

Chapitre 1
Vocabulaire de la statistique descriptive
Dernière mise à jour le 16 juillet juin 2009

« La vraie faute est celle que l'on ne corrige pas » - CONFUCIUS (551 av. J.-C. – 479 av. J.-C.)

0 – [Introduction](#)

1 – [Définition du champ de la statistique descriptive](#)

2 – [Terminologie](#)

- A – Population et unités statistiques
- B – Echantillons et sous-ensembles d'une population
 - 1) Echantillon et population
 - 2) Répartition des unités statistiques selon différents critères
- C - Critères de classification
 - 1) Critères quantitatifs
 - 2) Critères qualitatifs

3 – [Modes de regroupement des unités statistiques](#)

- 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

4 – [Résumé](#)

Exercices corrigés (PDF) :

TD5 2008 : [Sujets](#) - [Corrigés](#)

TD5 2009 : [Sujets](#) - [Corrigés](#)

[Voir aussi les sujets d'examens corrigés](#) (lien vers page HTML)

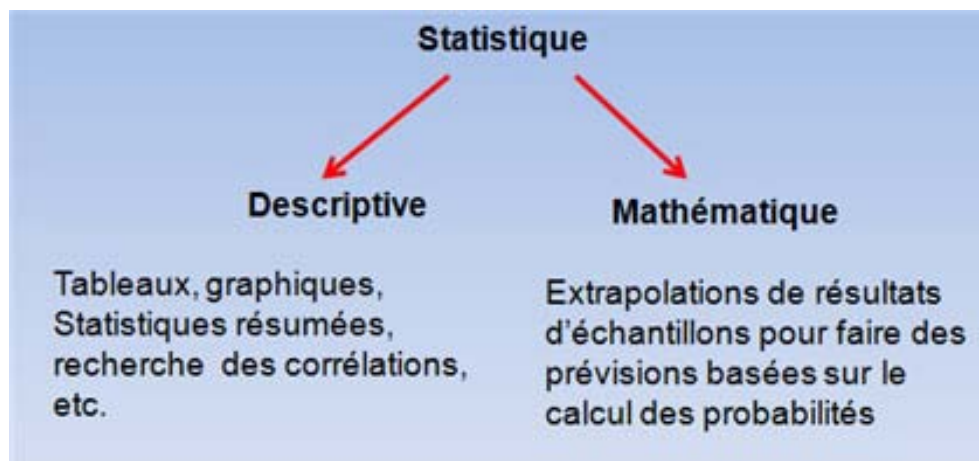
0 – Introduction

Ce chapitre présente le champ de la statistique descriptive et son vocabulaire de base.

1 – Définition du champ de la statistique descriptive

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 "unité statistique") 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 tirages limités effectués au sein d'une 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**.



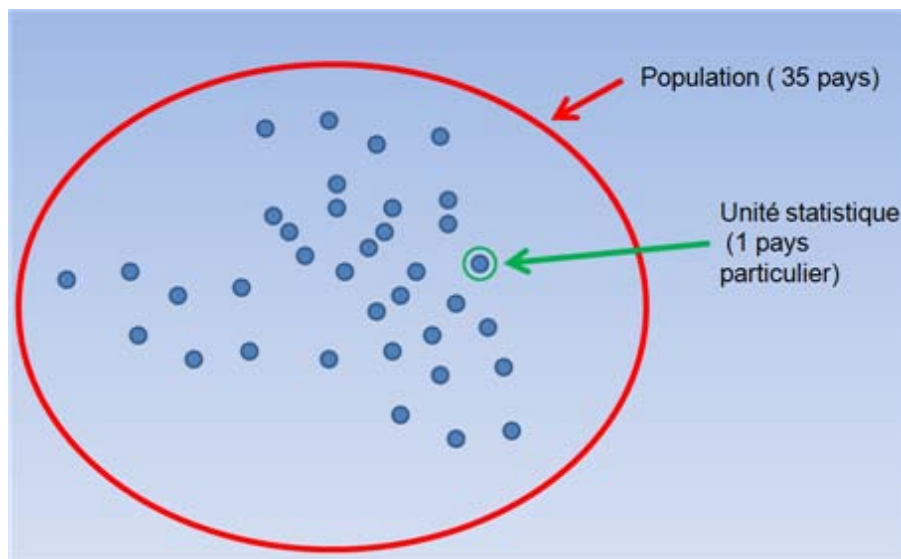
La statistique, qu'elle soit descriptive ou mathématique, est employée dans toutes les sciences, ainsi que dans la vie quotidienne. Son utilisation très intensive dans le champ de l'économie a fait naître une nouvelle expression : **L'économétrie**. **L'économétrie** est l'application de la statistique (descriptive et mathématique) à la mesure et à l'étude chiffrée des grandeurs économiques. De la même façon, on emploie parfois l'expression (un peu vieillie) de « sociométrie », pour parler de l'application de la statistique (descriptive et mathématique) à la mesure et à l'étude chiffrée des grandeurs sociologiques.

2 - Terminologie

A – Population et unités statistiques

En statistique, la **population** désigne un ensemble d'**unités statistiques**. Les unités statistiques sont des personnes, des animaux ou des objets concrets ou abstraits. Les premières populations ayant fait l'objet d'un recensement ayant été des populations humaines (d'où le lien étroit entre statistique et démographie) le terme "individu" est parfois employé comme synonyme du terme "unité statistique". La meilleure façon de se représenter une population statistique est de songer à une **base de données**.

Dans ce cours, sauf mention contraire, nous considérons donc que les expressions population et base de données sont synonymes. Autrement dit, nous nous occupons d'étudier des populations concrètes, dont le nombre, quoique pouvant être arbitrairement grand, est toujours fini. Les unités statistiques peuvent toujours être comptées avec précision et même identifiées individuellement, fût-ce par un numéro.



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 : Le tableau 1 ci-après qui contient des statistiques macroéconomiques relatives aux 27 pays de l'UE, plus huit de ses principaux partenaires commerciaux. L'unité statistique étant le pays, le tableau contient une population de 35 unités statistiques.

Tableau 1 : Quelques statistiques macroéconomiques des pays de l'UE à 27 et de huit de ses partenaires commerciaux – [Fichier EXCEL](#)

Pays	Monnaie (au 1er janvier 2009)	Superficie (km ²)	Année d'adhésion à l'UE (pour les pays concernés)	Nombre de frontières terrestres avec d'autres pays de l'UE à 27	PIB par habitant en dollars PPA (2008)
Allemagne	Euro	357021	1956	8	34800
Belgique	Euro	30528	1956	4	37500
France	Euro	547030	1956	5	32700
Italie	Euro	301320	1956	3	31000
Luxembourg	Euro	2586	1956	3	81100
Pays-Bas	Euro	41526	1956	2	40300
Danemark	Couronne Danoise	43094	1973	1	37400
Irlande	Euro	70280	1973	1	46200
Royaume-Uni	Livre britannique	244820	1973	1	36600
Grèce	Euro (2001)	131940	1981	1	32000
Espagne	Euro	504782	1986	2	34600
Portugal	Euro	92931	1986	1	22000
Autriche	Euro	83858	1995	6	39200
Finlande	Euro	337030	1995	1	37200
Suède	Couronne suédoise	449964	1995	1	38500
Chypre	Euro (2008)	9250	2004	0	28600
Estonie	Couronne estonien	45226	2004	1	21200
Hongrie	Forint	93030	2004	4	19800
Lettonie	Lat	64589	2004	2	17800
Lituanie	Litas	65200	2004	2	17700
Malte	Euro (2008)	316	2004	0	24200
Pologne	Zloty	312685	2004	4	17300
République tchèque	Couronne tchèque	78866	2004	4	26100
Slovaquie	Euro (2009)	48845	2004	4	21900
Slovénie	Euro (2007)	20253	2004	3	29500
Bulgarie	Lev	110910	2007	2	12900
Roumanie	Leu	238391	2007	2	12200
Suisse	Franc suisse	41290		4	40900
Etats-Unis	Dollar	9826630		0	47000
Chine	Yuan	9596960		0	6000
Inde	Roupie	3287592		0	2800
Japon	Yen	377835		0	34200
Russie	Rouble	17075200		5	15800
Taiwan	Dollar taiwanais	35980		0	31900
Hong Kong	Dollar de Hong Kong	1092		0	43800
Monde		510072000			10400

— Source : The World Fact Book, <https://www.cia.gov/library/publications/the-world-factbook/index.html>

— Note : Le PIB par habitant est exprimé en dollars PPA ou "Dollars de Parité de Pouvoir d'Achat" car cela reflète mieux la réalité (élimination des effets de taux de change.)

— Pour plus de détails sur la signification et la méthode de calcul du taux de change PPA,

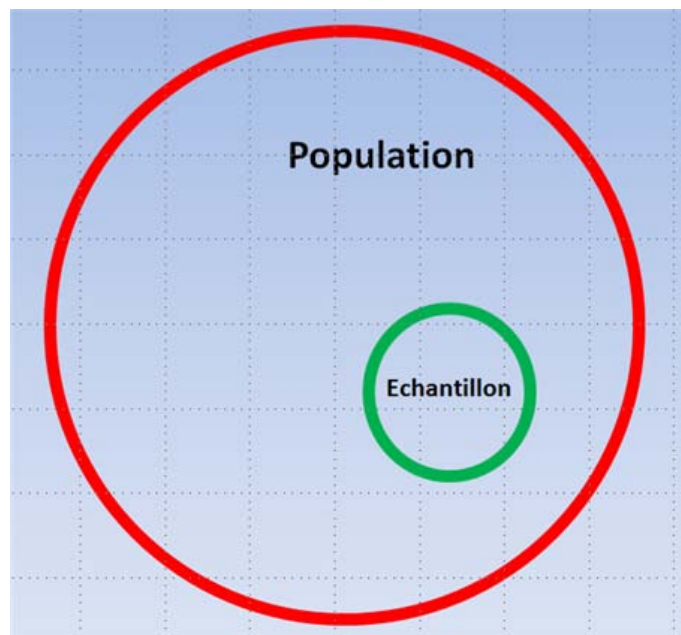
voir http://www.oecd.org/document/47/0,2340,en_2649_34357_36202863_1_1_1_1,00.html

B – Echantillons et sous-ensembles d'une population

1) Echantillon et population

Il est fréquent que l'on prélève un **échantillon** dans une population statistique. Le **diagramme d'EULER**¹ ci-après décrit le lien entre l'échantillon et la population.

Le lien entre l'échantillon et la population



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**.

2) Répartition des unités statistiques selon différents critères

L'objet de la statistique descriptive est de diviser les populations en sous-ensembles au moyen de certains critères (ou **dimensions** ou encore **caractéristiques**).

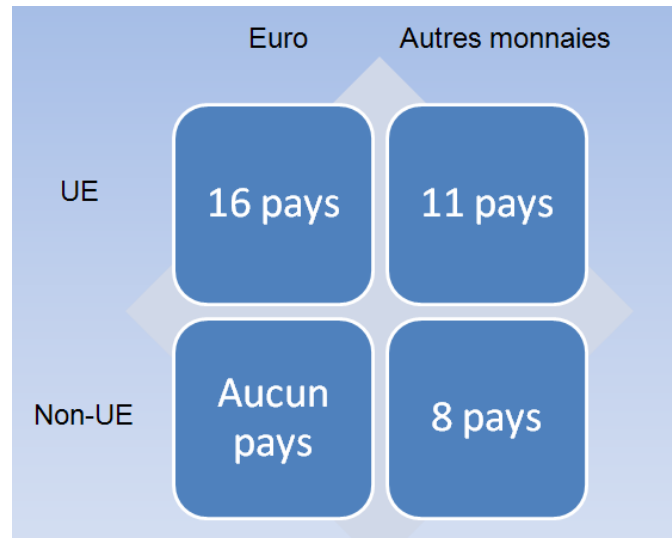
Prenons pour exemple la population des 35 pays du tableau 1. Ces 35 pays sont les unités statistiques et constituent, prises ensemble, la « population statistique ». Nous allons par exemple « découper » cette population entre trois sous ensembles,

¹ Leonhard Paul EULER, né le 15 avril 1707 à Bâle et mort le 18 septembre 1783 à Saint-Pétersbourg, est un mathématicien et physicien suisse, qui a passé la plupart de sa vie en Russie et en Allemagne. EULER a fait d'importantes découvertes dans des domaines aussi variés que le calcul infinitésimal et la théorie des graphes (comme le diagramme ci-dessus qui porte son nom). Il a aussi introduit une grande partie de la terminologie et de la notation des mathématiques modernes en particulier pour l'analyse mathématique, comme pour la notion d'une fonction mathématique. Il est également connu pour ses travaux en mécanique, en dynamique des fluides, en optique et en astronomie. EULER est considéré comme un éminent mathématicien du XVIII^e siècle et l'un des plus grands de tous les temps. En savoir plus : [http://fr.wikipedia.org/wiki/Leonhard Euler](http://fr.wikipedia.org/wiki/Leonhard_Euler)

suivant les critères de la monnaie utilisée et de l'appartenance à l'UE 27. On aura donc, comme l'illustre le schéma ci-après :

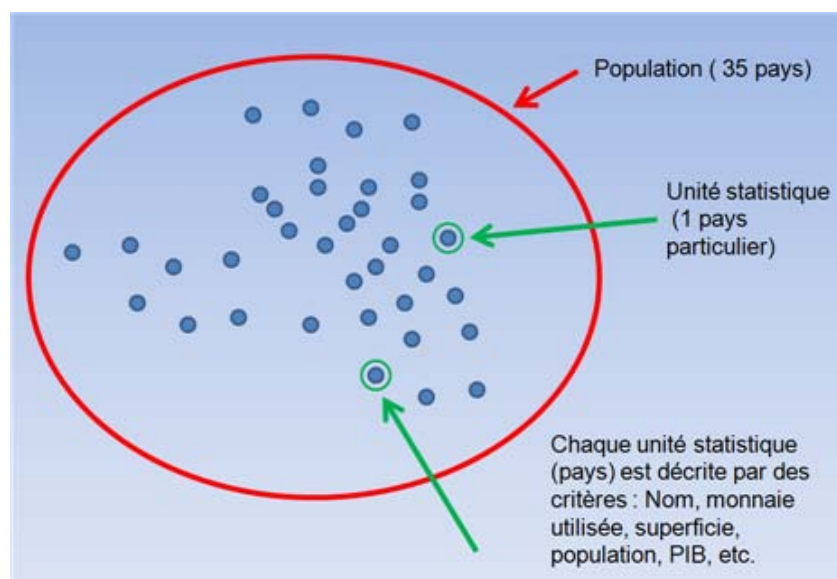
- 16 pays membres de l'UE 27 qui font partie de la zone Euro au 1^{er} janvier 2009,
- 11 pays membres de l'UE à 27 qui ne font pas (encore) partie de la zone Euro au 1^{er} janvier 2009
- 8 pays partenaires de l'UE 27 et qui utilisent d'autres monnaies.

Les 35 pays du tableau 1 répartis selon deux critères



C - Critères de classification

Nous avons vu dans l'exemple précédent que les unités statistiques d'une population pouvaient être regroupées suivant des **dimensions** ou **critères**, ou bien des **caractéristiques**. Ces critères sont choisis en fonction de ce qui intéresse le statisticien (ou la personne ou le groupe qui effectue l'étude).



On distingue deux sortes de critères² :

- Les critères quantitatifs
- Les critères qualitatifs

1) Critères quantitatifs

Les **critères quantitatifs** sont les critères qui sont représentés par des nombres et sur lesquels les opérations arithmétiques de base ont un sens. Les critères quantitatifs prennent des valeurs numériques, d'où l'idée de les appeler des « variables » (dans le dictionnaire, le mot variable est défini ainsi : quantité qui peut changer de valeur). Par exemple, dans le tableau 1, on peut voir que la superficie est un critère de classification quantitatif. Les différentes **occurrences** du critère quantitatif « superficie » sont des **valeurs numériques**.

Chacune des 35 unités statistiques de notre population est caractérisée par une valeur différente. La superficie est donc, dans notre cas particulier, un critère quantitatif qui prend 35 valeurs différentes. C'est un cas particulier dans lequel le nombre de valeurs 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 de la population, est grand) un regroupement par classes de valeurs est généralement inévitable.

Les critères quantitatifs peuvent faire l'objet de manipulations algébriques. On peut par exemple additionner les superficies, faire le rapport de la plus grande superficie à la plus petite, calculer une superficie moyenne, etc.

2) Critères qualitatifs

Les **critères qualitatifs** sont tous les critères qui ne sont pas représentés par des nombres³. Les caractères prennent des **modalités**⁴.

Par exemple, dans le tableau 1, on peut voir que la monnaie utilisée dans chaque pays est un critère qualitatif qui possède 20 modalités. Ces modalités sont les différentes monnaies. Il y a en effet 16 pays qui ont la modalité « euro » et les 19 autres qui ont chacun pour modalité une monnaie différente. On voit donc dans cet exemple que le nombre de modalités (20) est inférieur à celui de la population (35).

² Dans la suite, les mots « critères », « dimensions » et « caractéristiques » seront employés indifféremment.

³ Ils sont parfois codés par des nombres, nombres sur lesquels les opérations arithmétiques de base n'ont pas de sens (ou un sens très limité). Il est fréquent par exemple de coder des opinions.

⁴ Le fait de restreindre l'expression « modalité » aux seuls critères qualitatifs n'est qu'une suggestion qui est faite dans ce cours par souci de clarification, sans conséquence sur les différents calculs.

3 - Modes de regroupement des unités statistiques

A - Série simple

Le tableau 1 est un tableau dans lequel les unités statistiques n'ont pas été regroupées. C'est un tableau de données brutes. Nous pouvons lire sur chaque ligne du tableau 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 numériques ou de modalités.**

Par exemple, dans le cas de la dimension « superficie », il y a 35 valeurs différentes. Dans le cas du caractère « monnaie », il y a 20 modalités différentes. Dans le cas de la dimension « nombre de frontières terrestres avec d'autres pays de l'UE à 27 », les 35 pays se répartissent seulement sur 8 valeurs (si l'on exclut la valeur « 7 ») ou 9 valeurs (si l'on inclut la valeur « 7 »)⁵.

Mais une présentation exhaustive, dans laquelle aucun regroupement n'est effectué, n'est pas toujours pratique. Le plus souvent les données sont collectées et entrées dans l'ordinateur sous forme d'un tableau brut, puis elles sont **regroupées**.

⁵ La valeur « 7 » étant associée à un effectif nul, on peut décider de l'inclure (et dans ce cas il y a 9 valeurs) ou de ne pas l'inclure (et dans ce cas il y a 8 valeurs). Si la nomenclature des valeurs de la variable est amenée à servir pour plusieurs populations successives (ou une même population à différents temps), il vaut sans doute mieux, pour faire d'éventuelles comparaisons entre les populations ou les temps, inclure toutes les valeurs, y compris celles associées à un effectif nul, car l'effectif associé à une valeur peut changer selon la population et/ou le temps. Mais ceci ne vaut que si le nombre de valeurs possibles n'est **pas trop important** (sinon il faut procéder à un regroupement par classes de valeurs), **dénombrable** et **fini**. Certaines dimensions quantitatives peuvent avoir un nombre infini (dénombrable ou indénombrable) de valeurs. Elles sont alors dites **continues** (par opposition aux **dimensions quantitatives non continues** qui sont dites « discrètes » du fait d'une traduction abusive de l'anglais). Néanmoins, les **populations** (population = base de données) étudiées concrètement (populations pour lesquelles des données quantitatives ont été réunies) ont toujours **un nombre fini d'unités statistiques**. Par conséquent, au sein d'une base de données, le **nombre de valeurs** (c'est –à-dire le nombre de valeurs au sein de la population associées à un effectif non nul) qu'une dimension quantitative peut prendre est toujours dénombrable et fini et au maximum égal au nombre d'unités statistiques contenues dans la base de données (ou population). En effet, chaque unité statistique ne peut être caractérisée que par une seule valeur et donc le maximum que l'on puisse avoir est une distribution où chaque valeur a pour effectif 1, ce qui correspond au cas où chaque unité statistique est caractérisée par une valeur différente. La distinction entre dimension quantitative discrète et dimension quantitative continue s'appuie sur une deuxième justification, que nous mentionnons bien qu'elle nous paraisse source de confusion : une variable (dimension quantitative) est dite « à valeurs discrètes » **lorsque les valeurs peuvent être comptées** et est dite « à valeurs continues » **lorsque les valeurs peuvent être mesurées**. En se référant à cette distinction, on dira alors que le nombre de pièces d'un appartement est une variable discrète car on peut compter le nombre de pièces. Inversement, on dira que la taille des unités statistiques d'une population est une variable continue car pour savoir combien mesure chaque unité on devra procéder à une mesure de la taille (idem pour le poids), mesure qui d'ailleurs sera toujours imprécise, car relative à l'instrument de mesure utilisé. Finalement, la troisième et unique justification sérieuse de la distinction discret/continu s'applique lors de l'étude des **distributions théoriques** : certaines sont discrètes (loi uniforme discrète, loi binomiale, loi de POISSON, loi hypergéométrique, etc.) et d'autres sont continues (loi continue uniforme, loi normale, loi de PARETO, etc.).

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**. Dans notre exemple, 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	
	0	8	
	1	8	
	2	6	
	3	3	
	4	6	
	5	2	
	6	1	
	7	0	
	8	1	
		35	

Distribution sur 9 valeurs (la valeur 7 pourrait éventuellement être retirée du tableau, mais cela créerait une discontinuité)

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 la monnaie officielle utilisée dans chaque pays comme critère qualitatif pour effectuer un regroupement par modalités. Au 1^{er} janvier 2009, 16 sur les 35 pays sont dans la zone euro et les 19 autres utilisent leur monnaie nationale. Dans ces conditions, un regroupement par modalités, quoique peu utile, donnerait le résultat suivant :

Monnaie	Effectifs
Euro	16
Couronne danoise	1
Livre britannique	1
Couronne suédoise	1
Couronne estonienne	1
Forint	1
Lat	1
Litas	1
Zloty	1
Couronne tchèque	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

**35 unités
distribuées
sur 20
modalités**

C - Regroupement par catégories

Lorsqu'il y a beaucoup de valeurs ou de modalités, on peut procéder à un regroupement par **catégories de valeurs** ou par **catégories 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 pays (Malte) n'a qu'une superficie de 316 km^2 et le plus grand pays (La Russie) est caractérisé par une superficie de $17\,075\,200 \text{ km}^2$? Si l'on regarde les superficies des différents pays, on voit qu'un très grand nombre de pays ont des superficies inférieures à $600\,000 \text{ km}^2$. 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	377 835
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

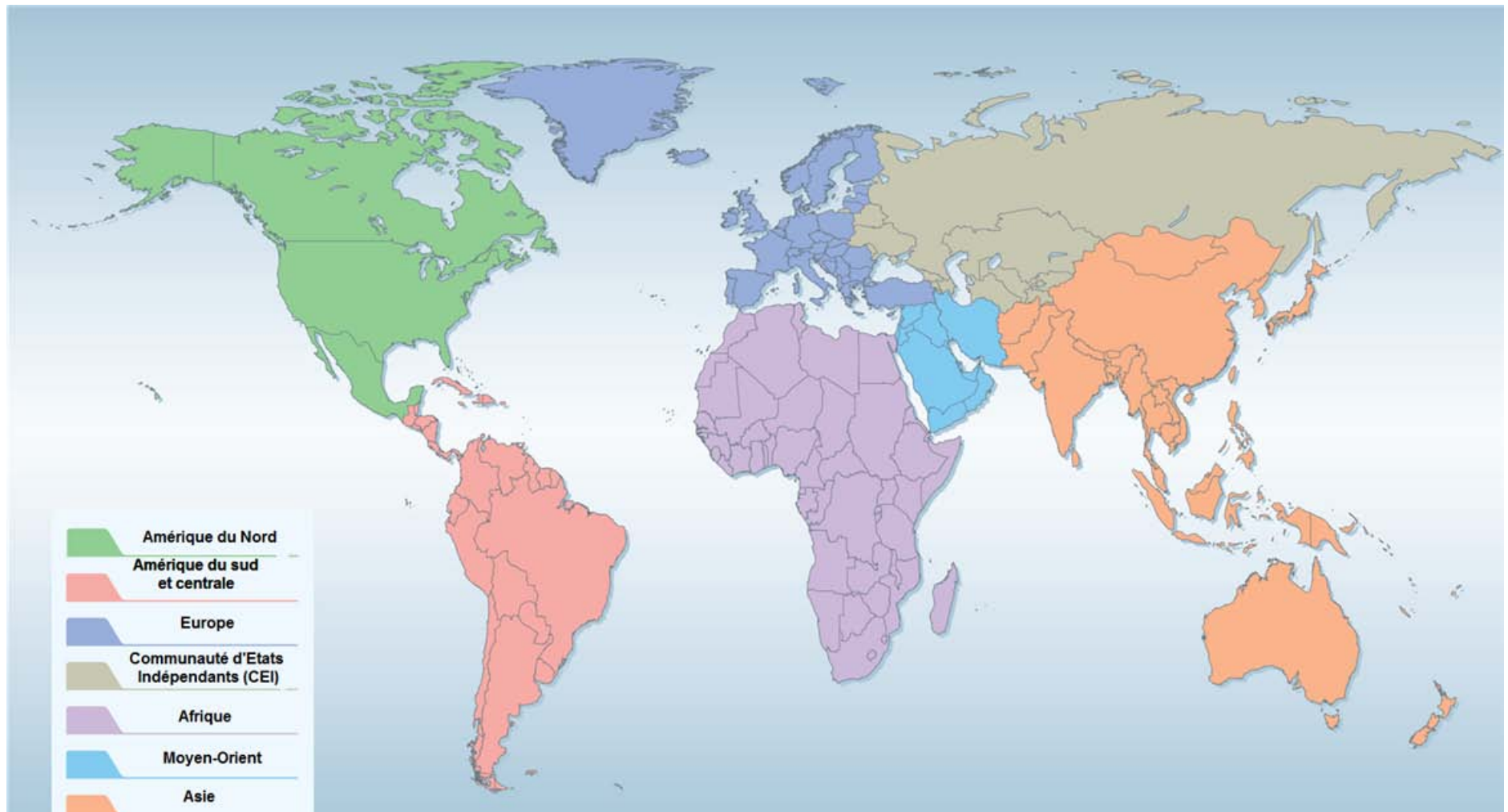
Transformation de la série simple en distribution par classes de valeurs

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 « nom du 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, Taiwan, 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.

4 – Résumé

La **statistique descriptive** est un ensemble de méthodes permettant de décrire les unités statistiques qui composent une population

La **statistique mathématique** s'appuie sur la statistique descriptive, ainsi que sur les **probabilités** pour formuler des lois à partir de l'observation d'échantillons.

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

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.

On parle d'**échantillon** d'une population statistique pour désigner le prélèvement, au hasard ou selon une méthode qui permet d'assurer la représentativité par rapport à la population totale, d'un petit nombre d'unités statistiques au sein de la population.

Par ailleurs, il est fréquent aussi que l'on divise une population en sous-ensembles au moyen de certains critères (ou **dimensions** ou encore **caractéristiques**).

On distingue deux sortes de critères : les **critères quantitatifs** et les **critères qualitatifs**. Les **critères quantitatifs** sont les critères qui sont représentés par des nombres et à condition que les opérations arithmétiques effectuées sur ces nombres aient un sens. C'est la raison pour laquelle on les appelle aussi parfois, par analogie avec les fonctions théoriques continues, des **variables**.

Les **critères qualitatifs** sont tous les critères qui ne sont pas représentés de façon numérique. Pour les distinguer des critères quantitatifs (désignés par l'expression « variables »), nous les appellerons des « caractères ». Et comme les variables prennent des valeurs numériques, nous dirons pour poursuivre la distinction mais conserver le parallèle, que les caractères prennent des **modalités**.

Les modalités ou les valeurs des critères associés à une population peuvent être représentées sous forme d'une **série simple** ou **regroupées**. Lorsqu'elles sont regroupées on les appelle des **distributions**. Ces distributions sont empiriques c'est-à-dire constituées par les modalités ou les valeurs particulières prises par les unités statistiques de la population concrètement étudiée.

Les unités d'une population peuvent être **distribuées par valeurs** (lorsque le critère de regroupement est numérique) ou **distribuées par modalités** (lorsque le critère de regroupement n'est pas numérique). On peut aussi effectuer des regroupements par **catégories (ou classes) de valeurs** ou par **catégories (ou classes) de modalités**.

Lorsqu'on effectue un regroupement des unités statistiques par catégories ou par classes de valeurs, on peut choisir des **classes d'égales amplitudes** ou des **classes d'inégales amplitudes**. L'**amplitude de classe** est la différence entre la valeur supérieure et la valeur inférieure de la classe. Le **centre de classe** est égal à la somme de la valeur inférieure et de la valeur supérieure, divisée par deux.

Exercices corrigés (PDF) :

TD5 2008 : [Sujets](#) - [Corrigés](#)

TD5 2009 : [Sujets](#) - [Corrigés](#)

[Voir aussi les sujets d'examens corrigés](#) (lien vers page HTML)