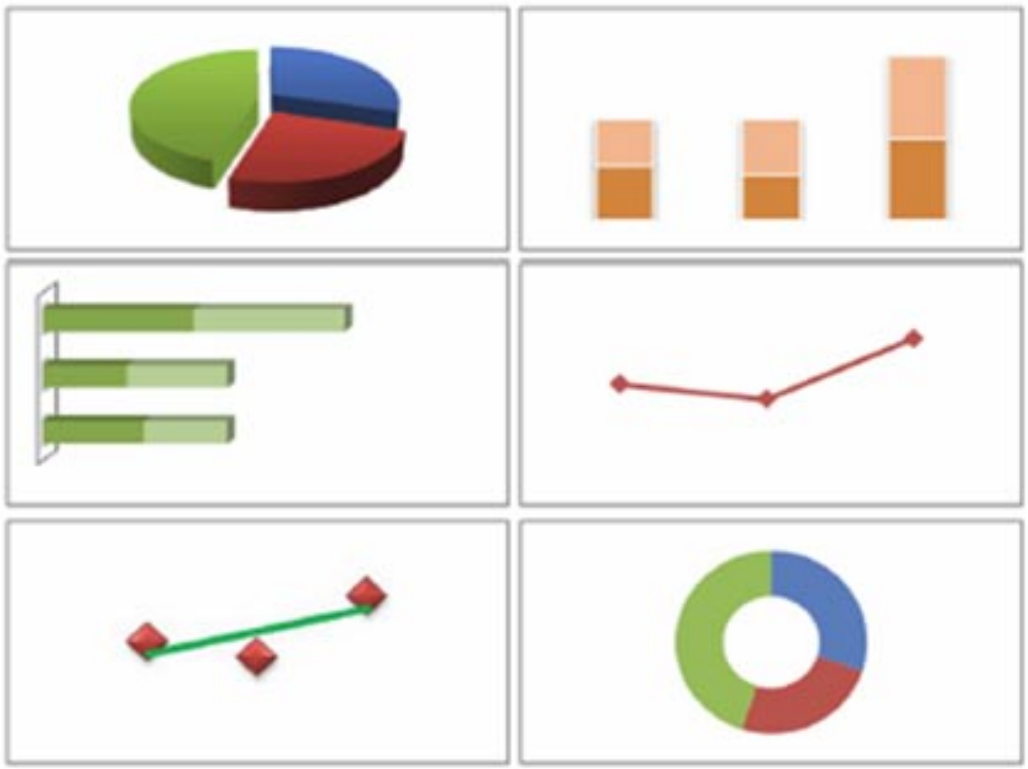
2008

Statistique descriptive



Fabrice MAZEROLLE

Statistique Descriptive

Fabrice MAZEROLLE

Notes de cours 2008

Dernière mise à jour le jeudi 13 mars 2008

1ère année de Licence AES – Marseille & Aix

Résumé du cours

Ce cours d'introduction à la statistique descriptive de niveau L1 a pour objet de donner les outils de bases permettant de décrire une population statistique. Les exemples décrits dans le cours sont essentiellement tirés de la vie économique.

Pour toute question relative à ce cours, merci de m'adresser un mail à fabrice@mazerolle.fr .

Sommaire

Chapitre 1 : Vocabulaire de la statistique descriptive

Chapitre 2 : Les tableaux statistiques

Chapitre 3 : Statistiques permettant de résumer une série

Chapitre 4 : Indices et progressions

Chapitre 5 : Diagrammes et graphiques

Chapitre 6 : Tendances et corrélations

Chapitre 7 : Courbe de LORENZ et coefficient de GINI

Bibliographie

Chapitre 1
Vocabulaire de la statistique descriptive

1 - Utilité de la statistique descriptive en économie
    A - Définition
    B - Exemples d'utilisation
2 - Terminologie
    A – Population et unités statistiques
    B – Echantillons et sous-ensembles d'une population
    C - Critères de classification
        1) Critères quantitatifs
        2) Critères qualitatifs
3 - Modes de regroupement des données
    A - Série simple
    B - Distribution par valeurs ou par modalités
        1) Distribution par valeurs
        2) Distribution par modalités
    C - Regroupement par catégories
        1) Catégories de valeurs
        2) Catégories de modalités

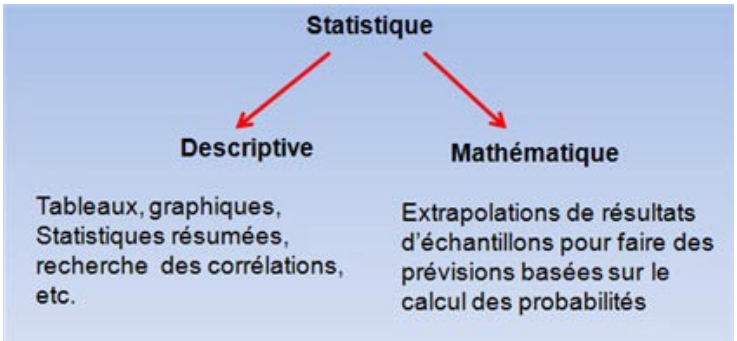
1 - Utilité de la statistique descriptive en économie

A – Définition

On divise généralement l'étude de la **statistique générale** en deux parties :

La **statistique descriptive**, qui est un ensemble de méthodes permettant de décrire les **unités statistiques** (voir la section 2 pour une définition plus précise du terme "unités statistiques") qui composent une **population** (voir la section 2 pour une définition plus précise du terme "population").

La **statistique mathématique** dont l'objet est de formuler des lois à partir de l'observation d'**échantillons**, c'est-à-dire de sous-ensemble d'une population statistique (voir la section 2 pour une définition plus précise du terme "population"). La statistique mathématique intervient dans les **enquêtes** et les **sondages**. Elle s'appuie sur la statistique descriptive, mais aussi sur le calcul des **probabilités**.



Enfin, **l'économétrie** est l'application de la statistique (descriptive et mathématique) à la mesure et à l'étude chiffrée des grandeurs économiques.

B - Exemples d'utilisation

Soit le tableau ci-après qui donnent quelques statistiques macroéconomiques des pays de l'UE à 27 ainsi que de certains de leurs partenaires commerciaux. Les chiffres contenus dans ce tableau permettent de répondre à de multiples questions.

Certaines des réponses sont immédiates, d'autres nécessitent de faire certains calculs ou bien de représenter les chiffres sous formes visuelles (graphique par exemple).

Supposons que l'on souhaite savoir **quel est le pays de l'UE à 27 qui a la superficie la plus élevée ?** La réponse est immédiate. Il suffit de regarder le tableau et de voir qu'il s'agit de la France avec une superficie égale à 643 427 km^2 (sachant que la France métropolitaine ne compte que 547 030 km^2, mais même dans ce cas, elle reste le plus grand pays de l'UE).

Bien qu'il suffise de regarder le tableau pour répondre à cette question, l'esprit est immédiatement attiré par la possibilité faire des comparaisons plus précises. De nombreuses autres questions peuvent alors être posées, mais qui vont nécessiter soit des calculs supplémentaires, soit un réagencement des données, soit une combinaisons de ces données avec d'autres données du tableau, etc.

Ainsi, on peut par exemple classer les pays du tableau (ou seulement ceux de l'UE à 27 par ordre de superficie croissante ou décroissante, ce qui permet alors d'un seul coup d'œil de se représenter l'ordre de grandeur des surfaces.

On peut également **calculer la surface totale** des pays de l'UE à 27 et ensuite diviser la surface de chaque pays par ce total et multiplier par cent. On obtient alors le **pourcentage** de la surface de chaque pays dans la surface de l'UE.

On peut aussi comparer la surface de l'UE à 27 avec la surface des Etats-Unis ou de la Chine, etc.

On peut encore, de façon très habituelle, comparer la superficie de chaque pays avec sa population. Par exemple, si on divise la population française totale en 2007 (métropolitaine et non métropolitaine) par la superficie totale de la France (métropolitaine et non métropolitaine), on obtient la densité de population ou nombre d'habitants par km2, soit :

$$\frac{63\,713\,926}{643427} = 99 \text{ habitants/km2}$$

On devra calculer ce chiffre pour tous les pays du tableau, ou seulement pour ceux de l'UE à 27 et faire une la moyenne des résultats obtenus. Ce qui permettra alors de savoir quel est l'écart de chaque pays par rapport à cette moyenne, etc.

Tableau 1 : Quelques statistiques macroéconomiques des pays de l'UE à 27 ainsi que de certains de leurs partenaires commerciaux

Pays	Monnaie	Superficie (km2)	Année d'adhésion	Nombre de frontières terrestres avec d'autres pays de l'UE à 27	Population (habitants) estimation de juillet 2007	Population (habitants) estimation de juillet 2006	Age médian (estimation de juillet 2007)	Indice de fécondité (en nombre d'enfants par femme, estimation 2007)	PIB en milliards d'USD (2006)	Exportations totales 2006 (milliards d'USD)	Exportations vers UE27 (2006) (Milliards d'USD)	Importations totales 2006 (milliards d'USD)	Importations venant de l'UE27 2006 (Milliards d'USD)
Allemagne	Euro	357 021	1956	8	82 400 996	82 422 299	43	1,4	2 875	1112	708	909	580
Belgique	Euro	30 528	1956	4	10 392 226	10 379 067	41,1	1,64	370	369	283	354	254
France	Euro	547 030	1956	5	63 713 926	63 195 000	39	1,98	2 151	490	320	535	369
Italie	Euro	301 320	1956	3	58 147 733	58 133 509	42,5	1,29	1 785	411	248	437	249
Luxembourg	Euro	2 586	1956	3	480 222	474 413	38,9	1,78	35	23	20	27	19
Pays-Bas	Euro	41 526	1956	2	16 570 613	16 491 461	39,7	1,66	613	462	367	416	207
Danemark	Couronne Danoise	43 094	1973	1	5 468 120	5 450 661	40,1	1,74	258	93	66	86	62
Irlande	Euro	70 280	1973	1	4 109 086	4 062 235	34,3	1,86	204	111	71	73	50
Royaume-Uni	Livre britannique	244 820	1973	1	60 776 238	60 609 153	39,6	1,66	2 346	448	282	619	361
Grèce	Euro (2001)	131 940	1981	1	10 706 290	10 688 058	41,2	1,35	224	21	13	63	36
Espagne	Euro	504 782	1986	2	40 448 191	40 397 842	40,3	1,29	1 084	205	146	316	193
Portugal	Euro	92 931	1986	1	10 642 836	10 605 870	38,8	1,48	177	43	34	67	50
Autriche	Euro	83 858	1995	6	8 199 783	8 192 880	41,3	1,37	310	140	102	140	112
Finlande	Euro	337 030	1995	1	5 238 460	5 231 372	41,6	1,73	198	77	44	69	44
Suède	Couronne suédoise	449 964	1995	1	9 031 088	9 016 596	41	1,66	373	147	89	127	88
Chypre	Euro (2008)	9 250	2004	0	788 457	784 301	35,1	1,8	16	1	1	7	5
Estonie	Couronne estonienne	45 226	2004	1	1 315 912	1 324 333	39,4	1,41	14	9	6	13	10
Hongrie	Forint	93 030	2004	4	9 956 108	9 981 334	38,9	1,33	113	74	59	77	54
Lettonie	Lat	64 589	2004	2	2 259 810	2 274 735	39,6	1,28	17	6	4	12	9
Lituanie	Litas	65 200	2004	2	3 575 439	3 585 906	38,6	1,21	30	14	9	19	12
Malte	Euro (2008)	316	2004	0	401 880	400 214	39	1,51	6	3	1	4	3
Pologne	Zloty	312 685	2004	4	38 518 241	38 536 869	37,3	1,26	337	110	87	126	92
République tchèque	Couronne tchèque	78 866	2004	4	10 228 744	10 235 455	39,5	1,22	119	95	81	93	75
Slovaquie	Couronne slovaque	48 845	2004	4	5 447 502	5 439 448	36,1	1,33	48	42	36	46	35
Slovénie	Euro (2007)	20 253	2004	3	2 009 245	2 022 331	41	1,26	38	23	16	24	19
Bulgarie	Lev	110 910	2007	2	7 322 858	7 385 367	40,9	1,39	28	15	9	23	14
Roumanie	Leu	238 391	2007	2	22 276 056	22 303 552	36,9	1,38	80	32	23	51	32
Suisse	Franc suisse	41 290		4	7 554 661	7 523 934	40,4	1,44	386	147	91	141	111
Etats-Unis	Dollar	9 826 630		0	301 139 947	298 444 215	36,6	2,09	13 160	1 038	215	1919	342
Chine	Yuan	9 596 960		0	1 321 851 741	1 313 973 713	33,2	1,75	2 527	969	190	791	91
Inde	Roupie	3 287 592		0	1 129 866 154	1 095 351 995	24,8	2,81	806	120	26	175	3
Japon	Yen	377835		0	127 433 494	127 463 611	43,5	1,23	4 883	650	95	580	60
Russie	Rouble	17 075 200		5	141 377 752	142 893 540	38,2	1,39	734	305	71	164	136
Taiwan	Dollar taïwanais	35 980		0	22 858 872	23 036 087	35,5	1,12	347	224	31	203	27
Hong Kong	Dollar de Hong Kong	1 092		0	6 980 412	6 940 432	41,2	0,98	189	323	47	336	24
Monde		510 072 000			6 602 224 175	6 525 170 264	28	2,59	46 770	12248		12248	

Notes et sources :

a - Les données concernant la superficie, le nombre de frontières terrestres avec les pays de l'UE27, la population, l'âge médian, l'indice de fécondité, le PIB sont extraites du World Fact Book, https://www.cia.gov/library/publications/the-world-factbook/index.html

b - Les données concernant les exportations et les importations totales de marchandises proviennent des statistiques de l'OMC : http://www.wto.org/french/res_f/statis_f/its2007_f/its07_toc_f.htm

c - Les chiffres des exportations vers les pays de l'UE à 27 et des importations en provenance des pays de l'UE à 27 ont été obtenus à partir des pourcentages donnés par Eurostat et ne concernent que les marchandises (pas les services commerciaux)

d - Pour l'Inde et la Russie, les chiffres des exportations vers l'UE et des importations en provenance de l'UE se rapportent à 2005 et ne concernent que l'UE 25.

e - Pour Taiwan et Hong-Kong les chiffres des exportations vers l'UE et des importations en provenance de l'UE ne concernent que l'UE25 et sont extraites des statistiques de l'OMC.

f - La coincidence entre le total des exportations et des importations mondiales a été calculé en faisant une moyenne simple des deux chiffres.

Bien souvent, pour répondre à certaines questions, les calculs précédents ne suffiront pas, où bien, s'ils suffisent, il faudra aussi créer un autre tableau, pour faire apparaître plus précisément certaines informations.

Supposons par exemple que l'on souhaite avoir une idée synthétique sur la question suivante : Combien y-a-t-il de pays qui sont membres de la zone Euro au premier janvier 2008, quels sont ces pays, combien représentent-ils en pourcentage du total des pays, et quels sont les autres pays.

Pour répondre à toutes ces questions, il faudra faire quelques calculs et ensuite récapituler ces résultats dans un **tableau** (pour plus de détails sur les tableaux, voir la section 3 de ce chapitre, ainsi que le chapitre 2 du cours) ou dans un **graphique**, ou encore sur une **carte**. Supposons ici, que pour simplifier, on se contente du tableau suivant :

Répartition des pays de l'UE à 27 entre membres et non-membres de la zone Euro au premier janvier 2008

		Nombre	Pourcentages
Pays membres de la zone euro au 1er janvier 2008	1999 : Allemagne, Belgique, France, Italie, Luxembourg, Pays-Bas, Irlande, Espagne, Portugal, Autriche, Finlande; 2001 : Grèce; 2007 : Slovénie; 2008 : Chypre et Malte	15	55,6
Pays non membres de la zone euro au 1er janvier 2008	Danemark, Estonie, Slovaquie, Suède, République Tchèque, Hongrie, Lettonie, Roumanie, Bulgarie,	12	44,4
Total		27	100,0

Ce tableau a donc nécessité quelques calculs statistiques simples :

Repérage des pays membres et non-membres
Comptage des pays appartenant à chaque catégorie
Calcul des pourcentages

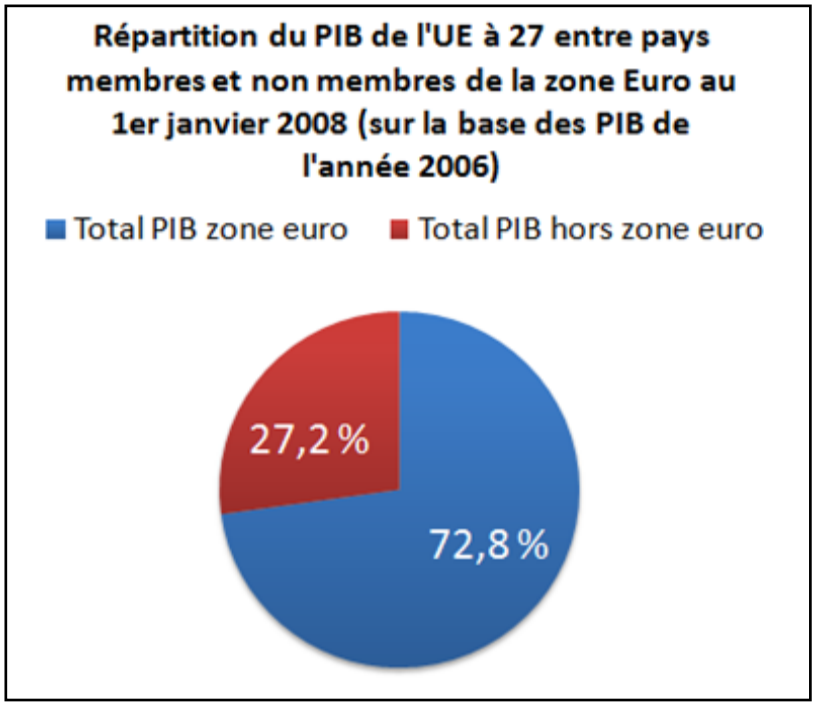
Des calculs plus complexes (mais aussi plus intéressants) peuvent être faits. Par exemple, plutôt que de calculer combien les pays membres et non-membres représentent en pourcentage des 27 de l'UE, on peut, sans doute de façon plus pertinente, se demander combien chaque groupe représente en pourcentage du Produit Intérieur Brut total de l'UE27. Pour obtenir un tel tableau (voir tableau ci-après), il faudra :

1) additionner les PIB des 15 pays membres de la zone euro au 1er janvier 2008
2) Additionner les PIB des 12 pays non membres de la zone euro au 1er janvier 2008.
3) calculer les pourcentages respectifs.

On obtient alors le tableau ci-après :

Pays	Monnaie	PIB en milliards d'USD (2006)	Pourcentages
Allemagne	Euro	2 875	
Belgique	Euro	370	
France	Euro	2 151	
Italie	Euro	1 785	
Luxembourg	Euro	35	
Pays-Bas	Euro	613	
Irlande	Euro	204	
Espagne	Euro	1 084	
Portugal	Euro	177	
Autriche	Euro	310	
Finlande	Euro	198	
Grèce	Euro (2001)	224	
Slovénie	Euro (2007)	38	
Chypre	Euro (2008)	16	
Malte	Euro (2008)	6	
Total PIB zone euro		**10 085**	**72,8**
Danemark	Couronne Danoise	258	
Royaume-Uni	Livre britannique	2 346	
Suède	Couronne suédoise	373	
Estonie	Couronne estonienne	14	
Hongrie	Forint	113	
Lettonie	Lat	17	
Lituanie	Litas	30	
Pologne	Zloty	337	
République tchèque	Couronne tchèque	119	
Slovaquie	Couronne slovaque	48	
Bulgarie	Lev	28	
Roumanie	Leu	80	
Total PIB hors zone euro		**3 762**	**27,2**
Total PIB UE 27		**13 847**	

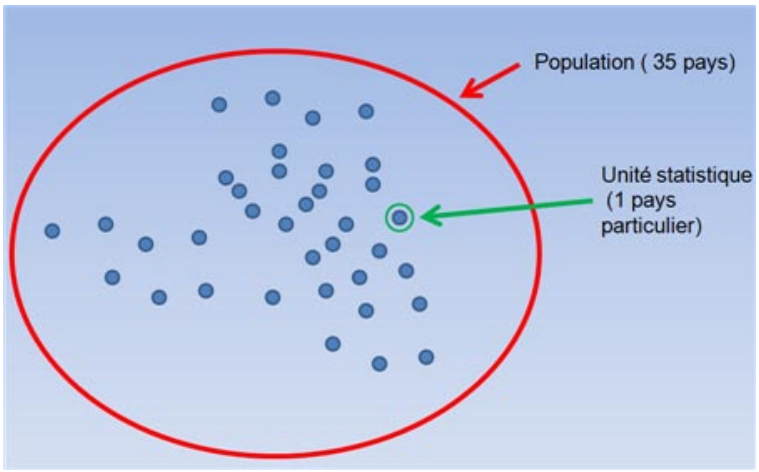
Plutôt que de représenter le résultat sous forme d'un tableau on peut choisir une **représentation visuelle**, par exemple une représentation en secteurs, dite aussi parfois en « **camembert** » :



2 - Terminologie

A – Population et unités statistiques

En statistique, la **population** désigne un ensemble d'**unités statistiques**. Les unités statistiques sont les entités abstraites qui représentent des personnes, des populations d'animaux ou des objets. Les premières populations ayant fait l'objet d'un recensement ayant été des populations humaines (d'où le lien étroit entre statistique et démographie) on emploie fréquemment le terme "individus" comme synonyme de "unités statistiques".



La statistique sert à décrire l'ensemble des unités statistiques qui composent la population. On commence par compter ces unités. La première information statistique que l'on tire d'une population est en effet le nombre de ses unités.

Exemple 1 : La population de la France, de ses régions, de ses départements, de ces communes de moins de 10 000 habitants et des communes de plus de 10 000 habitants (« grandes villes ») est estimée annuellement par l'INSEE. Les résultats sont disponibles sur son site internet[1]. On sait ainsi que les populations des 3 plus grandes villes de France sont, selon la dernière estimation publiée en janvier 2007 (qui porte sur l'année 2005) :

Villes	Nombre d'habitants (« population » ou « unités statistiques »)
Paris	2 153 600
Marseille	820 900
Lyon	466 400

Source : http://www.insee.fr/fr/recensement/nouv_recens/resultats/grandes-villes.htm#L

Exemple 2 : Le tableau 1 contient une population de 35 pays, donc 35 unités statistiques.

[1] Voir le lien suivant : http://www.insee.fr/fr/recensement/nouv_recens/resultats/premiers-resultats-recensement.htm

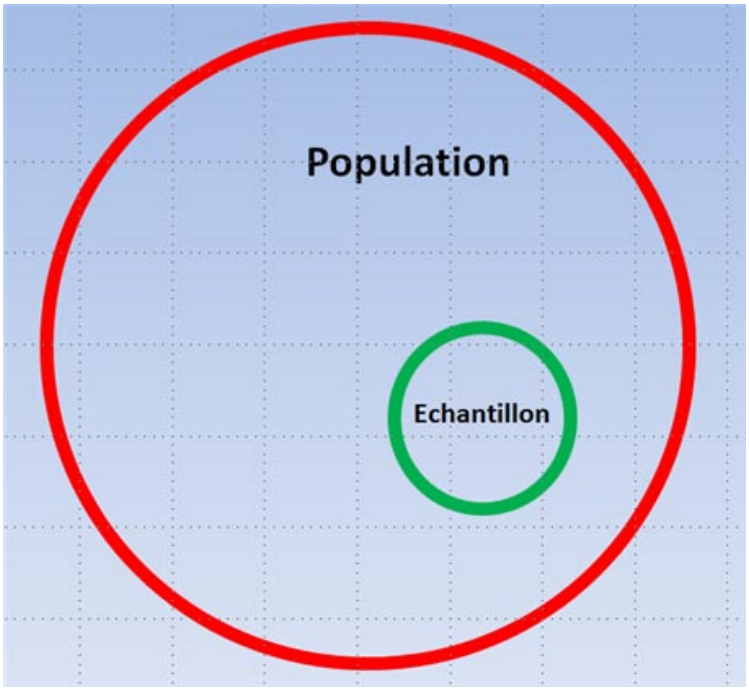
B – Echantillons et sous-ensembles d'une population

Il est fréquent que l'on prélève un **échantillon** dans une population statistique et/ou que l'on découpe la population selon certains **critères** (ou **dimensions** ou encore **caractéristiques**)

Pour comprendre le lien entre population et l'échantillon, prenons l'exemple du recensement de la population française. Chaque année, la population est réévaluée à partir des résultats d'une enquête qui porte sur le choix d'un échantillon. Ainsi, dans les 900 communes de 10 000 habitants ou plus, une partie de la population est recensée chaque année par tirage au sort (8%). Il y a donc un **sondage annuel** qui aboutit à recenser 40% de la population de ces villes en 5 ans. Cette opération est en fait un sondage à grande échelle. Concrètement, une ville de plus de 10 000 habitants est divisée en cinq groupes d'adresses réparties sur tout le territoire de la commune. Chaque année, l'INSEE prélève un échantillon de 8% d'adresses dans un des cinq groupes et on le recense. La détermination des échantillons de personnes interrogées est effectuée en utilisant les fichiers de taxe d'habitation et les registres d'assurance-maladie, ce qui permet l'extrapolation avec une grande fiabilité des données des sondages. Ainsi, tous les habitants d'une même rue ne seront pas recensés la même année[2].

Pour notre propos, la relation de la population à l'échantillon est facile à décrire à partir d'un **diagramme d'EULER** suivant.

Le lien entre l'échantillon et la population



[2] Pour plus de détails, voir le document de l'INSEE, Le plan de sondage dans les communes de 10 000 habitants ou plus, INSEE Méthodes, Pour comprendre le recensement de la population, numéro hors série.

En général, on parle d'échantillon d'une population statistique quand les unités statistiques sont tirées au sort ou alors choisies par une méthode qui permet d'assurer la représentativité de l'échantillon par rapport à la population totale. Cependant, ces définitions ne concernent plus directement la **statistique descriptive** mais plutôt la **statistique mathématique**.

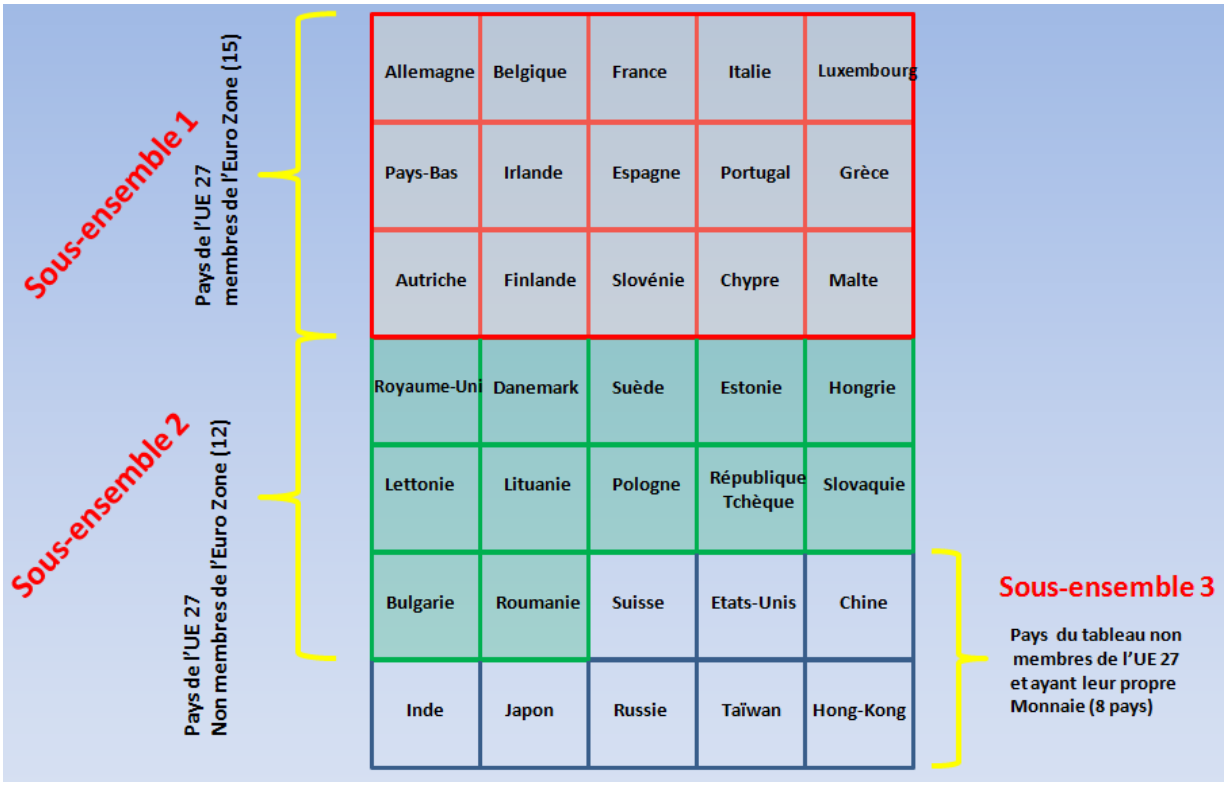
Ce qui nous intéresse ici, c'est la possibilité de « découper » une population en sous-populations en utilisant certains critères.

Prenons pour exemple la population des 35 pays du tableau 1. Ces 35 pays sont les unités statistiques du tableau. Nous souhaitons par exemple « découper » cette population entre trois sous ensembles, suivant les critères de la monnaie utilisée et l'appartenance à l'UE 27. On aura donc :

15 pays membres de l'UE 27 qui font partie de la zone Euro.
12 pays membres de l'UE à 27 qui ne font pas (encore) partie de la zone Euro
8 pays partenaires de l'UE 27 et qui utilisent d'autres monnaies.

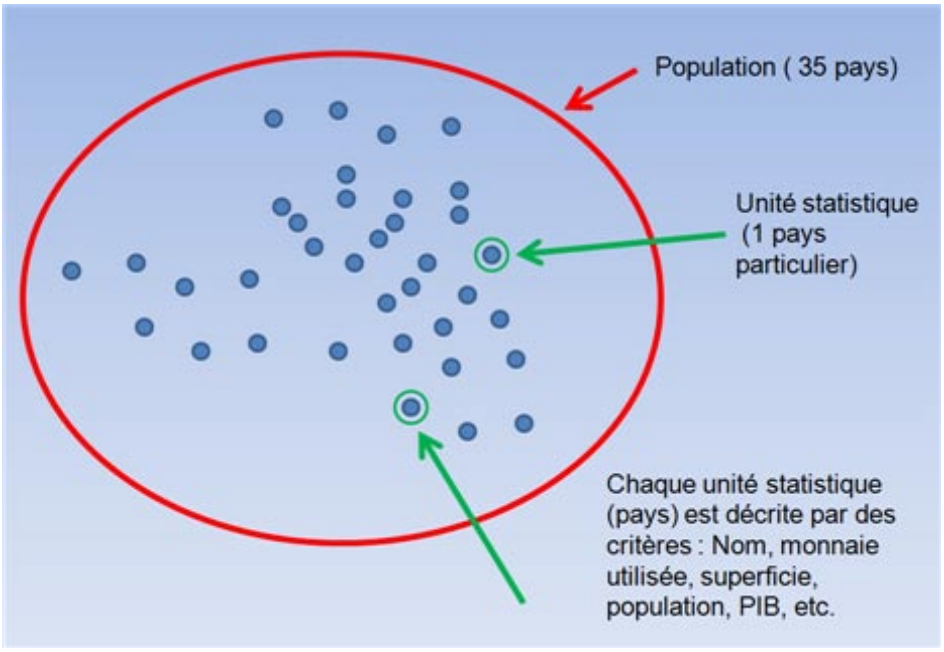
Pour concrétiser ce découpage en 3 sous populations, on peut par exemple construire un rectangle contenant 35 petits carrés, chaque carré représentant un pays. Puis « découper » les trois zones comme dans le graphique ci-dessous.

Découpage d'une population statistique (les 35 pays du tableau) en trois sous-populations, suivant deux critères (appartenance à l'UE27 et monnaie officielle).



C - Critères de classification

Nous avons vu dans l'exemple précédent que les unités statistiques d'une population pouvaient être regroupées suivant des **critères** ou **dimensions**. Ces critères sont choisis en fonction de ce qui nous intéresse. On parle de critère, mais aussi parfois de dimension.



On distingue deux sortes de critères :

- Les critères quantitatifs
- Les critères qualitatifs

1) Critères quantitatifs

Les **critères quantitatifs** sont les critères qui sont représentés par des chiffres. C'est la raison pour laquelle on les appelle aussi parfois des **variables**. Les variables prennent des **valeurs**.

Par exemple, dans le tableau 1 , on peut voir que la superficie est un critère de classification quantitatif. C'est une variable qui dont les différentes **occurrences** sont appelées **valeurs**. Chacune des 35 unités statistiques de notre population est ainsi caractérisée par une valeur. La superficie est donc ici une variable qui prend 35 valeurs différentes. C'est un cas particulier dans lequel le nombre de valeurs de la variable est égal au nombre des unités statistiques de la population. Nous verrons que dans des cas de ce type, ou bien lorsque le nombre de valeurs possibles, bien qu'inférieur au nombre d'unités statistiques, est grand, un regroupement par classes de valeurs peut être utile.

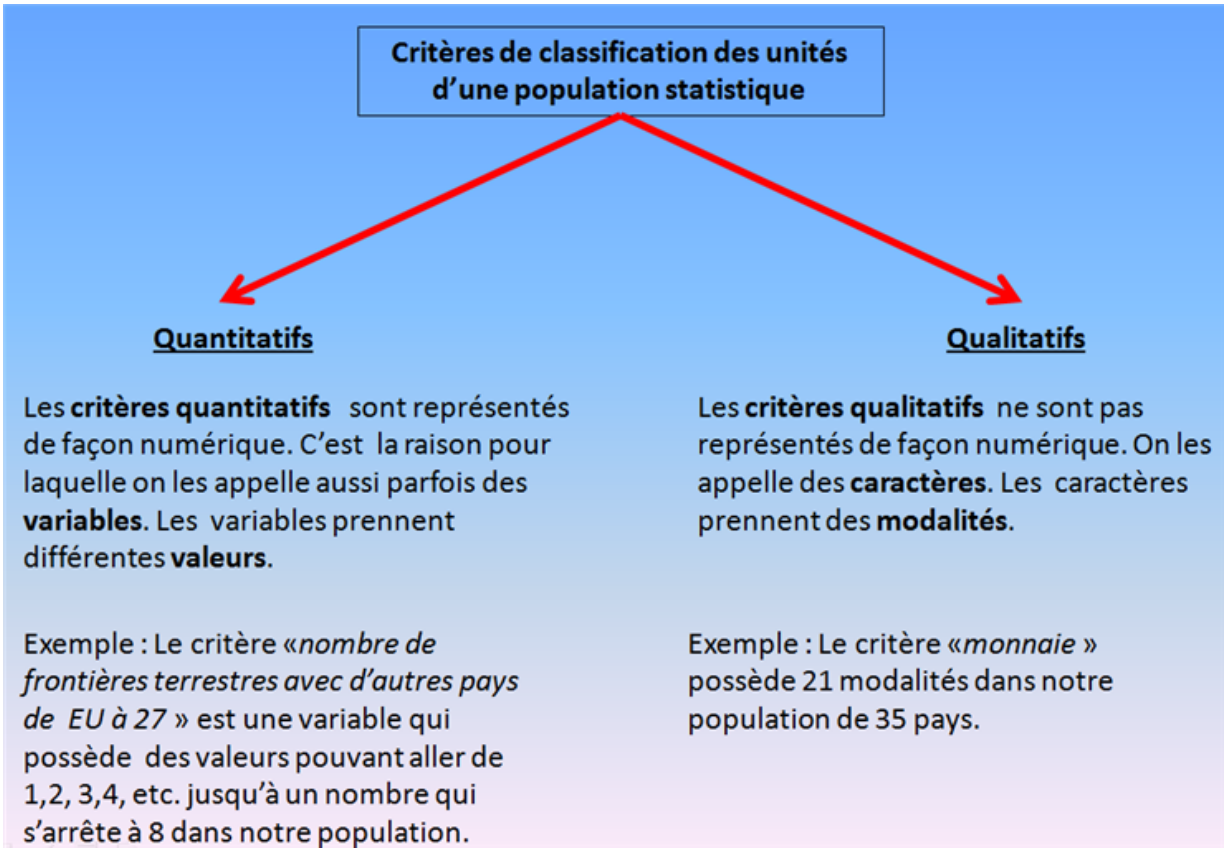
Les critères quantitatifs ou « variables » permettent de faire des calculs. On peut par exemple additionner les superficies, calculer des moyennes, etc.

Dans le tableau 1, la plupart des critères sont quantitatifs. On peut donc effectuer des calculs sur les valeurs. Il n'y a guère que la monnaie et le nom des pays qui ne sont pas des critères quantitatifs. Ce sont des critères qualitatifs.

2) Critères qualitatifs

Les **critères qualitatifs** sont tous les critères qui ne sont pas représentés de façon numérique. On les appelle des « caractères ». Les caractères prennent des **modalités**.

Par exemple, dans le tableau 1, on peut voir que la monnaie utilisée dans chaque pays est un critère qualitatif qui possède 21 modalités. Ces modalités sont les différentes monnaies. Il y a en effet 15 pays qui ont la modalité « euro » et les 20 autres qui ont chacun pour modalité une monnaie différente. On voit donc dans cette exemple que le nombre de modalités (21) est inférieur à celui de la population (35).



3 - Modes de regroupement des données

A - Série simple

Le tableau 1 est un tableau dans lequel les données n'ont pas été regroupées. C'est un tableau de données brutes. Nous pouvons lire pour chaque ligne les différentes valeurs ou modalités des variables ou des caractères associés à chacune des 35 unités statistiques de la population. **Chaque colonne correspond à une série simple de valeurs ou de modalités**.

Par exemple, dans le cas de la variable « superficie », il y a 35 valeurs différentes. Dans le cas du caractère « monnaie », il y a 21 modalités. Dans le cas de la variable « nombre de frontières terrestres avec d'autres pays de l'UE à 27 », les 35 pays se répartissent seulement sur 8 valeurs.

Mais on comprend facilement qu'une présentation exhaustive, dans laquelle aucun regroupement n'est effectué, n'est pas toujours pratique, même si l'on dispose d'un ordinateur, comme c'est le cas aujourd'hui, pour effectuer les calculs. En réalité, le plus souvent, les données sont collectées et entrées dans l'ordinateur sous forme d'un tableau brut de ce type (ou sous une autre forme), mais ensuite, elles sont regroupées.

B - Distribution par valeurs ou par modalités

Suivant que le critère est une variable ou un caractère, on peut effectuer un regroupement par valeurs ou un regroupement par modalités. Dans ce cas, on parle de distribution. En effet, la série initiale des 35 données va être distribuée sur un nombre généralement inférieur (ou au maximum égal), de valeurs ou de modalités.

1) Distribution par valeurs

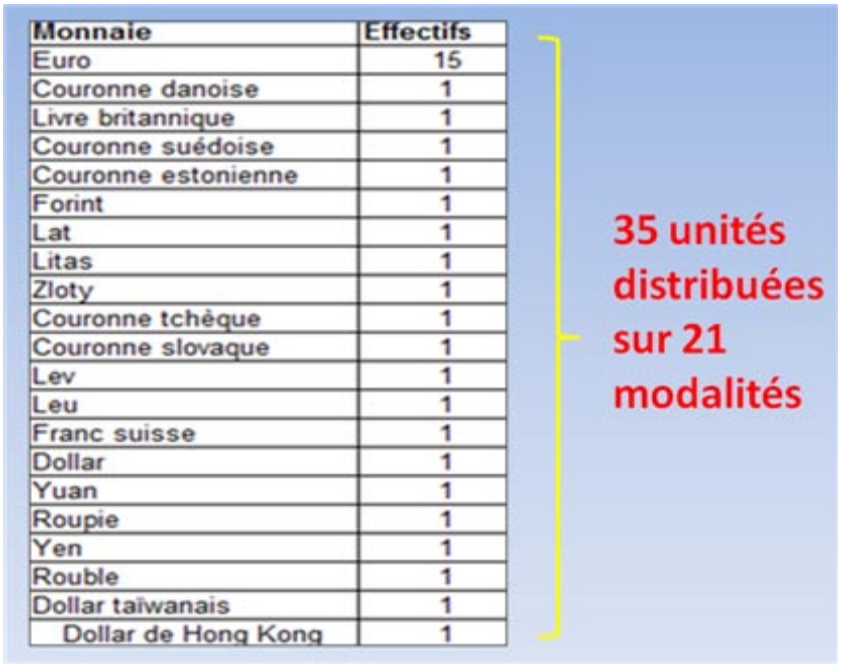
Prenons l'exemple de la variable « nombre de frontières terrestres avec d'autres pays de l'UE à 27 » dans le tableau 1. Un regroupement des 35 unités statistiques pour chacune des valeurs possibles de la variable donnera alors le tableau suivant :

Distribution des pays des pays du tableau 1
selon leur nombre de frontières terrestres avec les pays de l'UE à 27

	Nombre de frontières terrestres avec d'autres pays de l'UE à 27	Effectifs	
Distribution sur 9 valeurs (la valeur 7 pourrait éventuellement être retirée du tableau, mais cela créerait une discontinuité)	0	8	
	1	8	
	2	6	
	3	3	
	4	6	
	5	2	
	6	1	
	7	0	
	8	1	
		35	← La somme des effectifs de la distribution est égal à 35, la population totale

2) Distribution par modalités

Dans le tableau 1, nous allons choisir le seul critère qualitatif disponible pour effectuer un regroupement par modalités : la monnaie officielle utilisée dans chaque pays. On sait évidemment le résultat d'avance : En 2008, 15 pays sont dans la zone euro et les 20 autres utilisent toujours leur monnaie nationale. Dans ces conditions, un regroupement par modalités, quoique peu utile, donnerait le résultat suivant :



Monnaie	Effectifs
Euro	15
Couronne danoise	1
Livre britannique	1
Couronne suédoise	1
Couronne estonienne	1
Forint	1
Lat	1
Litas	1
Zloty	1
Couronne tchèque	1
Couronne slovaque	1
Lev	1
Leu	1
Franc suisse	1
Dollar	1
Yuan	1
Roupie	1
Yen	1
Rouble	1
Dollar taïwanais	1
Dollar de Hong Kong	1

C - Regroupement par catégories

Lorsqu'il y a trop de valeurs ou trop de modalités, on peut procéder à un regroupement par catégories de valeurs ou de modalités.

1) Catégories de valeurs

Prenons l'exemple de la variable « superficie » dans le tableau 1. Un regroupement des 35 unités statistiques pour chacune des valeurs possibles de la variable donnerait un tableau avec 35 valeurs, ce qui n'aurait aucun intérêt. En revanche, on peut créer des classes de valeurs pour les superficies et répartir les 35 pays à l'intérieur de ces classes. Comment procéder sachant que le plus petit (Malte) n'a qu'une superficie de 316 km^2 et le plus grand pays (La Russie) a une superficie de 17 075 200 km^2 ? Si l'on regarde les superficies des différents pays, on voit qu'un très grand nombre de pays ont des superficies inférieures à 600 000 km^2. Pour le faire apparaître, classons les pays par ordre croissant de superficies (voir le tableau ci-après)

Regroupement des pays par catégories de superficies

Pays	Superficie (km2)
Malte	316
Hong Kong	1 092
Luxembourg	2 586
Chypre	9 250
Slovénie	20 253
Belgique	30 528
Taiwan	35 980
Suisse	41 290
Pays-Bas	41 526
Danemark	43 094
Estonie	45 226
Slovaquie	48 845
Lettonie	64 589
Lituanie	65 200
Irlande	70 280
République tchèque	78 866
Autriche	83 858
Portugal	92 931
Hongrie	93 030
Bulgarie	110 910
Grèce	131 940
Roumanie	238 391
Royaume-Uni	244 820
Italie	301 320
Pologne	312 685
Finlande	337 030
Allemagne	357 021
Japon	377835
Suède	449 964
Espagne	504 782
France	547 030
Inde	3 287 592
Chine	9 596 960
Etats-Unis	9 826 630
Russie	17 075 200

[0 -50 000]

]50 000 - 100 000]

]100 000 - 600 000]

]600 000 - 18 000 000]

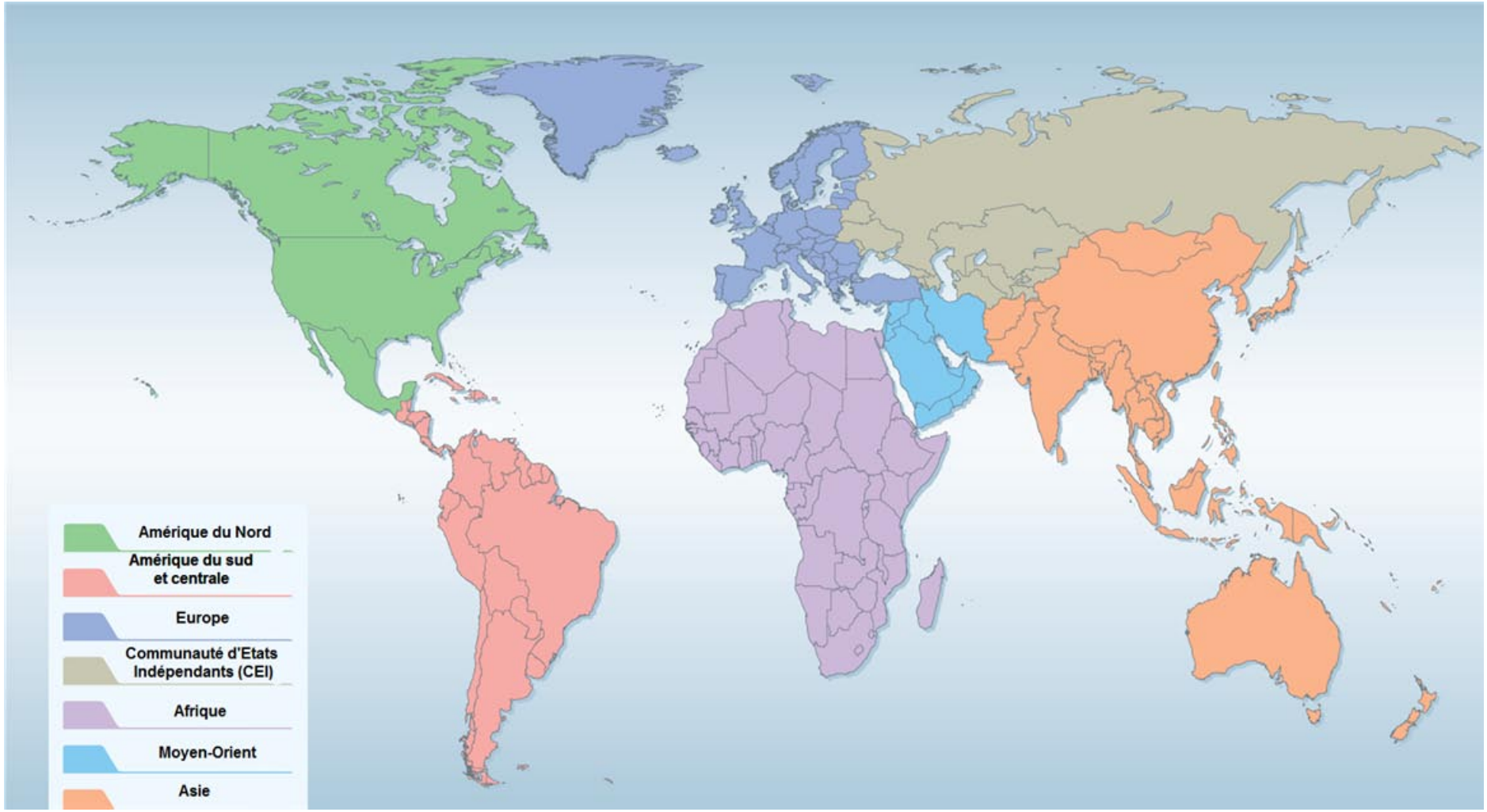
Transformation de la série simple en distribution par classes de valeurs

Les classes sont d'amplitudes inégales

Superficie (km2)	Effectifs	Amplitude de classe (km2)
[0 -50 000]	12	50 000
]50 000 - 100 000]	7	50 000
]100 000 - 600 000]	12	500 000
]600 000 - 18 000 000]	4	17 400 000

L'amplitude d'une classe est égale à la différence entre l'extrémité supérieure et l'extrémité inférieure

Découpage du monde en zones selon les catégories de l'OMC



Source : OMC, http://www.wto.org/english/res_e/statis_e/its2007_e/its07_world_maps_e.pdf

2) Catégories de modalités

Prenons l'exemple du caractère « pays » dans le tableau 1. Un regroupement des 35 unités statistiques pour chacune des modalités possibles du caractère donnerait un tableau avec 35 modalités, ce qui n'aurait aucun intérêt. En revanche, on peut créer des classes de modalités pour les pays. On peut par exemple répartir les 35 pays selon catégories proposée par l'Organisation Mondiale du Commerce (voir carte précédente). Si l'on regroupe nos 35 pays selon ces 6 catégories on obtient le tableau suivant :

Regroupement des pays selon des catégories géographiques

Catégories continentales	Effectifs	Pays inclus
Europe	28	UE à 27 + Suisse
Amérique du Nord	1	Etats-Unis
Communauté d'Etats Indépendants	1	Russie
Afrique	0	
Moyen-Orient	0	
Asie	5	Chine, Inde, Japon, Taïwan, Hong-Kong

A noter qu'il s'agit bien d'un regroupement par catégories de modalités car chaque pays est en lui-même une modalité.

Chapitre 2
Les tableaux statistiques

1 – Séries brutes ou vecteurs
- A - séries classées et non classées
- B - Séries identifiées et non identifiées

2 – Tableaux unidimensionnels
- A - Tableaux avec chiffres bruts
- B - Tableaux avec pourcentages
- C - Tableaux avec cumuls
  1) Cumuls des données brutes
  2) Cumuls des pourcentages

3 - Tableaux avec statistiques résumées

4 – Tableaux croisés
- A – Définition et exemple
  1) Définition
  2) Exemple
     a) Effectifs
     b) Pourcentages
- B – Distributions marginales
  1) Définition
  2) Exemple
     a) Effectifs
     b) Pourcentages
- C – Distributions conditionnelles
  1) Colonnes
     a) Effectifs
     b) Pourcentages
  2) Lignes
     a) Effectifs
     b) Pourcentages

1 – Séries brutes ou vecteurs

Avant même d'être présentées sous forme de tableau, les données sont parfois présentées sous formes de séries brutes.

Prenons l'exemple de la variable « nombre de frontières terrestres avec d'autres pays de l'UE à 27 » dans le tableau 1. On peut la représenter sous la forme d'un **vecteur de données**, également appelé **série.**

Série « nombre de frontières terrestres avec d'autres pays de l'UE à 27 » :

S1 = {8, 4, 5, 3, 3, 2,1, 1, 1, 1, 2, 1 ,6, 1, 1, 0, 1, 4, 2, 2, 0, 4, 4, 4, 3, 2, 2, 4, 0, 0, 0, 0, 5, 0,0}

A - séries classées et non classées

S1 est une **série non classée**. Considérons maintenant la série S2, qui elle, est une **série classée par ordre croissant**

S2 : {0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 6, 8}

B - Séries identifiées et non identifiées

En revanche, ni S1, ni S2 ne sont des **séries identifiées**. Pour qu'elles soient identifiées, il faudrait créer des couples de valeurs.

Ainsi, la série S3 ci-dessous est une **série identifiée, non classée** :

S3 = {{Allemagne, 8}, {Belgique, 4}, {France, 5}, {Italie, 3}, {Luxembourg, 3}, {Pays-Bas, 2}, {Danemark, 1}, {Irlande, 1}, {Royaume-Uni, 1}, {Grèce, 1}, {Espagne, 2}, {Portugal, 1}, {Autriche, 6}, {Finlande, 1}, {Suède, 1}, {Chypre, 0}, {Estonie, 1}, {Hongrie, 4}, {Lettonie, 2}, {Lituanie, 2}, {Malte, 0}, {Pologne, 4}, {République tchèque, 4}, {Slovaquie, 4}, {Slovénie, 3}, {Bulgarie, 2}, {Roumanie, 2}, {Suisse, 4}, {Etats-Unis, 0}, {Chine, 0}, {Inde, 0}, {Japon, 0}, {Russie, 5}, {Taiwan, 0}, {Hong Kong, 0}}

Enfin, la série S4 ci-dessous est une **série identifiée et classée** par ordre croissant du nombre de frontières terrestres avec d'autres pays de l'UE à 27 :

S4 = {{Chypre, 0}, {Malte,0}, {Etats-Unis,0}, {Chine,0},{Inde,0}, {Japon,0}, {Taiwan,0}, {Hong Kong, 0}, {Danemark,1}, {Irlande,1}, {Royaume-Uni,1}, {Grèce,1}, {Portugal, 1}, {Finlande,1}, {Suède,1}, {Estonie,1}, {Pays-Bas, 2}, {Espagne, 2}, {Lettonie, 2}, {Lituanie, 2}, {Bulgarie, 2}, {Roumanie,2}, {Italie ,3}, {Luxembourg, 3}, {Slovénie, 3}, {Belgique,4}, {Hongrie, 4}, {Pologne,4}, {République tchèque,4}, {Slovaquie, 4}, {Suisse,4}, {France, 5 }, {Russie, 5}, {Autriche,6}, {Allemagne, 8}}

2 – Tableaux unidimensionnels

La présentation sous forme de série est utile pour certains calculs, mais on utilise bien plus fréquemment les tableaux pour présenter les caractéristiques des unités d'une population statistique.

Le tableau est un outil statistique rébarbatif. La meilleure preuve est que l'on voit beaucoup moins de tableaux dans les médias que l'on ne voit de graphiques.

Néanmoins, pour des études précises, les tableaux sont souvent nécessaires et même plus utiles que les graphiques. Une remarque simple s'impose en effet ici : si l'on dispose d'un tableau, on peut faire un graphique. Inversement, si l'on dispose seulement d'un graphique, on ne peut pas revenir au tableau initial (ou très rarement). **Le tableau est donc une sorte de préalable au graphique**.

En outre :

- il est souvent possible de produire plusieurs graphiques à partir d'un même tableau.

- Il est aussi souvent possible de calculer plusieurs statistiques résumées à partir d'un tableau.

Ainsi, si le tableau est rébarbatif, il est cependant très utile.

A - Tableaux avec chiffres bruts

Le tableau 1 est un tableau de chiffres bruts qui contient plusieurs séries de chiffres caractérisant une population de 35 unités statistiques.

Le tableau ci-après, quant à lui, est également un tableau qui présente des chiffres bruts extraits du tableau 1 et regroupés par classes :

Distribution des 35 pays par classes de superficie
Chiffres bruts

Superficie (km2)	Effectifs
[0 -50 000]	12
]50 000 - 100 000]	7
]100 000 - 600 000]	12
]600 000 - 18 000 000]	4

B - Tableaux avec pourcentages

Souvent, on ajoute une colonne dans laquelle les données sont présentées en **pourcentages** comme ci-dessous :

Distribution des 35 pays par classes de superficie
Pourcentages

Superficie (km2)	Effectifs	Pourcentages	
[0 -50 000]	12	34,3	(12/35)*100
]50 000 - 100 000]	7	20,0	(7/35)*100
]100 000 - 600 000]	12	34,3	(12/35)*100
]600 000 - 18 000 000]	4	11,4	(4/35)*100
Total	35	100	

Effectif total

C - Tableaux avec cumuls

On peut aussi ajouter des colonnes avec les **cumuls**. Une colonne où les chiffres bruts sont cumulés et une autre où ce sont les pourcentages qui sont cumulés.

1) Cumuls des données brutes

Distribution des 35 pays par classes de superficie
Chiffres bruts et cumuls

Superficie (km2)	Effectifs	Effectifs cumulés
[0 -50 000]	12	12
]50 000 - 100 000]	7	19
]100 000 - 600 000]	12	31
]600 000 - 18 000 000]	4	35
Total	35	

2) Cumuls des pourcentages

Distribution des 35 pays par classes de superficie
Chiffres bruts et cumulés, pourcentages et pourcentages cumulés

Superficie (km2)	Effectifs	Effectifs cumulés	Pourcentages	Pourcentages cumulés
[0 -50 000]	12	12	34,3	34,3
]50 000 - 100 000]	7	19	20,0	54,3
]100 000 - 600 000]	12	31	34,3	88,6
]600 000 - 18 000 000]	4	35	11,4	100
Total	35		100	

3 - Tableaux avec statistiques résumées

Parfois, on préfère *résumer une série de chiffres par son total, par sa moyenne, par ses valeurs extrêmes, ou par différentes autres statistiques* que nous étudierons dans le chapitre 3. Le tableau ci-après indique les **moyennes simples** (voir le chapitre 5 pour une définition plus précise de la moyenne simple) de certaines des variables du tableau 1, ainsi que les valeurs minimales et maximales pour les séries correspondantes.

Tableau de statistiques résumées pour certaines des variables du tableau 1

	Superficie (km2)	Nombre de frontières terrestres avec d'autres pays de l'UE à 27	Population (habitants) estimation de juillet 2006	Age médian (estimation de juillet 2007)	Indice de fécondité (en nombre d'enfants par	PIB en milliards d'USD (2006)	Densité (Nombre d'habitants au km2) (2006)	PIB par habitant (en dollars) (2006)
Moyenne simple		2,20		38,80	1,52	1 053,66	357,91	24 975,06
Valeur minimale	316 (Malte)	0,00	400 214 (Malte)	24,8 (Inde)	0,98 (HK)	6 (Malte)	(8) Russie	735 (Inde)
Valeur maximale	17 075 200 (Russie)	8 (Allemagne)	1 313 973 713 (Chine)	43,5 (Japon)	2,81 (Inde)	13 160 (EU)	6356 (HK)	72785 (Luxembourg)

Note : Certaines moyennes ne sont pas reportées (superficie, population) car la dispersion est trop grande pour que la moyenne ait un sens.

4 -Tableaux croisés

A – Définition et exemple

1) Définition

Les tableaux croisés sont appelés ainsi car ils « croisent » deux distributions au sein d'un même tableau. Les possibilités de croisement sont multiples. En fait, comme l'illustre le tableau synoptique ci-après, il y a 16 possibilités.

Différentes possibilités de croisement de 2 distributions

1er critère / 2ème critère		Valeurs		Modalités	
		Simples	Regroupées	Simples	Regroupées
Valeurs	Simples	1	2	3	4
	Regroupées	5	6	7	8
Modalités	Simples	9	10	11	12
	Regroupées	13	14	15	16

2) Exemple

a) Effectifs

Le tableau croisé ci-après illustre le cas numéro 6. Les variables « Age médian » et « indice de fécondité » du tableau 1 ont en effet été regroupées par catégories de valeurs puis croisées dans le tableau. On a choisi de mettre les *catégories d'âges médians en lignes et les catégories d'indices de fécondité en colonne*, mais l'inverse aurait également été possible sans que cela ne change la signification du tableau.

Avant de construire le tableau croisé, on regarde les valeurs minimales et maximales des deux séries. On voit alors que l'âge médian varie de 24,8 ans (Inde) à 43,5 ans (Japon) et que l'indice de fécondité varie de 0,98 enfants par femme (Hong Kong) à 2,81 enfants par femme (Inde).

Il reste alors à former les catégories. **Il s'agit d'un choix arbitraire**. Pour simplifier, nous allons former les catégories suivantes :

Age médian (3 catégories) : [20-30 ans [; [30-40 ans [;[40-50]

Indice de fécondité (3 catégories) : [0-1,4 enfants/femme [; [1,4-2 enfants/femme [; [2-3 enfants/femme].

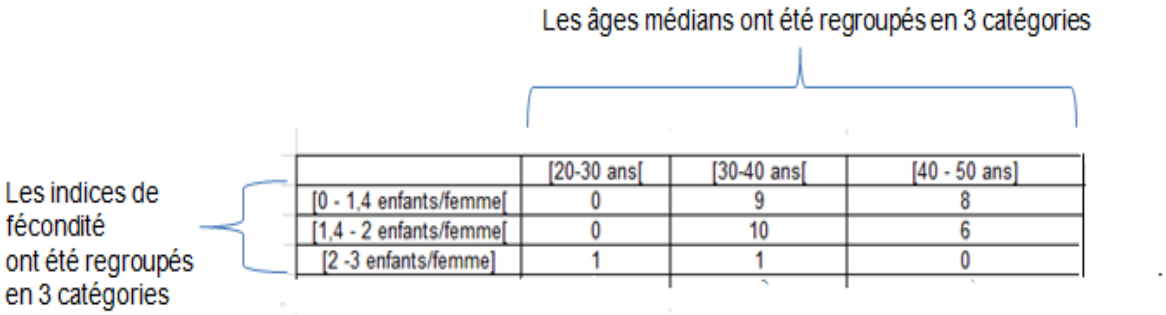
Ensuite on choisit une des 2 séries, par exemple la série des indices de fécondité en on la classe par ordre croissant.

Indice de fécondité (en nombre d'enfants par femme)	Age médian (estimation de juillet 2007)
0,98	41,2
1,12	35,5
1,21	38,6
1,22	39,5
1,23	43,5
1,26	37,3
1,26	41
1,28	39,6
1,29	40,3
1,29	42,5
1,33	36,1
1,33	38,9
1,35	41,2
1,37	41,3
1,38	36,9
1,39	38,2
1,39	40,9
1,4	43
1,41	39,4
1,44	40,4
1,48	38,8
1,51	39
1,64	41,1
1,66	39,6
1,66	39,7
1,66	41
1,73	41,6
1,74	40,1
1,75	33,2
1,78	38,9
1,8	35,1
1,86	34,3
1,98	39
2,09	36,6
2,81	24,8

On forme les 3 groupes de fécondité, en utilisant par exemple des couleurs différentes pour chaque groupe. Ensuite, il suffit de compter pour chaque groupe, combien de pays ont un âge médian compris dans les trois catégories d'âge médian que nous avons défini : [20-30 ans [; [30-40 ans [; [40-50]

On obtient alors le tableau suivant :

Tableau croisé « indice de fécondité/âge médian » - Effectifs



	[20-30 ans[	[30-40 ans[	[40 - 50 ans]
[0 - 1,4 enfants/femme[	0	9	8
[1,4 - 2 enfants/femme[	0	10	6
[2 -3 enfants/femme]	1	1	0

Notons bien que ce tableau croisé contient l'effectif des 35 pays. Autrement dit, si on fait la somme des 9 chiffres contenus dans le tableau, on trouve l'effectif total de la population, soit 35.

b) Pourcentages

Ce tableau peut être mis sous forme de pourcentages en divisant chacun des 9 chiffres par 35 et en multipliant par 100. On obtient alors une distribution croisée des 35 pays en fonction de l'âge médian et de l'indice de fécondité, mais contrairement au cas précédent, cette distribution croisée est exprimée en en pourcentages

Tableau croisé « indice de fécondité/âge médian » - Pourcentages

	[20-30 ans[	[30-40 ans[	[40 - 50 ans]
[0 - 1,4 enfants/femme[	0	25,71	22,86
[1,4 - 2 enfants/femme[	0	28,57	17,14
[2 -3 enfants/femme]	2,86	2,86	0,0

On peut facilement vérifier qu'il s'agit d'un tableau en pourcentages en additionnant les 9 chiffres pour obtenir 100 (en tenant compte des arrondis).

B – Distributions marginales

1) Définition

Lorsqu'on ajoute au tableau croisé une colonne pour la somme des valeurs en ligne et une ligne pour la somme des valeurs en colonnes, on appelle cette colonne et cette ligne les **distributions marginales**.

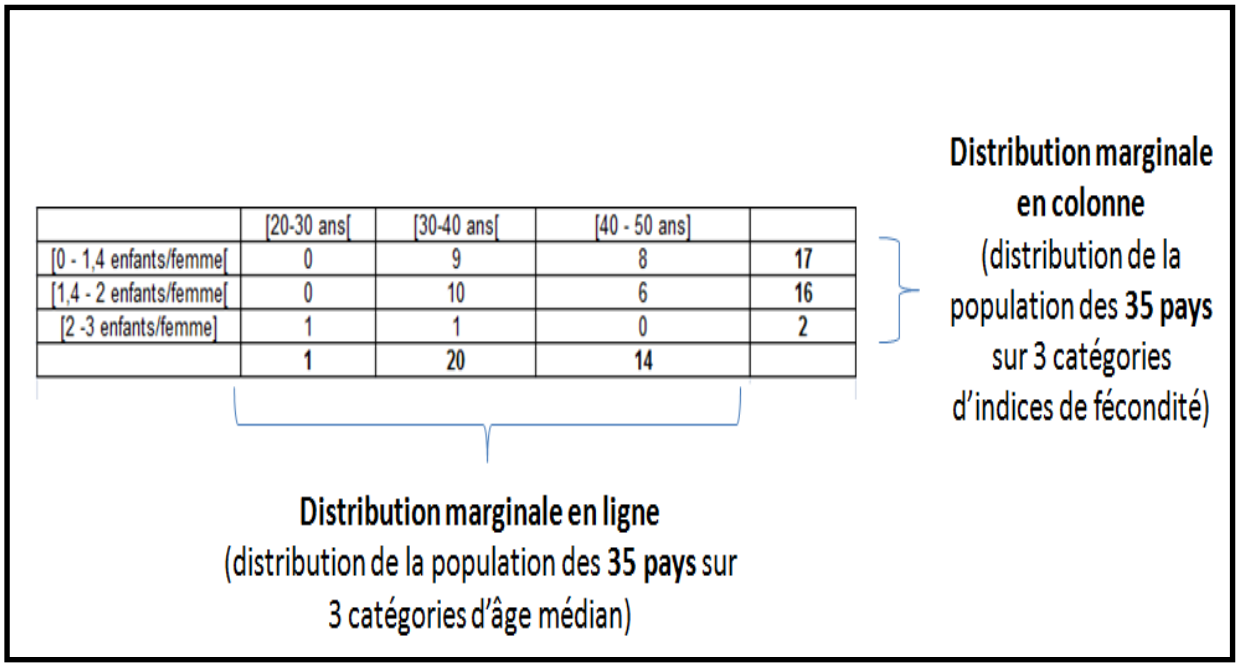
2) Exemple

a) Effectifs

Reprenons le tableau croisé « indice de fécondité/âge médian », mais ajoutons une ligne et une colonne.

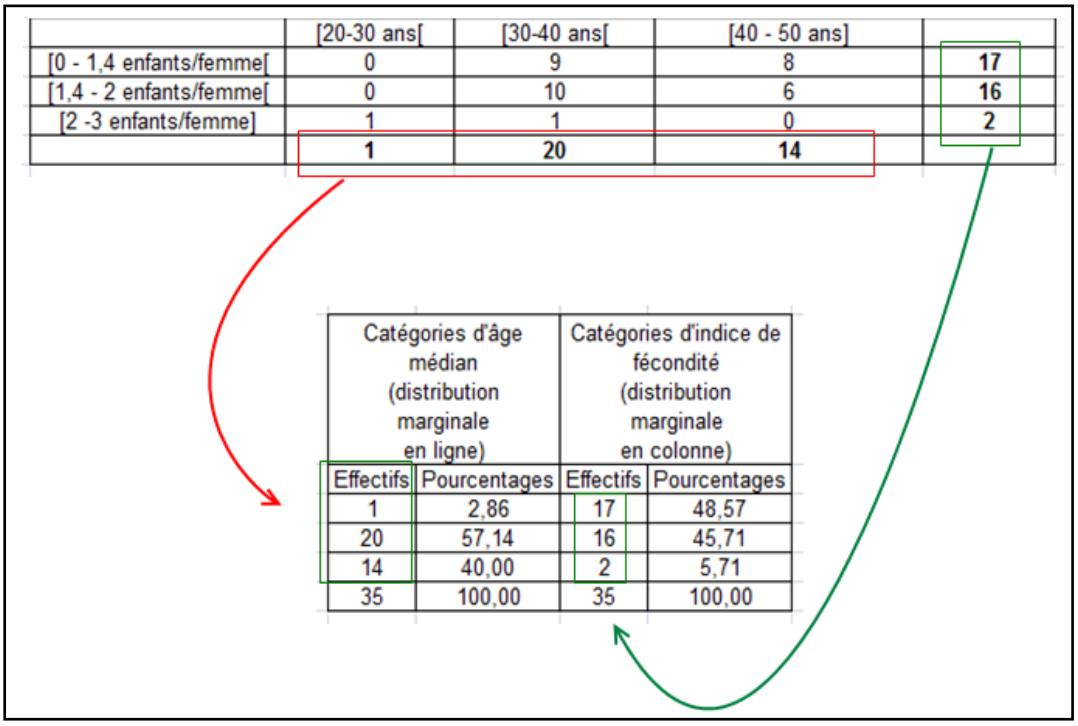
- Chaque chiffre de la dernière ligne ajoutée (en caractère gras) représente le total des effectifs de la colonne correspondante. C'est la distribution marginale en lignes ou distribution de la population des 35 pays sur 3 catégories d'âge médian. En effet 1+20+14 = 35.
- Chaque chiffre de la dernière colonne ajoutée représente le total des effectifs de la ligne correspondante. C'est la distribution marginale en colonnes ou distribution de la population des 35 pays sur 3 catégories d'indices de fécondité. En effet 17 +16 +2 = 35.

Les deux distributions marginales des effectifs



b) Pourcentages

La dernière ligne et la dernière colonne du tableau précédent peuvent s'exprimer en pourcentage de la façon suivante :



	[20-30 ans[	[30-40 ans[	[40 - 50 ans]	
[0 - 1,4 enfants/femme[	0	9	8	17
[1,4 - 2 enfants/femme[	0	10	6	16
[2 -3 enfants/femme]	1	1	0	2
	1	20	14	

Catégories d'âge médian (distribution marginale en ligne)		Catégories d'indice de fécondité (distribution marginale en colonne)	
Effectifs	Pourcentages	Effectifs	Pourcentages
1	2,86	17	48,57
20	57,14	16	45,71
14	40,00	2	5,71
35	100,00	35	100,00

C – Distributions conditionnelles

1) Colonnes

a) Effectifs

Reprenons le tableau croisé « indice de fécondité/âge médian », mais concentrons-nous sur les différentes colonnes. Considérons par exemple la colonne des âges médians compris dans l'intervalle [30-40[:

Exemple de distribution conditionnelle en colonne (effectifs)

	[20-30 ans[	[30-40 ans[	[40 - 50 ans]
[0 - 1,4 enfants/femme[	0	9	8
[1,4 - 2 enfants/femme[	0	10	6
[2 -3 enfants/femme]	1	1	0
Total	1	20	14

Distribution par catégories d'âge de fécondité des 20 pays dont l'âge médian est dans l'intervalle [30-40[

La distribution par catégories d'âge de fécondité des 20 pays dont l'âge médian est dans l'intervalle [30-40 ans [est appelée **distribution conditionnelle en colonne**. L'expression conditionnelle provient du fait que les 20 pays concernés sont une sous-population de la population totale et que cette sous-population correspond à tous les pays qui répondent à la condition « être dans l'intervalle [30-40[des âges médians ».

On voit qu'il y a 3 distributions conditionnelles possibles puisqu'il y a 3 catégories d'âges médians.

b) Pourcentages

L'effectif de la distribution conditionnelle précédente est de 20. Il est distribué selon les 3 catégories d'indices de fécondité. Si l'on fait abstraction du reste du tableau, on peut diviser chacun des chiffres de cette colonne par 20 et le multiplier par 100 de façon à exprimer la distribution conditionnelle en pourcentages. On aura alors :

Age médian [30-40[
(9/20) x 100 = 45%
(10/20) x 100 = 50%
(1/20) x 100 = 5%
Total (20/20) x 100 = 100 %

Si maintenant on effectue la même opération pour les trois colonnes on obtient le tableau des tableaux des **distributions conditionnelles en colonnes en pourcentages**.

Les 3 distributions conditionnelles en colonnes (pourcentages)

	[20-30 ans[	[30-40 ans[	[40 - 50 ans]
[0 - 1,4 enfants/femme[	0	45	57,1
[1,4 - 2 enfants/femme[	0	50	42,9
[2 -3 enfants/femme]	100	5	0
Total	**100**	**100**	**100**

Dans chaque colonne, l'effectif initial a été divisé par le chiffre correspondant de la sous population de pays associés à la catégorie d'âge médian correspondante.

2) Lignes

a) Effectifs

De la même façon qu'il y a des distributions conditionnelles en colonnes, il y a aussi des distributions conditionnelles en ligne. Cette fois, on isole 3 sous populations qui correspondent aux catégories d'indices. A titre d'exemple, dans le tableau ci-après, la catégorie d'indice de fécondité [1,4 – 2 enfants/femme [a été isolée, ce qui correspond à une sous population de pays égale à 16. La distribution de ces pays par catégories d'âges de fécondité est ensuite donnée par la ligne encadrée.

Naturellement, puisqu'il y a 3 catégories d'indice de fécondité, il y a 3 sous populations et trois distributions conditionnelles.

Exemple de distribution conditionnelle en ligne (effectifs)

	[20-30 ans[	[30-40 ans[	[40 - 50 ans]	
[0 - 1,4 enfants/femme[	0	9	8	**17**
[1,4 - 2 enfants/femme[	0	10	6	**16**
[2 -3 enfants/femme]	1	1	0	**2**

Distribution par catégories d'âge médian des 16 pays dont l'indice de fécondité tombre dans la tranche [1,4 - 2 enfants/femme[

b) Pourcentages

Suivant le même principe que pour les distributions conditionnelles en colonne, on peut transformer les distributions d'effectifs en distribution de pourcentages en divisant les chiffres de chaque ligne par le total de la ligne. On obtient alors le tableau suivant des **distributions conditionnelles en colonnes en pourcentages**.

Les 3 distributions conditionnelles en ligne (pourcentages)

	[20-30 ans[	[30-40 ans[	[40 - 50 ans]	
[0 - 1,4 enfants/femme[	0	52,9	47,1	100
[1,4 - 2 enfants/femme[	0	62,5	37,5	100
[2 -3 enfants/femme]	50	50	0	100

Chapitre 3 : Statistiques permettant de résumer une série

1 – Tendance centrale et dispersion des valeurs d'une variable
2 - Les statistiques de tendance centrale
- A - Le mode
  - 1) Définition
  - 2) Remarques à propos du mode
    - *a) Une série peut avoir plusieurs modes*
    - *b) Le mode n'existe pas forcément*
    - *c) Le mode n'est pas forcément la valeur la plus élevée*
    - *d) Variables et caractères peuvent avoir un mode*
    - *e) Mettre la série sous forme de distribution pour repérer le mode*
- B - La moyenne arithmétique
  - 1) La moyenne arithmétique simple
  - 2) La moyenne arithmétique pondérée
  - 3) Calcul de la moyenne sur des données catégorielles
- C - La médiane
  - 1) Origine du mot, sens géométrique
  - 2) Sens du mot en statistique descriptive
  - 3) Méthode de calcul
    - *a) n est pair*
    - *b) n est impair*

3 - Les statistiques de dispersion
- A - Minimum, maximum, intervalle de variation et rapport de variation
  - 1) Minimum et maximum d'une série
  - 2) Intervalle de variation ou étendue
  - 3) Rapport de variation
- B - Intervalle interquartile
- C - Variance, écart-type et coefficient de variation
  - 1) La variance
    - *a) Définition*
    - *b) Exemple*
    - *c) Utilité de la variance*
  - *2)* L'écart-type
    - *a) Définition*
    - *b) Méthode de calcul*
    - *c) Utilité de l'écart-type*
  - 3) Le coefficient de variation

Annexe : Méthode alternative pour le calcul des quartiles

1 – Tendance centrale et dispersion des valeurs d'une variable

Nous avons déjà vu dans le chapitre précédent, un exemple de tableau contenant des statistiques résumées, tableau qui est reproduit ci-dessous pour mémoire:

Tableau de statistiques résumées pour certaines des variables du tableau 1

	Superficie (km2)	Nombre de frontières terrestres avec d'autres pays de l'UE à 27	Population (habitants) estimation de juillet 2006	Age médian (estimation de juillet 2007)	Indice de fécondité (en nombre d'enfants par	PIB en milliards d'USD (2006)	Densité (Nombre d'habitants au km2) (2006)	PIB par habitant (en dollars) (2006)
Moyenne simple		2,20		38,80	1,52	1 053,66	357,91	24 975,06
Valeur minimale	316 (Malte)	0,00	400 214 (Malte)	24,8 (Inde)	0,98 (HK)	6 (Malte)	(8) Russie	735 (Inde)
Valeur maximale	17 075 200 (Russie)	8 (Allemagne)	1 313 973 713 (Chine)	43,5 (Japon)	2,81 (Inde)	13 160 (EU)	6356 (HK)	72785 (Luxembourg)

Ce tableau ne contient que trois statistiques résumant les différentes séries du tableau 1 : Il s'agit de la **moyenne simple**, ainsi que des **valeurs minimales** et **maximales** de chaque série. Nous allons maintenant étudier systématiquement les principales autres statistiques permettant de résumer une série. A chaque fois, on insistera sur la méthode de calcul (plutôt que sur la formule) et on s'appuiera sur des exemples extraits du tableau 1.

Il est habituel de distinguer **deux types de statistiques résumées**, quitte ensuite à les réunir dans des statistiques résumées plus élaborées :

- Les statistiques qui résument la **tendance « centrale »** d'une série ou d'une distribution :
  - **mode,**
  - **moyenne**
  - **médiane**.
- Les statistiques qui résument la **dispersion d'une série** :
  - **intervalle de variation (aussi appelée « étendue »)**
  - **intervalle interquartile**.

  Certaines de ces statistiques, tout en résumant la dispersion de la série, tiennent aussi compte de sa valeur centrale. C'est le cas de :
  - **l'écart-type,**
  - **de la variance**
  - **du coefficient de variation**[3].

Une dernière remarque : bien qu'il soit possible d'effectuer des calculs de statistiques résumées sur les données groupées en catégories, c'est déconseillé de

[3] Il existe aussi des statistiques qui résument la « forme » d'une distribution, mais celles-ci ne sont plus guère utilisées aujourd'hui dans la mesure où il est plus facile d'observer directement le graphique d'une distribution pour en apprécier la forme.

le faire quand on dispose des données brutes ou regroupées par valeurs ou modalités. C'est une question de bon sens. Si l'on effectue les calculs sur des données regroupées par catégories, on obtient des résultats approximatifs et même carrément faux quand on les compare aux calculs effectués sur les données brutes (sans parler des formules abstruses et abscondes qu'il faut employer pour effectuer les calculs de moyennes, médianes, quartiles ou variance sur des données catégorielles).

2 – Les statistiques de tendance centrale

A – Le mode

1) définition

Le mode d'une série est la valeur la plus fréquente d'une série.

Exemple : Soit la série {8,4,4,3,4,3,8,2,5}

La valeur la plus fréquente de cette série est 4. Le mode est donc égal à 4. L'effectif associé à ce mode est 3.

2) Remarques à propos du mode

a) Une série peut avoir plusieurs modes

Soit la série S = {4, 0, 1, 1, **2, 2, 2**, 3, 3, 4, **2**, 3, 4, 5, **2**, 1, 3, 3, 4, 5}, les "2" sont mis en gras et les "3" sont soulignés, car ce sont les valeurs qui reviennent le plus souvent : 5 fois chacune. Cette série a 2 modes, elle est **bimodale**. Ses deux modes sont : 2 et 3. L'effectif associé à chacun de ces modes est : 5. Bien entendu, on peut avoir des séries avec 3, 4, 5, etc. modes. Ce sont alors des **séries multimodales**.

b) Le mode n'existe pas forcément

C'est le cas lorsque toutes les valeurs ont le même effectif comme dans l'exemple suivant : {8,6,5,7,3,1}. Dans ce cas, on peut aussi dire que toutes les valeurs sont modales.

c) Le mode n'est pas la valeur la plus élevée

Il ne faut pas confondre le mode, qui est la valeur la plus fréquente, avec la valeur la plus élevée de la série. Dans la série {8,6,5,7,3,1}, il n'y a pas de mode, mais la valeur la plus élevée est 8. Il peut arriver que le mode soit aussi la valeur la plus élevée, mais ce n'est alors qu'une coïncidence.

d) Variables et caractères peuvent avoir un mode

La notion de mode existe aussi bien dans le cas d'une série qui se rapporte à une variable que dans le cas d'une série qui se rapporte à un caractère.

e) Mettre la série sous forme d'une distribution pour repérer le mode

Pour détecter le mode, il est souvent plus facile de distribuer les éléments de la série par valeurs (ou par modalités). Soit la série « nombre de frontières terrestres avec d'autres pays de l'UE à 27 » extraite du tableau 1 :

S1 = {8, 4, 5, 3, 3, 2,1, 1, 1, 1, 2, 1 ,6, 1, 1, 0, 1, 4, 2, 2, 0, 4, 4, 4, 3, 2, 2, 4, 0, 0, 0, 0, 5, 0,0}

Nous savons que cette série peut être mise sous forme d'une distribution par valeurs de la façon suivante :

Distribution des pays du tableau 1
selon leur nombre de frontières terrestres avec les pays de l'UE à 27

Nombre de frontières terrestres avec d'autres pays de l'UE à 27	Effectifs
0	8
1	8
2	6
3	3
4	6
5	2
6	1
7	0
8	1
	35

0 et 1 sont les deux valeurs modales de la distribution (ou la série correspondante), c'est-à-dire les plus fréquentes. Cette distribution (et la série correspondante) est donc bi-modale

Nous voyons alors plus facilement quels sont les deux modes de la série.

B - La moyenne arithmétique

Le mot moyenne a pour origine le latin "médius», mot signifiant "qui est au milieu". "Médius" est aussi l'origine du mot "médiane". Pourtant, en statistique, les deux mots conduisent à des définitions différentes. Ceci nous laisse supposer que la notion de milieu n'est pas toujours facile à définir.

1) La moyenne arithmétique simple

La moyenne arithmétique d'une série ou **moyenne arithmétique simple** se calcule par une formule qui est donnée par l'expression :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

Le "x" surmonté d'un trait désigne classiquement la moyenne. On remarque que la somme va de 1 à n où n désigne le nombre d'unités statistiques de la population. Appliquons cette définition au calcul de la moyenne de la série suivante = {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}. On a donc :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{4+0+1+1+2+2+2+3+3+4+2+3+4+5+2+1+3+3+4+5}{20} = \frac{54}{20} = 2,7$$

2) La moyenne arithmétique pondérée

La moyenne arithmétique d'une distribution ou **moyenne arithmétique pondérée** se calcule par une formule qui est donnée par l'expression :

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

La somme varie cette fois de 1 à k, avec k qui représente le nombre de valeurs de la série. Dans le cas où aucune valeur n'est répétée k=n. Sinon k<n. Remarquons que la somme va de 1 à k, mais que cette somme est divisée par n et non par k.

La notation n_i représente les effectifs ou fréquences absolues des valeurs. Appliquons cette définition au calcul de la moyenne de la distribution :

x_i	n_i
0	1
1	3
2	5
3	5
4	4
5	2

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{(0\times1)+(1\times3)+(2\times5)+(3\times5)+(4\times4)+(5\times2)}{20} = \frac{0+3+10+15+16+10}{20} = \frac{54}{20} = 2,7$$

Les différences entre la formule de la moyenne donnée en (1) et celle donnée en (2) sont importantes à noter, **quoique les deux formules donnent nécessairement le même résultat**.

- La première différence tient au fait que dans la formule (1) la somme se fait sur les n unités statistiques, alors que dans la formule (2) la somme se fait sur les k valeurs.
- La seconde différence tient au fait que dans la formule (1), chaque élément additionné compte pour un, c'est-à-dire que la pondération est de 1. Dans la formule 2, les pondérations sont différentes. La formule (1) est en fait un cas particulier de la formule (2). En effet si n = k et que $n_i = 1$ pour i = 1 à k alors les deux formules sont identiques.

En pratique, si le calcul de la moyenne doit être fait sans machine à calculer ni tableur et que les données sont peu nombreuses (inférieures à 10), mieux vaut utiliser la formule (1). Sinon, la formule (2) peut être préférée.

La moyenne arithmétique ne peut pas être calculée pour un caractère (dimension quantitative). Soit par exemple le caractère sexe, avec les deux modalités "F" et "H", dans une population de 10 personnes. On a la série suivante : {F,H,F,F,H,H, F,F,F,H}. La modalité "Femme" est plus fréquente (6 contre 4 pour la modalité "Homme") : c'est le mode. En revanche, on ne peut pas calculer de moyenne arithmétique. La même chose est vraie si l'on met cette série sous forme d'une distribution :

Sexe	Effectifs
H	4
F	6

On peut calculer les fréquences associées à chaque modalité. On voit alors que la modalité la plus fréquente est "F" (0,6 contre 0,4 pour la modalité "H"). Mais la notion de moyenne arithmétique n'a pas de sens pour un caractère.

3) Calcul de la moyenne sur des données catégorielles

Ainsi que précisé dans l'introduction à cette section consacrée à la moyenne, il faut à tout prix éviter de procéder à ce type de calcul. Nous ne le donnons ici qu'à titre d'information. Lorsque l'on a une distribution par classes de valeurs, la moyenne se calcule en prenant la formule de la moyenne pondérée et en remplaçant dans cette formule "x_i" par "c_i", où c_i représente le **centre de la classe i**, c'est-à-dire la moyenne arithmétique des extrémités de classe. A défaut d'avoir une valeur x_i on prend "c_i". Ceci explique que le calcul de la moyenne donne un résultat imprécis. On va le voir dans les deux exemples suivants :

Soit la série déjà utilisée précédemment : {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}. Nous savons que la moyenne arithmétique simple appliquée à cette série est :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{4+0+1+1+2+2+2+3+3+4+2+3+4+5+2+1+3+3+4+5}{20} = \frac{54}{20} = 2,7$$

Exemple 1 : Supposons maintenant que seule la distribution par classe de valeurs d'amplitudes égales nous soit donnée :

Classes	**n_i**
[0-2[	4
[2- 4[	10
[4- 6]	6

Pour calculer la moyenne, nous devons déterminer les centres de classe, puis faire la somme des "$n_i \times c_i$" et diviser par n. Autrement dit, nous devons appliquer la formule :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i$$

La notation c_i indique le centre de classe et où k représente le nombre de classes. Comme indiqué précédemment, le centre de classe est égal à la moyenne des extrémités de classe. On a donc le tableau de calcul suivant :

Classes	**n_i**	**c_i (moyenne des extrémités de classe)**	**$n_i \times c_i$**
[0-2[	4	1	4
[2- 4[	10	3	30
[4- 6]	6	5	30
			64

Et finalement :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \frac{1}{20}\sum_{i=1}^{3}\left[(4\times 1)+(10\times 3)+(6\times 5)\right] = \frac{4+30+30}{20} = \frac{64}{20} = 3,2$$

Nous avons donc une marge d'erreur non négligeable par rapport à la vraie moyenne, à savoir 2,7. La marge d'erreur en pourcentage est donnée par :

$$\frac{3,2-2,7}{2,7}\times 100 = \frac{0,5}{2,7}\times 100 = 18,5\%$$

La marge d'erreur dépend de la définition des classes.

Exemple 2 : Supposons que l'on ait maintenant deux classes d'amplitudes inégales. Le calcul se fait de la même façon, mais on obtient un résultat différent :

Classes	**n_i**	**c_i**	**$n_i \times c_i$**
[0-4[	14	2	28
[4- 6]	6	5	30
			58

La moyenne est donc :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \frac{1}{20}\sum_{i=1}^{2}\left[(14\times 2)+(6\times 5)\right] = \frac{28+30}{20} = \frac{58}{20} = 2,9$$

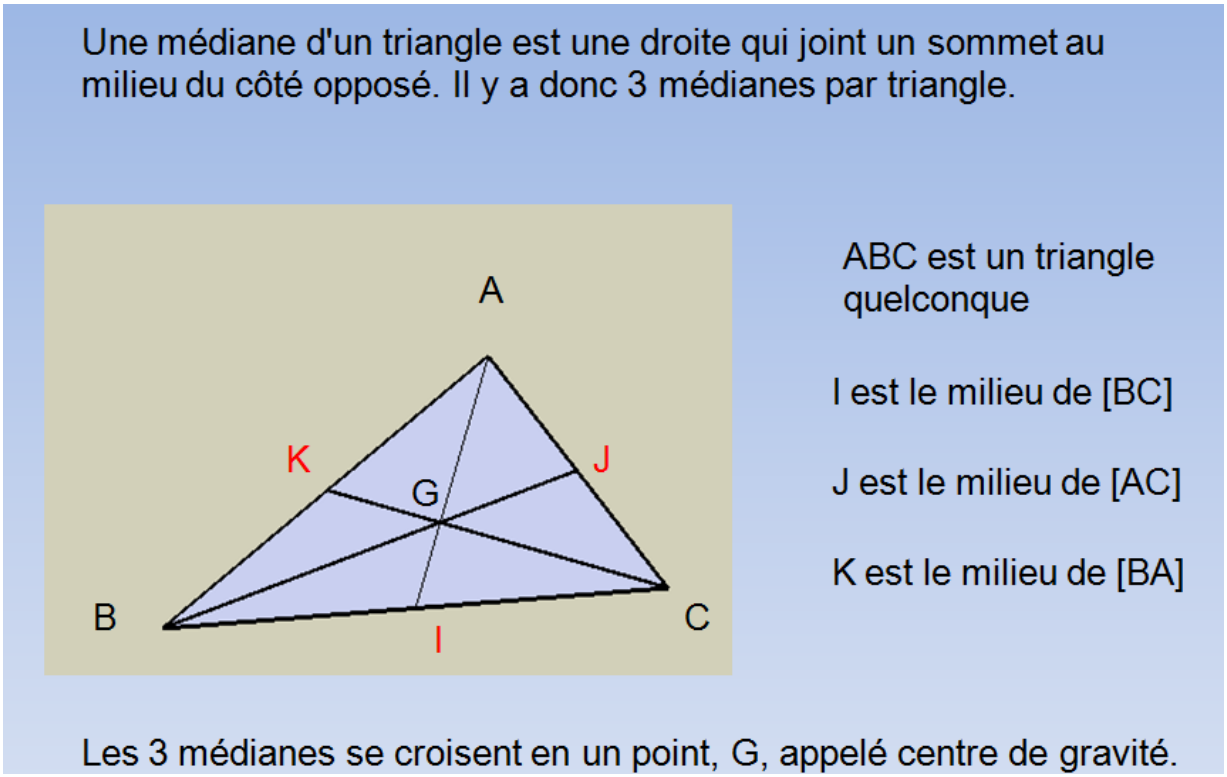
On voit donc que chaque fois que l'on change les classes ou que l'on modifie leur amplitude, on exerce un effet sur la moyenne par le jeu de la modification des centres de classe. Il est donc facile de manipuler la moyenne en choisissant les amplitudes de classe. **C'est pourquoi il est recommandé de ne calculer la moyenne à partir des centres de classe que lorsqu'on ne peut pas faire autrement, c'est-à-dire lorsque l'on ne dispose pas des données brutes.**

C - La médiane

1) Origine du mot, sens géométrique

Le mot « médiane » a pour origine le latin « médius», mot signifiant « qui est au milieu ». « Médius » est aussi l'origine du mot "moyenne". Pourtant, en statistique, les deux mots conduisent à des définitions différentes. Ceci nous laisse supposer que la notion de milieu n'est pas toujours facile à définir.

Avant d'avoir une définition statistique, la médiane a d'abord une définition géométrique. On définit ainsi, par exemple, les médianes d'un triangle. Une médiane d'un triangle est une droite qui joint un sommet au milieu du côté opposé. Il y a donc 3 médianes par triangle.



2) Sens du mot en statistique descriptive

La **médiane** est la valeur de la variable (et dans certains cas, la modalité du caractère) qui partage la population, dont les valeurs ont préalablement été classées par ordre croissant, en deux sous populations égales. **On la désigne par l'abréviation Me**.

3) Méthode de calcul

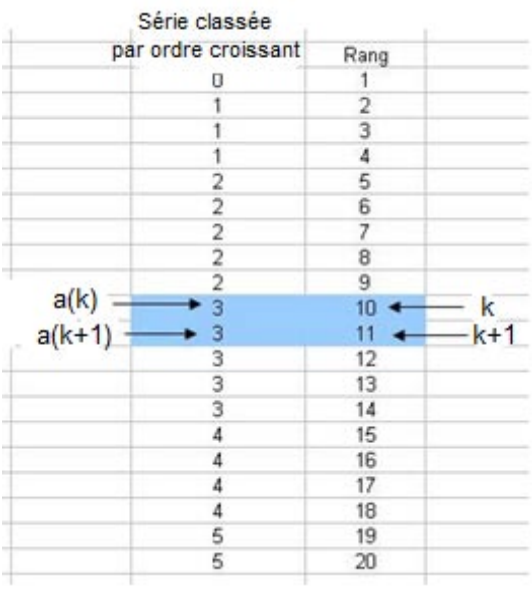
- **Toujours se ramener à une série :** Si les données se présentent sous forme d'une distribution, convertir la distribution en série.
- **Classer la série :** Une fois la série constituée, l'ordonner en classant les chiffres par ordre croissant. On désignera par *k* le rang d'une valeur dans la série (voir tableau ci-après).
- **Déterminer si la série comprend un nombre pair ou impair d'unités statistiques :** Soit *n* le nombre d'unités statistiques. Deux cas peuvent alors se présenter : celui ou *n* est pair et celui ou *n* est impair.

a) n est pair

Dans ce cas la médiane est égale à la moyenne arithmétique de a(k) et de a(k+1) où k est tel que n = 2k et où a(k) et a(k+1) sont les valeurs associées à k et à k+1

Exemple : si l'on prend la série S = {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}.

On a donc le tableau suivant (où les valeurs sont classées par ordre croissant):



Puisque n=20, n est pair. Donc n= 2k implique 20 = 2k soit k = 10 et k+1 = 11. Les valeurs associées sont a(k) = 3 et a(k+1) = 3. Par conséquent :

$$Me = [a(k)+a(k+1)]/2=(3+3)/2=3$$

On peut aussi appliquer la formule proposée par le tableur EXCEL de Microsoft. Voir la notice technique sur le site de Microsoft : http://support.microsoft.com/kb/103493/fr [4].Dans ce cas, on doit calculer Me de la façon suivante :

$$Me = a(k) + [f \times [a(k+1) - a(k)]]$$

Pour comprendre cette formule, il faut d'abord définir g :

$$g = [(1/2) (n-1)] +1.$$

Une fois que l'on a définit g, on peut en déduire k et f :

k représente le rang de la valeur dans le classement du tableau ci-dessus et est égal à la partie entière de g.

f est la partie décimale de g.

a(k) est la valeur associée au rang k

a(k+1) est la valeur associée au rang k+1.

Si l'on applique cette formule, on voit que :

$$g = [(1/2)(20-1)]+1=10,5$$

Par conséquent k= 10 et f = 0,5. On a donc a(k) = 3 et a(k+1)=3. Dès lors :

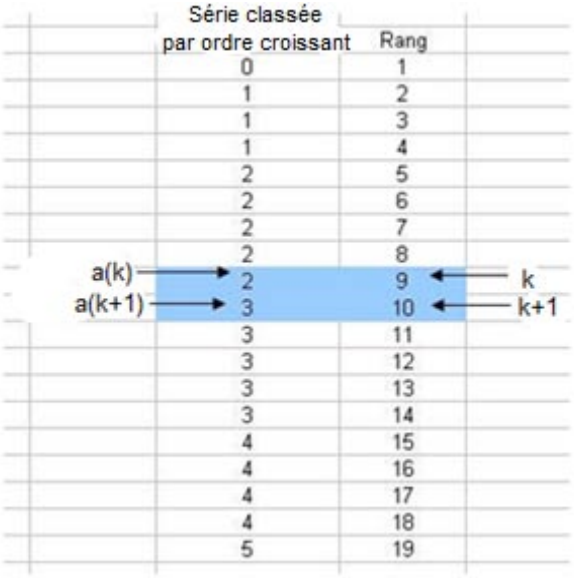
$$Me = 3 + [0,5(3-3)] = 3$$

[4] La notice technique sur le site de Microsoft (http://support.microsoft.com/kb/103493/fr) présente la formule différemment mais aboutit au même résultat.

b) n est impair

Dans ce cas la médiane est égale à a(k+1) où k est défini par 2k+1 = n et où a(k+1) est la valeur associée à k+1

Exemple : si l'on prend la série précédente mais que l'on enlève le 20ème élément, on alors la série { 4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4 } ou n = 19 et le tableau suivant (où les valeurs sont classées par ordre croissant) :



Série classée par ordre croissant	Rang
0	1
1	2
1	3
1	4
2	5
2	6
2	7
2	8
a(k) → 2	9 ← k
a(k+1) → 3	10 ← k+1
3	11
3	12
3	13
3	14
4	15
4	16
4	17
4	18
5	19

2k+1 = n donne donc 2k+1 = 19, c'est-à-dire k=9. Donc k+1 = 10 .

Par conséquent la médiane est égale à a(k+1) soit 3.

3 - Les statistiques de dispersion

A - Minimum, maximum, intervalle de variation et rapport de variation

1) Minimum et maximum d'une série

Revenons au tableau de statistiques résumées dans lequel figure, pour certaines des variables du tableau 1, les valeurs minimale et maximale de la série.

Tableau de statistiques résumées pour certaines des variables du tableau 1

	Superficie (km2)	Nombre de frontières terrestres avec d'autres pays de l'UE à 27	Population (habitants) estimation de juillet 2006	Age médian (estimation de juillet 2007)	Indice de fécondité (en nombre d'enfants par	PIB en milliards d'USD (2006)	Densité (Nombre d'habitants au km2) (2006)	PIB par habitant (en dollars) (2006)
Moyenne simple		2,20		38,80	1,52	1 053,66	357,91	24 975,06
Valeur minimale	316 (Malte)	0,00	400 214 (Malte)	24,8 (Inde)	0,98 (HK)	6 (Malte)	(8) Russie	735 (Inde)
Valeur maximale	17 075 200 (Russie)	8 (Allemagne)	1 313 973 713 (Chine)	43,5 (Japon)	2,81 (Inde)	13 160 (EU)	6356 (HK)	72785 (Luxembourg)

Ces deux valeurs donnent immédiatement une *certaine idée* de la dispersion. On voit par exemple que les écarts entre les superficies des pays, de même que l'écart entre les populations sont très importants.

2) Intervalle de variation (ou « étendue »)

L'intervalle de variation (IV) ou l'étendue de la série est simplement une façon de résumer le minimum et le maximum de la série en un seul chiffre. On l'obtient ainsi :

Intervalle de variation de la série = valeur maximale – Valeur minimale

Dans le tableau ci-après, l'intervalle de variation a été ajouté sur la dernière ligne pour les 8 séries :

Tableau de statistiques résumées pour certaines des variables du tableau 1

	Superficie (km2)	Nombre de frontières terrestres avec d'autres pays de l'UE à 27	Population (habitants) estimation de juillet 2006	Age médian (estimation de juillet 2007)	Indice de fécondité (en nombre d'enfants par	PIB en milliards d'USD (2006)	Densité (Nombre d'habitants au km2) (2006)	PIB par habitant (en dollars) (2006)
Moyenne simple		2,20		38,80	1,52	1 053,66	357,91	24 975,06
Valeur minimale	316 (Malte)	0,00	400 214 (Malte)	24,8 (Inde)	0,98 (HK)	6 (Malte)	(8) Russie	735 (Inde)
Valeur maximale	17 075 200 (Russie)	8 (Allemagne)	1 313 973 713 (Chine)	43,5 (Japon)	2,81 (Inde)	13 160 (EU)	6356 (HK)	72785 (Luxembourg)
Intervalle de variation	17 074 884	8,00	1 313 573 499	18,70	1,83	13 154,00	6 348,00	72 050,00

La dernière ligne donne l'intervalle de variation (arrondi au nombre entier le plus proche), c'est-à-dire la différence entre la valeur maximale et la valeur minimale.

3) Rapport de variation

Le rapport de variation est simplement le rapport de la valeur maximale à la valeur minimale. Ainsi, si l'on divise le PIB par habitant maximum par le PIB par habitant minimum en 2006, on obtient :

$$\frac{\text{PIB/hab du Luxembourg}}{\text{PIB/hab de l'Inde}} = \frac{72785}{735} = 99$$

On voit ainsi que l'écart est pratiquement de 1 à 100 puisque le PIB/habitant du Luxembourg est 99 fois supérieur à celui de l'Inde. Naturellement, il s'agit de chiffres exprimés en dollars courant. Il faudrait, pour être plus précis, les exprimer en parité de pouvoir d'achat.

B - Intervalle interquartile

L'intervalle de variation ne donne qu'une idée imprécise et parfois fausse de la dispersion de la série, car les valeurs extrêmes peuvent être exceptionnelles et le reste de la population statistique être concentré sur un intervalle beaucoup plus retreint. On peut parfaitement s'en rendre compte dans le tableau ci-dessus, où les intervalles de variation sont énormes pour les superficies et pour les populations, car de toutes petites îles (comme Malte) sont comparées avec des pays presque de la taille d'un continent (Chine, Inde). D'où l'idée de calculer **l'intervalle interquartile** qui donne une idée plus précise de la dispersion des valeurs d'une série (ou d'une distribution). Avant de définir l'intervalle interquartile, il convient cependant de définir les quartiles.

1) Quartiles

Les **quartiles** sont les **trois valeurs** qui partagent la population, dont les valeurs ont préalablement été classées par ordre croissant, en **quatre sous populations** de même taille. On les désigne respectivement par Q_1, Q_2 et Q_3 .

2) Calcul des quartiles

On notera que Q_2 = Me. Autrement dit, le deuxième quartile n'est autre que la médiane que nous avons déjà étudiée. **Il est important de noter qu'il n'existe pas d'algorithme universellement accepté pour déterminer les quartiles Q_1 et Q_3.** Dans ce qui suit, nous utiliserons la formule employée par le logiciel EXCEL de Microsoft[5].

Prenons l'exemple de la série {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}, on a n=20

Le calcul de Q_i (i=1,2 ou 3) s'effectuera dès lors au moyen la formule suivante :

$$Q_i = a(k) + [\, f_x \, [\, a(k+1) - a(k)] \,]$$

S'agissant du premier quartile on aura donc :

$$Q_1 = a(k) + [\, f_x \, [\, a(k+1) - a(k)] \,]$$

Pour comprendre cette formule, il faut d'abord définir g :

$$g = [(\mathbf{1/4})\ (n-1)] + 1.$$

---

[5] Voir la notice technique sur le site de Microsoft : http://support.microsoft.com/kb/103493/fr . La notice présente la formule différemment mais aboutit au même résultat. Les principales autres méthodes de calcul des quartiles sont résumées sur le site Mathworld. **L'avantage de la méthode Microsoft est qu'il n'est pas nécessaire d'appliquer une formule différente suivant que n est pair ou impair**.

Le logiciel **Mathematica 6**, détermine quant à lui les quartiles de la façon suivante :

- `Quartiles[list]` is equivalent to `Quantile[list, {1/4, 1/2, 3/4}, {{1/2, 0}, {0, 1}}]`. »
- The second quartile is equivalent to `Median[list]`. »
- For even `Length[list]`, the first quartile is equivalent to the median of the $\frac{n}{2}$ smallest elements in *list*.
- For odd `Length[list]`, the first quartile is equivalent to the average of the median of the $\frac{n-1}{2}$ smallest elements and the median of the $\frac{n+1}{2}$ smallest elements in *list*.
- The third quartile is defined as for the first, but with the largest rather than smallest elements.

S'agissant de la série précédente, on trouve ainsi Q_3 = 11,175 avec Mathematica :

```
In[4]:= Quartiles[{8.4, 8.5, 8.9, 9, 9.1, 9.5, 9.5, 9.6, 10.5, 11.1, 11.1, 11.2, 11.5, 11.6, 11.7}]
Out[4]= {9.025, 9.6, 11.175}
```

Un autre algorithme de calcul est proposé dans l'annexe à ce chapitre.

Une fois que l'on a définit g, on peut en déduire k et f :

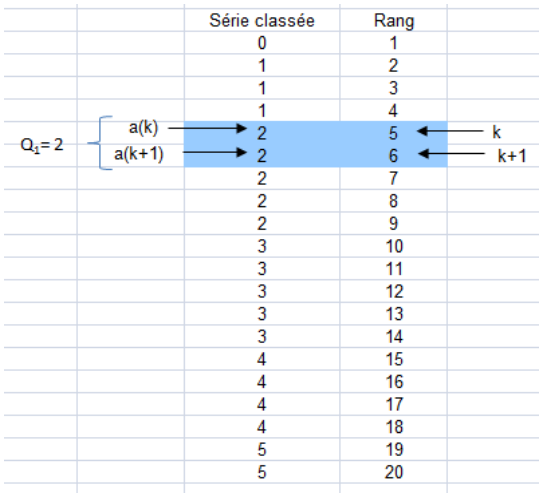
- k représente le rang de la valeur dans le classement du tableau ci-dessous et est égal à la partie entière de g
- f est la partie décimale de g.
- a(k) est la valeur associée au rang k et a(k+1) est la valeur associée au rang k+1 Si l'on applique cette formule, on voit que :

$$g = [(1/4)(20-1)]+1=5,75$$

Par conséquent k= 5 et f = 0,75. On a donc a(k) = 2 et a(k+1)=2. Dès lors :

$$Q_1 = 2 + [0,75 \times (2-2)] = 2$$

On peut suivre la formule sur le tableau ci-après :



	Série classée	Rang
	0	1
	1	2
	1	3
	1	4
a(k)	2	5
a(k+1)	2	6
	2	7
	2	8
	2	9
	3	10
	3	11
	3	12
	3	13
	3	14
	4	15
	4	16
	4	17
	4	18
	5	19
	5	20

Le calcul de Q_3 s'effectue au moyen de la même formule que pour Q_1, soit :

$$Q_3 = a(k) + [f \times [a\ (k+1) - a(k)]]$$

Mais avec un changement dans la définition de g. Désormais on a :

$$g = [(\mathbf{3/4})\ (n-1)] +1$$

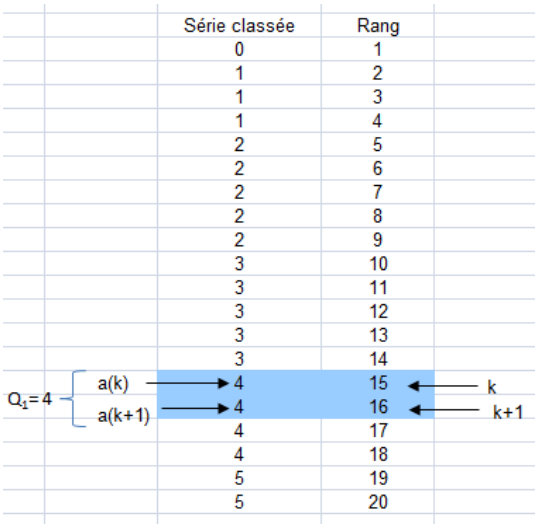
Prenons toujours l'exemple de la série {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}, on a n=20. Dès lors :

$$g = [(\mathbf{3/4})\ (n-1)] +1 = [(3/4)(20-1)]+1=15,25$$

Par conséquent k= 15 et f = 0,25. On a donc a(k) = 4 et a(k+1)=4. Dès lors :

$$Q_3 = 4 + [0,25 \times (4-4)] = 4$$

On peut suivre la formule sur le tableau ci-après :



3) Intervalle interquartile

L'intervalle interquartile (IIQ) est la différence entre le troisième quartile et le premier quartile. Il s'écrit :

$$IIQ = Q_3 - Q_1$$

L'intervalle interquartile sert à apprécier la dispersion de la série, de façon absolue, ou bien par comparaison avec une autre série (à condition que les valeurs de l'autre série soient exprimées dans la même unité). En effet, les valeurs Q_1 et Q_3 délimitent une plage au sein de laquelle **environ**[6] 50% des valeurs de la série sont concentrées. **Plus cet intervalle est large, plus la série est dispersée**. Dans l'exemple que nous avons utilisé, l'IIQ est égal à 4-2 = 2.

Le tableau ci-après donne la médiane, Q_1 et Q_3 , ainsi que l'intervalle interquartile pour certaines des variables du tableau 1 . Les calculs ont été effectués avec Microsoft EXCEL en utilisant la même formule que celle proposée ci-dessus et donnent par conséquent, sauf erreur, les mêmes résultats que si les calculs sont effectués avec une machine à calculer ou à la main (certains problèmes d'arrondis peuvent créer de légères divergences non significatives).

[6] C'est pour cette raison que les algorithmes de calcul diffèrent.

Tableau de statistiques résumées pour certaines des variables du tableau 1

	Superficie (km2)	Nombre de frontières terrestres avec d'autres pays de l'UE à 27	Population (habitants) estimation de juillet 2006	Age médian (estimation de juillet 2007)	Indice de fécondité (en nombre d'enfants par	PIB en milliards d'USD (2006)	Densité (Nombre d'habitants au km2) (2006)	PIB par habitant (en dollars) (2006)
Moyenne simple		2,20		38,80	1,52	1 053,66	357,91	24 975,06
Valeur minimale	316 (Malte)	0,00	400 214 (Malte)	24,8 (Inde)	0,98 (HK)	6 (Malte)	(8) Russie	735 (Inde)
Valeur maximale	17 075 200 (Russie)	8 (Allemagne)	1 313 973 713 (Chine)	43,5 (Japon)	2,81 (Inde)	13 160 (EU)	6356 (HK)	72785 (Luxembourg)
Intervalle de variation	17 074 884	8,00	1 313 573 499	18,70	1,83	13 154,00	6 348,00	72 050,00
Q1	42 310	1,00	5 335 410	37,75	1,00	63,91	74,00	9 630,50
Q3	347 026	4,00	49 265 676	41,00	2,00	769,55	212,00	37 858,00
Mediane	92 931	2,00	10 235 455	39,50	1,40	258,00	114,00	20 958,00
Intervalle Inter-quartile (IIQ)	304 716	3,00	43 930 266	3,25	1,00	705,64	138,00	28 227,50

Prenons l'exemple de la **densité de population**. Q_1, Q_3 et l'intervalle interquartile nous indiquent respectivement des chiffres égaux à 74, 212 et 138 (c'est-à-dire 212-74). Cela signifie qu'**environ la moitié** de nos 35 pays a une densité de population comprise entre 74 et 212, et que l'écart entre ces deux bornes est de 138. On peut également calculer le rapport Q_3/Q_1 qui est ici de 212/74 = 2,86, alors que le rapport de variation (valeur maximale/valeur minimale est de 6356/8 = 794.5). Ces résultats complètent ceux déjà indiqués par la moyenne et l'intervalle de variation. Ils nous montrent aussi la difficulté de résumer correctement une série statistique par un chiffre. C'est une des raisons pour lesquelles les graphiques sont de plus en plus utilisés de préférence aux statistiques résumées. En effet, non seulement ils sont plus parlants que les tableaux, mais aussi, ils résument mieux la série ou la distribution, qu'une kyrielle de statistiques telles que celles que nous sommes en train de calculer.

D'autant que ces statistiques résumées, bien qu'intéressantes et déjà fort nombreuses, restent encore insuffisantes. Elles peuvent en effet être complétées par trois autres indicateurs que nous allons étudier maintenant : La variance, l'écart-type et le coefficient de variation

C - Variance, écart-type et coefficient de variation

Ces trois statistiques sont liées entre elles. Elles sont toutes les trois des indicateurs de la dispersion d'une série par rapport à sa valeur moyenne. Le plus simple est de commencer par l'étude de la variance.

1) La variance

La variance est un indicateur de la dispersion d'une série par rapport à sa moyenne. De même que la moyenne, elle se résume à un seul chiffre qui s'obtient par un calcul que nous allons décomposer ci-après.

a) Définition

La définition de la variance d'une série de chiffres est donnée par la formule[7] :

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Où :

- V désigne la variance des *n* valeurs associées aux *n* unités statistiques de la population. Chacune de ces valeurs est désignée par x_i , le i étant un indice qui varie de 1 à n (i = 1 ….,n).

- $\overline{x}$ est la **moyenne arithmétique simple** des *n* valeurs associées aux unités statistiques x_i (*i* = 1, …, *n*).

[7] Nous donnons ici la formule de la variance des valeurs associées aux unités statistiques d'une population et non la variance d'un **échantillon** de cette population, dont la définition est légèrement différente. En effet, pour calculer la variance d'un échantillon, on divise par n-1 au lieu de diviser par n, mais dans ce cas le « n » de l'échantillon est évidemment beaucoup plus petit que le « n » de la population et l'on différencie alors les deux en désignant par *N* le nombre d'unités statistiques de la population et par *n* le nombre d'unités statistiques de l'échantillon. De plus, si l'on veut extraire plusieurs échantillons de la population, on est amené à rajouter un indice aux n pour les distinguer (on prendra alors l'indice *j* puisque l'indice *i* est déjà utilisé pour désigner les unités statistiques elles-mêmes).

b) Exemple

Soit la série S = {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5} ou *n*=20. Pour calculer la variance de cette série, on procède ainsi :

- **Toujours se ramener à une série** :par exemple, si au lieu d'avoir une série on avait la distribution suivante :

x_i	n_i
0	1
1	3
2	5
3	5
4	4
5	2

  Il faudrait d'abord la transformer en série.

- **Calculer la moyenne arithmétique simple** :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{4+0+1+1+2+2+2+3+3+4+2+3+4+5+2+1+3+3+4+5}{20} = \frac{54}{20} = 2,7$$

- **Retrancher ensuite cette moyenne de chacune des 20 valeurs de la série** (colonne 3 du tableau ci-après).

- On obtient ainsi une série qui comprend des valeurs négatives (car certaines valeurs de la variable sont inférieures à la moyenne et donc si on retranche la moyenne elles deviennent négatives) et des valeurs positives (car certaines valeurs de la variable sont supérieures à la moyenne et donc si on retranche la moyenne elles restent positives).

- Afin de tenir compte à la fois des distances positives et négatives, on ne peut pas additionner immédiatement les valeurs de la colonne 3. Il faut d'abord élever au carré chacune de ces valeurs, de façon à obtenir une série de valeurs positives (colonne 4).

- Cette série de valeurs positives reflète déjà en elle-même la dispersion par rapport à la moyenne. Mais il faut encore additionner ces valeurs pour avoir un chiffre unique (dernière valeur en gras dans la colonne 4)

- Diviser ensuite ce total par *n*, pour avoir en fait une sorte de moyenne. C'est pour cela que l'on dit que la variance n'est finalement que « la moyenne du carré des écarts à la moyenne ». Et l'on obtient la variance de notre série de chiffres, soit ici :

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{36,2}{20} = 1,81$$

Tableau pour la disposition du calcul de la variance

Colonne 1	Colonne 2	Colonne 3	Colonne 4
	Valeurs de la variable X (désignées par x_i)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
x_1	4	1,3	1,69
x_2	0	-2,7	7,29
x_3	1	-1,7	2,89
x_4	1	-1,7	2,89
x_5	2	-0,7	0,49
x_6	2	-0,7	0,49
x_7	2	-0,7	0,49
x_8	3	0,3	0,09
x_9	3	0,3	0,09
x_{10}	4	1,3	1,69
x_{11}	2	-0,7	0,49
x_{12}	3	0,3	0,09
x_{13}	4	1,3	1,69
x_{14}	5	2,3	5,29
x_{15}	2	-0,7	0,49
x_{16}	1	-1,7	2,89
x_{17}	3	0,3	0,09
x_{18}	3	0,3	0,09
x_{19}	4	1,3	1,69
x_{20}	5	2,3	5,29
			36,2 ← $\sum_{i=1}^{n}(x_i - \bar{x})^2$
	$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{36,2}{20} =$		**1,81**

c) Utilité de la variance

La variance n'est pas d'une grande utilité en soi. On peut seulement dire que plus elle est élevée, plus la dispersion autour de la moyenne est élevée. Mais comme les écarts à la moyenne ont été élevés au carré, le chiffre obtenu, quoiqu'exprimé dans l'unité de la variable, est généralement assez élevé et « encombrant ». C'est pourquoi, on utilise surtout la variance comme **calcul intermédiaire** pour obtenir **l'écart-type** et le **coefficient de variation**.

2) L'écart-type

a) Définition

La définition de **l'écart-type** d'une série de chiffres est donnée par la formule :

$$\sigma_x = \sqrt{V(x)}$$

En d'autres termes, l'écart-type est égal à la racine carrée de la variance.

b) Exemple

Soit la série S = {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5} ou *n*=20.

Nous voulons calculer l'écart-type. Nous avons déjà calculé la variance :

V(x)=1,81

Il suffit donc de prendre la racine carrée de la variance et l'on obtient :

$$\sigma_x = \sqrt{V(x)} = \sqrt{1,81} = 1,345$$

c) Utilité de l'écart-type

De façon générale :

- **si l'écart-type est faible**, cela signifie que les valeurs sont assez concentrées autour de la moyenne.
- **si l'écart-type est élevé**, cela veut dire au contraire que les valeurs sont plus dispersées autour de la moyenne.

Exemple : Dans une usine, le fait d'avoir un écart-type aussi bas que possible peut constituer un **objectif de contrôle de qualité**. Soit une entreprise qui fabrique un certain composant et qu'un des éléments du contrôle de la qualité consiste à mesurer le diamètre du composant. Chaque composant aura donc son diamètre mesuré. On calculera ensuite le diamètre moyen, puis l'écart-type. Si l'écart-type est faible, cela signifie que les pièces ont dans l'ensemble un diamètre proche de la moyenne, donc que leur diamètre se ressemble. *À la limite, un écart-type nul signifie que toutes les pièces ont le même diamètre*. Inversement, plus l'écart-type est élevé, plus il y a de pièces dont le diamètre s'écarte de la moyenne et qui risque de ne pas cadrer avec le système auxquelles elles sont destinées.

Une autre utilité de l'écart-type est de servir de bornes. On regardera par exemple combien de valeurs de la série sont comprises dans l'intervalle défini par :

$$[\bar{x} - \sigma_x ; \bar{x} + \sigma_x]$$

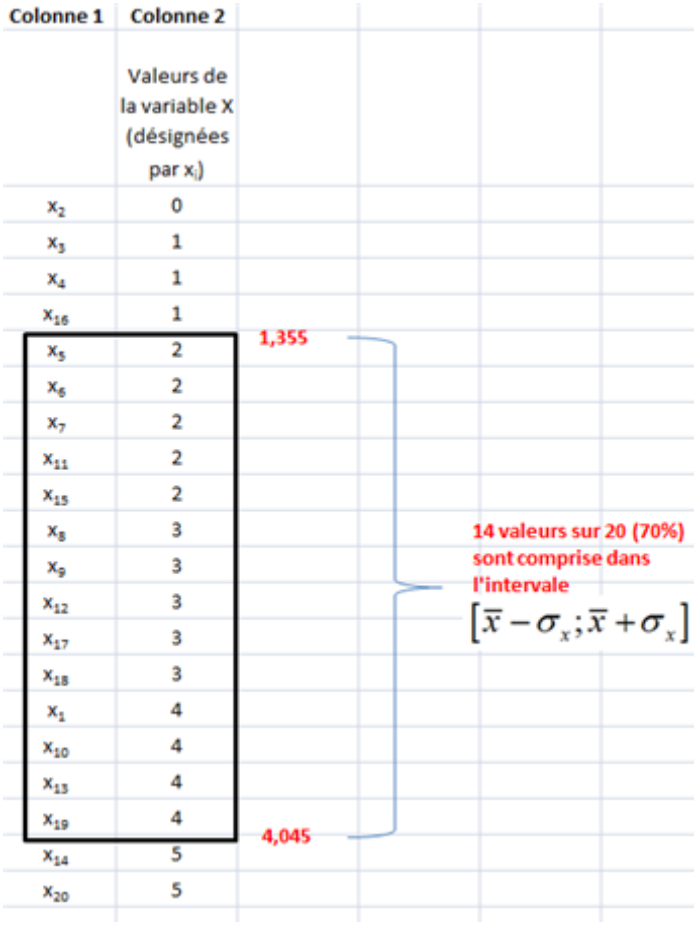
Dans l'exemple de S = {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5} , dont nous avons déjà calculé la moyenne simple (2,7), nous obtenons l'intervalle suivant :

[2,7 – 1,345 ; 2,7 + 1,345] = [1,355 ; 4,045]

Pour savoir combien d'unités ont des valeurs comprises dans cet intervalle, il faut classer la série par ordre croissant des valeurs de la variable et ensuite encadrer les valeurs comprises dans l'intervalle. On voit ainsi que 14 unités sur 20 ont des valeurs comprises dans l'intervalle, ce qui représente 70% de la population. Si ces valeurs se rapportaient aux caractéristiques des pièces d'un processus industriel on pourrait juger que la qualité est acceptable ou bien au contraire se fixer comme objectif d'augmenter le nombre de valeurs qui sont dans cet intervalle. On voit que cet intervalle est avec l'intervalle interquartile une autre façon de mesurer la dispersion d'une série.

Nombres d'unités statistiques dont la valeur est comprise dans l'intervalle

$$[\bar{x} - \sigma_x ; \bar{x} + \sigma_x]$$



3) Le coefficient de variation

L'écart-type est un outil intéressant pour mesurer la dispersion d'une série, ainsi que nous venons de le voir. Mais il possède une limite : **Il est est exprimé dans l'unité de la variable dont il mesure la dispersion des valeurs**.

Ainsi, par exemple, si on veut comparer les dispersions de deux séries qui sont exprimées dans des unités différentes, on ne peut pas.

Le tableau de statistiques résumées ci-après donne l'écart-type de certaines des variables du tableau 1. On peut regarder chaque écart-type et constater qu'il est élevé ou pas, mais on ne pourrait comparer chaque écart-type qu'à un autre écart-type exprimé dans la même unité.

Par exemple, si l'on avait aussi la série des âges médians de l'année 1997 pour les mêmes 35 pays, on pourrait regarder si la dispersion a augmenté ou diminué en 10 ans, car les deux écart-types seraient exprimés dans la même unité (l'année).

Par contre, on ne peut pas dire, en regardant l'écart-type des densités par habitant en 2006 (qui est de 138) que la distribution de valeurs est plus groupée autour de la moyenne que dans le cas des PIB par habitant de 2006, dont l'écart-type est pourtant beaucoup plus élevé (17 239). Car comme les deux séries sont exprimées dans des unités différentes, elles ne sont pas comparables.

Tableau de statistiques résumées pour certaines des variables du tableau 1

	Superficie (km2)	Nombre de frontières terrestres avec d'autres pays de l'UE à 27	Population (habitants) estimation de juillet 2006	Age médian (estimation de juillet 2007)	Indice de fécondité (en nombre d'enfants par	PIB en milliards d'USD (2006)	Densité (Nombre d'habitants au km2) (2006)	PIB par habitant (en dollars) (2006)
Moyenne simple		2,20		38,80	1,52	1 053,66	357,91	24 975,06
Valeur minimale	316 (Malte)	0,00	400 214 (Malte)	24,8 (Inde)	0,98 (HK)	6 (Malte)	(8) Russie	735 (Inde)
Valeur maximale	17 075 200 (Russie)	8 (Allemagne)	1 313 973 713 (Chine)	43,5 (Japon)	2,81 (Inde)	13 160 (EU)	6356 (HK)	72785 (Luxembourg)
Intervalle de variation	17 074 884	8,00	1 313 573 499	18,70	1,83	13 154,00	6 348,00	72 050,00
Q1	42 310	1,00	5 335 410	37,75	1,00	63,91	74,00	9 630,50
Q3	347 026	4,00	49 265 676	41,00	2,00	769,55	212,00	37 858,00
Mediane	92 931	2,00	10 235 455	39,50	1,40	258,00	114,00	20 958,00
Intervalle Inter-quartile (IIQ)	304 716	3,00	43 930 266	3,25	1,00	705,64	138,00	28 227,50
Ecart-type		1,97		3,37	0,33	2 329,55	1 053,25	17 238,92
Coefficient de variation		0,89		0,09	0,22	2,21	2,94	0,69

D'où l'idée de calculer le **coefficient de variation** qui est égal à l'écart-type divisé par la moyenne, c'est-à-dire :

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

Le tableau ci-dessus donne le coefficient de variation de certaines des variables du tableau 1. Cette fois on peut comparer les dispersions des différentes séries, car **le coefficient de variation est un nombre sans dimension**. S'il est égal à 0, c'est que toutes les valeurs de la série sont identiques. Plus il est élevé et plus les valeurs de la variable sont dispersées. Et si l'on compare par exemple la dispersion des densités à la dispersion des PIB, on peut dire que les densités par habitant sont au moins 4 fois plus dispersées que les PIB par habitant (2,94/0,69=4,26).

Annexe : Méthode alternative pour le calcul des quartiles

Cette méthode ne correspond pas à celle employée par EXCEL, ni par les autres logiciels de calcul. **Elle n'est donnée ici qu'à titre d'information parce qu'elle est la plus logique**. C'est aussi la méthode qui est proposée dans Wikipedia (voir http://fr.wikipedia.org/wiki/Crit%C3%A8res_de_position) :

i) **Toujours se ramener à une série :** Si les données se présentent sous forme d'une distribution par valeurs, convertir la distribution en série.

ii) **Classer la série :** Une fois la série constituée, l'ordonner en classant les chiffres par ordre croissant.

iii) **Déterminer le quotient et le reste de la division de n par 4 :** Soit n le nombre d'éléments de la série et p le quotient de la division de n par 4. Quatre cas peuvent se présenter, suivant les quatre valeurs possibles du reste de la division de n par 4. On peut en effet avoir n=4p (pas de reste) ; n=4p+1 (reste 1); n=4p+2 (reste 2); n=4p+3 (reste 3). Envisageons successivement ces quatre cas .

a) Cas où n = 4p

C'est le cas où, quand on divise n par 4, on trouve p et que le reste est nul. Dans ce cas, on a :

Q_1 = moyenne entre la p^e et $(p+1)^e$ valeur.

Q_2 = Me = moyenne entre la $(2p)^e$ valeur et la $(2p+1)^e$ valeur.

Q_3 = moyenne entre la $(3p)^e$ valeur et la $(3p+1)^e$ valeur

Exemple : si l'on prend la série {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}, on a n = 4 p = 20 $\Leftrightarrow$ p=5.

En classant cette série on obtient le tableau suivant :

	Nombre d'enfants	Rang	
	0	1	
	1	2	
	1	3	
	1	4	
$(p)^{ème}$ valeur →	2	5	← p
$(p+1)^{ème}$ valeur →	2	6	← p+1
	2	7	
	2	8	
	2	9	
$(2p)^{ème}$ valeur →	3	10	← 2p
$(2p+1)^{ème}$ valeur →	3	11	← 2p+1
	3	12	
	3	13	
	3	14	
$(3p)^{ème}$ valeur →	4	15	← 3p
$(3p+1)^{ème}$ valeur →	4	16	← 3p+1
	4	17	
	4	18	
	5	19	
	5	20	

Fichier EXCEL

Par conséquent, on a :

Q_1 = moyenne entre la p^e et la $(p+1)^e$ valeur =(2+2)/2=2

Q_2 = Me = moyenne entre la $(2p)^e$ valeur et la $(2p+1)^e$ valeur = (3+3)/2=3

Q_3 = moyenne entre la $(3p)^e$ valeur et la $(3p+1)^e$ valeur = (4+4)/2=4

Les 4 groupes de valeurs sont : {{0, 1, 1, 1, 2}, {2, 2, 2, 2, 3}, {3, 3, 3, 3, 4}, {4, 4, 4, 5, 5}}

b) Cas où n = 4p + 1

Dans ce cas, le reste de la division par 4 est 1 et l'on a :

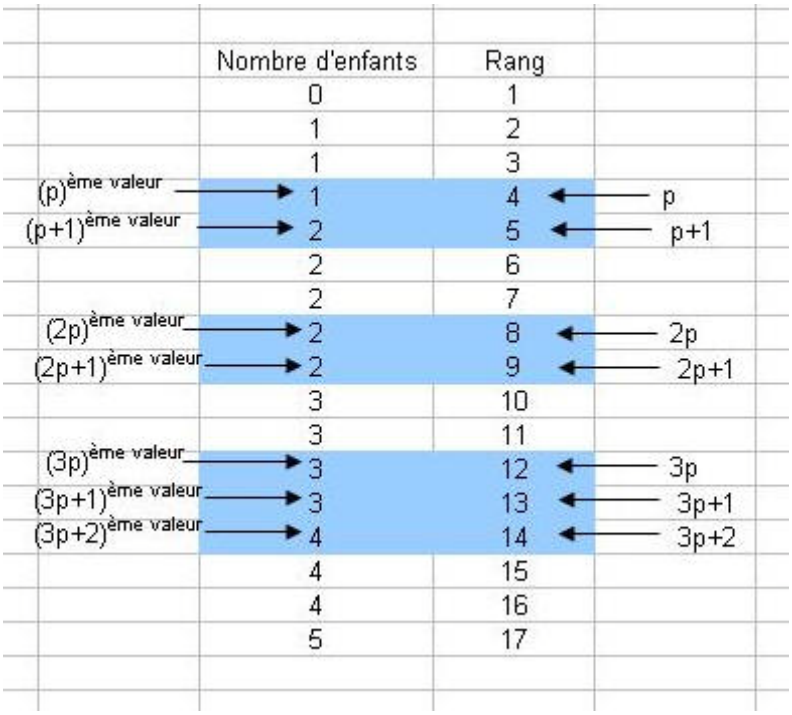
Q_1 = moyenne entre la p^e et la $(p+1)^e$ valeur.

Q_2 = $(2p+1)^e$ valeur.

Q_3 = moyenne entre la $(3p+1)^e$ valeur et la $(3p+2)^e$ valeur.

Exemple : si l'on prend la série {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3} on a n =17 et n = 4p + 1, avec p = 4.

En classant cette série on obtient le tableau suivant :



Fichier EXCEL

Par conséquent, on a :

Q_1 = moyenne entre la p^e et la $(p+1)^e$ valeur = (1+2)/2=1,5

Q_2 = $(2p+1)^e$ valeur = 2

Q_3 = moyenne entre la $(3p+1)^e$ valeur et la $(3p+2)^e$ valeur=(3+4)/2=3,5

Les 4 groupes de valeurs sont : {{0, 1, 1, 1}, {2, 2, 2, 2}, **2**, {3, 3, 3, 3}, {4, 4, 4, 5}}

On a exclu Q_2=Me pour obtenir 4 groupes égaux.

c) Cas où n = 4p + 2

Dans ce cas, le reste de la division par 4 est 2 et l'on a :

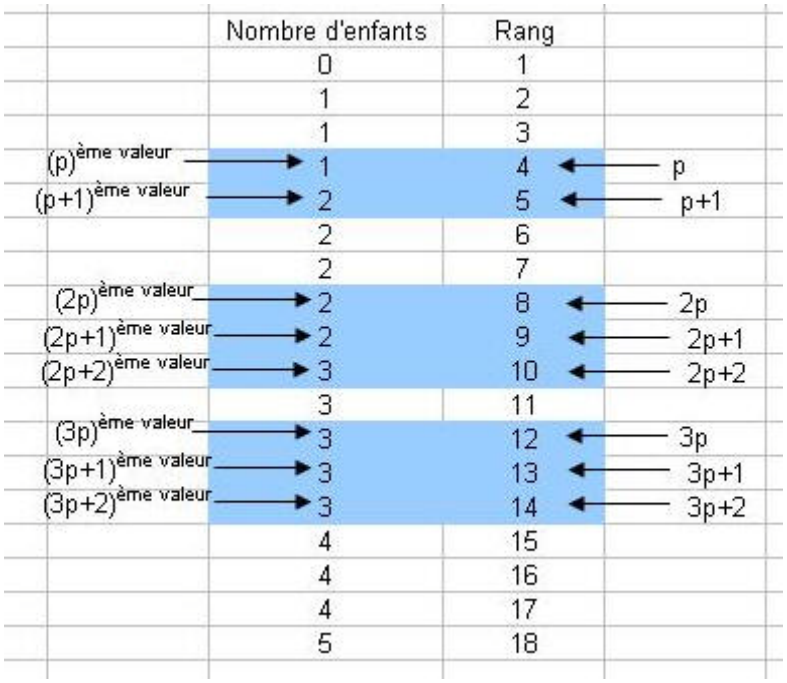
Q_1 =$(p+1)^e$ valeur.

Q_2 = moyenne entre la $(2p+1)^e$ valeur et la $(2p+2)^e$ valeur.

Q_3 = $(3p+2)^e$ valeur

Exemple : si l'on prend la série {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3}, on a n=18 et n = 4p+2, avec p= 4.

En classant cette série on obtient le tableau suivant :



Fichier EXCEL

Par conséquent, on a :

Q_1 =$(p+1)^e$ valeur = 2

Q_2 = moyenne entre la $(2p+1)^e$ valeur et la $(2p+2)^e$ valeur = (2+3)/2=2,5

Q_3 = $(3p+2)^e$ valeur =3

Les 4 groupes de valeurs sont : {{0, 1, 1, 1}, **2**, {2, 2, 2, 2}, {3, 3, 3, 3}, **3**, {4, 4, 4, 5}}

On a exclu Q_1 et Q_3 pour obtenir 4 groupes égaux.

d) Cas où n = 4p + 3

Dans ce cas, le reste de la division par 4 est 3 et l'on a :

Q_1 =$(p+1)^e$ valeur.

Q_2 = $(2p+2)^e$ valeur.

Q_3 = $(3p+3)^e$ valeur.

Exemple : si l'on prend la série {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4} on a n=19 et n = 4p+3, avec p= 4.

En classant cette série on obtient le tableau suivant :

	Nombre d'enfants	Rang	
	0	1	
	1	2	
	1	3	
$(p)^{ème\ valeur}$ →	1	4	← p
$(p+1)^{ème\ valeur}$ →	2	5	← p+1
	2	6	
	2	7	
$(2p)^{ème\ valeur}$ →	2	8	← 2p
$(2p+1)^{ème\ valeur}$ →	2	9	← 2p+1
$(2p+2)^{ème\ valeur}$ →	3	10	← 2p+2
	3	11	
$(3p)^{ème\ valeur}$ →	3	12	← 3p
$(3p+1)^{ème\ valeur}$ →	3	13	← 3p+1
$(3p+2)^{ème\ valeur}$ →	3	14	← 3p+2
$(3p+3)^{ème\ valeur}$ →	4	15	← 3p+3
	4	16	
	4	17	
	4	18	
	5	19	

Fichier EXCEL

Par conséquent, on a :

Q_1 =$(p+1)^e$ valeur = 2

Q_2 = $(2p+2)^e$ valeur = 3

Q_3 = $(3p+3)^e$ valeur = 4

Les 4 groupes de valeurs sont : {{0, 1, 1, 1}, **2**, {2, 2, 2, 2}, **3**, {3, 3, 3, 3}, **4**, {4, 4, 4, 5}}

On a exclu Q_1, Q_2 et Q_3 pour obtenir 4 groupes égaux.

Tableau récapitulatif

Fichier EXCEL

Formules de détermination des 3 quartiles (Q_1, Q_2 et Q_3) d'une série statistique

n= nombre d'unités statistiques de la série p = quotient de la division de n par 4

$n = 4p \begin{cases} Q_1 = \dfrac{p^{ème}\ \text{valeur} + (p+1)^{ème}\ \text{valeur}}{2} \\ Q_2 = \dfrac{(2p)^{ème}\ \text{valeur} + (2p+1)^{ème}\ \text{valeur}}{2} \\ Q_3 = \dfrac{(3p)^{ème}\ \text{valeur} + (3p+1)^{ème}\ \text{valeur}}{2} \end{cases}$	$n = 4p+1 \begin{cases} Q_1 = \dfrac{p^{ème}\ \text{valeur} + (p+1)^{ème}\ \text{valeur}}{2} \\ Q_2 = (2p+1)^{ème}\ \text{valeur} \\ Q_3 = \dfrac{(3p+1)^{ème}\ \text{valeur} + (3p+2)^{ème}\ \text{valeur}}{2} \end{cases}$
$n = 4p+2 \begin{cases} Q_1 = (p+1)^{ème}\ \text{valeur} \\ Q_2 = \dfrac{(2p+1)^{ème}\ \text{valeur} + (2p+2)^{ème}\ \text{valeur}}{2} \\ Q_3 = (3p+2)^{ème}\ \text{valeur} \end{cases}$	$n = 4p+3 \begin{cases} Q_1 = (p+1)^{ème}\ \text{valeur} \\ Q_2 = (2p+2)^{ème}\ \text{valeur} \\ Q_3 = (3p+3)^{ème}\ \text{valeur} \end{cases}$

Chapitre 4
Indices et progressions

1 – Indices
- A – Définitions
  - 1) Nombre indice
  - 2) Série indice
- B – Indice temporel et indice de situation
  - 1) Indice temporel
  - 2) Indice de situation
- C – Indice élémentaire et indice synthétique
  - 1) Indice élémentaire
  - 2) Indice synthétique
- D – Indice d'évolution de la valeur d'un panier de biens
  - 1) Définition de la valeur d'un panier de biens
  - 2) Indice de LASPEYRES
    - *a) Indice d'évolution des prix*
    - *b) Indice d'évolution des quantités*

2 – Progressions
- A – Variation absolue
- B – Taux de croissance sur une période
- C – Taux de croissance sur plusieurs périodes
  - 1) Formule directe (en passant par l'accroissement global)
  - 2) Formule indirecte (en passant par les accroissements successifs)
  - 3) Exemple numérique
  - 4) Lien avec la moyenne géométrique
- D – Taux de croissance du produit de 2 valeurs
- E – Taux de croissance du rapport de 2 valeurs
- F – Compléments
  - 1) Augmentations et/ou diminutions successives
  - 2) Augmentation en % suivie d'une diminution identique en pourcentage
  - 3) Temps de doublement d'une grandeur
  - 4) Exemple d'utilisation en économie

Nous avons vu au chapitre 2 que les chiffres d'une série pouvaient être présentés sous forme d'effectifs, de pourcentages et de pourcentages cumulés. Mais ce n'est pas tout : les chiffres d'une série peuvent aussi être mis sous forme d'**indices**. De plus, il est fréquent que l'on souhaite étudier une série dont les valeurs changent au cours du temps ou bien, plus simplement, que les valeurs d'une seule série correspondent à différentes valeurs dans le temps (il s'agit alors d'une série chronologique). Dans ces deux cas, le calcul d'un **indicateur de progression** (taux de variation, taux de croissance) va permettre de résumer l'évolution avec un seul chiffre. Les indices et les progressions sont aussi utilisés pour comparer des situations (généralement deux séries dont les valeurs changent selon le lieu).

1 - Indices

A - Définition

1) Nombre indice

Un nombre indice est une mesure de la variation d'une grandeur comparée à une valeur de référence égale à 100 et appelée « base ».

Exemple : En décembre 2007, l'Indice des prix à la consommation de l'INSEE (IPC), base 100 en 1998, s'établissait ainsi :

Indice des prix à la consommation, IPC (base 100 en 1998)	décembre 2006	novembre 2007	décembre 2007	evol. sur 1 mois	evol. sur 1 an
Ensemble des ménages, France entière (métropole et DOM)	114,73	117,26	117,7	0,44	2,53

Source : http://www.insee.fr/fr/indicateur/indic_conj/indconj_frame.asp?ind_id=29 .

La valeur de référence est ici la valeur 100 en 1998. A partir de ce tableau, on peut voir que :

- les prix ont augmenté de 17,70% entre 1998 et 2007 (en 10 ans)
- les prix ont augmenté de 117,26-114,73=2,53% entre décembre 2006 et novembre 2007 (en 1 an)
- Les prix ont augmenté de 117,7-117,26=0,44% entre novembre 2007 et décembre 2007 (en un mois)

Certains indices ne sont pas exprimés par rapport à une base 100, mais par rapport à une base 1. C'est le cas de l'« indice S.I.E.R »

Exemple : L'indice de trafic routier en île de France, dit « indice S.I.E.R ». (Service Interdépartemental d'Exploitation Routière) est égal à 1 quand le trafic est fluide, c'est-à-dire quand il faut en moyenne 1 minute pour faire un km. Si l'indice est égal à 2, cela signifie que les temps de parcours sur le réseau sont deux fois plus longs que si le trafic est fluide. S'il est égal à 3, ils sont 3 fois plus longs et ainsi de suite. (Source : www.sytadin.equipement.gouv.fr).

2) Série indice

Une **série indice** est une série divisée par une de ses valeurs et éventuellement multipliée par 100.

Exemple : Soit la série $S_1 = \{1, 3, 7, 4, 8, 6, 11, 9\}$

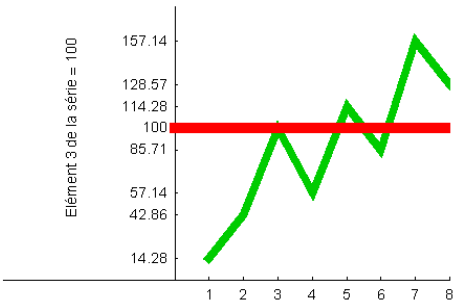
Pour transformer cette série en sérié indice, nous allons diviser chaque élément de la série par l'un des éléments, par exemple 7 (troisième élément) et ensuite multiplier chaque élément par 100. On obtient alors :

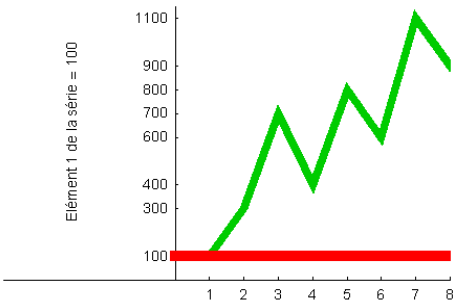
$$I_1 = \{14,3 ; 42,9 ; 100 ; 57,1 ; 114,3 ; 85,7 ; 157,1 ; 128,6\}$$

I_1 est une série indice. Sa base "100" est le troisième élément de la série. On voit ainsi que le choix de la "base" est arbitraire. On aurait pu aussi diviser par le premier élément de la série. Mais plutôt que de partir de la série initiale S_1, on peut aussi partir de I_1 et diviser chaque élément par 14,3 et multiplier par 100. On a alors effectué un **changement de base**, la nouvelle base étant le premier élément de la série.

$$I_2 = [I_1/I_1(1)]*100 = \{100, 300, 700, 400, 800, 600, 1100, 900\}$$

Les deux graphiques ci-dessous, du type "nuages de points reliés" représentent respectivement les séries indices I_1 et I_2. On remarque que le changement de base n'a pas d'incidence sur la forme de la courbe, mais seulement sur les valeurs indiquées par l'échelle de l'ordonnée (la position relative de ces valeurs les unes par rapport aux autres sur l'échelle demeurant inchangée).





B - Indice temporel et indice de situation

1) Indice temporel

Un **indice temporel** est un indice qui concerne une comparaison de valeurs dans le temps. La base est dans ce cas la date de référence.

Exemple : Une action a coté 54,10 euro à l'ouverture des marchés boursiers et 54 euros à la fermeture. L'indice de variation du cours de l'action sur la séance, donné par (54/54,1)*100 = 99,815, est un indice temporel, la base étant l'heure de l'ouverture du marché le jour considéré.

2) Indice de situation

Un **indice de situation**, également appelé **indice spatial**, est un indice qui concerne n'importe quelle comparaison de valeurs, hormis les comparaisons temporelles.

Exemple : En 2002, le nombre de victimes d'accidents de la route en France a été de 129 par million d'habitants, alors qu'au Portugal il a été de 165 par million d'habitants. L'indice de situation du nombre de victimes d'accidents est égal à (165/129)*100 = 127,9, si l'on prend le nombre d'accidents en France comme base.

C - Indice élémentaire et indice synthétique

1) Indice élémentaire

Un **indice élémentaire** est un indice qui renseigne sur l'évolution temporelle ou situationnelle (spatiale) d'**une seule valeur**. Il a pour définition :

$$I_{t/0} = \frac{V_t}{V_0} \times 100$$

Où V_0 représente la valeur de référence et V_t la valeur qui est comparée à la valeur de référence.

Dans le cas d'un indice temporel, « 0 » représente la période référence (la base) et « t » la période que l'on compare à la période de référence.

Dans le cas d'un indice de situation ou indice spatial, « 0 » représente la situation de référence (la base) et «t » la situation que l'on compare à la situation de référence.

Exemple : le « Ph », ou potentiel hydrogène de l'eau d'une piscine a été mesuré à 8h du matin. La mesure révèle qu'il est égal à sa valeur de neutralité (soit 7 sur une échelle qui varie de 1 à 14). Le soir à 18 h, on mesure à nouveau le Ph et cette valeur est alors de 5. L'indice élémentaire de la variation du Ph entre 8 h et 18 h est donné par :

$$I_{18h/8h} = \frac{5}{7} \times 100 = 71{,}43$$

2) Indice synthétique

Un **indice synthétique** est un indice qui résume l'évolution de plusieurs grandeurs : plusieurs prix, plusieurs quantités, plusieurs valeurs (prix x quantités), etc.

Exemple : Le prix d'un bien x est égal à 1,7 euros à la date 0. À la date t, il est égal à 2,5 euros. Le prix d'un bien y est égal à 3 euros à la date 0 et à 2 euros à la date t.

Les données sont résumées dans le tableau ci-après :

		Dates	
		0	**t**
Biens	**x**	1,7	2,5
	y	3	2

Nous pouvons calculer les indices élémentaires d'évolution des prix du bien x et du bien y. Mais nous pouvons aussi calculer l'indice synthétique d'évolution du prix des deux biens. Pour calculer cet indice synthétique, nous allons faire une **moyenne arithmétique** de chacun des indices élémentaires.

On obtient alors le tableau suivant :

		Dates				
		0	**t**	**Indice élémentaire**	**coefficients de pondération**	**Indice * Coefficient**
Biens	**x**	1,7	2,5	147,06	0,5	73,53
	y	3	2	66,67	0,5	33,33
						Indice synthétique 106,86

La colonne "indices élémentaires" est obtenue en divisant les prix en t par les prix en 0 et en multipliant ce rapport par 100 :

$Ix_{t/0}$ = Indice élémentaire du prix de x = (2,5/1,7)*100=147,06

$Iy_{t/0}$ = Indice élémentaire du prix de y = (2/3)*100= 66,67

Pour obtenir l'indice synthétique de l'évolution du prix des deux biens, on fait la moyenne des deux indices :

$I_{t/0}$ = Indice synthétique = (147,06 * 0,5) + (66,67*0,5)= 73,53+33,3=106,86

Lorsque l'on fait une moyenne simple, on suppose que la contribution de chaque bien à l'évolution totale est la même. En réalité cela revient à attribuer un coefficient 1/2 à chaque bien. Si l'on désigne par α_x le coefficient de pondération de x et par α_y le coefficient de pondération de y on aura α_x = 0,5 et α_y =0,5 et donc $\alpha_x + \alpha_y = 1$. L'indice synthétique d'évolution du prix des deux biens pourra alors s'écrire :

$$I_{t/0} = \alpha_x Ix_{t/0} + \alpha_y Iy_{t/0}$$

Dans le cas où α_x = 0,5 et α_y =0,5 on aura :

$$I_{t/0} = \frac{Ix_{t/0} + Iy_{t/0}}{2} = \frac{147,06 + 66,67}{2} = 106,86$$

Cependant, nous pouvons choisir de pondérer chaque bien par des coefficients α_x et α_y différents de ½ mais toujours tels que $\alpha_x + \alpha_y = 1$. Si l'on prend par exemple α_x = 1/4 et α_y = 3/4, on obtient :

$$I_{t/0} = \frac{1}{4} Ix_{t/0} + \frac{3}{4} Iy_{t/0} = \frac{1}{4} \times 147,06 + \frac{3}{4} \times 66,67 = 36,76 + 50 = 86,76$$

Et enfin si l'on prend α_x = 3/4 et α_y = 1/4, on obtient :

$$I_{t/0} = \frac{3}{4} Ix_{t/0} + \frac{1}{4} Iy_{t/0} = \frac{3}{4} \times 147,06 + \frac{1}{4} \times 66,67 = 110,29 + 16,67 = 126,96$$

L'intérêt du choix d'une pondération différente de la pondération 50/50 apparaît mieux si l'on étudie un cas particulier d'indice synthétique : l'indice d'évolution du prix d'un panier composé de plusieurs biens représentatifs, communément appelé **indice d'évolution des prix**.

D - Indice d'évolution de la valeur d'un panier de bien

En économie, on s'intéresse particulièrement à l'évolution du niveau général des prix. Cette question est délicate car chacun s'intéresse à des prix différents. Chacun a son propre panier représentatif de biens dont l'évolution des prix le préoccupe.

Malgré ces considérations qui pourraient conduire à renier la notion d'indice général des prix, la plupart des économistes se réfèrent à l'indice des prix calculé par l'INSEE (Institut National de la Statistique et des Études Économiques).

1) Définition de la valeur d'un panier de biens

La valeur de chaque produit d'un panier de bien est le produit d'un prix par une quantité. Soit $V_t^i = p_t^i x \ q_t^i$ la valeur du bien *i*, à la date t où p_t^i est le prix du bien *i* à la date t et q_t^i sa quantité. Par exemple, si $p_t^i = 3$ euros et que $q_t^i = 2$ unités, on a :

$$V_t^i = p_t^i q_t^i = 3 \times 2 = 6 \text{ euros}$$

S'il y a *n* produits dans le panier (*i= 1 à n*), la valeur totale du panier à la date *t* s'écrira :

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

Exemple : soit le tableau suivant qui donne le prix unitaire en euros et les quantités de 3 biens à la date t :

	p_t^i	q_t^i
Produit 1	15	3
Produit 2	7	9
Produit 3	3	11

La valeur du panier est alors donnée par :

$$V_t = \sum_{i=3}^{3} p_t^i q_t^i = p_t^1 q_t^1 + p_t^2 q_t^2 + p_t^3 q_t^3 = (15 \times 3) + (7 \times 9) + (3 \times 11)$$

La valeur du panier est donc égale à 141 euros.

L'évolution de la valeur du panier entre les deux dates 0 et *t* dépend de l'évolution du prix de chaque bien et de l'évolution de la quantité de chaque bien. Il faut les donc construire un **indice synthétique** qui permette d'imputer l'évolution de la valeur du panier au composant prix ou à la composante quantité. 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 fréquemment utilisé de nos jours est l'indice de LASPEYRES. C'est pourquoi nous n'étudierons que cet indice dans ce cours introductif. Le lecteur intéressé par les deux autres indices synthétiques peut se référer à l'ouvrage de Bernard PY, 2007, Statistique descriptive : nouvelle méthode pour comprendre et bien réussir 5ème édition, Economica.

L'indice de LASPEYRES permet de mesurer deux évolutions :

-- L'évolution des prix des produits composant un panier de biens (indice de LASPEYRES d'évolution des prix)

-- L'évolution des quantités de produits composant un panier de biens (indice de LASPEYRES d'évolution des quantités)

2) Indice de LASPEYRES

a) Indice d'évolution des prix

L'**indice de LASPEYRES d'évolution des prix** mesure l'évolution, entre deux dates 0 et t, des prix des biens qui composent un panier, en prenant comme référence la valeur du panier à la date initiale (t = 0) et en supposant que les quantités de biens dans le panier n'ont pas varié entre 0 et t.

Sa définition est la suivante :

$$L^{P}_{t/0} = \frac{\sum_{i=1}^{n} p_t^i q_0^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100$$

On voit ainsi que si les prix ne changent pas entre 0 et t (c'est-à-dire si $p_t^i = p_0^i$), l'indice synthétique de LASPEYRES des prix demeure égal à 100. Pour comprendre la signification de cet indice et voir comment on le calcule, prenons un exemple concret.

Exemple : Soit le tableau ci-après, qui donne les prix et les quantités de deux produits 1 et 2, aux dates 0 et t. On peut supposer que le produit 1 est un pantalon et le produit 2 un tee shirt (voir le fichier EXCEL).

	Date 0		Date t	
Produit 1	p_0^1=15	q_0^1=3	p_t^1=22	q_t^1=10
Produit 2	p_0^2=7	q_0^2=9	p_t^2=5	q_t^2=8

Dans cet exemple, le prix du bien 1 (pantalon) augmente (de 15 à 22 euros) tandis que celui du bien 2 (tee shirts) baisse (de 7 à 5 euros).

Mais les quantités aussi ont changé. Pour diverses raisons, les gens ont acheté plus de pantalons et moins de tee-shirts. Il n'est pas nécessaire que ces quantités évoluent en sens inverse des prix car il ne s'agit pas d'une relation instantanée, mais d'une évolution dans le temps. Pour mesurer l'évolution des prix, LASPEYRES suppose donc que les quantités ne changent pas. Il pose la question : quelle serait l'évolution de la valeur de ce panier si les quantités n'avaient pas changé ?

Pour répondre à cette question et savoir si l'indice synthétique des prix ainsi défini augmente ou baisse, appliquons la formule de LASPEYRES d'évolution des prix :

$$L^P_{t/0} = \frac{\sum_{i=1}^{n} p_t^i q_0^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100 = \frac{p_t^1 q_0^1 + p_t^2 q_0^2}{p_0^1 q_0^1 + p_0^2 q_0^2} \times 100 = \frac{(22\times 3)+(5\times 9)}{(15\times 3)+(7\times 9)} \times 100 = \frac{66+45}{45+63} \times 100 = \frac{111}{108} \times 100 = 102{,}8$$

On enregistre donc une évolution des prix du panier de bien de 2,8 % selon la formule de LASPEYRES.

b) Indice d'évolution des quantités

L'indice de LASPEYRES des quantités mesure l'évolution, entre deux dates 0 et t, des quantités des biens qui composent un panier, en prenant comme référence la valeur du panier à la date initiale (t=0) et en supposant que les prix des biens dans le panier n'ont pas varié entre 0 et t.

Sa définition est la suivante :

$$L^Q_{t/0} = \frac{\sum_{i=1}^{n} p_0^i q_t^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100$$

On voit ainsi que si les quantités ne changent pas entre 0 et t (c'est-à-dire si $q_t^i = q_0^i$), l'indice synthétique de LASPEYRES des quantités demeure égal à 100. Pour comprendre la signification de cet indice et voir comment on le calcule, prenons un exemple concret.

Exemple : reprenons le tableau précédent, qui donne les prix et les quantités de deux produits 1 et 2, aux dates 0 et t voir le fichier EXCEL).

	Date 0		Date t	
Produit 1	$p_0^1=15$	$q_0^1=3$	$p_t^1=22$	$q_t^1=10$
Produit 2	$p_0^2=7$	$q_0^2=9$	$p_t^2=5$	$q_t^2=8$

Dans cet exemple, la quantité du bien 1 augmente (de 3 à 10 unités) tandis que celle du bien 2 baisse (de 9 à 8 unités). Pour savoir si l'indice synthétique des volumes augmente ou baisse, appliquons la formule de LASPEYRES d'évolution des quantités :

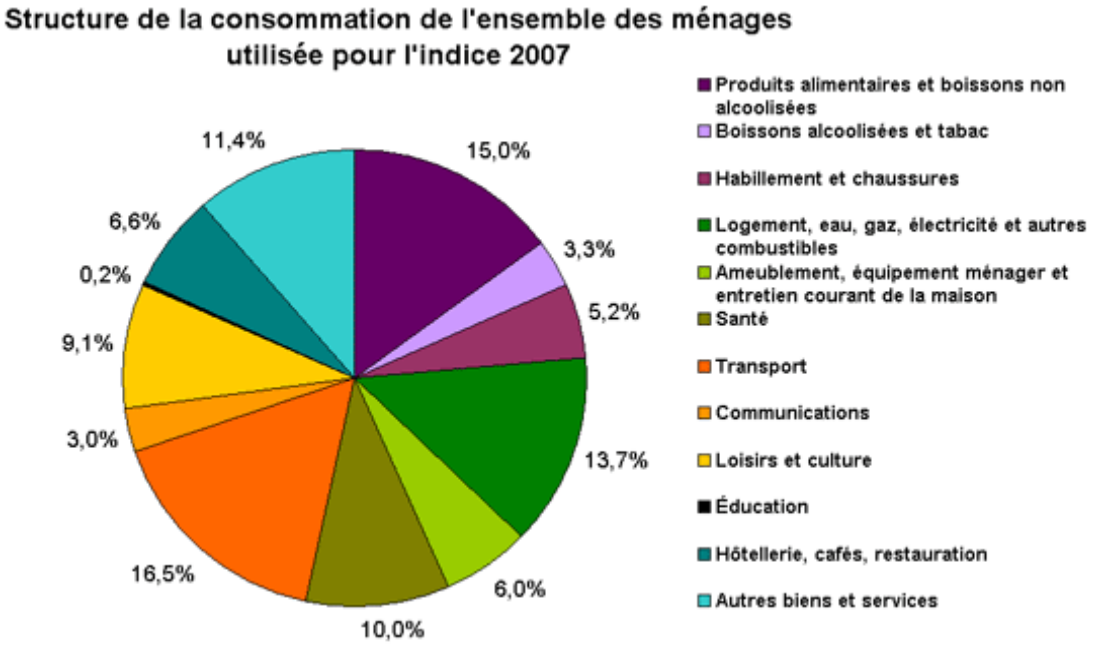
$$L^P_{t/0} = \frac{\sum_{i=1}^{n} p_0^i q_t^i}{\sum_{i=1}^{n} p_0^i q_0^i} \times 100 = \frac{p_0^1 q_t^1 + p_0^2 q_t^2}{p_0^1 q_0^1 + p_0^2 q_0^2} \times 100 = \frac{(15\times 10)+(7\times 8)}{(15\times 3)+(7\times 9)} \times 100 = \frac{150+56}{45+63} \times 100 = \frac{206}{108} \times 100 = 190{,}74$$

On enregistre donc une évolution des volumes du panier de bien de 90,74 % selon la formule de LASPEYRES.

3) Indice des prix de l'INSEE

L'un des indices synthétiques les plus connus et les plus utilisés est l'indice des prix à la consommation (IPC) publié chaque mois par l'INSEE. L'IPC permet de mesurer l'inflation, c'est-à-dire la variation du niveau général des prix des biens et des services consommés par les ménages sur le territoire français entre deux périodes données. C'est une mesure synthétique des évolutions de prix à qualité constante.

Pour le calculer, l'INSEE applique la formule de l'indice de LASPEYRES des prix à un échantillon de quelques 21000 indices élémentaires. Ces 21000 indices élémentaires sont calculés à partir de prix recueillis dans 106 agglomérations de plus de 2000 habitants réparties sur tout le territoire. L'indice couvre plus de 1000 variétés de produits, regroupées en 161 groupes. Pour éviter toute tentative de manipulation des prix, la liste précise de ces 1000 variétés de produits reste confidentielle. Actuellement, la période de référence, ou « base » de l'IPC, est 1998. Ci-après, le camembert qui donne la structure des pondérations par grandes catégories de consommation.



Source : Insee, http://www.insee.fr/fr/indicateur/indic_cons/info_ipc.htm

L'IPC est publié aux environs du 13 de chaque mois et porte sur l'évolution des prix du mois précédent. Ce chiffre, régulièrement relayé par les médias, est très attendu car il sert de multiples fonctions économiques parmi lesquelles la connaissance de l'inflation, la définition des objectifs de la politique monétaire, mais aussi le versement de pensions et de divers revenus, tels le SMIC, dont le montant est « indexé » sur l'évolution de l'IPC.

Depuis le milieu de l'année 2006, l'indice des prix a fait l'objet de critiques et de controverses. Selon l'économiste Florence JANY-CATRICE, "*Les réflexions les plus intéressantes qui ont été dressées pour éclairer ces critiques sont celles qui mettent en avant l'idée qu'avec la fin des classes moyennes, ou plus humblement, la fin du "Français moyen", il devient délicat pour les individus de s'identifier à l'espace de référence commun dressé par l'Insee (c'est-à-dire le panier moyen de la ménagère) qui, transformé en coefficients budgétaires, est un élément central du calcul de l'indice*" (Le Monde, 5 Mars 2007, "L'acte de naissance du chacun-pour-soi")

Pour répondre à ces critiques l'Insee met sur son site Internet, à disposition, un simulateur qui permet instantanément de mesurer "son" indice des prix personnels. Il est disponible à ce lien : http://www.insee.fr/fr/indicateur/indic_cons/indic_sip.htm

Selon l'économiste Florence JANY-CATRICE, c'est d'abord "l*a fin de l'indice des prix comme outil collectif de revendication d'un pouvoir d'achat négocié au sein de relations industrielles. Attaché à l'Etat-providence, l'indice des prix à la consommation était un des dispositifs majeurs de régulation du marché du travail salarié. Avec l'éclatement des lieux de négociation et plus généralement la faiblesse de la négociation collective comme mode de gouvernance, les usages principaux de l'indice des prix moyen sont en partie perdus*". Et c'est ensuite "*la prévalence d'une*

logique d'Etat néolibéral qui associe à son mode de gouvernance une démultiplication des centres de calcul, et une individualisation croissante des dispositifs évaluatifs, comme l'a décrit depuis longtemps déjà Alain DESROSIERES, de l'Insee. L'indice des prix calculé par chacun devient ici un mode de description statistique personnalisé, qui, sans avoir de statut légal, permet à chaque "agent" d'adopter un comportement rationnel, optimal".

Finalement, voici le tableau résumé de l'évolution des prix, tel que publié par l'INSEE en Décembre 2007 (régulièrement mis à jour et disponible à l'URL : http://www.insee.fr/fr/indicateur/indic_conj/indconj_frame.asp?ind_id=29) (voir le tableau ci-après)

	décembre 2006	novembre 2007	décembre 2007	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)	114,73	117,26	117,70	0,4	2,6
Ensemble cvs (00 C)	114,87	117,41	117,87	0,4	2,6
Alimentation (4000 E)	116,14	118,95	119,69	0,6	3,1
Tabac (0221 E)	178,81	189,94	189,95	0,0	6,2
Produits manufacturés (4003 E)	100,85	101,03	101,21	0,2	0,4
Énergie (4007 E)	134,57	148,11	148,88	0,5	10,6
Services (4009 E)	118,81	121,19	121,68	0,4	2,4
Alimentation y c. Tabac (4014 E)	121,97	125,37	126,07	0,6	3,4
Manufacturés y c. Energie (4015 E)	106,86	109,17	109,44	0,2	2,4
Manufacturés hors Habillement et chaussures (4016 E)	100,11	100,25	100,40	0,1	0,3
Ensemble hors loyers et hors tabac (5000 E)	113,20	115,56	116,00	0,4	2,5
Ensemble hors énergie (4017 E)	113,22	114,96	115,37	0,4	1,9
Ensemble hors tabac (4018 E)	113,59	116,02	116,46	0,4	2,5
Ménages urbains dont le chef est ouvrier ou employé, France entière (métropole et DOM)					
Ensemble hors Tabac (4018 D)	113,57	115,93	116,36	0,4	2,5
Ensemble (00 D)	115,25	117,76	118,19	0,4	2,6
Inflation sous-jacente					
Ensemble des ménages, France métropolitaine					
Ensemble «sous jacent» (4022 S)	112,17	113,91	114,17	0,2	1,8
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)	102,39	104,83	105,26	0,4	2,8

Les codes entre parenthèses correspondent à des regroupements d'intérêt pour l'analyse économique, appelés regroupements conjoncturels. Ils sont repris en particulier dans les tableaux détaillés de l'Informations Rapides sur les prix à la consommation et dans les fichiers téléchargeables du mois.

Source : Insee

2 - Progressions

Soit la valeur numérique V au temps t qui s'écrit respectivement V_0 quand t=0, V_1 quand t=1 et V_t quand t est une date quelconque.

A - Variation absolue

La variation absolue de la valeur V entre le temps 0 et le temps t s'écrit :

$$\Delta V = V_t - V_0$$

Exemple : Soit V_0 = 2 104 967 à fin novembre 2006 le nombre de chômeurs (de catégorie 1, c'est-à-dire ceux qui recherchent un emploi à temps plein, en contrat à durée indéterminée et qui n'ont pas travaillé plus de 78 heures au cours du mois- les demandeurs d'emploi peuvent en effet avoir une « activité réduite » tout en restant inscrits pour éviter un éloignement du marché du travail) en France et V_1 = 1 907 100 ce nombre à fin novembre 2007 (ce sont les chiffres donnés par le Ministère du travail, des relations sociales et de la solidarité : http://www.travail-solidarite.gouv.fr/etudes-recherche-statistiques-dares/statistiques/chomage/les-indicateurs-conjoncturels/les-dernieres-statistiques-mensuelles-novembre-2007-7187.html).

Le taux de variation entre fin novembre 2006 et fin novembre 2007 est alors égal à :

$$V_1 - V_0 = 1\,907\,100 - 2\,104\,967 = -197\,867$$

B - Taux de croissance sur une période

Le taux de croissance sur une période de la valeur V entre t = 0 (date du début de la période) et t =1 (date de fin de la période) s'écrit :

$$g_1 = \frac{V_1 - V_0}{V_0}$$

Calculons le taux de croissance du nombre de chômeurs entre fin novembre 2006 et fin novembre 2007 :

$$g_1 = (1\,907\,100 - 2\,104\,967)/\,2\,104\,967 = -197\,867/2\,104\,967 = -0{,}094$$

Le taux g s'exprime fréquemment en pourcentages. Dans ce cas on le multiplie par 100. On dira ainsi que le nombre de chômeurs a baissé de 9,4 % de fin novembre 2006 à fin novembre 2007.

C - Taux de croissance sur plusieurs périodes : formule du taux moyen

1) Formule de calcul directe (en passant par l'accroissement global)

La formule du taux de croissance moyen sur plusieurs périodes est donnée par l'expression :

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

où g = taux de croissance moyen, V_0 = grandeur à la date 0, V_t = grandeur à la date t et le t qui figure dans l'exposant 1/t est le nombre de périodes sur lequel le taux moyen est calculé. Pour voir comment on obtient cette formule, on va supposer que le taux moyen g se substitue au taux de croissance de chaque période g_1, g_2, ..., g_t dans la formule du taux de croissance sur une seule période. Autrement dit, supposons que $g= g_1, g_2, ..., g_t$. Dans ce cas :

$$g = g_1 = \frac{V_1 - V_0}{V_0}$$

pour la première période. Cette expression peut s'écrire :

$$V_1 = V_0(1+g)$$

Pour la période 2, on aura de même :

$$V_2 = V_1(1+g)$$

Ce qui peut s'écrire :

$$V_2 = V_1(1+g) = V_0(1+g)^2$$

Et ainsi de suite pour les autres périodes jusqu'à la période t pour laquelle on aura :

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

Donc, en reformulant cette expression :

$$V_t = V_0(1+g)^t \Leftrightarrow \frac{V_t}{V_0} = (1+g)^t \Leftrightarrow g = \left[\frac{V_t}{V_0}\right]^{\frac{1}{t}} - 1$$

2) Formule indirecte (en passant par les accroissements successifs)

Le taux de croissance moyen g peut aussi s'écrire sous forme d'une moyenne géométrique des taux de croissance de chaque période g_1, g_2, g_3, ..., g_t. En effet, on a :

$$g = \left[\frac{V_t}{V_0}\right]^{\frac{1}{t}} - 1 = \left[\frac{V_t}{V_{t-1}} \times \frac{V_{t-1}}{V_{t-2}} \ldots \frac{V_2}{V_1} \times \frac{V_1}{V_0}\right]^{\frac{1}{t}} - 1$$

Or :

$$\frac{V_t}{V_{t-1}} = 1 + g_t$$

.....

$$\frac{V_2}{V_1} = 1 + g_2$$

$$\frac{V_1}{V_0} = 1 + g_1$$

Donc, en remplaçant :

$$g = \left[(1+g_t)(1+g_{t-1})\ldots(1+g_2)(1+g_1)\right]^{\frac{1}{t}} - 1 = \left[\frac{V_t}{V_0}\right]^{\frac{1}{t}} - 1$$

Il y a donc deux façons équivalentes de calculer le **taux de croissance moyen**

- **La formule directe**, en passant par l'accroissement global de V_0 à V_t :

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

- **La formule indirecte**, en passant par les accroissements successifs de V_0 à V_1, de V_1 à V_2, jusqu'à V_t :

$$g = \left[(1+g_t)(1+g_{t-1})\ldots(1+g_2)(1+g_1)\right]^{\frac{1}{t}} - 1$$

3) Exemple numérique

Soit une entreprise dont le chiffre d'affaires en euros de 2001 à 2007 est donné par le tableau ci-dessous : (Fichier EXCEL)

Années	V	Chiffre d'affaires
2003	V_0	210000
2004	V_1	280000
2005	V_2	330000
2006	V_3	450000
2007	V_4	500000

a) Calcul du taux de croissance moyen par la formule directe (en passant par l'accroissement global) :

Ici, on pose par exemple V_0 = 210000, V_1 = 280000, V_2= 330000, V_3= 450000 et V_4 = 500000.

$$g = \left[\frac{V_4}{V_0}\right]^{\frac{1}{4}} - 1 = \left[\frac{500000}{210000}\right]^{\frac{1}{4}} - 1 = 1,242189 - 1 = 0,242189$$

(Fichier EXCEL)

Soit un taux de croissance moyen égal à 24,2%.

b) Calcul du taux de croissance moyen par la formule indirecte (en passant par les accroissements successifs) :

Calculons le taux de croissance annuel du CA entre 2003 et 2004, puis entre 2004 et 2005, 2005-06 et enfin 2006-07. Nous allons ainsi avoir 4 taux de croissance g_1, g_2, g_3 et g_4. Le tableau ci-dessous résume les calculs :

		gi	1+gi
2003- 04	g_1	0,33333333	1,33333333
2004- 05	g_2	0,17857143	1,17857143
2005- 06	g_3	0,36363636	1,36363636
2006- 07	g_4	0,11111111	1,11111111

(Fichier EXCEL)

Dans le tableau ci-dessus, chaque taux de croissance a été calculé conformément à la formule :

$$g_i = \frac{V_i - V_{i-1}}{V_i}$$

Ainsi, pour la première période, 2003-04, on aura :

$$g_1 = \frac{280000 - 210000}{210000} = \frac{70000}{210000} = 0,3333$$

Et ainsi de suite pour g_2 (2004-05), g_3 (2005-06) et g_4 (2006-07).

La dernière colonne du tableau donne $(1+g_1)$, $(1+g_2)$, $(1+g_3)$ et $(1+g_4)$. Effectuons le produit :

$$(1+g_1) \times (1+g_2) \times (1+g_3) \times (1+g_4) = 1,333 \times 1,17857 \times 1,36363 \times 1,11111 = 2,38095238$$

Puis élevons ce produit à la puissance 1/4 = 0,25 :

$$[(1+g_1) \times (1+g_2) \times (1+g_3) \times (1+g_4)]^{(1/4)} = (2,38095238)^{(1/4)} = 1,242189$$

Il nous reste à calculer g :

$$1+g = 1,242189 \Leftrightarrow \mathbf{g = 0,242189}$$

Soit un taux de croissance annuel moyen en pourcentage de 24,2 % (Fichier EXCEL).

4) Lien avec la moyenne géométrique

Nous avons ainsi calculé la moyenne géométrique simple des quatre valeurs $(1+g_1)$, $(1+g_2)$, $(1+g_3)$ et $(1+g_4)$. En effet, nous avons appliqué la formule :

$$\left[\prod_{i=1}^{4}(1+g_i)\right]^{\frac{1}{4}}$$

Cette formule est un cas particulier pour n=4 de la formule plus générale de la **moyenne géométrique simple** d'une série définie par les termes $\{(1+g_1), (1+g_2), (1+g_3), ..., (1+g_n)\}$:

$$\left[\prod_{i=1}^{n}(1+g_i)\right]^{\frac{1}{n}}$$

D - Taux de croissance du produit de 2 valeurs

En économie, on raisonne fréquemment sur des valeurs qui sont en fait le produit d'un prix et d'une quantité. C'est le cas par exemple de la recette totale, dont la définition est :

$$RT = p \times q$$

p est le prix d'un produit quelconque et q sa quantité. Dans ce cas, si le prix varie et que la quantité varie aussi, on peut souhaiter calculer le taux de croissance du produit des deux, c'est-à-dire le taux de croissance de la recette totale.

Soient le prix p_t et la quantité q_t. L'évolution du prix p_t et de la quantité q_t par rapport à la période précédente peut s'exprimer ainsi :

$$p_t = (1+g_p)p_{t-1}$$

$$q_t = (1+g_q)q_{t-1}$$

où g_p et g_q sont les taux de croissance respectivement du prix et de la quantité entre t -1 et t.

Partant de ces deux expressions, la recette totale en t s'écrit alors :

$$RT_t = p_t\, q_t = (1 + g_p)(1 + g_q)p_{t-1}\, q_{t-1} = (1 + g_p)(1 + g_q)RT_{t-1}$$

On en déduit son taux de croissance entre t-1 et t :

$$g_{RT} = (RT_t / RT_{t-1}) - 1 = (1 + g_p)(1 + g_q) - 1$$

Exemple : Un commerçant augmente le prix d'un produit de 1% (g_p = 0,01). À la suite de cette augmentation, la quantité vendue baisse de 4% (g_q=-0,04). Pour connaître l'évolution de la recette totale, on va calculer son taux de croissance à partir de la formule précédente:

$$g_{RT} = (RT_t / RT_{t-1}) - 1 = (1 + g_p)(1 + g_q) - 1$$

$$g_{RT} = (1 + 0,01)(1 - 0,04) - 1 = (1,01 \times 0,96) - 1 = 0,9696 - 1 = -0,0304$$

La recette totale a diminué de 3,04% à la suite de la hausse du prix de 1% et de la baisse de la quantité de 4%.

E - Taux de croissance du rapport de 2 valeurs

De la même façon que l'on a parfois besoin de connaître le taux de croissance du produit de 2 grandeurs, il arrive aussi que l'on ait besoin de connaître le taux de croissance du rapport de deux grandeurs. C'est le cas par exemple de la productivité apparente du travail, dont la définition est :

$$\text{Productivité} = \frac{Y}{L}$$

où Y = production exprimée en euros et L = nombre d'heures travaillées. Dans ce cas, si la production varie et que le nombre d'heures de travail varie aussi, on peut calculer le taux de croissance du rapport des deux, c'est-à-dire le taux de croissance de la productivité apparente du travail.

Soient la production Y_t et le travail L_t. Leur évolution par rapport à la période précédente peut s'exprimer ainsi :

$$Y_t = (1+g_Y)Y_{t-1}$$

$$L_t = (1+g_L)L_{t-1}$$

où g_Y et g_L sont les taux de croissance respectivement de la production et du travail entre t -1 et t.

Partant de ces deux expressions, la productivité à la date t s'écrit alors :

$$\frac{Y_t}{L_t} = \frac{(1+g_Y)Y_{t-1}}{(1+g_L)L_{t-1}}$$

On en déduit son taux de croissance entre t-1 et t :

$$g_{\text{productivité}} = \frac{\frac{Y_t}{L_t}}{\frac{Y_{t-1}}{L_{t-1}}} - 1 = \frac{(1+g_Y)}{(1+g_L)} - 1$$

Exemple : La production augmente de 10% et le nombre d'heures travaillées augmente de 4%. Quelle est l'augmentation de la productivité ?

$$g_{\text{productivité}} = \frac{(1+g_Y)}{(1+g_L)} - 1 = \frac{1+0,1}{1+0,04} - 1 = \frac{1,1}{1,04} - 1 = 1,057 - 1 = 0,0576923$$

La productivité a augmenté de 5,769%.

F - Compléments

1) Augmentations (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 : Soit V_0=10 et g_1=10%, g_2=12%, g_3=-5%. Déterminer V_3.

On a :

$$V_3 = V_0 (1 + g_1)(1 + g_2)(1 + g_3)$$

Donc :

$$V_3 = V_0 (1 + g_1)(1 + g_2)(1 + g_3) = 10(1 + 0,1)(1 + 0,12)(1 - 0,05) = 10 \times 1,1 \times 1,12 \times 0,95 = 11,704$$

On peut faire la vérification pas à pas :

V_0=10

$V_1 = V_0(1+g_1)= 10(1+0,1)=11$

$V_2=V_1(1+g_2)=11(1+0,12)=12,32$

$V_3=V2(1+g_3)=12,32(1-0,05)=11,704$

2) Augmentation en pourcentage suivie d'une diminution identique (ou diminution suivie d'une augmentation)

Lorsque l'on applique à une grandeur une augmentation d'un certain pourcentage, par exemple 10%, et qu'ensuite on applique au résultat un pourcentage identique de diminution, par exemple 10%, on ne retrouve pas le chiffre de départ.

Exemple : si l'on part de V_0 =10 et que l'on applique une augmentation de 10%, on obtient V_1=11. Si l'on applique une diminution de 10% à V_1, on obtient V_2=11 x (1-0,1)=11 x 0 ,9 = 9,9, parce que 10% de 11 = 1,1 alors que 10% de 10 =1. On ajoute donc 1 à 10, puis on retranche 1,1 à 11. On se retrouve donc avec 9,9.

De même, si on applique à une grandeur une diminution d'un certain pourcentage, par exemple 10%, et qu'ensuite on applique au résultat un pourcentage identique d'augmentation, par exemple 10%, on ne retrouve pas le chiffre de départ :

Exemple : si l'on part de V_0=10 et que l'on applique une diminution de 10%, on obtient V_1= 9. Si l'on applique une augmentation de 10% à V_1, on obtient V_2=9(1+0,1)=9 x 1,1 = 9,9.

3) Temps de doublement d'une grandeur

Le temps de doublement d'une grandeur qui croît à un taux moyen constant se calcule en appliquant la formule :

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

Exemple : En combien de temps un capital placé à 5% l'an double-t-il ?

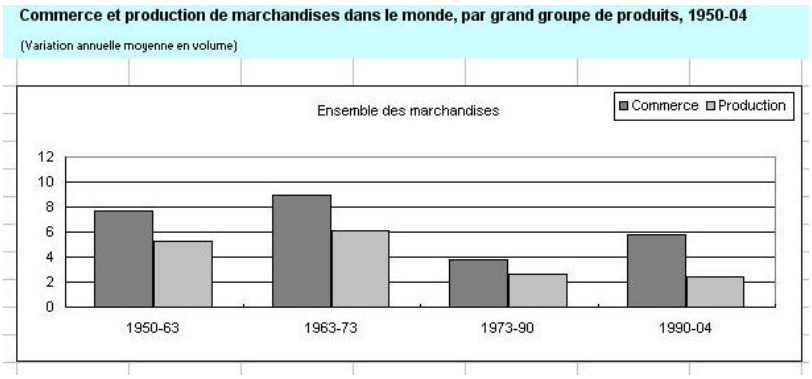
On a la formule :

$$2V_0 = V_0(1+0,05)^t \Leftrightarrow 2 = 1,05^t \Leftrightarrow \ln 2 = t \ln 1,05 \Leftrightarrow t = \frac{\ln 2}{\ln 1,05} = \frac{0,6931472}{0,0487902} = 14,206$$

Il faut donc 14 ans plus 0,2 x 12 mois = 2,4 mois pour qu'un capital placé à 5% l'an double. C'est-à-dire 14 ans, 2 mois et 0,4 x 30 jours = 12 jours. Soit 14 ans, 2 mois et 12 jours.

4) Exemple d'utilisation en économie

Les taux de croissance, et en particulier les taux de croissance annuels moyens, sont très utilisés en économie. On les représente parfois sous la forme de graphiques, comme dans l'exemple ci-dessous, tiré du site Internet de l'organisation mondiale du commerce, qui donne les taux de croissance du commerce mondial et de la production, par périodes. Le graphique illustre un point qui est fréquemment souligné : le commerce mondial est plus dynamique que la production. **Il a un taux de croissance annuel moyen plus rapide**.



Voir le fichier EXCEL complet

Source : OMC,
http://www.wto.org/french/res_f/statis_f/its2005_f/its05_longterm_f.htm

Chapitre 5
Diagrammes et graphiques

1 – Utilité des graphiques
- A – Qualités d'un bon graphique
- B – Quartet d'ANSCOMBE

2 – Les échelles graphiques
- A – Echelles numériques
- B – Echelle catégorielles
  1) catégories numériques
  2) catégories nominales
- C – Echelles ordinales
- D – Echelles verticales doubles
- E – Echelles logarithmiques
  1) Définition
  2) Calcul pratique du log décimal d'un nombre
  3) Rappels sur le log décimal
  4) Exemples
     a) L'échelle log pour mieux voir les différences de progression
     b) L'échelle log linéarise les évolutions à taux constant

3 – Diagrammes
- A - Pictogramme
- B - Cartogramme
- C – Diagramme de GANTT

4 – Graphiques usuels
- A - Graphique en barres
  1) Barres verticales
     a) Simples
     b) Multiples
     c) Tronçonnées
  2) Barres horizontales
     a) Simples
     b) Multiples
     c) Tronçonnées
- B – Courbes et aires
  1) Courbe simple
  2) Courbes multiples
  3) Aires délimitées par des courbes
- C - Graphique de dispersion ou nuage de points
- D - Secteurs
  1) Secteurs à 360 degrés
  2) Secteurs à 180 degrés
  3) Méthode de construction
     a) 360 degrés
     b) 180 degrés
  4) Anneaux
     a) Simples
     b) concentriques

5 – Autres graphiques
A – Graphiques en radar et toiles d'araignée
1) Radar
2) Toile d'araignée
B – Graphique à bulles
C – Graphiques boursiers
D – Graphique de TUKEY
E – Graphiques panachés
1) Secteur complété par une barre tronçonnée
2) Graphique de PARETO
a) De la loi de Pareto au graphique de Pareto
b) Définition, construction, exemple et interprétation
c) Interprétation
F – Histogramme
1) amplitudes de classes identiques
a) Histogramme d'effectifs
b) Histogramme de fréquences
2) Amplitudes de classes différentes
a) Histogramme d'effectifs
b) Histogramme de fréquences
G – Pyramide des âges
H – Graphique en cascade
I – Graphique en trois dimensions
1) Graphique en 2D avec ajout de « profondeur »
2) Graphique en barres avec 3 dimensions réelles

1 – Utilité des graphiques

A – Qualités d'un bon graphique

Selon Edward TUFTE[8], un des meilleurs spécialistes contemporains des graphiques, l'excellence en matière de graphiques statistiques consiste à communiquer avec clarté, précision et efficacité des idées complexes. Ensuite, il énumère neuf caractéristiques d'un « excellent graphique ».

Ainsi, un excellent graphique devrait-il avoir tout ou partie des qualités suivantes :

- Montrer les données.
- Attirer l'attention du lecteur ou de l'auditoire sur l'idée essentielle que le graphique vise à mettre en évidence plutôt que sur les qualités esthétiques du graphique lui-même.
- Eviter de déformer le message contenu dans les chiffres
- Présenter un grand nombre de chiffres dans un espace restreint
- Donner de la cohérence à de vastes ensembles de données
- Faciliter les comparaisons visuelles entre différents chiffres
- Révéler les chiffres à différents niveaux de détails, allant de la vision d'ensemble à une structure plus fine.
- Servir un objectif clair : décrire des données, explorer des données, ou simplement les tabuler ou leur donner un aspect esthétique.
- Etre étroitement corrélé avec la description purement statistique ou verbale des données.

B – Quartet d'ANSCOMBE

Les graphiques révèlent des informations sur la forme des séries que les tableaux et les statistiques résumées ne peuvent pas toujours montrer. La meilleure illustration en est le quartet d'ANSCOMBE.

Dans le tableau ci-après les quatre 4 paires de séries différentes de 11 chiffres ont toutes les mêmes statistiques résumées. La moyenne des X est égale à 9 et leur écart-type est 1,94 pour les 4 séries de X. La moyenne des Y est égale 7,5 et leur écart-type à 1,94 pour les 4 séries de Y. Sans les quatre graphiques ci-après, on pourrait déduire de façon erronée que comme les 4 paires de séries ont la même moyenne et la même dispersion (en outre, elles ont le même coefficient de corrélation et la même droite de régression Y = 3 + 0,5 X [sur le calcul de la droite de régression voir le chapitre 6]), elles sont très semblables. Or, comme le montrent les 4 graphiques dits « nuages de points » qui leurs sont associées, elles ont des formes très différentes. Et ceci confirme l'adage qui dit que « *un beau graphique vaut mieux qu'un long discours* » !

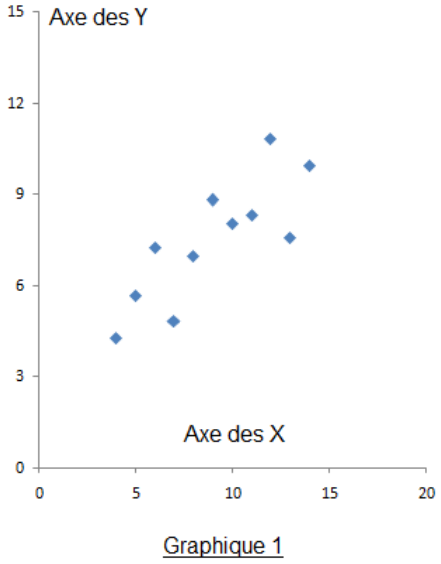
---

[8] TUFTE, Edward (2001), The Visual Display of Quantitative Information , Graphics Press. Voir le site internet de Edward TUFTE, page 13.

Quartet d'ANSCOMBE

Chiffres du graphique I

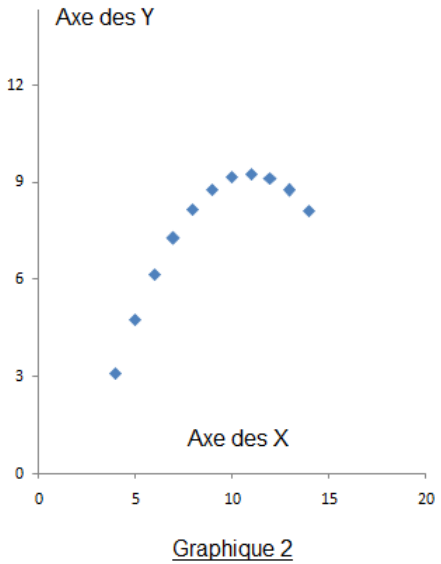
X	Y
10	8,04
8	6,95
13	7,58
9	8,81
11	8,33
14	9,96
6	7,24
4	4,26
12	10,84
7	4,82
5	5,68
Moyenne simple	9	7,5
Ecart-type	3,16	1,94



Graphique 1

Chiffres du graphique II

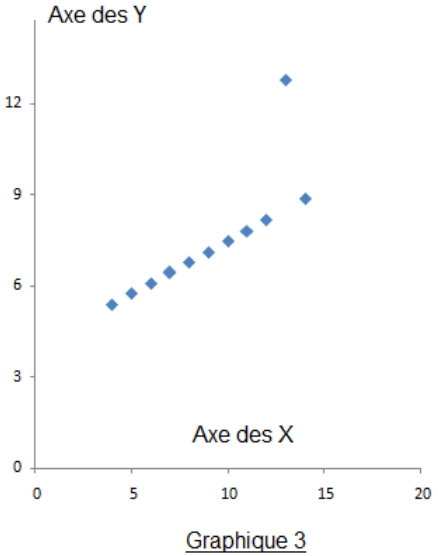
X	Y
10	9,14
8	8,14
13	8,74
9	8,77
11	9,26
14	8,1
6	6,13
4	3,1
12	9,13
7	7,26
5	4,74
Moyenne simple	9	7,5
Ecart-type	3,16	1,94



Graphique 2

Chiffres du graphique III

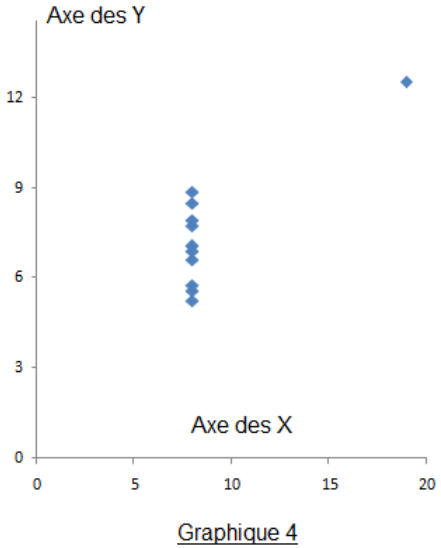
X	Y
10	7,46
8	6,77
13	12,74
9	7,11
11	7,81
14	8,84
6	6,08
4	5,39
12	8,15
7	6,42
5	5,73
Moyenne simple	9	7,5
Ecart-type	3,16	1,94



Graphique 3

Chiffres du graphique IV

X	Y
8	6,58
8	5,76
8	7,71
8	8,84
8	8,47
8	7,04
8	5,25
19	12,5
8	5,56
8	7,91
8	6,89
Moyenne simple	9	7,5
Ecart-type	3,16	1,94

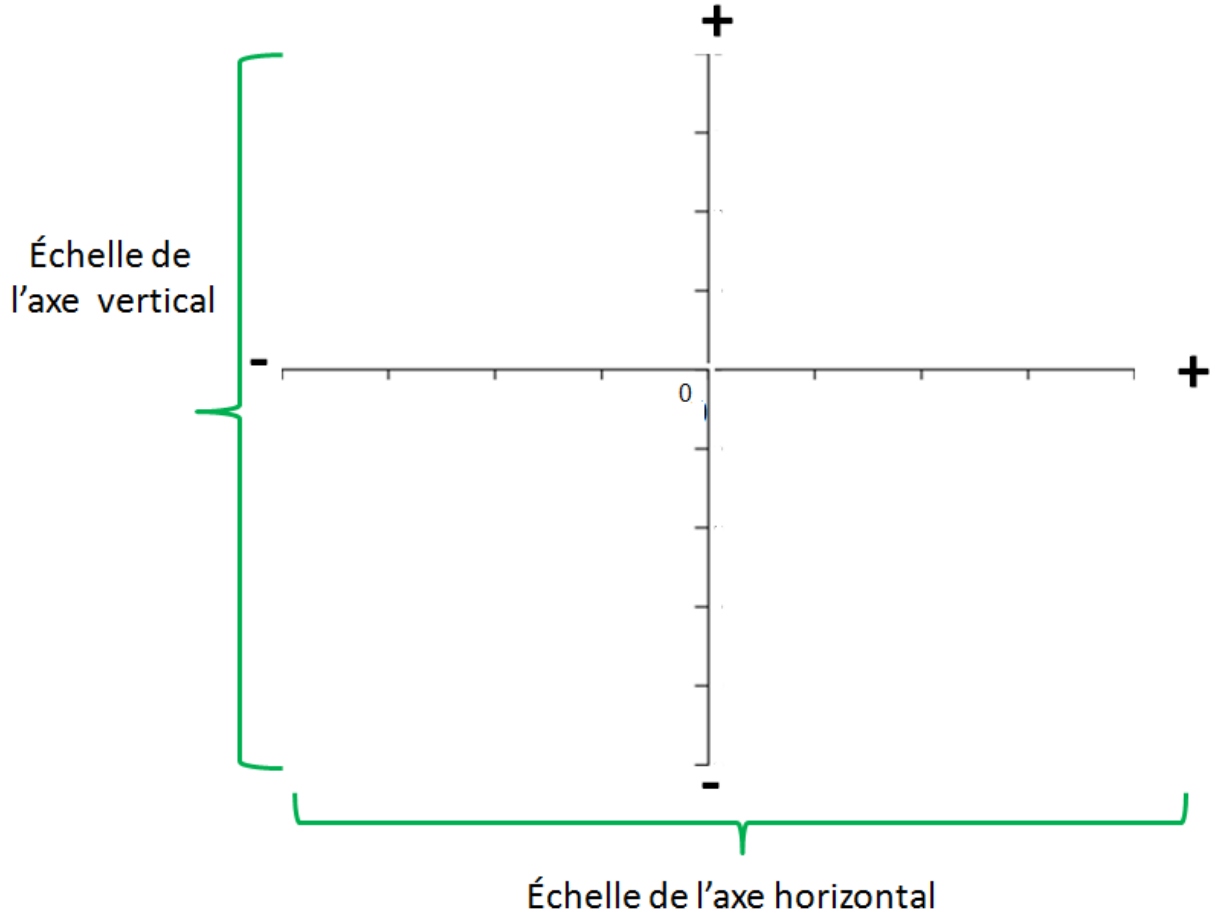


Graphique 4

Source des chiffres : F.J. ANSCOMBE, « Graphs in Statistical Analysis, » American Statistician, 27 février 1973, pages 17-21.

2 – Les échelles graphiques

Beaucoup de graphiques à deux dimensions se présentent sous la forme suivante dite d'un « système de coordonnées cartésiennes » :



Ce qui nous intéresse dans un premier temps ce sont les différentes sortes d'échelles propres à ce type de représentations graphiques.

A – Echelles numériques

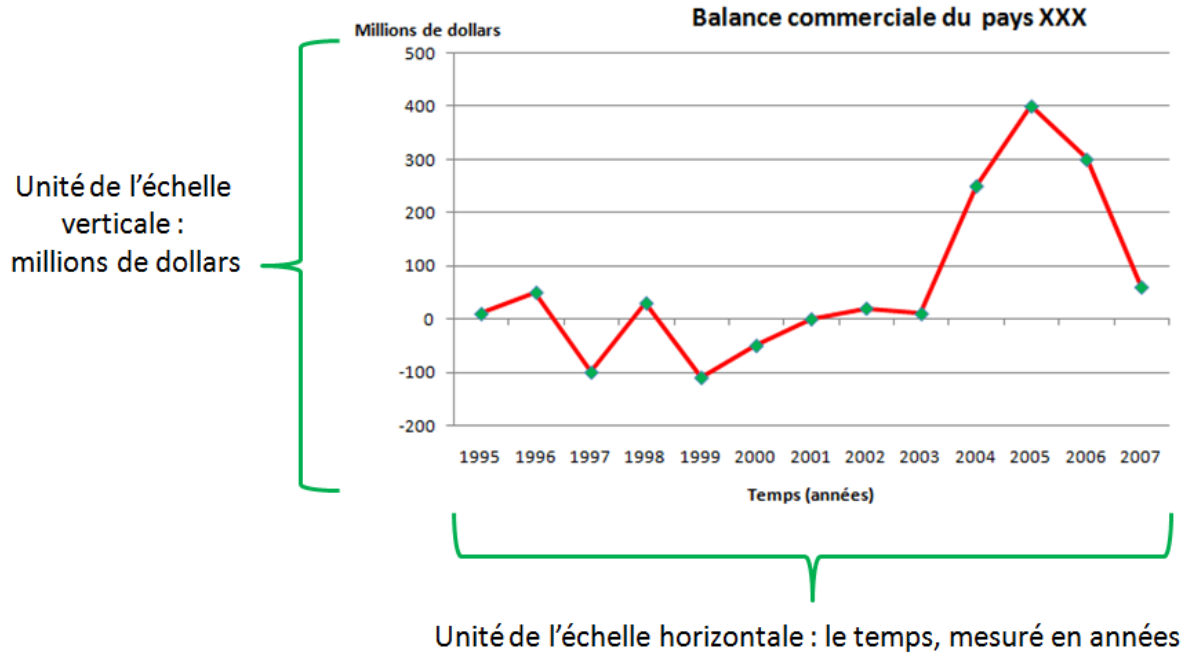
Une **échelle numérique** est une échelle qui mesure des valeurs qui peuvent varier de moins l'infini à plus l'infini. Ci-après, un graphique avec une échelle numérique sur l'axe horizontal et une échelle numérique sur l'axe vertical.

- Sur **l'axe horizontal**, l'unité de mesure numérique est l'année. Les valeurs s'échelonnent entre 1995 et 2007. Si l'on doit dessiner ce graphique à la main sur une feuille de papier, on prendra soin de définir la distance que l'on souhaite consacrer à une année (par exemple : 1 an = 1 cm). Si c'est un logiciel qui réalise le graphique, cette opération devient inutile car les dimensions du graphique seront choisies par défaut (il est possible cependant de les modifier à son gré en redimensionnant le graphique).

- Sur **l'axe vertical**, l'unité de mesure numérique est le déficit commercial d'un pays, mesuré en millions de dollars. Il varie de -150 millions à + 400 millions.

Si l'on doit dessiner ce graphique à la main sur une feuille de papier, on prendra soin de définir la distance que l'on souhaite consacrer à 100 millions de dollars (par exemple : 100 millions de dollars = 1 cm). Si c'est un logiciel qui réalise le graphique, cette opération devient inutile car les dimensions du graphiques seront choisies par défaut (il est possible cependant de les modifier à son gré).

Un exemple d'échelles numériques sur les deux axes



Remarque : il est important de toujours bien stipuler l'unité dans laquelle l'échelle est mesurée (ici l'unité est l'année pour l'axe horizontal et les millions de dollars pour l'axe horizontal)

B – Echelle catégorielles

Une **échelle catégorielle** est une échelle sur laquelle sont portées des catégories. Il peut s'agir :

- De catégories numériques (échelle quantitative)
- De catégories non numériques dites « nominales » (échelle qualitative)

1) Catégories numériques

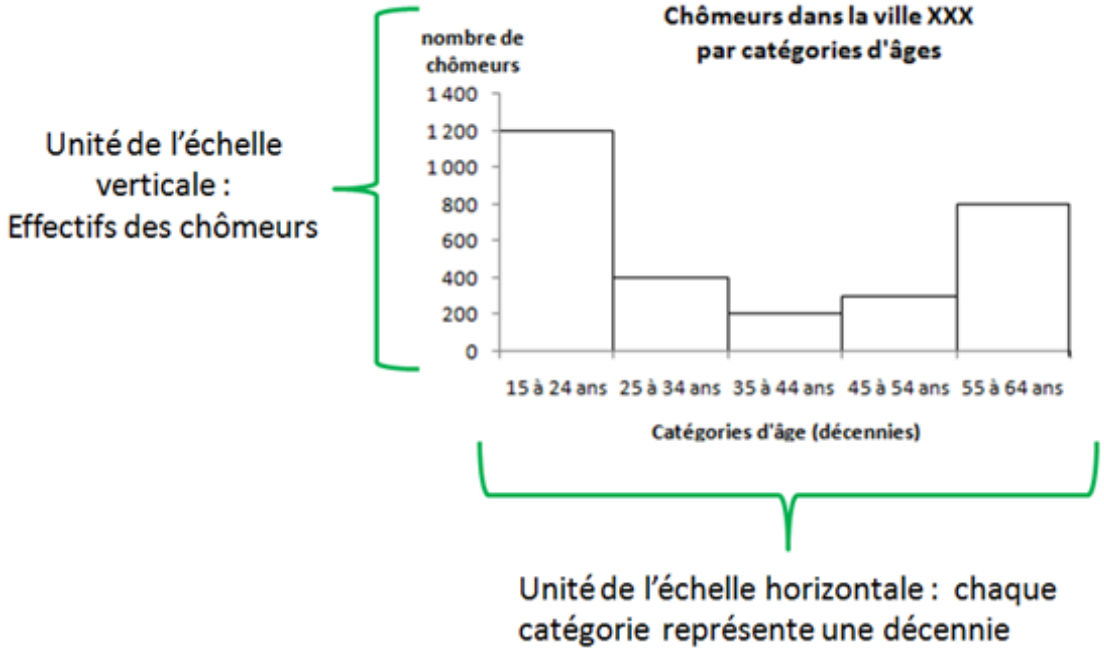
Ci-après un graphique où l'on a regroupé les chômeurs d'une ville par classes d'âges :

- **Sur l'axe horizontal figurent les catégories d'âges**. C'est une échelle de catégories ou catégorielle. L'unité est l'âge. On remarque aussi que les

classes d'âge ont la même amplitude c'est-à-dire que toutes les catégories d'âges ont le même nombre d'années. Il existe aussi des échelles de catégories d'amplitude différentes.

- **Sur l'axe vertical figurent l'effectif des chômeurs qui entrent dans chaque catégorie**. C'est une échelle numérique simple. L'unité est le nombre des chômeurs.

Exemple d'utilisation d'une échelle
avec catégories numériques sur l'axe horizontal



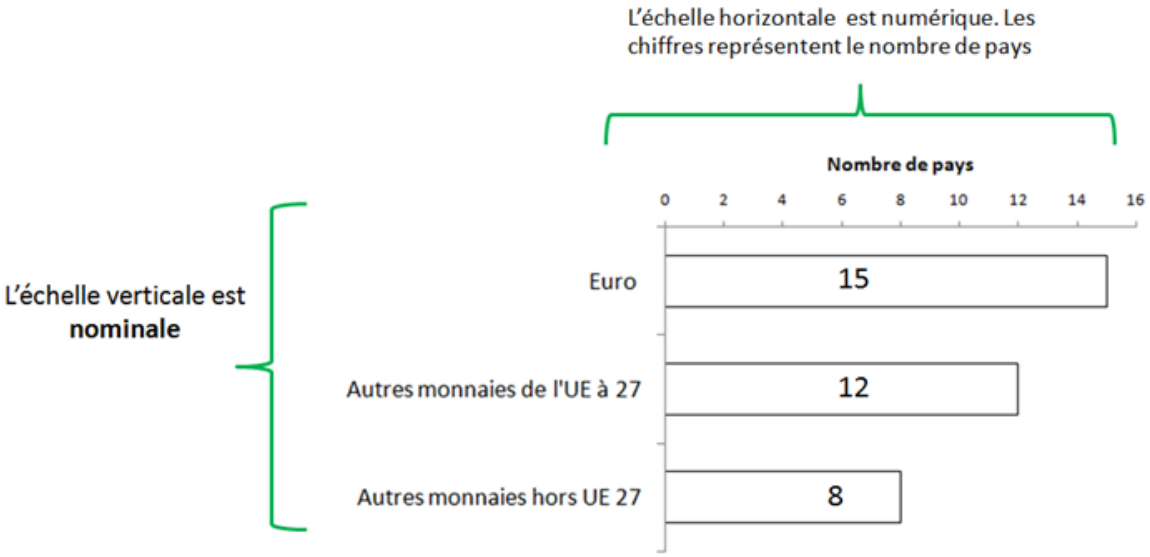
2) Catégories nominales

Ci-après un graphique où l'on a regroupé les 35 pays du tableau 1 selon la monnaie utilisée. On a créé trois catégories :

- La catégorie des pays qui font partie de la zone Euro en 2008
- La catégorie des pays de l'UE 27 qui n'en font pas partie et utilisent de ce fait d'autres monnaies
- La catégorie des pays hors UE qui sont représentés dans le tableau 1.

L'échelle de **l'axe vertical** est donc une échelle nominale. On a placé l'échelle nominale sur l'axe vertical car ainsi il est plus commode d'écrire ce que signifie chaque barre. L'échelle de **l'axe horizontal,** quant-à-elle, est numérique, elle mesure le nombre de pays appartenant à chaque catégorie.

Exemple d'utilisation d'une échelle
avec catégories nominales sur l'axe vertical



C – Echelles ordinales

Une **échelle ordinale** est une échelle sur laquelle un ordonnancement des modalités est concevable. Il peut s'agir :

- **D'un classement de préférences**. C'est souvent le cas dans les enquêtes et les sondages d'opinion.

- **D'un classement de rang**. On peut par exemple demander à des investisseurs de classer une liste de pays du plus attractif au moins attractif. Ce classement ne doit pas être confondu avec une échelle numérique simple. En effet, bien qu'il s'agisse de chiffres, l'écart entre les chiffres n'a pas de signification.

Le tableau et le graphique ci-après reproduisent des données fournies par le site internet de la Banque mondiale intitulé « Doing Business en 2008 »[9]). La Banque mondiale a classé 178 pays, parmi lesquels 33 des 35 pays du tableau 1 (Chypre et Malte ne figurent pas dans le classement de la Banque Mondiale), selon divers critères qui tentent d'appréhender l'attractivité des pays pour les investisseurs potentiels. Le fait que le classement soit représenté sous forme numérique ne doit pas laisser penser cependant qu'il s'agit d'une échelle numérique. En effet, ces chiffres ne représentent que des rangs.

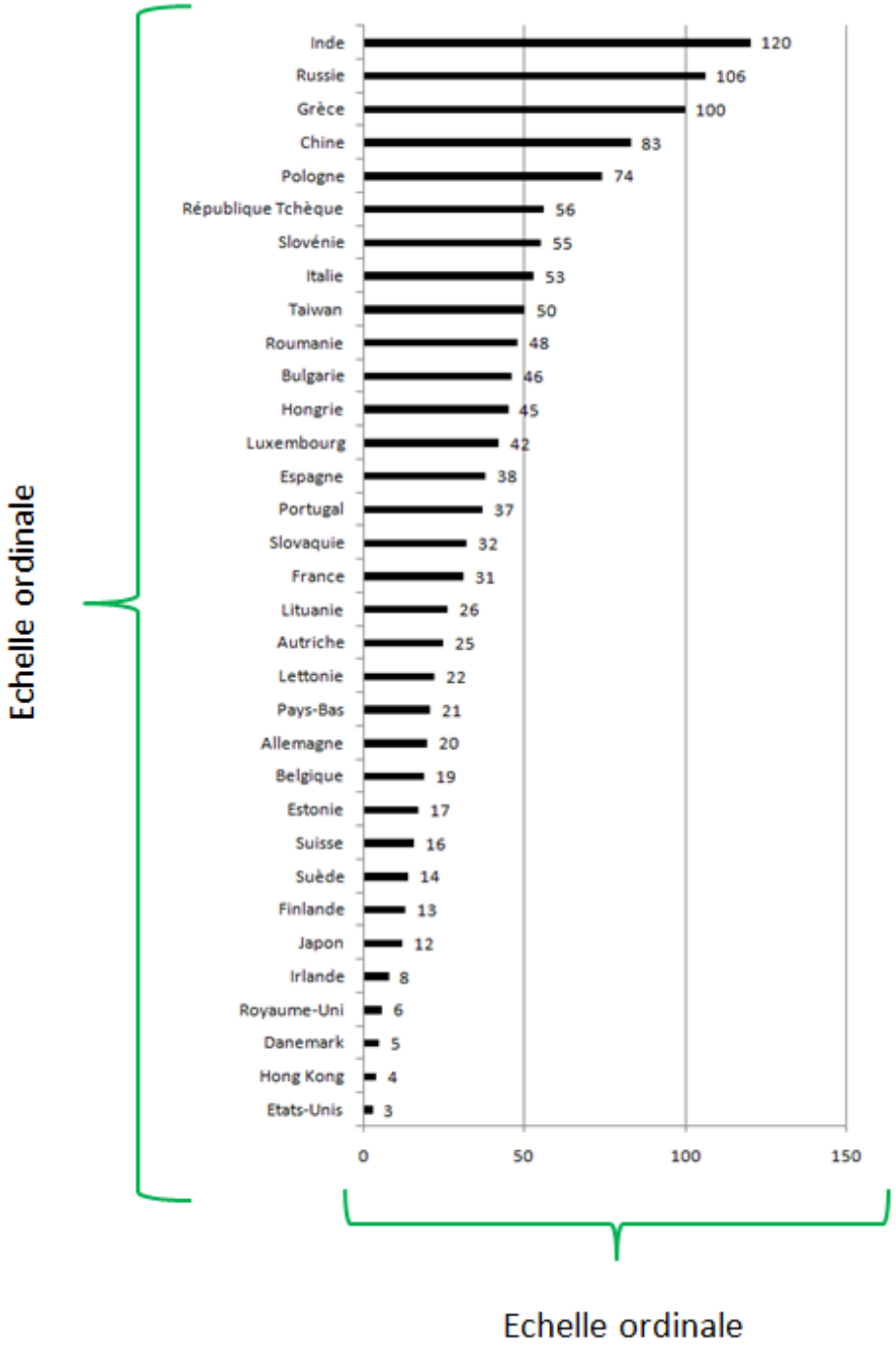
[9] Voir http://www.doingbusiness.org/economyrankings/ (Le classement qui figure dans cette version du cours a été relevé le 15/12/2007 et peut donc être différent de celui qui figure sur le site internet donné en référence).

Classement des pays du tableau 1 **selon le critère de la Banque Mondiale « Doing Business » - Année 2008**

Pays	Classement des pays selon le critère "Ease of Doing Business"
Etats-Unis	3
Hong Kong	4
Danemark	5
Royaume-Uni	6
Irlande	8
Japon	12
Finlande	13
Suède	14
Suisse	16
Estonie	17
Belgique	19
Allemagne	20
Pays-Bas	21
Lettonie	22
Autriche	25
Lituanie	26
France	31
Slovaquie	32
Portugal	37
Espagne	38
Luxembourg	42
Hongrie	45
Bulgarie	46
Roumanie	48
Taiwan	50
Italie	53
Slovénie	55
République Tchèque	56
Pologne	74
Chine	83
Grèce	100
Russie	106
Inde	120
Chypre	Non classé
Malte	Non classé

Source : http://www.doingbusiness.org/economyrankings/ (classement relevé le 15 décembre 2007)

Exemple d'échelles ordinales
(classement de pays sur l'axe vertical et sur l'axe horizontal)



Source : http://www.doingbusiness.org/economyrankings/ (classement relevé le 15 décembre 2007)

L'**échelle horizontale** du graphique ci-dessus n'est qu'apparemment numérique. En fait elle donne le classement du pays. Il s'agit donc bien d'une **échelle ordinale** car l'écart qui sépare les pays n'est pas quantifiable. Par exemple, les Etats-Unis sont classés au 3ème rang et l'Inde est classée au 120ème rang. Si l'on fait la différence

120-3 = 117, on ne peut pas en conclure que les Etats-Unis sont 117 fois plus attractifs que l'Inde du point de vue des investissements internationaux. Il ne s'agit pas d'une échelle réellement numérique, mais d'une échelle spéciale, dite « ordinale ».

L'**échelle verticale** du graphique ci-dessus est également ordinale : les pays y sont classés par ordre décroissant, du moins attractif au plus attractif.

D – Echelles verticales doubles

Lorsque que l'on veut représenter les données relatives à deux variables ou à deux caractères, on a parfois recours à une échelle double pour faciliter la lecture.

L'exemple suivant va permettre d'illustrer ce point. Le tableau ci-dessous montre deux séries mensuelles. La première indique le temps passé par une personne sur Internet chaque mois (en heures) et la seconde série indique le total de la somme dépensée sur différents sites marchands[10].

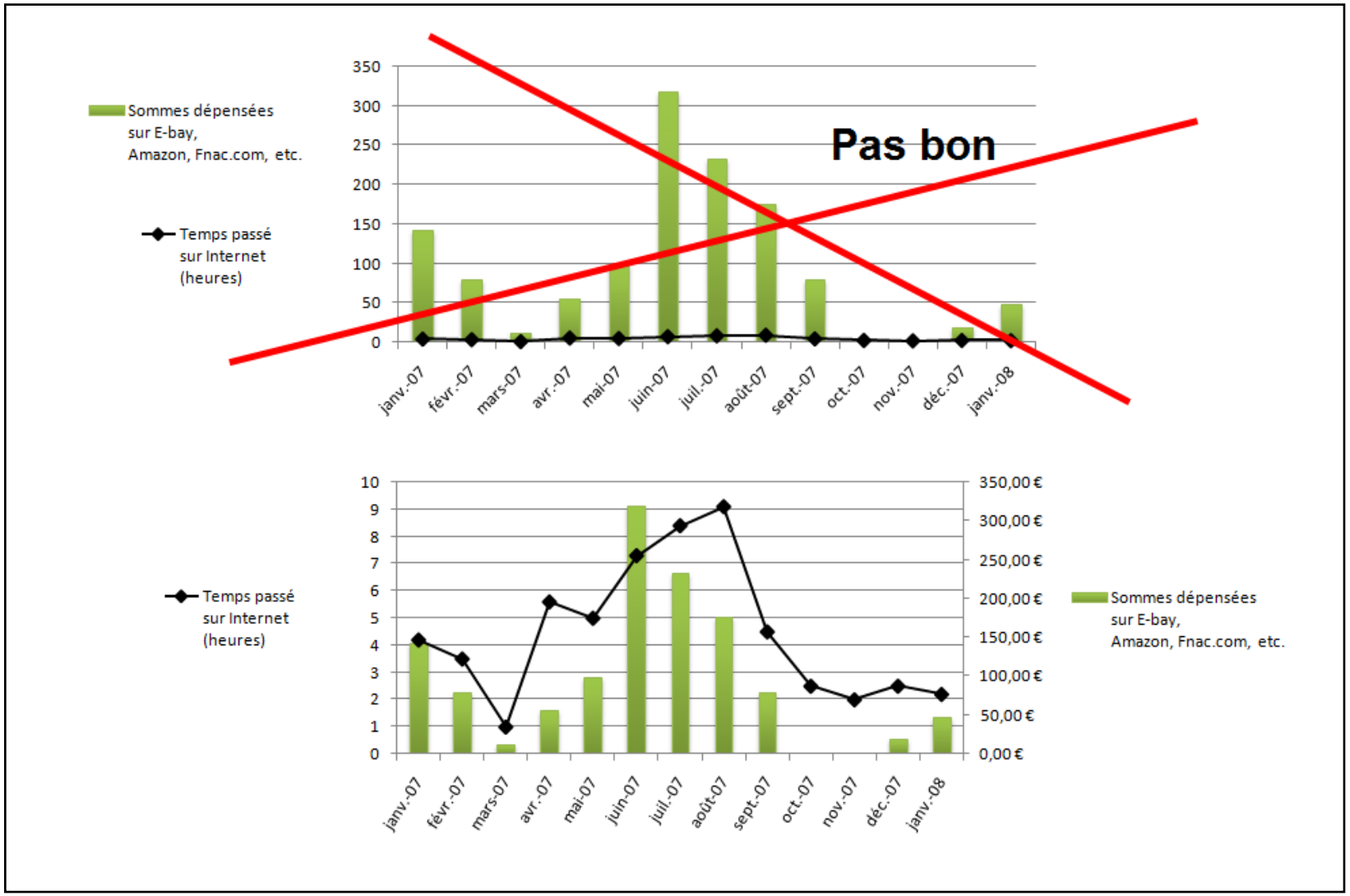
Temps passé sur Internet (heures/mois)
et sommes dépensées sur différents sites marchands (euros)

Mois	Temps passé sur Internet (heures)	Sommes dépensées sur E-bay, Amazon, Fnac.com, etc.
janv-07	4,20	142,25 €
févr-07	3,50	79,59 €
mars-07	1,00	12,50 €
avr-07	5,60	56,42 €
mai-07	5,00	98,74 €
juin-07	7,30	319,12 €
juil-07	8,40	232,58 €
août-07	9,10	175,91 €
sept-07	4,50	79,50 €
oct-07	2,50	0,00 €
nov-07	2,00	0,00 €
déc-07	2,50	19,93 €
janv-08	2,20	48,12 €

Source : Matthew McDONALD, « Creating a Combination Chart in EXCEL 2007 », Matthew McDONALD, video Youtube : http://fr.youtube.com/watch?v=WW2IDE4rPCc

[10] Exemple inspiré de « Creating a Combination Chart in EXCEL 2007 », par Matthew McDONALD, video Youtube : http://fr.youtube.com/watch?v=WW2IDE4rPCc

Exemple de l’utilité des échelles verticales doubles



Imaginons que l'on souhaite savoir s'il existe une covariation entre ces deux séries. Une bonne façon de procéder est de les mettre sous forme d'un graphique. Cependant, comme les deux échelles sont différentes, il faut réserver par exemple l'échelle verticale de gauche pour le temps passé sur internet (qui est exprimé en heures) et l'échelle verticale de droite pour les sommes dépensées (qui sont exprimées en euros). En effet, si l'on utilise seulement l'échelle verticale de gauche pour tracer les deux séries, la plus petite (celle des heures passées sur internet) sera écrasée par la plus grande (celle des euros dépensés) et le graphique ne révèlera rien du tout. Inversement, si l'on réserve une échelle pour chaque série, on obtient alors un graphique beaucoup plus lisible qui semble bel et bien révéler que plus cet individu a passé de temps sur Internet et plus il a dépensé d'argent (ceci n'est qu'un constat de covariation et non une relation de causalité, bien sûr).

E – Echelles logarithmiques

1) Définition

L'échelle logarithmique est une échelle qui mesure le logarithme décimal des valeurs de la variable. C'est un excellent moyen de mettre en évidence une idée ou un résultat grâce aux propriétés des logarithmes décimaux.

Avant de présenter le mode de construction d'un graphique à échelle semi-logarithmique, il convient cependant de faire un rappel sur les logarithmes décimaux.

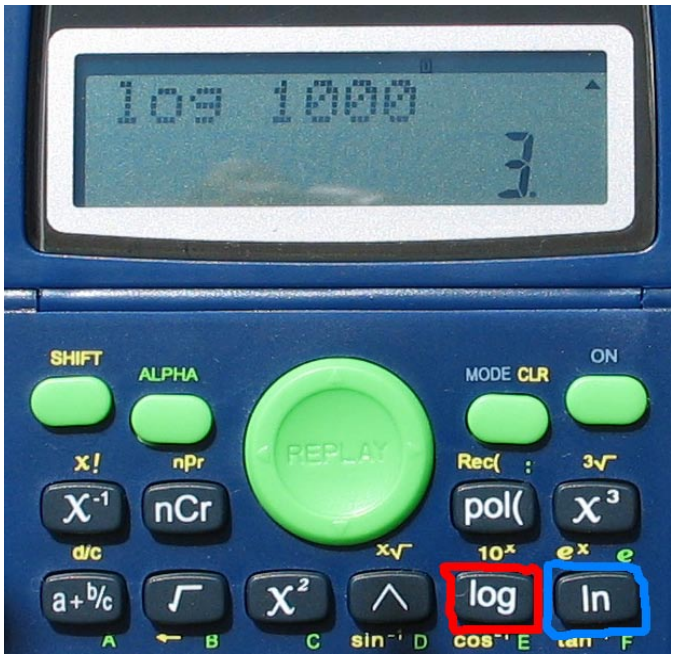
À ce sujet, sur un plan pratique, il est plus important de *savoir obtenir un logarithme décimal avec une machine à calculer*, que de c*omprendre le pourquoi et le comment des logarithmes décimaux*, ce qui est certainement passionnant mais relève d'un cours de mathématiques.

Nous allons donc commencer par voir comment on calcule un logarithme décimal avec une machine à calculer standard (ci-après la "SC-05 Plus") avant de faire un bref rappel sur les logarithmes décimaux.

2) Calcul pratique du log décimal d'un nombre

Il suffit d'appuyer sur la touche "log" d'une machine à calculer pour obtenir le log d'un nombre. Par exemple, l'image ci-dessous illustre le calcul du log décimal de 1000. L'écran indique que le log décimal de 1000 est égal à 3. Pour obtenir ce résultat, on procède ainsi :

1 - Allumer la machine
2 - appuyer sur la touche "log" entourée en rouge (attention de ne pas appuyer sur la touche "ln" qui est entourée en bleu juste à côté et qui sert à calculer les logarithmes naturels)
3 - Entrer le chiffre 1000
4 - Appuyer sur la touche "="
5 - Le résultat (ici le log de 1000 c'est 3) apparaît sur l'écran de la calculatrice.



3) Rappels sur le logarithme décimal

Le logarithme décimal d'un nombre est la puissance à laquelle il faut élever 10 pour obtenir ce nombre. Appliquons cette définition à quelques nombres. Quel est, par exemple, le logarithme décimal de 1 ? Autrement dit, à quelle puissance faut-il élever 10 pour obtenir 1 ?

La puissance à laquelle il faut élever 10 pour obtenir 1 est 0. Par conséquent, le logarithme décimal de 1 est égal à zéro:

$$10^0=1$$

On écrira donc :

$$\log 1 = 0$$

Quel est le logarithme décimal de 100 ? C'est la puissance à laquelle il faut élever 10 pour obtenir 100 :

$10^2=100$, parce qu'il faut élever 10 à la puissance 2 pour obtenir 100. Donc le logarithme décimal de 100 est égal à 2. On écrira par conséquent :

$$\log 100 =2$$

Inversement, si l'on demande "De quel chiffre 3 est-il le logarithme décimal ?", on fera le raisonnement inverse. Sachant que $10^3=1000$, la réponse est donc :

$$\log 3 = 1000$$

Autrement dit, le logarithme décimal de 1000 est égal à 3.

4) Exemples

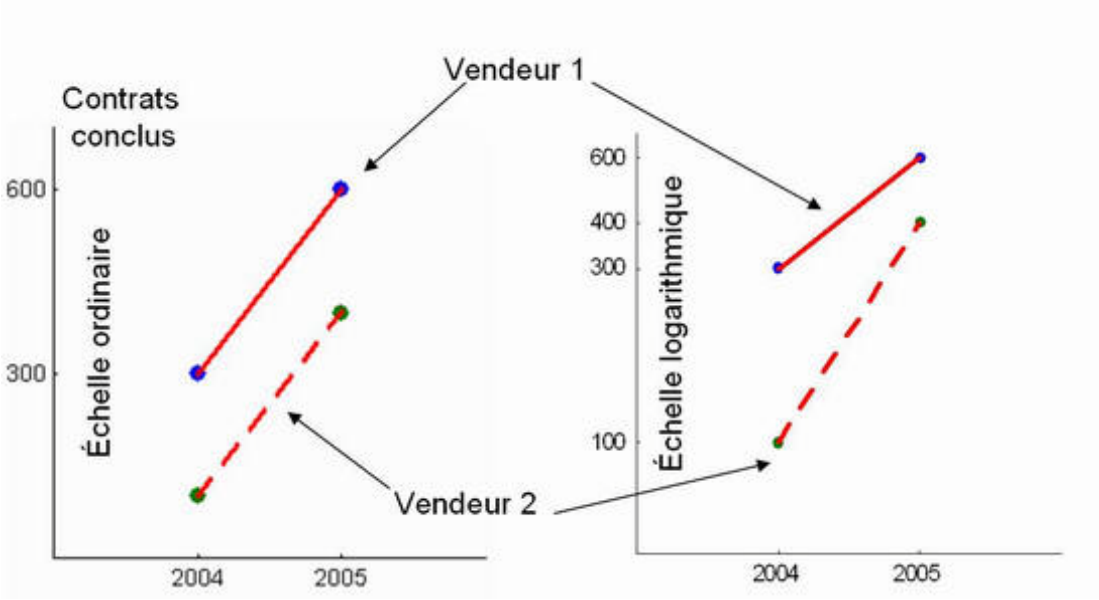
a) L'échelle logarithmique permet de mieux voir les différences de progression

On décide de comparer le nombre de contrats conclus par deux vendeurs en 2004 et 2005 :

	2004	2005	Progression
Vendeur 1	300	600	multiplié par 2
Vendeur 2	100	400	multiplié par 4

Le tableau montre que le nombre de contrats conclus par le vendeur 1 a été multiplié par 2 et que le nombre de contrats conclus par le vendeur 2 a été multiplié par 4.

Sur un graphique ordinaire (à gauche ci-dessous), les deux progressions sont parallèles. En revanche, sur un graphique avec une ordonnée logarithmique, on voit clairement que la progression du vendeur 2 est plus rapide que celle du vendeur 1.



Sur le graphique de droite, l'échelle de l'ordonnée est logarithmique, mais les chiffres indiqués (les nombres de contrats) sont les mêmes que sur le graphique de droite. Cependant, au lieu d'utiliser les valeurs elles-mêmes, le tracé utilise le logarithme décimal des valeurs, comme indiqué dans le tableau ci-dessous :

	2004	2005
Vendeur 1	log(300)=2,48	log(600)=2,78
Vendeur 2	log(100)=2	log(400)=2,6

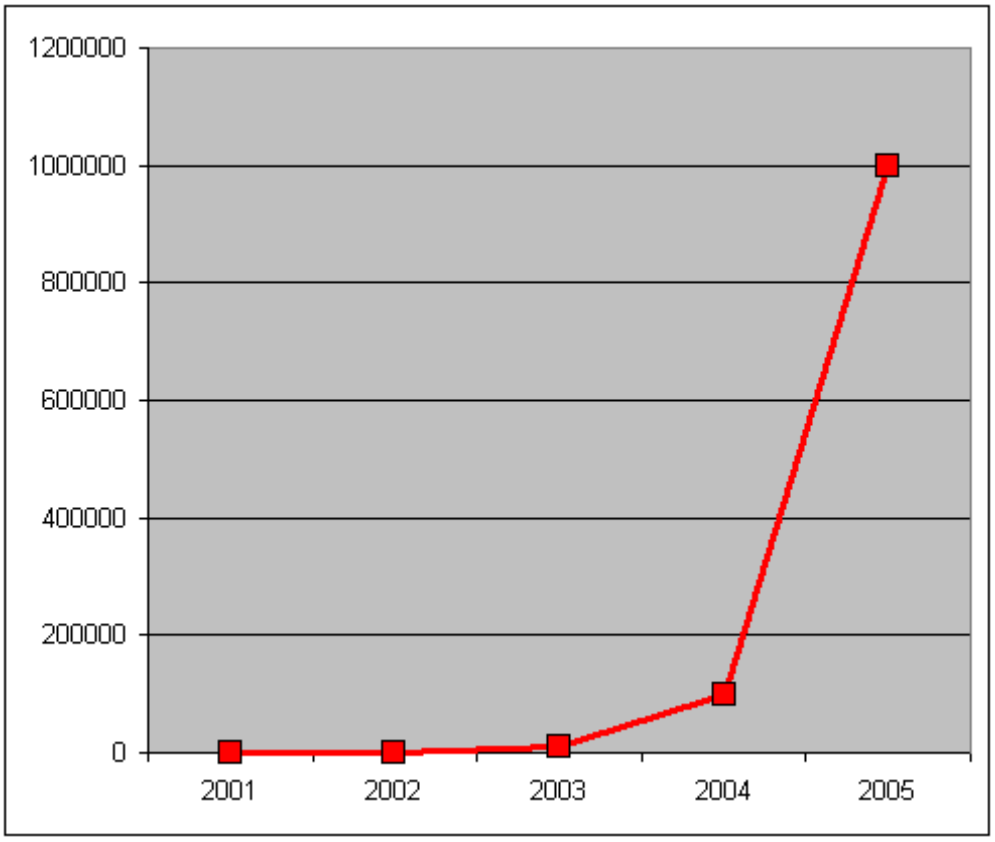
b) L'échelle log linéarise les évolutions à taux constant

On souhaite faire un graphique indiquant l'évolution du chiffre d'affaires d'une entreprise dont la croissance est très rapide :

Années	CA (en euros)
2001	100
2002	1000
2003	10000
2004	100000
2005	1000000

Comme on peut le voir sur le graphique ci-après les valeurs pour 2001, 2002 et 2003 sont écrasées par rapport à celles de 2004 et 2005 :

Graphique avec échelle des ordonnées ordinaires (sans transformation logarithmique)



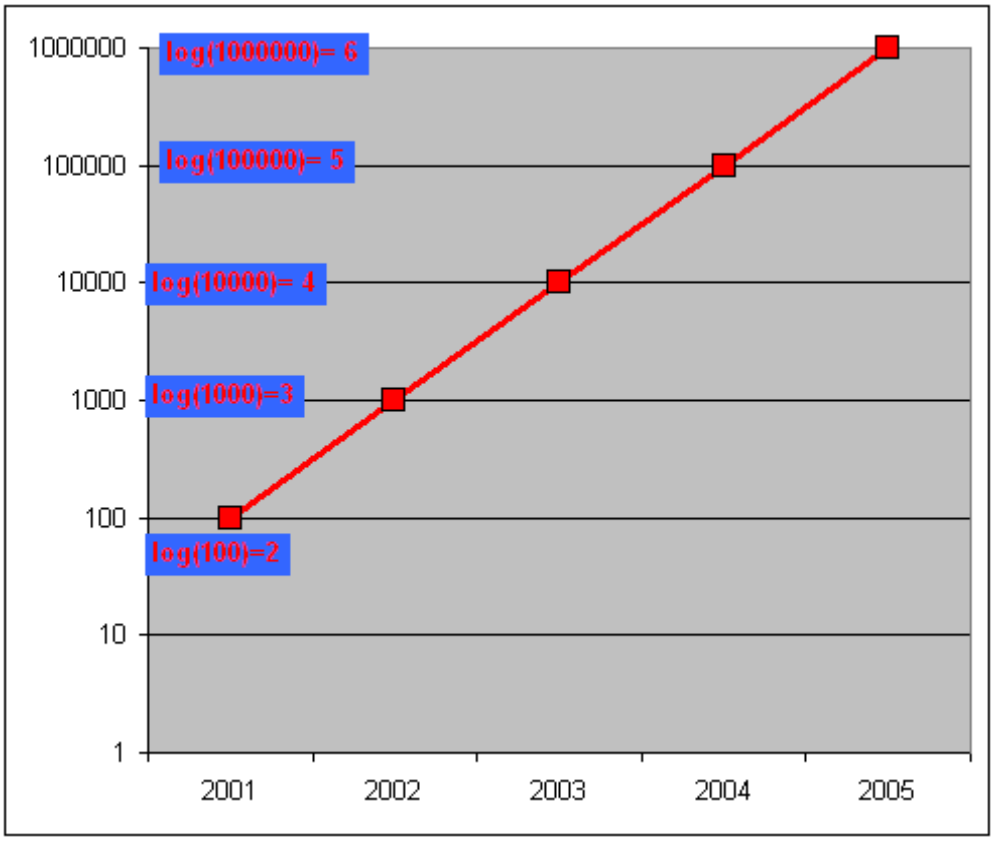
Fichier EXCEL

Appliquons une transformation logarithmique aux valeurs de l'ordonnée :

Années	**CA (en euros)**	
2001	100	log(100)=2
2002	1000	log(1000)=3
2003	10000	log(10000)=4
2004	100000	log(100000)=5
2005	1000000	log(1000000)=6

On obtient alors le graphique suivant où la progression devient linéaire :

Graphique « semi logarithmique» (l'échelle des abscisses est logarithmique)



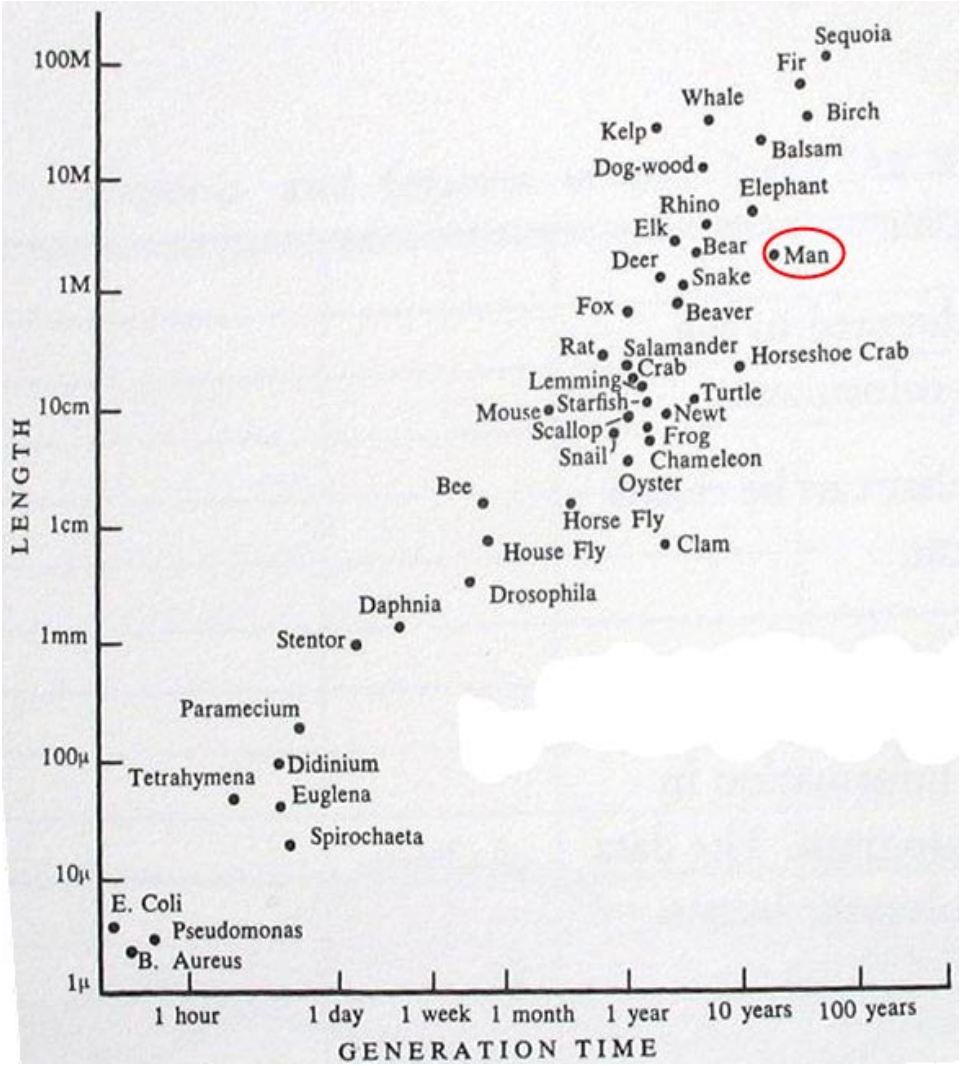
Fichier EXCEL

5 – Echelle doublement logarithmique

il existe aussi des graphiques avec échelle logarithmique sur les deux axes. Autrement dit, non seulement l'échelle des ordonnées est logarithmique, mais également l'échelle des abscisses. C'est assez peu fréquent en économie. L'exemple donné ci-après est celui de la relation entre le temps de génération (période allant de la naissance à l'âge moyen de reproduction) et la longueur de divers êtres vivants. On voit nettement sur ce graphique que le temps de génération croît avec la longueur. Mais on a ici un cas très intéressant où l'échelle de temps varie entre moins d'une heure et 100 ans et où l'échelle de longueur varie de l'infiniment petit à

100 m. Pour bien contraster ce graphique avec le graphique semi-logarithmique étudié précédemment, on parle de **graphique à échelle doublement logarithmique**.

Exemple d'un graphique à échelle doublement logarithmique



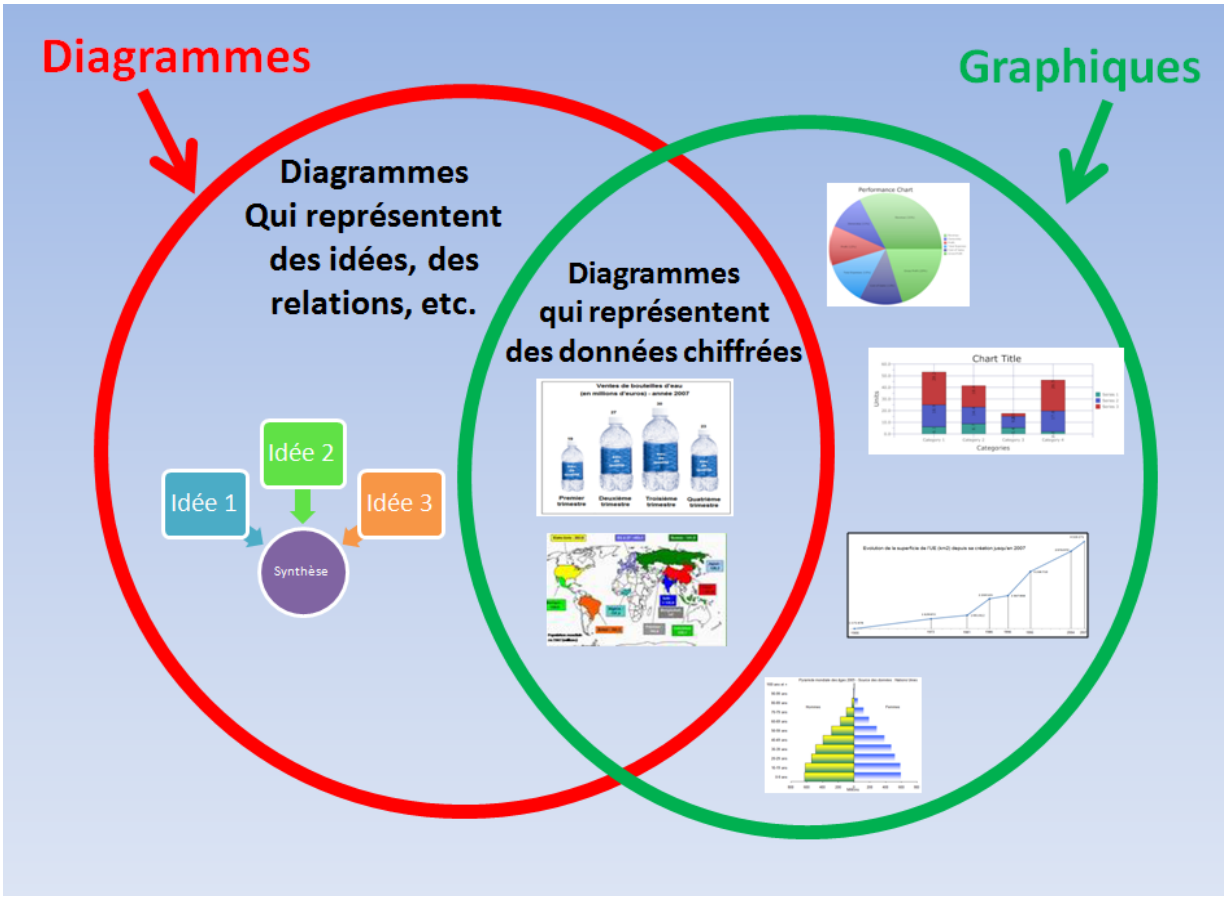
Source : John Tyler BONNER, *Size and Cycle : An Essay on the Structure of Biology* (Princeton, 1965), p.17. Reproduit dans Edward R. TUFTE, *The Visual Display of Quantitative Information* (Graphics Press LLC, 2004), p. 94

3 – Diagrammes

Un **diagramme** est une représentation visuelle simplifiée et structurée de concepts, d'idées, de constructions, des relations, de l'anatomie et aussi (et c'est en cela qu'il nous intéresse ici) de données statistiques. Il est employé dans tous les aspects des activités humaines pour visualiser et clarifier. Un diagramme permet aussi de décrire des phénomènes, de mettre en évidence des corrélations en certains facteurs ou de représenter des parties d'un ensemble.

Un **graphique** est aussi une représentation visuelle simplifiée, mais il représente principalement, voire exclusivement des chiffres.

Il existe donc une relation entre diagramme et graphique, que nous pouvons d'ailleurs illustrer par le diagramme de VENN ci-après (lequel est un diagramme non statistique !).

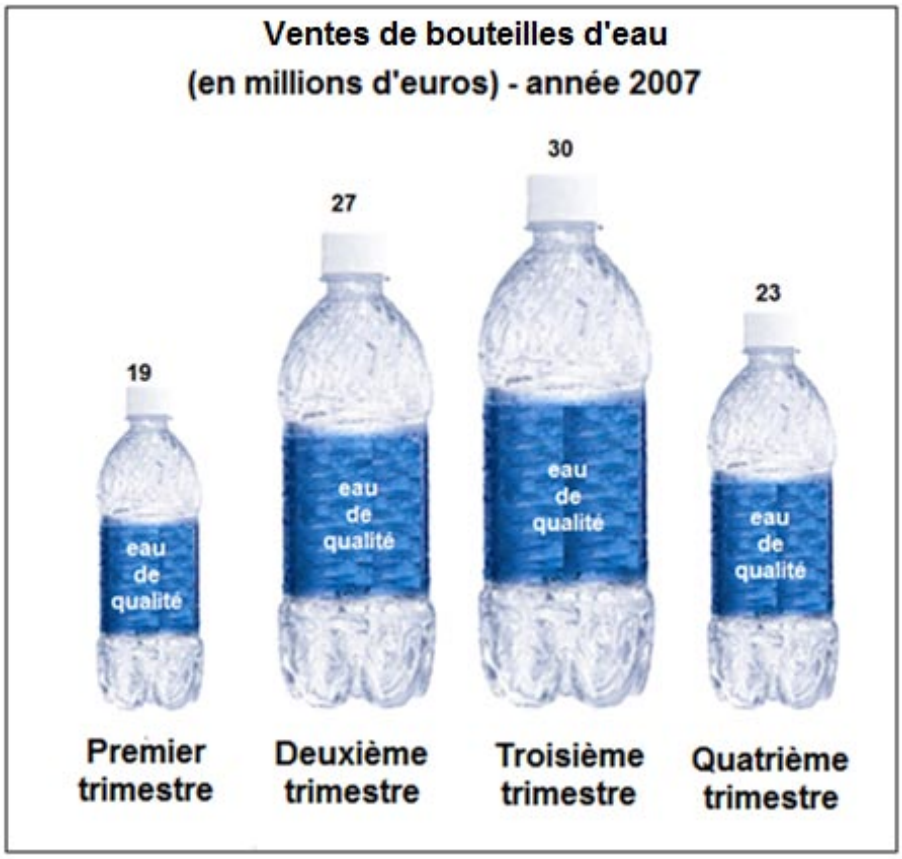


A – Pictogramme

Le pictogramme est un dessin généralement en couleur, conçu par un graphiste, qui essaie de traduire non seulement des données quantitatives, mais également des informations d'ordre commercial ou esthétique. Dans l'exemple suivant, 4 figurines qui représentent des bouteilles d'eau et la « corpulence » de chaque bouteille correspond aux ventes de la marque fictive « eau de qualité ». En réalité, il s'agit au départ d'un graphique en colonne qui a été redessiné pour remplacer les colonnes par des bouteilles. Ce qui compte ici, c'est la hauteur de la bouteille. La plus haute et la plus grosse montre que les ventes ont été particulièrement importantes au troisième semestre.

Remarquons qu'il s'agit d'un pictogramme qui traduit principalement une information statistique. Certains pictogrammes ne traduisent que des informations diverses, non statistiques.

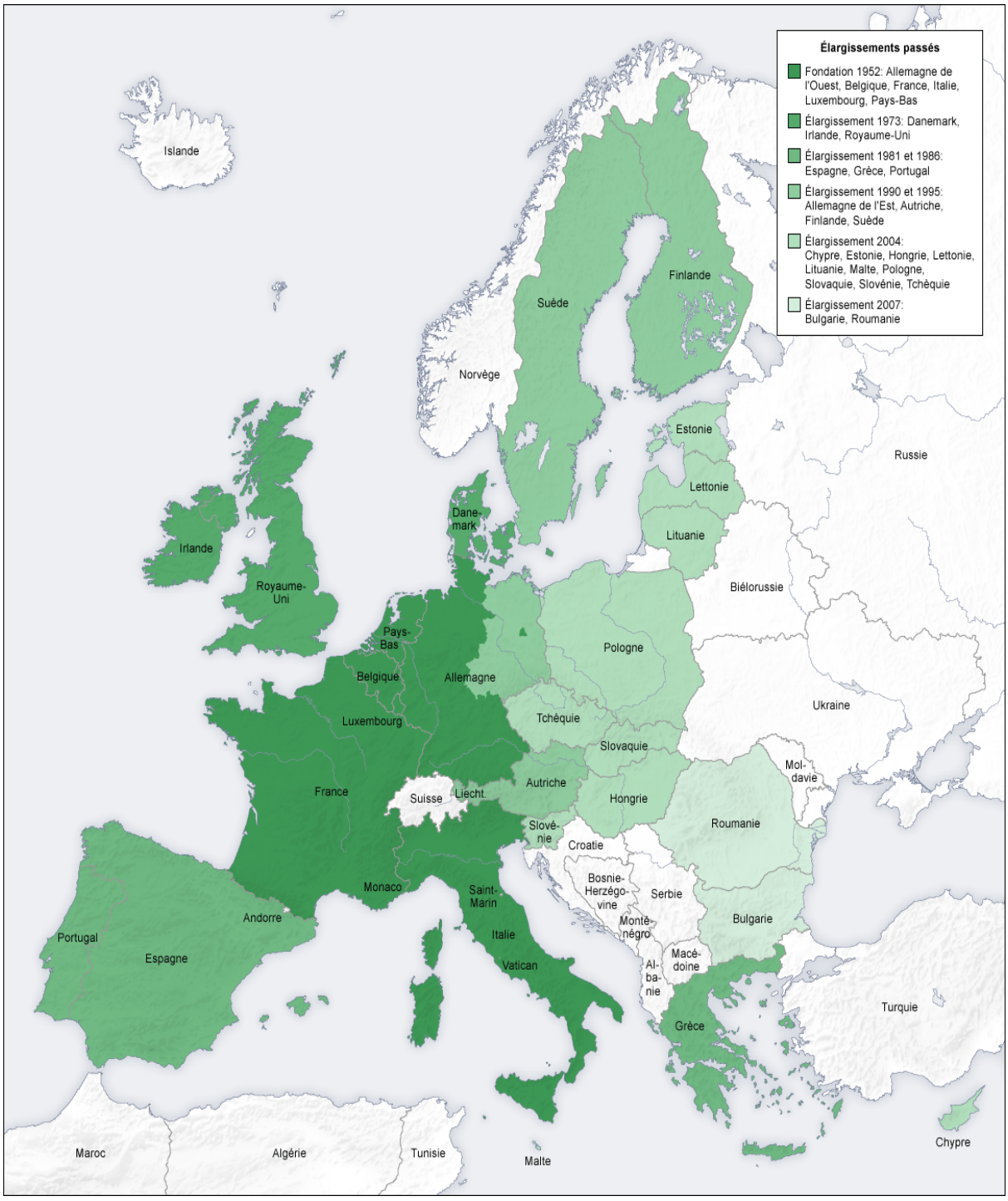
Un pictogramme qui représente des données chiffrées



B - Cartogramme

Les cartogrammes sont désormais très fréquemment utilisés pour représenter toutes sortes d'informations, de la météorologie, aux résultats électoraux, en passant naturellement par des informations économiques.

Un cartogramme statistique : De la CEE à 6 à l'UE à 27



Source : http://upload.wikimedia.org/wikipedia/commons/c/c4/European_union_past_enlargements_map_fr.png

Le cartogramme ci-après montre l'évolution de la construction européenne par date d'adhésion. On a d'abord :

En 1956 : les 6 pays fondateurs (France, Allemagne, Italie, Pays-Bas, Belgique et Luxembourg)
En 1973 : Le premier élargissement avec le Royaume-Uni, l'Irlande et le Danemark
En 1981 : La Grèce
En 1986 : L'Espagne et le Portugal
EN 1975 : L'Autriche, La Finlande et la Suède
En 2004 : 8 pays de l'Est (Estonie, Lettonie, Lituanie, République Tchèque, Slovaquie, Pologne, Slovénie, Hongrie) plus Malte et Chypre
En 2007 : La Bulgarie et la Roumanie

C - Diagramme de GANTT

Le diagramme de GANTT, du nom de l'ingénieur américain Henry Laurence GANTT (1861-1919) qui l'a popularisé, est un outil remarquable de gestion de projets. Il sert à visualiser dans le temps les différentes étapes qui composent un projet. Il existe des logiciels spécifiques qui permettent de produire ce diagramme, mais il est aussi possible d'utiliser EXCEL 2007 pour le tracer[11].

Nous allons partir d'un exemple simple sous forme d'un tableau décrivant la durée des différentes étapes de la réalisation d'un mémoire et nous transformerons ce tableau en diagramme de Gantt. Ci-après, le tableau qui va servir à faire le diagramme :

Tableau chronologique des tâches à effectuer pour réaliser un mémoire

Liste des tâches	Date de début	Durée (jours)	Date de fin
Choix du sujet	1/1	7	8/1
Documentation	8/1	10	18/1
Calculs, tableaux & graphiques	18/1	12	30/1
Plan	30/1	4	3/2
Rédaction des parties	3/2	25	28/2
Intro et conclusion	28/2	6	6/3
Bibliographie	6/3	7	13/3
TDM, sommaire, index, etc.	13/3	5	18/3
Charte graphique	18/3	5	23/3
Page de garde	23/3	3	26/3
Préparation pour l'impression	26/3	2	28/3
Impression	28/3	4	1/4

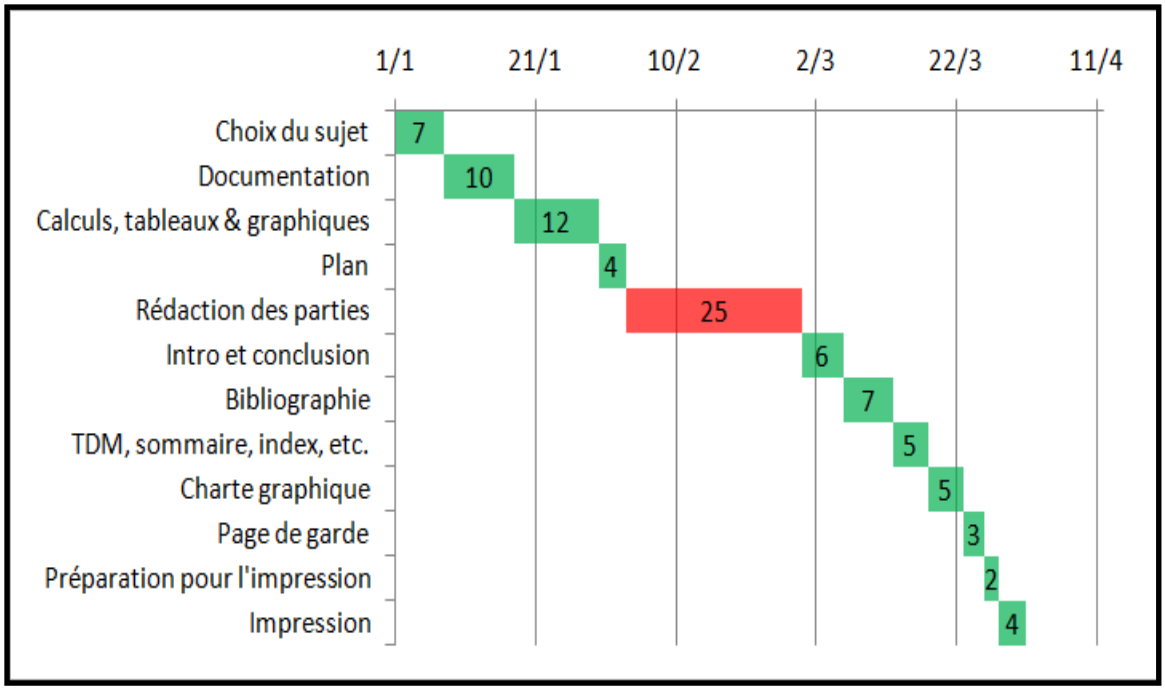
[11] Voir le tutorial vidéo http://fr.youtube.com/watch?v=_JfCMJ_s_Fk

Sur ce tableau figurent successivement :

- Dans la première colonne, l'énumération des tâches
- Dans la seconde colonne la date de début de chaque tâche
- Dans la troisième colonne la durée de chaque étape en jours
- Dans la dernière colonne, la date d'achèvement de chaque tâche

Voici maintenant le diagramme tel qu'il apparaît une fois terminé.

Diagramme de GANTT pour la réalisation d'un mémoire



Grâce à ce diagramme, nous pouvons :

- Visualiser la durée totale du projet
- Apprécier la durée de chaque étape et éventuellement réaffecter le temps total entre les différentes tâches
- Vérifier si les chevauchements sont réalistes (ici – pour simplifier- il n'y a pas de chevauchement)
- Ordonner les tâches dans un ordre chronologique

4 – Graphiques usuels

À l'ère du numérique et des télécommunications, les graphiques sont partout. Ces représentations visuelles colorées, aux formes très diverses sont plus agréables à regarder que les tableaux et permettent souvent de mieux faire passer un message au premier coup d'œil.

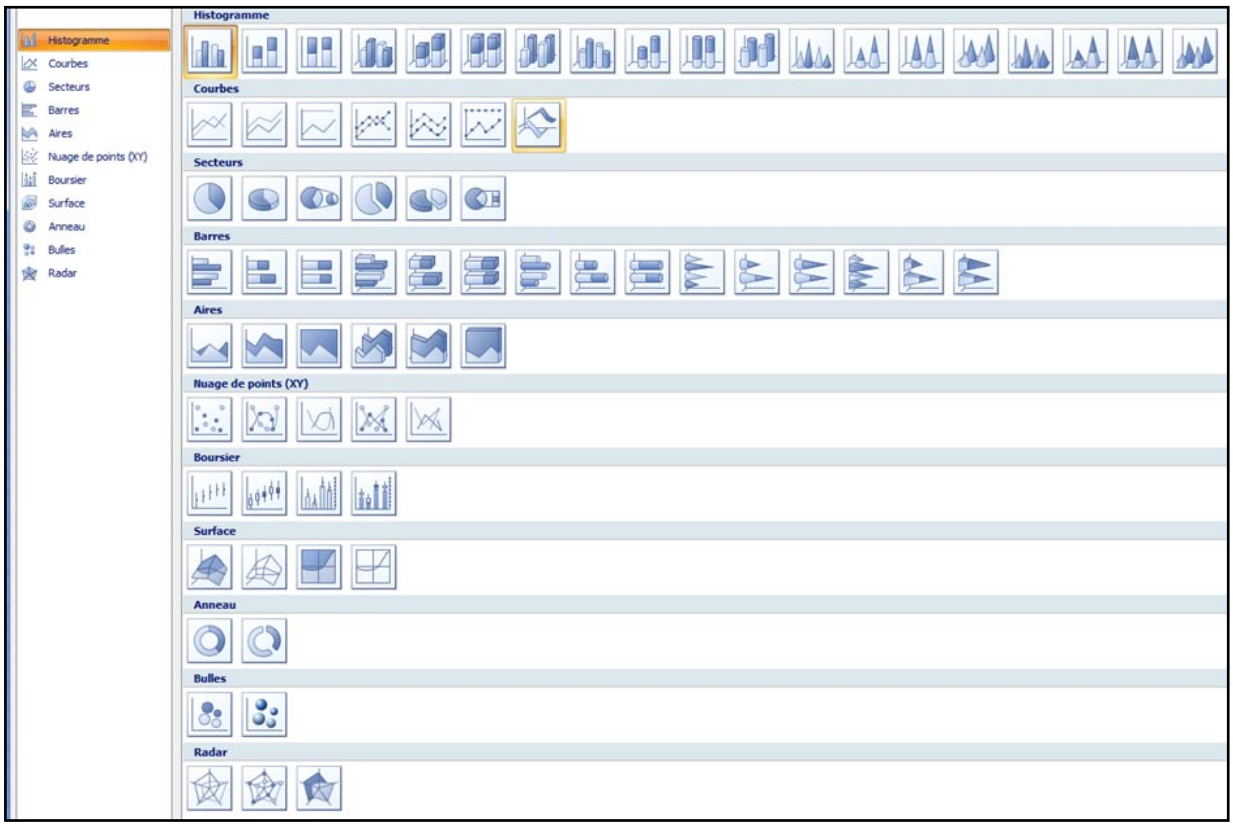
Il est également très facile, grâce aux logiciels tels que EXCEL 2007 de Microsoft de donner une « profondeur » aux différents graphiques, afin qu'ils apparaissent comme

ayant trois dimensions (Il est aussi possible dans EXCEL 2007, de représenter réellement 3 dimensions pour certains graphiques en barres ou pour les graphiques dits « de surface »).

Il ne faut cependant pas abuser de la possibilité qui nous est donnée aujourd'hui de réaliser des graphiques complexes. Car cette complexité peut finir par rendre le graphique difficilement compréhensible. Mieux vaut s'en tenir aux principales représentations graphiques connues et appréciées de tous : diagrammes en colonne, diagramme en barres, lignes, « camemberts », etc.…

Il existe une grande quantité de graphiques, tous plus imaginatifs les uns que les autres. Ainsi, à titre d'exemple, la figure ci-après montre les 73 possibilités de graphiques simples, regroupées en 11 catégories, qui peuvent être réalisés avec le logiciel EXCEL 2007. Mais en réalité, il est possible d'en faire beaucoup plus, soit en combinant ces formes de base, soit en utilisant certaines astuces.

Les 73 représentations graphiques de base dans EXCEL 2007, regroupées dans 11 catégories communes



De plus, grâce à des logiciels tels que FLASH d'Adobe, il est possible de réaliser des graphiques animés, ou des graphiques interactifs (pouvant par exemple être modifiés par l'utilisateur).

Les quatre formes graphiques les plus fréquemment utilisées sont :

- Barres verticales ou horizontales
- Courbes ou aires délimitées par des courbes
- Nuages de points
- Secteurs ou « camemberts »

A – Graphiques en barres

On distingue les graphiques en barres verticales et les graphiques en barres horizontales. Mais pour chacune de ces deux catégories, il existe 3 variétés : simple, multiples et tronçonnée. S'agissant des barres « tronçonnées », elles se subdivisent en barres d'effectifs ou en barres de pourcentages empilés à 100%. Le Tableau ci-après donne une représentation schématique de ces 8 variétés.

Pour créer les 8 graphiques les données ci-dessous ont été utilisées :

Ventes 2007 (euros)					
	Leila	Ahmed	Pierre	Elodie	Total
Marseille	13 225 478	20 154 287	17 892 555	15 897 233	67 169 553
Paris	37 895 214	35 877 421	32 558 741	22 044 687	128 376 063
Lyon	18 753 951	8 754 668	9 785 246	16 487 564	53 781 429
Total	69 874 643	64 786 376	60 236 542	54 429 484	**249 327 045**

.

Il s'agit du chiffre d'affaires hypothétique qu'une entreprise a réalisé en 2007 249 327 045 euros) répartis par ses 4 vendeurs et dans les trois villes où se trouvent ses clients. Les 8 graphiques ci-après sont juste les 8 principales façons de représenter ces données (ou une partie d'entre elles). Chaque graphique en barres fait ressortir ces chiffres d'une manière différente et pourra donc être préféré selon les circonstances dans lesquelles le graphique est utilisé.

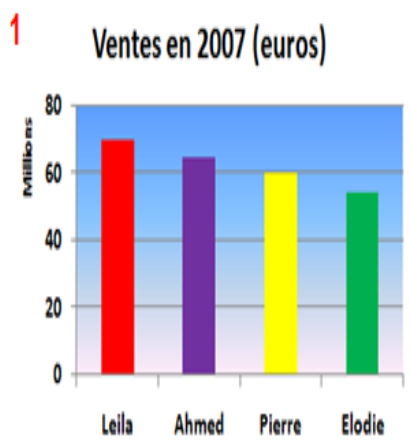
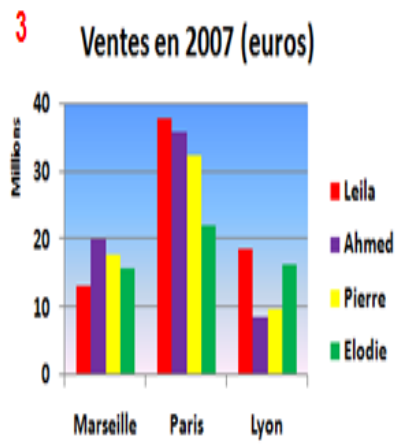
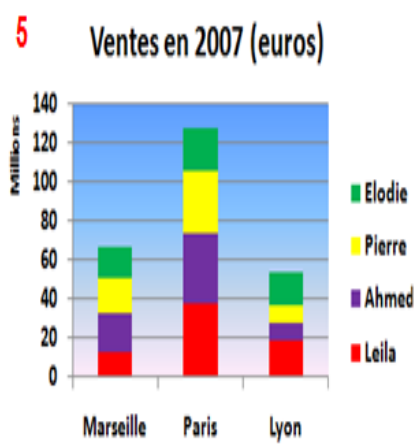
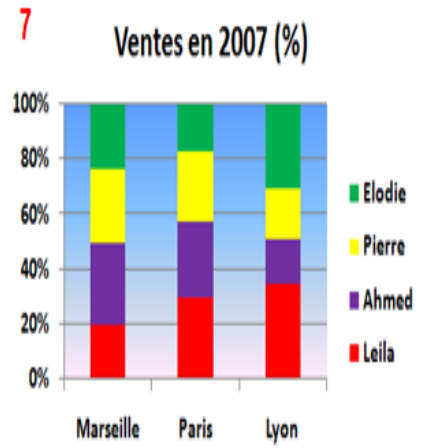
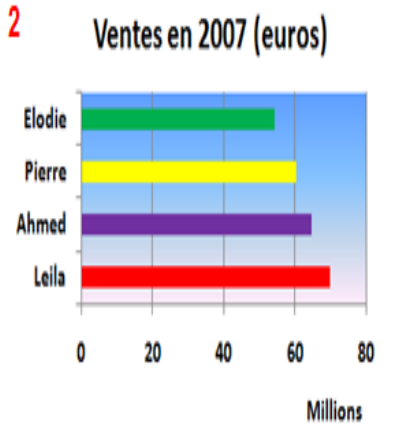
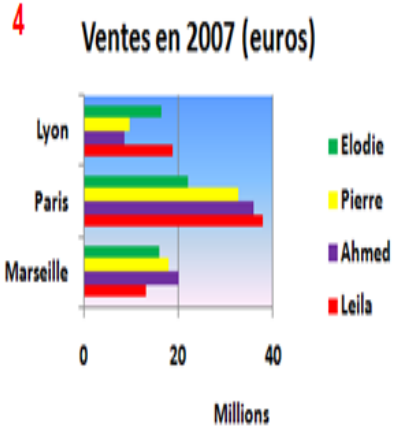
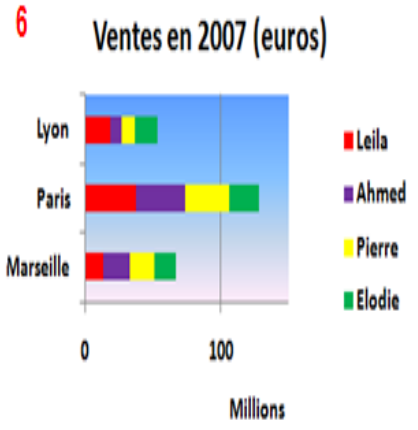
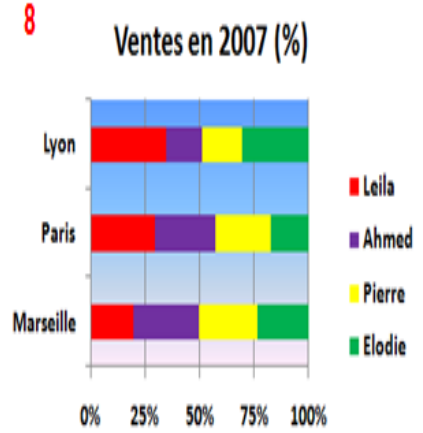
1) Barres verticales

La première ligne du tableau ci-après intitulé « les 8 principales variétés de graphiques en barres » représente les *4 principales façons de disposer les données du tableau sous forme de barres verticales* (ou colonnes). Comme l'axe horizontal représente des catégories nominales, la largeur des colonnes n'a pas d'importance, pourvu qu'elle soit identique pour toutes les colonnes. Cette largeur peut varier d'un minimum qui serait un simple trait vertical jusqu'à un maximum qui serait représenté par le fait que toutes les colonnes seraient « collées ».

a) Simple

Le type « barres verticales simples » est représenté par le graphique numéroté 1. Il permet de mettre en évidence le chiffre d'affaire réalisé par chaque vendeur. On voit du premier coup d'œil combien chaque vendeur à réalisé et quel est celui qui a réalisé le plus gros chiffre d'affaires. Les chiffres du graphique correspondent à la dernière ligne du tableau.

Les 8 principales variétés de graphiques en barres

	Simple	Multiples	Tronçonnées	
			valeurs absolues	Pourcentages
Barres verticales				
Barres horizontales				

b) Multiples

Le type « barres verticales multiples » est représenté par le graphique numéroté 3. Il permet de mettre en évidence le chiffre d'affaire réalisé par chaque vendeur dans chaque ville. Pour chacune des villes (Marseille, Paris, Lyon) on peut voir combien chaque vendeur à réalisé. Ce graphique permet de voir quel est le vendeur le plus performant dans chaque ville.

c) tronçonnées

Le type « barres verticales tronçonnée » est représenté par les graphiques numérotés 5 et 7.

- **Le graphique numéroté 5 montre les valeurs absolues :** il permet de voir d'une part quelle est la ville qui a produit le plus gros chiffre d'affaires. Mais il permet aussi de voir quelle est la contribution de chaque vendeur dans le chiffre d'affaires réalisé dans chaque ville.
- **Le graphique numéroté 7 montre les pourcentages :** il seulement de voir la contribution de chaque vendeur dans le CA de chaque ville. Et comme chaque barre verticale correspond à 100%, on peut comparer la performance de chaque vendeur dans chaque ville. On voit par exemple que la contribution de Leila (rouge) est plus forte à Lyon qu'à Marseille. Inversement, la contribution de Ahmed est plus forte à Marseille à qu'à Paris.

2) Barres horizontales

La deuxième ligne du tableau ci-avant intitulé « les 8 principales variétés de graphiques en barres » représentent *les 4 principales façons de disposer les données du tableau sous forme de barres horizontales.* Comme l'axe horizontal représente des catégories nominales, la largeur des barres n'a pas d'importance, pourvu qu'elle soit identique pour toutes les barres. Cette largeur peut varier d'un minimum qui serait un simple trait vertical jusqu'à un maximum qui serait représenté par le fait que toutes les colonnes seraient « collées ».

a) Simple

Le type « barres horizontales simples » est représenté par le graphique numéroté 2. Il est l'équivalent en barres horizontales du graphique numéroté 1

b) Multiples

Le type « barres horizontales multiples » est représenté par le graphique numéroté 4. Il est l'équivalent en barres horizontales du graphique numéroté 3.

c) tronçonnées

Le type « barres horizontales tronçonnées» est représenté par les graphiques numérotés 6 et 8. Il est l'équivalent en barres horizontales des graphiques numérotés 5 et 7.

B – Courbes et aires

La courbe est généralement le graphique le plus approprié pour montrer des évolutions dans le temps. Lorsque l'on veut montrer l'évolution de plusieurs séries dans le temps on utilisera également ce type de présentation. Parfois, on veut aussi montrer l'évolution de différentes composantes d'un ensemble dans le temps. Dans ce cas, on peut utiliser des aires délimitées par des courbes. Pour illustrer les différents usages possibles des courbes et des aires, soit le tableau suivant qui représente la répartition du chiffre d'affaires d'une entreprise sur les 3 villes où se situent ses clients.

Répartition du chiffre d'affaires de l'entreprise XXX par villes
(Millions d'euros)

Années	Marseille	Paris	Lyon	Total
2001	67,2	128	53,7	248,9
2002	50,8	140,7	40,2	231,7
2003	78,4	200,1	30,6	309,1
2004	80,7	250,2	90,1	421
2005	101,4	180,6	100,1	382,1
2006	202,8	170,1	70,2	443,1
2007	305,1	280,2	60,6	645,9

Pour compléter ce tableau, ajoutons le tableau en pourcentage suivant, qui donne, pour chaque année, la contribution en pourcentage de chaque ville au chiffre d'affaires total. Ce second tableau s'obtient en divisant les 3 premiers chiffres de chaque ligne du tableau initial par le quatrième chiffre de la ligne correspondante et en multipliant le résultat par 100

Contribution en % de chaque ville
au chiffre d'affaires total de chaque année

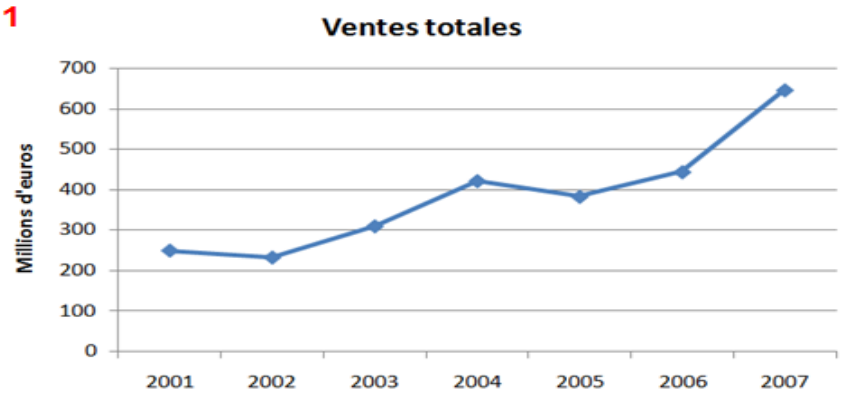
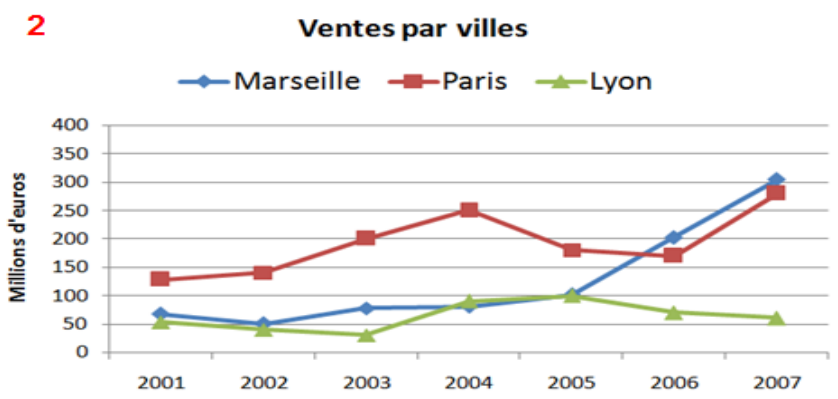
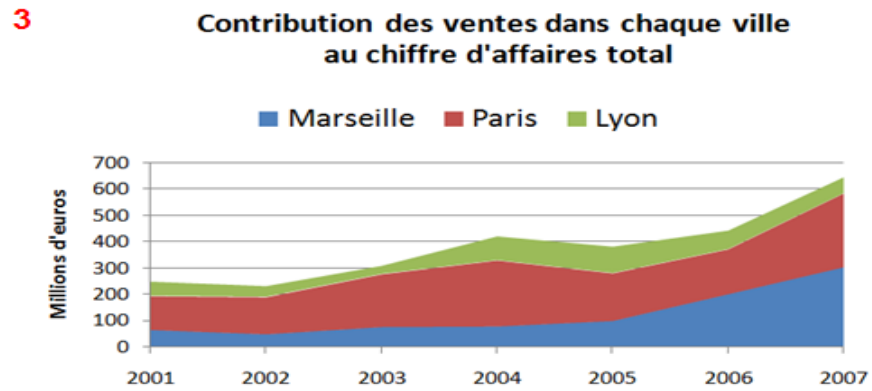
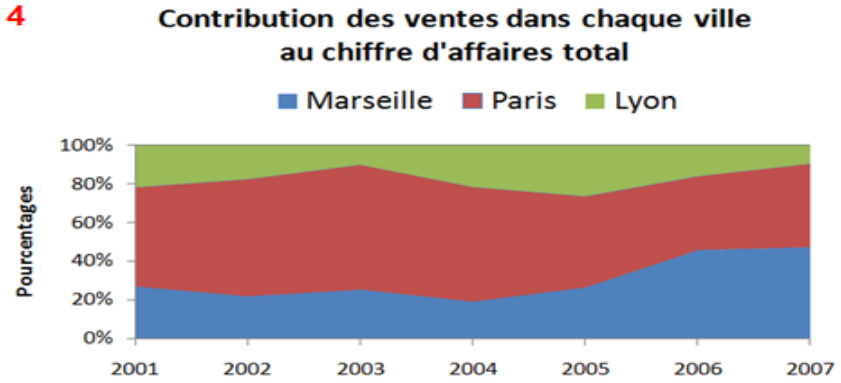
Années	Marseille	Paris	Lyon	Total
2001	27,0	51,4	21,6	100,0
2002	21,9	60,7	17,4	100,0
2003	25,4	64,7	9,9	100,0
2004	19,2	59,4	21,4	100,0
2005	26,5	47,3	26,2	100,0
2006	45,8	38,4	15,8	100,0
2007	47,2	43,4	9,4	100,0

Le tableau de graphiques ci-après illustre les 4 principales possibilités d'exploiter les graphiques en ligne où les aires délimitées par des lignes.

1) Courbes simples

Le graphique numéroté 1 représente l'évolution des ventes totales (ou chiffre d'affaires total) en millions d'euros de cette entreprise fictive. Il permet de lire à la fois l'augmentation et les périodes où l'évolution a marqué le pas. Il est parfaitement adapté pour illustrer l'évolution globale du chiffre d'affaires de l'entreprise.

Quatre principales façons d'utiliser les graphiques en courbes et aires

Courbe simple	Courbes multiples
1 	2 
Aires délimitées par des courbes (a) valeurs absolues	**Aires délimitées par des courbes (b) pourcentages**
3 	4 

2) Coubes multiples

Le graphique numéroté 2 représente l'évolution des ventes totales (ou chiffre d'affaires total) en millions d'euros que cette entreprise réalise pour chacune des 3 villes où se trouvent ses clients. Il est intéressant, mais en fait, on peut lui préférer un graphique qui allierait à la fois les avantages de la courbe simple (graphique numéroté 1) et la décomposition par ville (graphique numéroté 2). C'est ce que propose le graphique numéroté 3.

3) Aires délimitées par des courbes

On peut concevoir deux façons de présenter un graphique d'aires délimitées par des courbes :

- **Soit sous forme de valeurs absolues :** le graphique numéroté 3 représente ainsi à la fois l'évolution des ventes totales et la contribution de chaque ville à cette évolution. La contribution est représentée par le découpage en trois de la surface qui se trouve sous la courbe. La contribution de Marseille est en bleu, celle de Paris en rouge et celle de Lyon en vert. Les 3 contributions additionnées donnent l'évolution totale.

- **Soit sous forme de pourcentages :** le graphique numéroté 4 représente la contribution en pourcentage de chaque ville au chiffre total de l'entreprise. Ce graphique permet de voir par exemple que Paris était prépondérant en 2001 mais que sa part (ainsi que celle de Lyon) a été progressivement grignotée par Marseille.

La méthode de construction est simple : on part du tableau initial et l'on additionne d'abord la colonne de Marseille et de Paris, puis les colonnes Marseille, Parie et Lyon. Ensuite, on trace les 3 courbes sur un même graphique et on effectue un coloriage des 3 zones (voir schéma ci-après). Naturellement, si l'on dispose d'un tableur comme EXCEL 2007, le graphique est tracé automatiquement uniquement en sélectionnant les données du tableau initial en en choisissant le graphique adéquat.

Méthode construction d'un graphique
sous forme d'aires délimitées par des courbes

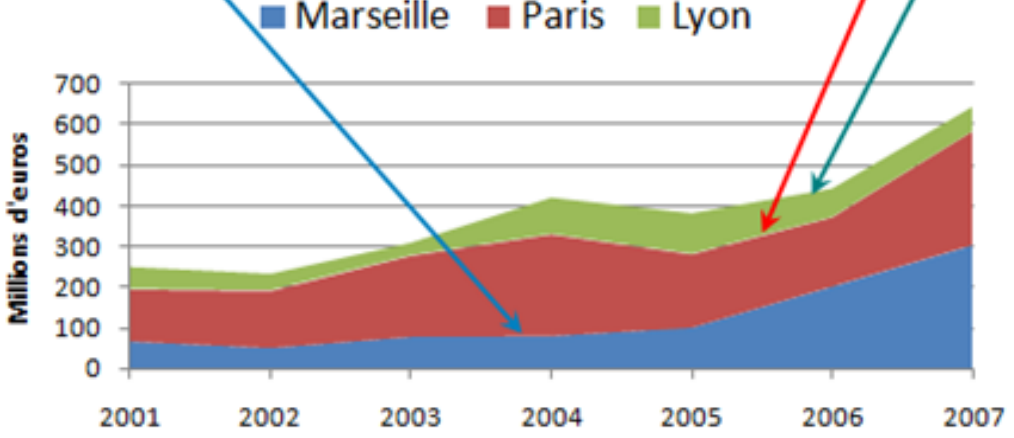
Tableau initial

Années	Marseille	Paris	Lyon	Total
2001	67,2	128	53,7	248,9
2002	50,8	140,7	40,2	231,7
2003	78,4	200,1	30,6	309,1
2004	80,7	250,2	90,1	421
2005	101,4	180,6	100,1	382,1
2006	202,8	170,1	70,2	443,1
2007	305,1	280,2	60,6	645,9

Tableau final pour tracer le graphique

Années	Marseille	Marseille et Paris	Marseille, Paris et Lyon
2001	67,2	195,2	248,9
2002	50,8	191,5	231,7
2003	78,4	278,5	309,1
2004	80,7	330,9	421
2005	101,4	282	382,1
2006	202,8	372,9	443,1
2007	305,1	585,3	645,9

Graphique

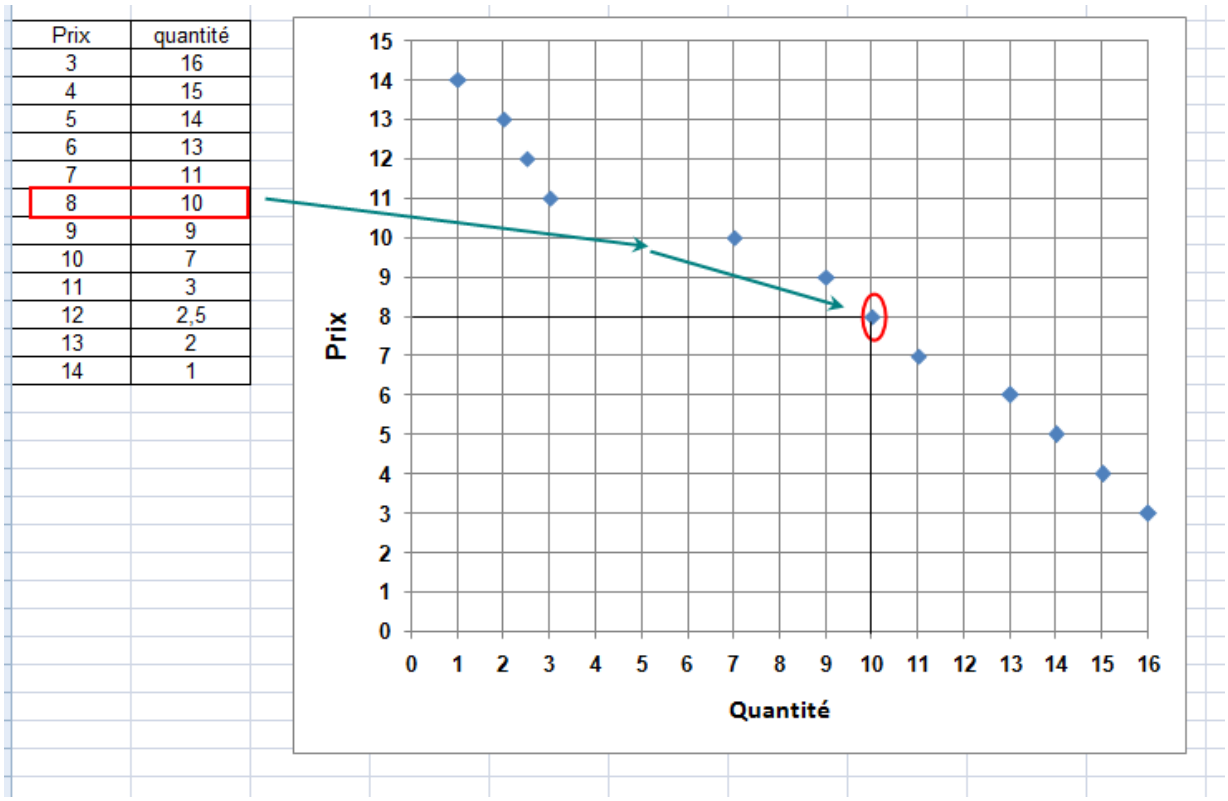


C - Graphique de dispersion ou nuage de points

Un graphique de dispersion ou nuage de points est un graphique qui met en relation les valeurs de deux variables sur un repère de coordonnées cartésiennes. Nous avons déjà rencontré ce type de graphique dans l'introduction à ce chapitre. En effet, les 4 graphiques du quartet d'ANSCOMBE sont des graphiques de dispersion.

Prenons un autre exemple. Ci-après, un tableau hypothétique qui illustre les différentes quantités d'un certain produit qu'un consommateur XXX est disposé à acheter suivant selon le prix du produit.

Exemple de graphique de dispersion ou « nuage de points »



Prix	quantité
3	16
4	15
5	14
6	13
7	11
8	10
9	9
10	7
11	3
12	2,5
13	2
14	1

Le graphique de dispersion correspondant permet d'associer à chaque une coordonnée d'abscisse (la quantité) et une coordonnée d'ordonnée (le prix).

Les graphiques de dispersion ou en nuage de points sont très utilisés pour l'étude des corrélations entre deux variables. (Voir le chapitre 6).

D - Secteurs

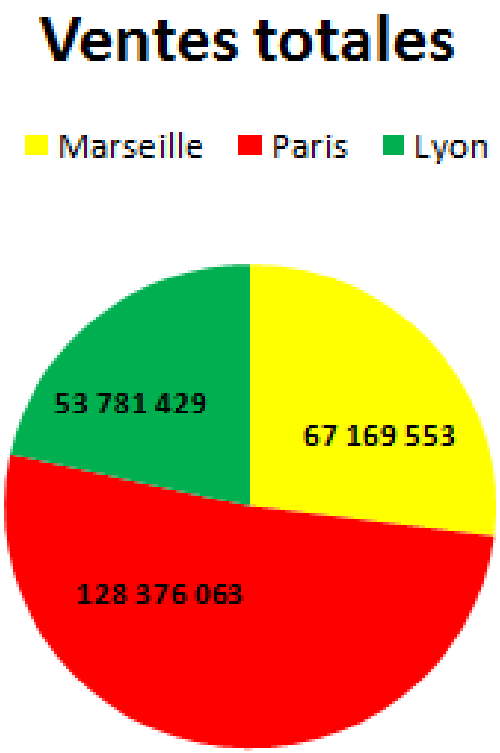
Les graphiques en secteurs sont utiles lorsque l'on veut représenter la relation entre une partie et un tout. On distingue les secteurs à 380° et ceux à 180°. Voyons un exemple de chacun d'eux avant de voir la méthode de construction qui repose sur la conversion des pourcentages en degrés.

Reprenons les données du chiffre d'affaires hypothétique qu'une entreprise a réalisé en 2007 249 327 045 euros) répartis par ses 4 vendeurs et dans les trois villes où se trouvent ses clients.

Ventes 2007 (euros)					
	Leila	Ahmed	Pierre	Elodie	Total
Marseille	13 225 478	20 154 287	17 892 555	15 897 233	67 169 553
Paris	37 895 214	35 877 421	32 558 741	22 044 687	128 376 063
Lyon	18 753 951	8 754 668	9 785 246	16 487 564	53 781 429
Total	69 874 643	64 786 376	60 236 542	54 429 484	**249 327 045**

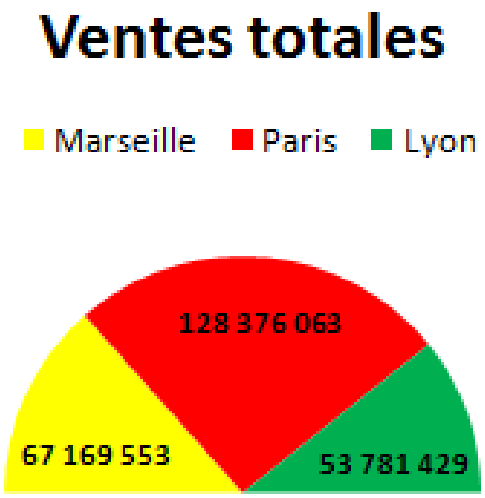
1) Secteurs à 360 degrés

Le secteur à 360° ci-dessous représente la répartition des ventes totales entre les trois villes (Marseille en jaune, Paris en rouge et Lyon en vert).



2) Secteurs à 180 degrés

Le secteur à 180° ci-dessous représente la répartition des ventes totales entre les trois villes (Marseille en jaune, Paris en rouge et Lyon en vert).



3) Méthode de construction

a) Secteur à 360 degrés

Pour obtenir la part du chiffre d'affaire réalisé dans chaque ville, on divise le chiffre d'affaires réalisé dans la ville par le chiffre d'affaires total et on multiplie par 360.

Par exemple, pour Marseille on a :

$$\frac{67169553}{249327045} \times 360 = 97$$

Le tableau ci-dessous donne les résultats en degrés pour les trois villes :

	Total	fréquences	degrés
Marseille	67169553	0,27	97,0
Paris	128376063	0,51	185,3
Lyon	53781429	0,22	77,7
Total	249327045	1,00	360,0

Une fois que l'on a calculé les degrés associés au chiffre d'affaire dans chaque ville, il faut tracer le secteur au moyen d'un compas, puis, avec un rapporteur, le diviser en 3 sous-secteurs ayant pour angle 97° (Marseille), 185,3° (Paris) et 77,7°(Lyon).

b) Secteur à 180 degrés

Pour obtenir la part du chiffre d'affaire réalisé dans chaque ville, on divise le chiffre d'affaires réalisé dans la ville par le chiffre d'affaires total et on multiplie par 180.

Par exemple, pour Marseille on a :

$$\frac{67169553}{249327045} \times 180 = 48,5$$

Le tableau ci-dessous donne les résultats en degrés pour les trois villes :

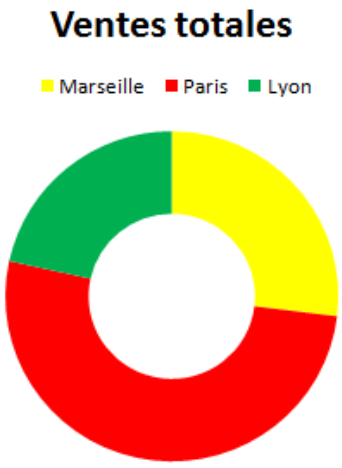
	Total	fréquences	degrés
Marseille	67 169 553	0,27	48,5
Paris	128 376 063	0,51	92,7
Lyon	53 781 429	0,22	38,8
Total	249 327 045	1,00	180,0

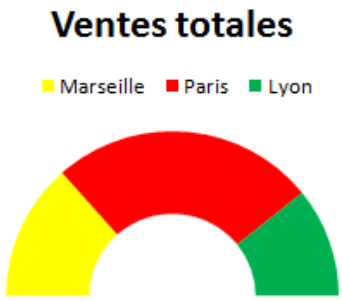
Un fois que l'on a calculé les degrés associés au chiffre d'affaires dans chaque ville, il faut tracer le secteur au moyen d'un compas, puis, avec un rapporteur, le diviser en 3 sous-secteurs ayant pour angle 48,5° (Marseille), 92,7° (Paris) et 38,8 degré (Lyon).

4) Anneaux

a) Simples

Les anneaux sont simplement des secteurs « troués » au milieu. Ainsi, par exemple, les deux secteurs précédents peuvent être représentés sous forme d'anneaux comme ci-après. La seule différence étant le cercle vide du milieu dont la taille peu être choisie de façon arbitraire en fonction de considération telles que l'esthétique ou l'économie d'encre (si le graphique est destiné à être imprimé) par exemple.

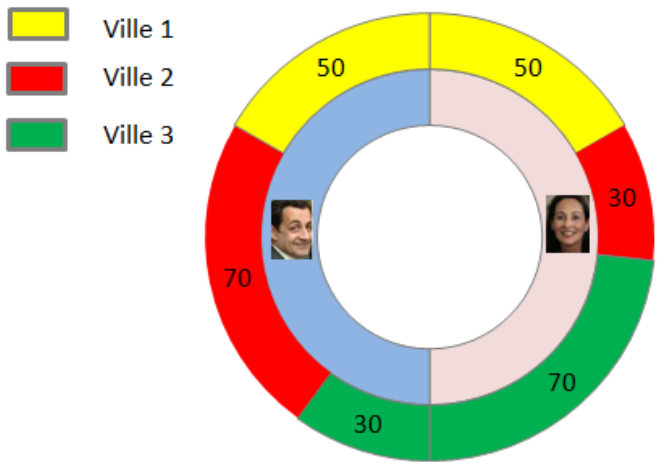




b) Concentriques

Un exemple typique d'**anneaux concentriques** est celui de la représentation des résultats du second tour de l'élection présidentielle 2007 pour 3 villes hypothétiques. Dans le cercle intérieur, on a la répartition des voix entre les deux candidats (dans cet exemple hypothétique, ils ont obtenu chacun 150 voix) et dans le cercle extérieur, on a la répartition des voix de chaque candidat dans chacune des villes.

	Voix
Ségo	**150**
Ville 1	50
Ville 2	30
Ville 3	70
Sarko	**150**
Ville 3	30
Ville 2	70
Ville 1	50
Total	**300**



5 – Autres graphiques

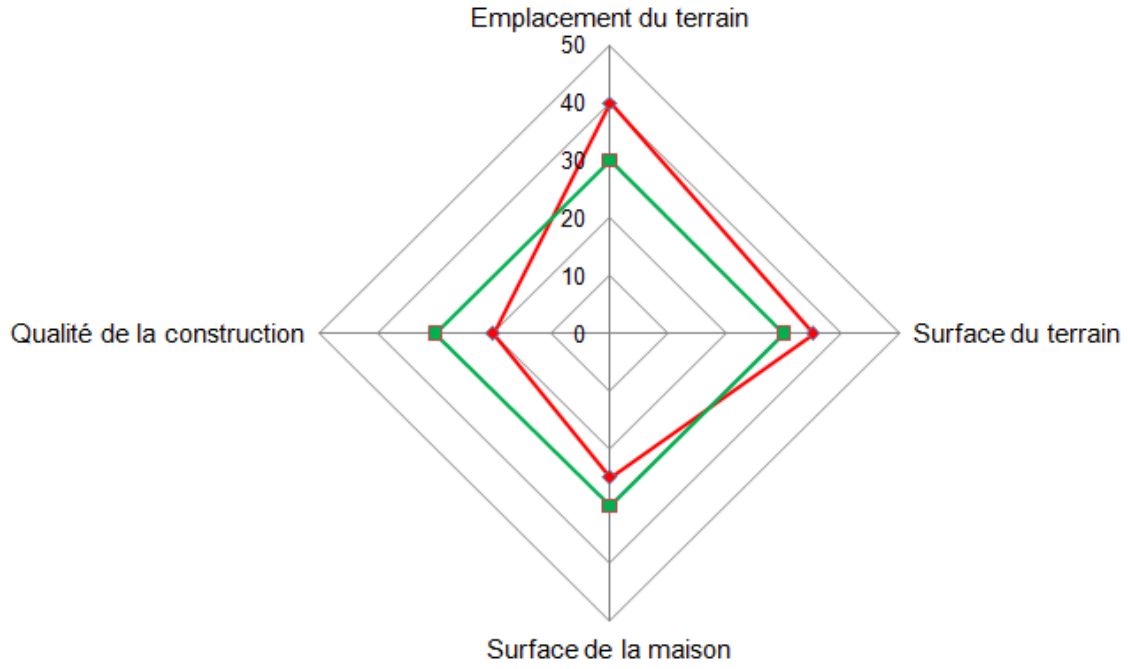
A – Graphiques en radar et toile d'araignée

1 – Radar

Soit par exemple 120 personnes qui sont interrogées dans la ville A et 120 personnes dans la ville B sur la caractéristique qu'elles placent en premier lors de l'achat d'une maison. Il s'agit d'une caractéristique avec 4 modalités. On a le tableau suivant :

	Ville A	Ville B
Modalités	Nombre de personnes	Nombre de personnes
Emplacement du terrain	40	30
Surface du terrain	35	30
Surface de la maison	25	30
Qualité de la construction	20	30
	120	**120**

On peut alors placer ces données sur un diagramme « en radar » de la façon suivante :



2 – Toile d'araignée

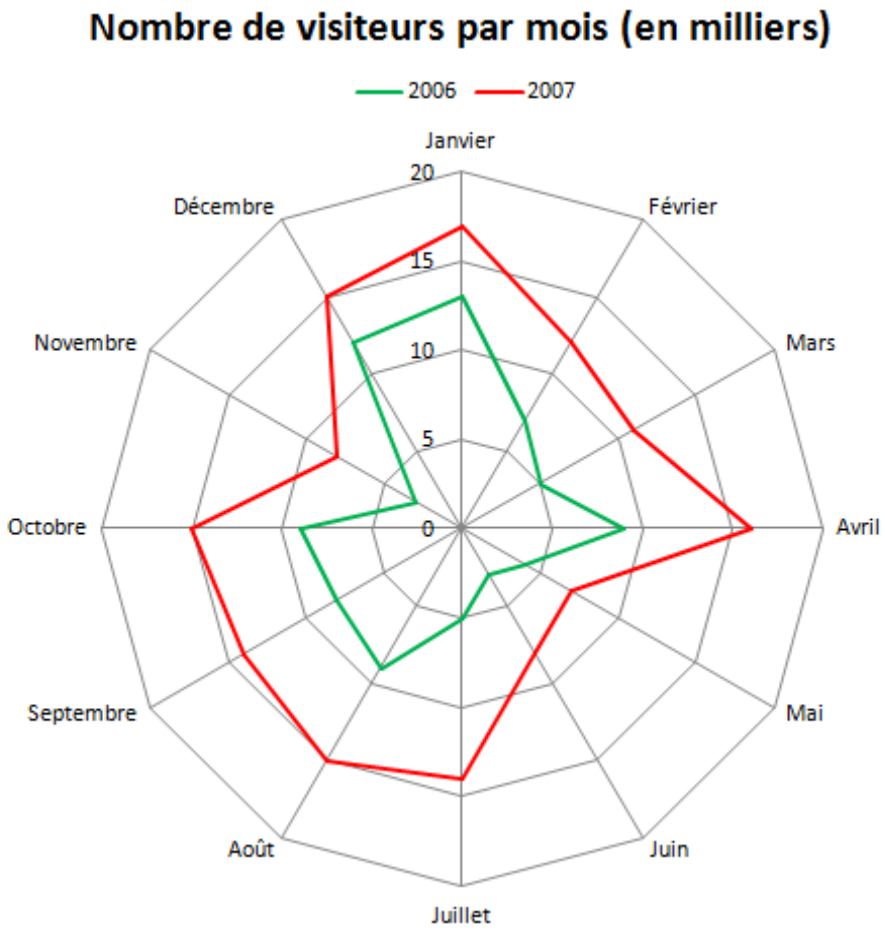
Le graphique en toile d'araignée est une variante du graphique en radar, mais avec un nombre d'axes plus grand. On l'utilise par exemple pour représenter des données chronologiques. Soit par exemple une entreprise qui souhaite comparer le nombre de visiteurs mensuels sur son site internet en 2006 et en 2007.

Nombre de visiteurs mensuels sur le site internet de l'entreprise XXX (milliers)

	2006	2007
Janvier	13	17
Février	7	12
Mars	5	11
Avril	9	16
Mai	4	7
Juin	3	8
Juillet	5	14
Août	9	15
Septembre	8	14
Octobre	9	15
Novembre	3	8
Décembre	12	15

Une bonne façon de présenter une comparaison visuelle est de faire un graphique en toile d'araignée» suivant :

Exemple de graphique en toile d'araignée



Ce graphique donne immédiatement deux informations :

- Les visites en 2007 ont chaque mois été supérieures aux visites en 2006
- Il y a un caractère cyclique dans les visites, car les mois « creux » et les mois « pleins » sont les mêmes en 2006 et en 2007.

B – Graphiques à bulles

Semblable au graphique de dispersion ou nuages de points, le **graphique (ou diagramme) en bulles** permet d'ajouter une troisième dimension à l'analyse. Les deux premières dimensions déterminent la position des bulles sur le diagramme tandis que la troisième fixe la surface de chacune des bulles.

Exemple 1 (à faire avec un logiciel) –

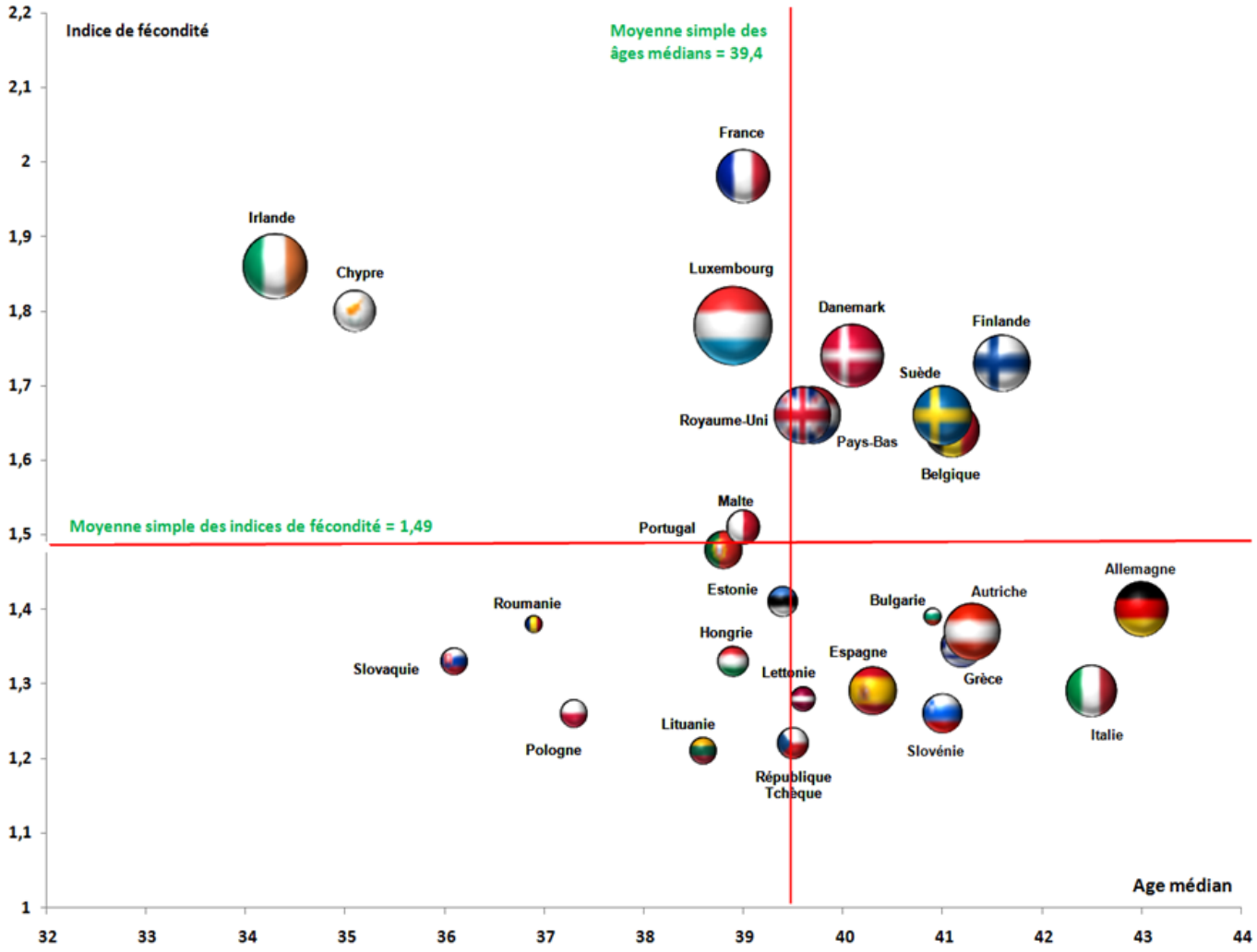
À titre d'exemple, soit les chiffres ci-dessous, extrait du tableau 1, qui donnent l'âge médian, l'indice de fécondité et le PIB par habitant des pays de l'UE à 27.

Age médian, Indice de fécondité et PIB/habitant

Pays	Age médian (estimation de juillet 2007)	Indice de fécondité (en nombre d'enfants par femme, estimation 2007)	PIB par habitant (en dollars) (Année 2006)
Allemagne	43	1,4	34881
Belgique	41,1	1,64	35610
France	39	1,98	34038
Italie	42,5	1,29	30705
Luxembourg	38,9	1,78	72785
Pays-Bas	39,7	1,66	37189
Danemark	40,1	1,74	47334
Irlande	34,3	1,86	50169
Royaume-Uni	39,6	1,66	38707
Grèce	41,2	1,35	20958
Espagne	40,3	1,29	26833
Portugal	38,8	1,48	16670
Autriche	41,3	1,37	37887
Finlande	41,6	1,73	37829
Suède	41	1,66	41313
Chypre	35,1	1,8	20872
Estonie	39,4	1,41	10488
Hongrie	38,9	1,33	11341
Lettonie	39,6	1,28	7254
Lituanie	38,6	1,21	8422
Malte	39	1,51	13743
Pologne	37,3	1,26	8745
République tchèque	39,5	1,22	11636
Slovaquie	36,1	1,33	8773
Slovénie	41	1,26	18751
Bulgarie	40,9	1,39	3793
Roumanie	36,9	1,38	3591

Dans le graphique à bulles ci-après, 3 dimensions sont représentées : l'indice de fécondité (axe vertical), l'âge médian (axe horizontal) et le PIB par habitant (surface de chaque « bulle »).

Exemple de graphique à bulles : Age médian (axe horizontal) , indice de fécondité (axe vertical)
et PIB par habitant (surface de la bulle) des pays de l'UE à 27. Années 2007 (âge médian et indice de fécondité) et 2006 (PIB/habitant en $)



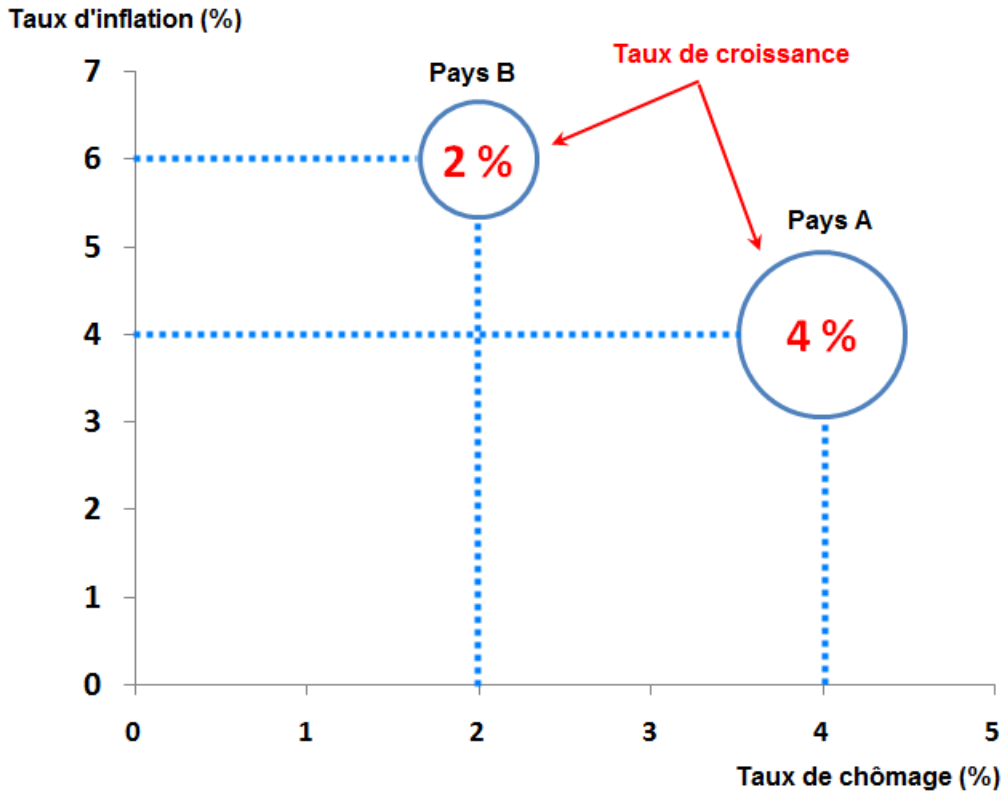
Exemple 2 : Le tableau ci-dessous donne les performances macroéconomiques de 2 pays hypothétiques A et B, en 2007.

	Chômage (%)	Inflation (%)	Croissance (%)
Pays A	4	4	4
Pays B	2	6	2

Représenter ces chiffres sur un graphique « en bulles » avec :

- Le taux de chômage en abscisse
- Le taux d'inflation en ordonnée
- La croissance par un cercle de surface proportionnelle au taux du pays considéré

On obtient alors le graphique suivant :



Dans cet exemple relativement simple, il suffit de faire en sorte que l'aire du disque qui représente le taux de croissance du pays A soit le double de l'aire qui représente le taux de croissance du pays B.

Pour assurer la proportionnalité des aires, il faut passer par la formule de l'aire du disque :

$$S = \pi R^2$$

On peut choisir pour référence la plus grosse valeur à représenter sous forme de disque.

Ensuite on lui attribue une aire arbitraire, par exemple 2 cm^2 et on en déduit le rayon :

$$S = \pi R^2 \Leftrightarrow 2 = 3,14 \times R^2 \Leftrightarrow R^2 = \frac{2}{3,14} \Leftrightarrow R = \frac{\sqrt{2}}{\sqrt{3,14}} = \frac{1,414}{1,772} = 0,798$$

Pour tracer le cercle, on arrondira le rayon à 8 mm.

On calcule ensuite l'aire de l'autre disque du graphique. Dans notre exemple, si l'aire du disque du pays A représente 4%, le disque du pays B, qui représente 2% doit avoir une aire 2 fois plus petite. Par conséquent, l'aire du disque représentatif de la croissance du pays B sera égale à la moitié de la surface du disque qui représente la croissance de A, soit 1 cm^2. On en déduit ensuite le rayon du disque de B par la formule :

$$S = \pi R^2 \Leftrightarrow 1 = 3,14 \times R^2 \Leftrightarrow R^2 = \frac{1}{3,14} \Leftrightarrow R = \frac{\sqrt{1}}{\sqrt{3,14}} = \frac{1}{1,772} = 0,56$$

Pour tracer le cercle, on arrondira le rayon à 5,5 mm.

En définitive, le disque de A aura un rayon de 8 mm (environ) et le disque de B aura un rayon de 5,5 mm (environ). Ceci n'est pas facile à tracer de manière précise ! C'est la raison pour laquelle les graphiques à bulles sont généralement réalisés avec un logiciel.

C – Graphiques boursiers

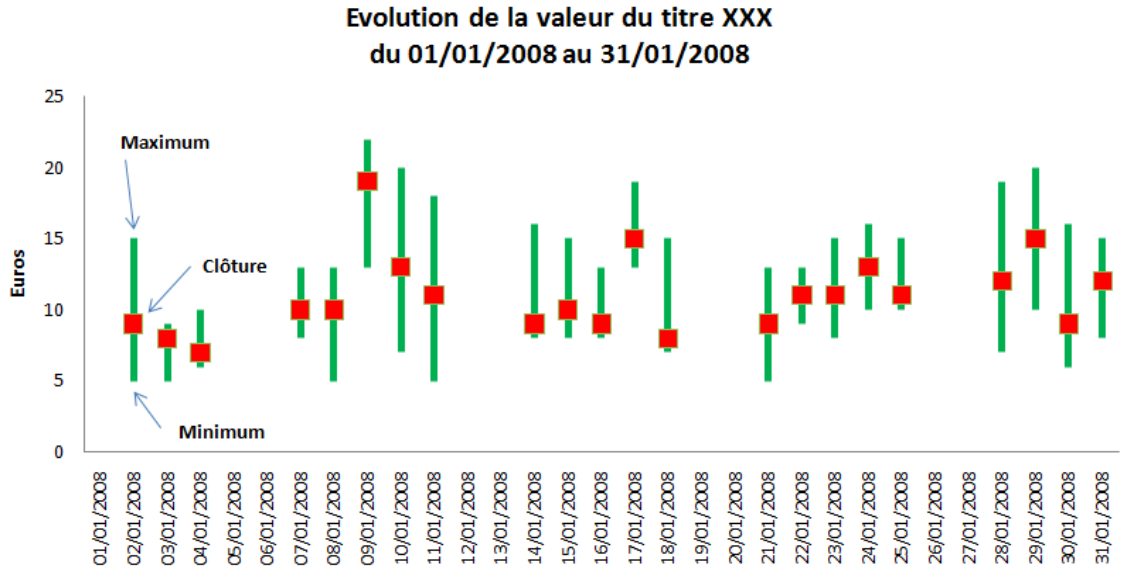
Les graphiques boursiers sont appelés ainsi car ils servent principalement à donner des indications sur l'évolution des cours boursiers. Dans l'exemple ci-après, nous allons voir la version la plus simple du graphique boursier, mais des versions plus complexes sont possibles et facilement réalisables dans EXCEL 2007 un fois que l'on a compris le principe de base.

Soit le tableau ci-dessous qui donne l'évolution du cours journalier d'un titre boursier (en euros) de 2 janvier au 31 janvier 2008, en en retenant que les jours ouvrables. On a relevé 3 informations chaque jour : le cours le plus bas, le cours le plus haut et le cours de clôture.

Evolution du cours du titre XXX

Date	Cours le plus bas	Cours le plus haut	Clôture
01/01/2008			
02/01/2008	5	15	9
03/01/2008	5	9	8
04/01/2008	6	10	7
05/01/2008			
06/01/2008			
07/01/2008	13	8	10
08/01/2008	5	13	10
09/01/2008	13	22	19
10/01/2008	7	20	13
11/01/2008	5	18	11
12/01/2008			
13/01/2008			
14/01/2008	8	16	9
15/01/2008	8	15	10
16/01/2008	13	8	9
17/01/2008	13	19	15
18/01/2008	15	7	8
19/01/2008			
20/01/2008			
21/01/2008	5	13	9
22/01/2008	9	13	11
23/01/2008	8	15	11
24/01/2008	10	16	13
25/01/2008	10	15	11
26/01/2008			
27/01/2008			
28/01/2008	7	19	12
29/01/2008	10	20	15
30/01/2008	6	16	9
31/01/2008	15	8	12

Le graphique ci-après, dit graphique boursier, permet de visualiser les 3 informations



D - Graphique de TUKEY

Le graphique « Boite à moustaches » ou diagramme en boîte (box plot) a été inventé en 1977 par le statisticien américain John TUKEY (1915-2000).

1) Les éléments constitutifs du graphique

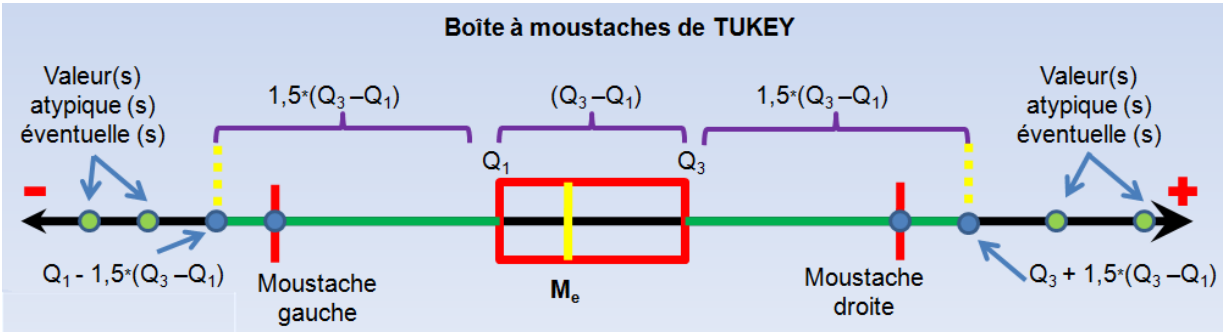
Sur le graphique ci-après, les éléments suivants apparaissent[12] :

L'intervalle interquartile : il est défini par $Q_3 - Q_1$ et donne les "dimensions" de la boîte.

Les moustaches : Ce sont les extrémités gauche et droite du graphique (parfois appelées "pattes").

> **La moustache gauche** est définie par la **valeur de la série** immédiatement supérieure à $Q_1 - 1{,}5\ (Q_3 - Q_1)$. S'il y a des valeurs inférieures à $Q_1 - 1{,}5\ (Q_3 - Q_1)$ dans la série, elles sont atypiques et représentées par des marqueurs.
>
> **La moustache droite** est définie par la **valeur de la série** immédiatement inférieure à $Q_3 + 1{,}5\ (Q_3 - Q_1)$. S'il y a des valeurs supérieures à $Q_3 + 1{,}5\ (Q_3 - Q_1)$, elles sont représentées par des marqueurs.

La médiane : la valeur de la variable qui partage la population en deux populations égales. On la désigne par l'abréviation Me.



2) Exemple

Soit la série des 20 éléments : {4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}

La médiane est égale à 3. De plus, $Q_1 = 2$ et $Q_3 = 4$. Enfin, la valeur minimale est 0 et la valeur maximale 5. Pour être complet, on peut ajouter la moyenne de la série qui est égale à 2,7.

Valeurs atypiques : Pour savoir s'il y a des valeurs atypiques il faut calculer $Q_1 - 1{,}5\ (Q_3 - Q_1) = 2 - 1{,}5 \times (4 - 2) = 2 - 1{,}5 \times 2 = 2 - 3 = -1 < 0$ et $Q_3 + 1{,}5\ (Q_3 - Q_1) = 4 + 1{,}5 \times (4 - 2) = 4 + 3 = 7 > 5$. Conclusion : puisque $-1 > 0$ et que $7 > 5$, n'y a pas de

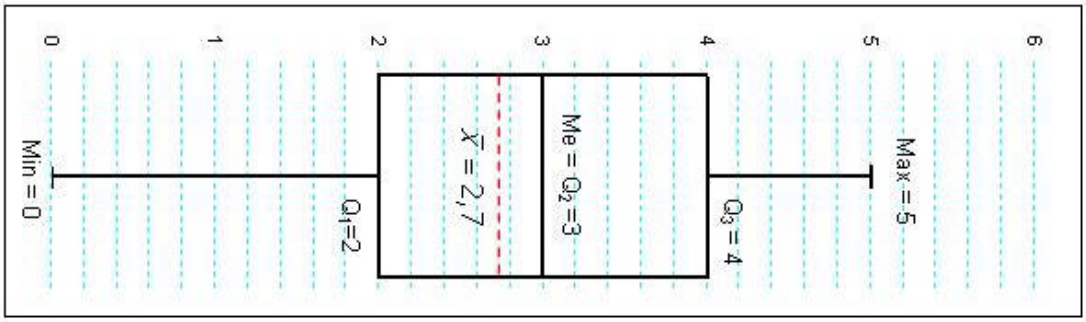
[12] Voir : http://www.sfds.asso.fr/groupes/statvotre/Boite-a-moustaches.pdf

valeurs atypiques[13]. Par conséquent, puisque la valeur immédiatement supérieure à $Q_1 - 1{,}5\ (Q_3 - Q_1)$ est 0 et que la valeur immédiatement inférieure à $Q_3 + 1{,}5\ (Q_3 - Q_1)$ est 5, les deux moustaches ont respectivement pour extrémité gauche le minimum de la série (soit 0) et pour extrémité droite le maximum de la série (soit 5).

On a donc le graphique boite à moustaches ci-après. La boite est délimitée par Q_1 et Q_3. À l'intérieur de la boite un premier trait noir indique la médiane (et un second trait rouge [en pointillé et facultatif] indique la moyenne). Les valeurs minimale et maximale constituent ici les moustaches, car elles sont comprises dans l'intervalle défini par $Q_1 - 1{,}5\ (Q_3 - Q_1)$ et $Q_3 + 1{,}5\ (Q_3 - Q_1)$.

Boite à moustaches de TUKEY pour la série

{4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5}



Voir le fichier EXCEL

[13] Selon Monique LE GUENN, « *La valeur 1.5 est selon TUKEY une valeur pragmatique (rule of thumb), qui a une raison probabiliste. Si une variable suit une distribution normale, alors la zone délimitée par la boîte et les moustaches devrait contenir 99,3 % des observations. On ne devrait donc trouver que 0.7% d'observations atypiques (outliers). Si le coefficient vaut 1, la probabilité serait de 0.957, et elle vaudrait 0.999 si le coefficient est égal à 2. Pour TUKEY la valeur 1.5 est donc un compromis pour retenir comme atypiques assez d'observations mais pas trop d'observations* ». http://www.sfds.asso.fr/groupes/statvotre/Boite-a-moustaches.pdf

E – Graphiques panachés

Il existe une infinité de façons de panacher les différents graphiques. Voyons quelques exemples.

1) Secteur complété par une barre tronçonnée

Soient les données déjà utilisées du chiffre d'affaires par ville et par vendeur d'une entreprise XXX en 2007.

Ventes 2007 (euros)

	Leila	Ahmed	Pierre	Elodie	Total
Marseille	13 225 478	20 154 287	17 892 555	15 897 233	67 169 553
Paris	37 895 214	35 877 421	32 558 741	22 044 687	128 376 063
Lyon	18 753 951	8 754 668	9 785 246	16 487 564	53 781 429
Total	69 874 643	64 786 376	60 236 542	54 429 484	**249 327 045**

Ventes 2007 (%)

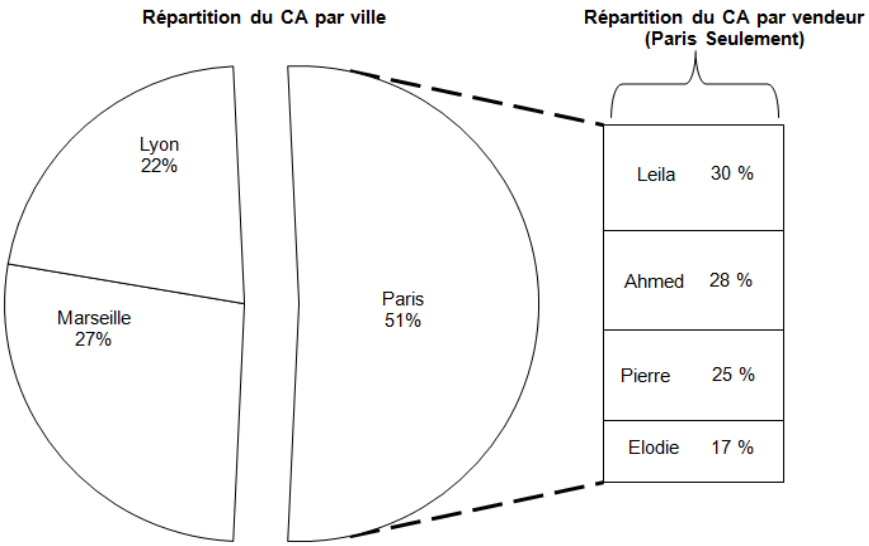
	Leila	Ahmed	Pierre	Elodie	Total
Marseille					27
Paris					51
Lyon					22
Total					100

Ventes 2007 (%)

	Leila	Ahmed	Pierre	Elodie	Total
Marseille					
Paris	30	28	25	17	100
Lyon					
Total					

On souhaite :

- utiliser un secteur à 360° pour faire apparaitre la répartition du chiffre d'affaires entre les trois villes.
- Décomposer le chiffre d'affaires réalisé sur Paris entre les quatre vendeurs et le faire apparaître au moyen d'une barre verticale tronçonnée à 100%.



2) Graphique de PARETO

Le graphique de PARETO est un cas particulier du graphique à échelle verticale double. Avant de voir comment il se construit, il convient de rappeler la « loi de PARETO » dont il sert à vérifier la validité. Le graphique a en fait été inventé par Joseph Moser DURAN (1904 -), fondateur de la démarche qualité. Le graphique de PARETO est en effet l'un des sept graphiques de base utilisé dans les contrôles de qualité. S'il est appelé graphique de PARETO et non graphique de DURAN (ou graphique de Kaoru ISHIKAWA (1915-1989), autre fondateur du management de la qualité), c'est en raison de la « loi de PARETO » que nous allons donc exposer pour commencer.

a) De la loi de PARETO au graphique de PARETO

Vilfredo PARETO, économiste et sociologue de la fin du 19ème siècle et du début du 20ème siècle (1848-1923) a notamment étudié la répartition des revenus dans différents pays. Une des principaux constats qui ressort de cette étude est une "loi", dite "**Loi de PARETO**": **dans presque tous les pays, 80% de la richesse sociale était possédée par 20% des individus**. Par la suite, on s'aperçut qu'un grand nombre de phénomènes étaient gouvernés par la loi du 80/20 :

20 % des clients représentent 80 % du chiffre d'affaires
20% des pièces stockées dans une entreprise représentent 80% de la valeur du stock
20% des fournisseurs représentent 80% du volume d'achat total
20% du personnel effectue 80% du travail
20% des salariés d'une entreprise empoche 80% de la masse salariale
20% des automobilistes causent 80% des accidents
20% des vêtements d'une garde-robe sont portés 80% du temps
20% des produits représentent 80% du chiffre d'affaires
20% des ventes représentent environ 80% de la marge bénéficiaire
20% du personnel total est concerné par 80% des accidents du travail
20% des causes peuvent être à l'origine de 80% des défauts
20% des clients sont à l'origine de 80% des réclamations
20% de la population paie 80% des impôts
20% des délinquants génèrent 80% des délits
20% des joueurs de foot marquent 80% des buts
20% des problèmes représentent 80% des préoccupations
20% des pays émettent 80% des gaz à effet de serre, etc.
20% des acteurs jouent dans 80% des films.

Source : http://www.ed-productions.com/leszed/index.php?80-20-pareto

Ces exemples illustrent la loi de PARETO. Quelques causes majeures, une fois isolées, permettent de résoudre la plus grande partie d'un problème de qualité. Une fois ces causes majeures identifiées, on peut concentrer les efforts et les ressources à les éliminer. Le **graphique de PARETO** est une façon de visualiser la loi des 80/20. Mais, comme on va le voir dans l'exemple étudié plus loin, cette loi n'est pas systématique, il arrive très souvent aussi que 20% des causes expliquent beaucoup moins que 80% des résultats.

b) Définition, construction, exemple et interprétation

Un **diagramme de PARETO** est un graphique qui combine un **graphique en barre** et une **courbe cumulative**. Il sert principalement à l'étude des données qualitatives, mais rien n'empêche de l'utiliser pour des données quantitatives.

Procédure pour construire le graphique :

i) Si les données sont sous forme d'une série, les regrouper par **modalités** (données) ou par **valeurs** (données quantitatives), de façon à obtenir une **distribution** par modalités ou par valeurs (éventuellement, une distribution par **classes de modalités** ou par **classes de valeurs**).

ii) Classer les **valeurs** ou les **modalités** par ordre décroissant des effectifs

iii) Ajouter une colonne pour la distribution en pourcentages

iv) Ajouter ensuite une colonne de pourcentages cumulés

v) Faire un graphique pour représenter simultanément :

- La distribution des pourcentages par un **graphique en barre**, en mettant l'axe des y à gauche

- Les pourcentages cumulés par une **courbe des pourcentages cumulés**, en mettant l'axe des y à droite

Exemple : Supposons que l'on veuille étudier les raisons de la résiliation d'un abonnement en ligne. Le problème ici est de comprendre pourquoi les abonnés résilient leur abonnement (afin de réduire le nombre de résiliation). On recherche donc les causes. Pour ce faire, lorsque les clients résilient leur abonnement, on leur propose un questionnaire (volontairement simplifié dans cet exemple) où il sont invité à cocher la case qui correspond à la raison de la résiliation de leur abonnement. Supposons que les 5 choix suivants leurs soient proposés (Remarque : nous sommes en présence de données qualitatives non hiérarchisables, les choix sont donc des **modalités nominales** : le contenu du site ne correspondait pas à mes attentes (réponse codifiée par "A"), le contenu n'est pas bon (réponse codifiée par "B"), le contenu n'est pas renouvelé assez souvent (réponse codifiée par "C"), difficultés techniques pour accéder au contenu (réponse codifiée par "D"), Autres (réponse codifiée par "E").

On a obtenu les résultats suivants en étudiant 15 cas de résiliations : {**E, B**,**D, E, D, E, A**, **B**, **B**,**C**, **D**, **A**, **B**,**B, E**}. Evidemment, en réalité, on étudierait un nombre de cas beaucoup plus grand, mais le principe de construction resterait identique.

Construction du graphique :

i) Si les données sont sous forme d'une série, les regrouper par **modalités** (données qualitatives) ou par **valeurs** (données quantitatives), de façon à obtenir une **distribution** :

Modalités	Nombre de réponses
Le contenu ne correspond pas à mes attentes (A)	2
Le contenu n'est pas bon (B)	5
Le contenu n'est pas renouvelé assez souvent (C)	1
Difficultés techniques pour accéder au contenu (D)	3
Autres (E)	4
Total	15

ii) Classer les **valeurs** ou les **modalités** par ordre décroissant des effectifs :

Modalités	Nombre de réponses
Le contenu n'est pas bon	5
Autres	4
Difficultés techniques pour accéder au contenu	3
Le contenu ne correspond pas à mes attentes	2
Le contenu n'est pas renouvelé assez souvent	1
Total	15

iii) Ajouter une colonne pour la distribution en pourcentages

Modalités	Nombre de réponses	Pourcentages
Le contenu n'est pas bon	5	33,3
Autres	4	26,7
Difficultés techniques pour accéder au contenu	3	20,0
Le contenu ne correspond pas à mes attentes	2	13,3
Le contenu n'est pas renouvelé assez souvent	1	6,7
Total	15	100,0

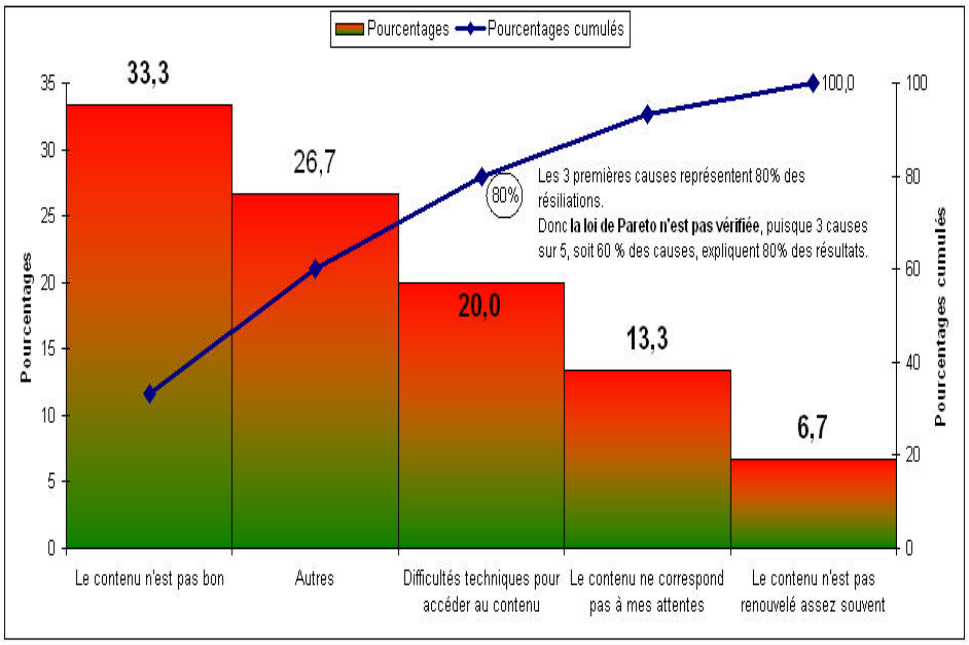
iv) Ajouter ensuite une colonne de pourcentages cumulés

Modalités	Nombre de réponses	Pourcentages	Pourcentages cumulés
Le contenu n'est pas bon	5	33,3	33,3
Autres	4	26,7	60,0
Difficultés techniques pour accéder au contenu	3	20,0	80,0
Le contenu ne correspond pas à mes attentes	2	13,3	93,3
Le contenu n'est pas renouvelé assez souvent	1	6,7	100,0
Total	15	100,0	

v) Faire un graphique pour représenter simultanément :

- La distribution des pourcentages par un **graphique en barre**, en mettant l'axe des y à gauche

- Les pourcentages cumulés par une **courbe des pourcentages cumulés**, en mettant l'axe des y à droite



Fichier EXCEL

c) Interprétation

Dans notre exemple, on voit que la loi de PARETO n'est pas vérifiée. En effet, la loi de PARETO veut que 20% des causes expliquent 80 % des résultats. Or ici, il y a 5 causes. Donc une cause représente à elle seule 20% des résultats. Pour que la loi de PARETO soit vérifiée, il faudrait qu'une seule cause (20% des causes) explique 80% des résultats (80% des résiliations). Or, ici, la première cause n'explique que 35% des résiliations et il faut 3 causes, soit 60% des causes, pour parvenir à expliquer 80% des résultats.

On peut se reporter aux sites internet suivants qui donnent des exemples intéressants :

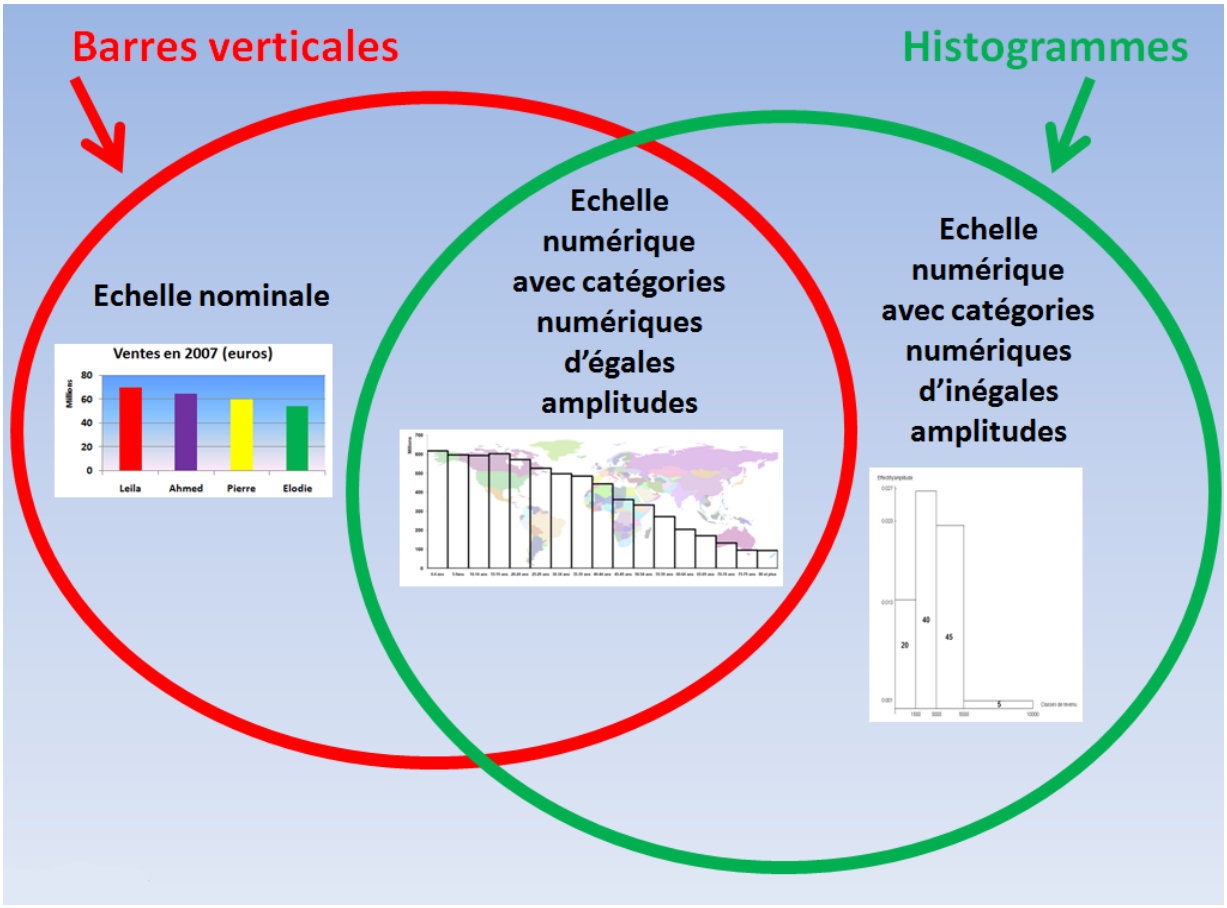
http://membres.lycos.fr/jflipes/pareto.htm (sécurité routière)

http://www.azaquar.com/outils/pareto.html (amélioration de la qualité dans une usine qui fabrique des conserves)

F – Histogramme

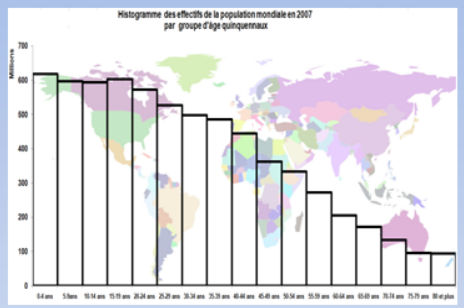
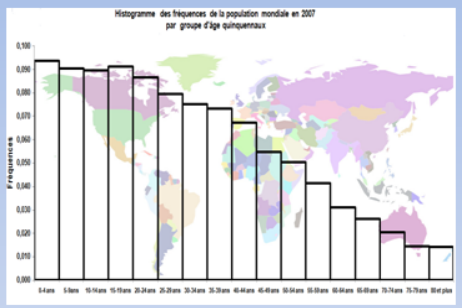
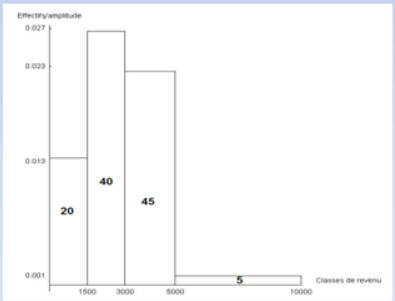
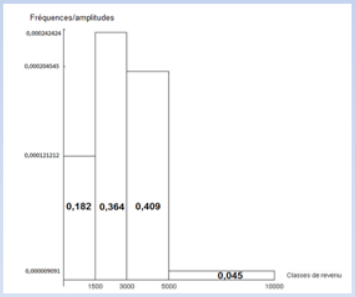
L'histogramme peut parfois être assimilé à un graphique en barre verticales. C'est le cas lorsque le graphique en barres verticales a pour échelle horizontale des catégories numériques d'amplitudes égales. Mais l'histogramme se distingue du graphique en barres verticales lorsque les catégories numériques de l'axe horizontal sont d'amplitudes inégales. Le diagramme de VENN ci-après permet d'illustrer ce point et les exemples qui suivent de le préciser.

Barres verticales et histogrammes : pas toujours la même chose !



S'agissant des histogrammes, il convient en outre de distinguer les histogrammes d'effectifs et les histogrammes de fréquences. Nous allons donc être amenés à étudier 4 types d'histogrammes comme indiqué dans le tableau ci-dessous.

Les 4 types d'histogrammes

	Histogramme d'effectifs	Histogramme de fréquences
Amplitudes de classes égales	1 	2 
Amplitudes de classes inégales	3 	4 

Nous allons construire chacun de ces 4 types d'histogrammes ci-après en les identifiant par leur numéro dans le tableau (de 1 à 4).

1) Amplitude de classes identiques

Soir le tableau ci-dessous qui donne a population mondiale en 2007 par groupes d'âges quinquennaux (hommes et femmes confondus). La dernière colonne, intitulées « fréquences » est simplement calculée en divisant l'effectif de chaque classe d'âge par la population mondiale totale. Par exemple, pour obtenir le premier chiffre de la colonne des fréquences, on a effectué le calcul suivant :

$$\frac{618068212}{6602236756} = 0,0936149725$$

Soit, en arrondissant : 0,094.

On remarque que toutes les classes d'âges sont identiques (5 ans)[14]. Les classes ont la même amplitude. A chaque classe d'âge est associé un effectif (colonne des

[14] La dernière classe va en fait de 80 à plus de 110 ans, mais pour simplifier, nos la supposons égale à 5 ans, en nous basant sur le fait que le nombre des 85 ans et plus reste encore minime comparé à l'ensemble de la population mondiale, même s'il est appelé à augmenter.

effectifs) ou une fréquence (colonne des fréquences). La somme des effectifs donne la population mondiale en 2007, tandis que la somme des fréquences est égale à 1.

Nous allons d'abord voir comment se présente l'histogramme des effectifs, puis ensuite l'histogramme des fréquences.

Population mondiale en 2007
par groupe d'âge quinquennaux
(Source : http://www.census.gov/cgi-bin/ipc/idbagg)

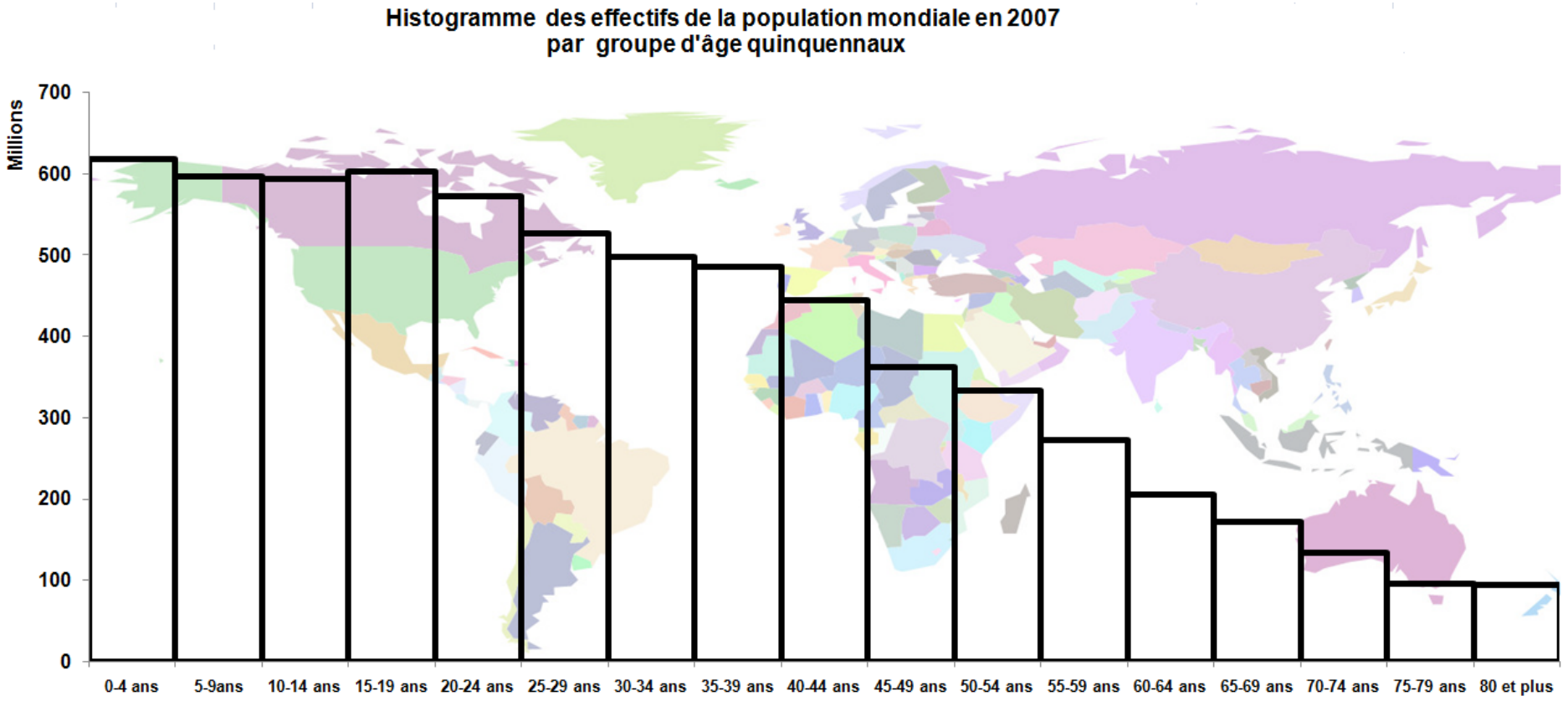
Groupes d'âge	Effectifs	Fréquences
0-4 ans	618 068 212	0,094
5-9ans	596 388 696	0,090
10-14 ans	592 749 377	0,090
15-19 ans	603 006 366	0,091
20-24 ans	571 675 796	0,087
25-29 ans	526 038 297	0,080
30-34 ans	497 161 565	0,075
35-39 ans	484 705 023	0,073
40-44 ans	443 952 422	0,067
45-49 ans	361 519 891	0,055
50-54 ans	333 131 650	0,050
55-59 ans	272 679 024	0,041
60-64 ans	205 231 425	0,031
65-69 ans	172 652 318	0,026
70-74 ans	134 109 563	0,020
75-79 ans	95 400 385	0,014
80 et plus	93 766 743	0,014
	6 602 236 753	1

a) Histogramme des effectifs

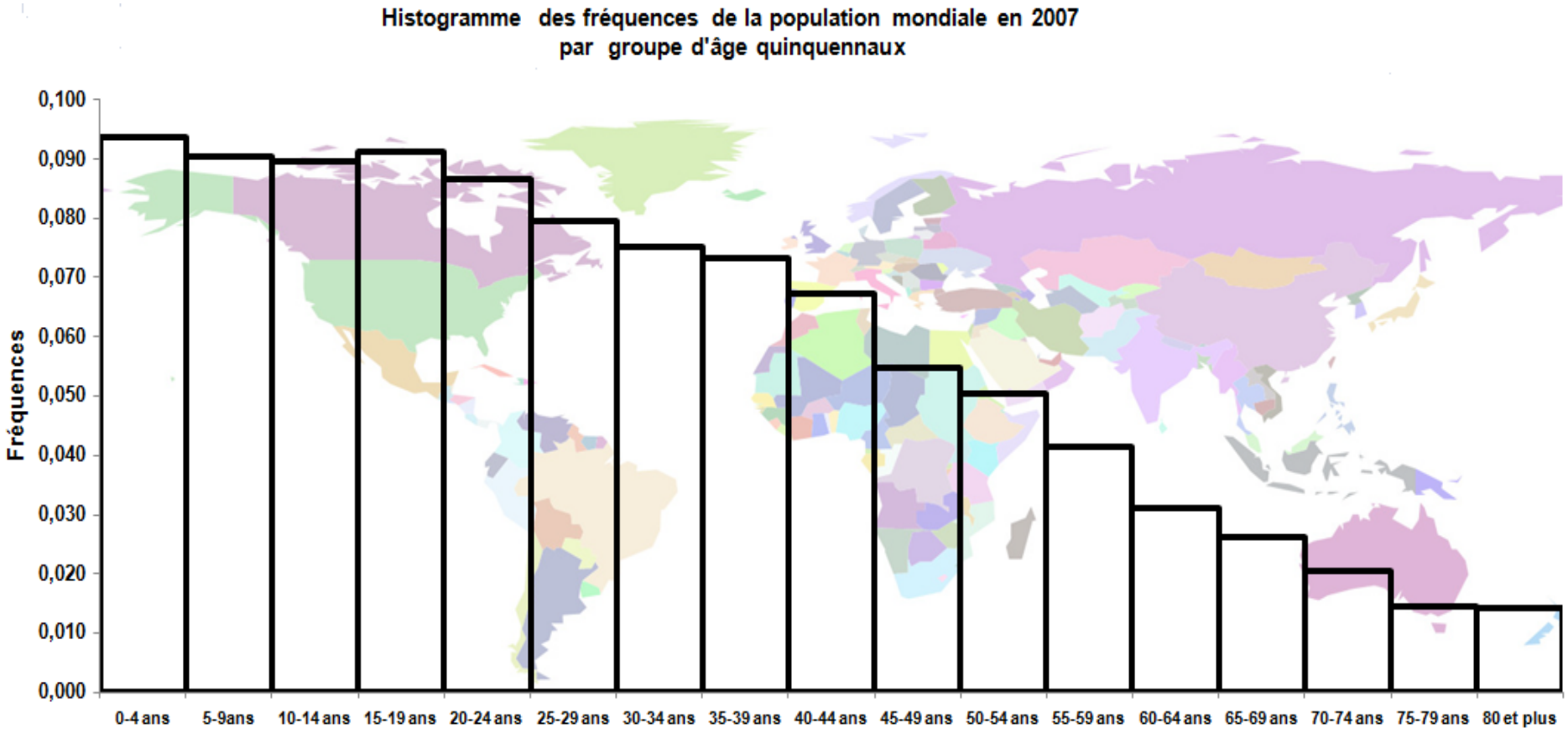
Sur l'histogramme des effectifs ci-après, on peut voir que :

- L'axe horizontal a pour échelle des catégories numériques d'amplitude égales (5 ans)
- L'axe vertical correspond à l'effectif de population associé à la classe d'âge correspondante.

Exemple d'histogramme d'effectifs quand les catégories numériques sont d'amplitudes égales
Correspond à l'histogramme numéroté 1 dans le tableau « Les 4 types d'histogrammes »



Exemple d'histogramme de fréquences quand les catégories numériques sont d'amplitudes égales
Correspond à l'histogramme numéroté 2 dans le tableau « Les 4 types d'histogrammes »



b) Histogramme des fréquences

Sur l'histogramme des fréquences ci-avant, on peut voir que :

- L'axe horizontal a pour échelle des catégories numériques d'amplitude égales (5 ans)
- L'axe vertical correspond à la fréquence associée à la classe d'âge correspondante.

Les deux histogrammes (celui des effectifs et celui des fréquences) ont la même forme, mais diffèrent par l'échelle de l'axe vertical.

2) Amplitude de classes différentes

Pour bien comprendre ce qu'est un histogramme (d'effectifs ou de fréquences) lorsque que les catégories numériques de l'axe horizontal sont d'amplitudes différentes, prenons l'exemple d'un échantillon de 110 ménages dont le revenu mensuel en euros est donné par la série classée ci-après :

Revenu = {1100, 1130, 1150, 1200, 1220, 1300, 1300, 1310, 1350, 1400, 1400, 1400, 1400, 1450, 1460, 1480, 1490, 1490, 1495,1495, 1500, 1500, 1550, 1600, 1600, 1630, 1640, 1700, 1900, 2000, 2020, 2050, 2070, 2090, 2100, 2200, 2220, 2400, 2500, 2540, 2560, 2600, 2710, 2730, 2750, 2800, 2810, 2810, 2820, 2840, 2850, 2850, 2850, 2870, 2890, 2900, 2920, 2960, 2980, 2990, 3000, 3000, 3000, 3000, 3000, 3030, 3050, 3070, 3080, 3090, 3090, 3090, 3095, 3100, 3200, 3210, 3250, 3280, 3300, 3350, 3400, 3400, 3400, 3400, 3420, 3450, 3500, 3550, 3560, 3570,3575, 3600, 3610, 3800, 4000, 4100, 4250, 4300, 4310, 4380, 4500, 4560, 4580, 4590, 4590, 5000, 6000, 7500, 9000, 9800}.

Supposons que l'on souhaite répartir ces ménages dans les catégories de revenu suivantes : [0 – 1500[; [1500 – 3000[;[3000 – 5000[; [5000 – 10000[. On va alors obtenir le tableau d'effectifs suivant :

Répartition des 110 ménages par classe de revenu

Classes de revenu	Effectifs
[0 -1500[	20
[1500-3000[	40
[3000-5000[	45
[5000-10000[	5
	110

L'amplitude de chaque classe n'est pas la même. Les 2 premières classes ont une amplitude de 1500 euros, la troisième a une amplitude de 2000 euros et la dernière classe a une amplitude de 5000 euros. Par conséquent, si l'on veut représenter ces données sous forme d'un histogramme d'effectifs, nous ne pouvons pas procéder de la même façon que lorsque les amplitudes de classes sont identiques. En effet, sur l'axe vertical, les « barres » n'auront plus la même longueur. L'échelle de l'axe horizontal est le suivant :

On voit que dans ces conditions, la hauteur des barres verticales ne peut plus être proportionnelle aux effectifs, car cela aboutirait à donner une image fausse de l'importance des effectifs inclus dans chaque classe.

a) Histogramme d'effectifs

Pour tracer l'histogramme des effectifs, il faut donc modifier l'échelle de l'axe vertical en divisant les effectifs de chaque classe par l'amplitude de classe correspondante. On ajoute pour cela deux colonnes au tableau précédent :

Calcul des effectifs corrigés (effectifs sur amplitudes)

Classes de revenu	Amplitudes de classe	Effectifs	Effectifs /amplitudes
[0 -1500[	1 500	20	0,013
[1500-3000[	1 500	40	0,027
[3000-5000[	2 000	45	0,023
[5000-10000[	5 000	5	0,001
		110	

La colonne « amplitude de classe » donne l'écart en euros entre les deux extrémités de chaque classe. La colonne « effectifs corrigés » se calcule en divisant chaque effectif par l'amplitude de classe qui lui correspond ; Ainsi, l'effectif corrigé de la classe de revenu [0 – 1500[s'obtient par l'opération suivante :

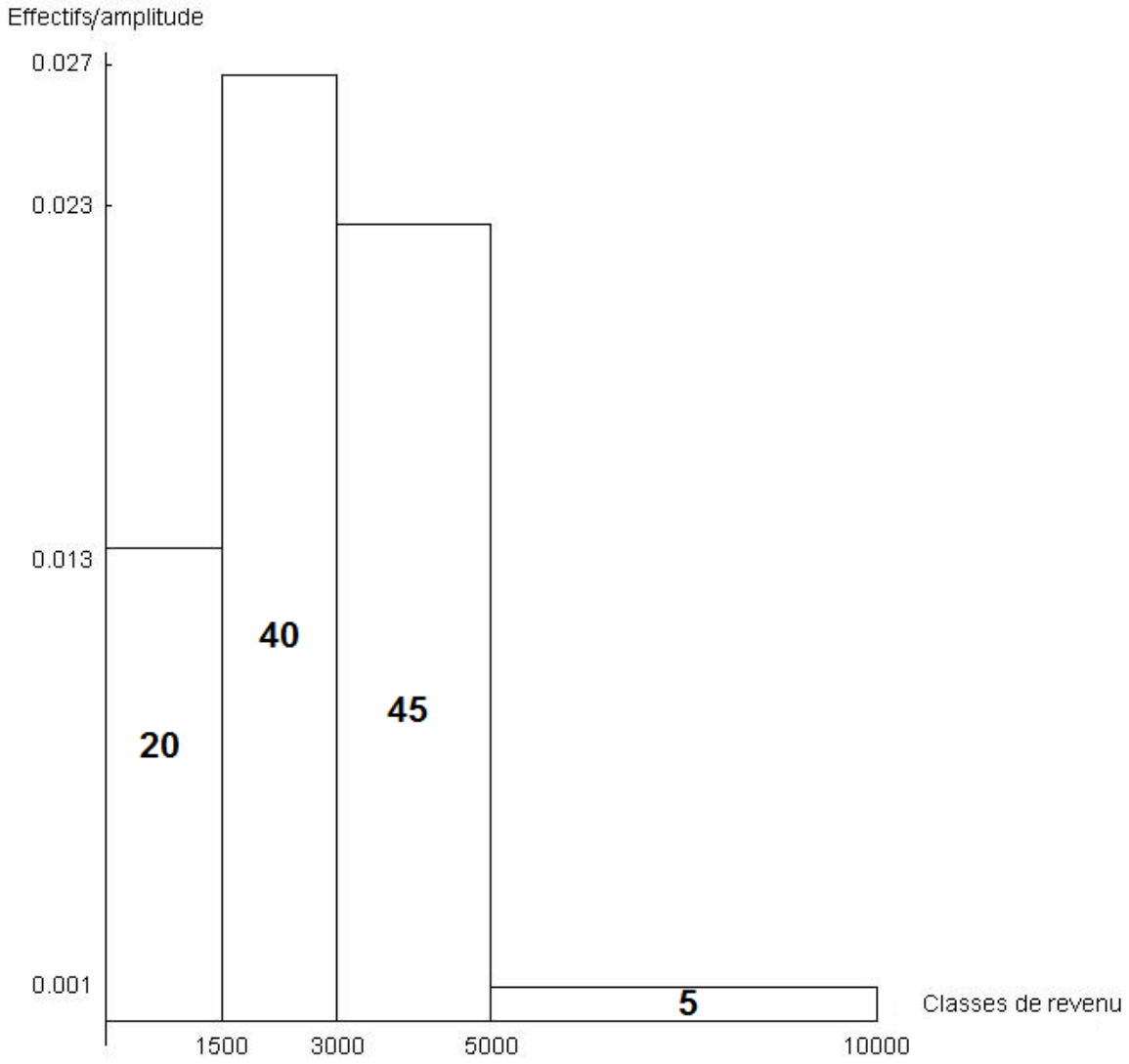
$$\frac{20}{1500} = 0,01333333$$

C'est-à-dire 0,013 en ne conservant que les 3 premières décimales. Les autres chiffres de la colonne s'obtiennent de la même façon.

Nous pouvons maintenant tracer l'histogramme des effectifs (voir graphique ci-après). Dans cet histogramme, ce n'est plus la **hauteur** de chaque barre qui indique l'effectif, mais sa **surface**. C'est la raison pour laquelle *l'effectif est reporté directement sur chaque barre*, tandis que l'axe vertical mesure **l'effectif corrigé**, ou plus précisément **l'effectif divisé par l'amplitude de classe**. Sur cet histogramme, ce n'est plus la hauteur qui correspond à l'effectif, mais la surface. On peut voir

facilement que la barre qui correspond à 40 a une surface double de celle qui correspond à 20. Et, bien que cela ne soit pas évident visuellement, la barre qui correspond à 45 a une surface qui est 45/40=1,125 plus grande que celle qui correspond à 40 et une surface 45/5 = 9 fois plus grande que celle qui correspond à 5.

Histogramme d'effectifs
avec catégories numériques d'amplitudes différentes
Correspond à l'histogramme numéroté 3
dans le tableau « Les 4 types d'histogrammes »



b) Histogramme de fréquences

Pour tracer l'histogramme des fréquences, il faut donc modifier l'échelle de l'axe vertical en divisant les fréquences de chaque classe par l'amplitude de classe correspondante.

On construit pour cela le tableau suivant :

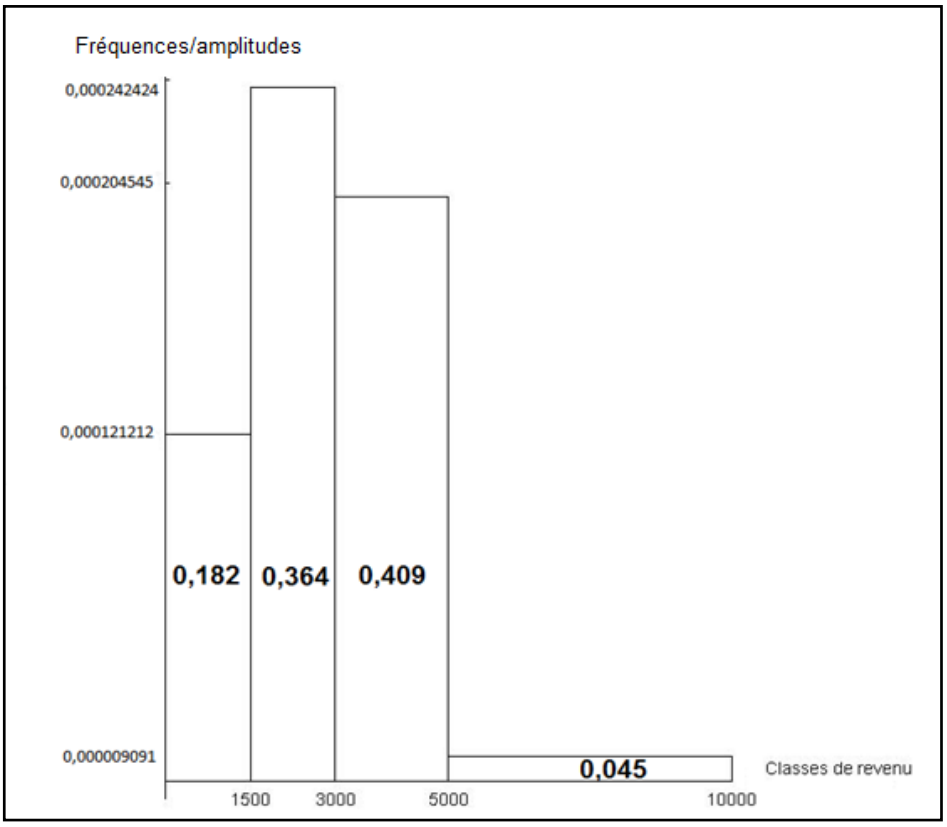
Calcul des fréquences corrigées (fréquences sur amplitudes)

Classes de revenu	Amplitudes de classe	Effectifs	Fréquences	Fréquences/amplitudes
[0 -1500[	1 500	20	0,182	0,000121212
[1500-3000[	1 500	40	0,364	0,000242424
[3000-5000[	2 000	45	0,409	0,000204545
[5000-10000[	5 000	5	0,045	0,000009091
		110	1,000	

Dans ce tableau, la colonne des fréquences correspond aux effectifs de chaque classe divisés par l'effectif total et la colonne des « fréquences corrigées » est égale au rapport de chaque fréquence à l'amplitude de classe correspondante.

On obtient un histogramme de fréquence qui a exactement la même forme que l'histogramme des effectifs, mais l'échelle verticale est graduée différemment (c'est l'échelle des amplitudes corrigées). Et dans chaque barre figure maintenant la fréquence qu'elle représente.

Histogramme de fréquences
avec catégories numériques d'amplitudes différentes
Correspond à l'histogramme numéroté 4 dans le tableau « Les 4 types d'histogrammes »



G – Pyramide des âges

La pyramide des âges est un outil de l'analyse démographique plus célèbre encore que le diagramme de LEXIS (et surtout plus populaire !). Ce graphique a été inventé en 1870 par le Général WALKER, alors directeur du *Bureau of Census*, organisme américain chargé du recensement de la population et des études démographiques.

C'est en fait un **double histogramme** qui indique la composition d'une population par classes d'âges et par sexe, à un moment donné.

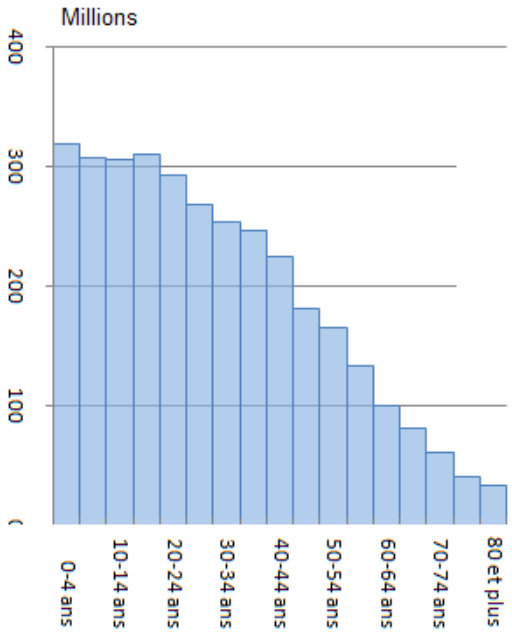
Faire une pyramide, c'est faire graphique du nombre d'hommes et de femmes de différents âges. Pour cela on place dos à dos, deux histogrammes décrivant la structure par classe d'âge d'une population masculine pour le premier et d'une population féminine pour le second. Le tableau ci-dessous répertorie la population mondiale par groupe d'âges quinquennaux et par âge en 2007. Nous allons l'utiliser à titre d'exemple pour construire la pyramide des âges.

Population mondiale en 2007
par groupe d'âge quinquennaux et par sexe
(Source : http://www.census.gov/cgi-bin/ipc/idbagg)

	Hommes	Femmes
0-4 ans	318 545 370	299 522 842
5-9ans	307 654 668	288 734 028
10-14 ans	305 356 133	287 393 244
15-19 ans	309 777 481	293 228 885
20-24 ans	292 268 618	279 407 178
25-29 ans	268 515 975	257 522 322
30-34 ans	253 351 999	243 809 566
35-39 ans	246 936 434	237 768 589
40-44 ans	224 153 326	219 799 096
45-49 ans	180 624 189	180 895 702
50-54 ans	165 121 279	168 010 371
55-59 ans	134 055 565	138 623 459
60-64 ans	99 801 631	105 429 794
65-69 ans	81 559 685	91 092 633
70-74 ans	61 202 989	72 906 574
75-79 ans	40 951 329	54 449 056
80 et plus	33 739 040	60 027 703

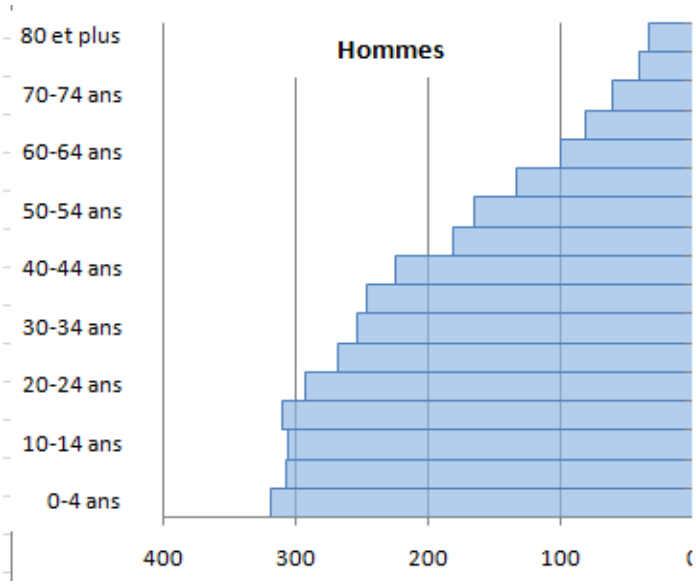
L'histogramme A ci-dessous représente la répartition de la population masculine mondiale en 2007 suivant les données du tableau précédent. Il s'agit d'un histogramme dont les amplitudes de classe sont identiques (5 ans) et qui peut donc être assimilé à un graphique en barres (voir le diagramme de VENN qui explique les différences et les similitudes entre barres verticales et histogramme).

Histogramme A



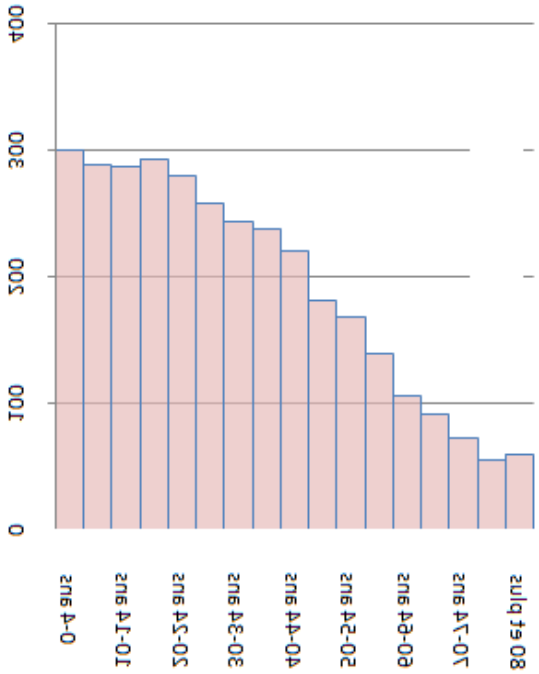
Nous pouvons ensuite faire effectuer une rotation à ce graphique de façon à obtenir l'histogramme B.

Histogramme B



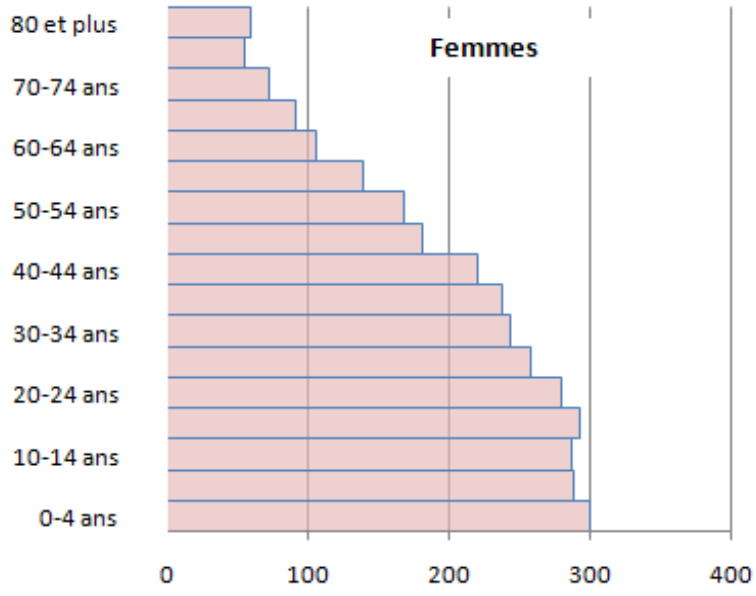
L'histogramme C ci-dessus représente la répartition de la population féminine mondiale en 2007.

Histogramme C

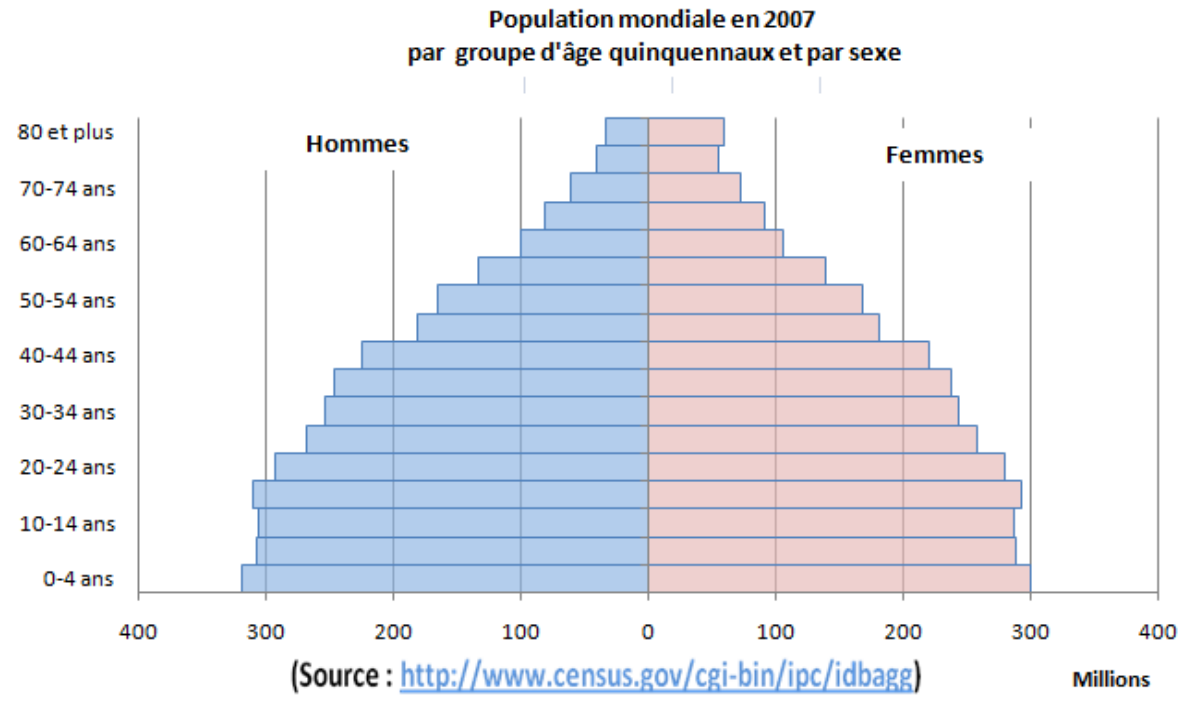


Nous pouvons ensuite faire effectuer une rotation à ce graphique de façon à obtenir l'histogramme D.

Histogramme D



Et finalement, en mettant côte à côte les histogrammes B et D, nous obtenons la pyramide classique des âges de la population mondiale en 2007 :



Il s'agit en fait d'un graphique qui représente TROIS dimension : l'âge, le sexe et les effectifs associés à ces deux catégories.

H – Graphique en cascade

Le graphique en cascade (waterfall graphic) est une variante du graphique en barres. Il sert à faire apparaître :

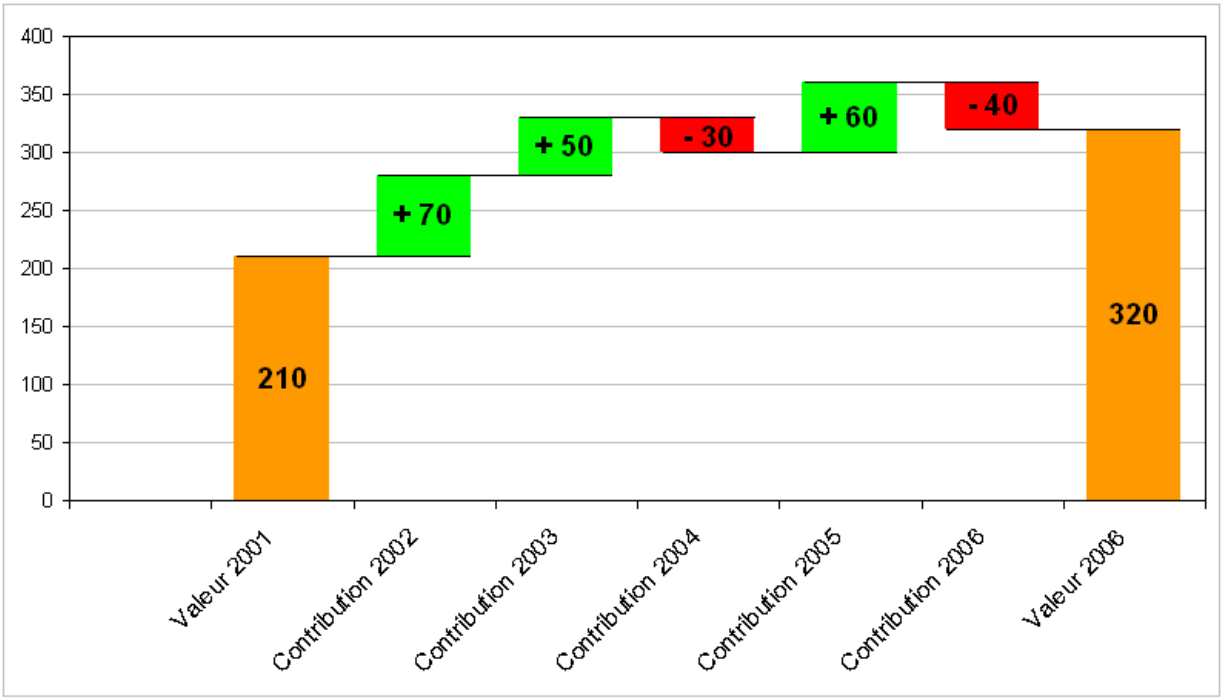
- les contributions positives et négatives à une grandeur

- les variations successives d'une grandeur.

Exemple : Le tableau ci-dessous indique l'évolution du chiffre d'affaire d'une entreprise de 2001 à 2006. La troisième colonne du tableau met en évidence les augmentations et la quatrième les diminutions. On voit ainsi que le CA a augmenté en 2002, 2003 et 2005, et qu'il a baissé en 2004 et 2006.

Années	CA (euros)	Plus	Moins
2001	210000		
2002	280000	70000	
2003	330000	50000	
2004	300000		-30000
2005	360000	60000	
2006	320000		-40000

Le graphique en cascade va permettre de faire apparaître ces variations (en milliers d'euros sur le graphique):



Voir le fichier Excel 2003 (il faut d'abord installer la macro : téléchargeable ici)

On voit ainsi immédiatement la contribution de chaque année et on peut visualiser :

- l'importance de la contribution
- Si la contribution est positive ou négative.
- comment on est passé de 210 à 320 par variations successives.

I – Graphique en trois dimensions

Grâce aux ordinateurs et aux logiciels il est devenu très facile aujourd'hui de réaliser de beaux graphiques en 3D. EXCEL 2007 offre diverses possibilités, tout comme d'autres logiciels, tels que Mathematica. En revanche, à moins d'être très bon dessinateur, il est impossible de réaliser ce type de graphique avec la règle, le rapporteur, le compas et les crayons de couleur (à l'inverse de tous les autres graphiques vu jusqu'à présent).

On peut distinguer 3 catégories de graphique en 3D :

- Les « faux » graphiques en 3D qui ne sont que des graphiques en 2D auxquels on ajoute une profondeur à des fins visuelles.
- Les graphiques en barres à 3 dimensions
- Les graphiques dits de surface

1) Graphiques en 2D avec ajout de profondeur

Pour illustrer ce type de graphique, reprenons les données du chiffre d'affaires hypothétique qu'une entreprise a réalisé en 2007 249 327 045 euros) répartis par ses 4 vendeurs et dans les trois villes où se trouvent ses clients.

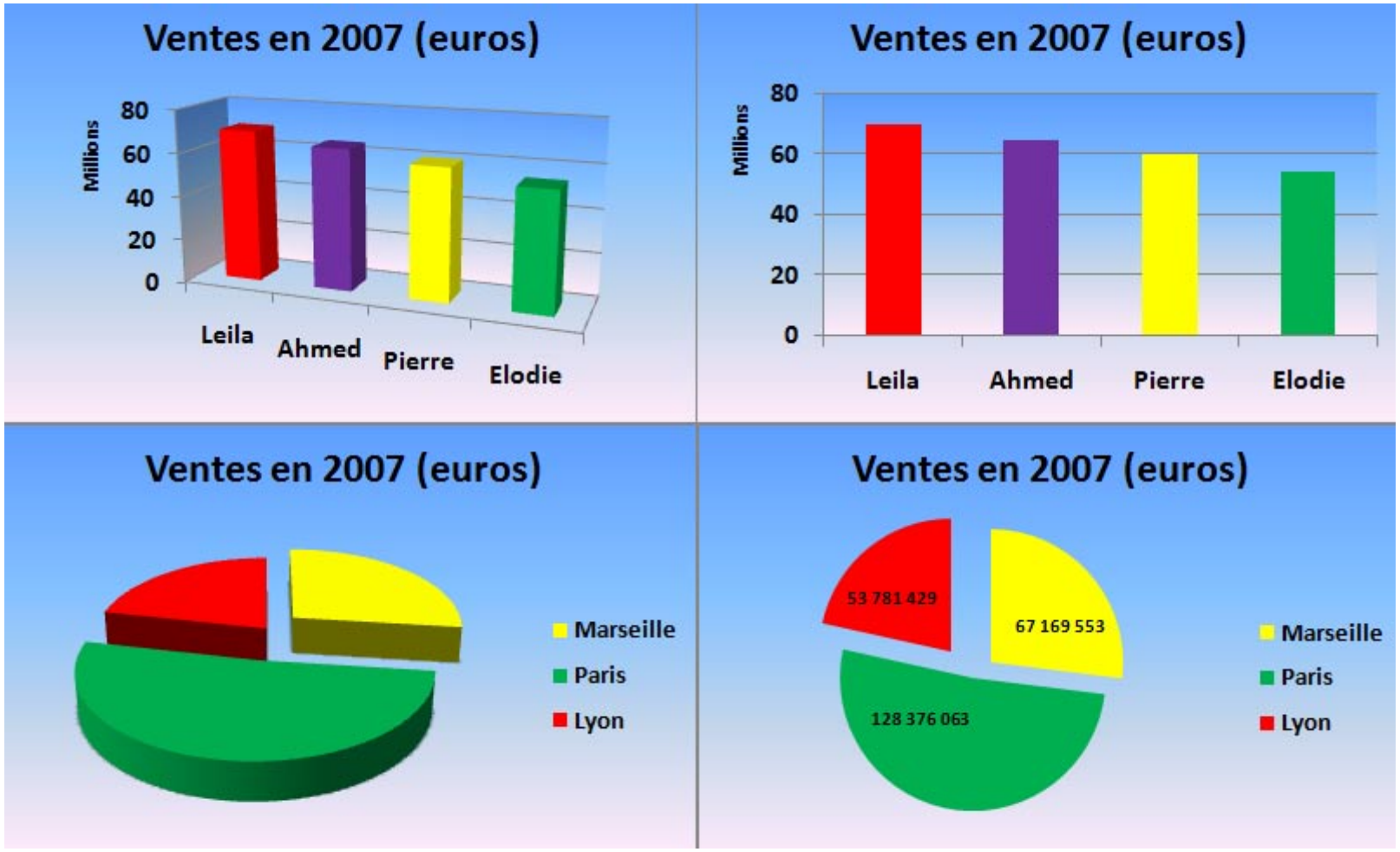
Ventes 2007 (euros)					
	Leila	Ahmed	Pierre	Elodie	Total
Marseille	13 225 478	20 154 287	17 892 555	15 897 233	67 169 553
Paris	37 895 214	35 877 421	32 558 741	22 044 687	128 376 063
Lyon	18 753 951	8 754 668	9 785 246	16 487 564	53 781 429
Total	69 874 643	64 786 376	60 236 542	54 429 484	**249 327 045**

Lorsque l'on ajoute de la « profondeur » ou de la « perspective » au graphique en barres verticales ou aux secteurs classiques, on obtient ce genre de résultats (voir graphiques ci-après).

Bien entendu, le nombre de dimensions n'a pas changé par rapport à l'équivalent 2D de ces deux graphiques qui n'ont que l'inconvénient de paraître « plats » par comparaison.

Il semble que les graphiques 2D avec ajout de profondeur ou de perspective attirent davantage les regards. Il ne faut donc pas se gêner pour les utiliser surtout étant donné la facilité avec laquelle on peut les réaliser grâce aux logiciels.

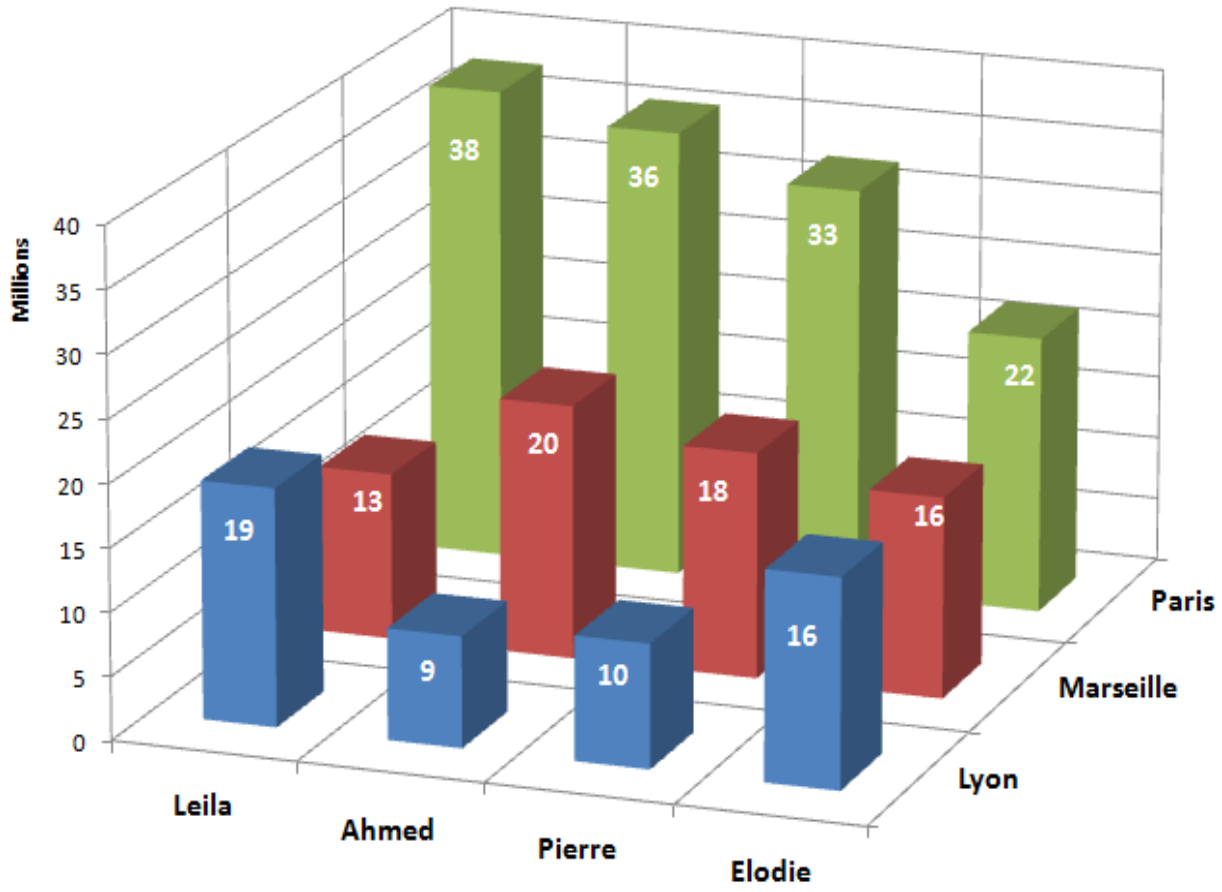
À gauche, graphiques en 2D avec ajout de « profondeur » ; à droite : leurs équivalents 2D



2) Graphique en barres avec 3 dimensions "réelles"

Par comparaison avec les graphiques en 2D avec ajout de profondeur, les graphiques en barres avec 3 dimensions ajoute réellement une dimension supplémentaire.

Cette fois, nous allons utiliser un seul graphique en barres verticales pour montrer à la fois la répartition du CA par villes et par vendeur en 2007.



Ventes 2007 (euros)

	Leila	Ahmed	Pierre	Elodie	Total
Lyon	18 753 951	8 754 668	9 785 246	16 487 564	53 781 429
Marseille	13 225 478	20 154 287	17 892 555	15 897 233	67 169 553
Paris	37 895 214	35 877 421	32 558 741	22 044 687	128 376 063
Total	13 225 478	20 154 287	17 892 555	15 897 233	**67 169 553**

Chapitre 6
Tendances et corrélations

1 – Introduction
2 – La détermination de la tendance d'une série chronologique
    A – Détermination graphique
    B – Détermination par la méthode des points extrêmes
    C – Détermination par la méthode des moindres carrés (MCO)
3 – L'étude de la corrélation entre deux variables
    A – L'exemple d'une fonction de demande
    B – L'équation de régression linéaire
    C – Le coefficient de détermination
4 – Le test du Khi-carré
    A – Introduction
    B – Exemple d'utilisation

1 – Introduction

Ce chapitre est consacré à l'utilisation d'un même outil statistique, l'ajustement linéaire, à deux cas de figure différents. L'ajustement linéaire, aussi appelé « méthode des Moindres Carrés Ordinaires (MCO)».

Il est appliqué successivement :

- A l'étude de la tendance d'une série chronologique
- À la mise en évidence d'une corrélation entre deux variables. La méthode des moindres carrés est également utilisée pour étudier l'existence d'une corrélation entre deux variables.

Ci-après, deux graphiques :

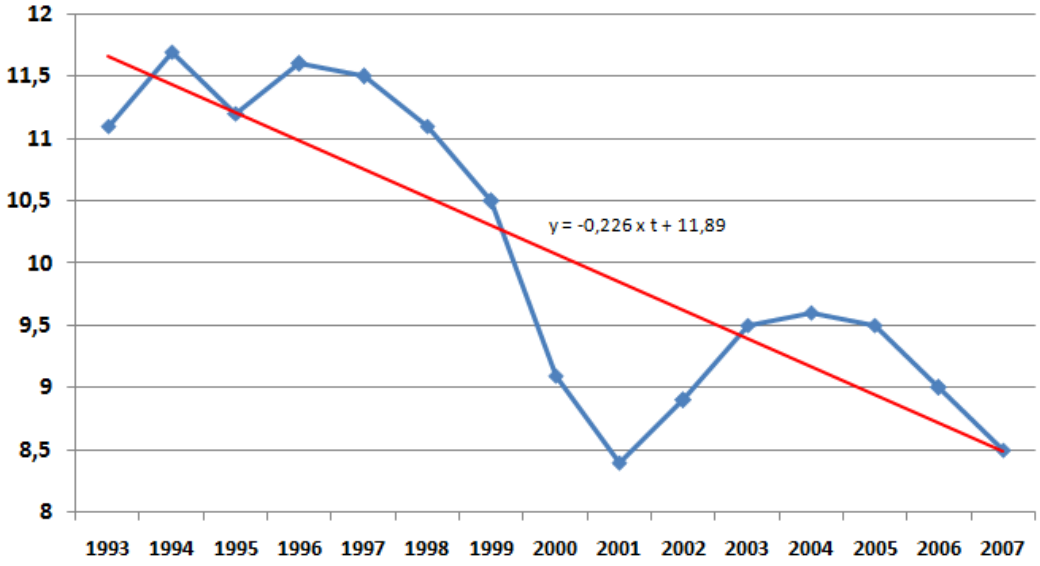
- l'un illustre le tracé d'un trend (« tendance ») linéaire à partir de l'étude d'une série chronologique
- l'autre illustrant le tracé d'une droite linéaire pour apprécier l'existence d'une relation entre deux variables.

Dans les deux cas, ces droites ont été obtenues à l'aide de la méthode des moindres carrés ordinaires :

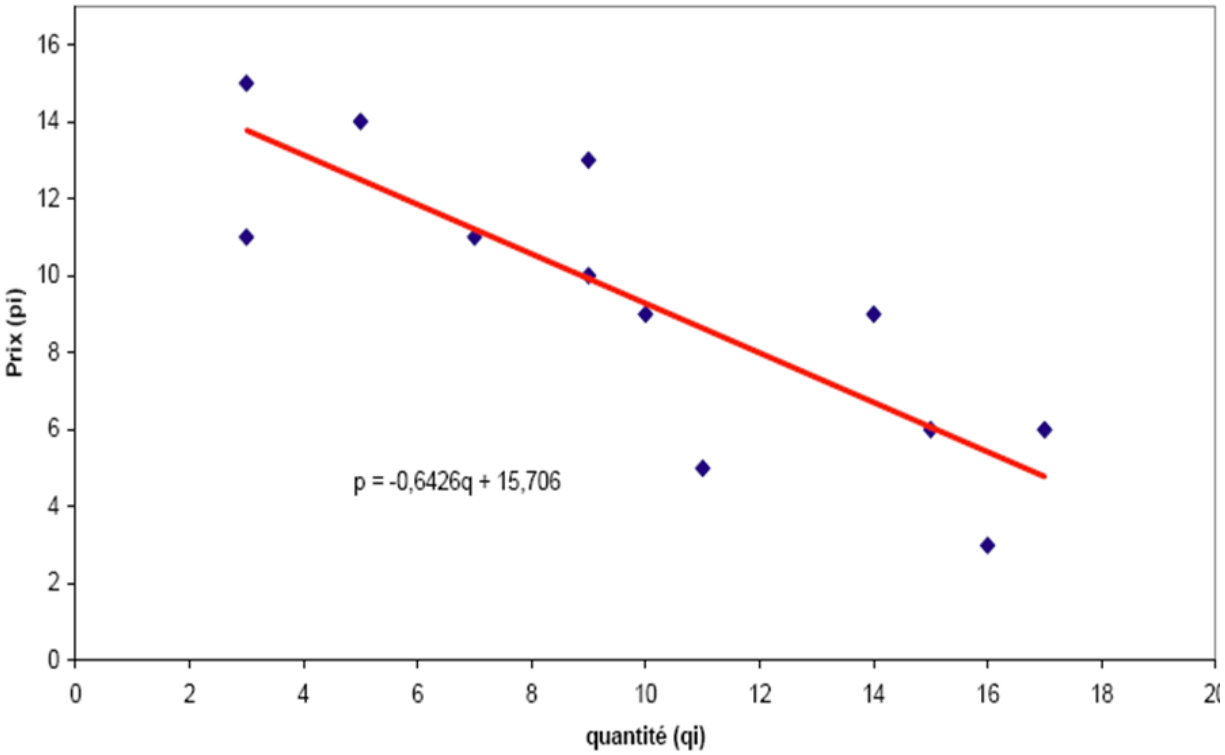
- pour les séries chronologiques, on la qualifie de « trend linéaire »,
- pour l'étude de la relation statistique entre deux variables, on parle plus volontiers de droite de régression.

Pour l'étude de la corrélation entre deux dimensions non quantitatives, c'est le test du Khi-carré qui remplace l'ajustement linéaire.

Trend linéaire d'une série chronologique



Ajustement linéaire de la relation entre prix et quantité d'un bien



2 – La détermination de la tendance d'une série chronologique

A – Détermination graphique

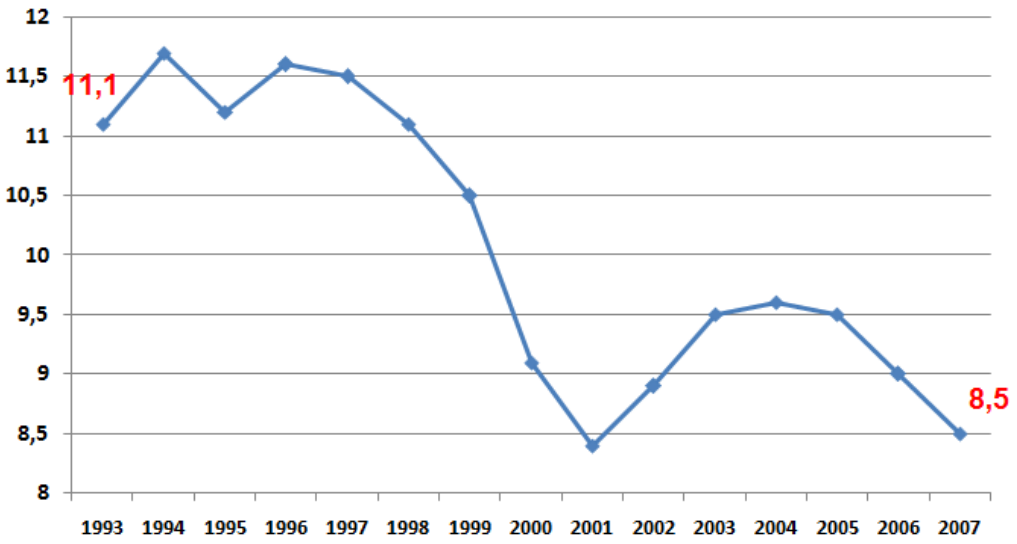
Soit le tableau suivant qui donne l'évolution du taux de chômage en France de 1993 à 2007.

Années	Taux de chômage (%) - France
1993	11,1
1994	11,7
1995	11,2
1996	11,6
1997	11,5
1998	11,1
1999	10,5
2000	9,1
2001	8,4
2002	8,9
2003	9,5
2004	9,6
2005	9,5
2006	9
2007	8,5

Source : FMI

Pour étudier l'évolution de cette série chronologique, le plus simple est de la représenter à l'aide d'un graphique en ligne :

Taux de chômage en France de 1993 à 2007 (selon données FMI)



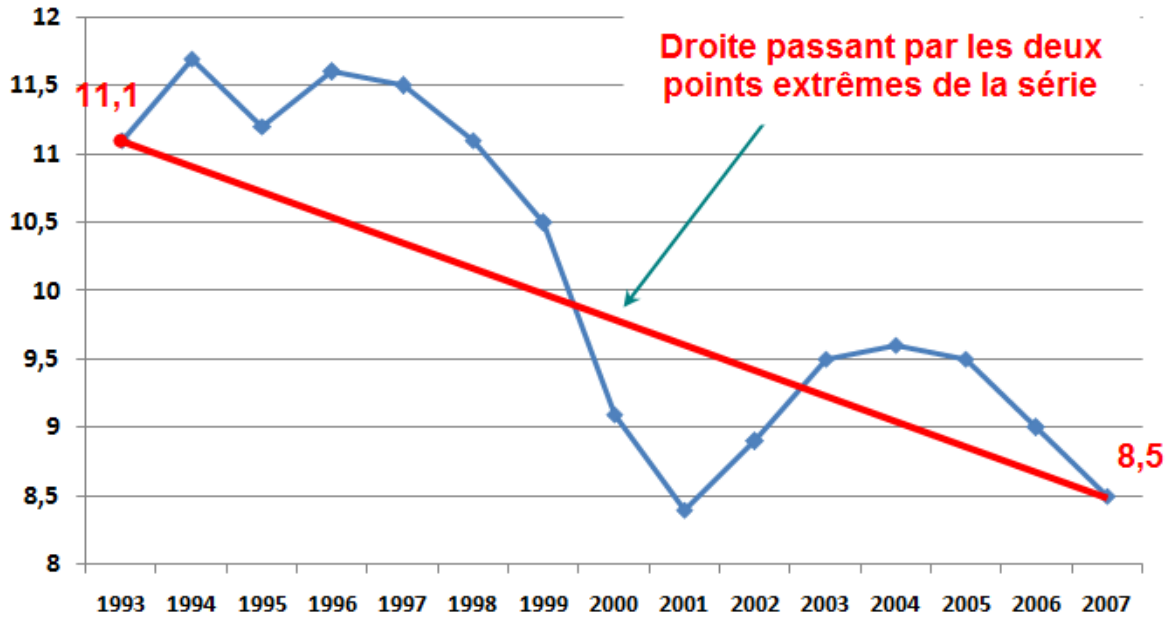
On constate alors que la « tendance » du taux de chômage sur les 15 dernières années est à la baisse. Il s'agit là d'une simple constatation visuelle, suffisante dans bien des cas. On constate aussi qu'après avoir baissé jusqu'en 2001, le taux de chômage a connu une légère remontée avec un pic en 2004 et a ensuite encore baissé pour atteindre 8,5 % en 2007.

Bien souvent, cette analyse graphique est suffisamment éloquente pour ne pas poursuivre l'analyse. Néanmoins, il est possible de poursuivre plus rigoureusement ce raisonnement et de déterminer mathématiquement une droite dont la pente nous donnera la « tendance ».

B – Détermination de la tendance par la méthode des points extrêmes

Puisqu'il faut deux points pour tracer une droite, une idée simple consiste à faire passer une droite par les deux points extrêmes de la série, soit {1993 ;11,1} et {2007 ; 8,5} d'autre part. On obtient alors une droite qui nous indique une tendance négative.

Détermination de la tendance par une droite passant par les deux points extrêmes



Cette méthode n'est cependant pas très satisfaisante car elle ne tient compte que de des deux points extrêmes. Une meilleure méthode est celle dite des « moindres carrés ordinaires » ou MCO en abrégé.

C – Détermination de la tendance par la méthode MCO

Ce nom bizarre vient du fait que la méthode consiste à déterminer la droite d'ajustement en **minimisant la somme du carré des écarts entre cette droite et les observations**. Les détails mathématiques de cette méthode importent peu dans un cours de statistique descriptive, car **l'essentiel est de savoir calculer les coordonnées de la droite**. De plus, les machines à calculer ainsi que les logiciels comme Excel permettent un calcul et un tracé facile de cette droite.

Cette droite, comme toutes les droites, a pour expression l'équation :

$$y_i = a.\, t_i + b$$

où i varie de 1 à n, et où n est le nombre des observations.

Les valeurs $\{t_1, t_2, ..., t_i, ... t_n\}$ sont les dates.

Dans notre exemple les chiffres 1 à 12 (le chiffre 1 correspond à 1993 et le chiffre 12 correspond à 2007). Les y_i , c'est-à-dire les valeurs tendancielles, ne peuvent être calculées qu'une fois que l'on connaît a et b. Pour calculer les coefficients a et b, nous allons donc utiliser les valeurs observées, à savoir la série :

{11,1 ;11,7 ;11,2 ;11,6 ;11,5 ;11,1 ;10,5 ;9,1 ;8,4 ;8,9 ;9,5 ; 9,6 ;9,5 ;8,5}.

Les formules de calcul des coefficients *a* et *b* sont alors données respectivement par[15] :

$$a = \frac{\sum_i t_i y_i - n \cdot \bar{t} \cdot \bar{y}}{\sum_i t_i^2 - n(\bar{t})^2}$$

$$b = \bar{y} - a\bar{t}$$

[15] Ces formules sont données ici sans démonstration, le lecteur intéressé par une démonstration rigoureuse pourra consulter avec profit le livre de PY, Bernard (2007), Statistique descriptive : nouvelle méthode pour comprendre et bien réussir 5ème édition, Economica.

Détermination du trend par la méthode des moindres carrés

t_i	y_i	$t_i y_i$	t_i^2
1	11,1	11,1	1
2	11,7	23,4	4
3	11,2	33,6	9
4	11,6	46,4	16
5	11,5	57,5	25
6	11,1	66,6	36
7	10,5	73,5	49
8	9,1	72,8	64
9	8,4	75,6	81
10	8,9	89	100
11	9,5	104,5	121
12	9,6	115,2	144
13	9,5	123,5	169
14	9	126	196
15	8,5	127,5	225
120	**151,2**	**1146,2**	**1240**

$\sum_{i=1}^{15} t_i$	$\sum_{i=1}^{15} y_i$	$\sum_{i=1}^{15} t_i y_i$	$\sum_{i=1}^{15} t_i^2$	$\bar{t} = \frac{1}{15}\sum_{i=1}^{15} t_i$	$\bar{t}^2$	$n \times \bar{t}^2$	$\bar{y} = \frac{1}{15}\sum_{i=1}^{15} y_i$	$n\,\bar{t}\,\bar{y}$
				8	64	960	10,08	1209,6

On calcule ensuite le coefficient a :

$$a = \frac{\sum_{i=1}^{n} t_i y_i - n \cdot \bar{t} \cdot \bar{y}}{\sum_{i=1}^{n} t_i^2 - n(\bar{t})^2} = \frac{1146,2 - 1209,6}{1240 - 960} = \frac{-63,4}{280} = -0,2264$$

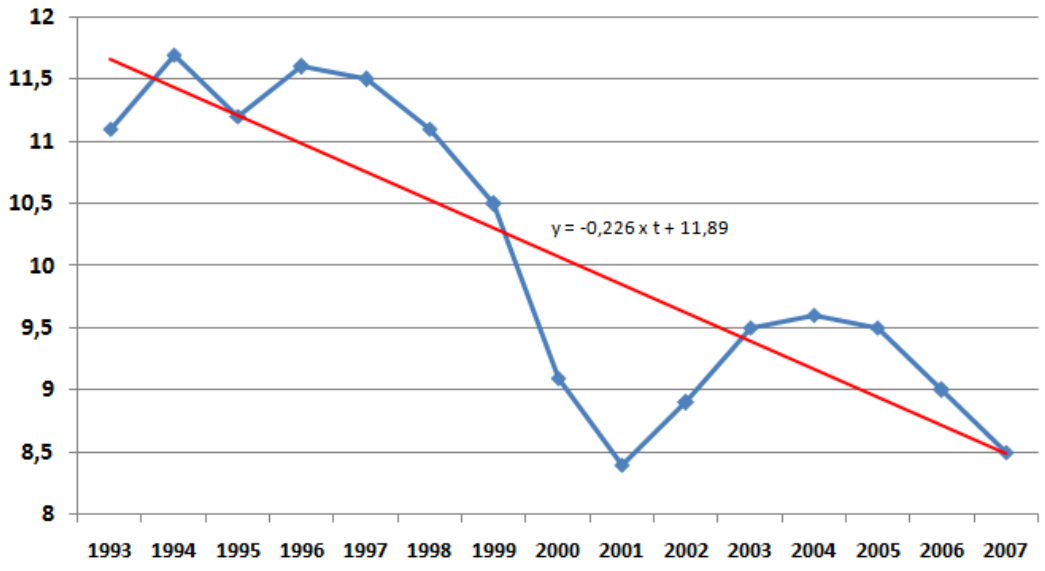
et on en déduit le coefficient b :

$$b = \bar{y} - a\bar{t} = 10,08 - (-0,2264 \times 8) = 10,08 + 1,8112 = 11,8912$$

On obtient donc l'équation du trend qui est :

$$y = -0,2264t + 11,891$$

Nous pouvons alors tracer la droite sur le graphique initial :



La méthode MCO est plus rigoureuse que la méthode car elle « calcule » la droite de tendance en tenant compte de toutes les observations.

3 – L'étude de la liaison statistique entre deux variables

Nous allons maintenant nous intéresser à la mise en évidence due relation statistique entre deux variables à partir de la méthode MCO. Nous prendrons comme exemple, l'étude de la relation entre prix et quantité d'un produit.

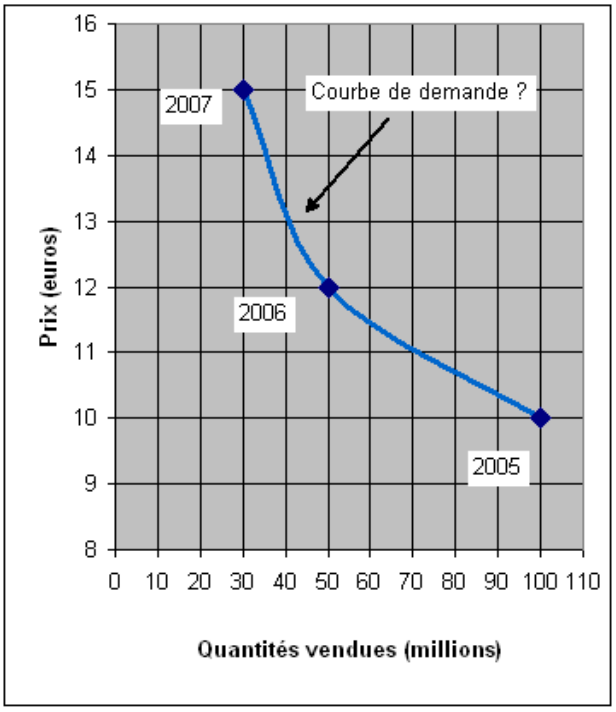
A – L'exemple d'une fonction de demande

Il est important pour une entreprise d'avoir une bonne idée de la demande qui s'adresse à son produit. Comment faire pour connaître la fonction de demande pour un produit ?

La première idée qui vient à l'esprit consiste à tracer un repère quantité/prix, avec la quantité en abscisse et le prix en ordonnée, comme ci-dessous. Supposons que l'on dispose pour cela des informations suivantes :

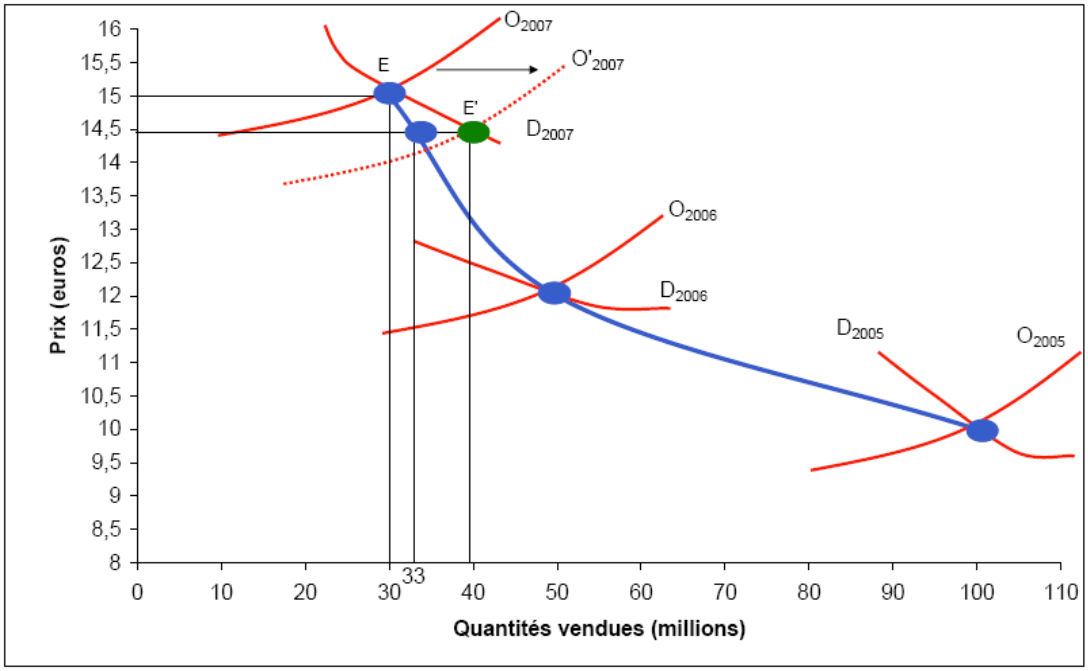
	Prix du produit (euros)	quantités vendues (millions d'unités)
2005	10	100
2006	12	50
2007	15	30

Un graphique basé sur ces informations donnerait le résultat suivant :



La courbe obtenue en joignant les 3 points est bien décroissante et suggère que plus le prix augmente, plus la quantité demandée diminue. S'agit-il pour autant d'une fonction de demande ? En fait, pas forcément. Le prix et la quantité d'un bien sont normalement déterminés à la fois par l'offre et la demande, du moins lorsque le marché est concurrentiel (si le marché n'est pas concurrentiel, les choses n'en sont que plus compliquées). Mais, quoiqu'il en soit, le prix et la quantité du produit s'établissent à l'intersection de l'offre et de la demande.

Ainsi, en fait, les 3 points du graphique précédent sont généralement interprétés par les économistes comme trois points d'équilibre, ainsi qu'illustré ci-dessous :

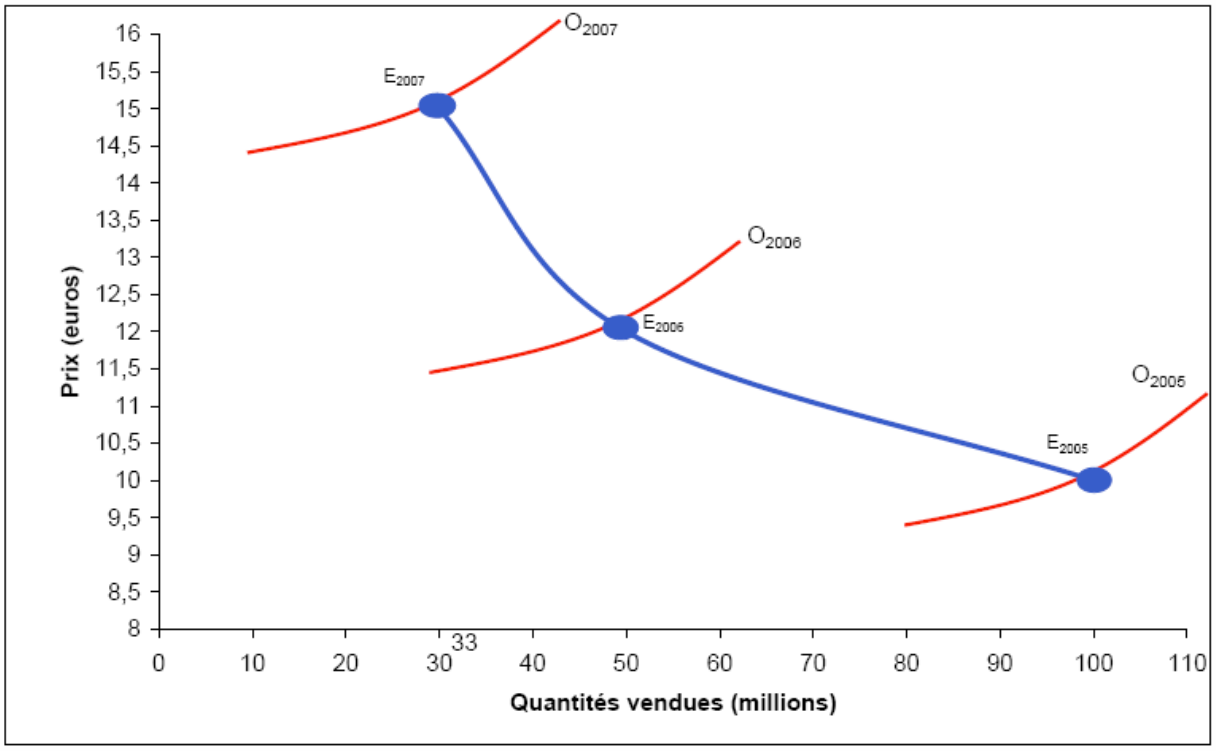


Sur ce graphique, nous voyons en fait que les 3 points précédents sont trois points d'équilibre qui résultent de l'intersection de courbes d'offre et de demande. Par exemple, en 2007, il s'est vendu 30 millions d'unités au prix unitaire de 15 euros, ce qui correspond au point d'équilibre E, qui est à l'intersection des courbes d'offre et de demande de l'année 2007.

En reliant les 3 observations temporelles pour en faire une courbe de demande, on commet sans doute une erreur. On est conduit ainsi à sous-estimer l'élasticité de la demande par rapport au prix. Ainsi, par exemple, on pourrait penser qu'une baisse du prix de 15 à 14,5 euros a pour effet d'augmenter la demande de 30 à 33 millions d'unités. Or en fait, ainsi qu'on peut le voir sur le graphique, une baisse du prix de 15 à 14,5 entraîne une augmentation bien plus importante de la demande (de 30 à 40 millions). Bien sûr, pour que la quantité vendue soit effectivement égale à 40 millions, il ne faut pas seulement qu'il y ait un déplacement le long de D_{2007}, il faut aussi que la courbe d'offre O_{2007} se déplace de façon à ce que le nouveau point d'intersection soit en E' (ce que nous supposons ici).

Cependant, comme illustré sur le graphique ci-dessous, on ne peut pas exclure que les 3 observations temporelles correspondent à 3 points sur la fonction de demande. Mais cela signifie en fait que la courbe de demande n'a pas changé, alors que la courbe d'offre s'est déplacée vers la gauche (en supposant que maintenant on commence en 2005, puis on continue avec 2006 et ensuite 2007).

Trois courbes d'offre successives, face à une courbe de demande inchangée constituent ainsi une justification simple de l'estimation d'une courbe de demande par un nuage de points constitués de couples prix/quantité observés à différents points du temps (de préférence en un même lieu). Il existe des analyses bien plus subtiles et le lecteur intéressé peut se rapporter pour plus de détails à un ouvrage d'économétrie.



B - L'équation de régression linéaire

Passons maintenant à la procédure d'estimation proprement dite, en supposant que le modèle approprié soit celui décrit par le schéma ci-dessus. Cependant, trois observations ne suffisent pas pour faire une estimation par la méthode des moindres carrés. Nous allons donc :

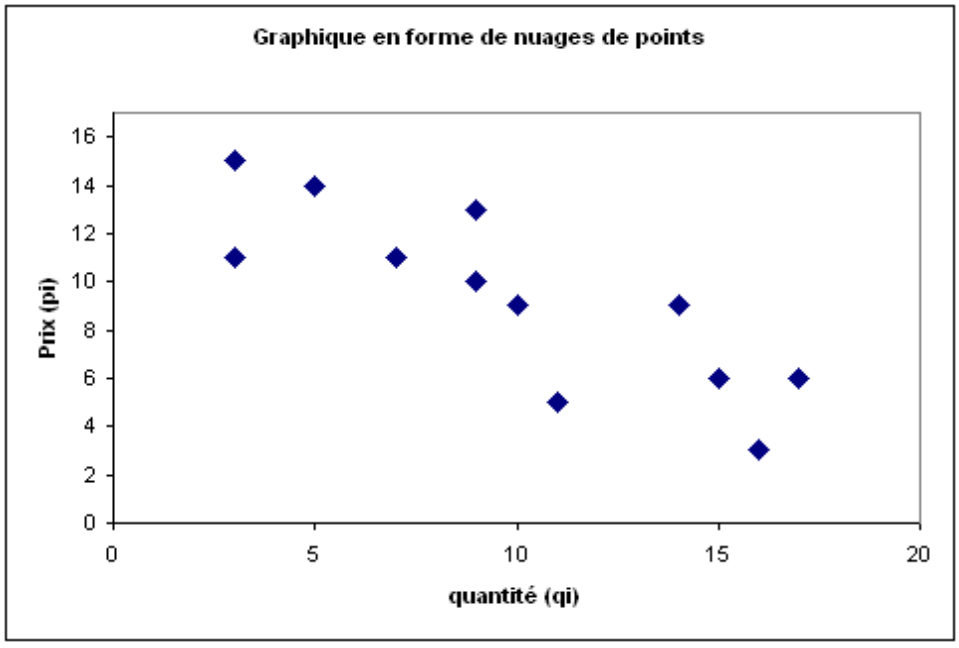
- ajouter des observations et pour ce faire
- changer de cadre temporel (nous allons passer de l'année au mois).

Le tableau ci-après donne les ventes mensuelles et le prix unitaire correspondant. Ces valeurs représentent les observations qui vont servir à l'estimation de la fonction de demande. Ces chiffres, sont inventés pour l'occasion et il serait sans doute plus

difficile d'obtenir une relation aussi évidente avec des chiffres réels. Cependant, ce qui nous intéresse principalement ici, c'est d'illustrer la méthode.

Dates	Quantité q_i (milliers d'unités)	Prix p_i (euros)
Janvier	5	14
Février	15	6
Mars	9	10
Avril	14	9
Mai	3	11
Juin	9	13
Juillet	10	9
Août	17	6
Septembre	11	5
Octobre	16	3
Novembre	7	11
Décembre	3	15

Le graphique en nuage de points (graphique de dispersion) des observations laisse présupposer l'existence d'une relation décroissante. Pour que notre graphique soit conforme à la présentation traditionnelle du diagramme de demande, nous avons mis le prix en ordonnée et la quantité en abscisse. *Toutefois, il faut garder présent à l'esprit le fait qu'au niveau de la causalité économique, c'est la quantité demandée qui est fonction du prix et non l'inverse*.



Appliquons maintenant la méthode des moindres carrés ordinaires introduite pour l'analyse du trend. Cette fois, nous devons estimer les coefficients a et b d'une équation de la forme $p = a * q+b$. Par rapport à la formule du trend temporel, il n'y a que les symboles qui changent

Nous nous attendons ici à ce que le coefficient *a* soit négatif. Les principaux calculs nécessaires sont donnés ci-après :

Quantité q_i	Prix p_i	p_iq_i	q_i^2
5	14	70	25
15	6	90	225
9	10	90	81
14	9	126	196
3	11	33	9
9	13	117	81
10	9	90	100
17	6	102	289
11	5	55	121
16	3	48	256
7	11	77	49
3	15	45	9
119	**112**	**943**	**1441**

9,91666667	98,3402778	1180,08333	9,333333333	1110,66667
$\bar{q} = \frac{1}{12}\sum_{i=1}^{12} q_i$	$\bar{q}^{\,2}$	$n \times \bar{q}^{\,2}$	$\bar{p} = \frac{1}{12}\sum_{i=1}^{12} p_i$	$n\,\bar{q}\,\bar{p}$
$\sum_{i=1}^{12} q_i$	$\sum_{i=1}^{12} p_i$	$\sum_{i=1}^{12} q_i p_i$	$\sum_{i=1}^{12} q_i^2$	

On calcule ensuite le coefficient a :

$$a = \frac{\sum_{i=1}^{n} q_i p_i - n \cdot \bar{q} \cdot \bar{p}}{\sum_{i=1}^{n} q_i^2 - n(\bar{q})^2} = \frac{943 - 1110,66667}{1441 - 1180,08333} = \frac{-167,66667}{260,91667} = -0,6426$$

et on en déduit le coefficient b :

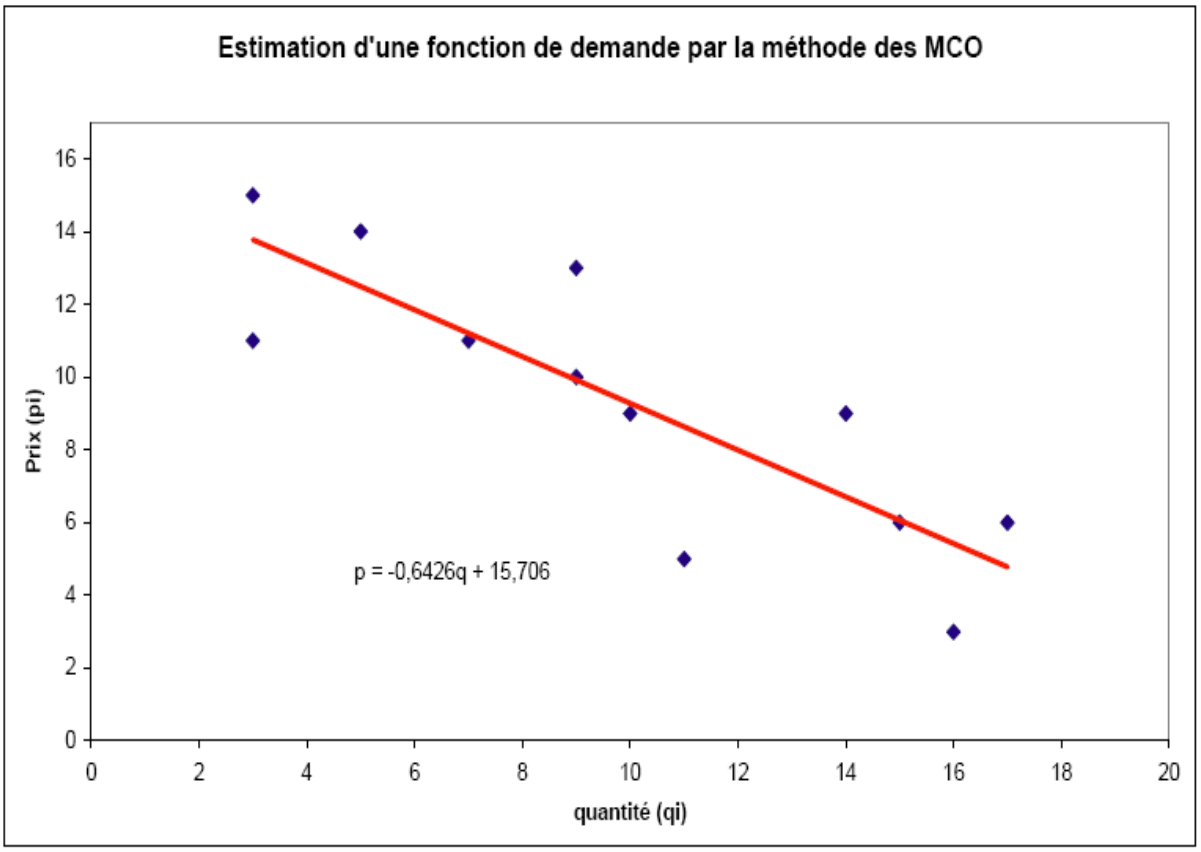
$$b = \bar{p} - a\bar{q} = 9,33333333 - (-0,642606 \times 9,91666667) = 15,706$$

On obtient donc l'équation du trend qui est :

$$p = -0,6426q + 15,706$$

Un didacticiel Microsoft est disponible ici pour l'utilisation directe de la fonction de régression sans faire soi-même les calculs : http://office.microsoft.com/fr-fr/excel/HA010877851036.aspx

On a ainsi la droite de demande décroissante comme illustré sur la figure :



On peut maintenant exprimer p en fonction de q si l'on préfère. Sachant que :

$$p = aq + b \Leftrightarrow q = \frac{1}{a}q - \frac{b}{a}$$

On a :

$$q = - 1{,}55618^{*}\ p + 24{,}4413$$

On peut ensuite se servir de la fonction de demande ainsi obtenue pour évaluer les conséquences d'une baisse du prix sur la quantité demandée et donc sur la recette totale.

C - Le coefficient de détermination

Lorsque l'on a estimé la droite de régression, on doit se demander ci cette estimation est de bonne qualité. On dispose d'un premier outil pour répondre à cette question : c'est le **coefficient de détermination** dont la formule est donnée par :

$$r^2 = \frac{\left[n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right]^2}{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}$$

Ce coefficient est compris entre 0 et 1. Plus il est proche de 1 et plus la qualité globale de la régression est bonne.

À titre d'exemple, calculons le coefficient de détermination de l'équation de régression de demande précédent. Remplaçons x par q et y par p dans la formule de r^2, afin d'avoir :

$$r^2 = \frac{\left[n\sum_{i=1}^{n} q_i p_i - \left(\sum_{i=1}^{n} q_i\right)\left(\sum_{i=1}^{n} p_i\right)\right]^2}{\left[n\sum_{i=1}^{n} q_i^2 - \left(\sum_{i=1}^{n} q_i\right)^2\right]\left[n\sum_{i=1}^{n} p_i^2 - \left(\sum_{i=1}^{n} p_i\right)^2\right]}$$

Un didacticiel Microsoft est disponible ici pour l'obtention directe du coefficient de détermination sans faire soi-même les calculs : http://office.microsoft.com/fr-fr/excel/HA010877851036.aspx

Pour faire les calculs, voici comment procéder :

	Quantité q_i	Prix p_i
1	5	14
2	15	6
3	9	10
4	14	9
5	3	11
6	9	13
7	10	9
8	17	6
9	11	5
10	16	3
11	7	11
12	3	15

La formule générale du coefficient de détermination est donnée par :

$$r^2 = \frac{\left[n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right]^2}{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}$$

En remplaçant xi par qi et yi par pi, on obtient :

$$r^2 = \frac{\left[n\sum_{i=1}^{n} q_i p_i - \left(\sum_{i=1}^{n} q_i\right)\left(\sum_{i=1}^{n} p_i\right)\right]^2}{\left[n\sum_{i=1}^{n} q_i^2 - \left(\sum_{i=1}^{n} q_i\right)^2\right]\left[n\sum_{i=1}^{n} p_i^2 - \left(\sum_{i=1}^{n} p_i\right)^2\right]}$$

Pour trouver r2, il faut d'abord calculer ses différentes composantes à l'aide du tableau ci-dessous :

	Quantité q_i	Prix p_i	$q_i \times p_i$	q_i^2	p_i^2
24					
25	5	14	70	25	196
26	15	6	90	225	36
27	9	10	90	81	100
28	14	9	126	196	81
29	3	11	33	9	121
30	9	13	117	81	169
31	10	9	90	100	81
32	17	6	102	289	36
33	11	5	55	121	25
34	16	3	48	256	9
35	7	11	77	49	121
36	3	15	45	9	225
37	**119**	**112**	**943**	**1441**	**1200**
38	↑	↑	↑	↑	↑
39					
40	$\sum_{i=1}^{n} q_i$	$\sum_{i=1}^{n} p_i$	$\sum_{i=1}^{n} q_i p_i$	$\sum_{i=1}^{n} q_i^2$	$\sum_{i=1}^{n} p_i^2$
41					
42					
43					
44	$\left(\sum_{i=1}^{n} q_i\right)^2$	$\left(\sum_{i=1}^{n} p_i\right)^2$	$n\times\sum_{i=1}^{n} q_i p_i$	$n\times\sum_{i=1}^{n} q_i^2$	$n\times\sum_{i=1}^{n} p_i^2$
45					
46					
47			↓	↓	↓
48	**14161**	**12544**	**11316**	**17292**	**14400**
49					
50	$\left(\sum_{i=1}^{n} q_i\right)\left(\sum_{i=1}^{n} p_i\right)$				
51					
52					
53	↓				
54					
55	**13328**				

En remplaçant dans la formule on obtient :

$$r^2 = \frac{\left[n\sum_{i=1}^{n} q_i p_i - \left(\sum_{i=1}^{n} q_i\right)\left(\sum_{i=1}^{n} p_i\right)\right]^2}{\left[n\sum_{i=1}^{n} q_i^2 - \left(\sum_{i=1}^{n} q_i\right)^2\right]\left[n\sum_{i=1}^{n} p_i^2 - \left(\sum_{i=1}^{n} p_i\right)^2\right]}$$

$$r^2 = \frac{\left[11316 - 13328\right]^2}{\left[17292 - 14161\right]\left[14400 - 12544\right]}$$

$$r^2 = \frac{4048144}{3131\times 1856} = \frac{4048144}{5811136} = 0,6966184$$

On peut éviter de faire les calculs ci-dessus en appliquant la fonction EXCEL :

=COEFFICIENT.DETERMINATION(A28:A39;B28:B39) → 0,6966184

Interprétation du résultat : Plus le coefficient r^2 tend vers 1, plus la qualité globale de la régression est bonne. Ici, le r^2 est proche de 0,7. **On peut juger que c'est insuffisant**. Il faut de toute manière compléter ce premier diagnostic par le calcul d'autres statistiques, mais ceci est l'objet d'un cours **d'économétrie** et non plus de statistique descriptive.

4 – Le test du Khi-carré

A - Introduction

L'expression khi-carré a deux significations en statistique :

- **L'expression khi-carré** est le nom du test statistique qui sert à apprécier l'existence ou non d'une relation entre deux dimensions au sein d'une population, lorsque ces dimensions sont mesurées sur des échelles qualitatives et que les modalités de ces échelles de mesure ont été regroupées en catégories. On peut bien sûr utiliser aussi le test du khi carré pour apprécier l'existence d'une relation entre deux dimensions mesurées sur des échelles quantitatives groupées en catégories. On peut enfin (et sans doute le plus souvent), l'utiliser pour étudier la relation entre une dimension quantitative et une autre qualitative.

- **L'expression khi carré** est aussi le nom d'une distribution théorique assez complexe, dont les valeurs servent à valider le test du chi-carré. Ces résultats sont présentés sous forme d'une table. De ce fait, en pratique, il suffit de se reporter à la table.

En résumé, l'expression khi-carré désigne à la fois le test et la distribution qui sert à valider le test.

B - Exemple d'utilisation

Nous allons maintenant montrer comment ce test peut-être utilisé en présence de deux dimensions dont l'une (l'âge) est mesurée sur une échelle quantitative mais où les âges ont été regroupés en 2 catégories (les moins de 20 ans et les plus de 20 ans) et l'autre (le produit) est mesurée sur une échelle qualitative « nominale » : quatre produits A, B, C et D.

La population de départ est constituée de 100 unités statistiques, ici 100 consommateurs à qui on a demandé :

- leur âge
- « P*armi ces 4 produits, quels est celui que vous avez acheté* ». Naturellement, seuls les consommateurs ayant acheté l'un des 4 produits en question ont été sélectionnés pour être interrogés.

On cherche à savoir si l'âge a une influence sur le type de produit acheté. Or ici, il y a deux dimensions dont l'une est quantitative (l'âge) et l'autre est qualitative (le produit acheté). On ne peut donc pas utiliser la méthode des moindres carrés. On peut en revanche utiliser un test du khi-carré.

Pour cela, les données sur l'âge ont été regroupées en 2 catégories : « jeunes » et « moins jeunes », pour savoir si le produit est également acheté par les plus de 20 ans par exemple. On va donc regrouper la population en deux catégories : les moins de 25 ans et ceux de 25 ans et plus (ici, cela se résume à la catégorie 20-45 ans). En ce qui concerne les catégories de produits, nous allons en revanche conserver les 4 catégories A, B, C et D. Les 100 personnes sont donc regroupées dans le tableau à deux dimensions ci-dessous :

	Moins de 25 ans	25 à 45 inclus	Total
A	10	15	25
B	10	25	35
C	15	5	20
D	20	0	20
Total	55	45	100

Le tableau se lit ainsi : 10 personnes de moins de 25 ans ont acheté le produit A, 15 personnes de 25 ans à 45 ans ont acheté le produit A, 25 en tout ont acheté le produit A. Et ainsi de suite pour les produits B, C et D.

On remarque avant même de procéder à un test que l'âge semble exercer une influence. **On remarque tout de suite que les produits A et B sont davantage achetés par les 25 à 45 ans et que les produits C et D sont en revanche davantage achetés par les moins de 25 ans**.

Le test du khi carré va apporter une information supplémentaire. Il va permettre de dire si les différences de comportement d'achat qui sont attribuées à l'âge sont le fait du hasard ou si elles sont réelles. Elles peuvent en effet être dues au hasard de l'échantillon. **Ce que le test va nous dire c'est dans quelle mesure la différence est indépendante de l'échantillon choisi**.

Pour effectuer le test, on part de la constatation suivante. Si l'âge n'a aucune influence sur le choix, les colonnes « moins de 25 ans » et « 25 à 45 ans » doivent être proportionnelles à la colonne « total ». Pour le vérifier, il suffit de **calculer deux colonnes fictives, dans lesquelles les individus choisissent les produits dans la même proportion quel que soit leur âge**.

Ainsi, les moins de 25 ans représentent 55% des effectifs totaux. Donc, si l'âge n'a pas d'influence sur le choix du produit, les choix de chaque produit A, B, C et D par les moins de 25 ans devraient correspondre à 55% de l'effectif total de chaque produit. Par exemple, il y a 25 personnes qui ont acheté le produit A. Si l'âge n'a pas d'influence sur le choix, le nombre de jeunes devrait être égal à 55% de 25, soit 13,75. Il s'agit évidemment d'un effectif « théorique » car les personnes ne peuvent pas être « coupées en morceaux ». En appliquant ce principe à toutes les données du tableau, on obtient :

	Moins de 25 ans (ni)	Effectifs théoriques (moins de 25 ans) (N_i)	25 à 45 inclus (ni)	Effectifs théoriques (25 à 45 ans inclus) (Ni)	Total
A	10	0,55 x 25 = 13,75	15	0,45 x 25 = 11,25	25
B	10	0,55 x 35 = 19,25	25	0,45 x 35 = 15,75	35
C	15	0,55 x 20 = 11	5	0,45 x 20 = 9	20
D	20	0,55 x 20 = 11	0	0,45 x 20 = 9	20
Total	55		45		100

On voit dès à présent que les effectifs réels ne sont pas égaux aux effectifs « attendus » ou « théoriques ». **Il se peut donc que l'âge ait une influence sur le choix du produit. Mais dans quelle mesure ?**

Il faut maintenant calculer les différences entre les effectifs réels et les effectifs attendus ou théoriques, puis porter cette différence au carré et la diviser par l'effectif théorique, et enfin additionner les huit chiffres ainsi obtenus.

En d'autres termes, il faut appliquer la formule suivante :

$$\chi^2\left(\text{calculé}\right) = \sum_{i=1}^{n} \frac{\left(N_i - n_i\right)^2}{N_i}$$

Pour faciliter les calculs, disposons-les dans le tableau suivant :

		Effectifs théoriques N_i	Effectifs réels n_i	$N_i - n_i$	$(N_i - n_i)^2$	$\frac{(N_i - n_i)^2}{N_i}$
Moins de 25 ans	A	13,75	10	3,75	14,0625	1,023
	B	19,25	10	9,25	85,5625	4,445
	C	11	15	-4	16	1,455
	D	11	20	-9	81	7,364
25 à 45 ans	A	11,25	15	-3,75	14,0625	1,250
	B	15,75	25	-9,25	85,5625	5,433
	C	9	5	4	16	1,778
	D	9	0	9	81	9,000
						31,746

Une fois que l'on connaît le khi carré calculé, on doit le comparer avec la valeur du khi-deux issue de la distribution du khi carré (voir le tableau ci-après).

Pour trouver cette valeur dans le tableau, nous devons prendre en compte deux informations supplémentaires :

- Le nombre de « degrés de liberté » qui se calcule ainsi :

Degrés de liberté = nombre d'observations - nombre de dimensions

Ici, il y a 8 observations (les huit chiffres du tableau) et deux dimensions (l'âge et le le produit choisi), ce qui donne un degré de liberté égal à 8-2 = 6.

- Ensuite, nous devons choisir la probabilité de fiabilité du test : 5% de chances de se tromper, 1% ou 1 pour 1000. Nous allons choisir 5%, soit P = 0,05.

Nous avons donc 6 degrés de liberté et une probabilité de fiabilité du test de P=0,05. Par conséquent, nous voyons dans la table que :

$$\chi^2_{0,05} = 12,59$$

Il nous reste maintenant à comparer le chi-carré calculé et le khi carré théorique, issu de la table :

$$\chi^2_{0,05} = 12,59 < \chi^2_{\text{calculé}} = 31,74$$

Degrés de liberté	P=0,05	P=0,01	P=0,001	Degrés de liberté	P=0,05	P=0,01	P=0,001
1	3.84	6.64	10.83	50	67.51	76.15	86.66
2	5.99	9.21	13.82	51	68.67	77.39	87.97
3	7.82	11.35	16.27	52	69.83	78.62	89.27
4	9.49	13.28	18.47	53	70.99	79.84	90.57
5	11.07	15.09	20.52	54	72.15	81.07	91.88
6	12.59	16.81	22.46	55	73.31	82.29	93.17
7	14.07	18.48	24.32	56	74.47	83.52	94.47
8	15.51	20.09	26.13	57	75.62	84.73	95.75
9	16.92	21.67	27.88	58	76.78	85.95	97.03
10	18.31	23.21	29.59	59	77.93	87.17	98.34
11	19.68	24.73	31.26	60	79.08	88.38	99.62
12	21.03	26.22	32.91	61	80.23	89.59	100.88
13	22.36	27.69	34.53	62	81.38	90.80	102.15
14	23.69	29.14	36.12	63	82.53	92.01	103.46
15	25.00	30.58	37.70	64	83.68	93.22	104.72
16	26.30	32.00	39.25	65	84.82	94.42	105.97
17	27.59	33.41	40.79	66	85.97	95.63	107.26
18	28.87	34.81	42.31	67	87.11	96.83	108.54
19	30.14	36.19	43.82	68	88.25	98.03	109.79
20	31.41	37.57	45.32	69	89.39	99.23	111.06
21	32.67	38.93	46.80	70	90.53	100.42	112.31
22	33.92	40.29	48.27	71	91.67	101.62	113.56
23	35.17	41.64	49.73	72	92.81	102.82	114.84
24	36.42	42.98	51.18	73	93.95	104.01	116.08
25	37.65	44.31	52.62	74	95.08	105.20	117.35
26	38.89	45.64	54.05	75	96.22	106.39	118.60
27	40.11	46.96	55.48	76	97.35	107.58	119.85
28	41.34	48.28	56.89	77	98.49	108.77	121.11
29	42.56	49.59	58.30	78	99.62	109.96	122.36
30	43.77	50.89	59.70	79	100.75	111.15	123.60
31	44.99	52.19	61.10	80	101.88	112.33	124.84
32	46.19	53.49	62.49	81	103.01	113.51	126.09
33	47.40	54.78	63.87	82	104.14	114.70	127.33
34	48.60	56.06	65.25	83	105.27	115.88	128.57
35	49.80	57.34	66.62	84	106.40	117.06	129.80
36	51.00	58.62	67.99	85	107.52	118.24	131.04
37	52.19	59.89	69.35	86	108.65	119.41	132.28
38	53.38	61.16	70.71	87	109.77	120.59	133.51
39	54.57	62.43	72.06	88	110.90	121.77	134.74
40	55.76	63.69	73.41	89	112.02	122.94	135.96
41	56.94	64.95	74.75	90	113.15	124.12	137.19
42	58.12	66.21	76.09	91	114.27	125.29	138.45
43	59.30	67.46	77.42	92	115.39	126.46	139.66
44	60.48	68.71	78.75	93	116.51	127.63	140.90
45	61.66	69.96	80.08	94	117.63	128.80	142.12
46	62.83	71.20	81.40	95	118.75	129.97	143.32
47	64.00	72.44	82.72	96	119.87	131.14	144.55
48	65.17	73.68	84.03	97	120.99	132.31	145.78
49	66.34	74.92	85.35	98	122.11	133.47	146.99
50	67.51	76.15	86.66	99	123.23	134.64	148.21
				100	124.34	135.81	149.48

Source de la table : http://www.ento.vt.edu/~sharov/PopEcol/tables/chisq.html

Etant donné que le chi-carré théorique est inférieur au khi carré calculé, nous pouvons conclure que la répartition des préférences est suffisamment différente d'une répartition homogène pour qu'on puisse raisonnablement se fier à l'idée que **l'âge a une influence sur le choix du produit.** Notre observation initiale sur la base de l'échantillon est donc probablement vraie à l'extérieur de l'échantillon (avec 5% de chances de nous tromper). Si l'on croît à ce type de test, on pourrait donc cibler des publicités pour les jeunes pour mieux vendre les produits C et D et cibler des publicités pour les moins jeunes pour mieux vendre les produits A et B. Il devrait alors en résulter une augmentation des ventes à la fois pour les produits destinés aux jeunes (C et D) et pour les produits destinés aux moins jeunes (A et B). Bien sûr, le test n'est pas fiable à 100%. Il y a toujours 5% de chances de se tromper. Mais l'on peut juger que cette marge d'erreur est suffisamment faible pour décider d'une politique de marketing en fonction des résultats. On peut aussi, choisir de refaire le test en choisissant P =0,001. Dans ce cas, il n'y a plus qu'une chance sur 1000 de se tromper et la valeur du khi-carré théorique passe à 22,46, valeur qui reste inférieure au chi-carré calculé et qui aboutit donc à la même conclusion, avec une fiabilité renforcée.

Si en revanche le chi-carré théorique avait été supérieur au khi carré calculé, alors nous n'aurions pas pu conclure avec suffisamment de certitude que, indépendamment de notre échantillon, **l'âge a une influence sur le choix du produit**

Chapitre 7
Courbe de LORENZ et coefficient de GINI

1 – Introduction
2 – La courbe de LORENZ
A – L'exemple de la répartition des PIB au sein de l'UE à 27
B – L'utilité de la courbe de LORENZ pour les comparaisons
C – Cas Général
3 – Le coefficient de GINI
A – Définition
B – Formules de calcul
C - Exemple

1 – Introduction

Max Otto LORENZ (1880 -1962) est l'économiste américain qui inventa le concept de courbe de Lorenz en 1905. Il s'agissait pour lui de décrire et de mesurer les inégalités de revenu. Par la suite, cette courbe qu'il fut le premier à utiliser servit plus généralement à mesurer la façon dont se répartit une masse (salariale, de revenus, de richesses, etc.) au sein d'un population pour se faire et une idée du caractère plus ou moins égalitaire de la répartition de ces masses au sein de la population et, partant de comparer différentes populations entre elles ou de comparer la distribution d'une masse au sein d'une population en deux ou plusieurs points du temps afin de savoir si l'inégalité augmente ou diminue.

Corrado GINI (1884 -1965) est le statisticien, démographe, ethnologue, sociologue et idéologue italien à qui on doit le coefficient de GINI une mesure de l'inégalité associé à la la courbe de LORENZ.

En pratique, lorsqu'on s'intéresse à la répartition d'une masse de revenus ou de richesse au sein d'une population, on trace d'abord une courbe de LORENZ afin d'avoir une idée visuelle de l'égalité ou de l'inégalité de cette répartition. Ensuite, si l'on désire résumer cette inégalité par un chiffre, on calcule le coefficient de GINI.

2 – La courbe de LORENZ

A – L'exemple de la répartition des PIB au sein de l'UE à 27

Pour introduire la courbe de LORENZ, prenons l'exemple de la répartition des PIB au sein de l'UE à 27. Ci-après, un extrait du tableau 1, où les 27 pays de l'UE ont été classés par ordre de PIB décroissant (colonne 2).

La colonne 3 est simplement un cumul des pays de 0 à 27. La colonne 4 est un cumul des PIB des pays.

La colonne 5 reprend les chiffres de la colonne 3 divisés par 27 (nombre total des pays) et multiplié par 100.

La colonne 6 reprend les chiffres de la colonne 4 divisés par 13847 (PIB total de l'UE à 27) et multiplié par 100.

Les chiffres des colonnes 5 et 6 nous permettent d'évaluer l'ampleur de l'inégalité de la production de richesse au sein des pays de l'UE à 27

Tableau pour la construction d'une courbe de LORENZ de la répartition des PIB au sein de l'UE à 27

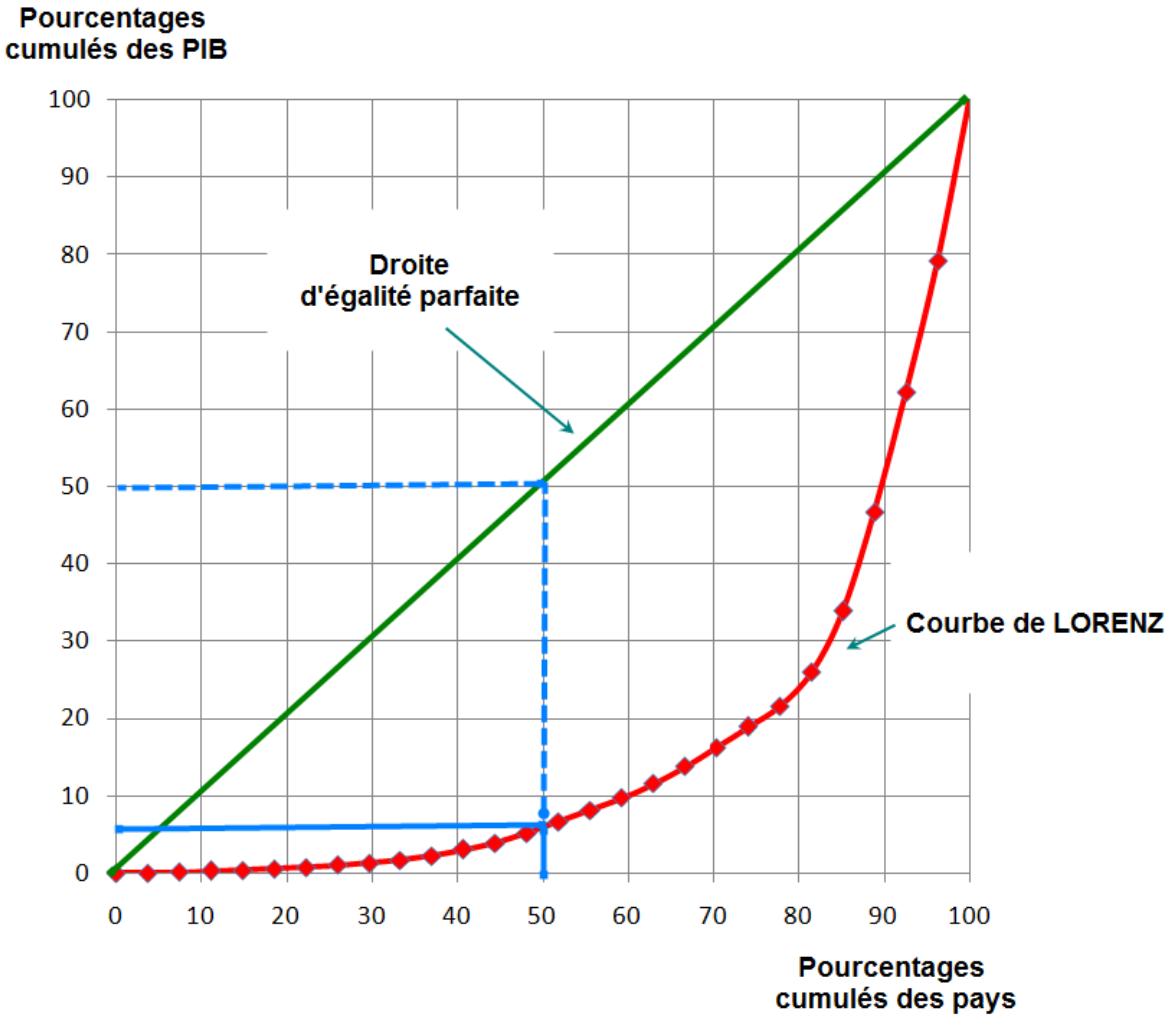
Pays	PIB en milliards d'USD (2006)	Cumul des pays	Cumul des PIB	Cumul des pays (en %)	Cumul des PIB (%)
		0	0	0,0	0,00
Malte	6	1	6	3,7	0,04
Estonie	14	2	19	7,4	0,14
Chypre	16	3	36	11,1	0,26
Lettonie	17	4	52	14,8	0,38
Bulgarie	28	5	80	18,5	0,58
Lituanie	30	6	110	22,2	0,80
Luxembourg	35	7	145	25,9	1,05
Slovénie	38	8	183	29,6	1,32
Slovaquie	48	9	231	33,3	1,67
Roumanie	80	10	311	37,0	2,24
Hongrie	113	11	424	40,7	3,06
République tchèque	119	12	543	44,4	3,92
Portugal	177	13	720	48,1	5,20
Finlande	198	14	918	51,9	6,63
Irlande	204	15	1122	55,6	8,10
Grèce	224	16	1346	59,3	9,72
Danemark	258	17	1604	63,0	11,58
Autriche	310	18	1914	66,7	13,82
Pologne	337	19	2251	70,4	16,26
Belgique	370	20	2621	74,1	18,92
Suède	373	21	2993	77,8	21,62
Pays-Bas	613	22	3606	81,5	26,04
Espagne	1 084	23	4690	85,2	33,87
Italie	1 785	24	6475	88,9	46,76
France	2 151	25	8626	92,6	62,30
Royaume-Uni	2 346	26	10972	96,3	79,24
Allemagne	2 875	27	13847	100,0	100,00

Le graphique ci-après représente une courbe de LORENZ. En abscisse, c'est le pourcentage cumulé de la population qui est mesuré (ici c'est le pourcentage cumulé des 27 pays). En ordonnée, c'est le pourcentage cumulé des PIB qui est mesuré.

La courbe de Lorenz s'inscrit donc dans un carré. Pour apprécier l'inégalité, on doit comparer cette courbe (en rouge sur le graphique) avec la droite d'égalité parfaite qui correspond à la diagonale (droite en vert).

Si les PIB étaient parfaitement distribués 50% des pays produiraient 50% du PIB total de l'UE à 27. Or, c'est loin d'être la cas puisque les 50% pays les plus pauvres en termes de PIB total ne produisent que péniblement environ 7% du PIB total de l'UE à 27. Pour atteindre 50% de la production de l'UE, il faut mettre à contribution 90% des pays ! A elle-seule, l'Allemagne créé plus de 20% de la richesse annuelle de l'UE à 27.

Courbe de LORENZ de la répartition des PIB de l'UE à 27 en 2006



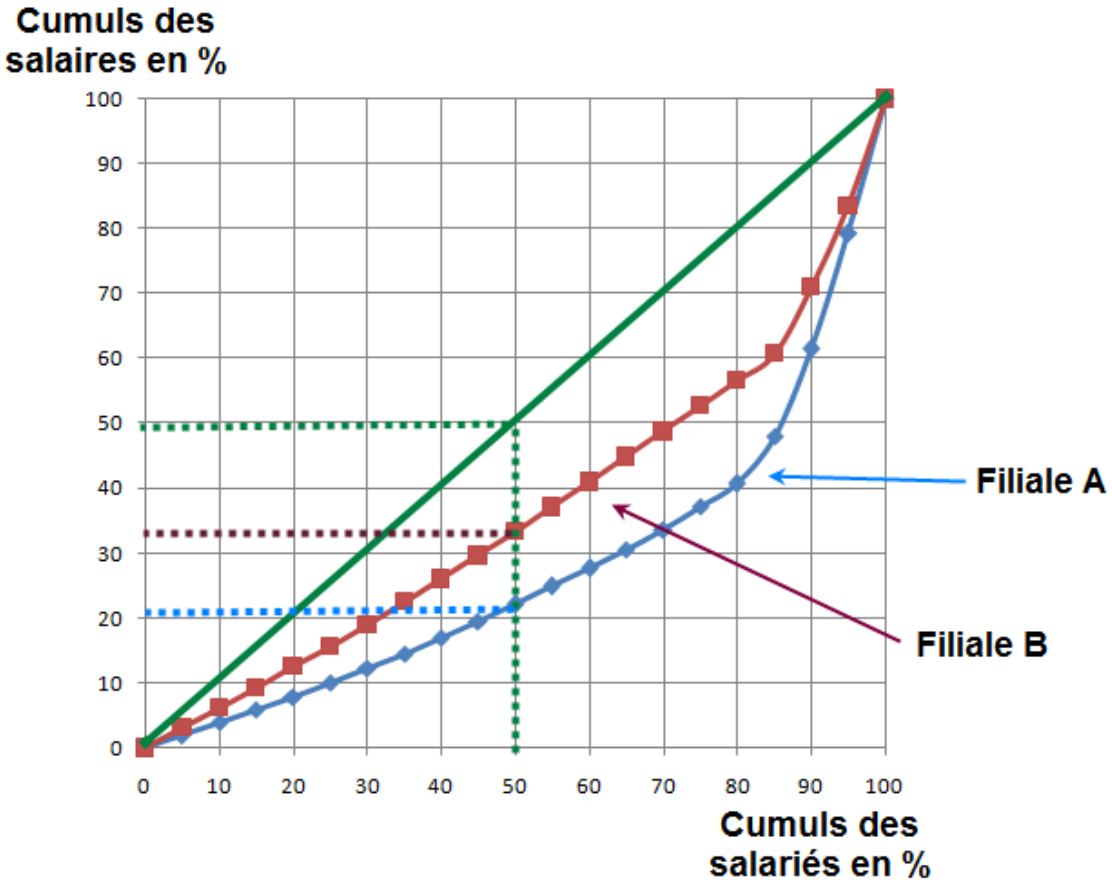
B – L'utilité de la courbe de LORENZ pour les comparaisons

Soit maintenant un autre exemple concernant cette fois la masse salariale des deux filiales A et B d'une entreprise et sa répartition entre les salariés. Pour simplifier, nous supposons qu'il y a 20 salariés dans chaque entreprise. Les salaires mensuels en euros sont donnés par le tableau ci-après qui détaille également les calculs des deux séries nécessaires au tracé de la courbe de LORENZ.

Comparaisons de la distribution des salaires dans les filiales A et B d'une entreprise XXX et calculs nécessaires pour le tracé de la courbe de LORENZ

Salaires mensuels nets (€)				Cumuls salaires		Cumuls salaires (%)	
Filiale A	Filiale B	Cumuls salariés	Cumuls salariés (%)	Filiale A	Filiale B	Filiale A	Filiale B
		0	0	0	0	0	0
1300	1500	1	5	1300	1500	2	3
1350	1500	2	10	2650	3000	4	6
1350	1500	3	15	4000	4500	6	9
1350	1550	4	20	5350	6050	8	12
1500	1550	5	25	6850	7600	10	16
1500	1600	6	30	8350	9200	12	19
1500	1700	7	35	9850	10900	15	23
1600	1700	8	40	11450	12600	17	26
1700	1740	9	45	13150	14340	19	30
1800	1800	10	50	14950	16140	22	33
1820	1840	11	55	16770	17980	25	37
1900	1850	12	60	18670	19830	28	41
2000	1870	13	65	20670	21700	31	45
2000	1900	14	70	22670	23600	34	49
2400	1920	15	75	25070	25520	37	53
2400	1940	16	80	27470	27460	41	57
5000	1960	17	85	32470	29420	48	61
9000	5000	18	90	41470	34420	61	71
12000	6000	19	95	53470	40420	79	83
14000	8000	20	100	67470	48420	100	100
			Abcisse de la courbe de LORENZ			**A**	**B**
						Ordonnées de la courbe de LORENZ	

Courbes de LORENZ des salaires des filiales A et B



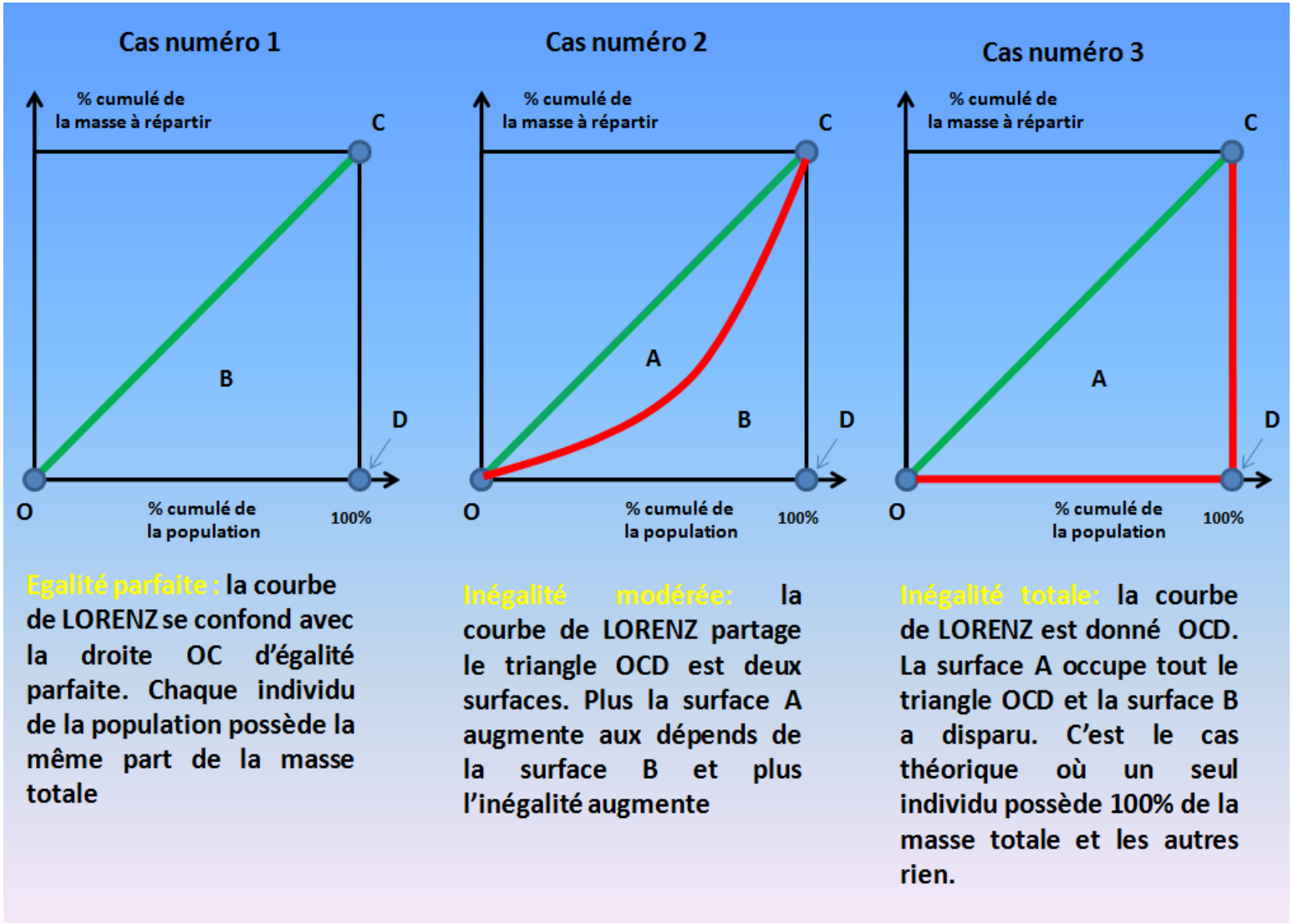
Sur le graphique ci-dessus, on constate que c'est la courbe de LORENZ de la filiale B qui est la plus proche de la droite d'égalité parfaite. C'est donc dans la filiale B que l'inégalité de la répartition des salaires est la moins grande. Par exemple, on voit que dans la filiale A, 50% des salariés reçoivent seulement 22% de la masse salariale, alors que dans la filiale B, 50 % des salariés reçoivent 33% de la masse salariale (voir les chiffres encadrés dans le tableau précédent). Si la distribution était parfaitement égalitaire, 50% des salariés devraient recevoir 50% de la masse salariale.

C – Cas général

De façon générale, plus une courbe de LORENZ se rapproche de la droite d'égalité parfaite et plus la répartition de la masse considérée au sein de la population est égalitaire. En effet, dans ce cas, la masse (des salaires, de la richesse, du revenu, etc.) est peu concentrée sur quelques uns.

Inversement, plus une courbe de LORENZ s'éloigne de la droite d'égalité parfaite et plus la répartition de la masse considérée au sein de la population est inégalitaire car la masse (des salaires, de la richesse, du revenu) est alors concentrée sur un petit nombre d'unités statistiques.

3 cas typiques, dont les deux cas limites, sont représentés par les graphiques ci-dessous



Egalité parfaite : la courbe de LORENZ se confond avec la droite OC d'égalité parfaite. Chaque individu de la population possède la même part de la masse totale

Inégalité modérée: la courbe de LORENZ partage le triangle OCD est deux surfaces. Plus la surface A augmente aux dépends de la surface B et plus l'inégalité augmente

Inégalité totale: la courbe de LORENZ est donné OCD. La surface A occupe tout le triangle OCD et la surface B a disparu. C'est le cas théorique où un seul individu possède 100% de la masse totale et les autres rien.

3 – Le coefficient de GINI

A – Définition

Le coefficient de Corrado GINI (1884 -1965) est une mesure de l'inégalité associé à la courbe de LORENZ. Il est donné par la formule :

$$G = \frac{A}{A+B}$$

Où A représente la surface comprise entre la courbe de LORENZ et la droite d'égalité parfaite et B représente la surface située sous la droite d'égalité parfaite *moins* la surface A.

Le coefficient de GINI est compris entre zéro et 1. En cas d'égalité parfaite, il est égal à zéro (car A=0). En cas d'inégalité totale il est égal à 1, car B=0. Par conséquent, à mesure que G augmente de zéro à 1, l'inégalité de la répartition augmente. Le coefficient de GINI permet ainsi de faire de nombreuses comparaisons.

Sachant que a courbe de LORENZ est inscrite dans un carré de 1 x 1, on voit que la surface A+B est égale à la moitié de cette surface. On a donc :

$$A+B=\frac{1}{2}$$

On peut de ce fait écrire :

$$G=\frac{A}{A+B}=\frac{A}{\frac{1}{2}}=2A$$

De plus, comme :

$$A+B=\frac{1}{2}\Rightarrow A=\frac{1}{2}-B$$

De ce fait on peut écrire que :

$$G = \frac{A}{A+B} = 2A = 2\left(\frac{1}{2} - B\right) = 1 - 2B$$

B – Formules de calcul

Il existe plusieurs formules de calcul du coefficient de GINI. Nous adopterons l'une des plus utilisées qui est donnée dans la notice explicative de la Banque mondiale pour le calcul de l'inégalité des répartitions[16], parfois dite « formule de BROWN ». Cette formule s'écrit :

$$G = 1 - \sum_{i=1}^{n} \left(X_i - X_{i-1}\right)\left(Y_i + Y_{i+1}\right)$$

où X est la part cumulée de la population, et Y la part cumulée de la masse à répartir. Dans le cas qui nous intéresse ici, comme les données sont connues individuellement cette formule peut se simplifier à :

$$G = 1 - \frac{1}{n} \sum_{i=1}^{n} \left(Y_i + Y_{i+1}\right)$$

Où n représente le nombre d'unités statistiques (la population).

Nous allons voir que ces deux formules donnent des résultats identiques.

C – Exemple

Reprenons l'exemple des deux filiales de la même entreprise et calculons d'abord les deux coefficients de GINI à l'aide la première formule, soit :

$$G = 1 - \sum_{i=1}^{n} \left(X_i - X_{i-1}\right)\left(Y_i + Y_{i+1}\right)$$

Les 2 tableaux ci-après montrent comment les calculs doivent être disposés pour parvenir rapidement au résultat.

[16] Voir http://info.worldbank.org/etools/docs/library/103072/ch6.pdf , page 3

Calcul du coefficient de GINI de la filiale A selon la formule :

$$G = 1 - \sum_{i=1}^{n} (X_i - X_{i-1})(Y_i + Y_{i+1})$$

A	B	C	D	E	F	G	H	I
Salaires mensuels filiale A	Fréquences salaires mensuels	Fréquences cumulées salaires mensuels (Yi)	$(Y_i + Y_{i+1})$	Nombre de salariés	Fréquence des salariés	Fréquence cumulée des salariés (X_i)	$(X_i - X_{i-1})$	$(X_i - X_{i-1})(Y_i + Y_{i+1})$
1300	0,019	0,019	0,019	1	0,05	0,05	0,05	0,00096
1350	0,020	0,039	0,059	1	0,05	0,1	0,05	0,00293
1350	0,020	0,059	0,099	1	0,05	0,15	0,05	0,00493
1350	0,020	0,079	0,139	1	0,05	0,2	0,05	0,00693
1500	0,022	0,102	0,181	1	0,05	0,25	0,05	0,00904
1500	0,022	0,124	0,225	1	0,05	0,3	0,05	0,01126
1500	0,022	0,146	0,270	1	0,05	0,35	0,05	0,01349
1600	0,024	0,170	0,316	1	0,05	0,4	0,05	0,01578
1700	0,025	0,195	0,365	1	0,05	0,45	0,05	0,01823
1800	0,027	0,222	0,416	1	0,05	0,5	0,05	0,02082
1820	0,027	0,249	0,470	1	0,05	0,55	0,05	0,02351
1900	0,028	0,277	0,525	1	0,05	0,6	0,05	0,02626
2000	0,030	0,306	0,583	1	0,05	0,65	0,05	0,02915
2000	0,030	0,336	0,642	1	0,05	0,7	0,05	0,03212
2400	0,036	0,372	0,708	1	0,05	0,75	0,05	0,03538
2400	0,036	0,407	0,779	1	0,05	0,8	0,05	0,03894
5000	0,074	0,481	0,888	1	0,05	0,85	0,05	0,04442
9000	0,133	0,615	1,096	1	0,05	0,9	0,05	0,05479
12000	0,178	0,793	1,407	1	0,05	0,95	0,05	0,07036
14000	0,207	1,000	1,793	1	0,05	1	0,05	0,08963
67470				**20**				**0,54893**
								0,45107

$$G = 1 - \sum_{i=1}^{20} (X_i - X_{i-1})(Y_i + Y_{i+1}) = 1 - 0,54893 = 0,45107$$

Calcul du coefficient de GINI de la filiale B selon la formule :

$$G = 1 - \sum_{i=1}^{n} (X_i - X_{i-1})(Y_i + Y_{i+1})$$

A	B	C	D	E	F	G	H	I
Salaires mensuels filiale B	Fréquences salaires mensuels	Fréquences cumulées salaires mensuels (Yi)	$(Y_i + Y_{i+1})$	Nombre de salariés	Fréquence des salariés	Fréquence cumulée des salariés (X_i)	$(X_i - X_{i-1})$	$(X_i - X_{i-1})(Y_i + Y_{i+1})$
1500	0,031	0,031	0,031	1	0,05	0,05	0,05	0,00155
1500	0,031	0,062	0,093	1	0,05	0,1	0,05	0,00465
1500	0,031	0,093	0,155	1	0,05	0,15	0,05	0,00774
1550	0,032	0,125	0,218	1	0,05	0,2	0,05	0,01089
1550	0,032	0,157	0,282	1	0,05	0,25	0,05	0,01410
1600	0,033	0,190	0,347	1	0,05	0,3	0,05	0,01735
1700	0,035	0,225	0,415	1	0,05	0,35	0,05	0,02076
1700	0,035	0,260	0,485	1	0,05	0,4	0,05	0,02427
1740	0,036	0,296	0,556	1	0,05	0,45	0,05	0,02782
1800	0,037	0,333	0,629	1	0,05	0,5	0,05	0,03147
1840	0,038	0,371	0,705	1	0,05	0,55	0,05	0,03523
1850	0,038	0,410	0,781	1	0,05	0,6	0,05	0,03904
1870	0,039	0,448	0,858	1	0,05	0,65	0,05	0,04289
1900	0,039	0,487	0,936	1	0,05	0,7	0,05	0,04678
1920	0,040	0,527	1,014	1	0,05	0,75	0,05	0,05072
1940	0,040	0,567	1,094	1	0,05	0,8	0,05	0,05471
1960	0,040	0,608	1,175	1	0,05	0,85	0,05	0,05874
5000	0,103	0,711	1,318	1	0,05	0,9	0,05	0,06592
6000	0,124	0,835	1,546	1	0,05	0,95	0,05	0,07728
8000	0,165	1,000	1,835	1	0,05	1	0,05	0,09174
48420				**20**				**0,72365**
								0,27635

$$G = 1 - \sum_{i=1}^{20} (X_i - X_{i-1})(Y_i + Y_{i+1}) = 1 - 0,72365 = 0,27635$$

On constate que le coefficient de GINI de la filiale A est beaucoup plus élevé que celui de la filiale B, indiquant que la distribution de la masse salariale y est plus inégalitaire. En effet, on a :

Coefficient de GINI de la filiale A = 0,45107
Coefficient de GINI de la filiale B = 0,27635

Voyons maintenant comment disposer les calculs en appliquant la deuxième formule (qui est plus simple et plus rapide tout en donnant les mêmes résultats)

Calcul du coefficient de GINI de la filiale A selon la formule :

$$G = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(Y_i + Y_{i+1}\right)$$

A	B	C	D	E
Salaires mensuels filiale A	Fréquences salaires mensuels	Fréquences cumulées salaires mensuels (Yi)	$(Y_i + Y_{i+1})$	
1300	0,019	0,01927	0,01927	
1350	0,020	0,03928	0,05854	
1350	0,020	0,05929	0,09856	
1350	0,020	0,07929	0,13858	
1500	0,022	0,10153	0,18082	
1500	0,022	0,12376	0,22529	
1500	0,022	0,14599	0,26975	
1600	0,024	0,16971	0,31570	
1700	0,025	0,19490	0,36461	
1800	0,027	0,22158	0,41648	
1820	0,027	0,24855	0,47013	
1900	0,028	0,27672	0,52527	
2000	0,030	0,30636	0,58307	
2000	0,030	0,33600	0,64236	
2400	0,036	0,37157	0,70757	
2400	0,036	0,40714	0,77872	
5000	0,074	0,48125	0,88839	
9000	0,133	0,61464	1,09589	
12000	0,178	0,79250	1,40714	
14000	0,207	1,00000	1,79250	
67470			10,97866	
			0,54893	
			0,45107	

$$G = 1 - \frac{1}{20}\sum_{i=1}^{20}\left(Y_i + Y_{i+1}\right) = 1 - 0,54893 = 0,45107$$

Calcul du coefficient de GINI de la filiale B selon la formule :

$$G = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(Y_i + Y_{i+1}\right)$$

A	B	C	D	E
Salaires mensuels filiale B	Fréquences salaires mensuels	Fréquences cumulées salaires mensuels (Yi)	$(Y_i + Y_{i+1})$	
1500	0,031	0,03098	0,03098	
1500	0,031	0,06196	0,09294	
1500	0,031	0,09294	0,15489	
1550	0,032	0,12495	0,21789	
1550	0,032	0,15696	0,28191	
1600	0,033	0,19000	0,34696	
1700	0,035	0,22511	0,41512	
1700	0,035	0,26022	0,48534	
1740	0,036	0,29616	0,55638	
1800	0,037	0,33333	0,62949	
1840	0,038	0,37133	0,70467	
1850	0,038	0,40954	0,78088	
1870	0,039	0,44816	0,85770	
1900	0,039	0,48740	0,93556	
1920	0,040	0,52705	1,01446	
1940	0,040	0,56712	1,09418	
1960	0,040	0,60760	1,17472	
5000	0,103	0,71086	1,31846	
6000	0,124	0,83478	1,54564	
8000	0,165	1,00000	1,83478	
48420			**14,47295**	
			0,72365	
			0,27635	

$$G = 1 - \frac{1}{20}\sum_{i=1}^{20}\left(Y_i + Y_{i+1}\right) = 1 - 0,72365 = 0,27635$$

On constate que les coefficients de GINI de la filiale A et de la filiale B obtenus avec la seconde formule sont identiques à ceux obtenus avec la première formule. **On pourra donc préférer utiliser la seconde formule dans les calculs (lorsque les données sont connues individuellement) car elle est la plus simple**.

Bibliographie

A

ABELL Martha L., James P. BRASELTON & John A. RAFTER (1998), Statistics with mathematica , Academic Press.

ALBARELLO, Luc, Jean-Luc GUYOT et Etienne BOURGEOIS (2002), Statistique descriptive, De Boeck

AVENEL, Jean-David (1999), Statistique descriptive : Cours et exercices corrigés, Dunod.

B

BADIA, Jacques, René BASTIDA et Jean-Robert HAIT (1997), Statistique sans mathématique, Ellipses

BAILLY, Pierre (1999), Statistique descriptive, Presses Universitaires de Grenoble

BEAUFILS, Béatrice (1996) , Statistiques appliquées à la psychologie. Statistiques descriptives, tome 1, éditions Bréal.

BLUMAN, Allan (2005), Elementary Statistics: A Step by Step Approach, Mc Graw Hill Publishing Company

BOUNDFORD, Trevor et Alaister CAMPBELL (2000), Digital Diagrams, Watson-Guptill Publications.

BOURSIN, Jean-Louis (2000), La statistique pour l'économie et la gestion: QCM, EJA/Gualino.

C

CALOT, Gérard (1969), Cours de statistique descriptive , Dunod.

CHAUVAT, Gérard et Jean-Philippe REAU (1995), Statistique descriptive, Hachette Supérieur.

D

DAGNELIE, Pierre (1998), Statistique théorique et appliquée. Statistique descriptive et bases de l'inférence statistique, tome 1 , De Boeck

DELMAS, Bernard (2005), Statistique descriptive, Armand Colin, Fac économie

de BERNONVILLE, Dugé (1939), Initiation à l'analyse statistique , Librairie de Droit et de Jurisprudence.

DUTHIL, Gérard (1998), Initiation à la statistique descriptive , Ellipse Marketing

G

GEORGIN, Jean-Pierre et Michel GOUET, Statistiques avec Excel : Descriptives, tests paramétriques et non paramétriques à partir de la version Excel 2000 (1Cédérom), Presses Universitaires de Rennes.

GONICK larry et Woollcott SMITH (1993), The Cartoon Guide to Statistics, HarperCollins Publishers

GOULET, DRETZKE (2004), Statistiques avec Microsoft Excel, Reynald et Goulet éditeur.

GRAIS, Bernard (2003), Statistique descriptive : Techniques statistiques , Dunod.

GUEGUEN, Nicolas (2005), Statistiques pour psychologues : Cours et exercices, Dunod.

H

HAND, D.J. (1993), A Handbook of Small Data Sets, Chapman & Hall.

HUFF, Darrell et Irving GEIS (1993), How to Lie With Statistics, W. W. Norton & Company

HOWELL, David (1998), Méthodes statistiques en sciences humaines, De Boeck.

I

INSEE (2005), Tableaux de l'économie française, INSEE Editeur, Collection "Références".

J

JAISINGH, Lloyd R. (2005), Statistics for the Utterly Confused, McGraw-Hill.

JANVIER, Michel (1999), Statistique descriptive : Avec ou sans tableur, cours et exercices corrigés, Dunod.

JONES, Gerald, E. (1995), How to Lie With Charts, Sybex

K

KAZMIER, Leonard (2003), Business Statistics: Based on Schaum's Outline of Theory and Problems of Business Statistics, Third Edition, Schaum/McGraw Hill Publishing Company.

L

LETHIELLEUX, Maurice (2003), Statistique descriptive, éditions Dunod, Collection "Express".

M

MASSONI, André (2002),Initiation aux statistiques descriptives avec Excel. 2ème édition, Vuibert.
.
MILNE, P. H. (1992), Presentation Graphics For Engineering, Science And Business, Spon (Chapman & Hall)

MONINO, Jean-Louis, Jean-Michel KOSIANSKI et François LE CORNU (2004), Statistique descriptive : Travaux dirigés, Dunod.

MOORE, David, S. et George P. McCABE (2002), Introduction to the Practice of Statistics , 4ème édition, W.H. Freeman & Company.

P

PIATIER, André (1966), Statistique, statistique descriptive et initiation à l'analyse, Puf, Presses Universitaires de France, Collection Thémis, Manuels Juridiques, Economiques et Politiques .

PILLER, Alain (2004), Statistique descriptive : Manuel d'exercices corrigés avec rappels de cours, éditions Premium.

PY, Bernard (2007), La statistique sans formule mathématique , 1ère édition, Pearson Education.

PY, Bernard (2007), Statistique descriptive : nouvelle méthode pour comprendre et bien réussir 5ème édition, Economica.

PY, Bernard (2007), Exercices corrigés de statistique descriptive : Problèmes, exercices et QCM , 3ème édition revue et augmentée, Economica.

R

REUCHLIN, Maurice (1998), Précis de statistique : Présentation notionnelle, 7e édition, PUF.

RODRIGUEZ, Marc et Michel TERRAZA (1998), Statistique descriptive: 30 exercices corrigés, EJA/gualino.

RUMSEY, Deborah (2003), Statistics for Dummies, Wiley Publishing inc. Site internet de la collection "... for dummies" : Etats-Unis. Voir aussi la page Web du livre.

S

SCHARLIG, Alain (1997), Faire parler les chiffres: La statistique descriptive au service de la gestion, Presses Polytechniques et Universitaires Romandes (PPUR)

SLAVIN, Steve (1998), Chances Are: The Only Statistics Book You'll Ever Need, Madison Books

SPIEGEL, Murray et Larry STEPHENS, Statistique: Cours et problèmes, 3ème édition, Série Schaum/McGraw Hill

T

TUFTE, Edward (2001), The Visual Display of Quantitative Information, Graphics Press. Voir le site internet de Edward TUFTE.

V

VOELKLER, David, Peter ORTON et Scott ADAMS (2001), Cliffsquickreview Statistics, Hungry Minds

W

WAINER, Howard (2005), Graphic Discovery: A Trout in the Milk and Other Visual Adventures, Princeton University Press.

WILKINSON Leyland, S. (1999), The Grammar of Graphics, Springer.

Z

ZELAZNY, Gene (2001), Say it with Charts : The Executive's Guide to Visual Communication, McGraw-Hill

Sites internet utiles

Le cours de statistiques descriptives de Daniel MIRZA :
http://perso.univ-rennes1.fr/daniel.mirza/

Le cours de Daniel GRAU. Très bien fait : http://www.iutbayonne.univ-pau.fr/~grau/ .
Il explique notamment comment tracer une courbe de LORENZ sous EXCEL

Le cours du Dr. Hossein ARSHAM, de l'Université de Baltimore :
http://home.ubalt.edu/ntsbarsh/

Un site pour la création de graphiques sous Excel : http://sn1.chez-alice.fr/presentation/excel.html

Statistics at square one : http://bmj.bmjjournals.com/collections/statsbk/index.shtml

Le cours de Statistiques & informatique de Jean VERONIS (Université de Provence), avec powerpoint téléchargeables :
http://www.up.univ-mrs.fr/~veronis/cours/index.html

Techniques d'analyse quantitative de données I de Gilles Dupuis - (Département de psychologie de l'Université du Québec à Montréal) :
http://www.er.uqam.ca/nobel/r16424/PSY7102/

Hyperstats Online TextBook : http://davidmlane.com/hyperstat/index.html

Le cours de B. ICARD (Université de Paris V) :
http://www.math-info.univ-paris5.fr/smel/cours/cadre_cours.html

A new view of statistics : http://www.sportsci.org/resource/stats/contents.html

Le PDF de Laurent DOYEN sur la statistique descriptive : http://www-lmc.imag.fr/lmc-sms/Laurent.Doyen/StatDesc2HTML.pdf

Statistics for economists, a beginning:
http://www.economics.utoronto.ca/archives/floyd_stats/

Le cours de Patrice BOUGETTE :
http://perso.orange.fr/patrice.bougette/HTML/tdstat.htm , avec notamment un lien vers un texte sur l'histoire des séries chronologiques (PDF).

Le cours de Pierre MAGAIN, Introduction aux méthodes quantitatives et éléments de statistiques, Institut d'Astrophysique, de géophysique et d'océanographie de Liège :
http://www.astro.ulg.ac.be/cours/magain/stat/index.html

Faire des graphiques avec EXCEL : Le cours de Christine CAMPIONI (Centre de Mathématiques et d'informatique de Chateau-Gombert)

http://www.cmi.univ-mrs.fr/~campioni/documents/MASS/cours/Graphiques.doc
(accès direct au document word).