CTU
Master AGPS

De la donnée à la connaissance : traitement, analyse et transmission

Introduction à la statistique descriptive

Prof. Marie-Hélène de Sède-Marceau

Année 2010/2011

Statistique 1

Introduction à la statistique descriptive

Plan

Introduction : Objectifs du cours

Court descriptif des objectifs et prétentions du cours et de son organisation

Chapitre 1 : Définitions, terminologie et notation

Qu'est-ce que la statistique ? Quelques définitions
Terminologie et notation standard. Exercices.

Chapitre 2 : Données et organisation des données : types de données et tableaux statistiques

Données qualitatives et données quantitatives, notion de classes, tableaux unidimensionnel, tableaux croisés à 2 dimensions, tableaux à n dimensions. Exercices.

Chapitre 3 : Modes de représentation des données : diagrammes et graphiques

Types de données et de tableaux et modes de représentation possibles, échelles graphiques, diagrammes, graphiques. Exercices.

Chapitre 4 : Caractériser une distribution et résumer des tableaux statistiques à l'aide de paramètres appropriés : tendance centrale et dispersion

Paramètres de tendance centrale (mode, moyenne, médiane, quantiles, etc.), paramètres de dispersion (variance, écart-type, coefficient de variation, standardisation, etc.). Exercices.

Chapitre 5 : Série Chronologique : progression et indices

Indices temporels et synthétiques, indice de Laspeyres, taux de croissance simple et successifs, etc. Exercices.

Chapitre 6 : Tendances et corrélations : relations entre deux variables, interpolation et extrapolation

Identifier et matérialiser une tendance par la méthode des moindres carrées, caractériser une relation entre deux variables (coefficient de corrélation), formuler une relation statistique entre deux variables (régression) en vue de l'interpolation ou de l'extrapolation. Exercices.

Annexes

Annexe 1 : Précision et explication sur une notation spécifique en statistique : somme et produit

Annexe 2 : Liste (non exhaustives) des fonctions Excel utiles en statistiques descriptive

Annexe 3 : Activer la macro « histogramme » dans Excel

Annexe 4 : Tableau croisé dynamique dans Excel : utilisation et compléments

Statistique 1

Introduction à la statistique descriptive

Introduction - Objectifs du cours

Ce cours est destiné en priorité à un public n'ayant aucune formation en statistique et cependant confronté de façon récurrente à la manipulation et à l'analyse de séries de données.

Aucun pré-requis en mathématique n'est exigé si ce n'est la connaissance des opérations mathématiques de base. Volonté, curiosité et ténacité permettront de maîtriser sans encombre les notions abordées qui, malgré leur complexité apparente, demeurent relativement simples.

Cette formation se présente davantage comme une initiation à la rigueur que nécessite la manipulation d'ensembles de données afin d'utiliser à bon escient les méthodes appropriées pour éviter de faire parler faussement les chiffres.

Les concepts et méthodes statistiques seront abordées au travers de nombreux exemples que viendront ponctués des exercices à réaliser dans le logiciel Excel dont la maîtrise de base est supposée acquise. Lorsque nécessaire, un point rouge | signalera la référence d'un exercice à réaliser.

Au final, il s'agira de se familiariser avec et de maîtriser la méthode statistique en général en vue de décrire, de résumer et d'analyser une population ou un ensemble de données.

Chapitre 1

1. Définitions, terminologie et notation

1.1 Qu'est-ce-que la statistique ?

Il n'existe pas de définition universelle et totalement aboutie de la statistique. Celles présentées ci-après donnent un aperçu des différentes facettes que peut revêtir le terme « statistique » en tant que science.

La statistique c'est la science des grands nombres regroupant l'ensemble de méthodes mathématiques qui, à partir du recueil et de l'analyse de données réelles, permettent l'élaboration de modèles probabilistes autorisant les prévisions. (Larousse).

On perçoit dans cette première définitions plusieurs termes et notions fondamentales propres à la statistique : le recueil sous-entend la collecte qui elle-même suppose dans bien des cas la réalisation d'une enquête ou d'un sondage. Enquête et sondage impose l'échantillonnage en vue de l'inférence[1] .
L'analyse des données suppose la manipulation de tableaux ou grands ensembles de données qu'il s'agira de décrire et de résumer tout en accompagnant cette opération de représentations graphiques et cartographiques.
La notion de modèles probabilistes sous-entend une certaine maîtrise de l'incertitude dans le but de réaliser des prévisions ou de pratiquer l'inférence.

Autre définition, moins académique celle-ci :
la statistique est un ensemble de méthodes permettant de prendre une bonne décision face à l'incertitude (Wallis & Roberts, The Nature of Statistics)

C'est aussi un ensemble d'outils et de méthodes qui permettent de synthétiser et de résumer des grands volumes de données, des grandes matrices d'informations.

On voit se dessiner ici les deux principales branches de la statistique :

- La statistique descriptive
- La statistique mathématique ou inférentielle

1 Inférence: Opération intellectuelle par laquelle on passe d'une vérité à une autre vérité, jugée telle en raison de son lien avec la première. *La déduction est une inférence.*
Règles d'inférence, celles qui permettent, dans une théorie déductive, de conclure à la vérité d'une proposition à partir d'une ou de plusieurs propositions, prises comme hypothèses. En statistique, l'inférence est une opération qui permet de généraliser à une population mère les propriétés et conclusions observées à partie d'un échantillon représentatif de cette population mère

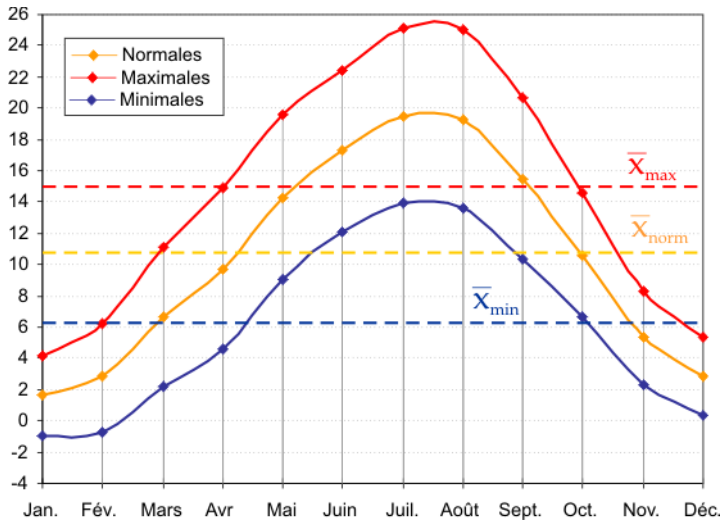
1.2 La statistique descriptive :

Ensemble des méthodes permettant de décrire une population par le biais des individus qui la composent. La statistique descriptive s'intéresse donc à décrire et caractériser un ensemble d'individus représenté la plupart du temps sous la forme de tableaux (tableaux de données), à résumer et synthétiser ces tableaux par l'intermédiaire de graphiques et de paramètres appropriés (fréquences, distribution, moyenne, dispersion, etc.). Elle s'attachera à éventuellement rechercher des corrélations (liaisons statistiques) entre les éléments de ces tableaux (variables et individus).

Exemple :

Les températures moyennes mensuelles à Strasbourg sur la période 1971-2000

Températures moyennes mensuelles (°C)	Jan.	Fév.	Mars	Avr.	Mai	Juin	Juil.	Août	Sept.	Oct.	Nov.	Déc.
Normales	1,6	2,8	6,7	9,7	14,3	17,3	19,5	19,3	15,5	10,6	5,3	2,8
Maximales	4,2	6,2	11,1	14,9	19,6	22,4	25,1	25,0	20,7	14,6	8,3	5,3
Minimales	-1,0	-0,7	2,2	4,6	9,0	12,1	13,9	13,6	10,3	6,6	2,3	0,3



$\overline{X}_{max}$ Moyenne annuelle maximale = 14,8 °C

$\overline{X}_{norm}$ Moyenne annuelle normales = 10,5 °C

$\overline{X}_{min}$ Moyenne annuelle minimales = 6,1 °C

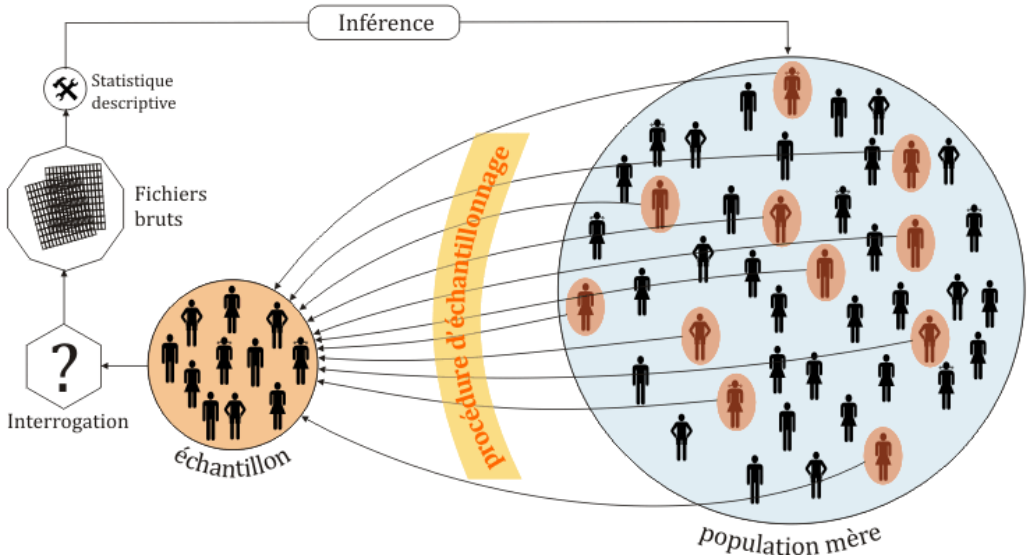
Le « simple » passage d'un tableau de données plus ou moins important à un graphique et/ou à quelques indicateurs pertinents telle que la moyenne constituent une opération relevant de la statistique descriptive.

1.3 La statistique mathématique ou inférentielle :

Cette branche des statistiques s'intéresse davantage à extrapoler des résultats issus d'échantillons en vue de caractériser une population mère inconnue, de faire des prévisions de comportements basées sur le calcul de probabilités.

Exemple :

En période électorale, on interroge 1 000 personnes sur leur intention de vote. A partir des résultats obtenus sur cet échantillon, on prévoit, avec une certaine précision, le comportement de l'ensemble des électeurs (population mère) et par là même, le résultat des élections. C'est ce qu'on appel l'inférence statistique et c'est le principe même du sondage d'opinion par exemple.



Le lien de complémentarité entre statistique inférentielle et statistique descriptive est évident : la première collecte et fournit à la seconde la « matière première » à décrire et à et analyser qui, retournée à la première est extrapolée.

Le présent cours sera consacré à la statistique descriptive. Mais avant de commencer, il convient de se familiariser avec le vocabulaire et la notation universelle de la statistique.

1.4 Terminologie et notation standard de la statistique

Terminologie et concepts fondamentaux

Population : ensemble des individus (ou unités statistiques) présentant un caractère commun. Pour une thématique donnée, la population regroupe toujours la totalité des individus relatif à cette thématique (notion d'exhaustivité).
Exemples :

- la population européenne : ensemble des individus résidant sur le territoire européen à un moment donné.
- Le parc automobile français: ensemble des automobiles immatriculées sur le territoire français.
- Le parc de logements de Toulouse : ensemble des logements de la ville de Toulouse.
- Le lot 9 718 du médicament « alpha » : ensemble boîtes de « alpha » produit sous le n°. de lot 9 718.
- Le cheptel bovin de l'exploitation Martin : ensemble des bovins femelles et mâles rattachés à l'exploitation agricole Martin.

La population est en général notée P
L'effectif total d'une population est noté N

Unité statistique (ou individu) : élément de base constitutif de la population à laquelle il appartient. Il est indivisible et peut être un animal, un végétal, un humain ou un objet. Exemples : une automobile, un logement, une vache, une ampoule, une ville, etc. noté i

Échantillon : sous-ensemble construit et représentatif d'une population donnée. Lorsque l'on parle d'échantillon on parle en général de population mère, c'est-à-dire de la population dont est issu l'échantillon. L'échantillon est fréquemment noté s

Dénombrement : comptage exhaustif des individus composant une population donnée. Le recensement de la population est un dénombrement.

Caractère(s) : caractéristique(s) de l'individu intégrant la population étudiée. Exemple : la couleur, le sexe, le poids, la taille, la marque, le modèle, l'espèce, le prix, la surface, etc.

Variable : une variable est une caractéristique pouvant prendre plusieurs des valeurs d'un ensemble d'observations possibles auquel une mesure ou une qualité peut être appliquée.

Modalité : valeur qualitative ou quantitative que peut prendre le caractère précédemment défini. Exemple : sexe féminin ou masculin, poids 45 kg, couleur verte, etc. Attention, les modalités sont exhaustives et mutuellement exclusives. Chaque individu doit pouvoir être classé dans une et une seule modalité.
Récapitulatif intermédiaire par l'exemple:

Population :	Le parc locatif privé loué vide de Cahors
Individu :	Un logement appartenant à ce parc
Caractère :	Taille du logement
Modalité :	Nombre de pièces de ce logement

Classe : il est fréquent qu'une population soit divisée en sous-ensembles cohérents construits à partir de critères déterminés de façon à réduire la taille des tableaux de données et à en faciliter la lecture, l'analyse et l'interprétation. Cette division induit une regroupement des individus et la formation de classes rassemblant chacune des individus présentant des caractères similaires.

Exemple : les classes d'âge d'une population, deux possibilités (suggestion)

Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6	Classe 7
0-19 ans	20-29 ans	30-39 ans	40-49 ans	50-59 ans	60-69 ans	70 ans et plus
13	21	32	31	26	19	14

Ou

Classe 1	Classe 2	Classe 3	Classe 4
0-19 ans	20-39 ans	40-59 ans	60 ans et plus
13	53	57	33

Plusieurs critères peuvent être utilisés simultanément pour former des classes ou sous-populations comme par exemple l'âge et le sexe :

Sexe \ Age	0-19 ans	20-39 ans	40-59 ans	60 ans et plus	/
	7	26	29	19	81
	6	27	28	14	75
&	13	53	57	33	N = 156

La seule contrainte réside dans le fait que la somme des effectifs par classe donne toujours l'effectif total N

Le découpage en classes d'une population selon un ou plusieurs critères est une opération appelée « discrétisation ». Celle-ci nécessite la plupart du temps une connaissance fine du phénomène étudié car sa réalisation, très sensible aux effets de seuils et de limites de classes, peut aboutir à des résultats dont l'interprétation peut être différente à totalement opposée notamment sur le plan cartographique. La discrétisation fera l'objet d'un paragraphe particulier dans ce cours.

Fréquence : Rapport du nombre d'individus d'une population ou d'un échantillon ayant un caractère commun (= modalité) au nombre total des individus de cette même population ou de ce même échantillon.

> Note : pour davantage de précisions et d'explication concernant la notation ci-après utilisée, on se reportera à l'**annexe 1** en fin du présent document

Exemple:

En 1999, une commune quelconque comptait 393 ménages. 108 d'entre eux étaient composés d'une seule personne soit une fréquence de : 108 / 393 = 0,275. Cette fréquence, également appelée fréquence relative, peut être exprimée en pourcentage soit 0,275 x 100 = 27,5 %. On la note F quand elle brute et $F\%$ quand elle est exprimée en pourcentage. L'effectif d'une modalité, ou nombre de fois qu'apparaît une modalité dans une population, est appelé fréquence absolue notée f. Dans notre cas, le nombre de fois où apparaît la modalité « ménage composé d'une seule personne » est 108. A noter que la somme des fréquences absolues des modalités donne le nombre total N d'individus d'une population (le symbole Σ signifiant somme (pour davantage de précision, se reporter à l'annexe 1)) :

$$\sum_{i=1}^{n} f_i = f_1 + f_2 + f_3 + \ldots + f_i + \ldots + f_n = N$$

La fréquence relative est donc le rapport de la fréquence absolue d'une modalité à la population totale (N) soit :

Fréquence relative : $F = \frac{f}{N}$

l'ensemble des fréquences pour toutes les modalités des individus d'une population ou d'un échantillon forme l'histogramme des fréquences. L'histogramme n'est autre chose que le graphique figurant la distribution des fréquences pour un phénomène donné. La somme des fréquences, pour une population ou un échantillon donné, est toujours égale à 1 :

$$\sum_{i=1}^{n} F_i = F_1 + F_2 + \ldots + F_i + \ldots + F_n = 1$$

F_1 représente la fréquence relative observée pour la modalité 1
F_2 représente la fréquence relative observée pour la modalité 2
F_i représente la fréquence relative observée pour la modalité i
F_n représente la fréquence relative observée pour la modalité n

Exemple:
Reprenons notre commune. Relativement à la variable « ménages » , 5 modalités ont été retenues:

Modalité 1 : ménages composés d' 1 personne
Modalité 2 : ménages composés de 2 personnes
Modalité 3 : ménages composés de 3 personnes
Modalité 4 : ménages composés de 4 personnes
Modalité 5 : ménages composés de 5 personnes et plus

Pour chacune de ces modalités nous avons une fréquence absolue et une fréquence relative

	Modalité 1 Ménages 1 pers.	Modalité 2 ménages 2 pers.	Modalité 3 ménages 3 pers.	Modalité 4 ménages 4 pers.	Modalité 5 ménages 5 pers. et plus	$\sum_{i=1}^{5} F_i$
Fréquence absolue f_i	f_1 = 108	f_2 =130	f_3 =72	f_4 =48	f_5 =35	$\sum_{i=1}^{5} f_i = 393$
Fréquence relative F_i	F_1 = 0,275 (108/393)	F_2 =0,331 (130/393)	F_3 =0,183 (72/393)	F_4 =0,122 (48/393)	F_5 =0,089 (35/393)	$\sum_{i=1}^{5} F_i = 1$
Fréquence relative en pourcentage $F_i\%$	$F_1\%$ = 27,5 % (F_1 x 100)	$F_2\%$ = 33,1 % (F_2 x 100)	$F_3\%$ = 18,3 % (F_3 x 100)	$F_4\%$ = 12,2 % (F_4 x 100)	$F_5\%$ = 8,9 % (F_5 x 100)	$\sum_{i=1}^{5} F_i\% = 100$

Pour ce qui est des fréquences absolues, on a :

$$\sum_{i=1}^{5} f_i = f_1 + f_2 + f_3 + f_4 + f_5 = 108 + 130 + 72 + 48 + 35 = 393$$

Avec la même formulation on peut écrire pour les fréquences relatives :

$$\sum_{i=1}^{5} F_i = F_1 + F_2 + F_3 + F_4 + F_5 = \frac{f_1}{N} + \frac{f_2}{N} + \frac{f_3}{N} + \frac{f_4}{N} + \frac{f_5}{N} = \frac{108}{393} + \frac{130}{393} + \frac{72}{393} + \frac{48}{393} + \frac{35}{393} = 0{,}275 + 0{,}331 + 0{,}183 + 0{,}122 + 0{,}089 = 1$$

Distribution : Selon le Petit Larousse, ensemble des données d'une série statistique associées à un ou à plusieurs caractères. Façon dont les individus d'une population se répartissent en fonction d'une ou plusieurs modalités.

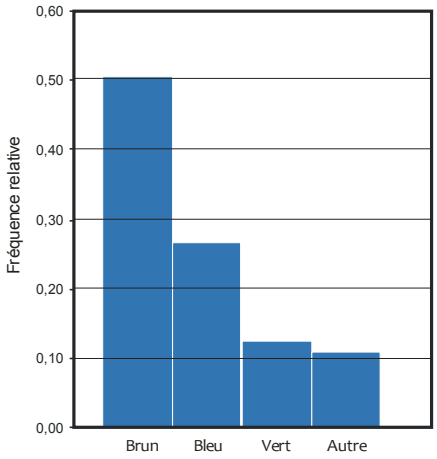
Exemple :

Distribution d'une population de 120 individus (Id) selon la couleur des yeux (Cy) :

Id	Cy	Id	Cy	Id	Cy	Id	Cy	Id	Cy	Id	Cy
1	Br	21	Bl	41	Vr	61	Vr	81	Bl	101	Br
2	Br	22	Au	42	Bl	62	Br	82	Vr	102	Br
3	Bl	23	Br	43	Au	63	Br	83	Br	103	Vr
4	Br	24	Br	44	Br	64	Vr	84	Br	104	Bl
5	Bl	25	Bl	45	Br	65	Au	85	Bl	105	Br
6	Br	26	Au	46	Br	66	Br	86	Au	106	Br
7	Br	27	Br	47	Bl	67	Bl	87	Br	107	Au
8	Vr	28	Bl	48	Br	68	Br	88	Br	108	Bl
9	Br	29	Br	49	Br	69	Au	89	Vr	109	Br
10	Bl	30	Br	50	Bl	70	Bl	90	Bl	110	Bl
11	Br	31	Br	51	Vr	71	Vr	91	Vr	111	Br
12	Bl	32	Br	52	Br	72	Br	92	Au	112	Vr
13	Au	33	Bl	53	Br	73	Br	93	Br	113	Bl
14	Br	34	Vr	54	Au	74	Au	94	Br	114	Br
15	Br	35	Bl	55	Bl	75	Br	95	Bl	115	Br
16	Bl	36	Br	56	Vr	76	Br	96	Br	116	Au
17	Au	37	Br	57	Br	77	Bl	97	Bl	117	Vr
18	Br	38	Bl	58	Bl	78	Br	98	Br	118	Bl
19	Vr	39	Br	59	Br	79	Bl	99	Br	119	Bl
20	Br	40	Bl	60	Br	80	Br	100	Br	120	Br

Couleur yeux	f_i	F_i%
Bruns (Br.)	61	50,8
Bleus (Bl.)	32	26,6
Verts (Vr.)	14	11,6
Autre (Au.)	13	10,8
$\sum$	120	100

Distribution de la population pour la variable « couleur des yeux » et son histogramme



Une distribution se représente la plupart du temps sous forme graphique soit à partir des données brutes, c'est-à-dire non regroupées en classes, soit à partir des données classifiées, discrétisées. Dans les deux cas, le graphique construit porte le même nom: l'Histogramme. Un histogramme figure toujours des fréquences, qu'elles soient absolues ou relatives.

Moyenne : pour une variable donnée, la moyenne correspond à la somme des valeurs d'une population $\sum x_i$ (ou d'une modalité) divisée par le nombre de valeurs N de ladite population (ou de ladite modalité).

Exemple: prix au m² du foncier à bâtir observé sur la commune de Besançon.

Parcelle	Prix de vente p_i TTC (€/m²)
p1	78,24
p2	81,15
p3	69,65
p4	101,54
p5	97,89
p6	77,23
p7	54,56
p8	98,21
p9	65,32
p10	113,33
p11	108,79
p12	93,66
p13	99,45
$N = 13$	$\sum_{i=1}^{13} p_i = 1139{,}02$

$$\text{Prix moyen} = \bar{P} = \frac{\text{somme des valeurs}}{\text{Nombre de valeurs}} = \frac{\sum_{i=1}^{13} p_i}{N} = \frac{1}{N}\sum_{i=1}^{13} p_i = 1139{,}02 \text{ over } 13 = 87{,}62 \text{ €/m²}$$

1.5 Notation standard

Concept / notion	Formulation / notation	lecture
Effectif total d'une population	N	Grand N
Effectif total d'un échantillon	n	Petit n
Moyenne de la variable x	$\bar{x}$	X barre
Somme des x	$\sum_{i=1}^{n} x_i$	Somme des x_i pour i = 1 jusqu'à n
Ecart-type de la variable x	σ_x	Ecart-type de x ou sigma x
Variance de la variable x	σ_x^2	Variance de x ou sigma carré x
Produite des x	$\prod_{i=1}^{n} x_i$	Produit des x_i pour i = 1 jusqu'à n
Coefficient de détermination	r^2	R carré
Coefficient de corrélation	r	r
Fréquence absolue	f	Petit f
Fréquence relative	F	Grand F ou F majuscule

Exercice 1 : fichier Excel associé « Exercice 1 - Somme et fréquences.xls »

Chapitre 2

2. Types de données et tableaux statistiques

2.1 Types et propriétés de la donnée

Les données manipulées en statistique (lors de la collecte et/ou lors du l'analyse) peuvent se présenter sous différentes formes. Ces formes, reflets des propriétés intrinsèques de la donnée, influent de façon décisive sur la manière de représenter celle-ci et sur les types de traitements qui pourront lui être appliqués en vue de son analyse. On distingue trois propriétés fondamentales qui permettent de caractériser précisément la donnée. Ce sont:

- Le type : qualitatif ou quantitatif
- L'échelle de mesure : nominale, ordinale, intervalle ou proportionnelle
- La nature : continue ou discrète

A chaque donnée, à chaque variable sont nécessairement rattachées ces trois propriétés.

2.1.1 Types, échelles de mesure et natures des données et variables

Les trois propriétés seront traitées simultanément tant elle sont indissociables. A toute variable ou toute donnée sont nécessairement rattachés un type, une échelle de mesure et une nature. Il existe cependant une hiérarchie naturelle entre les propriétés des variables et données et l'ordre dans lequel elles ont été précédemment évoquées en est le reflet et c'est celui que nous respecterons pour les décrire (Cf. figure 2).

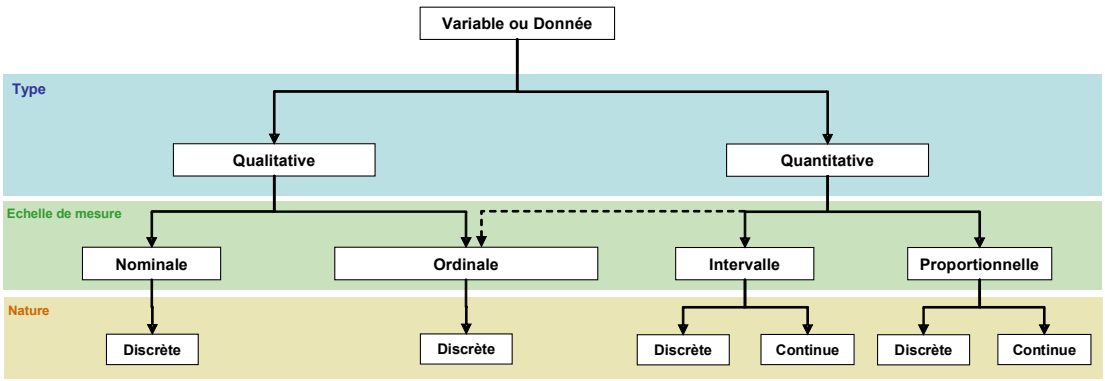


Figure 2 : propriétés des données et variables: inventaire et hiérarchie

Une donnée ou une variable est obligatoirement de type qualitatif ou de type quantitatif. Le type qualitatif est également appelé type « non-métrique » par opposition au type quantitatif dit type « métrique ».

2.1.1.1 Les données et variables qualitatives (ou variables non-métriques)

> **Définition :** les données ou variables qualitatives contiennent des valeurs qui expriment une qualité, un état, c'est-à-dire une condition, un statut unique et exclusif comme le *sexe*, la *couleur* ou bien encore la *catégorie socioprofessionnelle.*
>
> Les opérations arithmétiques que l'on peut réaliser sur ce type de variable sont relativement réduites et se limitent au comptage des effectifs par modalité (fréquences absolues) et au calcul de pourcentage (fréquences relatives) et le mode.

Exemple: le sexe

Une personne ne peut pas appartenir à des sexes différents en même temps et ne peut, en théorie, pas en changer (unicité) . Le fait d'être, par exemple, du sexe féminin l'exclue automatiquement des autres modalités que peut prendre la variable « sexe » (exclusivité).

Les données et variables qualitatives peuvent se présenter sous deux formes deux différentes: la forme nominale ou la forme ordinale.

La forme qualitative nominale : une variable est dite *qualitative nominale* quand ses valeurs sont des éléments d'une catégorie type nom non hiérarchique. En d'autres termes, ses éléments ne peuvent pas se ranger dans une gradation logique, selon une hiérarchie naturelle. La donnée qualitative nominale ne peut donc être appréhendée qu'à travers des modalités entre lesquelles il n'existe aucune relation d'ordre

Exemple : la variable « couleur » est de type qualitative nominale, les valeurs pouvant être prises par celle-ci étant bien de type nom (vert, jaune, noir, rouge, …) sans qu'aucune hiérarchie ne soit applicable entre les modalités recensées (on peut en aucun cas écrire jaune > rouge ou vert = noir).

La forme qualitative ordinale : une variable *qualitative ordinale* possède toutes les propriétés de la variable *qualitative nominale* avec en plus la possibilité de positionner et de hiérarchiser les individus entre eux selon la valeur attachée à leur caractère. En d'autres termes, il sera possible de ranger dans une gradation logique, selon une hiérarchie naturelle, les individus de la population étudiée pour le caractère retenu. D'une façon générale, la forme qualitative ordinale fait référence à des caractères non mesurables mais dont on sait que les modalités renferment une notion d'ordre, ou bien à des variables quantitatives ayant fait l'objet d'une classification. Les opérations autorisées pour l'échelle qualitative ordinale sont, en plus du comptage par modalité (fréquences absolues et fréquences relatives et mode), la médiane.

Exemple : la variable « niveau de confort d'un logement » est de type qualitative ordinale, les valeurs pouvant être prises par celle-ci étant bien de type nom (médiocre, moyen, bon, très bon) et une hiérarchie existe entre les modalités définies sans pour autant que l'on puisse mesurer de façon infaillible le niveau de confort : il n'existe pas de « conforomètre » ni d'unité de mesure du paramètre « confort » au demeurant très subjectif. Le caractère ordinal de la variable permet cependant d'écrire *bon* > *médiocre* ou *moyen* < *très bon*. L'époque de construction des logements est également une variable qualitative ordinale.

Une variable qualitative, qu'elle soit nominale ou ordinale, est toujours de nature discrète, contrairement à une variable quantitative qui peut être soit de nature *discrète*, soit de nature *continue*.

Définition : variable discrète

Une variable est dite *discrète* quand elle prendre un nombre fini ou dénombrable de valeurs. En d'autres termes, le passage d'une modalité à une autre est « brutal », sans continuité, sans glissement progressif. C'est typiquement le cas des variables qualitatives nominales et ordinales pour lesquelles la transitions entre modalités se réalise sans nuance, abruptement.

Exemple: la variable « catégorie socioprofessionnelle » est une variable *qualitative nominative discrète.* En effet, le nombre de valeurs qu'elle peut prendre est fini (ou dénombrable) et la transition entre modalité, par

exemple de la modalité « employé » à la modalité « agriculteur », se fait sans nuance, sans continuité, mais nettement.

Dans le même ordre d'idée, la variable « niveau d'éducation » avec les modalités « Analphabète, Primaire, Secondaire, Universitaire » est de type *qualitative ordinale discrète* pour les mêmes raisons qu'évoquées dans le cas précédent.

On verra le moment venu ce que recouvre la notion de continuité pour une variable, sachant que celle-ci ne s'applique qu'à la famille des données et variables quantitatives.

2.1.1.2 Les données et variables quantitatives (ou variables métriques)

Définition : les données ou variables quantitatives contiennent des valeurs numériques faisant référence à une unité de mesure reconnue. Pour cette raison, elles sont quelques fois qualifiées de variables métriques. La taille, le poids, la surface, la distance, le revenu, l'âge, le chiffre d'affaire ou bien encore la population (dans le sens du nombre d'habitants) sont des variables quantitatives.

Variables	Unité de mesure
Taille	Mètre
Poids	Kilogramme
Surface	Mètre carré
Distance	Mètre
Revenu	Euros
Age	Année
Chiffre d'affaire	Euros
Loyer	Euros/mois
Population	Nombre d'habitants

Toutes les opérations arithmétiques simples et complexes sont applicables aux variables quantitatives, du dénombrement (fréquences absolues) et autre calcul de pourcentage (fréquences relatives) en passant par la moyenne, la médiane et l'écart-type jusqu'à la modélisation numérique.

Exemple: le loyer d'un logement

Au-delà de la qualification d'un loyer (bon marché, correct, cher ou très cher) qui en fait alors une variable qualitative ordinale, le loyer demeure une variable mesurable objectivement selon une unité de mesure reconnue : le prix exprimé en euros par mois ou en euros par mois et par m². On peut l'additionner, en calculer la moyenne et l'écart-type, en regrouper les valeurs pour former des classes et même le modéliser.

Tout comme la donnée qualitative, la donnée quantitative peut se présenter sous différentes formes. On en dénombre trois, de la plus simple à la plus complexe : la forme (ou l'échelle) ordinale, l'échelle d'intervalles et l'échelle proportionnelle ou échelle de rapport.

La forme « quantitative » ordinale : Nous aborderons que succinctement l'échelle quantitative ordinale déjà évoquée dans le cas des variables qualitatives. Appliquée aux variables quantitatives, la forme ordinale revêt les mêmes caractéristiques. Elle s'applique en fait aux variables quantitatives pour lesquelles un regroupement par classes a été opéré (par ex. le regroupement d'individus par classes

d'âge ou classes de taille, le regroupement de villes selon leur taille ou bien encore le regroupement de parcelles foncières selon leur prix au m²). Même si l'échelle ordinale est abordée dans la paragraphe traitant des données quantitatives, il faut être conscient du fait que la transformation que l'on fait subir à une variable quantitative en en regroupant les valeurs à l'intérieur de classes a pour effet de transformer celle-ci en variable qualitative ordinale discrète

Exemple: le prix du foncier constructible par classe

Le prix du foncier au m² demeure fondamentalement une variable quantitative continue. Mais comme cela peut être le cas lorsque les données sont nombreuses et lorsque que l'on souhaite cartographier le phénomène, on est amené à regrouper ces valeurs sous forme de classes afin d'en améliorer la lecture et l'analyse. Cette transformation contribue à modifier les propriétés de la variable: de quantitative continue elle devient qualitative ordinale discrète

Parcelle	Prix de vente p_i TTC (€/m²)
p1	78,24
p2	81,15
p3	69,65
p4	101,54
p5	97,89
p6	77,23
p7	54,56
p8	98,21
p9	65,32
p10	113,33
...	...
p124	108,79
p125	93,66

Après regroupement, on obtient, par exemple :

Classe de prix (€/m²)	Effectif (fréquence absolue)	Fréquence relative (%)
< à 50 €/m²	13	10,4
de 50 à 74,99 €/m²	29	23,2
de 75 à 99,99 €/m²	57	45,6
>= à 100 €/m²	25	20,0
Total	125	100,0

Le processus qui vise à la fabrication des classes (ou discrétisation) est une opération délicate qui sera abordée plus avant.

L'échelle d'intervalle : cette forme concerne les données et variables se référant à des unités de mesure constantes mais dont le point zéro est fixé arbitrairement ne correspondant en rien à l'absence de phénomène. L'exemple le plus significatif pour ce cas est celui de la température: l'unité de mesure est constante une fois le système de référence défini (Celsius ou Fahrenheit) et le zéro est totalement arbitraire : dans le cas du système Celsuis °C le zéro correspond à la température de congélation de l'eau alors que dans le cas du système Fahrenheit °F, le zéro équivaut à la température de solidification d'un

mélange à part égal d'eau et de chlorure d'ammonium (Fahrenheit , 1724). Profitant du caractère quantitatif de la variable température, une relation peut cependant être établie entre les deux systèmes comme suit : °F = 1,8 °C +32 et inversement °C = (°F – 32) / 1,8. 0°C tout comme 0°F ne correspondent pas à une absence de température. Même en considérant le zéro absolu (0 °K = - 273,15 °C) , température la plus basse que l'on puisse observer dans l'univers et à laquelle tout mouvement moléculaire et atomique est stoppé compte tenu d'un état énergétique minimal, la température demeure une variable appartenant à l'échelle d'intervalle.

Une variable appartenant à l'échelle d'intervalle a ceci de spécifique que les valeurs qui la composent ne sont pas des multiples les unes de autres, et donc que les intervalles entre valeurs ne sont pas constants. Un exemple: on a relevé le 12/06/2008 à Moscou une température de 11°C. Le lendemain, on mesure une température de 22 °C à la même heure. Il a donc fait plus chaud le 13/12/2008 que la veille mais on ne peut cependant pas affirmer qu'il y a fait deux fois plus chaud.

L'échelle d'intervalles, en plus des opérations arithmétique s classiques, autorise la plupart des calculs statistiques : moyenne arithmétique, écart-type, coefficient de corrélation, variance, covariance, etc. Par contre, elle ne permet pas le calcul de la moyenne géométrique ou du coefficient de variation.

En dehors de la température, quantité d'autres variables se réfère à l'échelle d'intervalles. Parmi celle-ci, on peut citer l'échelle de Richter de mesure d'intensité des tremblements de terre, la mesure du temps via notre calendrier grégorien,

Les variables quantitatives d'intervalle peuvent être de nature *discrète* ou *continue*. On a vu plus haut à quoi correspondait la caractère « discret » de la données, voyons maintenant en quoi consiste sa nature « continue »

Définition : variable continue

Une variable *continue* peut, à l'inverse de la variable discrète, prendre un nombre infini ou non dénombrable de valeurs. Il n'y a, de ce fait, plus de modalité ou plutôt une infinité de modalités car entre deux valeurs données toutes les nuances de transitions sont possibles. Le cas « continu » ne concerne donc que les variables dites quantitatives pour lesquelles il peut y avoir autant de modalités qu'il y a d'individus.

> Exemple: la variable « température » est une variable *quantitative d'intervalle continue.* Celle-ci peut en effet prendre une infinité de valeurs quelles que soient les limites retenues. Par exemple, entre 10 et 12 °C, la variable peut prendre n'importe laquelle des innombrables valeurs existantes et mesurables : 10,007 °C, 11,11 °C ou bien encore 11,9999 °C si tant que l'on soit capable d'atteindre cette précision dans la mesure.

D'une façon générale, les valeurs que peut prendre une *variable quantitative continue* appartiennent à l'ensemble des nombres réels R alors que les valeurs caractérisant une appartiennent quant à elles à l'ensemble des nombres entiers N, comme par exemple le nombre d'habitants.

L'échelle proportionnelle ou échelle de rapport :

A la différence de l'échelle d'intervalle, l'échelle proportionnelle ou de rapport se caractérise par des proportions égales entre les valeurs mesurées de telle sorte qu'il existe entre ces valeurs une relation mathématique directe et constante. L'échelle proportionnelle possède en outre un zéro unique et

universel. Toutes les variables faisant référence au Système International d'Unité (SI – norme ISO 1000) appartiennent à l'échelle de mesure dite proportionnelle (ou de rapport): c'est le cas des longueurs, des surfaces, des poids et des comptages d'effectifs ainsi que la mesure du temps via le SI, et toutes les variables résultantes de la combinaison d'au moins deux des unités du SI telle que la vitesse (qui n'est qu'une expression de la distance par rapport au temps), la densité de population (effectif rapporté à une surface), etc. Le zéro y est universel et signifie absence de mesure ou mesure nulle, et chaque valeur non nulle mesurée est nécessairement le multiple de n'importe quelle autre valeur mesurée. Exemple: on pourra dire qu'une personne pesant 90 kg est deux fois plus lourde qu'une personne de 45 kg ou bien encore qu'un loyer de 337,50 €/mois est 1,5 fois (ou 50 %) plus élevé qu'un loyer de 225 €/mois.

L'échelle de rapport (ou échelle proportionnelle) possède toutes les propriétés et tous les niveaux d'informations des autres échelles plus l'immense avantage de se prêter à absolument toutes les opérations arithmétiques et statistiques pouvant exister.

Une variable quantitative proportionnelle (ou de rapport) peut également être de nature discrète ou de nature continue:

Exemple: une variable quantitative proportionnelle discrète : le nombre d'habitants.

Le nombre d'habitants d'un pays ou d'une ville est une *variable quantitative discrète à échelle proportionnelle*. La dimension quantitative de la variable n'est plus à démontrer. Le fait qu'elle appartienne à l'échelle proportionnelle se justifie par le fait qu'elle possède d'une part une zéro absolu universel (zéro habitant = pas d'habitant) et qu'il existe bien entre chaque modalité une relation mathématique de proportionnalité: un pays comptant 10 millions d'habitants est bien deux fois plus peuplé qu'un pays de 5 millions d'habitants ou bien encore 10 fois plus peuplé qu'un autre de 1 million d'âmes. La nature discrète de la variable se justifie par le caractère indivisible de l'élément de base, à savoir l'habitant: ainsi, l'ensemble des valeur que peut prendre la variable « nombre d'habitants » appartient bien à l'ensemble des entiers N. Il n'est donc pas possible d'écrire qu'une ville compte 12283,18 habitants. La variable « nombre d'habitant » est donc bien une *variable quantitative discrète à échelle de rapport (ou à échelle proportionnelle)*.

Exemple: une variable quantitative proportionnelle continue :le prix du foncier constructible au m².

Comme annoncé plus haut, le prix du foncier au m² demeure fondamentalement une *variable quantitative continue* . Elle se rapporte de plus à l'échelle proportionnelle (ou de rapport). En effet, son zéro est absolu (0 €/m² signifie bien absence de prix), la proportionnalité fonctionne puisqu'un terrain affichée à un prix de 90 €/m² est bien deux fois plus cher qu'une terrain offert à 45 €/m², et l'éventail des valeurs que peut prendre la variable est infini (entre 45 et 46 €/m², il existe une infinité de prix tous en théorie plausibles). La variable « prix du foncier au m² » est donc bien une *variable quantitative continue à échelle de rapport*.

2.2 Transformation de variables qualitatives (ou non-métriques) en variables quantitatives (ou métriques)

Certains traitements et analyses sur des données et variables qualitatives nécessitent voire exigent que ces dernières présentent une forme « pseudo quantitative » en lieu et place de leur forme « nominale ». C'est notamment le cas lorsqu'il s'agit d'utiliser des variables qualitatives dans un traitement multivarié ou simplement lorsque l'on désire les rendre manipulables et compatibles avec des logiciels statistiques. Il faut

donc faire subir à la variable une transformation lui conférant ce caractère « pseudo numérique », une transformation qui s'apparente davantage à un codage de l'information qualitative en information numérique.
Cette transformation doit cependant respecter certaines règles. En effet, dès lors que l'on introduit une dimension numérique, il s'instaure naturellement une hiérarchie qui doit respecter celle sous-jacente, si elle existe, à la dimension qualitative de la variable traitée. C'est le cas exclusivement des variables qualitatives ordinales. L'exemple qui suit illustre parfaitement cette règle.

> Exemple: la variable qualitative ordinale « moral des ménages français » propose les cinq modalités suivantes: Très bon, Bon, Moyen, Mauvais et Très mauvais. L'encodage numérique de la variable doit se faire en respectant son caractère ordinal initial. Ce faisant, on obtient le codage suivant:
>
> 5 = Très bon
> 4 = Bon
> 3 = Moyen
> 2 = Mauvais
> 1 = Très mauvais
>
> Cela dit, il s'agit d'un codage possible parmi d'autres.

Pour ce qui est des variables qualitative nominales, donc sans hiérarchie identifiable, cette règle ne s'applique plus comme le montre l'exemple qui suit:

> Exemple: la variable qualitative nominale « sexe » propose les deux modalités suivantes: Masculin et Féminin. Dans ce cas, l'encodage numérique n'a aucune hiérarchie à respecter mais doit seulement reproduire la distinction entre modalités. On peut ainsi indifféremment écrire:
>
> | | |
> |---|---|
> | 1 = Masculin | 1 = Féminin |
> | 2 = Féminin | 2 = Masculin |
>
> On évitera simplement l'utilisation du zéro davantage synonyme d'absence de phénomène.

Une autre règle est à respecter qui impose des « distances » ou intervalles égaux entre modalités lors de l'encodage numérique. Ainsi, pour reprendre un des exemples précédents, si 3 correspond à la modalité « Moyen » et 4 à la modalité « Bon », soit une « distance » de 1 entre les deux, on utilisera logiquement 5 pour « Très bon » et non 7 ou 8. De même, on affectera la valeur 2 à « Mauvais ».

Il est à noter que les nombres affectés aux modalités qualitatives en vue de leur transformation n'ont pas de signification et ne peuvent faire l'objet d'opérations arithmétiques comme par exemple le calcul d'une somme ou d'une moyenne. En réalité, ce sont des « *numéros* » qui ne modifient en rien les propriétés fondamentales rattachées aux variables qualitatives, qu'elles soient nominales ou ordinales. La transformation d'une variable qualitative en variable « numérique » ne lui confère en rien les propriétés de cette dernière. C'est pourquoi on parle davantage de transformation « pseudo-numérique ».

2.3 Transformation de variables quantitatives (ou métriques) en variables qualitatives (ou non-métriques)

L'opération inverse, c'est-à-dire la transformation d'une variable quantitative en variable qualitative, est également possible et même souhaitable dans certains cas de figures même si elle demeure plus délicate et impose de ce fait le respect de règles beaucoup plus strictes.

La plupart du temps la transformation d'une variable quantitative en une variable qualitative passe la constitution de classes à partir de la distribution observée. Cette opération est appelée discrétisation puisque, quelle que soit la nature des données quantitatives en amont (intervalle ou de rapport, discrète ou continue), elle aboutit inévitablement à la fabrication d'une variable qualitative ordinale discrète. Il est donc important d'avoir à l'esprit que cette transformation engendre une perte d'information et également une diminution de la capacité d'analyse et traitement des données puisque certains paramètres ne seront plus calculables précisément à partir d'une distribution discrète (moyenne, écart-type, etc.). En effet, chaque classe définie regroupe sous une même identité, selon un même caractère des individus qui à l'origine se distinguaient les uns des autres par des valeurs différentes. On soupçonne ici l'importance que revêt le processus d'élaboration des classes (définition des limites de classes, étendue des classes, nombre de classes, etc.), le but final étant de synthétiser un volume important d'informations en limitant la perte liée à la discrétisation. Autrement dit, il s'agit de maximiser la réduction de contenu informationnelle d'une distribution en en minimisant les pertes.

C'est un mal pour un bien et la transformation de données quantitatives en données qualitatives via la discrétisation demeure souvent incontournable. Il est en effet souvent bien plus commode et pertinent pour la lecture, l'analyse, l'interprétation ou la représentation d'un phénomène de regrouper les individus à l'intérieur de classes plutôt que de s'éreinter à essayer de lire et d'interpréter un tableau contenant des centaines voire des milliers de valeurs.

Il existe plusieurs méthodes plus ou moins complexes et élaborées en vue de la discrétisation d'une distribution de valeurs sachant que pour ce faire rien ne remplace le bon sens et la connaissance que l'on a du phénomène étudié. Lorsque cette expérience existe, les méthodes mises à disposition ne sont souvent là que pour assister l'utilisateur. Dans les autres cas, elles permettent d'orienter de façon objective la stratégie de discrétisation. Attention, certaines des méthodes présentées ci-après font appel à des notions qui ne seront vues que plus tard dans le cours : c'est le cas notamment de celle faisant appel à l'écart-type.

Il existe donc trois groupes de méthodes de discrétisation:

- les méthodes empiriques : basées sur l'expérience et la connaissance du phénomène étudié, elles utilisent en plus l'allure de la distribution pour y déceler des ruptures naturelles et ainsi délimiter les bornes des classes à créer. Cette méthode, pour partie visuelle, nécessite une bonne connaissance du phénomène à traiter.

  Exemple: on dispose des loyers surfaciques mensuels hors charges pour l'ensemble des logements locatifs sociaux d'un département, soit au total plus de 9 500 individus (= logements) avec, pour chacun d'eux, des valeurs dans 5 variables (loyers, nombres de pièces, surface, localisation, financement). Au total nous disposons donc de 9 500 x 6 = 57 000 valeurs. Hormis, dans un premier, le calcul de paramètres comme la moyenne, il semble difficile d'appréhender et d'analyser cette masse de données. Réduire la taille du tableau en opérant un regroupement des valeurs selon un système adapté de classes apparaît comme étant la meilleure solution pour y parvenir.
  La connaissance du phénomène que l'on a ainsi qu'une analyse de l'histogramme de la variable loyer nous permettent assez rapidement d'identifier les « cassures naturelles » pouvant servir de limites de classes (Cf. figure 3). Le nombre de classes alors défini est de 5, organisées de la façon suivante (il s'agit d'une possibilité parmi tant d'autres):

Numéro de classe	Borne (ou limite) inférieure	Borne (ou limite supérieure)	Étendue ou amplitude	Écriture
1	0	2,99	2,99	[0 ; 3,0[ou « Moins de 3»
2	3,0	3,99	0,99	[3,0 ; 4,0[ou « de 3,0 à 3,99 »
3	4,0	5,99	1,99	[4,0 ; 6,0[ou « de 4,50 à 5,99 »
4	6,0	8,99	3,99	[6,0 ; 9,0[ou « de 6,0 à 8,99 »
5	9,0	∞	∞	[9,0 ; ∞ [ou « Plus de 9,0 »

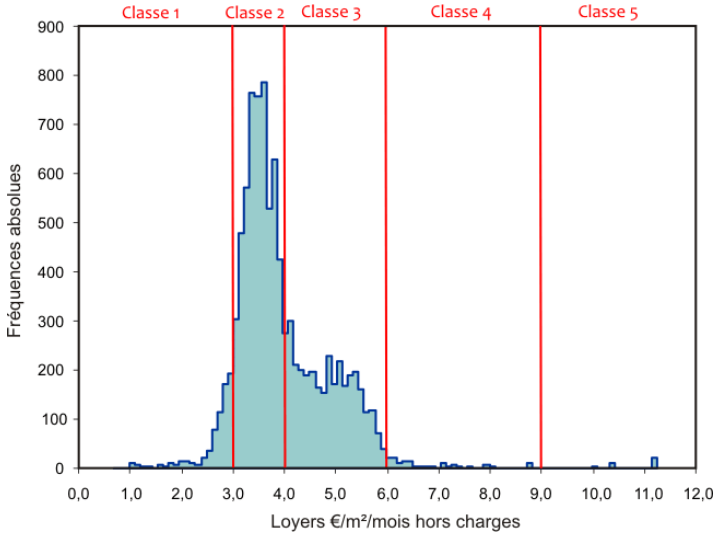


Figure 3 : histogramme de la variable loyers et discrétisation empirique

On pourrait affiner la discrétisation en subdivisant certaines des classes mais ce n'est pas forcément souhaitable, un nombre trop important de classes conduisant souvent à une dilution du phénomène et à une dispersion de la capacité d'analyse du lecteur. Autant que faire se peut, on choisit des valeurs de limites de classes correspondant à des nombres « repères » pour l'interprétation, c'est-à-dire se terminant en 0 ou 0,5 ou faute de mieux par un chiffre rond. Il faut éviter de préférence de définir des limites de classes avec des valeurs inutilement décimalisées ou éloignées des repères naturels de l'esprit (10, 25, 50, etc.) qui ne favorisent pas une interprétation immédiate (par ex. de 13,27 % à 21,86 % ou bien encore de 17 à 33) mais ce n'est pas toujours possible.

Une fois la discrétisation réalisée, la distribution du phénomène se présente comme suit:

Classes	Classes	Fréquence absolue	Fréquence relative (%)
[0 ; 2,5[	Moins de 3,0	685	7,2
[2,5 ; 4,5[	De 3,0 à 3,99	5 359	56,3
[4,5 ; 6,0[	de 4,0 à 5,99	3 287	34,5
[6,0 ; 9,0[	de 6,0 à 8,99 »	142	1,5
[9,0 ; ∞ [	Plus de 9,0	41	0,4

A partir de là, interprétation, graphiques et cartes deviennent plus aisées. Par contre, le contenu informationnel initial s'est fortement dégradé: là où il y avait une multitude de cas de figures entre les valeurs 3,0 et 3,99, il n'y en a plus qu'un seul après regroupement.

- les méthodes par défaut qui ne nécessitent ni une connaissance approfondie du phénomène ni une étude de la distribution. Leur simplicité est à la hauteur des approximations qu'elles génèrent et elles ont tendance, de fait, à lisser le phénomène étudié. Leur principe est simple: prenant en compte ou l'effectif total de la population étudiée ou l'amplitude totale de la distribution de la variable étudiée, ces méthodes proposent, dès lors qu'un nombre souhaité de classes est défini :

  - soit une discrétisation en classes d'égale amplitude,
  - soit une discrétisation en classes d'égal effectif.

  Exemple: reprenons l'exemple précédent. Nous disposions d'une population de 9 517 individus, en l'occurrence des logements locatifs privés, pour lesquels nous connaissons les loyers et d'autres caractéristiques. La simple consultation des données initiales nous permet de relever la valeur minimale et la valeur maximale de loyer pour calculer l'amplitude totale de la distribution:

  - Nombre total d'observations : 9 517
  - Valeur minimale observée de loyer: 0,68 €/m² mensuel hors charges
  - Valeur maximale observée de loyer: 11,26 €/m² mensuel hors charges
  - Amplitude totale de la distribution = 11,26 – 0,68 = 10,58

  Méthode des classes d'égale amplitude:
  Si l'on décide de créer 5 classes, la discrétisation en classes d'égale amplitude donnera des classes dont l'étendue sera identique et équivalente à: 10,58 / 5 = 2,11 €/m². Il suffit alors, pour former les limites de la première classe, de prendre la valeur minimale pour la borne inférieure et de lui ajouter 2,11 pour obtenir la borne supérieure. Pour la deuxième classe, on reprend la borne supérieure de la classe précédente en l'augmentant légèrement pour éviter le recouvrement (+ 0,01) et on lui ajoute toujours 2,11 pour obtenir la borne supérieure. On répète l'opération pour les classes suivantes:

	Borne inférieure	Borne supérieure	Fréquence absolue	Fréquence relative
Classe 1	0,68	0,68 + 2,11 = 2,79	351	3,7
Classe 2	2,80	2,80 + 2,11 = 4,91	7 520	79,0
Classe 3	4,92	4,92 + 2,11 = 7,03	1 549	16,3
Classe 4	7,04	7,04 + 2,11 = 9,15	54	0,6
Classes 5	9,16	9,16 + 2,11 = 11,27	40	0,4

  La dernière borne de la distribution doit être égale (aux arrondis prêts) à la valeur maximale observée dans la distribution (ici on 11,27 pour 11,26). Une fois les classes délimitées, il suffit, à partir de la distribution initiale, de « mettre » chaque individu dans la classe correspondant à sa valeur de loyer et de compter les effectifs par classe (voir tableau ci-dessus). On observe que le résultat obtenu est sensiblement différent de celui de la méthode empirique: les classes extrêmes notamment apparaissent plus faiblement représentées.

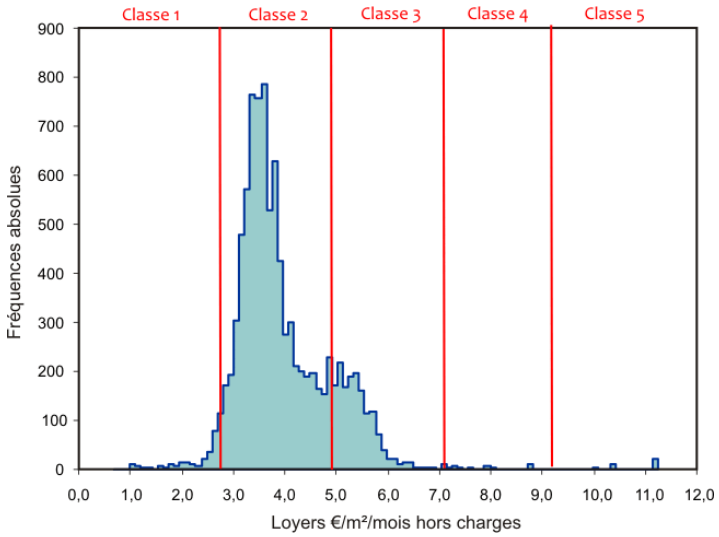


Figure 4: histogramme de la variable loyers et discrétisation selon la méthode des classes d'égale amplitude

Méthode des classes d'égal effectif:

Cette méthode est uniquement basée sur l'effectif total se rapportant à la distribution (dans notre exemple 9 517 logements). Avec un nombre de classe inchangé (5), la discrétisation en classes d'égale effectif donnera des classes contenant le même nombre d'individus, soit 9517 / 5 = 1903 logements. Pour déterminer les limites inférieure et supérieure d'une classe, il suffit de lire la valeur de loyer correspondant au rang du premier et du dernier individu la composant. Exemple, pour définir les bornes de la classe 1, on lit la valeur de loyer de l'individu de rang 1 (soit 0,68 €/m²) et la valeur de loyer de l'individu de rang 1903 (dans notre 3,31 €/m²) en ayant pris soin auparavant de classer les valeurs en ordre croissant. La borne inférieure de la deuxième classe correspondra à la valeur de l'individu de rang 1904 (également 3,31, on passe à 3,32 pour éviter le recouvrement), quant à la borne supérieure, elle correspondra à la valeur de loyer prise par l'individu de rang 1904+1903 = 3807 (ici 3,57). On répète l'opération pour les classes restantes et on obtient la classification suivante:

	Borne inférieure	Borne supérieure	Fréquence absolue	Fréquence relative
Classe 1	0,68	3,31	1903	20,0
Classe 2	3,32	3,57	1903	20,0
Classe 3	3,58	3,90	1903	20,0
Classe 4	3,91	7,79	1903	20,0
Classes 5	4,80	11,26	1903	20,0

Une discrétisation qui tranche avec celles obtenues précédemment

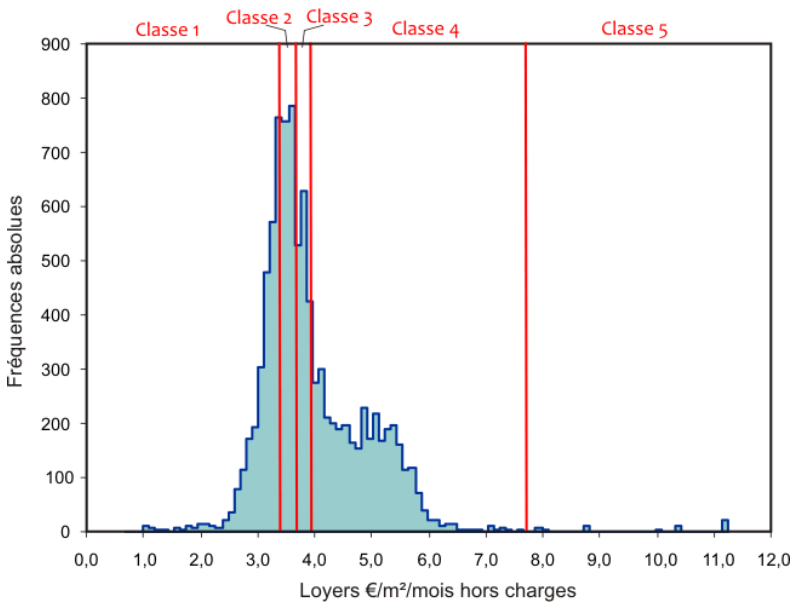


Figure 5: histogramme de la variable loyers et discrétisation selon la méthode des classes d'égal effectif

- les méthodes statistiques basées sur les paramètres de tendance centrale et de dispersion

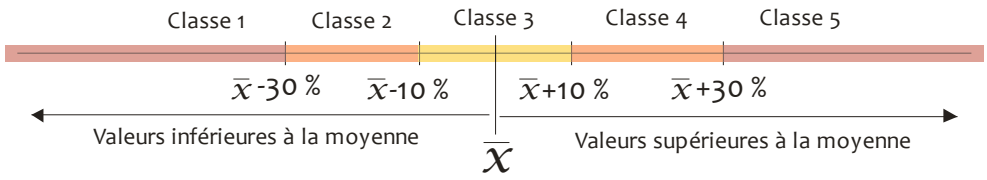
Sur la base de la moyenne

Les individus d'une distribution peuvent être répartis dans des classes en fonction de leur rapport à la moyenne. Cette approche permet souvent une comparaison plus facile des individus entre eux.

On crée une classe centrale regroupant les valeurs de la distribution proche de la moyenne à ± 10 % par exemple.

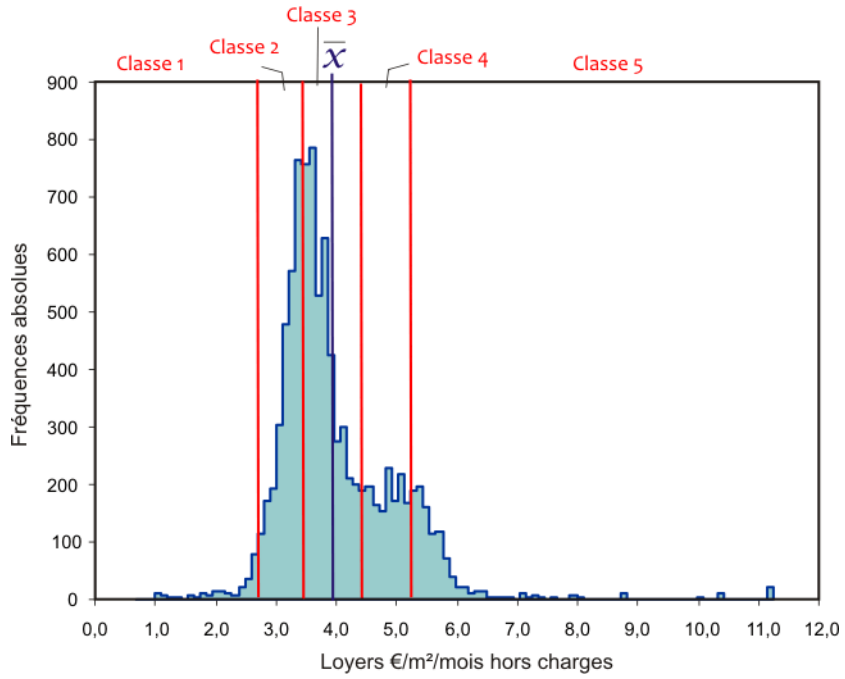
Les classes suivantes, dont les bornes restent libres de choix, contiennent quant à elles des individus dont la valeur est de plus en plus éloignée de celle de la moyenne.

On peut ainsi construire les 5 classes de la façon suivante:



Pour obtenir le découpage suivant :

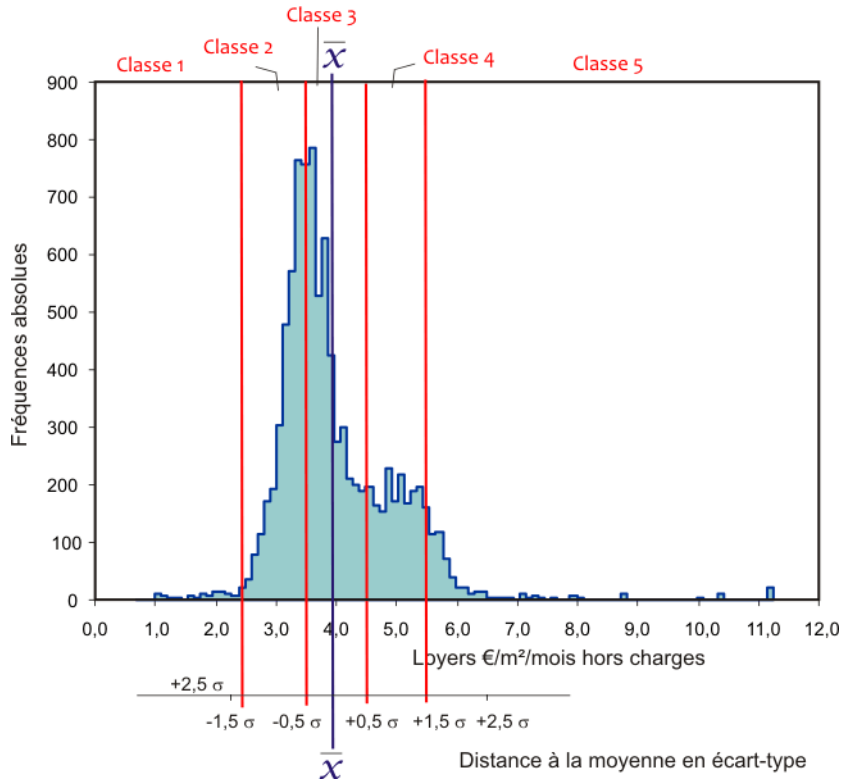
	Borne inférieure	Borne supérieure	Fréquence absolue	Fréquence relative
Classe 1	0,00	2,79	351	3,7
Classe 2	2,78	3,57	3 471	36,5
Classe 3	3,58	4,38	3 119	32,8
Classe 4	4,39	5,18	1 403	14,7
Classes 5	5,19	+∞	1 170	12,3



Sur la base de l'écart-type :

Les individus d'une distribution peuvent aussi être répartis dans des classes en fonction de leur distance rapport à la moyenne en unité d'écart-type de la distribution. Pour ce faire, il faut transformer la valeur de chaque individu en unité de distance à la moyenne en unité « écart-type ». Cette transformation est appelée standardisation et s'effectue de la façon suivante :

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$



Exercice 18 : fichier Excel associé « Exercice 20 - Discrétisation.xls ». Il est conseillé de réaliser cette exercice après avoir pris connaissance du chapitre 4 sur les mesures de tendance centrale et de dispersion.

2.2 Organiser les données: les tableaux statistiques

Tout traitement,toute représentation ou toute analyse d'un ensemble de données se rapportant aux unités d'une population impose que ces dernières aient été au préalable rangées et organisées dans une structure facilitant leur manipulation. La façon la plus simple et la plus efficace de structurer la données reste le tableau de son expression la plus basique (vecteur) à son expression la plus complexe (tableau multidimensionnel).

2.2.1 Vecteur ou série brute

Même si cette façon, la plus rudimentaire qui soit, d'organiser la donnée est rare et peu commode, il convient malgré tout d'en parler. Le vecteur ou la série brute consiste à énumérer les unes à la suite des autres en dans leur ordre d'apparition ou de collecte les données. Exemple: le nombre de villes de plus de 1 000 000 habitants sur chacun des 5continents en 2005 s'écrit :

$$S1 = \{82;181;75;40;54;6\}$$

On peut également écrire la série S1 en y ordonnant de façon croissante ou décroissante les données comme suit :

$$S2 = \{6;40;54;75;181\}$$

Mais ce type d'écriture ne permet pas de faire correspondre individu et donnée. On parle alors de série ou de vecteur non classé non identifié dans le premier cas (S1) et de série ou de vecteur classé non identifié dans le second (S2). Afin de réaliser la correspondance entre individus et données, il suffit d'accoler à la données concernée l'identifiant de l'individu auquel elle correspond.
Ainsi, une série non classée et non identifiée devient la série S3 non classée mais identifiée:

S3 = {(Europe,82);(Asie,181);(Amérique du Nord,75);(Amérique du Sud,40);(Afrique,54);(Océanie,6)}

Et la série classée non identifiée S2 devient la série S4 classée identifiée:

S4 = {(Asie,181);(Europe,82);(Amérique du Nord,75);(Afrique,54);(Amérique du Sud,40);(Océanie,6)}

Un modèle d'organisation qui peut encore fonctionner lorsque le nombre d'individus est réduit mais devient rapidement lourd et susceptible d'entraîner des erreurs à l'écriture c'est pourquoi on lui préfère une présentation des données sous forme de tableau.

2.2.2 Les tableaux

Quelques notions et définitions de base:
Un tableau est composé de lignes et colonnes. Par convention – mais ce n'est pas une obligation – les individus forment les lignes et les variables (ou caractères) les colonnes. La rencontre d'une ligne et d'une colonne constituant une cellule destinée à contenir la donnée caractéristique l'individu i pour la variable j . Chaque donnée est donc repérable dans un tableau par un couple de coordonnées (i,j) , i figurant la ligne et j la colonne.
La taille d'un tableau correspond au nombre de cellules qui le composent. Elle est obtenue en multipliant le nombre de lignes L par le nombre de colonnes C (L x C) (m x n)??.

Exemple: on interroge 10 individus sur leur taille, leur poids et leur sexe. Le tableau résultant de l'enquête comportera L = 10 lignes et C = 3 colonnes soit 10x3 = 40 cellules = 30 données. Chaque donnée est localisable dans le tableau par ses coordonnées comme suit:

Colonnes / Lignes	Taille	Poids	Sexe
Individu 1	(1,1)	(1,2)	(1,3)
Individu 2	(2,1)	(2,2)	(2,3)
Individu 3	(3,1)	(3,2)	(3,3)
Individu 4	(4,1)	(4,2)	(4,3)
Individu 5	(5,1)	(5,2)	(5,3)
Individu 6	(6,1)	(6,2)	(6,3)
Individu 7	(7,1)	(7,2)	(7,3)
Individu 8	(8,1)	(8,2)	(8,3)
Individu 9	(9,1)	(9,2)	(9,3)
Individu 10	(10,1)	(10,2)	(10,3)

Les coordonnées des données dans un tableau

La dimension d'un tableau est donnée par le nombre de variables se rapportant aux individus d'une même population. Dans notre exemple, le tableau comporte 3 dimensions (taille, poids et sexe).
Chaque individu est repérable dans l'espace de travail par un ensemble de coordonnées correspondant aux valeurs prises par celui-ci dans chacune des variables. On parle alors de coordonnées thématiques.
Exemple : dans le tableau qui suit, l'individu 1 a comme coordonnées thématiques : (1,82;78,M)

	Taille (m)	Poids (kg)	Sexe
Individu 1	1,82	78	M
Individu 2	1,67	61	F
Individu 3	1,71	70	F
Individu 4	1,75	69	M
Individu 5	1,88	82	M
Individu 6	1,69	55	F
Individu 7	1,72	71	M
Individu 8	1,90	92	M
Individu 9	1,85	88	F
Individu 10	1,64	59	F

Il est possible de représenter graphiquement les individus en fonction de leurs coordonnées thématiques dans un repère géométrique (x,y) pour 2 dimensions et (x,y,z) pour 3 dimensions, sachant qu'une représentation graphique n'est plus possible au-delà de 3 dimensions même si statistiquement et mathématiquement il demeure tout à fait possible de gérer et manipuler des tableaux dont la dimension est supérieure à 3.

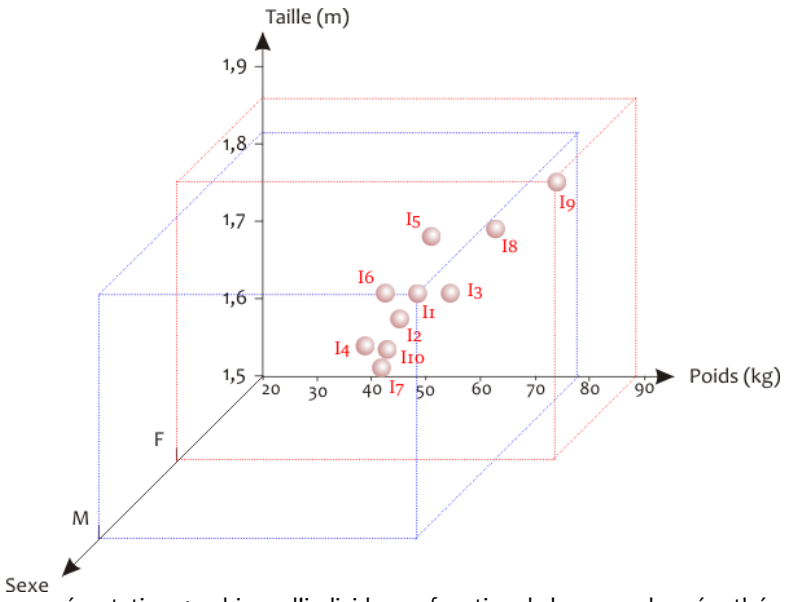


Figure : représentation graphique d'individus en fonction de leur coordonnées thématiques

Le tableau constitue l'étape intermédiaire entre la donnée brute et le graphique. Malgré un aspect quelque fois rébarbatif, le tableau véhicule souvent davantage d'informations que le graphique mais demeure, il est vrai, plus inaccessible à une lecture rapide et concise d'un phénomène. Une inaccessibilité qui va croissante avec sa taille.

Les différents types de tableaux :

2.2.2.1 Les tableaux unidimensionnels

Un tableau unidimensionnel ne concerne qu'une seule variable et par là même une seule distribution. Le tableau de l'exemple précédent comporte 3 dimensions. Il est possible de le scinder en 3 tableaux d'une seule dimension, en tableaux unidimensionnels. Chaque tableau correspond alors à une distribution (Cf. figure ci-dessous).

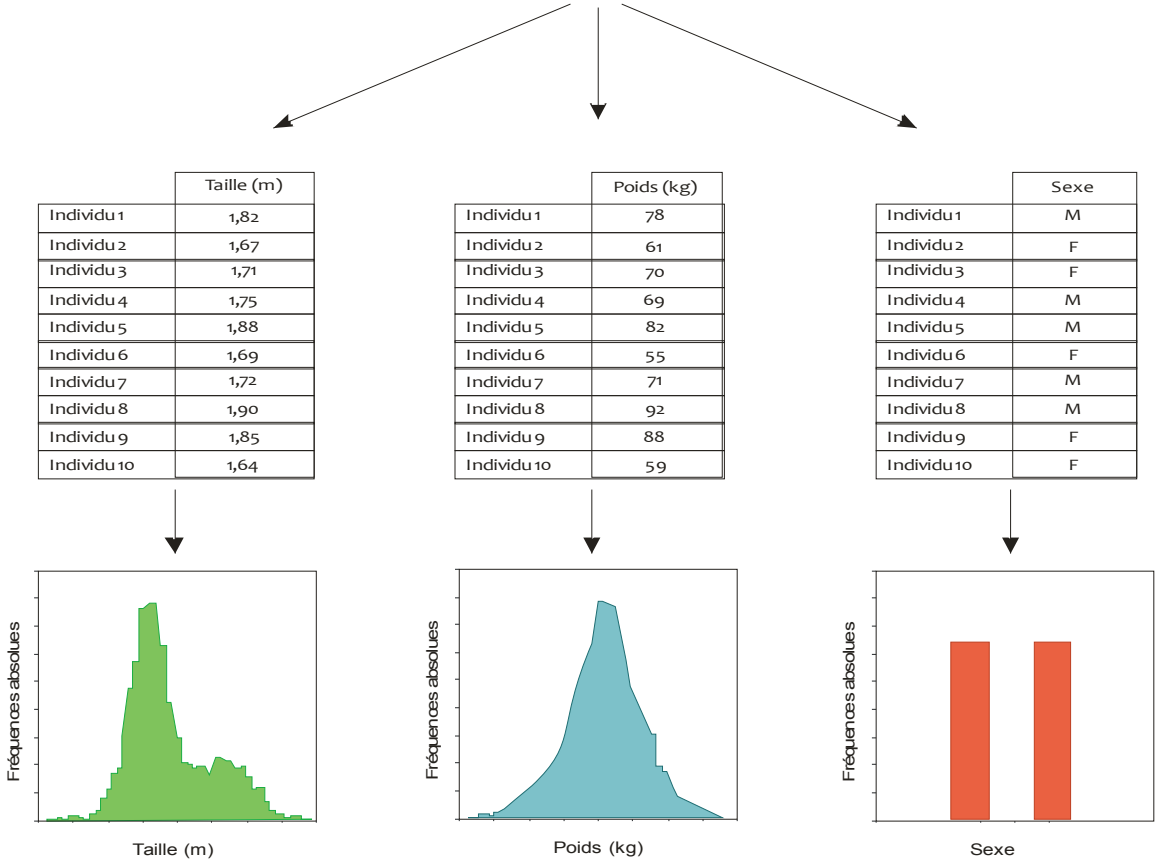
Le tableau unidimensionnel peut se présenter sous deux formes :

- une 1ère forme faisant correspondre individus et variable. Dans ce cas, chaque cellule du tableau contient la valeur de la variable prise par l'individu lui correspondant. On travaille ici sur des données brutes qui autorisent le calcul de la moyenne, de la médiane, des quartiles ou bien encore de la variance.

- Une 2ème forme où les lignes du tableau ne correspondent plus aux individus mais aux modalités de la variable étudiée (attention, s'il s'agit d'une variable continue, une discrétisation est nécessaire). Dans ce cas de figure, les cellules du tableau contiennent alors les effectifs relatifs à chaque modalité. Le regroupement des individus ne permet plus le calcul des paramètres de tendance centrale inhérents à la distribution initiale. Ce mode de représentation permet cependant, par ses aspects synthétique, une meilleure lisibilité du phénomène. Cela est surtout vrai lorsque le nombre d'individus est important.

Les formes restent complémentaires et nécessaires dans une approche globale et complète d'un phénomène

	Taille (m)	Poids (kg)	Sexe
Individu 1	1,82	78	M
Individu 2	1,67	61	F
Individu 3	1,71	70	F
Individu 4	1,75	69	M
Individu 5	1,88	82	M
Individu 6	1,69	55	F
Individu 7	1,72	71	M
Individu 8	1,90	92	M
Individu 9	1,85	88	F
Individu 10	1,64	59	F



	Taille (m)
Individu 1	1,82
Individu 2	1,67
Individu 3	1,71
Individu 4	1,75
Individu 5	1,88
Individu 6	1,69
Individu 7	1,72
Individu 8	1,90
Individu 9	1,85
Individu 10	1,64

	Fréquence absolue	Fréquence relative	Fréquence absolue cumulée	Fréquence relative cumulée
Moins de 1,70 m	3	30%	3	30%
de 1,70 à 1,79 m	3	30 %	6	60 %
1,80 m et plus	4	40 %	10	100 %
Total	10	100 %		

Figure : deux formes de présentation d'un tableau unidimensionnel

Exercice 2 : fichier Excel associé « Exercice 2 - Tableau à une dimension.xls ».

2.2.2.2 Les tableaux croisés à n dimensions ($n \geq 2$)

Le tableau croisé à n dimensions est appelé ainsi car il « croise » n distributions. Il va sans dire que plus n est grand, plus le tableau comporte de cellules et plus il devient difficile à lire. D'une manière générale, on considère qu'au-delà de 4 dimensions, la lecture d'une tableau croisé devient un exercice compliqué.
Un tableau croisé ne peut contenir que des effectifs (fréquences absolues ou relatives). Dans la quasi totalité des cas de figures, les variables figurées dans les tableaux croisés le sont sous forme discrète, représentées par un nombre fini de modalités. Lorsque que le nombre de dimensions est supérieur à 2, les variables et leurs modalités sont imbriquées. L'exemple qui suit devrait permettre de saisir toutes les propriétés et subtilités des tableaux croisés.

Exemple: tableau croisé relatif aux caractéristiques du parc locatif loué vide dans le département de Haute-Saône (données RGP 1999 – Insee):

Pour caractériser le parc locatif loué vide du département de la Haute-Saône, nous avons retenu les variables suivantes déclinées en modalités :

- Nombre de pièces (1 pièce, 2 pièces, 3 pièces, 4 pièces, 5 pièces et plus),
- Époque de construction (Avant 1915, de 1915 à 1948, de 1949 à 1967, de 1968 à 1981, de 1982 à 1989, 1990 et après)
- Type de logement (Individuel, Collectif)
- Statut (Parc Locatif Social, Parc Locatif Privé)

Tableau à 2 dimensions (ou tableau bidimensionnel) : retenons pour sa construction les 2 variables les plus représentatives de la caractéristique d'un parc de logement notamment dans la formation des loyers, à savoir le nombre de pièces et l'époque de construction. Leur croisement, époque de construction en ligne et nombre de pièces en colonnes, aboutit à un tableau croisé de dimension 2 comme suit :

	1 pièce	2 pièces	3 pièces	4 pièces	5 pièces et +
] Avant 1915 [	407	1457	2318	2094	1602
[1915 à 1948]	161	526	857	718	555
[1949 à 1967]	387	1132	2789	2889	1555
[1968 à 1981]	331	558	908	855	530
[1982 à 1989]	251	280	333	322	295
[1999 et après [	108	431	870	946	601

Chaque cellule du tableau croisé contient le nombre d'individus répondant strictement aux critères des modalités dont elle est issue. Ainsi, dans notre exemple, la cellule mise en valeur, de coordonnées (2,3), contient-elle le nombre de logements locatifs répondant à la fois au critère « 3 pièces » et au critère « de 1915 à 1948 » : 857 constitue le nombre de logements locatifs composés de 3 pièces et construits entre 1915 et 1948.

Tableau à 3 dimensions : au tableau précédent, il est possible de rajouter une dimension, c'est-à-dire une variable. Rajoutons la variable « type de logement » composées des modalités « Individuel » et « collectif ». Dans la mesure où nous sommes limités graphiquement par une représentation en 2 dimensions, la dimension supplémentaire doit être rajoutée soit en ligne, soit en colonne. On décide de la rajouter en ligne. Comme il y existe déjà une dimension (ou

variable), celle ajoutée doit y être déclinée pour chacune des modalités de la variable existante comme figuré dans le tableau qui suit; on dit alors que les dimensions sont imbriquées :

		1 pc	2 pc	3 pc	4 pc	5 pc+
] Avant 1915 [	Individuel	70	388	1075	1259	1207
	Collectif	337	1069	1243	835	395
[1915 à 1948]	Individuel	26	168	375	450	415
	Collectif	135	358	482	268	140
[1949 à 1967]	Individuel	21	116	387	758	729
	Collectif	366	1016	2402	2131	826
[1968 à 1981]	Individuel	14	48	186	429	409
	Collectif	317	510	722	426	121
[1982 à 1989]	Individuel	19	52	118	209	256
	Collectif	232	228	215	113	39
[1999 et après [	Individuel	10	91	335	627	474
	Collectif	98	340	535	319	127

Le nombre total de cellules s'accroît alors que les effectifs par cellule diminue. L'information devient plus précise mais se répartie en un nombre de cas de figures plus important. La cellule surlignée (3,3) renseigne sur le nombre de logements locatifs composés de 3 pièces sis dans un immeuble type maison individuelle construite entre 1915 et 1948.

Tableau à 4 dimensions : accroissons encore un petit peu la précision des informations en ajoutant une 4^{e} variable (ou dimension) à notre tableau. De la même façon, cette nouvelle variable peut être placée en ligne ou en colonne. Afin d'équilibrer le tableau, nous décidons de localiser la nouvelle variable « statut » et ses deux modalités (Parc Locatif Privé (PLP) et Parc Locatif Social (PLS)) en ligne selon le même principe que précédemment. On obtient le tableau qui suit. L'information devient encore plus précise mais parallèlement la lecture du tableau se complexifie, à l'image de l'intitulé de chacune des cellules le composant. La cellule « exemple » suivie depuis le début de l'exercice indique que 358 individus sont des logements locatifs ayant un statut privé et composés de 3 pièces sis dans une maison individuelle construite entre 1915 et 1948.

		1 pc		2 pc		3 pc		4 pc		5 pc+	
		PLS	PLP	PLS	PLP	PLS	PLP	PLS	PLP	PLS	PLP
] Avant 1915 [	Ind.	0	70	10	378	12	1063	15	1244	24	1183
	Coll.	33	304	67	1002	79	1164	41	794	9	386
[1915 à 1948]	Ind.	2	24	12	156	17	358	16	434	20	395
	Coll.	13	122	31	327	80	402	25	243	16	124
[1949 à 1967]	Ind.	9	12	23	93	86	301	277	481	137	592
	Coll.	169	197	701	315	1791	611	1611	520	615	211
[1968 à 1981]	Ind.	9	5	17	31	104	82	244	185	108	301
	Coll.	179	138	323	187	539	183	294	132	59	62
[1982 à 1989]	Ind.	12	7	23	29	26	92	43	166	40	216
	Coll.	41	191	99	129	97	118	49	64	19	20
[1999 et après [	Ind.	1	9	12	79	103	232	247	380	95	379
	Coll.	20	78	114	226	223	312	122	197	43	84

On a fait figurer dans ces tableaux successifs des effectifs (ou fréquences absolues) mais on aurait tout aussi bien pu y faire figurer des pourcentages (ou fréquences relatives). Les tableaux croisés permettent de confronter tous les

types de données entre eux (qualitatif et quantitatif) et ce, quelle que soit l'échelle de mesure (nominale, ordinale, intervalle ou de rapport).

2.2.2.2 Les distributions marginales

Les exemples précédents de tableaux croisés n'ont fait figurer que les effectifs cellulaires. Il est possible d'étendre la capacité informationnelle des tableaux en leur adjoignant une colonne terminale supplémentaire correspondant à la somme des valeurs en ligne et une ligne terminale supplémentaire correspondant à la somme des valeurs en colonne. Cette ligne et cette colonne sont appelées **distributions marginales**.

Exemple : en reprenant le dernier tableau croisé créé à 4 dimensions et y ajoutant les distributions marginales, on obtient le résultat suivant :

		1 pc		2 pc		3 pc		4 pc		5 pc+		Total par ligne
		PLS	PLP	PLS	PLP	PLS	PLP	PLS	PLP	PLS	PLP	
] Avant 1915 [	Ind.	0	70	10	378	12	1063	15	1244	24	1183	3999
	Coll.	33	304	67	1002	79	1164	41	794	9	386	3879
[1915 à 1948]	Ind.	2	24	12	156	17	358	16	434	20	395	1434
	Coll.	13	122	31	327	80	402	25	243	16	124	1383
[1949 à 1967]	Ind.	9	12	23	93	86	301	277	481	137	592	2011
	Coll.	169	197	701	315	1791	611	1611	520	615	211	6741
[1968 à 1981]	Ind.	9	5	17	31	104	82	244	185	108	301	1086
	Coll.	179	138	323	187	539	183	294	132	59	62	2096
[1982 à 1989]	Ind.	12	7	23	29	26	92	43	166	40	216	654
	Coll.	41	191	99	129	97	118	49	64	19	20	827
[1999 et après [	Ind.	1	9	12	79	103	232	247	380	95	379	1537
	Coll.	20	78	114	226	223	312	122	197	43	84	1419
Total par colonne		488	1157	1432	2952	3157	4918	2984	4840	1185	3953	27066

La lecture des distributions marginales distingue clairement lignes et colonnes: ainsi la lecture du total par ligne ne permettra plus de déceler les modalités de de la ou des variables figurant en colonnes et vice versa. Dans notre exemple, la cellule « total par ligne » allumée nous informe sur le nombre total de logements locatifs type maison individuelle construite entre 1915 et 1948(sous-entendu toute taille de logements et tous statuts confondus). On ne peut plus distinguer dans ce total ni la taille des logements ni leur statut. Cette remarque vaut pour les totaux calculés en colonnes. A noter que la cellule donne la somme des lignes, égale à la somme des colonnes et correspondant à l'effectif total de la distribution (27 066 logements locatifs).

Exercice 3 : fichier Excel associé « Exercice 3 - Tableau croisé dynamique.xls ». Utiliser l'annexe4 si vous n'êtes pas familier avec la fonction tableau croisé dynamique d'Excel (ou d'un autre tableur).

Chapitre 3
3. Modes de représentation des données : les graphiques

Graphiques et cartes sont les corollaires d'une bonne analyse et d'une interprétation la plus complète possible de séries statistiques ou de résultats sur des traitements de données. Ces modes de représentation de la donnée participent à la compréhension des phénomènes, au même titre que les tableaux simples ou élaborés, apportant une information certes agrégée, synthétique mais très visuelle et en cela plus facile à aborder et à interpréter que ne le ferait un tableau de chiffres.
Nous avons volontairement inclus dans ce chapitre le mode de représentation cartographique même s'il convient de préciser qu'il constitue à lui seul une technique et même une science digne d'un chapitre voire d'un ouvrage à part entière. C'est pourquoi il ne sera abordé que très superficiellement mais suffisante pour en acquérir les bases.

La représentation graphique comme cartographique de données s'accompagne nécessairement d'une simplification de la réalité à représenter. Cette perte d'information – car tout simplification se traduit par une perte d'information – est compensée, et quelques fois largement, par un gain indéniable en lisibilité et en compréhension, pour peu qu'un certain nombre de règles aient été respectées à l'occasion de l'élaboration du graphique ou de la carte.

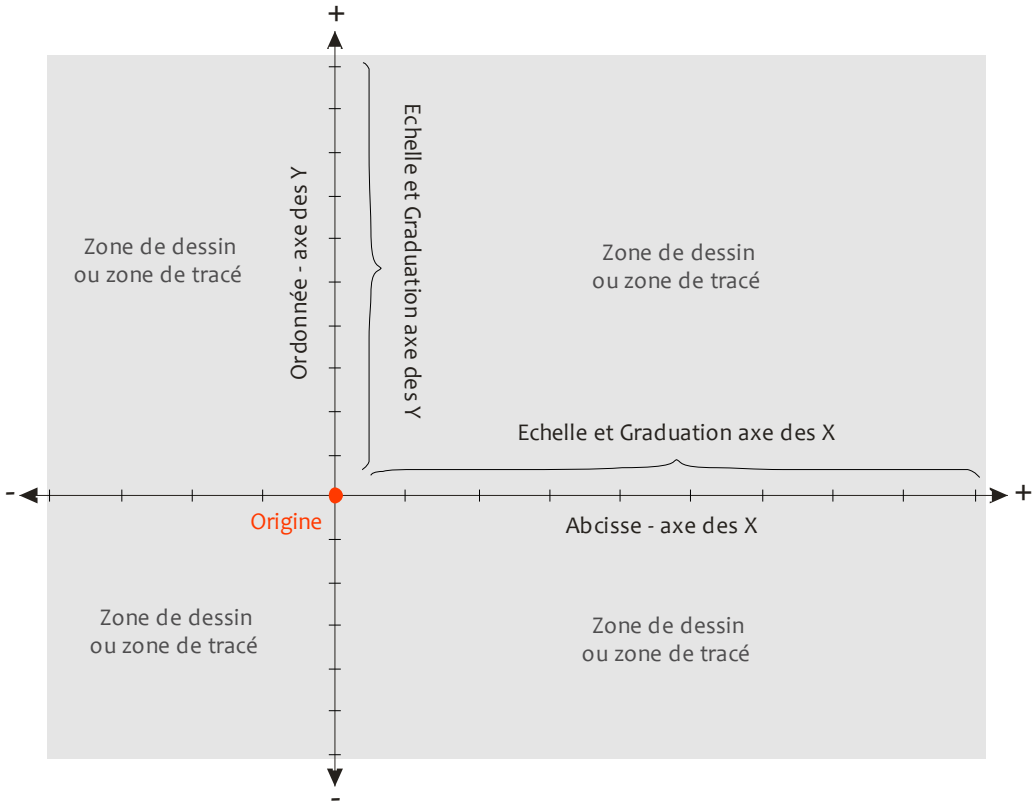
3.1 Graphiques : définition

Définition: un graphique - étymologiquement « qui figure par le dessin » - est une représentation visuelle et simplifiée d'une réalité appréhendée sous une forme essentiellement numérique (série, tableau).

Un graphique peut figurer une seule variable – au quel cas on parlera de graphique unidimensionnel – ou plusieurs variables. On parlera dans ce dernier cas de graphique multidimensionnel.

Un graphique est composé de plusieurs éléments incontournables de base. Ces sont :

- un système de coordonnées matérialisé par des axes (2 ou 3). Chaque axe représente selon les cas de figure soit une variable étudiée soit une fréquence (absolue ou relative), soit un repère temporelle (date). L'axe horizontale ou *abscisse* est par convention appelé axe des X, l'axe verticale, ou *ordonnée*, axe des Y. Les axes X et Y se croisent à angle droit en un point nommé origine. Chaque axe est gradué en fonction du type de la variable qu'il représente (qualitatif ou quantitatif, discret ou continu), de son unité de mesure et des valeurs ou modalités prises par celle-ci.
- Une zone de dessin (ou zone de tracé) à l'intérieur de laquelle est figuré le tracé issu des données X et Y. Le type de dessin ou le type de tracé dépend alors de la relation (X,Y), de ce que l'on souhaite montrer et de la nature des variables impliquées.



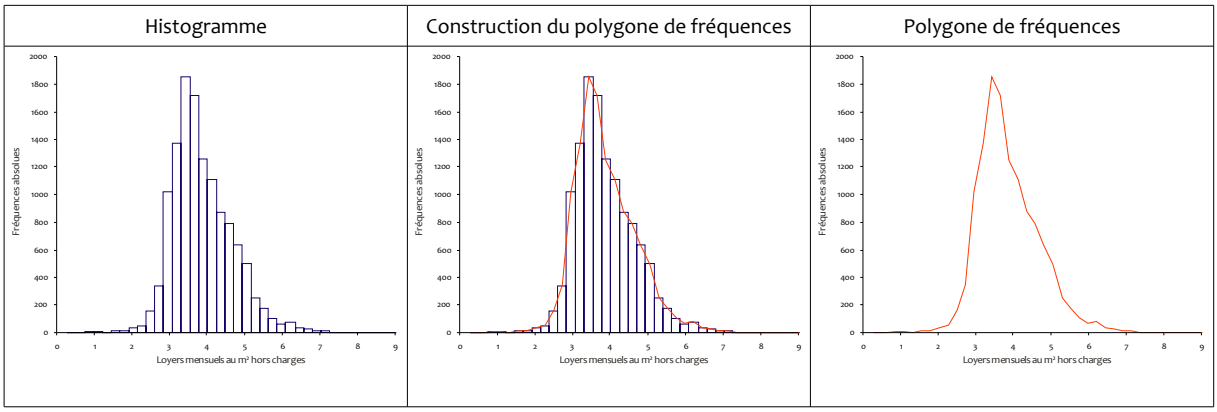
Chaque type de graphique est adapté à une ou plusieurs situation ou façon de représenter l'information. Selon la nature des données, le nombre de variables et ce que l'on souhaite montrer, il sera judicieux de choisir la représentation graphique la mieux adaptée.

3.2 Les histogrammes

C'est la seule représentation graphique habilitée à figurer une distribution statistique et ce, quelle que soit la nature de la variable. L'histogramme met toujours en relation les effectifs d'une population (fréquences absolues ou fréquences relatives) et les valeurs prises par les individus composant ladite population pour une variable donnée. Le résultat de cette confrontation est un graphique composé de barres ou bâtonnets jointifs dont la hauteur et la surface sont proportionnels à l'effectif qu'ils représentent. C'est là la grosse différence avec les graphiques en barres tels que sait les faire Excel : les bâtonnets ne se touchent pas et si leur hauteur est bien proportionnelle à l'effectif qu'ils représentent, ce n'est pas le cas de leur aire. Excel ne sait pas faire simplement un histogramme.

L'histogramme est un graphique fondamental dans l'approche statistique des caractéristiques d'une population et de la façon dont se distribue les individus qui la composent en fonction de leurs valeurs. C'est un peu le code génétique d'une population. Toute approche et analyse statistique d'un phénomène devraient être précédées d'un tracé et d'une étude de son histogramme. Il existe deux façons de « dessiner » un histogramme : soit sous la forme discrète de bâtonnets, soit sous une forme plus « continue » sorte de courbe, appelée *polygone de fréquences*, obtenue en joignant les points milieux des sommets de chaque bâtonnet.

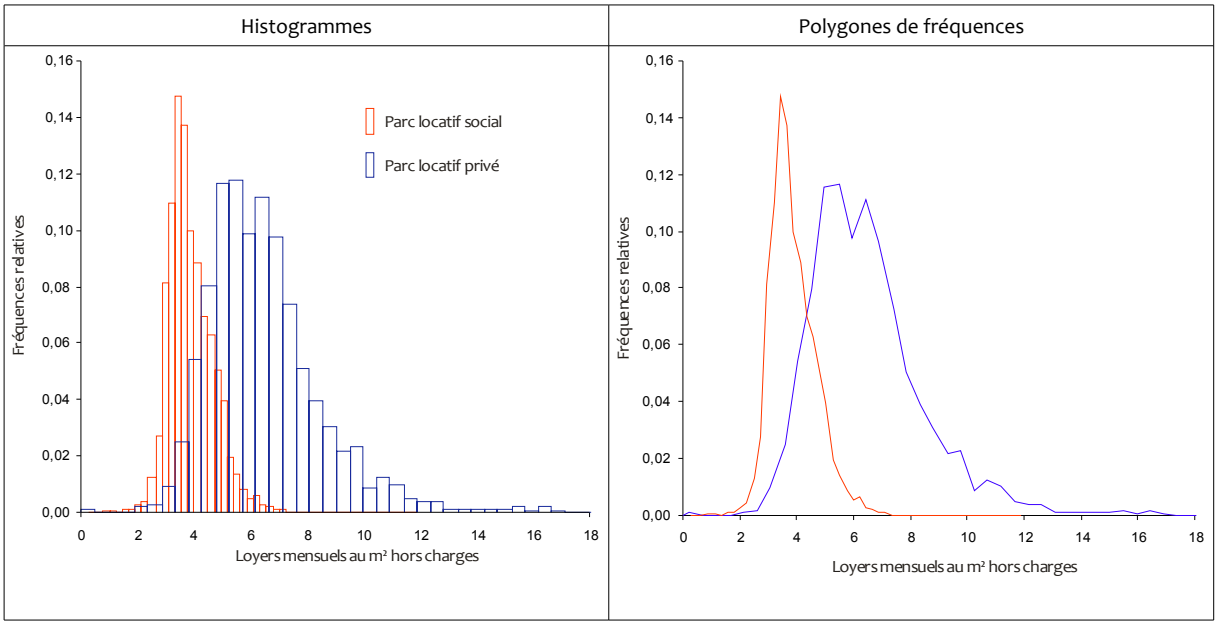
Exemple : la distribution des loyers mensuels au m² hors charges dans le parc social du département du Jura



Par convention, les fréquences sont inscrites en ordonnées, les valeurs de la variable en abscisse. Cette dernière peut être qualitative ou quantitative, discrète ou continue.

Il est possible de faire figurer plusieurs variables, donc plusieurs histogrammes ou polygones de fréquences sur un même graphique pour peu que les unités de mesure soient identiques et que les échelles de valeurs soient les mêmes ou à peu près. Il est également envisageable de faire figurer sur un même graphique plusieurs histogrammes d'une même population correspondant à son état à différentes dates.

Exemple : Comparaison des distributions des loyers mensuels au m² hors charges des parcs privé et public du département du Jura. En 2007.



Exercice 4 : fichier Excel associé « Exercice 4 - Histogramme.xls ».

3.3 Les Graphiques en barres

Même si en apparence les graphiques en barres ressemblent aux histogrammes, il a été dit précédemment en quels points ils en différaient. Les graphiques en barres permettent de comparer des effectifs ou des proportions selon les modalités retenues et ce, quelle que soit le nature des variables. Les possibilités qu'ils offrent en matière de représentation sont néanmoins beaucoup plus larges que ce que permet l'histogramme. Il est en effet relativement aisé de représenter plusieurs variables pour une même population, la même variable et ses variations dans le temps, plusieurs populations pour une même variable ainsi que plusieurs variables concernant plusieurs population. Imbriquées.

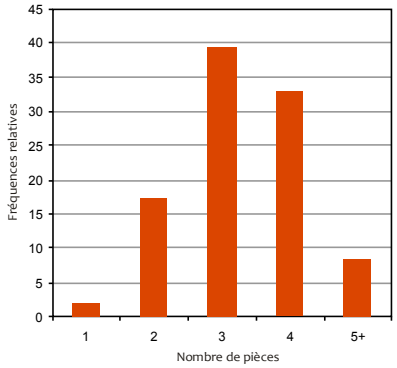
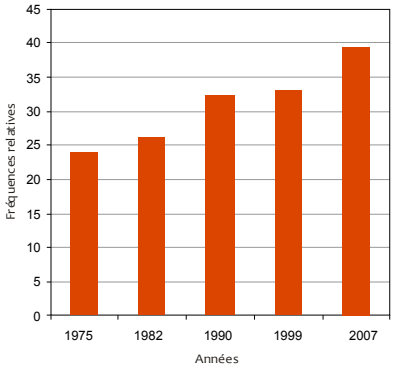
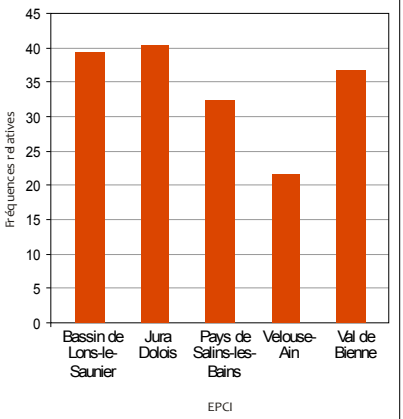
On distingue trois types de graphiques en barres :

- Les graphiques en barres simples
- Les graphiques en barres multiples
- Les graphiques en barres empilées

3.3.1 le graphique en barres simple :

Ils permettent de confronter individus, modalités ou populations à date fixe ou dans le temps

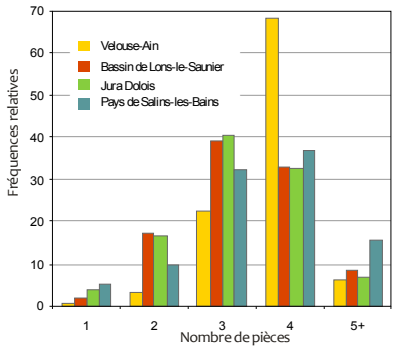
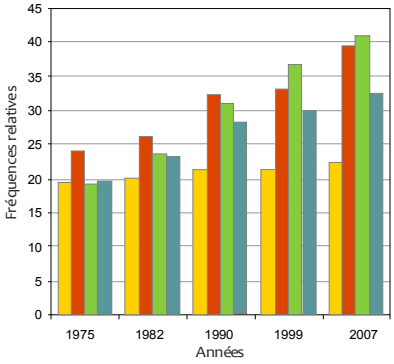
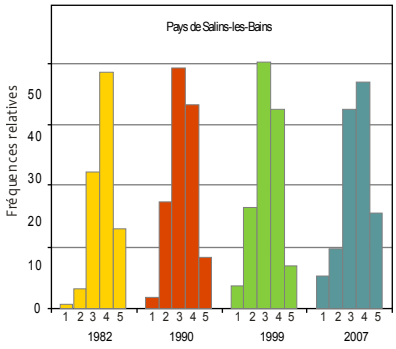
Exemple:

Plusieurs modalités une seule population	Une modalité plusieurs dates	Une modalité plusieurs individus
Structure du Parc Locatif Social de la CC Bassin de Lons-le-Saunier en 2007	Évolution de la proportion de logements de 3 pièces dans le Parc Locatif Social dans la CC Bassin de Lons-le-Saunier	Comparaison de la proportion de logements de 3 pièces dans le Parc Locatif Social en 2007 entre différents Epci
		

3.3.2 le graphique en barres multiple :

Ils permettent, sur une même zone de tracé, de confronter plusieurs individus et/ou plusieurs modalités à une ou plusieurs dates.

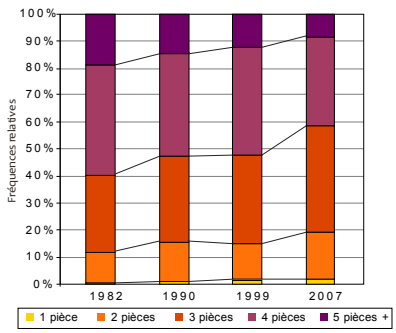
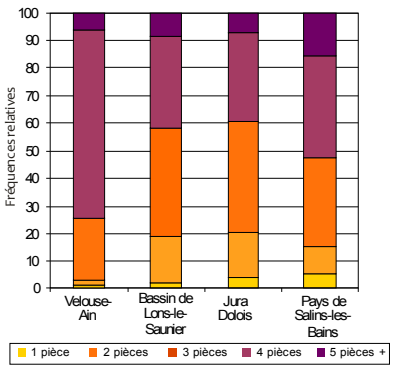
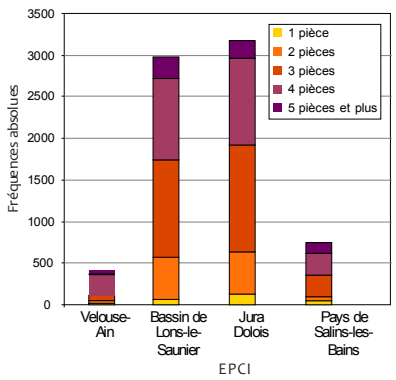
Exemple:

Plusieurs modalités plusieurs individus une seule date	Une seule modalité, plusieurs individus à plusieurs dates	Plusieurs modalités, plusieurs dates, une seule population
Structure du Parc Locatif Social dans différents EPCI du Jura en 2007	Évolution de la proportion de logements de 3 pièces dans le Parc Locatif Social de différents Epci	Évolution de la structure du Parc Locatif Social dans un EPCI du Jura
		

3.3.3 Le graphique en barres empilées :

Leur intérêt est indéniable mais ils présentent un inconvénient majeur : on a souvent quelques difficultés à apprécier précisément les proportions ou les effectifs réels → il faut souvent les noter sur le graphique occasionnant une surcharge susceptible de nuire à la lisibilité du graphique.

Exemple :

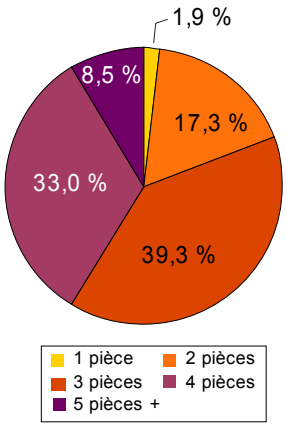
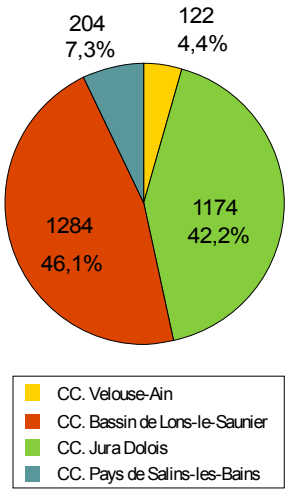
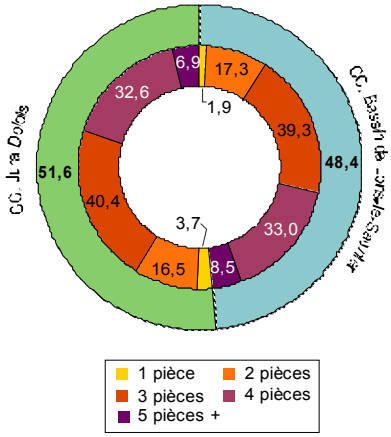
Plusieurs modalités, plusieurs dates une seule population	Plusieurs modalités, plusieurs populations, une seule date. Fréquences relatives	Plusieurs modalités, plusieurs individus, une seule date Fréquences absolues
		
Évolution de structure du parc locatif social d'un Epci	Structure du parc locatif sociale de différents Epci en 2007	Structure du parc locatif sociale de différents Epci en 2007

Tous ces graphiques peuvent se faire horizontalement et/ou avec effet 3D sans que ça ajoute à leur contenu informationnel.

Exercice 5 : fichier Excel associé « Exercice 5 - Graphiques en barres.xls » et Exercice 6 : fichier Excel associé « Exercice 6 - Graphiques en barres empilées.xls »

3.4 Les Graphiques en secteurs

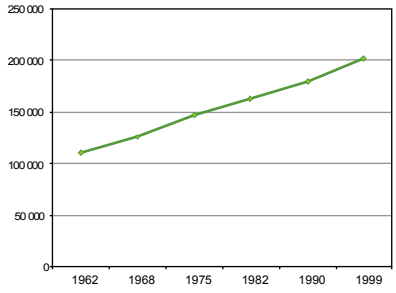
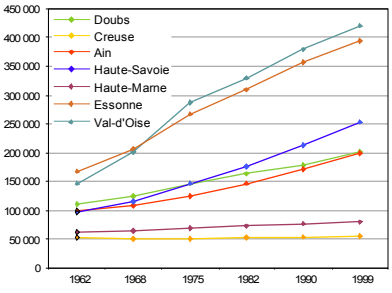
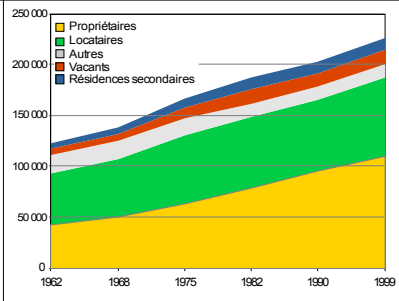
Leur rôle ou objectif est identique aux graphiques en barres avec cependant des possibilités graphiques moindres :il s'agit pour eux de figurer des effectifs en fonction d'individus et/ou de modalités. Pas de possibilité de représenter des évolutions.

Plusieurs modalités une seule population, une seule date	Plusieurs individus, une seule modalité, une seule date	Plusieurs individus, plusieurs modalités, une seule date
		
Structure du PLS de la CC de Lons en 2007	Le PLS des 3 pièces en 2007 : contribution des différents EPCI	Les Pls des principaux EPCI du département du Jura en 2007: poids et structure par taille des logements

Exercice 7 : fichier Excel associé « Exercice 7 - Graphiques en secteurs.xls ».

3.5 Les graphiques type courbes et aires

Essentiellement utilisée pour figurer des évolutions dans le temps d'un ou plusieurs phénomènes non plus seulement sous l'angle des effectifs mais aussi sous celui de la valeur même de la variable décrivant la population étudiée.

Une variable, une population ou un individu, plusieurs dates	Un variable, plusieurs populations ou individus et plusieurs dates	Plusieurs variables, une population et plusieurs dates
		
Évolution du nombre des résidences principales entre 1962 et 1999 dans le département du Doubs	Évolution du nombre des résidences principales entre 1962 et 1999 dans différents départements.	Évolution de la structure du parc de logements dans le département du Doubs entre 1962 et 1999.

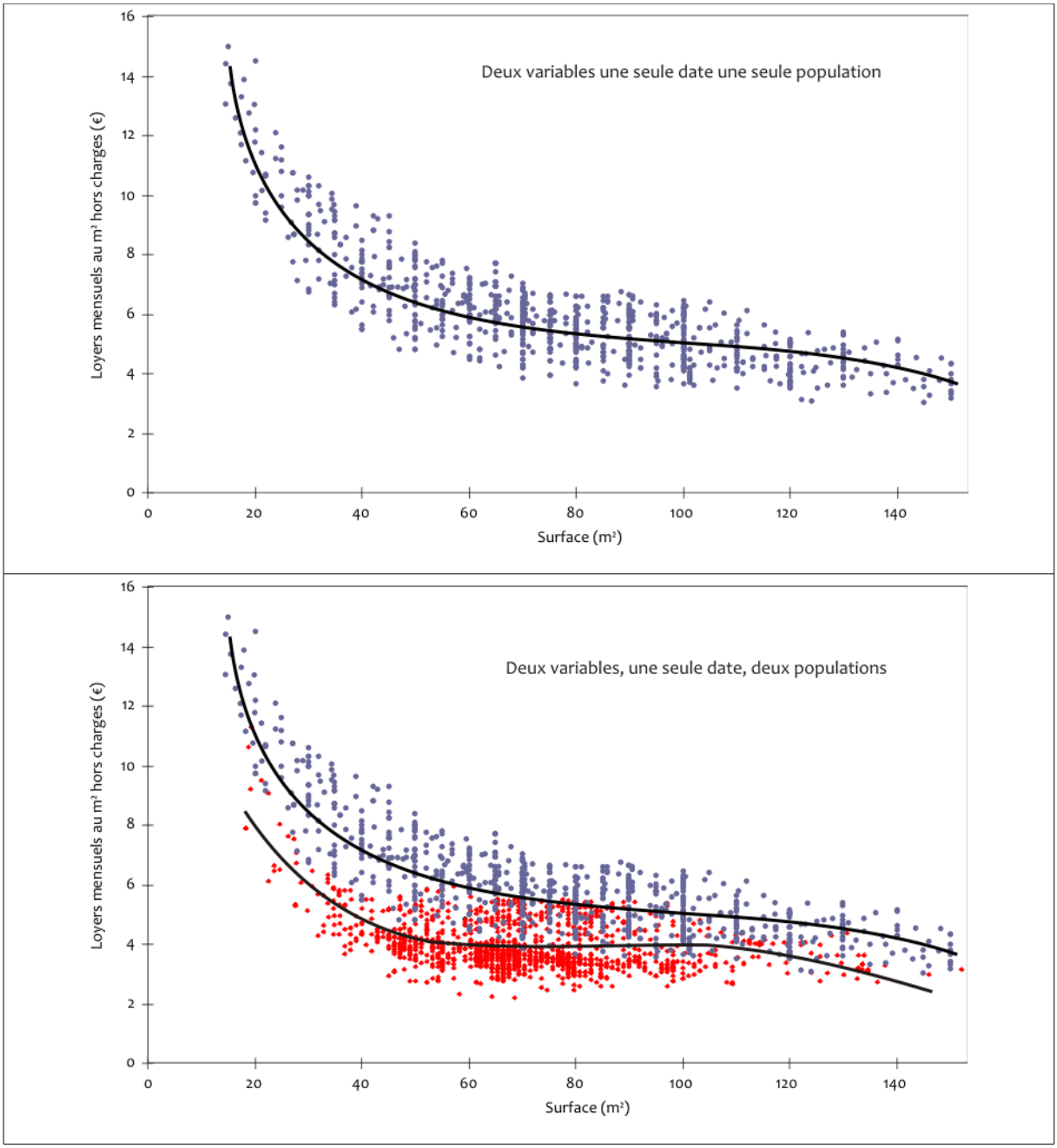
Exercice 8 : fichier Excel associé « Exercice 8 - Graphiques courbes et aires.xls ».

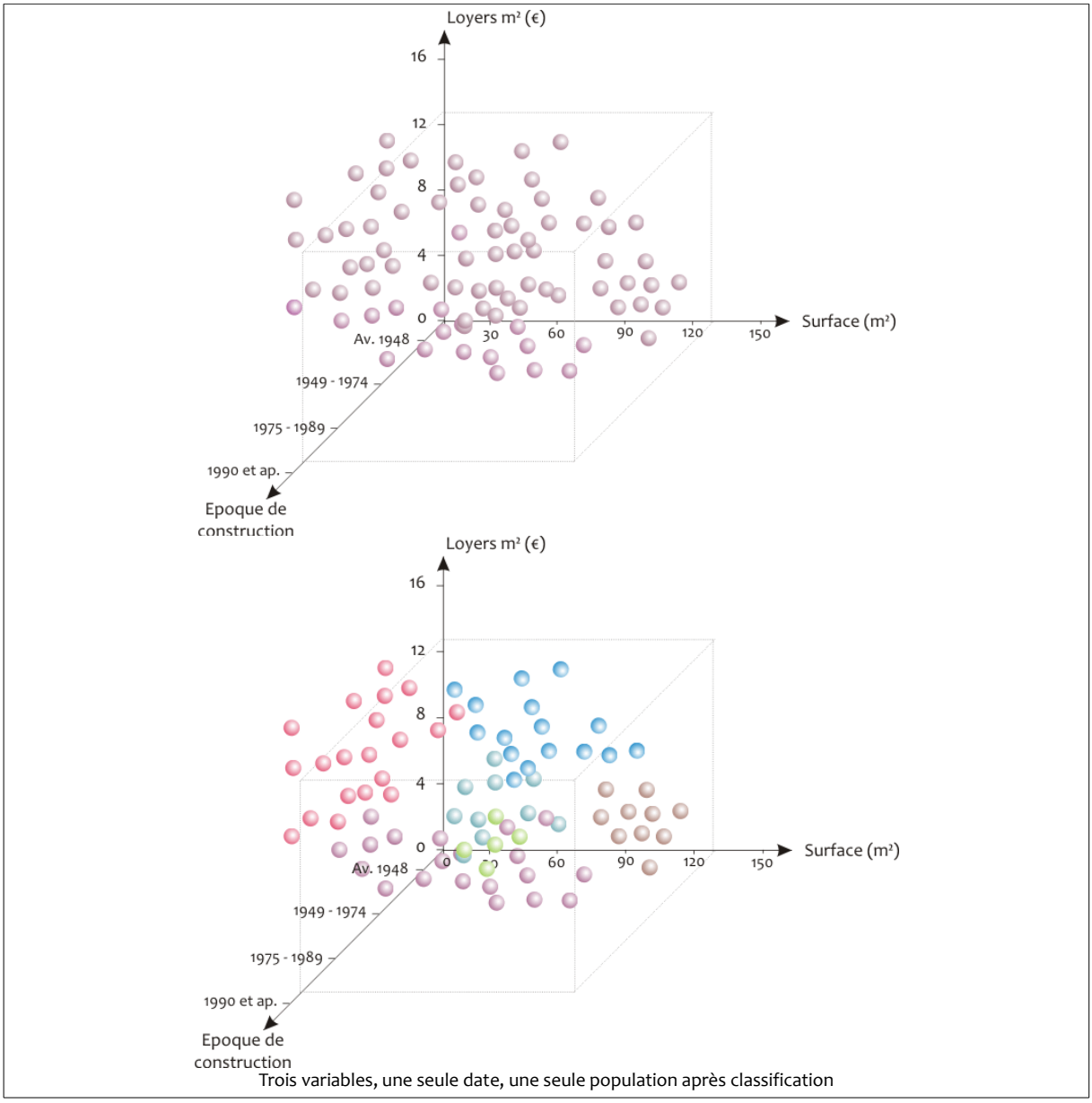
3.6 Les graphiques de dispersion ou nuages de points

Les graphiques de dispersion ou nuage de points mettent les valeurs de 2 ou 3 variables dans un repère de coordonnées cartésiennes en 2 ou 3 dimensions. On ne figure donc plus ici des effectifs mais des individus en fonctions des valeurs prises dans chacune des variables. Ce type de graphique revêt une importance fondamentale en statistique descriptive car il permet, entre autres choses, d'identifier et d'évaluer la relation entre deux variables et d'opérer une analyse sur les individus (hiérarchisation, regroupement, etc.).

Exemple:

Surface des logements locatifs et loyer mensuel au m² hors charges



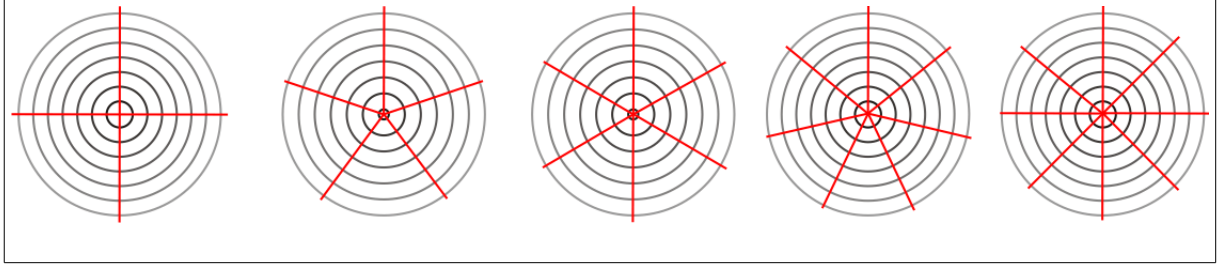


Trois variables, une seule date, une seule population après classification

Exercice 9 : fichier Excel associé « Exercice 9 - Graphiques de dispersion.xls ».

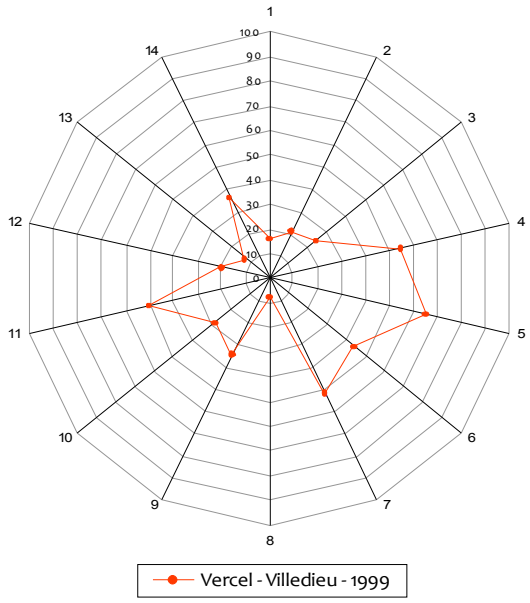
3.7 Les graphiques polaires ou radar

Très utiles et d'ailleurs très utilisés pour identifier des profils, des comportements (silhouettes) d'individus en fonction de leur comportement à l'égard de plusieurs variables (au moins 4 mais au plus 12/14 pour des questions de lisibilité). Le principe de ce type de graphique consiste à construire une figure comportant autant d'axes que de variables ou modalités étudiées dans un cercle virtuel avec une origine commune et un espacement égal à 360°/nombre de variables (d'où le nom polaire ou radar). Chaque axe possède une unité de mesure et une graduation qui lui sont propres relativement à la variable qu'il représente. Mais celles-ci doivent être identiques pour tous les individus. D'une façon générale, il est tout de même préférable d'avoir la même graduation pour l'ensemble des variables.

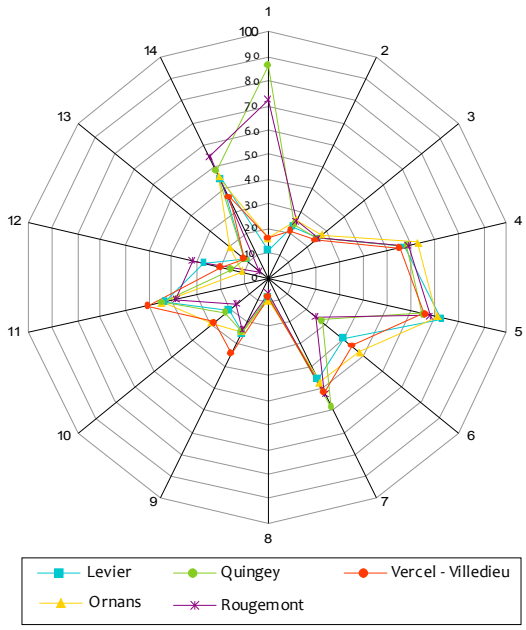


Exemple: profil socio-démographique de quelques communes

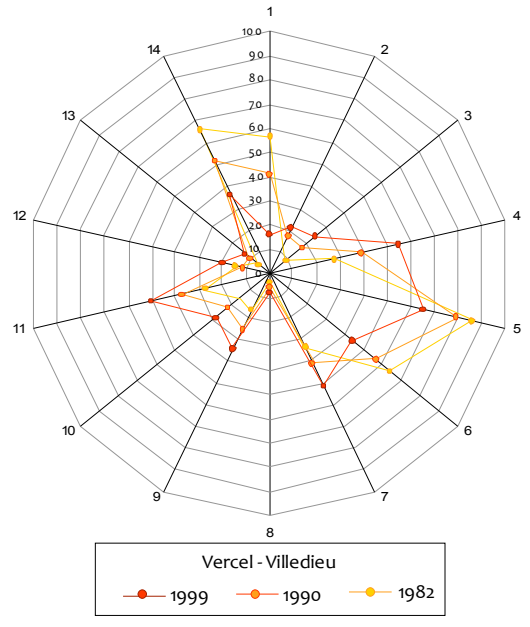
14 variables, un individu, une date



14 variables, 5 individus, une date



14 variables, un individus, plusieurs dates



Liste des variables utilisées

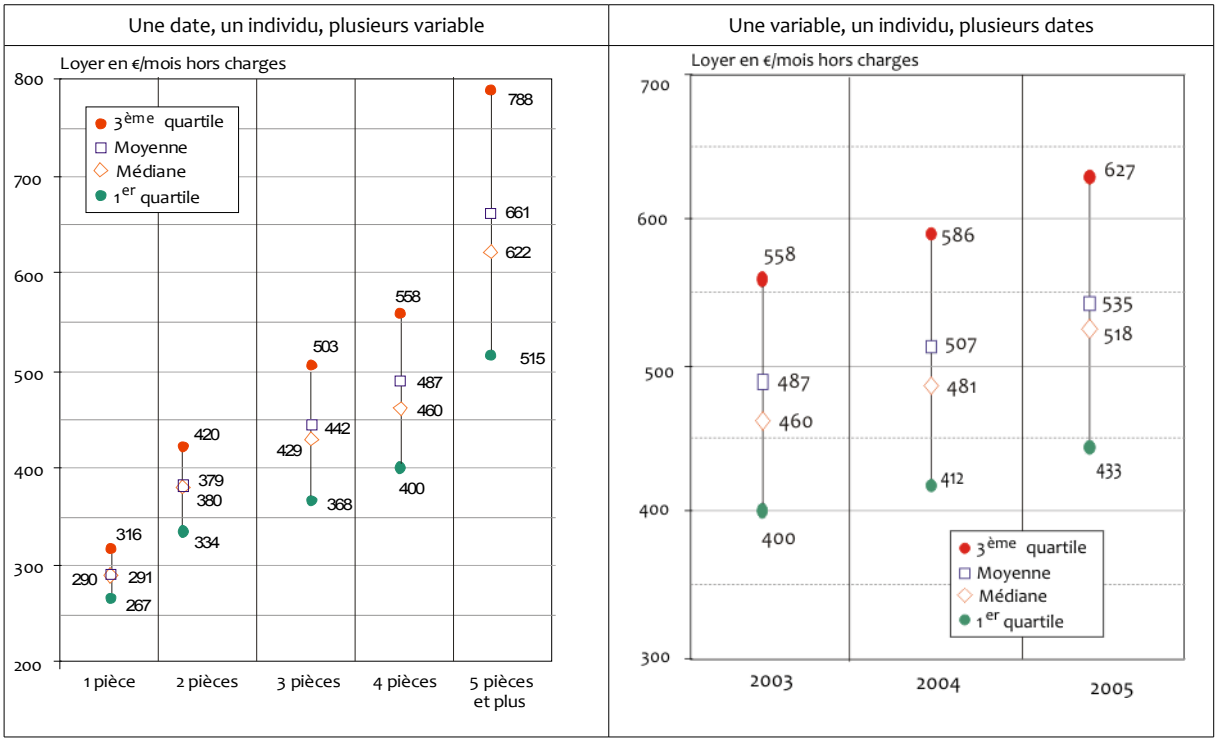
1. Part du Solde Migratoire dans le solde net de population 1999
2. Part de la population de + de 60 ans 1999
3. Part des ménages composés de 1 personne 1999
4. Part des ménages composés de 1 et 2 personnes 1999
5. Part de la population résident même commune 1999
6. Part des actifs ayant un emploi dans leur commune de résidence 1999
7. Part des 20-39 ans n'habitant pas dans commune en 1999 par rapport à 1990
8. Part des 1 et 2 pièces dans parc de résidences principales 1999
9. Part des résidences principales construites à partir de 1990 et occupées par un locataire en 1999
10. Part des résidences principales statut d'occupation locataires logement vide 1999
11. Mobilité totale dans le parc de résidence principale entre 1990 et 1999
12. Part des résidences principales construites à partir de 1990 et occupées par un locataire HLM en 1999
13. Part du collectif dans total des résidences principales en 1999
14. art des résidences principales construites avant 1945 en 1999

Exercice 10 : fichier Excel associé « Exercice 10 - Graphiques polaires.xls ».

3.8 Les graphiques boursiers ou graphiques « MinMax » ou graphique en « moustache »

Les traders et autres habitués des places boursières sont les principaux consommateurs de ce type de graphiques par ailleurs utilisables et utilisés dans bien d'autres contextes. C'est cependant de cette première utilisation qu'ils tirent leur nom car ils permettent en effet de renseigner sur l'évolution des cours boursiers au cours d'une période donnée en figurant 3 informations: le minimum et le maximum enregistrés au cours de ladite période ainsi que la valeur des cours en clôture. On peut facilement envisager une application de ce type de graphique à d'autres thématiques comme celle des loyers en considérant par exemple, par ville, par type de logements ou pour une type de logements par date, les loyers minimal et maximal mesurés ainsi que la moyenne (ou la médiane):

Exemple: les loyers dans le parc locatif privé de Besançon selon la taille des logements.

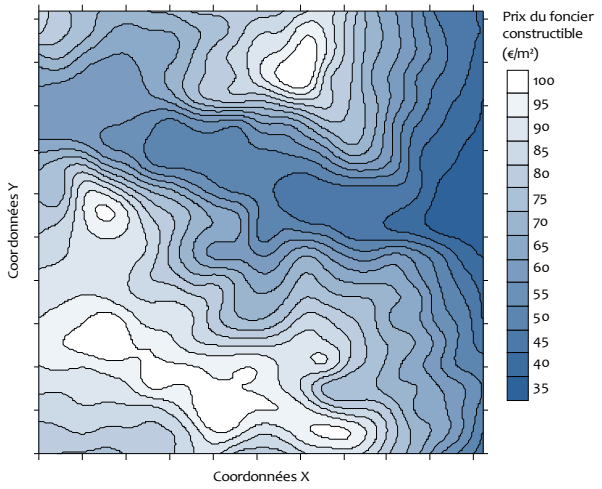
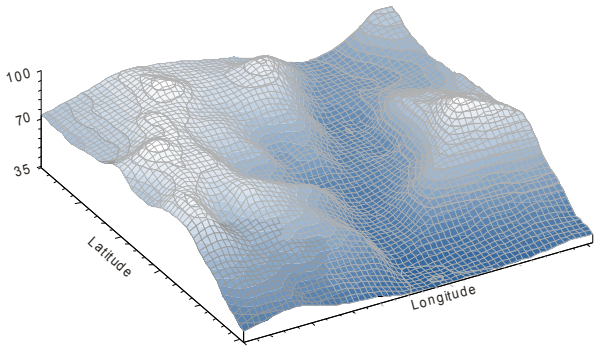


Exercice 11 : fichier Excel associé « Exercice 11 - Graphiques Min_Max.xls ».

3.9 Les graphiques spatiaux (xyz)

xyz représentent les 3 dimensions de l'espace: x et y les coordonnées géographiques, z la composante altimétrique que l'on peut fort bien remplacer par n'importe quelle variable pour peu qu'à celle-ci soit rattachée une dimension spatiale (ce qui n'est pas le cas de toutes les variables). Il en est de même des coordonnées géographiques lesquelles peuvent être substituées par d'autres variables. Le résultat est une surface

Exemple: la variabilité spatiale des prix du foncier constructible (€/m²).

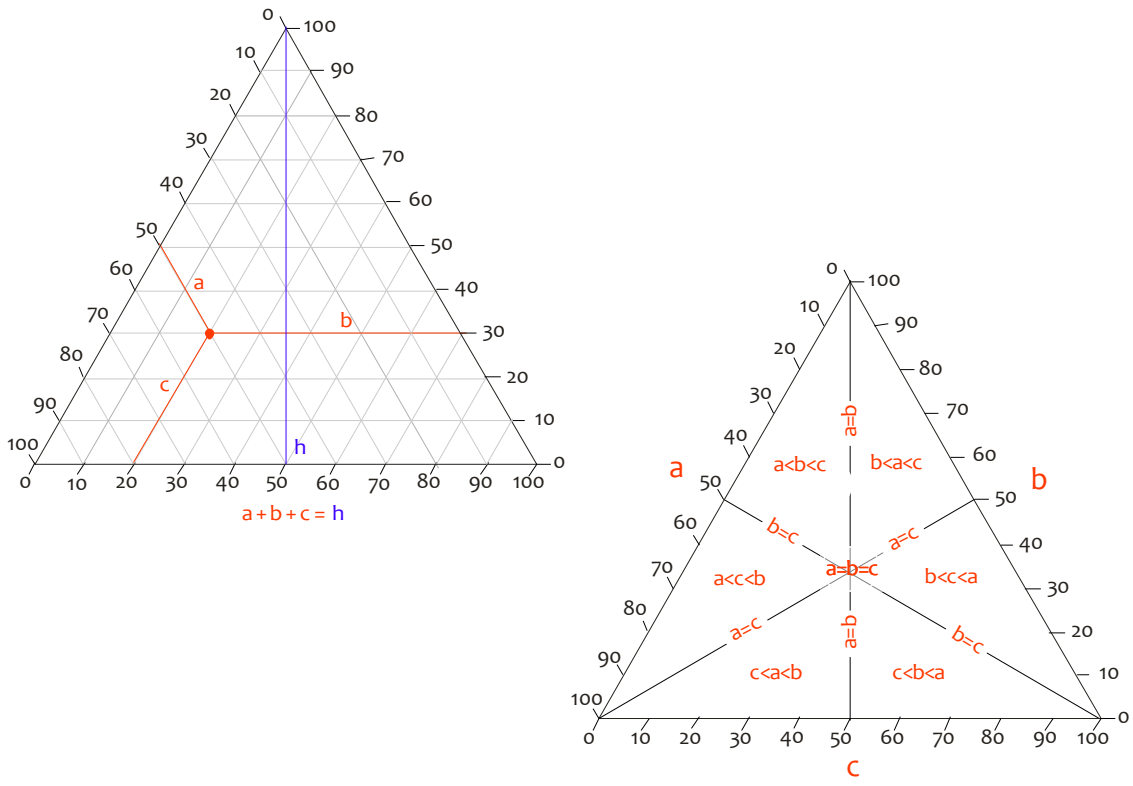
Une variable, une date, 2D	Une variable, une date, 3D
	

Précision : il n'est pas possible de réaliser ce type de graphique dans Excel

3.10 Les graphiques triangulaires ou triangle de d'Alembert

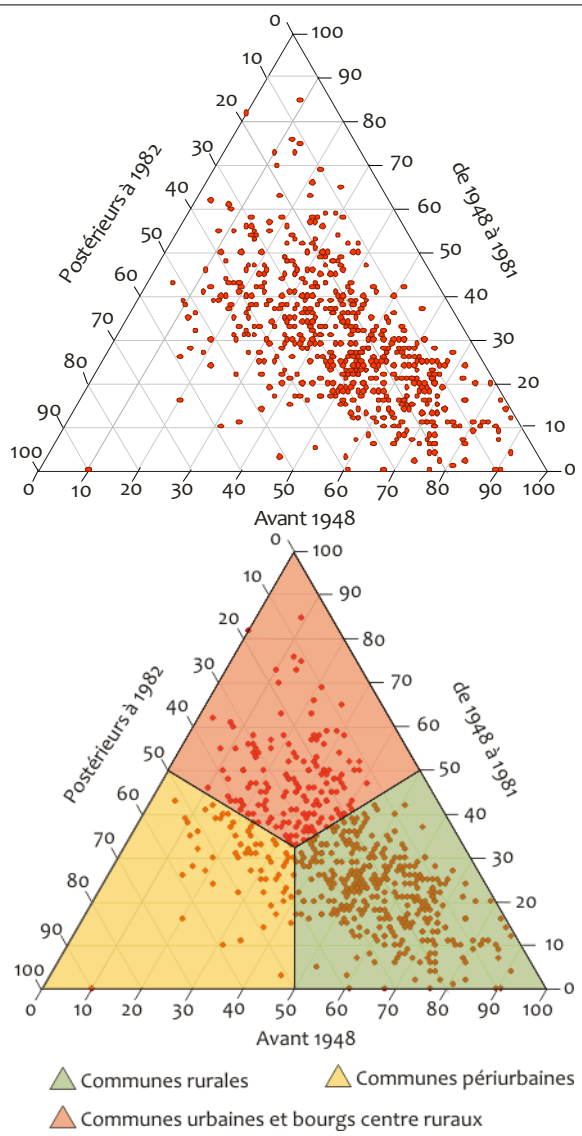
Le principe de ce type de graphique, au demeurant peu utilisé et c'est dommage car sa valeur heuristique est très forte, repose sur le fait que dans un triangle équilatéral la somme des distances d'un point s'y trouvant aux trois côtés est constante et égale à la hauteur dudit triangle. En utilisant et appliquant cette propriété, il devient possible de représenter un phénomène qui est la somme de trois grandeur représentées par des pourcentages.

Principe de construction du graphique triangulaire



Pour une même population, le graphique triangulaire permet, le cas échéant, de grouper les individus selon leur profil dans les 3 variables complémentaires retenues. Sur plusieurs dates, il permet de montrer l'évolution des profils. Il est, en ce sens, assez proche du graphique polaire ou radar.

Exemple : on s'intéresse à la structure par époque de construction du parc de logements de l'ensemble des communes d'un département français. Trois classes de périodes de construction considérées comme significativement discriminantes ont été retenues : Avant 1948, de 1948 à 1981, 1982 et après. Chaque commune est localisable à l'intérieur du graphique triangulaire au moyen de coordonnées triples correspondant aux valeurs prises dans chacune des modalités retenues. La projection de l'ensemble des individus dans le graphique triangulaire devrait permettre d'identifier des groupes composés de communes au profil semblable.



Graphiques triangulaires : utilisation en vue d'une classification des individus

Précision : il n'est pas possible de réaliser ce type de graphique dans Excel

Chapitre 4

4. Caractériser une distribution et résumer des tableaux statistiques à l'aide de paramètres appropriés : tendance centrale et mesure de dispersion

Paramètres de tendance centrale (mode, moyenne, médiane, quantiles, etc.), paramètres de dispersion (variance, écart-type, coefficient de variation, standardisation, etc.). Exercices.

C'est un des objectifs fondamentaux et LE défis de la statistique descriptive : résumer de façon simple de grandes séries statistiques tout en en conservant au mieux le contenu informationnel en limitant au maximum la perte d'informations inhérente à ce processus réducteur.

Afin d'y parvenir, la statistique a développé un certain nombre d'outils pour d'une part caractériser et résumer au mieux des distributions statistiques et pour d'autre part mettre en évidence, voire exacerber, le cas échéant, leurs différences.

Deux groupes complémentaires de paramètres permettent d'atteindre ces objectifs :

- Les paramètres de tendance centrale
- Les paramètres de dispersion

Ces deux groupes de paramètres sont complémentaires pour la description et le résumé de distributions statistiques et on ne saurait faire abstraction de l'un ou de l'autre pour ces opérations.

4.1 Les paramètres de tendance centrale

Les paramètres de tendance centrale ou « mesures de tendance centrale » sont des grandeurs susceptibles de représenter au mieux un ensemble de données. L'appellation « mesure de tendance centrale » vient du fait que ces paramètres donne une idée de ce qui se passe au centre d'une distribution, d'un ensemble de données.

On distingue trois mesures de tendance centrale :

- Le mode
- La médiane
- Le moyenne

Tous trois ne décrivent par la même chose et sont, de ce fait, complémentaires dans la description et l'analyse d'une distribution.

4.1.1 Le mode

Noté Mo, il correspond à la valeur qui apparaît le plus souvent dans une distribution, autrement la valeur qui a la fréquence (absolue ou relative) la plus élevée. S'il s'agit de données non groupées, la valeur modale est clairement identifiable. Par contre, si l'on est en présence de données groupées en classes, le mode se rapportera à la classe comportant le plus grand nombre d'individus : on parlera alors de classe modale.

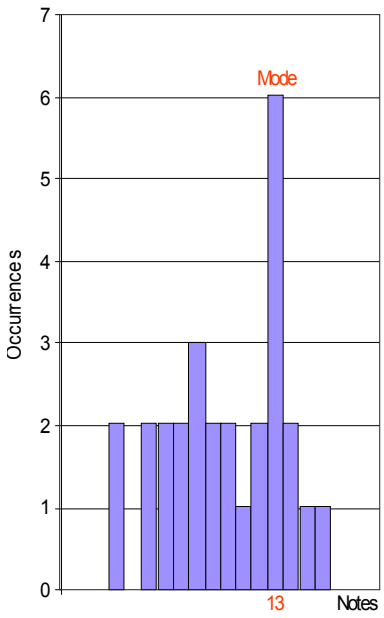
Attention ! Le mode est la seule mesure centrale qui peut être relevée et utilisée aussi bien pour des données qualitatives que quantitatives.

Exemple 1 : en relevant les notes à un examen d'une classe de 28 élèves, on obtient la série suivante :

S1 = {9;11;13;5;8;14;6;12;5;10;16;3;12;13;8;13;8;7;13;13;9;17;10;13;6;13;7;14} qui triée devient :

S1 = {3;3;5;5;6;6;7;7;8;8;8;9;9;10;10;11;12;12;13;13;13;13;13;13;14;14;16;17} à partir de laquelle on peut dresser le tableau de fréquences et l'histogramme suivants :

Note	Occurrences (fréquences absolues)	Fréquences relatives (%)
0	0	0,0
1	0	0,0
2	0	0,0
3	2	7,1
5	2	7,1
6	2	7,1
7	2	7,1
8	3	10,7
9	2	7,1
10	2	7,1
11	1	3,6
12	2	7,1
13	6	21,4
14	2	7,1
15	0	0,0
16	1	3,6
17	1	3,6
18	0	0,0
19	0	0,0
20	0	0,0



La note « 13 » apparaît 6 fois. Elle est, avec une fréquence relative de (6/28)*100 = 21,4 % la note la plus représentée de la distribution. Le mode Mo est donc ici égale à 13.

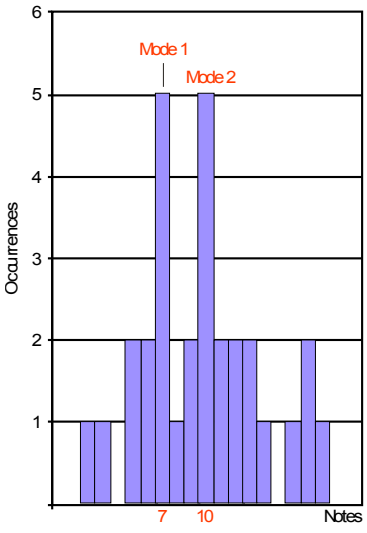
Exemple 2: Examinons les notes obtenues au même examen par la classe voisine composée de 30 élèves:

S2 = {**9;11;2;10;5;8;14;6;12;5;10;16;3;12;10;18;7;13;7;7;13;11;9;17;10;7;6;10;7;17**} qui une fois triée devient :

S2 = {**2;3;5;5;6;6;7;7;7;7;7;8;9;9;10;10;10;10;10;11;11;12;12;13;13;14;16;17;17;18**}

On obtient dès lors le tableau de fréquences et l'histogramme suivants:

Notes	Occurrences (fréquences absolues)	Fréquences relatives (%)
0	0	0,0
1	0	0,0
2	1	3,3
3	1	3,3
4	0	0,0
5	2	6,7
6	2	6,7
7	5	16,7
8	1	3,3
9	2	6,7
10	5	16,7
11	2	6,7
12	2	6,7
13	2	6,7
14	1	3,3
15	0	0,0
16	1	3,3
17	2	6,7
18	1	3,3
19	0	0,0
20	0	0,0



Dans ce cas-ci, deux modalités présentent les fréquences les plus élevées : les notes « 7 » et « 10 » avec toutes deux une fréquence relative de 16,6 % (5 occurrences chacune). La distribution comporte ici deux modes, $Mo_1 = 7$ et $Mo_2 = 10$. On parle alors de distribution bimodale.

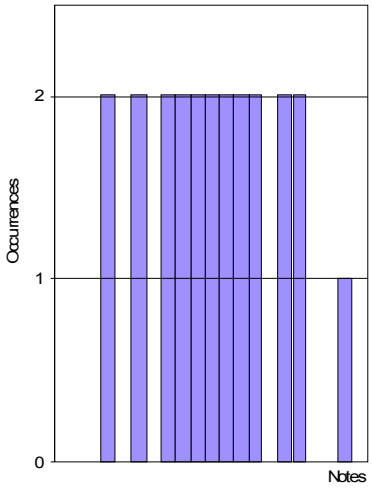
Exemple 3: Dans une troisième classe, composée de 24 élèves, les notes obtenues au même examen sont les suivantes:

S3 = {**3;12;16;5;3;7;10;7;16;5;11;13;11;9;13;9;10;12;8;15;15;8;19**} qui une fois triée devient :

S3 = {**3;3;5;5;7;7;8;8;9;9;10;10;11;11;12;12;13;13;15;15;16;16;19**}

On obtient par le fait le tableau de fréquences et l'histogramme suivants:

Notes	Occurrences	Fréquences relatives (%)
0	0	0,0
1	0	0,0
2	0	0,0
3	2	8,7
4	0	0,0
5	2	8,7
6	0	0,0
7	2	8,7
8	2	8,7
9	2	8,7
10	2	8,7
11	2	8,7
12	2	8,7
13	2	8,7
14	0	0,0
15	2	8,7
16	2	8,7
17	0	0,0
18	0	0,0
19	2	4,3
20	0	0,0



Plutôt que de parler de distribution multimodale (à plusieurs modes) on parlera davantage ici de distribution a-modale (sans réel mode). Dans cet exemple, le mode est une mesure non-significative. C'est souvent le cas lorsque l'on est en présence d'une distribution contenant peu de résultats.

Le mode n'est évidemment pas suffisant pour caractériser et résumer une distribution. Il l'est encore moins pour comparer et différencier des distributions. Deux distribution peuvent en effet avoir le même mode avec cependant des allures, et donc des caractéristiques, totalement différentes. On a donc inventer d 'autres paramètres, d'autres mesures susceptibles de mieux caractériser et/ou différencier des distribution. C'est le cas de la médiane.

Exercice 12 : fichier Excel associé « Exercice 12 - Mode.xls ».

4.1.2 La médiane

Étymologiquement « médiane » signifie milieu, et c'est bien de ça dont il s'agit car la médiane est réellement le milieu d'une distribution. Noté *Me*, la médiane correspond à la valeur de la distribution qui partage l'effectif total en deux sous-effectifs de même taille de telle sorte que l'on puisse dire que 50 % des individus d'une population

sont caractérisés par une valeur supérieure à celle de la médiane et que 50 % des individus de cette même population ont une valeur inférieure à la médiane.

> Exemple: la médiane des revenus pour une population donnée correspond à la valeur du revenu pour laquelle on a 50 % de ladite population dont le revenu est supérieur à cette valeur et 50 % dont le revenu est inférieur. On parle alors de revenu médian.
>
> Le revenu médian par ménage dans le département des Yvelines était, en 2002, de 34 506 € contre 17 640 pour le département de la Creuse.

Attention ! Contrairement au mode, la médiane est une mesure centrale qui ne peut être calculée et utilisée que pour des variables quantitatives, continues ou discrètes.

Comment calculer la médiane ?

Si le mode, pour être révélé, ne nécessite aucun calcul mais simplement de l 'observation, la médiane impose quant à elle, un certain nombre de manipulations voire de calcul pour sa mesure.

Reprenons pour ce faire l'exemple relatif aux notes relevées lors d'un même examen dans différents classes en ne retenant que deux séries :

Classe 1 28 élèves / notes	Classe 3 23 élèves / notes
9	3
11	12
13	16
5	5
8	3
14	7
6	10
12	7
5	19
10	16
16	5
3	11
12	13
13	11
8	9
13	13
8	9
7	10
13	12
13	8
9	15
17	15
10	8
13	
6	
13	
7	
14	

Quelle est, pour chacune des classes, la note médiane ?

Pour le calcul de la note médiane il faut:

1. Classer les valeurs de la série par ordre croissant. Cette opération a pour but d'affecter un rang à chaque valeur et ainsi de déterminer plus facilement le milieu de la série donc la médiane.

Rang	Classe 1 28 élèves / notes
1	3
2	5
3	5
4	6
5	6
6	7
7	7
8	8
9	8
10	8
11	9
12	9
13	10
14	10
15	11
16	12
17	12
18	13
19	13
20	13
21	13
22	13
23	13
24	13
25	14
26	14
27	16
28	17

Rang	Classe 3 23 élèves / notes
1	3
2	3
3	5
4	5
5	7
6	7
7	8
8	8
9	9
10	9
11	10
12	10
13	11
14	11
15	12
16	12
17	13
18	13
19	15
20	15
21	16
22	16
23	19

2. Déterminer si la série comporte un nombre n pair ou impair de valeurs. Deux cas peuvent alors se présenter:

- Si n est pair,il n'y a pas possibilité d'identifier simplement la valeur qui partage la population en deux effectifs égaux. Deux valeurs se situent au centre de la série et jouent ce rôle respectivement de rang (n/2) et [(n/2)+1]. La médiane est alors égale à la moyenne des valeurs encadrant le milieu de la série. C'est le cas dans la série de notes de la classe 1 composée de 28 valeurs. La médiane se situe entre le 14^{e} et le 15^{e} rang et sa valeur est donc comprise entre 10 et 11. L'application de la règle sus mentionnée nous donne donc une médiane ***Me*** de (10 + 11)/2 = 10,5
- Si n est impair alors il est possible d'identifier simplement la valeur qui partage la population en deux effectifs égaux. Le rang central étant égal à [(n+1)/2]. C'est le cas dans la série de notes de la classe 3 composée de 23 valeurs. La médiane se situe au niveau du 12^{e} rang et sa valeur est lue directement en face de ce 12^{e} rang, dans notre ***Me*** = 10

Rang	Classe 1 28 élèves /
1	3
2	5
3	5
4	6
5	6
6	7
7	7
8	8
9	8
10	8
11	9
12	9
13	10
14	10
15	11
16	12
17	12
18	13
19	13
20	13
21	13
22	13
23	13
24	13
25	14
26	14
27	16
28	17

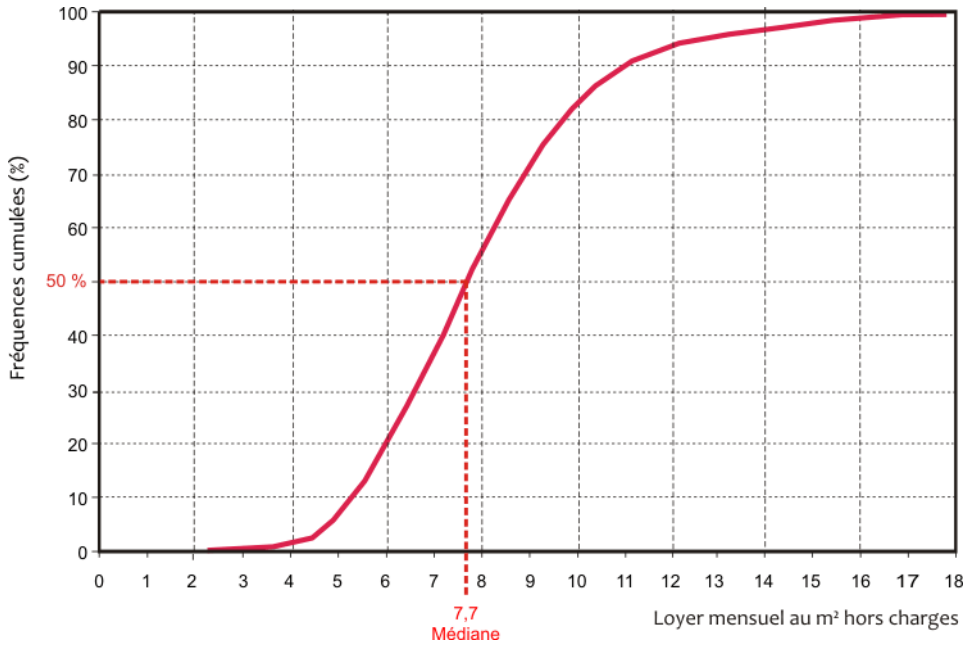
Valeurs encadrant le milieu

Milieu de la série
Me = (10+11)/2
=10,5

Rang	Classe 3 23 élèves / notes
1	3
2	3
3	5
4	5
5	7
6	7
7	8
8	8
9	9
10	9
11	10
12	10
13	11
14	11
15	12
16	12
17	13
18	13
19	15
20	15
21	16
22	16
23	19

Milieu de la série
Me = 10

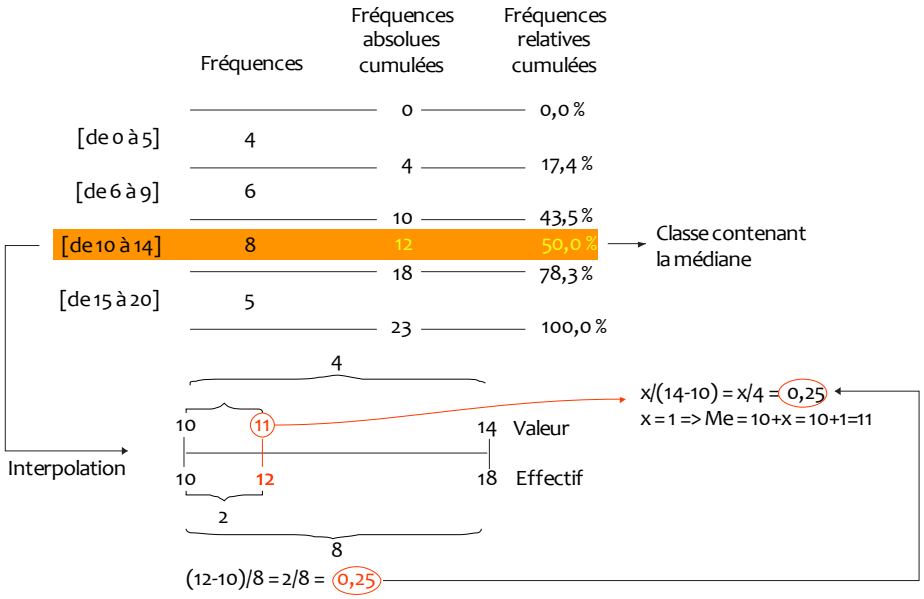
La médiane peut également être repérée graphiquement sur le courbe des fréquences cumulées comme suit :



Même si Excel ou d'autres applications disposent de fonctions capables de calculer automatiquement la médiane, il est bon de savoir comment ce calcul se fait.

La médiane de données groupées est également calculable ou plutôt estimable par interpolation. La médiane est trouver et à estimer dans le classe où se situe le rang divisant en deux parties égales la population.

Exemple: en regroupant les valeurs de la série de notes de la classe 3 en 4 groupes on obtient l'organisation suivante:



Pour chaque classe (ou groupe) on connaît la fréquence absolue ou relative que l'on cumule pour repérer plus facilement la classe ou le groupe devant contenir la médiane. Dans notre exemple, la classe contenant la note médiane est la classe [de 10 à 14] car c'est celle qui contient la fréquence cumulée 50 %. Connaissant n= 23 impair on sait que la médiane correspond au rang 12 qui se situe bien dans la classe [de 10 à 14]. Le rapport des différences effectif médian (12) – borne inférieure de la classe médiane (10) à borne supérieure de la classe médiane (18) – borne inférieure de la classe médiane (10) nous donne le rapport à appliquer aux valeurs pour trouver la note médiane :

(12-10)/(18-10) = 2/8 = 0,25 pour les effectifs. Pour la valeur médiane, on connaît l'amplitude de la classe médiane (14-10 = 4). Il nous reste donc à trouver la différence entre la médiane (Vm) et la borne inférieure de la classe de valeurs médiane (10). Cette différence est appelée x. A l'aide du rapport (0,25) calculé précédemment, on peut écrire:

$$\frac{(V_m - 10)}{(14-10)} = 0{,}25 \Rightarrow \frac{x}{4} = 0{,}25 \Rightarrow x = 1$$

La médiane *Me* est donc égale à la borne inférieure de la classe médiane + x soit 10 + 1 = 11

Exercice 13 : fichier Excel associé « Exercice 13 - Médiane.xls ».

4.1.3 La moyenne

La moyenne constitue un autre paramètre de tendance centrale fondamental mais non suffisant pour caractériser une distribution. Complémentaire du mode et surtout de la médiane, la moyenne constitue à n'en point douter, la mesure la plus calculée et la plus utilisée lors de la description de séries statistiques.

Il existe plusieurs types de moyennes, chacun adapté à des situations précises :

Dénomination	Notation courante
Moyenne arithmétique	$\bar{x}$
Moyenne géométrique	$\bar{G}\ ou\ \bar{x_G}$
Moyenne harmonique	$\bar{H}\ ou\ \bar{x_H}$
Moyenne quadratique	$\bar{Q}\ ou\ \bar{x_Q}$
Moyenne glissante	

La moyenne arithmétique :

C'est la plus simple et la communément utilisée et ce, pas toujours à bon escient. Elle se note $\bar{x}$ la plupart du temps. Elle peut être simple ou pondérée. Attention ! On ne peut pas calculer de moyenne arithmétique sur des données qualitatives.

La moyenne arithmétique simple

Sa version simple correspond à une somme de résultats divisée par le nombre de résultats et s'écrit :

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{\sum_{i=1}^{n}x_i}{n}=\frac{(x_1+x_2+x_3+...+x_i+...+x_n)}{n}$$

Avec : $n=nombre\ de\ résultats\ (ou\ nombre\ d'individus\ ou\ effectif\ total)$

$x_i=valeur\ pour\ i=1\ jusqu'à\ n$

Exemple : le loyer moyen dans le parc locatif privé de Besançon au 01/01/2008.

A la suite d'une enquête, on dispose de exactement 1 011 références de loyers représentatives ensemble de la structure du parc. La moyenne arithmétique simple des loyers mensuels au m² hors charges s'écrit donc :

$$\bar{L}=\frac{1}{n}\sum_{i=1}^{n}L_i=\frac{1}{1011}\sum_{i=1}^{1011}L_i=\frac{(L_1+L_2+L_3+...+L_i+...+L_{1011})}{1011}=\frac{7913{,}99}{1011}=7{,}83\,€/m^2$$

Le calcul nous donne un loyer mensuel moyen au m² hors charges de 7,83 €. Cependant, la moyenne simple, dans son principe de calcul, ne permet de tenir compte de la structure de la population étudiée et du poids éventuellement différent que peuvent avoir chacun des individus ou classes d'individus la composant.

La moyenne arithmétique pondérée

La moyenne arithmétique pondérée, autant le dire tout de suite, donne, dans son utilisation classique (c'est-à-dire lorsque tous les individus ont le même poids), le même résultat que la moyenne arithmétique simple. Sa formule est cependant différente puisqu'elle introduit la notion de poids via un terme supplémentaire qui peut

s'avérer utile dans certaines situations, notamment lorsque justement les individus composant une population n'ont pas le même poids ou coefficient : certains individus, pour diverses raisons, ont davantage d'influence dans ladite population que les autres. Ce peut être le cas par exemple lorsque l'on a affaire à une série de notes dont le coefficient n'est pas le même.

En considérant un ensemble de données

$$X = \{x_1, x_2, x_3, .., x_i, ..., x_n\}$$

et une ensemble de poids non négatifs correspondants :

$$W = \{w_1, w_2, w_3, .., w_i, ..., w_n\}$$

Dans le cas général le *poids* w_i représente l'influence de l'élément x_i par rapport aux autres. La formule de la moyenne pondérée s'écrit alors :

$$\bar{x_p} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3 + ... + w_i x_i + ... + w_n x_n}{w_1 + w_2 + w_3 + ... + w_1 + ... + w_n}$$

Exemple :

Reprenons l'exemple précédent pour lequel le calcul de la moyenne arithmétique simple sur l'ensemble des loyers attribuait par défaut un poids identique à chaque logement.

Or on sait que les loyers surfaciques sont inversement proportionnels à la taille des logements (nombre de pièces) et que les petits logements (1 et 2 pièces) constituent en général une part importante, voire la majoritaire du parc locatif privé. Dans ces conditions, la non pris en compte de la structure du parc et l'attribution de poids identiques à chaque logement se traduisent systématiquement par un sous estimation du loyer moyen.

L'attribution de poids différents à chaque logement en fonction de son nombre de pièces contribuera à rétablir la contribution vraie et réelle de chaque logement dans le calcul de la moyenne. Ainsi, dans notre échantillon de 1011 logements, on observe la structure suivante :

Catégories	Poids W_i par catégorie
1 pc	0,216
2 pc	0,244
3 pc	0,267
4 pc	0,197
5 pc+	0,076

Les poids par catégorie correspondent à la part de chaque catégorie dans le parc locatif total : à titre d'exemple, les logements de 3 pièces représentent, dans l'échantillon, 0,267 x 100 = 26,7 % du total des logements.

On attribue alors à chaque logement un coefficient pondérateur fonction de sa catégorie d'appartenance. Ainsi, à chaque logement de 1 pièce, on attribue le coefficient (ou poids) 0,216, à chaque logements composé de 2 pièces, le poids 0,244 et ainsi de suite.

Il est dès lors possible de calculer la moyenne pondérée pour l'ensemble de la distribution. On a :

$$\bar{x}_p = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{1\,769{,}34}{224{,}42} = 7{,}88\,€/m^2$$

Le résultat est au final assez peu différent de celui obtenu par la moyenne arithmétique simple car la moyenne arithmétique simple, vu le nombre important de valeurs et compte tenu de la structure de l'échantillon, tient compte, de façon presque naturelle, du poids des individus en attribuant implicitement à chaque individus le poids de sa catégorie.

On peut également utiliser la moyenne pondérée pour « corriger » et rectifier les biais et erreurs accumulés au cours de l'enquête et qui font que l'échantillon n'est au final plus tout à fait représentatif de la population mère, dans notre exemple du point de la structure du parc. En effet, lors de la fabrication de l'échantillon ou à l'issue du sondage, il se peut que certaines modalités soient sur- ou sous-représentées pour différentes raisons, au quel cas l'utilisation d'une moyenne arithmétique pondérée avec les poids tels que relevés dans l'échantillon biaisera inévitablement le résultat final. La connaissance de la structure de la population mère, rend alors possible l'introduction de nouveaux poids issus de la population mère qui, appliqués aux données collectées, viendront corriger le biais résident de l'échantillon en permettant le calcul d'une moyenne « moins fausse ».

	Échantillon	Population mère	
Catégories	Poids initiaux W_i par catégorie	Poids corrigés W'_i par catégorie	Statut
1 pc	0,216	0,256	sous -représentée
2 pc	0,244	0,272	sous -représentée
3 pc	0,267	0,227	sur -représentée
4 pc	0,197	0,172	sur -représentée
5 pc+	0,076	0,073	sur -représentée

Dans notre exemple, on observe que la structure de l'échantillon diffère sensiblement de la structure de la population mère. Certaines catégorie sont sur-représentées, comme par exemple les logements de 3 et 4 pièces, alors que d'autres sont sous-représentées, comme celles des logements de petite taille (1 et 2 pièces). Le calcul d'une moyenne à partir des données et poids du seul échantillon introduira un biais lié aux sur-représentations et aux sous-représentations évoquées en « tirant » la moyenne vers le bas, la catégorie des logements de taille moyenne (3 et 4 pièces) proposant en général des loyers surfaciques moins élevés que la catégorie sous-représentée des petits logements (1 et 2 pièces). En affectant aux individus de l'échantillon les poids relevés dans la population mère, on corrige en quelque sorte le biais de l'échantillon en donnant davantage de poids aux petits logements et en minorant celui des logements sur-représentés (3 et 4 pièces).

Appliqués à notre exemple ces nouveaux poids aboutissent au résultat suivant :

$$\bar{x}_p = \frac{\sum_{i=1}^{n} w'_i x_i}{\sum_{i=1}^{n} w'_i} = \frac{1\,812{,}54}{224{,}06} = 8{,}09\,€/m^2$$

Où w'_i représente les poids corrigés.

Nous obtenons ici un loyer moyen sensiblement différent de ceux calculés précédemment. Le rétablissement des contributions respectives vrais des différents catégories de logements et l'attribution de poids corrigés plus importants aux petits logements ont permis de faire disparaître le sous estimation inhérente à l'échantillon.

La moyenne arithmétique de données groupées

Autant que faire se peut, ce type de calcul est à éviter car source d'imprécision et d'erreur trop importantes. Cependant, on peut être confronter à une situation où seules des données groupées sont disponibles. Dans ce cas, et seulement dans celui-là, on peut être autorisé à calculer une moyenne à partir de classes. On agit alors comme si tous les résultats d'une classe se trouvaient au centre de celle-ci. La moyenne de la distribution est alors calculée à partir des valeurs centrales des classes pondérées par leurs effectifs respectifs.
Exemple :

Classe	Borne inférieure	Borne supérieure	Centre de classe	Fréquence absolue	fX
1	2,50	5,00	3,75	67	67 x 3,75 = 251,25
2	5,01	7,50	6,25	461	461 x 6,25 = 2 881,25
3	7,51	10,00	8,75	326	326 x 8,75 = 2 852,68
4	10,01	12,50	11,25	116	116 x 11,25 = 1 305,06
5	12,51	26,50	19,50	41	41 x 19,50 = 799,50
				1011	8090

$$\bar{x} = \frac{8\,090}{1\,011} = 8{,}00\,€/m^2$$

Exercice 14 : fichier Excel associé « Exercice 14 - Moyenne arithmétique.xls ».

La moyenne géométrique :

Sa définition purement mathématique est un peu rébarbative mais son utilité est grande comme nous allons le démontrer.

La moyenne géométrique de n valeurs positives x_i est la racine $n^{ième}$ du produit de ces valeurs. Notée $\bar{G}\ ou\ \bar{x}_G$, elle s'écrit :

$$\bar{G} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_i \cdot \ldots \cdot x_n}$$

La moyenne géométrique est un instrument permettant de calculer des taux moyens, notamment des taux moyens annuels. Son utilisation n'a un sens que si les valeurs ont un caractère multiplicatif.

Exemple : Les prix de l'immobilier ancien ont augmenté ces 10 dernières années de la façon suivante :

Année	Variation annuelle (%)
1	9,2
2	12,7
3	8,8
4	7,7
5	3,9
6	1,7
7	0,9
8	2,2
9	4,7
10	3,3

En utilisant la moyenne arithmétique simple, on obtiendrait une évolution moyenne de (13,8 + 12,7 + 12,8 + 11,2 + 7,7) / 5 = 55,1 / 10 = 5,51 % mais ce résultat est faux compte tenu de la relation entretenue par les taux d'une année sur l'autre.

L'utilisation de la moyenne géométrique permet de solutionner ce problème :

$$\bar{G} = \sqrt[5]{9,2 \cdot 12,7 \cdot 8,8 \cdot 7,7 \cdot 3,9 \cdot 1,7 \cdot 0,9 \cdot 2,2 \cdot 2,2 \cdot 4,7 \cdot 3,3}$$

$$\bar{G} = \sqrt[5]{1611964,46} = 1611964,46^{(\frac{1}{5})} = 4,18$$

Soit une hausse moyenne annuelle de 4,18 % contre 5,51 % avec la moyenne arithmétique.

Exercice 15 : fichier Excel associé « Exercice 15 - Moyenne géométrique.xls ».

La moyenne harmonique :

On utilise la moyenne harmonique lorsqu'on veut déterminer un rapport moyen dans des domaines ou ils existent des liens de proportionnalité inverse.

Exemples:

- Pour une distance donnée, le temps de trajet est d'autant plus court que la vitesse est élevée.
- Un loyer dans le parc privé est d'autant plus élevé que la taille ou la surface du logement est petite.

La moyenne harmonique de N valeurs est le nombre dont l'inverse est la moyenne arithmétique des inverses desdites valeurs. C'est un peu compliqué comme définition ! Voilà ce que ça donne sous une forme mathématique :

$$\bar{H} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_i} + \ldots + \frac{1}{x_n}}$$

La moyenne harmonique permet de calculer des moyennes sur des fractions si le dénominateur change. C'est le cas du calcul de la vitesse moyenne parcourue dans un trajet aller/retour, la vitesse étant la valeur représentée par distance / temps.

Exemple :

Dans un parc locatif privé, 3 logements ont respectivement un loyer surfacique de:

L1 = loyers surfacique Logement A : 7,49 €/m² pour 67 m²

L2 = loyers surfacique Logement B : 11,43 €/m² pour 28 m²

L3 = loyers surfacique Logement C : 6,18 €/m² pour 97 m²

La moyenne arithmétique des loyers donne:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{3}\sum_{i=1}^{3} L_i = \frac{(7{,}49 + 11{,}43 + 6{,}18)}{3} = \frac{25{,}1}{3} = 8{,}37\,€/m^2$$

La relation d'inverse proportionnalité qui existe entre surface des logements et loyer surfacique nous incite à utiliser la moyenne harmonique pour le calcul du loyer moyen. Pour cela il faut tenir compte du fait que la logement C est 3 fois plus grand que le logement B

La moyenne quadratique :

Une moyenne qui trouve des applications lorsque l'on a affaire à des phénomène présentant un caractère sinusoïdal avec alternance de valeurs positives et de valeurs négatives. Elle est, de ce fait, très utilisée en électricité. Elle permet notamment de calculer la grandeur d'un ensemble de nombre. A titre d'information, elle s'écrit :

$$\bar{Q} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Prenons un rapide exemple : considérons les nombre suivants {-2, 5, -8, 9, -4 }

Nous pouvons en calculer la moyenne arithmétique avec l'inconvénient de voir se neutraliser les valeurs positives et négatives et d'aboutir à un résultat nul sans que cela ne nous apprenne quoi que ce soit. En effet,

$\bar{x} = 0$

Le calcul de la moyenne quadratique pour la même série donne 6,16

La moyenne glissante ou moyenne mobile

La moyenne glissante, ou moyenne mobile trouve son application dans l'analyse des séries temporelles de données en permettant la suppression des fluctuations de façon à en souligner les tendances sur le long terme. Cette moyenne est dite *mobile* parce qu'elle est recalculée de façon perpétuelle, dès lors qu'une nouvelle donnée intègre la série en venant remplacer la plus ancienne, modifiant ainsi la date de référence. Cette façon de faire tend à lisser le phénomène étudié en noyant les valeurs extrêmes dans une masse de données davantage représentative d'une tendance moyenne.

Exemple : on dispose de données mensuelles concernant l'évolution des prix à la consommation (inflation) et on souhaite connaître pour chaque mois l'évolution mensuelle moyenne des prix sur un trimestre.

	Janv 08	Fev 08	Mars 08	Avr 08	Mai 0 8	Juin 08	Juil 08	Aout 08	Sept 08	Oct 08	Nov 08	Dec 08	Janv 09	Fev 09	mars 09	Avr 09	Mai 09
Evol% prix	0,3	0,4	0,6	0,9	0,5	0,2	-0,1	-0,3	0	0,1	0,4	0,5	0,4	0,3	0,5	0,7	0,6
Moy. glissante par trimestre	-	-	0,43	0,63	0,66	0,53	0,20	-0,07	-0,13	-0,07	0,17	0,33	0,43	0,40	0,40	0,50	0,60

La moyenne trimestrielle glissante calculée pour chaque mois tient compte tient de la valeur du mois de référence et des valeurs des 2 mois précédents. Ainsi, la moyenne trimestrielle calculée au mois de référence Juillet 2008 donnera donc : (-0,1 + 0,2 + 0,5) / 3 = 0,6 / 3 = 0,20. Celle du mois d'Août 2008 donnera (-0,3 + (-0,1) + 0,2) / 3 = -0,2 / 3 = -0,07. Remarque : on ne peut calculer la moyenne glissante pour les deux premiers mois de la série.

D'une façon générale, la moyenne glissante s'écrit :

$$\bar{x}_n = \frac{1}{N} \sum_{k=0}^{N-1} x_{n-k}$$

Où N représente le nombre de valeurs successives à prendre en compte. Dans notre exemple $N = 3$

x_n représente la valeur de référence. Dans notre exemple x_n soit x_3 et correspond à la valeur du mois de Juillet 2008 soit -0,1.

k représente le rang. Dans notre exemple, k = 0 pour juillet 2008 (référence), k = 1 pour Juin 2008, etc.

Dans notre exemple cela nous donne :

$$\bar{x}_n = \frac{1}{3} \sum_{k=0}^{2} x_{3-k} = \frac{1}{3}(x_{3-0} + x_{3-1} + x_{3-2}) = \frac{x_3 + x_2 + x_1}{3} = \frac{-0{,}1 + 0{,}2 + 0{,}5}{3} = \frac{0{,}6}{3} = 0{,}20$$

Relation entre les différentes moyennes

D'une façon générale, pour une même distribution, les résultats obtenus par les différentes moyennes décrites s'organisent de la façon suivante :

Moyenne Harmonique ≤ Moyenne Géométrique ≤ Moyenne Arithmétique ≤ Moyenne Quadratique

4.2 Les paramètres de dispersion

Pour caractériser et résumer une distribution il est nécessaire de fournir deux mesures : une reflétant le centre de la distribution (mesures de tendance centrale) et une autre renseignant sur la dispersion ou l'éparpillement des données autour notamment des paramètres de tendance centrale.

Nous étudierons quatre paramètres de dispersion parmi les principaux en mettant plus particulièrement l'accent sur la variance et l'écart-type :

- Minimum, maximum, étendue et rapport de variation
- L'intervalle interquartile
- La variance
- L'écart-type

4.2.1 Minimum, maximum, étendue et rapport de variation d'une distribution

Minimum et maximum d'une série statistique correspondent respectivement et comme leur nom l'indique à la valeur minimale et à la valeur maximale rencontrées dans ladite série. Ces deux paramètres ont une triple utilité: ils permettent,

1. de calculer l'étendue de la distribution, également appelée intervalle de variation (IV), c'est-à-dire l'écart entre le minimum et le maximum. La connaissance de ce paramètre est indispensable à toute opération de discrétisation. Il permet également, pour une même variable, de comparer plusieurs distributions

$$IV = Max - Min$$

2. de calculer le rapport de variation (V) , c'est-à-dire le rapport de la valeur maximale de la distribution à la valeur minimale de la même distribution. Utile également lorsque l'on souhaite comparer, pour une même variable, différentes distribution entre elles.

$$RV = \frac{V_{max}}{V_{min}}$$

3. de connaître les limites d'une distribution en vue de son éventuelle discrétisation

Exemple: les notes d'élèves de deux classes au même examen.

Classe 1 28 élèves / notes	Classe 3 23 élèves / notes
9	3
11	12
13	16
5	5
8	3
14	7
6	10
12	7
5	19
10	16
16	5
3	11
12	13
13	11
8	9
13	13
8	9
7	10
13	12
13	8
9	15
17	15
10	8
13	
6	
13	
7	
14	

	Classe 1	Classe 3
Minimum	3	3
Maximum	17	19
Étendue	(17 – 3) = 14	(19 – 3) = 16
Rapport de variation	17/3 = 5,7	19/3 = 6,3

Le rapport de variation nous apprend que dans la classe 1 la meilleure est 5,7 fois plus élevée que la note la plus faible. Ce rapport est plus important dans la classe 3 pour laquelle il est 6,3.

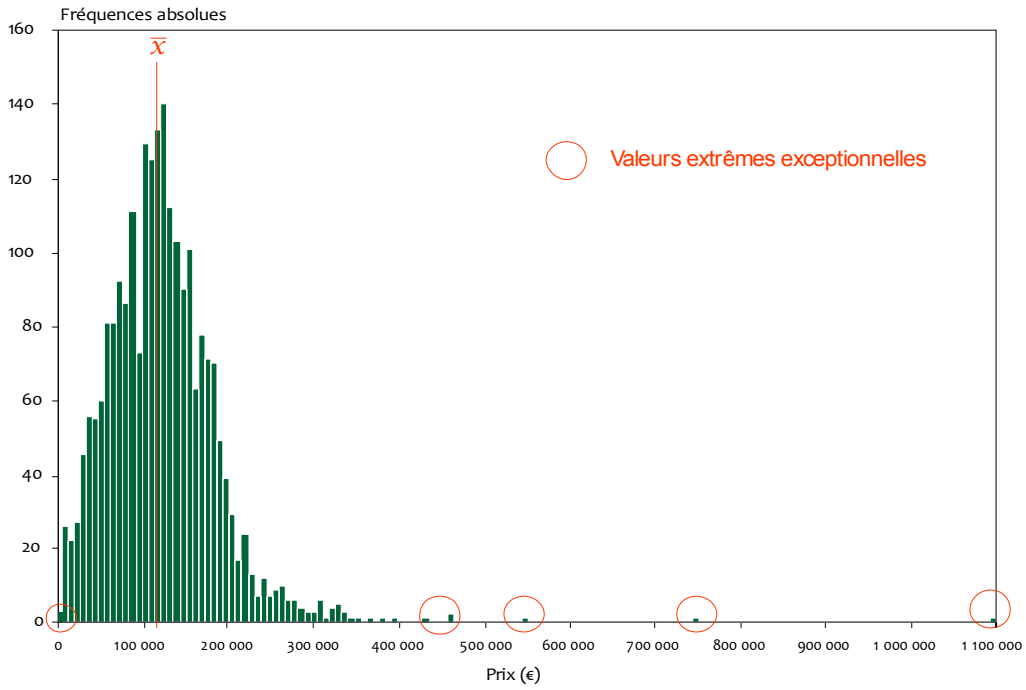
4.2.2 Intervalle interquartile

Étendue et rapport de variation ne renseignent que de façon imprécise voire trompeuse sur la dispersion des valeurs dans une distribution compte tenu notamment de la présence fréquente de valeurs extrêmes exceptionnelles, alors que le reste de la population demeure concentré sur une intervalle beaucoup plus restreint. Souvent peu nombreuses, ces valeurs extrêmes peuvent pourtant perturber de façon importante l'appréciation que l'on peut se faire des caractéristiques d'une distribution.
Pour s'en rendre compte, il suffit d'examiner l'exemple qui suit:

Exemple:
Nous disposons de l'ensemble des informations relatives aux transactions immobilières à titre onéreux dans l'ancien pour les maisons individuelles sur le département du Doubs pour l'année 2003. On cherche à étudier et résumer la distribution (env. 2 300 valeurs) afin d'en extraire les principales informations de prix en vue d'une présentation à des élus. On calcule donc les paramètres de tendance centrale et de dispersion connus jusqu'à ce stade de la présentation et on obtient :

Distribution brute	
Mesures de tendance centrale	
Médiane	117 427,50 €
Moyenne	122 164,57 €
Mesures de dispersion	
Minimum	3 811,00 €
Maximum	1 100 194,00 €
Étendue (Intervalle de variation)	1 096 383,00 €
Rapport de variation	288,7

Intervalle et rapport de variation sont très importants tant les individus qui composent la population étudiée diffèrent des uns des autres pour le caractère appréhendé (prix). Par contre, médiane et moyenne ne sont que très peu perturbées pas les valeurs extrêmes certes exceptionnelles par leur grandeur mais trop peu nombreuses au regard de la masse des valeurs dites « dans la norme » (voir histogramme). Preuve en est: si on retire ces valeurs extrêmes, moyenne et médiane ne bougent que très peu. A contrario, étendue et rapport de variation s'en trouve considérablement amoindris:



Distribution sans valeurs extrêmes	
Mesures de tendance centrale	
Médiane	117 400,00
Moyenne	120 613,16
Mesures de dispersion	
Minimum	4 600,00
Maximum	346 700,00
Étendue (Intervalle de variation)	342 100,00
Rapport de variation	75,37

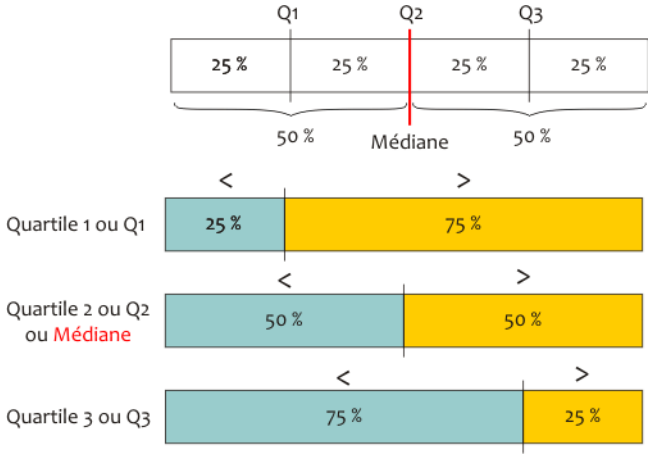
La nécessité se fait donc d'utiliser d'autres mesures de dispersion plus à même de prendre en compte de façon plus précise la dispersion d'une distribution comme par exemple l'intervalle interquartile. Auparavant il convient cependant de définir les quartiles.

4.2.3 Les quartiles, déciles et centiles

Dans une distribution dont les individus ont été au préalable triés par ordre croissant, les quartiles correspondent aux trois valeurs qui partagent une population en quatre sous-ensembles de même taille, c'est-à-dire d'effectifs égaux. Par convention, les quartiles sont respectivement par Q1, Q2 et Q3 de telle sorte que l'on peut écrire pour chacun d'eux :

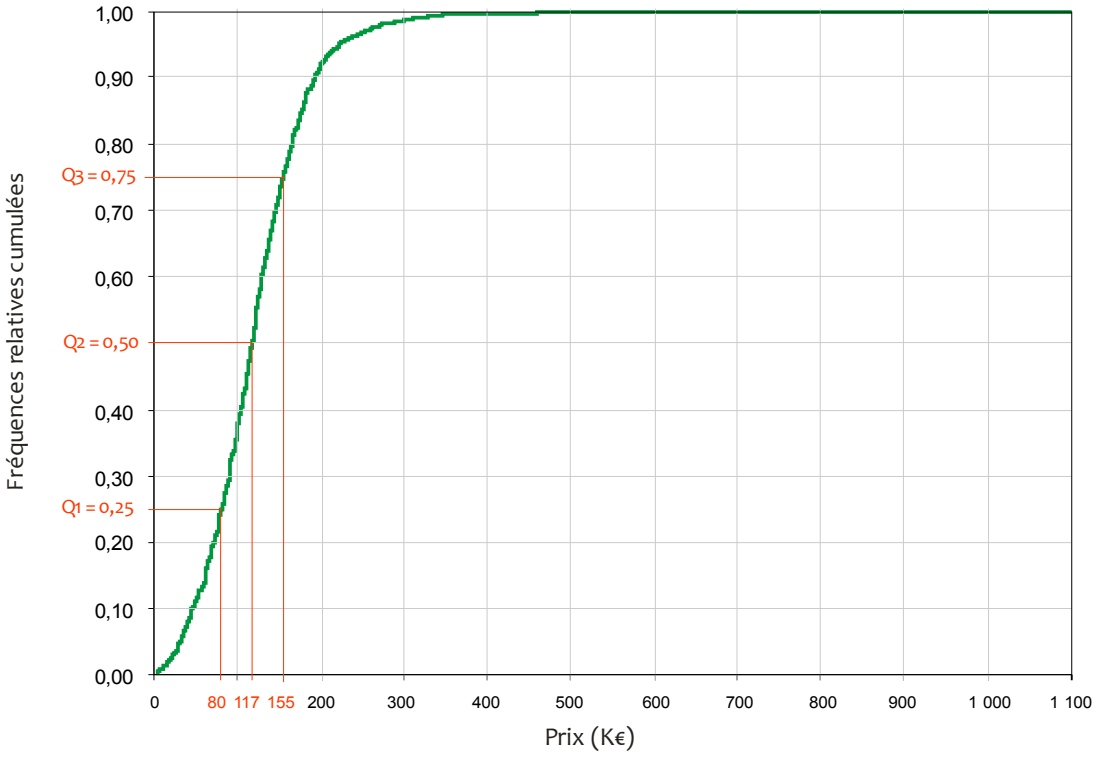
Quartile 1 ou Q1	25 % des effectifs de la population ont une valeur inférieur à Q1 et 75 % une valeur supérieur. Dans une distribution relative au revenu des ménages par exemple, Q1 marque la limite entre les ménages les 25% les plus modestes et les 75% les plus aisés

Quartile 2 ou Q2	50 % des effectifs de la population ont une valeur inférieur à Q1 et 50 % une valeur supérieur. Q2 correspond à la médiane.
Quartile 3 ou Q13	75 % des effectifs de la population ont une valeur inférieur à Q1 et 25 % une valeur supérieur. Dans une distribution relative au revenu des ménages par exemple, Q3 marque la limite entre les ménages les 25% les plus riches et les 75% restant de la population.



Les quartiles se déterminent de la même façon que la médiane et nécessitent, comme pour cette dernière, que les valeurs de la distribution aient été au préalable classées par ordre croissant. Il suffit alors de cumuler les fréquences (absolues ou relatives) et de se positionner à l'endroit où résident les seuils Q1 = 25%, Q2 = 50 % et Q3 = 75 % et de lire les valeurs correspondantes de la distribution.

De façon visuelle et approximative, il est toujours possible d'utiliser, après l'avoir tracée, la courbe des fréquences cumulées comme suit :



A noter qu'avec **Microsoft Excel** ainsi qu'avec **OpenOffice Calc** il est possible de déterminer automatiquement les quartiles d'une distribution (Fonction **QUARTILE** dans les deux cas).

De la même manière, et dans le but de préciser et d'affiner encore l'analyse de la dispersion d'une distribution, on peut faire appel aux notions de déciles et de centiles. Le principe demeure le même que pour les quartiles à la différence que la population est ici divisée respectivement en 10 et 100 sous-populations d'égal effectifs:

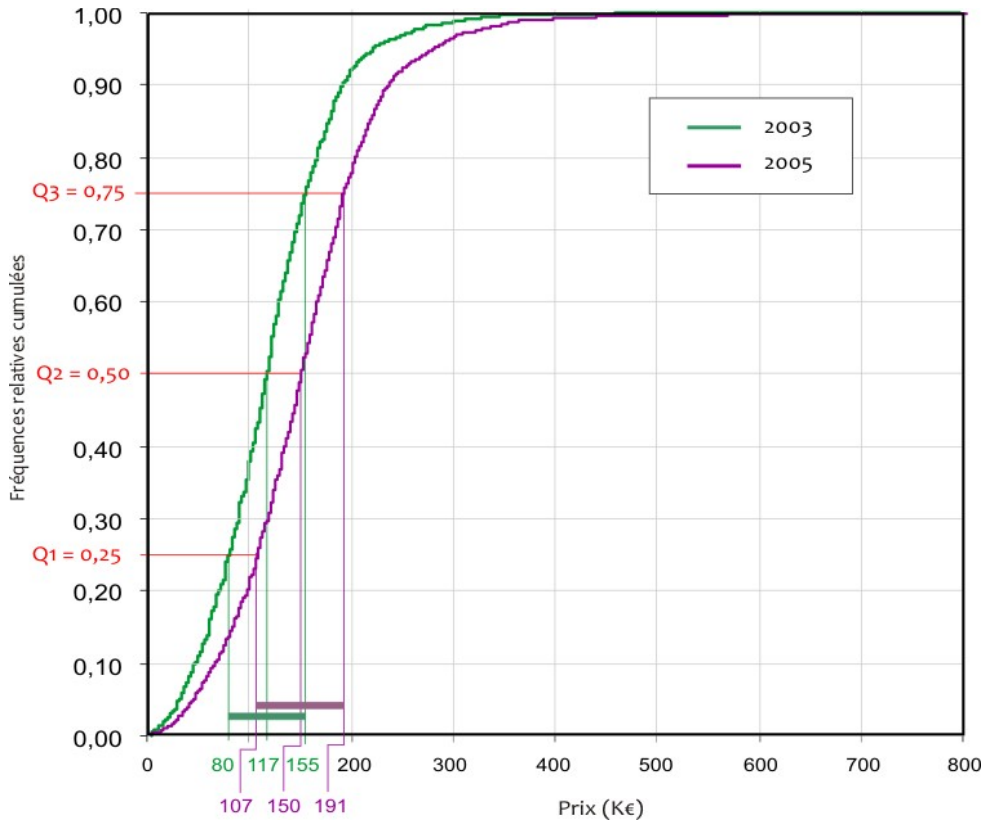
Décile 1 ou D1	10 % des effectifs de la population ont une valeur inférieur à D1 et 90 % une valeur supérieure.
Décile 2 ou D2	20 % des effectifs de la population ont une valeur inférieur à D2 et 80 % une valeur supérieure.
Décile 3 ou D3	30 % des effectifs de la population ont une valeur inférieur à D3 et 70 % une valeur supérieure.
Décile 4 ou D4	40 % des effectifs de la population ont une valeur inférieur à D4 et 60 % une valeur supérieure.
Décile 5 ou D5	50 % des effectifs de la population ont une valeur inférieur à Q1 et 50 % une valeur supérieure. D5 correspond à la médiane.
Décile 6 ou D6	60 % des effectifs de la population ont une valeur inférieur à D6 et 40 % une valeur supérieure.
Décile 7 ou D7	70 % des effectifs de la population ont une valeur inférieur à D7 et 30 % une valeur supérieure.
Décile 8 ou D8	80 % des effectifs de la population ont une valeur inférieur à D8 et 20 % une valeur supérieure.
Décile 9 ou D9	90 % des effectifs de la population ont une valeur inférieur à D9 et 10 % une valeur supérieure.

Grâce aux quartiles (comme aux déciles) il devient possible d'améliorer la description et l'analyse de la dispersion d'une distribution et de comparer de façon plus sûre et plus pertinente deux distributions entre elles ou bien encore la distribution d'une même population à deux dates différentes pour une même variable, en rappelant que la mesure de dispersion qui nous intéresse n'est pas nécessairement le quartile (qui n'est pas une mesure de dispersion) mais **l'intervalle interquartile**, c'est à dire la différence entre le troisième quartile (Q3) et le premier quartile (Q1).
Noté I_2Q il s'écrit : $I_2Q = Q_3 - Q_1$

L'intervalle interquartile contient toujours 50 % de la distribution. Plus il est large, plus la distribution est dispersée. Afin d'illustration, reprenons l'exemple précédent relatif aux prix des logements lors de transactions immobilières sur maisons individuelles dans le département du Doubs en 2003 et ajoutons l'année 2005 :

	2003	**2005**
Moyenne	122 165	154 220
Minimum	3 811	2 300
Maximum	1 100 194	800 000
Étendue (Intervalle de variation)	1 096 383	797 700
Rapport de variation	288,7	347,8
Q1	80 036	106 770
Q2 (médiane)	117 427	150 000
Q3	155 498	190 560
I_2Q	75 462	83 790

Graphiquement, cela donne :



Un certains nombre d'observations et de conclusions peuvent d'ores et déjà être tirées à partir des mesures effectuées et des graphiques établies qui permettent de décrire et de résumer un phénomène et sa distribution (rappelons qu'au départ nous avions une série de près de 3 000 valeurs) :

- En 2003, 50 % des biens vendus avaient une valeur de marché inférieure à 117 000 € (et de façon corollaire 50 % des biens vendus l'ont été à un pris supérieur à 117 000 €).

- En 2005, pour le même prix, 30 % des biens vendus avaient un prix inférieur et 70 % un prix supérieur: les prix ont monté. Le seuil de 50 % (médiane) est rendu à 150 000 € en 2005 soit plus élevé de 28,2 %. Sur les deux années, la moyenne passe de 122 165 à 154 220 soit une progression de 26,2 % moins importante que la médiane : de ce constat on peut en déduire que la dispersion des valeurs s'est aggravée ce que confirme l'intervalle interquartile calculé sur les deux dates

Malgré l'amélioration de la description et de la distribution et de la variable associée, il n'est cependant pas encore possible de décrire sans ambiguïté celle-ci et surtout de mesurer avec précision la dispersion des valeurs la composant. Alors que l'étendue (ou intervalle de variation) dépend uniquement des valeurs extrêmes, que l'intervalle interquartile dépend de 50 % des données situées au milieu de la distribution, il nous faut introduire un nouveau et ultime paramètre qui dépendra de tous les résultats. Cette mesure devra avoir la propriété d'être petite lorsque les valeurs seront proches les unes des autres, et grande lorsque ces mêmes valeurs seront très éparpillées. Cette mesure existe, elle se nomme écart-type.

Exercice 16 : fichier Excel associé « Exercice 16 - Quartiles et I2Q.xls ».

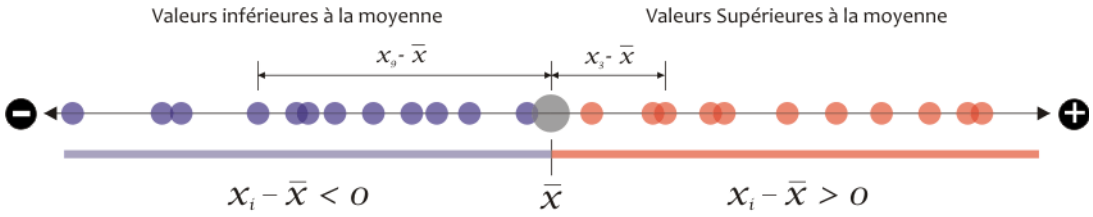
4.2.4 Variance et Écart-type et variance de données non groupées

Considérons une distribution pour laquelle on a calculé les paramètres de tendance centrale comme la médiane et la moyenne. Comme leurs noms l'indiquent, et comme mentionné plus haut, ces mesures caractérisent le centre de la distribution. Parmi celles-ci, considérons la moyenne comme une référence.
Que penser alors de l'écart entre chaque valeur de la distribution et cette moyenne ?

$$(x_i - \bar{x})$$

Plus cet écart sera faible, plus la valeur x_i sera proche de la moyenne et donc du centre de la distribution. A contrario, plus l'écart sera important et plus x_i sera éloignée du centre de la distribution. La prise en compte de la somme l'ensemble des écarts à la moyenne, c'est-à-dire de la somme de tous les écarts entre les x_i et la moyenne donne logiquement 0, la moyenne étant au centre de la distribution:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$



$$\underbrace{\sum_{[x_i<\bar{x}]} (x_i - \bar{x})}_{<0} + \underbrace{\sum_{[x_i>\bar{x}]} (x_i - \bar{x})}_{>0} = \sum (x_i - \bar{x}) = 0$$

avec : $\sum_{[x_i<\bar{x}]} |(x_i - \bar{x})| = \sum_{[x_i>\bar{x}]} (x_i - \bar{x})$

Si l'on veut tenir compte de l'ensemble des distances à la moyenne sans pâtir d'une somme nulle résultat de la compensation entre écarts négatifs et écarts positifs, il est nécessaire d'élever au carré [2] chaque écart de telle sorte que l'on est :

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 \neq 0$$

Que penser ensuite de la moyenne calculée de ces écarts élevés au carré ?

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

Ce paramètre écrit S^2 ou σ^2 c'est la **variance**, notion fondamentale en statistique. La variance satisfait à toutes les exigences énoncées plus haut relativement à la mesure de la dispersion d'une distribution. La variance pose toutefois le problème de proposer un résultat en unité élevée au carré. Si les données x_i sont en euros, la moyenne sera en euros, de même que l'écart $(x_i - \bar{x})$ alors que la variance sera en euros carrés.
Pour revenir à l'unité initiale il faut extraire la racine carrée de la variance ou **écart-type**. Ce dernier s'écrit :

2 Tout nombre, positif ou négatif, devient positif lorsqu'il est élevé au carré On préfèrera l'utilisation des puissances plutôt que les valeurs absolues, les premières se prêtant mieux au calcul algébrique que les secondes.

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

Exemple :

Individu	Intitulé	Revenu moyen (€/an)	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
x_1	BE - Belgique	19 129	4 166,52	17 359 888,9
x_2	CZ – Rép. Tchèque	6 139	-8 823,48	77 853 799,3
x_3	DK - Danemark	25 113	10 150,52	103 033 056,3
x_4	DE - Allemagne	20 208	5 245,52	27 515 480,1
x_5	EE - Estonie	5 304	-9 658,48	93 286 235,9
x_6	IE - Irlande	26 043	11 080,52	122 777 923,5
x_7	GR - Grèce	12 126	-2 836,48	8 045 618,8
x_8	ES - Espagne	13 613	-1 349,48	1 821 096,3
x_9	FR - France	18 481	3 518,52	12 379 983,0
x_{10}	IT - Italie	17 213	2 250,52	5 064 840,3
x_{11}	CY - Chypre	18 500	3 537,52	12 514 047,7
x_{12}	LV - Lettonie	4 086	-10 876,48	118 297 817,2
x_{13}	LT - Lituanie	3 939	-11 023,48	121 517 111,3
x_{14}	LU - Luxembourg	3 4213	19 250,52	370 582 520,3
x_{15}	HU - Hongrie	4 377	-10 585,48	112 052 386,8
x_{16}	MT - Malte	9 954	-5 008,48	25 084 871,9
x_{17}	NL - Pays-Bas	20 753	5 790,52	33 530 121,9
x_{18}	AT - Autriche	20 399	5 436,52	29 555 749,7
x_{19}	PL - Pologne	4 149	-10 813,48	116 931 349,7
x_{20}	PT - Portugal	9 918	-5 044,48	25 446 778,5
x_{21}	SI - Slovénie	10 719	-4 243,48	18 007 122,5
x_{22}	SK - Slovaquie	4 376	-10 586,48	112 073 558,8
x_{23}	FI - Finlande	20 787	5 824,52	33 925 033,2
x_{24}	SE - Suède	19 898	4 935,52	24 359 357,7
x_{25}	UK - Royaume-Uni	24 625	9 662,52	93 364 292,75
	Σ	**374 062**	**0,00**	**1 716 380 042,2**

Avec :

$$\bar{x} = 14\,962{,}48$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 1\,716\,380\,042{,}2$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = \frac{1\,716\,380\,042{,}2}{25} = 68\,655\,201{,}7$$

D'où $\sigma = \sqrt{\sigma^2} = \sqrt{68\,655\,201{,}7} = 8\,285{,}8\,€$

On mesure une dispersion élevée liée aux fortes différences de richesse entre pays de l'Union Européenne. Si l'on effectue le même travail sur le pays membre de l'union avant 2000, on obtient un écart-type réduit quasiment de moitié de 5 786,8 € ---> ensemble plus homogène de pays, dispersion moins grande. L'arrivée de nouveaux pays de l'Est plus pauvres a fait chuter la moyenne et augmenter l'écart-type.

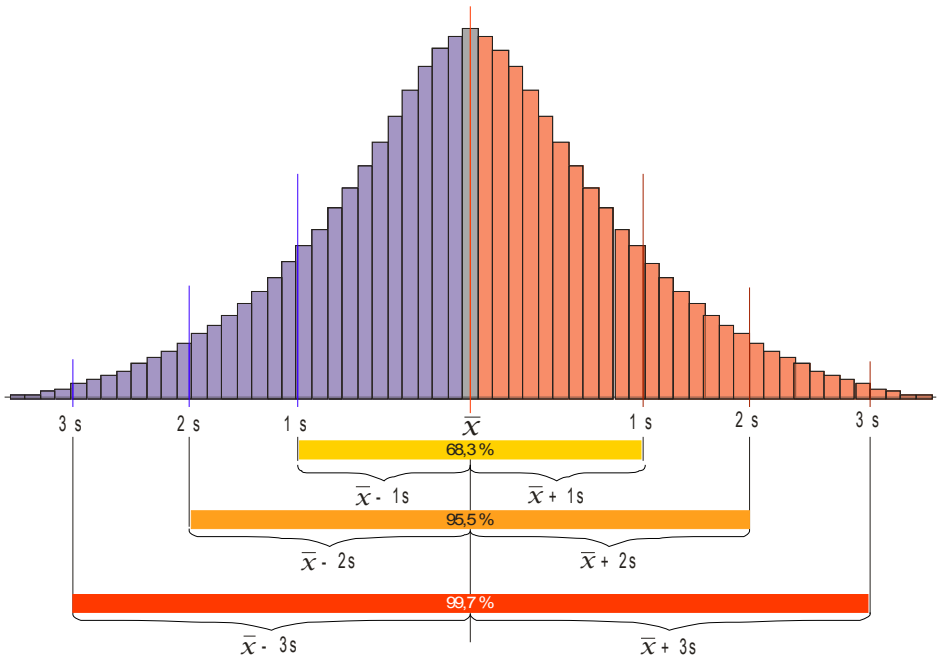
Propriétés et utilité de l'écart-type :

- Son unité est celle de la variable à laquelle il se rapporte. Si la variable étudiée est exprimée en euro (€), l'unité de l'écart-type sera l'euro.
- Un écart-type faible signifie que les valeurs sont relativement concentrées autour de la moyenne et que la population regroupe des individus aux caractéristiques relativement homogène.
- A contrario, un écart-type élevé est révélateur de valeurs très dispersées autour de la moyenne et d'une population hétérogène.
- L'écart-type peut servir de bornes pour délimiter une partie de la population, celle la plus proche des tendances centrales et donc la plus représentative du phénomène étudié et la plus pertinente à son interprétation, ou bien celle la plus éloignée. En prenant comme point de référence la moyenne d'une distribution et en considérant l'écart-type comme une unité de distance à cette moyenne, et de part et d'autre de celle-ci, il devient possible de mesurer la proportion de la population (ou le nombre d'individus) compris entre les limites ainsi définies qui s'écrivent :

$[\bar{x} - \sigma\ ;\ \bar{x} + \sigma]$	Contient tous les individus dont le caractère (la valeur) est comprise entre la moyenne – une fois écart-type et la moyenne + une fois l'écart-type.
$[\bar{x} - 1{,}5\,\sigma\ ;\ \bar{x} + 1{,}5\,\sigma]$	Contient tous les individus dont le caractère (la valeur) est comprise entre la moyenne – 1,5 fois écart-type et la moyenne + 1,5 l'écart-type.
$[\bar{x} - 2\,\sigma\ ;\ \bar{x} + 2\,\sigma]$	Contient tous les individus dont le caractère (la valeur) est comprise entre la moyenne – deux fois écart-type et la moyenne + deux fois l'écart-type.
$[\bar{x} - a\,\sigma\ ;\ \bar{x} + a\,\sigma]$	Contient tous les individus dont le caractère (la valeur) est comprise entre la moyenne – a fois écart-type et la moyenne + a fois l'écart-type.

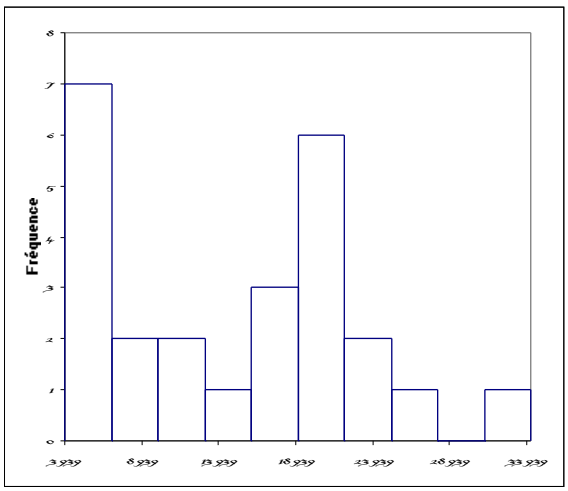
Dans les conditions statistiques idéales c'est-à-dire celle d'une population parfaitement bien distribuée autour des paramètres centraux, on sait que :

$[\bar{x} - \sigma\ ;\ \bar{x} + \sigma]$	Contient 68,3 % de l'ensemble des individus de la distribution.
$[\bar{x} - 2\,\sigma\ ;\ \bar{x} + 2\,\sigma]$	Contient 95,5 % de l'ensemble des individus de la distribution.
$[\bar{x} - 3\,\sigma\ ;\ \bar{x} + 3\,\sigma]$	Contient 99,7 % de l'ensemble des individus de la distribution.



Il est d'usage assez fréquent de considérer ces intervalles comme un moyen simple et efficace d'éliminer les valeurs extrêmes d'une distribution avant traitement et analyse statistique. Ce sujet sera abordé plus avant.

Dans l'exemple précédent, l'intervalle $[\bar{x}-\sigma\ ;\ \bar{x}+\sigma]$ correspond à l'intervalle de valeurs [14 962,5 – 8 285,8 ; 14 962,5 + 8 285,8] = [6 676,7 ; 23 248,3] et contient 14 unités statistiques soit 56 % de la distribution. Un chiffre bien en-dessous de ce que promet la distribution idéale évoquée. L'analyse de l'histogramme de la distribution permet d'élucider le mystère:



La structure bimodale et la forte dispersion des valeurs autour de la moyenne explique tout ou partie de la faible proportion d'individus compris dans ce premier intervalle. L'extension de l'intervalle à 1,5 σ de part et d'autre de la moyenne permet d'accroître la proportion de la population à 92 %. Cette dernière atteint 96 % lorsque les limites de l'intervalle sont repoussées à ± 2 σ.

Écart-type et variance de données groupées

De la même façon, il est possible, en respectant certaines règles, de calculer la variance et l'écart-type pour des données groupées, c'est-à-dire ayant fait l'objet d'une discrétisation.
Comme ce fut le cas pour le calcul de la moyenne de données groupées, il faut prendre en compte le centre de chaque classe et considérer que les individus d'une même classe ont tous la même valeur, celle du centre de leur classe.

Exemple : Trouver la variance et l'écart-type de la distribution suivante:

Classes	Fréquence absolue (f_i)	Centre de classe (x_i)	$(f_i \cdot x_i)$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i \cdot (x_i - \bar{x})^2$
[4 ; 6]	2	(6 + 4)/2 = 5	10	- 3,3	10,89	21,78
[7 ; 9]	5	(9 + 7)/2 = 8	40	- 0,3	0,09	0,45
[10 ; 12]	3	(12 + 10)/2 = 11	33	2,7	7,29	21,87
$\sum$	10		83	- 0,9		44,10

Cette distribution aura la même variance et le même écart-type que la série {5; 5; 8; 8; 8; 8; 8; 11; 11; 11}.

$$\bar{x} = \frac{83}{10} = 8{,}3 \quad \text{et} \qquad \sigma^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n-1} = \frac{44{,}10}{9} = 4{,}90 \quad \Rightarrow \quad \sigma = \sqrt{\sigma^2} = \sqrt{4{,}90} = 2{,}2$$

Quelques remarques :

- Contrairement au cas continue (données non groupées), $\sum (x_i - \bar{x}) \neq 0$. Cela est lié au fait que l'on ne travaille pas sur des valeurs justes mais sur des centres de classes. Dans ce cas, la compensation n'est pas automatique, d'où l'inégalité.

- Dans la formule de calcul de la variance (et indirectement dans celle de l'écart-type), n est remplacé par *(n-1)*. Cette substitution est courante lorsqu'il s'agit non pas de calculer sûrement variance et écart-type comme on peut le faire dans le cas d'une population à l'effectif connu et complet, mais lorsque l'on travaille sur un échantillon ou une population aux caractéristiques tronquée comme c'est le cas quand il y a eu discrétisation. On est alors amené à estimer la variance ou l'écart-type plutôt que de les calculer. Si n est le dénominateur de la variance d'un échantillon, l'estimation sera trop faible. En changeant n par *(n-1)* au dénominateur, la fraction augmente juste assez pour que la variance de l'échantillon devienne une bonne estimation de la variance de la population.

Exercice 17 : fichier Excel associé « Exercice 17 - Ecart-type.xls ».

4.2.5 Le coefficient de variation

L'écart-type, malgré sa pertinence dans la mesure de la dispersion d'une distribution, possède un inconvénient majeur: il est exprimé dans l'unité de la variable à laquelle il se rapporte. Il est alors impossible de comparer les dispersions de deux ou davantage distributions ayant un lien entre elles (lien de causalité ou autre) et dont les valeurs s'expriment dans des unités différentes.
Le coefficient de variation est une mesure de dispersion des observations d'une variable quantitative d'intervalle qui permet de s'affranchir de la notion d'unité et ainsi de comparer la dispersion de différentes distributions.

C'est une mesure neutre qui s'exprime la plupart du temps en pourcentage. Il se calcule en divisant l'écart-type par la moyenne et s'écrit donc :

Coefficient de variation :

$$C_v = \frac{\sigma}{\bar{x}}$$

Plus grand est le coefficient de variation, plus grande est la dispersion.

Exemple : considérons la surface des logements dits de petite taille (1 à 3 pièces) à celle des logements dits de grande taille (4 pièces et plus).

Pour le groupe de logements de 1, 2 et 3 pièces nous obtenons :

$$\bar{x} = 56{,}6\, m^2$$

$$\sigma = 12{,}4\, m^2$$

Pour le groupe de logements dits récents nous obtenons :

$$\bar{x} = 81{,}5\, m^2$$

$$\sigma = 13{,}2\, m^2$$

A première vue, et en examinant seulement les écarts types, on pourrait conclure que la dispersion de la surface des logements de grande taille est plus élevée que celle des petits logements. Le calcul des coefficients de variation respectifs montre qu'il n'en est rien :

Pour les petits logements $C_v = 0{,}219 (21{,}9\%)$

Pour les grands logements $C_v = 0{,}162 (16{,}2\%)$

On note que le coefficient de variation des logements de petite taille est plus élevé que celui des logements de grande taille. Contrairement à ce que laissait penser les écart-type calculés, la dispersion pour le groupe des petits logements est plus élevée que celle des grands logements.

Chapitre 5

5. Séries chronologiques : progression et indices

L'utilisation et l'analyse de séries chronologiques, c'est-à-dire de séries qui figurent l'évolution d'une variable statistique au cours du temps, s'appuient sur deux outils principaux : d'une part la progression et d'autre l'indice. Ces deux outils mesurent les variations d'une variable entre deux dates ou plus selon un pas de temps régulier ou non.

5.1 Progression

La progression mesure le sens et l'intensité du changement intervenu sur une variable numérique V à différents temps t. Au temps t la variable s'écrit par convention V_t (V indice t). Lorsque t = 0 la variable s'écrit V_0, quand t = 1 elle s'écrit V_1, lorsque t = n on a V_n. A chaque temps t la variable V est caractérisée par une valeur différente.

Exemple : considérons la population de la France à différentes dates ,comme suit :

t	Date	Population
0	1876	V_1 = 38 437 592
1	1901	V_2 = 40 681 415
2	1921	V_3 = 39 209 518
3	1946	V_4 = 40 506 639
4	1962	V_5 = 46 243 173
5	1982	V_6 = 54 334 871
6	1999	V_7 = 58 518 395
7	2007	V_8 = 62 106 000

Dès lors, plusieurs mesures de progression peuvent être appliquées afin de caractériser la variation de la variable « population ».

5.1.1 La variation absolue

La variation absolue correspond à la différence de valeurs de la variable V entre deux dates, deux temps t. Elle s'écrit :

Variation absolue : $\Delta V = V_t - V_0$

Reprenons notre exemple relatif à la population de la France à travers le temps. La variation absolue de population entre 1901 et 1946 s'écrit :

$V_0 =$ Population quand t = 0, c'est-à-dire à la date de départ, ici 1901 = 40 681 415

$V_1 =$ Population quand t = 1, c'est-à-dire à la date d'arrivée, ici 1946 = 40 506 639

$$\Delta V_{1901-1946} = V_1 - V_0 = Pop_{1946} - Pop_{1901} = 40\,506\,639 - 40\,681\,415 = -174\,776$$

La variation absolue de population en France entre 1901 et 1946 est donc négative signifiant une baisse des effectifs évaluée à – 174 776 habitants.

Le taux de croissance sur une période (entre deux dates)

La variation absolue mesure l'évolution brute et le sens de variation d'une quantité sans indication de son intensité par rapport à une situation de référence V_O. La mesure de cette intensité, en plus du sens de variation (positif ou négatif), se réalise par l'intermédiaire d'un taux, c'est-à-dire d'un rapport d'une différence sur une quantité de référence le tout exprimé en pourcentage. Ce taux s'écrit :

Taux de croissance sur une période : $g_t = \frac{V_1 - V_0}{V_0} = \frac{\Delta V}{V_0}$

Avec : V1 – Vo représente la différence. On reconnaît ici la variation absolue D V décrite plus en amont
Vo représente la quantité de référence, c'est-à-dire la valeur de la variable à la date initiale t = o

Le taux de croissance ainsi obtenu est sans unité et le résultat multiplié par 100 donne un pourcentage.

Exemple: nous avons vu dans dans le cas de la variation absolue que la population de la France entre 1901 et 1946 avait évolué à la baisse avec une perte de 174 776 habitants. Quelle est l'intensité de cette diminution par rapport à la situation initiale de 1901 ? en d'autres termes, quelle est le taux de croissance de la population en pourcentage entre 1901 et 1946 ?

$$g_{\%} = \frac{V_1 - V_0}{V_0} = \frac{Pop_{1946} - Pop_{1901}}{Pop_{1901}} = \frac{-174\,776}{40\,681\,415} = -0{,}0043 = -0{,}43\,\%$$

Entre 1901 et 1946, la population française a diminué de 174 776 habitants ce qui correspond à une baisse de -0,43 %.

5.1.2 Le taux de croissance sur plusieurs périodes ou taux de croissance moyen

Que se passe-t-il lorsque l'on dispose pour une même variable de plusieurs valeurs correspond à son état à plusieurs dates et que l'on souhaite connaître le taux de croissance moyen sur l'ensemble des périodes ? Ce cas de figure est similaire à celui abordé dans le paragraphe concernant la moyenne géométrique et la formule utilisée pour calculer le taux de croissance moyen sur plusieurs périodes en est identique. En voici la formulation adaptée :

Taux de croissance moyen sur plusieurs périodes :

$$\bar{g} = \left[\frac{V_t}{V_0}\right]^{\frac{1}{t}} - 1$$

Avec : $\bar{g}$ Taux de croissance moyen sur t périodes

V_0 Valeur de la variable étudiée à la date initiale

V_t Valeur de la variable étudiée à la date terminale

Exemple : reprenons le cas de la France et de sa population dans la première moitié du XXe siècle. Nous disposions des chiffres de la population pour les années 1901 et 1946. De nouvelles données sont disponibles à l'intérieur de cet intervalle pour les années 1906, 1911, 1921, 1926, 1931 et 1936 formant la série suivante :

t	Date	Population
0	1901	V_0 = 40 681 415
1	1906	V_1 = 41 066 809
2	1911	V_2 = 41 479 006
3	1921	V_3 = 39 209 518
4	1926	V_4 = 40 743 897
5	1931	V_5 = 41 834 923
6	1936	V_6 = 41 911 530
7	1946	V_7 = 40 506 639

Quel est le taux de croissance moyen de la population française entre 1901 et 1946 ?

$$\bar{g} = \left[\frac{V_t}{V_0}\right]^{\frac{1}{t}} - 1 = \left[\frac{V_7}{V_0}\right]^{\frac{1}{7}} - 1 = \left[\frac{40\,506\,639}{40\,680\,415}\right]^{\frac{1}{7}} - 1 = 0{,}9993 - 1 = -0{,}00061 = -0{,}061\%$$

5.1.3 Augmentations ou diminutions successives

Lorsque qu'une grandeur croît successivement à des taux différents à chaque période et que l'on veut connaître la valeur de la grandeur au terme des augmentations ou diminutions successives on applique la formule suivante:

$$V_t = V_0 \prod_{i=1}^{t} (1 + g_i)$$

Exemple : on connaît à un moment donné t = 0 le prix mensuel moyen des loyers au m² : V_0 = 7,48 €/m². Durant quatre années successives la seule information disponible concerne la hausse moyenne observée pour ces mêmes loyers soit : g_1 = 2,33 % pour le 1ère année, g_2 = -1,03 % pour le 2ième année, g_3 = 1,93 % pour le 3ième année et g_1 = 2,48 % pour le4ième année. Quelle la valeur du loyer mensuel moyen à l'issue de ces 4 années, autrement dit, déterminez V_4 ?

$$V_4 = V_0 \prod_{i=1}^{t} (1+g_i) = 7{,}48(1+g_1)(1+g_2)(1+g_3)(1+g_4)$$

$$V_4 = 7{,}48(1+0{,}0233)(1-0{,}0103)(1+0{,}0193)(1+0{,}0248) = 7{,}91\,€/m^2$$

A l'issue des 4 années, le loyer mensuel moyen au m² s'élève à 7,91 €/m² compte tenu des hausses successives enregistrées.

De la même façon, on peut, connaissant la valeur terminale V_t d'une variable et les taux de croissance successifs pour y aboutir, renverser le problème et calculer la valeur initiale V_0 selon la formule suivante:

$$V_0 = \frac{V_t}{\prod_{i=1}^{t}(1+g_i)}$$

Pour aller un peu plus loin : temps de doublement d'une grandeur

Quel est le temps nécessaire à une grandeur quelconque (population, prix, etc.) pour sa valeur double en considérant un taux de croissance moyen constant ?La solution est donnée par la formule suivante:

$$2V_0 = V_0(1+g)^t$$

L'inconnue est ici le paramètre t qui représente la quantité de temps nécessaire à la vérification de l'égalité posée. Sans faire étalage de démonstrations mathématiques superflues, on peut écrire que t est égal à :

$$t = \frac{\ln 2}{\ln(1+g)}$$

(*ln* représente le logarithme népérien)

Exemple : en 1850, la population de la France était d'environ 35 millions d'habitants. En considérant un taux de croissance annuel moyen de 0,53 % observé à l'époque, combien de temps aurait-il fallu pour que celle-ci double ?

$$2V_0 = V_0(1+g)^t \Leftrightarrow$$

$$2(35\,000\,000) = 35\,000\,000(1+0{,}0053)^t \Leftrightarrow$$

$$t = \frac{\ln 2}{\ln(1+0{,}0053)} = \frac{0{,}6934}{0{,}00258} = 131{,}1\,ans$$

Il aurait fallu au pays 131 ans pour voir sa population doubler, soit un doublement prévu en 1981. Or en 1981, la population de la France était de 54 millions d'âmes soit quelques 77 % de son objectif. C'est donc que le taux annuel moyen de croissance de la population n'a été constant sur la période et a été en moyenne inférieur à 0,53%. Les deux grandes guerres y sont probablement pour quelque chose.

Remarque : on observe que la valeur initiale V_0 n'intervient pas du tout dans le résultat finale et que la seule connaissance du taux de croissance annuel moyen (g) est nécessaire pour connaître, quel que soit le phénomène, un taux de doublement.

Exercice 19 : fichier Excel associé « Exercice 19 - Progression.xls ».

5.2 Indices

L'indice est avant toute chose un résumé d'informations. Il est une autre façon d'exprimer une variation relative, c'est-à-dire un rapport de valeurs absolues, en désignant dès le départ l'une d'elles comme référence ou *base* à laquelle on affecte par convention la valeur 100.

Exemple: plutôt que de dire que le prix d'un bien immobilier a augmenté de 12,5 % de 2006 à 2007, on peut écrire que sur base 100 en 2006, il était en 2007 à l'indice 112,5 (on note couramment 2006 = 100). Cette façon d'exposer une variation n'ajoute rien à la précédente si ce n'est qu'elle permet d'éviter les variations négatives : ainsi, au lieu de parler d'une baisse de – 20 % on écrira que l'indice est passé de 100 à 80.

Dans cet exemple, on a affaire à un *indice élémentaire* c'est-à-dire qui renseigne sur l'évolution temporelle ou spatiale d'une seule valeur, par opposition à un indice complexe ou indice synthétique qui résume quant à lui l'évolution de plusieurs grandeurs comme plusieurs prix, plusieurs quantités, plusieurs valeurs (prix x quantités), etc.

5.2.1 Les indices élémentaires

Définition : un indice élémentaire est un rapport entre deux valeurs d'une même grandeur dans deux situations différentes dont une est appelée « base » et adoptée comme valeur de référence, et l'autre situation « courante ». Si on note $I_{1/0}$ l'indice se rapportant à une grandeur simple g dans la situation 1 par rapport à la situation 0, on a :

$$I_{1/0} = \frac{g_1}{g_0}$$

Exemple : en 1876, la population française comptait 38,4 millions d'habitants. En 2007, cette même population était évaluée à 62,1 millions d'âmes. Calculer l'indice de variation de population en prenant comme référence l'année 1876.

$$I_{1/0} = I_{1876/2007} = \frac{62,1}{38,4} = 1,62$$

La situation de base, ou de référence (g_0), est toujours placée au dénominateur, le numérateur (g_1) étant occupé par la situation dite *courante*. Pour éviter de trainer trop de chiffres après la virgule, on a pour habitude de multiplier le résultat d'un indice par 100. Dans l'exemple précédent on obtient donc 1,62 x 100 = 162. En base 1876 = 100, la population française était en 2007 à l'indice 162, soit une population en progression de 62 % entre 1876 et 2007.

Les indices élémentaires ont trois propriétés:

La réversibilité: un indice élémentaire est réversible c'est-à-dire que l'on inverse les situations comme suit :

$$I_{0/1} = \frac{1}{I_{1/0}} = \frac{1}{\frac{g_1}{g_0}}$$

Cette propriété est peu utilisée dans les comparaisons chronologiques car il est peu fréquent de mettre au dénominateur une période postérieure à celle mise au numérateur. Elle l'est en revanche beaucoup plus et

même essentielle lorsqu'il s'agit de comparaisons géographiques pour lesquelles il n'existe aucune relation d'ordre entre les lieux comparés et où le choix du lieu de référence demeure parfaitement arbitraire.

Exemple: prenons le revenu moyen par ménage de 3 pays de l'Union Européenne en 2007 (Danemark, France et Hongrie) avec comme référence France = 100 et calculons les indices élémentaires. Nous obtenons :

Pays	Revenu moyen	Indice base France = 100
DK - Danemark	25 113	135,9
FR - France	18 481	100,0
HU - Hongrie	4 377	23,7

$$I_{DK/FR}=\frac{25\,113}{18\,481}\cdot 100=135{,}9 \qquad I_{HU/FR}=\frac{4\,377}{18\,481}\cdot 100=23{,}7$$

Le revenu moyen des ménages danois est supérieur de 35,9 % à celui des ménages français. En revanche le revenu moyen des ménages hongrois représente à peine le quart de celui des ménages français .
En appliquant la règle de réversibilité, on s'autorise à comparer la base France aux autres individus comme suit :

$$I_{FR/DK}=\frac{1}{I_{DK/FR}}=\frac{1}{\frac{25\,113}{18\,481}}\cdot 100=73{,}6 \qquad I_{FR/HU}=\frac{1}{I_{HU/FR}}=\frac{1}{\frac{4\,377}{18\,481}}\cdot 100=422{,}2$$

Ainsi le revenu moyen des ménages français ne représente-t-il que 73,6 % de celui des ménages danois. Par contre, un ménage français a en moyen un revenu plus de 4 fois supérieur à celui d'un ménage hongrois.

<u>La transitivité :</u> un indice élémentaire est transitif tel que :

$$I_{2/0}=I_{2/1}\cdot I_{1/0}$$ Plus souvent utilisée sous la forme $$I_{2/1}=\frac{I_{2/0}}{I_{1/0}}$$

L'intérêt de la transitivité se manifeste lorsque, une situation de référence 0 ayant été choisie, on souhaite pouvoir comparer deux situations différentes de celle prise pour référence.

Exemple : Considérons la série indicée de la variation de la population française à quelques dates clé entre 1876 et 2007 avec 1936 = 100 :

Date	Indice
1876	91,7
1901	97,1
1921	93,6
1936	100,0
1946	96,6
1962	110,3
1982	129,6
1999	139,6
2007	148,2

Chaque période est indicée par rapport à la situation de référence 1936 = 100. Mais dès lors que l'on ne possède plus les chiffres initiaux de population comment faire pour comparer une période avec une autre sur la base des seuls indices et toujours en considérant la référence 1936 = 100 ?
L'indice de variation de population entre 1936 et 1962 est 110,3 signifiant qu'entre ces deux dates le nombre d'habitants a cru de 10,3 %. Entre 1936 et 2007, l'indice donne une valeur de 148,3 indiquant une progression démographique de 48,3 %. Qu'en est-il de l'indice de variation de population entre 1962 et 2007 avec 1936 = 100 ? En posant $I_{1/0}$ = $I_{1962/1932}$ et $I_{2/0}$ = $I_{2007/1932}$ et en utilisant la propriété de transitivité, on peut écrire :

$$I_{2/1}=\frac{I_{2/0}}{I_{1/0}}=I_{2007/1692}(1932=100)=\frac{I_{2007/1932}}{I_{1962/1932}}=\frac{148{,}2}{110{,}3}=1{,}344 \ \ soit \ \ 134{,}4$$

En base 1936 = 100, l'indice de variation de population entre 1962 et 2007 est 134,4 traduisant un accroissement de population 34,4 % entre ces deux dates.

<u>La multiplication :</u> troisième propriété fondamentale des indices élémentaires que l'on peut énoncer ainsi : si une grandeur g est le produit de deux grandeurs h et k , l'indice élémentaire de la grandeur g est le produit des indices des grandeurs h et k pour une même période :

$$Si \ g=h\times k \ \ alors$$

$$I_{1/0}(g)=I_{1/0}(h)\cdot I_{1/0}(k)$$

Cette dernière propriété trouve une application essentielle en économie où la valeur est toujours considérée comme le produit d'une prix et d'une quantité.
Exemple :
L'Adil de Syldavie ne dispense qu'un type seul de conseil juridique qu'elle facture à ses consultants au prix unitaire de 8,50 €. En 2007, elle a délivré 8 573 conseils réalisant ainsi un chiffre d'affaire pour l'année de 72 870,5 €. En 2008, la crise aidant, le conseil d'administration de l'Adil de Syldavie propose de baisser le prix de la consultation à 7,80 € afin de permettre à davantage de personnes de profiter de la qualité et de la compétence toutes deux incomparables de son service juridique. A la fin de l'année 2008, le nombre de consultations donné atteint 9 788 représentant un chiffre d'affaire de 76 346,4 €. La baisse du prix de la consultation (PU) combinée à une hausse du nombre de contacts (NC) se sont traduites par un indice de variation du chiffre d'affaire (CA) calculé comme suit et correspondant à une augmentation de 4,8 % :

$$I_{1/0}(CA)=I_{1/0}(PU)\times I_{1/0}(NC)=$$

$$I_{2008/2007}(CA)=I_{2008/2007}(PU)\times I_{2008/2007}(NC)=$$

$$\frac{7{,}80}{8{,}50}\times\frac{9\ 788}{8\ 573}=0{,}918\times 1{,}142=1{,}048 \ \ soit \ l'indice \ 104{,}8$$

Quelques remarques supplémentaires concernant les indices élémentaires
Le choix de la base est totalement arbitraire.
La base n'est pas nécessairement la valeur initiale.
Par ailleurs, il est possible de prendre comme base la moyenne ou la médiane de la distribution.
Le recours aux indices n'est qu'un moyen parmi d'autres pour interpréter une évolution. Ainsi, dans notre exemple, il est autorisé de dire qu'entre 1936 et 2007 la population française a progressé de 148,2-100 = 48,2 % et qu'entre 1982 et 2007 elle a augmenté de (148,2 – 129,6) = 18,6 % et de 56,9 % entre 1876 et 2007 (148,2 – 91,7).

Lorsque l'on passe d'un
Il convient de toujours garder à l'esprit qu'un indice est une valeur relative sans unité

5.2.2 Les indices synthétiques

Selon l'Insee, un indice synthétique se définit comme suit :
« Un indice synthétique mesure la variation de la valeur d'une grandeur complexe définie comme l'agrégation d'un ensemble de grandeurs élémentaires. Ainsi, par exemple, l'Indice des Prix à la Consommation (IPC) mesure par un indice unique la variation des prix de 1.000 variétés de produits. L'indice de la grandeur complexe est alors une *moyenne pondérée* des indices des grandeurs élémentaires ; les pondérations sont les "masses" des grandeurs élémentaires (dans le cas des indices des prix, ces masses sont les dépenses). L'indice de Laspeyres pondère par les masses de la période de base. L'indice de Paasche pondère par les masses de la période courante.

Prenons un exemple concret pour aider à la formalisation de la notion d'indice synthétique :

Nous disposons du prix pour cinq biens a, b, c, d et e et ce à la date 0 et à la date t. Considérons que ces biens sont des logements locatifs et que le prix correspond au loyer surfacique. Nous obtenons le tableau suivant :

		Dates	
		0	t
Logements	a	11,16	11,57
	b	9,18	9,67
	c	7,73	8,09
	d	6,44	6,81
	e	5,56	5,91

Pour chacun des logements il demeure toujours possible de calculer l'indice élémentaire d'évolution des loyers. Mais l'intérêt existe de vouloir connaître l'évolution globale des loyers prenant en compte l'ensemble des logements, autrement dit, l'indice synthétique d'évolution des loyers de plusieurs logements.

Une première façon de procéder consisterait à calculer les indices élémentaires pour chacun des logements et à en faire la moyenne arithmétique comme suit :

	0	t	Indice élémentaire $I_{t/0}$	Coefficient de pondération	Indice x coefficient
a	11,16	11,57	$Ia_{t/0}$ = 103,7	0,20	20,73
b	9,18	9,67	$Ib_{t/0}$ = 105,3	0,20	21,07
c	7,73	8,09	$Ic_{t/0}$ = 104,7	0,20	20,93
d	6,44	6,81	$Id_{t/0}$ = 105,7	0,20	21,15
e	5,56	5,91	$Ie_{t/0}$ = 106,3	0,20	21,26
					Indice synthétique = 105,1

L'indice synthétique des loyers est égale à la moyenne arithmétique des indices élémentaires calculés :

$$I_{t/0}=\frac{Ia_{t/0}+Ib_{t/0}+Ic_{t/0}+Id_{t/0}+Ie_{t/0}}{n=5}=(Ia_{t/0}\cdot 0{,}2)+(Ib_{t/0}\cdot 0{,}2)+(Ic_{t/0}\cdot 0{,}2)+(Id_{t/0}\cdot 0{,}2)+(Ie_{t/0}\cdot 0{,}2)$$

(rappelons que diviser par 5 revient à multiplier par 0,2)

Dans notre exemple, cela donne :

$$I_{t/0}=\frac{103{,}7+105{,}3+104{,}7+105{,}7+106{,}3}{5} \quad \Rightarrow$$

$$I_{t/0}=(103{,}7\times 0{,}2)+(105{,}3\times 0{,}2)+(104{,}7\times 0{,}2)+(105{,}7\times 0{,}2)+(106{,}3\times 0{,}2)=105{,}1$$

0,2 est ici un facteur ou un *coefficient de pondération*, c'est-à-dire une valeur qui vise, le cas échéant, à attribuer à chaque individu, à chaque logement un poids correspond à son importance au sein de l'ensemble des individus concernés par le calcul. Dans notre cas, il n'y a aucune hiérarchie entre individu et chaque logement a donc le même poids. Le total des poids étant par convention égal à 1, et le nombre de logements étant de 5, chaque logement dispose donc d'un cinquième du poids total, c'est-à-dire 0,2.

Considérant ce coefficient de pondération, nous pouvons reformuler notre indice synthétique de la façon suivante :

$$I_{t/0}=(\alpha_a\cdot Ia_{t/0})+(\alpha_b\cdot Ib_{t/0})+(\alpha_c\cdot Ic_{t/0})+(\alpha_d\cdot Id_{t/0})+(\alpha_e\cdot Ie_{t/0})$$

Où
- α_a Représente le poids affecté au logement *a* soit dans notre exemple 0,2
- α_b Représente le poids affecté au logement *b* soit dans notre exemple 0,2
- α_c Représente le poids affecté au logement *c* soit dans notre exemple 0,2
- α_d Représente le poids affecté au logement *d* soit dans notre exemple 0,2
- α_e Représente le poids affecté au logement *e* soit dans notre exemple 0,2

Avec toujours $\sum \alpha=1$ *soit dans notre cas* $\alpha_a+\alpha_b+\alpha_c+\alpha_d+\alpha_e=1$

Évidemment, une partie de l'intérêt de l'indice synthétique réside dans la très vraisemblable variation du coefficient de pondération en fonction des situations observées.

En reprenant l'exemple précédent, on peut supposer que a, *b, c, d* et *e* ne sont pas des logements locatifs mais plutôt des catégories de logements constitutives d'un parc locatifs à l'échelle d'une ville par exemple. Ainsi, on peut imaginer que

- *a* représente le parc des logements de 1 pièce
- *b* représente le parc des logements de 2 pièces
- *c* représente le parc des logements de 3 pièces
- *d* représente le parc des logements de 4 pièces
- *e* représente le parc des logements de 5 pièces et plus

et que le loyer mesuré aux dates o et à la date t pour chacun des parcs est un loyer moyen et que l'on cherche à connaître l'indice d'évolution des loyers pour l'ensemble du parc locatif. Dans ce cas de figure-ci, les coefficients de pondération ont toutes les chances de ne plus être égaux d'abord parce que l'on souhaite que le calcul de l'indice synthétique d'évolution des loyers tienne compte de la structure existante du parc locatif et que de la sorte le poids de chaque parc soit respecté.

La structure du parc locatif sur la ville étudiée est la suivante :

	Catégories	Répartition (%)	Coefficient de pondération (poids)
a	1 pc	25,6 %	0,256
b	2 pc	27,2 %	0,272
c	3 pc	22,7 %	0,227
d	4 pc	17,2 %	0,172
e	5 pc+	7,3 %	0,073
		100 %	1

Reprenons la formule de l'indice synthétique précédemment utilisée et attribuons à chaque indice élémentaire le constituant les coefficients de pondération ainsi déterminés. Nous obtenons la formule suivante :

$$I_{t/0} = (\alpha_a \cdot Ia_{t/0}) + (\alpha_b \cdot Ib_{t/0}) + (\alpha_c \cdot Ic_{t/0}) + (\alpha_d \cdot Id_{t/0}) + (\alpha_e \cdot Ie_{t/0}) \Rightarrow$$

$$I_{t/0} = (103{,}7 \times 0{,}256) + (105{,}3 \times 0{,}272) + (104{,}7 \times 0{,}227) + (105{,}7 \times 0{,}172) + (106{,}3 \times 0{,}073) = 104{,}9$$

Quelques exemples d'indices synthétiques « célèbres » : les indices d'évolution de la valeurs d'un panier de biens.

Considérons un panier de biens courants, celui que les média appelle habituellement « le panier de la ménagère ». La valeur de chaque bien est le produit d'un prix et d'une quantité achetée. Si la panier contient n produits, la valeur du panier au temps t s'écrit :

$$V_t = p_t^1 q_t^1 + p_t^b q_t^b + \ldots + p_t^n q_t^n = \sum_{i=1}^{n} p_t^i q_t^i$$

Où V_t Valeur du panier au temps t

p_t^i Prix du bien i au temps t

q_t^1 Quantité du bien i au temps t

Exemple :

Examinons le panier de Madame Duraton au temps t . Il contient quatre produits dont le prix unitaire et les quantités achetées figurent dans le tableau qui suit :

	Prix p_t^i	Quantité q_t^i
Produit 1	3,88	7
Produit 2	7,50	4
Produit 3	12,45	3
Produit 4	4,40	12

La valeur du panier de Mme. Duraton au temps *t* s'écrit :

$$V_t = \sum_{i=1}^{4} p_t^i q_t^i = p_t^1 q_t^1 + p_t^2 q_t^2 + p_t^3 q_t^3 + p_t^4 q_t^4 \Rightarrow$$

$$V_t = (3{,}88 \times 7) + (7{,}50 \times 4) + (12{,}45 \times 3) + (4{,}40 \times 12) = 147{,}31$$

La partie intéressante du problème consiste à mesurer l'évolution de la valeur du panier de Mme. Duraton entre les deux dates 0 et t, sachant que cette évolution dépendra de l'évolution combinée de deux paramètres : le prix et la quantité de chaque bien. Toute la difficulté consiste à construire un indice synthétique capable de prendre en compte ces évolutions parallèles et combinées pour en déduire une évolution globale. Trois économistes, LASPEYRES, PAASCHE et FISHER, ont proposé des indices synthétiques différents pour mesurer l'évolution des composants prix et quantité au sein de la valeur du panier, le plus utilisé, en tous les cas en France, étant celui de Laspeyres. C'est celui que nous exposerons ici.

L'indice de Laspeyres :

L'indice de Laspeyres permet de mesurer deux évolutions : l'évolution des prix des biens composant le panier – on parlera alors d'indice d'évolution des prix de Laspeyres – et l'évolution des quantités des biens composant ce même panier – on parlera alors d'indice d'évolution des quantités de Laspeyres.

L'indice d'évolution des prix de Laspeyres

Cet indice mesure l'évolution des prix des biens composant un panier entre deux dates 0 et t en prenant comme référence la valeur du panier au temps initial $t = 0$ et en supposant que les quantités des biens du panier n'ont pas varié entre les deux dates. L'indice d'évolution des prix s'écrit alors :

$$L_{t/0}^{p} = \frac{V_t}{V_0} = \frac{\sum_{i=1}^{n} p_t^i q_t^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100$$

Où V_t Valeur du panier au temps *t*

V_0 Valeur du panier au temps initial *0*

p_0^i Prix du bien i au temps *0*

q_0^i Quantité du bien i au temps *0*

p_t^i Prix du bien i au temps *t*

q_t^i Quantité du bien i au temps *t*

Avec $q_t^i = q_0^i$ Puisque les quantités sont fixes

Exemple : reprenons le panier de Mme. Duraton à deux dates différentes :

	Date 0		Date t	
	p_0^i	q_0^i	p_t^i	q_t^i
Produit 1	3,88	7	4,13	8
Produit 2	7,50	4	8,42	3
Produit 3	12,45	3	11,71	5
Produit 4	4,40	12	4,89	9

Les prix ont évolué mais aussi les quantités. Or l'indice des prix de Laspeyres suppose que les quantités restent inchangées. Le calcul se fera donc à quantités égales avec comme référence les quantités au temps initial. On aura donc :

$$L_{t/0}^{p} = \frac{V_t}{V_0} = \frac{\sum_{i=1}^{4} p_t^i q_t^i}{\sum_{i=1}^{4} p_0^i q_0^i} \times 100 = \frac{p_t^1 q_t^1 + p_t^2 q_t^2 + p_t^3 q_t^3 + p_t^4 q_t^4}{p_0^1 q_0^1 + p_0^2 q_0^2 + p_0^3 q_0^3 + p_0^4 q_0^4} \times 100 \Rightarrow$$

$$L_{t/0}^{p} = \frac{p_t^1 q_0^1 + p_t^2 q_0^2 + p_t^3 q_0^3 + p_t^4 q_0^4}{p_0^1 q_0^1 + p_0^2 q_0^2 + p_0^3 q_0^3 + p_0^4 q_0^4} \times 100 = \frac{(4{,}13 \times 7) + (8{,}42 \times 4) + (11{,}71 \times 3) + (4{,}89 \times 12)}{(3{,}88 \times 7) + (7{,}50 \times 4) + (12{,}45 \times 3) + (4{,}40 \times 12)} \Rightarrow$$

$$L_{t/0}^{p} = \frac{156{,}4}{147{,}3} \times 100 = 106{,}2$$

Soit une progression de la valeur du panier de Mme. Duraton de 6,2 % selon les prix entre 0 et t.

L'indice d'évolution des quantité de Laspeyres

Cet indice mesure l'évolution des quantités des biens composant un panier entre deux dates 0 et t en prenant comme référence la valeur du panier au temps initial $t = 0$ et en supposant que les prix des biens du panier n'ont pas changé entre les deux dates. L'indice d'évolution des quantités s'écrit alors :

$$L_{t/0}^{q} = \frac{\sum_{i=1}^{n} p_t^i q_t^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100 = \frac{\sum_{i=1}^{n} p_0^i q_t^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100$$

Où

p_0^i Prix du bien i au temps 0

q_0^i Quantité du bien i au temps 0

p_t^i Prix du bien i au temps t

q_t^i Quantité du bien i au temps t

Avec $p_t^i = p_0^i$ Puisque les quantités sont fixes

Exemple : toujours le panier de Mme. Duraton à deux dates différentes :

	Date 0		Date t	
	p_0^i	q_0^1	p_t^i	q_t^1
Produit 1	3,88	7	4,13	8
Produit 2	7,50	4	8,42	3
Produit 3	12,45	3	11,71	5
Produit 4	4,40	12	4,89	9

Les quantités ont évolué mais aussi les prix. Or l'indice des quantités de Laspeyres suppose que les prix restent inchangées. Le calcul se fera donc à prix égaux avec comme référence les prix au temps initial. On aura donc :

$$L_{t/0}^q = \frac{V_t}{V_0} = \frac{\sum_{i=1}^{4} p_t^i q_t^i}{\sum_{i=1}^{4} p_0^i q_0^i} \times 100 = \frac{p_t^1 q_t^1 + p_t^2 q_t^2 + p_t^3 q_t^3 + p_t^4 q_t^4}{p_0^1 q_0^1 + p_0^2 q_0^2 + p_0^3 q_0^3 + p_0^4 q_0^4} \times 100 \Rightarrow$$

$$L_{t/0}^p = \frac{p_0^1 q_t^1 + p_0^2 q_t^2 + p_0^3 q_t^3 + p_0^4 q_t^4}{p_0^1 q_0^1 + p_0^2 q_0^2 + p_0^3 q_0^3 + p_0^4 q_0^4} \times 100 = \frac{(3{,}88 \times 8) + (7{,}50 \times 3) + (12{,}45 \times 5) + (4{,}40 \times 9)}{(3{,}88 \times 7) + (7{,}50 \times 4) + (12{,}45 \times 3) + (4{,}40 \times 12)} \times 100 \Rightarrow$$

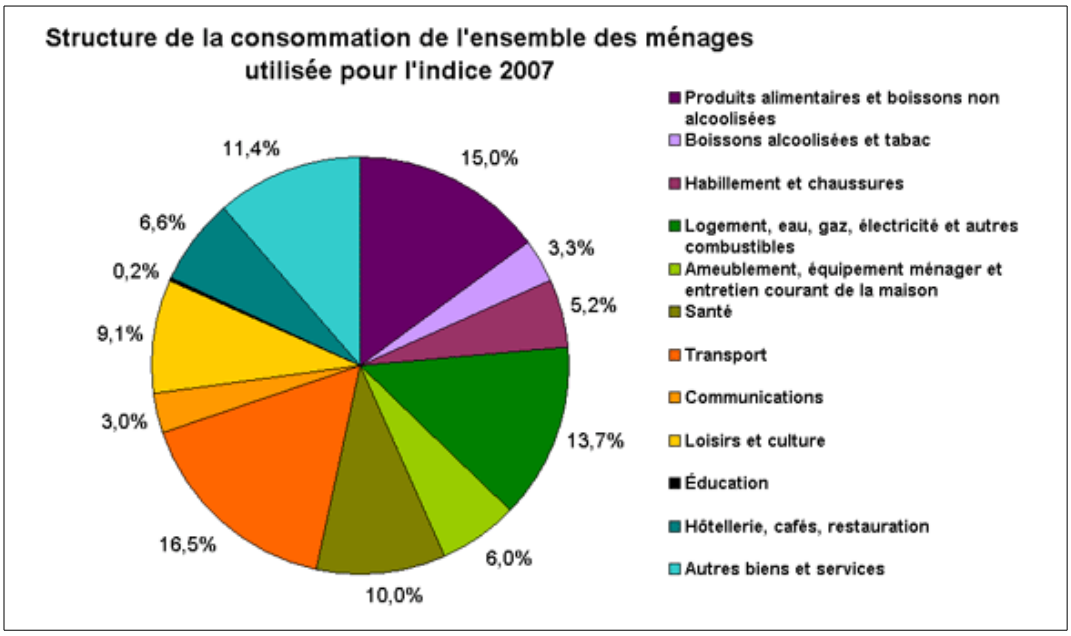
$$L_{t/0}^q = \frac{155{,}4}{147{,}3} \times 100 = 105{,}5$$

Soit une progression de la valeur du panier de Mme. Duraton de 5,5 % selon les quantités entre 0 et t.

L'indice des prix de l'INSEE

Cher au cœur – ou plutôt au porte-monnaie - des français, cet indice est sans doute le plus connu dans notre pays (peut-être juste derrière le CAC40 en ce moment). L'IPC – ou Indice des Prix à la Consommation – est calculé et publié mensuellement par l'INSEE. Il permet, sur la base d'un panel de produits et services consommés régulièrement et massivement par les ménages français, d'évaluer l'évolution des prix entre deux périodes. C'est une mesure synthétique d'évolution de prix à quantité constante.

Le calcul de l'IPC de l'INSEE utilise la formule de l'indice des prix Laspeyres qu'elle applique à un échantillon de quelques 21 000 indices élémentaires eux-mêmes calculés sur la base d'une collecte nationale de prix de produits dans 106 agglomérations de plus de 2 000 habitants réparties sur l'ensemble du territoire. LIPC couvre plus 1 000 variétés de biens et services regroupés en 161 catégories. La liste des biens et services enquêtés demeurent confidentielle afin d'éviter tout tentative éventuelle de manipulation des prix par les commerçants. Actuellement, la période de référence pour le calcul de l'IPC est 1998 = 100. Le graphique qui suit donne la structure du « panier » de l'IPC de l'INSEE et par conséquent les pondérations appliquées lors du calcul :



Source : http://www.insee.fr/fr/themes/indicateur.asp?id=29&type=1&page=info_ipc.htm#q2

Publié dans la première quinzaine de chaque mois et portant sur l'évolution des prix du mois précédent, l'IPC fait à chaque fois l'objet de commentaires et de débats passionnés tant sur sa valeur – il sert en effet de témoins à de nombreux paramètres économiques (inflation, revalorisation des pensions et du SMIC, etc.)- que sur sa composition et son mode calcul que certains jugent ne plus être en « phase » avec les vrais habitudes de consommation des français, notamment depuis 2006 avec la flambée des coût de l'énergie, des transport et du logement. Le tableau qui suit fait le point sur la valeur récente de l'indice général puis décliné par poste (source: http://www.insee.fr/fr/themes/indicateur.asp?type=1&id=29).

	février 2008	janvier 2009	février 2009	evol. sur 1 mois	evol. sur 1 an
Indice des prix à la consommation, IPC (base 100 en 1998)					
Ensemble des ménages, France entière (métropole et DOM)					
Ensemble (00 E)	117,81	118,39	118,84	0,4	0,9
Ensemble cvs (00 C)	118,31	119,11	119,41	0,3	0,9
Alimentation (4000 E)	121,64	124,30	124,36	0,0	2,2
Tabac (0221 E)	190,02	191,09	191,15	0,0	0,6
Produits manufacturés (4003 E)	99,97	99,76	100,09	0,3	0,1
Énergie (4007 E)	149,24	135,42	136,16	0,5	-8,8
Services (4009 E)	122,24	124,86	125,52	0,5	2,7
Alimentation y c. Tabac (4014 E)	127,92	130,52	130,58	0,0	2,1
Manufacturés y c. Energie (4015 E)	108,42	106,21	106,61	0,4	-1,7
Manufacturés hors Habillement et chaussures (4016 E)	100,26	99,93	100,10	0,2	-0,2
Ensemble hors loyers et hors tabac (5000 E)	116,11	116,59	117,06	0,4	0,8
Ensemble hors énergie (4017 E)	115,47	116,99	117,42	0,4	1,7
Ensemble hors tabac (4018 E)	116,57	117,13	117,59	0,4	0,9
Ménages urbains dont le chef est ouvrier ou employé, France entière (métropole et DOM)					
Ensemble hors Tabac (4018 D)	116,47	117,18	117,61	0,4	1,0
Ensemble (00 D)	118,30	119,02	119,45	0,4	1,0
Inflation sous-jacente					
Ensemble des ménages, France métropolitaine					
Ensemble «sous jacent» (4022 S)	114,47	116,25	116,59	0,3	1,9
Indice des prix à la consommation harmonisé de la France, IPCH (base 100 en 2005)					
Ensemble des ménages, France entière (métropole et DOM)					
Ensemble IPCH (00 H)	105,48	106,05	106,49	0,4	1,0

Exercice 20 : fichier Excel associé « Exercice 20 - Indices.xls ».

Chapitre 6

6. Relation entre deux variables : tendance, ajustement linéaire (ou régression linéaire) et corrélation

Dans les chapitres précédents nous avons énuméré et décrits les principaux outils statistiques à disposition pour caractériser et résumer des distributions de valeurs. Comment, en limitant au maximum la perte d'informations, passer d'un volume important de données difficilement manipulable à quelques indicateurs pertinents synthétisant l'allure et le contenu de la distribution de la population étudiée.

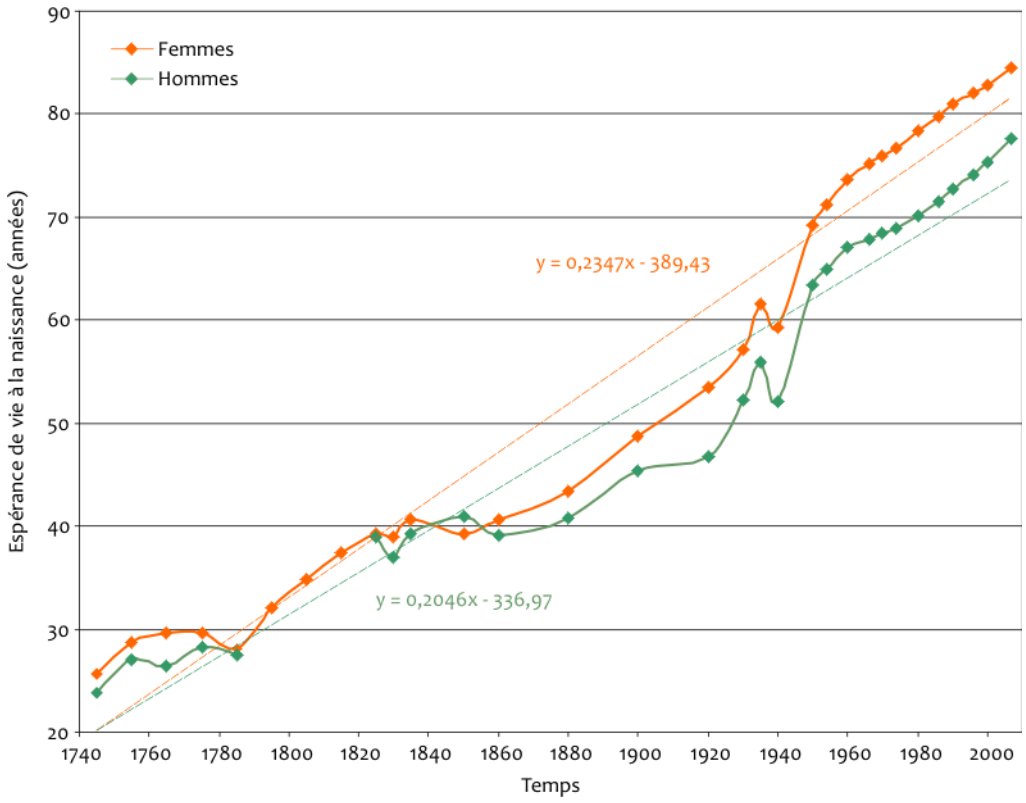
Le présent chapitre s'intéresse à un autre aspect de l'analyse statistique, celui qui décrit et étudie la relation pouvant exister entre deux variables. Il est en effet fréquent, lorsque l'on étudie et analyse un phénomène quel qu'il soit, que plusieurs variables ou facteurs entrent en ligne pour sa compréhension, variables qui de part la relation de cause à effet qu'elles entretiennent sont à même d'une part de permettre de mieux comprendre le phénomène et d'autre part d'en dégager tendance et projection compte tenu de la situation existante.

Les outils statistiques décrits ci-après ont un double objectif :

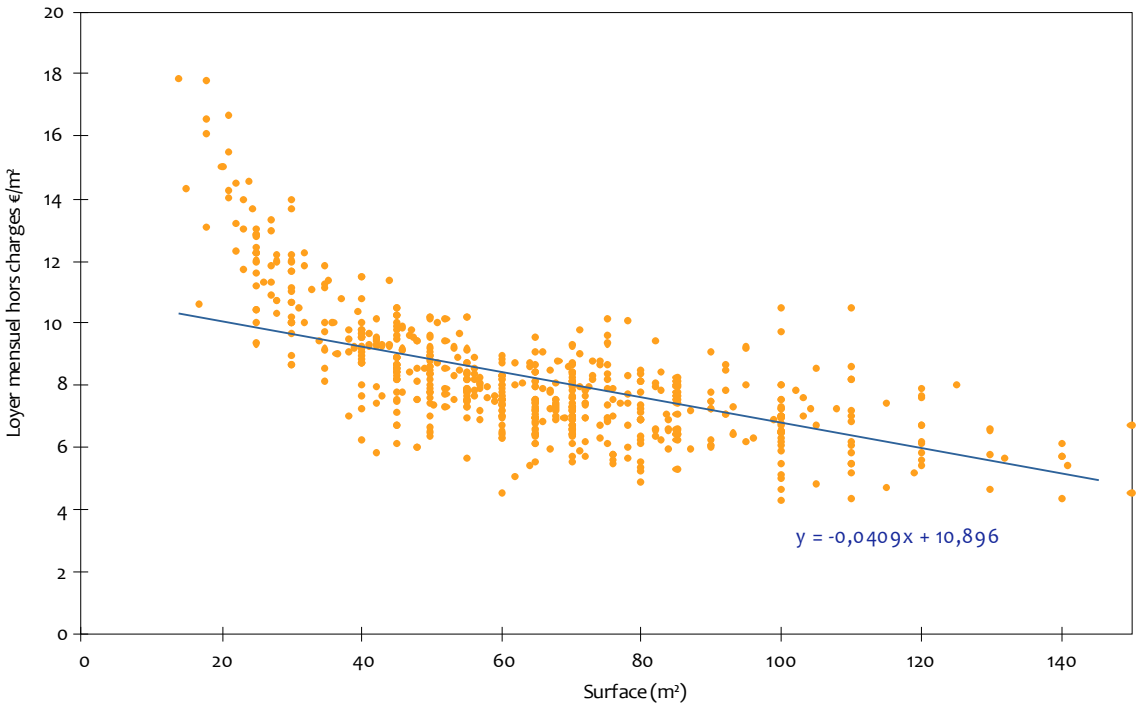
- D'une part résumer mathématiquement (par une équation) et graphiquement (par une droite) la relation pouvant exister entre deux variables : lorsqu'il s'agira d'une variable confrontée au temps ou à l'espace dans le cadre d'une relation non immuable , on parlera de tendance. Par contre lorsqu'il s'agira de deux variables entretenant une relation de cause à effet « immuable » on parlera davantage d'ajustement ou de régression.

- D'autre part qualifier et mesurer l'intensité de la relation (ou degré de liaison) entre les variables étudiées via des outils appropriés comme par exemple le coefficient de corrélation ou les test du Khi² lorsque l'on aura affaire à des variables non quantitatives.

Exemple de tendance linéaire : Évolution de l'espérance de vie à la naissance pour les femmes et les hommes en France de 1750 à nos jours et droites de tendance associées.



Exemple d'ajustement linéaire : Relation entre loyers mensuels moyens au m² hors charges et surface des logements dans le parc locatif privé du territoire de Belfort en 2007 avec la droite d'ajustement (ou droite de régression) associée.



6.1 Ajustement et régression linéaire

6.1.1 Énoncé et principes de la droite de tendance et de la droite d'ajustement (ou droite de régression) :

Droite de tendance et droite de régression matérialise la relation linéaire entre respectivement une variable et le temps ou entre deux variables. Cette matérialisation est double : d'abord sous une forme mathématique par l'intermédiaire d'une équation, puis sous une forme graphique puisque l'équation déterminée peut être figurée sur le graphique original.

L'équation de la droite est trouvée à partir des valeurs existantes de la distribution étudiée. Sa forme mathématique est relativement simple et s'écrit comme suit :

$$y = ax + b$$

Où y = Variable dépendante ou variable expliquée

x = Variable indépendante ou variable explicative

a = Pente de la droite de régression

b = Ordonnée à l'origine de la droite de régression

Cette relation suppose que y est une fonction de x, c'est-à-dire que la valeur de y dépend de celle de x, ou bien encore que la valeur de y est expliquée par la valeur de x. Ainsi, la façon dont évoluera la valeur de y dépendra de manière plus ou moins forte de la façon dont évoluera celle de x. x est alors appelée variable explicative, sa variation expliquant tout ou partie de la variation de y, elle-même appelée variable expliquée.

Exemple : reprenons les deux exemples brièvement évoqués précédemment.

- Dans le premier cas – tendance linéaire – l'évolution de l'espérance de vie dépend du temps (et non l'inverse). En effet, au fur et à mesure que l'on avance dans le temps, l'espérance de vie croît. La variable dépendante ou expliquée est ici « l'espérance de vie en année » et la variable indépendante ou explicative est « le temps », mais l'on sait pertinemment que ce n'est pas le temps qui explique l'accroissement l'espérance de vie mais davantage les progrès de l'alimentation et de la médecine qui eux s'améliore avec le temps. Le temps explique donc de façon indirecte l'augmentation de l'espérance de vie. Néanmoins, nous dirons que l'espérance vie est une fonction du temps et écrirons : Espérance de vie = f (temps)

  y = Variable expliquée = espérance de vie en années

  x = Variable explicative = temps

- Dans le deuxième exemple – ajustement ou régression linéaire – il semble raisonnable de supposer que ce sont les loyers qui dépendent de la surface du logement et non l'inverse. Les loyers représentent donc la variable à expliquer y et la surface la variable explicative x de telle sorte que l'on puisse dire que « les loyers sont une fonction de la surface » et écrire loyer = f (surface) :

y = Variable expliquée = loyer mensuel hors charges au m²

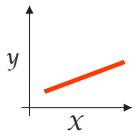
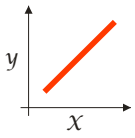
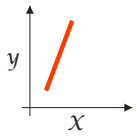
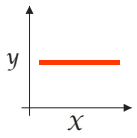
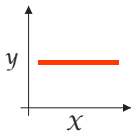
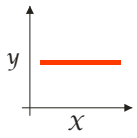
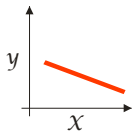
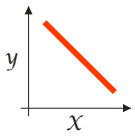
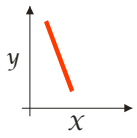
x = Variable explicative = surface des logements

Il nous reste à décrire les deux dernières composantes de l'équation de la droite de tendance et de la droite de régression à savoir a et b :

- a représente la pente de la droite (coefficient directeur en mathématique), c'est-à-dire son inclinaison ou, en d'autres termes, l'ampleur de la variation de la variable expliquée y quand la variable explicative x varie de une unité. Plus la variation de y sera importante pour la variation d'une unité de x, plus la pente sera importante. A l'opposé, moins la variation de y sera importante pour la variation d'une unité de x, moins la pente sera importante.

  Le sens de la pente, donc le signe du paramètre a, renseigne quant à lui sur le type de liaison qui unie les deux variables :

  - Si a est négatif, cela signifie que lorsque x augmente, y diminue. On parle alors de relation inversement proportionnelle;
  - Si a est nulle, cela signifie que lorsque x augmente, y demeure constant, ne varie pas;
  - Enfin si a est positif, cela signifie que lorsque x augmente, y augmente également (mais pas nécessairement dans les mêmes proportions). On parle alors de relation proportionnelle.

	$\Delta y = 2\Delta x$	$\Delta y = \Delta x$	$\Delta y = \frac{1}{2}\Delta x$
$a > 0$			
$a = 0$ y constant			
$a < 0$			

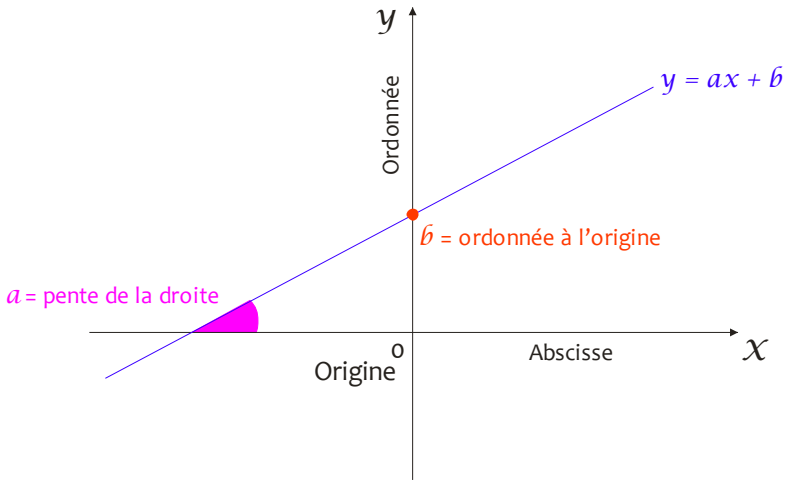
A titre d'exemple, $\Delta y = 2\Delta x$ signifie que lorsque x croît de 2 unités, y augmente d'une unité

- b représente l'ordonnée à l'origine, c'est-à-dire l'endroit où la droite de régression (ou de tendance) coupe l'axe des ordonnées (ou axe des y). b peut positif, négatif ou nul. Si $b = 0$ cela signifie que la droite de régression passe par l'origine. Ce peut être le cas quelle que soit la valeur de a. Lorsque la cas se produit, l'équation de la droite d'ajustement devient :

$$y = ax$$

On parle alors de fonction linéaire. Qui plus est, si $\Delta y = \Delta x$ alors l'équation s'écrit $y = x$.

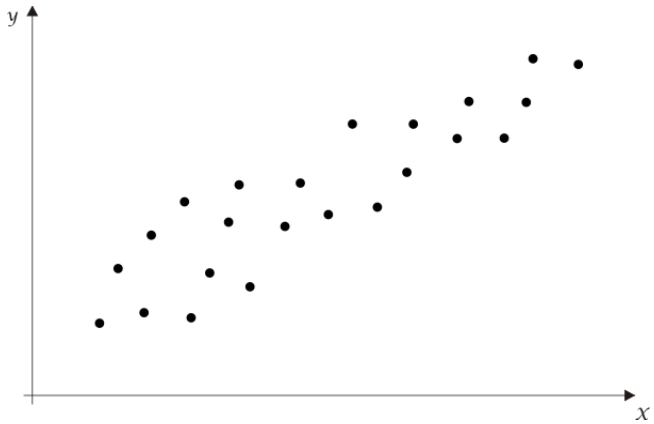
Récapitulatif : la droite de régression et ses composantes



6.1.2 Détermination des paramètres de la droite d'ajustement: la méthode des Moindres Carrés Ordinaires (MCO)

Équation et tracé de la droite d'ajustement pour une nuage de points donné nécessitent la découverte des deux paramètres fondamentaux que sont d'une part la pente a et d'autre part l'ordonnée à l'origine b. Rappelons que la droite telle qu'elle doit être mise en équation et tracée a pour objectif premier de résumer un nuage de points, c'est-à-dire la relation entre deux variables et ce, de façon qualitative (allure) et quantitative (intensité). Cette représentativité impose que la droite passe impérativement au plus près de tous les points du nuage. Cet ajustement, car il s'agit bien d'un ajustement, a logiquement donné son nom à la droite (droite d'ajustement). Il est réalisé par l'intermédiaire de la méthode dite des Moindres Carrés Ordinaires (MCO). La dénomination quelque peu rébarbative de la méthode découle directement de son principe : en effet l'ajustement s'effectue en minimisant la somme du carré des écarts entre la droite et les observations. En clair, cela signifie bien que la droite va passer au plus près de tous les points. Sur le plan graphique, le principe de la méthode MCO s'illustre de la façon suivante :

Soit un nuage de points matérialisant la relation entre la variable x et la variable y, les points représentant les observations :



Le tracé de la droite d'ajustement ($y = ax + b$) permet de comprendre la logique qui préside à sa construction. Il faut garder à l'esprit que cette droite est une représentation synthétique du nuage de points et qu'elle permet pour chaque x de calculer un nouvel y estimé à partir de son équation noté $\hat{y}$.

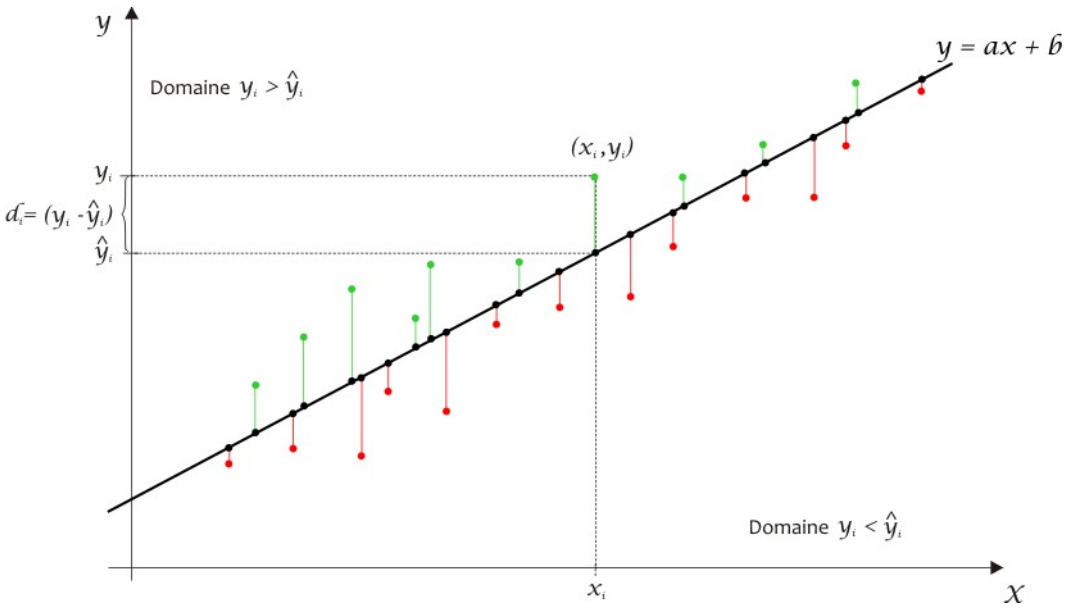
Le nuage initial de points se trouve maintenant distribué de part et d'autre de la droite et chaque observation se trouve dorénavant à une certaine distance de ladite droite. Lorsque que les observations se situent au-dessus de la droite d'ajustement (points verts), la distance est positive. Lorsque les observations se trouvent sous la droite d'ajustement (points rouges), la distance est négative. Ces distances correspondant aux écarts évoqués lors de la tentative de définition de la méthode des Moindres Carrées Ordinaires qu'il faudra, une fois élevés au carré, minimiser. Une distance, ou un écart, correspond à la différence entre la valeur observée de y et la valeur estimée de y soit :

$$(y_i - \hat{y}_i)$$

Rappelons que $\hat{y}_i$ représente la valeur estimée de y_i par l'équation de la droite d'ajustement , c'est-à-dire la valeur prise par la valeur observée y_i lorsqu'on projette sur la droite.

La distance ou l'écart ainsi déterminé est également appelé *résidu* en référence Plus le résidu est faible, plus la valeur observée est proche de la droite d'ajustement. De même, un résidu positif signifie que la valeur observée y_i est plus grande que sa valeur estimée $\hat{y}_i$ par l'équation. Par conséquent, un résidu négatif signifie que la valeur observée y_i est plus petite que sa valeur estimée $\hat{y}_i$ par l'équation. Si la valeur observée y_i est égale à la valeur estimée $\hat{y}_i$ alors le résidu est nul et la valeur observée se trouve exactement sur le droite d'ajustement.

Valeurs	Résidu
$y_i > \hat{y}_i$	$(y_i - \hat{y}_i) > 0$
$y_i < \hat{y}_i$	$(y_i - \hat{y}_i) < 0$
$y_i = \hat{y}_i$	$(y_i - \hat{y}_i) = 0$



Si les choses ont été faites dans les règles, c'est-à-dire si la droite passe bien au plus près de tous les points du nuage, alors la somme des résidus doit être égale à 0, la somme des résidus négatifs compensant exactement la somme des résidus positifs. Nous avons donc :

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

Le but étant à termes de minimiser les somme des résidus élevés au carré

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{minimum}$$

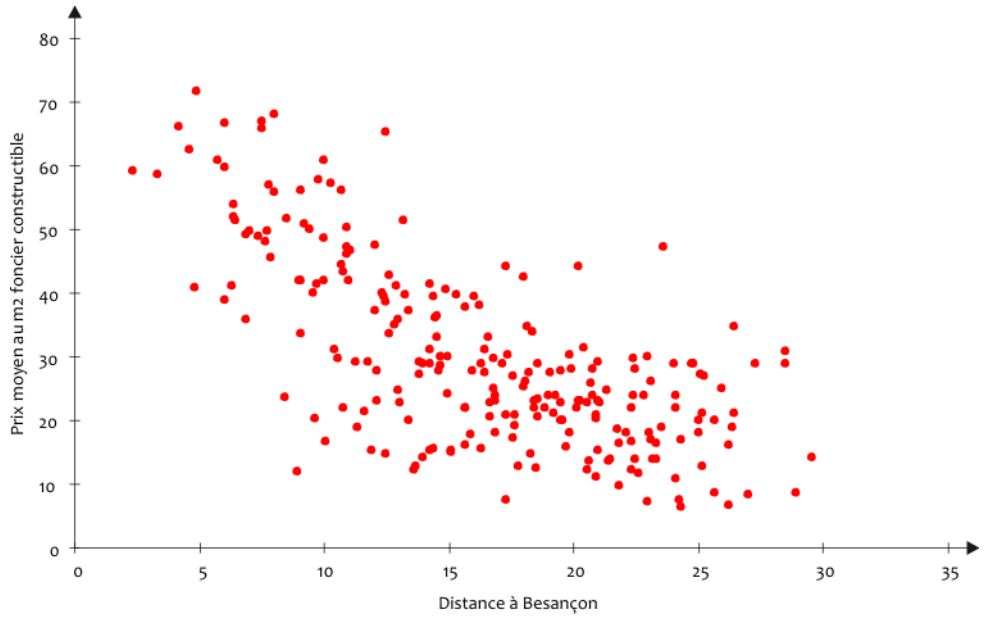
Pour atteindre cet objectif et respecter ainsi le postulat de départ, il nous faut trouver les paramètres a et b pour formaliser l'équation de la droite d'ajustement. Pour éviter de ce perdre dans des développements mathématiques inutiles à ce stades, on donnera ici les recettes permettant de déterminer directement a et b.

La pente de la droite d'ajustement : $$a = \frac{Cov(x,y)}{Var(x)} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

L'ordonnée à l'origine : $$b = \bar{y} - a\,\bar{x}$$

Exemple : on s'intéresse au prix moyen du foncier constructible et à la façon dont celui-ci varie au fur et mesure que l'on s'éloigne d'un pôle d'emplois et de services. On suppose logiquement que la valeur du foncier décroît en fonction de l'éloignement au pôle, autrement dit que le prix du foncier est inversement proportionnel à la distance au pôle. Ce postulat considère donc le prix du foncier constructible comme une fonction de la distance au pôle: prix du foncier constructible = f (distance au pôle).

Sur le terrain, on relève le prix moyen du foncier constructible par commune ainsi que la distance routière entre chacune de ces communes et la commune-pôle et on confronte les deux variables sur un même obtenant ainsi un nuage de points où chaque point représente une observation, c'est-à-dire une commune :



L'axe des x (abscisse) figure la variable explicative, en l'occurrence la distance au pôle, alors que l'axe des y correspond à la variable expliquée ou dépendante, le prix moyen au m² du foncier constructible. D'évidence, il existe bien une relation inversement proportionnelle entre les deux variables : plus la distance est grande, moins le prix du foncier semble élevé. La droite d'ajustement aura donc nécessairement une pente négative ($a < o$) .

Le calcul des paramètre de la droite de régression nous donne :

Pour la pente :

$$a = \frac{Cov(x,y)}{Var(x)} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{-61{,}19}{36{,}71} = -1{,}667$$

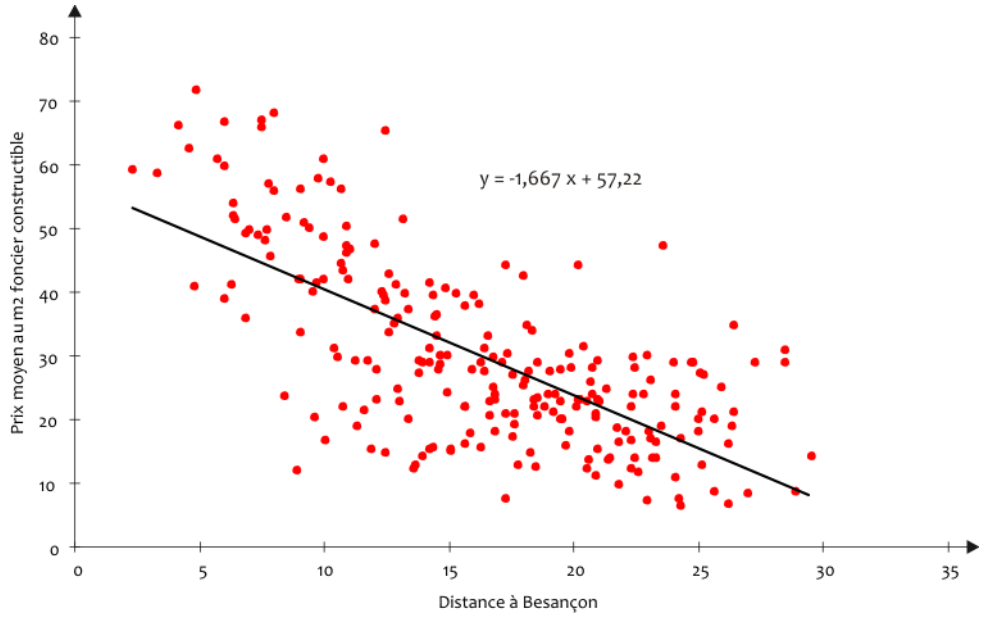
Pour l'ordonnée à l'origine:

$$b = \bar{y} - a\bar{x} = 29{,}99 - 1{,}667(16{,}34) = 57{,}22$$

Soit une droite d'ajustement d'équation :

$$y = -1{,}667\,x + 57{,}22 \Leftrightarrow prix\ foncier = -1{,}667(distance) + 57{,}22$$

Graphiquement, on obtient :



On vérifie assez aisément que le postulat de départ est vérifié, à savoir que la somme des résidus est égale à 0 :

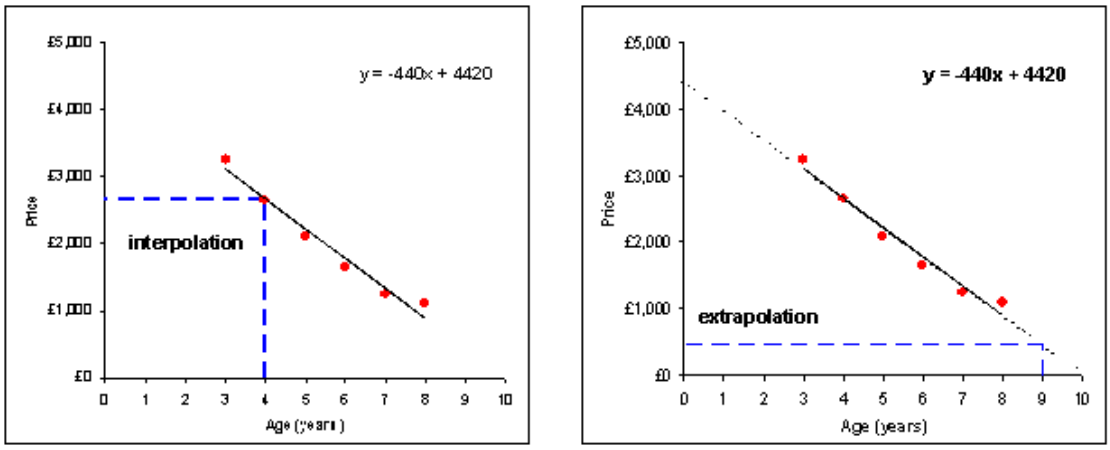
$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$$

Notre droite d'ajustement passe au plus près de tous les points du nuage.

6.1.3 Utilisations et limites de la régression linéaire:

Malgré sa simplicité apparente, le régression demeure une méthode puissante. Elle peut, dans certains cas et avec prudence, aider à la reconstitution de séries caractérisées par des lacunes. On parle alors d'interpolation. Elle permet également l'extrapolation, c'est-à-dire l'estimation de la variable y pour des valeurs de x qui sont en

dehors du champ de celles observées. Dans les deux cas, le fait de disposer d'une équation, qui n'est autre chose qu'un modèle mathématique, autorise la production de n'importe *y* pour n'importe quel *x*.

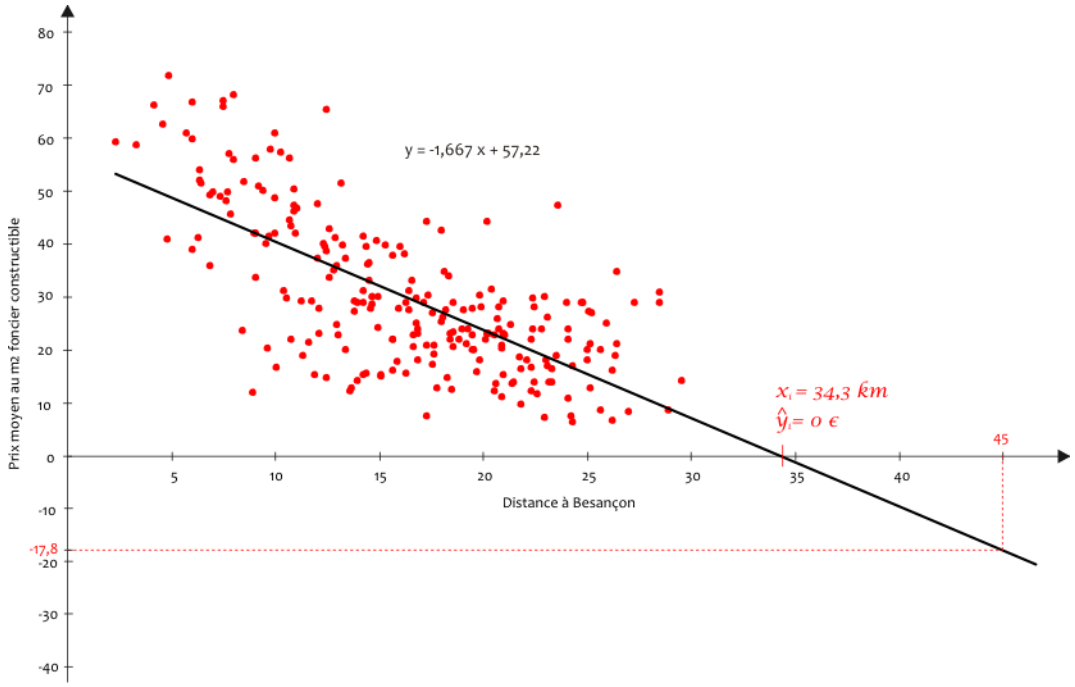


Interpolation Extrapolation

Source : http://www.coventry.ac.uk/ec/~nhunt/regress/pred1.html

Exemple : en reprenant l'exemple précédent, il est possible de calculer la valeur du foncier pour une distance au pôle de 30 km, 35 km, 40, 50 et même de 100 km:

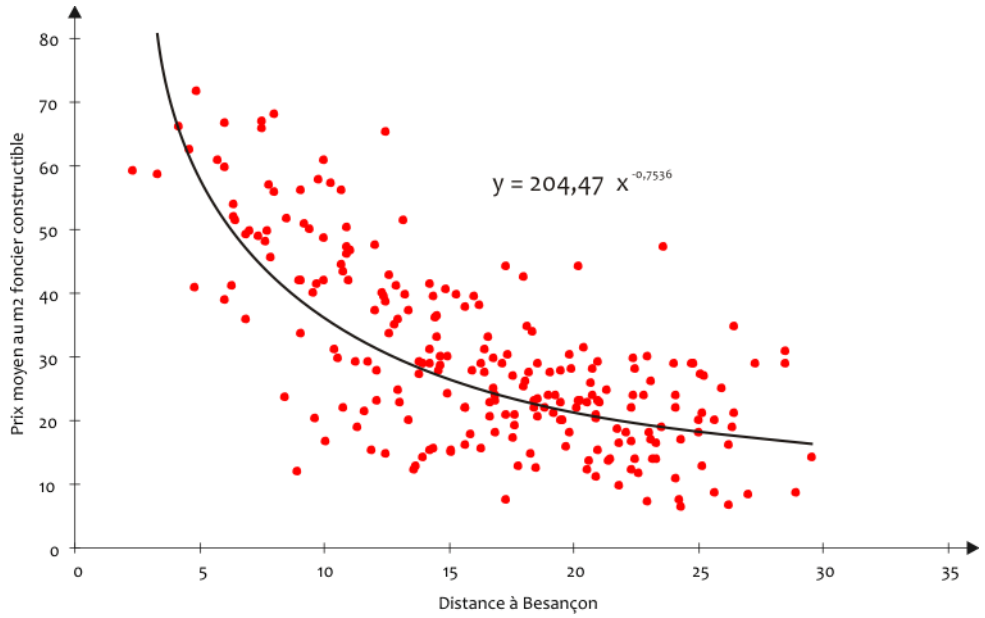
Distance (km)	$\hat{y}_i$ (€/m²)
30	7,2
35	-1,1
40	-9,5
45	- 17,8
50	-26,1
100	-109,5



Les résultats obtenus sont révélateurs de la limite du modèle de régression linéaire car à l'en croire, il suffirait de s'éloigner au-delà de 34,3 km pour que le prix du foncier constructible devienne nul puis négatif !

Un modèle plus juste devrait proposer une limite asymptotique, c'est-à-dire un prix qui demeure quasi constant, sans être nul ou négatif, à partir d'une certaine distance. Ces modèles existent mais ils ont la particularité de ne pas être linéaires: on parle alors de modèles polynomiaux, exponentiels ou bien encore logarithmiques.

Dans notre, il est possible d'affiner l'ajustement et par là même les estimations par le biais d'un modèle type « puissance »



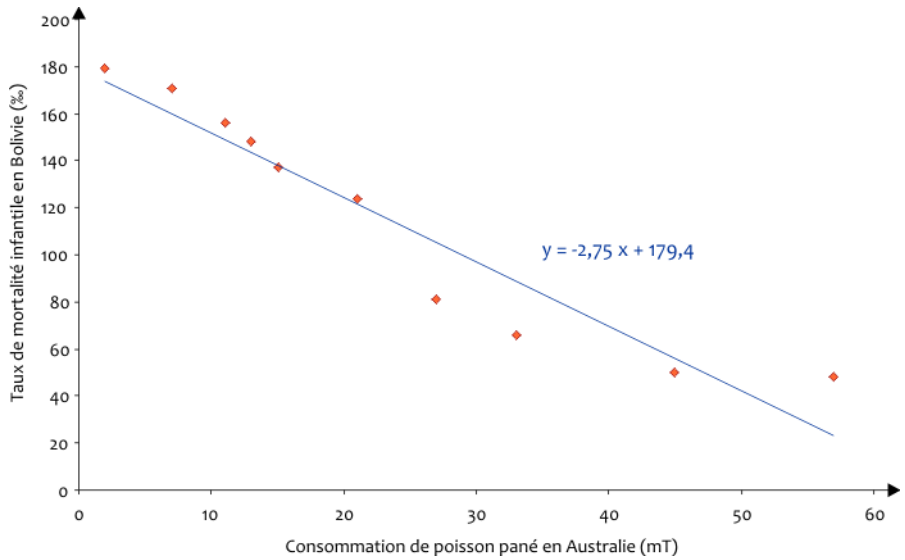
Comparons les estimations du modèle « puissance » à celles du modèle linéaire dans les mêmes conditions :

Distance (km)	$\hat{y}_i$ (€/m²) modèle linéaire	$\hat{y}_i$ (€/m²) modèle puissance
30	7,22	15,76
35	-1,12	14,03
40	-9,46	12,69
50	-26,13	10,72
100	-109,48	6,36

Il semblerait que les estimations obtenues soient plus « en phase » avec une supposée réalité. Cependant, rien ne nous garanti, dans le cadre d'une extrapolation, que les résultats fournis soient représentatif d'une quelconque réalité. En effet, rien n'interdit de penser qu'à partir d'une certaine distance les prix du foncier renouent avec la hausse du fait de l'influence d'un autre pôle d'emplois et de services.

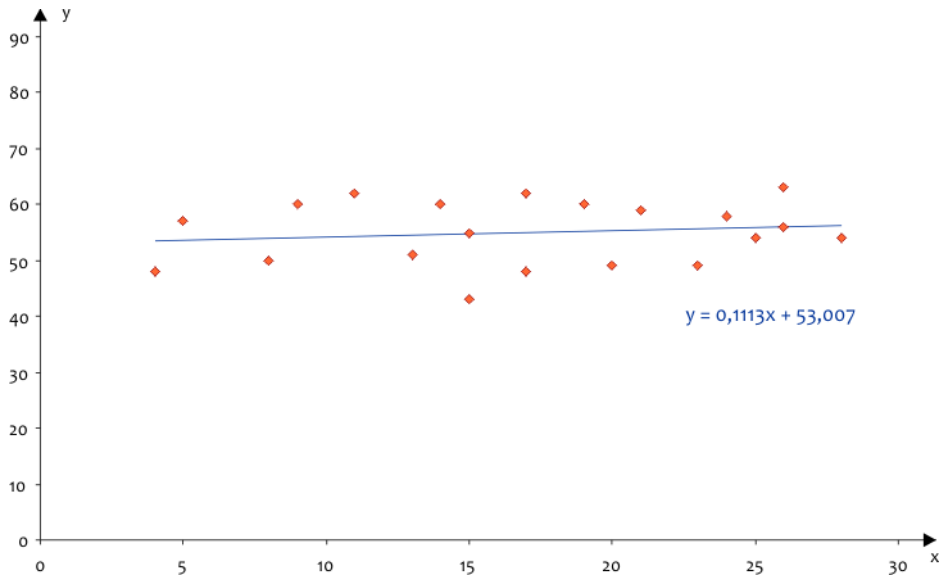
Le fait de pouvoir formaliser, sous forme d'une équation, une relation mathématique et statistique entre deux variables ne signifie pas nécessairement qu'il y ait, dans la réalité, une relation effective de cause à effet entre ces mêmes deux variables. Les exemples sont légion où, pour montrer les travers de la modélisation par régression, on décrit une relation entre variables qui dans la réalité n'a aucune chance ou raison de se réaliser.

Exemple : la consommation annuelle de poisson pané en Australie (x) et le taux de mortalité infantile (enfants de moins de 5 ans) en Bolivie (y) entre 1960 et 2005. A première vue, la relation entre les deux variable semble évidente. Elle est peut être mise en équation au même titre que n'importe quelle autre relation via le modèle de régression linéaire. Seulement elle n'existe tout simplement pas: il n'y a en effet aucune chance, et aucune raison, pour que l'augmentation de la consommation de poisson pané en Australie ait une quelconque influence sur le taux de mortalité infantile en Bolivie.



La formalisation de la relation par l'intermédiaire d'une équation ne renseigne pas non plus sur la qualité et l'intensité d'une supposée liaison entre deux variables. On peut en effet déterminer une équation matérialisant une relation que l'on juge probable et logique et qui cependant n'est pas ou peu marquée ou qui n'est pas systématique dans la réalité.

L'exemple qui suit montre, de façon certes caricaturale, le fait qu'une relation qui n'existe pas ou peu entre deux variables peut malgré tout être formalisée par une équation. Dans cet exemple, la variation de x n'a pour ainsi dire aucune conséquence sur la variation de y.



Nous sommes donc en présence de deux problèmes :

- d'une part un problème lié à l'identification de l'existence ou non d'une relation de cause à effet entre deux variables que l'on souhaite confronter : la plupart du temps, la solution de ce problème réside dans le bon sens (confrontation de variables dont on suppose qu'elles entretiennent un lien réel et logique) et/ou dans l'expérimentation (vérification d lien supposé par des méthodes statistiques);

- d'autre part, un problème lié à la mesure de la qualité et de l'intensité de la liaison entre deux variables. Pour ce faire, la statistique a développé des outils capables d'évaluer la qualité d'une liaison entre variables : le coefficient de corrélation et le coefficient de détermination.

6.2 Mesure de la qualité et de l'intensité d'une liaison entre deux variables : coefficient de corrélation et coefficient de détermination

Ces deux coefficients qualifient et mesurent la force de la relation mathématique et statistique entre deux variables. Pour les mêmes raisons que la droite de régression, l'obtention de coefficients jugés bons ne signifie pas l'existence d'une relation réelle entre les variables.
Coefficient de corrélation et coefficient de détermination sont intimement liés, le second n'étant ni plus ni moins que la carré du premier. Notés respectivement r et r^2, on écrit :

$$r = \frac{Cov(x,y)}{\sqrt{Var(x)}\cdot\sqrt{Var(y)}} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Cette formule, en apparence compliquée reprend, pour son calcul, beaucoup de paramètres normalement déjà connus lorsque les distributions des deux variables ont été étudiées et la droite d'ajustement déterminée.

6.2.1 Propriétés du coefficient de corrélation et du coefficient de détermination

Le coefficient de corrélation est toujours compris entre -1 et 1 : $-1 \leqslant r \leqslant 1$
Sa valeur mesure la force de la liaison tandis que son signe renseigne sur le sens de la corrélation :

- Lorsque r = 1 , la relation entre les variables x et y est proportionnelle et parfaite
- Lorsque r = -1, la relation entre les variables x et y est inversement proportionnelle et parfaite
- Lorsque r = 0, la relation entre les variables x et y est statistiquement inexistante

Entre ces bornes, tous les cas de figures sont possibles avec des degré de liaison variables. La forme du nuage de points permet déjà de se faire une idée quelques fois assez juste de la nature et de la force de liaison entre deux variables ainsi que l'illustrent les figures ci-après.

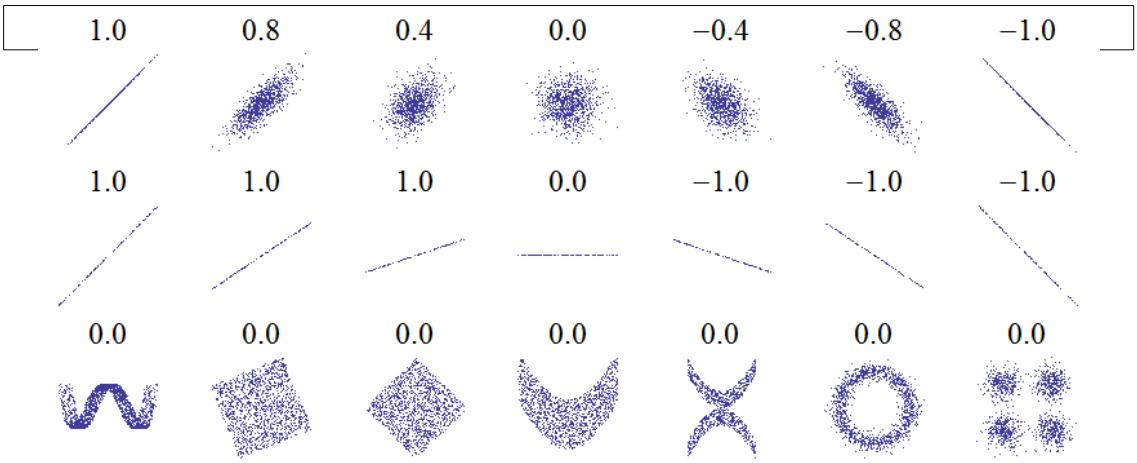
On considère que deux variables sont suffisamment liées pour pratiquer des interpolation et extrapolation lorsque $r \leqslant -0{,}75$ ou quand $r \geqslant +0{,}75$. En dehors de ces limites, la liaison se dégrade rapidement pour devenir insignifiante.

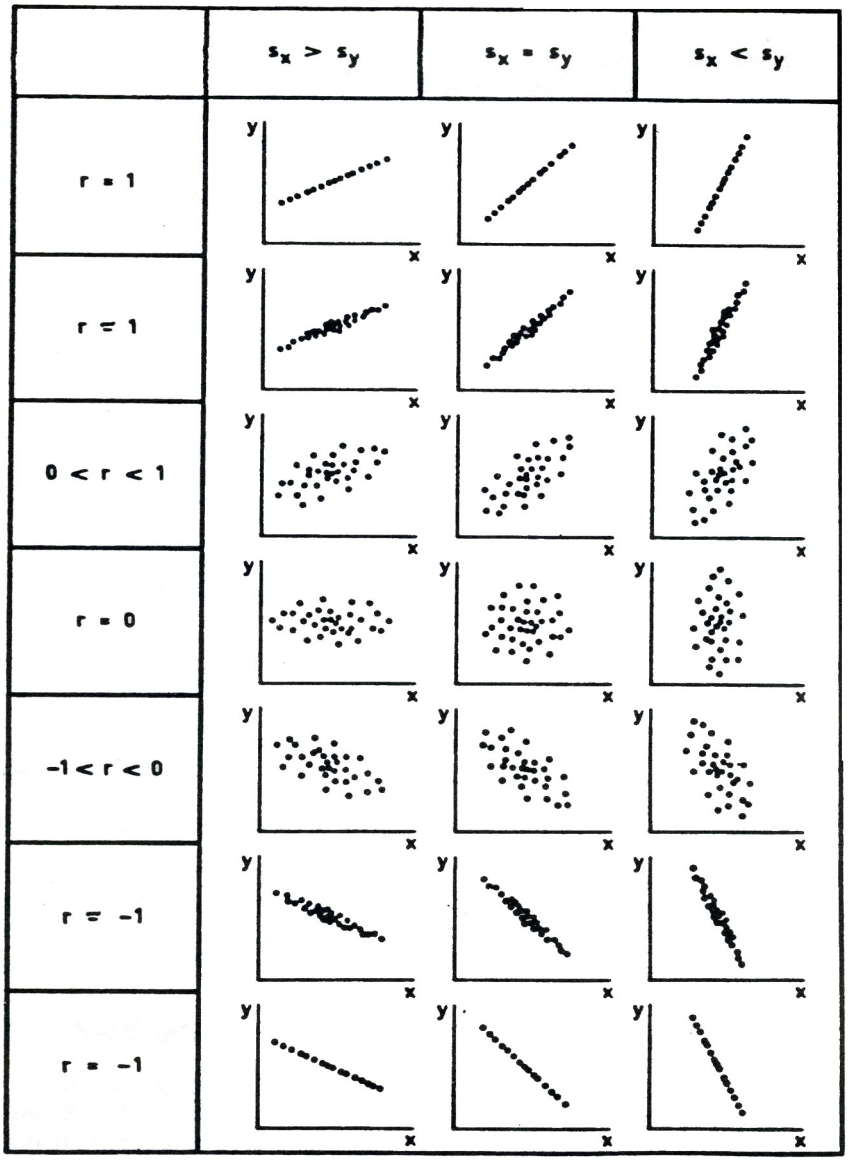
Mais attention, une corrélation significative ne démontre pas l'existence d'une relation systématique et réelle de causalité entre deux variables x et y, pas plus que l'absence d'une corrélation significative dans une seule étude ne démontre l'absence de lien causal.

Le coefficient de détermination r^2 renseigne sur la force de la liaison statistique entre deux variables. Il diffère assez peu du coefficient de corrélation r dont il est le carré. Pour cette raison, il est toujours de signe positif et n'informe donc pas sur le sens de la relation. C'est pourquoi on lui préfère le coefficient de corrélation .

Le coefficient de corrélation r , tout comme le coefficient de détermination r^2 , est un indicateur sans dimension aucune.

Il existe des tests (t**est de signification du r de Pearson par exemple)** permettant de juger objectivement de la signification statistique d'un coefficient de corrélation calculé et par là même de la signification de la liaison entre les variables étudiées.





Quelques formes typiques de nuages de points en relation avec les valeurs du coefficient de corrélation pour la régression linéaire (S_x et S_y représentent respectivement l'écart-type de la variable x et l'écart-type de la variable y)– Source : Guide pratique d'analyse des données p. 68, Crauser, Harvatopoulos et Sarnin, 1989.

Exemple:

Reprenons l'exercice qui avait consisté à confronter la variation du prix moyen du foncier constructible par commune en fonction de la distance des communes à un pôle d'emplois. L'hypothèse de départ avait supposé qu'une relation de cause à effet existait entre ces deux variables et qu'elle était inversement proportionnelle. En d'autres termes, le prix moyen du foncier constructible pour un ensemble de communes périphériques était inversement proportionnel à la distance entre ces communes et le pôle d'emploi.

Cette hypothèse avait été en partie confortée par la figuration du graphique de dispersion (nuage de points) dont l'allure montrait clairement une relation pouvant aller dans le sens des soupçons avancés. Partant de là, l'équation de la droite de régression avait été calculée permettant tout aussi bien, le croyait-on, inférence, interpolation et extrapolation. Les quelques « tests » réalisés sur la base du modèle linéaire nous avaient cependant interpelé sur la fragilité du modèle pour ce cas et il avait été suggéré du coup l'emploi d'un modèle non-linéaire plus approprié.

Mais absolument rien ne nous avait renseigné sur la qualité et l'intensité de cette supposée liaison. L'utilisation du coefficient de corrélation r doit nous permettre d'appréhender la qualité de la liaison. Son calcul nous donne le résultat suivant :

$$r=\frac{Cov(x,y)}{\sqrt{Var(x)}\cdot\sqrt{Var(y)}}=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{-61{,}19}{\sqrt{36{,}71}\cdot\sqrt{212{,}04}}=-0{,}695$$

et $r^2=0{,}48$

Il s'agit d'un score modeste. La relation existe bel et bien mais n'est pas convaincante sur le plan statistique. Plusieurs raisons à cela :

Si, pour le calcul des prix moyen on s'est basé sur un échantillon de communes et/ou de terrains, le biais introduit par la démarche peut affecter la distribution en favorisant quelque peu la dispersion : l'échantillon n'est pas forcément représentatif de la réalité et du coup la relation supposée entre les variables s'en trouve perturbée.

Le relation entre les deux variables existe mais n'est de type linéaire. Autrement dit, le prix moyen du foncier ne décroît pas linéairement au fur et à mesure que la distance augmente. L'ajustement trouve ses limites, comme déjà démontré et la force de la liaison statistique en est affectée, diminuée. Il faut trouver un autre modèle non linéaire plus adapté au phénomène.

L'exercice n'ayant été réalisé que sur année, il se peut que cette année ne soit pas représentative d'une tendance sur une longue période: des données « extraordinaires » pour cette année ont pu « polluer » la distribution et du même coup accroître la dispersion et fausser l'ajustement linéaire.

Il se peut également que le phénomène souffre d'anisotropie, c'est-à-dire qu'il présente des variations différentes selon les directions de l'espace : dans ce cas, la distance kilométrique n'est pas suffisante pour expliquer la décroissance des prix. La distance temps, variable pour une même distance kilométrique d'un point à un autre compte tenu de la qualité des axes de circulation et de l'intensité du trafic, serait peut-être plus appropriée pour expliquer la diminution des prix du foncier.

L'ensemble de ces remarques peuvent se combiner pour expliquer le score modeste obtenu.
Réitérons le calcul des coefficients r et r^2 en utilisant cette fois-ci le modèle de régression « puissance » et voyons ce que nous obtenons :

$$r=-0{,}66 \qquad\qquad r^2=0{,}44$$

Les résultats ne sont pas meilleurs, ils sont même moins bons qu'avec le modèle linéaire. Le problème ne réside donc pas dans le choix du modèle mais probablement davantage dans la construction de la distribution de la population ou de l'échantillon. Le modèle non linéaire (puissance) n'ajuste pas mieux le nuage de points que le

modèle linéaire. Pour ce qui est des interpolations, on lui préférera donc le modèle linéaire. Par contre il produit des extrapolations (ou des prédictions « meilleures » ou moins incohérentes que le modèle linéaire.

6.2.2 Erreur standard ou erreur-type de prédiction

On a vu comment le modèle de régression était à même, dans certaines conditions, d'autoriser l'interpolation comme la prédiction (extrapolation). On a également vu de quelle façon l'on pouvait caractériser et mesurer l'intensité de la relation entre deux variables. Les exemples exposés ont cependant montré de façon éclatante les pièges et dangers de la méthode et insisté sur les précautions à prendre afin au mieux de les éviter, au pire de réduire les risques d'erreur.
Quoiqu'il en soit, la prédiction, même avec les meilleurs modèles d'ajustement, demeure inévitablement entachée d'une certaine erreur que l'on peut tenter de mesurer par l'intermédiaire d'un paramètre nommé *erreur standard de prédiction* ou *erreur-type* notée $ES\hat{y}$. Ce paramètre peut-être considéré et interprété comme l'écart-type de la distribution (théorique) de toutes les erreurs qui seraient commises en effectuant la prédiction pour un grand nombre d'individus (distribution supposée normale et de moyenne nulle la plupart du temps).

L'erreur standard de la prédiction peut être estimée en appliquant la formule:

$$ES\,\hat{y} = \sigma_y \sqrt{1 - r^2}$$

Avec :

σ_y	Écart-type de la variable y
r^2	Coefficient de détermination

$ES\hat{y}$ renvoie l'erreur-type de la valeur y prévue pour chaque x de la régression. L'erreur type est une mesure du degré d'erreur dans la prévision de y à partir d'une valeur individuelle x. L'**erreur-type** est l'écart-type estimé de l'erreur de cette estimation. C'est donc une estimation de l'écart-type entre les valeurs mesurées ou estimées (d'une distribution d'échantillonnage) et les vraies valeurs.

Dans la formule ci-dessus, on note que:

- Si r = 1 ou si r = -1, $Es\hat{y} = 0$
- Si r = 0, $Es\hat{y} = \sigma_y$. La marge d'erreur est aussi importante que le dispersion de la distribution de y. La prédiction est, dans ce cas, prohibée.

Par conséquent, et afin de minimiser l'erreur-type synonyme d'un modèle de prédiction « fiable », on devrait toujours exiger de ce dernier qu'il produise un coefficient de corrélation au moins ≥ à 0,75 ou au moins ≤ à -0,75.

Exercice 21 : fichier Excel associé « Exercice 21 - Ajustement et corrélation.xls ».

Annexes

Annexe 1 : Précision et explication sur une notation spécifique en statistique : somme et produit

SOMME en statistique s'écrit avec le symbole Σ (sigma majuscule). Elle a la même signification qu'en mathématique : c'est une addition de termes.

Mais comme souvent en statistique, on est amené à additionner des séries relativement longues de valeurs (il n'est pas rare d'avoir à additionner 1 000 voir 10 000 valeurs). Plutôt que d'écrire les 1000 ou 10 000 valeurs les unes à la suite des autres séparées par un signe « + », il a été développé une notation synthétique ayant la même signification et produisant le même résultat.

Considérons une population composée de n = 10 individus. Chaque individu a, au sein de cette population, une place, un nom, un identifiant : il y a l'individu n°1, l'individu n°2, l'individu n°3, … jusqu'à l'individu n°10. On a vu que la notation standard pour les individus statistiques est i . On peut donc écrire que pour l'individu n°1, $i = 1$, que pour l'individu n°2, $i = 2$ et ainsi de suite jusqu'à $i = n = 10$.

A chaque individu i correspond également une valeur de la variable étudiée x. D'une façon générale, on a donc pour l'individu i la valeur de la variable x_i . On aura pour l'individu $i = 1$ la valeur x_1 , pour l'individu $i = 2$ la valeur x_2 et ainsi de suite jusqu'à $i = n = 10$ avec pour valeur x_{10} .

Si je veux sommer les valeurs des 10 individus composant ma population, je dois écrire :

$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$ ou plus simplement $\sum_{i=1}^{n} x_i = \sum_{i=1}^{10} x_i$ qui se dit :

« somme des x_i pour i = 1 jusqu'à 10 ». Je somme donc les valeurs de la variable x pour les 10 individus. Si notre population avait été composéé de 1 388 individus dont nous aurions souhaité faire la somme des valeurs pour la variable y, nous aurions écrit :

$$\sum_{i=1}^{1388} y_i$$

Dans les cas présentés, la somme s'est effectuée du 1er au dernier individu (de 1 à n). Mais elle peut très bien être sélective et se faire à n'importe où dans une population, comme par exemple du 21 ième individu au 133 ième. Au quel cas on écrira pour une variable x :

$$\sum_{i=21}^{133} x_i$$

La même logique s'applique à la notion de PRODUIT, notée Π (Pi majuscule) qui n'est autre chose que la multiplication de termes.
Ainsi, à partir du même exemple que précédemment, plutôt que d'écrire :

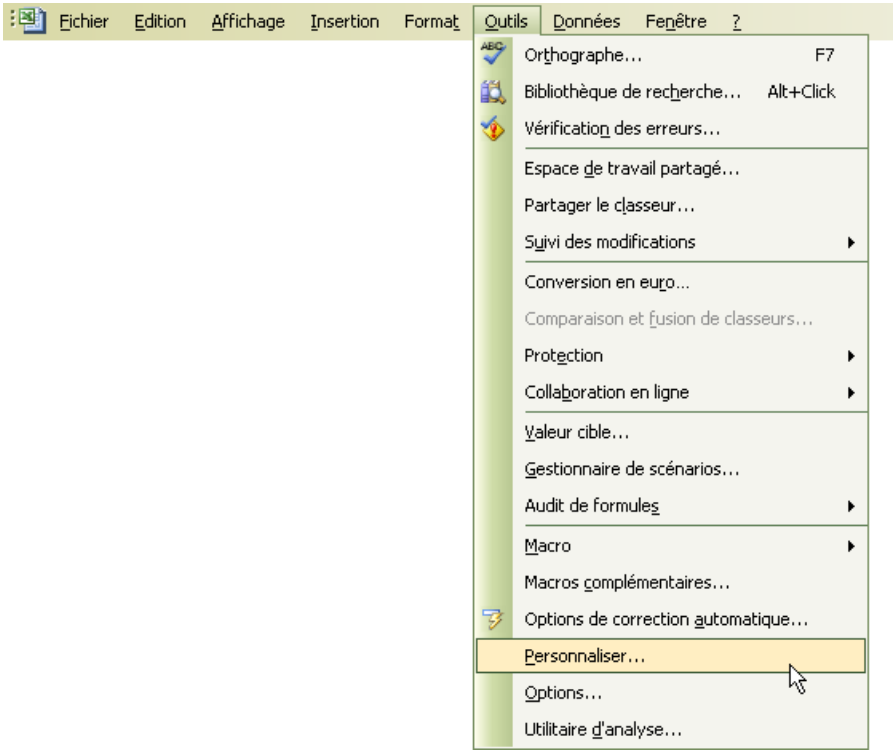
$x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot x_5 \cdot x_6 \cdot x_7 \cdot x_8 \cdot x_9 \cdot x_{10}$ on notera : $\prod_{i=1}^{n} x_i = \prod_{i=1}^{10} x_i$

Annexe 2 : Liste (non exhaustives) des fonctions Excel utiles en statistiques descriptive

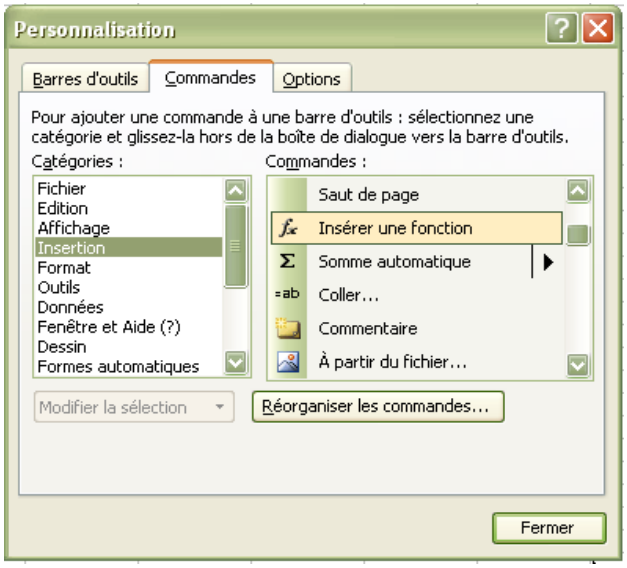
Pour accéder et insérer une fonction Excel à une feuille de calcul, il existe deux possibilités :

- Soit passer par le menu Insertion > Fonction
- Soit directement via l'icône  dans la barre d'outils. Si l'icône est absente de la barre d'outils, procéder comme suit pour l'y faire apparaître:

Dans la barre de menu choisir Outils > Personnaliser .



Sélectionner alors l'onglet Commande et dans la liste « Catégories » l'option « Insertion ». S'affiche alors en face les fonctionnalités disponibles et l'icône affectée. Choisir « Insérer une fonction » et faire glisser l'icône jusqu'à l'endroit de la barre d'outils où l'on souhaite la voir figurer définitivement.
Une fois l'opération terminée, fermer la fenêtre « Personnalisation ».



Les fonctions intéressantes en statistique :

Fonction (dénomination française)	Résultat
ABS	Valeur absolue d'une nombre
ARRONDI	Renvoie l'arrondi d'un nombre
CENTILE	Renvoie le k-ième centile d'une distribution
COEFFICIENT.CORRELATION	Renvoie le coefficient de corrélations d'une relation statistique entre deux variables
COEFFICIENT.DETERMINATION	Renvoie le coefficient de détermination d'une relation statistique entre deux variables
COVARIANCE	Calcule la covariance d'une relation statistique entre deux variables
CNUM	Transforme une chaîne de caractère représentant un nombre en un nombre
DROITEREG	Renvoie les paramètres de l'équation de la droite de régression (a et b)
ECARTYPE	Calcule l'écart-type d'une distribution
ERREUR.TYPE.XY	Renvoie l'erreur type (ou erreur standard) de prédiction d'un modèle de régression
FREQUENCE	Calcule la fréquence à laquelle des valeurs apparaissent dans une plage de valeurs
MAX	Renvoie le maximum d'une série de nombre
MEDIANE	Calcule la médiane d'une distribution
MIN	Renvoie le minimum d'une série de nombre
MODE	Calcule le mode d'une distribution
MOYENNE	Calcule la moyenne arithmétique d'une distribution
MOYENNE.GEOMETRIQUE	Calcule la moyenne géométrique d'une distribution
NB	Détermine le nombre de cellules contenant des nombres et les nombres compris dans la liste des arguments.
NB.SI	Détermine le nombre de cellules non vides d'une série répondant à la condition
NBVAL	Compte le nombre de cellules qui ne sont pas vides et les valeurs comprises dans la liste des arguments.
PLAFOND	Arrondi selon la précision demandée
ORDONNEE.ORIGINE	Calcule l'ordonnée à l'origine (b) d'une droite d'ajustement
PENTE	Renvois la pente (a) de la droite de régression
PRODUIT	Calcule le produit de plusieurs nombres
QUARTILE	Calcule le quartile 1, 2 ou 3 d'une distribution
RACINE	Renvoie la racine carré d'un nombre
SOMME	Calcule la somme de plusieurs nombres
SOMME.CARRES	Calcule la somme des carrés d'une série de nombre
SOMME.SI	Additionne des nombre si la condition est respectée
TENDANCE	Calcule les valeurs par rapport à une tendance linéaire.
VAR	Estime la variance sur le base d'un échantillon
VAR.P	Calcule la variance d'une population

Annexe 3 : Activer la macro « histogramme » dans Excel

Source : http://support.microsoft.com/kb/214269/fr

Cet article décrit étape par étape comment créer un histogramme avec un graphique à partir d'un ensemble de données d'exemple. L'utilitaire d'analyse compris dans Microsoft Excel inclut un outil Histogramme.

Vérifier l'installation de l'Utilitaire d'analyse

Avant d'utiliser l'outil Histogramme, vous devez vous assurer que le complément Utilitaire d'analyse est installé. Pour vérifier que l'Utilitaire d'analyse est installé, procédez comme suit :

1. Dans Microsoft Office Excel 2003 et dans les versions antérieures d'Excel, cliquez sur **Macros complémentaires** dans le menu **Outils**.

   Dans Microsoft Office Excel 2007, procédez comme suit :

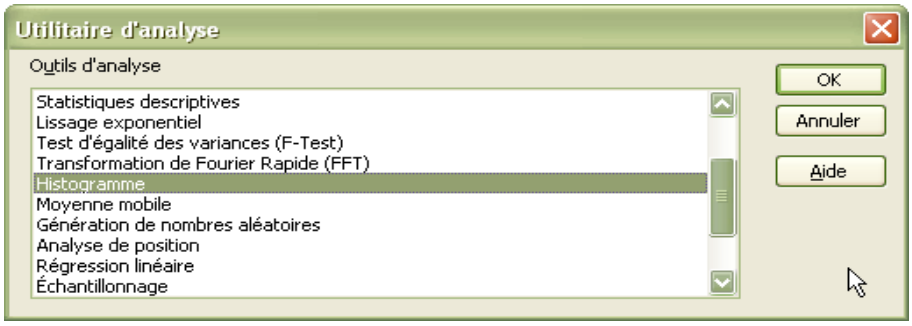
   1. Cliquez sur le **Bouton Microsoft Office**, puis sur **Options Excel**.
   2. Cliquez sur la catégorie **Compléments**.
   3. Dans la liste **Gérer**, sélectionnez **Compléments Excel**, puis cliquez sur **Rechercher**.
2. Dans la boîte de dialogue **Compléments**, assurez-vous que la case à cocher **Utilitaire d'analyse** est activée sous **Compléments disponibles**. Cliquez sur **OK**.

REMARQUE : pour que l'Utilitaire d'analyse s'affiche dans la boîte de dialogue **Compléments**, il doit être installé sur votre ordinateur. Si vous ne voyez pas l'**Utilitaire d'analyse** dans la boîte de dialogue **Compléments**, exécutez le programme d'installation de Microsoft Excel et ajoutez ce composant à la liste d'éléments installés.

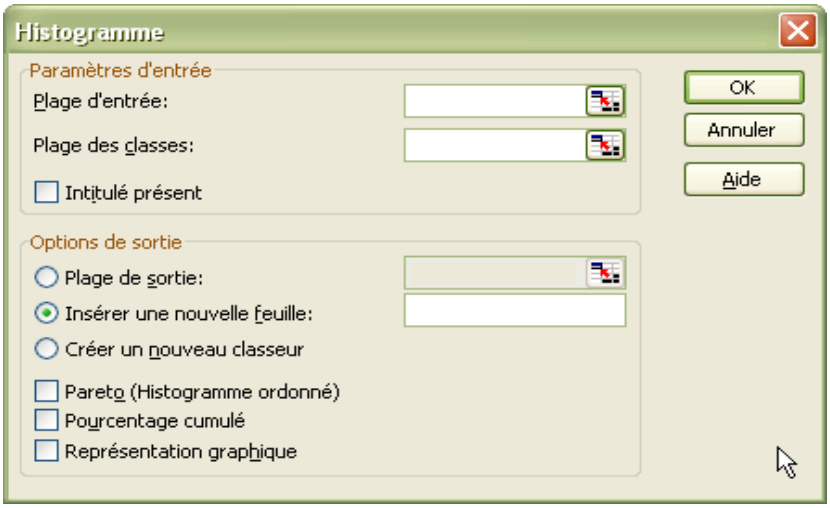
Créer un histogramme

Dans le menu principal, choisir « Outils » puis « Utilitaire d'analyse » comme suit (Dans Excel 2007, cliquez sur **Analyse des données** dans le groupe **Analyse** sous l'onglet **Données)** :

La liste suivante apparaît. Choisir l'option « Histogramme » puis cliquer « OK »



S'affiche alors la fenêtre paramétrique suivante qu'il convient de compléter en s'aidant eventuellement de l'aide proposée :



REMARQUE : vous ne serez pas en mesure de créer le graphique Histogramme si vous spécifiez les options (**Plage de sortie** ou **Nouvelle feuille de feuille de calcul**) qui créent le tableau d'histogramme dans le même classeur que vos données.

Pour plus d'informations, cliquez (Ctrl Clic gauche de la souris) sur le numéro ci-dessous pour afficher l'article correspondant dans la Base de connaissances Microsoft.
214029 (http://support.microsoft.com/kb/214029/) Utilisation d'outils d'analyse des données dans des feuilles regroupées

Annexe 4 : Tableau croisé dynamique dans Excel : utilisation et compléments

Introduction

Excel offre la possibilité de construire des tableaux de synthèse relativement élaborés dont le principe repose sur le croisement de plusieurs variables. L'appellation « tableaux croisés dynamiques » découle directement de ce principe, le qualificatif « dynamique » faisant référence au fait que toute modification opérées dans la série de données se traduit par une mise à jour quasi automatique du tableau croisé

Vocabulaire de base

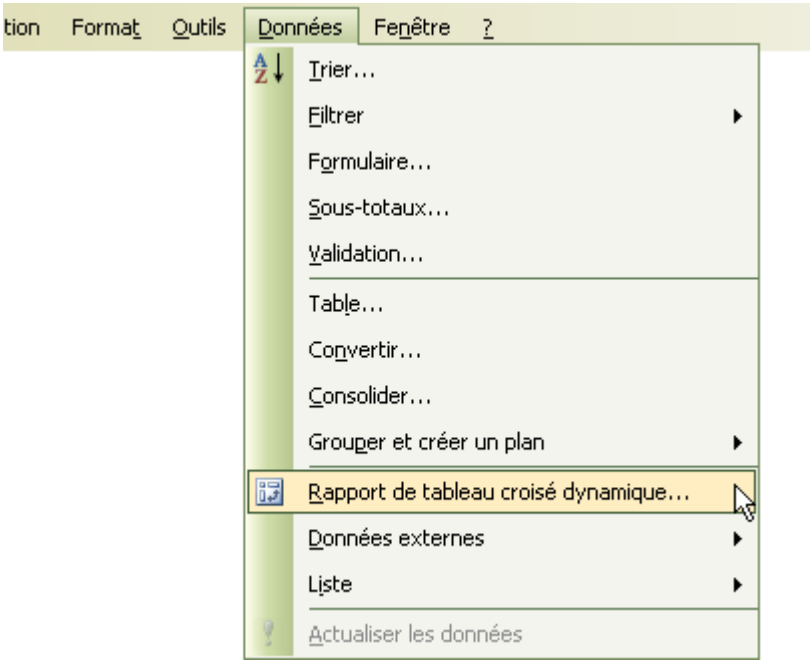
- Excel nomme « **champ** » les variables décrivant les individus de la population étudiée. Les champs constituent en général les colonnes du tableau.
- Les lignes du tableau Excel constituent les « enregistrements » qui décrivent les individus de la population étudiée. Une ligne = un individu.

Exemple de structure d'une tableau Excel :

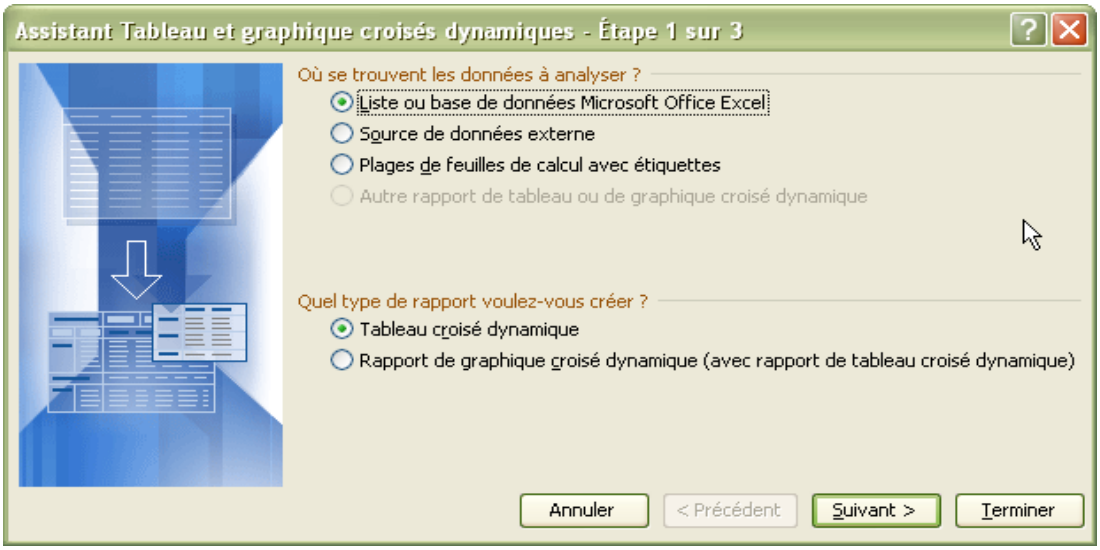
	A	B	C	D	E	F
1	**Epoque de construction**	**Catégorie de logement**	**Nombre de pièces**	**Nombre de personnes**	**Staut d'occupation**	**Type de logement**
2	1949 à 1967	RP	5	2	Propriétaire	Maison individuelle
3	1949 à 1967	RP	3	2	Propriétaire	Maison individuelle
4	1949 à 1967	RP	1	1	Loc. meublé	Sous-loc.
5	1949 à 1967	RP	1	1	Loc. meublé	Sous-loc.
6	1968 à 1974	RP	5	3	Propriétaire	Maison individuelle
7	1949 à 1967	RP	4	1	Propriétaire	Maison individuelle
8	1968 à 1974	RP	4	1	Propriétaire	Maison individuelle
9	1968 à 1974	RP	5	2	Propriétaire	Maison individuelle
10	1968 à 1974	RP	6	2	Propriétaire	Maison individuelle
11	1949 à 1967	RP	5	4	Logé gratuit	Maison individuelle
12	Av. 1945	RP	6	3	Propriétaire	Maison individuelle
13	1949 à 1967	RP	4	2	Propriétaire	Maison individuelle
14	Av. 1945	RP	2	2	Propriétaire	Maison individuelle
15	1949 à 1967	RP	3	1	Propriétaire	Autre
16	1982 à 1989	RP	6	2	Propriétaire	Maison individuelle
17	1949 à 1967	RP	3	3	Propriétaire	Maison individuelle
18	1945 à 1948	RP	3	1	Propriétaire	Maison individuelle
19	Av. 1945	RP	6	7	Propriétaire	Maison individuelle
20	1945 à 1948	RP	5	4	Propriétaire	Maison individuelle
21	Av. 1945	RP	6	4	Propriétaire	Maison individuelle

Création d'un tableau croisé dynamique

Du menu **Données,** sélectionnez l'option **Rapport de tableau croisé dynamique**.



L'écran suivant s'affiche :



Excel vous demande de précisez la localisation de la source des données qui servira à l'élaboration du tableau croisé dynamique. Plusieurs possibilités s'offrent à vous :

• Liste ou base de données Excel.	Les données proviennent d'une base de données Excel ou d'une série de cellules située sur une feuille de calcul d'Excel.

• Source de données externes	Les données proviennent d'autres logiciels tels qu'Access, dBASE, FoxPro ainsi que plusieurs autres.
• Plage de feuilles de calcul avec étiquette.	Créer automatiquement un tableau après lui avoir déterminé la plage de cellules à utiliser. Il utilise le contenu de la première ligne et de la première colonne pour déterminer le nom des champs du tableau.
• Autre tableau ou graphique croisé dynamique	Vous permet d'approfondir des analyses sur des tableaux et graphiques dynamiques qui ont déjà été conçus.

Excel vous demande ensuite quel type de rapport que vous souhaitez construire:

- tableau croisé dynamique : tableau croisé seul
- Rapport de graphique croisé dynamique : cette option vous permet de construire des graphiques élaborés à partir de tableau croisé. De ce fait, le choix de cette option s'accompagne également de la construction d'un tableau croisé dynamique

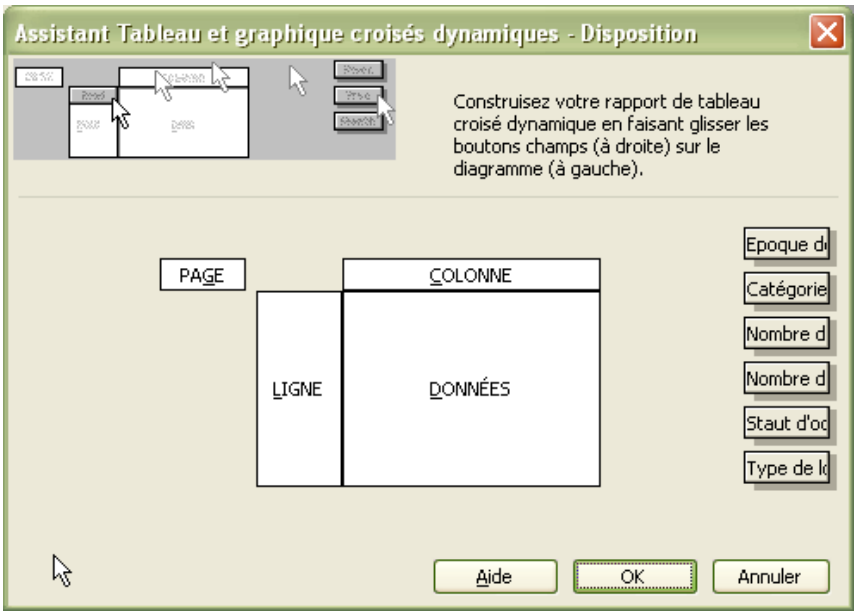
Une fois votre choix fait, appuyez sur le bouton **Suivant**.

Par défaut Excel sélectionne l'entièreté de la plage de données figurant sur la feuille active du fichier. Vous pouvez modifier cette sélection ou bien confirmer le choix d'Excel. Appuyer sur « Suivant ».

Excel vous demande ensuite de choisir l'endroit où vous voulez voir figurer les résultats. Vous avez le choix entre une nouvelle feuille et un endroit à préciser de la feuille active.

Si vous appuyez sur « Terminer », Excel s'exécute et produit la structure du tableau croisé à l'endroit précisé. Vous pouvez, avant cela, explorer les autres options proposées sur l'écran :

- L'option **Disposition vous permet de concevoir** immédiatement le tableau croisé dynamique (choix et disposition des champs à l'intérieur des différentes zones du tableau (page, ligne, colonne et données).



Données Cette zone affiche les résultats que vous voulez voir pour un champ. Par défaut, le tableau affiche la somme des valeurs si celui-ci est composé de chiffres. S'il est composé de texte, le tableau va afficher le nombre d'enregistrements qui répond au critère. Il y a d'autres fonctions qui sont disponibles tel que la moyenne, l'écart type et plusieurs autres. Une liste sera mentionnée à la fin de cette page.

Colonne Affiche chacune des valeurs d'un champ dans sa propre colonne.

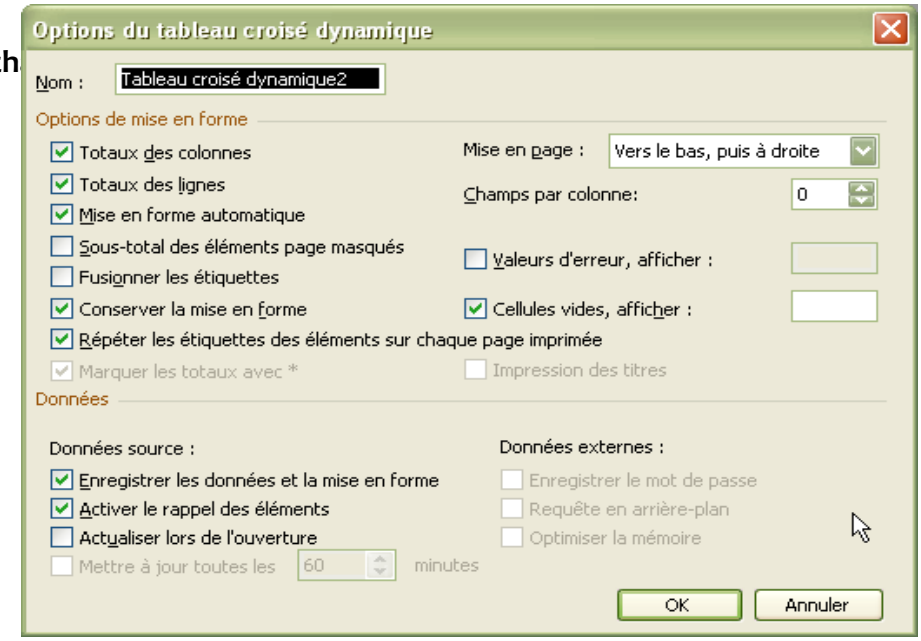
Ligne Affiche chacune des valeurs d'un champ sur sa propre ligne.

Page Permets de "filtrer" les valeurs du tableau par rapport aux valeurs d'un champ. Ceci permet de voir seulement les enregistrements qui répondent à un certain critère.

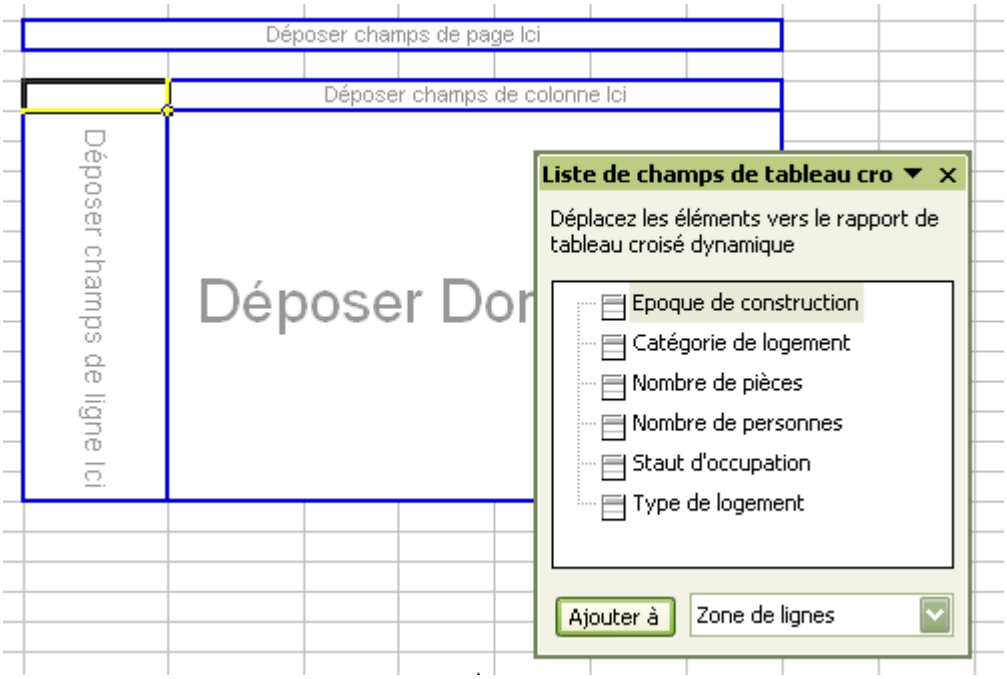
Le placement des champs peut être réalisé plus tard

- L'option **Option vous permet de personnaliser l'affichage de l'information dans le tableau. Vous pouvez, par exemple, décider d'afficher ou non les totaux par ligne ou par colonne, de nommer votre tableau, etc.**

Placer les ch



Dès lors que vous avez validez la création du tableau en ayant appuyer sur « Terminer » avec ou sans option mais avoir activer l'option « Disposition », Excel crée à l'endroit précisé, une structure vide de tableau dynamique qu'il vous appartient de compléter en y ajoutant les variables souhaitées dans les zones ad hoc.



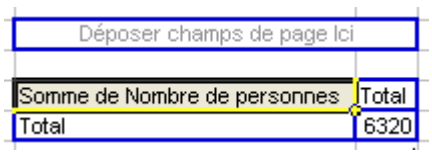
L'affichage de la structure du tableau s'accompagne normalement de l'apparition d'une nouvelle barre d'outils spécialement dédiée aux tableaux croisés dynamiques

Le remplissage du tableau peut alors s'effectuer

- A partir de la « liste de champs de tableau croisé dynamique », sélectionnez le champ « Nombre de personnes ».
- De la liste des zones du tableau, sélectionnez la **zone de données**.
- Appuyer sur le bouton « Ajouter à »

OU

- En gardant un doigt sur le bouton gauche de la souris, déplacez le champ dans la zone de données.
- Relâchez le bouton de la souris dès que le carré pour le champ **« Nombre de personnes »** est par-dessus la zone de données.



Le tableau indique maintenant que le nombre total des personnes habitant dans le parc de logement de la Ville de Gray s'élève à 6 320. La prochaine étape consiste à répartir cette population par type et taille de logements.

- De la barre d'outils Tableau croisé dynamique, sélectionnez le champ « Type de logements ».
- De la liste des zones du tableau, sélectionnez la **zone de colonnes**.
- Appuyez sur le bouton **Ajouter à**.

OU

- En gardant un doigt sur le bouton gauche de la souris, déplacez le champ dans la **zone de colonnes**.
- Relâchez le bouton de la souris dès que le carré pour le champ **« Type de logements »** est par-dessus la zone de colonnes.

Somme de Nombre de personnes	Type de logement							
	Autre	Chambre hotel	Collectif	Construction provisoire	Foyés	Maison individuelle	Sous-loc.	Total
Total	153	1	3660	1	83	2365	57	6320

La population est maintenant répartie en fonction du type de logement. Remarquez que le total des personnes est toujours de 6 320. Le tableau affiche chacune des valeurs du champ **« Type de logements »** avec le total des personnes pour celui-ci. L'étape suivante consiste à répartir le total des personnes par **« Type de logements »** et par « Nombre de pièces ».

- A partir de la barre d'outils Tableau croisé dynamique, sélectionnez le champ **« Nombre de pièces »**.
- En gardant un doigt sur le bouton gauche de la souris, déplacez le champ dans la zone de colonnes.
- Relâchez le bouton de la souris dès que le carré pour le champ **« Type de logements »** est par-dessus la zone de colonnes.

Somme de Nombre de personnes	Type de logement	Nombre de pièces											
	Collectif					Total Collectif	Maison individuelle					Total Maison individuelle	
	1	2	3	4	5+		1	2	3	4	5+		
Total	152	401	1103	1209	795	3660	13	97	316	654	1285	2365	

Le champ **« Nombre de pièces »** va être automatiquement placé devant le champ **Titre.** À cause de la longueur du tableau, seulement une partie est affichée à l'image ci-dessus. Il est possible aussi de changer l'ordre de présentation des champs. La prochaine opération consiste à donner la priorité au champ **« Type de logements »** par-dessus **« Nombre de pièces »**.

- Placez le pointeur par-dessus le champ **« Type de logements »** de la zone des colonnes du tableau croisé dynamique.
- En gardant un doigt sur le bouton gauche de la souris, déplacez le champ **« Type de logements »** devant le champ **« Nombre de pièces »**.
- Une fois devant le champ **« Nombre de pièces »**, relâchez le bouton de la souris.

Le tableau qui suit propose les mêmes informations mais avec un arrangement différent. On dispose maintenant du nombre de personnes par type de logement et selon le nombre de pièces, le total général restant inchangé. Pour ce faire, procéder comme suit :

- Placez le pointeur par-dessus le champ « **Type de logements** » de la zone des colonnes du tableau croisé dynamique.
- En gardant un doigt sur le bouton gauche de la souris, déplacez le champ « **Type de logements** » dans la zone des lignes du tableau croisé dynamique (par-dessus Somme de la ligne).
- Une fois le champ est dans la zone des lignes, relâchez le bouton de la souris.

Somme de Nombre de personnes	Nombre de pièces					
Type de logement	1	2	3	4	5+	Total
Autre	1	10	23	57	62	153
Chambre hotel	1					1
Collectif	152	401	1103	1209	795	3660
Construction provisoire	1					1
Foyés	81		2			83
Maison individuelle	13	97	316	654	1285	2365
Sous-loc.	12	11	22	4	8	57
Total	261	519	1466	1924	2150	**6320**

Voir les données

Excel vous permet de voir l'ensemble enregistrements qui composent les résultats du tableau. Pour ce faire, Excel génère automatiquement une nouvelle feuille. Vous pouvez obtenir le tableau des enregistrements pour n'importe quelle cellule du tableau croisé selon le même principa.

- Placez le pointeur sur la cellule contenant le total des personnes (6 320).
- Faites un double-clic sur la cellule

	A	B	C	D	E	F
1	Epoque de construction	Catégorie de logement	Nombre de pièces	Nombre de personnes	Staut d'occupation	Type de logement
2	Av. 1945	RP	1	1	Logé gratuit	Autre
3	1945 à 1948	RP	2	2	Loc. privé vide	Autre
4	1945 à 1948	RP	2	1	Loc. privé vide	Autre
5	Av. 1945	RP	2	1	Propriétaire	Autre
6	1945 à 1948	LV	2	0	Inoccupé	Autre
7	Av. 1945	RP	2	1	Loc. privé vide	Autre
8	Av. 1945	RP	2	1	Loc. privé vide	Autre
9	Av. 1945	RP	2	1	Loc. privé vide	Autre
10	Av. 1945	RP	2	1	Propriétaire	Autre
11	Av. 1945	RP	2	2	Propriétaire	Autre
12	1949 à 1967	RP	3	1	Propriétaire	Autre
13	1949 à 1967	RP	3	4	Logé gratuit	Autre
14	1949 à 1967	RP	3	1	Loc. HLM	Autre
15	1945 à 1948	RP	3	2	Loc. privé vide	Autre
16	1945 à 1948	RP	3	1	Loc. privé vide	Autre
17	Av. 1945	RP	3	4	Propriétaire	Autre
18	Av. 1945	RP	3	1	Loc. privé vide	Autre

Filtrer sur les champs

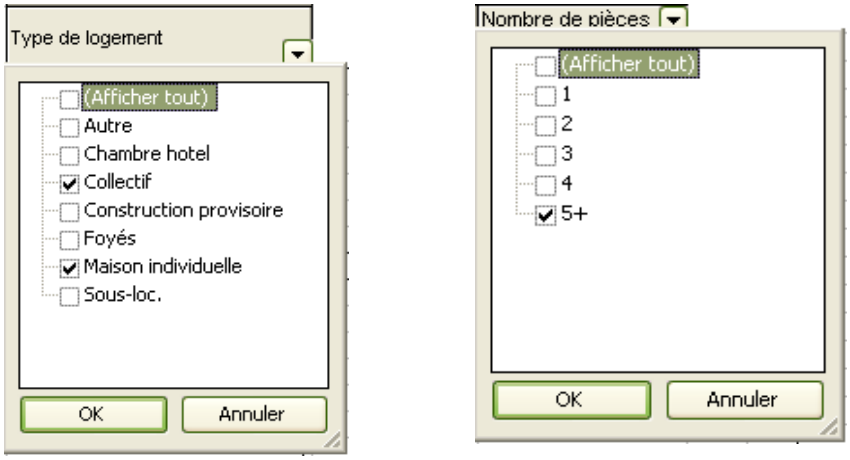
Excel vous permet de filtrer les données sur la base des modalités relatives à chaque champ (ou variable) en fonction de vos besoins. On peut ainsi masquer certaines modalités avec une mise à jour automatique du tableau, sachant que l'on peut à tout moment faire réapparaitre les champs occultés avec réactualisation du contenu du tableau.

Procédure de « masquage » de modalités:
À la droite du champ **« Type de logements »**, cliquez sur le bouton avec un triangle pointant vers le bas. Cette manipulation permet d'afficher l'ensemble des modalités disponibles et actives pour le champs **« Type de logements ».** Vous avez alors la possibilité de désactiver certains d'entre eux pour ne faire apparaître dans le tableau que les informations relatives à ceux encore actifs. Dans notre exemple, nous avons choisi de ne laisser actif que les modalités « collectif » et « maison individuelle ».

Le tableau est automatiquement mis à jour en tenant compte de vos choix. Vous pouvez à tout moment revenir à une situation affichant l'ensemble des informations pour l'ensemble des modalités. Le total général n'est évidemment plus le même puisque seule une partie de la population est maintenant prise en compte. Vous remarquez également que les modalités désactivées ne figurent plus dans le tableau.

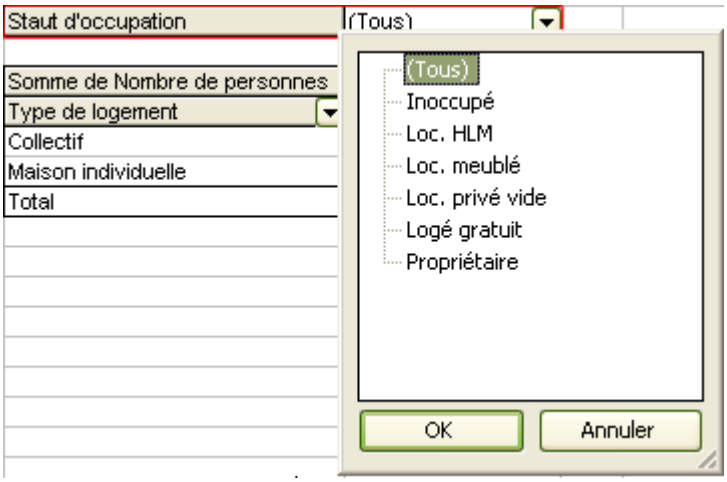
Somme de Nombre de personnes	Nombre de pièces					
Type de logement	1	2	3	4	5+	Total
Collectif	152	401	1103	1209	795	3660
Maison individuelle	13	97	316	654	1285	2365
Total	165	498	1419	1863	2080	**6025**

Le filtrage peut s'effectuer sur plusieurs champs simultanément, par exemple sur **« Type de logements » et « Nombre de pièces ».**



Somme de Nombre de personnes	Nombre de pièces	
Type de logement	5+	Total
Collectif	795	795
Maison individuelle	1285	1285
Total	2080	2080

Il est encore possible d'ajouter d'autres champs (ou variables) de manière à affiner, si nécessaire, le filtrage des informations. Cet ajout peut se faire dans la zone située au-dessus du tableau et dite « zone de page »



- A partir de la liste de champs de tableau croisé dynamique, sélectionnez le champ « Statut d'occupation ».
- Dans la même fenêtre, sélectionnez « **zone de pages** ».
- Appuyez sur le bouton **Ajouter à**.

OU

- En gardant un doigt sur le bouton gauche de la souris, déplacez le champ « Statut d'occupation » dans la zone de pages du tableau croisé dynamique.
- Une fois le champ est dans la zone de pages, relâchez le bouton de la souris.

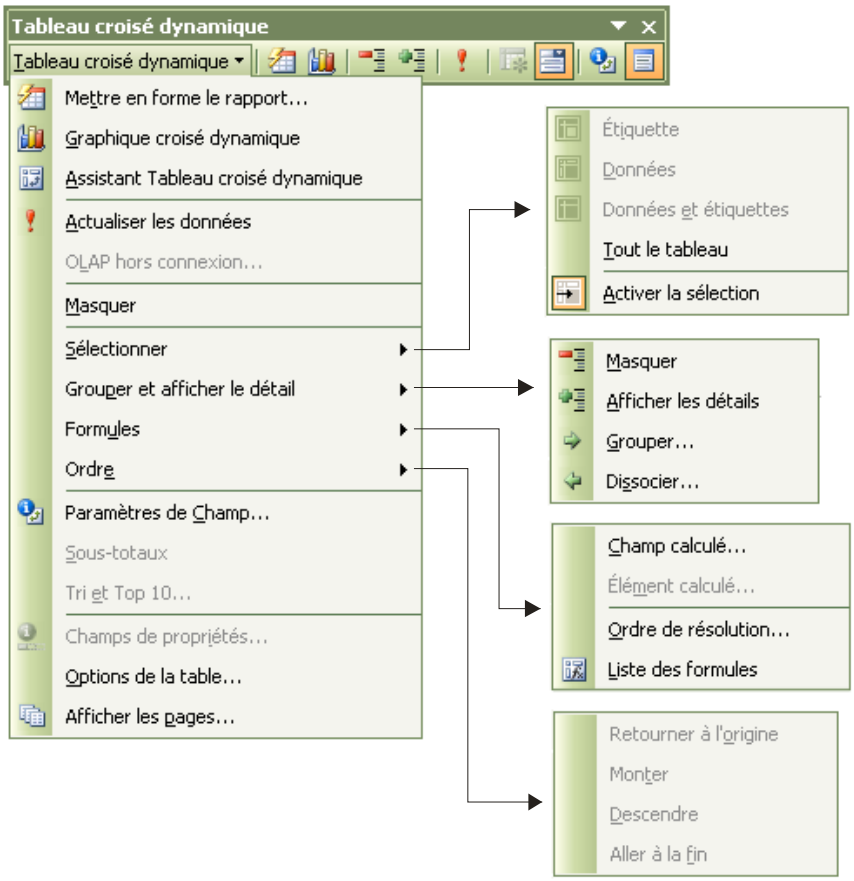
Dès lors que le champ « Statut d'occupation » est la zone de pages, il vous est possible de filtrer toutes les informations du tableau en activant uniquement par exemple la modalité « Propriétaire ».

Staut d'occupation	Propriétaire	
Somme de Nombre de personnes	Nombre de pièces	
Type de logement	5+	Total
Collectif	119	119
Maison individuelle	973	973
Total	1092	1092

De cette façon, nous pouvons connaître précisément la population ayant un statut de propriétaire, vivant dans des logements de 5 pièces et plus en habitat de type collectif ou maison individuelle.

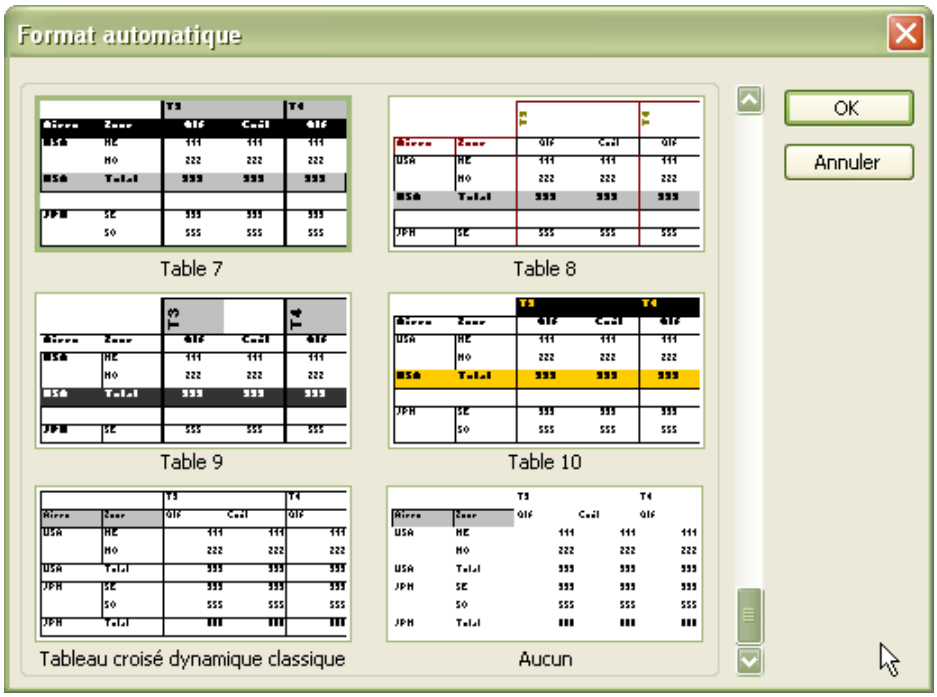
Les fonctions et options du menu et de la barre d'outils

La barre d'outils tableau croisé dynamique offre d'autres options pour notamment modifier et améliorer l'organisation et la présentation de l'information :

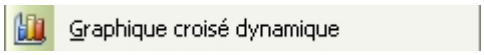


Mettre en forme le rapport 

A partir du tableau croisé dynamique créé, cette fonction vous permet de construire une présentation plus élaborée des résultats obtenus avec une meilleure maîtrise de la mise en page, des couleurs, de l'organisation en général du tableau. Excel propose en standard un certain nombre de modèle de mise en forme

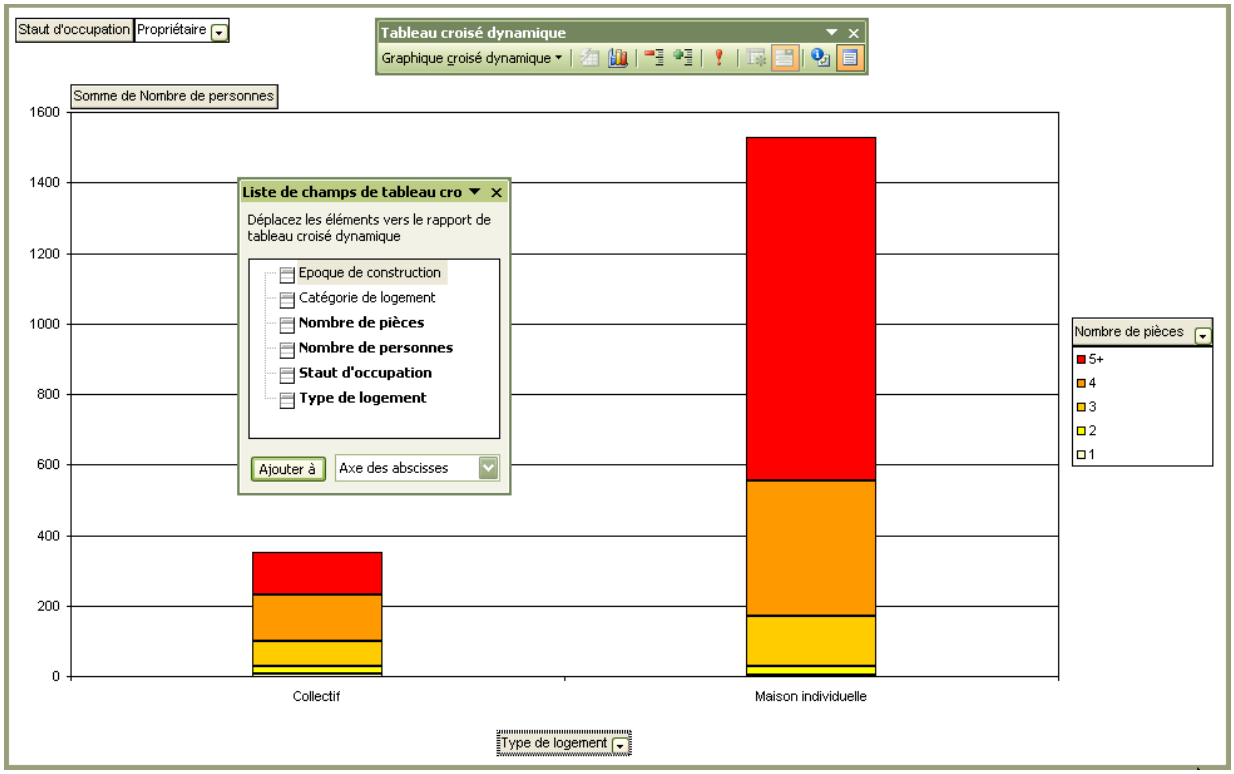


Créer des graphiques à partir du tableau croisé

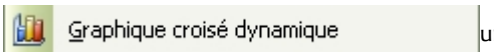


Il y a des situations où il est préférable de représenter une masse de données sous forme de graphique comme par exemple :

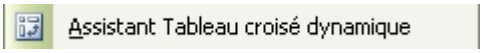
- Pour simplifier l'analyse d'une masse de données.
- Pour ressortir rapidement les tendances des séries de données.
- Pour pouvoir comparer les données.
- Pour ressortir des proportions.



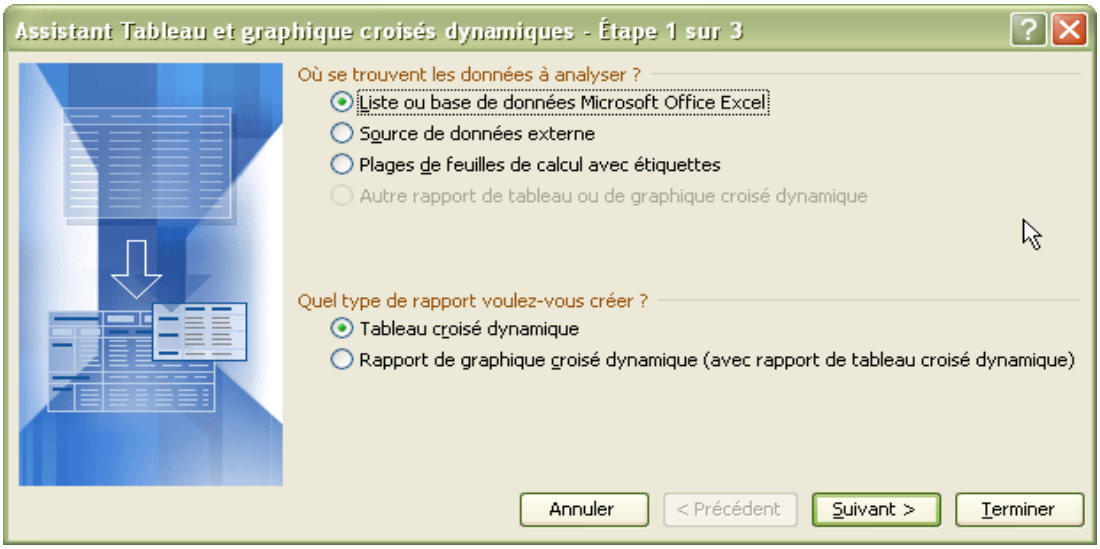
Excel génère par défaut un graphique en barres. Il est bien entendu possible de modifier le type de graphique en passant par la procédure classique prévue à cet effet. Relativement au graphique dynamique, Excel vous donne la possibilité de changer les variables à représenter, de modifier les filtres, etc. avec effets immédiat sur le graphique.

En appuyant sur le bouton 

une seconde fois, vous activez l'assistant graphique qui vous permet de modifier le type de graphique comme vous le feriez lors de la création d'un graphique classique.

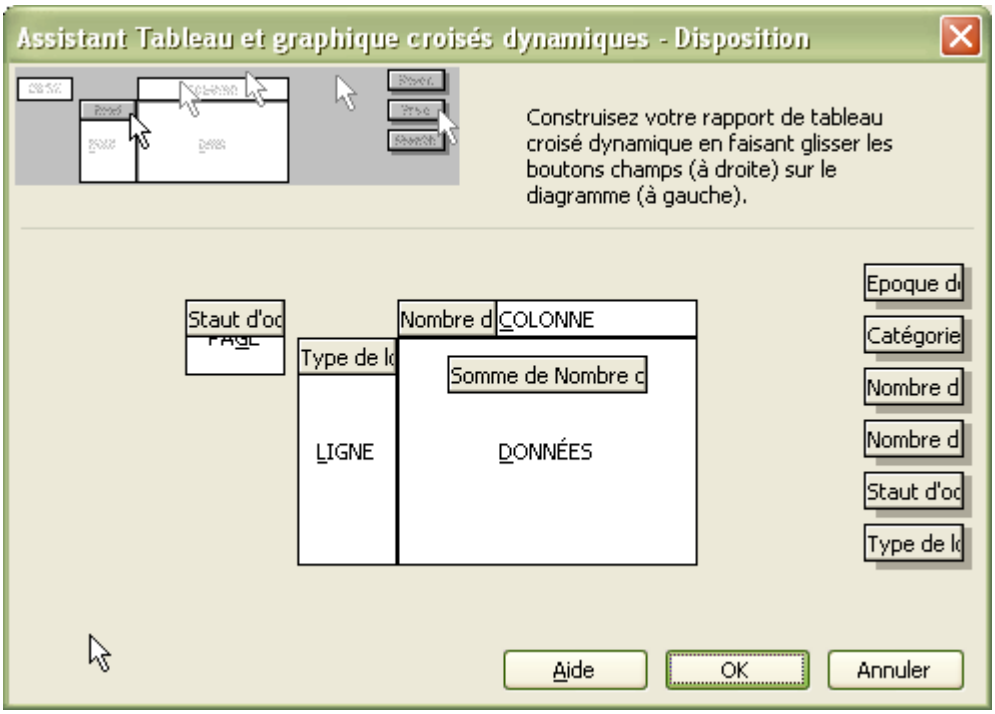
Assistant tableau croisé dynamique 

Cette option permet de changer la disposition des champs dans le tableau croisé dynamique. Assurez-vous de placer le pointeur à l'intérieur du tableau croisé dynamique.



Dans la barre d'outils du tableau croisé dynamique, sélectionnez l'option « **Assistant tableau croisé dynamique** ».

L'assistant affiche la structure actuelle du tableau en vous laissant la possibilité de la modifier à votre guise.



- Appuyez sur le bouton **OK.**

- Appuyez sur le bouton **Terminer.**

Staut d'occupation	Propriétaire					
Somme de Nombre de personnes	Nombre de pièces					
Type de logement	1	2	3	4	5+	Total
Collectif	6	21	72	133	119	351
Maison individuelle	2	28	141	385	973	1529
Total	8	49	213	518	1092	1880

Actualiser les données 

Cette option vous permet de remettre à jour les données du tableau croisé dynamique après qu'une modification de la base de données ait été effectuée.

- Placez le pointeur dans la feuille de calcul à n'importe quel endroit

- Changer une ou plusieurs données

- Retourner à la feuille de calcul ayant le tableau croisé dynamique.

- Appuyez sur le bouton  **et voyez le résultat notamment au niveau des sous-totaux.**

Masquer  ou afficher  les détails

Il est possible d'avoir dans une zone plusieurs champs pour mieux décrire les valeurs. Ces options permettent d'afficher ou de masquer les valeurs des champs qui sont à la droite du champ sélectionné. Si vous ne l'avez pas fait, ajoutez les champs Nom et Prénom à la zone des lignes.

Prenons le tableau suivant :

Staut d'occupation	(Tous)						
Somme de Nombre de personnes		Nombre de pièces					
Type de logement	Epoque de construction	1	2	3	4	5+	Total
Autre	1945 à 1948		3	5	12	23	43
	1949 à 1967			10	7	12	29
	1968 à 1974				1		1
	1975 à 1981				3	4	7
	Av. 1945	1	7	8	34	23	73
Total Autre		1	10	23	57	62	153
Chambre hotel	1945 à 1948	1					1
Total Chambre hotel		1					1
Collectif	1945 à 1948	24	47	94	41	67	273
	1949 à 1967	8	58	349	417	337	1169
	1968 à 1974	24	95	252	344	165	880
	1975 à 1981	3	23	75	97	19	217
	1982 à 1989	23	19	41	66	11	160
	1990 et ap.	1	2	13	28	35	79
	Av. 1945	69	157	279	216	161	882
Total Collectif		152	401	1103	1209	795	3660
Construction provisoire	Av. 1945	1					1
Total Construction provisoire		1					1
Foyés	1982 à 1989	81		2			83
Total Foyés		81		2			83
Maison individuelle	1945 à 1948	3	34	72	124	226	459
	1949 à 1967		5	68	170	234	477
	1968 à 1974		1	17	36	148	202
	1975 à 1981		1	12	67	135	215
	1982 à 1989		1	6	19	73	99
	1990 et ap.	1	8	24	30	45	108
	Av. 1945	9	47	117	208	424	805
Total Maison individuelle		13	97	316	654	1285	2365
Sous-loc.	1945 à 1948		1	6		7	14
	1949 à 1967	3	2				5
	1968 à 1974	1			4		5
	1975 à 1981			1			1
	1990 et ap.		1				1
	Av. 1945	8	7	15		1	31
Total Sous-loc.		12	11	22	4	8	57
Total		**261**	**519**	**1466**	**1924**	**2150**	**6320**

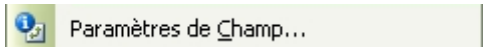
Placer le pointeur sur le champs « Époque de construction » et appuyez sur le bouton 

Bien que le nom du champs reste apparent, les informations s'y rattachant ont été masquées et ne sont plus affichées

Staut d'occupation	(Tous)						
Somme de Nombre de personnes		Nombre de pièces					
Type de logement	Epoque de construction	1	2	3	4	5+	Total
Autre		1	10	23	57	62	153
Chambre hotel		1					1
Collectif		152	401	1103	1209	795	3660
Construction provisoire		1					1
Foyés		81		2			83
Maison individuelle		13	97	316	654	1285	2365
Sous-loc.		12	11	22	4	8	57
Total		**261**	**519**	**1466**	**1924**	**2150**	**6320**

L'option  permet de réafficher les informations cachées dans les mêmes conditions. Sélectionnez le champs « Type de logement » et cliquer sur l'icône  pour faire réapparaître les informations relatives au champs **« Époque de construction ».**

En se positionnant à nouveau sur le champ « Époque de construction » et en cliquant sur  , Excel affiche la liste des champs non encore présents dans la partie du tableau concernée et que vous pouvez ajouter.

Changer les paramètre des champs 

Excel offre la possibilité de modifier les paramètres attachés à un champ. Par défaut, Excel produit pour un champ soit la somme soit le nombre. D'autres fonctions sont pourtant disponibles.

Sélectionner une des cases du tableau intitulée Total « nom du champ » et cliquez sur l'icône  . S'affiche alors la fenêtre suivante qui vous autorise à modifier les paramètres liés au champs sélectionné. De total ou somme vous pouvez passer à nombre, moyenne, minimum, maximum, produit, écart-type, etc. selon les besoins. En choisissant par exemple la paramètre « moyenne » en lieu et place de « total », Excel remplace le total en colonne et en ligne par une moyenne.

L'option « Avancé » vous permet de paramétrer plus en détail la procédure en vous donnant la possibilité d'effectuer par exemple des tris.

Par exemple, en sélectionnant le champs « somme de nombre de personnes » et le transformant en « moyenne de nombre de personnes » vous obtenez le nombre moyen de personnes occupant les logements selon la taille, le type et l'époque de construction.

Staut d'occupation	(Tous)						
Moyenne de Nombre de personnes		Nombre de pièces					
Type de logement	Epoque de construction	1	2	3	4	5+	Total
Autre	1949 à 1967			2,00	3,50	2,40	**2,42**
	1968 à 1974				1,00		**1,00**
	1975 à 1981				3,00	4,00	**3,50**
	1915 et av.	1,00	1,17	2,00	3,40	2,88	**2,52**
	1915 à 1948		1,00	1,67	4,00	2,56	**2,39**
Total Autre		1,00	1,11	1,92	3,35	2,70	**2,47**
Chambre hotel	1915 à 1948	1,00					**1,00**
Total Chambre hotel		1,00					**1,00**
Collectif	1949 à 1967	0,80	1,14	1,72	2,54	3,24	**2,20**
	1968 à 1974	0,89	0,91	1,75	2,59	4,34	**1,97**
	1975 à 1981	1,00	1,00	1,53	2,43	1,73	**1,72**
	1982 à 1989	0,96	0,83	1,32	2,36	2,75	**1,45**
	1990 et ap.	0,50	1,00	1,86	2,33	4,38	**2,55**
	1915 et av.	0,95	1,04	1,69	2,27	2,73	**1,62**
	1915 à 1948	0,89	1,15	1,77	2,28	3,94	**1,75**
Total Collectif		0,92	1,02	1,69	2,47	3,30	**1,88**
Construction provisoire	1915 et av.	1,00					**1,00**
Total Construction provisoire		1,00					**1,00**
Foyés	1982 à 1989	1,05		2,00			**1,06**
Total Foyés		1,05		2,00			**1,06**
Maison individuelle	1949 à 1967		0,71	1,94	2,00	2,41	**2,13**
	1968 à 1974		1,00	2,13	1,64	2,35	**2,15**
	1975 à 1981		1,00	2,00	2,48	2,33	**2,34**
	1982 à 1989		1,00	1,50	2,71	3,04	**2,75**
	1990 et ap.	1,00	1,33	2,40	2,73	2,81	**2,45**
	1915 et av.	0,82	0,92	1,63	2,24	2,48	**2,02**
	1915 à 1948	1,00	1,26	1,67	2,34	2,86	**2,24**
Total Maison individuelle		0,87	1,03	1,78	2,19	2,53	**2,16**
Sous-loc.	1949 à 1967	1,00	2,00				**1,25**
	1968 à 1974	1,00			4,00		**2,50**
	1975 à 1981			1,00			**1,00**
	1990 et ap.		1,00				**1,00**
	1915 et av.	1,00	1,00	2,14		1,00	**1,35**
	1915 à 1948		1,00	3,00		3,50	**2,80**
Total Sous-loc.		1,00	1,10	2,20	4,00	2,67	**1,58**
Total		**0,96**	**1,02**	**1,72**	**2,39**	**2,77**	**1,97**

La liste des transformations possibles

Somme	Affiche la somme de toutes les valeurs de ce champ.
Nbval	Affiche le nombre d'enregistrements dans cette catégorie.
Moyenne	Affiche la moyenne de toutes les valeurs de ce champ.
Max	Affiche la plus grande valeur du champ.
Min	Affiche la plus petite valeur du champ.
Produit	Affiche la multiplication de toutes les valeurs du champ.
Nb	Affiche le nombre d'enregistrements dans cette catégorie.
Ecartype	Affiche l'écart type du champ.
Ecartypep	Affiche l'écart type d'une population.
Var	Affiche la variance du champ.
Varp	Affiche la variance d'une population.

La fenêtre des paramètres du champ vous offre aussi d'autres options tel que démontré dans la prochaine partie.

Grouper ou dissocier des valeurs

Cette fonction vous permet de regrouper des modalités d'un même champ.

On peut, par exemple regrouper les logements construite « Av. 1915 » avec ceux « de 1915 à 1948 » de façon à former une catégorie « logements anciens ». les informations et totaux ou autres paramètres seront réajustés automatiquement.

Staut d'occupation	(Tous)							
Moyenne de Nombre de personnes			Nombre de pièces					
Type de logement	Epoque de construction2	Epoque de construction	1	2	3	4	5+	Total
Autre	Groupe2	1949 à 1967			2,00	3,50	2,40	**2,42**
		1968 à 1974				1,00		**1,00**
	Groupe3	1975 à 1981				3,00	4,00	**3,50**
	Groupe1	1915 et av.	1,00	1,17	2,00	3,40	2,88	**2,52**
		1915 à 1948		1,00	1,67	4,00	2,56	**2,39**
Total Autre			1,00	1,11	1,92	3,35	2,70	**2,47**
Chambre hotel	Groupe1	1915 à 1948	1,00					**1,00**
Total Chambre hotel			1,00					**1,00**
Collectif	Groupe2	1949 à 1967	0,80	1,14	1,72	2,54	3,24	**2,20**
		1968 à 1974	0,89	0,91	1,75	2,59	4,34	**1,97**
	Groupe3	1975 à 1981	1,00	1,00	1,53	2,43	1,73	**1,72**
		1982 à 1989	0,96	0,83	1,32	2,36	2,75	**1,45**
	1990 et ap.	1990 et ap.	0,50	1,00	1,86	2,33	4,38	**2,55**
	Groupe1	1915 et av.	0,95	1,04	1,69	2,27	2,73	**1,62**
		1915 à 1948	0,89	1,15	1,77	2,28	3,94	**1,75**
Total Collectif			0,92	1,02	1,69	2,47	3,30	**1,88**
Construction provisoire	Groupe1	1915 et av.	1,00					**1,00**
Total Construction provisoire			1,00					**1,00**
Foyés	Groupe3	1982 à 1989	1,05		2,00			**1,06**
Total Foyés			1,05		2,00			**1,06**
Maison individuelle	Groupe2	1949 à 1967		0,71	1,94	2,00	2,41	**2,13**
		1968 à 1974		1,00	2,13	1,64	2,35	**2,15**
	Groupe3	1975 à 1981		1,00	2,00	2,48	2,33	**2,34**
		1982 à 1989		1,00	1,50	2,71	3,04	**2,75**
	1990 et ap.	1990 et ap.	1,00	1,33	2,40	2,73	2,81	**2,45**
	Groupe1	1915 et av.	0,82	0,92	1,63	2,24	2,48	**2,02**
		1915 à 1948	1,00	1,26	1,67	2,34	2,86	**2,24**
Total Maison individuelle			0,87	1,03	1,78	2,19	2,53	**2,16**
Sous-loc.	Groupe2	1949 à 1967	1,00	2,00				**1,25**
		1968 à 1974	1,00			4,00		**2,50**
	Groupe3	1975 à 1981			1,00			**1,00**
	1990 et ap.	1990 et ap.		1,00				**1,00**
	Groupe1	1915 et av.	1,00	1,00	2,14		1,00	**1,35**
		1915 à 1948		1,00	3,00		3,50	**2,80**
Total Sous-loc.			1,00	1,10	2,20	4,00	2,67	**1,58**
Total			**0,96**	**1,02**	**1,72**	**2,39**	**2,77**	**1,97**

La fonction « Dissocier » aboutit au résultat inverse, dissociant les modalités groupées.

Changer le nom d'une cellule

-Placez le pointeur dans la cellule **Groupe1**.

-Cliquez dans la zone des formules.

-Changez le nom à **Administration**.

OU

-Appuyez sur la touche **F2**.

-Changez le nom à **Administration**.

-Placez le pointeur dans la cellule **Groupe2**.

-Cliquez dans la zone des formules.

-Changez le nom à **Terrain**.

OU

-Appuyez sur la touche **F2**.

-Changez le nom à **Terrain**.

Il reste qu'a changer le nom du champ Titre2 à Regroupement.

-Placez le pointeur sur le champ Regroupement.

-Appuyez sur le bouton  .

-Changez le nom du champ de **Titre2** à **Regroupement**.

L'employeur a besoin d'une synthèse qui n'inclut pas les champs Titre, Nom et Prénom. On pourrait retirer les champs inutiles. Mais nous allons simplement les masquer pour l'instant.

Placez le pointeur sur la cellule ayant le texte Administration.

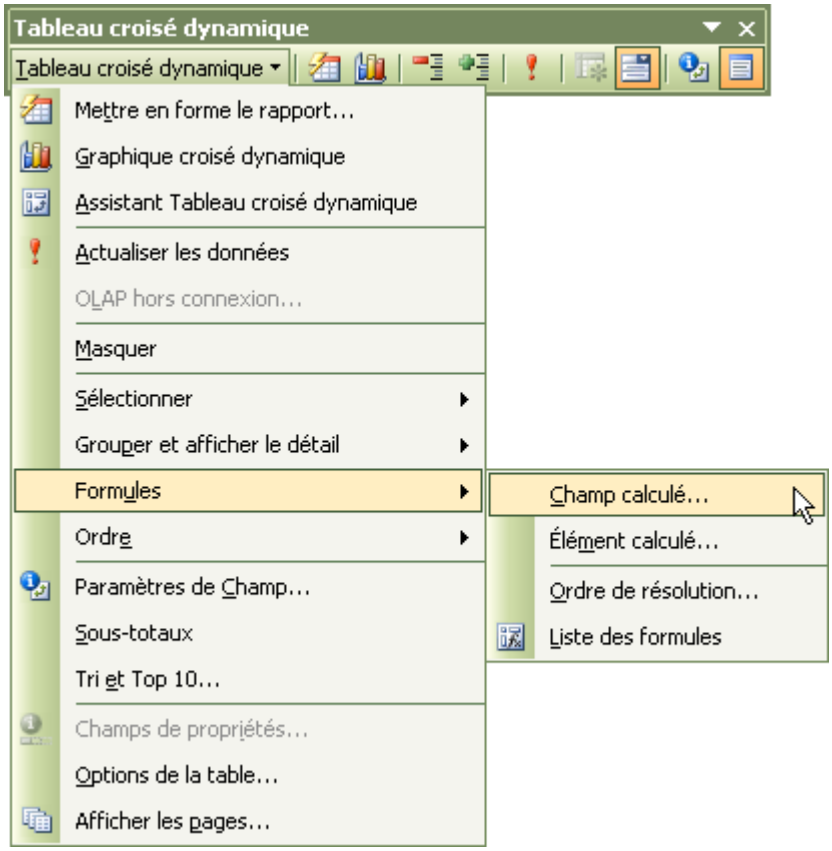
Appuyez sur le bouton . 

Placez le pointeur sur la cellule ayant le texte Terrain.

Appuyez sur le bouton . 

Création d'un champ calculé

Le tableau croisé dynamique vous permet en plus d'ajouter des champs calculés. On peut, par exemple, calculer le nombre moyen de pièces par personne en fonction des critères déjà présent dans le tableau croisé (époque de construction, type de logement, statut d'occupation)



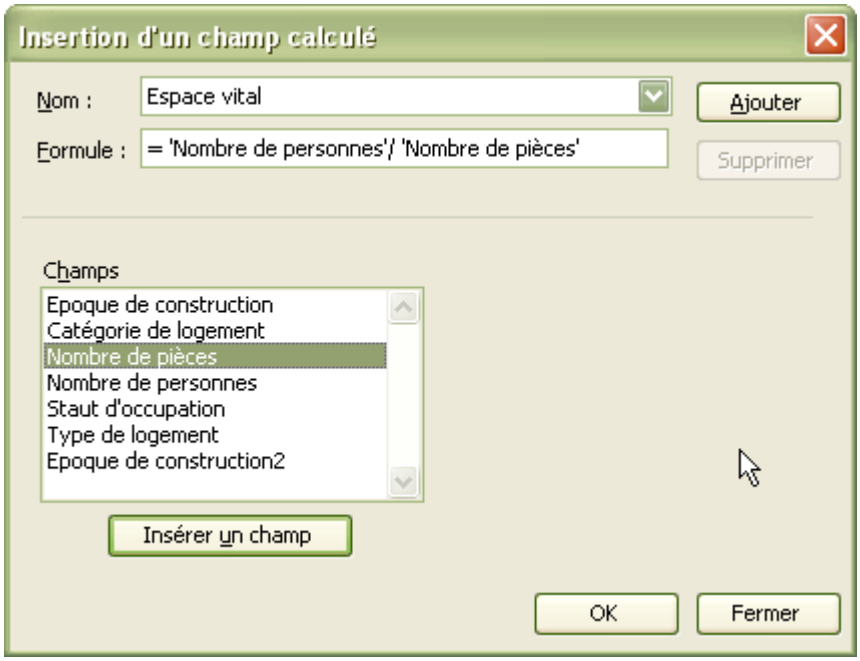
- Placez le pointeur sur le tableau croisé dynamique.

- A partir de la barre d'outils pour le tableau croisé dynamique, sélectionnez les options Formules et Champ calculé.

- Sélectionnez les champs concernés et la relation qui les liera dans la formule

- Donnez éventuellement un nom à votre champ calculé

- Cliquez « OK » et visualisez le résultat



Références

http://www.excel-online.net/tabl_crois.html