Pierre Bailly
Christine Carrère

Statistiques descriptives
Cours

Collection « Libres Cours Économie »

Presses universitaires de Grenoble
BP 47 – 38040 Grenoble cedex 9
Tél. : 04 76 82 56 52 – pug@pug.fr / www.pug.fr

Chapitre I

Les outils

Nous présentons quatre thèmes dans ce chapitre : les nomenclatures et les types de variable, les tableaux statistiques, les représentations graphiques, l'utilisation des pourcentages et des taux.

Les concepts de base

Avant tout calcul statistique, il est nécessaire de disposer de données. Pour atteindre cet objectif, il est impératif de définir très précisément la population sur laquelle s'effectue l'enquête et les variables que l'on cherche à appréhender. Le type de ces variables conditionne les traitements statistiques qu'elles seront susceptibles de subir.

La population et les unités statistiques

Dans le vocabulaire statistique, une population est un ensemble dont chaque élément est un individu ou une unité statistique. Les termes de population et d'individus sont employés aussi bien lorsqu'il s'agit d'un ensemble d'être humains : la population résidente en France, les salariés d'une entreprise… que d'un ensemble d'objets inanimés : la production automobile pour une année, le stock des machines à une date donnée, et même d'ensembles abstraits ou des événements : ensemble des jours d'une année, la série du revenu national depuis vingt ans… Chaque observation porte sur une unité statistique.

La population soumise à l'analyse statistique doit être définie avec précision afin que l'ensemble considéré soit déterminé sans ambiguïté, de sorte qu'un individu quelconque puisse y être affecté sans incertitude. La population française au premier janvier 1996 : il faut indiquer si les étrangers résidant en France sont inclus et comment sont comptabilisés les Français résidants à l'étranger. Il faudra alors préciser la signification de résider. Comment définir les personnes employées dans une entreprise au premier octobre 1995 ? Faut-il inclure les travailleurs à domicile, les travailleurs à temps partiel, les travailleurs intérimaires, les stagiaires, les apprentis, les travailleurs « au noir » ? Doit-on comprendre les travailleurs absents pour maladie, congé annuel ou détachement ? L'effectif présent

diffère en général de l'effectif théorique, celui des personnes juridiquement salariées de l'entreprise. Les règles qui définissent l'ensemble à étudier permettent de dire sans ambiguïté si une unité appartient ou non au domaine.

Les caractères et les modalités

Pour décrire une population, on classe les individus selon certains attributs que l'on appelle des caractères (sexe) ou des variables (âge). Il est indispensable de ne retenir que les caractères les plus pertinents pour pouvoir décrire une population correctement. Il convient de ne retenir qu'un nombre restreint de caractères pour obtenir une description synthétique. Le caractère est un critère de classement, il peut présenter plusieurs situations différentes, il prend plusieurs modalités. Les deux modalités du caractère sexe sont : masculin, féminin. Le nombre de modalités d'un caractère dépend de l'information disponible et du but de l'étude. Par exemple : l'état matrimonial peut comprendre quatre modalités : célibataire, marié, veuf, divorcé ou deux modalités marié, non marié. Chaque individu de la population présente une et une seulement des modalités du caractère. Les modalités d'un caractère constituent une nomenclature, elles sont incompatibles et exhaustives, elles déterminent une partition de l'ensemble. Une unité statistique peut présenter plusieurs caractères. Tous les individus appartenant à un même ensemble sont équivalents du point de vue du problème étudié. Le type de ces variables conditionne les traitements statistiques qu'elles seront susceptibles de subir.

Les caractères qualitatifs

Les caractères qualitatifs ou variables nominales ou variables catégorielles ont des attributs dont les différentes modalités ne sont ni mesurables, ni repérables Elles sont constatées. Avec l'usage de l'informatique, on utilise parfois le terme de données qualitatives. Le caractère se subdivise en catégories ou en modalités de la variable auxquelles seront attachées un effectif et une fréquence. C'est le cas pour le sexe, l'état matrimonial, la qualification professionnelle. Les modalités d'un caractère constituent les rubriques d'une nomenclature. Ce sont des noms ou des étiquettes permettant d'identifier une caractéristique de chaque élément. Il n'est pas toujours possible d'y établir un ordre. Les réponses peuvent être rangées dans une modalité particulière. Un caractère qualitatif peut-être nominal ou ordinal.

Les caractères qualitatifs nominaux

Une variable nominale décrit un nom ou une catégorie. Les différentes occurrences de la variable sont nominales. Les noms ou les catégories possibles ne suivent pas un ordre naturel. La plupart du temps, la présentation des modalités de la variable ne présuppose aucun ordre, si ce n'est l'ordre alphabétique.

Les caractères qualitatifs ordonnés ou variables qualitatives ordonnées

Certaines variables appellent naturellement un ordre dans le rangement de leurs catégories : niveau de formation, … Elles sont repérables selon un type d'échelle plus ou moins légitime. Les catégories pourront alors donner lieu à un codage par les rangs qui ouvrira une autre gamme de traitements possibles proches de ceux des variables quantitatives. Dans le cas d'une nomenclature de formation, le classement est fonction du nombre théorique d'années d'étude nécessaires pour acquérir le niveau de formation. C'est de ce point de vue une variable quantitative repérable, dans quelle mesure le niveau I est-il supérieur au niveau III (comparaison d'un doctorat et d'un BTS).

Un caractère ordinal est un caractère qualitatif dans lequel les modalités possibles peuvent être classées dans un ordre spécifique ou dans un ordre naturel quelconque. Dans le tableau, le caractère « comportement » est ordinal parce que la modalité « Excellent » est meilleure que la modalité « Très bon », etc. On n'y trouve un certain ordre naturel, mais celui-ci est limité par le fait que nous ne savons pas dans quelle mesure le comportement « Excellent » est meilleur que le comportement « Très bon ».

Classement des élèves selon le comportement

Comportement	Nombre d'élèves
Excellent	5
Très bon	12
Bon	10
Mauvais	2
Très mauvais	1

Variables textuelles

Une variable textuelle est une variable qui met en jeu des mots, des expressions langagières, voire des phrases qu'on ne peut réduire à des codes arbitraires, même si ceux-ci sont ordonnés. Il y a éventuellement un travail de préparation du texte, surtout s'il s'agit d'une transcription de textes oraux. En particulier, on peut lemmatiser c'est-à-dire restreindre aux lemmes (passer en minuscule, au masculin singulier, à l'infinitif…).

Une variable textuelle d'énonciation (ou semi textuelle) ne met en jeu que des expressions que l'on traitera par comptage alors qu'une variable textuelle « com-

plète » utilise des phrases, des segments et on calcule pour des mots, lemmes ou expressions à la fois des fréquences et des environnements. Ainsi la profession d'un adulte est une variable textuelle d'énonciation alors que la réponse à la question « *pourquoi y a-t-il du chômage en France ?* » est une variable textuelle « complète ».

La plupart des caractères qualitatifs requièrent une convention de définition ; c'est l'objet de la construction des nomenclatures.

Les caractères qualitatifs usuels et les nomenclatures

Elles constituent des outils de classement des caractères qualitatifs. Les différentes modalités d'un caractère constituent une nomenclature. Les nomenclatures sont le résultat d'un classement raisonné de modalités. La plupart du temps, la présentation des modalités de la variable ne présuppose aucun ordre, si ce n'est l'ordre alphabétique.

Les différentes occurrences de la variable sont nominales, nous utilisons le terme de modalité. Les différentes modalités d'un caractère constituent une nomenclature. Les nomenclatures sont le résultat d'un classement raisonné des modalités. Les organismes publics de statistiques ont défini, dans un but de clarté et d'homogénéité, les catégories des variables qu'ils utilisent régulièrement. Ces nomenclatures sont d'usage obligatoire au sein des administrations et recommandées pour les autres agents.

Les nomenclatures de l'INSEE

Elles sont nombreuses depuis la nomenclature des produits, d'activités, de catégories sociales ou de formation.

Les nomenclatures d'EUROSTAT

L'élargissement des mesures statistiques à l'Europe a nécessité la création d'un système de codage harmonisé. Le service des statistiques des Communautés européennes a construit des nomenclatures qui permettent de décrire les réalités économiques et sociales de l'ensemble de pays de l'Union européenne.

Exemple de nomenclature :

Architecture de la nomenclature (données de 1998)			
Nomenclature	**Niveau de ventilation**	**Codage**	**Nombre**
Système harmonisé (SH)	Section	Un chiffre	21
	Chapitre	Deux chiffres	99
	Position	Quatre chiffres	1241
	Sous-position	Six chiffres	5113
Nomenclature combinée (NC)	Sous-position	Huit chiffres	10587
CTCI	Section	Un chiffre	10
	Chapitre	Deux chiffres	67
	Position	Trois chiffres	261
	Sous-position	Quatre chiffres	1033
	Sous-position	Cinq chiffres	3118
Exemple de classement d'un produit dans la nomenclature combinée :			
Chapitre 10 du SH : céréales Position 10 06 du SH : riz Sous-position 10 06 20 du SH : riz décortiqué Sous-position 10 06 20 11 de la NC : riz décortiqué étuvé à grains ronds			

Les statistiques des échanges et des biens. Guide de l'utilisateur Eurostat 2000

Les variables quantitatives ou numériques

Les variables quantitatives ou variables statistiques ont des attributs dont les modalités sont exprimées sous forme numérique. Une variable quantitative est soit mesurable soit repérable. À chaque unité statistique est associée un nombre : la valeur de la variable. Pour l'analyse statistique, il est habituel de distinguer les variables discrètes et les variables continues. Une variable est discrète quand les valeurs sont obtenues par dénombrement, les modalités sont exprimées par un nombre et prennent un nombre fini de valeur. Lorsque la variable peut prendre toutes les valeurs à l'intérieur d'un intervalle, la variable est dite quantitative continue. Une variable statistique peut être discrète ou continue.

Variables numériques discrètes

Une variable dont les valeurs sont obtenues par dénombrement est une variable discrète. C'est par exemple le cas du nombre d'enfants. Une variable statistique est discrète ou discontinue lorsqu'elle ne peut prendre que certaines valeurs iso-

lées (valeurs prises dans $\mathbb{N}$ plus rarement dans $\mathbb{Z}$). C'est le cas du nombre de personnes qui composent un ménage. Un caractère discret peut prendre une infinité de valeurs dénombrables, il peut aussi n'en prendre que quelques-unes : le nombre d'enfants par familles qui est nécessairement un entier.

Dans cette situation, les modalités peuvent être exprimées par un nombre et prennent un nombre fini de valeurs, la variable est dite quantitative discrète. Certaines variables discrètes, comme le nombre de salariés d'une entreprise, pouvant prendre un très grand nombre de valeurs à l'intérieur d'un intervalle de grande amplitude, elles seront traitées comme des variables continues.

Variable statistique continue

Lorsque la variable peut prendre toutes les valeurs à l'intérieur d'un intervalle, la variable est dite quantitative continue. Les unités statistiques prenant sur ce type de variable un nombre très important de valeurs, il est nécessaire que les valeurs de la variable soient regroupées en classes. Certaines variables discrètes, comme le nombre de salariés d'une entreprise, pouvant prendre un très grand nombre de valeurs, elles seront traitées comme des variables continues.

Une variable statistique continue peut *a priori* prendre toutes les valeurs à l'intérieur d'un intervalle de $\mathbb{R}$: par exemple les salaires, le revenu par habitant. Le nombre de modalités possibles est alors infini. Pour obtenir un nombre fini de modalités, les valeurs sont regroupées en classe. La taille d'un individu est une variable continue, les revenus sont considérés comme continus ce qui n'est pas tout à fait juste puisqu'ils ne peuvent prendre que des valeurs exprimées en centimes.

Les valeurs d'une variable continue sont mesurables ou repérables, avec un degré de précision déterminé qui n'est pas toujours connu pour les données économiques et sociales.

En pratique, la distinction entre variables discrètes et variables continues est conventionnelle. La précision d'une mesure est toujours limitée et les résultats seront toujours donnés sous forme d'un nombre fini d'observations. La production d'acier, par exemple, sera donnée en millions de tonnes ou en milliers de tonnes. Inversement, si une variable discrète peut prendre un grand nombre de valeurs, deux valeurs voisines apparaissent comme proches. C'est le cas du nombre de salariés dans une entreprise. Elle sera alors traitée comme une variable continue. La distinction repose, dans la pratique, sur le fait que les variables se présentent ou non groupées en classe.

Les classes

Les unités statistiques prenant sur ce type de variable un nombre très important de valeurs, il est nécessaire que les valeurs de la variable soient regroupées en

classes avant tout traitement. Le choix des classes répond en général aux exigences suivantes :

– elles ne doivent pas être trop nombreuses sinon il y aurait une difficulté de compréhension ;

– elles ne doivent pas être trop peu nombreuses car il y aurait perte d'information ;

– il ne doit pas y avoir de classe vide.

Le rangement des données, selon un ordre précis, est insuffisant dès que le nombre de données est grand. Pour étudier une variable continue, il faudra parfois regrouper les données[1] sous une forme qui permette de ne pas perdre l'essentiel de l'information. Il sera nécessaire de construire des *classes* regroupant les valeurs en un nombre fini de modalités. Le regroupement ainsi effectué permet d'obtenir une distribution des fréquences ou des effectifs. Chaque classe aura un certain effectif ; certains auteurs utilisent le terme de fréquence absolue. Les calculs statistiques utiliseront les centres de classes comme représentatifs de l'ensemble de la classe. Les classes de valeurs possibles constituent les modalités du caractère étudié.

Les classes peuvent avoir une amplitude variable ou constante. L'effectif de chaque classe ne doit pas être trop réduit pour éviter les fluctuations accidentelles. La variable « âge » est souvent subdivisée en classes d'amplitude de 5 ans, 0 moins de 5 ans, 5 ans moins de 10 ans etc. 0, 5, 10 sont les extrémités des classes.

Le choix du nombre de classes et de leur amplitude est fonction de l'effectif de la population étudiée, sans que l'effectif de chacune soit trop faible afin d'éliminer les variations accidentelles. Il dépend aussi de la nature de l'étude. En pratique, l'application de quelques règles permet la construction des classes d'une distribution. Pour rendre les calculs significatifs, tout en préservant la compréhension de la distribution, le nombre de classes doit être compris entre 5 et 15. Les classes doivent être agencées de telle sorte que toutes les informations soient incluses et que chaque observation se retrouve dans une et une seule classe. Les classes constituent ainsi une partition de l'ensemble considéré. Les amplitudes des classes ne doivent pas être trop différentes.

La définition des classes

Les limites de classes doivent être sans équivoque. La présentation suivante est insatisfaisante.

1. Cela dépend de l'étude, pour certains indicateurs on utilise les données non groupées.

Nombre de salariés par entreprises :
– 0 à 10
– 10 à 50
– …

L'écriture la plus satisfaisante est la suivante :
– [0, 10[
– [10, 50[
– …

Le nombre de classes à retenir dépend de la précision des mesures et de l'effectif de la population étudiée.

L'amplitude de classe

Le choix des amplitudes de classe est déterminé par le souci d'obtenir des effectifs comparables d'une classe à l'autre.

La valeur de l'amplitude d'une classe est calculée par la différence entre les valeurs de la borne supérieure et celle de la borne inférieure. L'amplitude est donc pour la deuxième classe de $[10,50[= 40$ salariés. Il arrive que la borne inférieure de la première classe et la borne supérieure de la dernière ne soient pas données. Pour estimer les bornes absentes, nous disposons de deux solutions. Tout d'abord réfléchir à ce que pourrait être la valeur de cette borne (ici pour la première classe 0 semble une solution satisfaisante). Sinon, nous donnerons à la première classe l'amplitude de la seconde et à la dernière classe l'amplitude de l'avant-dernière (attention cependant à ne pas avoir des valeurs aberrantes).

Les centres de classe

Pour mener des calculs statistiques sur des séries classées, les classes sont réduites à une seule donnée : le centre de classe. Cela revient à considérer que tous les individus d'une classe peuvent être décrits par ce centre de classe. Le centre de classe c_i se calcule simplement :

$$c_i = \frac{x_i + x_{i+1}}{2}$$

avec x_i la borne inférieure de la classe i et x_{i+1} la borne supérieure de celle-ci.

Il faut faire attention aux extrémités de classe retenues, elles peuvent appartenir à la classe suivante ou la classe précédente.

L'amplitude de la dernière classe est supposée égale à l'avant-dernière, conformément à la règle énoncée. Le centre de classe de la classe i est obtenu en pre-

nant pour borne inférieure celle de la classe i et pour borne supérieure la borne inférieure de la classe $i+1$.

Les tableaux statistiques

Ils constituent le moyen le plus sûr de pouvoir répondre aux questions posées de par leur systématisme. Sauf cas exceptionnels, les données statistiques sont présentées sous forme de tableau. D'une part, cela permet d'appréhender l'information qui est synthétisée et d'autre part facilite ou rend possible les calculs.

Tableau statistique pour une variable qualitative

	Effectifs n_i	Fréquences f_i	Pourcentages p_i	Fréquences cumulées F_i
Catégorie 1	n_1	f_1	p_1	F_1
Catégorie i	n_i	$f_i = \frac{n_i}{N}$	$p_i = f_i \cdot 100$	$F_i = \sum_{k=1}^{k=i} f_k$
Catégorie m	n_m	f_m	p_m	$F_m = 1$
	$n = \sum_{i=1}^{i=m} n_i$	$\sum_{i=1}^{i=m} f_i = 1$	$\sum_{i=1}^{i=m} p_i = 100$	

Tableau statistique pour une variable quantitative discrète

Valeurs de la variable x_i	Effectifs n_i	Fréquences f_i	Pourcentages p_i	Fréquences cumulées F_i
x_1	n_1	f_1	p_1	F_1
x_i	n_i	$f_i = \frac{n_i}{N}$	$p_i = f_i \cdot 100$	$F_i = \sum_{k=1}^{k=i} f_k$
x_m	n_m	f_m	p_m	$F_m = 1$
	$n = \sum_{i=1}^{i=m} n_i$	$\sum_{i=1}^{i=m} f_i = 1$	$\sum_{i=1}^{i=m} p_i = 100$	

Tableau statistique pour une variable quantitative continue

Classes	Centres des classes c_i	Effectifs n_i	Fréquences f_i	Pourcentages p_i	Fréquences cumulées F_i
$[b_1 ; b_2[$	c_1	n_1	f_1	p_1	F_1
$[b_i ; b_{i+1}[$	c_i	n_i	$f_i = \frac{n_i}{N}$	$p_i = f_i \cdot 100$	$F_i = \sum_{k=1}^{k=i} f_k$
$[b_m ; b_{m+1}[$	c_m	n_m	f_m	p_m	$F_m = 1$
		$n = \sum_{i=1}^{i=m} n_i$	$\sum_{i=1}^{i=m} f_i = 1$	$\sum_{i=1}^{i=m} p_i = 100$	

QUELQUES CONVENTIONS

Chiffres significatifs

Les résultats statistiques provenant de calculs parfois réalisés à l'aide de micro-ordinateurs s'expriment sous formes de nombre d'une grande précision. Il n'est pas rare de trouver des résultats avec trois ou quatre décimales. Une telle précision dégage un caractère de scientificité qui éteint toute critique, alors qu'il ne s'agit que d'une précision illusoire qui n'apporte aucune information. La précision des observations est telle que généralement les résultats sont donnés avec une seule décimale.

On appelle chiffres significatifs d'un nombre les chiffres exacts constituant ce nombre : 5,32 a trois chiffres significatifs. La précision du résultat ne doit pas être supérieure à la précision des observations. Le résultat final d'un calcul ne peut avoir plus de chiffres significatifs que le nombre ayant le plus petit nombre de chiffres significatifs.

Exemple : $45{,}2 \cdot 65{,}324 = 2\,952{,}6$

Attention, ce n'est pas le cas pour les calculs intermédiaires où tous les chiffres doivent être impérativement conservés.

Les pourcentages sont beaucoup utilisés dans les calculs statistiques. En général, compte tenu de la précision des données, le résultat final sera fourni avec une seule décimale.

Les signes conventionnels

Dans un tableau statistique, certaines informations sont absentes, remplacées par des signes conventionnels qu'il est utile de connaître.

" ..... Le résultat n'existe pas faute d'enquête ou ne peut être obtenu

… ..... Résultat non disponible (pas encore publié, pas encore parvenu)

/// ..... Absence de résultat due à la nature des choses

– ..... Résultat rigoureusement nul

c ..... Résultat confidentiel par application des règles sur le secret statistique

ε ..... Résultat inférieur à la moitié de l'unité choisie

e ..... Estimation, évaluation

r ..... Nombre rectifié

p ..... Nombre provisoire

• ..... Rupture de série

Les notations indicées

À chaque modalité, il sera possible d'associer un certain nombre d'individus, ce nombre sera appelé l'effectif de la modalité. Celui de la modalité i sera noté n_i. Soit k le nombre de modalités prises par un caractère ; nous noterons I l'ensemble des valeurs $1,2,\ldots k$. L'ensemble constitué par les modalités et les effectifs associés à chacune des modalités est une série statistique ou, plus usuellement, une *distribution statistique*, du caractère pour la population considérée. Nous écrirons : $\{MO_i ; n_i\}$ où MO_i est la modalité i.

La notation somme (ou l'opérateur somme)

Soient les effectifs $n_1,n_2,\ldots,n_k$ de la distribution du caractère, nous noterons n la somme des effectifs.

$$n = n_1 + n_2 + \ldots + n_k$$

Cette écriture est peu maniable, nous remplacerons la somme précédente par la notation suivante :

$$\sum_{i=1}^{k} n_i = n \quad \text{avec} \quad i \in [1\,;k]$$

ou si la sommation est sans ambiguïté : $\sum n_i = n$

Le symbole $\sum$ se lit *sigma* et signifie somme des effectifs des k modalités de la distribution. C'est un opérateur linéaire.

Quelques propriétés de cet opérateur, (nous laissons au lecteur le soin de faire les démonstrations) :

$$\sum_{i=1}^{k}(x_i + y_i) = \sum_{i=1}^{k} x_i + \sum_{i=1}^{k} y_i$$

$$\sum_{i=1}^{k} a x_i = a \sum_{i=1}^{k} x_i$$

$$\text{si } a \text{ est une constante } \sum_{i=1}^{k} a = ka$$

$$\sum_{i=1}^{k}(x_i + b) = \sum_{i=1}^{k} x_i + kb$$

Quelques autres relations

$$\sum_{i=1}^{k} x_i^2 \neq \left(\sum_{i=1}^{k} x_i\right)^2$$

$$\sum_{i=1}^{k} \sqrt{x_i} \neq \sqrt{\sum_{i=1}^{k} x_i}$$

$$\sum_{i=1}^{k} \left(\frac{x_i}{y_i}\right) \neq \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} y_i}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{l} x_{ij} = \sum_{j=1}^{l}\sum_{i=1}^{k} x_{ij} = \sum_{i=1}^{k}\left[\sum_{j=1}^{l} x_{ij}\right] = \sum_{j=1}^{l}\left[\sum_{i=1}^{k} x_{ij}\right]$$

$$\frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} y_i} = \sum_{i=1}^{k} \frac{x_i}{\sum_{i=1}^{k} y_i}$$

La notation produit (opérateur produit)

De façon analogue à la notation somme, nous écrivons le produit de n nombres sous une forme abrégée.

$$n_1 \cdot n_2 \cdot \ldots \cdot n_p = \prod_{i=1}^{p} n_i$$

$$\prod_{i=1}^{p} a x_i = a^p \prod_{i=1}^{p} x_i$$

$$\prod_{i=1}^{p} a = a^p$$

$$\prod_{i=1}^{p} x_i y_i = \prod_{i=1}^{p} x_i \prod_{i=1}^{p} y_i$$

Les pourcentages et les fréquences

Le calcul d'une proportion ou d'une fréquence est l'acte statistique le plus élémentaire. Cette simple opération donne déjà une information plus accessible que la distribution statistique, elle permet de comparer des distributions dont les ordres de grandeur ne sont pas comparables. Les deux termes recouvrent des calculs formellement semblables.

Les proportions

Une répartition quantitative est le plus souvent exprimée sous forme de proportions. Une proportion indique l'importance relative d'une modalité dans l'ensemble des modalités. Une telle présentation permet de comparer des distributions statistiques dont les effectifs sont inégaux. Elle se calcule en faisant le rapport entre le nombre d'unités ayant le caractère et le nombre total d'unités.

Une forme très parlante de la présentation d'une proportion est de la donner comme une fraction du numérateur 1/2, 1/3, 1/10. L'inconvénient d'une telle présentation est qu'il est malaisé d'effectuer des additions, il faut réduire à un dénominateur commun.

Pour simplifier les opérations, mais aussi pour permettre des comparaisons plus immédiates on présente les proportions avec un dénominateur commun 10 ou plus généralement 100. Une proportion est généralement donnée en pourcentage. Une proportion sera comprise entre 0 et 1, un pourcentage sera compris entre 0 et 100 %. Par exemple, en 1981, 22,2 % de la population française avait

de 0 à 14 ans. Une remarque : en 1981, la population totale est évaluée en milliers de personnes à 53 838 et la population 0-14 ans dans la même unité à 11 932. Le rapport, la proportion de jeunes de 0 à 14 ans calculée « exactement » est de 22,1627 %. Un tel nombre n'a aucun sens, les pourcentages sont donnés avec un chiffre derrière la virgule, donc 22,2 %.

Le calcul d'un pourcentage consiste à appliquer le principe des proportions donc à poser l'équation suivante :

$$\frac{x}{100} = \frac{a}{b}, \; a \text{ et } b \text{ étant connus, il découle :}$$

$$x = \frac{100 \cdot a}{b}$$

La comparaison de deux nombres : les taux

Un taux mesure la modification relative d'une grandeur entre deux périodes. Il compare deux situations dans le temps. Soit Y une variable prenant les valeurs Y_0 et Y_1 aux temps t_0 et t_1. Le taux de croissance sera défini par : $r = \dfrac{Y_1 - Y_0}{Y_0}$ ou de façon plus générale $r = \dfrac{\Delta Y}{Y}$

L'application à un ensemble de grandeurs économiques, des salaires par exemple, d'un taux de croissance identique, conserve les proportions, mais accroît les écarts absolus. Une augmentation en valeur absolue conserve les écarts absolus mais réduit les écarts relatifs.

Pour exprimer la modification relative d'une grandeur, il est plus simple de l'exprimer à l'aide d'un multiplicateur ou d'un indice.

Nous pouvons écrire plus simplement :

$$r = \frac{X_1 - X_0}{X_0} = \frac{X_1}{X_0} - 1 \quad \text{ou} \quad \frac{X_1}{X_0} = 1 + r \quad \text{ou} \quad X_1 = X_0(1 + r)$$

avec $r =$ le taux de croissance et $1 + r$ le multiplicateur.

Dans le cas de taux de croissance successifs, le calcul en sera facilité. Soit une croissance de r_1 suivie d'une de r_2, le taux de croissance global n'est pas égal à $r_1 + r_2$. Le multiplicateur de croissance est :

$$(1 + r) = (1 + r_1)(1 + r_2)$$

donc le taux de croissance total r est égal à :

$$r = (1 + r_1)(1 + r_2) - 1$$

La comparaison de deux taux

Il est courant, en économie, de comparer l'évolution relative de deux taux : c'est le principe de l'élasticité. Par exemple, si nous voulons apprécier la variation relative de la demande d'un produit en réaction à une variation relative du prix de ce produit. Nous ferons le rapport de la variation relative de la quantité et de la variation relative des prix. Les deux mouvements sont, en général, de sens opposé ; l'élasticité est souvent négative.

$$e_p = \frac{\dfrac{\Delta q}{q}}{\dfrac{\Delta p}{p}} = \frac{\Delta q}{\Delta p}\frac{p}{q} = \frac{\Delta q}{q}\ \frac{p}{\Delta p}$$

$|e_p| < 1$ la demande est inélastique

$|e_p| > 1$ la demande est élastique

$|e_p| = 1$ aucune élasticité ou isoélasticité ou élasticité unitaire

Les fréquences relatives

En statistique, le terme de fréquence est utilisé plus souvent que celui de proportion. La fréquence d'une valeur dans une série statistique est son importance relative, elle est le plus souvent exprimée en pourcentage. Elle se calcule comme l'importance d'une modalité par rapport à l'ensemble des modalités. Pour un caractère K ayant M_i modalités $1 \leqslant i \leqslant k$, la fréquence de la modalité M_i sera notée f_i et se définit comme la proportion des individus de la population présentant la modalité M_i.

$$f_i = \frac{n_i}{n} = \frac{n_i}{\sum_{i=1}^{k} n_i}, \text{ avec } \sum_{i=1}^{k} f_i = 1$$

La fréquence est le plus souvent présentée en pourcentage. Les fréquences permettent de comparer les structures selon le caractère étudié de populations d'effectifs différents. Le calcul des fréquences permet d'éliminer l'effet de taille ; on énonce les jugements du type relativement plus ou relativement moins.

Les fréquences cumulées

Dans le cas des variables numériques, la présentation peut se faire par ordre croissant ou par ordre décroissant. On calcule les fréquences cumulées. Soit une

variable statistique prenant k modalités x_i, la fréquence cumulée F_i sera la somme des fréquences des valeurs inférieures à x_i.

$$F_1 = f_1,\ F_2 = f_1 + f_2\ \text{, plus généralement } F_j = \sum_{i=1}^{j} f_i$$

Les fréquences cumulées sont considérées comme les valeurs en des points connus d'une fonction de distribution $F(x)$.

LES REPRÉSENTATIONS GRAPHIQUES

Les graphiques permettent de donner une synthèse visuelle de la distribution d'une variable et de percevoir l'éventuelle relation entre les variables, cette section en présente quelques exemples. Les représentations peuvent être spécifiques à un type de variable ou de caractère. Sauf indication contraire tous les graphiques sont réalisables en effectifs ou en fréquences, ils sont superposables à l'échelle prés.

Ils constituent pour les pourcentages un moyen simple de comparer sur un base unique des données de valeurs différentes. Les taux permettent de suivre l'évolution de grandeurs.

Les graphiques permettent de mieux percevoir une relation entre des variables, ce chapitre présente quelques exemples.

Le cas d'une variable

Le choix des représentations graphiques dépend pour une large part du type du caractère statistique : caractère qualitatif, variable statistique discrète, variable statistique continue.

Les représentations des caractères qualitatifs

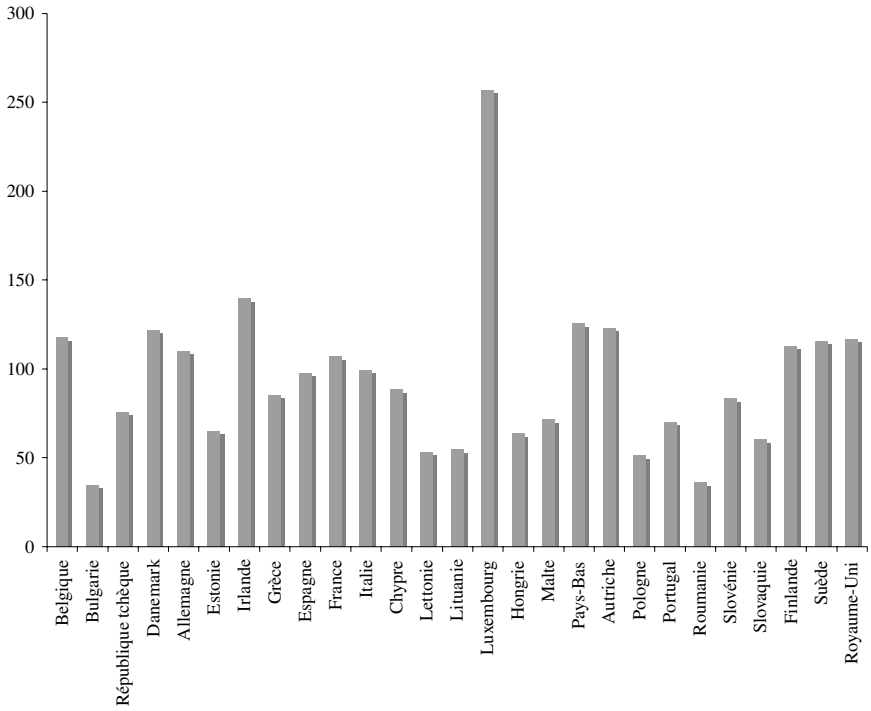
Les *diagrammes figuratifs*, les *pictogrammes* sont utilisés pour leur effet suggestif : des personnages pour une population humaine, des épis pour une production céréalière. La multiplication par deux des dimensions du diagramme indique une multiplication par quatre de la grandeur représentée. Les illustrations utilisées pour figurer la distribution de caractère qualificatif sont souvent imprécises. Le lecteur ne sait pas toujours s'il faut comparer les longueurs ou les surfaces. Pour qu'un diagramme figuratif soit significatif, il faut que les surfaces soient proportionnelles.

Les *cartogrammes* représentent les variations d'une grandeur sur un territoire géographique en assignant à chaque zone ses caractéristiques. Pour cela, on uti-

lise des fonds de cartes pour représenter les variables. Il existe deux grandes catégories de cartogrammes. Dans la première catégorie, les surfaces de chaque unité géographique sont hachurées ou coloriées ; dans la seconde catégorie, les phénomènes sont représentés par des surfaces centrées sur les unités géographiques et proportionnelles aux effectifs étudiés.

Le *diagramme en tuyaux d'orgue* ou en *barres* est constitué d'une suite de rectangles dont les hauteurs sont proportionnelles à l'effectif (ou à la fréquence) de la variable et dont les bases sont identiques. La représentation peut être horizontale ou verticale.

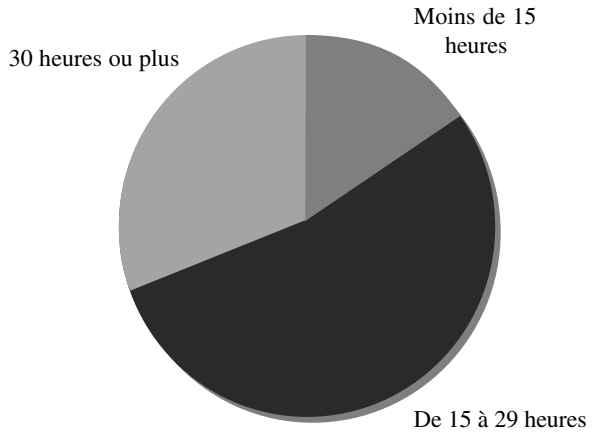
PIB par habitant en standards de pouvoir d'achat (SPA) (EU – 25 = 100)



Source : Eurostat

Le *diagramme en secteurs* ou en « camembert » visualise la part relative des catégories de la variable sur une population. Le cercle représente l'ensemble de la population, les différentes modalités seront représentées par des secteurs dont la surface est proportionnelle aux effectifs ou aux fréquences. Une telle représentation n'est significative que si le total des fréquences est de 100 %. Un demi-cercle peut jouer le même rôle.

Durée hebdomadaire moyenne de travail des femmes (2005)



Source : Insee

La représentation en secteurs visualise bien l'importance relative des différentes modalités. Cette représentation permet, pour des comparaisons dans le temps et dans l'espace, de rendre sensible les différences en valeur absolue. Les aires des cercles seront proportionnelles aux effectifs de chacune des populations. C'est-à-dire :

$$\frac{\pi r_1^2}{\pi r_2^2} = \frac{A_1}{A_2} \text{ autrement dit } \frac{r_1}{r_2} = \sqrt{\frac{A_1}{A_2}}$$

Les représentations des variables quantitatives

Dans certains cas, la représentation peut faire appel aux représentations décrites ci-dessus. Deux représentations graphiques retiendront plus particulièrement notre attention : la courbe cumulative des fréquences et l'histogramme.

Les nuages constituent une représentation où les modalités sont en abscisses et les effectifs ou les fréquences en ordonnées.

Variable quantitative discrète

Le *diagramme en bâtons* est la représentation graphique des effectifs ou des fréquences d'une variable discrète. À chaque valeur (x_i) en abscisse on fait correspondre un segment vertical de longueur proportionnelle soit à l'effectif (n_i), soit à la fréquence (f_i) de cette modalité. Ce graphique différentiel se distingue du graphique intégral ou cumulatif qui représente les fréquences cumulées. Le

graphique intégral représente la fonction cumulative ou fonction de répartition définie par $F(x_i) = F_i$, qui est une fonction étagée pour une variable discrète pour $x_i < x \leqslant x_{i+1}$.

L'exemple de la distribution du nombre d'enfants par famille nous permet d'illustrer ces définitions.

Répartition des familles selon le nombre d'enfants

Nombre d'enfants	0	1	2	3	4 et +	Total
Toutes structures familiales	7 492 332	3 615 859	3 255 259	1 267 979	465 353	16 096 782

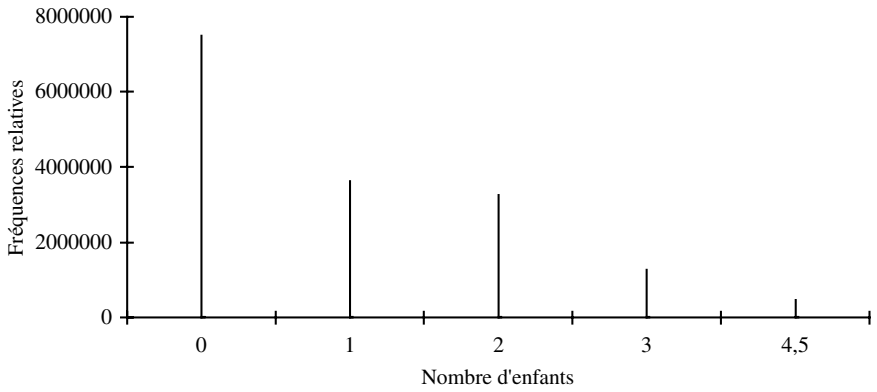
champ : France métropolitaine

Source : Insee, recensement 1999.

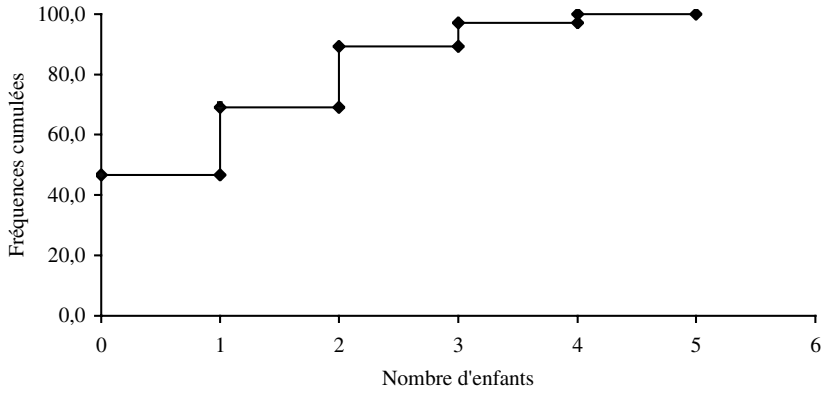
Tableau statistique

	Effectifs	Fréquences relatives	Fréquences cumulées
Nombre d'enfant	n_i	f_i	F_i
0	7 492 332	46,5	46,5
1	3 615 859	22,5	69,0
2	3 255 259	20,2	89,2
3	1 267 979	7,9	97,1
4,5	465 353	2,9	100,0
Ensemble	16 096 782	100,0	

Familles selon le nombre d'enfants (graphique différentiel)



Familles selon le nombre d'enfants (graphique intégral)



Les variables continues

Deux représentations graphiques retiendront plus particulièrement notre attention : l'histogramme et la courbe cumulative des fréquences.

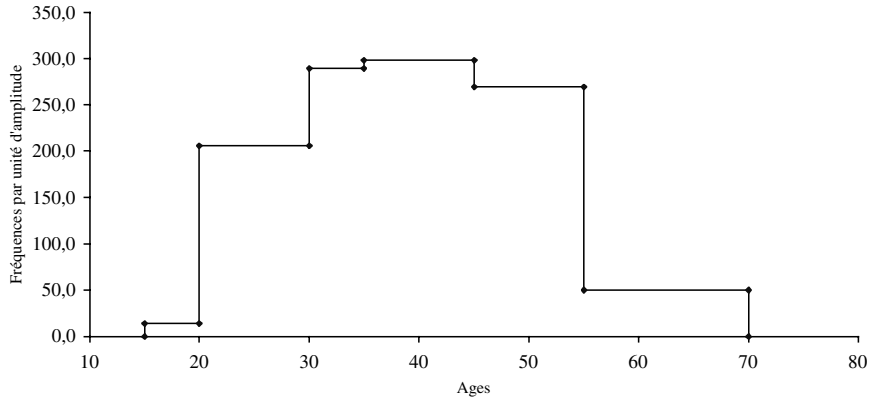
L'histogramme est réservé aux séries groupées en classes. Pour visualiser l'importance relative des classes, on préfère les représenter par des surfaces en construisant un histogramme. L'histogramme est une représentation graphique de la distribution des effectifs ou des fréquences d'une variable statistique continue ou considérée comme telle. À chaque classe de valeurs en abscisses, on fait correspondre un rectangle dont l'aire est proportionnelle à l'effectif de la classe (ou à la fréquence) : en abscisse l'amplitude de la classe, en ordonnée l'effectif (ou la fréquence) par unité d'amplitude. Soit une distribution $\{[b_i ; b_{i+1}[\, ; n_i\}$ d'une variable statistique continue, pour chaque classe, l'histogramme associe un rectangle de largeur $a_i = b_{i+1} - b_i$ et de hauteur $h_i = \dfrac{f_i}{a_i}$.

Emplois féminins par âge

$[b_i ; b_{i+1}[$	a_i	n_i (milliers)	f_i (en %)	$\dfrac{f_i}{a_i}$
[15 ; 20[	5	67	0,7	14,20
[20 ; 30[	10	1942	20,6	205,79
[30 ; 35[	5	1364	14,5	289,07
[35 ; 45[	10	2814	29,8	298,19
[45 ; 55[	10	2540	26,9	269,15
[55 ; 70]	15	710	7,5	50,16
Ensemble		9437	100,0	

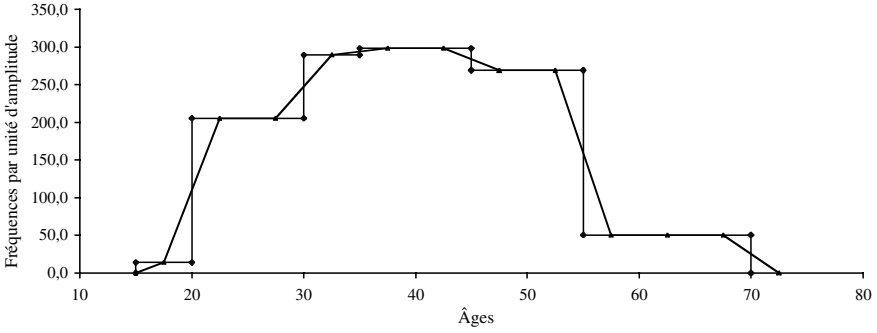
Source : Recensement de la population 1999 – INSEE

Histogramme de la distribution des femmes actives (graphique différentiel)



Le *polygone des fréquences* lisse l'histogramme de façon à éliminer les ruptures qui dépendent du choix du découpage en classe. L'histogramme est fidèle au tableau de départ, il donne l'impression, l'illusion, qu'au sein de chaque classe, les valeurs sont régulièrement distribuées et qu'apparaissent des modifications brusques. L'information paraît plus réaliste ; la courbe de fréquences respecte la compensation des aires, la surface incluse par la courbe est identique à celle de l'histogramme. Cette courbe des fréquences pourra être utilisée pour comparer la distribution réelle avec un modèle probabiliste connu.

Polygone des fréquences des activités féminines



Nous avons retenu comme limite inférieur de l'activité 15 qui correspond à l'âge légal, nous avons choisi 75 ans comme borne supérieure pour deux raisons tout d'abord pour la conservation des aires mais également par réalisme même si au-delà de 65 ans il s'agit en général d'activités à temps partiel.

La courbe cumulative des effectifs (ou des fréquences) représente graphiquement la fonction cumulative ou fonction de répartition définie par $F(x_i) = F_i$. La courbe cumulative des effectifs (ou des fréquences) s'obtient en joignant les points d'abscisse : la borne supérieure de la classe, et d'ordonnée : l'effectif cumulé croissant correspondant.

Courbe cumulative de la distribution des femmes actives (graphique intégral)

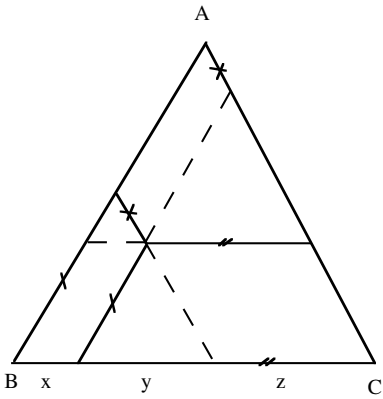
Il est possible de transformer une variable quantitative en variable qualitative, les valeurs de la variable ou les classes devenant alors les catégories de la variable qualitative. Les représentations graphiques préconisées pour les variables qualitatives sont alors applicables aux variables quantitatives transformées.

À ces représentations, nous pouvons ajouter les représentations triangulaires, les diagrammes polaires.

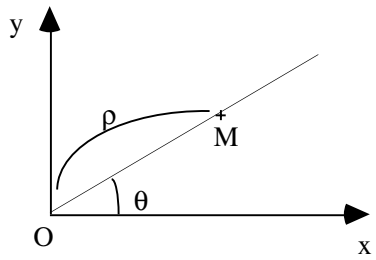
Le *graphique triangulaire* sert à représenter des phénomènes constitués de trois variables dont la somme est constante ; le plus souvent il s'agira de la décomposition en trois postes d'une grandeur variable. Le diagramme triangulaire compare trois données complémentaires pour visualiser leurs relations.

L'utilisation du diagramme triangulaire repose sur une propriété du triangle équilatéral. Si d'un point *M*, intérieur à un triangle équilatéral on trace les parallèles aux côtés, les longueurs des segments découpés sur ceux-ci ont une somme constante égale à la longueur du côté.

Le diagramme polaire permet de visualiser une phénomène sur plusieurs axes.



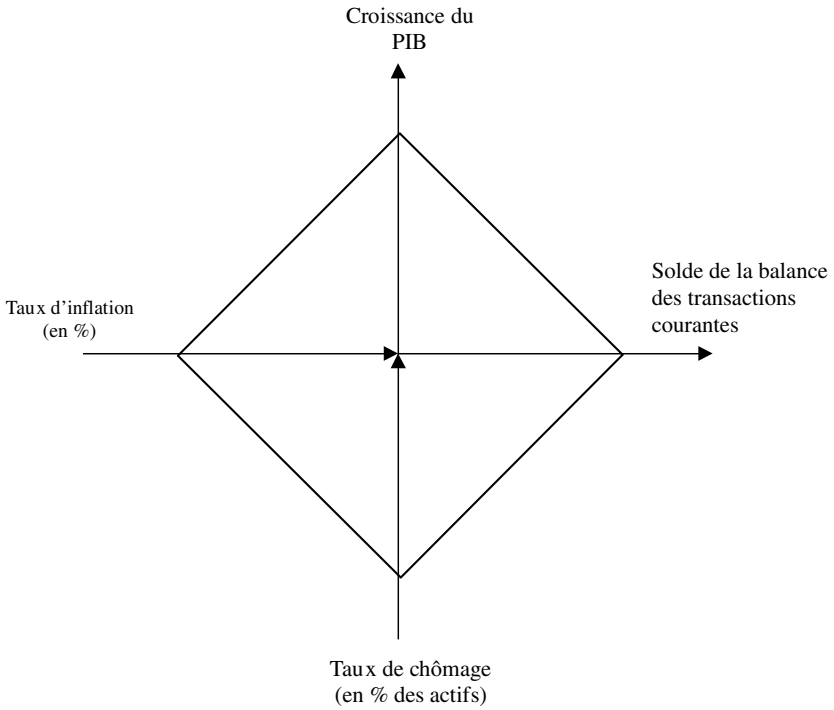
Dans un graphique à coordonnées cartésiennes un point *M* est repéré par ses coordonnées (*x* et *y*) ; dans un graphique polaire, il l'est par l'angle θ (angle polaire) et la mesure algébrique ρ du vecteur $\overrightarrow{OM}$.



Un exemple de ce type de graphique est connu sous le nom de Carré magique qui représente les quatre principaux objectifs de la politique économique qui sont :

– la croissance économique (évaluée par le taux de croissance du PIB) ;
– la situation de l'emploi (mesurée par le taux de chômage en % de la population active) ;
– la stabilité des prix (mesurée par le taux d'inflation en %) ;
– l'équilibre des comptes extérieurs (mesuré par le solde de la balance des paiements en % du PIB).

Le carré magique



Le cas de deux variables croisées

Croisement de deux variables qualitatives

Le tableau statistique des effectifs se présente sous la forme d'un tableau de contingence

Tableau des effectifs

Variable 2 / Variable 1	Modalité 1		Modalité j		Modalité p	Effectif marginal de la variable 1
Modalité 1	n_{11}		n_{1j}		n_{1p}	$n_{1.} = \sum_{k=1}^{k=p} n_{1k}$
Modalité i	n_{i1}		n_{ij}		n_{ip}	$n_{i.} = \sum_{k=1}^{k=p} n_{ik}$
Modalité m	n_{m1}		n_{mj}		n_{mp}	$n_{m.} = \sum_{k=1}^{k=p} n_{mk}$
Effectif marginal de la variable 2	$n_{.1} = \sum_{k=1}^{k=m} n_{k1}$		$n_{.j} = \sum_{k=1}^{k=m} n_{kj}$		$n_{.p} = \sum_{k=1}^{k=m} n_{kp}$	$n = \sum_{k=1}^{k=p} n_{.k} = \sum_{k=1}^{k=m} n_{k.}$

Comme représentation graphique, on utilise un diagramme en barre où les barres des modalités de la variables 1 sont partagées suivant les modalités de la variable 2.

Ce tableau est souvent assorti d'un tableau des fréquences conditionnelles :

Tableau des fréquences conditionnelles pour la variable 1

Variable 2 / Variable 1	Modalité 1		Modalité j		Modalité p	
Modalité 1	$f_{1/1} = \frac{n_{11}}{n_{1.}}$		$f_{j/1} = \frac{n_{1j}}{n_{1.}}$		$f_{p/1} = \frac{n_{1p}}{n_{1.}}$	1
Modalité i	$f_{1/i} = \frac{n_{i1}}{n_{i.}}$		$f_{j/i} = \frac{n_{ij}}{n_{i.}}$		$f_{p/i} = \frac{n_{ip}}{n_{i.}}$	1
Modalité m	$f_{1/m} = \frac{n_{m1}}{n_{m.}}$		$f_{j/m} = \frac{n_{mj}}{n_{m.}}$		$f_{p/m} = \frac{n_{mp}}{n_{m.}}$	1

Le tableau des les fréquences conditionnelles pour la variable 2 est analogue à celui ci-dessus.

La représentation utilisée est alors un diagramme à cumul interne où toutes les barres ont la même hauteur. Il permet de comparer la part relative des catégories de la variable 2 dans chacune des catégories de la variable 1.

Croisement d'une variable qualitative et d'une variable quantitative

Les diagrammes ne sont pas différents de ceux qui sont utilisés pour le croisement de deux variables qualitatives.

Croisement de deux variables quantitatives

Les nuages de points, où les valeurs de la variable 1 sont en abscisses et les valeurs de la variable 2 en ordonnées, constituent la représentation la plus utilisée pour des correspondances simples.