

**MÉTHODES QUANTITATIVES  
DES SCIENCES SOCIALES APPLIQUÉES  
AUX ÉTUDES URBAINES ET RÉGIONALES**

Nouvelle édition révisée

par  
ANDRÉ LEMELIN

INRS-Urbanisation, Culture et Société

Décembre 2004

Responsabilité scientifique :

[Andre.Lemelin@ucs.inrs.ca](mailto:Andre.Lemelin@ucs.inrs.ca)

INRS

Urbanisation, Culture et Société

(coordonnées ci-dessous)

Diffusion :

Institut national de la recherche scientifique

Urbanisation, Culture et société

3465, rue Durocher

Montréal H2X 2C6

Québec

Téléphone (514) 499-4000

Télécopieur (514) 499-4065

**MÉTHODES QUANTITATIVES DES SCIENCES SOCIALES  
APPLIQUÉES AUX ÉTUDES URBAINES ET RÉGIONALES**

**Nouvelle édition révisée**

par

**ANDRÉ LEMELIN**

INRS-Urbanisation, Culture et Société

Décembre 2004



## **AVIS IMPORTANT : CONDITIONS D'UTILISATION DE CET OUVRAGE**

Il me fait plaisir de mettre mon livre gratuitement à la disposition de la communauté scientifique et universitaire. Je n'en attends aucuns droits d'auteur. Je ne renonce pas pour autant à mes droits de propriété intellectuelle. Toute citation ou reproduction de cet ouvrage, totale ou partielle, en langue originale ou en traduction, devra en indiquer clairement la source. La référence exacte à donner est la suivante :

- pour la version papier :

**Lemelin, André (2004). *Méthodes quantitatives des sciences sociales appliquées aux études urbaines et régionales*, Montréal, INRS-UCS, édition révisée, pagination multiple. Copie papier disponible au centre de documentation de l'INRS-UCS.**

- pour la version électronique :

**Lemelin, André (2005). *Méthodes quantitatives des sciences sociales appliquées aux études urbaines et régionales*, en ligne :**

**[www.inrs-ucs.quebec.ca/Cours/Lemelin/EUR8213/index.htm](http://www.inrs-ucs.quebec.ca/Cours/Lemelin/EUR8213/index.htm)**

**(révision : le 3 mars 2005).**

En plus de la reconnaissance normale de mes droits de propriété intellectuelle, je prie les lecteurs et utilisateurs d'avoir l'obligeance de me signaler toute erreur ou inexactitude qu'ils pourraient détecter, ainsi que de me faire part de leurs commentaires, le cas échéant. D'ailleurs, même si vous n'avez aucune remarque à formuler, vous seriez très aimable de prendre la peine de me faire savoir que vous utilisez mon livre électronique.

Pour communiquer avec moi, je vous invite à utiliser le courrier électronique :

**andre\_lemelin@ucs.inrs.ca**

Je vous remercie d'avance de votre collaboration.

André Lemelin, Montréal, décembre 2004



## PRÉFACE

Ce livre est basé sur les cours que je dispense depuis 1992 dans le cadre du programme conjoint INRS-UQAM en Études urbaines. Son originalité est qu'il a été développé en interaction avec des étudiants des cycles supérieurs, qui sont cependant pour la plupart d'orientation « qualitative ». J'ai donc tenté de mettre au point un exposé de la matière qui soit parfaitement rigoureux, mais qui prenne appui sur la logique et l'épistémologie davantage que sur la formalisation mathématique. Les énoncés mathématiques sont néanmoins donnés, mais ils sont presque toujours accompagnés d'exemples numériques : car l'expérience m'a appris que même des étudiants qui ont une bonne compréhension de la logique et des procédures de calcul n'ont pas toujours la capacité de traduire avec assurance le symbolisme mathématique en opérations numériques.

J'ai donc cherché, par mes efforts pédagogiques, à rendre les méthodes quantitatives moins rébarbatives à de jeunes chercheurs qui sont d'emblée mieux disposés envers l'approche qualitative. Pour autant, ce n'est pas seulement à eux que je ne m'adresse. Je vise aussi les étudiants qui ont du goût et des aptitudes pour les méthodes quantitatives, mais qui, souvent, se laissent absorber par les aspects techniques, au détriment d'une compréhension plus fondamentale et plus critique. J'ai d'ailleurs constaté en plus d'une occasion que même les étudiants en science économique, pourtant rompus aux méthodes économétriques, sont parfois peu conscients des limites de la mesure. Ce que je leur propose ici, c'est un antidote contre ce biais.

Je souhaiterais enfin que cet ouvrage mette à la portée de tous les jeunes chercheurs en sciences sociales – et aussi de quelques chercheurs moins jeunes – une trousse d'outils d'analyse de données. Car ces outils sont aujourd'hui indispensables, pour peu que l'on fasse de la recherche appliquée. Les exemples utilisés sont surtout tirés du domaine des études régionales et urbaines, mais les méthodes présentées sont tout aussi pertinentes en géographie, en science politique, en sociologie...

Le livre comprend quatre parties, relativement indépendantes. La première, *Quantité et mesure*, est consacrée à la nature de la mesure, à sa portée et à ses limites, particulièrement dans le contexte des sciences sociales. La réflexion critique s'appuie sur une présentation détaillée de quelques-uns des outils de base en analyse des données : manipulation des tables de contingence, analyse de décomposition, construction de nombres indices, mesure de la

concentration (indice de Gini), mesure de la dissimilarité... Mais au-delà de l'apprentissage technique, l'objectif poursuivi dans cette première partie est d'ouvrir des pistes de réflexion critique qui, je l'espère, amèneront l'étudiant à devenir un lecteur averti et, éventuellement, un chercheur compétent, sachant pratiquer la critique et l'auto-critique et interpréter les résultats de recherche avec prudence.

La seconde partie du livre, *Le rôle de la statistique en science sociale* met l'accent sur la logique et le statut épistémologique de l'induction statistique. Je n'ai pas la prétention d'y présenter l'ensemble des outils statistiques que doit posséder un jeune chercheur : il ne manque pas de manuels fort bien faits, que l'on peut consulter à loisir. L'objectif poursuivi est plutôt d'amener l'étudiant à acquérir une maîtrise des principes fondamentaux, grâce auxquels il pourra ensuite utiliser à bon escient des méthodes plus avancées ou plus spécialisées ; grâce auxquels aussi il pourra devenir un lecteur averti lorsqu'il prendra connaissance de recherches qui s'appuient sur les méthodes statistiques.

La troisième partie du livre porte sur l'analyse de régression. En tant qu'outil passe-partout de l'analyse multivariée, la régression multiple est largement utilisée en sciences sociales et sa connaissance est indispensable aux chercheurs. En outre, cette méthode peut servir de référence, à partir de laquelle aborder les méthodes plus spécialisées et plus avancées. Ici encore, le point de vue épistémologique est privilégié.

La quatrième et dernière partie du livre s'intitule *L'analyse quantitative des données qualitatives*. Il s'agit d'une présentation des méthodes qui sont propres à l'analyse de variables catégoriques, une situation fréquente en sciences sociales. Trois thèmes sont abordés : l'analyse des tableaux de contingence, l'application de la régression multiple à l'analyse de la variance (variables indépendantes catégoriques) et enfin, les modèles à variable dépendante qualitative (logit et autres).

André Lemelin

Montréal, décembre 2003



## PLAN DE L'OUVRAGE

---

### **Première partie – Quantité et mesure**

- 1-0 Introduction à la première partie
- 1-1 L'approche quantitative et la mesure
- 1-2 L'interprétation des grandeurs
- 1-3 Le problème de la multidimensionnalité : les nombres indices
- 1-4 Mesure de l'inégalité et de la concentration
- 1-5 Mesure de la dissimilarité
- 1-6 En guise de conclusion...
- 1-A Quelques Outils mathématiques de base
- 1-B Principes et outils de gestion des données (notes schématiques)
- 1-CA Notes schématiques d'initiation à Excel (version avec commandes en anglais)
- 1-CF Notes schématiques d'initiation à Excel (version illustrée avec commandes en français)
- 1-D Notes schématiques d'initiation à SPSS
- 1-E Tableau de l'alphabet grec
- 1-F Développement de la formule de calcul de l'indice de Gini

---

### **Seconde partie – Le rôle de la statistique en sciences sociales**

- 2-0 Introduction à la seconde partie
- 2-1 Description et induction statistiques en sciences sociales
- 2-2 L'induction statistique
- 2-3 Les tests d'hypothèse
- 2-4 Conclusion de la seconde partie
- 2-A Rappel de quelques formules courantes en statistique

---

## **Troisième partie – L'analyse de régression**

3-0 Introduction – l'analyse multivariée : une classification des méthodes

3-1 Le modèle linéaire général et son estimation par la méthode des moindres carrés

3-2 L'induction statistique appliquée à la régression multiple

3-3 Conclusion de la troisième partie

3-A La lecture d'une sortie d'ordinateur

3-B L'allégorie de la caverne de Platon

3-C Note de terminologie : écart type, erreur type, etc.

---

## **Quatrième partie – L'analyse quantitative de données qualitatives**

4-0 Introduction : L'analyse quantitative de données qualitatives

4-1 Analyse des tableaux de contingence

4-2 Le modèle linéaire général et la régression multiple appliqués à l'analyse de variance

4-3 Modèles à variable dépendante qualitative

4-4 Conclusion de la quatrième partie

---

## **5-0 Postface**

## **6-0 Références**

## INTRODUCTION À LA PREMIÈRE PARTIE

Cette partie du cours aborde un problème ardu, celui de la mesure. Problème d'autant plus ardu en sciences sociales que les phénomènes étudiés sont complexes et se prêtent mal à l'approche expérimentale. Ce problème difficile, un scientifique y est confronté au double titre de « consommateur » et de producteur de recherche.

Comme lecteur, un scientifique doit garder son sens critique en éveil lorsqu'il prend connaissance de résultats de recherche. Bien que certaines faiblesses dans la construction des données soient suffisamment apparentes pour être facilement détectées, d'autres ne peuvent être mises à jour que par un examen plus technique. C'est d'ailleurs pour montrer cela qu'en dépit d'un parti pris d'accessibilité, cet ouvrage est quelque peu alourdi par des formules mathématiques et des exemples chiffrés.

Comme chercheur, pour peu que l'on fasse de la recherche appliquée, on finit inévitablement par devoir quantifier, c'est-à-dire mesurer. Cela ne va pas de soi : une réflexion sur la mesure est un aspect essentiel de la méthodologie :

- Que veut-on mesurer (définition du concept et de ses dimensions) ?
- Quels indicateurs peut-on ou veut-on utiliser ? (Que de compromis déchirants dans cette interrogation !)
- Quel est le modèle sous-jacent à la mesure ?
- À quel degré la mesure dépend-elle du « jugement du chercheur » ? (Cela n'est pas inacceptable, pourvu que l'on respecte les exigences de la transparence)
- Quelles sont les limites de la mesure, la marge d'incertitude des résultats ?

À travers et au-delà de l'apprentissage technique de quelques outils d'analyse quantitative, l'objectif poursuivi dans cette première partie est donc d'ouvrir des pistes de réflexion critique. L'étudiant est invité à suivre ces pistes parfois abruptes, pour devenir un lecteur averti et, éventuellement, un chercheur compétent, sachant pratiquer la critique et l'auto-critique et interpréter les résultats de recherche avec toute la prudence qu'exige la rigueur scientifique.

## CHAPITRE 1-1

### L'APPROCHE QUANTITATIVE ET LA MESURE

---

#### Plan

1-1.1 L'opérationnalisation des concepts : indicateurs, mesure et variables	2
1-1.2 Qu'est ce que la mesure ?	5
1-1.3 Échelles de mesure et types de variables	8
1-1.4 Types de données	12
1-1.5 Structure matricielle fondamentale des données	14

## CHAPITRE 1-1

### L'APPROCHE QUANTITATIVE ET LA MESURE

Références : Gilles (1994, Introduction et chap. 1 et 2) ; Bryman et Cramer, 1990, p. 61-74 ; Blalock (1972, chap. 2) ; Lazarsfeld (1971)

*Quantitatif* s'oppose à *qualitatif*. Non pas que les deux approches soient mutuellement exclusives : elles seraient plutôt complémentaires (sur les débats idéologiques et méthodologiques à propos des approches qualitative et quantitative, voir Gilles, 1994, Introduction, p. 1-9). Mais les deux termes s'opposent quant à leur définition : est quantitatif ce qui se mesure. Plus exactement, la quantité est la propriété de ce qui peut être *mesuré* ou compté, de ce qui est susceptible d'accroissement ou de diminution.

Mais que vient faire la mesure dans la démarche scientifique en sciences sociales ? Et puis, qu'est-ce que *mesurer* ?

#### 1-1.1 L'opérationnalisation des concepts : indicateurs, mesure et variables

En sciences – et cela est aussi vrai en sciences sociales – les théories et les hypothèses sont formulées au moyen de concepts et de relations entre des concepts. Un concept est une idée, une représentation mentale abstraite et générale d'un être, d'une manière d'être ou d'un rapport : c'est, en somme, un atome de pensée. Gilles (1994, p15) met l'accent sur l'opération qui crée le concept en explicitant sa compréhension et en fixant son extension<sup>1</sup> : pour lui, un concept est une « construction de la pensée résultant d'une opération par laquelle on individualise des traits permettant de rapprocher des objets différents ou de distinguer des objets autrement similaires », ou, autrement dit, par laquelle on définit les critères permettant de déterminer si tel ou tel objet fait partie ou non de l'extension du concept.

---

<sup>1</sup> L'*extension* logique est l'ensemble des objets concrets ou abstraits auxquels s'appliquent un concept, une proposition (ensemble des cas où elle est vraie) ou une relation (ensemble des systèmes qui la vérifient). L'extension d'un concept s'oppose à la *compréhension*, qui est l'ensemble des caractères qui appartiennent à un concept. Par exemple, le concept *homme* a une moindre extension, mais une plus grande compréhension que *mammifère*.

Exemple :

- « La consommation des ménages croît avec le revenu » ;  
cette proposition contient les concepts « consommation des ménages », « revenu » et le lien entre les deux est exprimé par les mots « croît avec ».

Pour rapprocher les propositions théoriques de la réalité, ou pour confronter les hypothèses à l'observation, il faut *opérationnaliser* les concepts, c'est-à-dire établir une relation systématique entre les concepts et la réalité observable, au moyen d'*indicateurs*. On peut définir les indicateurs comme des « signes, comportements ou réactions directement observables par lesquels on repère au niveau de la réalité les dimensions d'un concept » (Gilles, 1994, p. 27). Les dimensions sont les différentes composantes d'un concept (Gilles, 1994, p. 24) : nous reviendrons plus loin sur cette notion de dimension.

Opérationnaliser un concept, c'est donc lui associer un ou plusieurs *indicateurs* qui permettront de distinguer avec exactitude les variations observées dans la réalité par rapport au concept. Distinguer les variations, cela veut dire *mesurer* : l'opérationnalisation d'un concept conduit donc à la *mesure*.

Mentionnons que ce lien entre l'opérationnalisation et la mesure existe aussi bien dans l'approche qualitative que dans l'approche quantitative. Car, même dans l'approche qualitative, il faut bien classifier et compter les sujets, ce qui constitue une opération de mesure, comme nous le verrons plus loin. À cet égard, Gilles écrit : « D'une manière générale, les méthodes dites qualitatives (histoire de vie, analyse de récit, observation participante, entrevue en profondeur, étude de cas...) font, elles aussi, usage de la statistique à des fins descriptives » (1994, p. 3) ; il distingue : « Méthodes qualitatives et données qualitatives ne procèdent pas de la même logique. Les premières reposent sur une conception humaniste, herméneutique ou interprétative des sciences sociales qui, dans cette perspective, deviennent sciences humaines. Quant aux secondes, elles peuvent être prises dans le sens de données ne permettant que l'utilisation de certaines techniques statistiques dites "robustes" » (1994, p. 3, note 5). Nous verrons d'ailleurs sous peu qu'en un certain sens, on peut mesurer des propriétés qualitatives : c'est pourquoi il n'est pas absurde de parler de l'analyse quantitative de données qualitatives (Quatrième partie de l'ouvrage).

Tout cela est résumé par Gilles (1994, p. 24), qui se réfère au schéma classique de Lazarsfeld (1971) : opérationnaliser, c'est « soumettre les concepts, par l'analyse, à un processus qui les

transforme en dimensions, puis en indicateurs permettant de les observer, de les mesurer ou de les quantifier ».

Exemple :

- Pour opérationnaliser le concept « consommation des ménages », c'est-à-dire pour mesurer la consommation d'un ménage, on peut utiliser le montant déclaré, dans une enquête auprès des ménages, en réponse à une question comme « La semaine dernière, combien ont dépensé l'ensemble des personnes qui composent le ménage ? »
- Mais on pourrait aussi mesurer la consommation d'un ménage par calcul, en soustrayant de ses revenus le montant d'impôt sur le revenu qu'il a payé et la somme qu'il a épargné durant une année donnée.

En général, un concept peut se traduire par plusieurs indicateurs. Le choix des indicateurs est très important en recherche. Les indicateurs retenus doivent être valides et fiables (on dit aussi fidèles).

- Un indicateur est *valide* lorsqu'il mesure bien ce que l'on veut mesurer, c'est-à-dire lorsqu'il reflète les variations relatives au concept même qu'il est censé représenter. Pour examiner la validité d'un indicateur, il faut évidemment qu'au préalable le concept ait été clairement défini.
- Un indicateur est *fiable* ou *fidèle* lorsque les variations dans la mesure correspondent à des variations véritables.

Exemple :

- Le montant dépensé la semaine dernière n'est peut-être pas une mesure valide de la consommation, parce que ce montant inclut peut-être des dépenses en capital (investissement résidentiel), alors que la définition du concept théorique « consommation » exclut le coût d'acquisition de biens durables.
- La réponse à propos du montant dépensé la semaine dernière n'est peut-être pas fiable, parce que la personne qui répond au questionnaire ne sait peut-être pas ce qu'ont dépensé les autres membres du ménage.

Le résultat de l'application d'un indicateur à un ensemble d'objets est une *variable*. Une variable est donc définie par la chose à mesurer (le concept et ses dimensions), par la façon de la mesurer (l'indicateur) et par son domaine d'application (les objets auxquels s'applique la mesure).

Soulignons enfin la distinction qu'il faut faire entre une *variable* et les différentes *valeurs* qu'elle peut prendre.

Exemple :

- Dans une enquête auprès d'un échantillon de ménages, la réponse à la question « La semaine dernière, combien ont dépensé l'ensemble des personnes qui composent le ménage ? » est une variable qui prend une valeur différente pour chacun des ménages de l'enquête.

### 1-1.2 Qu'est ce que la mesure ?

Mesurer, c'est comparer. Mais encore ? Dans la langue courante, on définit la mesure comme « l'évaluation d'une grandeur par comparaison avec une autre de la même espèce prise pour unité » (Dictionnaire Larousse de la langue française, cédérom, 1996). Nous verrons dans un moment que cette définition est très étroite. *Encyclopaedia Britannica* (cédérom, 1998) propose : « measurement : the process of associating numbers with physical quantities and phenomena ». Dans le même esprit, Gilles (1994, p. 34) écrit : « Mesurer, c'est établir une correspondance entre l'ensemble que constitue le phénomène à mesurer et un ensemble de nombres que l'on choisit en fonction de la nature du phénomène ». Ces deux dernières définitions, plus larges, sont cependant incomplètes tant que l'on ne spécifie pas quelles sont les conditions que doit remplir une correspondance numérique pour constituer une mesure. C'est l'objet de la *théorie de la mesure*.

Pour nos fins, nous retiendrons ceci<sup>2</sup> : une correspondance constitue une mesure si elle permet de *comparer* deux *objets quelconques* par rapport à une *propriété donnée*.

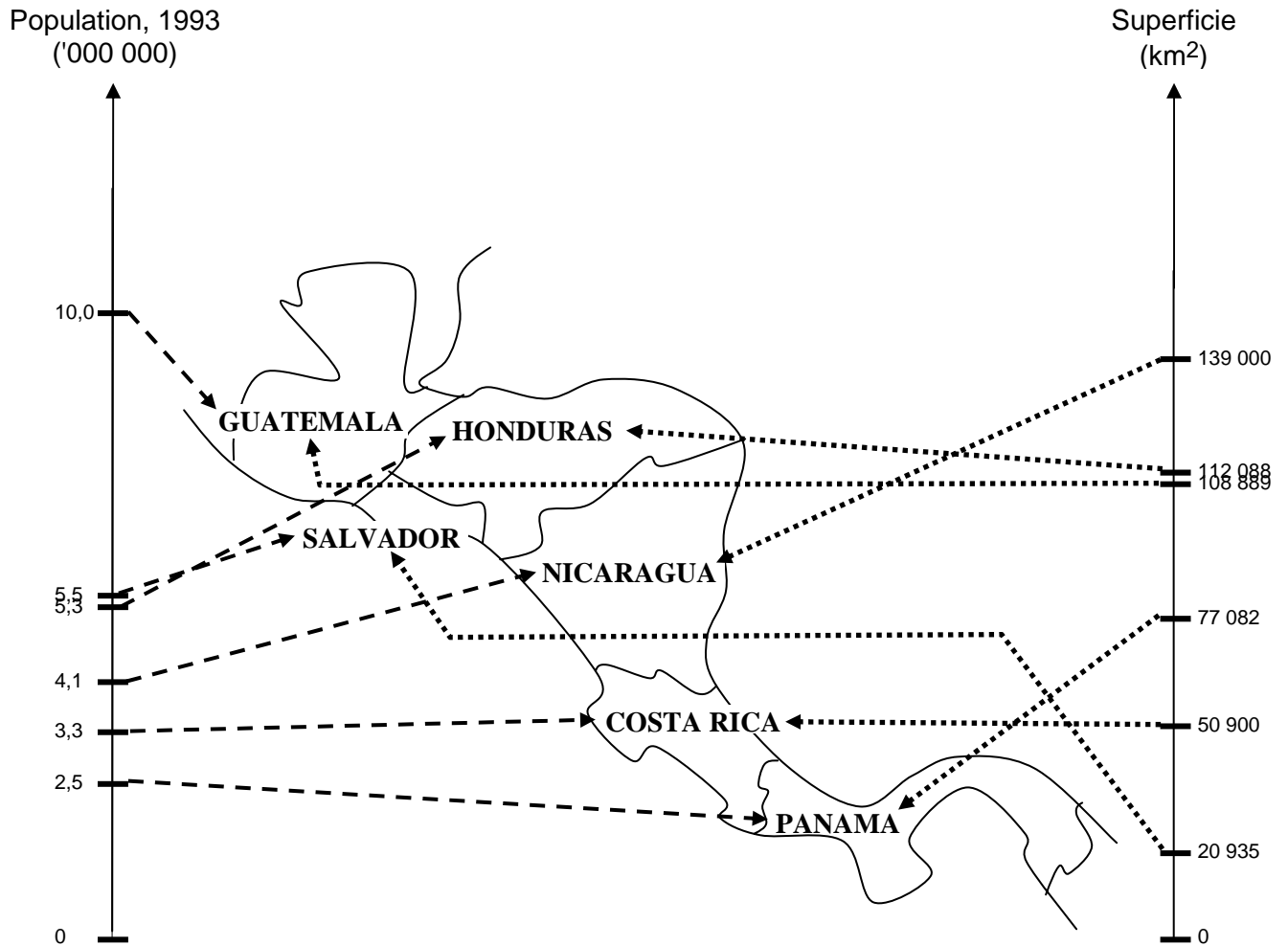
Par exemple, supposons que la propriété à mesurer soit la superficie. Les pays, les chambres à coucher et les mouchoirs de poche sont des objets pour lesquels la propriété « superficie » est définie. Pour qu'elle puisse constituer une mesure de superficie, une correspondance doit permettre de comparer quant à leur superficie deux pays, ou deux mouchoirs de poche, ou même un pays et un mouchoir de poche. Mais qu'est-ce que *comparer* ? Dans le contexte de la mesure, comparer, c'est déterminer si, par rapport à une propriété donnée, les deux objets comparés sont semblables ou non, et, s'ils ne sont pas semblables, lequel possède la propriété mesurée à un degré plus grand que l'autre.

---

<sup>2</sup> Ce qui suit est inspiré de Taylor, 1977, chap. 2, p. 38-41, mais la notation utilisée ici est différente.



## UNE MESURE EST UNE CORRESPONDANCE...



Source des données : Facultad Latino Americana de Ciencias Sociales  
FLACSO, Sede Costa Rica, San José, Costa Rica

On peut formaliser ce qui précède ainsi. Désignons par  $A$  et  $B$  deux objets quelconques (deux pays, par exemple) qui possèdent la propriété à mesurer (la superficie, par exemple). Puisqu'une mesure associe un nombre à chaque objet, un peu comme une fonction mathématique, il est naturel de représenter la mesure de la même façon : convenons alors que  $f(A)$  est le chiffre de la superficie de  $A$  ; de même,  $f(B)$  est le chiffre de la superficie de  $B$ . La comparaison examine les relations suivantes :

$$f(A) = f(B)$$

$$f(A) \neq f(B)$$

$$f(A) < f(B)$$

$$f(A) > f(B)$$

Une *mesure* est une correspondance qui permet, pour au moins l'une des relations qui précèdent, de déterminer si elle est vraie ou fausse. Dans le cas de la superficie, on peut établir une correspondance entre chaque pays et le nombre de kilomètres carrés qui sont compris à l'intérieur de ses frontières, ou entre chaque mouchoir de poche et la fraction de kilomètre carré qu'il recouvre. Lorsqu'on compare les chiffres donnés par cette correspondance, on peut décider s'il est vrai que  $f(A) = f(B)$  ( $A$  et  $B$  ont même superficie), ou  $f(A) \neq f(B)$  ( $A$  et  $B$  n'ont pas même superficie), auquel cas, ou bien  $f(A) < f(B)$  ( $A$  est plus petit que  $B$ ), ou bien  $f(A) > f(B)$  ( $A$  est plus grand que  $B$ ).

Dans l'exemple de la superficie, la mesure permet de déterminer la valeur de vérité (vraie ou fausse) de *chacune* des quatre relations =,  $\neq$ , < et >. Mais la définition de la mesure n'exige pas que l'on puisse déterminer la valeur de vérité des *quatre* relations. Par exemple, supposons que la propriété examinée soit la nationalité. On pourrait définir la correspondance suivante :

$$f(X) = 0 \text{ si la personne } X \text{ est de nationalité costaricaine ;}$$

$$f(X) = 1 \text{ si la personne } X \text{ est d'une autre nationalité centraméricaine ;}$$

$$f(X) = 2 \text{ dans tous les autres cas.}$$

Alors

$f(A) = f(B)$  signifie que la personne  $A$  et la personne  $B$  sont de même nationalité (dans la classification retenue) ;

$f(A) \neq f(B)$  signifie que la personne  $A$  et la personne  $B$  ne sont pas de même nationalité.

Par contre, les relations  $f(A) < f(B)$  et  $f(A) > f(B)$  n'ont aucune signification. La correspondance constitue néanmoins une mesure au sens large : c'est une mesure de la nationalité. En un certain sens, donc, on peut mesurer des propriétés qualitatives.

Note : Les valeurs numériques de la correspondance n'ont aucune signification et elles sont parfaitement arbitraires. On pourrait même définir la correspondance en termes de symboles autres que des nombres. Par exemple, on aurait pu définir

$f(X) = \text{'CR'}$  si la personne X est de nationalité costaricaine ;

$f(X) = \text{'CA'}$  si la personne est d'une autre nationalité centraméricaine ;

$f(X) = \text{'OT'}$  dans tous les autres cas.

### 1-1.3 Échelles de mesure et types de variables

Même si l'on admet qu'une propriété qualitative comme la nationalité d'une personne peut être mesurée, il n'en demeure pas moins que la mesure d'une telle propriété semble imparfaite, en comparaison de la mesure de propriétés comme la superficie ou le revenu. En effet, pour une propriété comme la nationalité, on ne peut pas décider s'il est vrai ou faux que  $f(A) < f(B)$  ou  $f(A) > f(B)$  : cela n'aurait pas de signification. Par contre, pour la superficie d'un territoire ou le revenu d'un ménage, on peut déterminer s'il est vrai ou faux que  $f(A) < f(B)$  ou  $f(A) > f(B)$  : la mesure est plus complète.

C'est pourquoi on distingue plusieurs types de variables, selon l'échelle de mesure qui leur est associée <sup>3</sup> :

1. Variables *catégoriques*
2. Variables *ordinales*
3. Variables *d'intervalle*
4. Variables *rationnelles*

#### **Variables catégoriques**

Les variables catégoriques (« nominal » en anglais) résultent de l'application d'une échelle de mesure qui ne permet de décider que des relations = et  $\neq$ . La valeur que prend une variable catégorique s'appelle « modalité » : elle indique à quelle catégorie appartient l'individu auquel

---

<sup>3</sup> On trouve une classification similaire chez Legendre et Legendre (1998, p. 28 et suivantes).

elle se rapporte : une variable catégorique permet donc de classer les individus en groupes. On distingue

- Variables *dichotomiques* : 2 catégories possibles ;
- Variables *polytomiques* : plus de 2 catégories.

Exemples :

- Sexe (homme/femme) : variable catégorique dichotomique ;
- Nationalité : variable catégorique polytomique (lorsque l'on distingue plus de deux nationalités).

On peut remplacer une variable polytomique par plusieurs variables dichotomiques. D'ailleurs, certaines méthodes d'analyse l'exigent. Par exemple, considérons une variable polytomique de nationalité :

$NAT = 0$  si la personne  $X$  est de nationalité costaricaine ;

$NAT = 1$  si la personne est d'une autre nationalité centraméricaine ;

$NAT = 2$  dans tous les autres cas.

On peut remplacer cette variable par deux variables dichotomiques, comme

$COR = 1$  si la personne est citoyenne du Costa Rica et  $COR = 0$  autrement

$CAM = 1$  si la personne est citoyenne d'un pays d'Amérique Centrale autre que le Costa Rica et  $CAM = 0$  autrement.

Question : pourquoi seulement deux variables dichotomiques, alors que la variable polytomique peut prendre trois valeurs ?

### **Variables ordinales**

Les variables ordinales résultent de l'application d'une échelle de mesure qui permet de décider que de chacune des quatre relations  $=$ ,  $\neq$ ,  $<$  et  $>$ . Les valeurs que prend une variable ordinale pour différents individus permettent donc de ranger les individus en ordre croissant ou décroissant par rapport à la propriété mesurée. On distingue les ordres *faibles* ou *réduits* – incomplets, par classes d'équivalence – et les ordres *complets*.

Exemples :

- Nombre de points obtenus à un test d'aptitudes (ordre complet : si deux sujets obtiennent le même nombre de points, la mesure indique qu'ils possèdent le même degré d'aptitude selon ce test) ;

- Variable définie par : 1 si l'étudiant réussit un examen donné et 0 s'il échoue (ordre faible : si deux étudiants ont réussi, ça ne veut pas dire qu'ils sont d'égale force).

Les mesures ordinales sont définies « à une transformation monotone croissante près », c'est-à-dire que l'on ne change pas la mesure si l'on applique à la variable une transformation mathématique, pourvu que l'on ne change pas l'ordre numérique des valeurs. Par exemple, on pourrait remplacer le nombre de points obtenus par le logarithme du nombre de points, ou par le carré du nombre de points, ou on pourrait ajouter un million de points à tous les sujets.

### **Variables d'intervalle**

Les variables d'intervalle sont similaires aux variables ordinales, mais en plus de permettre de ranger les individus en ordre croissant ou décroissant, elles permettent de comparer les *différences* entre individus.

Exemples :

- La température : s'il fait  $-15^{\circ}\text{C}$  à Montréal,  $+24^{\circ}\text{C}$  à San José (Costa Rica) et  $+18^{\circ}\text{C}$  à Miami, on peut dire que la différence de température est moins grande entre San José et Miami ( $6^{\circ}\text{C}$ ) qu'entre Miami et Montréal ( $33^{\circ}\text{C}$ ). Avec une variable ordinale, ce genre de comparaison n'a pas de sens.

Formellement, les variables d'intervalle résultent de l'application d'une échelle de mesure où les *différences* entre les valeurs sont aussi des mesures, ordinales : l'échelle de mesure permet de déterminer la valeur de vérité de *chacune* des relations suivantes

$$f(A) - f(B) = f(C) - f(D)$$

$$f(A) - f(B) \neq f(C) - f(D)$$

$$f(A) - f(B) < f(C) - f(D)$$

$$f(A) - f(B) > f(C) - f(D)$$

Avec une variable d'intervalle, le zéro de l'échelle de mesure est arbitraire mais les transformations de l'échelle doivent préserver la comparaison entre les écarts. C'est pourquoi les échelles d'intervalle sont définies « à une transformation linéaire près ». Par exemple, on passe de l'échelle Celsius à l'échelle Fahrenheit au moyen de la transformation linéaire

$$F = 32 + 1,8 \times C$$

Autre exemple, en géographie, la direction est donnée en degrés, calculés dans le sens des aiguilles d'une montre à partir de la direction nord. Le zéro (franc nord) est arbitraire.

### **Variables rationnelles**

Les variables rationnelles (aussi appelées *proportionnelles*) sont similaires aux variables d'intervalle, sauf qu'avec les variables rationnelles, il existe un zéro naturel<sup>4</sup>. Il en découle que le rapport entre deux valeurs a un sens (rationnel vient de « ratio », rapport).

Exemple :

- Le revenu est une variable rationnelle. Si une personne gagne 60 000 \$, on peut dire qu'elle gagne deux fois plus qu'une personne qui gagne 30 000 \$. Par contre, il serait absurde de prétendre qu'il fait deux fois plus chaud à 20° C qu'à 10° C (20° C = 68° F et 10° C = 50° F).

Formellement, les variables rationnelles résultent de l'application d'une échelle de mesure où les *rapports* entre les valeurs sont aussi des mesures, ordinales : l'échelle de mesure permet de déterminer la valeur de vérité de *chacune* des relations suivantes

$$\frac{f(A)}{f(B)} = \frac{f(C)}{f(D)}$$

$$\frac{f(A)}{f(B)} \neq \frac{f(C)}{f(D)}$$

$$\frac{f(A)}{f(B)} < \frac{f(C)}{f(D)}$$

$$\frac{f(A)}{f(B)} > \frac{f(C)}{f(D)}$$

Si l'on revient à la définition de la mesure selon Larousse comme « l'évaluation d'une grandeur par comparaison avec une autre de la même espèce prise pour unité », on constate que cette définition ne s'applique en vérité qu'aux échelles de mesure rationnelles. La définition du Larousse est donc restrictive.

Il arrive souvent que les valeurs observées de variables rationnelles ou d'intervalle soient regroupées en classes. Par exemple, une variable « revenu » pourrait prendre la forme suivante :

$$REV = 1 \text{ si revenu} < 10\,000 \$$$

$$REV = 2 \text{ si } 10\,000 \$ \leq \text{revenu} < 25\,000 \$$$

---

<sup>4</sup> Pour cette raison, la température mesurée en degrés Kelvin est une variable rationnelle, puisqu'il existe un zéro naturel, le « zéro absolu ». Le zéro absolu, qui équivaut à environ -273,16 à l'échelle Celsius, est la température

$REV = 3$  si  $25\ 000 \$ \leq \text{revenu} < 50\ 000 \$$

$REV = 4$  si  $\text{revenu} \geq 50\ 000 \$$

Une variable de ce type est une variable ordinaire qui définit un ordre faible. Le fait de regrouper les valeurs observées en classes a donc pour effet de transformer une variable rationnelle (ou d'intervalle) en variable ordinaire d'ordre faible. On passe ainsi à une échelle de mesure plus « primitive » et on perd de l'information. Il est donc préférable, lorsque c'est possible, d'utiliser les données sous leur forme originale.

### ***Échelles de mesure et méthodes quantitatives***

Il existe des méthodes d'analyse quantitative adaptées à tous les types de variables. On peut donc appliquer des méthodes *quantitatives* à l'analyse de données *qualitatives*, lorsque celles-ci peuvent être mesurées au moyen de variables catégoriques ou ordinaires.

#### **1-1.4 Types de données**

Cette partie du cours porte sur les méthodes quantitatives d'analyse des données. Mais la qualité de l'analyse dépend d'abord et avant tout de la qualité des données analysées. Les données ne sont jamais « parfaites » et l'analyste compétent doit adapter ses méthodes à la qualité des données qu'il doit traiter.

On peut distinguer trois types de données :

- les données primaires
- les données secondaires non publiées
- les données secondaires publiées

Il y a des problèmes de qualité spécifiques à chaque type de données.

#### **1. Données primaires (enquêtes)**

Le contrôle de la qualité doit se faire à toutes les étapes :

- préparation des instruments de cueillette des données (questionnaires)
- cueillette
- codification

- saisie, validation, correction et organisation
- évaluation *ex post* de la qualité

## 2. Données secondaires non publiées

Ces données sont souvent collectées pour des fins, administratives ou autres, différentes de la recherche (les rôles d'évaluation foncière, par exemple) : les concepts sont souvent mal définis ou différents de ceux qu'on cherche à mesurer (les variables formées à partir de ces données ne sont qu'imparfaitement valides).

Le contrôle de la qualité des données secondaires non publiées pose souvent des problèmes proches de ceux qui se posent lorsqu'il s'agit de données primaires. Cependant, dans le cas de données secondaires, l'analyste ne peut pas veiller lui-même au contrôle de la qualité à toutes les étapes.

## 3. Données secondaires publiées

L'utilisation judicieuse de données secondaires publiées exige que l'on tienne compte de toute l'information pertinente qui accompagne les données (métadonnées) <sup>5</sup> :

- définitions et concepts
- méthodes de cueillette et de compilation
- évaluation de la qualité par l'émetteur
- crédibilité des sources

En plus de dépendre de la qualité des données, la qualité de l'analyse peut être compromise par des erreurs dans le traitement préalable qu'on fait subir aux données avant de leur appliquer les méthodes d'analyse.

Exemples :

- Erreur de variable lors de l'extraction de données d'une banque de données (masse salariale au lieu du salaire horaire).
- Erreur de programmation lors de l'appariement de deux fichiers (« merge ») : dédoublement de certaines observations.
- Erreur de formule dans un tableur (adresses relatives ou absolues...) ; ces erreurs résultent souvent d'opérations de « copier-coller ».

---

<sup>5</sup> Atkinson et Brandolini (2001) examinent les avantages et les pièges des données secondaires dans le contexte de l'analyse des inégalités de revenu dans les pays de l'OCDE.



Si vous ne trouvez pas d'erreur dans les données,  
c'est parce que vous ne cherchez pas bien ...

### 1-1.5 Structure matricielle fondamentale des données

Pour être utilisables, les données doivent d'abord être organisées de façon à ce que l'on sache à quoi réfère chaque nombre. Il y a plusieurs manières d'organiser les données, mais toutes découlent de la structure fondamentale des données. Cette structure fondamentale est celle d'une matrice, ou d'un tableau où, par convention,

- les *colonnes* correspondent habituellement à différentes *variables* (caractéristiques, propriétés, attributs, indicateurs, descripteurs, ...);
- les *lignes* correspondent habituellement à différentes *observations* (cas, individus, objets).

Il arrive que les observations se rapportent à des moments ou à des périodes successives : on est alors en présence de *séries chronologiques* ou *temporelles*. On a une situation analogue lorsque les observations se rapportent aux différents lieux d'un ensemble géographique donné (pays d'un continent, villes ou régions d'un pays, quartiers d'une ville, zones...) : on pourrait parler alors de *séries spatiales*. Les données spatiales ne sont pas toujours des *séries complètes*, qui comportent une observation, et une seule, pour chacun des lieux que l'on distingue dans un espace donné. Qu'elles constituent ou non des séries complètes, on dit que des données sont *géoréférencées* lorsqu'elles contiennent une ou plusieurs variables qui permettent de situer chaque observation dans l'espace géographique.

La structure matricielle fondamentale se généralise à plus de deux dimensions<sup>6</sup> quand certaines des variables sont catégoriques (variables de classification). Par exemple, supposons que l'on ait réalisé une enquête auprès d'un échantillon de personnes et que, parmi les variables pour lesquelles on a recueilli des données se trouve la profession du répondant. Dans ce cas, le reste des données peut être organisé en plusieurs tableaux à deux dimensions, un par profession. Si l'on superpose ces tableaux, on peut concevoir l'organisation des données comme un cube dont les couches successives correspondent aux différentes professions, les lignes à différents répondants, et les colonnes aux différentes autres variables. Naturellement, avec plus d'une variable catégorique, on peut imaginer un « hypercube » de données à quatre

---

<sup>6</sup> Attention : le mot « dimensions » est utilisé dans un contexte différent pour désigner les dimensions d'un concept.

dimensions ou plus. La représentation mentale la plus appropriée dépend des analyses que l'on veut faire.

Ajoutons que la distinction entre observations et variables n'est pas étanche. Il arrive que les observations et les variables soient interchangeables, notamment lorsque les observations correspondent aux différentes catégories d'une variable de classification, tandis que les variables sont des attributs se rapportant aux différentes catégories d'une autre variable de classification<sup>7</sup>. Tel est le cas, par exemple, d'un tableau du nombre d'emplois par branche d'activité et par ville dans une région donnée. Dans ces conditions, on peut considérer

- soit que chaque observation correspond à une ville, et que les variables sont les nombres d'emplois des différentes branches d'activité dans cette ville ;
- soit que chaque observation correspond à une branche d'activité, et que les variables sont les nombres d'emplois de cette branche dans les différentes villes.

Là encore, la représentation mentale que l'on privilégie dépend des analyses que l'on veut faire.

Revenons au modèle d'organisation élémentaire, celui d'un tableau à deux dimensions. On distingue deux points de vue, ou modes d'analyse, selon que l'on s'attache aux relations entre les observations ou aux relations entre les variables (Jayet, 1993, p. 1-2 ; Legendre et Legendre, 1998, p. 248, reprennent une terminologie de Cattell, 1952, et distinguent l'analyse « en mode R », qui est l'analyse des relations entre les descripteurs, et l'analyse « en mode Q », qui est l'analyse des relations entre les objets). Cette distinction permet de classer les types d'analyses et les méthodes qui leur sont associées (voir le tableau qui suit).

---

<sup>7</sup> Comme le montre l'exemple donné dans les lignes qui suivent, une telle ambivalence est généralement le fait de données qui ont fait l'objet d'un premier traitement et qui sont constituées en tableau de contingence ou en tableau d'analyse de variance.

---

**Point de vue « horizontal » : entre les variables**

- Combiner plusieurs variables en une seule, qui les résume : construction de nombres indices
- Comparer deux variables : mesure de la similarité/dissimilarité
- Étudier les relations de dépendance
  - entre deux variables : corrélation, régression simple
  - entre une variable dépendante et plusieurs variables indépendantes : régression multiple et autres méthodes multivariées comportant une variable dépendante
  - entre plusieurs variables parmi lesquelles on ne distingue pas de variable dépendante : méthodes multivariées

**Point de vue « vertical » : entre les observations ou objets**

- Caractériser la distribution d'une variable : mesure de l'inégalité ou de la concentration, méthodes statistiques univariées
  - Lorsqu'il existe un ordre naturel entre les observations, étudier les relations entre les différentes observations d'une même variable : mesure et modélisation de l'évolution des séries temporelles, analyse de l'autocorrélation (temporelle, spatiale)
  - Comparer deux objets : mesure de la similarité/dissimilarité
-

## CHAPITRE 1-2

### L'INTERPRÉTATION DES GRANDEURS

---

#### Plan

1-2.1 Mesures relatives : l'exemple du quotient de localisation	2
Le quotient de localisation	3
Estimation de l'emploi exportateur au moyen du quotient de localisation	11
1-2.2 L'analyse de décomposition additive et multiplicative des variations	13
Principe	13
Application à l'analyse « shift-share »	14
1-2.3 La mesure de la croissance (le calcul du taux de variation dans le temps)	21
Taux de croissance par période	21
Moyenne des taux de croissance par période	23
Calcul d'un taux de croissance exponentiel	24
Entre deux maux...	26
Ajustement d'une courbe de tendance	27
Que retenir ?	28

## CHAPITRE 1-2

### L'INTERPRÉTATION DES GRANDEURS

Au-delà du problème de la mesure surgit celui de l'interprétation des grandeurs, c'est-à-dire du sens à donner aux nombres. Nous aborderons ici trois sujets. Nous examinerons d'abord deux techniques numériques fréquemment utilisées pour faciliter l'interprétation des grandeurs : la construction d'une mesure relative et l'analyse de décomposition. Dans chaque cas, la méthode sera illustrée par une technique largement répandue en sciences régionales et en études urbaines. L'accent sera mis sur les limites de ces outils. Ensuite, nous traiterons de la mesure de la croissance ou, plus généralement, de la manière de résumer l'évolution d'une grandeur dans le temps.

#### 1-2.1 Mesures relatives : l'exemple du quotient de localisation

- En 1600, l'Angleterre comptait une population de l'ordre de cinq millions d'habitants<sup>1</sup>. Peut-on dire que l'Angleterre était alors densément peuplée ?
- À Berlin, vers 1800, une famille de maçon de cinq personnes consacrait 44,2 % de son budget à l'achat de pain<sup>2</sup>. Pour l'époque, était-ce normal ?
- Le prix réel du blé en France, traduit en heures de travail de manoeuvre, a été de moins de 100 heures le quintal au 15<sup>e</sup> siècle et au 16<sup>e</sup> jusqu'en 1543, puis au-dessus de 100 jusqu'en 1883, environ<sup>3</sup>. Qu'est-ce que cela signifie pour le niveau de vie ?

En somme, « Est-ce beaucoup ? » : c'est la réaction de la plupart d'entre nous quand on nous cite un chiffre dans un domaine qui ne nous est pas familier. Dans les exemples qui précèdent, ce ne sont pas les unités de mesure qui font problème. C'est le manque de point de référence.

---

<sup>1</sup> Braudel (1979, p. 49). Braudel compare l'Angleterre à la France (20 millions à la même époque) et conclut que la France était surpeuplée, puisque « Si l'un et l'autre pays s'étaient agrandis au rythme moyen du monde, l'Angleterre devrait compter aujourd'hui 40 millions d'habitants, la France 160 », ce qui est loin des chiffres actuels : en 2001, la France comptait 59,6 millions d'habitants et le Royaume-Uni, 58,9 (PNUD, 2003)

<sup>2</sup> Braudel (1979, p. 142). L'ensemble de la nourriture représente 72,7 % du budget. Le pain compte donc pour 60,8 % de la dépense alimentaire de la famille, « proportion énorme étant donné le prix relatif des céréales ». L'auteur fait la comparaison avec la dépense alimentaire du Parisien en 1788 et 1854 : « Le blé, premier fournisseur d'énergie, n'arrive qu'au troisième poste des dépenses, après la viande et le vin (17 % seulement, dans les deux cas, de la dépense totale) » (p. 143-144).

<sup>3</sup> Braudel (1979, p. 145) explique : « Un travailleur accomplit *approximativement* 3 000 heures de travail, chaque année; sa famille (4 personnes) consomme *approximativement* 12 quintaux par an... Franchir la ligne des 100 heures pour un quintal [1 quintal = 100 kg] est toujours grave; celle des 200 signale une cote d'alerte; à 300, c'est la famine ».

En somme, si mesurer, c'est comparer, l'interprétation des grandeurs requiert une « méta-comparaison », une comparaison avec une grandeur qui a un sens pour l'observateur, afin de mettre les données en perspective et de saisir l'ordre de grandeur des chiffres.

La méthode qui est peut-être la plus courante pour donner à l'observateur un point de repère pour l'interprétation des grandeurs est la représentation graphique avec comparaison. Le lecteur est invité à consulter à cet égard le passage « Du bon et du mauvais usage des graphes » dans Wonnacott et Wonnacott (1992, p. 61-69).

Il est également important, pour interpréter correctement une grandeur, de connaître son domaine de variation. Nous reviendrons sur ce thème lorsque nous examinerons les mesures d'inégalité et les mesures de similarité/dissimilarité.

Cela dit, il est parfois utile d'aller plus loin et de formaliser davantage cette méta-comparaison en construisant une mesure relative, qui est le rapport de deux valeurs. C'est ce que fait le quotient de localisation.

#### LE QUOTIENT DE LOCALISATION

Réf. : Page-Patton, 1991, ch. 14 ; Polèse, 1994, p. 128-129

Les quotients de localisation, aussi appelés *indices de concentration relative*, sont des mesures de l'importance relative de l'emploi d'une branche d'activité dans une ville ou une région<sup>4</sup>. Ils s'appliquent donc aux données d'un tableau de l'emploi par branche et par ville ou région. Voici un exemple numérique fictif :

BRANCHE	B1	B2	B3	Total
ZONE				
Z1	48	325	287	660
Z2	27	185	148	360
Z3	45	90	45	180
Total	120	600	480	1200

Un tableau de ce type s'appelle un tableau de contingence (voir Gilles, 1994, section 6.3). Nous voulons pouvoir répondre à des questions comme : « 48 emplois de la branche B1 dans la zone Z1, est-ce peu ? » ou « 325 emplois de la branche B2 dans la zone Z1, est-ce beaucoup ? ».

<sup>4</sup> Ils appartiennent à la famille de ce que Jayet (1993, p. 18) appelle les « indicateurs de spécificité ».

Comme première étape, nous pouvons examiner les distributions.

### Distribution de l'emploi des branches entre zones

BRANCHE	B1	B2	B3	Total
ZONE				
Z1	0,400	0,542	0,598	0,550
Z2	0,225	0,308	0,308	0,300
Z3	0,375	0,150	0,094	0,150
Total	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>

48 emplois de la branche *B1* dans la zone *Z1*, est-ce peu ? L'examen de la distribution de l'emploi entre les zones montre que ces 48 emplois constituent 40 % de l'emploi total de la branche *B1* : la zone *Z1* est celle où l'on trouve le plus grand nombre d'emplois de cette branche. Par contre, la zone *Z1* contient 55 % de l'emploi, toutes branches confondues : 48 emplois, ce n'est donc pas beaucoup, compte tenu de la taille de la zone *Z1*.

### Distribution de l'emploi des zones entre branches

BRANCHE	B1	B2	B3	Total
ZONE				
Z1	0,073	0,492	0,435	<b>1,000</b>
Z2	0,075	0,514	0,411	<b>1,000</b>
Z3	0,250	0,500	0,250	<b>1,000</b>
Total	0,100	0,500	0,400	<b>1,000</b>

325 emplois de la branche *B2* dans la zone *Z1*, est-ce beaucoup ? L'examen de la distribution de l'emploi entre les branches montre que ces 325 emplois représentent près de la moitié (49 %) de l'emploi de la zone *Z1*. Mais dans l'ensemble de l'économie, la branche *B2* compte pour la moitié : 325 emplois, c'est donc « normal ».

Le calcul des quotients de localisation est une manière de formaliser ce genre de raisonnement. Dans le premier cas (48 emplois de la branche *B1* dans la zone *Z1*), l'importance de la zone est mesurée par la part de cette zone dans l'emploi total de la branche ( $48 / 120 = 0,4$  ou 40 %) <sup>5</sup>. Mais pour interpréter ce 40 %, nous nous sommes référés au pourcentage correspondant de l'ensemble des activités ( $660 / 1200 = 0,55$  ou 55 %). La mesure relative que nous avons utilisée implicitement pour apprécier l'importance de *Z1* pour *B1* est le rapport  $0,40 / 0,55$  : un quotient de localisation, c'est ça ! Dans le second cas (325 emplois de la branche *B2* dans la zone *Z1*), nous avons procédé de manière analogue. L'importance de la branche est mesurée par la part de cette branche dans l'emploi total de la zone ( $325 / 660 = 0,492$  ou 49 %). Mais

<sup>5</sup> En un sens, c'est déjà là une mesure relative, puisque cette part est donnée par le rapport de deux nombres comparables.

pour interpréter de 49 %, nous nous sommes référés au pourcentage correspondant de l'ensemble des zones ( $600 / 1200 = 0,5$  ou 50 %). La mesure relative que nous avons utilisée implicitement pour apprécier l'importance de  $B2$  pour  $Z1$  est le rapport  $0,49 / 0,5$  : ça aussi, c'est un quotient de localisation. Ainsi, le quotient de localisation compare deux points correspondants sur deux distributions (deux points correspondants, et non pas deux distributions : les distributions sont des objets multidimensionnels ; nous verrons au chapitre 1-4 comment on peut les comparer). Voyons cela plus formellement.

### **Tableaux de contingence : notation et identités fondamentales**

Avant de faire une présentation plus formelle des quotients de localisation, établissons un système de notation approprié et rappelons les identités fondamentales qui se vérifient dans un tableau de contingence comme celle de l'emploi par zone et par branche.

#### **Notation**

$x_{ij}$	nombre d'emplois de la branche $j$ dans la zone $i$
$x_{\bullet j} = \sum_i x_{ij}$	nombre total d'emplois de la branche $j$
$x_{i\bullet} = \sum_j x_{ij}$	nombre total d'emplois dans la zone $i$
$x_{\bullet\bullet} = \sum_i \sum_j x_{ij}$	nombre total d'emplois de toutes branches dans toutes zones
$p_{ij} = \frac{x_{ij}}{x_{\bullet\bullet}}$	fraction de l'emploi total global qui appartient à la branche $j$ et est situé dans la zone $i$
$p_{\bullet j} = \sum_i p_{ij}$	fraction de l'emploi total global qui appartient à la branche $j$
$p_{i\bullet} = \sum_j p_{ij}$	fraction de l'emploi total global qui est situé dans la zone $i$
$p_{j i\bullet} = \frac{p_{ij}}{p_{i\bullet}}$	fraction de l'emploi total de la zone $i$ qui appartient à la branche $j$
$p_{i \bullet j} = \frac{p_{ij}}{p_{\bullet j}}$	fraction de l'emploi total de la branche $j$ qui est situé dans la zone $i$



On aura reconnu dans cette notation que  $p_{\bullet j}$  et  $p_{i\bullet}$  sont des *probabilités marginales* :  $p_{\bullet j}$  est la probabilité qu'un emploi tiré au hasard parmi les  $x_{\bullet\bullet}$  emplois dénombrés appartienne à la branche  $j$ ;  $p_{i\bullet}$  est la probabilité qu'un emploi tiré au hasard soit situé dans la zone  $i$ . On aura aussi reconnu que  $p_{j/i\bullet}$  et  $p_{i/\bullet j}$  sont des *probabilités conditionnelles* :  $p_{j/i\bullet}$  est la probabilité qu'un emploi tiré au hasard appartienne à la branche  $j$ , étant donné qu'il est situé dans la zone  $i$ ;  $p_{i/\bullet j}$  est la probabilité qu'un emploi tiré au hasard soit situé dans la zone  $i$ , étant donné qu'il appartient à la branche  $j$ .

On a naturellement les identités suivantes :

$$p_{\bullet j} = \sum_i p_{ij} = \sum_i \frac{x_{ij}}{x_{\bullet\bullet}} = \frac{x_{\bullet j}}{x_{\bullet\bullet}} \quad \text{et} \quad p_{i\bullet} = \sum_j p_{ij} = \sum_j \frac{x_{ij}}{x_{\bullet\bullet}} = \frac{x_{i\bullet}}{x_{\bullet\bullet}}$$

$$p_{j/i\bullet} = \frac{p_{ij}}{p_{i\bullet}} = \frac{x_{ij}/x_{\bullet\bullet}}{x_{i\bullet}/x_{\bullet\bullet}} \quad \text{et} \quad p_{i/\bullet j} = \frac{p_{ij}}{p_{\bullet j}} = \frac{x_{ij}/x_{\bullet\bullet}}{x_{\bullet j}/x_{\bullet\bullet}}$$

$$\sum_i \sum_j p_{ij} = \sum_i p_{i\bullet} = \sum_j p_{\bullet j} = 1$$

$$\sum_j p_{j/i\bullet} = \frac{\sum_j p_{ij}}{p_{i\bullet}} = 1 \quad \text{et} \quad \sum_i p_{i/\bullet j} = \frac{\sum_i p_{ij}}{p_{\bullet j}} = 1$$

### **Le quotient de localisation : formalisation**

Le quotient de localisation peut être défini aussi bien à partir de la distribution de l'emploi entre branches qu'à partir de la distribution entre zones. À partir de la distribution entre zones, le quotient de localisation de l'activité  $j$  dans la zone  $i$  est défini comme

$$QL_{ij} = \frac{\text{Fraction de l'emploi total de la branche } j \text{ situé dans la zone } i}{\text{Fraction de l'emploi total global situé dans la zone } i}$$

$$QL_{ij} = \frac{p_{i/\bullet j}}{p_{i\bullet}} = \frac{\frac{x_{ij}}{x_{\bullet\bullet}}}{\frac{x_{i\bullet}}{x_{\bullet\bullet}}}$$

Par exemple,

$$QL_{21} = 0,225 / 0,300 = 0,750$$

De façon équivalente, à partir de la distribution entre branches, le quotient de localisation de l'activité  $j$  dans la zone  $i$  est défini comme

$$QL_{ij} = \frac{\text{Fraction de l'emploi total de la zone } i \text{ appartenant à la branche } j}{\text{Fraction de l'emploi total global appartenant à la branche } j}$$

$$QL_{ij} = \frac{p_{j/i\bullet}}{p_{\bullet j}} = \frac{\frac{x_{ij}/x_{i\bullet}}{x_{\bullet j}/x_{\bullet\bullet}}}{\frac{x_{\bullet j}/x_{\bullet\bullet}}{x_{\bullet j}/x_{\bullet\bullet}}} = \frac{x_{ij}/x_{i\bullet}}{x_{\bullet j}/x_{\bullet\bullet}}$$

Par exemple,

$$QL_{21} = 0,075 / 0,100 = 0,750$$

Ce n'est pas par accident que les deux calculs donnent le même résultat, puisque

$$\frac{\frac{x_{ij}/x_{\bullet j}}{x_{i\bullet}/x_{\bullet\bullet}}}{\frac{x_{\bullet j}/x_{\bullet\bullet}}{x_{\bullet j}/x_{\bullet\bullet}}} = \frac{\frac{x_{ij}/x_{i\bullet}}{x_{\bullet j}/x_{\bullet\bullet}}}{\frac{x_{\bullet j}/x_{\bullet\bullet}}{x_{\bullet j}/x_{\bullet\bullet}}} = \frac{x_{ij} x_{\bullet\bullet}}{x_{i\bullet} x_{\bullet j}}$$

Dans notre exemple,

$$QL_{21} = \frac{\frac{x_{21}/x_{\bullet 1}}{x_{2\bullet}/x_{\bullet\bullet}}}{\frac{x_{\bullet 1}/x_{\bullet\bullet}}{x_{\bullet 1}/x_{\bullet\bullet}}} = \frac{\frac{x_{21}/x_{2\bullet}}{x_{\bullet 1}/x_{\bullet\bullet}}}{\frac{x_{\bullet 1}/x_{\bullet\bullet}}{x_{\bullet 1}/x_{\bullet\bullet}}} = \frac{x_{21} x_{\bullet\bullet}}{x_{2\bullet} x_{\bullet 1}}$$

$$QL_{21} = \frac{27/120}{360/1200} = \frac{27/360}{120/1200} = \frac{27 \times 1200}{360 \times 120} = 0,75$$

Il y a bel et bien équivalence.

**Quotients de localisation**

BRANCHE	B1	B2	B3
ZONE			
Z1	0,727	0,985	1,087
Z2	0,750	1,028	1,028
Z3	2,500	1,000	0,625

Les quotients de localisation peuvent prendre des valeurs entre zéro et l'infini <sup>6</sup>. Lorsque  $x_{ij} = 0$ , le quotient de localisation atteint sa valeur minimum :  $QL_{ij} = 0$ . Par ailleurs, il atteint la valeur la plus élevée possible lorsque  $x_{ij} = x_{\bullet j} = x_{i\bullet}$ , c'est-à-dire lorsque la totalité des emplois de l'activité  $j$  sont situés dans la zone  $i$  et qu'il ne se trouve aucune autre activité dans cette zone ;

dans ces conditions,  $QL_{ij} = \frac{x_{i\bullet}}{x_{ij}}$

Dans l'expression qui précède,  $x_{ij} = x_{\bullet j} = x_{i\bullet} \geq 1$ , sans quoi la branche  $j$  n'existerait pas. Il s'ensuit que la valeur maximale de  $QL_{ij}$  est  $x_{i\bullet}$  : cette valeur n'a pas de limite théorique et c'est pourquoi on dit que le quotient de localisation peut prendre des valeurs jusqu'à l'infini ; en pratique, le maximum est néanmoins limité par les valeurs observées.

Le point de repère naturel pour interpréter le quotient de localisation, est 1,0. Et comme le montrent les formules qui précèdent, on peut faire deux lectures du quotient de localisation :

- selon la première lecture, si  $QL_{ij} > 1$ , on dit que l'activité  $j$  est *relativement* concentrée <sup>7</sup> dans la zone  $i$  ; « relativement », c'est-à-dire en comparaison des autres activités, parce que la fraction de l'emploi qui est situé dans la zone  $i$  est *plus* importante pour l'activité  $j$  que pour les autres activités ; plus exactement, on dirait que la zone  $i$  est une zone de concentration relative pour cette activité, parce qu'il peut y avoir d'autres zones de concentration relative de cette même activité ;

Par exemple,  $QL_{23} = 1,028$  : l'activité  $B3$  est *relativement* concentrée dans la zone  $Z2$  ; pour autant, ce n'est pas dans la zone  $Z2$  qu'il y a le plus d'emploi de cette branche : c'est dans la zone  $Z1$ .

<sup>6</sup> Certains auteurs normalisent le quotient de localisation à l'aide de la transformation  $\frac{QL_{ij} - 1}{QL_{ij} + 1}$ . Ce rapport varie de -1 à +1.

<sup>7</sup> D'où, la justesse de l'expression « indice de concentration relative » pour désigner le quotient de localisation.

- selon la seconde lecture, si  $QL_{ij} > 1$ , on dira aussi que la zone  $i$  est *relativement* spécialisée dans l'activité  $j$ ; « relativement », c'est-à-dire en comparaison des autres zones, parce que l'activité  $j$  occupe dans la zone  $i$  une place *plus* importante *qu'ailleurs* ;

Par exemple,  $QL_{31} = 2,500$  : la zone  $Z3$  est *relativement* spécialisée dans l'activité  $B1$  ; pour autant, ce n'est pas la branche  $B1$  qui compte le plus grand nombre d'emplois de la zone  $Z3$  : c'est la branche  $B2$ .

- si au contraire  $QL_{ij} < 1$ , on dit que l'activité  $j$  est *relativement* moins présente dans la zone  $i$  qu'ailleurs : l'activité  $j$  n'est pas *relativement* concentrée dans la zone  $i$ , et la zone  $i$  n'est pas *relativement* spécialisée dans l'activité  $j$ .

Par exemple,  $QL_{12} = 0,985$  : la branche  $B2$  est *relativement* moins présente dans la zone  $Z1$ , bien qu'elle y soit la branche avec le plus grand nombre d'emplois et bien que ce soit en  $Z1$  que cette branche ait le plus grand nombre d'emplois (325 est le nombre le plus grand de sa ligne et de sa colonne).

Les exemples donnés montrent que l'adverbe « relativement » est important dans les énoncés d'interprétation qui précèdent. Voyons cela d'un point de vue plus général. Si la zone  $i$  est petite par rapport aux autres zones ( $p_{i\bullet}$  petit), il se peut, même quand  $QL_{ij} > 1$ , que la fraction de l'emploi de l'activité  $j$  qui se trouve dans la zone  $i$  ( $p_{i/\bullet j}$ ) ne soit pas importante. En effet,

$$QL_{ij} = \frac{\text{Fraction de l'emploi total de la branche } j \text{ situé dans la zone } i}{\text{Fraction de l'emploi total global situé dans la zone } i}$$

$$QL_{ij} = \frac{p_{i/\bullet j}}{p_{i\bullet}} = \frac{\frac{x_{ij}}{x_{\bullet j}}}{\frac{x_{i\bullet}}{x_{\bullet\bullet}}}$$

de sorte que si  $p_{i\bullet}$  est petit, il est possible que  $QL_{ij} > 1$  même si  $p_{i/\bullet j}$  est petit, pourvu que  $p_{i\bullet}$  soit encore plus petit. Dans une telle situation, il serait évidemment inexact de prétendre que l'activité  $j$  est concentrée (en termes *absolus*) dans la zone  $i$ .

De même, si l'activité  $j$  est d'importance mineure dans l'économie ( $p_{\bullet j}$  petit), il se peut, même quand  $QL_{ij} > 1$ , que la part de l'emploi de l'activité  $j$  dans l'emploi total de la zone  $i$  ( $p_{j/i\bullet}$ ) ne soit pas importante. En effet,

$$QL_{ij} = \frac{\text{Fraction de l'emploi total de la zone } i \text{ appartenant à la branche } j}{\text{Fraction de l'emploi total global appartenant à la branche } j}$$

$$QL_{ij} = \frac{p_{j/i\bullet}}{p_{\bullet j}} = \frac{x_{ij}/x_{i\bullet}}{x_{\bullet j}/x_{\bullet\bullet}}$$

de sorte que si  $p_{\bullet j}$  est petit, il est possible que  $QL_{ij} > 1$  même si  $p_{j/i\bullet}$  est petit, pourvu que  $p_{\bullet j}$  soit encore plus petit. Dans une telle situation, il serait évidemment inexact de prétendre que la zone  $i$  est spécialisée (en termes *absolus*) dans l'activité  $j$ .

Correctement interprétés, les quotients de localisation peuvent servir notamment à l'analyse descriptive de données d'emploi (voir Lemelin et Polèse, 1993).

NOTE : Il est mathématiquement impossible que  $QL_{ik} > 1$  pour toutes les zones  $i$  en même temps (ou, symétriquement que  $QL_{ik} < 1$  pour toutes les zones  $i$  en même temps). En effet,

puisque  $QL_{ik} = \frac{p_{i/\bullet k}}{p_{i\bullet}}$ , cela impliquerait que  $p_{i/\bullet k} > p_{i\bullet}$  pour chaque  $i$ , de sorte que l'on aurait

$\sum_i p_{i/\bullet k} > \sum_i p_{i\bullet}$ , ce qui est manifestement impossible, étant donné que les deux sommations doivent être égales à 1.

De même, il est mathématiquement impossible que  $QL_{kj} > 1$  pour toutes les activités  $j$  en même temps (ou, symétriquement, que  $QL_{kj} < 1$  pour toutes les activités  $j$  en même temps). En effet,

puisque  $QL_{kj} = \frac{p_{j/k\bullet}}{p_{\bullet j}}$ , cela impliquerait que  $p_{j/k\bullet} > p_{\bullet j}$  pour chaque  $j$ , de sorte que l'on aurait

$\sum_j p_{j/k\bullet} > \sum_j p_{\bullet j}$ , ce qui est manifestement impossible, étant donné que les deux termes de la

comparaison doivent être égaux à 1.

Il est utile de se rappeler ces règles : si l'on obtient un tel résultat, c'est qu'il y a une erreur dans les calculs...

### ESTIMATION DE L'EMPLOI EXPORTATEUR AU MOYEN DU QUOTIENT DE LOCALISATION

On utilise aussi les quotients de localisation dans le cadre de la théorie de la base économique (Polèse, 1994, p. 125-138), pour estimer l'emploi « exportateur » (pour un exemple, voir Polèse et Stafford, 1982). Vu la rareté de données sur les échanges interrégionaux, cette possibilité est attrayante. Mais l'estimation de l'emploi exportateur au moyen des quotients de localisation repose sur des hypothèses plutôt restrictives (Isserman, 1980, p. 157) :

1. La productivité du travail est égale entre villes ou régions.
2. L'absorption (utilisation locale) du produit par emploi dans l'économie locale est égale entre villes ou régions <sup>8</sup>.
3. Il n'y a pas d'importations ou d'exportations nettes de l'ensemble du pays.
4. La demande locale s'approvisionne en priorité auprès des producteurs locaux ; cela implique qu'il n'y a pas de flux croisés entre villes ou régions (« cross-hauling »).

Sous ces conditions, on peut interpréter l'excédent du quotient de localisation par rapport à 1,0 comme une mesure de l'emploi exportateur. Plus exactement,  $EXP_{ij}$ , l'emploi « exportateur » de la branche  $j$ , qui appartient à la *base économique* de la région  $i$ , peut alors s'estimer au moyen de la formule

$$EXP_{ij} = \begin{cases} x_{ij} \frac{(QL_{ij} - 1)}{QL_{ij}}, & \text{si } QL_{ij} > 1 \\ 0 & \text{autrement} \end{cases}$$

Par exemple,  $EXP_{31} = 45 \times \frac{2,5 - 1}{2,5} = 27$  des 45 emplois de la branche  $B1$  dans la zone  $Z3$ .

La fraction de  $x_{ij}$  qui appartient à l'emploi exportateur est la fraction de  $QL_{ij}$  qui excède 1. Quand  $QL_{ij} < 1$ , il n'y a pas d'exportations de l'activité  $j$  à partir de la région  $i$  et, en conséquence, l'emploi exportateur est nul.

Pour comprendre plus facilement la signification de ce calcul, on substitue  $QL_{ij}$  et on simplifie, de façon à obtenir

$$EXP_{ij} = x_{ij} - \left( \frac{x_{i\bullet}}{x_{\bullet\bullet}} \right) x_{\bullet j} = x_{ij} - p_{i\bullet} x_{\bullet j}, \text{ si } x_{ij} > p_{i\bullet} x_{\bullet j}$$

<sup>8</sup> Isserman (1980) et Norcliffe (1983) utilisent le terme « consommation » pour désigner l'utilisation à la fois par la demande finale et par la demande intermédiaire. Le terme « absorption » semble plus exact.

Par exemple,  $EXP_{31} = 45 - \left(\frac{180}{1200}\right)120 = 27$

On voit alors que l'emploi exportateur est la différence entre la valeur observée  $x_{ij}$  et la valeur hypothétique que prendrait le chiffre de l'emploi si la région  $i$  produisait seulement « sa part » de  $j$  (c'est-à-dire  $p_{i\bullet}$ , auquel cas le quotient de localisation  $QL_{ij}$  serait égal à 1).

Pour voir comment interviennent les hypothèses énoncées précédemment, récrivons la formule sous la forme suivante :

$$EXP_{ij} = \left[ \left( \frac{x_{ij}}{x_{\bullet j}} \right) - \left( \frac{x_{i\bullet}}{x_{\bullet\bullet}} \right) \right] x_{\bullet j} = (p_{i\bullet j} - p_{i\bullet}) x_{\bullet j}, \text{ si } p_{i\bullet j} > p_{i\bullet}$$

Par exemple,  $EXP_{31} = \left[ \left( \frac{45}{120} \right) - \left( \frac{180}{1200} \right) \right] 120 = 27$

La **première hypothèse** concerne le rapport  $p_{i\bullet j}$  ou  $\frac{x_{ij}}{x_{\bullet j}}$  ; ce rapport est la part de la région  $i$  dans l'emploi de l'activité  $j$  ; la première hypothèse permet de considérer ce rapport comme une approximation de la part de la région dans la production du bien  $j$ .

La **seconde hypothèse** concerne le rapport  $p_{i\bullet}$  ou  $\frac{x_{i\bullet}}{x_{\bullet\bullet}}$  ; ce rapport est la part de la région  $i$  dans l'emploi total ; la seconde hypothèse permet de considérer ce rapport comme une approximation de la part de la région dans l'utilisation totale (absorption) du bien  $i$ .

Les deux autres hypothèses permettent d'interpréter la différence comme la part de l'emploi national de la branche  $j$  qui appartient à la base économique de la région  $i$ . Ainsi, la **troisième hypothèse** dit que les importations et exportations internationales sont nulles : il s'ensuit que l'identité

<i>Production</i> + <i>Importations des autres régions</i> + <i>Importations internationales</i>	=	<i>Absorption</i> + <i>Exportations aux autres régions</i> + <i>Exportations internationales</i>
--	---	--

devient

$$\boxed{\begin{array}{c} \textit{Production} \\ + \\ \textit{Importations des autres régions} \end{array}} = \boxed{\begin{array}{c} \textit{Absorption} \\ + \\ \textit{Exportations aux autres régions} \end{array}}$$

c'est-à-dire

$$\boxed{\textit{Production} - \textit{Absorption}} = \boxed{\textit{Exportations nettes aux autres régions}}$$

lorsque l'excédent de la production sur l'absorption est positif.

La **quatrième hypothèse** enfin dit que, s'il y a des exportations vers les autres régions, il n'y a pas d'importations en provenance d'autres régions, et vice-versa. Par conséquent, quand les exportations nettes sont positives, elles sont égales aux exportations brutes.

Après avoir estimé l'emploi exportateur de chaque branche, il suffit de faire la somme pour obtenir un estimé de la « base » économique de la région  $i$  (on dit aussi emploi *basique*) :

$$\text{Base exportatrice} = \sum_{j \text{ lorsque } QL_{ij} > 1} EXP_{ij}$$

Dans le modèle de la base économique, on fait l'hypothèse que le rapport

$$\theta = \frac{\sum_j x_{ij}}{\sum_j EXP_{ij}}$$

est constant. Le modèle prédit que pour chaque emploi exportateur qui se crée (ou qui disparaît), l'emploi total augmente (ou diminue) de  $\theta$  emplois. Le facteur  $\theta$  s'appelle le *multiplicateur de la base économique* <sup>9</sup>.

## 1-2.2 L'analyse de décomposition additive et multiplicative des variations

### PRINCIPE

L'analyse « shift-share » est un cas particulier d'une technique plus générale, l'analyse de décomposition des variations <sup>10</sup>. L'analyse de décomposition des variations peut s'appliquer en

<sup>9</sup> On a proposé de multiples variantes du quotient de localisation pour relâcher les hypothèses très contraignantes sur lesquelles repose la méthode.

<sup>10</sup> L'article, bien connu en sciences régionales, de Williamson (1965) en donne un autre exemple : il propose une décomposition de l'évolution dans le temps d'une mesure d'inégalité interrégionale.



principe à tout écart entre deux valeurs observées d'une même variable. Il peut s'agir de deux observations d'un même objet à des moments différents dans le temps, ou d'observations effectuées sur deux objets distincts.

L'analyse de décomposition des variations consiste à décomposer l'écart entre deux valeurs d'une mesure en une somme de termes (décomposition additive) ou en un produit de facteurs (décomposition multiplicative). Une telle décomposition est toujours une tautologie du genre

$$x - y = (x - a) + (a - b) + (b - c) + (c - y)$$

ou

$$x/y = (x/a) (a/b) (b/c) (c/y)$$

c'est-à-dire

$$\log x - \log y = (\log x - \log a) + (\log a - \log b) + (\log b - \log c) + (\log c - \log y)$$

L'utilité de la décomposition vient donc de l'interprétation que l'on peut donner aux termes d'une décomposition additive ou aux facteurs d'une décomposition multiplicative. Cette interprétation repose sur un modèle, qui demeure souvent implicite. Le langage utilisé (« effet ceci », « facteur cela ») laisse parfois percer des connotations de causalité qui ne sont pas toujours justifiées.

#### **APPLICATION À L'ANALYSE « SHIFT-SHARE »**

Réf. : Page-Patton, ch.9 ; Coffey et Polèse (1988) ; Polèse, 1994, p. 349-357

L'analyse « shift-share »<sup>11</sup> est une méthode d'analyse de décomposition bien connue des praticiens des sciences régionales. Elle consiste à décomposer la variation de l'emploi d'une ville ou d'une région. Nous allons examiner successivement la méthode de décomposition de la variation de l'emploi d'une activité, puis celle de la variation de l'emploi d'un ensemble d'activités.

#### ***Décomposition de la variation de l'emploi d'une activité***

Pour illustrer l'analyse shift-share, nous allons nous servir de l'exemple numérique fictif suivant.

---

<sup>11</sup> Jayet (1993, p. 29-34) emploie l'expression « analyse structurelle-géographique ». Je préfère encore l'expression de Bonnet (1995) : « analyse structurelle-résiduelle ».

### Emploi par zone et par branche

BRANCHE	An 1				An 2			
	B1	B2	B3	Total	B1	B2	B3	Total
ZONE								
Z1	48	325	287	660	24	388	300	712
Z2	27	185	148	360	11	173	200	384
Z3	45	90	45	180	25	99	52	176
Total	120	600	480	1200	60	660	552	1272

### Variation de l'emploi par zone et par branche entre l'An 1 et l'An 2

BRANCHE	Différences				Taux de variation			
	B1	B2	B3	Total	B1	B2	B3	Total
ZONE								
Z1	-24	63	13	52	-50,00%	19,38%	4,53%	7,88%
Z2	-16	-12	52	24	-59,26%	-6,49%	35,14%	6,67%
Z3	-20	9	7	-4	-44,44%	10,00%	15,56%	-2,22%
Total	-60	60	72	72	-50,00%	10,00%	15,00%	6,00%

Penchons-nous sur la variation de l'emploi de la branche *B1* dans la zone *Z2*.

L'analyse shift-share est basée sur la comparaison de trois scénarios :

- Quelle aurait été la variation si l'emploi de *B1* en *Z2* avait évolué au même taux que l'emploi total (toutes branches et toutes zones) ?
  - Taux = 6 %
  - Nombre = 6 % de 27 = 1,62
- Quelle aurait été la variation si l'emploi de *B1* en *Z2* avait évolué au même taux que l'emploi de l'ensemble de la branche *B1* ?
  - Taux = -50 %
  - Nombre = -50 % de 27 = -13,50
- Quelle a été la variation observée de l'emploi de *B1* en *Z2* ?
  - Taux = -59,26 %
  - Nombre = -59,26 % de 27 = -16

La comparaison de ces trois scénarios conduit à la décomposition additive suivante :

- Effet national = scénario 1 :
  - Taux = 6 %
  - Nombre = 6 % de 27 = 1,62
- Effet proportionnel (ou sectoriel) = écart entre scénario 2 et scénario 1 :
  - Taux = -50 % - 6 % = -56 %

– Nombre =  $-56\%$  de  $27 = -15,12 = -13,5 - 1,62$

3. Effet résiduel (ou régional) = écart entre scénario 3 et scénario 2 :

– Taux =  $-59,26\% - (-50\%) = -9,26\%$

– Nombre =  $-9,26\%$  de  $27 = -2,5 = -16 - (-13,5)$

On peut vérifier que la somme des trois « effets » est bien égale à la variation observée :

– Taux =  $6\% + (-56\%) + (-9,26\%) = -59,26\%$

– Nombre =  $1,62 + (-15,12) + (-2,5) = -16$

Cette méthode de décomposition peut se formaliser à l'aide de la notation suivante :

$x_{ij}^t$	l'emploi de la branche $j$ dans la région $i$ au temps $t$
$x_{\bullet j}^t = \sum_i x_{ij}^t$	l'emploi de la branche $j$ dans l'ensemble des régions au temps $t$
$x_{\bullet\bullet}^t = \sum_i \sum_j x_{ij}^t$	l'emploi de toutes les branches dans l'ensemble des régions au temps $t$

On a l'identité suivante :

$$\frac{x_{ij}^t}{x_{ij}^0} = \left( \frac{x_{ij}^t}{x_{ij}^0} - \frac{x_{\bullet j}^t}{x_{\bullet j}^0} \right) + \left( \frac{x_{\bullet j}^t}{x_{\bullet j}^0} - \frac{x_{\bullet\bullet}^t}{x_{\bullet\bullet}^0} \right) + \frac{x_{\bullet\bullet}^t}{x_{\bullet\bullet}^0}$$

Dénotons les taux d'accroissement par

$$r_{ij} = \frac{x_{ij}^t}{x_{ij}^0} - 1$$

L'identité précédente peut alors se récrire en termes des taux d'accroissement

$$\left( \frac{x_{ij}^t}{x_{ij}^0} - 1 \right) = \left[ \left( \frac{x_{ij}^t}{x_{ij}^0} - 1 \right) - \left( \frac{x_{\bullet j}^t}{x_{\bullet j}^0} - 1 \right) \right] - \left[ \left( \frac{x_{\bullet j}^t}{x_{\bullet j}^0} - 1 \right) - \left( \frac{x_{\bullet\bullet}^t}{x_{\bullet\bullet}^0} - 1 \right) \right] - \left( \frac{x_{\bullet\bullet}^t}{x_{\bullet\bullet}^0} - 1 \right)$$

c'est-à-dire

$$r_{ij} = (r_{ij} - r_{\bullet j}) + (r_{\bullet j} - r_{\bullet\bullet}) + r_{\bullet\bullet}$$

ou alors en nombres d'emplois

$$r_{ij} x_{ij}^0 = (r_{ij} - r_{\bullet j}) x_{ij}^0 + (r_{\bullet j} - r_{\bullet\bullet}) x_{ij}^0 + r_{\bullet\bullet} x_{ij}^0$$

Dans cette décomposition,

- $r_{\bullet\bullet} x_{ij}^0$  est l'effet national (« national share effect ») : c'est l'accroissement qui se serait réalisé si l'emploi de la branche  $j$  dans la région  $i$  avait augmenté au même taux que l'emploi total au pays (scénario 1) ;
- $(r_{\bullet j} - r_{\bullet\bullet}) x_{ij}^0$  est l'effet sectoriel ou effet de déplacement proportionnel (« proportional shift effect ») : c'est l'accroissement *supplémentaire* (positif ou négatif) de l'emploi qui se serait réalisé si l'emploi de la branche  $j$  dans la région  $i$  avait augmenté au même taux que l'emploi de la branche  $j$  dans l'ensemble du pays (c'est donc la différence entre le scénario 2 et le scénario 1) ; l'effet sectoriel se rattache à la question de savoir si, comparée au reste de l'économie, la branche  $j$  est dynamique, si elle jouit d'une croissance accélérée ;
- $(r_{ij} - r_{\bullet j}) x_{ij}^0$  est l'effet régional ou effet de déplacement différentiel (« differential shift effect ») : c'est l'écart résiduel entre l'accroissement observé et l'accroissement résultant de l'application de l'effet de part et l'effet de déplacement proportionnel.

La somme de l'effet de déplacement proportionnel et de l'effet de déplacement différentiel est l'effet de déplacement total ou net (« total shift » ou « net shift ») pour une activité  $j$  dans une région  $i$ .

Le déplacement différentiel est souvent interprété comme une mesure de la *compétitivité* de la branche  $j$  de la région  $i$  (« competitive effect »). Cette utilisation est fort contestable. En effet, supposons que l'emploi de la branche  $j$  croît plus rapidement dans la région  $i$  que dans la région  $k$  ; cela veut-il dire que la *production* de cette branche croît plus rapidement dans la région  $i$  que dans la région  $k$  ? Pas nécessairement, si la proportion entre la main-d'oeuvre et les autres facteurs de production varie d'une région à l'autre et dans le temps (en réponse aux changements des prix relatifs) <sup>12</sup>. Nous verrons dans un moment que ce n'est pas la seule raison de douter de la validité de l'interprétation de l'effet régional comme mesure de la compétitivité.

### **Décomposition de la variation de l'emploi total d'une région**

En faisant la sommation sur l'ensemble des branches de chacun des trois termes de la décomposition, on obtient

---

<sup>12</sup> Une démonstration formelle de cette proposition exigerait que l'on développe un modèle d'équilibre général de deux économies.

$$\sum_j r_{ij} x_{ij}^0 = \sum_j (r_{ij} - r_{\cdot j}) x_{ij}^0 + \sum_j (r_{\cdot j} - r_{\cdot\cdot}) x_{ij}^0 + \sum_j r_{\cdot\cdot} x_{ij}^0$$

où

$$\sum_j r_{\cdot\cdot} x_{ij}^0 = r_{\cdot\cdot} \sum_j x_{ij}^0 \text{ est l'effet national}$$

$$\sum_j (r_{\cdot j} - r_{\cdot\cdot}) x_{ij}^0 \text{ est l'effet de structure}$$

$$\sum_j (r_{ij} - r_{\cdot j}) x_{ij}^0 \text{ est l'effet régional}$$

Les quatre tableaux qui suivent complètent les calculs pour notre exemple.

## Analyse shift-share par branche

### Branche B1

ZONE	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	2,88	-26,88	0	-24	6,00%	-56,00%	0,00%	-50,00%
Z2	1,62	-15,12	-2,5	-16	6,00%	-56,00%	-9,26%	-59,26%
Z3	2,7	-25,2	2,5	-20	6,00%	-56,00%	5,56%	-44,44%

### Branche B2

ZONE	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	19,5	13	30,5	63	6,00%	4,00%	9,38%	19,38%
Z2	11,1	7,4	-30,5	-12	6,00%	4,00%	-16,49%	-6,49%
Z3	5,4	3,6	0,0	9	6,00%	4,00%	0,00%	10,00%

### Branche B3

p	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	17,22	25,83	-30,05	13	6,00%	9,00%	-10,47%	4,53%
Z2	8,88	13,32	29,8	52	6,00%	9,00%	20,14%	35,14%
Z3	2,7	4,05	0,25	7	6,00%	9,00%	0,56%	15,56%

### Ensemble des branches (classification à trois branches)

ZONE	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	39,6	11,95	0,45	52	6,00%	1,81%	0,07%	7,88%
Z2	21,6	5,6	-3,2	24	6,00%	1,56%	-0,89%	6,67%
Z3	10,8	-17,55	2,75	-4	6,00%	-9,75%	1,53%	-2,22%

L'effet de structure est interprété comme l'effet de la structure économique ou industrielle de la région, c'est-à-dire de la composition de sa production industrielle (« industry mix effect »). L'effet régional est souvent interprété comme une mesure de la compétitivité de la région  $i$  : cette interprétation appelle les mêmes réserves que précédemment. Mais il y a plus grave.

En effet, lorsqu'on décompose la variation du niveau *global* de l'emploi d'une région, l'importance de l'effet de structure dépend du niveau d'agrégation de la classification des

activités. Et comme l'effet régional est calculé de façon résiduelle, ce dernier dépend aussi du niveau d'agrégation. Pour une région particulière, le passage d'une classification à une autre peut augmenter ou diminuer l'effet régional, de sorte qu'il peut en résulter des interversions de rang entre les régions : une région qui paraissait plus compétitive qu'une autre dans une classification donnée peut paraître moins compétitive dans une autre classification ! Cela restreint encore la validité de l'interprétation de l'effet régional comme mesure de la compétitivité d'une région.

Ce phénomène est illustré au moyen de l'agrégation des branches *B1* et *B2*. On peut constater que les résultats de la décomposition pour l'ensemble des branches sont différents de ceux qui ont été obtenus précédemment avec une classification à trois branches.

#### Branches *B1* et *B2* agrégées

ZONE	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	22,38	-22,38	39	39	6,00%	-6,00%	10,46%	10,46%
Z2	12,72	-12,72	-28	-28	6,00%	-6,00%	-13,21%	-13,21%
Z3	8,1	-8,1	-11	-11	6,00%	-6,00%	-8,15%	-8,15%

#### Ensemble des branches (classification à deux branches)

ZONE	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	39,6	<b>3,45</b>	<b>8,95</b>	52	6,00%	<b>0,52%</b>	<b>1,36%</b>	7,88%
Z2	21,6	<b>0,6</b>	<b>1,8</b>	24	6,00%	<b>0,17%</b>	<b>0,50%</b>	6,67%
Z3	10,8	<b>-4,05</b>	<b>-10,75</b>	-4	6,00%	<b>-2,25%</b>	<b>-5,97%</b>	-2,22%

Cette faiblesse touche également l'analyse de décomposition de la variation du niveau d'une seule activité, puisque l'activité en question est définie selon une classification qui comporte inévitablement un certain degré d'agrégation. En d'autres mots, s'agissant d'une seule activité, l'effet régional calculé contient l'effet structurel associé aux déplacements entre les sous-branches qui composent l'activité considérée. Il est donc abusif, même lorsqu'on ne considère qu'une seule branche, d'interpréter l'effet résiduel comme une mesure de la compétitivité.

### 1-2.3 La mesure de la croissance (le calcul du taux de variation dans le temps)

D'une certaine manière, l'analyse d'une série chronologique pose le problème de la multidimensionnalité, dont il sera question plus loin. En effet, le concept de « croissance » ou de « variation dans le temps » comporte de multiples dimensions. Car sauf dans le cas où le taux de variation est constant (la croissance est uniforme), le concept a autant de dimensions que l'on a d'observations sur la croissance.

#### TAUX DE CROISSANCE PAR PÉRIODE

Considérons par exemple l'évolution de l'indice des prix à la consommation (IPC) au Canada de 1984 à 1992 <sup>13</sup> :

1984	92,4
1985	96,0
1986	100,0
1987	104,4
1988	108,6
1989	114,0
1990	119,5
1991	126,2
1992	128,1

Entre chaque moment et le suivant, on peut calculer un *taux de croissance pour cet intervalle*. Ainsi, entre 1984 et 1985, le taux de croissance pour la période a été de

$$\frac{96,0 - 92,4}{92,4} = 0,039$$

soit 3,9 %. Avec neuf observations consécutives (de 1984 à 1992), on peut calculer huit taux de croissance par période :

de...	à...	taux
1984	1985	0,039
1985	1986	0,042
1986	1987	0,044
1987	1988	0,040
1988	1989	0,050
1989	1990	0,048
1990	1991	0,056
1991	1992	0,015

---

<sup>13</sup> Moyenne annuelle non désaisonnalisée (Statistique Canada 62-210).



En général, avec une série de  $T+1$  observations, de la période 0 à la période  $T$ ,

$$x_0, x_1, x_2, \dots, x_t, \dots, x_T$$

on peut calculer  $T$  valeurs du taux de croissance  $r_t$  d'une période par rapport à la précédente :

$$r_t = \frac{x_t - x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - \frac{x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - 1$$

Par exemple, pour  $t = 1985$ ,

$$r_{1985} = \frac{96,0}{92,4} - 1 = 0,039 = 3,9 \%$$

où  $t$  varie de 1 à  $T$ . On dit alors qu'entre  $t-1$  et  $t$ ,  $x$  s'est accru de  $R_t$  pourcent, où

$$R_t = 100 \times r_t$$

Note : si  $x_t$  est inférieur à  $x_{t-1}$ , il y a eu *décroissance* :  $r_t$  est *néglatif*. On parle alors de *croissance négative*.

La série

$$r_1, r_2, \dots, r_t, \dots, r_T$$

décrit l'évolution dans le temps de la variable  $x$ . En effet, si on inverse la formule de calcul du taux de croissance par période, on obtient

$$x_t = (1+r_t) x_{t-1}$$

Mais on a également

$$x_{t-1} = (1+r_{t-1}) x_{t-2}$$

de sorte que

$$x_t = (1+r_t) (1+r_{t-1}) x_{t-2}$$

et, en procédant à des substitutions successives, on a

$$x_t = (1+r_t) (1+r_{t-1}) \dots (1+r_2) (1+r_1) x_0$$

En somme, si l'on connaît les  $r_t$  et  $x_0$ , on peut reconstituer la série des  $x_t$ . Il est donc bien vrai que la série des taux de croissance par période

$$r_1, r_2, \dots, r_t, \dots, r_T$$

décrit l'évolution dans le temps de la variable  $x$ <sup>14</sup>. Mais comment peut-on résumer cette évolution en un seul chiffre ?

D'une certaine manière, les  $T$  valeurs de  $r_t$  constituent autant de *dimensions* de l'évolution dans le temps de la variable  $x$ . C'est en ce sens que la mesure de la croissance s'apparente au problème de la multidimensionnalité et de la construction des nombres indices.

### MOYENNE DES TAUX DE CROISSANCE PAR PÉRIODE

Pour résumer l'évolution de la variable  $x$  dans le temps, on peut prendre la moyenne arithmétique des taux de croissance par période :

$$\frac{1}{T} \sum_{t=1}^T r_t = \frac{r_1 + r_2 + \dots + r_t + \dots + r_T}{T}$$

Dans le cas de l'IPC entre 1984 et 1992, la moyenne des taux de croissance par période est de 0,042 (c'est-à-dire 4,2 %).

Cette façon de résumer l'évolution de la variable  $x$  dans le temps a cependant l'inconvénient de ne tenir aucun compte de la variabilité des taux de croissance. Or, pour un même taux moyen, plus les taux sont uniformes, plus la croissance cumulée est forte<sup>15</sup>. Nous n'allons pas démontrer cette propriété : un exemple suffira à l'illustrer. Comparons les deux séries suivantes :

100, 110, 121

et

100, 100, 120

Dans les deux cas, la moyenne des taux de croissance par période est égale à 0,1 (c'est-à-dire 10 %). Pourtant, la croissance cumulée sur les deux périodes est de 21 % dans le premier cas, mais de seulement 20 % dans le second. La question plus générale que cela soulève est donc la suivante : dans quelle mesure la moyenne des taux par période est-elle représentative de l'évolution d'une série lorsque les taux par période sont variables ?

---

<sup>14</sup> En pratique cependant, on utilise pour les taux de croissance des valeurs arrondies, de sorte que l'on ne pourrait pas reconstituer avec exactitude la série originale.

<sup>15</sup> Si la croissance cumulée dépend de la variabilité des taux, elle ne dépend pas en revanche de l'ordre chronologique entre les différents taux de croissance. On peut le voir en constatant dans la formule suivante que la valeur de  $x_t$  demeure inchangée si l'on change l'ordre des facteurs du membre de droite :

$$x_t = (1+r_t) (1+r_{t-1}) \dots (1+r_2) (1+r_1) x_0.$$

### CALCUL D'UN TAUX DE CROISSANCE EXPONENTIEL

Le taux de croissance exponentiel est une autre façon de résumer l'évolution de la variable  $x$  dans le temps. On peut le définir de deux façons équivalentes.

La première définition du taux de croissance exponentiel fait appel à la moyenne géométrique<sup>16</sup>. C'est le taux de croissance  $r$  obtenu à partir de la moyenne géométrique des facteurs<sup>17</sup> de croissance par période :

$$1 + r = \left[ (1 + r_T)(1 + r_{T-1}) \cdots (1 + r_2)(1 + r_1) \right]^{1/T} = \sqrt[T]{(1 + r_T)(1 + r_{T-1}) \cdots (1 + r_2)(1 + r_1)}$$

C'est-à-dire, sous forme logarithmique :

$$\log(1 + r) = \frac{1}{T} \sum_{t=1}^T \log(1 + r_t)$$

La mesure précédente, au contraire, était donnée par la moyenne *arithmétique* des taux de croissance par période. On peut simplifier le calcul du taux de croissance exponentiel en exploitant la relation

$$x_T = (1 + r_T)(1 + r_{T-1}) \cdots (1 + r_2)(1 + r_1) x_0$$

Or selon la définition de la moyenne géométrique

$$(1 + r)^T = (1 + r_T)(1 + r_{T-1}) \cdots (1 + r_2)(1 + r_1)$$

de sorte que

$$x_T = (1 + r)^T x_0$$

où  $x_T$  et  $x_0$  sont des valeurs connues et où  $r$  est l'inconnue. En développant cette formule, on arrive à une méthode de calcul du taux de croissance exponentiel.

$$(1 + r)^T = \frac{x_T}{x_0}$$

$$\log(1 + r)^T = \log\left(\frac{x_T}{x_0}\right)$$

$$T \log(1 + r) = \log(x_T) - \log(x_0)$$

---

<sup>16</sup> À propos de la moyenne géométrique et de ses applications, voir Wonnacott et Wonnacott (1991, p. 755).

<sup>17</sup> Noter la distinction entre le *taux* de croissance  $r$  et le *facteur* de croissance  $(1 + r)$ .

$$\log(1+r) = \frac{\log(x_T) - \log(x_0)}{T}$$

$$1+r = \text{antilog} \left( \frac{\log(x_T) - \log(x_0)}{T} \right)$$

où antilog  $z = e^z$  ou  $10^z$ , selon qu'on a pris le logarithme népérien de base  $e (= 2,71828\dots)$  ou le logarithme commun de base 10.

$$r = \text{antilog} \left( \frac{\log(x_T) - \log(x_0)}{T} \right) - 1$$

C'est-à-dire, avec les logarithmes communs,

$$r = 10^{\frac{\log x_T - \log x_0}{T}} - 1$$

et avec les logarithmes népériens.

$$r = e^{\frac{\ln x_T - \ln x_0}{T}} - 1 = \exp \left( \frac{\ln x_T - \ln x_0}{T} \right) - 1$$

Par exemple, le taux de croissance exponentiel de l'IPC de 1984 à 1992 se calcule de la façon suivante :

$$x_0 = 92,4 \text{ et } \log_e x_0 = 4,526126979$$

$$x_T = 128,1 \text{ et } \log_e x_T = 4,852811209$$

$$T = 8$$

$$r = \exp \left( \frac{4,852811209 - 4,526126979}{8} \right) - 1 = 0,042, \text{ c'est-à-dire } 4,2 \%$$

Il y a une autre définition du taux de croissance exponentiel, qui contient sa propre interprétation. En effet, nous venons de voir que le taux de croissance exponentiel  $r$  est la solution de l'équation

$$x_T = (1+r)^T x_0$$

Le taux de croissance exponentiel peut donc être vu comme un taux hypothétique : c'est la réponse à la question « si la variable  $x$  avait évolué à un taux par période constant, à quel taux eût-il fallu qu'elle crût pour que sa valeur terminale soit égale à la valeur terminale observée ? ». Par définition, le taux de croissance exponentiel est donc le taux de croissance par période uniforme qui donne la même croissance cumulée que la série

$$r_1, r_2, \dots, r_t, \dots, r_T$$

En ce sens, le taux de croissance exponentiel résume, comme la moyenne arithmétique des taux par période, l'évolution de la variable  $x$  dans le temps. Mais le taux de croissance exponentiel ne tient compte que de la première et de la dernière valeur de la série. Cela constitue un inconvénient si la première valeur de la série,  $x_0$ , ou la dernière,  $x_T$ , est exceptionnelle (hors tendance) : dans ce cas, le taux de croissance exponentiel pourra être trompeur.

En tant qu'indice, c'est en quelque sorte un indice tronqué, puisqu'il n'utilise pas toute l'information disponible. Les indices tronqués peuvent parfois avoir leur utilité et leurs faibles exigences en information leur font certainement des adeptes : qu'on pense au « Big Mac cost-of-living index » du périodique *The Economist*.

#### **ENTRE DEUX MAUX...**

En tant qu'indices de l'évolution chronologique d'une variable, la moyenne des taux par période et le taux de croissance exponentiel présentent tous deux des inconvénients. Il s'agit donc de choisir le moindre mal. Lequel ? Cela dépend évidemment de l'utilisation que l'on veut en faire. Par exemple, est-il important que le taux de croissance retenu permette de « prédire » avec exactitude (c'est-à-dire de reproduire) la valeur terminale à partir de la valeur initiale ? À cet égard, le taux de croissance exponentiel semble préférable. Ou est-il plus important au contraire que le taux de croissance retenu soit représentatif de la tendance ? Dans ce cas, rien ne garantit *a priori* que les valeurs initiale et terminale ne soient pas hors tendance et que, par conséquent, le taux exponentiel ne soit pas trompeur.

Mais nous avons vu que, de son côté, la moyenne des taux par période a le défaut de ne pas tenir compte de la variabilité des taux par période. Ce défaut est-il important ? Ça dépend. Ainsi, dans le cas de l'IPC 1984-1992 au Canada, si le taux de croissance avait été constant, égal à la moyenne des taux de croissance par période, la valeur de l'IPC en 1992 aurait été de 128,16, au lieu de 128,1. L'écart est minime (six centièmes d'un pourcent !).

L'écart serait-il plus considérable sur un grand nombre de périodes, avec un taux de croissance plus volatil ? Par exemple, la valeur de fermeture de l'indice boursier Standard & Poor's 500 était de 470,34 le 17 février 1994, et de 656,37 le 9 février 1996, 499 séances de marché plus

tard <sup>18</sup>. La moyenne des taux de croissance par période (d'une séance de marché sur la précédente) a été de 0,0007 (0,07 %), avec un écart type de 0,0057 (0,57 %), ce qui représente une volatilité importante (le coefficient de variation est de 8,37). Or quelle aurait été la valeur de clôture le 9 février 1996 si l'indice avait crû à un taux constant égal à la moyenne des taux par période ? Elle aurait été de 661,76... Encore une fois, l'écart est mince (0,82 %) : dans ce cas encore, la moyenne des taux de croissance par période est assez représentative de la tendance.

### AJUSTEMENT D'UNE COURBE DE TENDANCE

Quoi qu'il en soit, il y a une manière plus exacte de résumer l'évolution d'une série : c'est de lui ajuster une courbe de tendance. On choisit pour cela un *modèle* de l'évolution de la série. Par exemple, on peut prendre le modèle linéaire simple

$$x_t = a + bt$$

ou le modèle exponentiel simple

$$x_t = a b^t$$

qui devient un modèle linéaire simple lorsqu'on prend les logarithmes :

$$\log x_t = \log a + (\log b) t$$

Le modèle exponentiel simple peut être considéré comme une version améliorée du taux de croissance exponentiel. En effet, on peut établir un parallèle entre le paramètre  $a$  et la valeur initiale  $x_0$  et entre  $b$  et le facteur de croissance exponentielle  $(1+r)$ . L'estimation du modèle exponentiel simple consiste à rechercher les valeurs de  $x_0^*$  ( $= a^*$ ) et de  $(1+r^*)$  ( $= b^*$ ) qui feront que les valeurs  $x_t^*$  « prédites » par la relation

$$x_t^* = x_0^* (1+r^*)^t = a^* (b^*)^t$$

« colleront » le mieux possible aux valeurs observées. De cette manière, on aura calculé un taux de croissance exponentiel qui risque moins d'être influencé par des valeurs hors tendance <sup>19</sup>.

---

<sup>18</sup> Source : [www.fortitude.com/data.htm](http://www.fortitude.com/data.htm).

<sup>19</sup> Voir Wonnacott et Wonnacott (1992, p. 513-523).

Plus généralement, après avoir choisi un modèle, on choisit les valeurs des paramètres qui rapprochent le modèle le plus possible de la réalité observée, en appliquant les méthodes statistiques appropriées. La régression linéaire est une technique qui permet de trouver les « meilleures » valeurs pour  $a^*$  et  $b^*$ . Nous y reviendrons dans la troisième partie de cet ouvrage.

### **QUE RETENIR ?**

Nous examinerons bientôt la question de la multidimensionnalité dans la mesure. La mesure de la croissance, apparemment simple, illustre déjà quelques-unes des difficultés soulevées par la multidimensionnalité. Nous avons examiné deux façons simples de résumer en un taux unique l'évolution d'une variable dans le temps (la moyenne des taux de croissance par période et le taux de croissance exponentiel) : chacune des deux mesures présente des inconvénients. Nous avons évoqué une autre technique, l'ajustement d'une courbe de tendance, qui semble exempte des défauts des deux autres. Mais l'utilisation de cette technique a son prix : une plus grande complexité, une moindre transparence et des calculs plus lourds. Voilà bien une illustration des difficultés qui font que la mesure n'est pas qu'une science, elle est aussi un art :

- Il n'y a guère de mesure parfaite.
- Les mesures moins imparfaites sont généralement plus complexes et plus lourdes à utiliser.

## CHAPITRE 1-3

### LE PROBLÈME DE LA MULTIDIMENSIONNALITÉ : LES NOMBRES INDICES

---

<b>Plan</b>	
1-3.0 Problématique de la multidimensionnalité	2
1-3.1 Illustration No 1 : les indices de prix	3
1-3.1.1 L'indice de Laspeyres	3
1-3.1.2 L'indice de Paasche	7
1-3.1.3 Utilisations des indices de prix	8
1-3.1.4 Indices de prix et coût de la vie	12
1-3.1.5 Conclusion : indices et modèles	15
1-3.2 Illustration No 2 : l'indicateur de développement humain (IDH) du Programme des Nations-Unies pour le Développement (PNUD)	16
1-3.2.1 Dimensions du concept et variables	16
1-3.2.2 Ajustement du PIB réel par habitant	17
1-3.2.3 Calcul de l'IDH	18
1-3.2.4 Réflexions sur l'IDH	19
1-3.3 Pour en savoir plus...	23
Les indicateurs urbains	23
Un indice de statut socio-économique (Renaud et Mayer)	24
Et plus...	26



## CHAPITRE 1-3

### LE PROBLÈME DE LA MULTIDIMENSIONNALITÉ : LES NOMBRES INDICES

#### 1-3.0 Problématique de la multidimensionnalité

Nous avons déjà cité Gilles (1994, p. 24), qui, se référant au schéma classique de Lazarsfeld (1971), définit l'opérationnalisation comme le fait de « soumettre les concepts, par l'analyse, à un processus qui les transforme en dimensions, puis en indicateurs permettant de les observer, de les mesurer ou de les quantifier ». La réflexion théorique qui conduit à identifier les dimensions d'un concept relève de la discipline ou du champ d'étude pertinent. Ici, nous prenons acte du fait que la plupart des concepts ont des dimensions multiples et nous examinons les implications de ce fait quant à la construction de mesures associées à ces concepts.

Il ne manque pas d'exemples de concepts ayant des dimensions multiples, à chacune desquelles on peut associer une mesure distincte :

1. Le concept politique de « niveau de satisfaction à l'égard du gouvernement » peut se décomposer en plusieurs dimensions, comme « satisfaction quant à la politique économique », « satisfaction quant à la politique sociale », « satisfaction quant à la politique étrangère », etc.
2. Le concept de « coût de la vie » peut se décomposer en « coût de l'alimentation », « coût du logement », etc.

Lorsqu'un même concept comporte plusieurs dimensions, mais qu'on veut néanmoins traiter le concept comme un tout, il faut trouver une façon de combiner les mesures associées aux différentes dimensions en une seule mesure qui les résume toutes. Le problème, pour ainsi dire, est d'additionner des bananes et des oranges.

La manière la plus répandue d'aborder ce problème consiste à construire un *nombre indice*. Un nombre indice est une règle (une formule) pour combiner plusieurs mesures en un seul chiffre. Les différentes mesures qui entrent dans la composition de l'indice se rapportent à différentes dimensions d'un concept ; l'indice lui-même est utilisé comme mesure globale du concept étudié. En général, il n'y a pas d'indice qui soit parfaitement fiable (c'est-à-dire dont les variations reflètent des variations réelles). Il est même souvent difficile de construire un indice

valide (qui mesure bien ce que l'on veut mesurer). C'est l'un des points importants que cherche à illustrer ce chapitre, à l'aide de deux exemples : les indices de prix (notamment l'indice des prix à la consommation) et l'Indicateur de Développement Humain du Programme des Nations Unies pour le Développement (PNUD).

### 1-3.1 Illustration No 1 : les indices de prix

Réf. : Wonnacott et Wonnacott (1991, chap. 22) ; Statistique Canada (1996 et 1997)

Le concept « niveau général des prix » a autant de dimensions qu'il y a de prix différents, c'est-à-dire autant de dimensions qu'il y a de biens ou services différents. Un indice de prix sert à comparer les prix d'un groupe de biens et services à deux moments ou plus rarement en deux endroits différents.

L'exemple suivant illustre le problème <sup>1</sup>. Il s'agit des prix de l'alimentation pour un pays fictif où le régime alimentaire se compose uniquement de trois aliments : le steak, le poivre et le pain.

Biens	Prix (\$)		Indices des prix des biens individuels
	1980	1985	
	$p_{0i}$	$p_{ti}$	$p_{ti} / p_{0i}$
Steak (kg)	4,85 \$	6,60 \$	1,36
Poivre (g)	0,07 \$	0,07 \$	1,00
Pain (kg)	1,10 \$	1,32 \$	1,20

Les trois indices de prix individuels constituent les trois dimensions du concept « prix de l'alimentation ». Comment peut-on combiner ces trois indices de prix individuels en une mesure unique ? Les deux indices de prix les plus couramment utilisés sont l'indice de *Laspeyres* et l'indice de *Paasche*.

#### 1-3.1.1 L'INDICE DE LASPEYRES

Un indice de *Laspeyres* mesure les variations du niveau général des prix en comparant le coût d'acquisition d'un panier représentatif de biens et services à deux moments différents dans le temps. La plupart des indices des prix à la consommation sont des indices de Laspeyres.

<sup>1</sup> Source : adapté de Wonnacott et Wonnacott (1991, p. 753).

Notation :

- $p_{ti}$  prix du bien  $i$  à la période  $t$
- $p_{0i}$  prix du bien  $i$  à la période 0
- $q_{0i}$  quantité du bien  $i$  achetée par un ménage typique à la période 0

Dans l'indice de Laspeyres, le panier représentatif est défini par les  $q_{0i}$ , les quantités achetées par un ménage typique durant la période 0, appelée période *de référence*, ou année *de base*. Supposons qu'il y a  $n$  biens et services dans ce panier représentatif. En utilisant l'opérateur sommation, le coût d'acquisition du panier représentatif peut s'écrire :

Coût du panier de référence aux prix de la période 0

$$= p_{01}q_{01} + p_{02}q_{02} + \dots + p_{0n}q_{0n} = \sum_{i=1}^n p_{0i}q_{0i}$$

Coût du panier de référence aux prix de la période  $t$

$$= p_{t1}q_{01} + p_{t2}q_{02} + \dots + p_{tn}q_{0n} = \sum_{i=1}^n p_{ti}q_{0i}$$

L'indice de Laspeyres compare ces deux valeurs,  $\sum_{i=1}^n p_{0i}q_{0i}$  et  $\sum_{i=1}^n p_{ti}q_{0i}$ . La valeur de l'indice de Laspeyres pour la période  $t$  est donnée par le rapport

$$I_t^L = \frac{\sum_{i=1}^n p_{ti}q_{0i}}{\sum_{i=1}^n p_{0i}q_{0i}}$$

L'indice de Laspeyres est donc le rapport du coût du panier représentatif lorsque les prix sont ceux de la période  $t$  sur son coût lorsque les prix sont ceux de la période 0.

Illustrons le calcul à l'aide des données fictives de l'exemple précédent.

Biens	Données				Calculs	
	Prix (\$)		Quantités		Coût du panier	
	1980	1985	1980	1985	1980	1985
	$p_{0i}$	$p_{ti}$	$q_{0i}$	$q_{ti}$	$p_{0i} q_{0i}$	$p_{ti} q_{0i}$
Steak (kg)	4,85 \$	6,60 \$	23	18	111,55 \$	151,80 \$
Poivre (g)	0,07 \$	0,07 \$	57	85	3,99 \$	3,99 \$
Pain (kg)	1,10 \$	1,32 \$	36	45	39,60 \$	47,52 \$
					155,14 \$	203,31 \$

L'indice de Laspeyres des prix de 1985, base 1980, est donné par

$$I_t^L = \frac{203,31}{155,14} = 1,31$$

NOTE : Les indices de prix publiés, notamment par Statistique Canada, sont généralement exprimés en pourcentage, de sorte que l'on aurait habituellement

$$I_t^L = 100 \times \frac{203,31}{155,14} = 131$$

Nous ignorons cette convention ici, pour alléger l'écriture des formules.

Nous allons montrer que l'indice de Laspeyres est une *moyenne pondérée* des indices de prix des biens individuels. Développons la formule de l'indice de Laspeyres :

$$I_t^L = \frac{\sum_{i=1}^n p_{ti} q_{0i}}{\sum_{i=1}^n p_{0i} q_{0i}} = \frac{\sum_{i=1}^n p_{ti} q_{0i}}{\sum_{k=1}^n p_{0k} q_{0k}} = \sum_{i=1}^n \left( \frac{p_{ti} q_{0i}}{\sum_{k=1}^n p_{0k} q_{0k}} \right) = \sum_{i=1}^n \left[ \left( \frac{q_{0i}}{\sum_{k=1}^n p_{0k} q_{0k}} \right) p_{ti} \right]$$

$$I_t^L = \sum_{i=1}^n \left( \frac{p_{0i} q_{0i}}{\sum_{k=1}^n p_{0k} q_{0k}} \right) \left( \frac{p_{ti}}{p_{0i}} \right) = \text{moyenne pondérée des indices de prix des biens}$$

L'indice de Laspeyres peut donc s'interpréter comme une *moyenne pondérée* des indices de prix  $\left( \frac{p_{ti}}{p_{0i}} \right)$  des biens individuels, où le poids du bien  $i$

$$w_{0i} = \frac{p_{0i} q_{0i}}{\sum_{k=1}^n p_{0k} q_{0k}}$$

est la part de ce bien dans le budget du ménage typique à la période 0.

Voici comment cela s'applique à notre exemple.

Biens	Données				Indice de prix de Laspeyres		
	Prix			Quant.	Coût du panier	Poids	Calcul de l'indice
	1980	1985	Rapport	1980	1980	1980	1985
	$p_{0i}$	$p_{ti}$	$\left(\frac{p_{ti}}{p_{0i}}\right)$	$q_{0i}$	$p_{0i} q_{0i}$	$w_{0i}$	$w_{0i} \left(\frac{p_{ti}}{p_{0i}}\right)$
Steak (kg)	4,85 \$	6,60 \$	1,36	23	111,55 \$	0,719	0,978
Poivre (g)	0,07 \$	0,07 \$	1,00	57	3,99 \$	0,026	0,026
Pain (kg)	1,10 \$	1,32 \$	1,20	36	39,60 \$	0,255	0,306
					155,14 \$	1,000	1,310

En réalité, le calcul de l'indice des prix à la consommation de Statistique Canada est plus compliqué, pour plusieurs raisons :

- Pour maintenir la représentativité de l'indice, le panier de référence est mis à jour entre les changements d'année de base (=100) de l'indice.
- Les poids sont calculés avec les quantités de chaque panier de référence et les prix d'une autre période.
- Le calcul de l'indice est fait de manière à ce que ses valeurs s'enchaînent sans rupture lorsqu'on passe d'un panier de référence au suivant.

Par exemple, en 2003, l'indice des prix à la consommation est calculé au moyen du panier de référence de 1996, évalué aux prix de décembre 1997 et son année de base est 1992 (1992 = 100). Pour plus de détails, voir le *Document de référence de l'indice des prix à la consommation*, No 62-553 au catalogue de Statistique Canada.

### 1-3.1.2 L'INDICE DE PAASCHE

Quelle est la différence entre un indice de Laspeyres et un indice de *Paasche* ? C'est le choix du panier représentatif : dans l'indice de Paasche, le panier représentatif est donné par les dépenses d'un ménage typique, non pas à la période de base 0, mais à la période  $t$ . La valeur de l'indice de Paasche pour la période  $t$  est donc donnée par

$$I_t^P = \frac{\sum_{i=1}^n p_{ti} q_{ti}}{\sum_{k=1}^n p_{0k} q_{tk}}$$

Reprenons notre exemple pour illustrer le calcul.

Biens	Données				Calculs	
	Prix (\$)		Quantités		Coût du panier	
	1980	1985	1980	1985	1980	1985
	$p_{0i}$	$p_{ti}$	$q_{0i}$	$q_{ti}$	$p_{0i} q_{ti}$	$p_{ti} q_{ti}$
Steak (kg)	4,85 \$	6,60 \$	23	18	87,30 \$	118,80 \$
Poivre (g)	0,07 \$	0,07 \$	57	85	5,95 \$	5,95 \$
Pain (kg)	1,10 \$	1,32 \$	36	45	49,50 \$	59,40 \$
					142,75 \$	184,15 \$

L'indice de Paasche des prix de 1985, base 1980, est donné par

$$I_t^P = \frac{184,15}{142,75} = 1,29$$

L'indice de Paasche peut lui aussi s'interpréter comme une *moyenne pondérée* des indices de

prix  $\left( \frac{p_{ti}}{p_{0i}} \right)$  des biens individuels. Dans le cas de l'indice de Paasche, le poids du bien  $i$

$$w_{ti} = \frac{p_{0i} q_{ti}}{\sum_{k=1}^n p_{0k} q_{tk}}$$

est la part de ce bien dans le budget fictif d'un ménage qui aurait consommé les quantités  $q_{ti}$  du panier de consommation d'un ménage typique à la période  $t$ , avec les prix  $p_{0i}$  de la période de base. Voici le développement qui conduit à ce résultat :

$$I_t^P = \frac{\sum_{i=1}^n p_{ti} q_{ti}}{\sum_{i=1}^n p_{0i} q_{ti}} = \frac{\sum_{i=1}^n p_{ti} q_{ti}}{\sum_{k=1}^n p_{0k} q_{tk}} = \sum_{i=1}^n \left( \frac{p_{ti} q_{ti}}{\sum_{k=1}^n p_{0k} q_{tk}} \right) = \sum_{i=1}^n \left[ \left( \frac{q_{ti}}{\sum_{k=1}^n p_{0k} q_{tk}} \right) p_{ti} \right]$$

$$I_t^P = \sum_{i=1}^n \left( \frac{p_{0i} q_{ti}}{\sum_{k=1}^n p_{0k} q_{tk}} \right) \left( \frac{p_{ti}}{p_{0i}} \right) = \text{moyenne pondérée des indices de prix des biens}$$

Voici comment cela s'applique à notre exemple.

Biens	Données				Indice de prix de Paasche		
	Prix		Rapport	Quant.	Coût du panier	Poids	Calcul de l'indice
	1980	1985			NAP <sup>2</sup>	NAP	1985
	$p_{0i}$	$p_{ti}$	$\left( \frac{p_{ti}}{p_{0i}} \right)$	$q_{ti}$	$p_{0i} q_{ti}$	$w_{ti}$	$w_{ti} \left( \frac{p_{ti}}{p_{0i}} \right)$
Steak (kg)	4,85 \$	6,60 \$	1,36	18	87,30 \$	0,612	0,832
Poivre (g)	0,07 \$	0,07 \$	1,00	85	5,95 \$	0,042	0,042
Pain (kg)	1,10 \$	1,32 \$	1,20	45	49,50 \$	0,347	0,416
					142,75 \$	1,000	1,290

L'indice de Paasche est moins utilisé que l'indice de Laspeyres, parce qu'il exige de refaire à chaque période une enquête auprès des ménages pour définir le panier représentatif de biens et services (les  $q_{ti}$ ).

### 1-3.1.3 UTILISATIONS DES INDICES DE PRIX

#### Utilisation comme dégonfleurs

Les indices de prix sont souvent utilisés comme *dégonfleurs*<sup>3</sup> dans l'analyse des séries temporelles. Lorsque des chiffres sont exprimés en unités monétaires et qu'ils se rapportent à

des années différentes, il est difficile de les comparer si les prix ont évolué. Pour rendre les chiffres comparables, il faut leur appliquer une correction pour tenir compte de l'évolution des prix <sup>4</sup>.

Considérons par exemple l'évolution des dépenses personnelles de consommation au Canada de 1991 à 1999 <sup>5</sup> :

Année	Dép.pers.
1991	398 314
1992	411 167
1993	428 219
1994	445 857
1995	460 906
1996	480 427
1997	510 695
1998	531 169
1999	560 954

Ces données sont exprimées en « dollars courants », c'est-à-dire qu'elles ne sont pas corrigées pour tenir compte de l'évolution des prix. Comment peut-on comparer, par exemple, le 480 milliards de 1996 avec le 411 milliards de 1992, alors que les prix ont sensiblement augmenté entre ces deux années ? On peut le faire en utilisant un indice de prix comme dégonfleur. Cet indice de prix pourrait être un indice de Laspeyres, un indice de Paasche, ou tout autre indice, à condition qu'il soit approprié à la nature des chiffres à dégonfler. En l'occurrence par exemple, il ne serait pas approprié d'appliquer un indice des prix industriels à une série de chiffres sur les dépenses personnelles de consommation. L'indice des prix le plus approprié ici est l'indice de prix des dépenses personnelles calculé par Statistique Canada <sup>6</sup>.

Une fois que l'on a obtenu un indice de prix approprié, il suffit, pour obtenir des chiffres comparables, de diviser chaque chiffre par l'indice des prix de l'année correspondante : si  $x_t$  est un chiffre en dollars courants, alors

---

<sup>2</sup> Ne s'applique pas, c'est-à-dire que les chiffres de la colonne ne s'appliquent à aucune année.

<sup>3</sup> L'utilisation de ce mot se justifie par le fait que dans l'histoire économique récente, les périodes d'augmentation des prix (inflation) ont été plus longues et plus fréquentes que les périodes de diminution des prix.

<sup>4</sup> Voir aussi « L'utilisation de l'IPC pour comparer des valeurs en dollars à travers le temps » (Statistique Canada, *Votre guide d'utilisation de l'indice des prix à la consommation*, 62-557-XPB, p. 12.

<sup>5</sup> Dépenses personnelles en biens et services de consommation (millions de \$) selon les comptes nationaux (Statistique Canada, *L'observateur économique canadien, Supplément statistique historique 2001/02*, No 11-210-XPB).

<sup>6</sup> Statistique Canada, *L'observateur économique canadien, Supplément statistique historique 2001/02*, No 11-210-XPB. Cet indice est spécifiquement construit pour les dépenses personnelles dans le produit intérieur brut. Il est donc préférable en l'occurrence à l'indice des prix à la consommation.



$$y_t = \frac{x_t}{I_t}$$

est un chiffre *dégonflé*, exprimé en dollars constants de l'année de référence de l'indice. On dit encore que  $y_t$  est une donnée en valeur « réelle », par opposition à  $x_t$ , qui est en valeur « nominale ».

Dans le cas d'un indice exprimé en pourcentage, la formule précédente devient  $y_t = 100 \frac{x_t}{I_t}$ .

Lorsque l'on veut exprimer un montant en dollars constants d'une année  $\theta$  autre que l'année de base, il faut appliquer la formule plus générale  $y_t = x_t \frac{I_\theta}{I_t}$ .

Voici comment on pourrait utiliser un indice des prix comme dégonfleur pour les dépenses personnelles de consommation.

**Dépenses personnelles de consommation au Canada de 1991 à 1999, en millions de dollars courants, avec l'indice des prix correspondant**

	Indice de prix des dép. pers. dans le PIB <sup>7</sup>	Dép. pers.
1991	91,0	398 314
1992	92,5	411 167
1993	94,6	428 219
1994	95,6	445 857
1995	96,8	460 906
1996	98,4	480 427
1997	100,0	510 695
1998	101,2	531 169
1999	102,9	560 954

Source : Statistique Canada, *L'observateur économique canadien*, Supplément statistique historique 2001/02, No 11-210-XPB.

En dollars constants de 1997, la valeur des dépenses personnelles de consommation de 1999 est égale à<sup>8</sup>

$$100 \times \frac{560\,954}{102,9} = \frac{560\,954}{1,029} = 545\,145$$

<sup>7</sup> Cet indice de prix est différent de l'indice des prix à la consommation.

<sup>8</sup> Le chiffre publié par Statistique Canada pour les dépenses personnelles de 1999 en dollars constants de 1997 est plutôt de 545 162 millions de \$. La différence est simplement due au fait que nous utilisons une valeur arrondie de l'indice de prix.

Questions pièges :

- Quelle est la valeur des dépenses personnelles de 1992 en dollars constants de 1997 ?
- Quelle est la valeur des dépenses personnelles de consommation de 1999 en dollars constants de 1992 ?

### **Utilisation aux fins d'indexation**

Une autre utilisation bien connue des indices de prix, très proche de la précédente, est l'*indexation*<sup>9</sup>. L'indexation vise à maintenir la valeur d'un paiement d'année en année lorsque les prix évoluent.

Exemples :

- Dans la convention collective entre un employeur et le syndicat de ses employés, il arrive que les salaires soient fixés pour la première année seulement ; pour les années suivantes, on s'entend pour ajuster les salaires en fonction de l'évolution des prix à la consommation, selon une formule d'indexation convenue.
- Les pensions versées par l'État à certaines catégories de citoyens (fonctionnaires à la retraite, personnes âgées, ...) sont souvent fixées de la même manière : le montant initial est fixé et le montant des années subséquentes est calculé selon une formule d'indexation.

Il y a plusieurs formules d'indexation ; la plupart de celles qui sont utilisées concrètement sont des formules d'indexation partielle et certaines d'entre elles sont passablement compliquées. Voyons ici la méthode d'indexation la plus simple. Un montant  $m_0$  fixé à l'année zéro est indexé d'année en année au moyen de la formule

$$m_t = I_t m_0$$

où  $m_t$  est le montant indexé pour l'année  $t$  et  $I_t$  est un indice de prix approprié ayant pour base l'année zéro ( $I_0 = 1$ )<sup>10</sup>. Si l'indice de prix a pour base une année autre que l'année zéro, la formule devient simplement

$$m_t = m_0 \frac{I_t}{I_0}$$

---

<sup>9</sup> Voir aussi « L'utilisation de l'IPC pour comparer des valeurs en dollars à travers le temps » (Statistique Canada, *Votre guide d'utilisation de l'indice des prix à la consommation*, 62-557-XPB, p. 11).

<sup>10</sup> Dans le cas d'un indice exprimé en pourcentage ( $I_0 = 100$ ), la formule devient  $m_t = \frac{I_t m_0}{100}$ .

Par exemple, un montant de 35 000 \$ en dollars de 1997, indexé pour l'année 1999, est égal à  
 $35\,000 \$ \times 1,029 = 36\,015 \$$

Et un montant de 35 000 \$ en dollars de 1994, indexé pour l'année 1999, est égal à

$$35\,000 \$ \times \frac{1,029}{0,956} = 37\,673 \$$$

#### 1-3.1.4 INDICES DE PRIX ET COÛT DE LA VIE

Les indices de prix sont-ils des mesures fiables du coût de la vie ?

Rappel : une variable est *fiable* lorsque les variations dans la mesure correspondent à des variations véritables.

La réponse est non : Statistique Canada affirme que « L'indice des prix à la consommation n'est pas un indice du coût de la vie, bien que l'on ait trop souvent tendance à l'appeler ainsi » (1996, p. 3). Voyons pourquoi.

Considérons le cas particulier où un indice de prix est utilisé pour indexer un revenu (salaire, pension, etc.). Si l'indice utilisé est fiable, le revenu indexé

$$m_t = m_0 \frac{I_t}{I_0}$$

permettra à la personne qui le reçoit de vivre à la période  $t$ , avec un revenu de  $m_t$ , aussi bien qu'elle vivait (ou qu'elle aurait vécu) à la période 0 avec un revenu de  $m_0$ .

D'abord, il est clair que l'indexation en direct n'est pas praticable, puisque la valeur de l'indice de prix n'est connue qu'après un certain délai. Il s'ensuit qu'un revenu ne peut être indexé qu'avec un retard (il est vrai que l'indexation pourrait être rétroactive). Nous parlons donc d'une situation théorique. De plus, les préférences et les choix de consommation sont propres à chacun. Ce serait le plus grand des hasards si le panier de référence, qui est le reflet du comportement général, correspondait à la consommation d'un individu ou d'une famille en particulier.

Mais laissons de côté ces objections pratiques pour nous demander si, au plan théorique, les indices de Laspeyres et de Paasche peuvent être parfaitement fiables. Peuvent-ils mesurer avec exactitude les variations du coût de la vie ? Non. Quand les prix augmentent, l'indice de *Laspeyres sur-estime* l'accroissement du coût de la vie, tandis que l'indice de *Paasche* le *sous-*

*estime*. Inversement, lorsque les prix baissent, l'indice de *Laspeyres sous-estime* l'importance de la diminution du coût de la vie, tandis que l'indice de *Paasche* la *sur-estime*. Il s'ensuit qu'aussi bien en temps d'inflation qu'en temps de déflation (diminution générale des prix), un revenu indexé à l'aide d'un indice de Laspeyres sera un peu plus élevé que nécessaire, tandis qu'un revenu indexé à l'aide d'un indice de Paasche ne sera pas tout à fait suffisant.

On peut démontrer ce qui précède au moyen de la théorie économique de la consommation. Mais on peut aussi le voir intuitivement. Les indices de prix de Laspeyres et de Paasche ne mesurent pas avec exactitude l'évolution du coût de la vie parce qu'ils ne tiennent pas compte du comportement adaptatif des consommateurs. Voici comment.

Quand les prix changent, ils ne changent pas tous dans la même proportion : certains prix augmentent ou diminuent plus que d'autres (en fait, il arrive que des prix évoluent en sens contraire, les uns augmentant tandis que d'autres diminuent). Il s'ensuit que les prix *relatifs* (c'est-à-dire les prix des biens les uns par rapport aux autres) changent <sup>11</sup>.

Exemple :

Supposons que le café coûte 20 ¢ la tasse et le thé, 10 ¢ : le prix du café est deux fois celui du thé. Supposons que le prix du café augmente de 35 % et celui du thé, de 50 % : le nouveau prix du café est de 27 ¢, ce qui est 1,8 fois le nouveau prix du thé (15 ¢). Même si les deux prix ont augmenté, le prix du café a diminué *relativement* à celui du thé, parce qu'il est passé de 2 à 1,8 fois le prix du thé.

Lorsque deux biens sont des *substituts* et que leurs prix relatifs changent, comment réagissent les consommateurs ? Ils s'adaptent en réorientant une partie de leur consommation vers celui des deux biens dont le prix relatif a diminué. Voyons cela à l'aide d'un exemple numérique. Nous distinguons trois biens : le thé, le café, et un bien composite qui comprend tout le reste et qui est désigné par l'étiquette *Et cœtera*.

---

<sup>11</sup> La traduction française de Wonnacott et Wonnacott (1991) utilise l'expression « prix relatifs » dans un sens différent au tableau 22-1 (p. 752), pour désigner les indices de prix des biens individuels.

		Thé	Café	<i>Et cœtera</i>	Total
<b>Année 0</b>	Quantité	1250	800	10000	
<b>Année 0</b>	Prix	0,40	0,50	1,00	
	Dépense	500	400	10000	10900
<b>Année t</b>	Prix	4,00	0,50	1,00	
	Dépense	5000	400	10000	15400

Calculons l'Indice de Laspeyres :

$$I_t^L = \frac{(4,00 \times 1250) + (0,50 \times 800) + (1,00 \times 10000)}{(0,40 \times 1250) + (0,50 \times 800) + (1,00 \times 10000)} = 1,41284$$

Revenu indexé pour l'année  $t = 1,41284 \times 10900 = 15400$

Ce revenu indexé permet à un consommateur typique d'acheter pour l'année  $t$  les mêmes quantités que ce consommateur typique achetait durant l'année 0. Donc, avec un revenu indexé et les nouveaux prix, le consommateur typique pourrait vivre exactement comme avant.

En voyant le prix très élevé du thé, la plupart d'entre nous choisiraient de dépenser autrement leur revenu indexé : acheter un peu moins de thé, boire du café à la place, et acheter un peu plus d'*et cœtera*. Nous ferions ce choix parce qu'il nous permettrait de vivre mieux *en nous adaptant* aux changements des prix relatifs par la substitution. Mais si le revenu indexé permet de vivre mieux, c'est qu'il est plus élevé que nécessaire (si vous n'êtes pas encore convaincus, demandez-vous quelle situation vous préféreriez : le revenu et les prix de l'année 0 ou le revenu indexé et les prix de l'année  $t$ ).

En termes techniques, ce qui fait que l'indice de Laspeyres n'est pas exact, c'est que son numérateur, qui ne tient pas compte des possibilités de substitution, est trop grand.

Évidemment, les changements de prix relatifs sont rarement aussi dramatiques que ceux de l'exemple numérique. Le principe de la substitution demeure cependant le même et c'est en vertu de ce principe que l'on peut affirmer qu'un revenu indexé selon un indice de Laspeyres est toujours un peu plus élevé que nécessaire.

Ce biais est-il important ? Pour le Canada, Crawford (1993) estime que le biais dû à la substitution est de l'ordre de 0,1 ou 0,2 % (c'est-à-dire que l'IPC surestimerait l'augmentation du coût de la vie d'un ou deux dixièmes de point de pourcentage par année). Aux États-Unis, l'évaluation plus récente de la *CPI Commission* est d'environ un d'un demi-point de pourcentage par année (Boskin *et al.*, 1998). Ce n'est pas tout à fait négligeable. Par contre, les études

citées identifient d'autres biais, plus importants, comme le biais dû aux changements dans la qualité des biens. L'effet de l'*ensemble* des biais est évalué à un demi-point de pourcentage par Crawford, et par la *CPI Commission* à quelque chose de l'ordre de 0,8 à 1,6 points de pourcentage.

Contrairement à ce qui se passe avec l'indice de Laspeyres, un revenu indexé selon un indice de Paasche n'est jamais tout à fait suffisant. On peut le voir en écrivant l'indice sous la forme suivante :

$$I_t^P = \frac{\sum_{i=1}^n p_{ti} q_{ti}}{\sum_{k=1}^n p_{0k} q_{tk}}$$

Dans ce cas, c'est le *dénominateur* de l'indice qui est trop grand, parce qu'il ne tient pas compte des possibilités de substitution : hypothétiquement confronté aux prix  $p_{0k}$ , le ménage typique choisirait des quantités différentes des  $q_{tk}$ .

C'est pour tenter de corriger les distorsions de l'un et l'autre indice que l'on a proposé l'indice « idéal » de *Fisher*, qui est la moyenne géométrique de l'indice de Laspeyres et de celui de Paasche :

$$I_t^F = \sqrt{I_t^L \times I_t^P}$$

### 1-3.1.5 CONCLUSION : INDICES ET MODÈLES

Nous avons constaté que les indices de prix de Laspeyres et de Paasche n'étaient pas des mesures parfaitement fiables de l'évolution du coût de la vie. Nous l'avons démontré en mettant en évidence les lacunes du *modèle de comportement* sous-jacent à ces indices : ce modèle ne tient pas compte du comportement adaptatif des consommateurs.

Il y a une leçon à tirer de cet examen des indices de Laspeyres et de Paasche. Tous les indices reposent sur un modèle sous-jacent, tantôt explicite, tantôt implicite <sup>12</sup>. Les indices ne sont des mesures fiables qu'en autant que les modèles sous-jacents représentent fidèlement la réalité

---

<sup>12</sup> On dit qu'un indice est « exact » par rapport à un modèle donné lorsqu'il est parfaitement cohérent avec ce modèle.

que l'on cherche à mesurer. Avant d'utiliser un indice, l'analyste doit donc s'interroger sur le modèle sous-jacent.

Or les indices de prix de Laspeyres et de Paasche sont « typiques » de beaucoup d'indices couramment utilisés dans toutes les disciplines. En effet, la moyenne pondérée est une forme d'indice très répandue, non seulement pour mesurer l'évolution des prix d'un groupe de biens, mais aussi pour mesurer toutes sortes de concepts aux dimensions multiples. Le calcul de tels indices est facile et leur interprétation, accessible : les poids des composantes représentent l'importance relative de chacune d'elles. Toutefois, les indices construits comme des moyennes pondérées, à cause de leur simplicité, reposent le plus souvent, comme les indices de Laspeyres et de Paasche, sur des hypothèses particulièrement restrictives. Par contre, les indices qui sont dérivés de modèles plus élaborés peuvent être fort compliqués <sup>13</sup>.

## **1-3.2 Illustration No 2 : l'indicateur de développement humain (IDH) du Programme des Nations-Unies pour le Développement (PNUD)**

### **1-3.2.1 DIMENSIONS DU CONCEPT ET VARIABLES**

Réf. : Programme des Nations-Unies pour le Développement (PNUD), *Rapport mondial sur le développement humain*, (annuel, à partir de 1990); disponible sur le site web :  
<http://hdr.undp.org/>

L'indicateur de développement humain (IDH) du Programme des Nations-Unies pour le Développement (PNUD) est proposé comme indicateur de développement en remplacement (les plus modérés disent « en complément ») du produit intérieur brut (PIB) utilisé par le Fonds Monétaire International (FMI) et la Banque Mondiale. Car le PIB est fort critiqué comme mesure du développement, parce qu'il ignore plusieurs dimensions du développement humain. C'est pourquoi l'IDH comprend trois composantes (dimensions du concept de développement humain) :

- Longévité
- Savoir
- Niveau de vie

---

<sup>13</sup> Par exemple, l'indice de Törnqvist associé à la fonction translog (« transcendantale logarithmique »).

L'opérationnalisation de ces trois dimensions du concept de développement humain a conduit à choisir les variables suivantes :

- Longévité : espérance de vie à la naissance
- Savoir : taux d'alphabétisation des adultes et taux de scolarisation (tous niveaux confondus) <sup>14</sup> (2 variables)
- Niveau de vie : produit intérieur brut (PIB) réel par habitant, en dollars ajustés en fonction du coût de la vie (parité de pouvoir d'achat)

Pour chaque variable, on mesure le progrès réalisé pour atteindre un niveau maximum de l'indicateur, par rapport à la distance totale entre le niveau maximum et le niveau minimum <sup>15</sup>.

#### Maximums et minimums :

Les versions de l'IDH antérieures à 1994 prenaient pour valeurs maximums et minimums les niveaux les plus élevés et les plus bas observés cette année-là parmi les pays. Cela rendait impossibles les comparaisons d'année en année. La version courante de l'IDH prend pour minimums les valeurs les plus faibles observées au cours des trente dernières années <sup>16</sup> et pour maximums les valeurs les plus élevées que l'on prévoit pour les trente prochaines :

Variable	Maximum	Minimum
Espérance de vie	85 ans	25 ans
Taux d'alphabétisation	100 %	0 %
Taux de scolarisation	100 %	0 %
PIB réel/habitant	40 000 \$	100 \$

#### 1-3.2.2 AJUSTEMENT DU PIB RÉEL PAR HABITANT

##### a) Taux de change appliqué aux conversions monétaires

Pour que l'on puisse comparer les pays entre eux, les données du PIB réel par habitant doivent être exprimées dans une même unité monétaire. Les chiffres exprimés en Yens japonais, en Deutsche Marks ou en Colones costaricains sont donc tous convertis en dollars U.S. Cependant, pour faire cette conversion, on n'utilise pas simplement le taux de change ; à la place, on utilise un taux de conversion qui reflète le pouvoir d'achat relatif des monnaies.

---

<sup>14</sup> Jusqu'en 1994, c'était le *nombre moyen d'années* de scolarité. On a abandonné cette variable pour des raisons de disponibilité de données.

<sup>15</sup> Il s'agit donc de mesures relatives.

<sup>16</sup> La variable de revenu constitue une exception : sa valeur minimum devrait être de 200 \$ ; mais, pour pouvoir inclure dans les analyses les valeurs inférieures de la variante sexospécifique (voir ci-après) de la variable de revenu, le niveau minimum de PIB réel par habitant a été réduit à 100 \$.



Exemple :

Jusqu'à récemment, si le revenu moyen des japonais était converti en dollars U.S. au taux de change courant, ce revenu pouvait paraître très élevé. Mais le coût de la vie au Japon est beaucoup plus élevé qu'aux États-Unis lorsqu'on fait la comparaison en appliquant le taux de change courant. Pour pouvoir comparer le revenu moyen japonais au revenu moyen états-unien, il faut tenir compte de ce facteur.

Dans le même ordre d'idées, on dit qu'à 0,76 \$ U.S., le dollar Canadien est sous-évalué et qu'un taux d'environ 0,85 \$ U.S. refléterait mieux son pouvoir d'achat relatif <sup>17</sup>.

C'est pourquoi, aux fins de calcul de l'IDH, le niveau de vie est mesuré par le PIB réel par habitant exprimé en « PPA », c'est-à-dire en « parités de pouvoir d'achat ».

*b) Correction relative à l'effet décroissant des accroissements de revenu sur le développement humain*

De plus, « l'IDH met plus l'accent sur la suffisance que sur la satiété » (PNUD, 1994, p. 97). C'est pourquoi, en plus d'être corrigé pour tenir compte du coût de la vie, le PIB réel par habitant est aussi ajusté pour refléter l'hypothèse que les accroissements successifs du revenu per capita contribuent de moins en moins à l'épanouissement humain. L'application de ce principe se traduit depuis le Rapport de 1999 par une transformation logarithmique : l'indicateur du niveau de vie utilisé dans le calcul de l'IDH est donc le *logarithme* du PIB réel par habitant exprimé en « PPA », c'est-à-dire en « parités de pouvoir d'achat » <sup>18</sup>.

### 1-3.2.3 CALCUL DE L'IDH

Calcul de l'IDH pour le pays  $j$  (on trouve un exemple numérique du calcul dans PNUD, 1999, p. 60 ou dans PNUD, 2001, p. 239-240) :

1. Pour chacune des quatre variables, on calcule un sous-indice qui est le rapport du progrès réalisé sur le chemin à parcourir :

$$I_{ij} = \frac{x_{ij} - \min x_i}{\max x_i - \min x_i}$$

où

---

<sup>17</sup> Lafrance et Schembri (2002).

$x_{ij}$  est la valeur de l'indicateur  $i$  dans le pays  $j$  ;

$\max x_i$  est la valeur maximum de l'indicateur  $i$  ;

$\min x_i$  est la valeur minimum de l'indicateur  $i$ .

2. L'indicateur retenu pour le savoir est une moyenne pondérée des deux variables utilisées (taux d'alphabétisation et taux brut de scolarisation) ; le poids accordé à l'alphabétisation est le double de celui accordé à la scolarité :

$$I_{\text{savoir},j} = 0,67 \times I_{\text{alpha},j} + 0,33 \times I_{\text{scolar},j}$$

3. L'indicateur de développement humain est une moyenne arithmétique des indicateurs associés aux trois composantes :

$$I_j = \frac{1}{3} \sum_{i=1}^3 I_{ij}$$

#### 1-3.2.4 RÉFLEXIONS SUR L'IDH

##### 1. Fondements théoriques et éthiques

L'IDH s'inspire des idées d'Amartya Sen <sup>19</sup> sur la justice sociale (voir Sugden, 1993). Les concepts proposés par Sen sont difficiles à cerner et encore plus difficiles à traduire de façon opérationnelle. Il n'est pas évident non plus que l'IDH constitue une traduction empirique fidèle des concepts mis de l'avant par Sen. Pour cette raison, plusieurs font à l'IDH le reproche d'être dénué de fondement théorique. Les partisans de l'IDH admettent qu'il constitue une mesure imparfaite mais ils affirment que cette mesure est utile et qu'elle contribue à renouveler la réflexion sur le développement. L'utilité de l'IDH n'est pas reconnue par tous, loin de là. Le débat sur les fondements théoriques et la validité de l'IDH est bien vivant (Aturupane, Glewne et Iseman, 1994 ; Srinivasan, 1994 ; Streeten, 1994 ; Ravallion, 1997).

##### 2. Formes fonctionnelles et pondérations arbitraires

Lorsque d'aucuns affirment que l'IDH est dépourvu de fondements théoriques, ils disent en fait que le lien entre le concept théorique et son opérationnalisation n'est pas suffisamment solide.

---

<sup>18</sup> Jusqu'à 1998, le PIB réel PPA par habitant était ajusté suivant la soi-disant « formule d'Atkinson », éminemment critiquable (voir PNUD, 1999, p. 159 et Lemelin, 1999).

En particulier, le PNUD ne présente aucun argument pour fonder théoriquement le choix de donner un poids égal aux trois composantes de l'IDH, ni pour justifier la pondération utilisée pour calculer l'indicateur relatif à l'éducation (deux tiers-un tiers).

Or, lorsqu'on applique des pondérations arbitraires pour construire un indice ordinal (comme l'IDH), il s'ensuit que l'ordre que cet indice établit entre les observations (les pays) est lui aussi arbitraire, ce qui a pour effet de dépouiller l'indice de son statut de mesure.

Avant 1999, la forme de la relation entre le PIB réel par habitant et le PIB corrigé (« formule d'Atkinson ») était en contradiction avec le principe selon lequel, au-delà d'un certain seuil, l'accroissement du revenu contribue de moins en moins à l'épanouissement humain (Lemelin, 1999). La correction logarithmique appliquée dans le Rapport de 1999 constitue une nette amélioration, mais le choix de cette formule de correction semble néanmoins arbitraire.

### **3. Prise en compte des disparités à l'intérieur des pays**

Par ailleurs, l'IDH est une mesure moyenne et il peut cacher de fortes disparités à l'intérieur d'un pays, entre les régions, les sexes, les groupes raciaux ou les classes socio-économiques, par exemple. Pour obtenir une image plus fidèle de la réalité, l'approche la plus précise est évidemment de calculer la valeur de l'IDH séparément pour les différentes régions ou pour les différents groupes d'un pays. C'est ce que fait le PNUD dans certaines études de cas (PNUD, 1994, « Décompositions de l'IDH », p. 104-107). Il sied d'ajouter qu'au Mexique, le Conseil National de la Population (CONAPO) calcule la valeur de l'IDH par État et par « municipio »<sup>20</sup>.

Certains chercheurs ont calculé la valeur de l'IDH pour des régions particulières ou pour des groupes particuliers. Par exemple, à l'intérieur du Canada, la valeur de l'IDH pour le Québec est moins élevée que sa valeur pour l'ensemble du Canada. De même, en Ontario, l'IDH des franco-ontariens est inférieur à celui de la province dans son ensemble (essentiellement à cause d'un taux d'alphabétisation plus faible).

Quant aux inégalités économiques, elles étaient prises en compte, jusqu'en 1996, grâce au calcul d'un IDH ajusté en fonction de la répartition du revenu. Mais cette variante de l'IDH est obtenue de façon quelque peu mécanique, en multipliant l'IDH global d'un pays par un

---

<sup>19</sup> Amartya Sen a remporté le prix Nobel d'économie en 1998. Voir <http://www.nobel.se/announcement-98/economics98.html>.

<sup>20</sup> [http://www.conapo.gob.mx/m\\_en\\_cifras/principal.html](http://www.conapo.gob.mx/m_en_cifras/principal.html) (Población de México en cifras); dans « Menú de sección », voir « Índices de desarrollo humano ».

« coefficient de disparité » qui est le rapport de la part du revenu obtenu par les 20 % de la population situés au bas de l'échelle sur la part de revenu obtenue par les 20 % situés en haut de l'échelle. Depuis 1997, le PNUD tente de prendre en compte les inégalités économiques au moyen d'indicateurs complémentaires, les « Indicateurs de la pauvreté humaine » (IPH) <sup>21</sup>. L'année suivante, le PNUD a commencé à présenter deux indicateurs différents, l'un (IPH-1) pour les pays en développement et l'autre (IPH-2) pour les pays industrialisés.

En vue de calculer l'IDH séparément pour les femmes et pour les hommes, on dispose pour plusieurs pays de données par sexe sur l'espérance de vie, l'alphabétisation et la scolarisation, mais rarement sur le partage du PIB. Avant 1995, le PIB per capita ajusté pour les femmes était obtenu en multipliant le PIB per capita ajusté par un « rapport global revenu féminin-revenu masculin » ; ce rapport était lui-même obtenu en multipliant les deux rapports suivants :

- le rapport du taux de salaire des femmes dans l'industrie sur celui des hommes ;
- le rapport du taux de participation des femmes à la population active en dehors de l'agriculture sur celui des hommes.

Selon le PNUD, ce rapport revenu féminin-revenu masculin « sous-estime l'importance de la discrimination dans la mesure où les différences entre les femmes et les hommes sont généralement plus grandes dans l'agriculture et les services que dans l'industrie » (PNUD, 1994, p. 103).

Depuis le rapport de 1995, le PNUD a tenté de mieux prendre en compte les différences socio-économiques entre les sexes. Malgré la complication des formules utilisées, le principe derrière l'IDH « sexospécifique » (ISDH) <sup>22</sup> est simple. Chacune des variables qui entre dans le calcul de l'IDH est en fait une moyenne <sup>23</sup> entre la valeur de cette variable pour les hommes et sa valeur pour les femmes. Dans l'IDH sexospécifique, la moyenne arithmétique est remplacée par une moyenne harmonique <sup>24</sup>, une formule qui dégonfle la valeur de la moyenne selon le degré d'inégalité entre hommes et femmes.

---

<sup>21</sup> On trouve un énoncé de la théorie mathématique sous-jacente dans les *Notes techniques* du Rapport de 1997.

<sup>22</sup> Indicateur « sexospécifique » de développement humain; *gender related development index* (GDI) en anglais.

<sup>23</sup> Plus précisément, une moyenne pondérée, où les poids sont proportionnels aux nombres de personnes de chaque sexe.

<sup>24</sup> La formule utilisée est la suivante :

$$X = (p_f X_f^{1-\epsilon} + p_m X_m^{1-\epsilon})^{1/(1-\epsilon)}$$

où  $p_f$  est la proportion de femmes dans la population et  $p_m$ , la proportion d'hommes. Le paramètre  $\epsilon$  doit être non négatif et différent de 1. La valeur choisie par le PNUD est 2. Si  $X_f = X_m$ , alors  $X = X_f = X_m$  ; autrement, la valeur de  $X$  se situe quelque part entre les deux.

L'IDH sexospécifique tente d'améliorer l'IDH, mais il utilise lui aussi des formes fonctionnelles *a priori* et des pondérations arbitraires (encore que sa complexité puisse le faire paraître davantage « scientifique »). La raison fondamentale de cela est simplement qu'il n'y a pas de réponse unique à des questions comme « Comment faudrait-il mesurer l'inégalité ? » et « Quel poids devrions-nous donner à l'inégalité ? ».

#### **4. Dimensions ignorées**

Malgré les travaux exploratoires faits dans ce sens, les chercheurs du PNUD ne sont pas parvenus à proposer une façon satisfaisante de prendre en compte de la performance des pays en matière d'environnement.

Par ailleurs, l'IDH vise, conformément aux idées de Sen, à mesurer le développement des « capacités » des êtres humains. À cet égard, le taux de chômage devrait sans doute entrer en ligne de compte, puisqu'il n'est pas indifférent qu'un revenu soit gagné en salaire ou qu'il soit reçu en prestations sociales (comme l'ont démontré en particulier les études sur la relation entre la santé et le chômage). Le taux de chômage de longue durée est d'ailleurs pris en compte dans le calcul de l'Indice de la pauvreté humaine IPH-2, celui qui est appliqué aux pays industrialisés.

Dans le Rapport de 2003, basé sur les données de 2001, le Canada est huitième quant à l'IDH, après la Norvège, l'Islande, la Suède, l'Australie, les Pays-Bas, la Belgique et les États-Unis; le Canada était troisième en 2001 et premier jusqu'en 2000. Pour ce qui est de l'IPH-2, parmi les 17 pays à développement humain élevé pour lequel l'IPH-2 est calculé, le Canada se classe 12e rang quant à la pauvreté humaine, après la Suède, la Norvège, la Finlande, les Pays-Bas, le Danemark, l'Allemagne, le Luxembourg, la France, l'Espagne, le Japon et l'Italie.

#### **Que conclure ?**

L'IDH est un projet qui mobilise des ressources pour recueillir et organiser des données relatives à des dimensions du développement humain qui sont absentes des données strictement économiques. En attirant l'attention sur ces dimensions négligées du développement, l'IDH met en évidence les lacunes du point de vue strictement économique,

conduit à une compréhension plus juste de la situation et contribue à élargir la discussion sur le développement <sup>25</sup>. C'est aussi un instrument de mobilisation politique.

Mais jamais on ne trouvera de solution définitive au problème de la mesure du développement et du progrès social, puisque cette solution n'existe pas. Il serait même pernicieux que la complexité des calculs nous fasse succomber à l'illusion de « scientificité ». Néanmoins, il peut être utile de continuer à affronter le défi intellectuel de tenter de mesurer ce qui ne peut pas se mesurer, pourvu que cela conduise à une meilleure compréhension des limites de la mesure et à une amélioration des méthodes de mesure des réalités sociales.

### **1-3.3 Pour en savoir plus...**

En sciences sociales, et en études urbaines en particulier, la question de la construction d'indices est d'une grande actualité. Car, à mesure que s'estompe le souvenir des « Trente Glorieuses », la nécessité s'impose de gérer au plus serré. Pour les politiques sociales, cela veut dire bien identifier les besoins pour cibler les interventions avec précision et, ensuite, mesurer les résultats pour évaluer le succès des programmes. Ces tâches font appel à la mesure de réalités complexes comme la pauvreté, la qualité de vie, l'accessibilité au logement, etc. Bref, l'IDH du PNUD fait des petits...

Dans ce contexte, il est de la responsabilité des scientifiques de soumettre à l'examen critique les multiples indicateurs proposés, qui, tout comme l'IDH du PNUD, essaient tant bien que mal de mesurer ce qui n'est pas vraiment mesurable. Voici quelques suggestions pour le lecteur intéressé à entreprendre une exploration des écrits sur le sujet : les références complètes sont données dans la bibliographie de l'ouvrage.

#### **LES INDICATEURS URBAINS**

Le survol publié par l'OCDE (1997) constitue un excellent point de départ. Voir aussi Collin, Séguin et Pelletier (1999). La revue *Real Estate Economics* a produit un numéro spécial à l'occasion de la conférence *Habitat II* à Istanbul en 1997. Enfin, l'article de Coombes et Wong

---

<sup>25</sup> Voir à ce sujet les propos d'Amartya Sen lui-même (PNUD, 1999, p. 23). On trouve aussi dans le *Rapport 2000* les propos suivants : « L'information et les statistiques constituent un instrument puissant pour forger une culture de la responsabilité et réaliser les droits de l'homme. Les militants, les juristes, les statisticiens et les spécialistes du développement ont besoin de coopérer avec la population et les communautés. L'objectif : produire des données et des preuves destinées à faire tomber les barrières de l'incrédulité et à inciter au changement des politiques et des comportements » (p. 10).

(1994), quoique moins récent, adopte le point de vue méthodologique et présente une espèce de *How to...* : intéressant, bien que trop flou à mon goût sur certains points de la méthode.

### **UN INDICE DE STATUT SOCIO-ÉCONOMIQUE (RENAUD ET MAYER)**

Jean Renaud et Francine Mayer <sup>26</sup> travaillent depuis plusieurs années au développement d'un indice de statut socio-économique des quartiers urbains basé sur les données du recensement quinquennal. Leur travail s'apparente à la construction d'indicateurs urbains.

Le modèle théorique sous-jacent à la construction de cet indice de statut socio-économique est le modèle de cohabitation de l'écologie sociale urbaine (Renaud *et al.*, 1996, chap.1). Selon ce modèle, les gens semblables tendent à se regrouper dans les mêmes quartiers (« qui s'assemblent se ressemblent »), ce qui crée en milieu urbain une différenciation spatiale où chaque quartier est caractérisé par le type de gens qui y habitent. De façon répétée, les résultats d'études empiriques ont fait ressortir trois dimensions « classiques » qui caractérisent la répartition spatiale de la population :

- le statut socio-économique (richesse/pauvreté)
- le statut familial (présence ou non d'enfants, âge)
- l'appartenance ethnique ou linguistique

L'indice de statut socio-économique traite plus particulièrement de la première de ces trois dimensions (Renaud *et al.*, 1996, p.35-51 et Annexe C, p.133-138). La démarche suivie comprend quatre grandes étapes :

1. Analyse d'écologie factorielle <sup>27</sup> des données du Recensement pour confirmer empiriquement les dimensions qui caractérisent la répartition spatiale de la population <sup>28</sup>.
2. Identification du contenu du ou des facteurs d'ordre socio-économique : l'examen des facteurs socio-économiques révèle que les variables socio-économiques qui contribuent le

---

<sup>26</sup> Renaud, Mayer et Lebeau (1996), Mayer-Renaud et Renaud (1989), Mayer-Renaud (1986).

<sup>27</sup> Renaud *et al.* (1996, chap.2). En gros, l'analyse factorielle est une méthode statistique d'analyse multivariée où l'on résume l'information en réduisant le nombre de variables par la création de variables « composites » (les « facteurs », qui sont analogues à des indices formés de sommes pondérées des variables originales). En examinant la composition des facteurs, on cherche à leur donner une interprétation, c'est-à-dire à leur associer un concept. C'est, pour ainsi dire, la démarche inverse de la construction d'un indice : le concept émerge de l'interprétation de la composition des facteurs, au lieu d'être le point de départ de la construction de l'indice.

<sup>28</sup> « [L'indice] se base, pour le choix des variables et de la méthodologie, sur les résultats de l'écologie factorielle mais sans utiliser le score factoriel, et ce pour éviter la contamination par les variables qui appartiennent à d'autres dimensions » (Renaud *et al.*, 1996, p.38).

plus à caractériser la répartition spatiale sont le revenu, la scolarité et la profession. Les variables retenues pour construire l'indice sont donc :

- revenu des ménages
- scolarité des individus

### 3. Construction de l'indice basé sur le modèle de la cohabitation.

Le lien entre l'indice de statut socio-économique et le modèle sous-jacent n'est donc pas mathématique au sens où, comme pour les indices de prix, on peut déduire la formule de l'indice mathématiquement à partir d'un modèle <sup>29</sup>. Le lien est plutôt « associatif », c'est-à-dire basé sur des mesures d'association statistique entre variables.

Toutefois, l'indice de statut socio-économique n'a pas le caractère quelque peu arbitraire de l'IDH. Ici, les concepts émergent de l'analyse statistique : on a d'abord fait une analyse des données (l'analyse factorielle) ; ensuite, on a interprété les résultats de cette analyse à la lumière d'une hypothèse théorique (le modèle d'écologie factorielle) ; c'est sur cette base que l'on a défini les dimensions du concept de statut socio-économique. Dans l'élaboration de l'IDH, au contraire, on a défini *a priori* le concept de développement humain et on a fixé d'emblée ses dimensions (longévité, savoir et niveau de vie).

Il n'en demeure pas moins que l'indice de statut socio-économique tente de mesurer une réalité complexe, qui comporte par surcroît des dimensions ordinales (c'est-à-dire qu'on ne peut pas mesurer au moyen de variables d'intervalle ou rationnelles). Pour agréger les multiples dimensions en une seule (ici, deux variables en un indice), il faut traiter des variables ordinales <sup>30</sup> comme si elles étaient rationnelles. C'est pourquoi la construction de l'indice repose lourdement sur le jugement de valeur du chercheur, qui doit attribuer des valeurs numériques aux catégories.

---

<sup>29</sup> Il est vrai que les modèles sous-jacents aux indices de Laspeyres ou de Paasche reposent sur des hypothèses extrêmement restrictives, mais la formule de calcul des indices se déduit néanmoins mathématiquement des modèles.

<sup>30</sup> La scolarité, bien sûr, mais aussi le revenu, puisque ce dernier n'est connu que par tranche (voir au chapitre 1-1, à propos des échelles de mesure, plus précisément, à propos des variables rationnelles d'intervalle regroupées en classes).



### **ET PLUS...**

Mentionons enfin qu'au Mexique, le Conseil National de la Population (CONAPO) produit un indice de marginalisation des « municipios » et même des « comunidades » (villages) <sup>31</sup>. Cet *índice de marginación* présente une certaine ressemblance avec l'indice de statut socio-économique de Renaud et Mayer, mais sa formulation est plus étroitement liée aux résultats de l'analyse factorielle.

---

<sup>31</sup> [http://www.conapo.gob.mx/m\\_en\\_cifras/principal.html](http://www.conapo.gob.mx/m_en_cifras/principal.html) (Población de México en cifras); dans « Menú de sección », voir « Marginación ».

## CHAPITRE 1-4

### MESURE DE L'INÉGALITÉ ET DE LA CONCENTRATION

---

#### Plan

1-4.1 Le coefficient de concentration de l'économie industrielle	4
1-4.2 L'indice de concentration de Hirschman-Herfindahl	4
1-4.3 La courbe de Lorenz et l'indice de concentration de Gini	5
La différence moyenne de Gini	5
Calcul de l'indice de concentration de Gini	6
La courbe de Lorenz	10
Calcul géométrique de l'indice de Gini au moyen de la courbe de Lorenz	15
Propriétés de l'indice de concentration de Gini	16
1-4.4 Pour conclure à propos de la mesure de l'inégalité...	18

## CHAPITRE 1-4

### MESURE DE L'INÉGALITÉ ET DE LA CONCENTRATION

Références : Arriaga, 1975, p. 65-71 ; Taylor, 1977, 179-185 ; Mills et Hamilton, 1989, p. 413-414 ; Kendall et Stuart (1991, p. 58) ; Jayet (1993, p. 18-29) ; Valeyre (1993) ; MacLachlan et Sawada (1997).

Nous nous attachons dans ce chapitre à l'examen des différentes valeurs d'une même variable dans un ensemble d'observations. Une mesure d'inégalité (on dit aussi « de disparité ») indique à quel degré les valeurs diffèrent les unes des autres. Prenons, par exemple, les revenus des habitants d'un pays ; une mesure d'inégalité du revenu sert à quantifier le degré d'inégalité de la distribution du revenu entre les habitants du pays, de façon à pouvoir le comparer à celui d'autres pays. Dans cet exemple, la variable examinée est le revenu et les observations correspondent aux habitants du pays.

Lorsque les observations correspondent à des catégories et que la variable examinée est le nombre d'individus (d'objets) d'une population donnée qui se trouve dans chaque catégorie, alors une mesure d'inégalité est aussi une mesure de concentration. Par exemple, si l'on considère la distribution de la population humaine entre les régions d'un pays, une mesure d'inégalité indique à quel point la population du pays est concentrée.

En sciences sociales, on s'est intéressé à la mesure d'inégalité dans plusieurs contextes différents : inégalité dans la distribution du revenu, concentration des parts de marché (mesure inverse du degré de concurrence), concentration spatiale des populations ou des activités économiques, etc.

La construction de mesures d'inégalité ou de concentration pose un problème analogue à celui de la multidimensionnalité dans la définition de nombres indices : il s'agit de résumer en un seul chiffre une caractéristique possédée par l'ensemble des valeurs que prend une variable. On peut donc s'attendre à ce qu'il n'y ait pas de solution unique.

En général, une mesure de l'inégalité compare la distribution observée avec une distribution de référence, qui représente l'égalité parfaite. Souvent, la distribution de référence reste implicite. Mais il est parfois nécessaire de l'explicitier. Par exemple, s'agissant de la répartition spatiale d'une population entre des régions, une concentration nulle correspond-elle à la situation où le nombre d'habitants est le même dans toutes les régions ? Ou correspond-elle plutôt à la

situation où le nombre d'habitants est proportionnel à la superficie des régions ? Ou encore, à la superficie habitable ?

Quelles sont les propriétés désirables d'une mesure d'inégalité ? Valeyre (1993) propose les six propriétés suivantes :

1. Une mesure d'inégalité doit prendre des valeurs non négatives, puisqu'il s'agit d'une mesure de l'éloignement de la distribution observée par rapport à la distribution de référence.
2. Une mesure d'inégalité doit prendre la valeur zéro si, et seulement si, la distribution observée est identique à la distribution de référence.
3. Toutes les observations doivent être traitées de la même manière.
4. Une mesure d'inégalité doit être indépendante de la valeur moyenne de la variable examinée ; une mesure de concentration doit être indépendante de la taille de la population dont on étudie la distribution.
5. L'agrégation d'observations affichant le même degré de spécificité ne doit pas changer la valeur de la mesure <sup>1</sup>.
6. Principe de transfert de Pigou-Dalton : une mesure d'inégalité doit diminuer si la distribution est modifiée d'une façon qui réduit incontestablement l'inégalité <sup>2</sup>.

Ces principes permettent d'évaluer la validité des différentes mesures d'inégalité qui se proposent. Ainsi, l'écart-type ou la variance ne possèdent pas la propriété 4. Par contre, le coefficient de variation possède les six propriétés énoncées : correctement utilisé, il constitue donc une bonne mesure d'inégalité.

Voyons maintenant quelques autres exemples de mesures d'inégalité ou de concentration.

---

<sup>1</sup> La spécificité réfère au rapport entre une valeur observée de la variable étudiée et la valeur correspondante dans la distribution de référence. Par exemple, les quotients de localisation sont des indicateurs de spécificité. Une mesure de la concentration géographique de l'emploi d'une branche d'activité donnée ne devrait pas être affectée si l'on agrège deux régions dont les quotients de localisation sont égaux.

<sup>2</sup> Techniquement, cela se traduit par la condition suivante : si la valeur de la variable diminue pour une observation  $i$  et augmente d'un même montant pour une autre observation  $j$ , et si le degré de spécificité de l'observation  $i$  est supérieur à celui de l'observation  $j$ , alors la mesure d'inégalité doit diminuer.

### 1-4.1 Le coefficient de concentration de l'économie industrielle

Cette mesure est surtout utilisée en économie industrielle, mais elle a aussi été utilisée pour mesurer le degré de concentration de la distribution par taille des villes. C'est tout simplement la somme des parts des  $n$  plus grandes entités. Par exemple, Rosen et Resnick (1980) mesurent le degré de concentration d'une hiérarchie urbaine au moyen la fraction de la population urbaine totale qui se trouve dans les trois plus grandes villes. En économie industrielle, on mesure souvent la concentration de marché au moyen de la somme des parts des quatre plus grandes entreprises.

Cette mesure a l'avantage de ne pas être très exigeante en termes de données, mais il lui manque la plupart des propriétés désirables : elle ne possède que la première et la quatrième.

### 1-4.2 L'indice de concentration de Hirschman-Herfindahl

Cet indice est simplement par la somme des carrés des parts. Par exemple, pour mesurer le degré de concentration dans un système urbain qui comporte  $n$  villes, on peut calculer

$$H = \sum_{i=1}^n s_i^2$$

où  $s_i$  est la fraction de la population urbaine totale qui se trouve dans la ville  $i$ . L'indice  $H$  varie entre  $\frac{1}{n}$  et 1 : il prend la valeur  $\frac{1}{n}$  quand toutes les villes sont de taille égale, et la valeur 1 dans le cas où toute la population urbaine est concentrée dans une seule ville. On interprète parfois l'indice  $H$  en termes de « nombre équivalent », notamment en économie industrielle : dans un marché de, disons quarante entreprises, si l'indice  $H$  a une valeur de  $x$ , on dit que le degré de concentration « équivaut » à celle d'un marché de  $\frac{1}{x}$  firmes ayant des parts de marché égales.

L'indice de Hirschman-Herfindahl ne possède pas les propriétés 2 et 5. En outre, il dépend du nombre d'observations  $n$ . Il est à noter enfin que l'indice  $H$  est très étroitement lié à la variance des parts : celle-ci est en effet égale à

$$\frac{1}{n} \sum_{i=1}^n \left( s_i - \frac{1}{n} \right)^2 = \frac{H}{n} - \frac{1}{n^2}$$

### 1-4.3 La courbe de Lorenz et l'indice de concentration de Gini

#### LA DIFFÉRENCE MOYENNE DE GINI

L'indice de concentration de Gini est ainsi nommé en l'honneur du statisticien italien Corrado Gini (1884-1965). Il mesure l'inégalité au moyen des différences entre toutes les paires d'observations ( $y_j, y_k$ ). La somme pondérée des différences s'appelle la « différence moyenne de Gini » et elle se calcule, pour des données groupées, selon la formule suivante<sup>3</sup> :

$$\Delta = \frac{1}{N^2} \sum_{j=1}^n \sum_{k=1}^n |y_j - y_k| f_j f_k$$

où

$n$  est le nombre de valeurs distinctes observées

où  $f_j$  est la fréquence de la valeur  $y_j$  dans la distribution, de sorte que

$$N = \sum_{j=1}^n f_j \text{ est le nombre d'observations}$$

Par exemple, s'agissant de mesurer l'inégalité de la distribution du revenu au Québec,  $f_j$  serait le nombre de personnes qui ont un revenu de  $y_j$ ;  $N$  est le nombre de personnes dans la population.

Lorsque les observations sont groupées par classes, la valeur  $y_j$  est la valeur *moyenne* de la variable  $Y$  dans la classe  $j$  (et non pas le point milieu de l'intervalle de revenu de la classe  $j$ ).

Écrivons

$$v_j = \frac{f_j}{N}, \text{ la fraction de la population appartenant à la classe } j.$$

La valeur moyenne de la variable  $Y$  s'écrit alors

$$\mu = \frac{1}{N} \sum_{j=1}^n f_j y_j = \sum_{j=1}^n v_j y_j$$

---

<sup>3</sup> Dans cette formule chaque observation est comparée à chacune des observations, y compris à elle-même ; c'est la différence moyenne avec répétition. Kendall et Stuart (1991, p. 58) donnent aussi la formule *sans* répétition. Lorsque  $N$  est grand, la différence est négligeable.

Soit

$$M = \sum_{j=1}^n f_j y_j, \text{ la somme des valeurs de la variable } Y, \text{ et}$$

$$w_j = \frac{f_j y_j}{\sum_{k=1}^n f_k y_k} = \frac{f_j y_j}{N\mu} = \frac{v_j y_j}{\mu}, \text{ la fraction de la somme allouée à la classe } j.$$

Rangeons ensuite les observations, en vue de la construction d'une courbe de Lorenz (voir ci-après), par ordre croissant des rapports  $w_j/v_j$ . Complétons la notation en posant

$$Cw_j = \sum_{k=1}^j w_k$$

$Cw_j$  est la fraction cumulée des classes de 1 à  $j$ .

En développant la formule de calcul de la différence moyenne de Gini, on obtient :

$$\Delta = 2\mu \left( 1 - \sum_{j=1}^n v_j Cw_j - \sum_{j=1}^n v_j Cw_{j-1} \right)$$

Cela est démontré à l'annexe 1-F.

#### CALCUL DE L'INDICE DE CONCENTRATION DE GINI

L'indice de concentration de Gini est simplement le rapport de la différence moyenne de Gini sur deux fois la moyenne :

$$G = \frac{\Delta}{2\mu} = 1 - \left( \sum_{j=1}^n v_j Cw_j + \sum_{j=1}^n v_j Cw_{j-1} \right) = 1 - \sum_{j=1}^n v_j (Cw_j + Cw_{j-1})$$

Arriaga (1975, p. 65-71), comme aussi plusieurs géographes, définit le coefficient Gini comme

$$G = \sum_{i=2}^n Cw_i Cv_{i-1} - \sum_{i=2}^n Cw_{i-1} Cv_i$$

où  $Cv_j = \sum_{k=1}^j v_k$

Cette formule peut se déduire de la précédente.

$$G = 1 - \sum_{j=1}^n v_j (Cw_j + Cw_{j-1})$$

$$G = 1 - \sum_{j=1}^n (Cv_j - Cv_{j-1})(Cw_j + Cw_{j-1})$$

$$G = 1 - \sum_{j=1}^n Cv_j Cw_j - \sum_{j=1}^n Cv_j Cw_{j-1} + \sum_{j=1}^n Cv_{j-1} Cw_j + \sum_{j=1}^n Cv_{j-1} Cw_{j-1}$$

où

$$\sum_{j=1}^n Cv_{j-1} Cw_{j-1} = \sum_{j=0}^{n-1} Cv_j Cw_j \text{ et } Cv_0 = Cw_0 = 0, \text{ de sorte que}$$

$$\begin{aligned} - \sum_{j=1}^n Cv_j Cw_j + \sum_{j=1}^n Cv_{j-1} Cw_{j-1} &= - \sum_{j=1}^n Cv_j Cw_j + \sum_{j=1}^{n-1} Cv_j Cw_j \\ &= - Cv_n Cw_n = -1 \end{aligned}$$

Et

$$G = 1 - 1 - \sum_{j=1}^n Cv_j Cw_{j-1} + \sum_{j=1}^n Cv_{j-1} Cw_j$$

ce qui, puisque  $Cv_0 = Cw_0 = 0$ , équivaut à

$$G = - \sum_{j=2}^n Cv_j Cw_{j-1} + \sum_{j=2}^n Cv_{j-1} Cw_j$$

$$G = \sum_{j=2}^n Cw_j Cv_{j-1} - \sum_{j=2}^n Cw_{j-1} Cv_j$$

Lorsque les observations sont groupées, il suffit donc, pour calculer l'indice de concentration de Gini, de connaître la répartition entre les classes de la population (les  $v_j$ ) et de la somme des valeurs de la variable  $Y$  (les  $w_j$ ).

Le plus souvent, la population (ou les ménages) sont d'abord rangés en ordre croissant de revenu ; on définit ensuite des catégories de tailles égales : quartiles, quintiles, déciles, etc. On dira ainsi : « Les 20 % de la population avec les revenus les plus élevés (le quintile supérieur) accaparent xx % du revenu global, tandis que les 20 % avec les



revenus les moins élevés (le quintile inférieur) n'en reçoivent que zz % ». De tels énoncés donnent aussi une mesure de la concentration, mais, contrairement au coefficient de Gini, ce sont des mesures partielles, qui ne tiennent compte que d'une partie de la distribution.

Voyons, par exemple, la répartition du revenu (Y) entre les familles au Canada en 1995 (la population considérée est donc celle des familles, et non des individus). Statistique Canada a récemment diffusé les données suivantes, tirées du Recensement de la population de 1996 <sup>4</sup>.

---

<sup>4</sup> *Le Quotidien*, 3 mars 1999. Il est à noter que les données du Recensement de 1996 sur les revenus annuels se rapportent à l'année précédente.

**Tableau: Limites supérieures (en \$ de 1995) des déciles du revenu familial et répartition du revenu global familial par décile, 1995**

Décile	Limite supérieure	Part du revenu global (%)
Premier	15158	1,45
Deuxième	23184	3,55
Troisième	31097	4,96
Quatrième	38988	6,42
Cinquième	46951	7,86
Sixième	55355	9,37
Septième	64997	10,91
Huitième	77501	13,11
Neuvième	98253	15,85
Dixième		26,53

Les données de ce tableau peuvent être présentées autrement, comme ceci :

Classe de revenus (\$ de 1995)	Fraction du nombre de familles (%)	Part du revenu global (%)
0-15158	10,00	1,45
15159-23184	10,00	3,55
23185-31097	10,00	4,96
31098-38988	10,00	6,42
38989-46951	10,00	7,86
46952-55355	10,00	9,37
55356-64997	10,00	10,91
64998-77501	10,00	13,11
77502-98253	10,00	15,85
98254 et plus	10,00	26,53

Dans le tableau qui précède, les  $w_j$  sont les parts du revenu global ; les  $v_j$  sont tous égaux à 10 %. À partir de ce tableau, on peut effectuer les calculs préliminaires comme dans le tableau suivant.

Classe de revenus (\$ de 1995)	Fraction du nombre de familles (%)	Part du revenu global (%)			
	$v_j$	$w_j$	$Cw_j$	$v_j Cw_j$	$v_j Cw_{j-1}$
0-15158	10,00	1,45	0,0145	0,0015	0,0000
15159-23184	10,00	3,55	0,0500	0,0050	0,0015
23185-31097	10,00	4,96	0,0996	0,0100	0,0050
31098-38988	10,00	6,42	0,1638	0,0164	0,0100
38989-46951	10,00	7,86	0,2424	0,0242	0,0164
46952-55355	10,00	9,37	0,3361	0,0336	0,0242
55356-64997	10,00	10,91	0,4452	0,0445	0,0336
64998-77501	10,00	13,11	0,5763	0,0576	0,0445
77502-98253	10,00	15,85	0,7348	0,0735	0,0576
98254 et plus	10,00	26,53	1,0001	0,1000	0,0735
Total	100,00	100,00		0,3663	0,2663

L'indice de concentration de Gini du revenu familial au Canada par déciles, en 1995, est donc égal à :

$$G = 1 - (0,3663 + 0,2663) = 0,3675$$

Deux remarques s'imposent ici :

- Les données utilisées ici étaient d'emblée rangées par ordre croissant des rapports  $w_j/v_j$ .  
Ce n'est pas toujours le cas ! En général, avant de calculer l'indice de Gini, il faut préalablement ranger les données dans le bon ordre (voir l'exemple tiré de Taylor, 1977, ci-après).
- Avec des données groupées, l'indice de concentration de Gini dépend du groupement ou schème de classement utilisé. Si la population des familles avait été groupée par quintiles, ou par centiles, le résultat du calcul aurait été différent. Nous reviendrons sur ce point.

## LA COURBE DE LORENZ

La courbe de Lorenz est un instrument de comparaison graphique entre deux distributions. Rappelons que

$$Cv_j = \sum_{k=1}^j v_k = \text{fraction cumulée de } X \text{ (par exemple, ci-haut, des familles)}$$

$$Cw_j = \sum_{k=1}^j w_k = \text{fraction cumulée de } Y \text{ (par exemple, ci-haut, des revenus)}$$

On a naturellement :

$$Cv_n = Cw_n = 1$$

Méthode de construction de la courbe de Lorenz (Voir l'exemple numérique ci-après, tiré de Taylor, 1977, p. 179) :

1. Calculer les rapports  $\frac{w_i}{v_i}$  <sup>5</sup>.
2. Réordonner les catégories en ordre croissant de  $\frac{w_i}{v_i}$  :  $\frac{w_1}{v_1} < \frac{w_2}{v_2} < \dots < \frac{w_n}{v_n}$
3. Calculer les fractions cumulatives  $Cv_i$  et  $Cw_i$
4. La courbe de Lorenz est l'ensemble des points  $(Cv_i, Cw_i)$ , où les  $Cv_i$  sont repérés sur l'axe horizontal.

La courbe de Lorenz a les propriétés suivantes :

1.  $Cv_0 = Cw_0 = 0$  (par définition de  $Cv_i$  et de  $Cw_i$ ) : la courbe part de l'origine ;
2.  $Cv_n = Cw_n = 1$  (par définition de  $Cv_i$  et de  $Cw_i$ ) : la courbe aboutit au point de coordonnées  $[1,1]$  (ou  $[100\%, 100\%]$ ) ;
3. Lorsque les deux distributions sont identiques, on a, pour tout  $i$ ,  
 $Cv_i = Cw_i$   
c'est-à-dire que la courbe de Lorenz coïncide avec la diagonale.
4.  $Cv_i \geq Cw_i$  pour  $i$  différent de 0 et de  $n$  (par construction, étant donné le réordonnement des catégories) : la courbe se situe sous la diagonale ou coïncide avec elle ;
5. La pente de chaque segment de la courbe de Lorenz est égale à la valeur l'indicateur de spécificité associé à l'observation correspondante :

$$\text{pente du segment } i = \frac{Cw_i - Cw_{i-1}}{Cv_i - Cv_{i-1}} = \frac{w_i}{v_i}$$

---

<sup>5</sup> Ces rapports ne sont autre que les *spécificités* associées aux observations.

6. La courbe de Lorenz est concave vers le haut, c'est-à-dire que chaque segment a une pente plus abrupte que le précédent : cela découle de 5, puisque, par construction,  $\frac{w_i}{v_i} < \frac{w_{i+1}}{v_{i+1}}$

**CONSTRUCTION D'UNE COURBE DE LORENZ (EXEMPLE NUMÉRIQUE TIRÉ DE TAYLOR, 1977, P. 179)**

**Première étape : calcul des  $w_i/v_i$**

Zone	$x_i$ Nombre de ménages de classe moyenne	$v_i$ Distrib. de x	$y_i$ Nombre de votes du parti Républi- cain	$w_i$ Distrib. de y	$w_i/v_i$
A	30	0,25	30	0,30	1,20
B	20	0,17	15	0,15	0,90
C	10	0,08	8	0,08	0,96
D	10	0,08	5	0,05	0,60
E	20	0,17	19	0,19	1,14
F	30	0,25	23	0,23	0,92
<b>Total</b>	<b>120</b>	<b>1,00</b>	<b>100</b>	<b>1,00</b>	

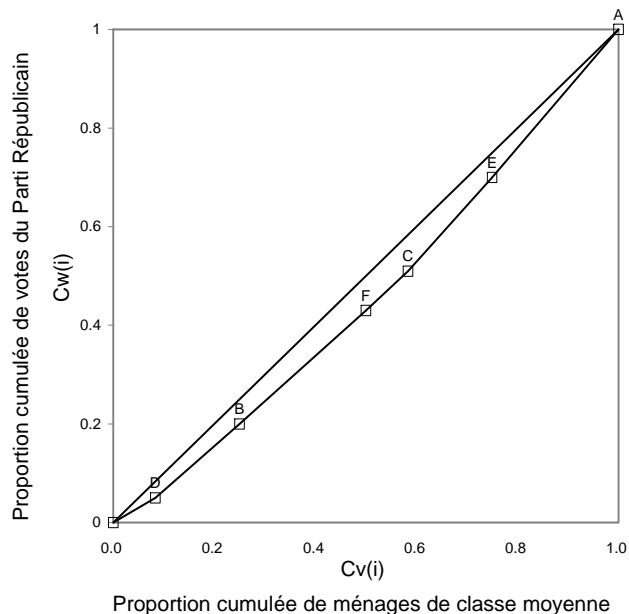
**Deuxième étape : tri par ordre croissant des  $w_i/v_i$**

**Troisième étape : calcul des  $Cv_i$  (abscisses) et des  $Cw_i$  (ordonnées)**

Zone	$x_i$	$v_i$	$y_i$	$w_i$	$w_i/v_i$	$Cv_i$ Abscisse	$Cw_i$ Ordonnée	Écart ( $Cv_i - Cw_i$ )	Écart $ v_i - w_i $
						0,00	0,00		
D	10	0,08	5	0,05	0,60	0,08	0,05	0,033	0,033
B	20	0,17	15	0,15	0,90	0,25	0,20	0,050	0,017
F	30	0,25	23	0,23	0,92	0,50	0,43	0,070	0,020
C	10	0,08	8	0,08	0,96	0,58	0,51	<b>0,073</b>	0,003
E	20	0,17	19	0,19	1,14	0,75	0,70	0,050	0,023
A	30	0,25	30	0,30	1,20	1,00	1,00	0,000	0,050
<b>Total</b>	<b>120</b>	<b>1,00</b>	<b>100</b>	<b>1,00</b>					<b>0,147</b>

Note : on peut constater que l'écart maximum entre la courbe de Lorenz et la diagonale est égal à  $\frac{1}{2} \sum_i |v_i - w_i|$ .

### Courbe de Lorenz



#### Quatrième étape : calcul de l'indice de concentration de Gini

Zone	$x_j$	$v_j$	$y_j$	$w_j$	$w_j/v_j$	$Cv_j$ Abscisse	$Cw_j$ Ordonnée	$v_j Cw_j$	$v_j Cw_{j-1}$
						0,00	0,00		
D	10	0,08	5	0,05	0,60	0,08	0,05	0,004	0,000
B	20	0,17	15	0,15	0,90	0,25	0,20	0,033	0,008
F	30	0,25	23	0,23	0,92	0,50	0,43	0,108	0,050
C	10	0,08	8	0,08	0,96	0,58	0,51	0,043	0,036
E	20	0,17	19	0,19	1,14	0,75	0,70	0,117	0,085
A	30	0,25	30	0,30	1,20	1,00	1,00	0,250	0,175
<b>Total</b>	<b>120</b>	<b>1,00</b>	<b>100</b>	<b>1,00</b>				<b>0,554</b>	<b>0,354</b>

$$G = 1 - (0,554 + 0,354) = 0,092$$

### CALCUL GÉOMÉTRIQUE DE L'INDICE DE GINI AU MOYEN DE LA COURBE DE LORENZ

Ce fut une réalisation remarquable de Corrado Gini que de démontrer, en 1914, que l'indice de concentration qui porte son nom est égal au rapport entre (1) la superficie comprise entre la diagonale et la courbe de Lorenz et (2) la superficie totale sous la diagonale :

$$G = \frac{\text{Superficie comprise entre la diagonale et la courbe de Lorenz}}{\text{Superficie totale sous la diagonale}}$$

La superficie totale du triangle sous la diagonale est donnée par

$$\frac{Cw_n \times Cv_n}{2} = \frac{1 \times 1}{2} = \frac{1}{2}$$

La superficie comprise entre la diagonale et la courbe de Lorenz est calculée comme la différence entre la superficie totale du triangle sous la diagonale ( $= \frac{1}{2}$ ) et la superficie sous la courbe de Lorenz. La superficie sous la courbe de Lorenz (voir l'exemple numérique précédent et la figure ci-après) est la somme de  $n$  trapèzes dont chacun a une surface égale à

$$\frac{1}{2} v_i (Cw_i + Cw_{i-1})$$

La superficie sous la courbe de Lorenz est donc la somme de ces  $n$  surfaces :

$$\frac{1}{2} \sum_{i=1}^n v_i (Cw_i + Cw_{i-1})$$

Et le coefficient Gini est donné par

$$G = \frac{\left(\frac{1}{2}\right) - \left(\frac{1}{2} \sum_{i=1}^n v_i (Cw_i + Cw_{i-1})\right)}{\left(\frac{1}{2}\right)} = 1 - \sum_{i=1}^n v_i (Cw_i + Cw_{i-1}) = \frac{\Delta}{2\mu}$$

ce qui correspond bien à la formule énoncée précédemment.

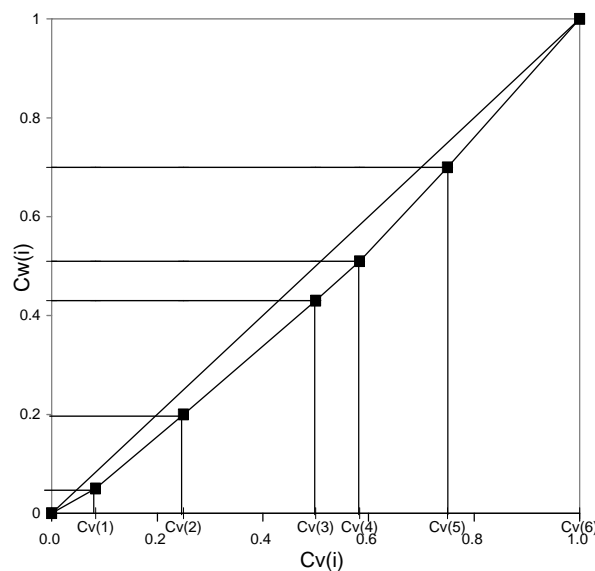
Pour faciliter l'interprétation de la courbe de Lorenz et de l'indice de Gini qui lui est associé, il est utile de se rappeler que c'est la distribution  $V$  qui joue le rôle de distribution de référence (i.e. d'égalité parfaite ou de concentration nulle). Dans la courbe de Lorenz, les  $Cv_i$  sont repérés sur l'axe horizontal et les  $Cw_i$ , sur l'axe vertical.



Exemples :

- Si  $V$  est la répartition du territoire entre les zones et  $W$ , la répartition de la population, le coefficient Gini est une mesure de la concentration géographique de la population.
- Si  $V$  est une distribution de la population (ou des ménages) en  $n$  catégories et  $W$ , la distribution du revenu agrégé par catégorie, alors le coefficient Gini est une mesure de la concentration du revenu.

### Calcul géométrique de l'indice de concentration de Gini



### PROPRIÉTÉS DE L'INDICE DE CONCENTRATION DE GINI

L'indice de concentration de Gini possède les six propriétés que doit avoir une mesure d'inégalité, telles qu'énoncées au début de ce chapitre. Il a en outre les propriétés suivantes :

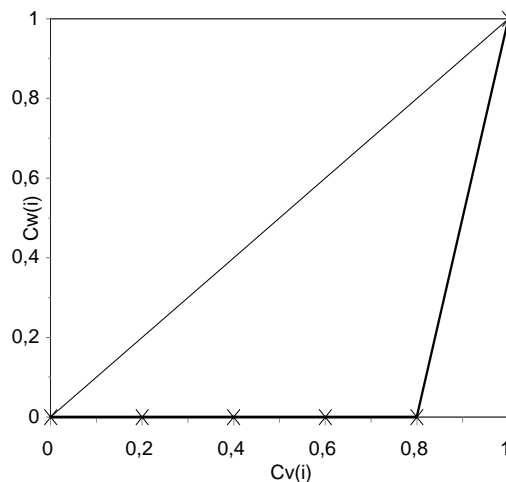
1. L'indice de Gini varie entre zéro et 1<sup>6</sup>. La valeur minimum du coefficient Gini, zéro, est atteinte lorsque les deux distributions sont identiques. Sa valeur maximum théorique, 1, est atteinte lorsque la courbe de Lorenz suit la base et le côté droit de la « boîte » ; mais pour

<sup>6</sup> Ou entre 0 % et 100 % lorsqu'on l'exprime en pourcentage.

que ce maximum théorique soit atteint, il faut que le nombre de catégories tende vers l'infini, de sorte que  $v_n$  tende vers zéro <sup>7</sup>.

2. On peut démontrer que l'indice de Gini est symétrique, c'est-à-dire que les rôles des deux distributions sont interchangeables ; en d'autres mots, si l'on intervertit les rôles, la valeur du coefficient Gini est inchangée.
3. Lorsque les données sont groupées, l'indice de Gini est sensible à la définition et au nombre des catégories utilisées (classes, zones).
4. Lorsqu'on l'utilise comme mesure de concentration spatiale, le Gini ne tient aucun compte de la proximité dans l'espace des différentes zones de forte densité (l'espace est traité comme un puzzle défectif).

À propos de la valeur maximum que peut atteindre le coefficient de Gini lorsque le nombre de catégories n'est pas infini, précisons qu'elle est égale à  $1 - v_n$ . Cette propriété est illustrée dans la figure suivante.



Dans cet exemple, le lecteur peut facilement vérifier, en appliquant la méthode de calcul géométrique, que  $v_n = 0,2$  et  $G = (1 - 0,2) = 0,8$ .

La troisième propriété mérite une attention plus poussée. Elle se manifeste notamment par ceci : l'agrégation de deux ou plusieurs catégories a *toujours* pour effet de réduire la valeur calculée du coefficient Gini (à moins que les deux catégories n'aient la même spécificité, auquel

---

<sup>7</sup> Autrement, lorsque  $v_n > 0$ , la valeur maximum de  $G$  est égale à  $1 - v_n$ .

cas la propriété 5 des mesures d'inégalité se manifeste). Cela se vérifie aisément si l'on pense au calcul géométrique fait au moyen de la courbe de Lorenz : l'agrégation de deux catégories voisines réduit l'espace compris entre la courbe de Lorenz et la diagonale. Cela est également conforme à l'intuition que l'agrégation de catégories a pour effet de gommer une partie des différences.

Cette sensibilité du Gini à la définition des catégories peut sérieusement compromettre sa fiabilité comme mesure de la concentration, notamment lorsque les catégories sont de tailles inégales. Pour illustrer ce phénomène, imaginons que l'on veuille comparer la concentration de la population à deux moments dans le temps, sur un territoire découpé en trois zones de même superficie (disons égale à 1) :

	Superf.	Population		Densité	
		au temps 0	au temps $t$	au temps 0	au temps $t$
Zone 1	1	10	80	10	80
Zone 2	1	80	10	80	10
Zone 3	1	10	10	10	10

Il est évident dans cet exemple que la concentration est restée la même à l'échelle considérée ( $G = 0,47$ ), même si le centre de gravité de la population s'est déplacé vers la Zone 1. Supposons maintenant que l'on ait agrégé les Zones 2 et 3 :

	Superf.	Population		Densité	
		au temps 0	au temps $t$	au temps 0	au temps $t$
Zone 1	1	10	80	10	80
Zones 2 et 3	2	90	20	45	10

Les données agrégées donnent l'illusion que la concentration a augmenté, puisqu'on a  $G = 0,23$  au temps 0 et  $G = 0,47$  au temps  $t$  (noter que l'indice de Gini calculé est plus faible avec les données agrégées au temps 0, mais qu'il est le même au temps  $t$ , puisque dans ce dernier cas, les zones agrégées sont de même densité, c'est-à-dire de même spécificité).

#### 1-4.4 Pour conclure à propos de la mesure de l'inégalité...

Nous n'avons évoqué ici que quelques-unes des multiples mesures d'inégalités qui sont maintenant proposés. Parmi les mesures que nous avons laissées de côté, mentionnons celles qui sont des mesures d'entropie, comme la mesure de Shannon, ou la mesure du gain d'information de Kullback-Leibler (aussi associée au nom de Theil). Le lecteur intéressé pourra consulter le survol de Valeyre (1993).

Rappelons enfin que les mesures d'inégalité sont des mesures d'éloignement d'une distribution observée par rapport à une distribution de référence. Elles sont en cela étroitement apparentées aux mesures de dissimilarité, qui comparent deux distributions, dont les rôles sont toutefois symétriques (aucune des deux ne joue le rôle de référence).

## CHAPITRE 1-5

### MESURE DE LA DISSIMILARITÉ

---

#### Plan

1-5.1 Multidimensionnalité, dissimilarité et concentration	2
Problématique de la mesure de la dissimilarité	2
La mesure de la dissimilarité entre des distributions	6
Dissimilarité et inégalité-concentration : quelle différence ?	7
1-5.2 L'indice de dissimilarité	8
Un exemple numérique	8
Définition de l'indice de dissimilarité	9
L'indice de dissimilarité comme mesure de concentration ou d'inégalité	12
Propriétés de l'indice de dissimilarité	15
Application de l'indice de dissimilarité à une dichotomie	22
Un dernier regard critique	27
1-5.3 Distance et dissimilarité	28
1-5.4 La mesure de la similarité en statistique	31
1-5.5 Autres mesures de similarité et de dissimilarité	31

## CHAPITRE 1-5

### MESURE DE LA DISSIMILARITÉ

#### 1-5.1 Multidimensionnalité, dissimilarité et concentration

##### PROBLÉMATIQUE DE LA MESURE DE LA DISSIMILARITÉ

Nous avons vu qu'une mesure associée à un concept établit une correspondance entre les objets et des nombres, ce qui permet de comparer les objets et de déterminer la valeur de vérité d'une ou de plusieurs des relations  $=$ ,  $\neq$ ,  $>$  ou  $<$ . Si, comme cela arrive souvent, un concept comprend plusieurs dimensions, et que l'on veut néanmoins le traiter comme un tout, nous avons vu qu'il faut surmonter le problème de la multidimensionnalité et qu'on peut le faire en construisant un indice.

Mais il arrive que l'on soit confronté à des concepts auxquels on ne peut pas associer de mesure autre que catégorique. Impossible, alors, d'envisager la construction d'un indice. Considérons par exemple le concept de structure économique d'une ville ou d'une région. On a beau se satisfaire de définir la structure économique comme la répartition de l'emploi entre les branches d'activité, on ne peut guère associer à ce concept d'autre mesure qu'une classification (variable catégorique) : ville mono-industrielle, ville de services, etc. Mais comment arrive-t-on à construire une classification qui permette de bien saisir la réalité ? Une manière de procéder consiste à comparer les objets (en l'occurrence, les structures économiques observées) pour constituer des groupes d'objets assez similaires entre eux, et nettement différents des objets des autres groupes. Une telle classification peut ensuite servir de base à l'élaboration d'une typologie et à la définition d'une variable catégorique associée au concept.

Mentionnons en passant que, même lorsque la construction d'un indice est possible en principe, l'approche qui vient d'être évoquée peut être souhaitable s'il s'avère impossible de construire un indice qui soit satisfaisant au plan théorique. Ne pourrait-on pas, par exemple, étudier le développement humain en constituant une typologie des pays ? Une telle typologie permettrait de définir un indice approprié à chaque type de pays, de manière à ne comparer que des pays comparables, et avec des mesures adaptées aux caractéristiques de ces pays (c'est ce que fait déjà le PNUD par rapport à la mesure de la pauvreté : il calcule deux « Indices de la pauvreté humaine », l'un pour les pays en développement et l'autre, pour les pays développés).

La démarche qui consiste à classer les objets pour dégager des types ne peut qu'être grandement facilitée si l'on peut formaliser le concept de similarité et lui associer une mesure. Il existe d'ailleurs des procédures de classification automatique fondées sur des mesures de similarité<sup>1</sup>. En outre, on souhaitera parfois s'en tenir à une démarche heuristique, plus informelle, et examiner le degré de similarité entre des objets sans aller jusqu'à construire une typologie. Là encore, une mesure de la similarité peut être un outil précieux. C'est donc de la mesure de la similarité qu'il est question ici.

Notons d'abord que le concept de la similarité s'applique à une *paire* d'objets. La similarité n'est donc une propriété d'aucun des deux objets : elle est une propriété de la paire<sup>2</sup>. Ensuite, le concept de similarité est un concept général, qui recouvre une myriade de concepts spécifiques : car lorsqu'on examine la similarité entre deux objets, c'est toujours *par rapport* à un attribut donné. Un concept de similarité spécifique est défini par l'attribut auquel on se réfère pour comparer les objets dont on veut mesurer la similarité. S'agissant de villes, par exemple, on peut considérer la similarité par rapport à la structure démographique, par rapport au taux de criminalité, par rapport à la qualité de vie, etc.

Convenons d'emblée que la mesure de la similarité par rapport à un attribut unidimensionnel est une affaire triviale : il n'y a pas de problème particulier à mesurer, par exemple, la similarité entre deux pays quant au chiffre de leur population, au taux de criminalité ou à la valeur de l'IDH du PNUD<sup>3</sup>. Par contre, lorsqu'on veut mesurer la similarité par rapport à une propriété multidimensionnelle que l'on n'a pas au préalable résumée en un indice<sup>4</sup>, on est confronté au même problème que dans la construction d'un nombre indice. Par exemple,

- Par rapport à leur structure économique, quel est le degré de similarité entre le Québec et l'Ontario ?
- Par rapport à leur répartition sur le territoire, quel est le degré de similarité entre la culture bananière et l'élevage au Costa Rica ?

---

<sup>1</sup> Dendrogrammes, algorithmes de partition automatique, etc. Voir Legendre et Legendre (1984 et 1998).

<sup>2</sup> On pourrait dire que l'objet auquel s'applique la similarité est une paire d'objets.

<sup>3</sup> Cet exemple est délibérément paradoxal : alors que l'IDH est un indicateur qui cherche à mesurer une réalité multidimensionnelle, la comparaison de deux pays quant à la valeur de cet indicateur est, elle, unidimensionnelle.

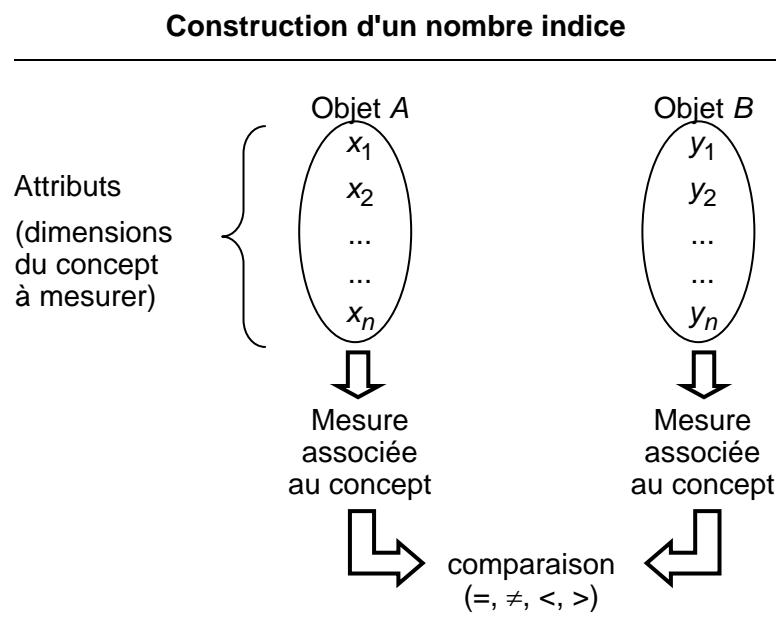
<sup>4</sup> Ou, ce qui revient au même, mesurer simultanément la similarité sous plusieurs aspects ou, pour le dire de façon elliptique, mesurer la similarité entre deux objets multidimensionnels.

Pour mesurer la similarité dans les exemples qui précèdent, on doit tenir compte de plus d'une dimension, parce que le rapport sous lequel on examine la similarité réfère à un concept qui comprend plus d'une dimension :

- S'agissant de la similarité entre pays quant à leur structure économique, il faut tenir compte des différentes branches de la production.
- S'agissant de la similarité entre activités quant à la répartition spatiale, il faut tenir compte des différentes parties du territoire (zones, districts, provinces, ou autres, selon le découpage géographique utilisé).

En un sens, donc, les mesures de la similarité entre objets multidimensionnels s'apparentent à des indices. Pour bien expliciter les différences, nous allons nous attacher dans les lignes qui suivent à faire ressortir ce qu'il y a de spécifique à la mesure de la similarité.

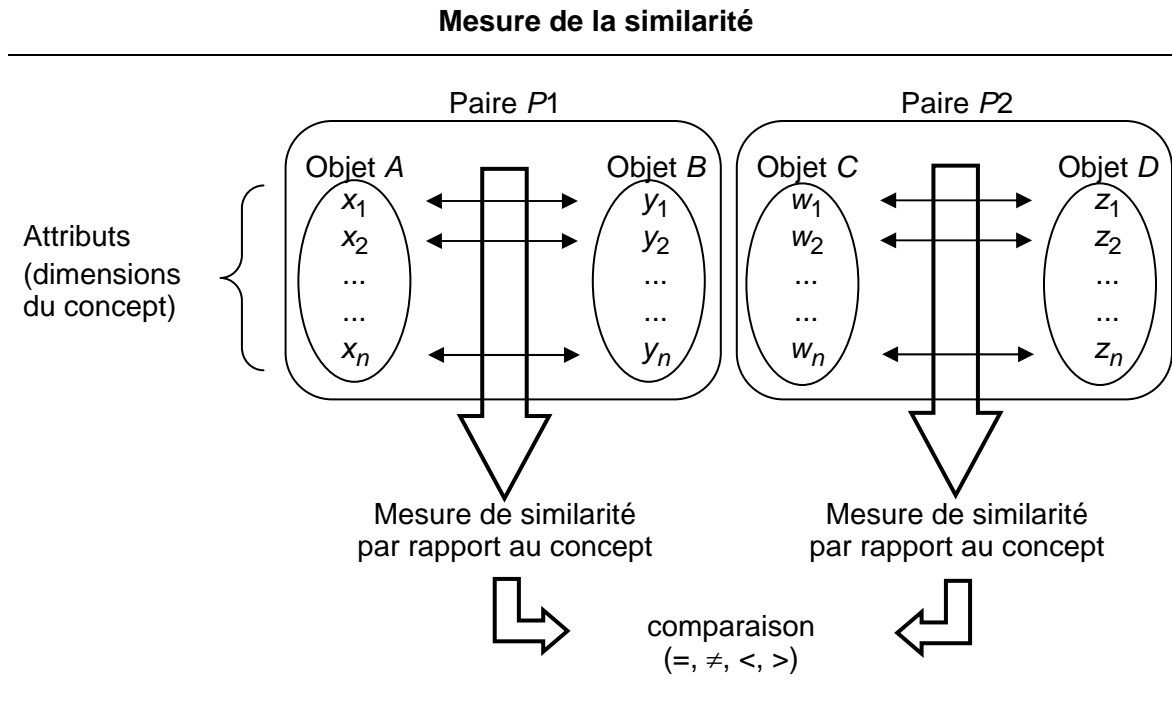
Un indice, comme nous l'avons vu, résume en un seul chiffre les valeurs des indicateurs associés aux multiples dimensions d'un concept. L'indice est une mesure, parce qu'il permet de comparer deux objets quant au degré auquel ils possèdent la propriété définie par le concept. Cela est résumé dans le schémas suivant.



Pour mesurer la similarité, en revanche, on compare d'abord deux objets trait pour trait. On obtient ainsi autant de mesures partielles de similarité qu'il y a de dimensions à la comparaison. Il faut ensuite agréger toutes ces mesures partielles en une seule. Le résultat est une mesure



de la similarité : cette mesure permet de comparer deux *paires* d'objets quant à leur similarité par rapport à un attribut multidimensionnel donné. Cela est résumé dans le schéma qui suit. La comparaison entre les deux schémas montre en quoi la mesure de la similarité entre objets multidimensionnels est différente de la construction d'un indice.



Néanmoins, il s'agit bel et bien d'une mesure au sens où nous l'avons définie au chapitre 1-1. Rappelons qu'une mesure associée à un concept établit une correspondance entre les objets et des nombres, ce qui permet de comparer les objets et de déterminer la valeur de vérité d'une ou de plusieurs des relations  $=$ ,  $\neq$ ,  $>$  ou  $<$ . Une mesure de la similarité est donc une correspondance qui permet de comparer deux *paires* d'objets quelconques du point de vue de leur similarité quant à un attribut donné. Formellement, si l'on convient que  $f(A,B)$  est la mesure de la similarité entre les objets de la paire  $[A,B]$  et que  $f(C,D)$  est la mesure de la similarité entre les objets de la paire  $[C,D]$ , alors, une mesure de similarité permet de décider d'une ou de plusieurs des relations suivantes :

- $f(A,B) = f(C,D)$
- $f(A,B) \neq f(C,D)$
- $f(A,B) < f(C,D)$
- $f(A,B) > f(C,D)$

Par exemple, si  $A$  est le Nicaragua,  $B$  est le Costa Rica,  $C$  est le Costa Rica et  $D$  est le Canada <sup>5</sup>, une mesure de similarité permet de répondre à la question « Par rapport à la composition de sa production, le Costa Rica ressemble-t-il davantage au Nicaragua ou au Canada ? ». De même, si  $A$  est la culture bananière,  $B$  est l'élevage,  $C$  est la culture bananière et  $D$  est la culture des agrumes, une mesure de la similarité permet de répondre à la question « Par rapport à sa répartition géographique au Costa Rica, la culture bananière ressemble-t-elle davantage à l'élevage ou à la culture des agrumes ? ».

Il est à noter que rien de ce qui précède n'implique que l'on mesure toujours la similarité selon une échelle rationnelle. Il est vrai que les variables utilisées comme mesures de similarité ou de dissimilarité sont souvent des variables rationnelles. Mais le problème de la multidimensionnalité fait qu'en général, il y a plusieurs mesures possibles et il n'y en a aucune qui puisse être considérée d'emblée comme la meilleure. C'est pourquoi, sauf dans des contextes particuliers, les mesures de similarité doivent normalement être interprétées comme des mesures ordinales : il faut se garder de leur donner une interprétation abusive de mesure d'intervalle ou rationnelle.

En outre, les mesures de similarité, comme nous le verrons, sont le plus souvent des *mesures inverses*, c'est-à-dire qu'elles sont en fait des mesures de *dissimilarité*. Il faut y être attentif, car cela peut causer la confusion.

### LA MESURE DE LA DISSIMILARITÉ ENTRE DES DISTRIBUTIONS

Une distribution, ou une répartition, est une propriété (multidimensionnelle) d'une population (au sens général de collection de personnes ou d'objets), lorsque cette population est classée en catégories : c'est le nombre d'individus ou la fraction de la population qui se trouve dans chacune des catégories. Dans les exemples déjà évoqués,

- Les personnes employées dans une économie constituent une « population », que l'on peut classer entre les « catégories » que sont les branches d'activité. La structure économique du pays peut être décrite par une distribution : le nombre de personnes employées par branche d'activités.
- Les hectares de terre consacrés à une activité donnée (la culture bananière, par exemple) constituent une « population », que l'on peut classer entre les « catégories » que sont les

---

<sup>5</sup> Comme le montre cet exemple, il peut arriver que  $B=C$  (ou  $B=D$ , ou  $A=C$ , ou  $A=D$ ), mais ce n'est pas nécessairement le cas.

subdivisions (provinces ou autres) d'un territoire. La répartition spatiale de l'activité peut être décrite par une distribution : le nombre d'hectares qui lui sont consacrés dans chaque subdivision du territoire.

Une distribution est donc un objet multidimensionnel. Mais la comparaison entre les distributions est grandement facilitée du fait qu'il existe une « règle de normalisation » naturelle : la mesure associée à chacune des dimensions de la distribution est simplement la fraction de la population appartenant à la catégorie correspondante. Or dans une distribution, la somme des parts est nécessairement égale à 1. Cela élimine d'emblée une partie du problème de la multidimensionnalité, celui, déjà mentionné à propos des nombres indices, du poids à accorder à chacune des dimensions.

Par contre, lorsqu'on tente de comparer deux objets qui ne sont pas des distributions, le choix de l'unité de mesure de chaque dimension de la comparaison détermine implicitement quel sera son poids dans la mesure de dissimilarité. Se pose alors dans toute son intensité le problème de multidimensionnalité évoqué à propos des nombres indices.

#### **DISSIMILARITÉ ET INÉGALITÉ-CONCENTRATION : QUELLE DIFFÉRENCE ?**

Dans les exemples donnés jusqu'ici, il s'est agi simplement d'examiner le degré d'association entre deux phénomènes ou inversement, le degré de ségrégation entre eux. Mais il y a une autre utilisation des mesures de dissimilarité entre deux distributions : c'est la mesure de la concentration ou de la dispersion. Une mesure de dissimilarité devient une mesure de concentration lorsqu'on compare la distribution étudiée avec une distribution de référence ou *théorique*. Cette distribution théorique, qui sert de point de référence, représente une concentration nulle et elle sert en quelque sorte d'étalon de mesure (nous verrons un exemple de cela plus loin).

Cela est cohérent avec ce que nous avons vu au chapitre 1-4 : en général, une mesure de l'inégalité compare la distribution observée avec une distribution de référence, qui représente l'égalité parfaite. Une mesure d'inégalité est donc une mesure de dissimilarité entre la distribution observée et la distribution de référence.

Il s'ensuit que l'indice de Gini est tout aussi approprié comme mesure de dissimilarité que comme mesure d'inégalité. D'ailleurs, nous avons déjà signalé parmi les propriétés de l'indice de Gini que celui-ci est symétrique, c'est-à-dire que les rôles de la distribution examinée et de la

distribution de référence sont interchangeables ; en d'autres mots, si l'on intervertit les rôles, la valeur du coefficient Gini est inchangée.

## 1-5.2 L'indice de dissimilarité

### UN EXEMPLE NUMÉRIQUE

Nous considérons maintenant une mesure de dissimilarité largement utilisée, qui s'applique aux distributions comme, par exemple, la répartition géographique de l'emploi. Voici un exemple numérique fictif :

**Emploi par zone et par branche**

BRANCHE	B1	B2	B3	Total
ZONE				
Z1	48	325	287	660
Z2	27	185	148	360
Z3	45	90	45	180
Total	120	600	480	1200

Il s'agit de mesurer la similarité entre les branches d'activité quant à leur répartition géographique. On s'intéresse donc à la fraction de l'emploi de chaque branche dans chaque zone :

**Distribution de l'emploi entre zones**

BRANCHE	B1	B2	B3	Total
ZONE				
Z1	0,400	0,542	0,598	0,550
Z2	0,225	0,308	0,308	0,300
Z3	0,375	0,150	0,094	0,150
Total	1,000	1,000	1,000	1,000

La façon la plus simple qui se puisse imaginer d'examiner la similarité entre deux distributions consiste à regarder les différences entre ces fractions zone par zone. Faisons la comparaison entre les branches *B1* et *B2* :

**Comparaison de la répartition géographique  
 des branches B1 et B2**

BRANCHE	B1	B2	Écart
ZONE			
Z1	0,400	0,542	0,142
Z2	0,225	0,308	0,083
Z3	0,375	0,150	-0,225
Total	1,000	1,000	0,000

Chacun des écarts calculés constitue l'une des dimensions de la dissimilarité entre les deux répartitions géographiques. Pour mesurer la dissimilarité, il faut combiner les écarts en un chiffre unique. Il va de soi qu'une simple addition donnera toujours le même résultat, zéro <sup>6</sup>. C'est pourquoi l'on fera la somme des valeurs absolues :

$$|0,142| + |0,083| + |-0,225| = 0,142 + 0,083 + 0,225 \text{ (et non pas } -0,225)$$

Pour des raisons qui deviendront évidentes plus loin, on divise le résultat par deux et on obtient :

$$\frac{|0,142| + |0,083| + |-0,225|}{2} = 0,225$$

**DÉFINITION DE L'INDICE DE DISSIMILARITÉ**

***La mesure de la dissimilarité dans une table de contingence : rappel de la notation***

Pour formaliser la présentation, nous reprenons, en la généralisant, la notation développée à la section 1-2.1 <sup>7</sup>. Nous traitons une table de contingence à deux dimensions. Convenons que les colonnes correspondent à  $n$  groupes différents, alors que les lignes correspondent à  $m$  catégories différentes (dans notre exemple, comme à la section 1-2.1, les  $n$  « groupes » sont les 3 branches d'activité, tandis que les  $m$  « catégories » sont les 3 zones.

<sup>6</sup> Puisque  $\sum_i v_i = \sum_i w_i = 1$ , alors  $\sum_i (v_i - w_i) = \sum_i v_i - \sum_i w_i = 0$ .

<sup>7</sup> Le lecteur est invité à se référer à la section 1-2.1 pour un énoncé des identités fondamentales qui se vérifient dans une table de contingence.

$x_{ij}$	nombre d'emplois de la branche $j$ dans la zone $i$
$x_{\bullet j} = \sum_i x_{ij}$	nombre total d'emplois de la branche $j$
$x_{i\bullet} = \sum_j x_{ij}$	nombre total d'emplois dans la zone $i$
$x_{\bullet\bullet} = \sum_i \sum_j x_{ij}$	nombre total d'emplois de toutes branches dans toutes zones
$p_{ij} = \frac{x_{ij}}{x_{\bullet\bullet}}$	fraction de l'emploi total global qui appartient à la branche $j$ et est situé dans la zone $i$
$p_{\bullet j} = \sum_i p_{ij}$	fraction de l'emploi total global qui appartient à la branche $j$
$p_{i\bullet} = \sum_j p_{ij}$	fraction de l'emploi total global qui est situé dans la zone $i$
$p_{j/i\bullet} = \frac{p_{ij}}{p_{i\bullet}}$	fraction de l'emploi total de la zone $i$ qui appartient à la branche $j$
$p_{i/\bullet j} = \frac{p_{ij}}{p_{\bullet j}}$	fraction de l'emploi total de la branche $j$ qui est situé dans la zone $i$

Dans l'exemple numérique ci-haut, nous avons appliqué une mesure de dissimilarité entre deux répartitions géographiques, celle de la branche  $B1$  et celle de la branche  $B2$ . Selon la notation courante, cela correspond à l'application d'une mesure de dissimilarité aux distributions données par les vecteurs

$$Q_1 = \begin{bmatrix} p_{1/\bullet 1} \\ p_{2/\bullet 1} \\ \vdots \\ p_{m/\bullet 1} \end{bmatrix} \text{ et } Q_2 = \begin{bmatrix} p_{1/\bullet 2} \\ p_{2/\bullet 2} \\ \vdots \\ p_{m/\bullet 2} \end{bmatrix}$$

Plus généralement, on compare les distributions

$$Q_h = \begin{bmatrix} p_{1/\bullet h} \\ p_{2/\bullet h} \\ \vdots \\ p_{m/\bullet h} \end{bmatrix} \text{ et } Q_k = \begin{bmatrix} p_{1/\bullet k} \\ p_{2/\bullet k} \\ \vdots \\ p_{m/\bullet k} \end{bmatrix}$$

ou encore les distributions

$$R_g = [p_{1/g\bullet} \quad p_{2/g\bullet} \quad \cdots \quad p_{n/g\bullet}] \text{ et } R_i = [p_{1/i\bullet} \quad p_{2/i\bullet} \quad \cdots \quad p_{n/i\bullet}]$$

NOTE : On peut travailler soit avec des fractions, comme dans la notation ci-haut, soit avec des pourcentages, obtenus en multipliant les fractions par 100. Ici, nous convenons de travailler avec des fractions, parce que cela allège l'écriture des formules. Mais la pratique courante dans la présentation des résultats consiste à présenter des pourcentages, ce qui allège les tableaux, grâce à l'élimination de la virgule décimale.

### **Définition**

Dans ce qui suit, nous appliquons l'indice de dissimilarité à une comparaison des distributions  $Q_h$  et  $Q_k$ . Tous les développements peuvent se transposer aisément à une comparaison entre les distributions  $R_g$  et  $R_j$  ou, en vérité, à n'importe quelle paire de distributions formellement comparables (c'est-à-dire ayant le même nombre de possibilités).

L'indice de dissimilarité se définit comme

$$D = \frac{1}{2} \sum_i |p_{i \bullet h} - p_{i \bullet k}|$$

Dans l'exemple numérique donné ci-haut,

$$Q_1 = \begin{bmatrix} 0,400 \\ 0,225 \\ 0,375 \end{bmatrix} \text{ et } Q_2 = \begin{bmatrix} 0,542 \\ 0,308 \\ 0,150 \end{bmatrix}$$

et

$$D = \frac{|0,400 - 0,542| + |0,225 - 0,308| + |0,375 - 0,150|}{2} = 0,225$$

Cet indice de dissimilarité et ses proches variantes apparaissent sous divers noms dans différentes disciplines. Par exemple,

- On désigne aussi l'indice de dissimilarité par les expressions « indice de différenciation » et « indicateur de dissociation ».
- Lorsque l'une des distributions est la répartition spatiale d'une activité économique et l'autre, celle de l'ensemble des activités, cette mesure correspond à ce que l'on appelle en science régionale le *coefficient de localisation*. Il est à noter toutefois que le coefficient de localisation, bien qu'il se calcule au moyen de la même formule, n'est pas à proprement parler un indice de dissimilarité : nous verrons pourquoi plus loin.

- On connaît aussi, en sciences régionales, le coefficient de *spécialisation*, qui compare la structure économique d'une zone (répartition de l'emploi entre les branches d'activité) avec celle de l'ensemble du territoire à l'étude. Cette mesure non plus, n'est pas à proprement parler un indice de dissimilarité.
- En géographie, Taylor (1977, p. 180) cite une multitude d'appellations pour l'indice de dissimilarité ; parmi ces appellations, « coefficient d'association géographique » est particulièrement déroutante, puisque  $D$  est une mesure de *dissimilarité*, ou de *dissociation*. On trouve même des géographes qui utilisent le terme « indice de Gini » pour désigner l'indice de dissimilarité...
- Les démographes et les sociologues utilisent ce même indice, sous le nom de « coefficient de ségrégation résidentielle » ou « indice de discrimination », pour comparer les distributions spatiales résidentielles de différents groupes ethniques ou raciaux (Mills et Hamilton, 1989, p.233-239 ; Waldorf, 1993).

Que faut-il retenir de cette confusion terminologique ? Ceci : lorsque vous prenez connaissance de résultats de recherche qui font appel à des indices de ce type, assurez-vous de bien vérifier quelle est la formule mathématique utilisée.

Au delà des particularités propres à chaque discipline, examinons cet indice de dissimilarité en tant que mesure de dissimilarité entre deux distributions.

### **L'INDICE DE DISSIMILARITÉ COMME MESURE DE CONCENTRATION OU D'INÉGALITÉ**

Jusqu'à maintenant, nous avons discuté des utilisations de l'indice de dissimilarité pour mesurer la dissimilarité entre deux distributions observées. Mais on peut aussi utiliser l'indice de dissimilarité pour mesurer l'inégalité ou la concentration. D'ailleurs, répétons-le, les mesures d'inégalité ou de concentration sont généralement des mesures de dissimilarité entre une distribution observée et une distribution de référence. Pour mesurer l'inégalité ou la concentration, il s'agit donc de comparer une distribution observée avec une distribution de référence, qui représente l'égalité parfaite ou une concentration nulle (évidemment, dans ce cas, le tableau des données n'est pas une table de contingence).

#### **Exemple**

Supposons que l'on veuille mesurer le degré de concentration géographique de la population sur un territoire donné, préalablement découpé en zones (provinces, districts, ...). Une



concentration nulle correspond à une situation où la densité de la population (habitants/km<sup>2</sup>) est partout la même. On peut donc dire que la concentration est nulle si la fraction de la population dans chaque zone est égale à la fraction du territoire compris dans cette zone.

Soit  $V$ , la distribution de la superficie du territoire et  $W$ , celle de la population.

$$V = [v_1 \quad v_2 \quad \dots \quad v_n] \text{ et } W = [w_1 \quad w_2 \quad \dots \quad w_n]$$

$v_i$  est la fraction de la superficie totale qui est comprise dans la zone  $i$  et  $w_i$  est la fraction de la population qui se trouve dans la zone  $i$ .

La concentration est nulle si

$$w_i = v_i \text{ pour tout } i$$

Dans ce cas, la distribution *observée* du territoire sert de distribution *de référence* à la population : elle est la distribution théorique ou hypothétique d'une population de concentration nulle<sup>8</sup>. On peut alors utiliser l'indice de dissimilarité entre la distribution du territoire et la distribution de la population comme mesure de la concentration géographique de la population.

On aura

$$D = \frac{1}{2} \sum_i |w_i - v_i|$$

Le tableau qui suit illustre cette utilisation de l'indice de dissimilarité. On y mesure le degré de concentration de la population de la Ville de Montréal. Les données de population sont celles du Recensement de 1991. Le territoire est découpé selon les 54 quartiers de planification de la Ville, rangés par ordre décroissant de densité. On obtient  $D = 0,2361$ , c'est-à-dire que, pour obtenir une densité uniforme, il faudrait déplacer 23,61 % de la population d'un quartier à un autre.

---

<sup>8</sup> En d'autres mots, la distribution  $V$  est observée quand il s'agit du territoire, mais elle devient hypothétique quand on l'applique à la population

**Mesure de la concentration de la population au moyen de l'indice de dissimilarité :**

**Ville de Montréal (54 quartiers de planification), population Recensement 1991**

Quartier	Données		Densité hab/km <sup>2</sup>	Répartitions		Écart absolu
	Pop. 1991	Superf. km <sup>2</sup>		Pop.	Superf.	
11	29469	1,65	17860	2,90%	0,88%	0,0201
8	10604	0,72	14728	1,04%	0,38%	0,0066
18	27022	2,03	13311	2,66%	1,08%	0,0157
34	24258	1,85	13112	2,38%	0,99%	0,0140
13	30314	2,39	12684	2,98%	1,28%	0,0170
35	14187	1,24	11441	1,39%	0,66%	0,0073
31	19652	1,73	11360	1,93%	0,92%	0,0101
33	15752	1,40	11251	1,55%	0,75%	0,0080
42	25495	2,32	10989	2,51%	1,24%	0,0127
15	19126	1,75	10929	1,88%	0,93%	0,0095
16	15030	1,38	10891	1,48%	0,74%	0,0074
29	15606	1,46	10689	1,53%	0,78%	0,0075
9	21348	2,02	10568	2,10%	1,08%	0,0102
32	14737	1,48	9957	1,45%	0,79%	0,0066
40	20350	2,15	9465	2,00%	1,15%	0,0085
14	15973	1,80	8874	1,57%	0,96%	0,0061
10	14165	1,65	8585	1,39%	0,88%	0,0051
27	11592	1,41	8221	1,14%	0,75%	0,0039
17	16167	2,00	8084	1,59%	1,07%	0,0052
30	29664	3,69	8039	2,91%	1,97%	0,0095
45	24738	3,23	7659	2,43%	1,72%	0,0071
46	19880	2,60	7646	1,95%	1,39%	0,0057
39	34906	4,85	7197	3,43%	2,59%	0,0084
51	8452	1,20	7043	0,83%	0,64%	0,0019
23	18672	2,67	6993	1,83%	1,43%	0,0041
12	14980	2,21	6778	1,47%	1,18%	0,0029
6	16785	2,48	6768	1,65%	1,32%	0,0033
19	11499	1,75	6571	1,13%	0,93%	0,0020
4	23636	3,70	6388	2,32%	1,98%	0,0035
44	18699	2,96	6317	1,84%	1,58%	0,0026
24	13665	2,22	6155	1,34%	1,19%	0,0016
21	20564	3,62	5681	2,02%	1,93%	0,0009
48	17038	3,02	5642	1,67%	1,61%	0,0006
41	20092	3,59	5597	1,97%	1,92%	0,0006
5	18478	3,36	5499	1,82%	1,79%	0,0002
49	14687	2,73	5380	1,44%	1,46%	0,0001
20	27819	5,22	5329	2,73%	2,79%	0,0005
43	24957	4,84	5156	2,45%	2,58%	0,0013
3	18052	3,56	5071	1,77%	1,90%	0,0013
28	17764	3,56	4990	1,75%	1,90%	0,0015
2	25181	5,25	4796	2,47%	2,80%	0,0033
26	19073	4,01	4756	1,87%	2,14%	0,0027
22	9651	2,18	4427	0,95%	1,16%	0,0022
38	12512	3,16	3959	1,23%	1,69%	0,0046
7	22660	5,84	3880	2,23%	3,12%	0,0089
1	22613	5,85	3865	2,22%	3,12%	0,0090
52	35098	9,50	3695	3,45%	5,07%	0,0162
50	14403	4,07	3539	1,42%	2,17%	0,0076
47	13111	4,45	2946	1,29%	2,38%	0,0109
54	47534	19,04	2497	4,67%	10,16%	0,0549
37	3546	2,06	1721	0,35%	1,10%	0,0075
25	4009	4,28	937	0,39%	2,28%	0,0189
53	11970	13,92	860	1,18%	7,43%	0,0625
36	431	4,24	102	0,04%	2,26%	0,0222
Total	1017666	187,34	5432	100,00%	100,00%	0,472

**Indice de dissimilarité : 0,2361**

## PROPRIÉTÉS DE L'INDICE DE DISSIMILARITÉ

### ***L'indice de dissimilarité et les propriétés d'une mesure d'inégalité***

Puisqu'une mesure d'inégalité est une mesure de dissimilarité entre la distribution observée et une distribution de référence, les propriétés désirables d'une mesure d'inégalité sont également désirables d'une mesure de dissimilarité. Qu'en est-il donc de l'indice de dissimilarité  $D$  ?

Rappelons les propriétés désirables d'une mesure d'inégalité selon Valeyre (1993) :

1. Une mesure d'inégalité doit prendre des valeurs non négatives, puisqu'il s'agit d'une mesure de l'éloignement de la distribution observée par rapport à la distribution de référence.
2. Une mesure d'inégalité doit prendre la valeur zéro si, et seulement si, la distribution observée est identique à la distribution de référence.
3. Toutes les observations doivent être traitées de la même manière.
4. Une mesure d'inégalité doit être indépendante de la valeur moyenne de la variable examinée ; une mesure de concentration doit être indépendante de la taille de la population dont on étudie la distribution.
5. L'agrégation d'observations affichant le même degré de spécificité ne doit pas changer la valeur de la mesure.
6. Une mesure d'inégalité doit diminuer si la distribution est modifiée d'une façon qui réduit incontestablement l'inégalité (Principe de transfert de Pigou-Dalton).

L'indice de dissimilarité possède les propriétés 1 à 5, mais pas la propriété 6 : sa valeur demeure inchangée après un transfert entre deux catégories dont les spécificités sont toutes deux supérieures ou toutes deux inférieures à 1<sup>9</sup>.

### ***Domaine de variation***

Si l'on vous dit que vous avez obtenu la note 18 à un examen, serez-vous content ? Cette note est-elle une bonne, ou une mauvaise note ? Pour le savoir, il faut d'abord savoir quelle est la

---

<sup>9</sup> On peut démontrer cette caractéristique de façon très simple à l'aide de l'interprétation géométrique de l'indice de dissimilarité comme distance verticale maximum entre la courbe de Lorenz et la diagonale. Voir ci-après.

note maximale <sup>10</sup>. Si l'examen est noté sur 20, la note 18 est probablement une bonne note ; s'il est noté sur 100, vous ne serez sans doute pas content...

C'est pour cette raison que l'on s'intéresse au domaine de variation d'une mesure. Le domaine de variation d'une mesure est l'ensemble des valeurs qu'elle peut prendre. Pour une mesure continue, le domaine de variation est défini par la valeur minimum et la valeur maximum que peut prendre la mesure. Pour pouvoir savoir si une valeur donnée est « grande » ou non, il faut au moins connaître son domaine de variation, pour voir si cette valeur est plus proche du maximum ou du minimum.

Dans le cas de l'indice de dissimilarité, sa valeur minimum est zéro : cet indice prend la valeur zéro quand  $p_{i/\bullet h} = p_{i/\bullet k}$  pour tout  $i$ , c'est-à-dire quand les distributions sont identiques.

Quelle est sa valeur maximum ?

Lorsque l'on compare les distributions de deux populations parfaitement distinctes <sup>11</sup>, la valeur maximum que peut prendre l'indice est 1 : cela se produit quand  $p_{i/\bullet h} = 0$  lorsque  $p_{i/\bullet k} > 0$  et vice-versa, c'est-à-dire quand la séparation entre les deux populations est complète : elles ne sont jamais présentes ensemble dans la même catégorie. Dans cette situation en effet, pour chaque catégorie  $i$ , on a

SOIT  $p_{i/\bullet h} = 0$ , et alors

$$|p_{i/\bullet h} - p_{i/\bullet k}| = |0 - p_{i/\bullet k}| = p_{i/\bullet k} = 0 + p_{i/\bullet k} = p_{i/\bullet h} + p_{i/\bullet k}$$

SOIT  $p_{i/\bullet k} = 0$ , et alors

$$|p_{i/\bullet h} - p_{i/\bullet k}| = |p_{i/\bullet h} - 0| = p_{i/\bullet h} = p_{i/\bullet h} + 0 = p_{i/\bullet h} + p_{i/\bullet k}$$

On a donc

$$D^{\max} = \frac{1}{2} \sum_i |p_{i/\bullet h} - p_{i/\bullet k}| = \frac{1}{2} \sum_i (p_{i/\bullet h} + p_{i/\bullet k})$$

$$D^{\max} = \frac{1}{2} \left( \sum_i p_{i/\bullet h} + \sum_i p_{i/\bullet k} \right) = \frac{1+1}{2} = 1$$

---

<sup>10</sup> Ce n'est pas la seule considération. L'interprétation de la note dépend aussi de la note obtenue par les autres et des critères qui sont communément utilisés pour l'interpréter (comme la note de passage).

La division par 2, dans la formule de calcul de l'indice de dissimilarité, a donc pour effet de « normaliser » son domaine de variation à l'intervalle [0, 1].

L'indice de dissimilarité ne peut-il pas prendre une valeur supérieure à 1 ? Non. Pour s'en convaincre, il suffit de se demander, à partir de la situation de séparation complète décrite ci-haut, quelle serait la conséquence de déplacer un individu d'une catégorie à une autre (effet nul si cet individu reste avec ceux de son espèce ; autrement, la valeur de l'indicateur diminue). L'exemple numérique suivant illustre le cas de la ségrégation totale.

**Indice de dissimilarité : exemple de ségrégation totale**

ETHNIE	Nombres			Répartitions			Écart $ p_{i \cdot h} - p_{i \cdot k} $
	Martiens $x_{i1}$	Terriens $x_{i2}$	Total $x_{i1} + x_{i2}$	Martiens $p_{i \cdot 1}$	Terriens $p_{i \cdot 2}$	Total $p_{i \cdot}$	
PLANÈTE							
TERRE	0	6	6	0,00	0,75	0,40	0,75
LUNE	0	2	2	0,00	0,25	0,13	0,25
MARS	3	0	3	0,43	0,00	0,20	0,43
JUPITER	4	0	4	0,57	0,00	0,27	0,57
TOTAL	7	8	15	1,00	1,00	1,00	

Indice de dissimilarité :

$$\frac{0,75 + 0,25 + 0,43 + 0,57}{2} = 1,00$$

**Interprétation métaphorique**

Même si l'on connaît parfaitement le domaine de variation d'une mesure, il est parfois difficile d'avoir une intuition concrète de ce qu'est une « grande » valeur. D'où l'utilité d'une interprétation métaphorique. Une interprétation métaphorique, comme son nom l'indique, repose sur une comparaison, une métaphore : « C'est comme si »... Il faut bien se garder de prendre ces interprétations métaphoriques au pied de la lettre.

Pour ce qui est de l'indice de dissimilarité, il compare la distribution de deux groupes parfaitement distincts <sup>12</sup>, disons *h* et *k*. On peut interpréter l'indice comme la fraction du groupe *h* qu'il faudrait déplacer d'une catégorie à l'autre, pour que sa distribution soit identique à celle du groupe *k*.

<sup>11</sup> On entend par là qu'aucun individu n'appartient aux deux populations à la fois.

<sup>12</sup> On entend par là qu'aucun individu n'appartient aux deux populations à la fois.

Ainsi, dans l'exemple numérique donné au début de cette section, l'indice de dissimilarité entre la répartition spatiale des emplois de la branche  $B1$  et ceux de la branche  $B2$  est de 0,225. Cela signifie que, pour rendre la répartition spatiale de  $B1$  identique à celle de  $B2$ , il faudrait déplacer 22,5 % des emplois de  $B1$ .

Ce résultat est facile à démontrer. Commençons par déterminer quelle est la fraction du groupe  $h$  qu'il faudrait déplacer pour passer de la distribution représentée par les  $p_{i/\bullet h}$  à la distribution représentée par les  $p_{i/\bullet k}$ . Il suffit pour cela d'additionner les fractions de population à retirer des catégories (zones, régions,...) « excédentaires » pour les redistribuer dans des catégories « déficitaires ». Désignons par  $A$  l'ensemble des catégories « excédentaires », c'est-à-dire où  $p_{i/\bullet h} > p_{i/\bullet k}$ . Pour chacune des catégories appartenant à l'ensemble  $A$ , la fraction « excédentaire » de la population  $h$  est égale à  $p_{i/\bullet h} - p_{i/\bullet k}$ . Au total, la fraction de la population  $h$  à retirer des catégories « excédentaires » est donc donnée par

$$\sum_{i \in A} (p_{i/\bullet h} - p_{i/\bullet k})$$

On peut de façon équivalente additionner les fractions de population à ajouter aux catégories « déficitaires », c'est-à-dire

$$\sum_{i \notin A} (p_{i/\bullet k} - p_{i/\bullet h})$$

Naturellement,  $\sum_{i \in A} (p_{i/\bullet h} - p_{i/\bullet k}) = \sum_{i \notin A} (p_{i/\bullet k} - p_{i/\bullet h})$

puisque  $\sum_{i \in A} (p_{i/\bullet h} - p_{i/\bullet k}) - \sum_{i \notin A} (p_{i/\bullet k} - p_{i/\bullet h})$

$$= \sum_{i \in A} p_{i/\bullet h} - \sum_{i \in A} p_{i/\bullet k} - \sum_{i \notin A} p_{i/\bullet k} + \sum_{i \notin A} p_{i/\bullet h} = \sum_i p_{i/\bullet h} - \sum_i p_{i/\bullet k} = 0$$

Quel rapport avec l'indice de dissimilarité ? Eh bien, si l'on additionne les deux sommations du membre de gauche de l'équation précédente (au lieu de soustraire la seconde de la première), on obtient

$$\sum_{i \in A} (p_{i/\bullet h} - p_{i/\bullet k}) + \sum_{i \notin A} (p_{i/\bullet k} - p_{i/\bullet h}) = \sum_i |p_{i/\bullet h} - p_{i/\bullet k}|$$

Et puisque les deux termes du membre de droite sont égaux, on a donc

$$\sum_{i \in A} (p_{i/\bullet h} - p_{i/\bullet k}) = \sum_{i \notin A} (p_{i/\bullet k} - p_{i/\bullet h}) = \frac{1}{2} \sum_i |p_{i/\bullet h} - p_{i/\bullet k}| = D$$

Et voilà une autre bonne raison de diviser la somme par 2 !

### **Symétrie**

Il est à noter que l'indice de dissimilarité  $D$  est symétrique par rapport aux groupes  $h$  et  $k$  :

$$D = \frac{1}{2} \sum_i |p_{i/\bullet h} - p_{i/\bullet k}| = \frac{1}{2} \sum_i |p_{i/\bullet k} - p_{i/\bullet h}|$$

Par conséquent, on peut tout aussi bien interpréter l'indicateur comme la fraction du groupe  $k$  qu'il faudrait déplacer pour que sa distribution soit identique à celle du groupe  $h$  : quel que soit le groupe que l'on imagine de déplacer pour rendre sa distribution identique à celle de l'autre, la fraction à déplacer est la même (ainsi, on pourrait dire qu'il faudrait déplacer 22,5 % de l'emploi de la branche  $B2$  pour rendre sa distribution identique à celle de  $B1$ ). Pour ce qui est toutefois du *nombre* d'individus à déplacer, il est naturellement égal à cette fraction, multipliée par le chiffre de la population. Si les deux groupes sont de taille différente, le nombre d'individus à déplacer (hypothétiquement) sera différent selon que l'on imagine de déplacer une fraction de l'un ou de l'autre.

Insistons de nouveau sur le caractère métaphorique de cette interprétation. D'abord, la similarité des distributions n'est pas nécessairement une bonne chose (qu'on pense à la controverse à propos du *busing* qui a été pratiqué aux États-Unis pour réaliser l'intégration scolaire des Blancs et des Noirs). Ensuite, le déplacement (forcé) des populations n'est décidément pas un moyen acceptable lorsqu'il s'agit de populations humaines.

### **Autres propriétés**

L'indice de dissimilarité, comme tout indice, a ses limites. Outre le non-respect du principe de transfert de Pigou-Dalton, mentionnons :

- Quand les données sont groupées, l'indice de dissimilarité, comme l'indice de Gini, est sensible à la définition et au nombre des catégories utilisées (classes, zones). Cette faiblesse n'est pas trop grave si le découpage choisi est assez fin – s'il comprend un grand

nombre de catégories – mais les comparaisons entre découpages différents sont sans signification<sup>13</sup>.

- Lorsqu'il est utilisé comme mesure de concentration spatiale, l'indice de dissimilarité, comme l'indice de Gini, ne tient aucun compte de la contiguïté ou de la proximité des unités spatiales.
- L'indice de dissimilarité n'admet pas de données négatives. Par exemple, on ne pourrait pas utiliser l'indice de dissimilarité pour mesurer la similarité entre deux branches d'activité quant aux *variations* du nombre d'emplois par zone, parce que ces variations peuvent être négatives.

### **L'indice de dissimilarité et la courbe de Lorenz**

Nous venons de voir que, comme l'indice de Gini, l'indice de dissimilarité peut servir à mesurer la concentration, bien qu'il ne possède pas toutes les propriétés désirables de l'indice de Gini (il lui manque le principe de transfert de Pigou-Dalton). Nous avons vu aussi que l'indice de Gini peut se calculer géométriquement, à partir de la courbe de Lorenz. Existe-t-il un rapport entre l'indice de dissimilarité et la courbe de Lorenz ? Oui !

Il se trouve en effet que l'indice de dissimilarité est égal à l'écart vertical maximum entre la courbe de Lorenz et la diagonale

$$D = \text{MAX}_k [Cv_k - Cw_k]$$

#### **Démonstration :**

Puisque

$$\sum_i (v_i - w_i) = \sum_i v_i - \sum_i w_i = 1 - 1 = 0 ,$$

cette somme contient des termes positifs et des termes négatifs (à moins que tous les termes ne soient nuls). Or l'ordonnancement des observations en ordre croissant des rapports  $w_i / v_i$  fait en sorte que les termes  $(v_i - w_i)$  qui sont positifs précèdent ceux qui sont négatifs. Alors, il est évident que l'écart vertical

---

<sup>13</sup> Cette question est discutée dans les écrits en géographie sous la rubrique MAUP, c'est-à-dire « Modifiable Areal Unit Problem ».



$$Cv_k - Cw_k = \sum_{i=1}^k v_i - \sum_{i=1}^k w_i = \sum_{i=1}^k (v_i - w_i)$$

atteint sa valeur maximum quand on choisit  $k$  de façon à inclure dans la sommation tous les termes positifs, tout en excluant tous ceux qui sont négatifs. Donc

$$\text{MAX}_k [Cv_k - Cw_k] = \sum_{\substack{i \text{ tel que} \\ v_i > w_i}} (v_i - w_i)$$

où, puisque  $\sum_i (v_i - w_i) = 0$ ,

$$\sum_{\substack{i \text{ tel que} \\ v_i > w_i}} (v_i - w_i) = \sum_{\substack{i \text{ tel que} \\ v_i < w_i}} |v_i - w_i| = \frac{1}{2} \sum_i |v_i - w_i| = D$$

L'indice de dissimilarité  $D$  trouve ainsi une interprétation géométrique : c'est la distance maximum entre la diagonale et la courbe de Lorenz associée à la distribution  $V$  (voir l'exemple numérique tiré de Taylor, 1977, et discuté en 1-4.3).

Il est aisé de constater, à l'aide de cette interprétation, que l'indice de dissimilarité est insensible à toute redistribution qui ne réduit pas l'écart vertical maximum mais qui rapproche néanmoins la courbe de Lorenz de la diagonale. C'est cette insensibilité qui viole le principe de transfert de Pigou-Dalton.

### **Sommaire des propriétés de l'indice de dissimilarité**

1. Possède les 5 premières propriétés désirables d'une mesure d'inégalité, mais pas la dernière (il manque le principe de transfert de Pigou-Dalton ; Valeyre, 1993)
2. Domaine de variation (valeurs maximum et minimum)
  - $D = 0$  quand  $p_{i/\bullet h} = p_{i/\bullet k}$  pour tout  $i$  (les deux distributions sont identiques)
  - $D = 1$  quand il y a ségrégation complète :
    - soit  $p_{i/\bullet k} > 0$ , et alors,  $p_{i/\bullet h} = 0$
    - soit  $p_{i/\bullet h} > 0$ , et alors,  $p_{i/\bullet k} = 0$

3.  $D$  est symétrique par rapport aux groupes  $h$  et  $k$  :

$$D = \frac{1}{2} \sum_i |p_{i/\bullet h} - p_{i/\bullet k}| = \frac{1}{2} \sum_i |p_{i/\bullet k} - p_{i/\bullet h}|$$

4. Interprétation métaphorique (groupes parfaitement distincts) :

$D$  = fraction du groupe  $h$  qu'il faudrait déplacer pour que sa distribution soit identique à celle du groupe  $k$  ou vice-versa.

5. Quand les données sont groupées,  $D$ , aussi bien que  $G$ , est sensible à la définition et au nombre de catégories utilisées (classes, zones).

Cela implique notamment que l'agrégation d'une ou de plusieurs catégories peut entraîner une diminution de la valeur de l'indice de dissimilarité.

6. En tant que mesure de concentration spatiale, l'indice de dissimilarité, comme le Gini, ne tient aucun compte de la proximité dans l'espace des différentes zones de forte densité.
7. Ne s'applique pas à des données négatives (ex. : comparaison des variations de l'emploi).
8.  $D$  est égal à l'écart vertical maximum entre la courbe de Lorenz et la diagonale.

#### APPLICATION DE L'INDICE DE DISSIMILARITÉ À UNE DICHOTOMIE

##### *Équivalence de la formule de Duncan et Duncan (1955)*

Lorsqu'on ne distingue que deux groupes, on a affaire à une dichotomie : on compare alors un groupe  $h$  avec le reste de la population (qui joue le rôle du groupe  $k$ ). Pour le groupe  $k$ , on a alors

$$p_{i/\bullet k} = \frac{x_{i\bullet} - x_{ih}}{x_{\bullet\bullet} - x_{\bullet h}} = \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}}$$

Et dans ce cas, on peut écrire l'indice de dissimilarité sous la forme

$$D = \frac{\sum_i p_{i\bullet} |p_{h/i\bullet} - p_{\bullet h}|}{2 p_{\bullet h} (1 - p_{\bullet h})}$$

Cette seconde définition, qui est celle donnée dans l'article classique de Duncan et Duncan (1955), est équivalente à celle que nous avons donnée précédemment, lorsqu'on l'applique à une dichotomie.

L'équivalence entre les deux définitions dans le cas d'une dichotomie se démontre comme suit :

$$D = \frac{1}{2} \sum_i |p_{i/\bullet h} - p_{i/\bullet k}| = \frac{1}{2} \sum_i \left| p_{i/\bullet h} - \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}} \right| = \frac{1}{2} \sum_i \left| \frac{p_{ih}}{p_{\bullet h}} - \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}} \right|$$

$$D = \frac{1}{2} \sum_i p_{i\bullet} \left| \frac{p_{ih}/p_{i\bullet}}{p_{\bullet h}} - \frac{1 - (p_{ih}/p_{i\bullet})}{1 - p_{\bullet h}} \right| = \frac{1}{2} \sum_i p_{i\bullet} \left| \frac{p_{h/i\bullet}}{p_{\bullet h}} - \frac{1 - p_{h/i\bullet}}{1 - p_{\bullet h}} \right|$$

$$D = \frac{\sum_i p_{i\bullet} |p_{h/i\bullet}(1 - p_{\bullet h}) - (1 - p_{h/i\bullet})p_{\bullet h}|}{2 p_{\bullet h}(1 - p_{\bullet h})}$$

$$D = \frac{\sum_i p_{i\bullet} |p_{h/i\bullet} - p_{h/i\bullet} p_{\bullet h} - p_{\bullet h} + p_{h/i\bullet} p_{\bullet h}|}{2 p_{\bullet h}(1 - p_{\bullet h})} = \frac{\sum_i p_{i\bullet} |p_{h/i\bullet} - p_{\bullet h}|}{2 p_{\bullet h}(1 - p_{\bullet h})}$$

Cette formule se prête à une interprétation intéressante.

Au numérateur, on a une moyenne pondérée des écarts absolus  $|p_{h/i\bullet} - p_{\bullet h}|$  entre, d'une part, la proportion  $p_{h/i\bullet}$  du groupe  $h$  dans chaque catégorie  $i$  et, d'autre part, la proportion  $p_{\bullet h}$  du groupe  $h$  dans l'ensemble de la population : le poids  $p_{i\bullet}$  de chaque catégorie est proportionnel à sa population, tous groupes confondus.

Quant à l'expression au dénominateur, elle est égale à l'écart absolu moyen entre les *individus* (et non pas entre les catégories) de la variable dichotomique d'appartenance au groupe  $h$ . Cet écart absolu moyen est égal à deux fois la variance de la même variable.

Soit en effet la variable dichotomique d'appartenance  $g_t$  :

$$g_t \begin{cases} = 1 \text{ si l'individu } t \text{ appartient au groupe } h \\ = 0 \text{ autrement} \end{cases}$$

où l'indice  $t$  se rapporte aux individus des deux groupes :  $t$  varie de 1 à  $x_{\bullet\bullet}$ .

La variable  $g_t$  a une distribution binomiale, dont la moyenne est donnée par

$$\mu_g = \frac{\sum_t g_t}{x_{\bullet\bullet}} = \frac{\sum_i x_{ih}}{x_{\bullet\bullet}} = \frac{x_{\bullet h}}{x_{\bullet\bullet}} = p_{\bullet h}$$

L'écart absolu moyen (mean deviation) est donné par

$$d_g = \frac{\sum_t |g_t - \mu_g|}{x_{..}} = \frac{\sum_t |g_t - p_{.h}|}{x_{..}} = \frac{\sum_{t \text{ tel que } g_t=1} |g_t - p_{.h}| + \sum_{t \text{ tel que } g_t=0} |g_t - p_{.h}|}{x_{..}}$$

$$d_g = \frac{p_{.h}x_{..}|1 - p_{.h}| + (1 - p_{.h})x_{..}|0 - p_{.h}|}{x_{..}} = p_{.h}|1 - p_{.h}| + (1 - p_{.h})|0 - p_{.h}|$$

$$d_g = 2p_{.h}(1 - p_{.h})$$

La variance, quant à elle, est donnée par

$$\sigma_g^2 = \frac{\sum_t (g_t - p_{.h})^2}{x_{..}} = \frac{\sum_t (g_t^2 - 2s_t p_{.h} + p_{.h}^2)}{x_{..}} = \frac{\sum_t g_t^2 - 2p_{.h} \sum_t g_t + \sum_t p_{.h}^2}{x_{..}}$$

$$\sigma_g^2 = \frac{\sum_t g_t - 2p_{.h} \sum_t g_t + \sum_t p_{.h}^2}{x_{..}} = \frac{p_{.h}x_{..} - 2p_{.h}(p_{.h}x_{..}) + p_{.h}^2x_{..}}{x_{..}}$$

$$\sigma_g^2 = p_{.h} - p_{.h}^2 = p_{.h}(1 - p_{.h})$$

**Le coefficient de localisation et l'indice de dissimilarité : pas la même chose !**

En science régionale, le coefficient de localisation <sup>14</sup> est largement utilisé pour mesurer le degré de spécificité de la répartition spatiale d'une activité économique par rapport à l'ensemble.

Dans une table de contingence de l'emploi par zone et par branche,  $p_{i/.h}$  désigne la fraction de l'emploi total de la branche  $h$  qui est situé dans la zone  $i$ ; et  $p_{i.}$  désigne la fraction de l'emploi total de l'ensemble des branches qui est situé dans la zone  $i$ . Le coefficient de localisation se définit comme

$$CL = \frac{1}{2} \sum_i |p_{i/.h} - p_{i.}|$$

<sup>14</sup> Selon Isard (1960, p. 251), c'est à P. Sargant Florence que l'on doit l'introduction du coefficient de localisation parmi les outils de la science régionale; Duncan et Duncan (1955) citent P. Sargant FLORENCE, W. G. FRITZ et R. C. GILLES, « Measures of industrial distribution », chap. 5 dans : National Resources Planning Board, *Industrial Location and National Resources*, Washington, Government Printing Office, 1943.

À première vue, c'est un indice de dissimilarité. Mais non ! En vérité, la relation entre le coefficient de localisation  $CL$  et l'indice de dissimilarité  $D$  est donnée par

$$CL = (1 - p_{\bullet h})D$$

Démonstration :

$D$  étant appliqué à une dichotomie, on a

$$D = \frac{1}{2} \sum_i |p_{i/\bullet h} - p_{i/\bullet k}| = \frac{1}{2} \sum_i \left| p_{i/\bullet h} - \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}} \right|$$

$$D = \frac{1}{2(1 - p_{\bullet h})} \sum_i |p_{i/\bullet h}(1 - p_{\bullet h}) - p_{i\bullet} + p_{ih}|$$

$$D = \frac{1}{2(1 - p_{\bullet h})} \sum_i |p_{i/\bullet h} - p_{i/\bullet h} p_{\bullet h} - p_{i\bullet} + p_{ih}|$$

$$D = \frac{1}{2(1 - p_{\bullet h})} \sum_i |p_{i/\bullet h} - p_{ih} - p_{i\bullet} + p_{ih}|$$

$$D = \frac{1}{2(1 - p_{\bullet h})} \sum_i |p_{i/\bullet h} - p_{i\bullet}| = \frac{CL}{(1 - p_{\bullet h})}$$

La différence vient de ce que le coefficient de localisation compare la distribution d'un groupe (une branche d'activité) avec celle de l'ensemble dont ce groupe fait partie, alors que l'indice de dissimilarité compare la distribution d'un groupe avec celle du reste de la population (les *autres* activités). Il s'ensuit que l'on ne peut pas donner au coefficient de localisation l'interprétation métaphorique que l'on donne à l'indice de dissimilarité, en termes de la fraction du groupe à déplacer pour obtenir des distributions identiques. En outre, le domaine de variation de  $CL$ , de zéro à  $(1 - p_{\bullet h})$ , est plus étroit pour les branches plus importantes, de sorte qu'il est difficile de comparer les coefficients de localisation de branches de différentes tailles. Par contre, si l'on veut mesurer à quel point la répartition spatiale de chaque activité économique est particulière à cette activité-là, l'indice de dissimilarité présente l'inconvénient d'utiliser une distribution de référence différente pour chaque branche : c'est celle de l'ensemble des autres branches, un ensemble qui est défini différemment pour chaque branche, évidemment.

On peut illustrer ces différences à l'aide de l'exemple utilisé au début de ce chapitre.

**Emploi par zone et par branche et distribution de l'emploi entre zones**

BRANCHE	<i>Emploi</i>					<i>Distribution entre zones</i>				
	B1	B2	B3	B1+2	Total	B1	B2	B3	B1+2	Total
ZONE										
Z1	48	325	287	373	660	0,400	0,542	0,598	0,518	0,550
Z2	27	185	148	212	360	0,225	0,308	0,308	0,294	0,300
Z3	45	90	45	135	180	0,375	0,150	0,094	0,188	0,150
Total	120	600	480	720	1200	1,000	1,000	1,000	1,000	1,000

**Comparaison de la distribution géographique de la branche B3**

**avec celle de l'ensemble des trois branches, puis avec la somme de B1 et B2**

BRANCHE	B3	Total	Dif.absol.	B1+2	Dif.absol.
ZONE					
Z1	0,598	0,550	0,048	0,518	0,080
Z2	0,308	0,300	0,008	0,294	0,014
Z3	0,094	0,150	0,056	0,188	0,094
Total	1,000	1,000	0,113	1,000	0,188

Appliquons donc la formule de calcul de l'indice de dissimilarité à chacune des deux comparaisons. Dans le premier cas (B3 et total), on obtient le coefficient de localisation :

$$CL = \frac{|0,048| + |0,008| + |-0,056|}{2} = 0,056$$

Dans le second cas (B3 et B1+2), on obtient l'indice de dissimilarité :

$$D = \frac{|0,080| + |0,014| + |-0,094|}{2} = 0,094$$

Les résultats numériques sont bel et bien différents, comme prévu. Mais ils sont néanmoins liés par la relation

$$CL = \left(1 - \frac{480}{1200}\right) D = 0,6 \times 0,094 = 0,056$$

où le facteur 0,6 est égal à la part de l'emploi des branches *autres que B3*.

Lorsque la sous-population ne représente qu'une petite fraction de la population parente,  $p_{\bullet h}$  est petit et la valeur du coefficient de localisation est proche de celle de l'indice de dissimilarité.

Dans le cas particulier où il y a ségrégation totale, l'indice de dissimilarité  $D$  est égal à 1 et le coefficient de localisation est égal à la part de l'emploi des branches *autres que B3*. On peut illustrer ce dernier point à l'aide de l'exemple de ségrégation totale déjà étudié.

**Coefficient de localisation : exemple de ségrégation totale**

ETHNIE	Nombres		Répartitions		Écart $ v_i - w_i $
	Martiens $x_i$	Total $y_i$	Martiens $v_i$	Total $w_i$	
PLANÈTE					
TERRE	0	6	0,00	0,40	0,40
LUNE	0	2	0,00	0,13	0,13
MARS	3	3	0,43	0,20	0,23
JUPITER	4	4	0,57	0,27	0,30
TOTAL	7	15	1,00	1,00	

Coefficient de localisation :

$$\frac{0,40 + 0,13 + 0,23 + 0,30}{2} = 0,53 = 1 - \frac{7}{15}$$

= fraction de non-Martiens dans la population = fraction de Terriens

On obtiendrait de même pour les Terriens un coefficient de localisation de

$$0,47 = 1 - \frac{8}{15}$$

**Post scriptum : le coefficient de localisation et les quotients de localisation**

À cause de la ressemblance entre leur noms, on peut être porté à confondre le coefficient de localisation et le quotient de localisation. Mais alors que le coefficient de localisation compare deux distributions, le quotient de localisation compare deux parts (voir ci-haut), c'est-à-dire deux points correspondants sur deux distributions. Il y a cependant une relation entre les deux, que l'on trouve en développant la définition du coefficient de localisation :

$$CL = \frac{1}{2} \sum_i |p_{i/\bullet h} - p_{i\bullet}| = \frac{1}{2} \sum_i p_{i\bullet} \left| \left( \frac{p_{i/\bullet h}}{p_{i\bullet}} \right) - 1 \right| = \frac{1}{2} \sum_i p_{i\bullet} |QL_{ih} - 1|$$

Le coefficient de localisation est une moyenne pondérée des écarts absolus entre les quotients de localisation et la valeur repère 1.

**UN DERNIER REGARD CRITIQUE**

De même que l'on peut construire des indices de prix dont les fondements théoriques sont plus satisfaisants que ceux des indices de Laspeyres et de Paasche, moyennant un degré accru de complication, on peut définir des indicateurs de dissimilarité plus raffinés. Waldorf (1993) en fournit un exemple. Il sied cependant de s'interroger, selon le contexte, sur la pertinence de tels

raffinements et sur leur portée concrète. En outre, la présentation de Waldorf (1993) n'évite pas complètement le piège qui consiste à glisser de la métaphore à l'interprétation littérale : dans le contexte d'une étude de la ségrégation raciale aux États-Unis, l'auteur évoque une mesure de l'« effort requis » par un déplacement de la population.

### 1-5.3 Distance et dissimilarité

Parmi les mesures de dissimilarité, certaines sont des mesures de distances, généralisées à plus de deux ou trois dimensions, en ce sens qu'elles possèdent les propriétés que doit avoir une mesure de distance. Réciproquement, on peut considérer la distance comme un cas particulier de la dissimilarité : la distance est une dissimilarité entre deux objets par rapport à leur situation dans l'espace ou plus simplement entre deux lieux dans l'espace.

Une surface (comme la surface de la terre, si l'on ignore le relief <sup>15</sup>) est un espace à deux dimensions. La spécification d'une situation dans l'espace comporte donc deux dimensions : longitude et latitude, ou coordonnées cartésiennes  $(x,y)$ . Par conséquent, la mesure de la distance géographique comprend elle aussi deux dimensions. Et, même si, dans la vie courante, on utilise sans y penser la distance euclidienne, il y a plus d'une façon de mesurer la distance <sup>16</sup>.

Une mesure de distance doit satisfaire certaines conditions. La fonction  $d(a,b)$  est une fonction de distance si et seulement si, pour tout ensemble de lieux  $a$ ,  $b$  et  $c$ , elle satisfait les quatre conditions suivantes :

(c1) non négativité :

$$d(a,b) \geq 0$$

(c2) identité :

$$d(a,b) = 0 \text{ si, et seulement si, } a = b$$

(c3) symétrie :

$$d(a,b) = d(b,a)$$

(c4) inégalité triangulaire :

$$d(a,c) \leq d(a,b) + d(b,c)$$

---

<sup>15</sup> Si l'on tient compte du relief, on a un espace tri-dimensionnel.

<sup>16</sup> Voir Huriot et Perreur (1990 et 1994).



La mesure de distance la plus familière est la *distance euclidienne*. La distance euclidienne entre le point  $a$ , de coordonnées  $(x_a, y_a)$ , et le point  $b$ , de coordonnées  $(x_b, y_b)$ , est donnée par :

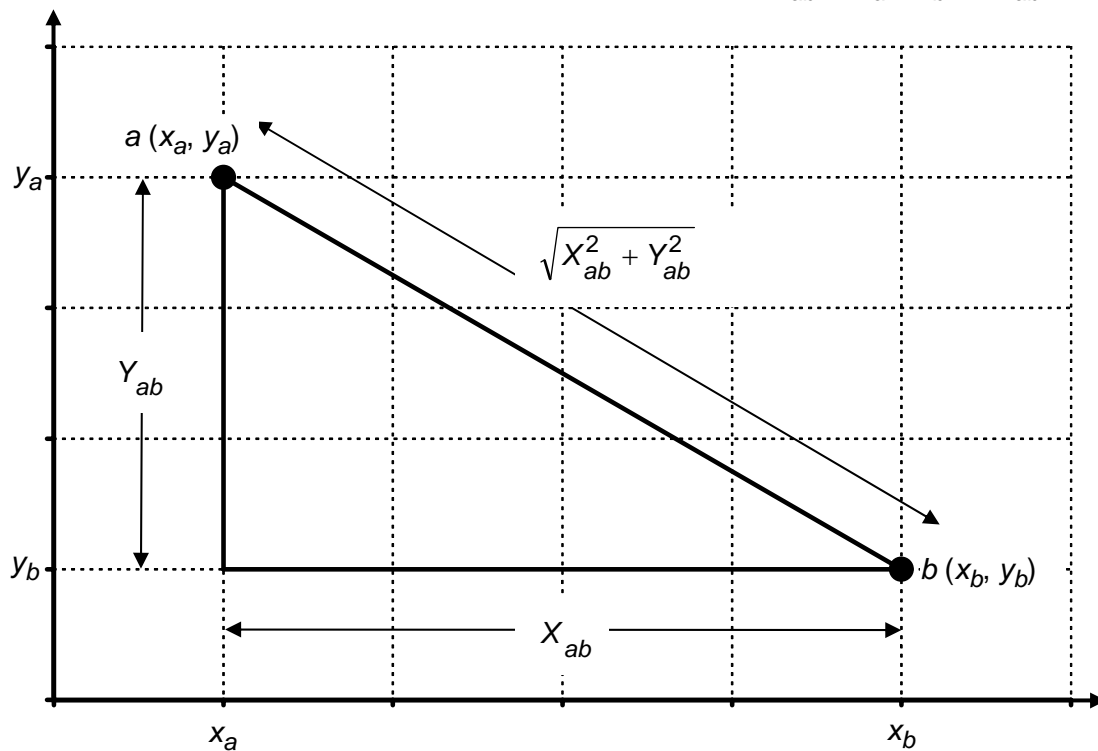
$$d_e(a, b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

Parmi les autres mesures de distance, signalons la *distance rectilinéaire* [aussi appelée distance *rectangulaire* ou distance (*selon la métrique*) de *Manhattan* ; voir Huriot et Perreur, 1994, p. 44] :

$$d_r(a, b) = |x_a - x_b| + |y_a - y_b|$$

La distance selon la métrique de Manhattan est la distance qu'il faut parcourir pour aller de  $a$  à  $b$  en suivant le tracé des rues, lorsque celles-ci forment une grille rectangulaire comme à Manhattan.

Les deux métriques sont illustrées dans la figure qui suit, où  $X_{ab} = |x_a - x_b|$  et  $Y_{ab} = |y_a - y_b|$



Puisque la distance géographique peut être interprétée comme une dissimilarité, la réciproque est aussi vraie : les mesures de la distance peuvent être utilisées pour mesurer des dissimilarités qui ne sont pas des distances géographiques.

Ainsi, considérons deux objets que l'on décrit à l'aide de  $n$  variables, dont chacune mesure une caractéristique (dimension) pertinente :

$x_{11}, x_{12}, \dots, x_{1n}$  pour le premier objet et

$x_{21}, x_{22}, \dots, x_{2n}$  pour le second.

Exemple :

Si les deux objets étaient deux quartiers d'une ville, les caractéristiques pertinentes pourraient être la densité de la population, la proportion de la population ayant moins de quinze ans, la proportion de la population ayant complété l'école primaire, le revenu moyen des ménages, etc.

Pour mesurer la dissimilarité entre deux objets multidimensionnels, on utilise souvent la distance euclidienne généralisée, qui est définie par :

$$\sqrt{\sum_i (x_{1i} - x_{2i})^2}$$

On utilise aussi la distance rectilinéaire généralisée ou distance généralisée selon la métrique de Manhattan, qui est donnée par

$$\sum_i |x_{1i} - x_{2i}|$$

Le lecteur perspicace aura remarqué la parenté entre l'indice de dissimilarité  $D$  et la distance généralisée selon la métrique de Manhattan :  $D$  est égal à la moitié de la distance rectilinéaire généralisée. Dans le présent contexte cependant, les deux objets comparés ne sont pas nécessairement des distributions. Il s'ensuit notamment qu'il n'y a pas de valeur maximum inhérente à la distance rectilinéaire généralisée (ni à la distance euclidienne généralisée, d'ailleurs).

De façon générale, la valeur d'une mesure de distance n'est pas indépendante des unités de mesure des variables sous-jacentes. C'est pourquoi, lorsqu'on compare des objets multidimensionnels à l'aide d'une distance généralisée, on doit affronter un problème analogue à celui auquel on fait face dans la construction d'un nombre indice. En effet, le choix de l'unité de mesure de chaque variable détermine implicitement quel sera son poids dans la mesure de

distance-dissimilarité. C'est seulement quand les objets comparés sont des distributions que le problème des échelles de mesure ne se pose pas.

### 1-5.4 La mesure de la similarité en statistique

Le problème de la mesure de la similarité se pose souvent en statistique. Par exemple, considérons deux séries d'observations sur deux variables :

$$x_1, x_2, \dots, x_n \text{ et } y_1, y_2, \dots, y_n$$

Le coefficient de corrélation simple est une mesure de la similarité entre ces deux séries de données <sup>17</sup>.

De même, pour évaluer l'exactitude d'un modèle par rapport aux données qui ont servi à estimer ses paramètres, on mesure la similarité entre les valeurs observées et les valeurs prédites par le modèle. L'une des mesures les plus utilisées pour cela est le coefficient de détermination multiple  $R^2$  (dont il sera question dans la troisième partie de cet ouvrage).

Enfin, le Khi-deux de Pearson <sup>18</sup> est une mesure de la dissimilarité entre les effectifs observés et les effectifs « théoriques » prédits par une hypothèse.

Toutes ces mesures appartiennent à la grande famille des mesures de similarité et de dissimilarité.

On trouve dans Webber (1984, p. 41-45) une discussion intéressante de la pertinence de différentes mesures d'ajustement (dans le contexte de l'évaluation de l'exactitude du modèle de répartition spatiale de Lowry).

### 1-5.5 Autres mesures de similarité et de dissimilarité

Il existe une grande abondance de mesures de similarité et de dissimilarité. Legendre et Legendre (1984, tome 2, chap. 6 et 1998, chap. 7) présentent et discutent une multitude de mesures, utilisées en écologie numérique et qui pourraient être employées pour l'analyse spatiale en sciences sociales.

---

<sup>17</sup> Voir l'annexe 2-A « Rappel de quelques formules courantes en statistique ». Le coefficient de corrélation mesure plus exactement la similarité entre les valeurs observées d'une variable et ses valeurs prédites à l'aide de l'autre variable.

<sup>18</sup> Voir 4-1. Le Khi-deux n'est cependant pas une mesure symétrique : sa valeur change si l'on intervertit les rôles des valeurs observées et des valeurs théoriques.

## EN GUISE DE CONCLUSION...

Que restera-t-il de toutes les formules, de tous les chiffres et de tous les mots qui constituent cette partie du cours ? Quelques idées clés, peut-être...

- L'approche quantitative repose sur la mesure et mesurer, c'est comparer. Il y a des degrés dans la mesure, dépendamment du type de comparaisons que l'on peut faire (=, ≠, < ou >). Rares sont les cas de mesure parfaitement valide et fiable. En général, on peut associer plus d'une mesure à une même dimension d'un concept : même pour résumer (mesurer) l'évolution temporelle d'une série, il y a plus d'une possibilité.
- Ce n'est pas tout de mesurer, encore faut-il pouvoir attacher une signification aux chiffres. L'interprétation des grandeurs fait souvent appel à une « méta-comparaison », grâce à laquelle une donnée peut être mise en perspective. L'analyse de décomposition est aussi une technique utile d'examen des données, mais il faut se garder de confondre les parties d'une décomposition avec des causes, *a fortiori* lorsque l'une des parties est un résidu...
- Les concepts qui comportent plus d'une dimension soulèvent un problème de mesure qui n'a pas en général de solution unique. Par la construction d'indices, on cherche à apporter au problème de la multidimensionalité la solution la moins imparfaite possible. La validité des indices dépend largement de la validité du modèle sous-jacent. En particulier, les indices qui sont des moyennes pondérées reposent souvent sur des modèles réducteurs (et parfois ne reposent sur rien du tout !) ; qui plus est, les pondérations appliquées sont parfois arbitraires, ce qui a pour effet de dépouiller un indice de son statut de mesure (puisque, en dépit de son apparente « scientificité », l'ordre qu'il établit entre les observations est aussi arbitraire).
- La mesure de l'inégalité ou de la concentration et celle de la dissimilarité sont étroitement apparentées : la plupart des mesures d'inégalité sont des mesures de la dissimilarité entre une distribution observée et une distribution de référence. Il existe toute une panoplie de mesures de ce type. Certaines sont préférables à d'autres parce qu'elles possèdent plusieurs ou même la totalité des propriétés désirables d'une telle mesure. La connaissance de leurs propriétés est une condition préalable à l'utilisation judicieuse de ces divers indices.

## ANNEXE 1-A : QUELQUES OUTILS MATHÉMATIQUES DE BASE

---

### Plan

1. L'opérateur sommation	2
1.1 Définition	2
1.2 Règles de base pour les sommes finies	4
1.3 Sommations doubles	5
Note : L'opérateur produit	7
Exercices sur l'opérateur sommation	7
2. Les logarithmes et la fonction exponentielle	10
2.1 Les exposants	10
2.2 Les logarithmes	11
2.3 La fonction exponentielle	14
2.4 Pourquoi les logarithmes népériens ?	17
Solutions des exercices sur l'opérateur sommation	18

## 1. L'opérateur sommation <sup>1</sup>

### 1.1 DÉFINITION

L'opérateur sommation est simplement une façon compacte d'écrire une somme lorsque les termes successifs peuvent s'écrire sous la forme d'une expression générale qui varie en fonction d'un indice. Par exemple, la somme

$$x_1 + x_2 + x_3 + x_4 + x_5$$

peut s'écrire

$$\sum_{i=1}^5 x_i$$

Dans cette expression,  $i$  est une variable qui prend successivement les valeurs 1, 2, 3, 4 et 5 : le «  $i=1$  » qu'on trouve sous le  $\Sigma$  indique que la valeur initiale de la variable  $i$  est 1 ; le « 5 » qu'on trouve au-dessus du  $\Sigma$  indique que la valeur terminale de la variable  $i$  est 5. La variable  $x_i$  est une fonction de la variable  $i$ , c'est-à-dire que sa valeur dépend de la valeur de  $i$  : quand  $i = 1$ ,  $x_i = x_1$  ; quand  $i = 2$ ,  $x_i = x_2$  ; et ainsi de suite. Enfin, le signe  $\Sigma$  indique qu'il faut *additionner*  $x_1$ ,  $x_2$ ,  $x_3$ ,  $x_4$  et  $x_5$ , les valeurs successives de  $x_i$ . On lit cette expression de la façon suivante : « la somme des  $x_i$  pour  $i$  variant de 1 à 5 ».

De façon plus générale, on aura

$$\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_n$$

De plus, lorsqu'il n'y a pas d'ambiguïté possible sur les valeurs initiale et terminale de l'indice, on peut écrire de façon elliptique

$$\sum_i x_i = x_1 + x_2 + \cdots + x_n$$

---

<sup>1</sup> Ce qui suit est largement inspiré de HOHN, Franz E. (1964) *Elementary matrix algebra*, 2nd ed., MacMillan, New York, Annexe I.

Il est à noter que l'indice  $i$  est un indice muet (*dummy index*). Le choix de la lettre qui sert à représenter l'indice muet est parfaitement arbitraire :

$$\sum_{i=1}^n x_i = \sum_{j=1}^n x_j = \sum_{k=1}^n x_k = x_1 + x_2 + \dots + x_n$$

Il y a aussi un certain degré d'arbitraire dans le choix des valeurs initiale et terminale, comme le montre l'exemple suivant :

$$\sum_{i=1}^n x_i = \sum_{i=0}^{n-1} x_{i+1} = \sum_{i=2}^{n+1} x_{i-1} = x_1 + x_2 + \dots + x_n$$

Il est parfois commode dans les développements mathématiques de pouvoir ainsi décaler l'indice muet.

\* \* \*

Pour calculer la valeur numérique de l'expression

$$\sum_j x_j = x_1 + x_2 + \dots + x_n$$

il faut connaître les valeurs de  $x_1, x_2, \dots, x_n$ . Dans certains cas, la notation permet de connaître directement la valeur de chacun des termes de la sommation. Voici quelques exemples :

$$\sum_{t=1}^n t^2 = 1^2 + 2^2 + \dots + n^2$$

$$\sum_{k=1}^K \left(\frac{1}{k}\right) = \left(\frac{1}{1}\right) + \left(\frac{1}{2}\right) + \left(\frac{1}{3}\right) + \dots + \left(\frac{1}{K}\right)$$

On trouve aussi des expressions comme

$$\sum_{j=0}^n a_j x^j = a_0 x^0 + a_1 x^1 + a_2 x^2 + \dots + a_n x^n$$

où l'indice muet joue à la fois un rôle d'indice proprement dit (dans  $a_j$ ) et un rôle numérique (comme exposant dans  $x^j$ ).

On utilise aussi l'opérateur sommation pour traiter des sommes infinies comme

$$x_1 + x_2 + \dots + x_n + \dots$$

On emploie alors le symbole  $\infty$  pour désigner la valeur terminale de l'indice :

$$\sum_{j=1}^{\infty} x_j = x_1 + x_2 + \cdots + x_n + \cdots$$

## 1.2 RÈGLES DE BASE POUR LES SOMMES FINIES

Les règles de base d'utilisation de l'opérateur sommation sont les suivantes :

1.  $\sum_{i=1}^n c = nc$
2.  $\sum_{i=1}^k x_i + \sum_{i=k+1}^n x_i = \sum_{i=1}^n x_i$
3.  $\sum_{i=1}^n (c x_i) = c \left( \sum_{i=1}^n x_i \right)$
4.  $\sum_{i=1}^t (x_i + y_i) = \sum_{i=1}^t x_i + \sum_{i=1}^t y_i$

Toutes ces règles sauf la première peuvent se déduire de la définition

$$\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_n$$

La première règle est plutôt une convention, que l'on justifie de la façon suivante. Supposons que la variable  $x_j$  soit une constante :

$$x_1 = x_2 = \dots = x_n = c$$

Alors la valeur de la somme est donnée par

$$\sum_{j=1}^n x_j = x_1 + x_2 + \cdots + x_n = c + c + \cdots + c = nc$$

L'expression «  $\sum_{i=1}^n c$  » est donc interprétée comme «  $\sum_{i=1}^n x_i$  où  $x_i = c$  pour tout  $i$  ». De là vient la première règle. Ainsi

$$\sum_{i=1}^5 7 = 5 \times 7 = 35$$



### 1.3 SOMMATIONS DOUBLES

Supposons que l'on ait à traiter un ensemble de  $n \times m$  quantités  $t_{ij}$ , avec  $i = 1, 2, \dots, n$  et  $j = 1, 2, \dots, m$ . Ces quantités peuvent être disposées sous forme de tableau :

$$\begin{array}{cccc} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nm} \end{array}$$

Pour faire la somme de tous les  $t_{ij}$ , on peut d'abord faire le total des termes de chaque ligne, puis additionner les totaux de lignes, ce qui donne

$$\sum_{j=1}^m t_{1j} + \sum_{j=1}^m t_{2j} + \cdots + \sum_{j=1}^m t_{nj}$$

Cette expression peut s'écrire de manière plus compacte à l'aide d'un second opérateur sommation :

$$\sum_{i=1}^n \left( \sum_{j=1}^m t_{ij} \right)$$

On aurait pu, de façon équivalente, faire d'abord le total des termes de chaque colonne, puis additionner les totaux de colonnes :

$$\sum_{j=1}^m \left( \sum_{i=1}^n t_{ij} \right)$$

Puisque, de toute évidence, les deux calculs donnent le même résultat, on a :

$$\sum_{i=1}^n \left( \sum_{j=1}^m t_{ij} \right) = \sum_{j=1}^m \left( \sum_{i=1}^n t_{ij} \right)$$

Pour cette raison, on omet généralement les parenthèses. On a donc la règle

$$5. \quad \sum_{i=1}^n \sum_{j=1}^m t_{ij} = \sum_{j=1}^m \sum_{i=1}^n t_{ij}$$

(La règle 5 ne s'applique pas toujours aux sommations infinies)

On peut évidemment généraliser cette règle à des sommations triples, quadruples, etc.

Il est à noter que dans une double sommation, l'indice de la sommation extérieure peut apparaître comme valeur initiale ou terminale de la sommation intérieure. Dans ce cas cependant, on ne peut pas intervertir les sommations comme le permet la règle 5. Par exemple, supposons que l'on veuille faire la somme des valeurs du tableau triangulaire suivant

$$\begin{array}{cccc} a_{11} & & & \\ a_{21} & a_{22} & & \\ a_{31} & a_{32} & a_{33} & \ddots \\ \vdots & \vdots & \vdots & \ddots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{array}$$

On peut écrire la somme des totaux des lignes

$$\sum_{i=1}^n \sum_{j=1}^i a_{ij}$$

et, de façon plus elliptique,

$$\sum_i \sum_{j \leq i} a_{ij}$$

Ou encore on peut écrire la somme des totaux des colonnes

$$\sum_{j=1}^n \sum_{i=j}^n a_{ij}$$

et, de façon plus elliptique,

$$\sum_j \sum_{i \geq j} a_{ij}$$

Toutefois, on **ne pourrait pas** écrire  $\sum_{i \geq j} \sum_j a_{ij}$  : cela n'aurait aucun sens. En effet, l'expression

$\sum_j \sum_{i \geq j} a_{ij}$  signifie  $\sum_j \left( \sum_{i \geq j} a_{ij} \right)$  : la seconde sommation s'effectue à l'intérieur de la première.

C'est pourquoi la valeur initiale de la première sommation ne peut pas dépendre de l'indice de la sommation intérieure (la seconde sommation).

Supposons que l'on veuille exclure de la somme les termes de la diagonale du tableau ( $a_{11}$ ,  $a_{22}$ , ...,  $a_{nn}$ ). On écrit alors (noter la différence entre  $<$  et  $\leq$ )

$$\sum_{i=1}^n \sum_{j=1}^i a_{ij} - \sum_{i=1}^n a_{ii} = \sum_i \sum_{j \leq i} a_{ij} - \sum_i a_{ii} = \sum_i \sum_{j < i} a_{ij}$$

ou (noter la différence entre  $>$  et  $\geq$ )

$$\sum_{j=1}^n \sum_{i=j}^n a_{ij} - \sum_{i=1}^n a_{ii} = \sum_j \sum_{i \geq j} a_{ij} - \sum_i a_{ii} = \sum_i \sum_{j > i} a_{ij}$$

### NOTE : L'OPÉRATEUR PRODUIT

L'opérateur produit est analogue à l'opérateur sommation. Il sert à écrire de façon compacte des produits :

$$\prod_{i=1}^n x_i = x_1 \times x_2 \times \dots \times x_n$$

L'emploi de l'opérateur produit suit les règles suivantes :

1.  $\prod_{j=1}^n c = c^n$
2.  $\left( \prod_{j=1}^k x_j \right) \left( \prod_{j=k+1}^n x_j \right) = \left( \prod_{j=1}^n x_j \right)$
3.  $\prod_{j=1}^n k x_j = k^n \left( \prod_{j=1}^n x_j \right)$
4.  $\prod_{j=1}^n x_j y_j = \left( \prod_{j=1}^n x_j \right) \left( \prod_{j=1}^n y_j \right)$
5.  $\prod_{i=1}^m \prod_{j=1}^n x_{ij} = \prod_{j=1}^n \prod_{i=1}^m x_{ij}$

### EXERCICES SUR L'OPÉRATEUR SOMMATION

1. Dans ce qui suit, les valeurs des  $x_i$  sont données par l'équation

$$x_i = 5 + 3i$$

Évaluez les expressions suivantes.

$$1.1 \quad \sum_{k=1}^4 x_k$$

$$1.2 \quad \sum_{i=0}^3 x_i$$

$$1.3 \quad \sum_{i=0}^{n-1} x_{i+1} \text{ pour } n = 4$$

2. Calculez

$$2.1 \quad \sum_{x=2}^3 x^3$$

$$2.2 \quad \sum_{i=1}^4 \left(\frac{1}{i}\right)$$

$$2.3 \quad \sum_{j=1}^{10} a, \text{ pour } a = 345$$

3. Démontrez les règles suivantes en explicitant les expressions à partir de la définition de l'opérateur sommation :

$$3.1 \quad \sum_{i=1}^k x_i + \sum_{i=k+1}^n x_i = \sum_{i=1}^n x_i$$

$$3.2 \quad \sum_{i=1}^n (c x_i) = c \left( \sum_{i=1}^n x_i \right)$$

$$3.3 \quad \sum_{i=1}^t (x_i + y_i) = \sum_{i=1}^t x_i + \sum_{i=1}^t y_i$$

4. Dans ce qui suit, on manipule des données disposées sous forme de tableau :

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} = \begin{bmatrix} 4 & 50 \\ 5 & 30 \\ 6 & 10 \end{bmatrix}$$

Évaluez les expressions suivantes :

$$4.1 \quad \sum_i \sum_j a_{ij}$$

$$4.2 \quad \sum_i a_{2i}$$

$$4.3 \quad \sum_{j=1}^2 \sum_{i=1}^2 a_{ij}$$

$$4.4 \quad \sum_{i=1}^2 \sum_{j=1}^2 a_{ij}$$

$$4.5 \quad \sum_{j=1}^2 \sum_{i=1}^j a_{ij}$$

$$4.6 \quad \sum_{i=1}^2 \sum_{j=1}^i a_{ij}$$

$$4.7 \quad \sum_i \sum_{j>i} a_{ij}$$

$$4.8 \quad \sum_i \sum_{j \leq i} a_{ij}$$

Les solutions à ces exercices sont données à la fin de l'annexe.

## 2. Les logarithmes et la fonction exponentielle

### 2.1 LES EXPOSANTS

Pour un nombre réel positif  $b$  et un nombre entier positif  $n$ , l'expression

$$b^n$$

signifie par définition

$$b \times b \times \dots \times b$$

où le nombre réel  $b$  apparaît  $n$  fois. De cette définition, il découle :

1.  $b^m \times b^n = b^{m+n}$
2.  $b^m \div b^n = b^{m-n}$  lorsque  $m > n$
3.  $(b^n)^m = b^{m \times n}$

Suivant la définition que nous en avons donnée initialement, l'expression  $b^n$  n'a de sens que pour  $n$  entier positif. Les trois règles qui précèdent conduisent toutefois aux généralisations suivantes.

Lorsque  $m=n$ ,

$$b^0 = b^{m-n} = b^m \div b^n = 1$$

On a aussi

$$b^{-n} = b^{0-n} = b^0 \div b^n = 1 \div b^n = \frac{1}{b^n}$$

Enfin, si  $a$  est la racine  $n^{\text{ème}}$  de  $b$ , c'est-à-dire si  $a = \sqrt[n]{b}$ , alors  $a^n = b$ . On a donc

$$b^{(\frac{1}{n})} = (a^n)^{(\frac{1}{n})} = (a^{n \times (\frac{1}{n})}) = a^{(\frac{n}{n})} = a = \sqrt[n]{b}$$

Et plus généralement

$$b^{(\frac{m}{n})} = \sqrt[n]{b^m}$$

À la suite de ces généralisations, l'expression  $b^r$  est définie pour tout nombre réel positif  $b$  et pour tout nombre rationnel  $r = \frac{m}{n}$ . Quant aux nombres irrationnels, ils peuvent être atteints par une suite convergente de nombres rationnels, ce qui permet de donner un sens à l'expression

$b^r$ , non seulement lorsque  $r$  est un nombre rationnel, mais plus généralement lorsque  $r$  est un nombre réel, que ce nombre soit rationnel ou irrationnel.

## 2.2 LES LOGARITHMES

Les logarithmes, comme l'opérateur sommation, ne sont rien d'autre qu'une convention d'écriture. L'expression

$$x = \log_b y$$

se lit «  $x$  est le logarithme de  $y$  dans le système de base  $b$  » ou, plus simplement, «  $x$  est le logarithme de  $y$  à base  $b$  ». Elle signifie tout simplement que si l'on élève  $b$  à la puissance  $x$ , on obtient  $y$  :

$$y = b^x = b^{\log_b y}$$

Par exemple,

$$\log_{10} 1 = 0$$

$$\log_{10} 10 = 1$$

$$\log_{10} 100 = 2$$

$$\log_{10} 1000 = 3$$

et ainsi de suite.

Le logarithme à base 10 d'un nombre qui n'est pas une puissance entière de 10 ne sera pas un nombre entier. Par exemple <sup>2</sup>,

$$\log_{10} 2 = 0,30103 \text{ signifie } 10^{0,30103} = 2$$

$$\log_{10} 12 = 1,07918 \text{ signifie } 10^{1,07918} = 12$$

Les bases de logarithmes les plus fréquemment utilisées sont 10 et le nombre irrationnel  $e = 2,71828\dots$ . Les logarithmes à base 10 sont appelés « communs », alors que les logarithmes à base  $e$  sont appelés « naturels » ou « népériens » <sup>3</sup>. On utilise souvent la notation  $\ln y$  au lieu de  $\log_e y$  pour désigner le logarithme népérien de  $y$ .

---

<sup>2</sup> Dans les tables de logarithmes que l'on utilisait avant les Lotus et autres Excel, le logarithme était décomposé en deux parties : la partie entière s'appelait la *caractéristique*, et la partie fractionnelle, la *mantisse*. Dans  $\log 12$ , la caractéristique est égale à 1 et la mantisse, à 07918.

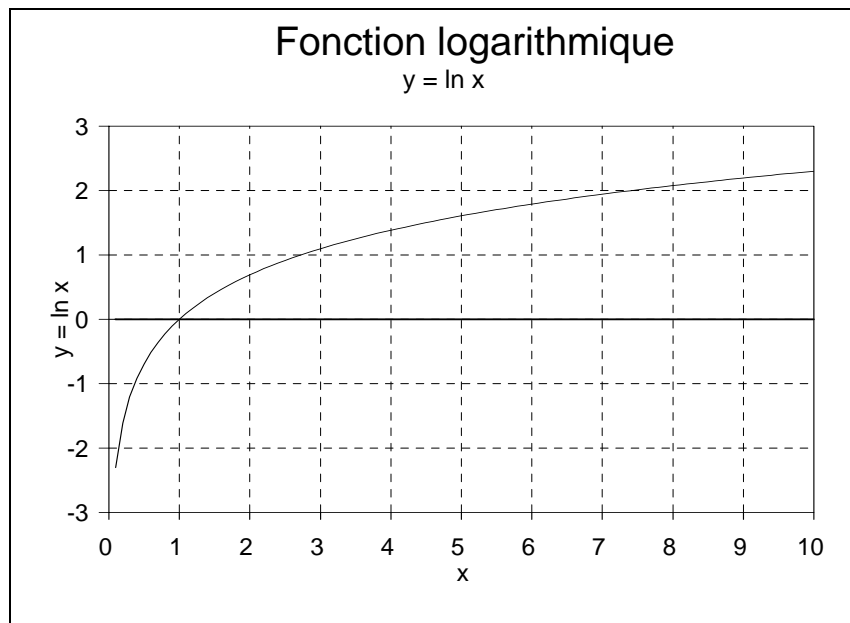
<sup>3</sup> Du nom de leur inventeur, le théologien et mathématicien écossais John Napier (1550-1617), dont le nom s'écrit aussi « Neper ».

L'opération qui consiste à trouver un nombre à partir de son logarithme est parfois désignée par l'expression « antilog » ; ainsi, on a l'équivalence suivante

$$\text{antilog}_b x = b^x$$

Donc,  $\text{antilog } x = e^x$  ou  $10^x$ , selon que l'on considère  $x$  comme un logarithme népérien à base  $e$  ou comme un logarithme commun à base 10.

La figure suivante illustre la relation entre un nombre et son logarithme.



On remarque que le logarithme est une transformation « monotone croissante » : si  $y_1 > y_2$ , alors  $\log y_1 > \log y_2$  ; cela est évident si l'on se rappelle que, par définition,  $y$  et  $\log y$  sont liés par la relation  $y = b^{\log y}$ .

Les règles qui s'appliquent aux exposants se transposent aux logarithmes.

La règle  $b^m \times b^n = b^{m+n}$  implique

$$\log (y \times z) = \log y + \log z$$

La règle  $b^m \div b^n = b^{m-n}$  implique

$$\log \left( \frac{y}{z} \right) = \log y - \log z$$

La règle  $(b^n)^m = b^{m \times n}$  implique

$$\log y^r = r \times \log y$$

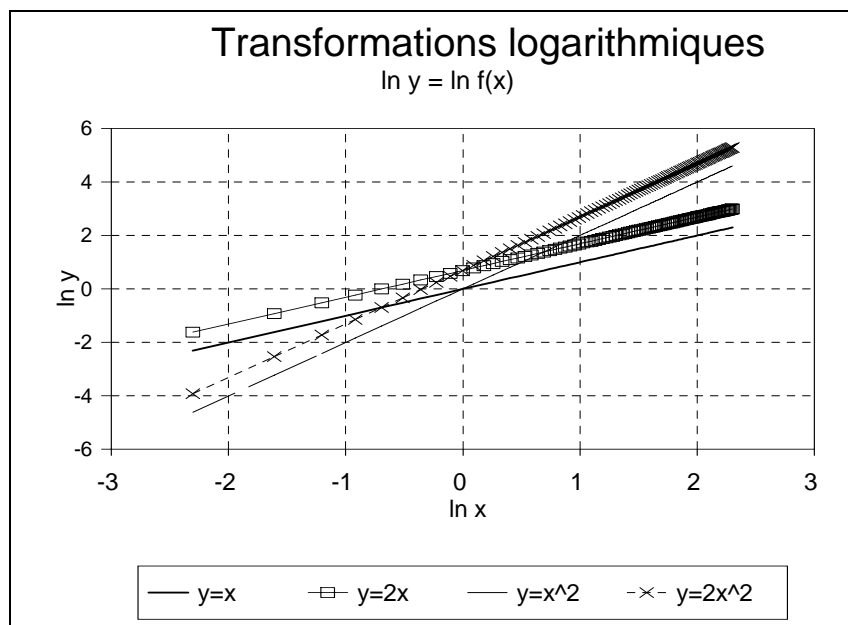


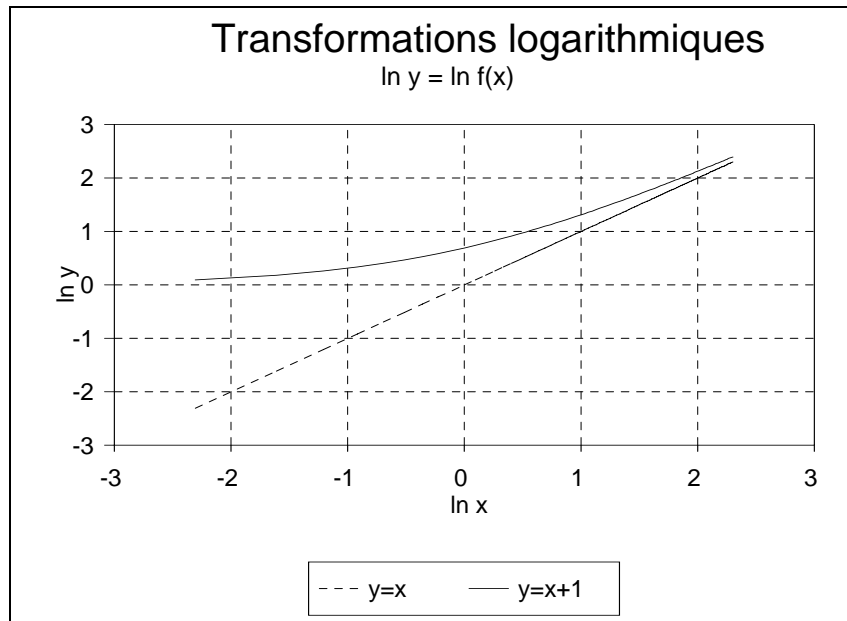
De cette dernière règle, on peut déduire celle du passage d'une base à l'autre. Supposons qu'on veuille passer du logarithme commun à base 10 au logarithme népérien à base  $e$ . Puisque  $x = \log_{10} y$  signifie  $10^x = y$ , on a

$$\log_e y = \log_e 10^x = x \log_e 10 = \log_{10} y \times \log_e 10$$

Les figures suivantes montrent comment les transformations logarithmiques changent la forme des relations entre variables. Les relations représentées sont :

- $y = x \quad \Rightarrow \quad \ln y = \ln x$
- $y = 2x \quad \Rightarrow \quad \ln y = \ln 2 + \ln x$
- $y = x^2 \quad \Rightarrow \quad \ln y = 2 \ln x$
- $y = 2x^2 \quad \Rightarrow \quad \ln y = \ln 2 + 2 \ln x$
- $y = x + 1 \quad \Rightarrow \quad \ln y = \ln(x + 1)$





On constate qu'après transformation logarithmique, des relations non linéaires deviennent linéaires et des relations linéaires deviennent non linéaires <sup>4</sup>.

### 2.3 LA FONCTION EXPONENTIELLE

La fonction

$$y = e^x$$

aussi dénotée  $y = \exp(x)$ , ou, plus rarement,  $y = \text{antilog}_e x$ , s'appelle la fonction exponentielle.

On emploie aussi l'exponentielle négative, de la forme

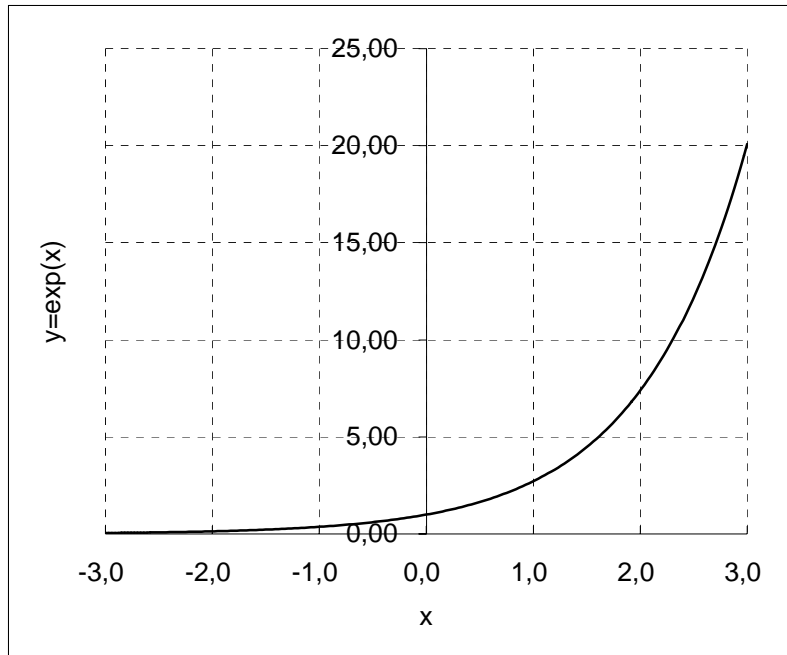
$$y = \exp(-x) = e^{-x} = \frac{1}{e^x}$$

Les figures suivantes illustrent la fonction exponentielle.

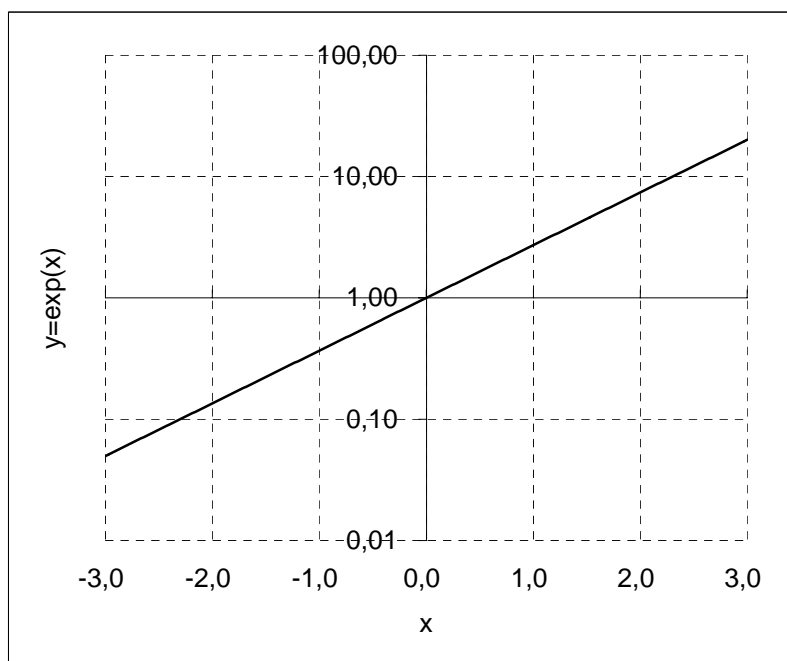
---

<sup>4</sup> Voir WONNACOTT, Thomas H. et WONNACOTT, Ronald J (1992) *Statistique : économie, gestion, sciences, médecine*, 4e éd., Economica, p 513-523, « La non-linéarité résolue grâce aux logarithmes ».

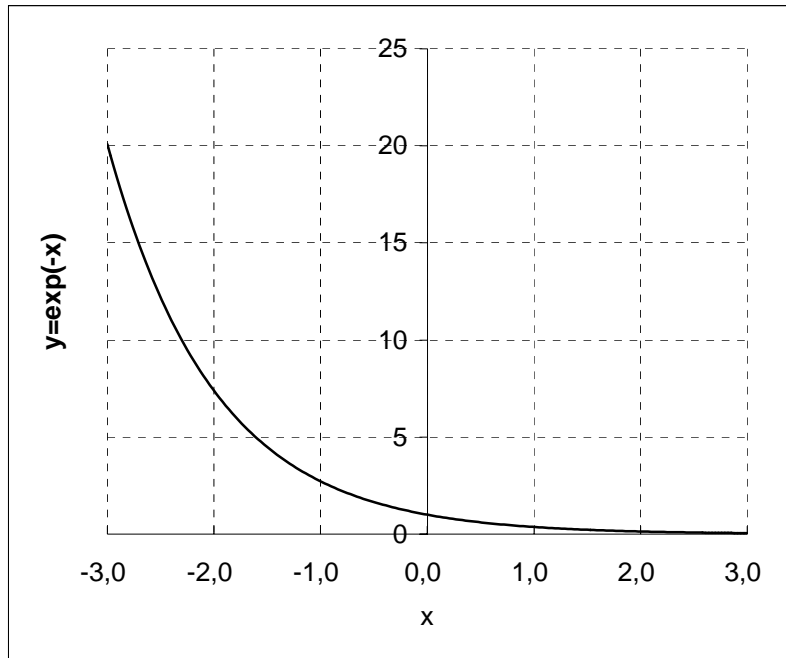
### Fonction exponentielle $y = \exp(x)$



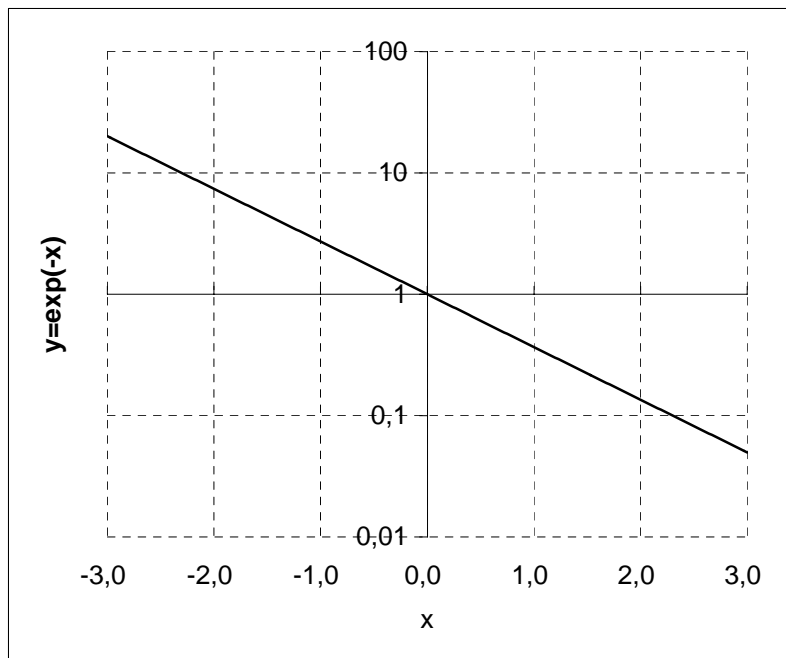
### Fonction exponentielle $y = \exp(x)$ Échelle verticale logarithmique



### Fonction exponentielle négative $y = \exp(-x)$



### Fonction exponentielle négative $y = \exp(-x)$ Échelle verticale logarithmique



## 2.4 POURQUOI LES LOGARITHMES NÉPÉRIENS ?

Le nombre  $e$  est défini comme la limite d'une suite infinie :

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$$

L'intérêt de cette constante népérienne vient de l'analyse de la croissance exponentielle. Si, à partir d'une valeur initiale  $q_0$ , une quantité  $q$  est multipliée à chaque période par un facteur  $(1+r)$ , au bout de  $t$  périodes, cette quantité sera de

$$q = q_0 (1+r)^t$$

C'est la formule de la croissance géométrique d'un montant auquel on applique un intérêt composé une fois par période. Supposons que l'on multiplie par  $n$  la fréquence à laquelle l'intérêt est composé. On a

$$q' = q_0 \left(1 + \frac{r}{n}\right)^{nt}$$

Qu'arrive-t-il lorsque  $n$  devient très grand (c'est-à-dire lorsque  $n$  tend vers l'infini et que l'intérêt est composé de façon continue) ? Pour le voir, récrivons l'équation précédente sous la forme suivante

$$q' = q_0 \left\{ \left[ 1 + \frac{1}{\left(\frac{n}{r}\right)} \right]^{\frac{n}{r}} \right\}^{rt}$$

Lorsque  $n$  tend vers l'infini,  $\frac{n}{r}$  tend aussi vers l'infini et l'expression entre accolades tend vers la constante  $e$ . On obtient donc

$$\lim_{n \rightarrow \infty} q' = q_0 e^{r t}$$

C'est la formule de la croissance exponentielle, qui est la version continue de la croissance géométrique.

## Solutions des exercices sur l'opérateur sommation

$$1.1 \quad \sum_{k=1}^4 x_k = (5 + 3) + (5 + 6) + (5 + 9) + (5 + 12) = 50$$

$$1.2 \quad \sum_{i=0}^3 x_i = (5 + 0) + (5 + 3) + (5 + 6) + (5 + 9) = 38$$

$$1.3 \quad \text{Pour } n = 4, \quad \sum_{i=0}^{n-1} x_{i+1} = \sum_{i=0}^3 x_{i+1} = \sum_{i=1}^4 x_i = 50$$

$$2.1 \quad \sum_{x=2}^3 x^3 = 2^3 + 3^3 = 8 + 27 = 35$$

$$2.2 \quad \sum_{i=1}^4 \frac{1}{i} = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \frac{25}{12}$$

$$2.3 \quad \text{Pour } a = 345, \quad \sum_{j=1}^{10} a = 10 a = 3450$$

$$3.1 \quad \sum_{i=1}^k x_i + \sum_{i=k+1}^n x_i = (x_1 + x_2 + \cdots + x_k) + (x_{k+1} + x_{k+2} + \cdots + x_n)$$

$$\sum_{i=1}^k x_i + \sum_{i=k+1}^n x_i = x_1 + x_2 + \cdots + x_k + x_{k+1} + x_{k+2} + \cdots + x_n = \sum_{i=1}^n x_i$$

$$3.2 \quad \sum_{i=1}^n (cx_i) = cx_1 + cx_2 + \cdots + cx_n = c(x_1 + x_2 + \cdots + x_n) = c \sum_{i=1}^n x_i$$

$$3.3 \quad \sum_{i=1}^t (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \cdots + (x_t + y_t)$$

$$\sum_{i=1}^t (x_i + y_i) = (x_1 + x_2 + \cdots + x_t) + (y_1 + y_2 + \cdots + y_t) = \sum_{i=1}^t x_i + \sum_{i=1}^t y_i$$

$$4.1 \quad \sum_i \sum_j a_{ij} = 4 + 50 + 5 + 30 + 6 + 10 = 105$$

$$4.2 \quad \sum_i a_{2i} = 5 + 30 = 35$$

$$4.3 \quad \sum_{j=1}^2 \sum_{i=1}^2 a_{ij} = (4 + 5) + (50 + 30) = 89$$

$$4.4 \quad \sum_{i=1}^2 \sum_{j=1}^2 a_{ij} = (4 + 50) + (5 + 30) = 89$$

$$4.5 \quad \sum_{j=1}^2 \sum_{i=1}^j a_{ij} = (4) + (50 + 30) = 84$$

$$4.6 \quad \sum_{i=1}^2 \sum_{j=1}^i a_{ij} = (4) + (5 + 30) = 39$$

$$4.7 \quad \sum_i \sum_{j>i} a_{ij} = 50$$

$$4.8 \quad \sum_i \sum_{j \leq i} a_{ij} = (4) + (5 + 30) + (6 + 10) = 55$$

## ANNEXE 1-B

# Principes et outils de gestion des données

1. Types de données
2. Métadonnées
3. Tâches générales de gestion de données
4. Étapes de la gestion de données
5. Aspects géographiques/spatiaux

---

## 1. Types de données

- Données primaires  
ex. : Enquête dans le cadre d'un projet
- Données secondaires publiées  
ex. : Données de Statistique Canada
- Données secondaires non publiées  
ex. : Données de fichiers administratifs  
(comme le rôle d'évaluation)



# 1. Types de données

- Recherche
  - Sources
  - Accès
- Gestion

# 2. Métadonnées

- « Au-delà » des données :  
Données sur les données
  - 2.1 Source
  - 2.2 Portée, univers de référence
  - 2.3 Concepts, définitions, variables
  - 2.4 Structure, organisation
  - 2.5 Méthode de collecte
  - 2.6 Évaluation de la qualité

## 2. Métadonnées

### 2.1 Source

- Quelle est l'agence, l'organisation qui a produit ces données ?
- Quelle est la crédibilité de l'émetteur ?
- Comment les données ont-elles été obtenues ?

---

## 2. Métadonnées

### 2.2 Portée, univers de référence

- Population couverte
  - « Population » s'entend dans le sens statistique  
= ensemble de personnes ou de choses...  
(ex. : « Population » d'entreprises, de villes...)
- Limites spatiales, géographiques
  - Tout le pays, une région, une municipalité...
- Période(s) de référence
  - À quelle ou à quelles années ou mois se rapportent les données ?

## 2. Métadonnées

### 2.3 Concepts, définitions, variables

- Quelles sont les unités d'observation ?  
(cela fait aussi partie de la définition de la population couverte)  
ex. : Familles ou individus ? « Familles », cela inclut-il les personnes vivant seules ?
- Quelle information (quelles variables) sur chaque unité d'observation ?

10

Méthodes quantitatives © André Lemelin, 2002

---

## 2. Métadonnées

### 2.3 Concepts, définitions, variables (suite)

- Quelle est la définition exacte de chaque variable ? De quel type de variable s'agit-il ?  
ex. : Le « Revenu » inclut-il les revenus autres que de travail ?
- Quelles sont les unités de mesure de chaque variable ?  
Exemples :
  - milliers ou millions ?
  - dollars canadiens ou dollars états-uniens ?
  - dollars courants ou dollars constants ? de quelle année ?

11

Méthodes quantitatives © André Lemelin, 2002

## 2. Métadonnées

### 2.3 Types de variables

- Variables *catégoriques* («nominal» en anglais)
  - à quelle catégorie appartient l'individu ?
  - Variable *dichotomique* : 2 catégories possibles
  - Variable *polytomique* : plus de 2 catégories
- Variables *ordinales*
  - classer les individus en ordre croissant ou décroissant
- Variables *d'intervalle*
  - comparer les différences entre individus
- Variables *rationnelles*
  - il y a un zéro naturel et le rapport entre deux valeurs a une signification

12

Méthodes quantitatives © André Lemelin, 2002

---

## 2. Métadonnées

### 2.4 Structure, organisation

- Format informatique (csv,dbf, excel...)
- Structure
  - Comment sont constitués les tableaux ?
  - Nous allons parler d'organisation plus loin

13

Méthodes quantitatives © André Lemelin, 2002

## 2. Métadonnées

### 2.5 Méthode de collecte

- Recensement ou échantillon ?
- Si échantillon, tiré comment ?
- Instrument(s) de collecte
  - questionnaire postal
  - questionnaire pour entrevue téléphonique
  - questionnaire pour entrevue en personne
  - entrevue semi-dirigée
  - etc.
- Conditions de la cueillette

14

Méthodes quantitatives © André Lemelin, 2002

---

## 2. Métadonnées

### 2.6 Évaluation de la qualité

- Couverture complète ?
- Données manquantes ?
- Données fiables ?

15

Méthodes quantitatives © André Lemelin, 2002

## 3. Tâches générales

- 3.1 Planifier
- 3.2 Exécuter
- 3.3 Mémoriser
- 3.4 Vérifier/valider
- 3.5 Documenter

---

## 3. Tâches générales

- 3.1 Planifier
  - Certaines décisions sont irréversibles
  - Planifier le « quoi ? » :  
flexibilité c. économie
    - Qu'est-ce qui pourrait être utile ?  
... parce qu'on ne peut pas tout prévoir
    - Qu'est-ce qui sera réellement utile ?  
... parce que l'exécution coûte du travail
  - Planifier le « comment ? »  
... parce que l'exécution mal planifiée  
coûte davantage de travail

## 3. Tâches générales

### 3.2 Exécuter

- Les difficultés lors de l'exécution peuvent amener à revoir la planification
- Revoir la planification, plutôt que d'improviser une solution !

---

## 3. Tâches générales

### 3.3 Mémoriser

- Que mémoriser ? Tout !
  - Les données elles-mêmes
  - Les métadonnées
  - La structure d'organisation
  - La liste des fichiers et leur localisation
  - Les traitements ou transformations

## 3. Tâches générales

### 3.3 Mémoriser

#### – Pourquoi mémoriser ?

##### Sécurité

contre la destruction, la suppression accidentelle...

##### Transparence

pour que d'autres puissent les utiliser

##### Flexibilité

pouvoir revenir en arrière et vérifier ce que l'on a fait  
pouvoir corriger les erreurs

21

Méthodes quantitatives © André Lemelin, 2002

---

## 3. Tâches générales

### 3.3 Mémoriser

#### – Comment mémoriser ?

##### Prendre des notes

Tenir un journal

##### Conserver des archives

Documents que l'on ne touche pas,  
généralement sous forme de fichiers électroniques

ex. : original, original nettoyé, versions  
successives, traitements, résultats...

Avec des noms différents

Avec un catalogue des fichiers et de leur description

22

Méthodes quantitatives © André Lemelin, 2002



## 3. Tâches générales

### 3.3 Mémoriser

- Méthodes de sécurité
  - Copies de sécurité
  - Disquettes verrouillées
  - Fichiers protégés contre l'écriture
  - Cellules protégées dans Excel

23

Méthodes quantitatives © André Lemelin, 2002

---

## 3. Tâches générales

### 3.4 Vérifier : procédures de validation

- Tests de cohérence
  - maximum, minimum, moyenne, fréquences
  - valeurs compatibles avec le type de variable
  - ...et avec son domaine de variation
- Totaux de contrôle
- Comparaison de calculs différents qui doivent logiquement donner le même résultat

**Valider, c'est pratiquer la méfiance  
systématique de l'Inspecteur Colombo !**

24

Méthodes quantitatives © André Lemelin, 2002

## 3. Tâches générales

### 3.5 Documenter

- Documentation active : produire l'information
- Documentation passive : générée par les procédures de traitement de données (programmes, fichiers « log », formules d'Excel...) ...grâce à
  - la systématisation
  - la standardisation
  - l'automatisation

25

Méthodes quantitatives © André Lemelin, 2002

---

## 4. Étapes de la gestion de données

4.1 À la réception des données...

4.2 Organisation

4.3 Partage

4.4 Exploitation

27

Méthodes quantitatives © André Lemelin, 2002

## 4. Étapes de la gestion de données

### 4.1 À la réception des données...

- Sauvegarder les données originales
- Prendre connaissance des métadonnées
- Valider une première fois
  - Sont-ce bien les données que l'on attendait ?
  - Sont-elles conformes aux métadonnées ?
- Documenter le processus de réception des données

28

Méthodes quantitatives © André Lemelin, 2002

---

## 4. Étapes de la gestion de données

### 4.2 Organisation



29

Méthodes quantitatives © André Lemelin, 2002

## 4. Étapes de la gestion de données

### 4.2 Organisation

- Le « data modelling » : un monde !  
Il existe plusieurs modèles d'organisation
- Réfléchir et planifier la structure d'organisation:  
le modèle d'organisation des données adéquat est celui qui reflète le modèle *ex ante* de la réalité que l'on veut étudier
- L'organisation inclut les métadonnées !

30

Méthodes quantitatives © André Lemelin, 2002

---

## 4. Étapes de la gestion de données

### 4.2 Organisation

- Structure fondamentale : un tableau
  - Lignes = enregistrements, observations, objets
  - Colonnes = champs, variables, attributs
- Une clé d'identification unique pour les objets
  - éviter la confusion entre les homonymes
  - éviter les noms multiples pour un même objet (données multilingues, variations d'orthographes, erreurs...)

31

Méthodes quantitatives © André Lemelin, 2002

## 4. Étapes de la gestion de données

### 4.2 Organisation

– Structure fondamentale : un tableau

		Variables				
		$X_1$	$X_2$	$X_3$	...	$X_k$
Observations	1	$x_{11}$	$x_{12}$	$x_{13}$	...	$x_{1k}$
	2	$x_{21}$	$x_{22}$	$x_{23}$	...	$x_{2k}$
	3	$x_{31}$	$x_{32}$	$x_{33}$	...	$x_{3k}$
	4	$x_{41}$	$x_{42}$	$x_{43}$	...	$x_{4k}$
	...	...	...	...	...	...
	$n$	$x_{n1}$	$x_{n2}$	$x_{n3}$	...	$x_{nk}$

32

## 4. Étapes de la gestion de données

### 4.2 Organisation : le cas des SIG

(Systèmes d'Information Géographique)

- Un SIG est une base de données...
- organisées en « couches » multiples d'information
- Représentation de l'espace :  
fond de carte
- Représentation « dans » l'espace :  
éléments géo-référencés  
(associés à des lieux identifiés)

33

## 4. Étapes de la gestion de données

### 4.2 Organisation : le cas des SIG (Systèmes d'Information Géographique)

- Intégration de données de sources diverses
- Possibilités de croisements, de liens

34

Méthodes quantitatives © André Lemelin, 2002

---

## 4. Étapes de la gestion de données

### 4.3 Partage

- Pourquoi le partage ?
  - La recherche est de moins en moins individuelle, de plus en plus en équipe
  - Même le chercheur solitaire doit rendre compte
    - Rapport de recherche, mémoire, article scientifique...
    - Une exigence de plus en plus courante : donner accès à ses données pour que d'autres puissent confirmer les résultats en les reproduisant, selon l'exigence méthodologique de la reproductibilité
  - Partager... avec soi-même !  
(quand on met de côté pour quelque temps, on oublie)

35

Méthodes quantitatives © André Lemelin, 2002

## 4. Étapes de la gestion de données

### 4.3 Partage

- Pour partager des données, il faut partager les métadonnées
- Établir des procédures pour...
  - modification des données
  - mémorisation
  - validation
  - documentation
  - etc.

36

## 4. Étapes de la gestion de données

### 4.4 Exploitation

Les analyses nécessitent une préparation adéquate des données

37

## Conclusion

- L'importance des métadonnées  
y compris la documentation
- L'importance de la mémoire  
Tout mémoriser !



## ANNEXE 1-C : NOTES SCHÉMATIQUES D'INITIATION À EXCEL

### Version avec commandes en anglais

#### 1. Ouverture

Faire démarrer le logiciel

Workbook et Worksheet

Cellules et adresses

Menus

1. souris
2. *Alt*+...
3. Raccourcis : *Ctrl*-...

#### 2. Manipulation des fichiers : menu *FICHIER (FILE)*

**Nouveau (New)**

**Ouvrir (Open)**

Ouvrir le fichier *ShShExem.xls*

**Fermer (Close)**

**Sauvegarder (Save)**

**se)**

**Sauvegarder Sous (Save As)**

Nommer les fichiers

Versions successives à conserver, copies de sécurité

**Print Area/Set Print Area**

Sélectionner avec la souris

1. *Shift* = sélection continue
2. *Ctrl* = sélection discontinue

#### 3. Se déplacer

1. Souris
2. Clavier :

1.	←	↑	→	↓
2.	<i>Home</i>	<i>Ctrl-Home</i>	<i>Ctrl-End</i> ou <i>End+Home</i>	
3.	<i>End+←</i>	<i>End+↑</i>	<i>End+→</i>	<i>End+↓</i>
4.	<i>PageUp</i>	<i>PageDown</i>		
5.	<i>Alt-PageUp</i>	<i>Alt-PageDown</i>		

- Pour voir la structure de la feuille : *View/Zoom/...*, puis *Ctrl-End* ou *End+Home*

#### 3. Information supplémentaire :

- Si l'«Assistant» est présent, faites-le disparaître :
  - Cliquez sur l'«Assistant»
  - Parmi les «Options», choisissez «Ne pas utiliser l'assistant» ou, selon la version d'Excel, fermez la petite fenêtre en cliquant sur le «x» en haut à droite
- Alors, vous pouvez accéder à...
  - *Help/ContentsAndIndex/Find/«move»/«move and scroll through a worksheet»*

- *Help/ContentsAndIndex/Find/«shortcut keys»/«move and scroll through a worksheet using shortcut keys»*
- 4. Un truc : les autoroutes (déplacements rapides à l'aide de *End*+flèche le long de corridors)
  - (voir le fichier *BdeDonn.xls*)

Faire une autoroute

## 4. Les formules

### 4.1 OPÉRATEURS DE BASE

1. Les 4 opérations : + - \* /

Sélectionner B7 : voir la fenêtre où s'affiche la formule

2. Exponentiation :  $3^2=9$

### 4.2 OPÉRATIONS DONT LES ARGUMENTS SONT DES CELLULES

Voir B16

Mises à jour automatiques : avantage-clé !

Arguments d'un autre feuillet ou d'un autre fichier

Voir '2 Branches'!B7 à 9

Écrire des formules à l'aide du clavier et de la souris

- Combinaison efficace : clavier pour les opérations, souris pour les adresses

### 4.3 FONCTIONS

Voir I7

Menu : *Insert/Function*

Menu : *Help/ContentsAndIndex/Find/nom\_de\_la\_fonction*

### 4.4 CORRECTIONS DANS LES CELLULES

- Sélectionner la cellule + F2
- Sélectionner la cellule, placer le curseur dans la fenêtre et écrire
- *Enter* pour confirmer; *Escape* pour maintenir le statu quo

## 5. Copier-coller des formules : le menu *EDIT*

### 5.1 PROCÉDURE STANDARD

#### **Copy**

Copier des valeurs ou insérer des formules ? Préserver les mises à jour automatiques !

#### **Paste**

### 5.2 ADRESSES ABSOLUES ET RELATIVES

Voir B27 à 29.

#### **Cut**

L'original disparaît !

... suivi de *Paste* : les adresses relatives demeurent inchangées (renvoient aux mêmes cellules qu'avant).

### **Paste Special**

*PasteSpecial/Value*

- Attention ! On perd le bénéfice de la mise à jour automatique.

*PasteSpecial/Transpose*

*PasteSpecial/Formats*

## **6. Le menu *FORMAT***

*Format/Cells*

- Pour fixer le nombre de décimales affichées : *Format/Cells/Number/Number (sic !)*
- Pour transformer des fractions en pourcentages : *Format/Cells/Number/Percentage*

Voir F16-I19

*Format/Column/Hide ou Unhide*

*Format/Style* : pour définir l'apparence par défaut

## **7. Protection contre les changements non désirés et réglages divers : le menu *TOOLS***

### **7.1 PROTECTION**

***Pour protéger ou déprotéger le feuillet contre l'écriture :***

*Tools/Protection*, puis choisir «*Protect sheet*» ou «*Unprotect sheet*»

*Format/Cells/Protection*, puis cocher ou annuler «*Locked*»

Lorsque le feuillet est protégé contre l'écriture, le menu *Format/Cells* n'est pas accessible. Donc, avant de changer la protection d'une ou de plusieurs cellules, il faut déprotéger la feuille.

***Pour protéger seulement certaines cellules :***

1. *Tools/Protection*, puis choisir «*Unprotect sheet*»
2. Sélectionner les cellules à protéger, et exécuter :  
*Format/Cells/Protection*, puis cocher «*Locked*»
3. Sélectionner les cellules à NE PAS protéger, et exécuter :  
*Format/Cells/Protection*, puis annuler «*Locked*»  
(Le statut par défaut – protégé ou non – est défini dans *Format/Style*)
4. *Tools/Protection*, puis choisir «*Protect sheet*»

### **7.2 RÉGLAGES DIVERS : *TOOLS/OPTIONS***

*Tools/Options/View* : cocher «*Gridlines*» ou non

*Tools/Options/Edit* : cocher ou non l'option «*Move selector after Enter*»

## **8. Les fenêtres : le menu *WINDOW***

***New***

***Arrange***

***Split***

***Freeze Panes (voir le fichier BdeDonn.xls)***

## 9. Présentation des données

### 9.1 TRI

1. Sélectionner l'ensemble à trier  
Un truc : créer d'abord une colonne avec des numéros séquentiels et l'inclure dans le tri ;  
on pourra ensuite reconstituer l'ordre initial
2. *Data/Sort* (si la première ligne contient des entêtes, cocher *Header Row*)

### 9.2 AU-DELÀ DU TRI : LA «LISTE» COMME BASE DE DONNÉES

Le tri crée automatiquement une «Liste» qui peut être gérée comme une base de données rudimentaire (voir le fichier *BdeDonn.xls*) :

Ensemble sélectionné pour tri = Liste = Base de données

Colonnes = champs (*fields*), c'est-à-dire variables

Entêtes de colonnes (incluses dans le tri) = noms de champs (*field names*)

Lignes = enregistrements (*records*), c'est-à-dire observations

Fonctions à explorer :

*Data/Filter/Autofilter*

*Data/PivotTable Report*

Voir :

*Help/Contents and Index/Find/«List»*

et consulter les rubriques suivantes

*About using a list as a database*

*Guidelines for creating a list on a worksheet*

*Database functions*

*About creating a PivotTable from a Microsoft Excel list or database*

### **Procédure pour créer un tableau de contingence à partir d'une liste**

Voir le fichier *EdefaQ92.xls*

1. Sélectionner la liste, y compris la colonne des numéros séquentiels et la ligne des entêtes
2. Créer la liste à l'aide de *Data/Sort*
3. *Data/PivotTableReport* et répondez aux questions :
  - Étape 1 : Source de données = liste
  - Étape 2 : L'ensemble de cellules se nomme «*Database*» par défaut
  - Utiliser une table existante comme base de la nouvelle ? Généralement, NON.
  - Étape 3 : Nouveau feuillet ou non
  - Étape 4 : cliquer sur le bouton *Layout*
    - Choisir (*Drag & Drop*) la variable de classification de la ligne
    - Choisir la variable de classification de la colonne
    - Choisir la variable de classification de la «page», si table à 3 dimensions  
(On peut aussi construire une table à plus de deux dimensions simplement en sélectionnant plus d'une variable de classification en ligne et/ou en colonne)
    - Choisir la variable de contenu de la table
    - Double cliquer sur la variable de contenu pour choisir la façon de traiter la variable, le format de représentation, etc.
  - Étape 5 : cliquer sur le bouton *Options*
    - « For empty cells, show... »

NOTES :

1. Pour créer un tableau de contingence, on peut choisir n'importe quelle variable de contenu (y compris l'une des variables de classification, pourvu que l'on spécifie que l'on veut faire un décompte (*Count*) **et pourvu que cette variable ne comporte aucun blanc**.
2. Si l'on veut utiliser le tableau de contingence pour faire d'autres calculs (par exemple, faire un test du Khi-2 au moyen de la fonction *Test.Khideux* ou *Chitest*), il faut s'assurer de choisir l'option qui inscrit des zéros dans les cellules où il n'y a pas d'observation, sans quoi les cellules vides ne seront pas prises en compte dans certains calculs).

### 9.3 GRAPHIQUES

#### *Insert/Chart*

Attention ! Step 4 of 4 :

- *As new sheet*
- *As object*

## 10 Échanges de données avec d'autres logiciels

### 10.1 COPIER DES TABLEAUX DE OU VERS UN TRAITEMENT DE TEXTE

1. Sélectionner
2. Copier
3. Se rendre à destination
4. Coller

Idem pour les graphiques !

### 10.2 TRANSFORMER DES DONNÉES EN FORMAT «TEXTE» EN DONNÉES NUMÉRIQUES

#### *Data/Text-to-columns*

Exemple :

1. *Le Quotidien* de Stat Can en format pdf (Acrobat) ou htm <sup>1</sup>
2. Sauvegarde en format texte
3. Traitement de texte
4. Changer de police au besoin pour aligner les colonnes : *Courrier New*
5. Ôter les séparateurs de milliers (sélection verticale Alt-souris)
6. Attention au séparateur décimal (. ou ,) : on peut faire le changement, s'il y a lieu, avant de passer à Excel, dans le panneau de configuration (*Regional settings*), **ou** on peut faire un remplacement global dans le traitement de texte, **ou** on peut faire le changement dans Excel (voir ci-après, étape 11)
7. Dans le fichier texte ou Word, sélectionner le tableau
8. Copier et coller dans Excel
9. Le tableau s'affiche comme une colonne d'entrées texte, qui débordent la cellule à droite (s'assurer que l'option «Wrap text» est désactivée pour ces cellules **et** que les cellules voisines à droite sont vides)
10. Avec la colonne sélectionnée, choisir *Courrier New* comme police de caractère (c'est une police à espacement constant)

---

<sup>1</sup> En format htm, la configuration du tableau est respectée; pas en pdf.

11. Si le séparateur décimal d'Excel est la virgule et que c'est le point qui est utilisé dans le tableau copié, sélectionner toute la colonne et remplacer tous les points par des virgules : menu *Edit/Replace* «.» par «,» et cliquer *All*.
12. Avec la colonne sélectionnée, aller dans le menu *Data/Text-to-Columns* et suivre les instructions; habituellement, il suffit de choisir «Fixed width», puis «Finish».

## ANNEXE 1-C : NOTES SCHÉMATIQUES D'INITIATION À EXCEL

### Version avec commandes en français

#### 1. Ouverture

Faire démarrer le logiciel

Classeur (Workbook) et Feuille (Worksheet)

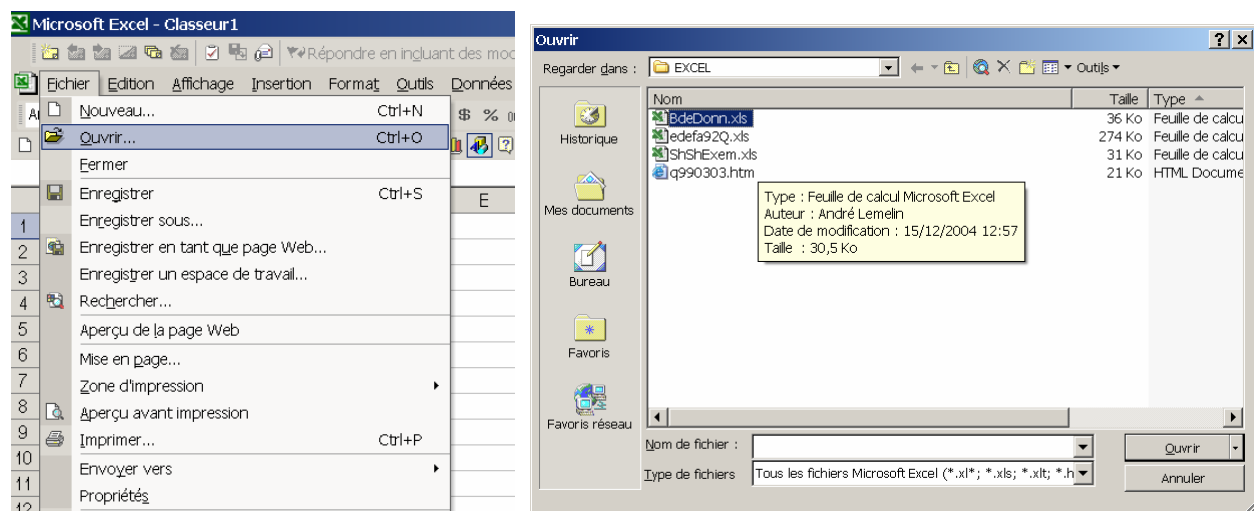
Cellules et adresses

Menus

1. souris
2. *Alt*+...
3. Raccourcis : *Ctrl*-...

#### 2. Manipulation des fichiers : menu *FICHIER* (FILE)

Ouvrir le fichier *ShShExem.xls*



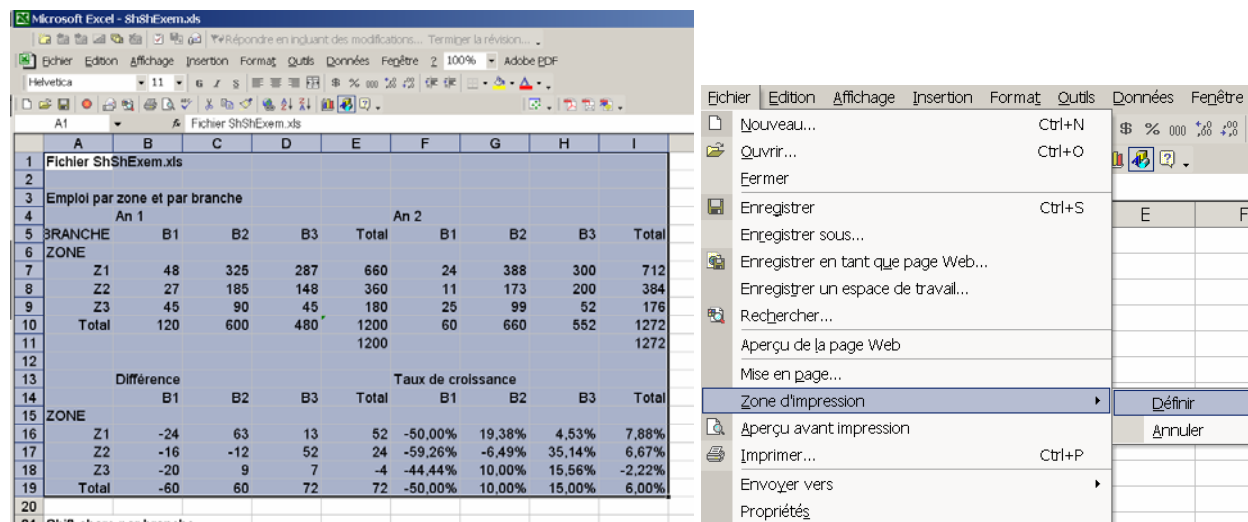
Nommer les fichiers

Versions successives à conserver, copies de sécurité



### Zone d'impression/Définir (Print Area/Set Print Area)

Dans Excel, vous pouvez définir une zone d'impression en deux étapes : 1) en sélectionnant un ensemble de cellules avec la souris; 2) en activant l'item *Zone d'impression / Définir*.



### 3. Se déplacer

1. Souris
2. Clavier :

1.	←	↑	→	↓
2.	Home	Ctrl-Home	Ctrl-End ou End+Home	
3.	End+←	End+↑	End+→	End+↓
4.	PageUp	PageDown		
5.	Alt-PageUp	Alt-PageDown		

- Pour voir la structure de la feuille : *Affichage/Zoom/...*, puis *Ctrl-End* ou *End+Home*

### 3. Information supplémentaire :

- Si l'«Assistant» (aussi appelé «Compagnon Office») est présent, faites-le disparaître :
  - Cliquez sur l'«Assistant»
  - Parmi les «Options», choisissez «Ne pas utiliser l'assistant» ou, selon la version d'Excel, fermez la petite fenêtre en cliquant sur le «x» en haut à droite
- Alors, vous pouvez accéder par le menu déroulant à *?/Aide sur Microsoft Excel/Index*, puis...
  - chercher le mot-clé «Déplacement» et choisir dans la liste la rubrique «Faire défiler les données dans une feuille de calcul» (cliquez sur *Afficher tout*)
  - chercher le mot-clé «Clavier» et choisir dans la liste la rubrique «Raccourcis clavier» et aller voir dans «Touches des classeurs et feuilles de calcul»

### 4. Un truc : les autoroutes (déplacements rapides à l'aide de *End+flèche* le long de corridors)

- (voir le fichier *BdeDonn.xls*)

Faire une autoroute

## 4. Les formules

### 4.1 OPÉRATEURS DE BASE

1. Les 4 opérations : + - \* /

Sélectionner B7 : voir la fenêtre où s'affiche la formule

	A	B	C
1	Fichier ShShExem.xls		
2			
3	Emploi par zone et par branche		
4		An 1	
5	BRANCHE	B1	B2
6	ZONE		
7	Z1	48	325
8	Z2	27	185

2. Exponentiation :  $3^2=9$

### 4.2 OPÉRATIONS DONT LES ARGUMENTS SONT DES CELLULES

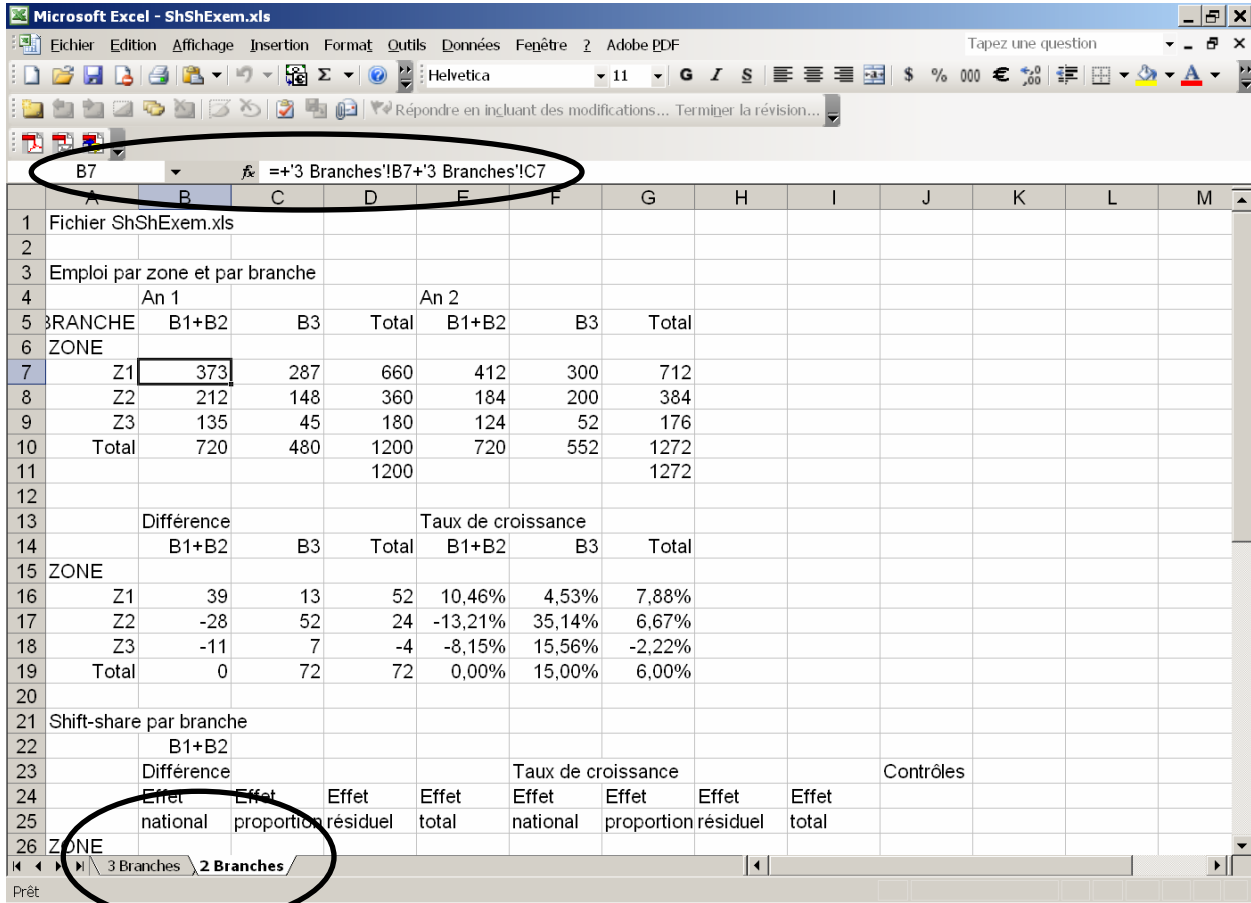
Voir B16

Mises à jour automatiques : avantage-clé !

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Fichier ShShExem.xls												
2													
3	Emploi par zone et par branche												
4		An 1				An 2							
5	BRANCHE	B1	B2	B3	Total	B1	B2	B3	Total				
6	ZONE												
7	Z1	48	325	287	660	24	388	300	712				
8	Z2	27	185	148	360	11	173	200	384				
9	Z3	45	90	45	180	25	99	52	176				
10	Total	120	600	480	1200	60	660	552	1272				
11					1200				1272				
12													
13		Différence				Taux de croissance							
14		B1	B2	B3	Total	B1	B2	B3	Total				
15	ZONE												
16	Z1	-24	63	13	52	-50,00%	19,38%	4,53%	7,88%				
17	Z2	-16	-12	52	24	-59,26%	-6,49%	35,14%	6,67%				
18	Z3	-20	9	7	-4	-44,44%	10,00%	15,56%	-2,22%				

Arguments d'un autre feuillet ou d'un autre fichier

Voir '2 Branches'!B7 à 9



Écrire des formules à l'aide du clavier et de la souris

- Combinaison efficace : clavier pour les opérations, souris pour les adresses

### 4.3 FONCTIONS

Voir I7

Microsoft Excel - ShShExem.xls

Repondre en incluant des modifications... Terminer la révision...

Echier Edition Affichage Insertion Format Outils Données Fenêtre ? 100% Adobe PDF

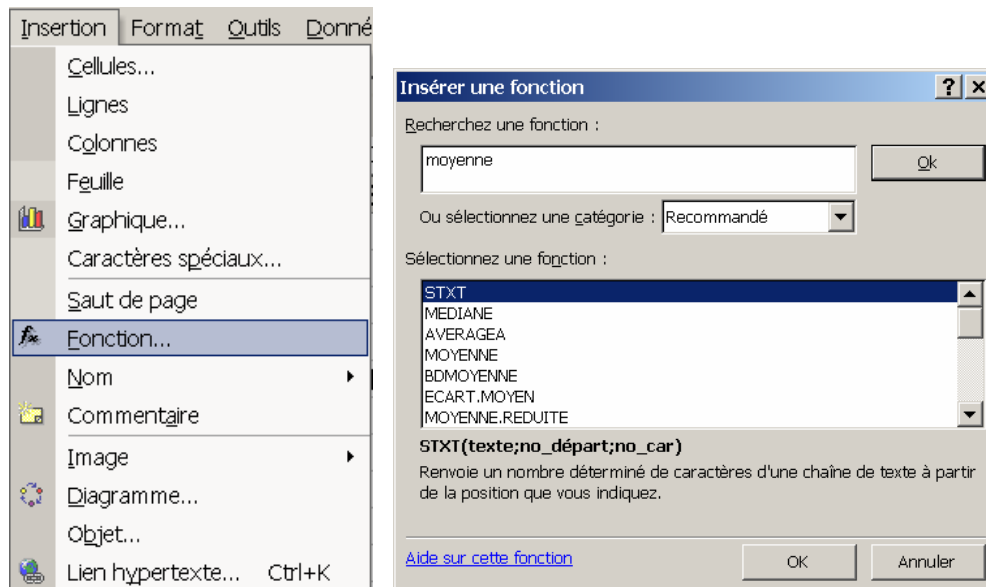
Helvetica 11

=SOMME(F7:H7)

	A	B	C	D	E	F	G	H	I	
1	Fichier ShShExem.xls									
2										
3	Emploi par zone et par branche									
4		An 1			An 2					
5	BRANCHE	B1	B2	B3	Total	B1	B2	B3	Total	
6	ZONE									
7	Z1	48	325	287	660	24	388	300	712	
8	Z2	27	185	148	360	11	173	200	384	
9	Z3	45	90	45	180	25	99	52	176	
10	Total	120	600	480	1200	60	660	552	1272	
11					1200				1272	

Menu : Insertion/Fonction

Menu : ?/Aide sur Microsoft Excel/Index : chercher le nom de la fonction comme mot-clé  
Pour trouver une fonction, tapez un mot-clé dans la zone « rechercher une fonction » de la boîte de dialogue. Par exemple, tapez le mot *moyenne*



### 4.4 CORRECTIONS DANS LES CELLULES

- Sélectionner la cellule + F2
- Sélectionner la cellule, placer le curseur dans la fenêtre et écrire
- *Enter* pour confirmer; *Escape* pour maintenir le statu quo
- Oups ! Pour revenir en arrière après confirmation : menu *Édition* / *Annuler* ou CTRL-z

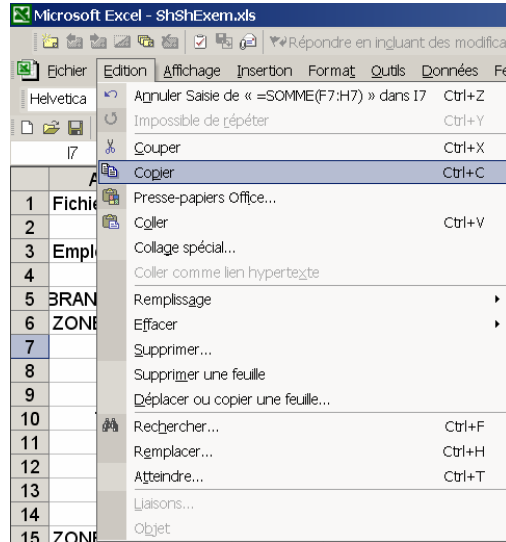
## 5. Copier-coller des formules : le menu *ÉDITION*

### 5.1 PROCÉDURE STANDARD

#### **Copier**

Copier des valeurs ou insérer des formules ? Préserver les mises à jour automatiques !

#### **Coller**



### 5.2 ADRESSES ABSOLUES ET RELATIVES

Voir B27 à 29.

	A	B	C
26	ZONE		
27	Z1	2,88	-26,88
28	Z2	1,62	-15,12
29	Z3	2,7	-25,2

#### **Couper**

L'original disparaît !

... suivi de *Coller* : les adresses relatives demeurent inchangées (renvoient aux mêmes cellules qu'avant).

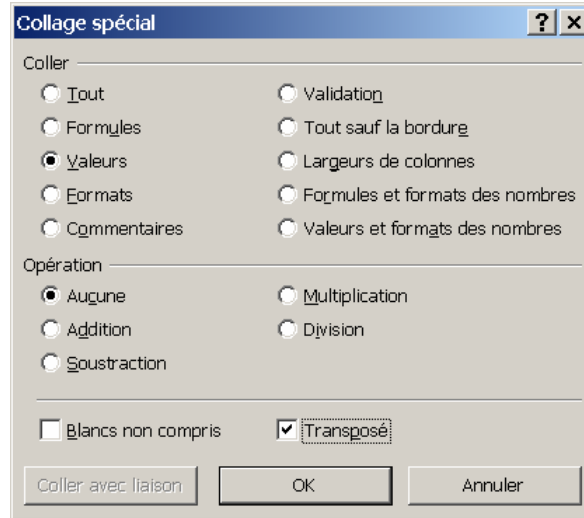
#### **Collage spécial**

*Collage spécial/Valeurs*

- Attention ! On perd le bénéfice de la mise à jour automatique.

*Collage spécial/Transposé*

*Collage spécial/Formats*

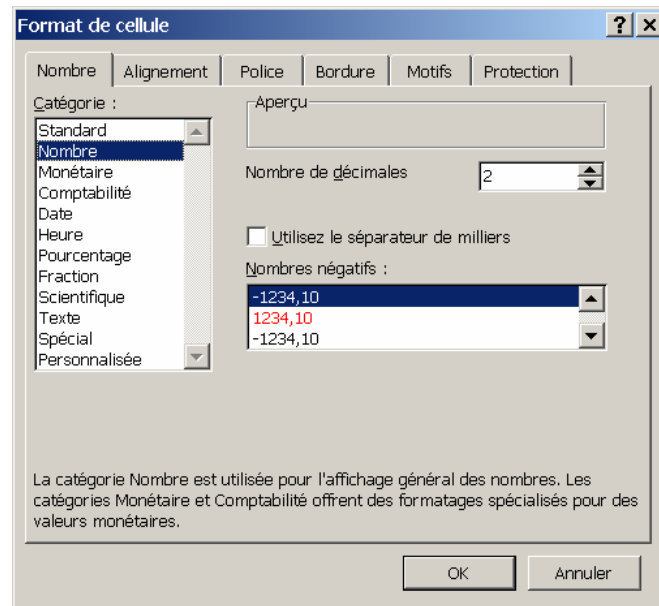
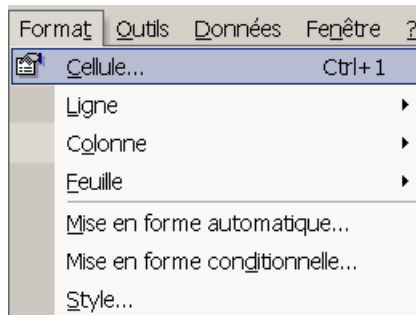


## 6. Le menu *FORMAT*

### *Format/Cellule*

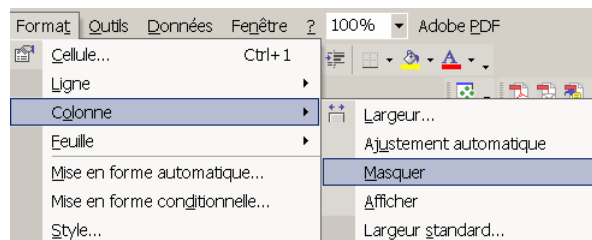
- Pour fixer le nombre de décimales affichées : *Format/Cellule/Nombre/Nombre* [sic !]
- Pour transformer des fractions en pourcentages : *Format/Cellule/Nombre/Pourcentage*

Voir F16-I19



## Format/Colonne/Masquer ou Afficher

Format/Style : pour définir l'apparence par défaut



## 7. Protection contre les changements non désirés et réglages divers : le menu OUTILS

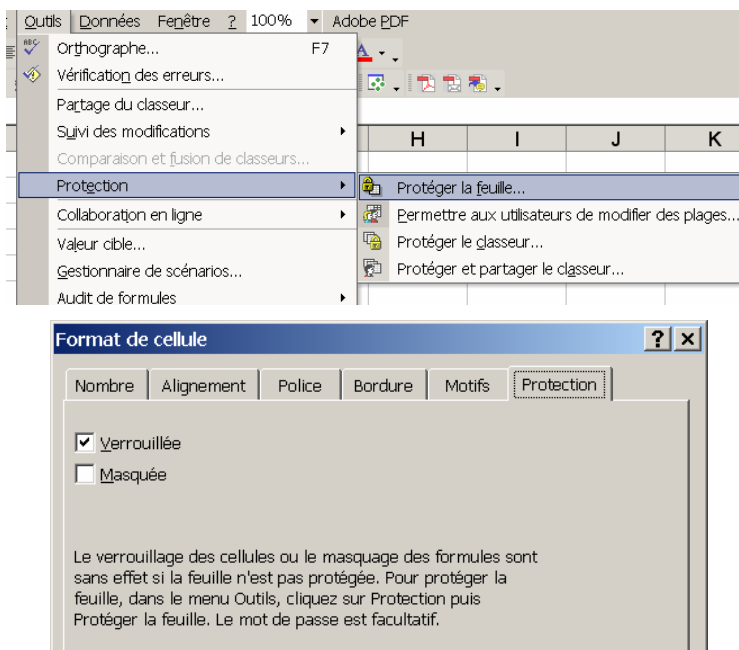
### 7.1 PROTECTION

**Pour protéger ou déprotéger le feuillet contre l'écriture :**

*Outils/Protection*, puis choisir «*Protéger la feuille*» ou «*Ôter la protection de la feuille*»

*Format/Cellule/Protection*, puis cocher ou annuler «*Verrouillée*»

Lorsque le feuillet est protégé contre l'écriture, le menu *Format/Cellule* n'est pas accessible. Donc, avant de changer la protection d'une ou de plusieurs cellules, il faut déprotéger la feuille.



**Pour protéger seulement certaines cellules :**

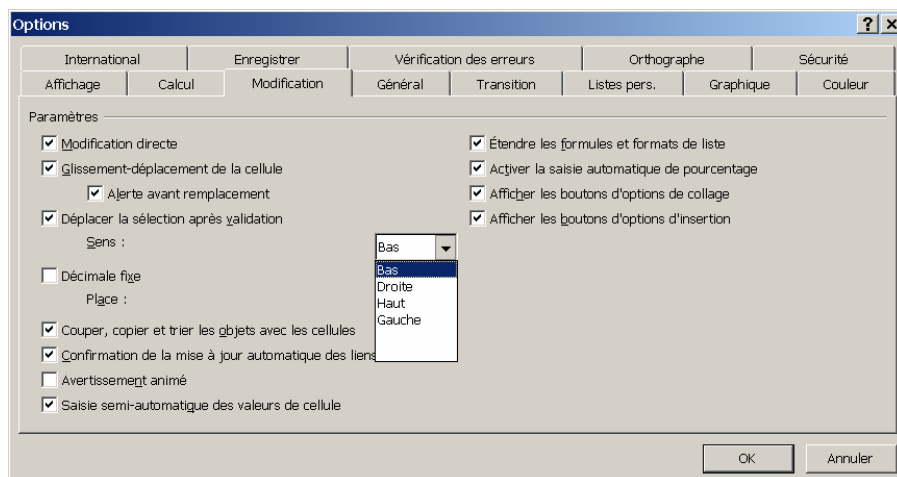
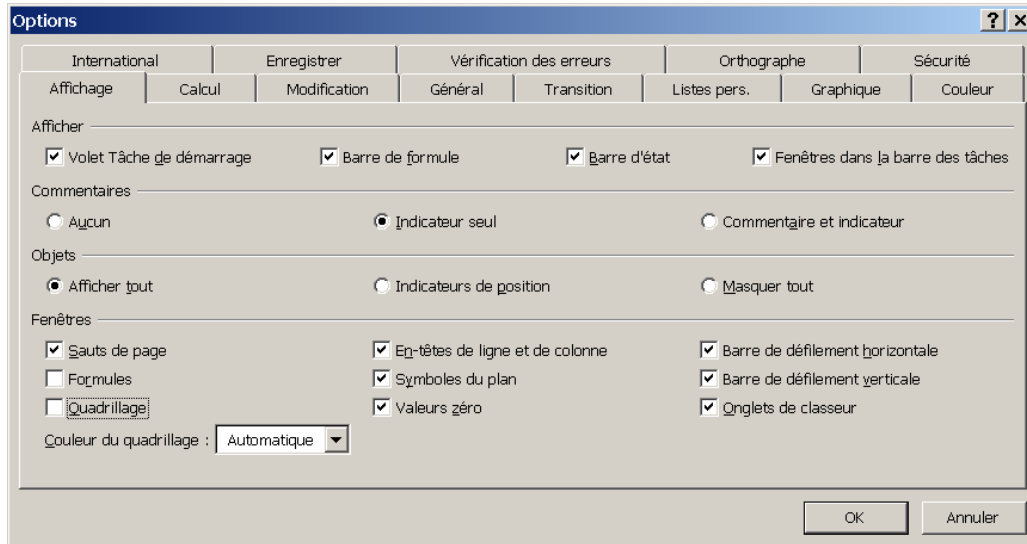
1. *Outils/Protection*, puis choisir «*Ôter la protection de la feuille*»
2. Sélectionner les cellules à protéger, et exécuter :

- Format/Cellule/Protection*, puis cocher «*Verrouillée*»
- Sélectionner les cellules à NE PAS protéger, et exécuter :  
*Format/Cellule/Protection*, puis annuler «*Verrouillée*»  
(Le statut par défaut – protégé ou non – est défini dans *Format/Style*)
  - Outils/Protection*, puis choisir «*Protéger la feuille*»

## 7.2 RÉGLAGES DIVERS : *OUTILS/OPTIONS*

*Outils/Options/Affichage* : cocher ou non «*Quadrillage*»

*Outils/Options/Modification* : cocher ou non l'option «*Déplacer la sélection après validation*»



## 8. Les fenêtres : le menu *FENÊTRE*

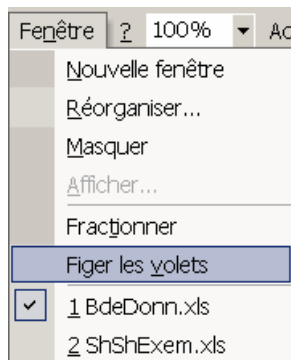
***Nouvelle fenêtre***

***Réorganiser***

***Fractionner***

***Figurer les volets (voir le fichier *BdeDonn.xls*)***





## 9. Présentation des données

### 9.1 TRI

#### 1. Sélectionner l'ensemble à trier

Un truc : créer d'abord une colonne avec des numéros séquentiels et l'inclure dans le tri ; on pourra ensuite reconstituer l'ordre initial

#### 2. Données/Trier (si la première ligne contient des entêtes, cocher *Ligne de titres*)

### 9.2 AU-DELÀ DU TRI : LA «LISTE» COMME BASE DE DONNÉES

Le tri crée automatiquement une «Liste» qui peut être gérée comme une base de données rudimentaire (voir le fichier *BdeDonn.xls*) :

Ensemble sélectionné pour tri = Liste = Base de données

Colonnes = champs (*fields*), c'est-à-dire variables

Entêtes de colonnes (incluses dans le tri) = noms de champs (*field names*)

Lignes = enregistrements (*records*), c'est-à-dire observations

No	Ville	Population	Primaire	Manufac.	Construc.	Tr-Com-SP	Commerce	ServProd	ServAutr
1	Toronto	2 998 947	14 335	390 540	89 350	130 655	297 955	254 570	456 365
2	Montreal	2 828 347	8 760	225 555	50 025	123 875	239 505	157 910	429 795
3	Vancouver	1 268 187					9 635	126 490	89 365
4	Ottawa_Hull	717 977					5 885	52 020	43 130
5	Edmonton	657 057					4 125	66 615	43 770
6	Calgary	592 747					8 835	57 835	55 090
7	Winnipeg	584 847					5 915	57 485	30 895
8	Quebec	576 077					6 810	42 940	26 355
9	Hamilton	542 097					4 360	45 365	23 140
10	St_Catherines_Niagara	304 357					8 725	22 750	10 720
11	Kitchener	287 807					6 340	24 885	13 525
12	Halifax	277 727					2 305	24 645	15 265
13	London	254 287					9 010	26 845	16 755
14	Windsor	246 117					6 590	18 320	9 010
15	Victoria	233 487					6 950	17 790	10 935
16	Regina	164 317					8 805	16 970	9 910
17	St_John's	154 827					6 320	13 215	6 380
18	Oshawa	154 217	1 070	26 555	3 600	5 440	12 290	6 470	20 990
19	Saskatoon	154 210	3 750	7 445	6 285	6 815	16 420	7 765	31 170
20	Sudbury	149 923	12 035	7 840	4 120	4 450	11 520	4 295	22 520
21	Chicoutimi_Jonquiere	135 172	1 455	12 025	3 570	3 055	9 190	3 770	19 340
22	Thunder_Bay	121 379	2 310	10 305	3 970	8 185	10 620	3 780	21 685
23	Saint_John	114 048	525	7 830	4 175	6 635	9 540	4 675	16 930
24	Trois_Rivieres	111 453	505	11 855	2 385	3 645	7 680	3 210	17 560

Fonctions à explorer :

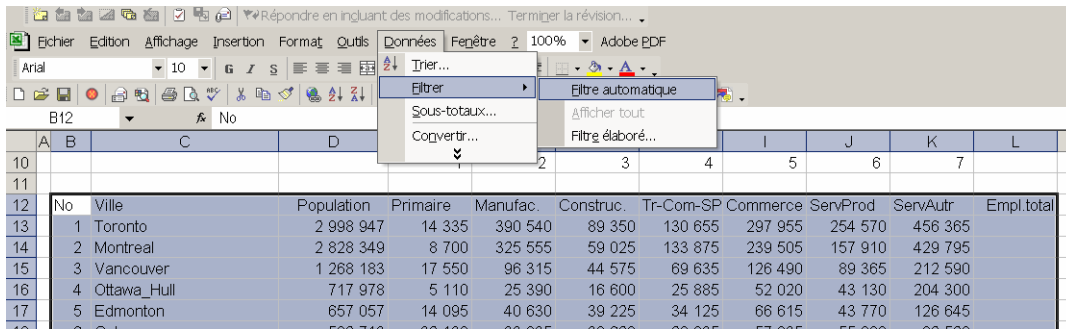
*Données/Filtrer/Filtre automatique*

*Données/Rapport de tableau croisé dynamique...*

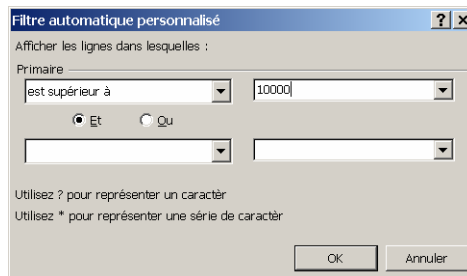
Voir :

*?/Aide sur Microsoft Excel/Index, chercher le mot-clé «Liste» et consulter les rubriques suivantes :*

- *À propos des données sources de tableaux et graphiques croisés dynamiques : voir Listes ou bases de données Microsoft Excel*
- *Instructions pour créer une liste dans une feuille de calcul*
- *Fonctions de feuille de calcul répertoriées par catégorie : voir base de données*
- *Fonctions de synthèse pour l'analyse de données*



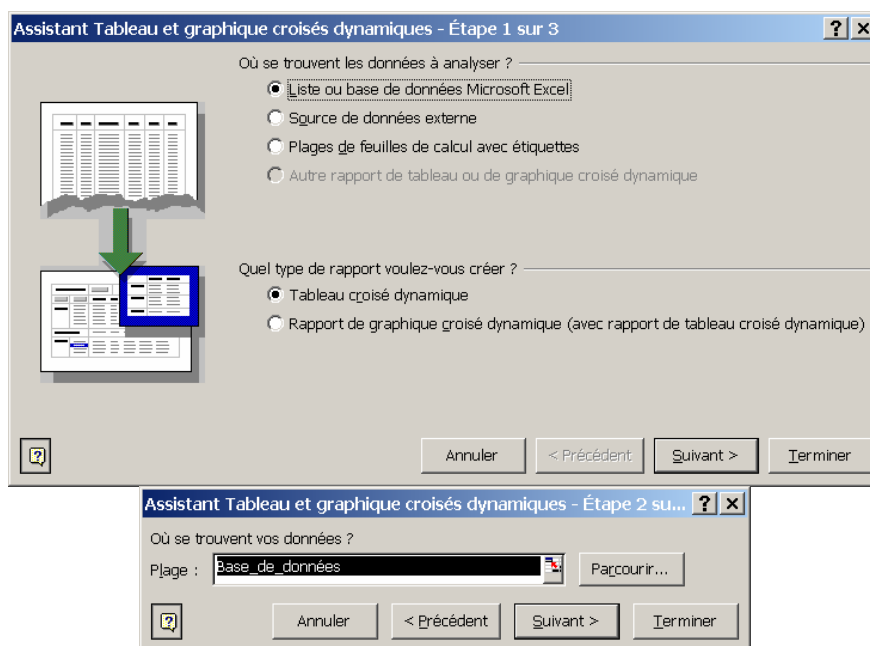
No	Ville	Population	Primaire	Manufac	Construc	Tr-Com-SP	Commerci	ServProd	ServAutr	Empl.total
1	Toronto	2 998 947	14 335	390 540	89 350	130 655	297 955	254 570	456 365	
2	Montreal	2 828 349	8 700	325 555	59 025	133 875	239 505	157 910	429 795	
3	Vancouver	1 268 183	17 550	96 315	44 575	69 635	126 490	89 365	212 590	
4	Ottawa_Hull	717 978	5 110	25 390	16 600	25 885	52 020	43 130	204 300	
5	Edmonton	657 057	14 095	40 630	39 225	34 125	66 615	43 770	126 645	
6	Calgary	592 743	1 725	33 035	39 620	28 835	57 835	55 090	92 520	
7	Winnipeg	584 841	1 835	51 460	14 695	35 915	57 485	30 895	108 235	
8	Quebec	576 074	2 035	28 310	13 670	16 810	42 940	26 355	134 355	
9	Hamilton	542 094	2 310	87 285	16 025	14 360	45 365	23 140	78 395	
10	St_Catherines_Niagara	304 354	2 610	44 505	7 955	8 725	22 750	10 720	44 940	
11	Kitchener	287 801	2 925	55 015	7 945	6 340	24 885	13 525	39 760	
12	Halifax	277 721	3 510	10 875	7 250	12 305	24 645	15 265	67 515	
13	London	254 286	3 750	30 470	7 945	9 010	26 845	16 755	53 630	
14	Windsor	246 116	4 575	34 125	6 085	6 590	18 320	9 010	35 130	
15	Victoria	233 481	5 110	8 110	8 475	6 950	17 790	10 935	57 260	

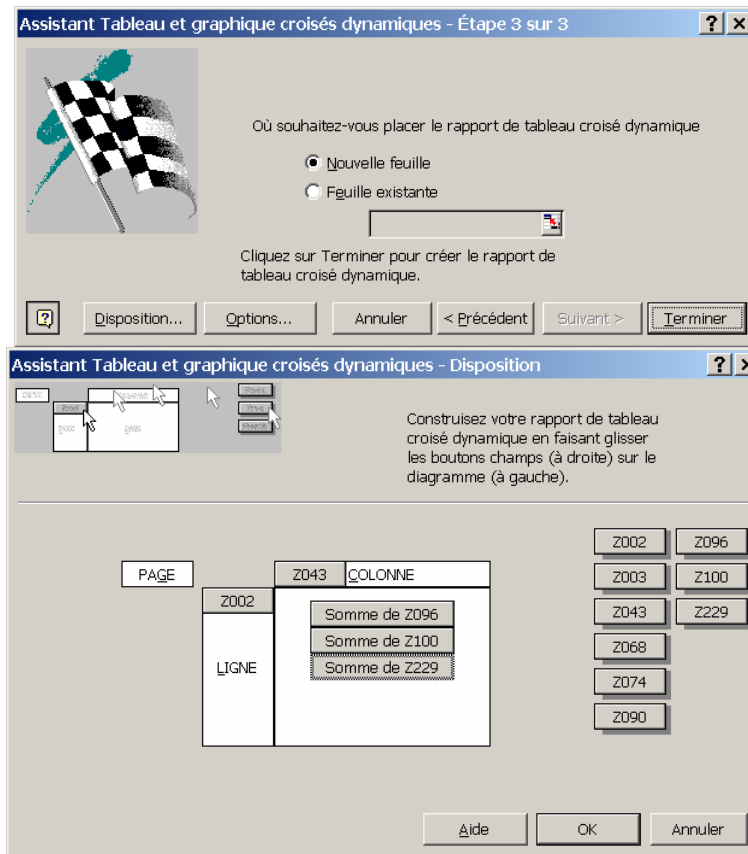


### Procédure pour créer un tableau de contingence à partir d'une liste

Voir le fichier *EdefaQ92.xls*

1. Sélectionner la liste, y compris la colonne des numéros séquentiels et la ligne des entêtes
2. Créer la liste à l'aide de *Données/Trier*
3. *Données/Rapport de tableau croisé dynamique* et répondez aux questions :
  - Étape 1 : Source de données = liste
  - Étape 2 : L'ensemble de cellules se nomme «*Base\_de\_données*» par défaut
  - Utiliser une table existante comme base de la nouvelle ? Généralement, NON.
  - Étape 3 : Nouveau feuillet ou non
  - Étape 4 : cliquer sur le bouton *Disposition*
    - Choisir (*Drag & Drop*) la variable de classification de la ligne
    - Choisir la variable de classification de la colonne
    - Choisir la variable de classification de la «page», si table à 3 dimensions  
(On peut aussi construire une table à plus de deux dimensions simplement en sélectionnant plus d'une variable de classification en ligne et/ou en colonne)
    - Choisir la variable de contenu de la table
    - Double cliquer sur la variable de contenu pour choisir la façon de traiter la variable, le format de représentation, etc.
  - Étape 5 : cliquer sur le bouton *Options*
    - « Cellules vides, afficher... »





#### NOTES :

1. Pour créer un tableau de contingence, on peut choisir n'importe quelle variable de contenu (y compris l'une des variables de classification, pourvu que l'on spécifie que l'on veut faire un décompte (*Nombre*) et **pourvu que cette variable ne comporte aucun blanc**.
2. Si l'on veut utiliser le tableau de contingence pour faire d'autres calculs (par exemple, faire un test du Khi-2 au moyen de la fonction *Test.Khideux* ou *Chitest*), il faut s'assurer de choisir l'option qui inscrit des zéros dans les cellules où il n'y a pas d'observation, sans quoi les cellules vides ne seront pas prises en compte dans certains calculs).

### 9.3 GRAPHIQUES

#### Insertion/Graphique

Attention ! à la 4e étape de 4 :

- Sur une nouvelle feuille
- En tant qu'objet dans...

## 10 Échanges de données avec d'autres logiciels

### 10.1 COPIER DES TABLEAUX DE OU VERS UN TRAITEMENT DE TEXTE

1. Sélectionner
2. Copier
3. Se rendre à destination

#### 4. Coller

Idem pour les graphiques !

### 10.2 TRANSFORMER DES DONNÉES EN FORMAT «TEXTE» EN DONNÉES NUMÉRIQUES

#### *Données/Convertir*

Exemple :

1. *Le Quotidien* de Stat Can en format pdf (Acrobat) ou htm <sup>1</sup>
2. Sauvegarde en format texte
3. Traitement de texte
4. Changer de police au besoin pour aligner les colonnes : `Courrier New`
5. Ôter les séparateurs de milliers (sélection verticale Alt-souris)
6. Attention au séparateur décimal (. ou ,) : on peut faire le changement, s'il y a lieu, avant de passer à Excel, dans le panneau de configuration (*Options régionales*), **ou** on peut faire un remplacement global dans le traitement de texte, **ou** on peut faire le changement dans Excel (voir ci-après, étape 11)
7. Dans le fichier texte ou Word, sélectionner le tableau
8. Copier et coller dans Excel
9. Le tableau s'affiche comme une colonne d'entrées texte, qui débordent la cellule à droite (s'assurer que l'option «Renvoyer à la ligne automatiquement» est désactivée pour ces cellules **et** que les cellules voisines à droite sont vides)
10. Avec la colonne sélectionnée, choisir `Courrier New` comme police de caractère (c'est une police à espacement constant)
11. Si le séparateur décimal d'Excel est la virgule et que c'est le point qui est utilisé dans le tableau copié, sélectionner toute la colonne et remplacer tous les points par des virgules : menu *Édition/Remplacer* «.» par «,» et cliquer *Remplacer tout*.
12. Avec la colonne sélectionnée, aller dans le menu *Données/Convertir* et suivre les instructions; habituellement, il suffit de choisir «Largeur fixe», puis «Terminer».

---

<sup>1</sup> En format htm, la configuration du tableau est respectée; pas en pdf.

## ANNEXE 1-D : NOTES SCHÉMATIQUES D'INITIATION À SPSS

### 1. Importation d'un fichier de données Excel

#### FICHER DE DONNÉES EXCEL

X\_DonEnqH05.xls

Ce fichier contient certaines données de l'enquête réalisée dans le cadre du cours EUR-8112 à Place Versailles, à l'automne 2000. Les données ont été modifiées de façon aléatoire pour respecter les exigences de confidentialité et pour illustrer la présence de valeurs aberrantes.

L'enquête a été réalisée du 19 au 23 octobre 2000. Elle a porté uniquement sur les personnes de 18 ans et plus et de moins de 45 ans, à l'exclusion des personnes qui travaillaient à Place Versailles. Les répondants ont été recrutés sur place par les enquêteurs, qui ont interrogé les sujets et rempli le questionnaire. Les blancs correspondent à des données manquantes.

Les variables du fichier sont les suivantes :

**NoSeq** : Numéro séquentiel du questionnaire.

**Age** : Dans quelle catégorie d'âge vous situez-vous ?

1	moins de 18 ans
2	18-24 ans
3	25-29 ans
4	30-34 ans
5	35-39 ans
6	40-44 ans
7	45 ans et plus

**Sexe** :

1	Un homme
2	Une femme

**Distance** : Distance parcourue, en kilomètres, pour venir à Place Versailles.

**Date** : Date à laquelle l'entrevue a été réalisée.

Valeur de Date	jour de la semaine	jour du mois	mois	année
19	jeudi	19	octobre	2000
20	vendredi	20	octobre	2000
21	samedi	21	octobre	2000
22	dimanche	22	octobre	2000
23	lundi	23	octobre	2000

**Motif** : Variable construite à partir des réponses à la question « Pour quel motif, êtes-vous venu à Place Versailles ? Vous pouvez identifier plus d'un motif ».

0	s'il n'y a pas de réponse
1	si l'unique motif est « Pour faire des achats »
2	s'il y a plus d'un motif ET que l'un d'eux est « Pour faire des achats »
3	si l'unique motif est « Pour voir la marchandise... »
4	s'il y a plus d'un motif ET que l'un d'eux est « Pour voir la marchandise... » ET que l'un d'eux n'est PAS « Pour faire des achats »
5	dans tous les autres cas

**Mode** : Variable construite à partir des réponses à la question « Aujourd'hui, quel(s) mode(s) de transport avez-vous utilisé pour venir à Place Versailles ? Vous pouvez nommer plus d'un moyen de transport ».

0	s'il n'y a pas de réponse
1	si l'unique mode est « Auto comme conducteur »
2	si l'unique mode est « Auto comme passager »
3	si l'unique mode est l'autobus
4	si l'unique mode est le métro
5	si l'on a utilisé l'autobus ET le métro
6	dans tous les autres cas

Voici comment se présente le fichier Excel :

The screenshot shows a Microsoft Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K	L
1	NoSeq	Age	Sexe	Distan	Date	Motif	Mode					
2	1	2	1	32,25	20	1	2					
3	2	3	2	9,25	20	1	4					
4	3	1	2	5,00	21	2	2					
5	4	5	2	23,00	20	5	5					
6	5	5	2	9,50	20	1	5					
7	6	3	2	9,75	20	5	1					

N.B.

1. Entêtes de colonnes
2. blancs du fichier Excel = données manquantes (ex. : NoSeq=15)

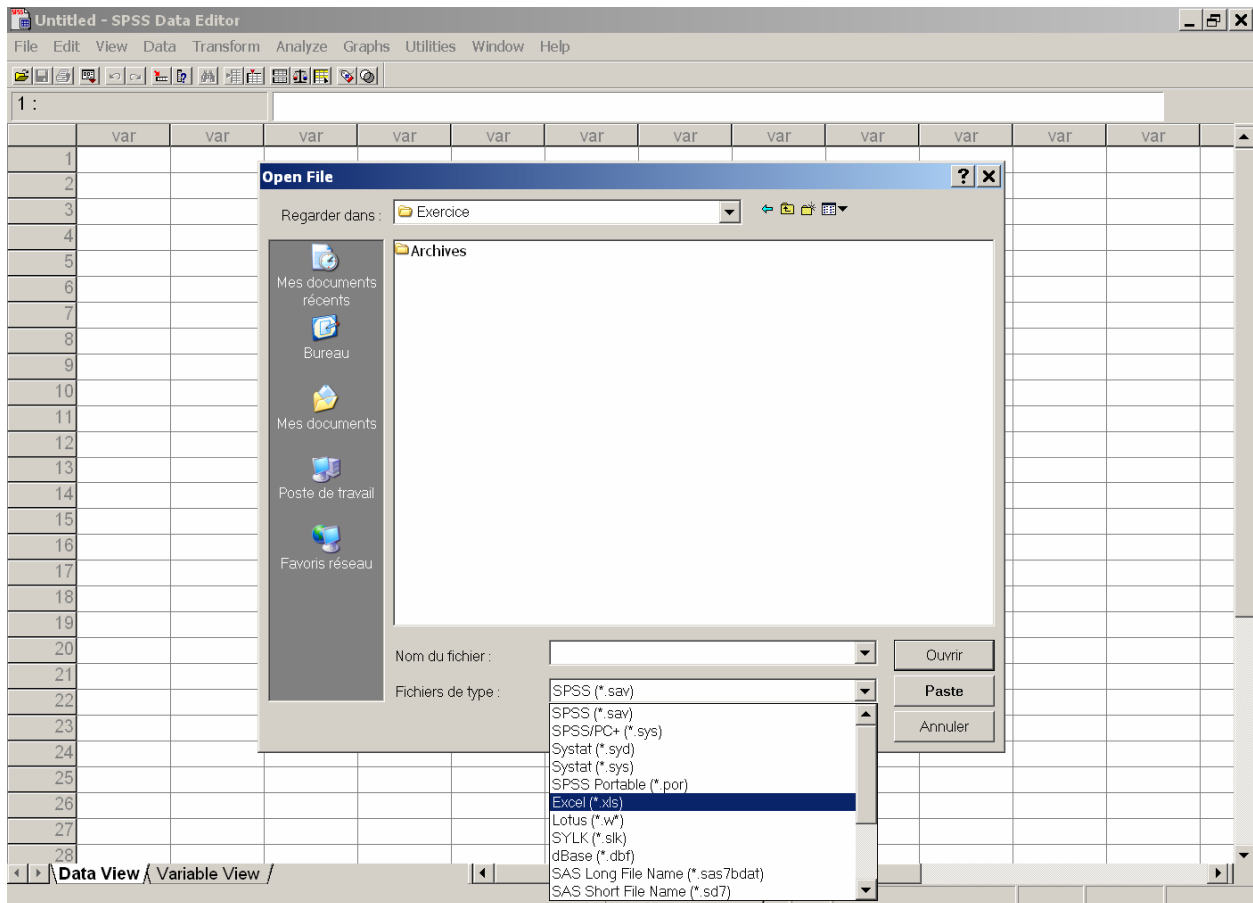
Fermer le fichier Excel

## OUVRIR SPSS ET IMPORTER LES DONNÉES DU FICHIER EXCEL

- File/Open/Data

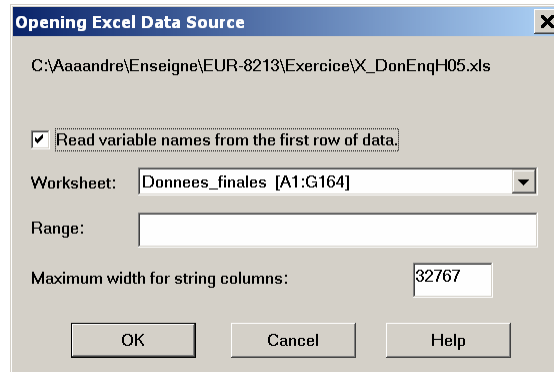
- trouver le répertoire
- type = \*.xls

N.B. pour être ouvert dans SPSS, le fichier Excel ne doit pas être déjà ouvert dans Excel





- Read variable names = oui (coché)
- Worksheet = nom de la feuille
- Range : laisser en blanc parce qu'on prend tout le fichier



Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : NoSeq 1

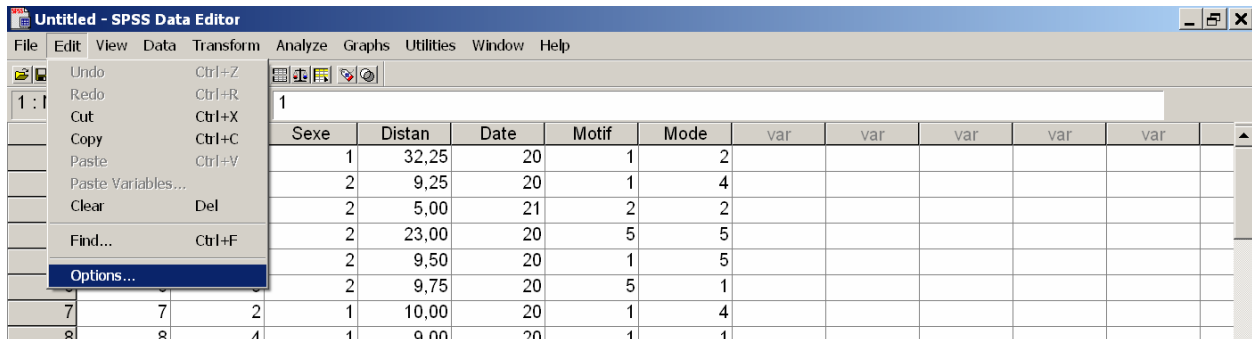
	NoSeq	Age	Sexe	Distan	Date	Motif	Mode	var	var	var	var	var	var	var
1	1	2	1	32,25	20	1	2							
2	2	3	2	9,25	20	1	4							
3	3	1	2	5,00	21	2	2							
4	4	5	2	23,00	20	5	5							
5	5	5	2	9,50	20	1	5							
6	6	3	2	9,75	20	5	1							
7	7	2	1	10,00	20	1	4							
8	8	4	1	9,00	20	1	1							
9	9	2	2	11,00	19	5	2							
10	10	5	1	6,00	20	1	5							
11	11	2	2	20,00	22	5	4							
12	12	3	2	3,00	20	1	2							
13	13	3	1	6,75	20	1	2							
14	14	2	2	22,25	20	0	5							
15	15	4	1	.	21	3	1							
16	16	2	1	3,25	19	1	6							
17	17	3	2	16,25	19	1	1							
18	18	2	3	3,25	20	1	1							
19	19	5	1	-2,50	22	2	1							
20	20	2	1	9,50	22	1	1							
21	21	2	2	3,00	21	2	1							
22	22	6	2	.	20	1	1							
23	23	3	2	8,75	20	1	5							
24	24	6	1	.	20	5	1							
25	25	3	1	2,75	19	1	5							
26	26	4	1	28,00	21	2	1							
27	27	6	1	3,00	21	5	5							
28	28	4	.	.	21	2	1							

Data View Variable View /

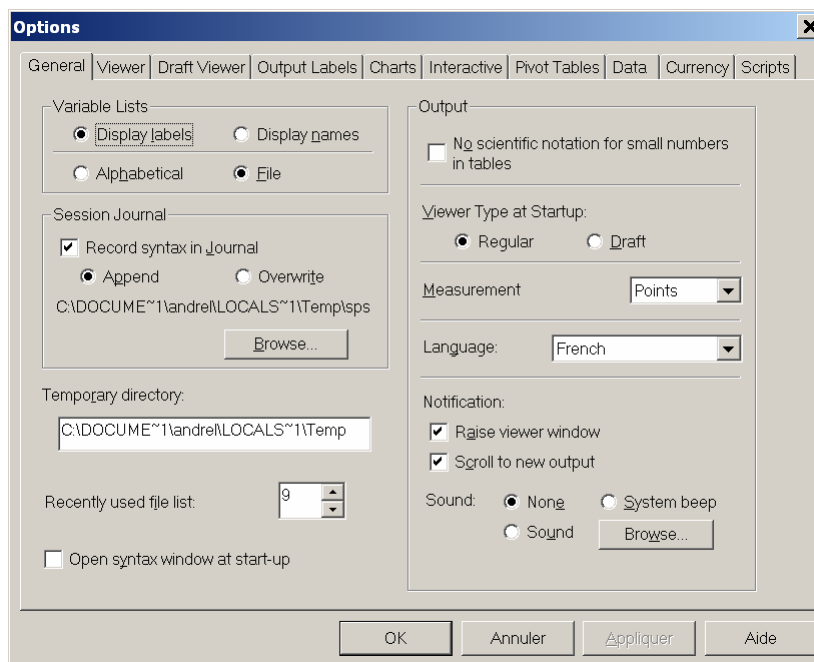
SPSS Processor is ready

## 2. Quelques réglages préliminaires

- Edit/Options



- Voir à l'onglet «General» où est le journal : ...\\spss.jnl
- Choisir la langue de sortie (vous aurez un mélange d'anglais et de mauvais français...)  
Cliquer sur « Appliquer »

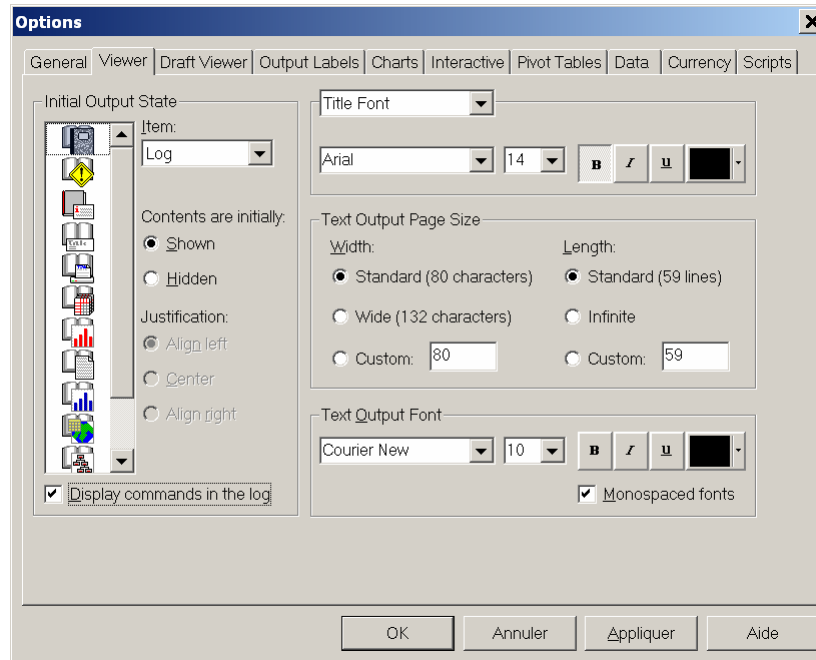


- Onglet « Viewer » :

Sous « Initial output state », choisir « Log » dans le menu déroulant sous « Item » :

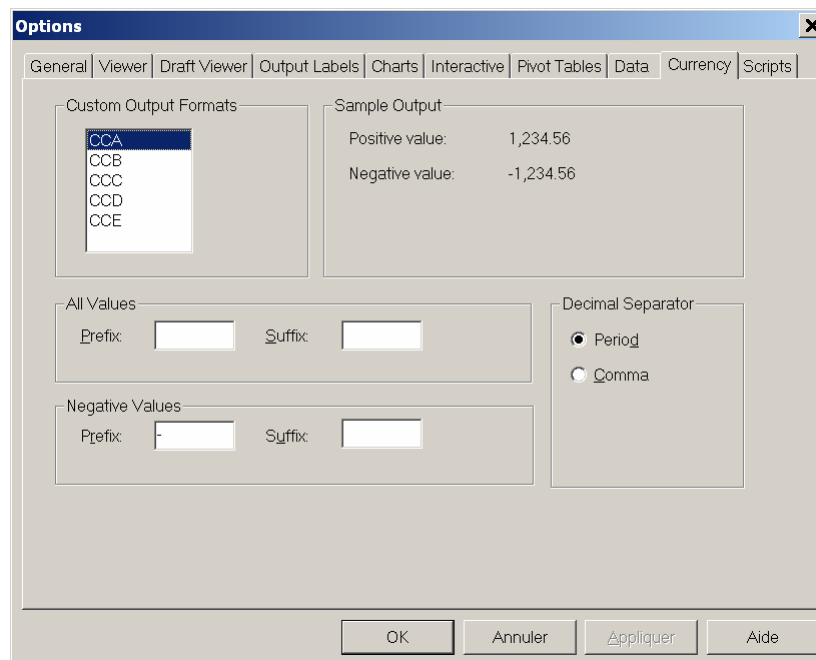
cocher « Shown » et cocher « Display commands in the log »

Cliquer sur « Appliquer »



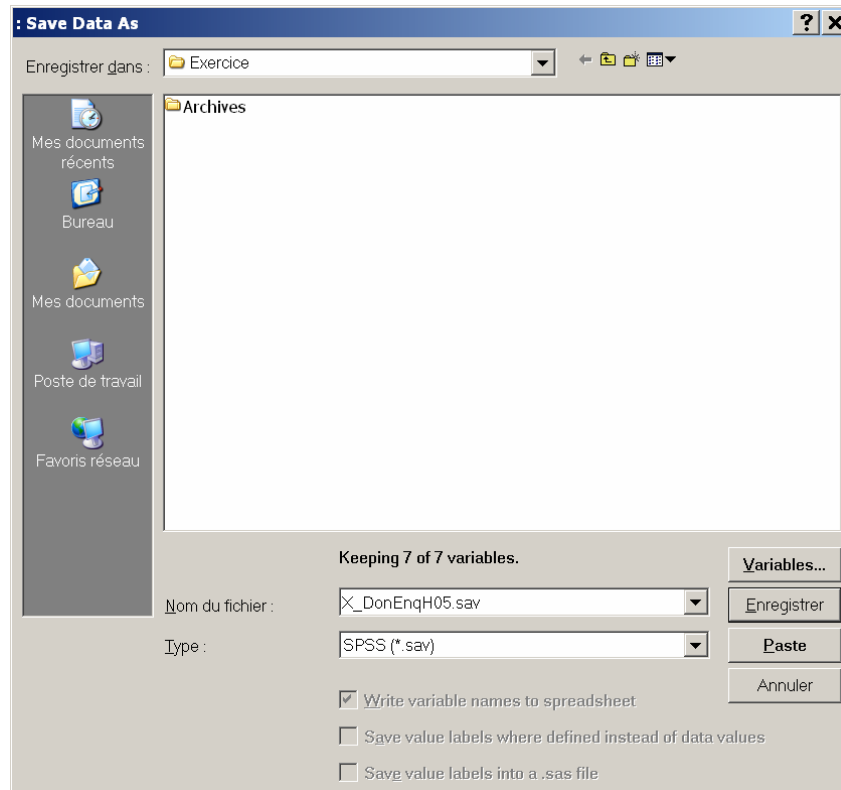
- Onglet « Currency » :

choisir entre « Period » (point) et « Comma » (virgule) comme séparateur décimal



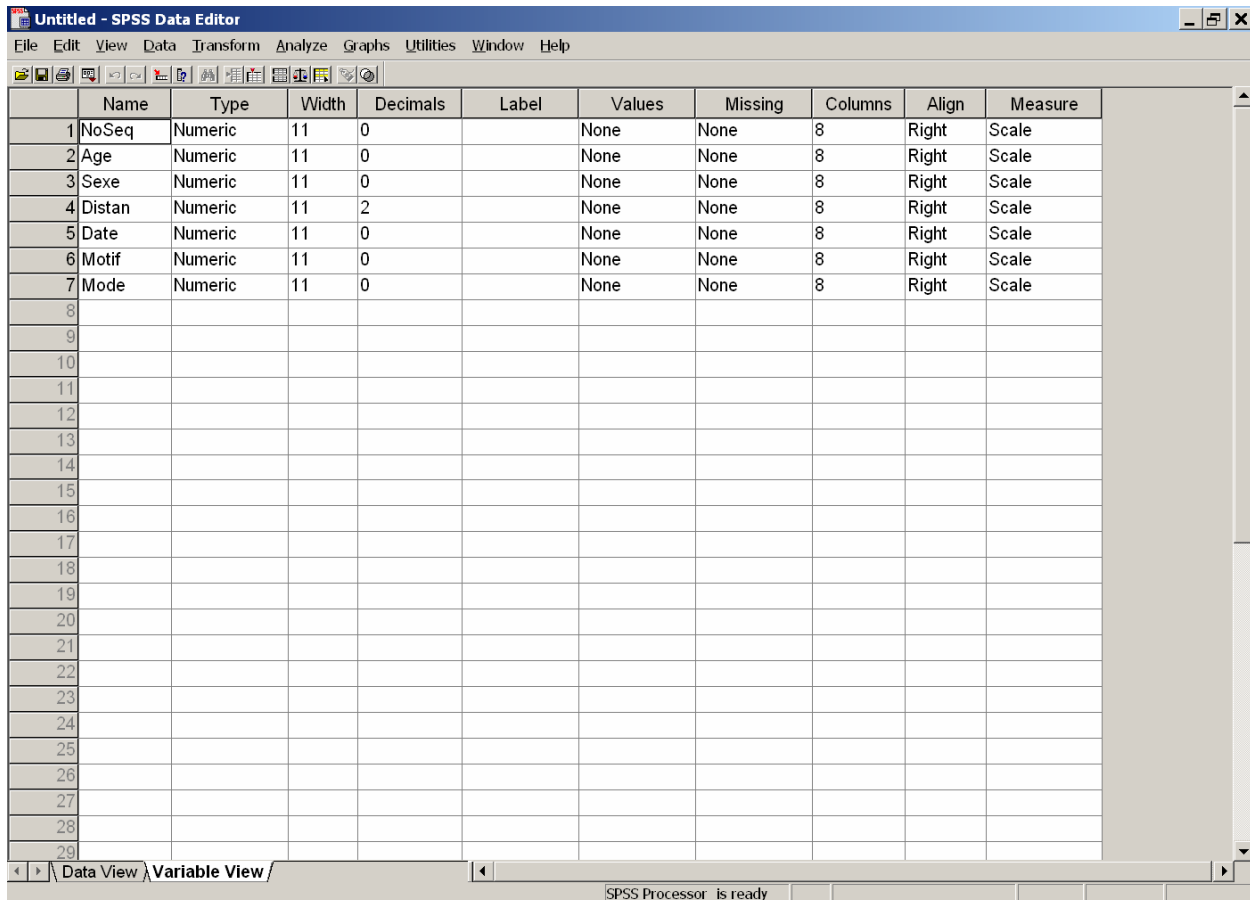
### 3. Sauvegarde des données en format SPSS

- Retour fenêtre « SPSS Data Editor », onglet « Data View »  
Sauvegarder le fichier en format SPSS : File/Save As...



Les fichiers de données SPSS ont le suffixe « \*.sav ».

## 4. L'onglet « Variable View » de la fenêtre « SPSS Data Editor » : liste des variables



The screenshot shows the SPSS Data Editor window in Variable View. The window title is 'Untitled - SPSS Data Editor'. The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Window, and Help. The toolbar contains various icons for file operations and data manipulation. The main area is a table with the following columns: Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, and Measure. The table contains 7 rows of variables:

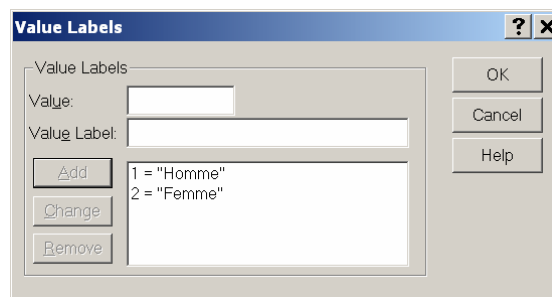
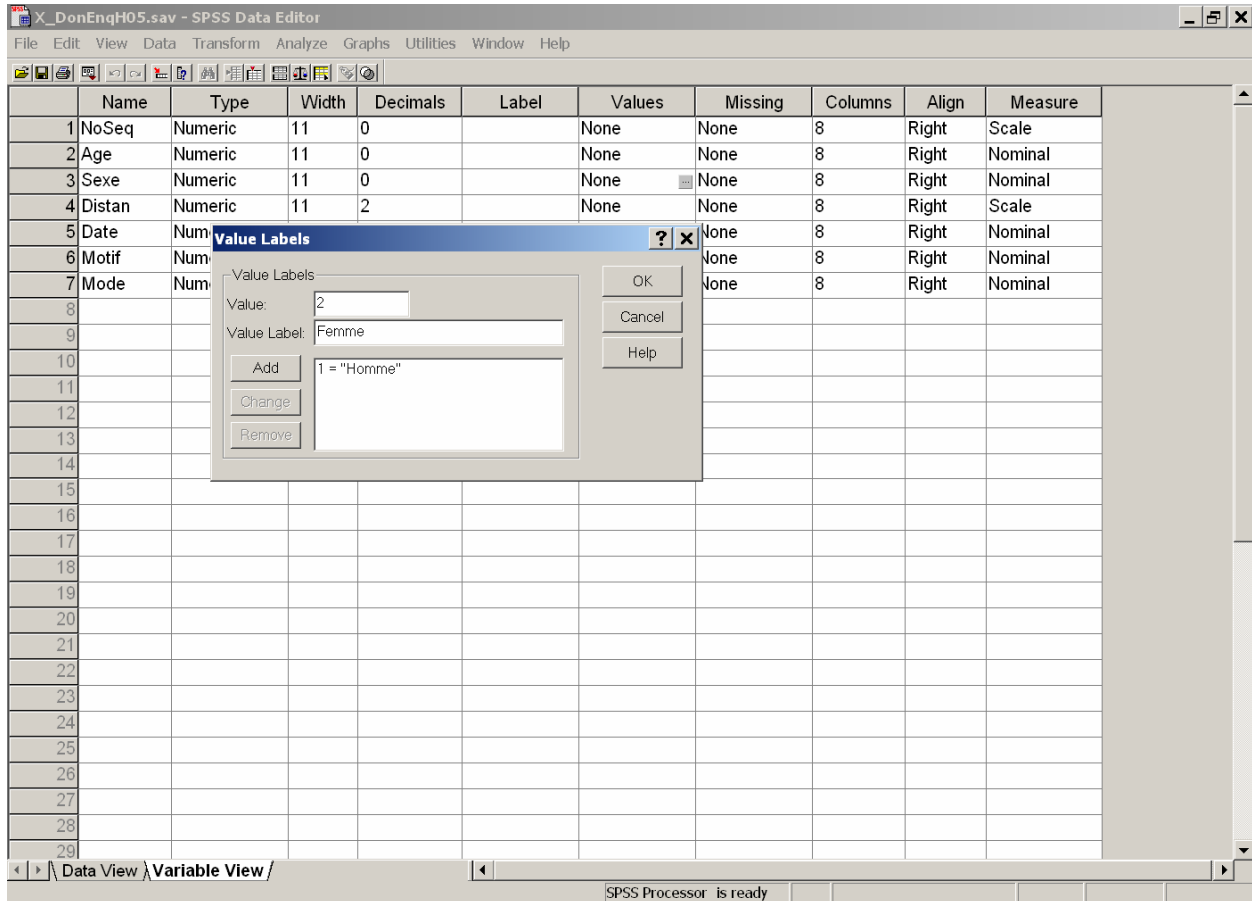
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	NoSeq	Numeric	11	0		None	None	8	Right	Scale
2	Age	Numeric	11	0		None	None	8	Right	Scale
3	Sexe	Numeric	11	0		None	None	8	Right	Scale
4	Distan	Numeric	11	2		None	None	8	Right	Scale
5	Date	Numeric	11	0		None	None	8	Right	Scale
6	Motif	Numeric	11	0		None	None	8	Right	Scale
7	Mode	Numeric	11	0		None	None	8	Right	Scale
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										
26										
27										
28										
29										

The status bar at the bottom indicates 'SPSS Processor is ready'.

- Colonne « Type » :
  - Sélectionner une cellule et cliquer sur
  - Valeur numérique (Numeric) ou alphanumérique (chaîne de caractères, *string*)
- Colonne « Label » : inscrire les descriptions de variables  
(les noms de variables dans SPSS doivent avoir 8 caractères ou moins)  
On peut voir les descriptions simplement en posant le pointeur sur l'entête, dans la feuille « Data View »

- Colonne « Values » : inscrire les étiquettes qu'on veut associer à chaque valeur  
(Ne pas oublier de cliquer « Add » à chaque coup)

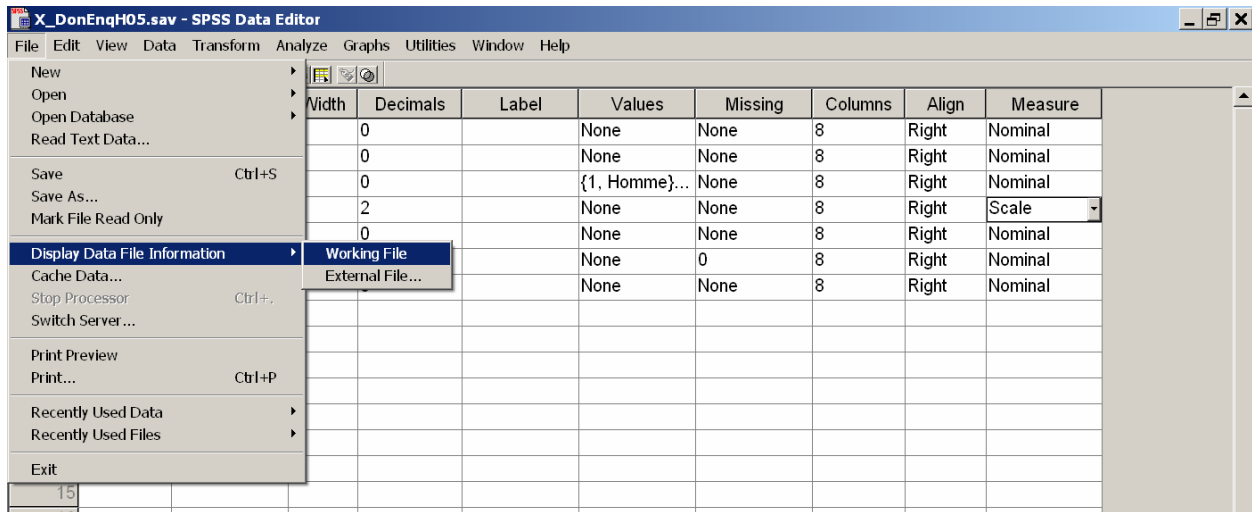
ex. : Pour la variable « Sexe », 1 = « Homme » et 2 = « Femme »



Dans la feuille « Data View », on fait apparaître/disparaître les étiquettes au moyen de View/Value Labels



- Menu File/Display Data File Information/Working File : compte rendu de l'information de la feuille « Variable View »



On est automatiquement transféré à la feuille « Output », où l'on voit :

**Informations de la variable**

Variable	Position	Etiquette	Niveau de mesure	Largeur des colonnes	Alignement	Format d'impression	Format d'écriture	Valeurs manquantes
NoSeq	1	<none>	Nominal	8	Right	F11	F11	
Age	2	<none>	Nominal	8	Right	F11	F11	
Sexe	3	<none>	Nominal	8	Right	F11	F11	
Distan	4	<none>	Scale	8	Right	F11.2	F11.2	
Date	5	<none>	Nominal	8	Right	F11	F11	
Motif	6	<none>	Nominal	8	Right	F11	F11	0
Mode	7	<none>	Nominal	8	Right	F11	F11	

Variables du fichier de travail

**Valeurs des variables**

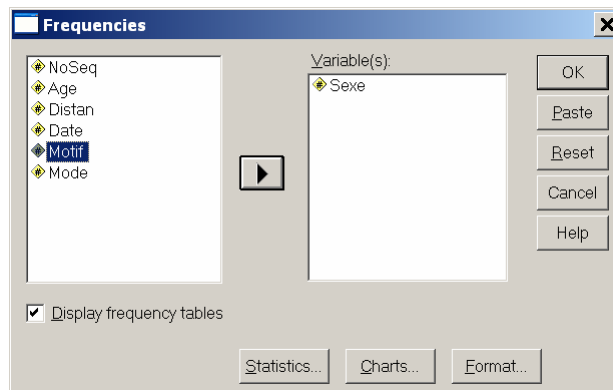
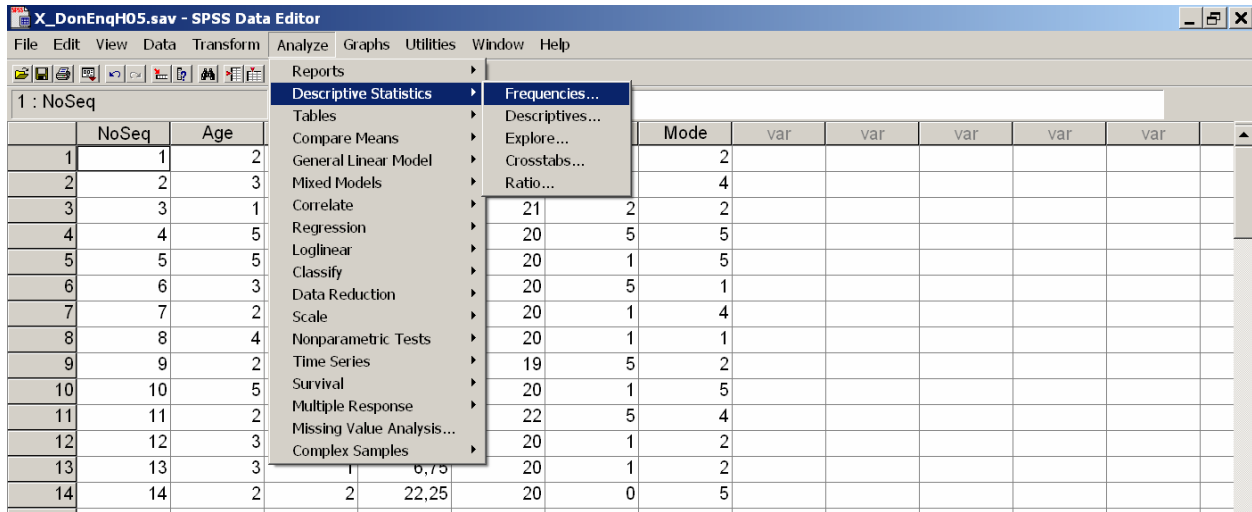
Valeur	Etiquette
Sexe 1	Homme
2	Femme

Note : C'est dans la feuille « Output » que sont affichés les résultats, ainsi que l'historique des commandes (le « log »), si on choisit cette option. Cette feuille est un fichier distinct du fichier de données : si l'on veut le conserver, il faut donc l'enregistrer (suffixe « \*.spo »). Normalement, au moment de fermer le logiciel, celui-ci vous demande si vous voulez enregistrer le fichier de résultats. Pour lui attribuer le nom de fichier que vous désirez, faites « Save As... ».

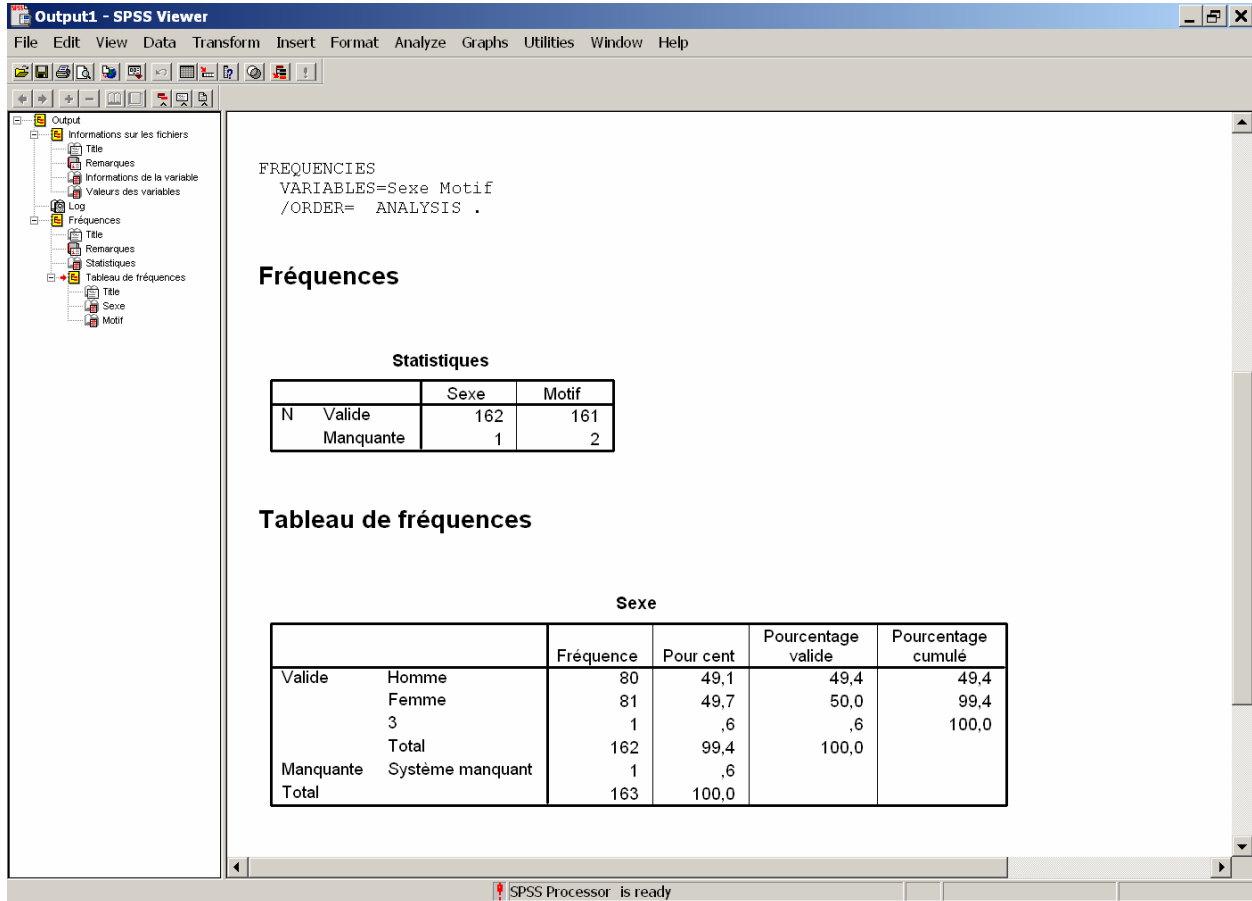


## 5. Tableau de fréquences

- Menu Analyze/Descriptive Statistics/Frequencies



**RÉSULTAT DANS LA FENÊTRE « OUTPUT »**



- Variable Sexe : 1 valeur manquante de système (« Missing – System »), dans le tableau de fréquences;
- noter aussi une variable aberrante : Sexe = 3

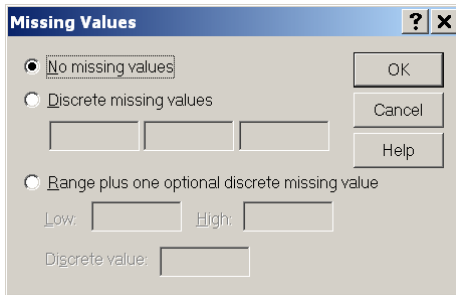
**Motif**

		Fréquence	Pour cent	Pourcentage valide	Pourcentage cumulé
Valide	1	71	43,6	44,1	44,1
	2	23	14,1	14,3	58,4
	3	23	14,1	14,3	72,7
	4	17	10,4	10,6	83,2
	5	27	16,6	16,8	100,0
	Total	161	98,8	100,0	
Manquante	0	2	1,2		
Total		163	100,0		

- Motif : 2 valeurs manquantes (« Missing ») mais pas « Système »

- Refaire pour Motif après avoir supprimé 0=« Discrete missing »

À la colonne « Missing » de la feuille « Variable View », cocher « No missing values » : il n'y a plus de valeurs manquantes



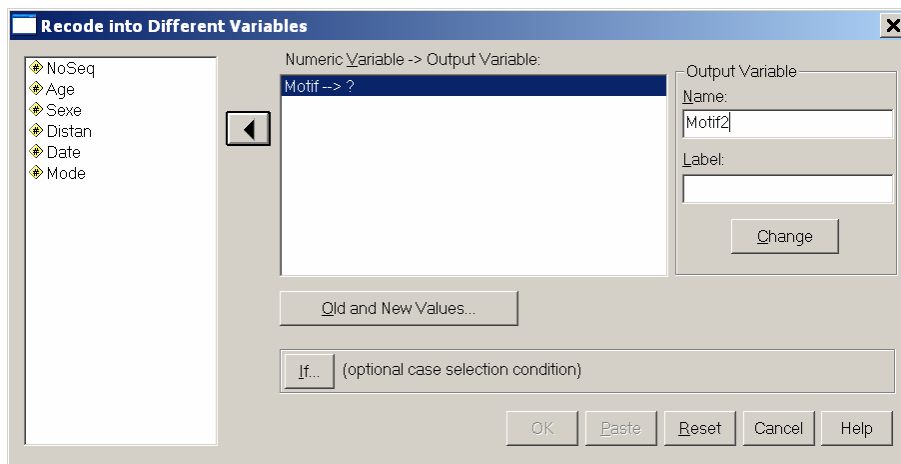
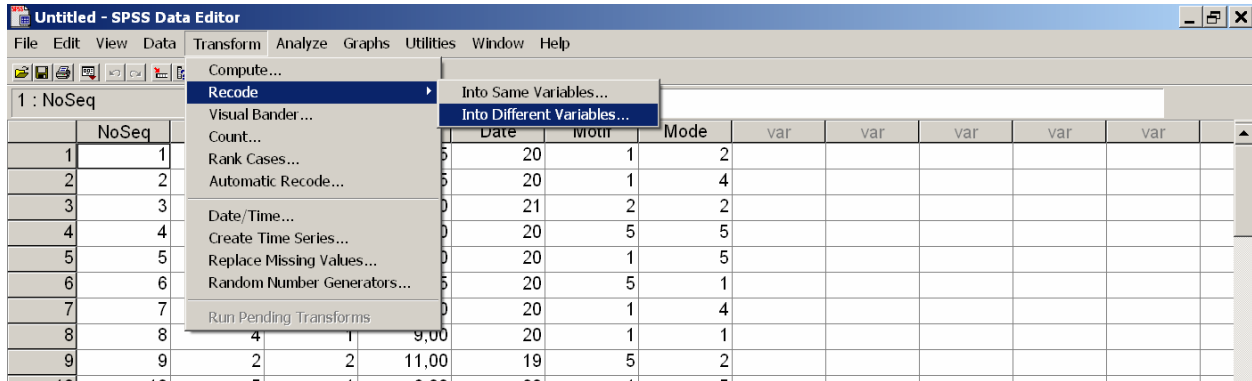
On obtient le tableau de fréquences suivant :

**Motif**

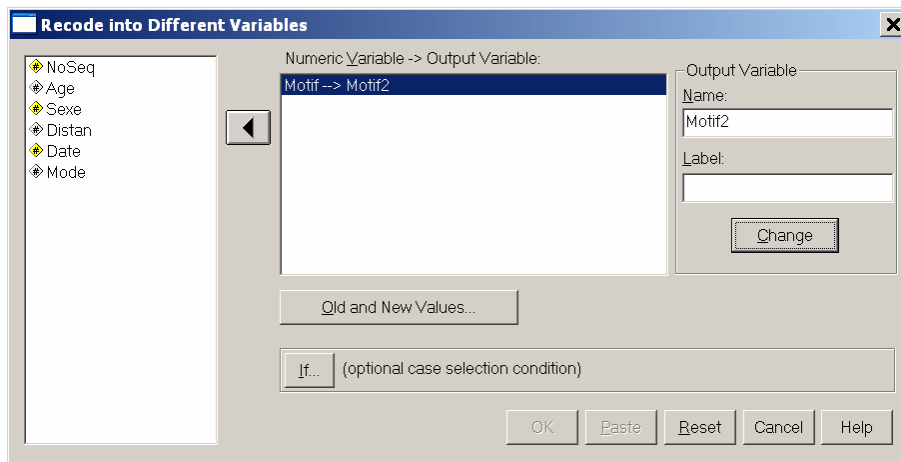
		Fréquence	Pour cent	Pourcentage valide	Pourcentage cumulé
Valide	0	2	1,2	1,2	1,2
	1	71	43,6	43,6	44,8
	2	23	14,1	14,1	58,9
	3	23	14,1	14,1	73,0
	4	17	10,4	10,4	83,4
	5	27	16,6	16,6	100,0
	Total	163	100,0	100,0	

## 6. Recodification de variables

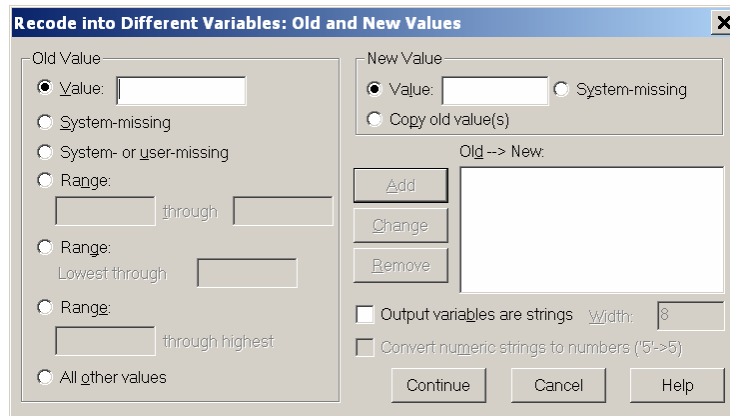
Menu Transform/Recode : créer la nouvelle variable Mode2 à partir de Mode



Cliquer « Change »

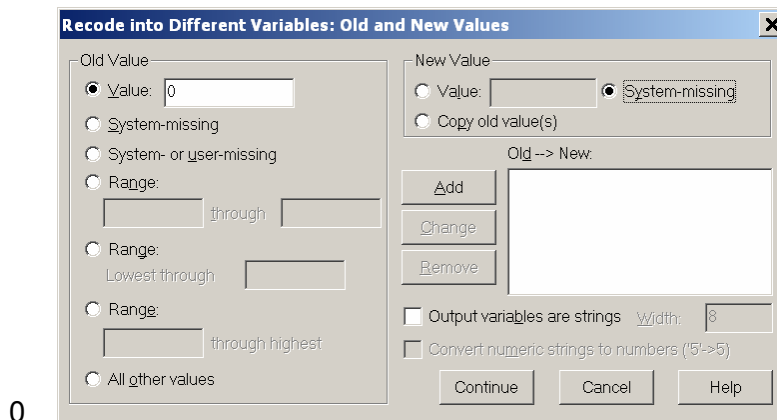


Cliquer sur « Old and New Values »

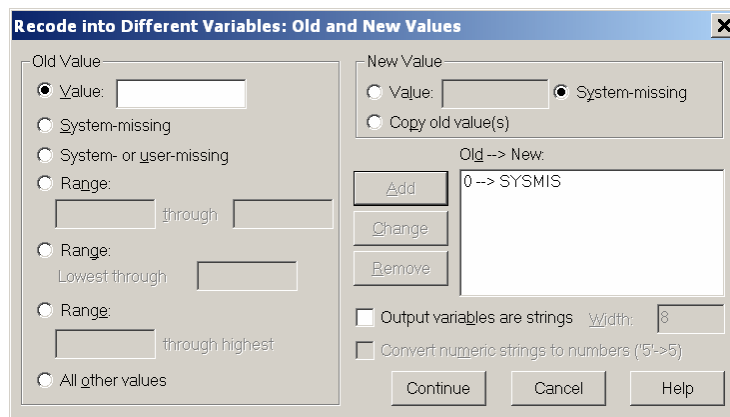


Inscrire la nouvelle codification :

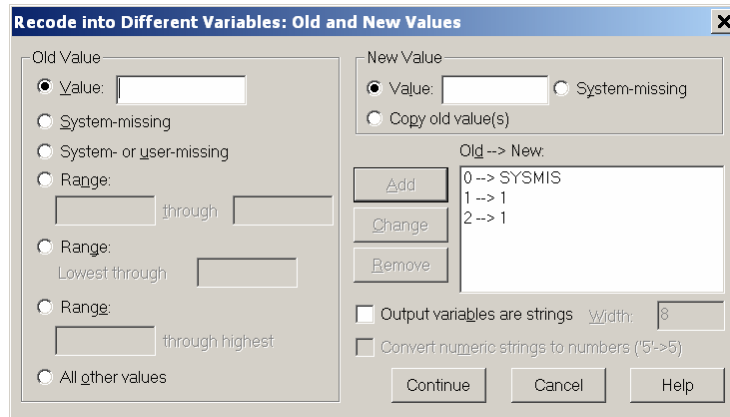
- 0 → missing (system)



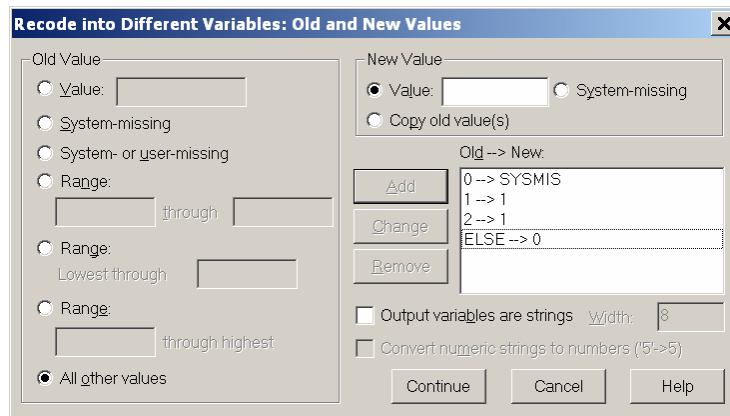
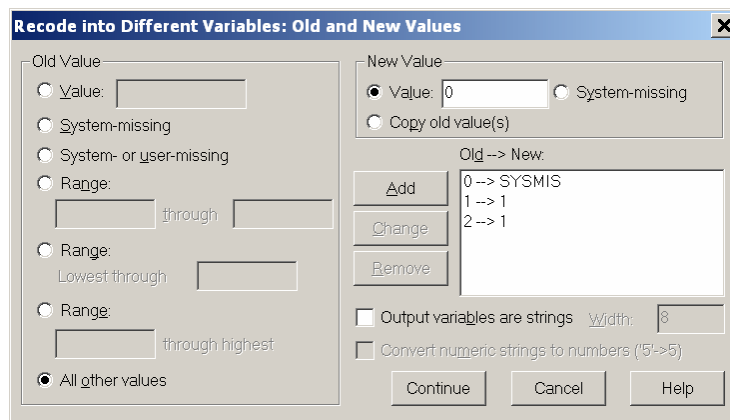
- Cliquer sur « Add »



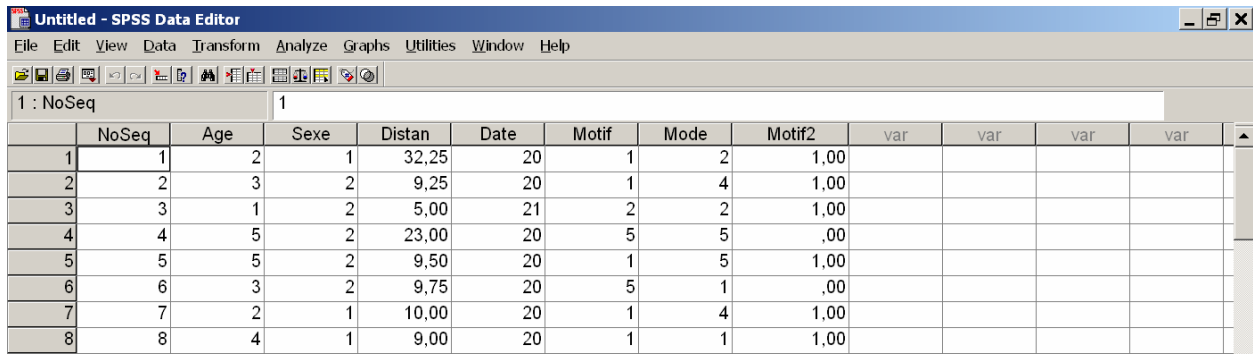
- 1 et 2 → 1 (automobile)



- Autres → 0 (pas automobile)  
Cocher « All other values »  
OU « Range » : 3 through 6  
OU « Range » 3 through highest



- Cliquer à chaque fois « Add »
- À la fin, cliquer «Continue», puis «OK»



The screenshot shows the SPSS Data Editor window titled "Untitled - SPSS Data Editor". The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Window, and Help. The toolbar contains various icons for file operations and data manipulation. The data grid shows a table with 8 rows and 14 columns. The first column is labeled "1" and contains values 1 through 8. The second column is labeled "NoSeq" and contains values 1 through 8. The third column is labeled "Age" and contains values 2, 3, 1, 5, 5, 3, 2, 4. The fourth column is labeled "Sexe" and contains values 1, 2, 2, 2, 2, 2, 1, 1. The fifth column is labeled "Distan" and contains values 32,25, 9,25, 5,00, 23,00, 9,50, 9,75, 10,00, 9,00. The sixth column is labeled "Date" and contains values 20, 20, 21, 20, 20, 20, 20, 20. The seventh column is labeled "Motif" and contains values 1, 1, 2, 5, 1, 5, 1, 1. The eighth column is labeled "Mode" and contains values 2, 4, 2, 5, 5, 1, 4, 1. The ninth column is labeled "Motif2" and contains values 1,00, 1,00, 1,00, ,00, 1,00, ,00, 1,00, 1,00. The last four columns are labeled "var" and contain empty cells.

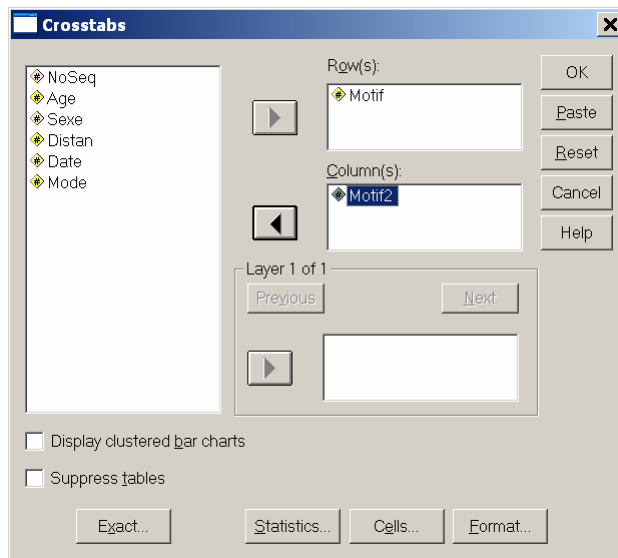
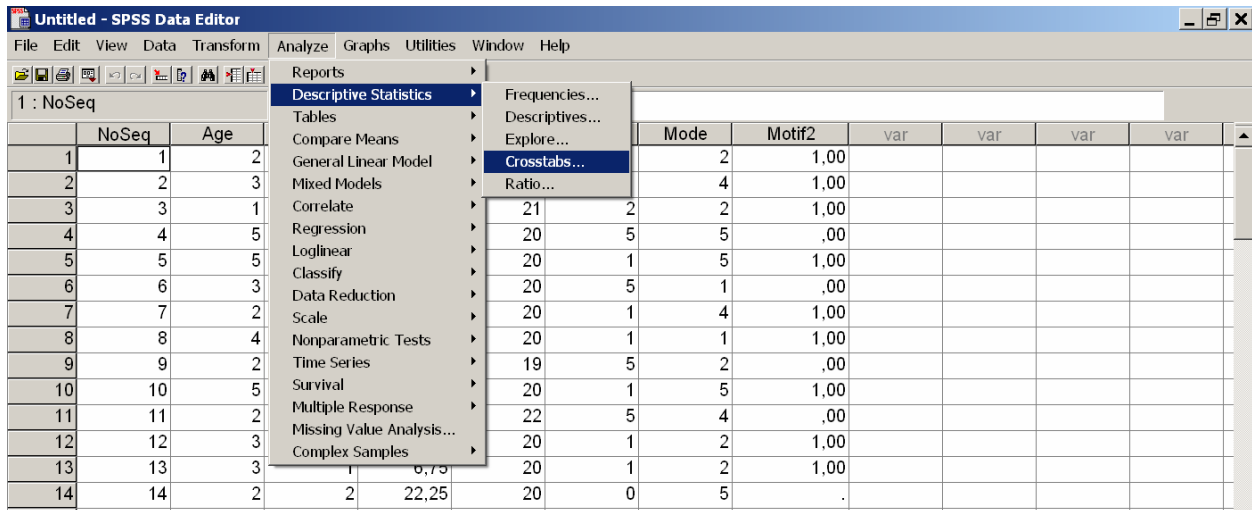
	NoSeq	Age	Sexe	Distan	Date	Motif	Mode	Motif2	var	var	var	var
1	1	2	1	32,25	20	1	2	1,00				
2	2	3	2	9,25	20	1	4	1,00				
3	3	1	2	5,00	21	2	2	1,00				
4	4	5	2	23,00	20	5	5	,00				
5	5	5	2	9,50	20	1	5	1,00				
6	6	3	2	9,75	20	5	1	,00				
7	7	2	1	10,00	20	1	4	1,00				
8	8	4	1	9,00	20	1	1	1,00				

N.B. propriétés de la nouvelle variable à corriger au besoin

- On vérifie le résultat grâce à un...

## 7. Tableau de contingence (1)

- Menu Analyze/Descriptive Statistics/Crosstabs





Output1 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Output  
 Log  
 Tableaux croisés  
 Titre  
 Remarques  
 Récapitulatif du traitement des observations  
 Tableau croisé Motif \* Motif2

```

RECODE
  Motif
  (0=SYSMIS) (1=1) (2=1) (ELSE=0) INTO Motif2 .
EXECUTE .
CROSSTABS
  /TABLES=Motif BY Motif2
  /FORMAT= AVALUE TABLES
  /CELLS= COUNT
  /COUNT ROUND CELL .
    
```

→ **Tableaux croisés**

**Récapitulatif du traitement des observations**

	Observations					
	Valide		Manquante		Total	
	N	Pourcent	N	Pourcent	N	Pourcent
Motif * Motif2	161	98,8%	2	1,2%	163	100,0%

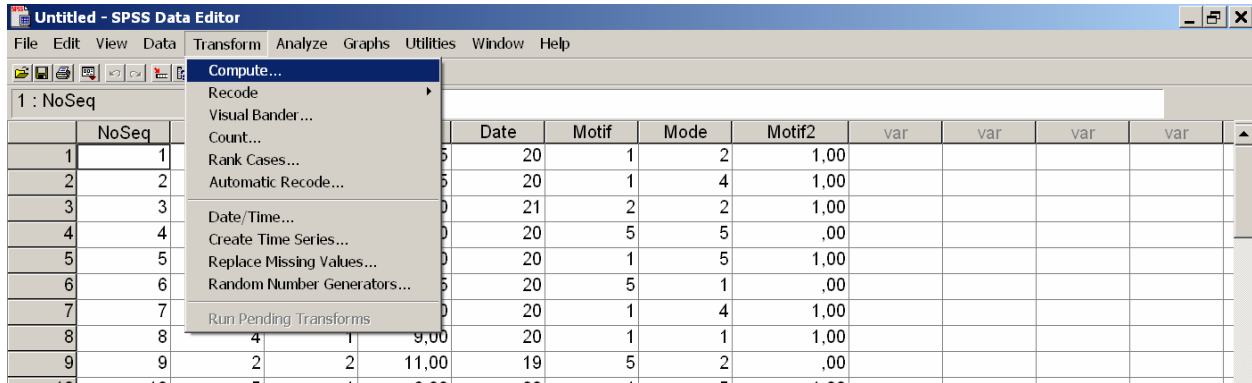
**Tableau croisé Motif \* Motif2**

Effectif		Motif2		Total
		,00	1,00	
Motif	1	0	71	71
	2	0	23	23
	3	23	0	23
	4	17	0	17
	5	27	0	27
Total		67	94	161

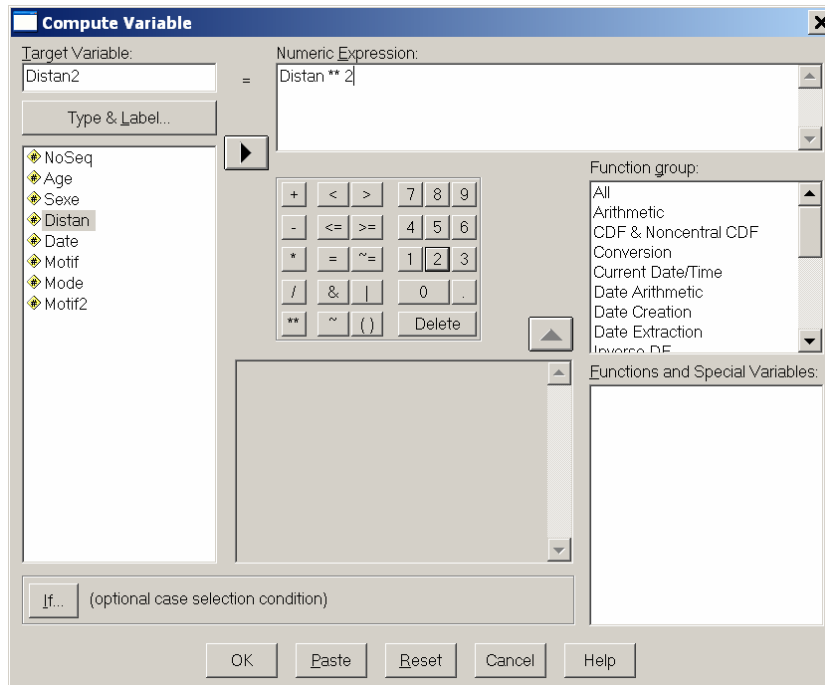
SPSS Processor is ready

## 8. Transformation de variables

### Menu Transform/Compute



- $Distan2 = (Distan)^2$



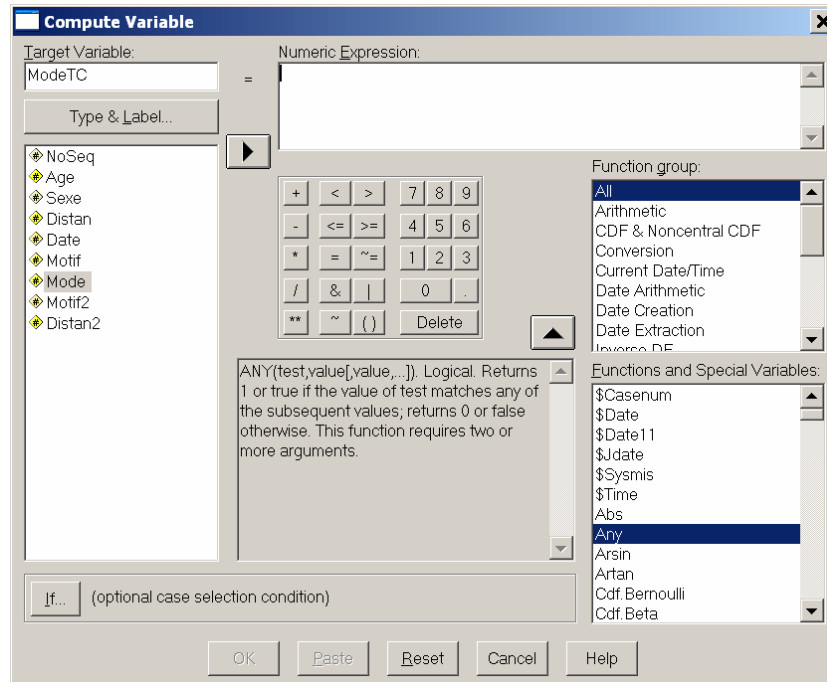
The screenshot shows the SPSS Data Editor window with the completed data table. The new variable 'Distan2' has been calculated and added to the table.

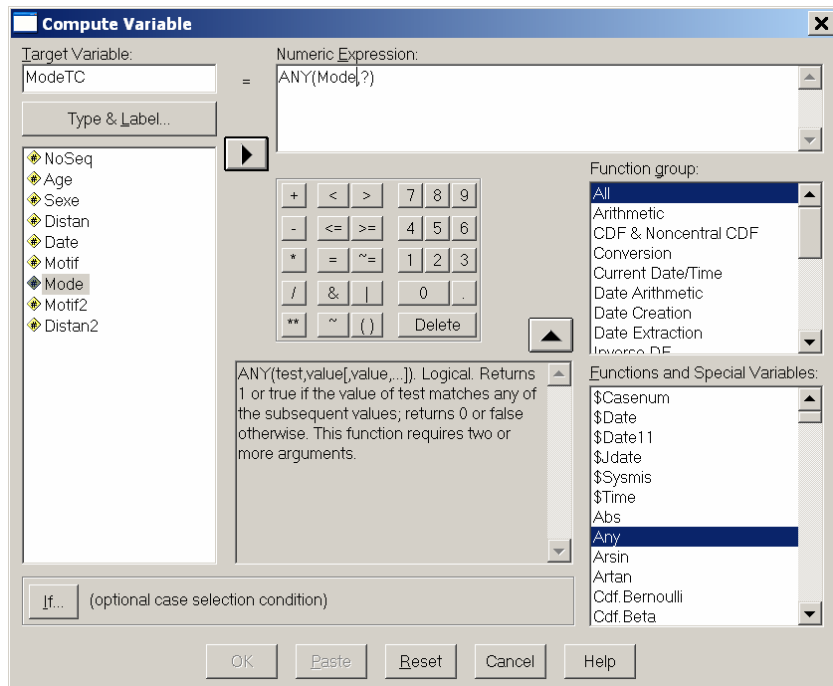
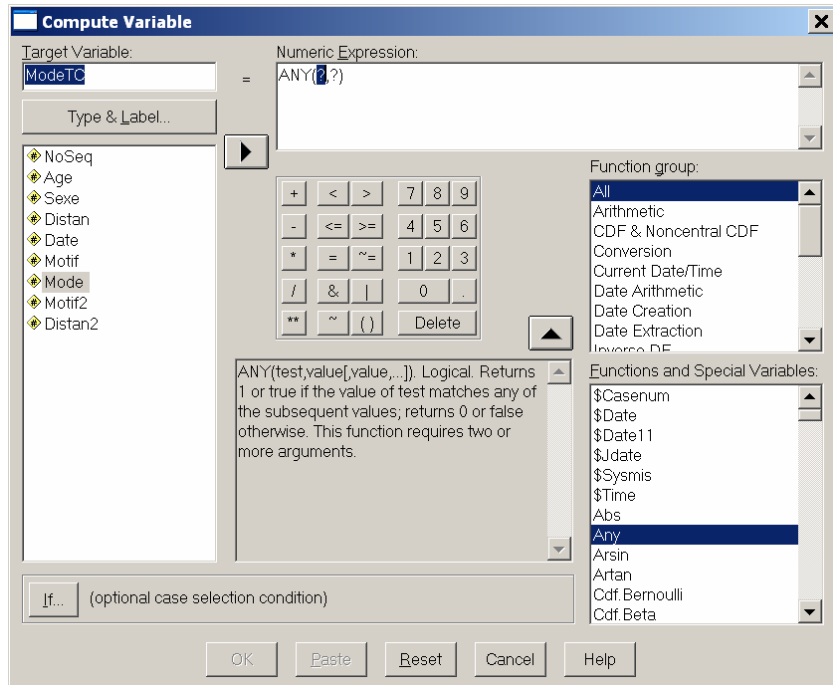
	NoSeq	Age	Sexe	Distan	Date	Motif	Mode	Motif2	Distan2	var	var	var
1	1	2	1	32,25	20	1	2	1,00	1040,06			
2	2	3	2	9,25	20	1	4	1,00	85,56			
3	3	1	2	5,00	21	2	2	1,00	25,00			
4	4	5	2	23,00	20	5	5	,00	529,00			
5	5	5	2	9,50	20	1	5	1,00	90,25			
6	6	3	2	9,75	20	5	1	,00	95,06			
7	7	2	1	10,00	20	1	4	1,00	100,00			
8	8	4	1	9,00	20	1	1	1,00	81,00			

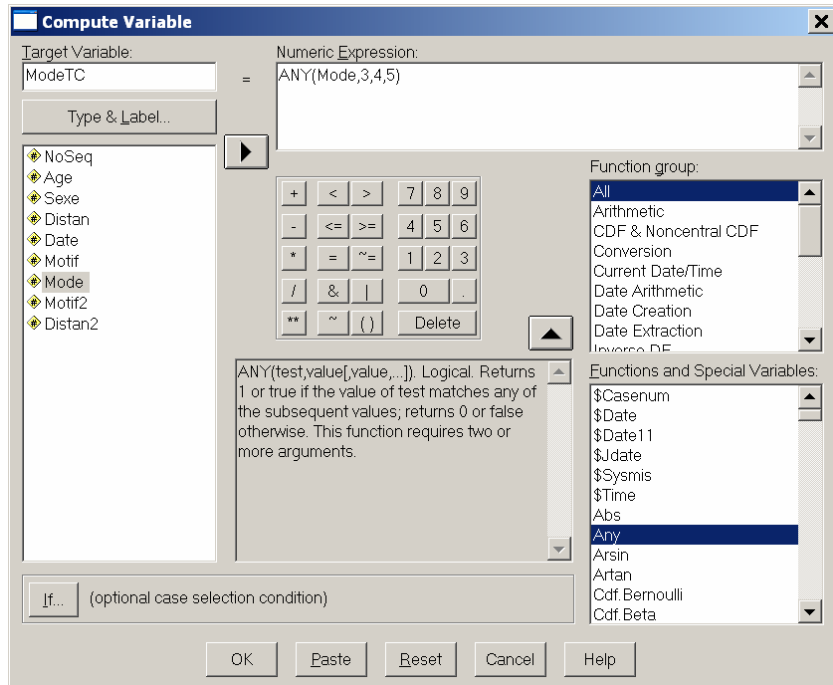
- ... au lieu de « Recode » : pour créer ModeTC (=1 si transport en commun)

Avec « Recode » :	
3, 4 et 5	→ 1 (TC)
Autres	→ 0 (pas TC)

- Avec « Compute », 2 possibilités
  - ModeTC = ANY(Mode, 3,4,5)







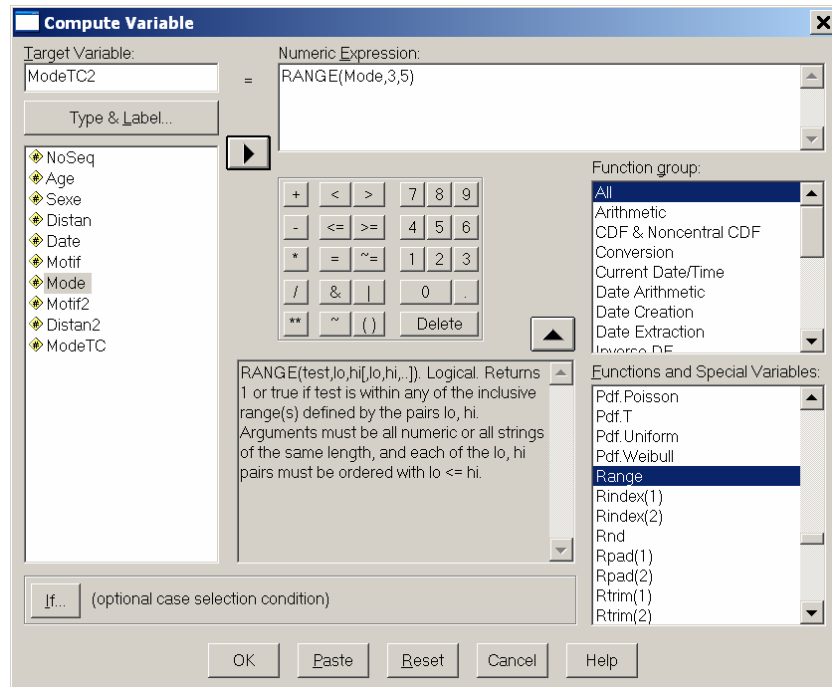
Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : Sexe 1

	NoSeq	Age	Sexe	Distan	Date	Motif	Mode	Motif2	Distan2	ModeTC	var	var
1	1	2	1	32,25	20	1	2	1,00	1040,06	,00		
2	2	3	2	9,25	20	1	4	1,00	85,56	1,00		
3	3	1	2	5,00	21	2	2	1,00	25,00	,00		
4	4	5	2	23,00	20	5	5	,00	529,00	1,00		
5	5	5	2	9,50	20	1	5	1,00	90,25	1,00		
6	6	3	2	9,75	20	5	1	,00	95,06	,00		
7	7	2	1	10,00	20	1	4	1,00	100,00	1,00		
8	8	4	1	9,00	20	1	1	1,00	81,00	,00		

- ModeTC = RANGE(Mode, 3,5)



## 9. Méthodes de sélection des cas

### POUR ÉCARTER LES RÉPONSES « NON SIGNIFIANTES » (« PAS DE RÉPONSE », « AUTRES »...)

1. Trier les observations et supprimer : **À ÉVITER !**

On peut les traiter comme données manquantes.

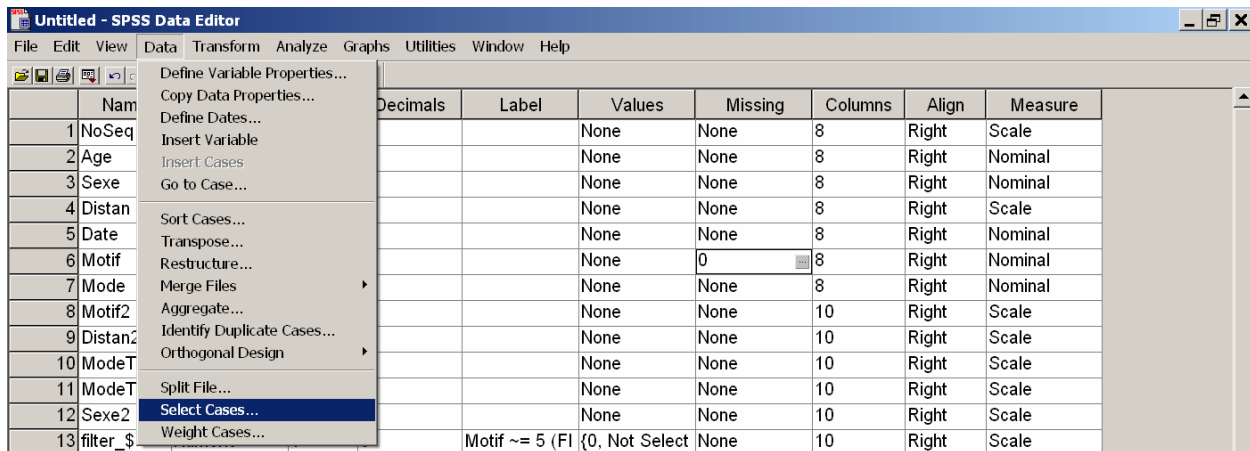
2. Onglet « Variable View », dans la colonne « Missing », spécifier les valeurs à écarter de l'analyse.

3. Transform/Recode : recoder les valeurs à écarter comme manquantes.

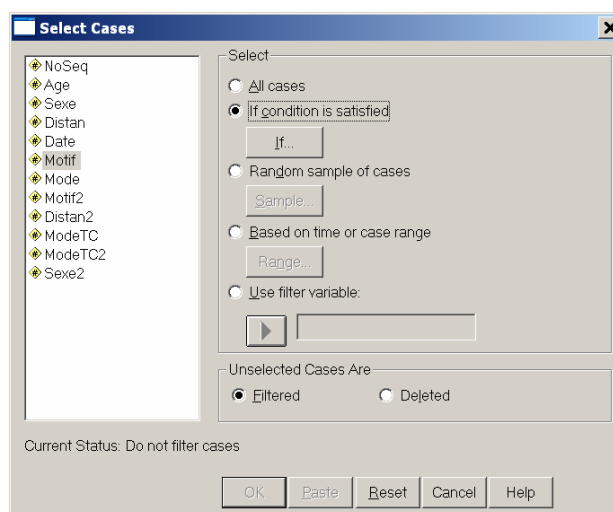
On peut aussi sélectionner les observations à inclure dans l'analyse.

4. Select case if...

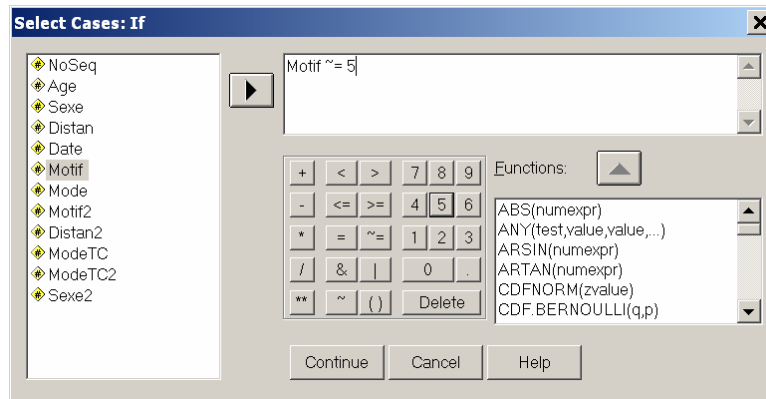
Menu Data/Select Cases



Cocher « If condition is satisfied »



Définir la condition : Motif~5 (Motif ≠ 5)



Cela crée une variable filtre qui peut être réutilisée par la suite :

	NoSeq	Age	Sexe	Distan	Date	Motif	Mode	Motif2	Distan2	ModeTC	filter_\$
1	1	2	1	32,25	20	1	2	1,00	1040,06	,00	1
2	2	3	2	9,25	20	1	4	1,00	85,56	1,00	1
3	3	1	2	5,00	21	2	2	1,00	25,00	,00	1
4	4	5	2	23,00	20	5	5	,00	529,00	1,00	0
5	5	5	2	9,50	20	1	5	1,00	90,25	1,00	1
6	6	3	2	9,75	20	5	1	,00	95,06	,00	0
7	7	2	1	10,00	20	1	4	1,00	100,00	1,00	1
8	8	4	1	9,00	20	1	1	1,00	81,00	,00	1

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	NoSeq	Numeric	11	0		None	None	8	Right	Scale
2	Age	Numeric	11	0		None	None	8	Right	Nominal
3	Sexe	Numeric	11	0		None	None	8	Right	Nominal
4	Distan	Numeric	11	2		None	None	8	Right	Scale
5	Date	Numeric	11	0		None	None	8	Right	Nominal
6	Motif	Numeric	11	0		None	0	8	Right	Nominal
7	Mode	Numeric	11	0		None	None	8	Right	Nominal
8	Motif2	Numeric	8	2		None	None	10	Right	Scale
9	Distan2	Numeric	8	2		None	None	10	Right	Scale
10	ModeTC	Numeric	8	2		None	None	10	Right	Scale
11	filter_\$	Numeric	1	0	Motif <math>\neq</math> 5 (FI)	{0, Not Select}	None	10	Right	Scale

5. Créer une variable de sélection, puis avec « Data/Select Cases », utiliser « Use filter variable »

ex. :  $OBS = 1 - ANY(MOTIF, 0, 5)$

Vérifier avec un tableau de contingence Motif x OBS

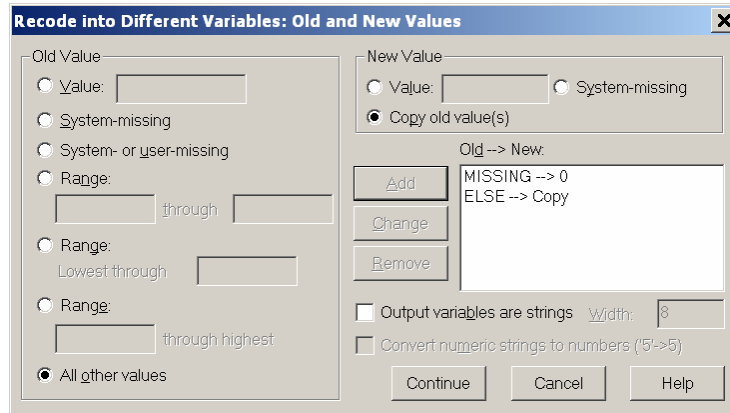


## POUR INCLURE LES DONNÉES MANQUANTES DANS L'ANALYSE

Par exemple, pour les compter dans un tableau de contingence.

- On peut remplacer les valeurs manquantes par une valeur numérique

Recode Sexe → Sexe2



- Comparer Motif x Sexe et Motif x Sexe2

**Récapitulatif du traitement des observations**

	Observations					
	Valide		Manquante		Total	
	N	Pourcent	N	Pourcent	N	Pourcent
Motif * Sexe	162	99,4%	1	,6%	163	100,0%

**Tableau croisé Motif \* Sexe**

Effectif

		Sexe			Total
		1	2	3	
Motif	0	1	1	0	2
	1	34	36	1	71
	2	8	14	0	22
	3	14	9	0	23
	4	6	11	0	17
	5	17	10	0	27
Total		80	81	1	162

**Récapitulatif du traitement des observations**

	Observations					
	Valide		Manquante		Total	
	N	Pourcent	N	Pourcent	N	Pourcent
Motif * Sexe2	163	100,0%	0	,0%	163	100,0%

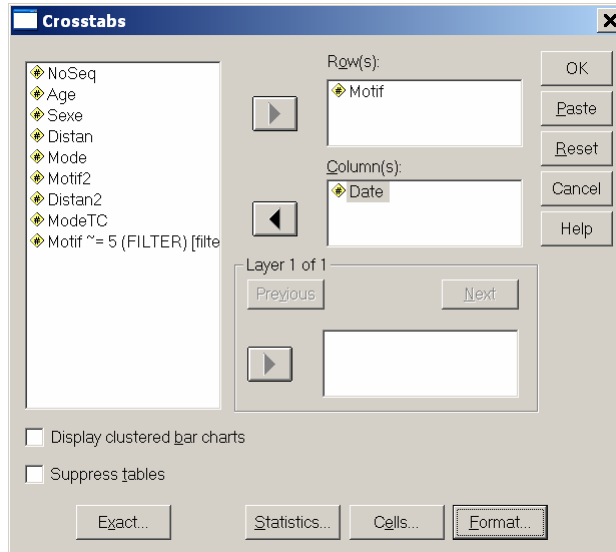
**Tableau croisé Motif \* Sexe2**

Effectif

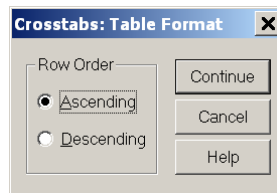
		Sexe2				Total
		,00	1,00	2,00	3,00	
Motif	0	0	1	1	0	2
	1	0	34	36	1	71
	2	1	8	14	0	23
	3	0	14	9	0	23
	4	0	6	11	0	17
	5	0	17	10	0	27
Total		1	80	81	1	163

## 10. Tableaux de contingence (2)

### OPTIONS DES TABLEAUX DE CONTINGENCE

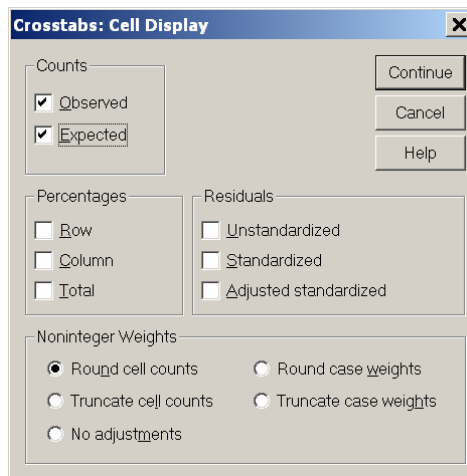


- Format : ordre des catégories (ascendant ou descendant)

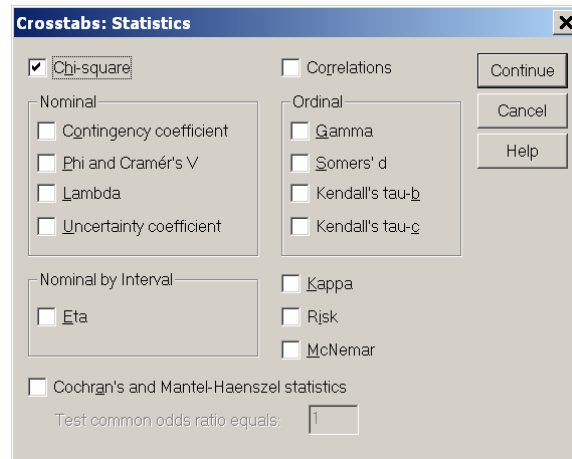


- Cells : que veut-on dans le tableau

N.B. « Expected » = fréquences théoriques



- Statistics :
  - « Chi-Square » = donne le test d'indépendance du Khi-deux de Pearson
  - Mesures d'association, dont Phi, V2...
  - Cliquer sur « Help » pour la description



RÉSULTAT

Récapitulatif du traitement des observations

	Observations					
	Valide		Manquante		Total	
	N	Pourcent	N	Pourcent	N	Pourcent
Motif * Date	134	98,5%	2	1,5%	136	100,0%

Tableau croisé Motif \* Date

			Date					Total
			19	20	21	22	23	
Motif 1	Effectif		9	35	8	19	0	71
	Effectif théorique		9,0	29,1	14,8	17,5	,5	71,0
2	Effectif		4	5	8	6	0	23
	Effectif théorique		2,9	9,4	4,8	5,7	,2	23,0
3	Effectif		2	9	6	6	0	23
	Effectif théorique		2,9	9,4	4,8	5,7	,2	23,0
4	Effectif		2	6	6	2	1	17
	Effectif théorique		2,2	7,0	3,6	4,2	,1	17,0
Total	Effectif		17	55	28	33	1	134
	Effectif théorique		17,0	55,0	28,0	33,0	1,0	134,0

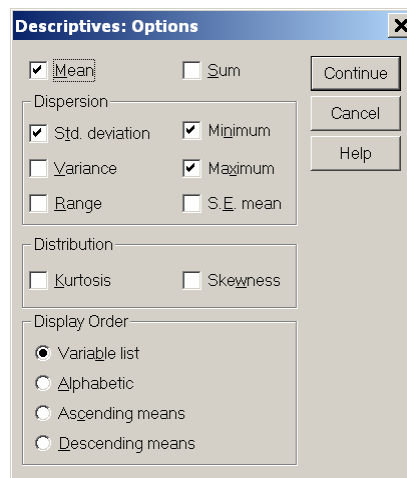
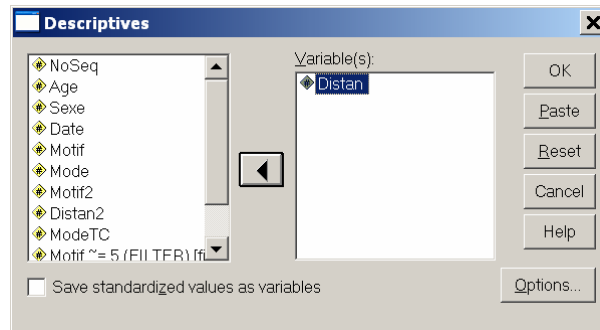
Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	19,576 <sup>a</sup>	12	,076
Rapport de vraisemblance	17,430	12	,134
Association linéaire par linéaire	,525	1	,469
Nombre d'observations valides	134		

a. 11 cellules (55,0%) ont un effectif théorique inférieur à 5.  
L'effectif théorique minimum est de ,13.

## 11. Autres analyses, notamment pour variables continues

- Analyze/Descriptive Statistics/Descriptives



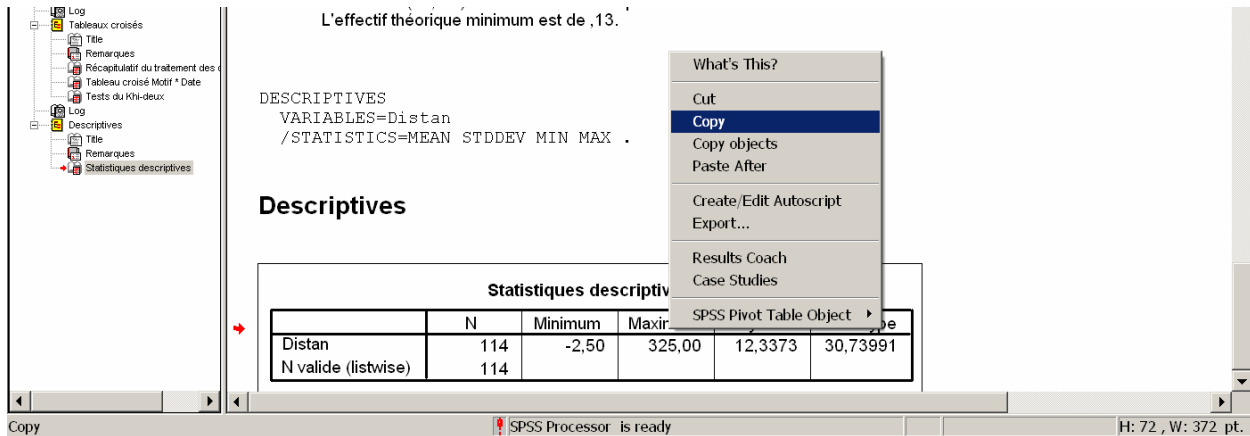
**Statistiques descriptives**

	N	Minimum	Maximum	Moyenne	Ecart type
Distan	114	-2,50	325,00	12,3373	30,73991
N valide (listwise)	114				

- Compare Means :
  - Tests de Student de différence entre moyennes
  - ANOVA à une dimension
- Corrélations
- Régression

## 12. Transfert des résultats vers Word ou Excel

- Sélectionner l'« image » d'un tableau, puis copier et coller  
Si on clique avec le bouton de droite de la souris, on a le choix entre « Copy » et « Copy objects »



The screenshot shows the SPSS interface with the 'Descriptives' output window. A context menu is open over the 'Statistiques descriptives' table. The menu options are: What's This?, Cut, Copy (highlighted), Copy objects, Paste After, Create/Edit Autoscript, Export..., Results Coach, Case Studies, and SPSS Pivot Table Object. The table data is as follows:

	N	Minimum	Maximum	Mean	Std. Dev.
Distan	114	-2,50	325,00	12,3373	30,73991
N valide (listwise)	114				

- Avec « Copy », les tableaux deviennent des tableaux dans Word; dans Excel, les données sont prêtes à être traitées.
- Avec « Copy objects », les tableaux s'inscrivent dans Word comme images

## 13. Aide et didacticiel

- Menu Help/Tutorial
- Le didacticiel (*tutorial*) peut se visionner de façon séquentielle, à la manière d'un cours, ou...
  - on peut accéder à la rubrique pertinente en passant par...
    - la table des matières (*contents*)
    - l'index
    - le moteur de recherche (*search*)
- Menu Help/Topics : présentation sous forme texte

## EXERCICE FACULTATIF

1. Vérifiez, parmi les propriétés attribuées à chacune des variables par SPSS, que l'échelle de mesure (le type de variable) est la bonne. Corrigez au besoin. SPSS utilise l'expression « *measurement level* » pour désigner l'échelle de mesure. SPSS distingue :

- *Scale* pour les variables d'intervalle ou rationnelles
- *Ordinal* pour les variables ordinales
- *Nominal* pour les variables catégoriques

À l'aide de la commande *Utilities/File Info*, produisez une liste des variables, avec leurs échelles de mesure, puis, au moyen d'un copier-coller, inscrivez cette liste dans votre rapport. Vous pouvez également produire cette liste à la main si vous le désirez.

Dans le cas des variables ordinales, indiquez s'il s'agit de variables ordinales d'ordre complet ou d'ordre faible ou réduit.

2. Validation sommaire des données.

a) Pour chacune des variables catégoriques et ordinales d'ordre incomplet, construisez un tableau de fréquences et indiquez dans votre rapport si toutes les valeurs font partie du domaine de variation de cette variable. Indiquez aussi le nombre d'observations pour lesquelles la valeur de cette variable est manquante.

b) Pour les autres variables, vérifiez que toutes les valeurs font partie du domaine de variation en repérant parmi les statistiques descriptives la valeur maximum et la valeur minimum. Incluez ces tableaux de fréquences et statistiques descriptives dans votre rapport en respectant les règles de présentation. Indiquez aussi le nombre d'observations pour lesquelles la valeur de cette variable est manquante.

3. Construisez les deux variables suivantes :

- *Vendredi* = 1 si *Date* = 20  
*Vendredi* = 0 autrement
- *AchaSeul* = 1 si *Motif* = 1  
*AchaSeul* = 0 autrement

Vérifiez que vos nouvelles variables sont construites correctement en produisant les tableaux croisés suivants :

- *Vendredi* par *Date*
- *AchaSeul* par *Motif*

Incluez ces deux tableaux de contingence dans votre rapport en respectant les règles de présentation applicables aux tableaux de contingence, tel que détaillées dans le manuel, à l'alinéa 4-1.1.3, p. 4-1.6).

4. À l'aide de SPSS, construisez deux tableaux de contingence : (a) *Motif* et *Date* (jour de la semaine) et (b) *AchaSeul* et *Vendredi*. Incluez ces deux tableaux de contingence dans votre rapport en respectant les règles de présentation.



## ANNEXE 1-E : TABLEAU DE L'ALPHABET GREC

Rang dans l'alphabet grec	Lettre majuscule	Lettre minuscule	Correspondance clavier (police symbole)	Nom de la lettre grecque	Phonétique grecque moderne
1	A	α	a	alpha	a
2	B	β	b	beta	v
3	Γ	γ	g	gamma	g
4	Δ	δ	d	delta	dh
5	E	ε	e	epsilon	e
6	Z	ζ	z	zeta	z
7	H	η	h	eta	i
8	Θ	θ	q	theta	th
9	I	ι	i	iota	i
10	K	κ	k	kappa	k
11	Λ	λ	l	lambda	l
12	M	μ	m	mu	m
13	N	ν	n	nu	n
14	Ξ	ξ	x	xi	x
15	O	ο	o	omicron	o
16	Π	π	p	pi	p
17	P	ρ	r	rho	r
18	Σ	σ	s	sigma	s
19	T	τ	t	tau	t
20	Υ	υ	u	upsilon	i
21	Φ	φ	f	phi	f
22	X	χ	c	chi ou khi	kh
23	Ψ	ψ	y	psi	«ps»
24	Ω	ω	w	omega	ô
	ϑ	φ	j	Formes archaïques de theta et phi	
	ς	ϖ	v	ς est la forme de sigma comme lettre finale	

## ANNEXE 1-F : DÉVELOPPEMENT DE LA FORMULE DE CALCUL DE L'INDICE DE GINI

Dans le cas où le nombre d'observations est fini (cas discret), la différence moyenne de Gini s'écrit <sup>1</sup> :

$$\Delta = \frac{1}{N^2} \sum_{j=1}^n \sum_{k=1}^n |y_j - y_k| f_j f_k$$

où

$n$  est le nombre de valeurs distinctes observées

où  $f_j$  est la fréquence de la valeur  $y_j$  dans la distribution, de sorte que

$$N = \sum_{j=1}^n f_j \text{ est le nombre d'observations}$$

Lorsque les observations sont groupées par classes, la valeur  $y_j$  est la valeur moyenne de la variable  $Y$  dans la classe  $j$ .

Écrivons

$$v_j = \frac{f_j}{N}, \text{ la fraction de la population appartenant à la classe } j.$$

La valeur moyenne de la variable  $Y$  s'écrit alors

$$\mu = \frac{1}{N} \sum_{j=1}^n f_j y_j = \sum_{j=1}^n v_j y_j$$

Soit

$$M = \sum_{j=1}^n f_j y_j, \text{ la somme des valeurs de la variable } Y, \text{ et}$$

$$w_j = \frac{f_j y_j}{\sum_{k=1}^n f_k y_k} = \frac{f_j y_j}{N \mu} = \frac{v_j y_j}{\mu}, \text{ la fraction de la somme allouée à la classe } j.$$

---

<sup>1</sup> Dans cette formule chaque observation est comparée à chacune des observations, y compris à elle-même ; c'est la différence moyenne avec répétition. Kendall et Stuart (1991, p. 58) donnent aussi la formule *sans* répétition. Lorsque  $N$  est grand, la différence est négligeable.

Développons maintenant la formule de la différence moyenne de Gini :

$$\Delta = \frac{1}{N^2} \sum_{j=1}^n \sum_{k=1}^n |y_j - y_k| f_j f_k$$

$$\Delta = \sum_{j=1}^n \sum_{k=1}^n |y_j - y_k| v_j v_k$$

$$\Delta = \sum_{j=1}^n \sum_{k=1}^n |v_j v_k y_j - v_j v_k y_k|$$

$$\Delta = \sum_{j=1}^n \sum_{k=1}^n |v_k \mu w_j - v_j \mu w_k|$$

$$\Delta = \mu \sum_{j=1}^n \sum_{k=1}^n |v_k w_j - v_j w_k|$$

La suite du développement s'appuie sur deux observations. D'abord,

$$v_k w_j - v_j w_k = 0 \text{ pour } k = j.$$

Ensuite, si les observations ont été rangées, en vue de la construction d'une courbe de Lorenz, par ordre croissant des rapports  $w_j/v_j$ , alors

$$[k < j] \Rightarrow \left[ \frac{w_j}{v_j} \geq \frac{w_k}{v_k} \right] \Rightarrow [v_k w_j - v_j w_k \geq 0]$$

et on peut écrire

$$\Delta = 2\mu \sum_{j=1}^n \sum_{k < j} |v_k w_j - v_j w_k| = 2\mu \sum_{j=1}^n \sum_{k < j} (v_k w_j - v_j w_k)$$

$$\Delta = 2\mu \left( \sum_{j=1}^n \sum_{k < j} v_k w_j - \sum_{j=1}^n \sum_{k < j} v_j w_k \right)$$

$$\Delta = 2\mu \left( \sum_{j=1}^n w_j \sum_{k < j} v_k - \sum_{j=1}^n v_j \sum_{k < j} w_k \right)$$

Développons le premier des deux termes entre parenthèses :

$$\sum_{j=1}^n w_j \sum_{k < j} v_k = w_1 v_0 + w_2 (v_1) + w_3 (v_1 + v_2) + w_4 (v_1 + v_2 + v_3) + \dots$$

où  $v_0 = 0$ .

$$\sum_{j=1}^n w_j \sum_{k < j} v_k = \sum_{k=1}^n v_k \sum_{j > k} w_j = \sum_{k=1}^n v_k \left( 1 - \sum_{j \leq k} w_j \right) = \sum_{k=1}^n v_k \left( 1 - \sum_{j=1}^k w_j \right)$$

Pour finir, complétons la notation en posant

$$Cw_j = \sum_{k=1}^j w_k$$

On obtient alors

$$\sum_{j=1}^n w_j \sum_{k < j} v_k = \sum_{k=1}^n v_k (1 - Cw_k) = 1 - \sum_{k=1}^n v_k Cw_k = 1 - \sum_{j=1}^n v_j Cw_j$$

Et il résulte

$$\Delta = 2\mu \left( 1 - \sum_{j=1}^n v_j Cw_j - \sum_{j=1}^n v_j Cw_{j-1} \right)$$

L'indice de concentration de Gini est simplement le rapport de la différence moyenne de Gini sur deux fois la moyenne :

$$G = \frac{\Delta}{2\mu} = 1 - \left( \sum_{j=1}^n v_j Cw_j + \sum_{j=1}^n v_j Cw_{j-1} \right) = 1 - \sum_{j=1}^n v_j (Cw_j + Cw_{j-1})$$

## QUANTITÉ ET MESURE - DÉFINITIONS PRÉLIMINAIRES

### **Quantitatif s'oppose à qualitatif**

- pas mutuellement exclusifs : méthodologiquement complémentaires
- s'opposent quant à la définition

### **Quantitatif = ce qui se mesure**

- La quantité est la propriété de ce qui peut être *mesuré* ou *compté*.

### **Concept**

Théories et hypothèses sont formulées au moyen de concepts et de relations entre concepts.

Un **concept** est une idée, comme « pauvreté ». Plus formellement,

Concept = représentation mentale abstraite et générale d'une chose

- extension = ensemble des objets auxquels s'applique le concept
- compréhension = ensemble des caractères qui appartiennent à ce concept

Un concept comprend habituellement plusieurs **dimensions**

ex. : la pauvreté dans les pays en développement a différents aspects, comme :

- un bas niveau de revenu
- la faim
- la maladie, faute d'accès aux services de santé
- un faible niveau de scolarité, par manque d'écoles ou parce que les enfants doivent travailler
- etc.

### **Opérationnalisation des concepts**

Pour rapprocher les propositions théoriques de la réalité, il faut **opérationnaliser** les concepts :

→ identifier ses **dimensions**

→ associer des **indicateurs** (mesures) à ses différentes dimensions

→ appliquer les indicateurs à une situation donnée → **variables**

## CONCEPTS, INDICATEURS ET VARIABLES

**Opérationnaliser** les concepts : **dimensions** → **indicateurs** → **variables**

- Un indicateur, en sciences sociales, est comme un instrument de mesure dans les sciences physiques (thermomètre, voltmètre, compteur de vitesse, odomètre, etc.).

Un instrument de mesure produit un chiffre qui caractérise un aspect donné d'un phénomène.

Quand le compteur de vitesse d'une automobile pointe le chiffre « 70 », cela veut dire que l'auto se déplace à 70 km/heure. Quand la vitesse change, l'aiguille se déplace pour pointer le chiffre correspondant (dans le cas d'un instrument à affichage numérique, le chiffre affiché change). Le compteur de vitesse nous informe de l'aspect « vitesse » du phénomène de déplacement du véhicule. Mais il ne dit rien de la direction...

- Une **variable** est le résultat de l'application d'un indicateur à une situation donnée.  
Par exemple :
  - Chaque jour, durant 30 jours, prenons à l'aide d'un thermomètre la température de l'air à 08h00 sur le palier devant la porte principale de l'INRS-UCS.
  - À la fin, nous aurons 30 valeurs de la variable « température de l'air à 08h00 devant la porte principale de l'INRS-UCS ».
  - Une variable s'appelle « variable » parce qu'elle est le résultat de la mesure des *variations* de ce que l'on observe.

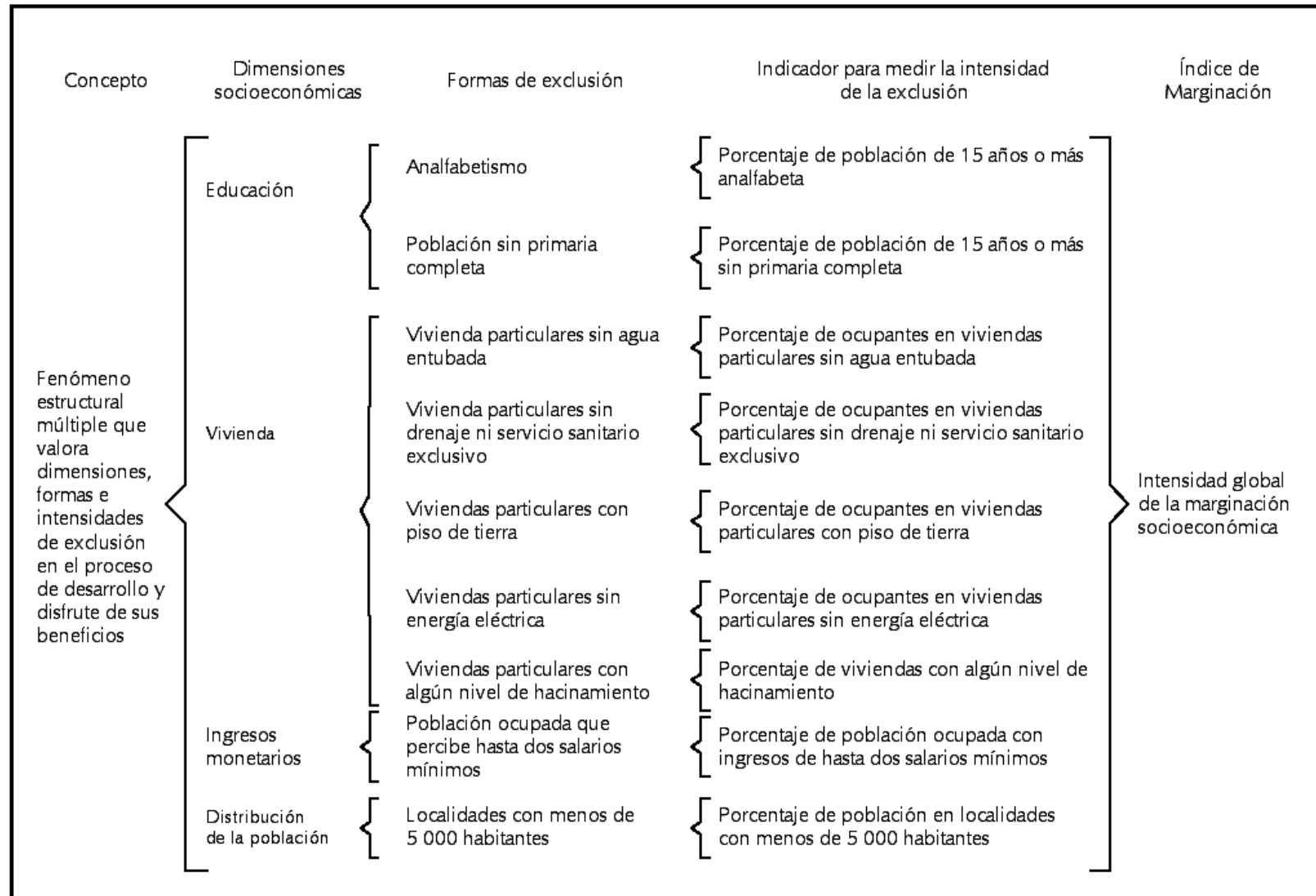
### Propriétés d'un bon indicateur

- **valide** = reflète les variations relatives au concept même que la variable est censée représenter (et non à un autre concept).
- **fiable** = les différences mesurées correspondent à de véritables différences.

### Exemples du passage du concept aux variables

- Indice et degré de marginalisation du Conseil National de la Population (CONAPO) au Mexique
- Indicateur de développement humain (IDH) du Programme des Nations Unies pour le Développement (PNUD)

**Figura 1.1. Esquema conceptual de la marginación**



## QU'EST-CE QUE LA MESURE ?

### Mesure

Mesurer, c'est comparer

Une *mesure* est une correspondance qui permet de comparer deux objets par rapport à une propriété donnée.

Et plus précisément,

Une *mesure* est une correspondance qui permet, pour au moins l'une des relations qui suivent, de déterminer si elle est vraie ou fausse (Voir carte schématique d'Amérique Centrale).

$$f(A) = f(B)$$

$$f(A) \neq f(B)$$

$$f(A) < f(B)$$

$$f(A) > f(B)$$

**Peut-on mesurer la nationalité ? – En un certain sens, oui !**

### Échelles de mesure et types de variables

1. Variables *catégoriques* («nominal» en anglais) : à quelle catégorie appartient l'individu.
  - Variable *dichotomiques* : 2 catégories possibles ;
  - Variable *polytomique* : plus de 2 catégories.
2. Variables *ordinales* : classier les individus en ordre croissant ou décroissant.
  - ordre *réduit* ou *faible* – par classes d'équivalence ;
  - ordre *complet*.

Définies «à une transformation monotone croissante près».

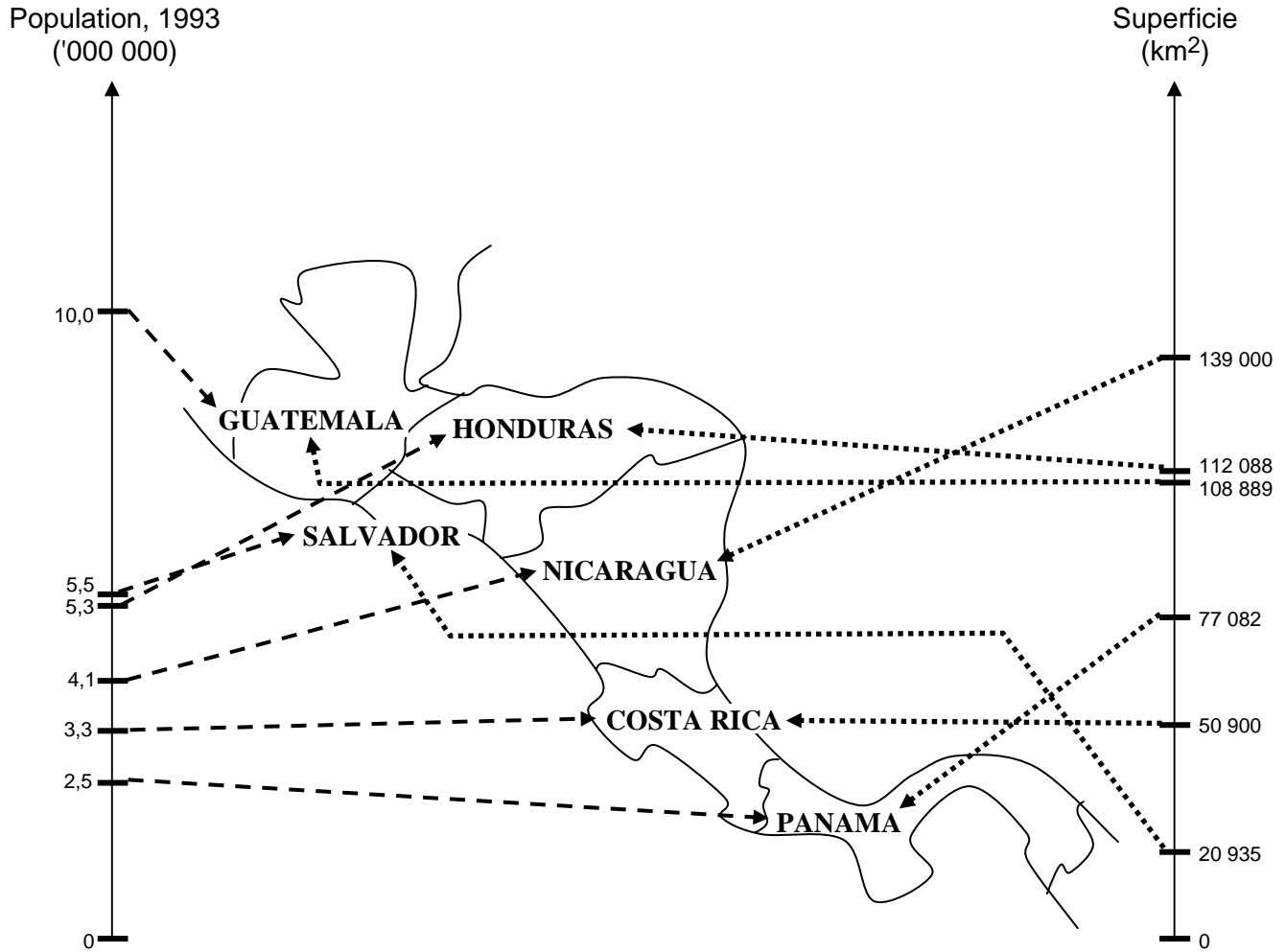
3. Variables *d'intervalle* : permettent de comparer les différences entre individus.

Définies «à une transformation linéaire près».

4. Variables *rationnelles* : il existe un zéro naturel et le rapport entre deux valeurs a un sens.



## QU'EST-CE QUE LA MESURE ? UNE MESURE EST UNE CORRESPONDANCE...



Source des données : Facultad Latino Americana de Ciencias Sociales  
FLACSO, Sede Costa Rica, San José, Costa Rica

## PEUT-ON MESURER LA NATIONALITÉ ?

- Superficie  $\Rightarrow$  mesure permet de déterminer  $=$ ,  $\neq$ ,  $<$  et  $>$ .
- Mais la définition de la mesure n'exige pas que l'on puisse déterminer la valeur de vérité des *quatre* relations.
- Exemple : nationalité
  - $f(X) = 0$  si la personne  $X$  est de nationalité costaricaine ;
  - $f(X) = 1$  si la personne  $X$  est d'une autre nationalité centraméricaine ;
  - $f(X) = 2$  dans tous les autres cas.

Alors

$f(A) = f(B) \Rightarrow A$  et  $B$  sont de même nationalité (dans la classification retenue)

$f(A) \neq f(B) \Rightarrow A$  et  $B$  ne sont pas de même nationalité.

$f(A) < f(B)$  et  $f(A) > f(B)$  n'ont aucune signification.

La correspondance constitue néanmoins une mesure au sens large

En un certain sens, donc, on peut mesurer des propriétés qualitatives.

Note : Les valeurs numériques de la correspondance n'ont aucune signification et elles sont parfaitement arbitraires. On pourrait même définir la correspondance en termes de symboles autres que des nombres. Par exemple, on aurait pu définir

$f(X) = \text{'CR'}$  si la personne  $X$  est de nationalité costaricaine ;

$f(X) = \text{'CA'}$  si la personne est d'une autre nationalité centraméricaine ;

$f(X) = \text{'OT'}$  dans tous les autres cas.

## TYPES DE DONNÉES

Chaque type de données comporte ses difficultés quant au contrôle de la qualité.

**Données primaires**

**Données secondaires non publiées**

**Données secondaires publiées**

Si vous ne trouvez pas d'erreur dans les données, c'est parce que vous ne cherchez pas bien ...

## STRUCTURE DES DONNÉES (1)

Structure fondamentale : matrice ou tableau

		<b>Variables</b>				
		(indicateurs, caractéristiques, propriétés, attributs, descripteurs...)				
		$X_1$	$X_2$	$X_3$	...	$X_k$
<b>Observations</b> (cas, individus, objets)	1	$x_{11}$	$x_{12}$	$x_{13}$	...	$x_{1k}$
	2	$x_{21}$	$x_{22}$	$x_{23}$	...	$x_{2k}$
	3	$x_{31}$	$x_{32}$	$x_{33}$	...	$x_{3k}$
	4	$x_{41}$	$x_{42}$	$x_{43}$	...	$x_{4k}$
	...	...	...	...	...	...
	$n$	$x_{n1}$	$x_{n2}$	$x_{n3}$	...	$x_{nk}$

Les observations se rapportent à des moments ou périodes successives

→ Séries chronologiques ou temporelles

Les observations se rapportent à différents lieux d'un ensemble géographique donné  
(pays d'un continent, villes ou régions d'un pays, quartiers d'une ville...)

→ Séries « spatiales » ?

Une ou plusieurs variables permettant de situer chaque observation dans l'espace géographique

→ Données géoréférencées

Observations classifiées selon une ou plusieurs variables catégoriques

→ Possibilité de structure matricielle à plus de 2 dimensions

## STRUCTURE DES DONNÉES : POINTS DE VUE HORIZONTAL ET VERTICAL

---

### ***Point de vue «horizontal» : entre les variables***

- Combiner plusieurs variables en une seule, qui les résume : construction de nombres indices
- Comparer deux variables : mesure de la similarité/dissimilarité
- Étudier les relations de dépendance
  - entre deux variables : corrélation, régression simple
  - entre une variable dépendante et plusieurs variables indépendantes : régression multiple et autres méthodes multivariées
  - entre plusieurs variables parmi lesquelles on ne distingue pas de variable dépendante : méthodes multivariées

### ***Point de vue «vertical» : entre les observations ou objets***

- Caractériser la distribution d'une variable : mesure de l'inégalité ou de la disparité, méthodes statistiques univariées
  - Lorsqu'il existe un ordre naturel entre les observations, étudier les relations entre les différentes observations d'une même variable : mesure et modélisation de l'évolution des séries temporelles, analyse de l'autocorrélation (temporelle, spatiale)
  - Comparer deux objets : mesure de la similarité/dissimilarité
-

## STRUCTURE DES DONNÉES (2)

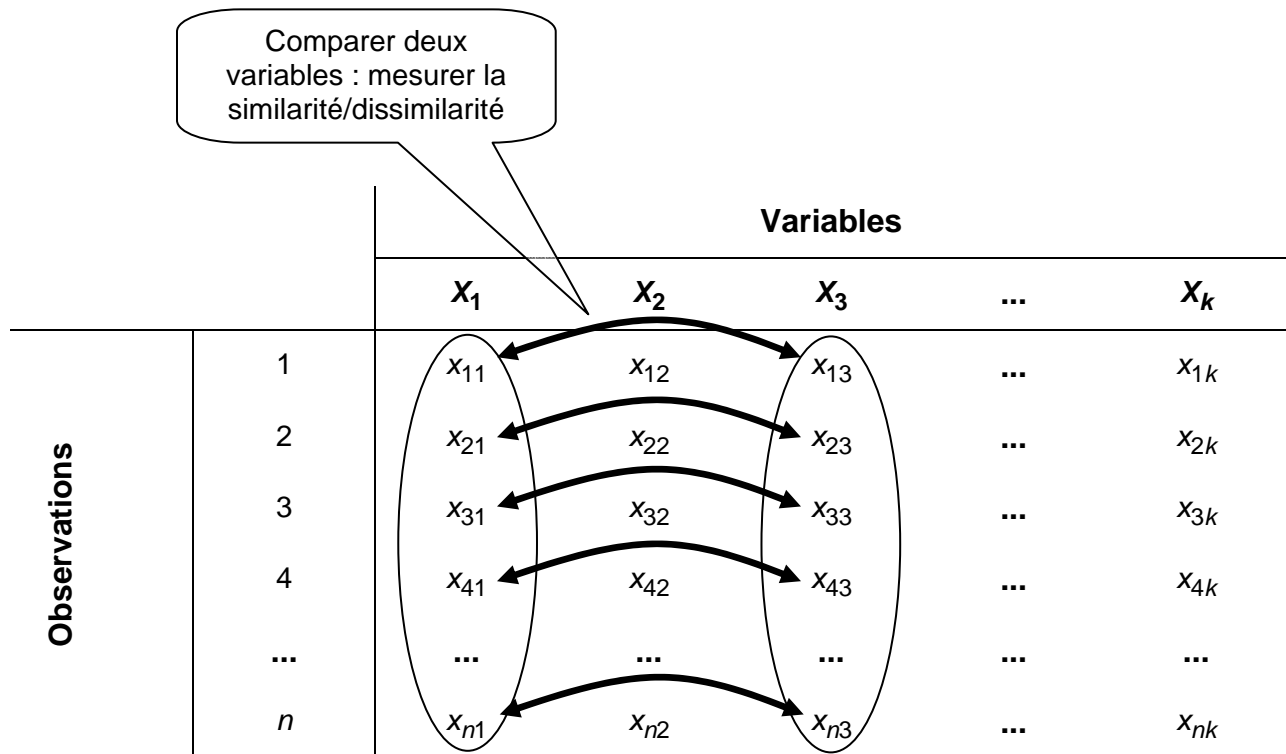
### POINT DE VUE HORIZONTAL : NOMBRES INDICES

Combiner plusieurs variables en une seule, qui les résume : construction de nombres indices

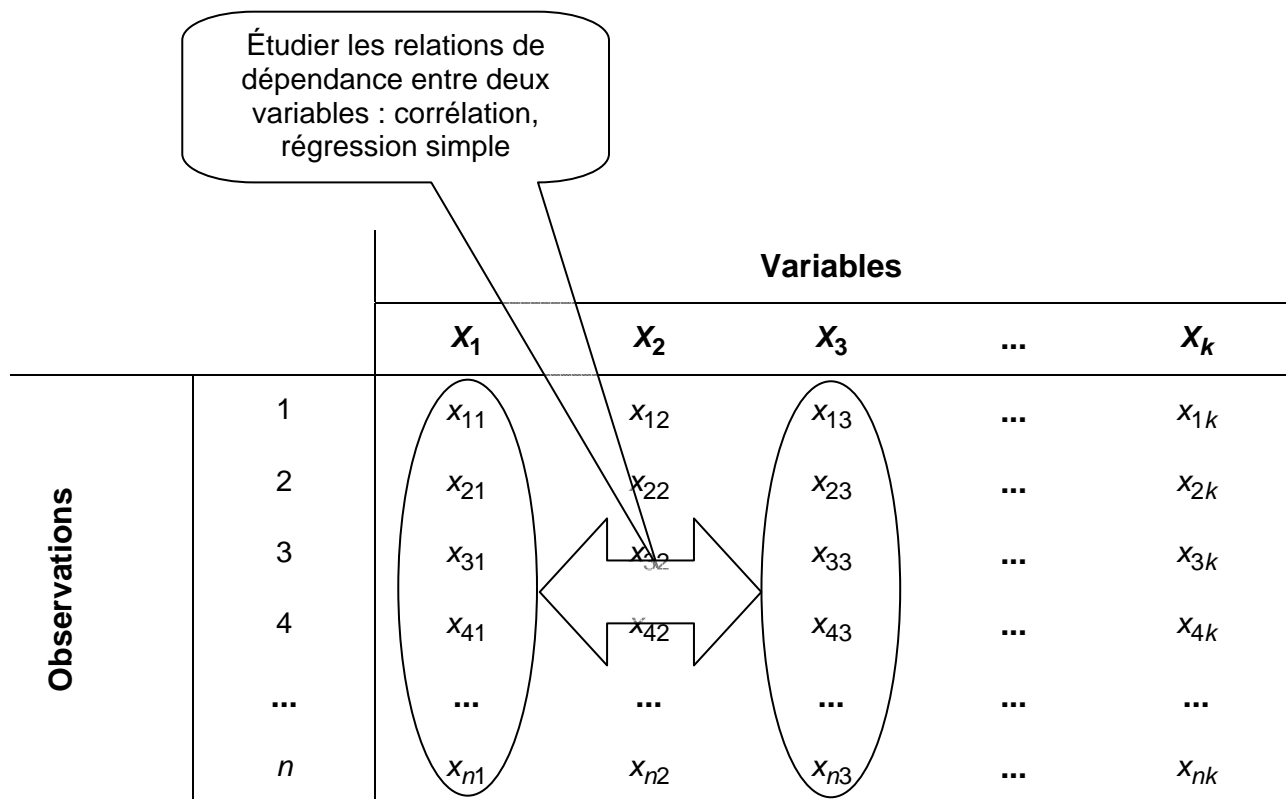
		Variables				
		$X_1$	$X_2$	$X_3$	...	$X_k$
Observations	1	$x_{11}$	$x_{12}$	$x_{13}$	...	$x_{1k}$
	2	$x_{21}$	$x_{22}$	$x_{23}$	...	$x_{2k}$
	3	$x_{31}$	$x_{32}$	$x_{33}$	...	$x_{3k}$
	4	$x_{41}$	$x_{42}$	$x_{43}$	...	$x_{4k}$
	...	...	...	...	...	...
	$n$	$x_{n1}$	$x_{n2}$	$x_{n3}$	...	$x_{nk}$

### STRUCTURE DES DONNÉES (3)

#### POINT DE VUE HORIZONTAL : SIMILARITÉ/DISSIMILARITÉ



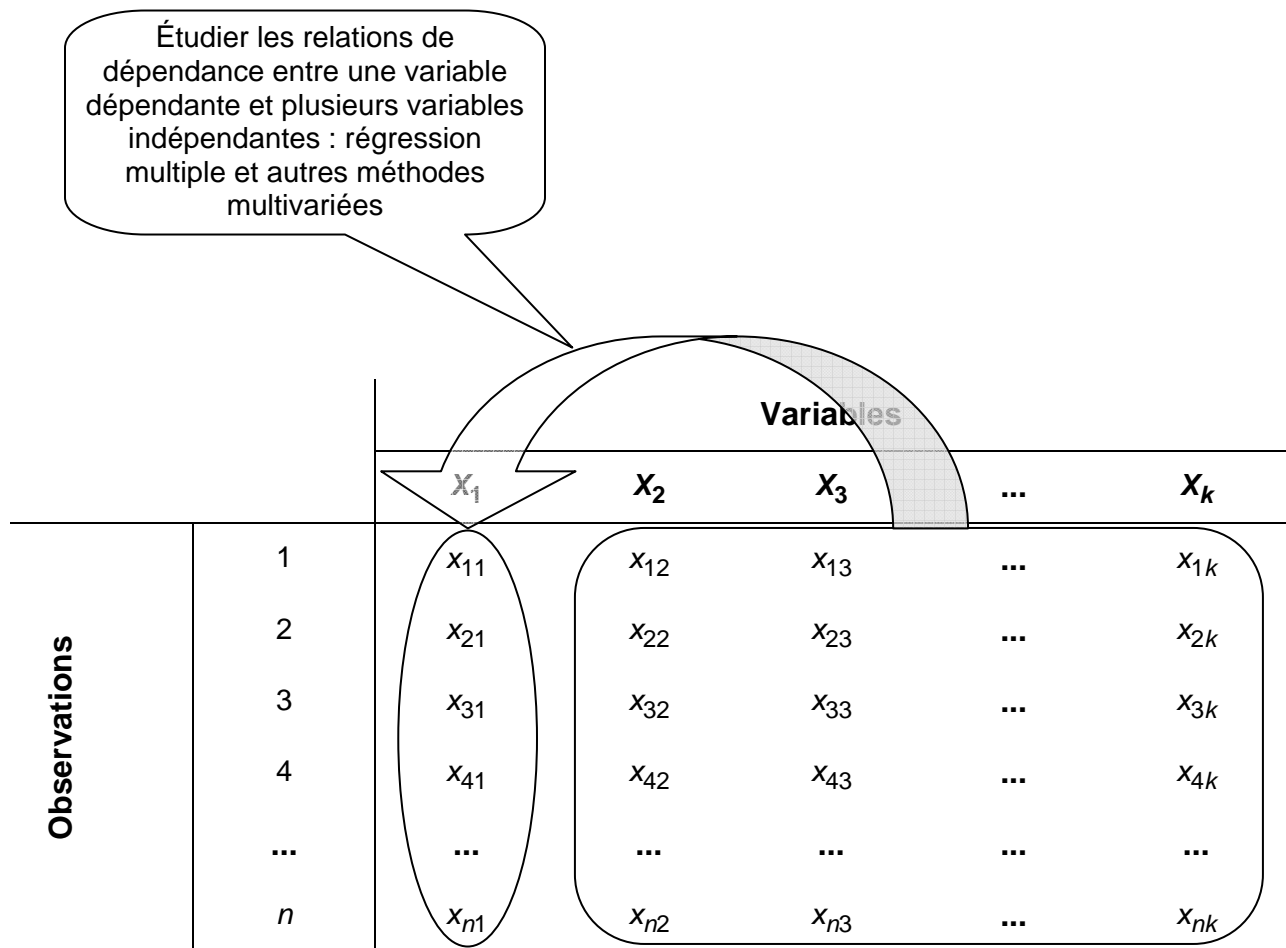
## STRUCTURE DES DONNÉES (4) POINT DE VUE HORIZONTAL : DÉPENDANCE ENTRE DEUX VARIABLES





## STRUCTURE DES DONNÉES (5)

### POINT DE VUE HORIZONTAL : DÉPENDANCE ENTRE UNE VARIABLE DÉPENDANTE ET PLUSIEURS VARIABLES INDÉPENDANTES



## STRUCTURE DES DONNÉES (6)

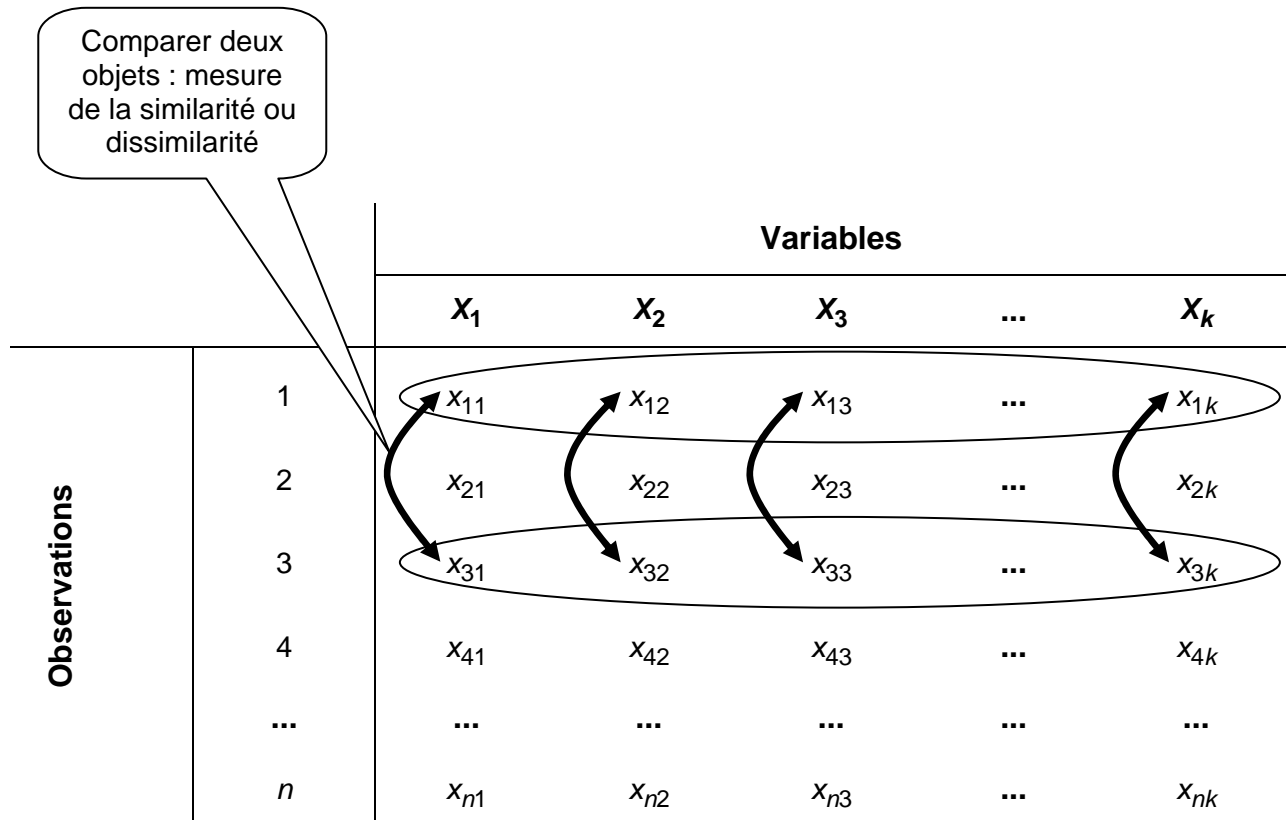
### POINT DE VUE VERTICAL : INÉGALITÉ, DISTRIBUTION

- Caractériser la distribution
  - Mesurer l'inégalité
- S'il existe un ordre naturel des observations :
- Analyser des séries temporelles
  - Analyser l'autocorrélation temporelle ou spatiale

		Variables				
		$X_1$	$X_2$	$X_3$	...	$X_k$
Observations	1	$x_{11}$	$x_{12}$	$x_{13}$	...	$x_{1k}$
	2	$x_{21}$	$x_{22}$	$x_{23}$	...	$x_{2k}$
	3	$x_{31}$	$x_{32}$	$x_{33}$	...	$x_{3k}$
	4	$x_{41}$	$x_{42}$	$x_{43}$	...	$x_{4k}$
	...	...	...	...	...	...
	$n$	$x_{n1}$	$x_{n2}$	$x_{n3}$	...	$x_{nk}$

## STRUCTURE DES DONNÉES (7)

### POINT DE VUE VERTICAL : SIMILARITÉ/DISSIMILARITÉ (BIS)



## INDICATEURS DE SPÉCIFICITÉ (QUOTIENTS DE LOCALISATION)

### EXEMPLE NUMÉRIQUE

#### Emploi par zone et par branche

BRANCHE	B1	B2	B3	Total
ZONE				
Z1	48	325	287	<b>660</b>
Z2	27	185	148	<b>360</b>
Z3	45	90	45	<b>180</b>
<b>Total</b>	<b>120</b>	<b>600</b>	<b>480</b>	<b>1200</b>

#### Distribution de l'emploi entre zones

BRANCHE	B1	B2	B3	Total	BRANCHE	B1	B2	B3	Total
ZONE					ZONE				
Z1	0,400	0,542	0,598	0,550	Z1	40,0 %	54,2 %	59,8 %	55,0 %
Z2	0,225	0,308	0,308	0,300	Z2	22,5 %	30,8 %	30,8 %	30,0 %
Z3	0,375	0,150	0,094	0,150	Z3	37,5 %	15,0 %	9,4 %	15,0 %
<b>Total</b>	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>	<b>1,000</b>	<b>Total</b>	<b>100,0 %</b>	<b>100,0 %</b>	<b>100,0 %</b>	<b>100,0 %</b>

#### Distribution de l'emploi entre branches

BRANCHE	B1	B2	B3	Total	BRANCHE	B1	B2	B3	Total
ZONE					ZONE				
Z1	0,073	0,492	0,435	<b>1,000</b>	Z1	7,3 %	49,2 %	43,5 %	<b>100,0 %</b>
Z2	0,075	0,514	0,411	<b>1,000</b>	Z2	7,5 %	51,4 %	41,1 %	<b>100,0 %</b>
Z3	0,250	0,500	0,250	<b>1,000</b>	Z3	25,0 %	50,0 %	25,0 %	<b>100,0 %</b>
<b>Total</b>	<b>0,100</b>	<b>0,500</b>	<b>0,400</b>	<b>1,000</b>	<b>Total</b>	<b>10,0 %</b>	<b>50,0 %</b>	<b>40,0 %</b>	<b>100,0 %</b>

#### Quotients de localisation

BRANCHE	B1	B2	B3
ZONE			
Z1	0,727	0,985	1,087
Z2	0,750	1,028	1,028
Z3	2,500	1,000	0,625

## INDICATEURS DE SPÉCIFICITÉ (QUOTIENTS DE LOCALISATION) MÉTHODE DE CALCUL

**Exemple de calcul : deux méthodes équivalentes**

$$QL_{21} = \frac{x_{21} / x_{\cdot 1}}{x_{2\cdot} / x_{\cdot\cdot}} = \frac{x_{21} x_{\cdot\cdot}}{x_{2\cdot} x_{\cdot 1}} \quad \text{ou} \quad QL_{21} = \frac{x_{21} / x_{2\cdot}}{x_{\cdot 1} / x_{\cdot\cdot}} = \frac{x_{21} x_{\cdot\cdot}}{x_{2\cdot} x_{\cdot 1}}$$

$$QL_{21} = \frac{27 / 120}{360 / 1200} = \frac{0,225}{0,300} = 0,75 \quad \text{ou} \quad QL_{21} = \frac{27 / 360}{120 / 1200} = \frac{0,075}{0,100} = 0,75$$

**Formule générale**

$x_{ij}$	nombre d'emplois de la branche $j$ dans la zone $i$
$x_{\cdot j} = \sum_i x_{ij}$	nombre total d'emplois de la branche $j$
$x_{i\cdot} = \sum_j x_{ij}$	nombre total d'emplois dans la zone $i$
$x_{\cdot\cdot} = \sum_i \sum_j x_{ij}$	nombre total d'emplois de toutes les branches dans toutes les zones

$$QL_{ij} = \frac{\text{Part de la zone } i \text{ dans l'emploi de la branche } j}{\text{Part de la zone } i \text{ dans l'emploi total}} = \frac{x_{ij} / x_{\cdot j}}{x_{i\cdot} / x_{\cdot\cdot}} = \frac{x_{ij} x_{\cdot\cdot}}{x_{i\cdot} x_{\cdot j}}$$

$$QL_{ij} = \frac{\text{Part de la branche } j \text{ dans l'emploi de la zone } i}{\text{Part de la branche } j \text{ dans l'emploi total}} = \frac{x_{ij} / x_{i\cdot}}{x_{\cdot j} / x_{\cdot\cdot}} = \frac{x_{ij} x_{\cdot\cdot}}{x_{i\cdot} x_{\cdot j}}$$

**Ce n'est pas par accident que les deux calculs donnent le même résultat !**

## INDICATEURS DE SPÉCIFICITÉ (QUOTIENTS DE LOCALISATION) PROPRIÉTÉS ET INTERPRÉTATION

### Domaine de variation

#### LA PLUS PETITE VALEUR POSSIBLE

Lorsque  $x_{ij} = 0$ ,

$$\rightarrow QL_{ij} = 0$$

#### LA PLUS GRANDE VALEUR POSSIBLE

Lorsque  $x_{ij}$  est la seule valeur non nulle dans sa ligne et dans sa colonne :

$$x_{i\bullet} = x_{\bullet j} = x_{ij}$$

$$\rightarrow QL_{ij} = \frac{x_{\bullet\bullet}}{x_{ij}}$$

Cette valeur n'a pas de limite théorique.

NOTE : il est mathématiquement impossible que tous les  $QL$  d'une même ligne ou d'une même colonne soient tous  $> 1$  ou qu'ils soient tous  $< 1$ .

### Forme normalisée

$$\frac{QL_{ij} - 1}{QL_{ij} + 1} \text{ varie de } -1 \text{ à } +1$$

### Interprétation

POINT DE REPÈRE  $QL_{ij} = 1$

DEUX MÉTHODES DE CALCUL, DEUX « LECTURES » : SI  $QL_{ij} > 1$ ...

- l'activité  $j$  est *relativement* concentrée dans la zone  $i$  parce que la fraction de l'emploi qui est situé dans la zone  $i$  est *plus* importante pour l'activité  $j$  que pour les autres activités.
- la zone  $i$  est *relativement* spécialisée dans l'activité  $j$  parce que l'activité  $j$  occupe dans la zone  $i$  une place *plus* importante qu'ailleurs.

## INDICATEURS DE SPÉCIFICITÉ (QUOTIENTS DE LOCALISATION) DES INDICES DE CONCENTRATION *RELATIVE*

### L'importance de « relativement »

Si on l'oublie, on peut se tromper lourdement !

### Un exemple fictif

**Emploi par zone et par branche**

BRANCHE	B1	B2	B3	Total
ZONE				
Z1	50	250	250	550
Z2	20	400	350	770
Z3	5	100	45	150
<b>Total</b>	<b>75</b>	<b>750</b>	<b>645</b>	<b>1470</b>

**Quotients de localisation**

BRANCHE	B1	B2	B3
ZONE			
Z1	1,782	0,891	1,036
Z2	0,509	1,018	1,036
Z3	0,653	1,307	0,684

Dans cet exemple fictif, peut-on dire que...

- La zone Z1 est spécialisée dans la branche B1 ?
- La branche B2 est concentrée dans la zone Z3 ?

Oui, mais **relativement** ! Parce que...

- La branche B1 n'est **pas** celle qui compte **le plus grand nombre** d'emplois dans la zone Z1 (au contraire, dans cet exemple particulier, elle est celle qui en compte le plus petit nombre) :  
il est donc **faux** de dire que la zone Z1 est spécialisée dans l'activité B1 en termes **absolus**.
- La zone Z3 n'est **pas** celle qui contient **le plus grand nombre** d'emplois de la branche B2 (au contraire, dans cet exemple, elle est celle qui en contient le plus petit nombre) :  
il est donc **faux** de dire que la branche B2 est concentrée dans la zone Z3 en termes **absolus**.

## INDICATEURS DE SPÉCIFICITÉ (QUOTIENTS DE LOCALISATION) ET PROBABILITÉS DANS UN TABLEAU DE CONTINGENCE

$x_{ij}$	nombre d'emplois de la branche $j$ dans la zone $i$
$x_{\bullet j} = \sum_i x_{ij}$	nombre total d'emplois de la branche $j$
$x_{i\bullet} = \sum_j x_{ij}$	nombre total d'emplois dans la zone $i$
$x_{\bullet\bullet} = \sum_i \sum_j x_{ij}$	nombre total d'emplois de toutes les branches dans toutes les zones
$p_{ij} = \frac{x_{ij}}{x_{\bullet\bullet}}$	fraction de l'emploi total global qui appartient à la branche $j$ et qui se trouve dans la zone $i$
$p_{\bullet j} = \sum_i p_{ij}$	fraction de l'emploi total global qui appartient à la branche $j$
$p_{i\bullet} = \sum_j p_{ij}$	fraction de l'emploi total global qui se trouve dans la zone $i$
$p_{j/i\bullet} = \frac{p_{ij}}{p_{i\bullet}}$	fraction de l'emploi total dans la zone $i$ qui appartient à la branche $j$
$p_{i/\bullet j} = \frac{p_{ij}}{p_{\bullet j}}$	fraction de l'emploi total de la branche $j$ qui se trouve dans la zone $i$

**Identités :**

$$p_{\bullet j} = \sum_i p_{ij} = \sum_i \frac{x_{ij}}{x_{\bullet\bullet}} = \frac{x_{\bullet j}}{x_{\bullet\bullet}} \quad \text{et} \quad p_{i\bullet} = \sum_j p_{ij} = \sum_j \frac{x_{ij}}{x_{\bullet\bullet}} = \frac{x_{i\bullet}}{x_{\bullet\bullet}}$$

$$p_{j/i\bullet} = \frac{p_{ij}}{p_{i\bullet}} = \frac{x_{ij}/x_{\bullet\bullet}}{x_{i\bullet}/x_{\bullet\bullet}} = \frac{x_{ij}}{x_{i\bullet}} \quad \text{et} \quad p_{i/\bullet j} = \frac{p_{ij}}{p_{\bullet j}} = \frac{x_{ij}/x_{\bullet\bullet}}{x_{\bullet j}/x_{\bullet\bullet}} = \frac{x_{ij}}{x_{\bullet j}}$$

$$\sum_i \sum_j p_{ij} = \sum_i p_{i\bullet} = \sum_j p_{\bullet j} = 1 ; \quad \text{aussi} \quad \sum_j p_{j/i\bullet} = \frac{\sum_j p_{ij}}{p_{i\bullet}} = 1 \quad \text{et} \quad \sum_i p_{i/\bullet j} = \frac{\sum_i p_{ij}}{p_{\bullet j}} = 1$$

**Quotient de localisation :**

$$QL_{ij} = \frac{\text{Part de la zone } i \text{ dans l'emploi de la branche } j}{\text{Part de la zone } i \text{ dans l'emploi total}} = \frac{p_{i/\bullet j}}{p_{i\bullet}} = \frac{\frac{x_{ij}}{x_{\bullet j}}}{\frac{x_{i\bullet}}{x_{\bullet\bullet}}} = \frac{x_{ij} x_{\bullet\bullet}}{x_{i\bullet} x_{\bullet j}}$$

$$QL_{ij} = \frac{\text{Part de la branche } j \text{ dans l'emploi de la zone } i}{\text{Part de la branche } j \text{ dans l'emploi total}} = \frac{p_{j/i\bullet}}{p_{\bullet j}} = \frac{\frac{x_{ij}}{x_{i\bullet}}}{\frac{x_{\bullet j}}{x_{\bullet\bullet}}} = \frac{x_{ij} x_{\bullet\bullet}}{x_{i\bullet} x_{\bullet j}}$$

On peut voir l'indicateur de spécificité comme le rapport d'une probabilité conditionnelle sur une probabilité marginale.



## ESTIMATION DE L'EMPLOI EXPORTATEUR AU MOYEN DU QUOTIENT DE LOCALISATION

### Hypothèses

1. La productivité du travail est égale entre villes ou régions ;
2. L'absorption (utilisation locale) du produit par emploi dans l'économie locale est égale entre villes ou régions ;
3. Il n'y a pas d'importations ou d'exportations nettes de l'ensemble du pays ;
4. La demande locale s'approvisionne en priorité auprès des producteurs locaux ; cela implique qu'il n'y a pas de flux croisés entre villes ou régions («cross-hauling»).

**Estimation de l'emploi exportateur de la branche  $j$  dans la région  $i$  :**

$$EXP_{ij} = \begin{cases} x_{ij} \frac{(QL_{ij} - 1)}{QL_{ij}}, & \text{si } QL_{ij} > 1 \\ 0 & \text{autrement} \end{cases}$$

ex. : pour la branche  $B1$  dans la zone  $Z3$ ,  $EXP_{31} = 45 \times \frac{2,5 - 1}{2,5} = 27$

**Développement de l'expression  $x_{ij} \frac{(QL_{ij} - 1)}{QL_{ij}}$**

$$1) \quad x_{ij} \frac{(QL_{ij} - 1)}{QL_{ij}} = x_{ij} \left( \frac{QL_{ij}}{QL_{ij}} - \frac{1}{QL_{ij}} \right)$$

$$2) \quad x_{ij} \frac{(QL_{ij} - 1)}{QL_{ij}} = x_{ij} - x_{ij} \frac{1}{\left( \frac{x_{ij} x_{..}}{x_{i.} x_{.j}} \right)}$$

$$3) \quad x_{ij} \frac{(QL_{ij} - 1)}{QL_{ij}} = x_{ij} - x_{ij} \frac{x_{i.} x_{.j}}{x_{ij} x_{..}}$$

$$4) \quad x_{ij} \frac{(QL_{ij} - 1)}{QL_{ij}} = x_{ij} - \frac{x_{i.} x_{.j}}{x_{..}}$$

$$5) \quad x_{ij} \frac{(QL_{ij} - 1)}{QL_{ij}} = x_{ij} \frac{x_{.j}}{x_{.j}} - \frac{x_{i.} x_{.j}}{x_{..}}$$

$$6) \quad x_{ij} \frac{(QL_{ij} - 1)}{QL_{ij}} = \left( \frac{x_{ij}}{x_{.j}} - \frac{x_{i.}}{x_{..}} \right) x_{.j}$$

**Base exportatrice =**  $\sum_{j, \text{ lorsque } QL_{ij} > 1} EXP_{ij}$

**Modèle de la base économique :** hypothèse que  $\theta = \frac{\sum_j x_{ij}}{\sum_j EXP_{ij}}$  est constant.

## ESTIMATION DE L'EMPLOI EXPORTATEUR AU MOYEN DU QUOTIENT DE LOCALISATION (SUITE)

$$EXP_{ij} = \begin{cases} x_{ij} \frac{(QL_{ij} - 1)}{QL_{ij}} = \left( \frac{x_{ij}}{x_{\bullet j}} - \frac{x_{i\bullet}}{x_{\bullet\bullet}} \right) x_{\bullet j}, & \text{si } QL_{ij} > 1 \\ 0 & \text{autrement} \end{cases}$$

### Rôle des hypothèses

**Hypothèse 1** – La productivité du travail est égale entre villes ou régions :

$$\frac{x_{ij}}{x_{\bullet j}} \quad (\text{la part de la région } i \text{ dans l'emploi de la branche } j)$$

= la part de la région  $i$  dans la **production** de la branche  $j$

**Hypothèse 2** – L'absorption (utilisation locale) du produit par emploi dans l'économie locale est égale entre villes ou régions :

$$\frac{x_{i\bullet}}{x_{\bullet\bullet}} \quad (\text{la part de la région } i \text{ dans l'emploi total}) = \text{la part de la région } i \text{ dans l'absorption}$$

**Hypothèse 3** – Il n'y a pas d'importations ou d'exportations nettes de l'ensemble du pays :

<i>Production</i> + <i>Importations des autres régions</i> + <i>Importations internationales</i>	=	<i>Absorption</i> + <i>Exportations aux autres régions</i> + <i>Exportations internationales</i>
--	---	--

devient

<i>Production</i> + <i>Importations des autres régions</i>	=	<i>Absorption</i> + <i>Exportations aux autres régions</i>
--	---	--

c'est-à-dire

<i>Production – Absorption</i>	=	<i>Exportations nettes aux autres régions</i>
--------------------------------	---	---

**Hypothèse 4** – Il n'y a pas de flux croisés entre villes ou régions («cross-hauling»). Alors, quand les exportations nettes sont positives,

<i>Exportations nettes aux autres régions</i>	=	<i>Exportations totales aux autres régions</i>
---	---	--

## L'ANALYSE DE DÉCOMPOSITION ADDITIVE ET MULTIPLICATIVE DES VARIATIONS

### Principe

#### *Décomposition additive :*

$$x - y = (x - a) + (a - b) + (b - c) + (c - y)$$

#### *Décomposition multiplicative :*

$$x/y = (x/a) (a/b) (b/c) (c/y)$$

c'est-à-dire

$$\log x - \log y = (\log x - \log a) + (\log a - \log b) + (\log b - \log c) + (\log c - \log y)$$

#### *Exemples*

- Analyse «shift-share»
- Williamson (1965)

## L'ANALYSE SHIFT-SHARE (EXEMPLE NUMÉRIQUE)

### Emploi par zone et par branche

BRANCHE	An 1				An 2			
	B1	B2	B3	Total	B1	B2	B3	Total
ZONE								
Z1	48	325	287	660	24	388	300	712
Z2	27	185	148	360	11	173	200	384
Z3	45	90	45	180	25	99	52	176
Total	120	600	480	1200	60	660	552	1272

### Variation de l'emploi par zone et par branche

BRANCHE	Différences				Taux			
	B1	B2	B3	Total	B1	B2	B3	Total
ZONE								
Z1	-24	63	13	52	-50,00%	19,38%	4,53%	7,88%
Z2	-16	-12	52	24	-59,26%	-6,49%	35,14%	6,67%
Z3	-20	9	7	-4	-44,44%	10,00%	15,56%	-2,22%
Total	-60	60	72	72	-50,00%	10,00%	15,00%	6,00%

## L'ANALYSE SHIFT-SHARE (EXEMPLE NUMÉRIQUE)

### Analyse de la variation de l'emploi de la branche B1 dans la zone Z2

#### Trois scénarios :

1. Quelle aurait été la variation si l'emploi de B1 en Z2 avait évolué au même taux que l'emploi total (toutes branches et toutes zones) ?
  - Taux = 6 %
  - Nombre = 6 % de 27 = 1,62
2. Quelle aurait été la variation si l'emploi de B1 en Z2 avait évolué au même taux que l'emploi de l'ensemble de la branche B1 ?
  - Taux = -50 %
  - Nombre = -50 % de 27 = -13,50
3. Quelle a été la variation observée de l'emploi de B1 en Z2 ?
  - Taux = -59,26 %
  - Nombre = -59,26 % de 27 = -16

#### Décomposition additive :

4. Effet national = scénario 1 :
  - Taux = 6 %
  - Nombre = 6 % de 27 = 1,62
5. Effet proportionnel (ou sectoriel) = écart entre scénario 2 et scénario 1 :
  - Taux = -50 % - 6 % = -56 %
  - Nombre = -56 % de 27 = -15,12 = -13,5 - 1,62
6. Effet résiduel (ou régional) = écart entre scénario 3 et scénario 2 :
  - Taux = -59,26 % - (-50 %) = -9,26 %
  - Nombre = -9,26 % de 27 = -2,5 = -16 - (-13,5)

#### Vérification de la somme :

- Taux = 6 % + (-56 %) + (-9,26 %) = -59,26 %
- Nombre = 1,62 + (-15,12) + (-2,5) = -16

## L'ANALYSE SHIFT-SHARE (EXEMPLE NUMÉRIQUE)

### Analyse shift-share par branche

#### Branche B1

ZONE	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	2,88	-26,88	0	-24	6,00%	-56,00%	0,00%	-50,00%
Z2	1,62	-15,12	-2,5	-16	6,00%	-56,00%	-9,26%	-59,26%
Z3	2,7	-25,2	2,5	-20	6,00%	-56,00%	5,56%	-44,44%

#### Branche B2

ZONE	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	19,5	13	30,5	63	6,00%	4,00%	9,38%	19,38%
Z2	11,1	7,4	-30,5	-12	6,00%	4,00%	-16,49%	-6,49%
Z3	5,4	3,6	0,0	9	6,00%	4,00%	0,00%	10,00%

#### Branche B3

ZONE	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	17,22	25,83	-30,05	13	6,00%	9,00%	-10,47%	4,53%
Z2	8,88	13,32	29,8	52	6,00%	9,00%	20,14%	35,14%
Z3	2,7	4,05	0,25	7	6,00%	9,00%	0,56%	15,56%

#### Ensemble des branches (classification à trois branches)

ZONE	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	39,6	11,95	0,45	52	6,00%	1,81%	0,07%	7,88%
Z2	21,6	5,6	-3,2	24	6,00%	1,56%	-0,89%	6,67%
Z3	10,8	-17,55	2,75	-4	6,00%	-9,75%	1,53%	-2,22%

## L'ANALYSE SHIFT-SHARE (EXEMPLE NUMÉRIQUE)

### Effet de l'agrégation

#### Branches B1 et B2 agrégées

ZONE	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	22,38	-22,38	39	39	6,00%	-6,00%	10,46%	10,46%
Z2	12,72	-12,72	-28	-28	6,00%	-6,00%	-13,21%	-13,21%
Z3	8,1	-8,1	-11	-11	6,00%	-6,00%	-8,15%	-8,15%

#### Ensemble des branches (classification à deux branches)

ZONE	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	39,6	<b>3,45</b>	<b>8,95</b>	52	6,00%	<b>0,52%</b>	<b>1,36%</b>	7,88%
Z2	21,6	<b>0,6</b>	<b>1,8</b>	24	6,00%	<b>0,17%</b>	<b>0,50%</b>	6,67%
Z3	10,8	<b>-4,05</b>	<b>-10,75</b>	-4	6,00%	<b>-2,25%</b>	<b>-5,97%</b>	-2,22%

#### Ensemble des branches (classification à trois branches)

ZONE	Nombre d'emplois				Taux de variation			
	Effet national	Effet sectoriel	Effet résiduel	Effet total	Effet national	Effet sectoriel	Effet résiduel	Effet total
Z1	39,6	<b>11,95</b>	<b>0,45</b>	52	6,00%	<b>1,81%</b>	<b>0,07%</b>	7,88%
Z2	21,6	<b>5,6</b>	<b>-3,2</b>	24	6,00%	<b>1,56%</b>	<b>-0,89%</b>	6,67%
Z3	10,8	<b>-17,55</b>	<b>2,75</b>	-4	6,00%	<b>-9,75%</b>	<b>1,53%</b>	-2,22%

## LA MESURE DE LA CROISSANCE (EXEMPLE NUMÉRIQUE)

### Calcul du taux de croissance par période

$$r_t = \frac{x_t - x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - \frac{x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - 1$$

Année	IPC	Taux de croissance par période
1984	92,4	
1985	96,0	0,039
1986	100,0	0,042
1987	104,4	0,044
1988	108,6	0,040
1989	114,0	0,050
1990	119,5	0,048
1991	126,2	0,056
1992	128,1	0,015

Par exemple, pour  $t = 1985$ ,

$$r_{1985} = \frac{96,0}{92,4} - 1 = 0,039 = 3,9 \%$$

Si l'on connaît les  $r_t$  et  $x_0$ , on peut reconstituer la série des  $x_t$  :

$$x_t = (1+r_t) x_{t-1} \text{ et } x_{t-1} = (1+r_{t-1}) x_{t-2} \quad \Rightarrow \quad x_t = (1+r_t) (1+r_{t-1}) x_{t-2}$$

... par substitutions successives,

$$x_t = (1+r_t) (1+r_{t-1}) \dots (1+r_2) (1+r_1) x_0$$

Exemple :

$$x_{1987} = (1+0,044) \times (1+0,042) \times (1+0,039) \times 92,4 = 104,4$$

### Moyenne arithmétique des taux de croissance par période

$$\frac{r_1 + r_2 + \dots + r_t + \dots + r_T}{T} = \frac{1}{T} \sum_{t=1}^T r_t$$

### Taux de croissance exponentiel

= Moyenne géométrique des facteurs de croissance périodiques :

$$1 + r = \left[ (1+r_T) (1+r_{T-1}) \dots (1+r_2) (1+r_1) \right]^{\frac{1}{T}} = \sqrt[T]{(1+r_T) (1+r_{T-1}) \dots (1+r_2) (1+r_1)}$$

$$\log(1+r) = \frac{1}{T} \sum_{t=1}^T \log(1+r_t)$$



## CROISSANCE CUMULÉE ET VARIABILITÉ DES TAUX DE CROISSANCE PAR PÉRIODE

Période	Série 1		Série 2	
	Taux	Valeur	Taux	Valeur
0		100		100
1	0,10	110	0,00	100
2	0,10	121	0,20	120
Taux moyen	0,10		0,10	

## LA MESURE DE LA CROISSANCE

### CALCUL PRATIQUE DU TAUX DE CROISSANCE EXPONENTIEL

$$x_T = (1+r_T) (1+r_{T-1}) \dots (1+r_2) (1+r_1) x_0$$

$$(1+r_T) (1+r_{T-1}) \dots (1+r_2) (1+r_1) = (1+r)^T$$

$$x_T = (1+r)^T x_0$$

$$(1+r)^T = \frac{x_T}{x_0}$$

$$\log(1+r)^T = \log\left(\frac{x_T}{x_0}\right)$$

$$T \log(1+r) = \log(x_T) - \log(x_0)$$

$$\log(1+r) = \frac{\log(x_T) - \log(x_0)}{T}$$

$$(1+r) = \text{antilog}\left[\frac{\log(x_T) - \log(x_0)}{T}\right]$$

Avec les logarithmes communs,

$$r = 10^{\frac{\log x_T - \log x_0}{T}} - 1$$

Avec les logarithmes népériens.

$$r = e^{\frac{\ln x_T - \ln x_0}{T}} - 1 = \exp\left(\frac{\ln x_T - \ln x_0}{T}\right) - 1$$

## STRUCTURE DES DONNÉES

### POINT DE VUE HORIZONTAL : NOMBRES INDICES

Combiner plusieurs variables en une seule, qui les résume :  
construction de nombres indices

		Variables				
		$x_1$	$x_2$	$x_3$	...	$x_k$
Observations	1	$x_{11}$	$x_{12}$	$x_{13}$	...	$x_{1k}$
	2	$x_{21}$	$x_{22}$	$x_{23}$	...	$x_{2k}$
	3	$x_{31}$	$x_{32}$	$x_{33}$	...	$x_{3k}$
	4	$x_{41}$	$x_{42}$	$x_{43}$	...	$x_{4k}$
	...	...	...	...	...	...
	$n$	$x_{n1}$	$x_{n2}$	$x_{n3}$	...	$x_{nk}$

## LES INDICES DE PRIX (EXEMPLE NUMÉRIQUE – LASPEYRES)

### Indices de prix individuels

Biens	Prix (\$)		Indices des prix des biens individuels
	1980	1985	
	$p_{0i}$	$p_{ti}$	$p_{ti} / p_{0i}$
Steak (kg)	4,85 \$	6,60 \$	1,36
Poivre (g)	0,07 \$	0,07 \$	1,00
Pain (kg)	1,10 \$	1,32 \$	1,20

### Calcul de l'indice de prix de Laspeyres (2 méthodes)

Première méthode	Biens	Données				Calculs	
		Prix (\$)		Quantités		Coût du panier	
		1980	1985	1980	1985	1980	1985
		$p_{0i}$	$p_{ti}$	$q_{0i}$	$q_{ti}$	$p_{0i} q_{0i}$	$p_{ti} q_{0i}$
	Steak (kg)	4,85 \$	6,60 \$	23	18	111,55 \$	151,80 \$
	Poivre (g)	0,07 \$	0,07 \$	57	85	3,99 \$	3,99 \$
	Pain (kg)	1,10 \$	1,32 \$	36	45	39,60 \$	47,52 \$
						155,14 \$	203,31 \$

$$\text{Indice de Laspeyres} = \frac{203,31}{155,14} = 1,310$$

Deuxième méthode	Biens	Données				Indice de prix de Laspeyres		
		Prix			Quant.	Coût du panier	Poids	Calcul de l'indice
		1980	1985	Rapport	1980	1980	1980	1985
		$p_{0i}$	$p_{ti}$	$\left(\frac{p_{ti}}{p_{0i}}\right)$	$q_{0i}$	$p_{0i} q_{0i}$	$w_{0i}$	$w_{0i} \left(\frac{p_{ti}}{p_{0i}}\right)$
	Steak (kg)	4,85 \$	6,60 \$	1,36	23	111,55 \$	0,719	0,978
	Poivre (g)	0,07 \$	0,07 \$	1,00	57	3,99 \$	0,026	0,026
	Pain (kg)	1,10 \$	1,32 \$	1,20	36	39,60 \$	0,255	0,306
						155,14 \$	1,000	1,310

## QU'EST-CE QU'UNE MOYENNE PONDÉRÉE ?

Exemple : quel est le PIB per capita moyen au sein de l'ALENA ?

### Données

	<b>PIB per capita 2000 \$U.S. PPA</b>	<b>Population (millions)</b>	<b>Poids selon la population</b>	<b>Calcul de la moyenne pondérée</b>
États-Unis	34 142	283,2	68,6%	23 417
Canada	27 840	30,8	7,5%	2 077
Mexique	9 023	98,9	24,0%	2 161
<b>Total</b>		<b>412,9</b>	<b>100,0%</b>	<b>27 655</b>

Moyenne 23 668

Source : PNUD, *Rapport mondial sur le développement humain 2002*.

### Calcul de la moyenne simple

$$23\,668 = \frac{34\,142 + 27\,840 + 9\,023}{3} = \frac{1}{3} 34\,142 + \frac{1}{3} 27\,840 + \frac{1}{3} 9\,023$$

### Calcul de la moyenne pondérée selon la population

#### POIDS

$$0,686 = \frac{283,2}{412,9} ; 0,075 = \frac{30,8}{412,9} ; 0,240 = \frac{98,9}{412,9}$$

#### MOYENNE PONDÉRÉE

$$27\,655 = (0,686 \times 34\,142) + (0,075 \times 27\,840) + (0,240 \times 9\,023)$$

## LES INDICES DE PRIX

### Notation :

- $p_{ti}$  prix du bien  $i$  à la période  $t$
- $p_{0i}$  prix du bien  $i$  à la période 0
- $q_{0i}$  quantité du bien  $i$  achetée par un ménage typique à la période 0

### Indice de Laspeyres

#### Définition :

$$\text{Coût du panier de référence à la période 0} = p_{01}q_{01} + p_{02}q_{02} + \dots + p_{0n}q_{0n} = \sum_{i=1}^n p_{0i}q_{0i}$$

$$\text{Coût du panier de référence à la période } t = p_{t1}q_{01} + p_{t2}q_{02} + \dots + p_{tn}q_{0n} = \sum_{i=1}^n p_{ti}q_{0i}$$

$$I_t^L = \frac{\sum_{i=1}^n p_{ti}q_{0i}}{\sum_{i=1}^n p_{0i}q_{0i}}$$

#### Transformation et interprétation :

$$I_t^L = \frac{\sum_{i=1}^n p_{ti}q_{0i}}{\sum_{i=1}^n p_{0i}q_{0i}} = \frac{\sum_{i=1}^n p_{ti}q_{0i}}{\sum_{k=1}^n p_{0k}q_{0k}} = \sum_{i=1}^n \left( \frac{p_{ti}q_{0i}}{\sum_{k=1}^n p_{0k}q_{0k}} \right) = \sum_{i=1}^n \left( \frac{q_{0i}}{\sum_{k=1}^n p_{0k}q_{0k}} \right) p_{ti}$$

$$I_t^L = \sum_{i=1}^n \left( \frac{p_{0i}q_{0i}}{\sum_{k=1}^n p_{0k}q_{0k}} \right) \left( \frac{p_{ti}}{p_{0i}} \right) = \text{moyenne pondérée des indices de prix des biens}$$

### Indice de Paasche

$$I_t^P = \frac{\sum_{i=1}^n p_{ti}q_{tk}}{\sum_{k=1}^n p_{0k}q_{tk}} = \sum_{i=1}^n \left( \frac{p_{0i}q_{tk}}{\sum_{k=1}^n p_{0k}q_{tk}} \right) \left( \frac{p_{ti}}{p_{0i}} \right)$$

## LES INDICES DE PRIX (EXEMPLE NUMÉRIQUE – PAASCHE)

### Calcul de l'indice de prix de Paasche

Biens	Données				Calculs	
	Prix (\$)		Quantités		Coût du panier	
	1980	1985	1980	1985	1980	1985
	$p_{0i}$	$p_{ti}$	$q_{0i}$	$q_{ti}$	$p_{0i} q_{ti}$	$p_{ti} q_{ti}$
Steak (kg)	4,85 \$	6,60 \$	23	18	87,30 \$	118,80 \$
Poivre (g)	0,07 \$	0,07 \$	57	85	5,95 \$	5,95 \$
Pain (kg)	1,10 \$	1,32 \$	36	45	49,50 \$	59,40 \$
					142,75 \$	184,15 \$

Biens	Données				Indice de prix de Paasche		
	Prix			Quant.	Coût du panier	Poids	Calcul de l'indice
	1980	1985	Rapport	1985	NAP <sup>1</sup>	NAP	1985
	$p_{0i}$	$p_{ti}$	$\left(\frac{p_{ti}}{p_{0i}}\right)$	$q_{ti}$	$p_{0i} q_{ti}$	$w_{ti}$	$w_{ti} \left(\frac{p_{ti}}{p_{0i}}\right)$
Steak (kg)	4,85 \$	6,60 \$	1,36	18	87,30 \$	0,612	0,832
Poivre (g)	0,07 \$	0,07 \$	1,00	85	5,95 \$	0,042	0,042
Pain (kg)	1,10 \$	1,32 \$	1,20	45	49,50 \$	0,347	0,416
					142,75 \$	1,000	1,290

<sup>1</sup> Ne s'applique pas, c'est-à-dire que les chiffres de la colonne ne s'appliquent à aucune année.

## UTILISATIONS DES INDICES DE PRIX

**Dépenses personnelles de consommation au Canada de 1991 à 1999, en millions de dollars courants, avec l'indice des prix correspondant**

	Indice de prix des dép. pers. dans le PIB	Dép. pers.
1991	91,0	398 314
1992	92,5	411 167
1993	94,6	428 219
1994	95,6	445 857
1995	96,8	460 906
1996	98,4	480 427
1997	100,0	510 695
1998	101,2	531 169
1999	102,9	560 954

Source : Statistique Canada, *L'observateur économique canadien*, Supplément statistique historique 2001/02, No 11-210-XPB.

### Utilisation d'un indice de prix comme dégonfleur

**Valeur en dollars constants de l'année 0 (année de base de l'indice)**

$$y_t = \frac{x_t}{I_t}$$

Dépenses personnelles de consommation de 1999 en dollars constants de 1997

$$100 \times \frac{560\,954}{102,9} = \frac{560\,954}{1,029} = 545\,145$$

**Généralisation : valeur en dollars constants d'une année  $\theta$  autre que l'année de base**

$$y_t = x_t \frac{I_\theta}{I_t}$$

Dépenses personnelles de consommation de 1999 en dollars constants de 1992

$$\frac{92,5}{102,9} 560\,954 = \frac{0,925}{1,029} 560\,954 = 504\,259$$

### Utilisation d'un indice de prix pour indexer un montant

**Indexation pour l'année  $t$  d'un montant  $m_0$  fixé à l'année zéro (année de base de l'indice)**

$$m_t = I_t m_0$$

35 000 \$ en dollars de 1997, indexé pour l'année 1998 :

$$35\,000 \$ \times 1,012 = 35\,420 \$$$

**Indexation pour l'année  $t$  d'un montant  $m_\theta$  fixé à une année  $\theta$  autre que l'année de base**

$$m_t = m_\theta \frac{I_t}{I_\theta}$$

35 000 \$ en dollars de 1998, indexé pour l'année 1999 :

$$35\,000 \$ \times \frac{1,029}{1,012} = 35\,588 \$$$



## INDICES DE PRIX ET COÛT DE LA VIE

	Thé	Café	<i>Et coetera</i>	Total
Quantité	1250	800	10000	
<b>Année 0</b>				
Prix	0,40	0,50	1,00	
Dépense	500	400	10000	10900
<b>Année t</b>				
Prix	4,00	0,50	1,00	
Dépense	5000	400	10000	15400

Indice de Laspeyres :

$$I_t^L = \frac{4,00 * 1250 + 0,50 * 800 + 1,00 * 10000}{0,40 * 1250 + 0,50 * 800 + 1,00 * 10000} = 1,41284$$

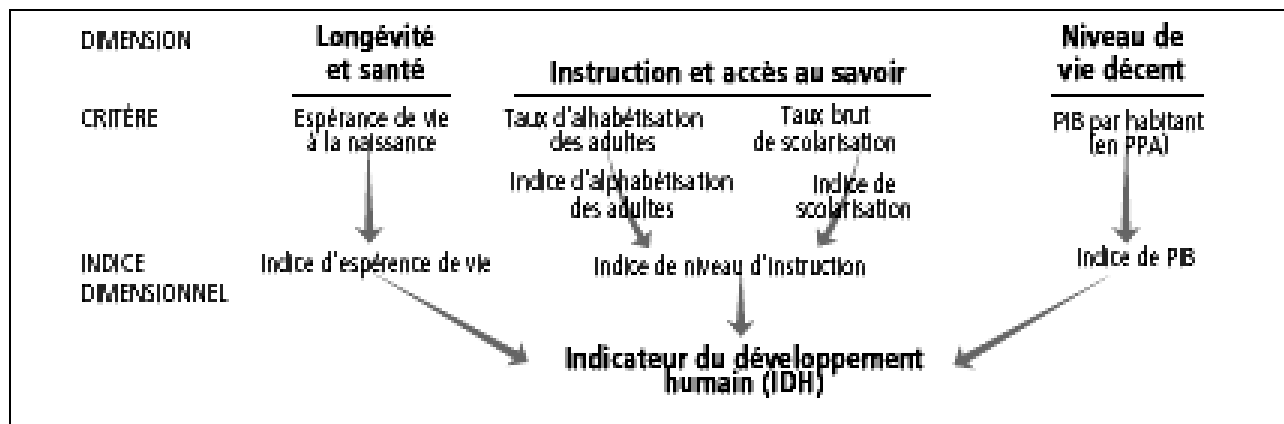
## L'INDICATEUR DE DÉVELOPPEMENT HUMAIN (IDH) DU PNUD <sup>2</sup>

Trois composantes (dimensions du concept de développement humain) :

- Longévité
- Savoir
- Niveau de vie

Variables (opérationnalisation des dimensions choisies) :

- Longévité : espérance de vie à la naissance
- Savoir : taux d'alphabétisation des adultes et taux de scolarisation (2 variables)
- Niveau de vie : produit intérieur brut (PIB) réel par habitant, en dollars ajustés en fonction du coût de la vie («PPA», c'est-à-dire en « parité de pouvoir d'achat »)



Source : PNUD, *Rapport mondial sur le développement humain 2003 - Les objectifs du millénaire pour le développement : Un pacte entre les pays pour vaincre la pauvreté humaine*, p. 340.

### Étapes du calcul de l'IDH

1. Recueillir les données sur les indicateurs associés aux dimensions du concept
2. Calculer l'indice relatif à chaque composante
3. Calculer l'IDH

<sup>2</sup> Programme des Nations Unies pour le Développement : <http://hdr.undp.org/>

## L'IDH DU PNUD (SUITE)

Pour chacun des indicateurs, le chemin à parcourir sur la voie du développement commence au niveau le plus bas qui se puisse observer dans le monde et aboutit au niveau le plus élevé qui se puisse espérer. Le progrès réalisé par un pays se mesure comme la fraction du chemin parcouru.

### Maximums et minimums (1995) :

Variable	Maximum	Minimum
Espérance de vie	85 ans	25 ans
Taux d'alphabétisation	100 %	0 %
Taux de scolarisation	100 %	0 %
PIB réel/habitant	40 000 \$	100 \$

NOTE : L'indicateur du niveau de vie utilisé dans le calcul de l'IDH est le *logarithme* du PIB réel par habitant exprimé en «PPA», c'est-à-dire en « parité de pouvoir d'achat ».

### Calcul de l'IDH pour le pays $j$ :

1. Pour chacune des quatre variables :

$$I_{ij} = \frac{x_{ij} - \min x_i}{\max x_i - \min x_i}, \text{ c'est-à-dire : Indicateur} = \frac{\text{Valeur réelle} - \text{Valeur minimale}}{\text{Valeur maximale} - \text{Valeur minimale}}$$

2. Indicateur retenu pour le savoir :

$$I_{\text{savoir},j} = 0,67 \times I_{\text{alpha},j} + 0,33 \times I_{\text{scolar},j}$$

3. Indicateur de développement humain :

$$I_j = \frac{1}{3} \sum_{i=1}^3 I_{ij}$$

## L'INDICATEUR DE DÉVELOPPEMENT HUMAIN (IDH) DU PNUD CALCUL POUR LE MEXIQUE, 2001

### CALCUL DE L'INDICE DE L'ESPÉRANCE DE VIE

$$\text{Indice de l'espérance de vie} = \frac{73,1 - 25}{85 - 25} = 0,802$$

### CALCUL DE L'INDICE DU NIVEAU D'ÉDUCATION

$$\text{Indice d'alphabétisation des adultes} = \frac{91,4 - 0}{100 - 0} = 0,914$$

$$\text{Indice de scolarisation} = \frac{74 - 0}{100 - 0} = 0,74$$

Indice de niveau d'éducation =

$$\frac{2}{3} (\text{Indice d'alphabétisation des adultes}) + \frac{1}{3} (\text{Indice de scolarisation})$$

$$\text{Indice de niveau d'éducation} = \frac{2}{3} 0,914 + \frac{1}{3} 0,74 = 0,86$$

### CALCUL DE L'INDICE DU PIB

$$\text{Indice du PIB} = \frac{\log(8\,430) - \log(100)}{\log(40\,000) - \log(100)} = 0,74$$

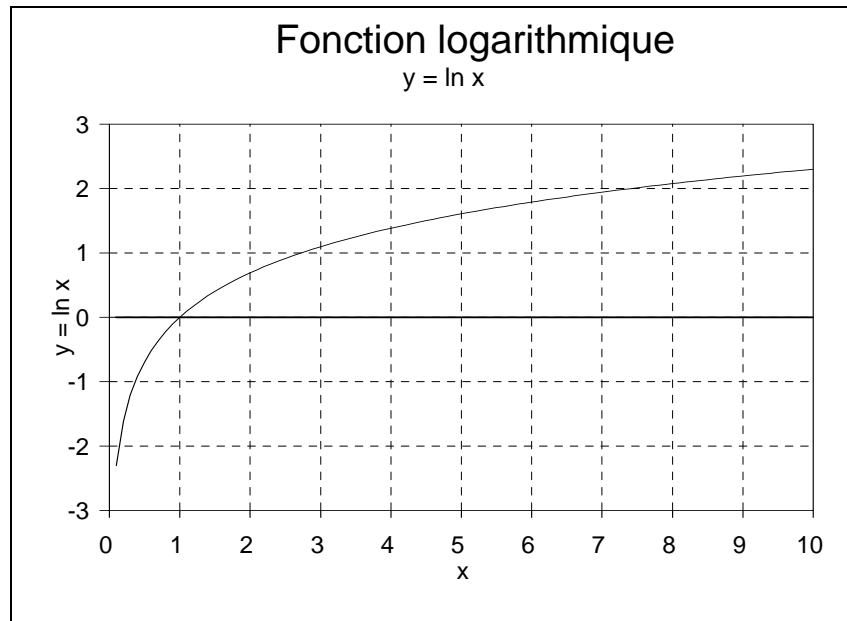
### CALCUL DE L'IDH

L'IDH est la moyenne des trois indices. Pour le Mexique en 2001 :

$$\text{Indice de développement humain (IDH)} = \frac{0,80 + 0,86 + 0,74}{3} = 0,800$$

## LA TRANSFORMATION LOGARITHMIQUE

### Relation entre un nombre $x$ et son logarithme $y$



C'est une transformation *monotone croissante* :

si  $y_1 > y_2$ , alors  $\log y_1 > \log y_2$ , puisque  $y = b^{\log y}$ .

## PROBLÉMATIQUE DE LA MESURE DE LA SIMILARITÉ/DISSIMILARITÉ

### Mesure

Mesurer, c'est comparer

Une *mesure* est une correspondance qui permet de comparer deux objets par rapport à une propriété donnée.

### Le problème de la multidimensionnalité : les nombres indices

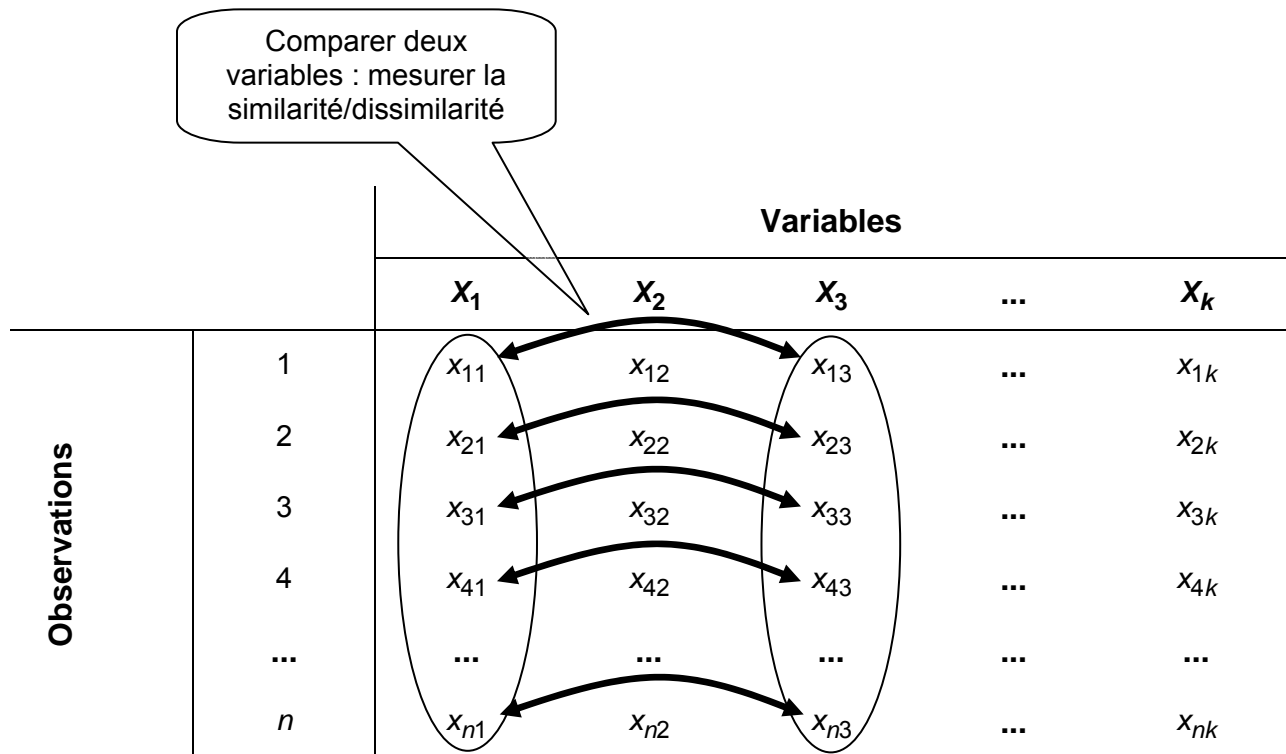
- Lazarsfeld : concept → dimensions → indicateurs (mesures)
- Problème :  
objet ou concept multidimensionnel  
**mais** on veut le traiter comme un tout  
⇒ il faut combiner les mesures partielles en une seule mesure globale, qui les résume
- Un nombre indice est une mesure : permet de comparer par rapport au concept
  - Fiabilité ?
  - Validité ?

### Comparer sans indice : mesure de la similarité/dissimilarité

- Certains concepts ne sont pas réductibles à un indice  
On ne peut pas associer une mesure unique au concept
- Dimensions multiples → indicateurs multiples → comparaisons multiples
- Comment faire la synthèse des comparaisons ?  
Au moyen d'un « indice des comparaisons »
- On mesure
  - **non pas** le degré auquel chaque objet « possède » le concept
  - **mais plutôt** le degré de similarité entre les objets par rapport au concept

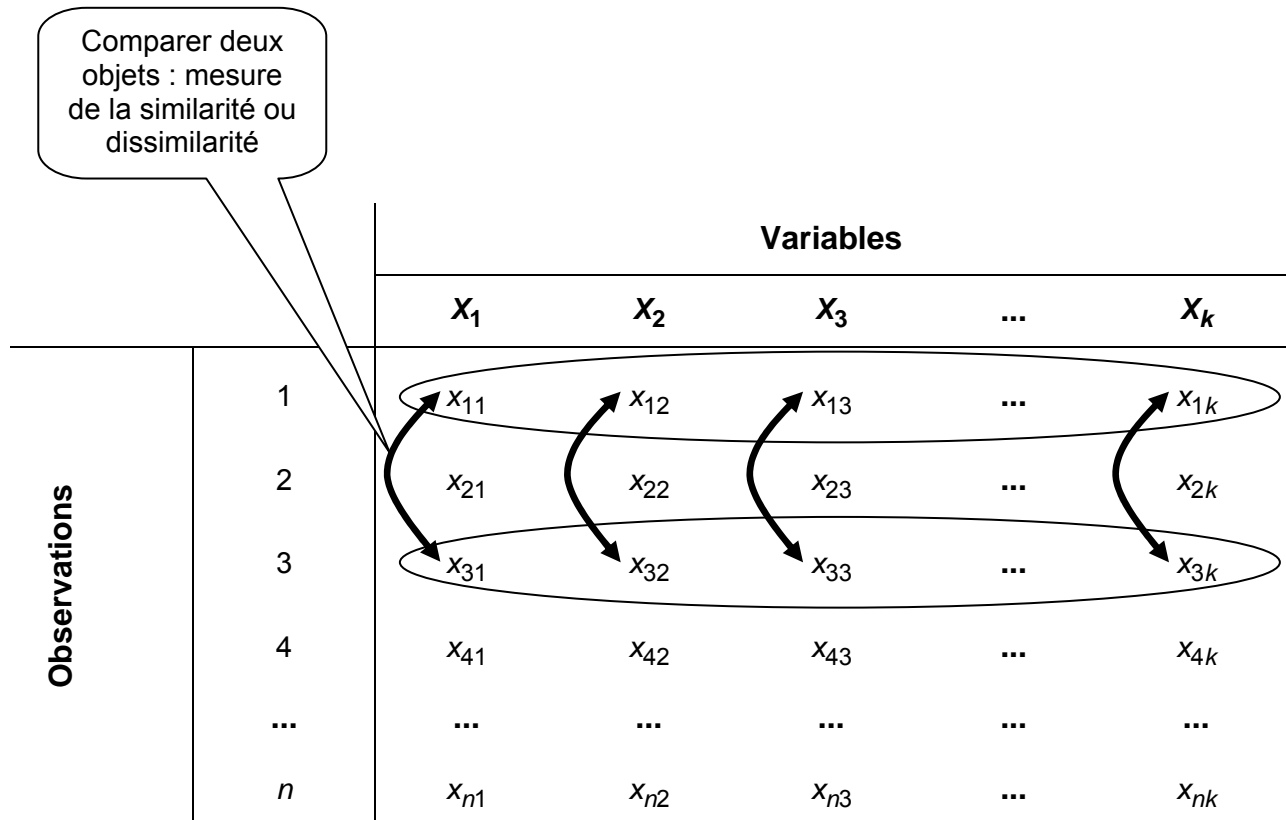
### STRUCTURE DES DONNÉES (3)

#### POINT DE VUE HORIZONTAL : SIMILARITÉ/DISSIMILARITÉ



## STRUCTURE DES DONNÉES (7)

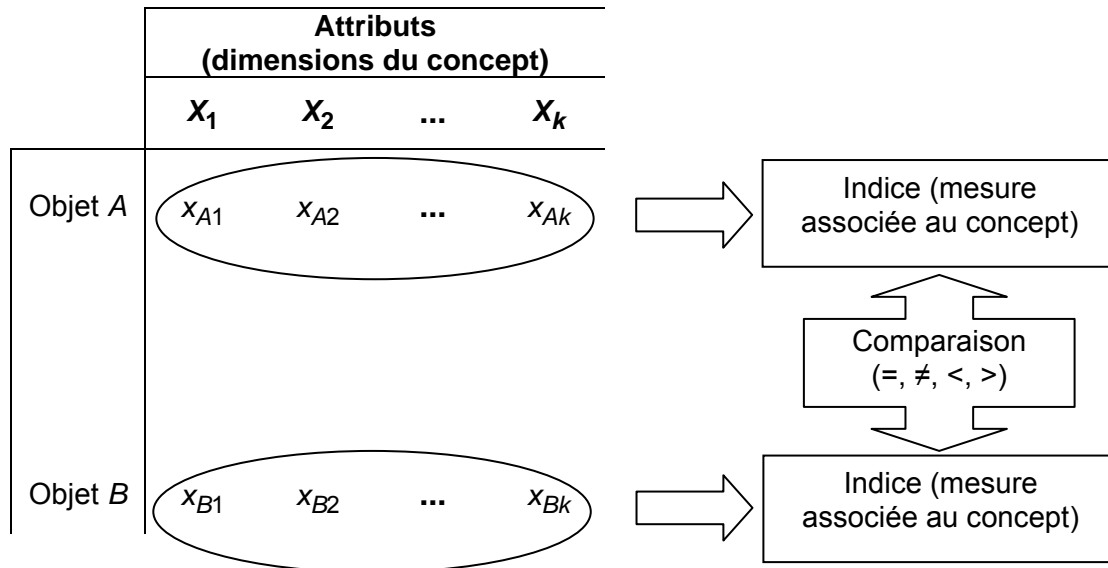
### POINT DE VUE VERTICAL : SIMILARITÉ/DISSIMILARITÉ (BIS)



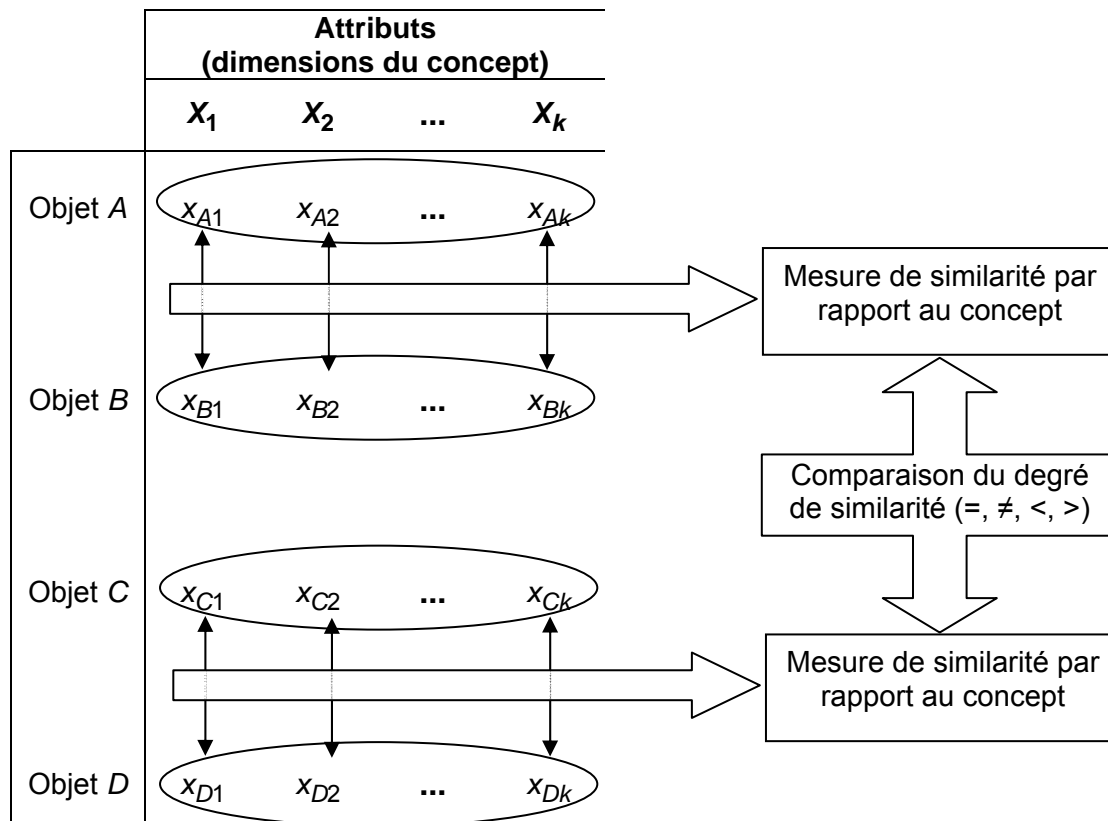


## DIFFÉRENCE ENTRE LA MESURE DE LA SIMILARITÉ ET LA CONSTRUCTION D'UN NOMBRE INDICE

### Construction d'un nombre indice



### Mesure de la similarité



## DANS QUELLES CIRCONSTANCES... ?

### Mesure de la similarité/dissimilarité en général

- Construction de typologies, algorithmes de classification
- Mesures d'ajustement (« goodness of fit ») en statistique : similarité entre les observations et les prédictions d'un modèle  
(ex. : fréquences observées et théoriques d'un tableau de contingence;  
coefficient de détermination multiple de la régression)

### Mesure de la similarité/dissimilarité dans un tableau de contingence (exemple)

- entre lignes ou entre colonnes...
- quant à la **structure**, c'est-à-dire quant à la répartition

L'indicateur de spécificité (quotient de localisation) s'applique à chaque cellule séparément.

L'analyse des tableaux de contingence, le test d'indépendance et les mesures d'intensité s'appliquent à l'ensemble du tableau.

La mesure de similarité/dissimilarité s'applique à chaque paire de lignes ou de colonnes.

Dans un tableau de contingence : cas particulier...

### Plus généralement : Mesure de la similarité/dissimilarité entre deux distributions

- Distribution de fréquences ou distribution d'une variable continue
- En particulier, distributions spatiales
- Dans une distribution, la somme des parts est égale à 1 (100 %) :  $\sum_1 p_i = 1$

Cela règle le problème de la pondération

### Distribution observée et distribution théorique

- Approche analogue à la construction du test d'hypothèse d'indépendance
- **mais** ici, la distribution théorique n'est pas une hypothèse à tester, c'est la représentation du degré maximum d'une propriété
- Cette approche s'applique notamment à la...

### Mesure de l'inégalité ou de la concentration

- La concentration est le contraire de l'égalité dans une distribution
- Elle peut se mesurer par le degré de dissimilarité par rapport à une distribution de référence qui représente l'égalité parfaite
- Nombreux champs d'application de la mesure de l'inégalité :
  - géographie : concentration spatiale des phénomènes
  - économie : inégalités de revenu et questions de justice sociale; concentration de marché

## POPULATION ACTIVE EMPLOYÉE DANS LA RÉGION MÉTROPOLITAINE DE MONTRÉAL ZONE DE RÉSIDENCE, SELON LE SEXE ET LA PROFESSION, 1991

Zone de résidence	Professions					TOTAL toutes professions
	Directeurs, gérants, administrateurs et assimilés	Professionnels, enseignants et cols blancs spécialisés	Employés de bureau et travailleurs dans la vente	Ouvriers	Travailleurs spécialisés dans les services, personnel d'exploitation des transports, etc.	
<b>Femmes</b>						
Ville de Montréal	24 025	58 204	76 450	24 385	28 825	<b>211 889</b>
Reste de la CUM	22 575	42 207	70 003	14 065	17 435	<b>166 285</b>
Couronne Nord	16 785	31 699	63 491	11 975	18 630	<b>142 580</b>
Couronne Sud	18 365	35 674	65 290	10 485	19 380	<b>149 194</b>
Hors RMR	3 265	7 535	11 089	3 190	3 565	<b>28 644</b>
<b>Total Femmes</b>	<b>85 015</b>	<b>175 319</b>	<b>286 323</b>	<b>64 100</b>	<b>87 835</b>	<b>698 592</b>
<b>Hommes</b>						
Ville de Montréal	32 336	55 045	43 546	65 340	46 850	<b>243 117</b>
Reste de la CUM	39 146	39 920	37 819	46 173	28 749	<b>191 807</b>
Couronne Nord	33 287	27 560	31 170	62 852	29 329	<b>184 198</b>
Couronne Sud	36 006	32 464	30 600	58 778	29 721	<b>187 569</b>
Hors RMR	8 270	8 590	8 270	22 305	9 099	<b>56 534</b>
<b>Total Hommes</b>	<b>149 045</b>	<b>163 579</b>	<b>151 405</b>	<b>255 448</b>	<b>143 748</b>	<b>863 225</b>
<b>Total hommes et femmes</b>						
Ville de Montréal	56 361	113 249	119 996	89 725	75 675	<b>455 006</b>
Reste de la CUM	61 721	82 127	107 822	60 238	46 184	<b>358 092</b>
Couronne Nord	50 072	59 259	94 661	74 827	47 959	<b>326 778</b>
Couronne Sud	54 371	68 138	95 890	69 263	49 101	<b>336 763</b>
Hors RMR	11 535	16 125	19 359	25 495	12 664	<b>85 178</b>
<b>Total H + F</b>	<b>234 060</b>	<b>338 898</b>	<b>437 728</b>	<b>319 548</b>	<b>231 583</b>	<b>1 561 817</b>

Source : Statistique Canada, Recensement de 1991

## STRUCTURE DES DONNÉES (6)

### POINT DE VUE VERTICAL : INÉGALITÉ, DISTRIBUTION

- Caractériser la distribution
  - Mesurer l'inégalité ou la concentration
- S'il existe un ordre naturel des observations :
- Analyser des séries temporelles
  - Analyser l'autocorrélation temporelle ou spatiale

		Variables				
		$X_1$	$X_2$	$X_3$	...	$X_k$
Observations	1	$x_{11}$	$x_{12}$	$x_{13}$	...	$x_{1k}$
	2	$x_{21}$	$x_{22}$	$x_{23}$	...	$x_{2k}$
	3	$x_{31}$	$x_{32}$	$x_{33}$	...	$x_{3k}$
	4	$x_{41}$	$x_{42}$	$x_{43}$	...	$x_{4k}$
	...	...	...	...	...	...
	$n$	$x_{n1}$	$x_{n2}$	$x_{n3}$	...	$x_{nk}$

## COMMENT MESURER L'INÉGALITÉ ?

### Qu'est-ce que l'inégalité ? Exemple du revenu

- Entre 2 personnes :
  - Si  $R_1 = R_2 \Rightarrow$  égalité
  - Si  $R_1 \neq R_2 \Rightarrow$  inégalité : elle peut se mesurer par la différence ( $R_1 - R_2$ ), le rapport  $\frac{R_1}{R_2}$  ou une transformation de ceux-ci.
- Entre plus de 2 personnes :
  - Si  $R_1 = R_2 = R_3 = \dots \Rightarrow$  égalité
  - Sinon, il n'y a pas égalité **mais** comment mesurer le degré d'inégalité ?

### Propriétés désirables d'une mesure d'inégalité (Valeyre, 1993)

1. Non négative
2. Égale à zéro si, et seulement si la distribution observée identique à distribution de référence.
3. Toutes observations traitées de la même manière.
4. Indépendante de la valeur moyenne de la variable.  
Indépendante de la taille de la population.
5. L'agrégation d'observations affichant le même degré de spécificité ne doit pas changer la valeur de la mesure.
6. Principe de transfert de Pigou-Dalton : une mesure d'inégalité doit diminuer si la distribution est modifiée d'une façon qui réduit incontestablement l'inégalité.

## COMMENT MESURER L'INÉGALITÉ ? (SUITE)

**La dispersion des valeurs observées s'interprète souvent comme le reflet de la concentration ou de l'inégalité de la propriété mesurée.**

- Exemple : avec des données sur le revenu, si tout le monde a le même revenu, il n'y a pas de dispersion (la variance est nulle), il n'y a pas d'inégalité entre les individus (observations) et le revenu n'est pas concentré ; plus il y a de différences entre les revenus, plus la variance est grande.

### Rappel : mesures de dispersion en statistique descriptive

- Domaine de variation : valeur minimum et valeur maximum
- Écart inter-quartile
- Variance :  $\sigma_x^2 = \frac{1}{n} \sum_i (x_i - \mu_x)^2$

NOTE : Cette formule est celle qui s'applique à une population, puisque la statistique descriptive ne distingue pas entre population et échantillon.

- Écart-type :  $\sigma_x = \sqrt{\sigma_x^2}$
- Coefficient de variation :  $C_x = \frac{\sigma_x}{\mu_x}$

Seul le coefficient de variation possède les 6 propriétés désirées.

### Peut-on mesurer l'inégalité ou la concentration sans référer à la moyenne ?

- Oui ! Corrado Gini (1884-1965) a proposé de comparer chacun des individus avec chacun des autres : cela donne la différence moyenne de Gini.

### Autres mesures d'inégalité ou de concentration

- Le coefficient de concentration de l'économie industrielle

$$C4 = \sum_{i=1}^4 s_i, \text{ où } s_i \text{ est la part de } i \text{ dans le total}$$

- L'indice de concentration de Hirschman-Herfindahl

$$H = \sum_{i=1}^n s_i^2$$

$$\frac{1}{n} \leq H \leq 1$$

Interprétation en «nombre équivalent»

$$\text{Variance des parts} = \frac{1}{n} \sum_{i=1}^n \left( s_i - \frac{1}{n} \right)^2 = \frac{H}{n} - \frac{1}{n^2}$$

- Mesure d'entropie

## L'INDICE DE CONCENTRATION DE GINI : LA DIFFÉRENCE MOYENNE DE GINI (EXEMPLE NUMÉRIQUE)

**Données  
(fictives)**

Individus	Revenu
<b>A</b>	100
<b>B</b>	40
<b>C</b>	30
<b>D</b>	20
<b>E</b>	20
<b>F</b>	20
<b>G</b>	20
<b>H</b>	20
<b>I</b>	20
<b>J</b>	10

<b>Total</b>	300
<b>Moyenne</b>	30
<b>Éc. type</b>	24,49
<b>Coef. var.</b>	0,816

Calcul de la différence (absolue) moyenne  $|x_i - x_j|$

	A	B	C	D	E	F	G	H	I	J
	<b>100</b>	<b>40</b>	<b>30</b>	<b>20</b>	<b>20</b>	<b>20</b>	<b>20</b>	<b>20</b>	<b>20</b>	<b>10</b>
<b>100</b>	0	60	70	80	80	80	80	80	80	90
<b>40</b>	60	0	10	20	20	20	20	20	20	30
<b>30</b>	70	10	0	10	10	10	10	10	10	20
<b>20</b>	80	20	10	0	0	0	0	0	0	10
<b>20</b>	80	20	10	0	0	0	0	0	0	10
<b>20</b>	80	20	10	0	0	0	0	0	0	10
<b>20</b>	80	20	10	0	0	0	0	0	0	10
<b>20</b>	80	20	10	0	0	0	0	0	0	10
<b>20</b>	80	20	10	0	0	0	0	0	0	10
<b>10</b>	90	30	20	10	10	10	10	10	10	0

<b>Somme</b>	2000
<b>Dif. Moy. Gini</b>	20
<b>Coef. Gini</b>	0,333

$$\Delta = \frac{1}{N^2} \sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|, \text{ avec } N=10 : \Delta = \frac{2000}{10^2} = 20$$

$$G = \frac{\Delta}{2\mu} = \frac{20}{2 \times 30} = 0,333$$

## L'INDICE DE CONCENTRATION DE GINI : LA DIFFÉRENCE MOYENNE DE GINI (EXEMPLE NUMÉRIQUE AVEC DONNÉES GROUPÉES)

Données groupées

		$f_j$	$y_j$
Cat.	Rev.	N.	R/N
>25	170	3	56,67
15-25	120	6	20
<15	10	1	10
<b>Tot.</b>	300	10	
<b>Moy.</b>	30	(pondérée)	

Écarts

$$|y_i - y_j|$$

		>25	15-25	<15
		56,67	20	10
>25	56,67	0	36,67	46,67
15-25	20	36,67	0	10
<15	10	46,67	10	0

Poids

$$f_i f_j$$

		>25	15-25	<15
		3	6	1
>25	3	9	18	3
15-25	6	18	36	6
<15	1	3	6	1
<b>Tot.</b>		<b>100</b>		

Écarts pondérés

$$|y_i - y_j| f_i f_j$$

		>25	15-25	<15
		0	660	140
>25	0	660	0	60
15-25	660	0	60	0
<15	140	60	0	0
<b>Tot.</b>		<b>1720</b>		

$$\Delta = \frac{1}{N^2} \sum_{i=1}^n \sum_{j=1}^n |y_i - y_j| f_i f_j = \frac{1720}{100} = 17,2$$

$$G = \frac{\Delta}{2\mu} = \frac{17,2}{2 \times 30} = 0,287$$



## L'INDICE DE CONCENTRATION DE GINI : LA DIFFÉRENCE MOYENNE DE GINI (FORMULE ALGÈBRIQUE)

### Notation

$n$  = le nombre de valeurs distinctes observées

$f_j$  = fréquence de la valeur  $y_j$  dans la distribution

$$N = \sum_{j=1}^n f_j = \text{nombre d'observations}$$

### Définition de la différence moyenne de Gini

$$\Delta = \frac{1}{N^2} \sum_{j=1}^n \sum_{k=1}^n |y_j - y_k| f_j f_k$$

### Observations groupées par classes

$y_j$  = valeur *moyenne* de la variable  $Y$  dans la classe  $j$

$v_j = \frac{f_j}{N}$ , la fraction de la population appartenant à la classe  $j$ .

La valeur moyenne de la variable  $Y$  s'écrit alors

$$\mu = \frac{1}{N} \sum_{j=1}^n f_j y_j = \sum_{j=1}^n v_j y_j$$

### Notation

$$M = \mu N = \sum_{j=1}^n f_j y_j = \text{somme des valeurs de la variable } Y$$

$$w_j = \frac{f_j y_j}{\sum_{k=1}^n f_k y_k} = \frac{f_j y_j}{N \mu} = \frac{v_j y_j}{\mu} = \text{fraction de la somme allouée à la classe } j.$$

$$Cw_j = \sum_{k=1}^j w_k, \text{ avec observations par ordre croissant des } w_j/v_j$$

$$\Delta = 2\mu \left( 1 - \sum_{j=1}^n v_j Cw_j - \sum_{j=1}^n v_j Cw_{j-1} \right)$$

## CALCUL DE L'INDICE DE CONCENTRATION DE GINI

$v_j = \frac{f_j}{N}$ , la fraction de la population appartenant à la classe  $j$ .

$Cw_j = \sum_{k=1}^j w_k$ , avec observations par ordre croissant des  $w_k/v_k$

$$G = \frac{\Delta}{2\mu} = 1 - \left( \sum_{j=1}^n v_j Cw_j + \sum_{j=1}^n v_j Cw_{j-1} \right) = 1 - \sum_{j=1}^n v_j (Cw_j + Cw_{j-1})$$

**Calcul équivalent d'Arriaga (1975, p. 65-71)**

$$G = \sum_{i=2}^n Cw_i Cw_{i-1} - \sum_{i=2}^n Cw_{i-1} Cw_i$$

où  $Cv_j = \sum_{k=1}^j v_k$

## LA COURBE DE LORENZ

### La courbe de Lorenz

#### Notation supplémentaire

$$CV_j = \sum_{k=1}^j v_k = \text{fraction cumulée de la population } X$$

$$CW_j = \sum_{k=1}^j w_k = \text{fraction cumulée de la population } Y$$

$$CV_n = CW_n = 1$$

#### Méthode de construction de la courbe de Lorenz

1. Calculer les rapports  $\frac{w_i}{v_i}$ . Ce sont les *spécificités* associées aux observations.
2. Réordonner les catégories en ordre croissant de  $\frac{w_i}{v_i}$  :  $\frac{w_1}{v_1} < \frac{w_2}{v_2} < \dots < \frac{w_n}{v_n}$
3. La courbe de Lorenz est l'ensemble des points  $(CV_i, CW_i)$ , où les  $CV_i$  sont repérés sur l'axe horizontal.

#### Propriétés de la courbe de Lorenz

1.  $CV_0 = CW_0 = 0$
2.  $CV_n = CW_n = 1$
3. Lorsque les deux distributions sont identiques, on a, pour tout  $i$ ,  
 $CV_i = CW_i$   
La courbe de Lorenz coïncide avec la diagonale.
4.  $CV_i \geq CW_i$  pour  $i$  différent de 0 et de  $n$
5. La pente de chaque segment de la courbe de Lorenz est égale à la valeur l'indicateur de spécificité associé à l'observation correspondante :  
$$\text{pente du segment } i = \frac{CW_i - CW_{i-1}}{CV_i - CV_{i-1}} = \frac{w_i}{v_i}$$
6. La courbe de Lorenz est concave vers le haut, c'est-à-dire que chaque segment a une pente plus abrupte que le précédent : cela découle de 5, puisque, par construction,  $\frac{w_i}{v_i} < \frac{w_{i+1}}{v_{i+1}}$

## CONSTRUCTION D'UNE COURBE DE LORENZ (EXEMPLE NUMÉRIQUE TIRÉ DE TAYLOR, 1977, P. 180)

### Première étape : calcul des $w_j/v_j$

Zone	$x_j$ Nombre de ménages de classe moyenne	$v_j$ Distrib. de x	$y_j$ Nombre de votes du parti Républ.	$w_j$ Distrib. de y	$w_j/v_j$
A	30	0,25	30	0,30	1,20
B	20	0,17	15	0,15	0,90
C	10	0,08	8	0,08	0,96
D	10	0,08	5	0,05	0,60
E	20	0,17	19	0,19	1,14
F	30	0,25	23	0,23	0,92
<b>Tot.</b>	<b>120</b>	<b>1,00</b>	<b>100</b>	<b>1,00</b>	

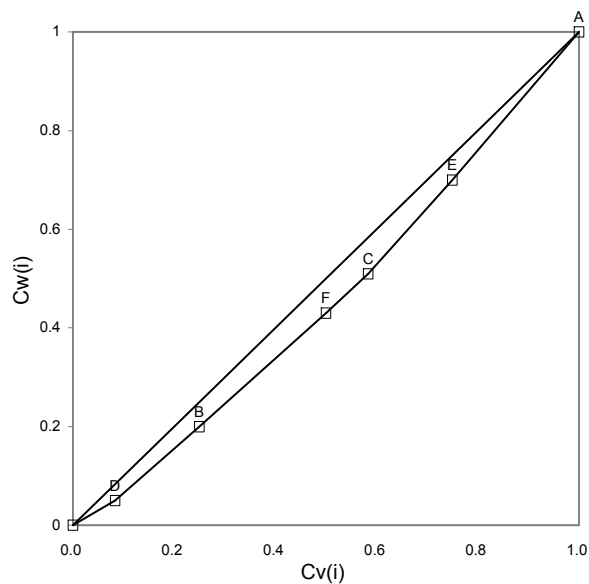
### Deuxième étape : tri par ordre croissant des $w_j/v_j$

### Troisième étape : calcul des $Cv_j$ (abscisses) et des $Cw_j$ (ordonnées)

Zone	$x_j$	$v_j$	$y_j$	$w_j$	$w_j/v_j$	$Cv_j$ Abscisse	$Cw_j$ Ordonnée	Différ. ( $Cv_j - Cw_j$ )	Différ.abs. $ v_j - w_j $
						0,00	0,00		
D	10	0,08	5	0,05	0,60	0,08	0,05	0,033	0,033
B	20	0,17	15	0,15	0,90	0,25	0,20	0,050	0,017
F	30	0,25	23	0,23	0,92	0,50	0,43	0,070	0,020
C	10	0,08	8	0,08	0,96	0,58	0,51	<b>0,073</b>	0,003
E	20	0,17	19	0,19	1,14	0,75	0,70	0,050	0,023
A	30	0,25	30	0,30	1,20	1,00	1,00	0,000	0,050
<b>Tot.</b>	<b>120</b>	<b>1,00</b>	<b>100</b>	<b>1,00</b>					<b>0,147</b>

Note : on peut voir que le maximum de la différence absolue entre la courbe de Lorenz et la diagonale est égal à  $\frac{1}{2} \sum_i |v_i - w_i|$ .

### Courbe de Lorenz

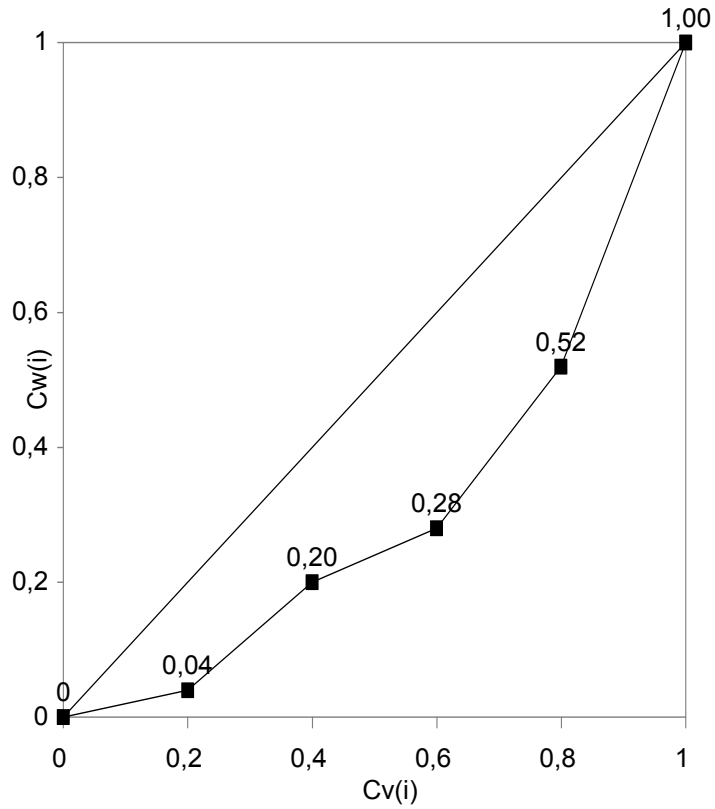


#### Quatrième étape : calcul de l'indice de concentration de Gini

Zone	$x_i$	$v_i$	$y_i$	$w_i$	$w_i/v_i$	$Cv_i$	$Cw_i$	$v_i Cw_i$	$v_i Cw_{i-1}$
						Abscisse	Ordonnée		
						0,00	0,00		
D	10	0,08	5	0,05	0,60	0,08	0,05	0,004	0,000
B	20	0,17	15	0,15	0,90	0,25	0,20	0,033	0,008
F	30	0,25	23	0,23	0,92	0,50	0,43	0,108	0,050
C	10	0,08	8	0,08	0,96	0,58	0,51	0,043	0,036
E	20	0,17	19	0,19	1,14	0,75	0,70	0,117	0,085
A	30	0,25	30	0,30	1,20	1,00	1,00	0,250	0,175
<b>Tot.</b>	<b>120</b>	<b>1,00</b>	<b>100</b>	<b>1,00</b>				<b>0,554</b>	<b>0,354</b>

$$G = 1 - (0,554 + 0,354) = 0,092$$

## LA COURBE DE LORENZ ET LE COEFFICIENT GINI : CECI N'EST PAS UNE COURBE DE LORENZ !



## CALCUL GÉOMÉTRIQUE DE L'INDICE DE CONCENTRATION DE GINI

### Définition géométrique de l'indice de concentration de Gini

$$G = \frac{\text{Superficie comprise entre la diagonale et la courbe de Lorenz}}{\text{Superficie totale sous la diagonale}}$$

### Calcul

$$\text{Superficie totale du triangle sous la diagonale} = \frac{Cw_n \times Cv_n}{2} = \frac{1 \times 1}{2} = \frac{1}{2}$$

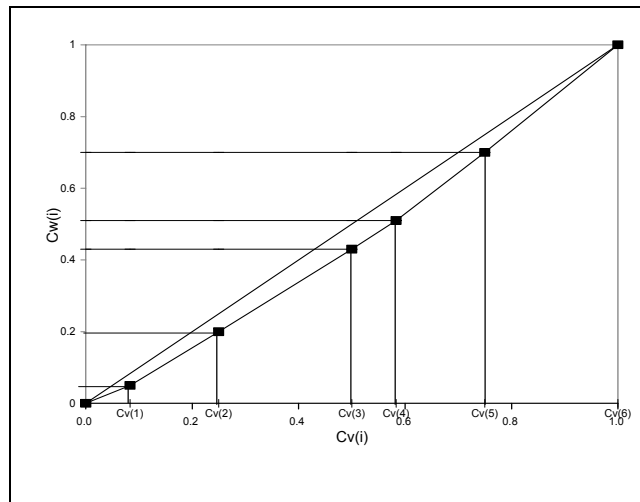
Superficie entre diagonale et courbe de Lorenz = différence entre :

Superficie totale du triangle sous la diagonale (=1/2) et

Superficie sous la courbe de Lorenz

Superficie sous la courbe de Lorenz = somme de n trapèzes :

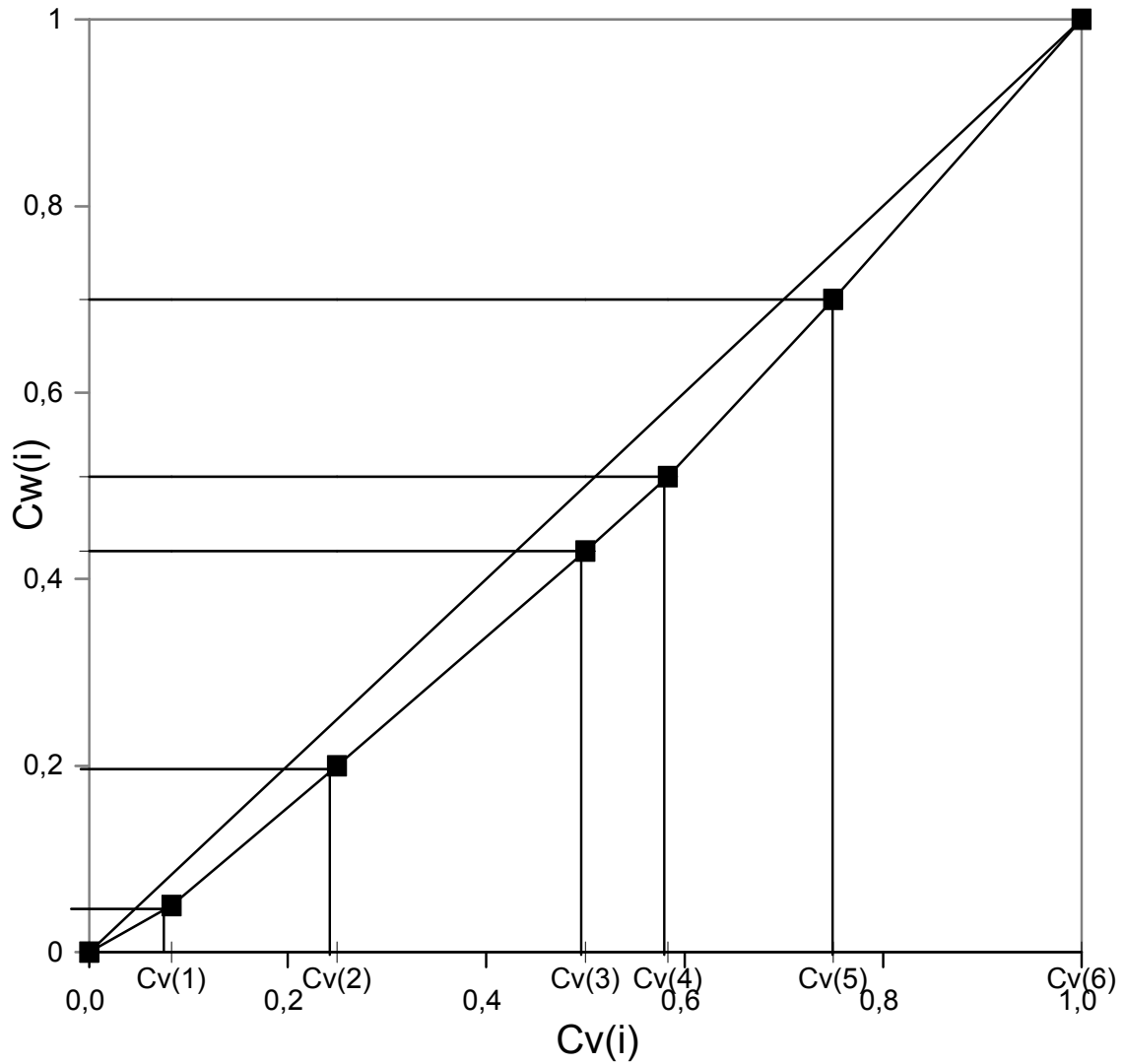
$$\frac{1}{2} v_i (Cw_i + Cw_{i-1})$$



### Coefficient Gini :

$$G = \frac{\left(\frac{1}{2}\right) - \left(\frac{1}{2} \sum_{i=1}^n v_i (Cw_i + Cw_{i-1})\right)}{\left(\frac{1}{2}\right)} = 1 - \sum_{i=1}^n v_i (Cw_i + Cw_{i-1}) = \frac{\Delta}{2\mu}$$

## LA COURBE DE LORENZ ET L'INDICE DE CONCENTRATION DE GINI : CALCUL GÉOMÉTRIQUE





## INTERPRÉTATION ET PROPRIÉTÉS DE L'INDICE DE CONCENTRATION DE GINI

### *Interprétation*

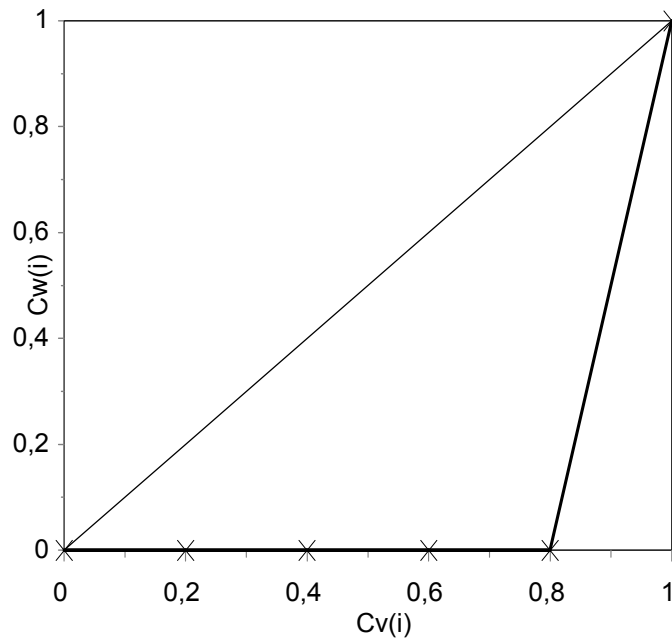
1. Mesure de dissimilarité entre deux distributions
2. Mesure de concentration :
  - $V$  = distribution de référence (axe horizontal)
  - $W$  = distribution dont on veut mesurer la concentration (axe vertical)

### *Propriétés de l'indice de concentration de Gini*

1. Possède les 6 propriétés désirables d'une mesure d'inégalité (Valeyre, 1993)
2.  $0 \leq G \leq 1$ , ou plus exactement  $0 \leq G \leq 1 - v_n$
3.  $G$  est symétrique.
4. Quand les données sont regroupées,  $G$  est sensible à la définition et au nombre des catégories utilisées (classes, zones).

Cela se manifeste notamment par : l'agrégation de deux ou plusieurs catégories entraîne une diminution de la valeur de l'indice de Gini, **sauf** si les catégories ont la même spécificité.
5. En tant que mesure de concentration spatiale, le Gini ne tient aucun compte de la proximité dans l'espace des différentes zones de forte densité.

## LA VALEUR MAXIMUM DU COEFFICIENT GINI



Zone	$v(i)$	$w(i)$	$w(i)/v(i)$	$Cv(i)$	$Cw(i)$	$Cv(i)-Cw(i)$	$ v(i)-w(i) $
A	0,20	0,00	0,00	0,20	0,00	0,20	0,20
B	0,20	0,00	0,00	0,40	0,00	0,40	0,20
C	0,20	0,00	0,00	0,60	0,00	0,60	0,20
D	0,20	0,00	0,00	0,80	0,00	0,80	0,20
E	0,20	1,00	5,00	1,00	1,00	0,00	0,80
Total	1,00	1,00					1,60

Indice de dissimilarité ( $D$ ) = 0,80

Coefficient Gini = 0,80

## EXEMPLE NUMÉRIQUE DE L'EFFET DE L'AGRÉGATION

### Données initiales (« détaillées »)

	Superf.	Population		Densité	
		Période 0	Période $t$	Période 0	Période $t$
Zone 1	1	10	80	10	80
Zone 2	1	80	10	80	10
Zone 3	1	10	10	10	10

$G_0 = G_t = 0,47$ , même si le centre de gravité de la population s'est déplacé vers la Zone 1.

### Agrégation des zones 2 et 3 (découpage A)

	Superf.	Population		Densité	
		Période 0	Période $t$	Période 0	Période $t$
Zone 1	1	10	80	10	80
Zones 2 et 3	2	90	20	45	10

$G'_0 = 0,23$  ;  $G'_t = 0,47$

$G'_t = G_t = 0,47$ , puisque les zones agrégées sont d'égale densité (spécificité) à la période  $t$ .

### Agrégation des zones 1 et 2 (découpage B)

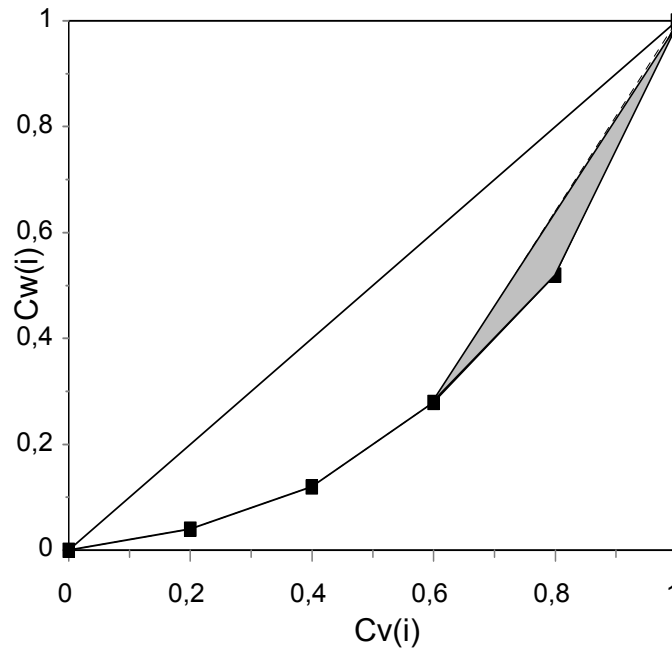
	Superf.	Population		Densité	
		Période 0	Période $t$	Période 0	Période $t$
Zones 1 et 2	2	90	90	45	45
Zone 3	1	10	10	10	10

$G''_0 = G''_t = 0,23 < G_0 = G_t = 0,47$

### Conclusions

- Sensibilité au découpage : les résultats « détaillés », ceux du découpage A et ceux du découpage B sont différents.
- Effet de l'agrégation : la valeur de l'indice de Gini diminue lorsqu'on agrège, **sauf** si on agrège des catégories (zones) de même spécificité (densité).

## EFFET DE L'AGRÉGATION SUR LE COEFFICIENT GINI



Zona	$v(i)$	$w(i)$	$w(i)/v(i)$	$Cv(i)$	$Cw(i)$	$Cv(i) - Cw(i)$	$ v(i) - w(i) $
A	0,20	0,04	0,2	0,20	0,04	0,16	0,16
B	0,20	0,08	0,4	0,40	0,12	0,28	0,12
C	0,20	0,16	0,8	0,60	0,28	0,32	0,04
D	0,20	0,24	1,2	0,80	0,52	0,28	0,04
E	0,20	0,48	2,4	1,00	1,00	0,00	0,28
Total	1,00	1,00					0,64
<b>Agregation des catégories D et E</b>							
D+E	0,40	0,72	1,80	1,00	1,00	0,00	0,32
Total	1,00	1,00					0,64

Indice de dissimilarité ( $D$ ) = 0,32

Coefficient Gini = 0,416 avant l'agrégation

Coefficient Gini = 0,368 après l'agrégation

## DISTANCE ET DISSIMILARITÉ

### **Propriétés d'une fonction de distance :**

(c1) non négativité :

$$d(a,b) \geq 0$$

(c2) identité :

$$d(a,b) = 0 \text{ si, et seulement si } a = b$$

(c3) symétrie :

$$d(a,b) = d(b,a)$$

(c4) inégalité triangulaire :

$$d(a,c) \leq d(a,b) + d(b,c)$$

### **Distance euclidienne**

$$d_e(a,b) = \sqrt{X_{ab}^2 + Y_{ab}^2}$$

où

$$X_{ab} = |x_a - x_b|$$

$$Y_{ab} = |y_a - y_b|$$

### **Distance rectilinéaire (métrique de Manhattan) :**

$$d_r(a,b) = X_{ab} + Y_{ab}$$

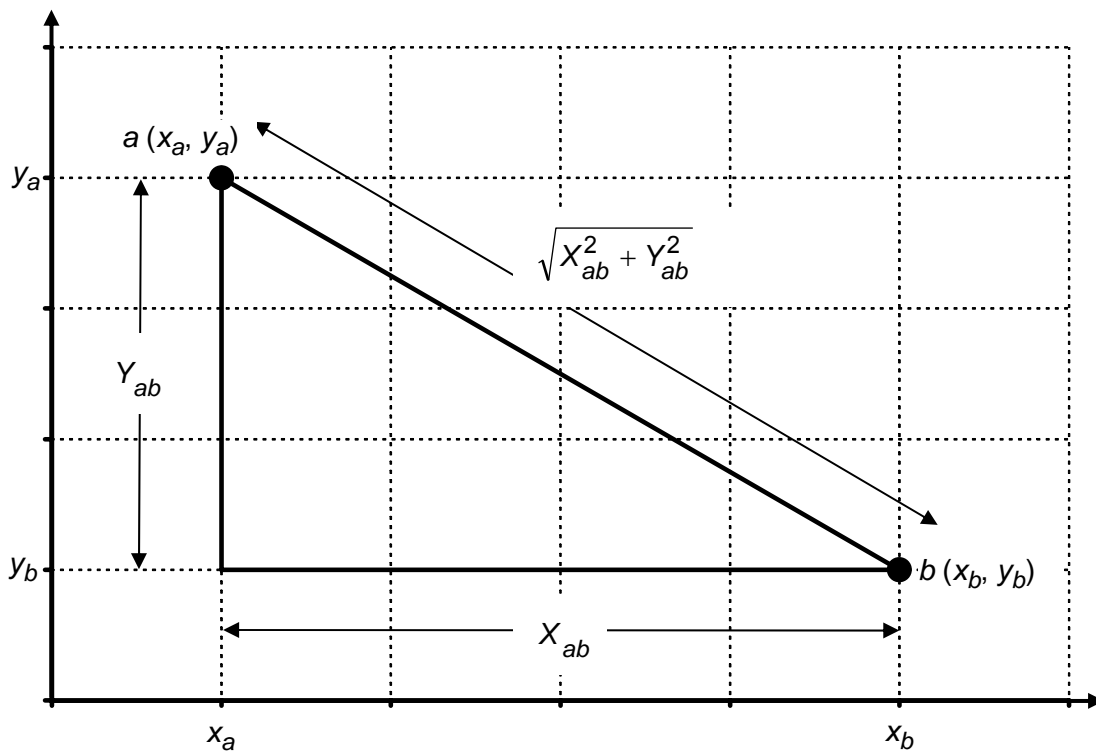
## DISTANCES

Soit les points  $a$  et  $b$ , de coordonnées cartésiennes  $(x_a, y_a)$  et  $(x_b, y_b)$  respectivement.

Définissons

$$X_{ab} = |x_a - x_b|$$

$$Y_{ab} = |y_a - y_b|$$



### **Distance euclidienne**

$$d_e(a,b) = \sqrt{X_{ab}^2 + Y_{ab}^2}$$

### **Distance rectilinéaire (métrique de Manhattan) :**

$$d_r(a,b) = X_{ab} + Y_{ab}$$

## DISTANCE ET MESURE DE LA DISSIMILARITÉ

La mesure de la distance est une mesure de la dissimilarité quant à la situation dans l'espace.

La situation dans un espace à 2 dimensions est décrite par 2 coordonnées :

	Latitude x	Longitude y
Point a	$x_a$	$y_a$
Point b	$x_b$	$y_b$
Différence	$x_a - x_b$	$y_a - y_b$

Définir une mesure de distance, c'est définir la façon de combiner les différences en une seule mesure.

La mesure de distance permet ensuite de déterminer, parmi les relations suivantes, lesquelles sont vraies :

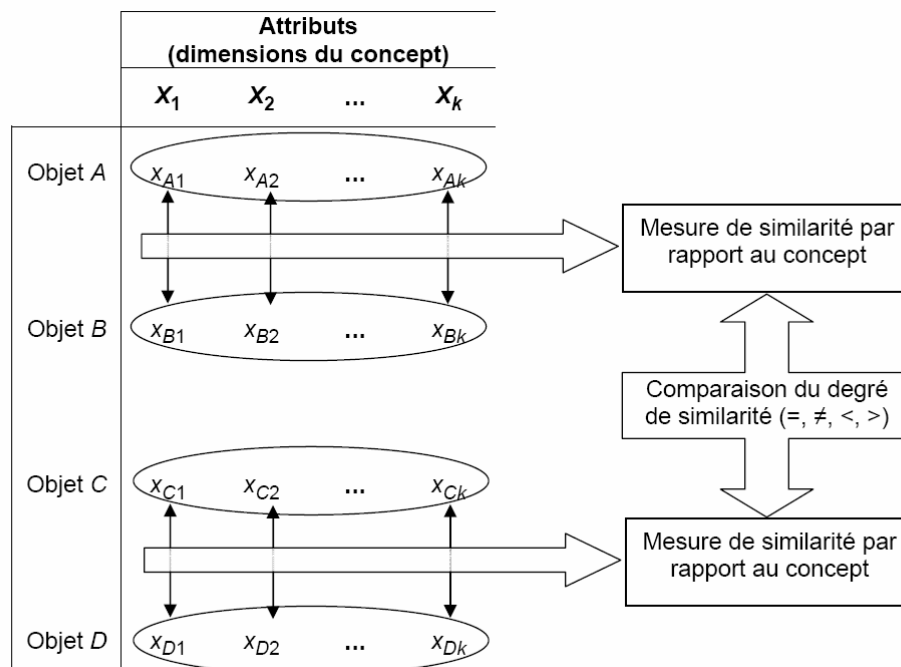
$$d_{\lambda}(a,b) = d_{\lambda}(a,b)$$

$$d_{\lambda}(a,b) \neq d_{\lambda}(a,b)$$

$$d_{\lambda}(a,b) > d_{\lambda}(a,b)$$

$$d_{\lambda}(a,b) < d_{\lambda}(a,b)$$

### Mesure de la similarité



## DISTANCES GÉNÉRALISÉES

### Distances généralisées entre des distributions

Distributions comparées :

$$p_{11}, p_{12}, \dots, p_{1n}$$

$$p_{21}, p_{22}, \dots, p_{2n}$$

$$\text{avec } \sum_i p_{ki} = 1$$

- Distance rectilinéaire généralisée

$$\sum_i |p_{1i} - p_{2i}|$$

(formule similaire à celle de l'indice de dissimilarité  $D$ , mais sans la division par 2)

- Distance euclidienne généralisée

$$\sqrt{\sum_i (p_{1i} - p_{2i})^2}$$

### Distances généralisées entre des vecteurs d'attributs quelconques

Attributs des objets comparés :

$x_{11}, x_{12}, \dots, x_{1n}$  pour le premier

$x_{21}, x_{22}, \dots, x_{2n}$  pour le second

Par exemple, une comparaison de quartiers d'une ville, caractérisés par...

$x_{j1}$  = pourcentage de la population de moins de 15 ans

$x_{j2}$  = taux de chômage

$x_{j3}$  = revenu familial moyen

etc.

- Distance rectilinéaire généralisée

$$\sum_i |x_{1i} - x_{2i}|$$

(formule similaire à celle de l'indice de dissimilarité, mais sans la division par 2)

- Distance euclidienne généralisée

$$\sqrt{\sum_i (x_{1i} - x_{2i})^2}$$

Mais si  $\{x_{1i}\}$  et  $\{x_{2i}\}$  ne sont pas des **distributions**, problème des **poids** (arbitraires ?)

Or le poids est fixé implicitement par les unités de mesure utilisées...



## L'INDICE DE DISSIMILARITÉ (EXEMPLE NUMÉRIQUE)

**Tableau de contingence : Emploi par zone et par branche**

BRANCHE	B1	B2	B3	Total
ZONE				
Z1	48	325	287	660
Z2	27	185	148	360
Z3	45	90	45	180
Total	120	600	480	1200

**Distribution de l'emploi entre zones**

BRANCHE	B1	B2	B3	Total
ZONE				
Z1	0,400	0,542	0,598	0,550
Z2	0,225	0,308	0,308	0,300
Z3	0,375	0,150	0,094	0,150
Total	1,000	1,000	1,000	1,000

**Comparaison de la répartition géographique  
des branches B1 et B2**

BRANCHE	B1	B2	Écart
ZONE			
Z1	0,400	0,542	0,142
Z2	0,225	0,308	0,083
Z3	0,375	0,150	-0,225
Total	1,000	1,000	0,000

**Mesure de la dissimilarité :**

$$D = \frac{1}{2} \sum_i |v_i - w_i|$$

$$D = \frac{|0,400 - 0,542| + |0,225 - 0,308| + |0,375 - 0,150|}{2} = 0,225$$

$$D = \frac{|0,142| + |0,083| + |-0,225|}{2} = 0,225$$

$D$  = la moitié de la distance de Manhattan (distance rectilinéaire)

## MESURE DE LA DISSIMILITUDE DANS UN TABLEAU DE CONTINGENCE

$x_{ij}$	nombre d'emplois de la branche $j$ dans la zone $i$
$x_{\bullet j} = \sum_i x_{ij}$	nombre total d'emplois de la branche $j$
$x_{i\bullet} = \sum_j x_{ij}$	nombre total d'emplois dans la zone $i$
$x_{\bullet\bullet} = \sum_i \sum_j x_{ij}$	nombre total d'emplois de toutes les branches dans toutes les zones
$p_{ij} = \frac{x_{ij}}{x_{\bullet\bullet}}$	fraction de l'emploi total global qui appartient à la branche $j$ et qui se trouve dans la zone $i$
$p_{\bullet j} = \sum_i p_{ij}$	fraction de l'emploi total global qui appartient à la branche $j$
$p_{i\bullet} = \sum_j p_{ij}$	fraction de l'emploi total global qui se trouve dans la zone $i$
$p_{j/i\bullet} = \frac{p_{ij}}{p_{i\bullet}}$	fraction de l'emploi total dans la zone $i$ qui appartient à la branche $j$
$p_{i/\bullet j} = \frac{p_{ij}}{p_{\bullet j}}$	fraction de l'emploi total de la branche $j$ qui se trouve dans la zone $i$

Dans l'exemple précédent, on mesure la dissimilitude entre

$$Q_1 = \begin{bmatrix} p_{1/\bullet 1} \\ p_{2/\bullet 1} \\ \vdots \\ p_{m/\bullet 1} \end{bmatrix} \text{ et } Q_2 = \begin{bmatrix} p_{1/\bullet 2} \\ p_{2/\bullet 2} \\ \vdots \\ p_{m/\bullet 2} \end{bmatrix}$$

En général, on compare les distributions

$$Q_h = \begin{bmatrix} p_{1/\bullet h} \\ p_{2/\bullet h} \\ \vdots \\ p_{m/\bullet h} \end{bmatrix} \text{ et } Q_k = \begin{bmatrix} p_{1/\bullet k} \\ p_{2/\bullet k} \\ \vdots \\ p_{m/\bullet k} \end{bmatrix}$$

ou les distributions

$$R_g = [p_{1/g\bullet} \quad p_{2/g\bullet} \quad \cdots \quad p_{n/g\bullet}] \text{ et } R_i = [p_{1/i\bullet} \quad p_{2/i\bullet} \quad \cdots \quad p_{n/i\bullet}]$$

## PROPRIÉTÉS DE L'INDICE DE DISSIMILARITÉ

1. Remplit les conditions d'une mesure de distance (c'est la moitié de la distance rectilinéaire)
2. Possède les 5 premières propriétés désirables d'une mesure d'inégalité, mais pas la dernière (il manque le principe de transfert de Pigou-Dalton ; Valeyre, 1993)
3. Domaine de variation (valeurs maximum et minimum)
  - $D = 0$  quand  $v_i = w_i$  pour tout  $i$  (les deux distributions sont identiques)
  - $D = 1$  quand il y a ségrégation complète :
    - soit  $v_i > 0$ , et alors,  $w_i = 0$
    - soit  $w_i > 0$ , et alors,  $v_i = 0$
4. Interprétation métaphorique (groupes parfaitement distincts) :  
 $D =$  fraction du groupe  $h$  qu'il faudrait déplacer pour que sa distribution soit identique à celle du groupe  $k$  ou vice-versa.
5.  $D$  est égal à l'écart vertical maximum entre la courbe de Lorenz et la diagonale.
6. Quand les données sont groupées,  $D$ , aussi bien que  $G$ , est sensible à la définition et au nombre de catégories utilisées (classes, zones).  
Cela implique notamment que l'agrégation d'une ou de plusieurs catégories peut entraîner une diminution de la valeur de l'indice de dissimilarité.
7. En tant que mesure de concentration spatiale, l'indice de dissimilarité, comme le Gini, ne tient aucun compte de la proximité dans l'espace des différentes zones de forte densité.
8. Ne s'applique pas à des données négatives (ex. : comparaison des variations de l'emploi).

## L'INDICE DE DISSIMILARITÉ ET LES PROPRIÉTÉS D'UNE MESURE DE DISTANCE

Propriétés d'une distance	Indice de dissimilarité $D$
Non négativité : $d(a,b) \geq 0$	OUI
Identité : $d(a,b) = 0$ si, et seulement si $a = b$	OUI
Symétrie : $d(a,b) = d(b,a)$	OUI $D = \frac{1}{2} \sum_i  v_i - w_i  = \frac{1}{2} \sum_i  w_i - v_i $
inégalité triangulaire : $d(a,c) \leq d(a,b) + d(b,c)$	OUI

**Normal :  $D$  est la demie de la distance rectilinéaire généralisée (distance de Manhattan)**

## L'INDICE DE DISSIMILARITÉ ET LES PROPRIÉTÉS D'UNE MESURE D'INÉGALITÉ

Propriétés d'une mesure d'inégalité	Indice de dissimilarité $D$
Une mesure d'inégalité doit prendre des valeurs non négatives.	OUI
Une mesure d'inégalité doit prendre la valeur zéro si, et seulement si, la distribution observée est identique à la distribution de référence.	OUI
Toutes les observations doivent être traitées de la même manière.	OUI
Mesure indépendante de la valeur moyenne de la variable examinée ou de la taille de la population dont on étudie la distribution.	OUI, puisque $D$ est calculé à partir de la distribution.
L'agrégation d'observations affichant le même degré de spécificité ne doit pas changer la valeur de la mesure.	OUI
Principe de transfert de Pigou-Dalton	NON

## L'INDICE DE DISSIMILARITÉ EXEMPLE DE SÉGRÉGATION TOTALE

ETHNIE	Indice de dissimilarité						Écart $ v_i - w_i $
	Nombres			Répartitions			
	Martiens	Terriens	Total	Martiens $v_i$	Terriens $w_i$	Total	
PLANÈTE							
TERRE	0	6	6	0,00	0,75	0,40	0,75
LUNE	0	2	2	0,00	0,25	0,13	0,25
MARS	3	0	3	0,43	0,00	0,20	0,43
JUPITER	4	0	4	0,57	0,00	0,27	0,57
TOTAL	7	8	15	1,00	1,00	1,00	

Indice de dissimilarité :

$$\frac{0,75 + 0,25 + 0,43 + 0,57}{2} = 1,00$$

Ainsi,  $D$  varie entre 0 et 1

**Et voilà pourquoi on divise par 2 !**

### INDICE DE DISSIMILARITÉ VALEUR MAXIMUM

**Démonstration que  $D = 1$  lorsqu'il y a ségrégation complète**

SOIT  $v_i = 0$ , et alors  $|v_i - w_i| = |0 - w_i| = w_i = 0 + w_i = v_i + w_i$

SOIT  $w_i = 0$ , et alors  $|v_i - w_i| = |v_i - 0| = v_i = v_i + 0 = v_i + w_i$

Il s'ensuit

$$D^{\max} = \frac{1}{2} \sum_i |v_i - w_i| = \frac{1}{2} \sum_i (v_i + w_i)$$

$$D^{\max} = \frac{1}{2} \left( \sum_i v_i + \sum_i w_i \right) = \frac{1+1}{2} = 1$$

## INTERPRÉTATION MÉTAPHORIQUE RENDRE LA DISTRIBUTION *B2* IDENTIQUE À *B1* (EXEMPLE NUMÉRIQUE)

### Comparaison de la répartition géographique des branches *B1* et *B2*

BRANCHE	<i>B1</i>	<i>B2</i>	Écart
ZONE			
Z1	0,400	0,542	0,142
Z2	0,225	0,308	0,083
Z3	0,375	0,150	-0,225
Total	1,000	1,000	0,000

« Excédents » de *B2* sur *B1* :

$$= 0,142 + 0,083 = 0,225 \text{ (Z1 et Z2)}$$

« Déficits » de *B2* par rapport à *B1* :

$$= 0,225 \text{ (Z3)}$$

**Interprétation métaphorique :**

« Il faut prendre 22,5 % (= *D*) des emplois de *B2*, dont 14,2 % dans *Z1* et 8,3 % dans *Z2* et il faut les déplacer vers *Z3* ».

**Ou, réciproquement :**

« Il faut prendre 22,5 % des emplois de *B1* dans *Z3* ("excédentaires") et les déplacer vers les autres zones : 14,2 % dans *Z1* et 8,3 % dans *Z2* ».

**Ou, en nombre d'emplois :**

- Si on déplace les emplois de *B2*, ce sont 22,5 % de 600 emplois = 135 emplois.
- Si on déplace les emplois de *B1*, ce sont 22,5 % de 120 emplois = 27 emplois.

<b>Mais il ne faut pas prendre la métaphore au pied de la lettre !</b>
--

## L'INDICE DE DISSIMILARITÉ ET LA COURBE DE LORENZ ÉCART MAXIMUM ENTRE LA COURBE ET LA DIAGONALE

L'écart entre la courbe de Lorenz et la diagonale est donné par  $Cv_k - Cw_k$

Pour quel  $k$  atteint-on la valeur maximum de  $Cv_k - Cw_k$  ?

Pour chaque  $k$ , on a  $Cv_k - Cw_k = \sum_{i=1}^k v_i - \sum_{i=1}^k w_i = \sum_{i=1}^k (v_i - w_i)$

Lorsque les observations sont en ordre croissant de spécificité, on a

$$\frac{w_1}{v_1} < \frac{w_2}{v_2} < \dots < \frac{w_n}{v_n}$$

Donc, pour les premières observations,  $v_i \geq w_i$  et pour les dernières,  $w_i \geq v_i$

Par conséquent, tant que  $v_i \geq w_i$ ,  $Cv_i - Cw_i \geq Cv_{i-1} - Cw_{i-1}$

Pour trouver le maximum, il suffit de n'additionner que les valeurs positives (qui viennent toutes avant les négatives) :  $MAX_k [Cv_k - Cw_k] = \sum_{\substack{i \text{ lorsque} \\ v_i > w_i}} (v_i - w_i)$

Mais puisque  $\sum_{i=1}^n (v_i - w_i) = 0$ , on a  $\sum_{\substack{i \text{ lorsque} \\ v_i > w_i}} (v_i - w_i) = - \sum_{\substack{i \text{ lorsque} \\ v_i < w_i}} (v_i - w_i)$

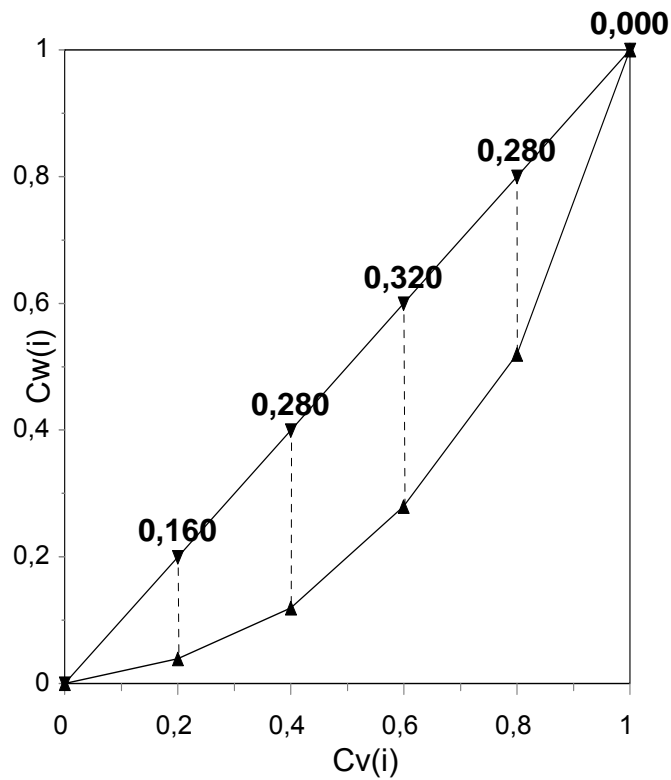
De plus,  $\sum_{\substack{i \text{ lorsque} \\ v_i > w_i}} (v_i - w_i) - \sum_{\substack{i \text{ lorsque} \\ v_i < w_i}} (v_i - w_i) = \sum_i |v_i - w_i|$ ,

de sorte que  $\sum_{\substack{i \text{ lorsque} \\ v_i > w_i}} (v_i - w_i) = - \sum_{\substack{i \text{ lorsque} \\ v_i < w_i}} (v_i - w_i) = \frac{1}{2} \sum_i |v_i - w_i|$

**Donc,**

$$MAX_k [Cv_k - Cw_k] = \sum_{\substack{i \text{ lorsque} \\ v_i > w_i}} (v_i - w_i) = \frac{1}{2} \sum_i |v_i - w_i| = D$$

## ÉCART ENTRE LA COURBE DE LORENZ ET LA DIAGONALE



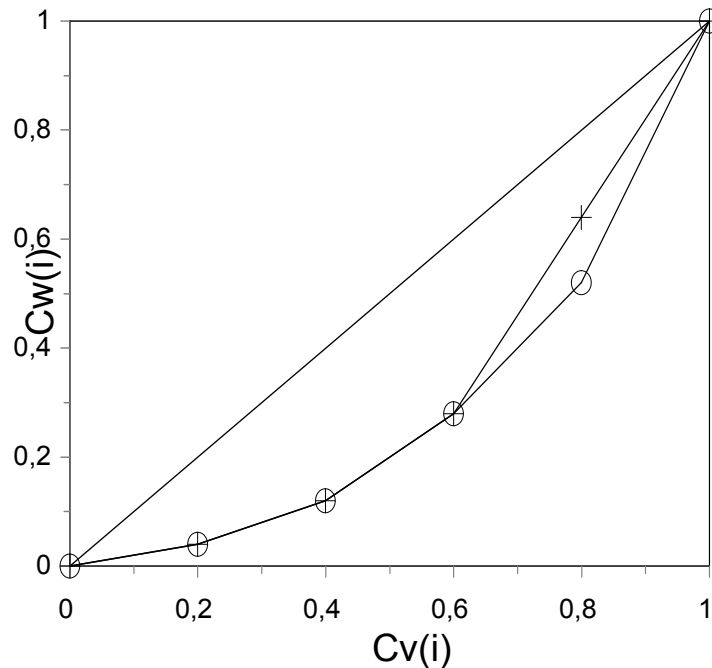
Zone	v(i)	w(i)	w(i)/v(i)	Cv(i)	Cw(i)	Cv(i)-Cw(i)	v(i)-w(i)
A	0,20	0,04	0,60	0,20	0,04	0,16	0,16
B	0,20	0,08	0,90	0,40	0,12	0,28	0,12
C	0,20	0,16	0,92	0,60	0,28	0,32	0,04
D	0,20	0,24	0,96	0,80	0,52	0,28	0,04
E	0,20	0,48	1,14	1,00	1,00	0,00	0,28
Total	1,00	1,00					0,64

Indice de dissimilarité  $D = 0,32$

Coefficient Gini = 0,416



## INSENSIBILITÉ DE L'INDICE DE DISSIMILARITÉ À CERTAINS CHANGEMENTS



Distribution «O»							
Zone	$v(i)$	$w(i)$	$w(i)/v(i)$	$Cv(i)$	$Cw(i)$	$Cv(i)-Cw(i)$	$ v(i)-w(i) $
A	0,20	0,04	0,2	0,20	0,04	0,16	0,16
B	0,20	0,08	0,4	0,40	0,12	0,28	0,12
C	0,20	0,16	0,8	0,60	0,28	0,32	0,04
D	0,20	0,24	1,2	0,80	0,52	0,28	0,04
E	0,20	0,48	2,4	1,00	1,00	0,00	0,28
Total	1,00	1,00					0,64

Distribution «+»							
Zone	$v(i)$	$w(i)$	$w(i)/v(i)$	$Cv(i)$	$Cw(i)$	$Cv(i)-Cw(i)$	$ v(i)-w(i) $
A	0,20	0,04	0,60	0,20	0,04	0,16	0,16
B	0,20	0,08	0,90	0,40	0,12	0,28	0,12
C	0,20	0,16	0,92	0,60	0,28	0,32	0,04
D	0,20	0,36	1,80	0,80	0,64	0,16	0,16
E	0,20	0,36	1,80	1,00	1,00	0,00	0,16
Total	1,00	1,00					0,64

Indice de dissimilarité  $D = 0,32$

Indice de Gini = 0,416 pour la distribution «O»

Indice de Gini = 0,368 pour la distribution «+»

## L'INDICE DE DISSIMILARITÉ COMME MESURE DE LA CONCENTRATION DE LA POPULATION

Ville de Montréal (54 quartiers de planification), population Recensement 1991

Quartier	Données		Densité hab/km <sup>2</sup>	Répartitions		Écart absolu
	Pop. 1991	Superf. km <sup>2</sup>		Pop.	Superf.	
11	29469	1,65	17860	2,90%	0,88%	0,0201
8	10604	0,72	14728	1,04%	0,38%	0,0066
18	27022	2,03	13311	2,66%	1,08%	0,0157
34	24258	1,85	13112	2,38%	0,99%	0,0140
13	30314	2,39	12684	2,98%	1,28%	0,0170
35	14187	1,24	11441	1,39%	0,66%	0,0073
31	19652	1,73	11360	1,93%	0,92%	0,0101
33	15752	1,40	11251	1,55%	0,75%	0,0080
42	25495	2,32	10989	2,51%	1,24%	0,0127
15	19126	1,75	10929	1,88%	0,93%	0,0095
16	15030	1,38	10891	1,48%	0,74%	0,0074
29	15606	1,46	10689	1,53%	0,78%	0,0075
9	21348	2,02	10568	2,10%	1,08%	0,0102
32	14737	1,48	9957	1,45%	0,79%	0,0066
40	20350	2,15	9465	2,00%	1,15%	0,0085
14	15973	1,80	8874	1,57%	0,96%	0,0061
10	14165	1,65	8585	1,39%	0,88%	0,0051
27	11592	1,41	8221	1,14%	0,75%	0,0039
17	16167	2,00	8084	1,59%	1,07%	0,0052
30	29664	3,69	8039	2,91%	1,97%	0,0095
45	24738	3,23	7659	2,43%	1,72%	0,0071
46	19880	2,60	7646	1,95%	1,39%	0,0057
39	34906	4,85	7197	3,43%	2,59%	0,0084
51	8452	1,20	7043	0,83%	0,64%	0,0019
23	18672	2,67	6993	1,83%	1,43%	0,0041
12	14980	2,21	6778	1,47%	1,18%	0,0029
6	16785	2,48	6768	1,65%	1,32%	0,0033
19	11499	1,75	6571	1,13%	0,93%	0,0020
4	23636	3,70	6388	2,32%	1,98%	0,0035
44	18699	2,96	6317	1,84%	1,58%	0,0026
24	13665	2,22	6155	1,34%	1,19%	0,0016
21	20564	3,62	5681	2,02%	1,93%	0,0009
48	17038	3,02	5642	1,67%	1,61%	0,0006
41	20092	3,59	5597	1,97%	1,92%	0,0006
5	18478	3,36	5499	1,82%	1,79%	0,0002
49	14687	2,73	5380	1,44%	1,46%	0,0001
20	27819	5,22	5329	2,73%	2,79%	0,0005
43	24957	4,84	5156	2,45%	2,58%	0,0013
3	18052	3,56	5071	1,77%	1,90%	0,0013
28	17764	3,56	4990	1,75%	1,90%	0,0015
2	25181	5,25	4796	2,47%	2,80%	0,0033
26	19073	4,01	4756	1,87%	2,14%	0,0027
22	9651	2,18	4427	0,95%	1,16%	0,0022
38	12512	3,16	3959	1,23%	1,69%	0,0046
7	22660	5,84	3880	2,23%	3,12%	0,0089
1	22613	5,85	3865	2,22%	3,12%	0,0090
52	35098	9,50	3695	3,45%	5,07%	0,0162
50	14403	4,07	3539	1,42%	2,17%	0,0076
47	13111	4,45	2946	1,29%	2,38%	0,0109
54	47534	19,04	2497	4,67%	10,16%	0,0549
37	3546	2,06	1721	0,35%	1,10%	0,0075
25	4009	4,28	937	0,39%	2,28%	0,0189
53	11970	13,92	860	1,18%	7,43%	0,0625
36	431	4,24	102	0,04%	2,26%	0,0222
<b>Total</b>	<b>1017666</b>	<b>187,34</b>	<b>5432</b>	<b>100,00%</b>	<b>100,00%</b>	<b>0,472</b>

**Indice de dissimilarité : 0,236**

## LE COEFFICIENT DE LOCALISATION N'EST PAS L'INDICE DE DISSIMILARITÉ

**Bien qu'ils se calculent de la même manière, l'indice de dissimilarité et le coefficient de localisation sont différents !**

### Emploi par zone et par branche

BRANCHE	B1	B2	B3	B1 + B2	Total
ZONE					
Z1	48	325	287	373	660
Z2	27	185	148	212	360
Z3	45	90	45	135	180
Total	120	600	480	720	1200

### Distribution de l'emploi entre zones

BRANCHE	B1	B2	B3	B1 + B2	Total
ZONE					
Z1	0,400	0,542	0,598	0,518	0,550
Z2	0,225	0,308	0,308	0,294	0,300
Z3	0,375	0,150	0,094	0,188	0,150
Total	1,000	1,000	1,000	1,000	1,000

**Comparaison de la distribution géographique de la branche B3 avec celle de l'ensemble des trois branches, puis avec la somme de B1 et B2**

BRANCHE	B3	Total	Dif.absol.	B1 + B2	Dif.absol.
ZONE					
Z1	0,598	0,550	0,048	0,518	0,080
Z2	0,308	0,300	0,008	0,294	0,014
Z3	0,094	0,150	0,056	0,188	0,094
Total	1,000	1,000	0,113	1,000	0,188

#### Mesure de la dissimilarité :

$$\text{Indice de dissimilarité } D = \frac{|0,080| + |0,014| + |-0,094|}{2} = 0,094$$

$$\text{Coef. de localisation } CL = \frac{|0,048| + |0,008| + |-0,056|}{2} = 0,056$$

$$CL = \left(1 - \frac{480}{1200}\right) D = 0,6 \times 0,094 = 0,056$$

## APPLICATION DE L'INDICE DE DISSIMILARITÉ À UNE DICHOTOMIE

**Bien qu'ils se calculent de la même manière,  
 l'indice de dissimilarité et le coefficient de localisation sont différents !**

$$CL = \frac{1}{2} \sum_i |p_{i \bullet h} - p_{i \bullet}|$$

$$D = \frac{1}{2} \sum_i |p_{i \bullet h} - p_{i \bullet k}|$$

$$CL = (1 - p_{\bullet h})D$$

**Démonstration :**

$D$  est appliqué à une dichotomie. Donc

$$D = \frac{1}{2} \sum_i |p_{i \bullet h} - p_{i \bullet k}| = \frac{1}{2} \sum_i \left| p_{i \bullet h} - \frac{p_{i \bullet} - p_{ih}}{1 - p_{\bullet h}} \right|$$

$$D = \frac{1}{2(1 - p_{\bullet h})} \sum_i |p_{i \bullet h}(1 - p_{\bullet h}) - p_{i \bullet} + p_{ih}|$$

$$D = \frac{1}{2(1 - p_{\bullet h})} \sum_i |p_{i \bullet h} - p_{i \bullet h} p_{\bullet h} - p_{i \bullet} + p_{ih}|$$

$$D = \frac{1}{2(1 - p_{\bullet h})} \sum_i |p_{i \bullet h} - p_{ih} - p_{i \bullet} + p_{ih}|$$

$$D = \frac{1}{2(1 - p_{\bullet h})} \sum_i |p_{i \bullet h} - p_{i \bullet}| = \frac{CL}{(1 - p_{\bullet h})}$$

## L'INDICE DE DISSIMILARITÉ ET COEFFICIENT DE LOCALISATION

### EXEMPLE DE SÉGRÉGATION TOTALE

#### Indice de dissimilarité

ETHNIE	Nombres			Répartitions			Écart $ p_{i/\bullet h} - p_{i/\bullet k} $
	Martiens $x_{11}$	Terriens $x_{12}$	Total $x_{11} + x_{12}$	Martiens $p_{i/\bullet 1}$	Terriens $p_{i/\bullet 2}$	Total $p_{i\bullet}$	
PLANÈTE							
TERRE	0	6	6	0,00	0,75	0,40	0,75
LUNE	0	2	2	0,00	0,25	0,13	0,25
MARS	3	0	3	0,43	0,00	0,20	0,43
JUPITER	4	0	4	0,57	0,00	0,27	0,57
TOTAL	7	8	15	1,00	1,00	1,00	

Indice de dissimilarité :

$$\frac{0,75 + 0,25 + 0,43 + 0,57}{2} = 1,00$$

#### Coefficient de localisation

ETHNIE	Nombres		Répartitions		Écart $ v_i - w_i $
	Martiens $x_i$	Total $y_i$	Martiens $v_i$	Total $w_i$	
PLANÈTE					
TERRE	0	6	0,00	0,40	0,40
LUNE	0	2	0,00	0,13	0,13
MARS	3	3	0,43	0,20	0,23
JUPITER	4	4	0,57	0,27	0,30
TOTAL	7	15	1,00	1,00	

Coefficient de localisation :

$$\frac{0,40 + 0,13 + 0,23 + 0,30}{2} = 0,53 = 1 - \frac{7}{15}$$

= fraction de non-Martiens dans la population = fraction de Terriens

#### Indice de discrimination

Indice de discrimination :

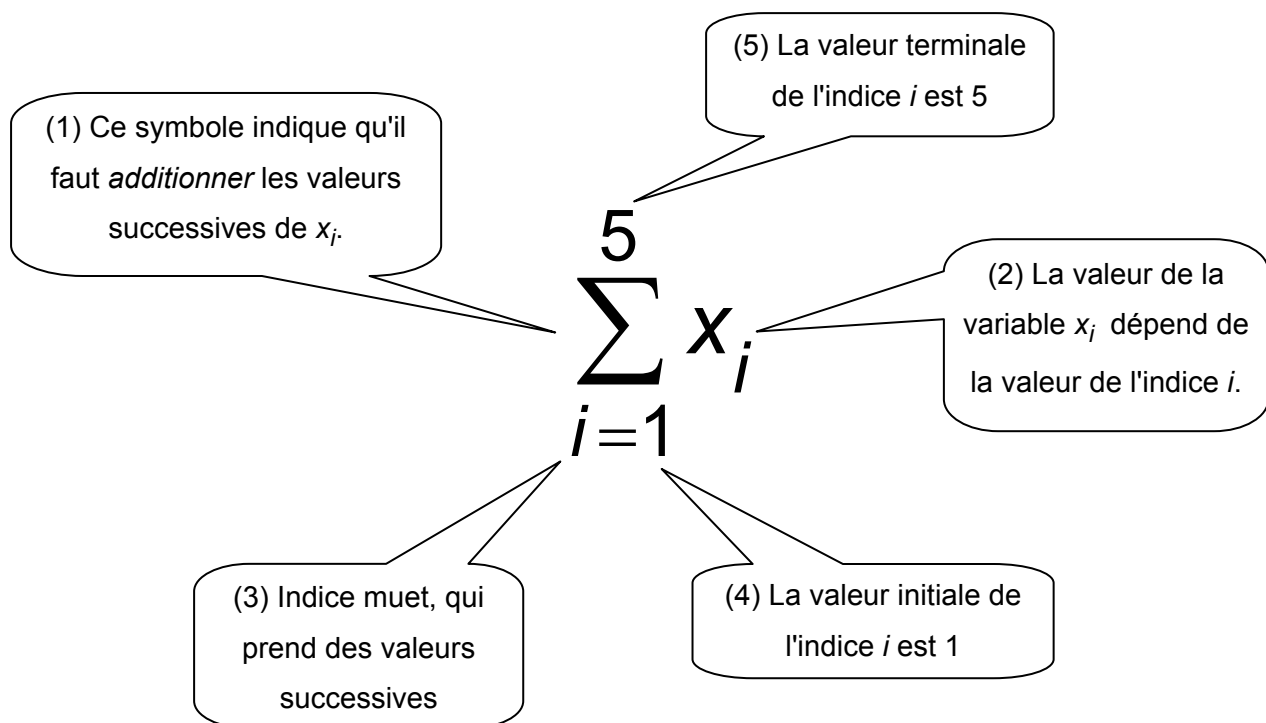
$$\frac{\left( \frac{0,40 + 0,13 + 0,23 + 0,30}{2} \right)}{0,53} = 1,00$$

## L'OPÉRATEUR SOMMATION (1)

L'opérateur sommation est...

- une façon compacte d'écrire une somme
- lorsque les termes successifs peuvent s'écrire sous la forme d'une expression générale
- qui varie en fonction d'un indice.

« La somme des  $x_i$  pour  $i$  variant de 1 à 5 »



**Exemples :**

$$\sum_{i=1}^5 x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

Et de façon elliptique

$$\sum_i x_i = x_1 + x_2 + \dots + x_n$$

## L'OPÉRATEUR SOMMATION (2)

Le choix de la lettre qui sert à représenter l'indice muet est parfaitement arbitraire :

$$\sum_{i=1}^n x_i = \sum_{j=1}^n x_j = \sum_{k=1}^n x_k = x_1 + x_2 + \dots + x_n$$

Le choix des valeurs initiale et terminale est arbitraire :

$$\sum_{i=1}^n x_i = \sum_{i=0}^{n-1} x_{i+1} = \sum_{i=2}^{n+1} x_{i-1} = x_1 + x_2 + \dots + x_n$$

Dans certains cas, la notation permet de connaître directement la valeur de chacun des termes de la sommation :

$$\sum_{t=1}^n t^2 = 1^2 + 2^2 + \dots + n^2$$

$$\sum_{k=1}^K \binom{1}{k} = \binom{1}{1} + \binom{1}{2} + \binom{1}{3} + \dots + \binom{1}{K}$$

Expressions où l'indice joue un double rôle :

$$\sum_{j=0}^n a_j x^j = a_0 x^0 + a_1 x^1 + a_2 x^2 + \dots + a_n x^n$$

Sommes infinies

$$\sum_{j=1}^{\infty} x_j = x_1 + x_2 + \dots + x_n + \dots$$

**Règles de base d'utilisation de l'opérateur sommation sont les suivantes**

1.  $\sum_{i=1}^n c = n c$

2.  $\sum_{i=1}^k x_i + \sum_{i=k+1}^n x_i = \sum_{i=1}^n x_i$

3.  $\sum_{i=1}^n (c x_i) = c \left( \sum_{i=1}^n x_i \right)$

4.  $\sum_{i=1}^t (x_i + y_i) = \sum_{i=1}^t x_i + \sum_{i=1}^t y_i$

## L'OPÉRATEUR SOMMATION (3)

### Sommations doubles

$$\begin{array}{cccc} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nm} \end{array}$$

### La cinquième règle

$$5. \sum_{i=1}^n \sum_{j=1}^m t_{ij} = \sum_{j=1}^m \sum_{i=1}^n t_{ij}$$

### Sommations doubles de tableaux triangulaires

$$\begin{array}{cccc} a_{11} & & & \\ a_{21} & a_{22} & & \\ a_{31} & a_{32} & a_{33} & \ddots \\ \vdots & \vdots & \vdots & \ddots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{array}$$

$$\sum_{i=1}^n \sum_{j=1}^i a_{ij} = \sum_i \sum_{j \leq i} a_{ij} = \sum_{j=1}^n \sum_{i=j}^n a_{ij} = \sum_j \sum_{i \geq j} a_{ij}$$

(noter la différence entre  $>$  et  $\geq$ )

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^i a_{ij} - \sum_{i=1}^n a_{ii} &= \sum_i \sum_{j \leq i} a_{ij} - \sum_i a_{ii} = \sum_i \sum_{j < i} a_{ij} \\ &= \sum_{j=1}^n \sum_{i=j}^n a_{ij} - \sum_{i=1}^n a_{ii} = \sum_j \sum_{i \geq j} a_{ij} - \sum_i a_{ii} = \sum_i \sum_{j > i} a_{ij} \end{aligned}$$