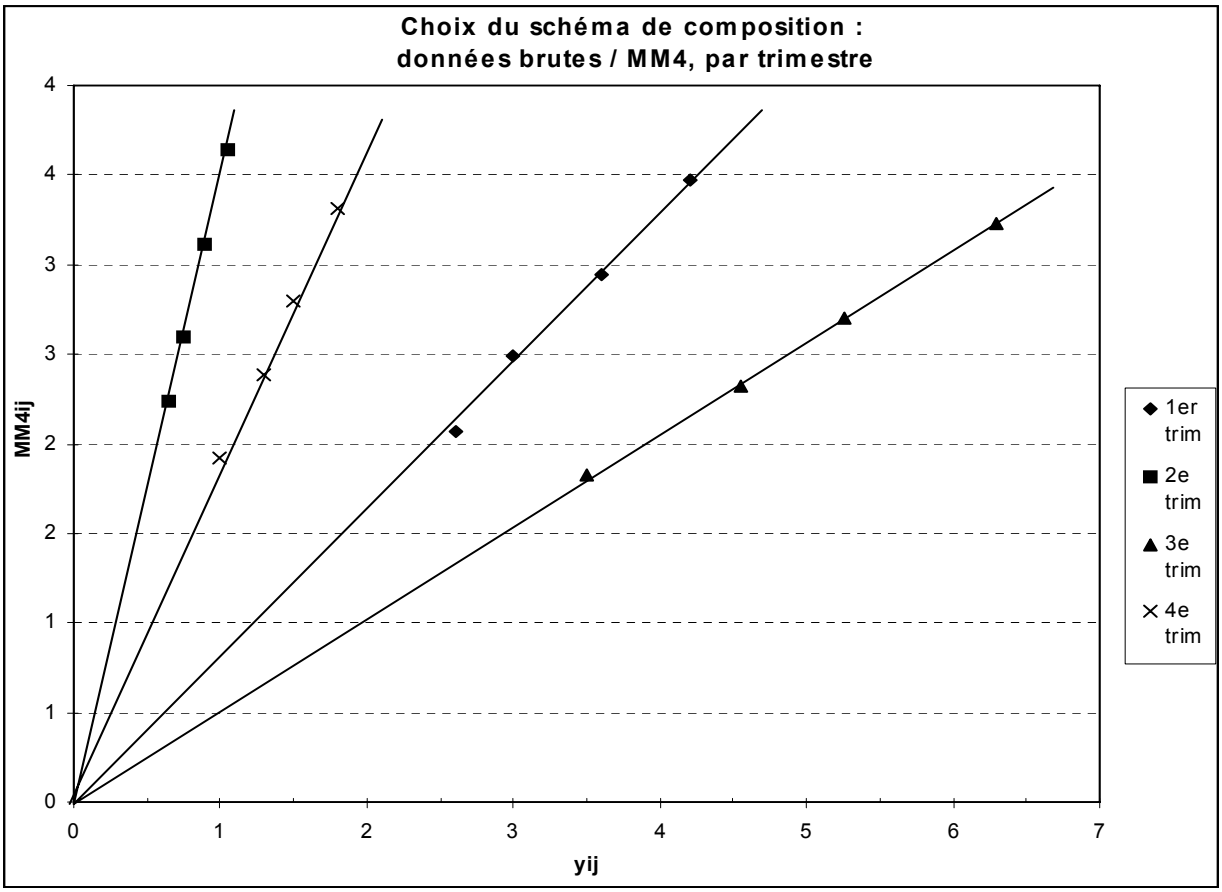


Avec une ordonnée arithmétique, on obtient :



Sur le graphique suivant, qui met en jeu des MM4, les droites ajustées sur les nuages de points correspondant à chacun des 4 trimestres, convergent toutes les quatre vers l'origine du repère, ce qui confirme encore le choix du schéma multiplicatif :




**Désaisonnalisation par la méthode des moyennes mobiles**

On commence par le calcul des moyennes mobiles d'ordre 4 (MM4), à partir du tableau initial des données brutes, afin d'estimer la composante extra-saisonnière de la série (étape 3).

On remarque qu'on a une perte d'information de 4 trimestres :

**Calcul des moyennes mobiles MMij (MM4)**

| | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| **1er trim** | | 2,1 | 2,5 | 2,9 | 3,5 |
| **2e trim** | | 2,2 | 2,6 | 3,1 | 3,6 |
| **3e trim** | 1,8 | 2,3 | 2,7 | 3,2 | |
| **4e trim** | 1,9 | 2,4 | 2,8 | 3,3 | |

Comme dans l'exemple précédent, pour des raisons pédagogiques, on réalise ici le calcul des **rapports saisonniers**, selon les 2 méthodes (médiane et moyenne). Pour cela, une fois déterminées les variations saisonnières, on lit le tableau en ligne et l'on calcule soit la médiane ou la moyenne des 4 valeurs correspondantes. Ensuite, on somme les 4 coefficients obtenus pour vérifier le principe de conservation des aires. Ici, on peut assimiler à 4 les valeurs trouvées, mais pour des raisons pédagogiques, on réalise tout de même une correction des rapports obtenus :

**Calcul des variat saisonnières : $1 + s_{ij} = y_{ij} / MM_{ij}$**

| | 1994 | 1995 | 1996 | 1997 | 1998 | Rapports saisonniers 1+s' j Méd | 1+sbar j Moy |
|---|---|---|---|---|---|---|---|
| **1er trim** | | 1,3 | 1,2 | 1,2 | 1,2 | 1,217 | 1,224 |
| **2e trim** | | 0,3 | 0,3 | 0,3 | 0,3 | 0,289 | 0,289 |
| **3e trim** | 1,9 | 2,0 | 1,9 | 2,0 | | 1,949 | 1,943 |
| **4e trim** | 0,5 | 0,5 | 0,5 | 0,5 | | 0,540 | 0,536 |
| | | | | **Somme** | | 3,994 | 3,993 |
| | | | | **Somme / 4** | | 0,999 | 0,998 |

Les valeurs corrigées du tableau suivant sont obtenues en divisant par 0,999 la valeur de chaque rapport saisonnier pour $s'_j$ et par 0,998 pour $sbar_j$. Si l'on effectue les sommes en colonnes sur les rapports corrigés, on trouve évidemment 4 :

| | 1 + s'j corr. | 1 + sbarj corr. |
|---|---|---|
| **1er trim.** | 1,22 | 1,23 |
| **2ème trim.** | 0,29 | 0,29 |
| **3ème trim.** | 1,95 | 1,95 |
| **4ème trim.** | 0,54 | 0,54 |

Enfin, on calcule le rapport de chaque donnée brute au rapport saisonnier corrigé correspondant (selon le trimestre) et l'on obtient les séries CVS (selon la médiane et selon la moyenne) dans les 2 tableaux suivants :

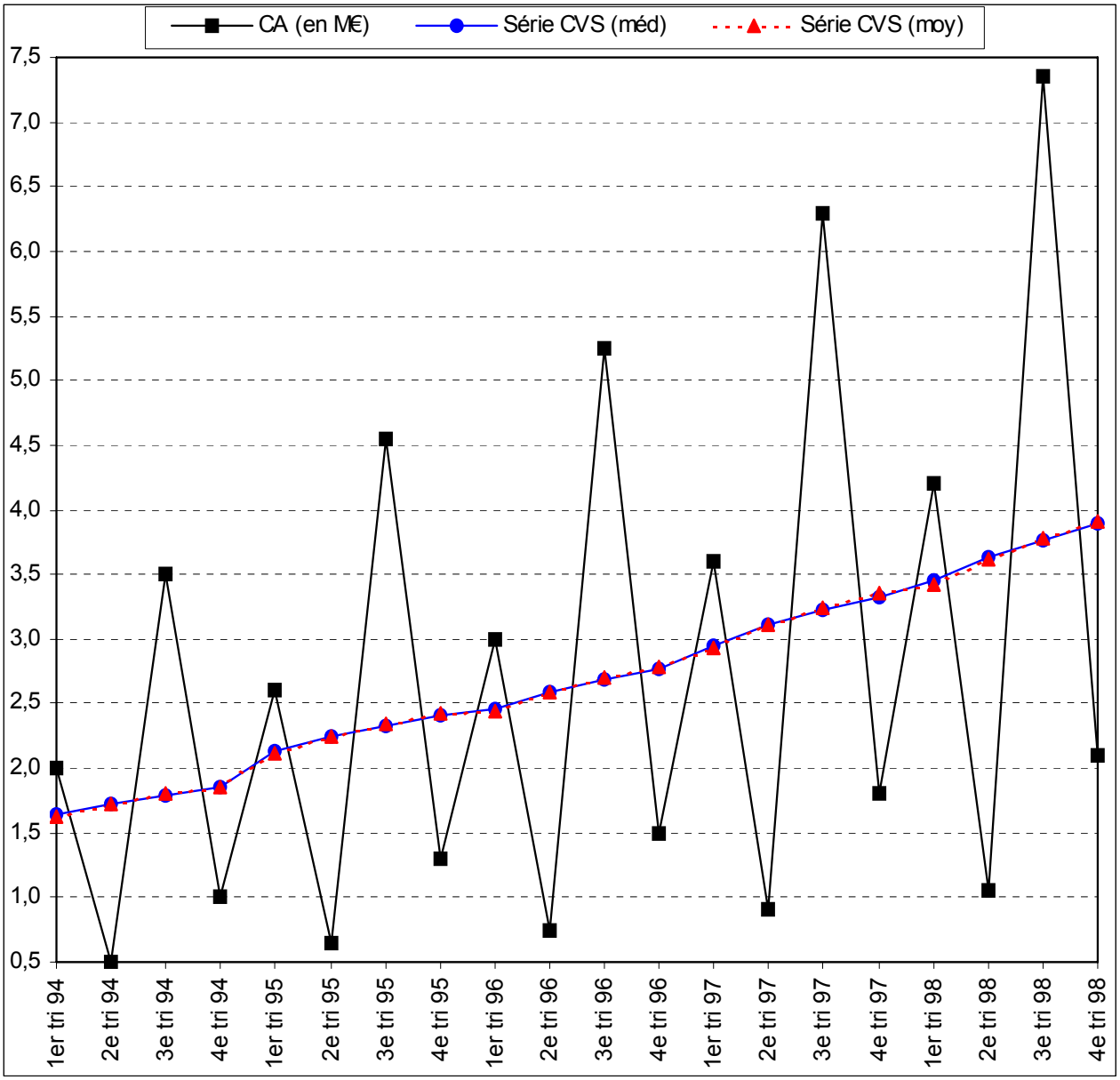
| CVS (méd.) | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| **1er trim** | 1,64 | 2,13 | 2,46 | 2,95 | 3,45 |
| **2e trim** | 1,73 | 2,25 | 2,59 | 3,11 | 3,63 |
| **3e trim** | 1,79 | 2,33 | 2,69 | 3,23 | 3,77 |
| **4e trim** | 1,85 | 2,41 | 2,78 | 3,33 | 3,89 |

| CVS (moy.) | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| **1er trim** | 1,63 | 2,12 | 2,45 | 2,94 | 3,42 |
| **2e trim** | 1,73 | 2,24 | 2,59 | 3,11 | 3,62 |
| **3e trim** | 1,80 | 2,34 | 2,70 | 3,24 | 3,78 |
| **4e trim** | 1,86 | 2,42 | 2,79 | 3,35 | 3,91 |

Le graphique suivant indique la série brute et les séries CVS, calculées à partir des médianes et des moyennes.

On constate ici qu'entre les 2 séries CVS il n'y a pas de différences, mais ce n'est pas toujours le cas.

Par ailleurs, la tendance générale de la série apparaît nettement, montrant une croissance régulière du chiffre d'affaires :



**Désaisonnalisation par la méthode analytique**

Afin d'opter pour le meilleur modèle (linéaire ou exponentiel) d'estimation de la composante extra-saisonnière de la série, on calcule les coefficients de corrélation $\rho_{t,y_t}$ et $\rho_{t,\log y_t}$, à l'aide des 2 tableaux suivants :

| Modèle linéaire | | | | | |
|---|---|---|---|---|---|
| **Tri ($x_t$)** | **CA ($y_t$)** | $x_t^2$ | $y_t^2$ | $x_t\, y_t$ | $y^*_t$ |
| **1** | 2,0 | 1 | 4,0 | 2,0 | 1,6 |
| **2** | 0,5 | 4 | 0,3 | 1,0 | 1,7 |
| **3** | 3,5 | 9 | 12,3 | 10,5 | 1,8 |
| **4** | 1,0 | 16 | 1,0 | 4,0 | 2,0 |
| **5** | 2,6 | 25 | 6,8 | 13,0 | 2,1 |
| **6** | 0,7 | 36 | 0,4 | 3,9 | 2,2 |
| **7** | 4,6 | 49 | 20,7 | 31,9 | 2,3 |
| **8** | 1,3 | 64 | 1,7 | 10,4 | 2,4 |
| **9** | 3,0 | 81 | 9,0 | 27,0 | 2,5 |
| **10** | 0,8 | 100 | 0,6 | 7,5 | 2,6 |
| **11** | 5,3 | 121 | 27,6 | 57,8 | 2,8 |
| **12** | 1,5 | 144 | 2,3 | 18,0 | 2,9 |
| **13** | 3,6 | 169 | 13,0 | 46,8 | 3,0 |
| **14** | 0,9 | 196 | 0,8 | 12,6 | 3,1 |
| **15** | 6,3 | 225 | 39,7 | 94,5 | 3,2 |
| **16** | 1,8 | 256 | 3,2 | 28,8 | 3,3 |
| **17** | 4,2 | 289 | 17,6 | 71,4 | 3,4 |
| **18** | 1,1 | 324 | 1,1 | 18,9 | 3,5 |
| **19** | 7,4 | 361 | 54,0 | 139,7 | 3,7 |
| **20** | 2,1 | 400 | 4,4 | 42,0 | 3,8 |
| **210** | **53,9** | **2870** | **220,3** | **641,6** | |

xbar = 10,50 ectyp x = 5,77

ybar = 2,70 ectyp y = 1,94

**rho xy = 0,34**

cov xy = 3,78

rho² xy = 0,11

a = 0,11 b = 1,50

$$y_t^* = \hat{f}_t = 0{,}11t + 1{,}50$$

| Modèle exponentiel | | |
|---|---|---|
| $\log (y_t)$ | $\log (y_t)^2$ | $x_t \log (y_t)$ |
| 0,30103 | 0,09062 | 0,30103 |
| -0,30103 | 0,09062 | -0,60206 |
| 0,54407 | 0,29601 | 1,63220 |
| 0,00000 | 0,00000 | 0,00000 |
| 0,41497 | 0,17220 | 2,07487 |
| -0,18709 | 0,03500 | -1,12252 |
| 0,65801 | 0,43298 | 4,60608 |
| 0,11394 | 0,01298 | 0,91155 |
| 0,47712 | 0,22764 | 4,29409 |
| -0,12494 | 0,01561 | -1,24939 |
| 0,72016 | 0,51863 | 7,92175 |
| 0,17609 | 0,03101 | 2,11310 |
| 0,55630 | 0,30947 | 7,23193 |
| -0,04576 | 0,00209 | -0,64060 |
| 0,79934 | 0,63895 | 11,99011 |
| 0,25527 | 0,06516 | 4,08436 |
| 0,62325 | 0,38844 | 10,59524 |
| 0,02119 | 0,00045 | 0,38141 |
| 0,86629 | 0,75045 | 16,45946 |
| 0,32222 | 0,10383 | 6,44439 |
| **6,190** | **4,182** | **77,427** |

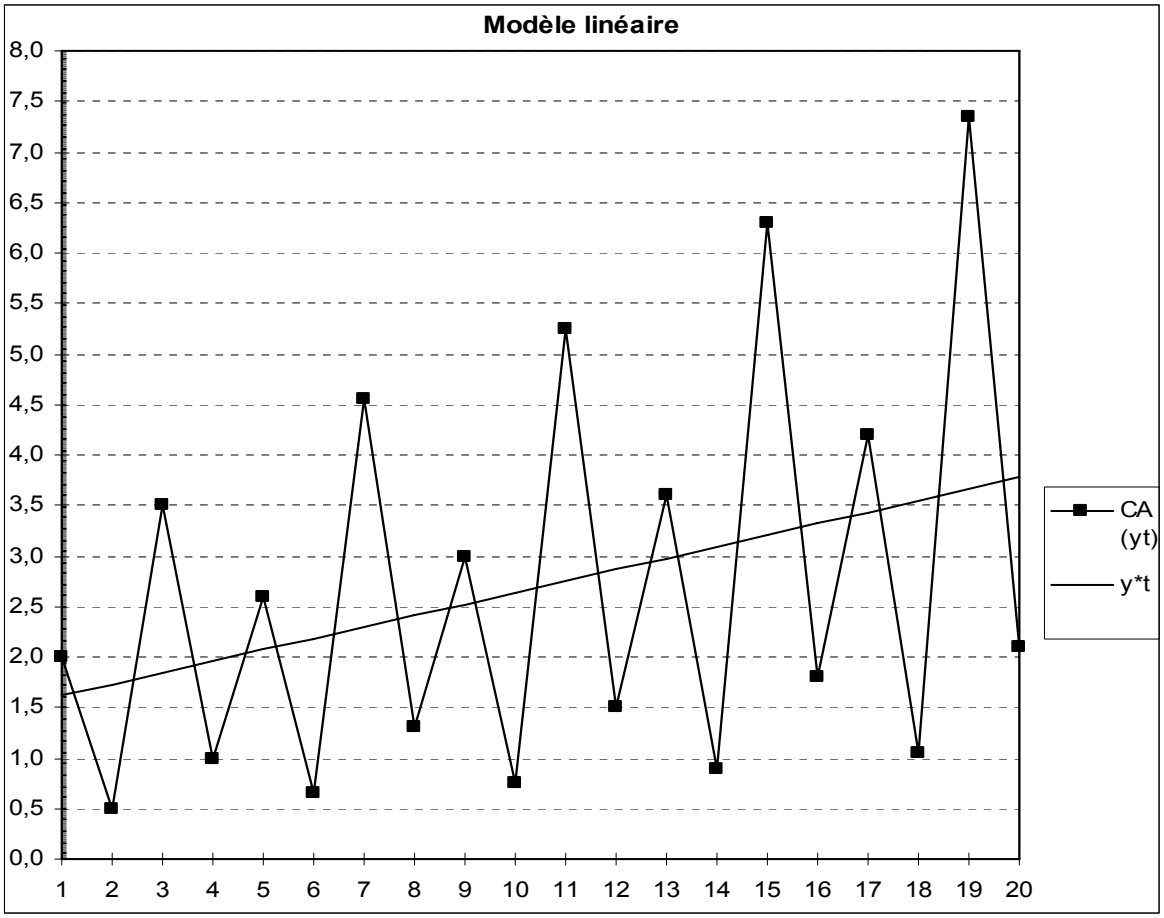
| | | | |
|---|---|---|---|
| xbar = | 10,50 | ectyp x = | 5,77 |
| (log y)bar = | 0,31 | ectyp logy = | 0,34 |
| cov x logy = | 0,62 | **rho x logy =** | **0,32** |
| | | rho² x logy = | 0,10 |
| a* = | 0,019 | b* = | 0,11 |

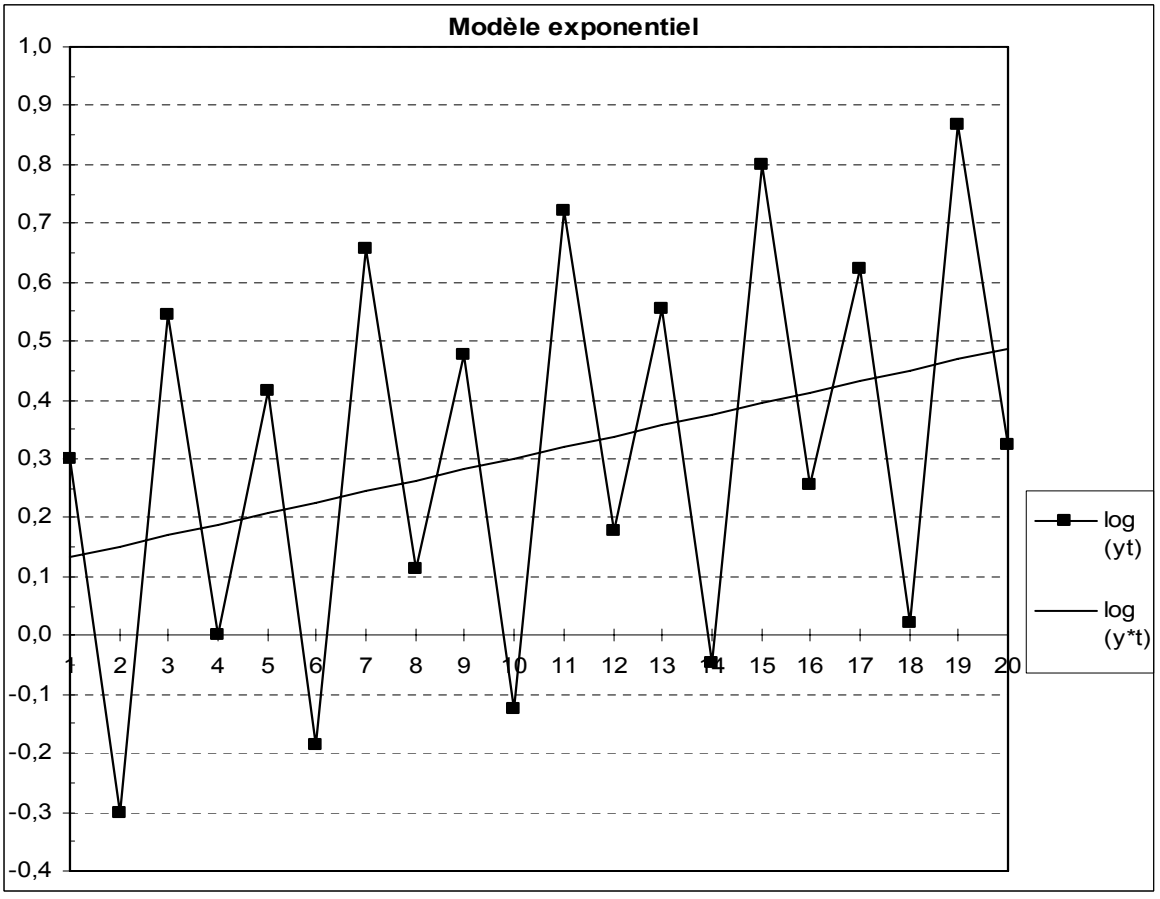
Conclusion :

Le coefficient de corrélation linéaire rho xy (ou le coefficient de détermination rho² xy) du modèle linéaire est ici supérieur à celui du modèle exponentiel. On a :

$$\rho_{t,y_t} \geq \rho_{t,\log y_t} \Leftrightarrow 0,34 > 0,32$$

On retient donc un **ajustement linéaire** pour estimer le trend de la série étudiée.


Modèle linéaire
8,0
7,5
7,0
6,5
6,0
5,5
5,0
4,5
4,0
3,5
3,0
2,5
2,0
1,5
1,0
0,5
0,0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
CA (yt)
y*t


Modèle exponentiel
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
-0,1
-0,2
-0,3
-0,4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
log (yt)
log (y*t)

Pour la suite de la désaisonnalisation, on retrouve les mêmes étapes qu'avec la méthode des moyennes mobiles.

Ici cependant, on note que **l'on ne perd pas d'information**. Pour cette raison, on dispose de 5 valeurs par ligne.

C'est pourquoi nous calculons ici une **moyenne tronquée** (sur 3 valeurs, après avoir, dans chaque ligne, soustrait au calcul les valeurs la plus haute et la plus basse), dans la colonne de droite :

**Calcul des variations saisonnières sij = yij / y*ij**

| | 1994 | 1995 | 1996 | 1997 | 1998 | Rapports saisonniers 1+s' j Méd | Rapports saisonniers 1+sbar j tr Moy tr |
|---|---|---|---|---|---|---|---|
| **1er tri** | 1,238 | 1,256 | 1,188 | 1,208 | 1,223 | 1,2231 | 1,2233 |
| **2e tri** | 0,289 | 0,298 | 0,284 | 0,291 | 0,296 | 0,2910 | 0,2921 |
| **3e tri** | 1,900 | 1,981 | 1,908 | 1,965 | 2,007 | 1,9647 | 1,9511 |
| **4e tri** | 0,511 | 0,539 | 0,523 | 0,542 | 0,556 | 0,5392 | 0,5349 |
| | | | | **Somme** | | 4,0180 | 4,0014 |
| | | | | **Somme / 4** | | 1,0045 | 1,0003 |

Ici encore, on somme les 4 coefficients obtenus pour vérifier le principe de conservation des aires et l'on peut assimiler à 4 les valeurs trouvées, mais pour des raisons pédagogiques, on réalise tout de même une correction des coefficients obtenus :

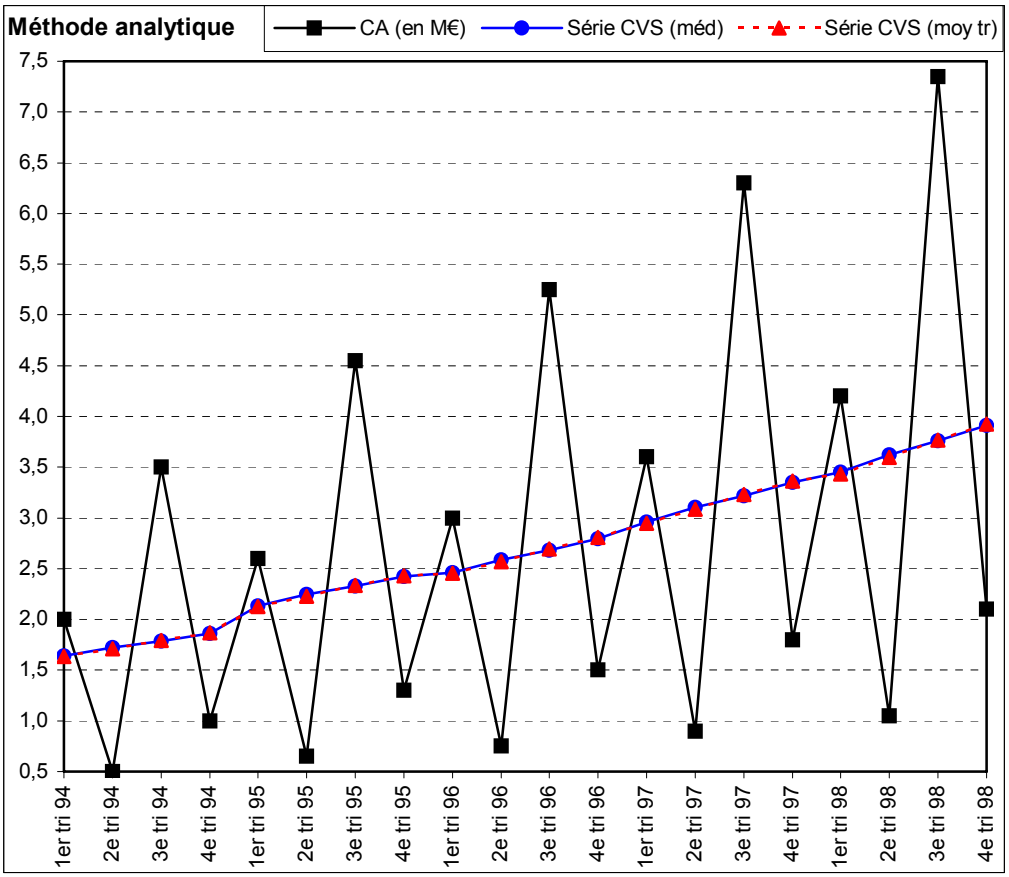
| | 1 + s'j corr. | 1 + sbarj corr. |
|---|---|---|
| **1er trim.** | 1,22 | 1,22 |
| **2ème trim.** | 0,29 | 0,29 |
| **3ème trim.** | 1,96 | 1,95 |
| **4ème trim.** | 0,54 | 0,53 |

En procédant de la même façon que pour les moyennes mobiles, on obtient les séries CVS (selon la médiane et selon la moyenne tronquée) dans les deux tableaux suivants :

| CVS (méd.) | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| **1er trim** | 1,64 | 2,14 | 2,46 | 2,96 | 3,45 |
| **2e trim** | 1,73 | 2,24 | 2,59 | 3,11 | 3,62 |
| **3e trim** | 1,79 | 2,33 | 2,68 | 3,22 | 3,76 |
| **4e trim** | 1,86 | 2,42 | 2,79 | 3,35 | 3,91 |

| CVS (moy.) | 1994 | 1995 | 1996 | 1997 | 1998 |
|---|---|---|---|---|---|
| **1er trim** | 1,64 | 2,13 | 2,45 | 2,94 | 3,43 |
| **2e trim** | 1,71 | 2,23 | 2,57 | 3,08 | 3,60 |
| **3e trim** | 1,79 | 2,33 | 2,69 | 3,23 | 3,77 |
| **4e trim** | 1,87 | 2,43 | 2,81 | 3,37 | 3,93 |

Dans cet exemple, les courbes CVS calculées selon la méthode des MM et selon la méthode analytique sont très peu différentes, car la série ne comporte pas de fortes ruptures de tendance sur l'ensemble de la période d'étude :




Remarque terminale : exemple de **prévision à court terme**.

Supposons que le directeur du magasin souhaite savoir à quel niveau le chiffre d'affaires se situera au 3ème trimestre de 1999 (soit t = 23 trimestres, depuis le début de la série).

En posant l'hypothèse que la tendance croissante du chiffre d'affaires se maintienne dans le futur, on peut utiliser la droite d'ajustement calculée plus haut : $\hat{y}_t = 0{,}11t + 1{,}50$ .

On aura donc :

$$\hat{y}_{23} = 0{,}11 \times 23 + 1{,}50 \Leftrightarrow \hat{y}_{3\text{è tri. }99} \approx 4{,}0 \text{ millions d'euros}$$

## 43. 2ème exemple de mise en œuvre d'un schéma de composition additif

### I. CHOIX DU SCHEMA DE COMPOSITION

**Données brutes $y_{ij}$**

| | CA (en K€) |
|---|---|
| **1er trim 03** | 305 |
| **2e trim 03** | 390 |
| **3e trim 03** | 211 |
| **4e trim 03** | 532 |
| **1er trim 04** | 345 |
| **2e trim 04** | 388 |
| **3e trim 04** | 255 |
| **4e trim 04** | 576 |
| **1er trim 05** | 367 |
| **2e trim 05** | 410 |
| **3e trim 05** | 278 |
| **4e trim 05** | 596 |
| **1er trim 06** | 382 |
| **2e trim 06** | 424 |
| **3e trim 06** | 291 |
| **4e trim 06** | 612 |
| **1er trim 07** | 444 |
| **2e trim 07** | 474 |
| **3e trim 07** | 348 |
| **4e trim 07** | 672 |

**Données brutes $y_{ij}$**

| | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|
| **1er trim** | 305 | 345 | 367 | 382 | 444 |
| **2e trim** | 390 | 388 | 410 | 424 | 474 |
| **3e trim** | 211 | 255 | 278 | 291 | 348 |
| **4e trim** | 532 | 576 | 596 | 612 | 672 |

**Etape 1 : repérage des valeurs "aberrantes"**

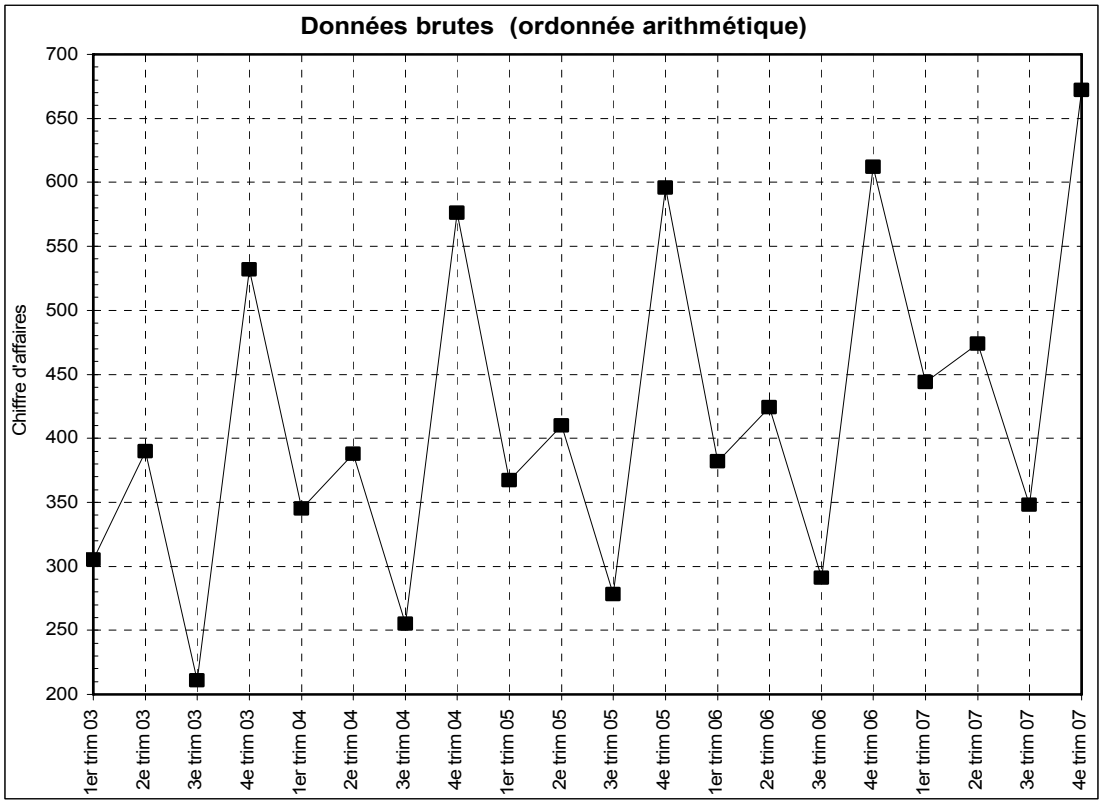
| **Années** | **2003** | **2004** | **2005** | **2006** | **2007** |
|---|---|---|---|---|---|
| **Ecarts-types** | 118,0 | 117,1 | 116,0 | 117,0 | 117,8 |

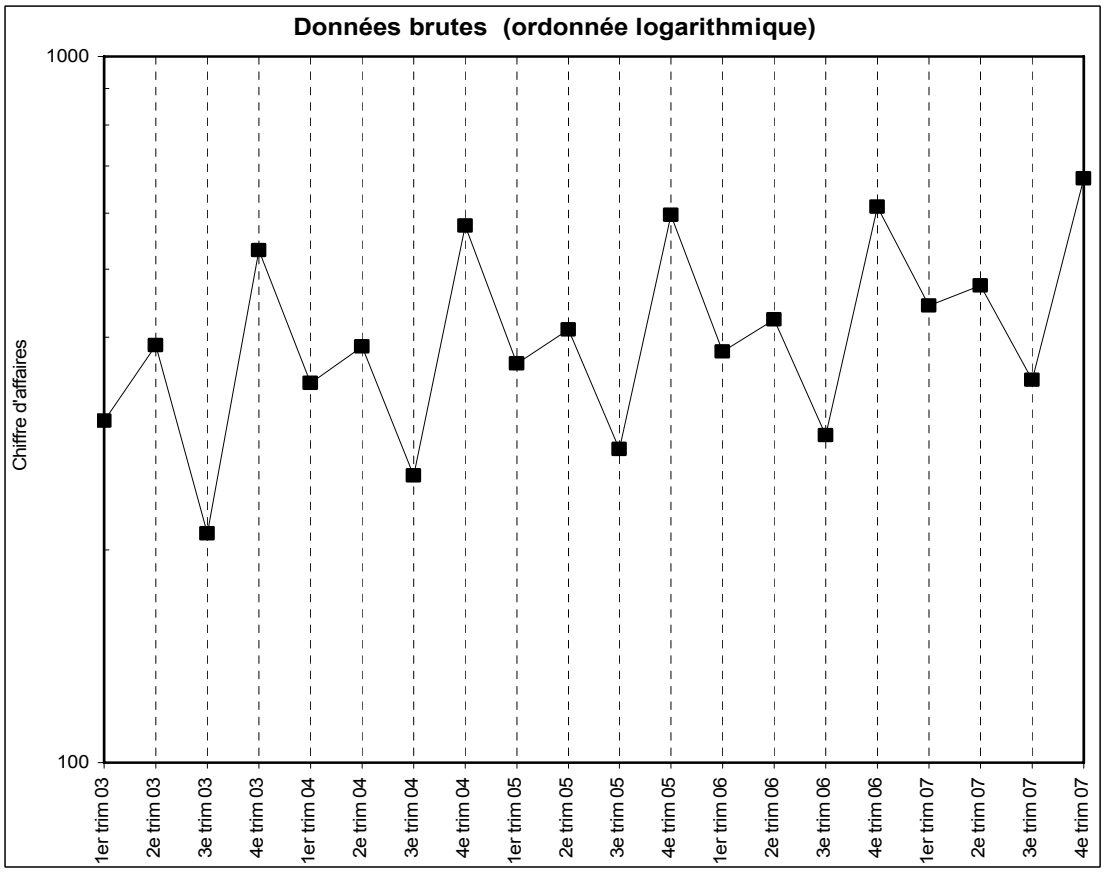
**Etape 2 : choix du schéma de composition**

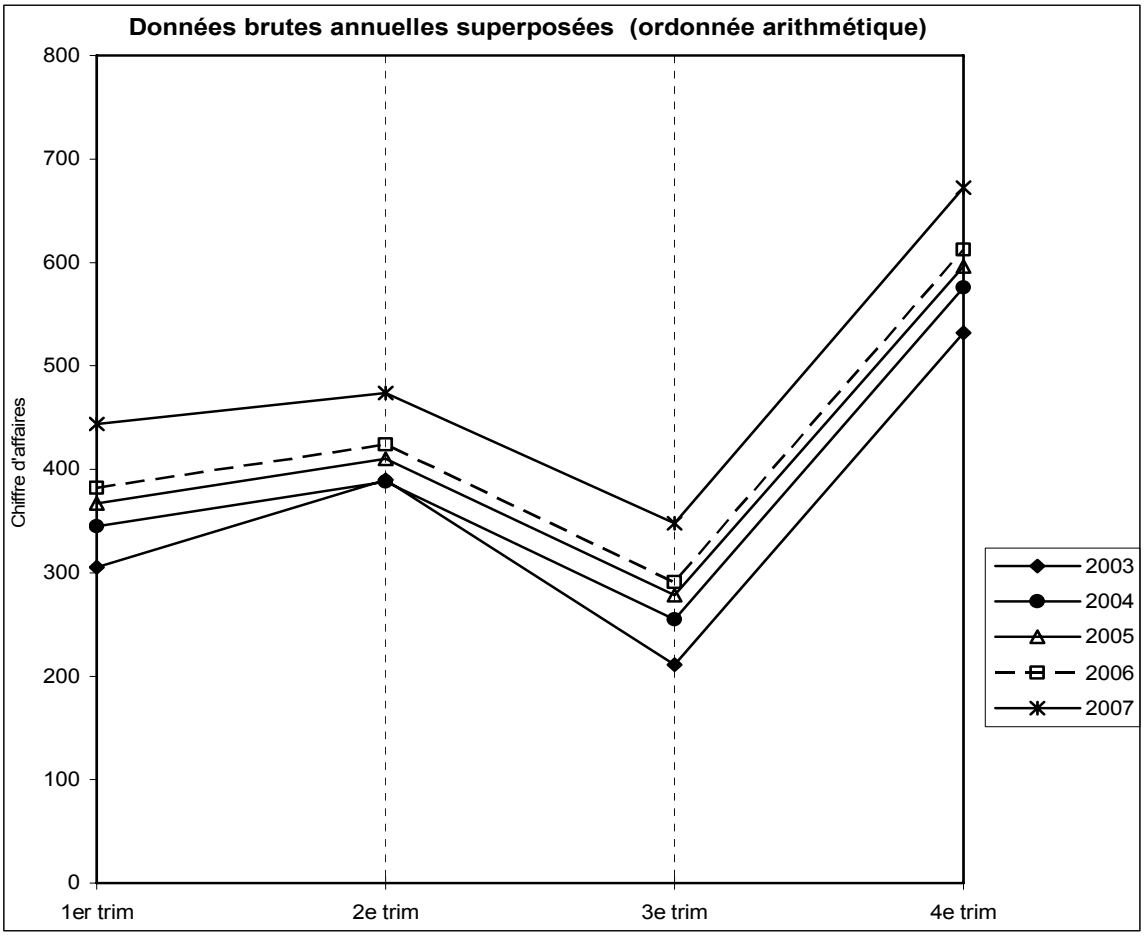
La représentation graphique des données brutes met en évidence des mouvements de même amplitude à des périodes (trimestres) identiques d'une année sur l'autre : baisse au 3ème trimestre et hausse au 4ème trimestre.
Elle ne permet pas une analyse satisfaisante de la tendance à long terme de la série. La correction des variations saisonnières apparaît donc indispensable.
Il est possible de recourir à une représentation en coordonnées polaires ou en données annuelles superposées.

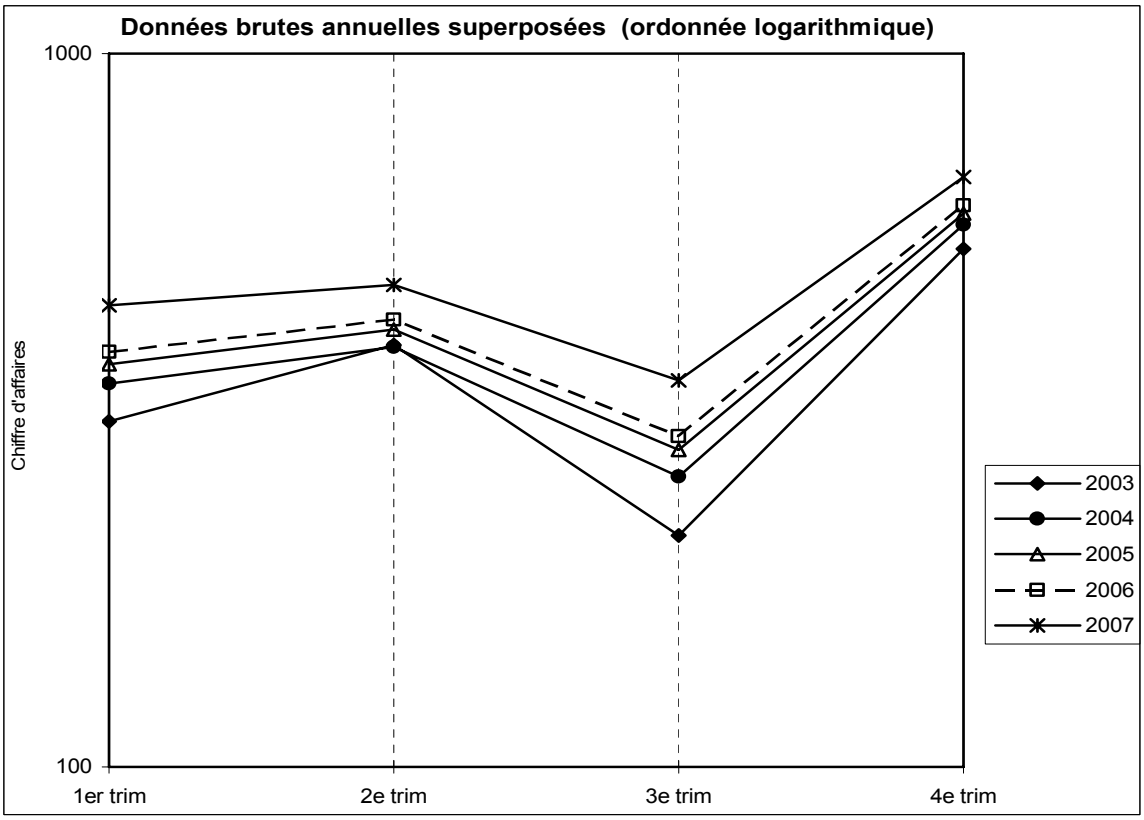
Les courbes en données superposées font apparaître un mouvement saisonnier dont l'amplitude est constante avec le niveau de la série. Il convient de noter que le niveau du C.A. du 2ème trimestre de 2003 est relativement plus élevé que lors des autres années. On pourrait éventuellement corriger cette donnée, pour le calcul du trend et des variations saisonnières (variation accidentelle exceptionnelle).

Le calcul des écarts-types annuels corrobore l'impression donnée par le graphique ci-dessus. On utilisera donc ici un schéma de composition additif.


Données brutes (ordonnée arithmétique)
Chiffre d'affaires
700
650
600
550
500
450
400
350
300
250
200
1er trim 03
2e trim 03
3e trim 03
4e trim 03
1er trim 04
2e trim 04
3e trim 04
4e trim 04
1er trim 05
2e trim 05
3e trim 05
4e trim 05
1er trim 06
2e trim 06
3e trim 06
4e trim 06
1er trim 07
2e trim 07
3e trim 07
4e trim 07


Données brutes (ordonnée logarithmique)
Chiffre d'affaires
1000
100
1er trim 03
2e trim 03
3e trim 03
4e trim 03
1er trim 04
2e trim 04
3e trim 04
4e trim 04
1er trim 05
2e trim 05
3e trim 05
4e trim 05
1er trim 06
2e trim 06
3e trim 06
4e trim 06
1er trim 07
2e trim 07
3e trim 07
4e trim 07


Données brutes annuelles superposées (ordonnée arithmétique)
Chiffre d'affaires
800
700
600
500
400
300
200
100
0
1er trim
2e trim
3e trim
4e trim
2003
2004
2005
2006
2007


Données brutes annuelles superposées (ordonnée logarithmique)
Chiffre d'affaires
1000
100
1er trim
2e trim
3e trim
4e trim
2003
2004
2005
2006
2007

**Etape 3 : estimation du trend**

**Calcul des moyennes mobiles MMij (MM4)**

| | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|
| **1er trim** | | 374,5 | 404,9 | 421,6 | 462,4 |
| **2e trim** | | 385,5 | 410,3 | 425,3 | 477,0 |
| **3e trim** | 364,5 | 393,8 | 414,6 | 435,0 | |
| **4e trim** | 369,3 | 399,3 | 418,3 | 449,0 | |

**Etape 4 : calcul des variations saisonnières**
**sij = yij - MMij**

| | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|
| **1er trim** | | -29,5 | -37,9 | -39,6 | -18,4 |
| **2e trim** | | 2,5 | -0,3 | -1,3 | -3,0 |
| **3e trim** | -153,5 | -138,8 | -136,6 | -144,0 | |
| **4e trim** | 162,8 | 176,8 | 177,8 | 163,0 | |

**Etape 5 : calcul des coefficients saisonniers**

| | **s' j** Méd | **sbar j** Moy |
|---|---|---|
| 1er trim | -33,7 | -31,3 |
| 2e trim | -0,8 | -0,5 |
| 3e trim | -141,4 | -143,2 |
| 4e trim | 169,9 | 170,1 |
| **Somme** | -5,9 | -5,0 |
| **Somme / 4** | -1,5 | -1,3 |

Le calcul des coefficients saisonniers est fait sur la base de la médiane (s'j), puis de la moyenne (sbar j) . Comme le principe de conservation des aires nécessite une somme des s'j ou des sbar j nulle, il faut corriger ces coefficients.

**Méthode 1 :** on divise par 4 (trimestres) la somme trouvée (-5,9) ou (-5,0) et l'on retranche le résultat à chaque coefficient.

**Méthode 2 :** on peut utiliser une méthode plus précise qui consiste à calculer un écart-type pour chaque trimestre, sur les variations saisonnières, de façon à répartir l'écart à zéro de la somme s'j ou sbar j au prorata de la dispersion observée pour chaque trimestre.

**Etape 6 : calcul des coefficients saisonniers médians corrigés s'j corr.**
**Méthode 1 : équirépartition de la différence à zéro entre les coefficients.**

| | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|
| **1er trim** | **-32,2** | -32,2 | -32,2 | -32,2 | -32,2 |
| **2e trim** | **0,7** | 0,7 | 0,7 | 0,7 | 0,7 |
| **3e trim** | **-139,9** | -139,9 | -139,9 | -139,9 | -139,9 |
| **4e trim** | **171,4** | 171,4 | 171,4 | 171,4 | 171,4 |
| **Somme** | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |

**Etape 6 : calcul des coefficients saisonniers moyens corrigés sbarj corr.**
**Méthode 1 : équirépartition de la différence à zéro entre les coefficients.**

| | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|
| **1er trim** | **-30,1** | -30,1 | -30,1 | -30,1 | -30,1 |
| **2e trim** | **0,8** | 0,8 | 0,8 | 0,8 | 0,8 |
| **3e trim** | **-142,0** | -142,0 | -142,0 | -142,0 | -142,0 |
| **4e trim** | **171,3** | 171,3 | 171,3 | 171,3 | 171,3 |
| **Somme** | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |

**Etape 6 : calcul des coefficients saisonniers médians corrigés s'j corr.**
**Méthode 2 : répartition de la différence à zéro entre les coefficients, au prorata de la dispersion observée chaque trimestre.**

| | Ectyp (s ij) | Clé de répart. | s' j<br>s j corr. |
|---|---|---|---|
| **1er trim** | 8,4 | 34,9% | **-31,6** |
| **2e trim** | 2,0 | 8,3% | **-0,3** |
| **3e trim** | 6,5 | 27,0% | **-139,8** |
| **4e trim** | 7,2 | 29,8% | **171,6** |
| **Total** | 24,1 | 100,0% | 0,0 |

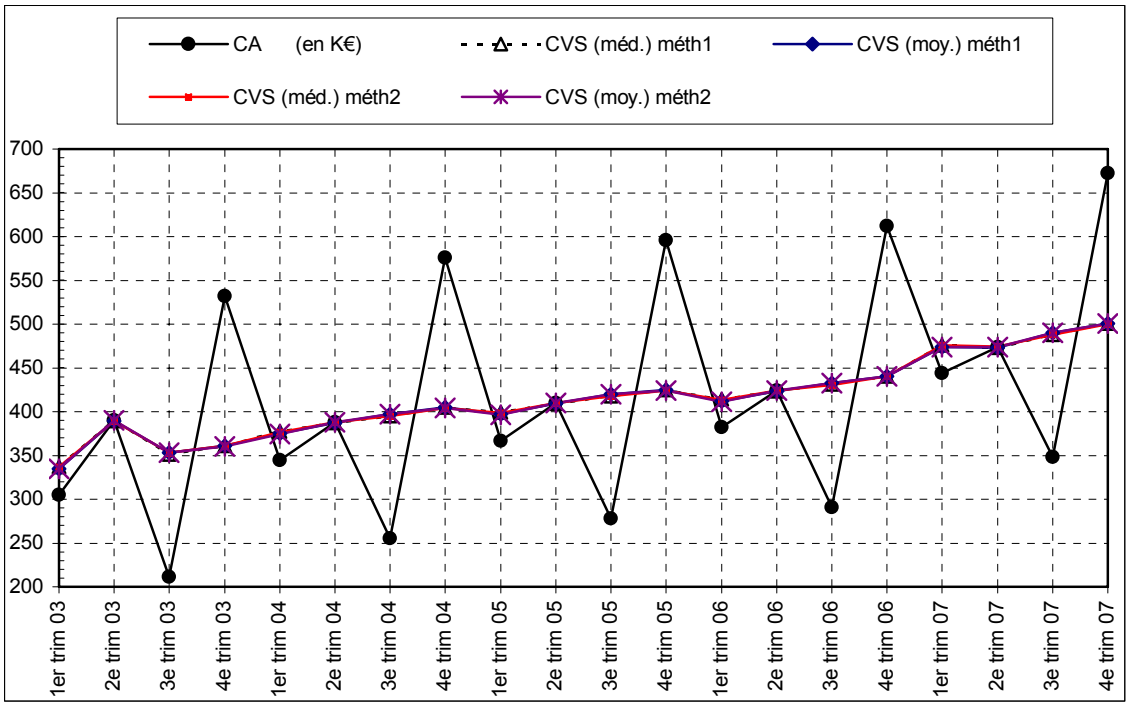
**Etape 6 : calcul des coefficients saisonniers moyens corrigés sbarj corr.**
**Méthode 2 : répartition de la différence à zéro entre les coefficients, au prorata de la dispersion observée chaque trimestre.**

| | Ectyp (s ij) | Clé de répart. | sbar j<br>s j corr. |
|---|---|---|---|
| **1er trim** | 8,4 | 34,9% | **-29,6** |
| **2e trim** | 2,0 | 8,3% | **-0,1** |
| **3e trim** | 6,5 | 27,0% | **-141,9** |
| **4e trim** | 7,2 | 29,8% | **171,6** |
| **Total** | 24,1 | 100,0% | 0,0 |

**Etape 7 : calcul de la série corrigée des variations saisonnières (CVS)**

Les données CVS sont obtenues en retranchant aux données brutes, avant correction des variations accidentelles (le cas échéant), le coefficient saisonnier du trimestre correspondant.

| | CA (en K€) | CVS (méd.) méth1 | CVS (moy.) méth1 | CVS (méd.) méth2 | CVS (moy.) méth2 |
|---|---|---|---|---|---|
| **1er trim 03** | **305** | 337,2 | 335,1 | 336,6 | 334,6 |
| **2e trim 03** | **390** | 389,3 | 389,3 | 389,3 | 390,1 |
| **3e trim 03** | **211** | 350,9 | 353,0 | 353,0 | 352,9 |
| **4e trim 03** | **532** | 360,6 | 360,7 | 360,7 | 360,4 |
| **1er trim 04** | **345** | 377,2 | 375,1 | 376,6 | 374,6 |
| **2e trim 04** | **388** | 387,3 | 387,3 | 388,3 | 388,1 |
| **3e trim 04** | **255** | 394,9 | 397,0 | 394,8 | 396,9 |
| **4e trim 04** | **576** | 404,6 | 404,7 | 404,4 | 404,4 |
| **1er trim 05** | **367** | 399,2 | 397,1 | 398,6 | 396,6 |
| **2e trim 05** | **410** | 409,3 | 409,3 | 410,3 | 410,1 |
| **3e trim 05** | **278** | 417,9 | 420,0 | 417,8 | 419,9 |
| **4e trim 05** | **596** | 424,6 | 424,7 | 424,4 | 424,4 |
| **1er trim 06** | **382** | 414,2 | 412,1 | 413,6 | 411,6 |
| **2e trim 06** | **424** | 423,3 | 423,3 | 424,3 | 424,1 |
| **3e trim 06** | **291** | 430,9 | 433,0 | 430,8 | 432,9 |
| **4e trim 06** | **612** | 440,6 | 440,7 | 440,4 | 440,4 |
| **1er trim 07** | **444** | 476,2 | 474,1 | 475,6 | 473,6 |
| **2e trim 07** | **474** | 473,3 | 473,3 | 474,3 | 474,1 |
| **3e trim 07** | **348** | 487,9 | 490,0 | 487,8 | 489,9 |
| **4e trim 07** | **672** | 500,6 | 500,7 | 500,4 | 500,4 |



**Complément de l'exercice 43 : méthode analytique**

**Estimation de la composante extra-saisonnière**

| | CA (en K€) | CVS (méd.) | CVS (moy.) |
|---|---|---|---|
| **1er trim 03** | 305 | 337,2 | 336,2 |
| **2e trim 03** | 390 | 387,7 | 387,8 |
| **3e trim 03** | 211 | 351,9 | 352,5 |
| **4e trim 03** | 532 | 361,3 | 361,5 |
| **1er trim 04** | 345 | 377,2 | 376,2 |
| **2e trim 04** | 388 | 385,7 | 385,8 |
| **3e trim 04** | 255 | 395,9 | 396,5 |
| **4e trim 04** | 576 | 405,3 | 405,5 |
| **1er trim 05** | 367 | 399,2 | 398,2 |
| **2e trim 05** | 410 | 407,7 | 407,8 |
| **3e trim 05** | 278 | 418,9 | 419,5 |
| **4e trim 05** | 596 | 425,3 | 425,5 |
| **1er trim 06** | 382 | 414,2 | 413,2 |
| **2e trim 06** | 424 | 421,7 | 421,8 |
| **3e trim 06** | 291 | 431,9 | 432,5 |
| **4e trim 06** | 612 | 441,3 | 441,5 |
| **1er trim 07** | 444 | 476,2 | 475,2 |
| **2e trim 07** | 474 | 471,7 | 471,8 |
| **3e trim 07** | 348 | 488,9 | 489,5 |
| **4e trim 07** | 672 | 501,3 | 501,5 |

**Données brutes $y_{ij}$**

| | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|
| **1er trim** | 305 | 345 | 367 | 382 | 444 |
| **2e trim** | 390 | 388 | 410 | 424 | 474 |
| **3e trim** | 211 | 255 | 278 | 291 | 348 |
| **4e trim** | 532 | 576 | 596 | 612 | 672 |

| Modèle linéaire | | | | | | |
|---|---|---|---|---|---|---|
| **Tri ($x_t$)** | **CA ($y_t$)** | $x_t^2$ | $y_t^2$ | $x_t\ y_t$ | $y^*_t$ | |
| **1** | 305 | 1 | 93025 | 305 | 330,1 | |
| **2** | 390 | 4 | 152100 | 780 | 339,0 | |
| **3** | 211 | 9 | 44521 | 633 | 347,9 | |
| **4** | 532 | 16 | 283024 | 2128 | 356,9 | |
| **5** | 345 | 25 | 119025 | 1725 | 365,8 | |
| **6** | 388 | 36 | 150544 | 2328 | 374,8 | **Nb val = 20** |
| **7** | 255 | 49 | 65025 | 1785 | 383,7 | |
| **8** | 576 | 64 | 331776 | 4608 | 392,6 | |
| **9** | 367 | 81 | 134689 | 3303 | 401,6 | |
| **10** | 410 | 100 | 168100 | 4100 | 410,5 | |
| **11** | 278 | 121 | 77284 | 3058 | 419,5 | |
| **12** | 596 | 144 | 355216 | 7152 | 428,4 | |
| **13** | 382 | 169 | 145924 | 4966 | 437,4 | |
| **14** | 424 | 196 | 179776 | 5936 | 446,3 | |
| **15** | 291 | 225 | 84681 | 4365 | 455,2 | |
| **16** | 612 | 256 | 374544 | 9792 | 464,2 | |
| **17** | 444 | 289 | 197136 | 7548 | 473,1 | |
| **18** | 474 | 324 | 224676 | 8532 | 482,1 | |
| **19** | 348 | 361 | 121104 | 6612 | 491,0 | |
| **20** | 672 | 400 | 451584 | 13440 | 499,9 | |
| **210** | **8300** | **2870** | **3753754** | **93096** | | |

xbar = 10,50 ectyp x = 5,77

ybar = 415,00 ectyp y = 124,35

**rho xy = 0,41**

cov xy = 297,30

rho² xy = 0,17

a = 8,94 b = 321,12

**Modèle exponentiel**

| $\log (y_t)$ | | $\log (y_t)^2$ | $x_t \log (y_t)$ | $\log (y^*_t)$ | $y^*_t$ |
|---|---|---|---|---|---|
| 2,48430 | | 6,17175 | 2,48430 | **2,508** | **322,1** |
| 2,59106 | | 6,71362 | 5,18213 | 2,517 | 329,2 |
| 2,32428 | | 5,40229 | 6,97285 | 2,527 | 336,5 |
| 2,72591 | | 7,43059 | 10,90365 | 2,537 | 344,0 |
| 2,53782 | | 6,44053 | 12,68910 | 2,546 | 351,6 |
| 2,58883 | | 6,70205 | 15,53299 | 2,556 | 359,4 |
| 2,40654 | | 5,79144 | 16,84578 | 2,565 | 367,4 |
| 2,76042 | | 7,61993 | 22,08338 | 2,575 | 375,5 |
| 2,56467 | | 6,57751 | 23,08199 | 2,584 | 383,8 |
| 2,61278 | | 6,82664 | 26,12784 | 2,594 | 392,3 |
| 2,44404 | | 5,97335 | 26,88449 | 2,603 | 401,0 |
| 2,77525 | | 7,70199 | 33,30296 | 2,613 | 409,9 |
| 2,58206 | | 6,66705 | 33,56682 | 2,622 | 419,0 |
| 2,62737 | | 6,90305 | 36,78312 | 2,632 | 428,3 |
| 2,46389 | | 6,07077 | 36,95839 | 2,641 | 437,8 |
| 2,78675 | | 7,76598 | 44,58802 | 2,651 | 447,5 |
| 2,64738 | | 7,00864 | 45,00551 | 2,660 | 457,4 |
| 2,67578 | | 7,15979 | 48,16401 | 2,670 | 467,5 |
| 2,54158 | | 6,45963 | 48,29001 | 2,679 | 477,9 |
| 2,82737 | | 7,99402 | 56,54739 | 2,689 | 488,5 |
| **51,968** | | **135,381** | **551,995** | | |

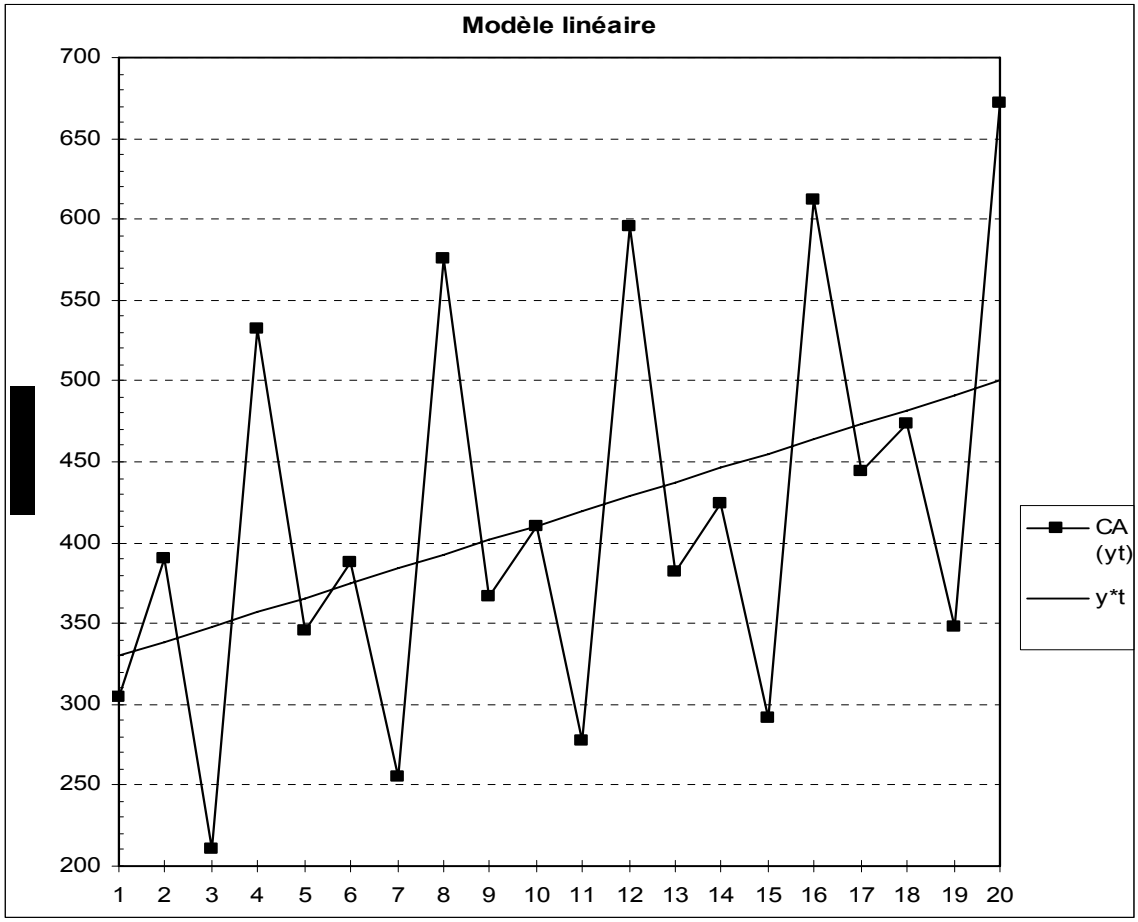
**Nb val = 20**

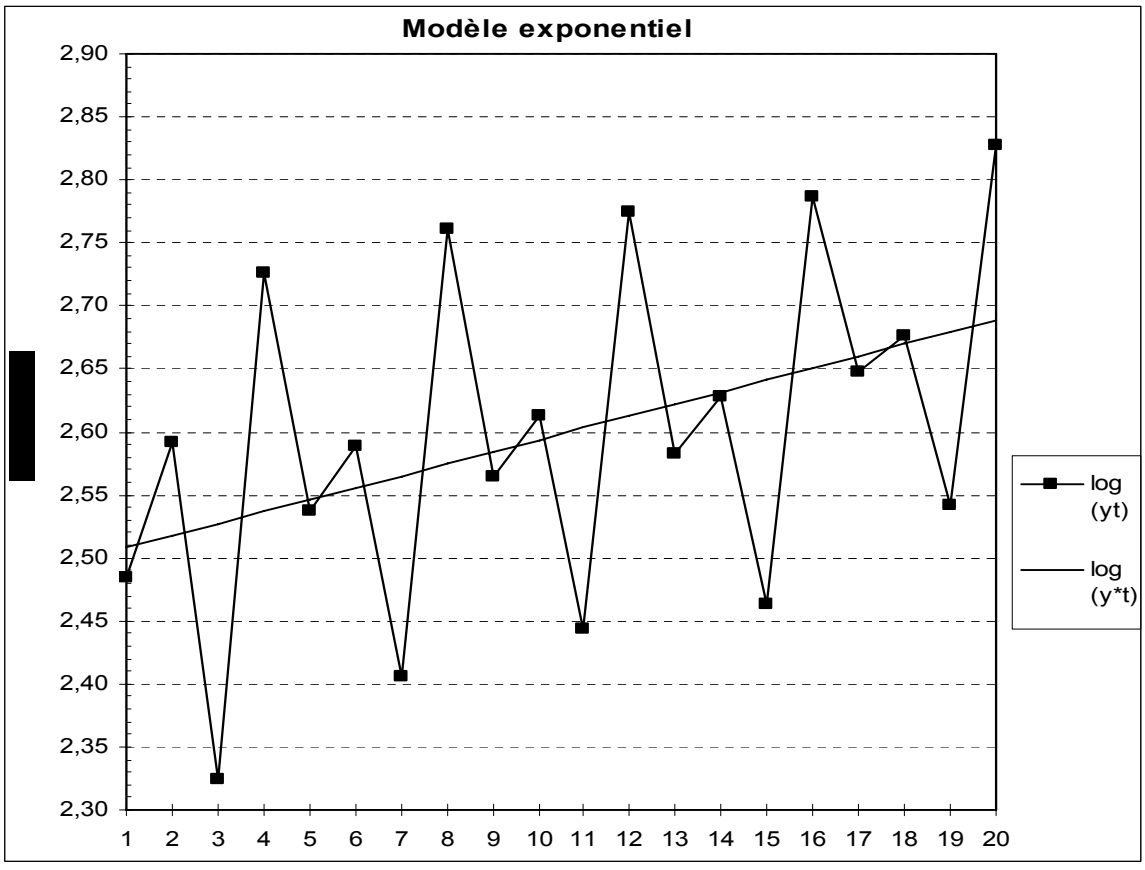
| | | | |
|---|---|---|---|
| xbar = | 10,50 | ectyp x = | 5,77 |
| log (ybar) = | 2,60 | ectyp logy = | 0,13 |
| | | **rho x logy =** | **0,42** |
| cov x logy = | 0,32 | | |
| | | rho² x logy = | 0,17 |
| a* = | 0,010 | b* = | 2,50 |

On a : log ft estimé = a*.t + b*

D'où : ft estimé = trend estimé = 10^(a*.t + b*) = 10^a*.t . 10^b*

Par exemple, pour le 1er trimestre 2003, on a :
log ft estimé = a*.t + b* = 2,508
D'où : ft estimé = trend estimé = 10^(a*.t + b*) = 10^2,508 = 322,1


Modèle linéaire
700
650
600
550
500
450
400
350
300
250
200
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
CA (yt)
y*t


Modèle exponentiel
2,90
2,85
2,80
2,75
2,70
2,65
2,60
2,55
2,50
2,45
2,40
2,35
2,30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
log (yt)
log (y*t)

**Calcul des variations saisonnières sij = yij - y*ij**

**Coeff. saisonniers s' j sbar j tr**

| | 2003 | 2004 | 2005 | 2006 | 2007 | Méd | Moy tr |
|---|---|---|---|---|---|---|---|
| **1er trim** | -17,1 | -6,6 | -16,8 | -37,0 | -13,4 | -16,8 | -15,8 |
| **2e trim** | 60,8 | 28,6 | 17,7 | -4,3 | 6,5 | 17,7 | 17,6 |
| **3e trim** | -125,5 | -112,4 | -123,0 | -146,8 | -129,9 | -125,5 | -126,1 |
| **4e trim** | 188,0 | 200,5 | 186,1 | 164,5 | 183,5 | 186,1 | 185,9 |
| | | | | **Somme** | | 61,4 | 61,6 |
| | | | | **Somme / 4** | | 15,4 | 15,4 |

Les différences à zéro sont significatives : les coefficients saisonniers doivent être corrigés.

**Calcul des coefficients saisonniers médians corrigés s'j corr.**

| | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|
| **1er trim** | -32,2 | -32,2 | -32,2 | -32,2 | -32,2 |
| **2e trim** | 2,3 | 2,3 | 2,3 | 2,3 | 2,3 |
| **3e trim** | -140,9 | -140,9 | -140,9 | -140,9 | -140,9 |
| **4e trim** | 170,7 | 170,7 | 170,7 | 170,7 | 170,7 |
| **Som** | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |

| | |
|---|---|
| **1er tri** | **-32,2** |
| **2e tri** | **2,3** |
| **3e tri** | **-140,9** |
| **4e tri** | **170,7** |

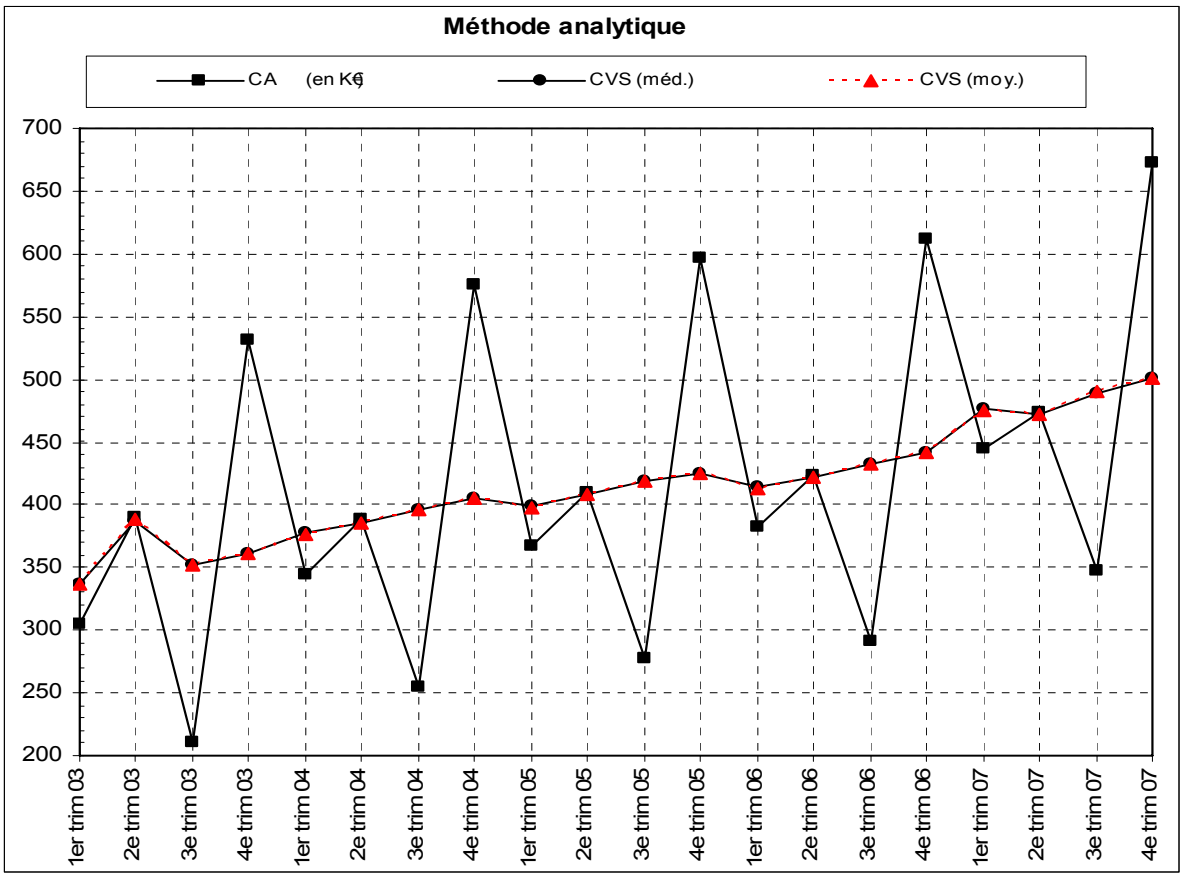
**s'j corr.**

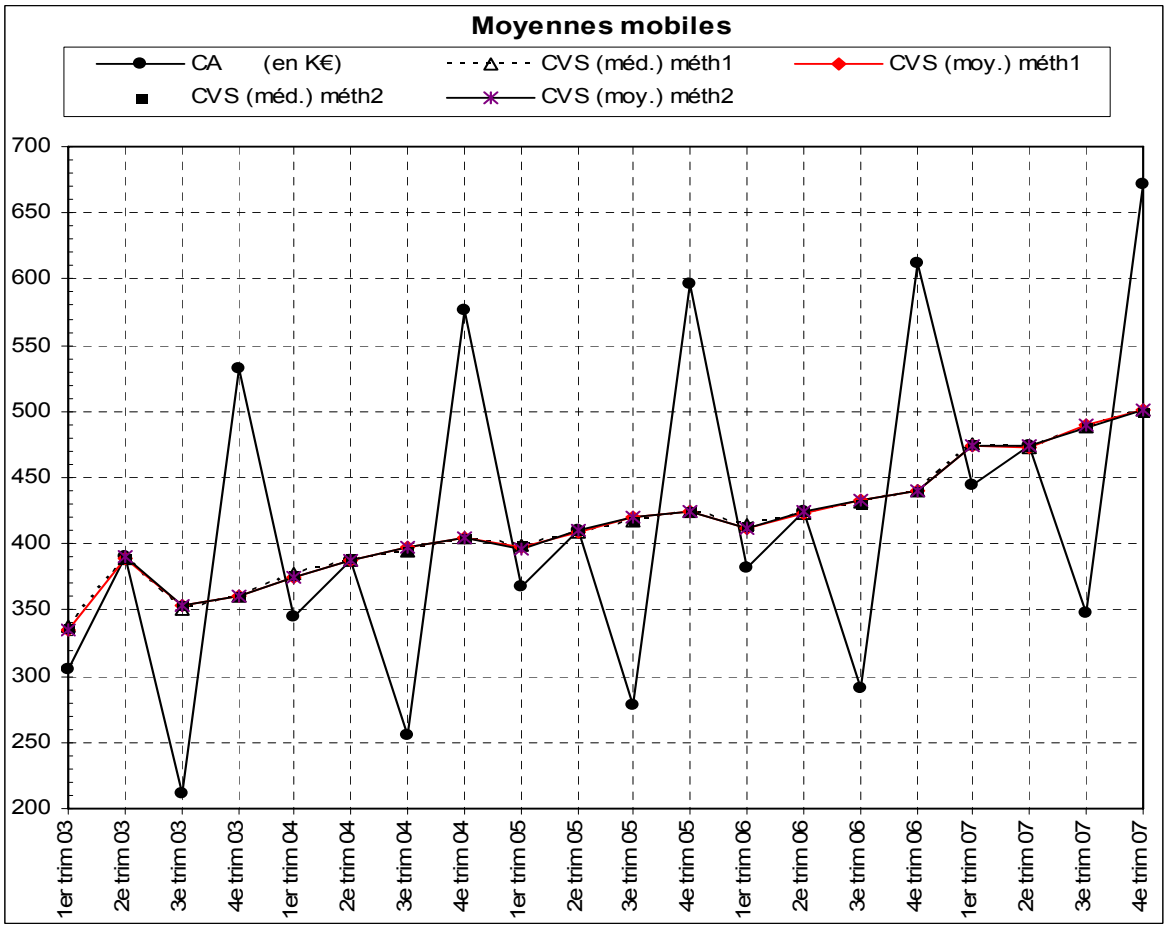
**Calcul des coefficients saisonniers moyens tr. corrigés sbarj tr. corr.**

| | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|
| **1er trim** | -31,2 | -31,2 | -31,2 | -31,2 | -31,2 |
| **2e trim** | 2,2 | 2,2 | 2,2 | 2,2 | 2,2 |
| **3e trim** | -141,5 | -141,5 | -141,5 | -141,5 | -141,5 |
| **4e trim** | 170,5 | 170,5 | 170,5 | 170,5 | 170,5 |
| **Som** | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |

| | |
|---|---|
| **1er trim** | **-31,2** |
| **2e trim** | **2,2** |
| **3e trim** | **-141,5** |
| **4e trim** | **170,5** |

**sbarj tr. corr.**


Méthode analytique
CA (en K€)
CVS (méd.)
CVS (moy.)
700
650
600
550
500
450
400
350
300
250
200
1er trim 03
2e trim 03
3e trim 03
4e trim 03
1er trim 04
2e trim 04
3e trim 04
4e trim 04
1er trim 05
2e trim 05
3e trim 05
4e trim 05
1er trim 06
2e trim 06
3e trim 06
4e trim 06
1er trim 07
2e trim 07
3e trim 07
4e trim 07


Moyennes mobiles
CA (en K€)
CVS (méd.) méth1
CVS (moy.) méth1
CVS (méd.) méth2
CVS (moy.) méth2
700
650
600
550
500
450
400
350
300
250
200
1er trim 03
2e trim 03
3e trim 03
4e trim 03
1er trim 04
2e trim 04
3e trim 04
4e trim 04
1er trim 05
2e trim 05
3e trim 05
4e trim 05
1er trim 06
2e trim 06
3e trim 06
4e trim 06
1er trim 07
2e trim 07
3e trim 07
4e trim 07

**Exemple terminal** : représentations graphiques mensuelles

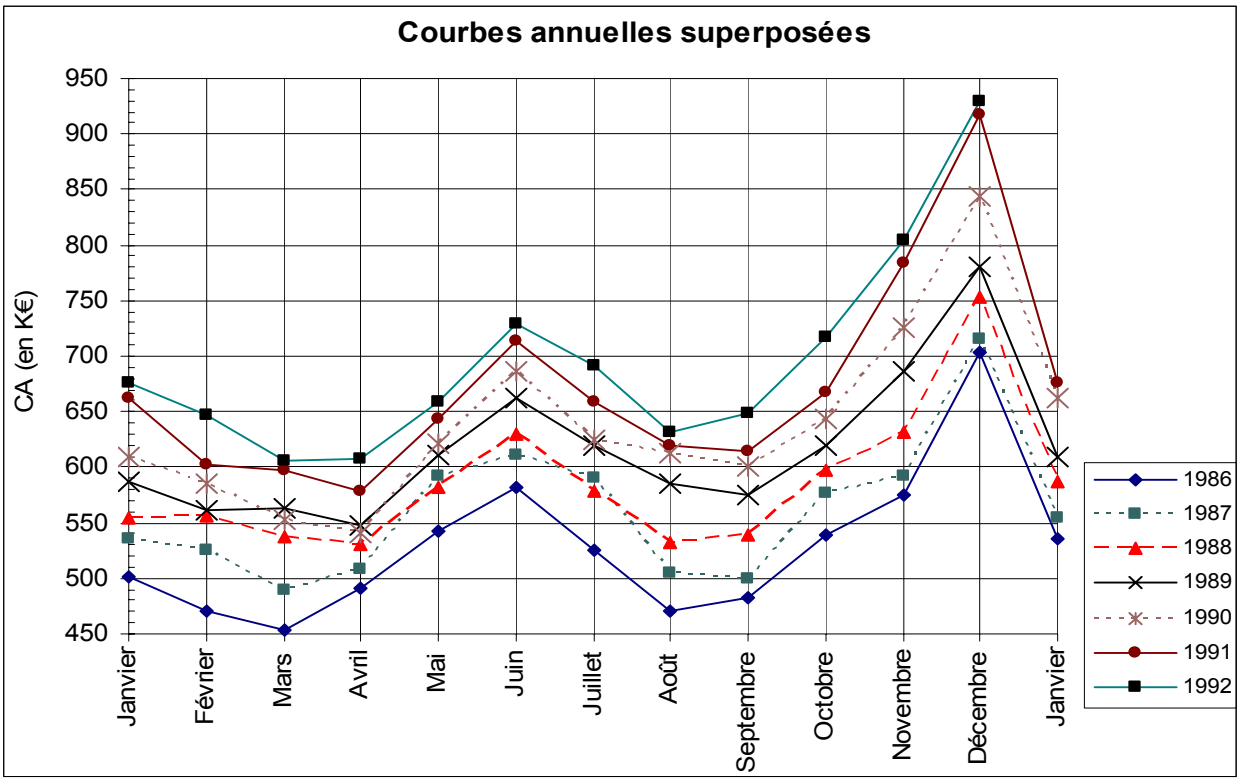
Chiffre d'affaires d'un supermarché (en K€)

| Mois \ Ans | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 | 1992 |
|---|---|---|---|---|---|---|---|
| **Janvier** | 502 | 536 | 554 | 587 | 610 | 662 | 676 |
| **Février** | 470 | 525 | 556 | 561 | 585 | 603 | 647 |
| **Mars** | 454 | 489 | 538 | 563 | 552 | 598 | 606 |
| **Avril** | 491 | 509 | 530 | 547 | 541 | 579 | 607 |
| **Mai** | 542 | 592 | 581 | 611 | 622 | 643 | 659 |
| **Juin** | 581 | 611 | 630 | 663 | 686 | 714 | 729 |
| **Juillet** | 525 | 591 | 579 | 619 | 625 | 659 | 692 |
| **Août** | 470 | 505 | 533 | 586 | 612 | 619 | 632 |
| **Septembre** | 483 | 500 | 539 | 575 | 601 | 614 | 648 |
| **Octobre** | 539 | 577 | 597 | 620 | 643 | 667 | 717 |
| **Novembre** | 575 | 592 | 631 | 687 | 726 | 784 | 805 |
| **Décembre** | 704 | 716 | 753 | 780 | 843 | 917 | 929 |
| **Janvier** | 536 | 554 | 587 | 610 | 662 | 676 | |

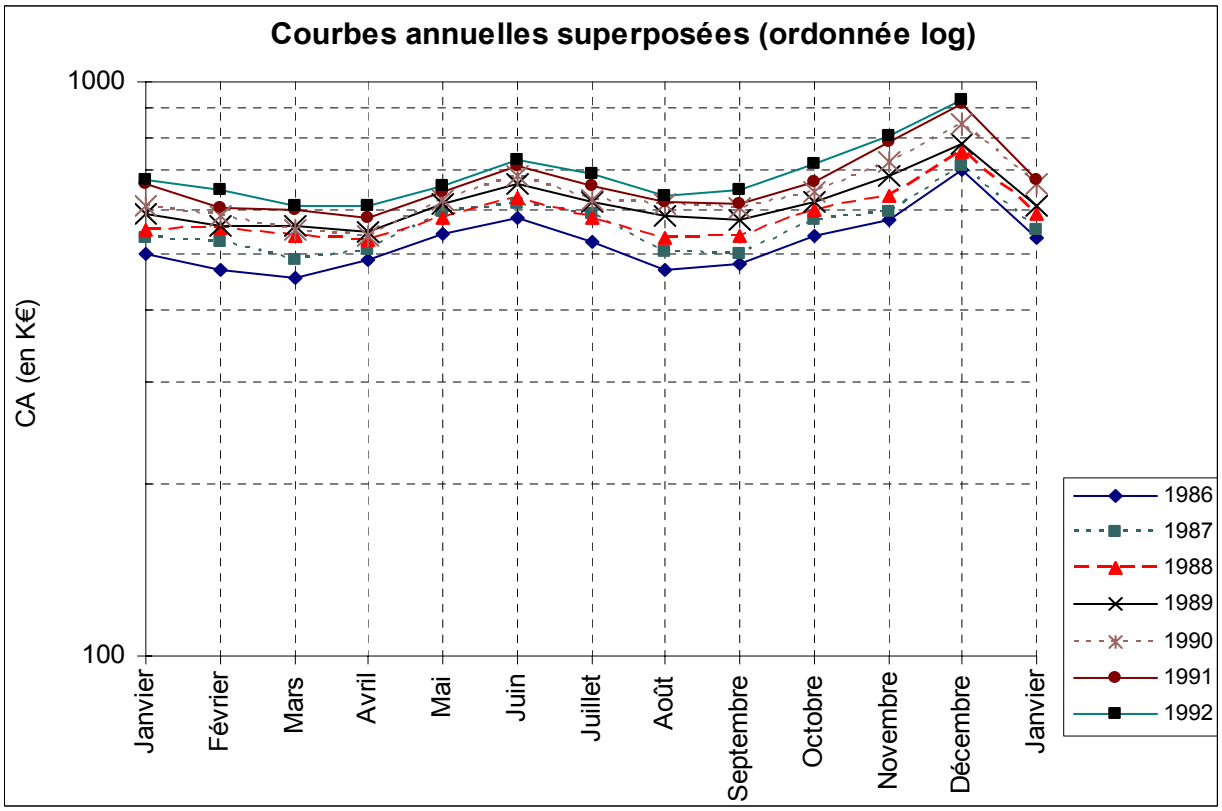
| | |
|---|---|
| janv-86 | 502 |
| févr-86 | 470 |
| mars-86 | 454 |
| avr-86 | 491 |
| mai-86 | 542 |
| juin-86 | 581 |
| juil-86 | 525 |
| août-86 | 470 |
| sept-86 | 483 |
| oct-86 | 539 |
| nov-86 | 575 |
| déc-86 | 704 |
| janv-87 | 536 |
| févr-87 | 525 |
| mars-87 | 489 |
| avr-87 | 509 |
| mai-87 | 592 |
| juin-87 | 611 |
| juil-87 | 591 |
| août-87 | 505 |
| sept-87 | 500 |
| oct-87 | 577 |
| nov-87 | 592 |
| déc-87 | 716 |
| janv-88 | 554 |
| févr-88 | 556 |
| mars-88 | 538 |
| avr-88 | 530 |
| mai-88 | 581 |
| juin-88 | 630 |
| juil-88 | 579 |
| août-88 | 533 |
| sept-88 | 539 |
| oct-88 | 597 |
| nov-88 | 631 |
| déc-88 | 753 |
| janv-89 | 587 |
| févr-89 | 561 |
| mars-89 | 563 |
| avr-89 | 547 |
| mai-89 | 611 |
| juin-89 | 663 |
| juil-89 | 619 |
| août-89 | 586 |
| sept-89 | 575 |
| oct-89 | 620 |
| nov-89 | 687 |
| déc-89 | 780 |

| | |
|---|---|
| janv-90 | 610 |
| févr-90 | 585 |
| mars-90 | 552 |
| avr-90 | 541 |
| mai-90 | 622 |
| juin-90 | 686 |
| juil-90 | 625 |
| août-90 | 612 |
| sept-90 | 601 |
| oct-90 | 643 |
| nov-90 | 726 |
| déc-90 | 843 |
| janv-91 | 662 |
| févr-91 | 603 |
| mars-91 | 598 |
| avr-91 | 579 |
| mai-91 | 643 |
| juin-91 | 714 |
| juil-91 | 659 |
| août-91 | 619 |
| sept-91 | 614 |
| oct-91 | 667 |
| nov-91 | 784 |
| déc-91 | 917 |
| janv-92 | 676 |
| févr-92 | 647 |
| mars-92 | 606 |
| avr-92 | 607 |
| mai-92 | 659 |
| juin-92 | 729 |
| juil-92 | 692 |
| août-92 | 632 |
| sept-92 | 648 |
| oct-92 | 717 |
| nov-92 | 805 |
| déc-92 | 929 |

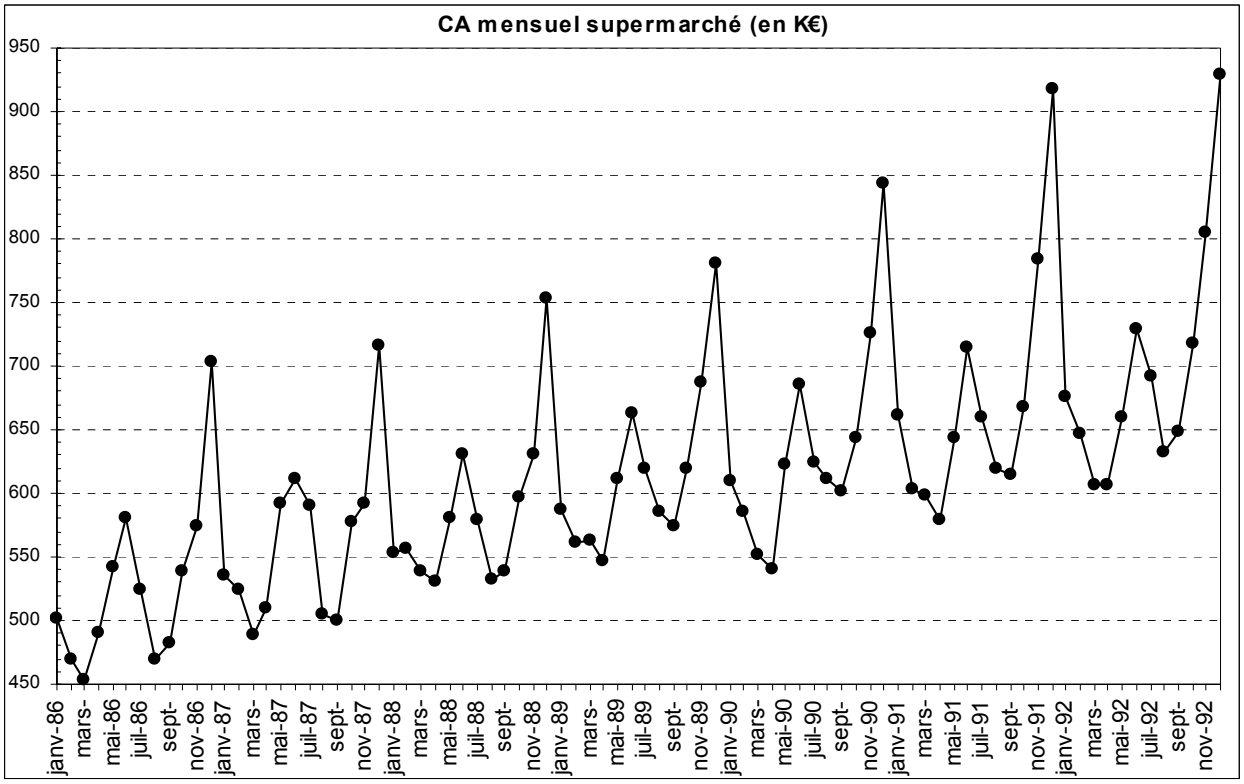
## Courbes annuelles superposées



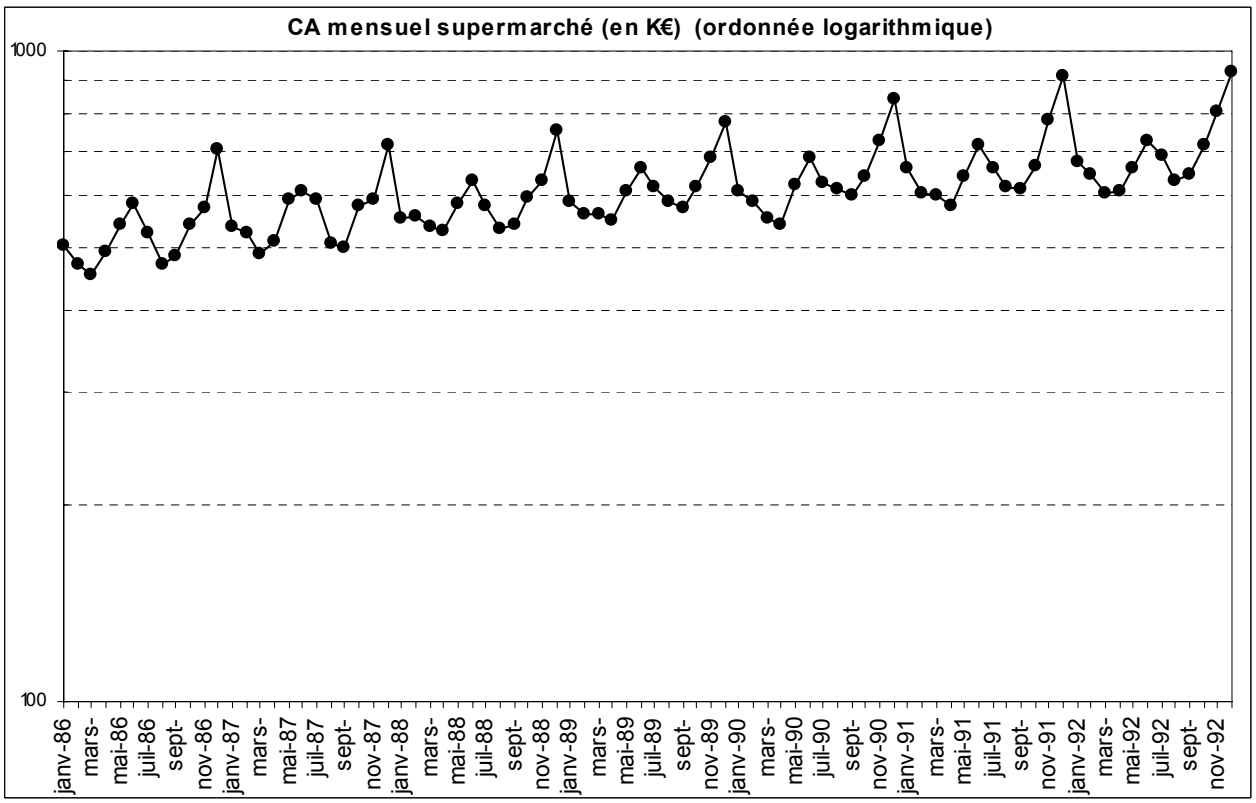

Les amplitudes des fluctuations saisonnières tendent à augmenter en même temps que l'accroissement de la tendance générale à long terme, surtout pour décembre. Cela justifie l'utilisation d'un schéma multiplicatif cf. graphique à ordonnée logarithmique ci-dessous.



## Série chronologique brute

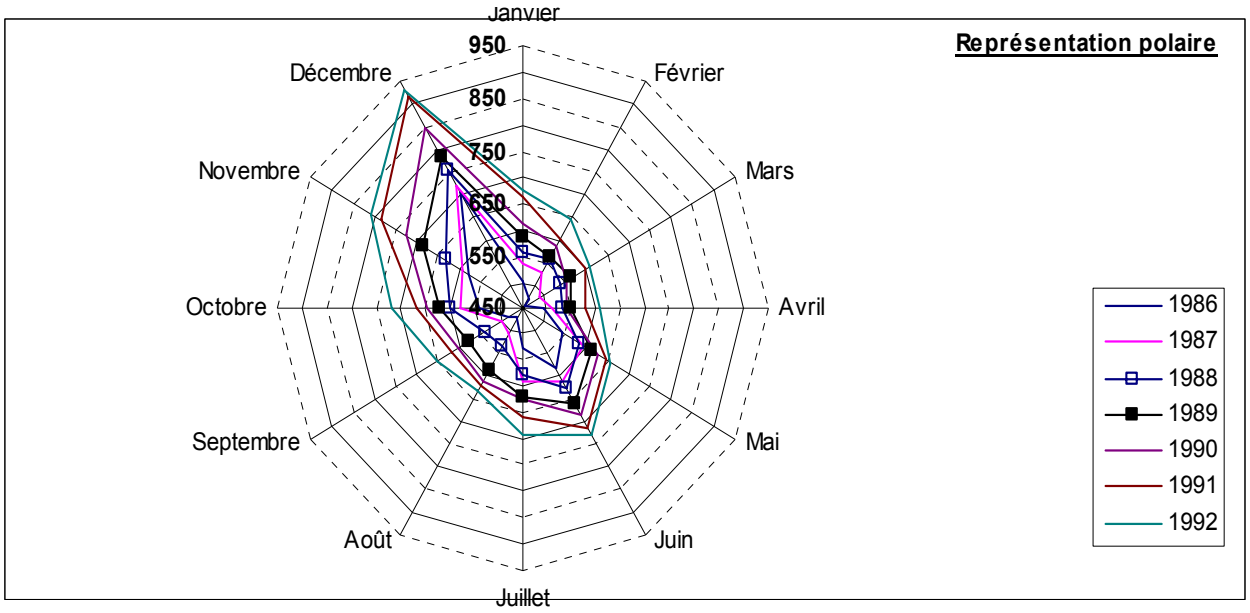


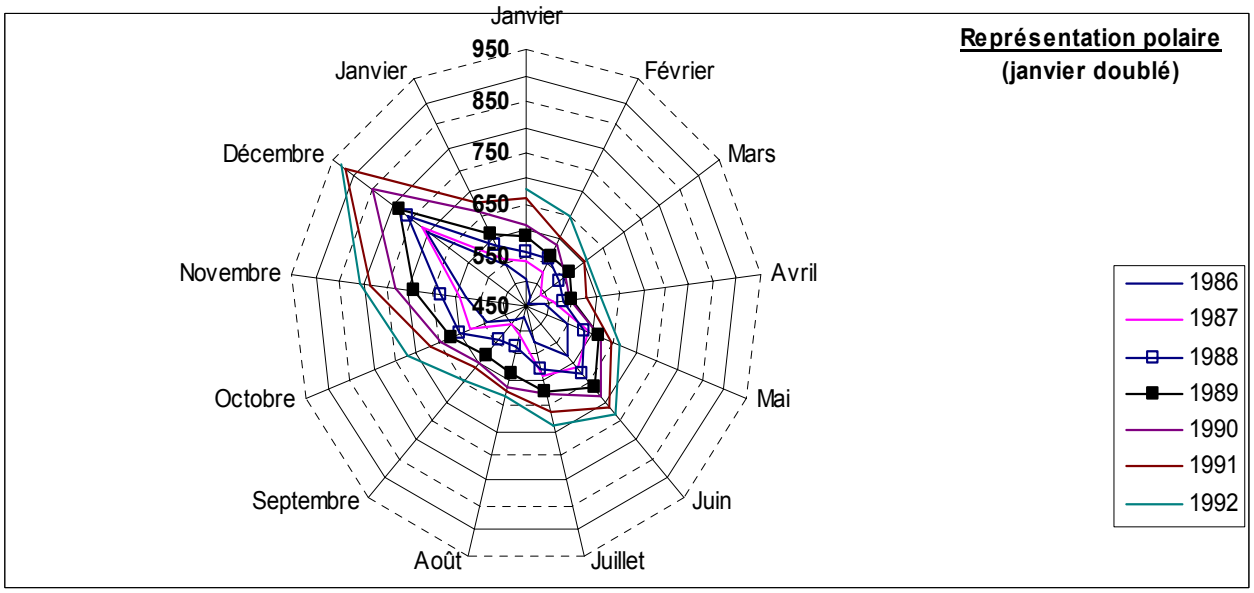

Comme sur le graphique des données annuelles superposées, on constate un accroissement des amplitudes des fluctuations saisonnières au fil du temps, au fur et à mesure où la tendance générale s'élève. cf. graphique à ordonnée logarithmique ci-dessous.



## Graphiques polaires

Graphiques polaires avec ordonnée arithmétique



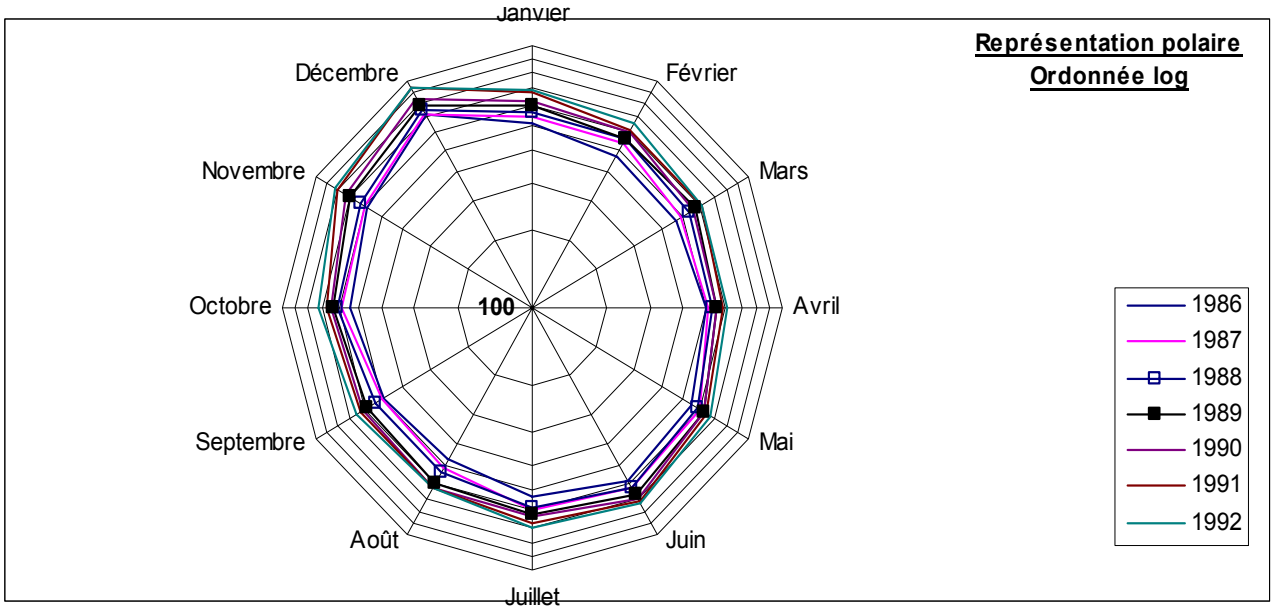



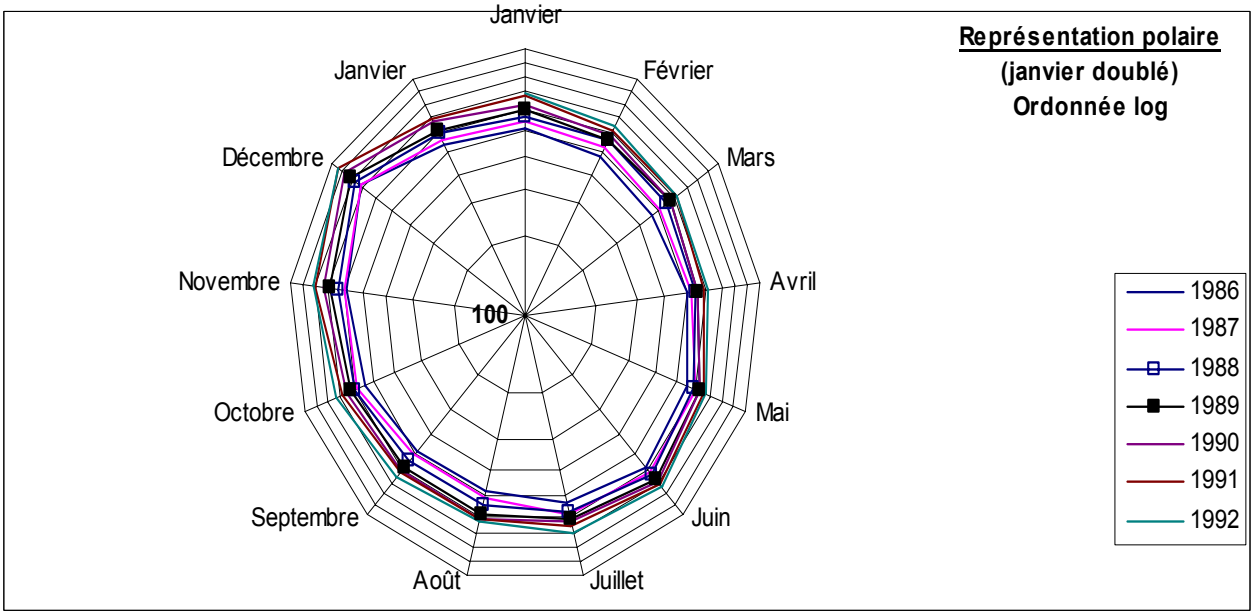

Difficulté sous Excel : les séries annuelles superposées ne sont pas reliées entre elles, dans ce type de représentation graphique. Sur le premier graphique, pour chaque année, on a un polygone fermé sur lui-même.

Pour éviter ce problème, sur le deuxième graphique, on utilise un subterfuge qui consiste à créer 13 axes, dont deux mois de janvier.

On peut réaliser le même constat qu'avec les autres types de représentation graphique.

Graphiques polaires avec ordonnée logarithmique



## 5. Quelques compléments

### 51. Évolution d'une grandeur en moyenne annuelle et en glissement

En général, on raisonne sur des indices plutôt que sur les données brutes.

**Principe général**

Supposons qu'on a affaire à une série mensuelle portant sur plusieurs années (5 ans, par ex.).

On pose également qu'il s'agit d'une variable de stock (le repérage de la valeur de la grandeur étudiée est donc effectué le dernier jour de chaque mois). Notamment, en décembre, la donnée vaut pour la date du 31 décembre de l'année correspondante.

À cette date, chaque année, on peut calculer une valeur moyenne de la grandeur étudiée pour l'ensemble de l'année. Ainsi, sur une période de cinq ans, on peut déterminer 5 moyennes annuelles.

À partir de ces cinq valeurs, on peut calculer l'**évolution de la grandeur en moyenne annuelle**, en faisant le rapport de 2 moyennes annuelles successives :

**valeur moyenne année t / valeur moyenne année t - 1**

Mais on peut également calculer l'évolution de la grandeur en posant le rapport :

**valeur décembre année t / valeur décembre année t - 1**

Dans ce cas, on dit qu'on calcule l'**évolution de la grandeur en glissement**, et l'on peut envisager des calculs d'évolution en glissement d'un mois quelconque d'une année t au mois correspondant de l'année t + 1. Il y a donc 12 possibilités d'effectuer des calculs d'évolutions en glissement (un par mois de l'année).

La mesure de l'évolution d'une grandeur en moyenne annuelle fait intervenir les informations dont on dispose sur l'ensemble de l'année précédente, c'est-à-dire le profil moyen de l'année t - 1. Les valeurs moyennes résument "l'histoire" de l'année t - 1 au moyen d'un nombre unique.

En comptabilité nationale, on utilise cette technique sur les indices de prix, de manière à déflater (c'est-à-dire annuler l'effet des hausses ou des baisses de prix sur les valeurs des grandeurs considérées) les séries chronologiques annuelles, relatives notamment aux salaires, à la consommation ou à l'investissement. Ainsi, on peut comparer ces différents flux annuels à ceux de l'année t - 1, sans perturbations occasionnées par les variations de prix.

Lorsqu'on effectue une mesure de l'évolution d'une grandeur en glissement, on ne prend pas en compte les informations qui caractérisent l'évolution d'une série sur l'ensemble d'une année. Cette technique de calcul est intéressante lorsqu'on mène des analyses conjoncturelles. En effet, dans ce cas on raisonne sur des taux de croissance instantanés. Cela permet d'obtenir des "photos" immédiates de l'état de l'actualité économique de la grandeur étudiée.

**Exemple**

Supposons que, dans un pays, les indices trimestriels de la branche d'activité "Biens de consommation" soient les suivants (base 100 en 1993) :

| | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 |
|---|---|---|---|---|---|---|
| **Trim 1** | | 101,8 | 106,3 | 108,5 | 105,2 | 106,8 |
| **Trim 2** | | 102,7 | 107,2 | 109,2 | 107,1 | 107,0 |
| **Trim 3** | | 103,5 | 106,8 | 105,8 | 106,8 | 103,8 |
| **Trim 4** | 103,2 | 105,6 | 107,8 | 105,1 | 105,3 | 103,7 |

Dans une première étape, on peut calculer des indices annuels moyens. Pour cela, on calcule la moyenne des quatre trimestres d'une même année. On obtient le tableau suivant :

| 1996 | 1997 | 1998 | 1999 | 2000 |
|---|---|---|---|---|
| 103,4 | 107,0 | 107,2 | 106,1 | 105,3 |

À partir de ces résultats, dans une deuxième étape, on peut suivre l'évolution de la production industrielle pour les biens de consommation en moyenne annuelle. Pour cela, on calcule en pourcentage, dans le tableau suivant, la valeur de l'accroissement ou de la diminution d'un indice par rapport au précédent :

| 1997/96 | 1998/97 | 1999/98 | 2000/99 |
|---|---|---|---|
| 3,51 % | 0,12 % | - 0,98 % | - 0,73 % |

On a ainsi une idée de la façon dont a évolué l'activité "Biens de consommation" durant la période, sachant qu'on a tenu compte de ce qui s'est passé durant chaque année. En effet, la valeur moyenne pour une année donnée prend en compte les valeurs des quatre trimestres correspondants (effet "mémoire" relatif à chacune des années).

La troisième étape consiste à calculer, en pourcentage, les évolutions de la production industrielle des biens de consommation en glissement (sur les 4$^{èmes}$ trimestres, de façon à ce que la comparaison avec la 1$^{ère}$ méthode soit pertinente) :

| 1996/95 | 1997/96 | 1998/97 | 1999/98 | 2000/99 |
|---|---|---|---|---|
| 2,33 % | 2,08 % | - 2,50 % | 0,19 % | - 1,52 % |

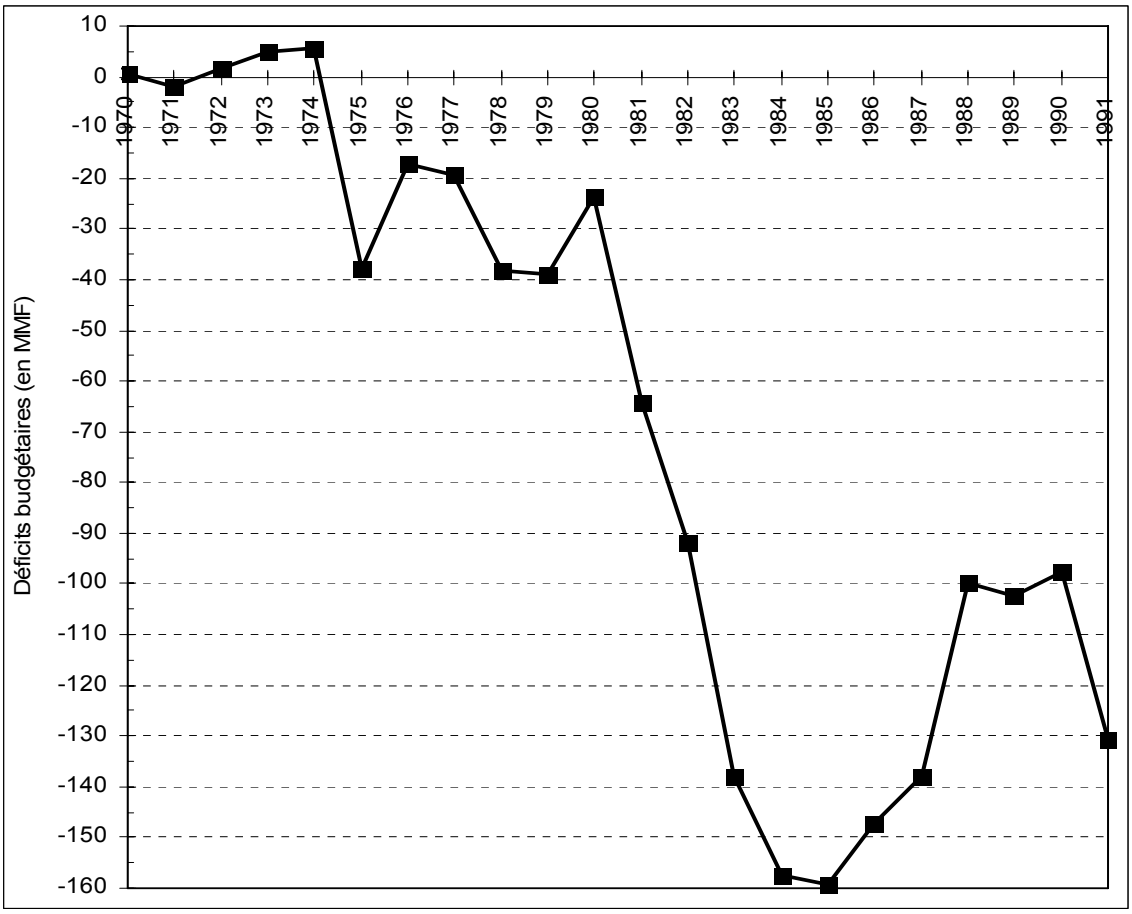
Dans la mesure où l'on dispose de l'indice du 4$^{ème}$ trimestre pour l'année 1995, on gagne une année puisqu'on peut calculer l'évolution instantanée de la production entre les 4$^{èmes}$ trimestres de 1995 et de 1996. Ici, on a une photo instantanée (pas d'effet "mémoire" dans ce cas) qui permet de comparer, en glissement, les indices de la production de chaque 4$^{ème}$ trimestre.

Dans cet exemple, comme dans le cas général, on constate que les résultats sont très différents entre les deux méthodes d'évaluation. Cela provient du fait que la 1$^{ère}$ méthode met en jeu des valeurs qui "mémorisent" l'ensemble de l'année correspondante, ce qui n'est pas le cas de la seconde.

## 52. Exemple d'utilisation de la méthode analytique de lissage d'une série chronologique

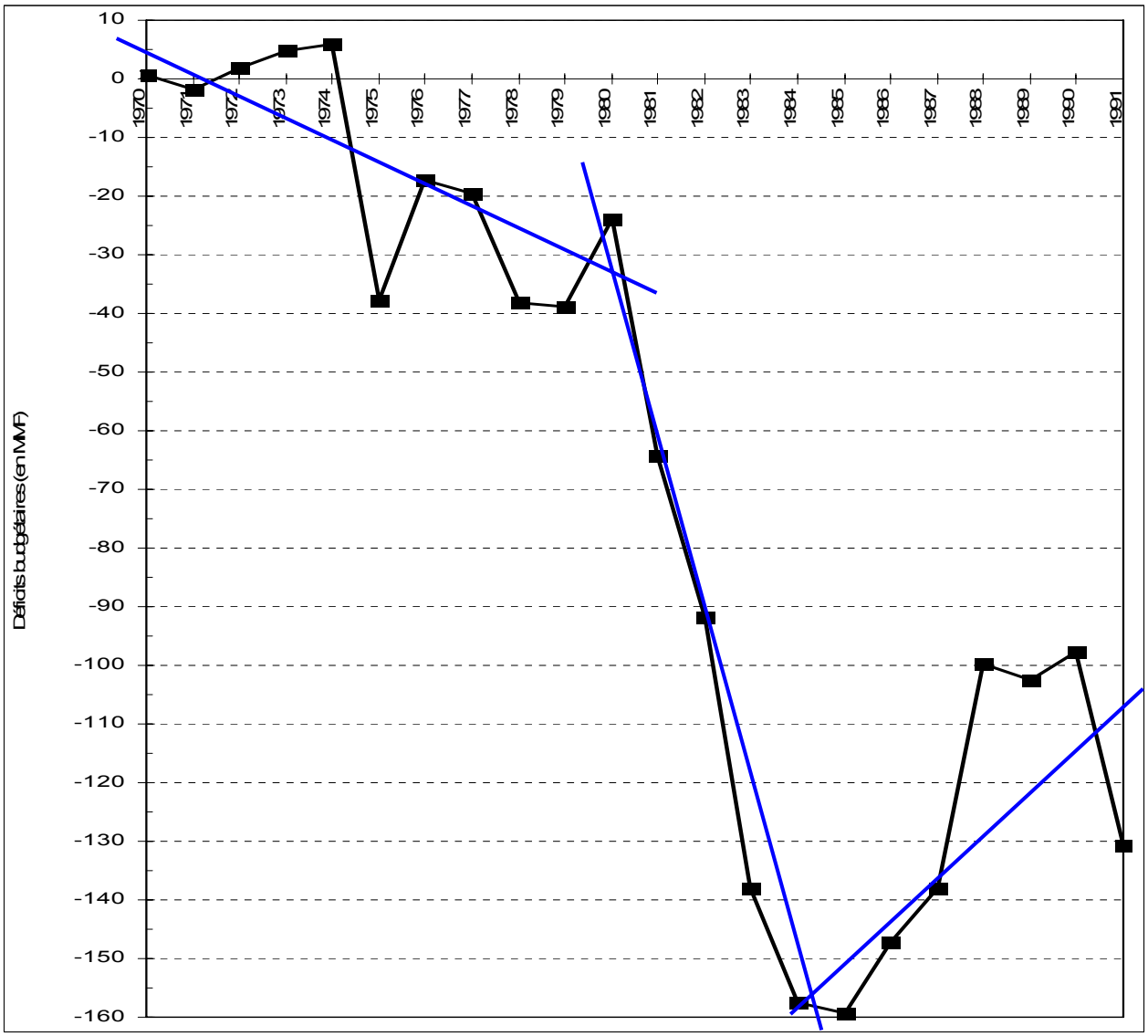
En complément des éléments précisés au point 22 du chapitre 2, reprenons l'exemple des soldes d'exécution des lois de finances en France, entre 1970 et 1991 (en MMF), utilisé pour illustrer le fonctionnement des moyennes mobiles au point 213 du chapitre 2 :

| $x_t$ | $y_t$ | $x_t^2$ | $y_t^2$ | $x_t\, y_t$ | $y^*_t$ |
|---|---|---|---|---|---|
| **1970** | 0,6 | 3 880 900 | 0,4 | 1 182,0 | 14,9 |
| **1971** | -1,8 | 3 884 841 | 3,2 | -3 547,8 | 7,0 |
| **1972** | 1,8 | 3 888 784 | 3,2 | 3 549,6 | -0,9 |
| **1973** | 4,8 | 3 892 729 | 23,0 | 9 470,4 | -8,8 |
| **1974** | 5,8 | 3 896 676 | 33,6 | 11 449,2 | -16,6 |
| **1975** | -37,8 | 3 900 625 | 1 428,8 | -74 655,0 | -24,5 |
| **1976** | -17,2 | 3 904 576 | 295,8 | -33 987,2 | -32,4 |
| **1977** | -19,5 | 3 908 529 | 380,3 | -38 551,5 | -40,2 |
| **1978** | -38,2 | 3 912 484 | 1 459,2 | -75 559,6 | -48,1 |
| **1979** | -38,9 | 3 916 441 | 1 513,2 | -76 983,1 | -56,0 |
| **1980** | -23,8 | 3 920 400 | 566,4 | -47 124,0 | -63,8 |
| **1981** | -64,3 | 3 924 361 | 4 134,5 | -127 378,3 | -71,7 |
| **1982** | -91,8 | 3 928 324 | 8 427,2 | -181 947,6 | -79,6 |
| **1983** | -137,9 | 3 932 289 | 19 016,4 | -273 455,7 | -87,4 |
| **1984** | -157,5 | 3 936 256 | 24 806,3 | -312 480,0 | -95,3 |
| **1985** | -159,4 | 3 940 225 | 25 408,4 | -316 409,0 | -103,2 |
| **1986** | -147,3 | 3 944 196 | 21 697,3 | -292 537,8 | -111,0 |
| **1987** | -137,9 | 3 948 169 | 19 016,4 | -274 007,3 | -118,9 |
| **1988** | -99,8 | 3 952 144 | 9 960,0 | -198 402,4 | -126,8 |
| **1989** | -102,4 | 3 956 121 | 10 485,8 | -203 673,6 | -134,7 |
| **1990** | -97,7 | 3 960 100 | 9 545,3 | -194 423,0 | -142,5 |
| **1991** | -130,7 | 3 964 081 | 17 082,5 | -260 223,7 | -150,4 |
| **43 571** | **-1 490,9** | **86 293 251** | **175 287,4** | **-2 959 695,4** | |




Remarque importante : ici, on constate sur le graphique que la courbe fait apparaître des retournements de tendance importants. Dans un tel cas, il est évidemment plus pertinent d'utiliser la méthode empirique des moyennes mobiles.

Dans ce cas précis, pour utiliser la méthode analytique de manière pertinente, donc plus efficace (surtout en fin de période), il convient de repérer les grandes tendances décelables sur le graphique de la série brute (1970-80, 1980-85, 1985-90), puis de réaliser trois régressions linéaires successives :




Le tableau de calculs précédent a pour objectif de déterminer les coefficients d'une droite d'ajustement linéaire associée à la série chronologique. Dans la dernière colonne du tableau ($y^*_t$), on trouve les valeurs prises par la droite d'ajustement pour chacune des 22 années étudiées.

**Remarque importante** : rappelons que, contrairement à deux caractères quantitatifs ordinaires, ici, en soi, la valeur du coefficient de corrélation linéaire (ou du coefficient de détermination) n'a pas d'importance, car y est fonctionnellement liée au temps t.

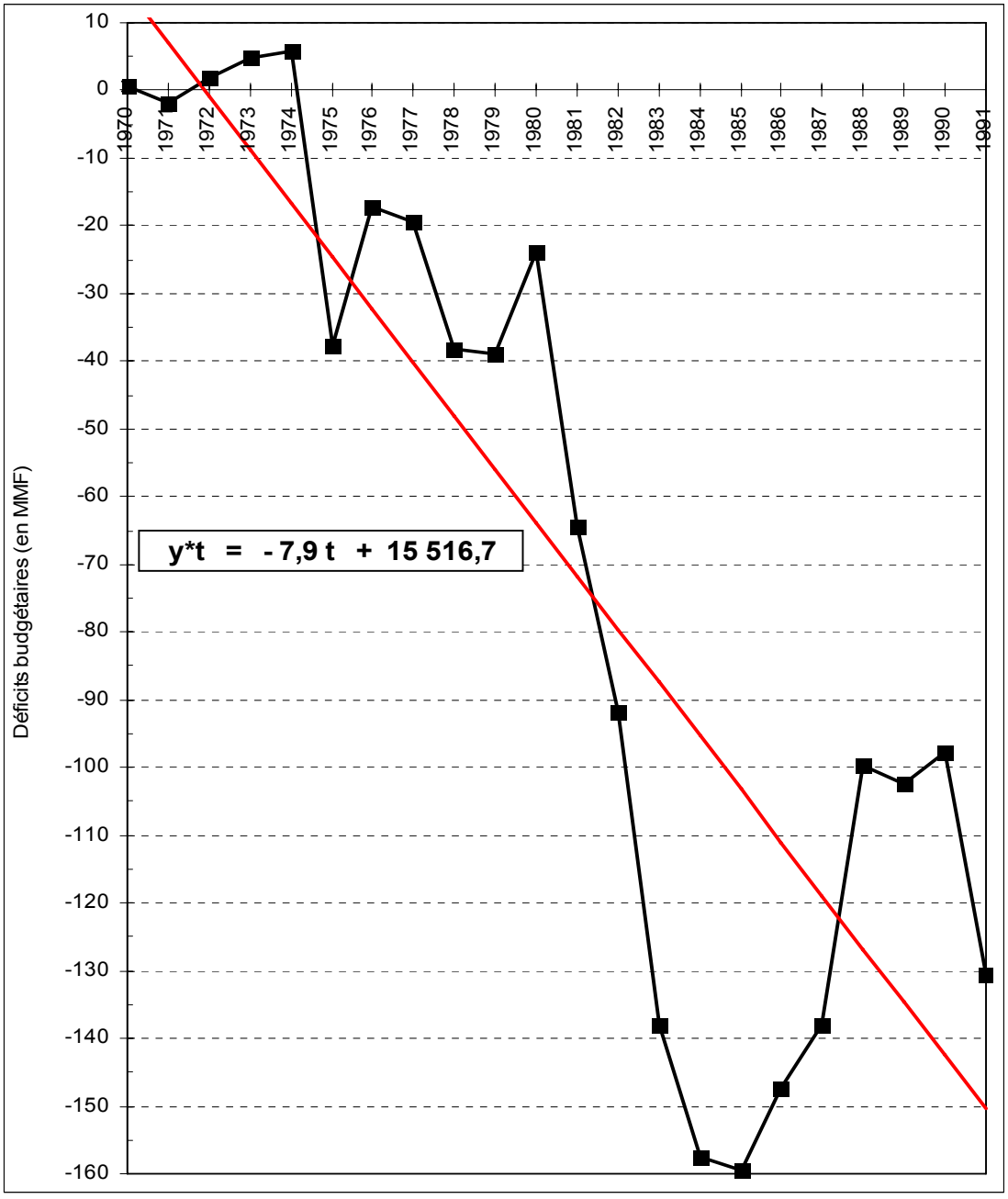
Cependant, plus cette valeur est élevée, plus la pertinence dans l'utilisation de la méthode analytique est renforcée. On cherche seulement à tracer une droite d'ajustement linéaire associée à la série chronologique, afin d'estimer le trend de cette série, sachant que ce n'est pas le temps qui explique la croissance du déficit budgétaire sur la période d'analyse (corrélation ≠ relation de causalité).

Dans le cas d'une désaisonnalisation cependant, la valeur du coefficient de corrélation joue un rôle dans le choix du modèle à mettre en œuvre pour une estimation analytique de la composante extra-saisonnière de la série.

Ici, l'on obtient :

| | | | |
|---|---|---|---|
| xbar = | 1980,5 | ectyp x = | 6,3 |
| ybar = | -67,8 | ectyp y = | 58,1 |
| cov xy = | -316,7 | rho xy = | -0,86 |
| | | rho² xy = | 0,74 |
| a = | -7,9 | b = | 15516,7 |

**y*t = - 7,9 t + 15 516,7**



## 53. Filtrage d'une série chronologique

La procédure de filtrage cherche à mettre en évidence des **cycles**, dont la période est supérieure à l'année. Le plus souvent, on utilise les moyennes mobiles pour réaliser cette mise en évidence.

**Principe du filtrage** : supposons qu'un phénomène connaisse un cycle quinquennal. Dans ce cas, si l'on retient une moyenne mobile d'ordre 5 (MM5) et si l'on pose la différence :

valeur du trend - valeur de la MM5 correspondante, alors cette différence est égale à 0.

On en déduit que la moyenne mobile d'ordre 5 (MM5) filtre le mouvement cyclique de période 5. On peut aussi dire que le filtre MM5 supprime (arrête) le cycle de période 5.

Exemple : considérons le modèle (d'école) suivant: :

a) le trend est fonction de sa valeur à la période t - 1, selon l'expression : $T_t = T_{t-1} + 10$.

b) la valeur de la série brute $Y_t$ est fonction de la valeur du trend $T_t$ et de la valeur du cycle $C_t$ , tel qu'on ait : $Y_t = T_t + C_t$ .

c) on pose l'hypothèse qu'il existe un cycle rigoureux, tel que C prenne les valeurs successives suivantes tous les 5 ans : 0, + 20, - 15, - 10, + 5 .

On se propose, à titre de comparaison, de calculer les moyennes mobiles MM5, MM6 et MM12. Les résultats, consignés dans le tableau suivant :

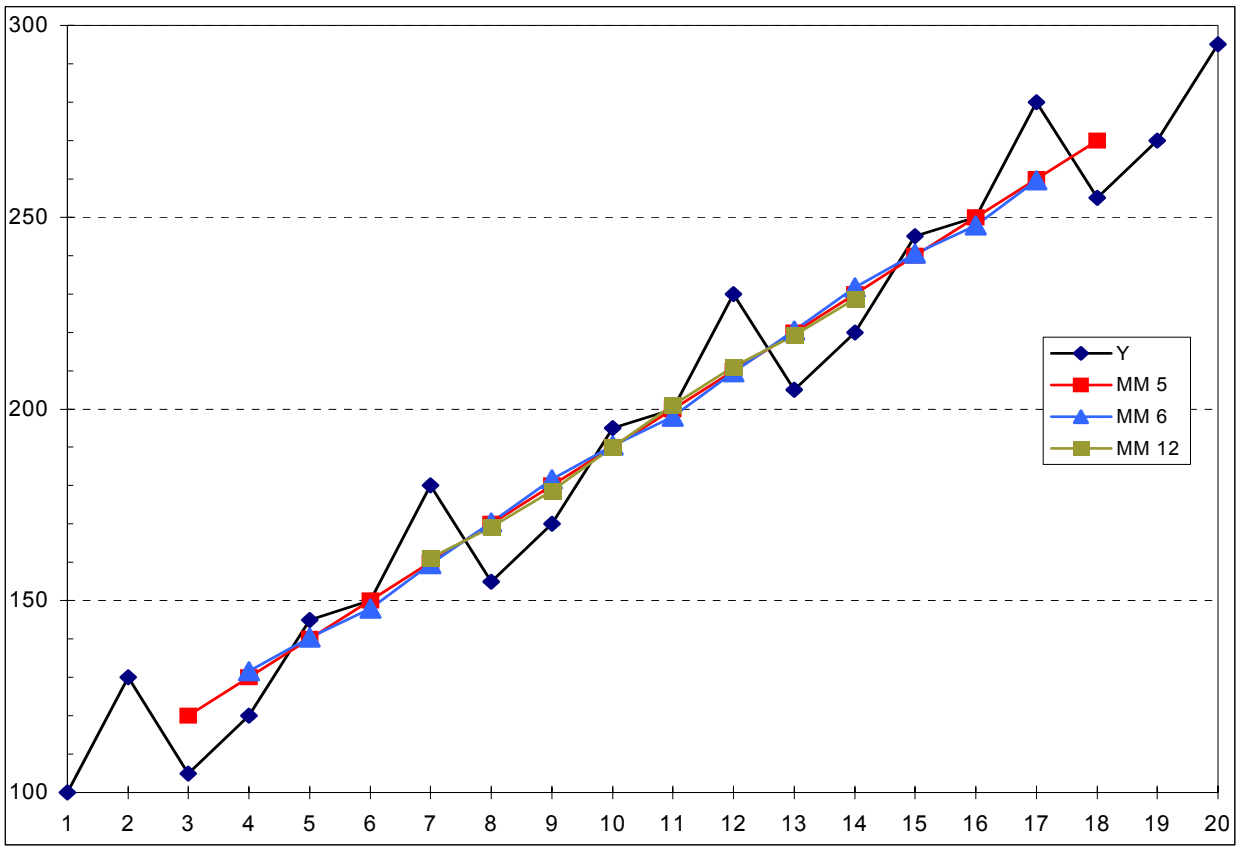
- confirment que la MM5 filtre le mouvement cyclique de période 5 ;
- montrent que si l'on calcule des moyennes mobiles d'ordre différent de 5 (ici, MM6 et MM12), il se produit des **artefacts**, c'est-à-dire qu'on observe des mouvements périodiques qui n'existent pas, mais qui sont induits par la méthode de filtrage par les moyennes mobiles.

| t | T | C | Y | MM 5 | MM 6 | MM 12 | dMM5 | dMM6 | dMM12 |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 100 | 0 | 100 | | | | | | |
| **2** | 110 | 20 | 130 | | | | | | |
| **3** | 120 | -15 | 105 | 120 | | | 0 | | |
| **4** | 130 | -10 | 120 | 130 | 131,67 | | 0 | 1,67 | |
| **5** | 140 | 5 | 145 | 140 | 140,42 | | 0 | 0,42 | |
| **6** | 150 | 0 | 150 | 150 | 147,92 | | 0 | -2,08 | |
| **7** | 160 | 20 | 180 | 160 | 159,58 | 161,04 | 0 | -0,42 | 1,04 |
| **8** | 170 | -15 | 155 | 170 | 170,42 | 169,17 | 0 | 0,42 | -0,83 |
| **9** | 180 | -10 | 170 | 180 | 181,67 | 178,75 | 0 | 1,67 | -1,25 |
| **10** | 190 | 5 | 195 | 190 | 190,42 | 190,00 | 0 | 0,42 | 0,00 |
| **11** | 200 | 0 | 200 | 200 | 197,92 | 201,04 | 0 | -2,08 | 1,04 |
| **12** | 210 | 20 | 230 | 210 | 209,58 | 211,04 | 0 | -0,42 | 1,04 |
| **13** | 220 | -15 | 205 | 220 | 220,42 | 219,17 | 0 | 0,42 | -0,83 |
| **14** | 230 | -10 | 220 | 230 | 231,67 | 228,75 | 0 | 1,67 | -1,25 |
| **15** | 240 | 5 | 245 | 240 | 240,42 | | 0 | 0,42 | |
| **16** | 250 | 0 | 250 | 250 | 247,92 | | 0 | -2,08 | |
| **17** | 260 | 20 | 280 | 260 | 259,58 | | 0 | -0,42 | |
| **18** | 270 | -15 | 255 | 270 | | | 0 | | |
| **19** | 280 | -10 | 270 | | | | | | |
| **20** | 290 | 5 | 295 | | | | | | |

Les colonnes dMM5, dMM6 et dMM12 sont égales à la différence entre une valeur de la moyenne mobile et la valeur correspondante du trend.

On constate que les MM6 et MM12 génèrent des mouvements périodiques artificiels qui n'apparaissent pas dans la série brute. On remarque aussi que la perte d'information augmente en même temps que l'ordre des moyennes mobiles.

Sur le graphique suivant, on note que les valeurs de MM5 sont alignées et correspondent exactement au trend de la série brute :




Remarque terminale . sur une série chronologique réelle, comment peut-on repérer un cycle ?

Si l'on observe des variations prononcées (à la hausse ou à la baisse) toutes les i années (périodes), alors on peut calculer une moyenne mobile d'ordre i (MMi).

Si la différence :

$$\text{valeur du trend} - \text{valeur de la MMi} \sim 0$$

sur l'ensemble de la période d'analyse, alors on peut raisonnablement avancer l'hypothèse qu'il existe un cycle de période i.