# Chapitre 1
# Vocabulaire de la statistique descriptive

*Dernière mise à jour le 1er octobre 2009*

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 caractères
- C – Caractères quantitatifs et non quantitatifs
  - 1) Caractères quantitatifs (continus et non continus [discrets])
  - 2) Caractères qualitatifs (ordinaux ou nominaux)
  - 3) tableau récapitulatif
  - 4) Pertinence de la classification des caractères

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) :**

TD1 2008 : Sujets - Corrigés
TD1 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.

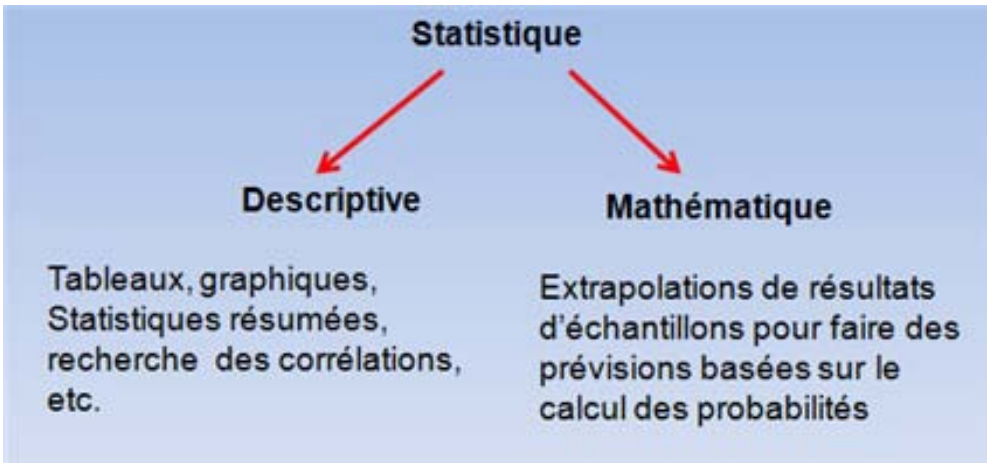
## 1 – Définition du champ de la statistique descriptive

Il existe tellement de définitions différentes de la statistique qu'on pourrait presque en faire une étude … statistique. Ainsi, R. DUMAS, dans son ouvrage datant de 1967, en dénombre-t-il déjà une centaine[1].

Gérard CALOT, dans son *Cours de statistique descriptive* (DUNOD, 1965), définit la statistique comme une « *méthode qui vise à la description quantitative des ensembles nombreux* ». C'est donc une méthode et non une théorie, une méthode quantitative et non qualitative, tournée vers le grand nombre, la généralité, plutôt que vers l'accidentel et le particulier.

On divise souvent aujourd'hui l'étude de la **statistique,** au sens général, 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

[1] R. DUMAS, L'entreprise et la statistique, tome 1, page 1, DUNOD Editeur, Paris (1967).

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**.

C'est pourquoi, 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 très grand, est toujours fini. Les unités statistiques peuvent toujours être comptées avec précision et même identifiées individuellement, fût-ce par un numéro.

En assimilant population et base de données, nous restreignons notre étude aux populations statistiques qui ont un nombre fini d'unités.

La statistique descriptive va alors nous permettre de 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, désigné par $n$.

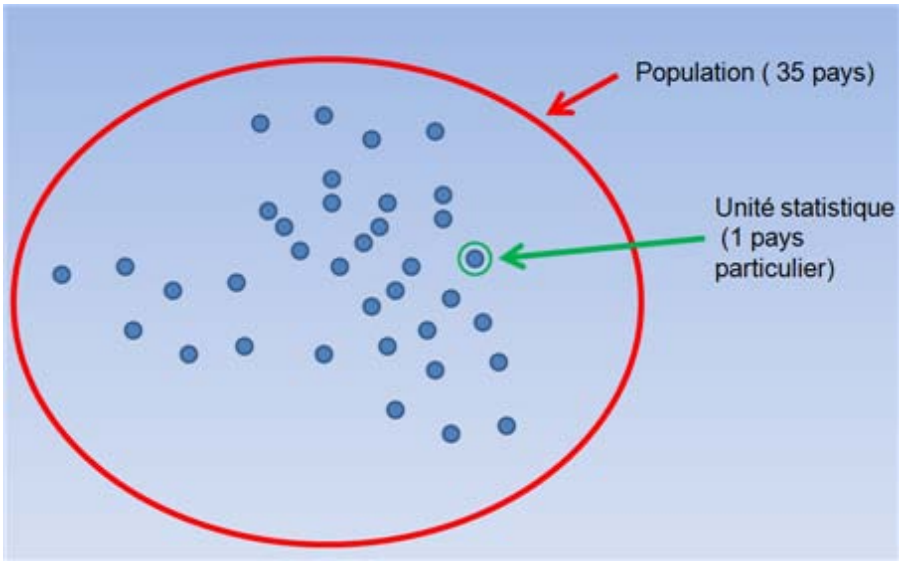
**Exemple** : Le tableau 1 ci-après qui contient des statistiques macroéconomiques relatives aux 27 pays de l'UE, plus huit de ses principaux partenaires commerciaux.

**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 (km2) | 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 estonienr | 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 taïwanais | 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 effet sde 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

L'unité statistique étant le pays, le tableau contient une population de *n* = 35 unités statistiques.

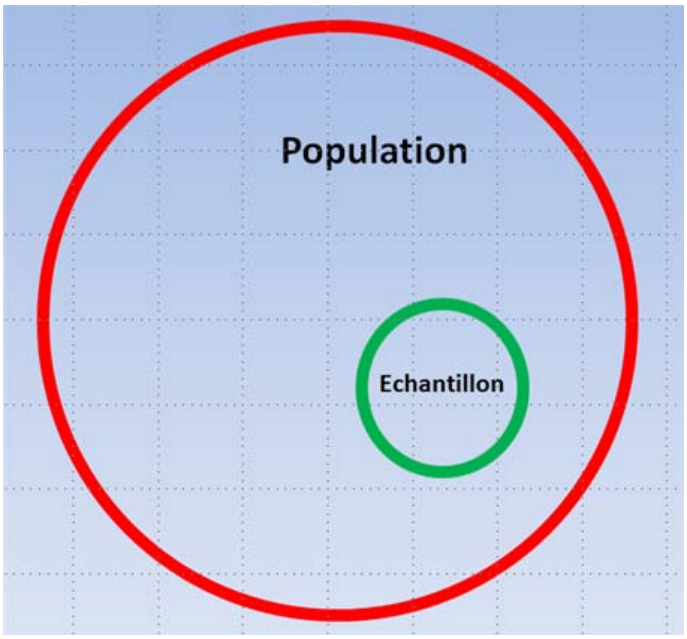


## B – Echantillons et sous-ensembles d'une population

### 1) Echantillon et population

Il est fréquent que l'on prélève un **échantillon** dans une population statistique. Le **diagramme d'EULER**[2] 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] 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 XVIIIe siècle et l'un des plus grands de tous les temps. En savoir plus : http://fr.wikipedia.org/wiki/Leonhard_Euler

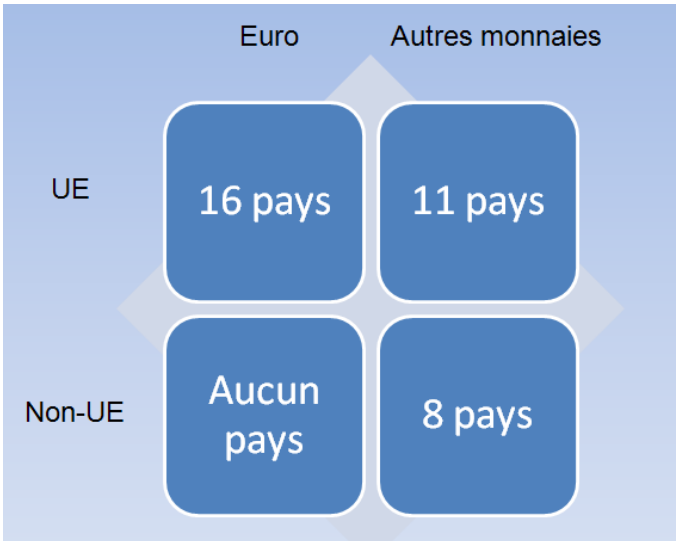
## 2) Répartition des unités statistiques selon différents caractères

On divise les populations en sous-ensembles au moyen de certains critères appelés **caractères**[3] (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 (quatre si l'on compte celui qui est vide), 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 1er janvier 2009,
- 11 pays membres de l'UE à 27 qui ne font pas (encore) partie de la zone Euro au 1er janvier 2009
- 8 pays partenaires de l'UE 27 et qui utilisent d'autres monnaies.

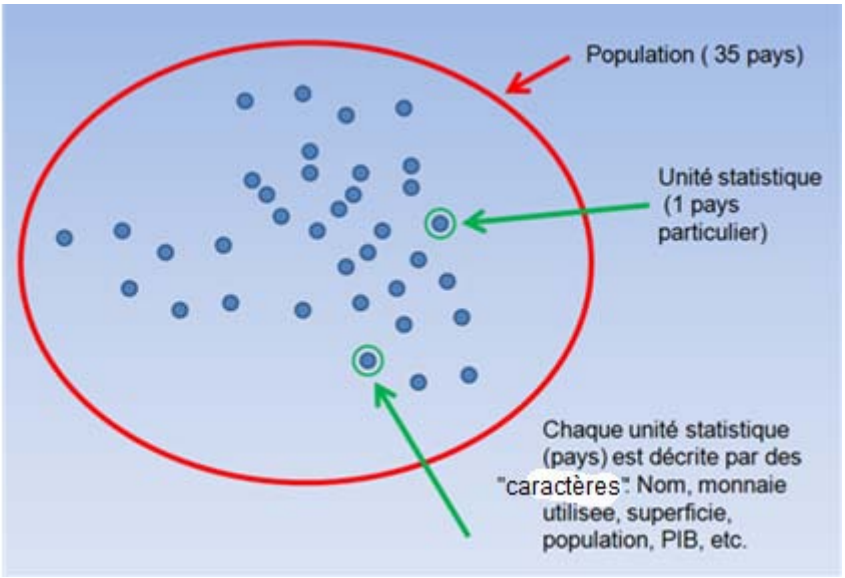
**Les 35 pays du tableau 1 répartis selon deux caractères : la monnaie utilisée et le groupe de pays**




[3] Quand le critère est quantitatif (voir définition plus loin) de très nombreux auteurs (même les plus puristes), utilisent l'expression « variable » à la place de l'expression consacrée « caractère ».

## C – Caractères quantitatifs[4] et non quantitatifs

Nous avons vu dans l'exemple précédent que les unités statistiques d'une population pouvaient être regroupées suivant des **caractères** ou **dimensions** (ou **critères** ou **caractéristiques**). Ces caractères 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 caractères[5] :

- Les caractères quantitatifs (aussi appelés « variables statistiques »)
- Les caractères qualitatifs

### 1) Caractères quantitatifs (continus et non continus [discrets])

Les **caractères quantitatifs** sont ceux qui sont mesurés par des nombres et sur lesquels les opérations arithmétiques de base ont un sens. Les caractères 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 caractère 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 caractère 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 plus loin que dans des cas de ce type (ou bien lorsque le nombre de valeurs

[4] Souvent appelés « variables », ce qui a l'avantage de réserver l'emploi du terme « caractère », pour les critères qualitatifs.
[5] Dans la suite, les mots « caractères », « critères », « dimensions » et « caractéristiques » seront employés indifféremment.

possibles, bien qu'inférieur au nombre d'unités statistiques de la population, est grand) un regroupement par classes de valeurs est généralement inévitable.

Les caractères 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.

**Certains caractères quantitatifs peuvent avoir un nombre infini (dénombrable ou indénombrable) de valeurs sur un intervalle**. On dit alors qu'ils sont **continus**. C'est le cas de la taille et du poids, par exemple.

Inversement, lorsqu'un caractère quantitatif ne peut prendre qu'un nombre fini de valeurs sur un intervalle, il est appelé **caractère quantitatif non continu** ou « **discret** » du fait d'une traduction abusive de l'anglais. C'est le cas du caractère « nombre de pièces du logement », par exemple.

La distinction entre caractère quantitatif continu et non continu peut donc être présentée ainsi : un caractère quantitatif est dit « à valeurs discrètes » **lorsque les valeurs peuvent être comptées** (on peut savoir combien il y en a à l'avance) et est dit « à valeurs continues » **lorsque les valeurs doivent être mesurées** (on sait les mesurer – avec une plus ou moins grande précision – mais on ne peut pas dire combien il y en a). En se référant à cette distinction, on dira alors que le nombre de pièces d'un appartement est un caractère quantitatif discret car on peut compter le nombre de pièces. Inversement, on dira que la taille des unités statistiques d'une population est un caractère quantitatif continu 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.

La distinction discret/continu joue un rôle important dans 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.)

En économie, **certains caractères quantitatifs sont traités comme s'ils étaient continus**, alors qu'ils ne le sont pas : **c'est le cas du revenu**. Celui-ci est exprimé en unité monétaires. Or les unités monétaires se ramènent à des unités discrètes, les centimes[6]. On peut donc toujours compter les valeurs comprises dans un intervalle donné. Mais comme ces valeurs sont très nombreuses, on procède par regroupement (la section 2 est consacrée aux divers modes de regroupement). **En d'autres termes, on traite le revenu comme un caractère quantitatif continu, quoiqu'il n'en soit pas un.**

Inversement, on a parfois du mal à reconnaître des caractères continus car il peut arriver qu'on les traite comme s'ils étaient non continus. C'est souvent le cas du caractère « âge », qui est continu par nature, mais qu'on regroupe cependant

[6] On peut cependant objecter que lorsqu'on convertit une unité monétaire en une autre, on crée parfois (et même souvent) une valeur avec plus de 2 chiffres après la virgule, que l'on arrondi ensuite, par défaut ou par excès.

souvent par années d'âge (0 an, 1 an, 2 an, … 99 ans, etc.). On pourrait donc avoir l'impression trompeuse qu'il s'agit d'un caractère quantitatif discret alors qu'en fait il s'agit d'un caractère quantitatif continu dont on a regroupé les valeurs par intervalles d'un an (la section 2 est consacrée aux divers modes de regroupement). La même remarque s'applique aux séries temporelles. On aura ainsi, par exemple, le taux de change de l'euro « en 2009 », « en 2010 », etc. Les années 2009, 2010, etc. correspondent en fait à des intervalles de temps (durant lesquels le taux de change en question a parfois varié davantage que d'une année à l'autre) et la valeur unique du taux de change qui leur correspond est soit une valeur choisie arbitrairement, soit une moyenne des valeurs journalières, voire horaires, sur l'année.

## 2) Caractères qualitatifs (ordinaux et nominaux)

Les **caractères qualitatifs** sont tous les caractères qui ne sont pas représentés par des nombres[7]. Les caractères qualitatifs prennent des **modalités**[8]. Par exemple, les deux « modalités » du caractère qualitatif « sexe », sont « masculin » et « féminin ».

Dans le tableau 1, on peut voir que la monnaie utilisée dans chaque pays est un caractère 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).

Les modalités de certains caractères qualitatifs peuvent être ordonnées ou hiérarchisées. On dit alors que le caractère est **ordinal**. C'est le cas par exemple de tous les caractères qualitatifs dont les modalités sont des opinions.

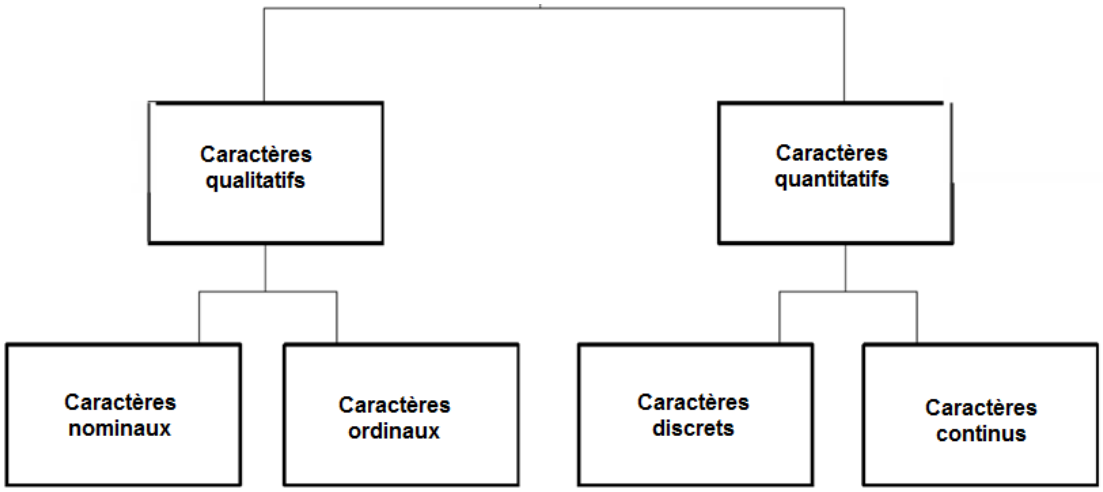
Lorsque les modalités d'un caractère ne peuvent pas être ordonnées, le caractère est dit **nominal**.

[7] Les modalités des caractères qualitatifs peuvent cependant être codées par des chiffres. Par exemple, soit le caractère « sexe », étudié dans une population. On peut remplacer la modalité « Masculin » par « 1 » et la modalité « féminin » par zéro. Cela ne fait pas pour autant du sexe caractère quantitatif. En effet, une opération telle que (0+1)/2=0,5, ne donne en aucun cas la moyenne du caractère. Si l'on prend maintenant le caractère « taille des vêtements », on a en revanche un caractère quantitatif qui est codé par des modalités qualitatives de type « S », « M » et « XL ».

[8] 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) Tableau récapitulatif

Le schéma ci-après récapitule les quatre types de caractères rencontrés en statistique descriptive.




### 4) Questions sur la pertinence de la classification des caractères

La classification des caractères ne doit pas être tenue pour trop absolue. Soit une population des 30 élèves d'une classe de philosophie. Ils font une dissertation sur le thème : « Comment reconnaitre ce qui est continu et ce qui ne l'est pas ? ». Le professeur note la dissertation de chaque élève de 0 à 20. Il y a ainsi 4 dissertations notées pendant l'année (sur d'autres sujets comme par exemple : « Est-il censé de calculer une moyenne sur des données groupées par intervalles ? »). Chaque élève obtient donc 4 notes à la fin de l'année (on suppose qu'aucun élève ne manque lors des épreuves ou bien que des rattrapages sont organisés).

**Question 1 :** Le caractère « note » est-il quantitatif ou qualitatif ?

**Réponse possible :** Le caractère « note » revêt toutes les apparences d'un caractère quantitatif. Les occurrences sont des valeurs numériques et les opérations sur ces valeurs ont apparemment un sens. On peut par exemple, pour chacune des quatre épreuves successives, calculer la moyenne des notes de la classe, ainsi que toute une batterie d'indicateurs de position et de dispersion (voir le chapitre 3). On peut aussi, pour un élève donné, calculer la moyenne de ses quatre notes, ainsi que toute une batterie d'indicateurs de position et de dispersion. Cependant, est-il réaliste de soutenir que la différence entre un élève qui a eu la note « 9 » et un autre qui a eu la note « 11 », différence qui est de 11-9=2, a un sens ? S'agit-il par exemple de la même différence qu'entre deux notes (9 et 11) obtenues par un même élève, à deux épreuves ? Plus important : si deux professeurs devaient corriger à chaque fois les copies, il est pratiquement certains que des différences apparaîtraient. Le caractère « note » ne serait donc qu'apparemment quantitatif. En fait, on **prétend** mesurer, alors qu'en réalité on **juge** et on **évalue**. Peut-être vaudrait-il mieux considérer que l'on est en présence d'un caractère qualitatif.

**Question 2 :** Si le caractère « note » est quantitatif, est-il quantitatif continu ou quantitatif discret ?

**Réponse possible :** Admettons que le caractère « note » soit quantitatif (après tout, il répond à la définition statistique du caractère quantitatif donnée plus haut : les caractères quantitatifs sont ceux qui sont mesurés par des nombres et sur lesquels les opérations arithmétiques de base ont un sens). Il faut alors décider s'il s'agit d'un caractère quantitatif discret ou continu. On pourra soutenir, comme le bon sens nous y invite, qu'il s'agit d'un caractère quantitatif discret, car les plus pointilleux des professeurs ne notent qu'exceptionnellement avec une précision supérieure au quart de point. On aura donc un caractère quantitatif discret dont le nombre de valeurs possibles sera égal à (20 x 4)+1 = 81. Mais si on fait par exemple la moyenne des 3 notes {10,25 ; 11,75 ; 13,5}, on obtient un nombre décimal au développement illimité quoique périodique : (10,25 + 11,75 + 13,5)/3 = 71/6 = 11,83333333. Bien que cette moyenne ne fasse pas partie de l'ensemble initial des valeurs qui définissent le caractère discret, on peut toutefois se demander dans quelle mesure son existence ne remet pas en question la notion de caractère « discret ».

**Question 3 :** Si le caractère « note » est en réalité qualitatif, est-il qualitatif nominal ou qualitatif ordinal ?

**Réponse possible :** Le caractère note est à première vue qualitatif ordinal. Par exemple, on peut remplacer les notes de 0 à 20 par des lettres A, B, C, D et E) et parfois faire des sous-catégories plus ou moins précises au sein de chaque lettre, en ajoutant des « plus » ou des « moins ». Cela à l'avantage de barrer la route à la tentation de calculer des moyennes par élève et par classe (on devra se rabattre sur des notes « modales »).

**Question 4 :** De façon générale, que peut-on dire des classements ordinaux ?

**Réponse possible :** Nous avons vu qu'en mettant une note, on prétend mesurer, alors qu'en réalité on juge (ou au mieux on évalue sur la base de critères qui, même s'ils sont clairement explicités, restent toujours contestables). Mais on mesure quoi et pourquoi ? Déplaçons le problème de la notation des élèves en classe à celui de la notation des pays, comme le fait par exemple la Banque mondiale dans son classement annuel de l'attractivité des pays selon le climat des affaires[9]. On voit par exemple qu'en 2010 c'est Singapour qui est jugé comme le meilleur pays pour faire des affaires, parmi 183 « élèves », moins méritants ! Une autre année, ce sera Hong Kong, où Dubaï, etc. Bref, on arrive à un classement, qui a été établi sur la base d'une moyenne pondérée de notes qui résultent de critères objectifs et d'évaluations données par des panels représentatifs d'hommes d'affaires[10].

Prenons un autre exemple, encore plus scabreux, celui des sondages de satisfaction. On demande à un échantillon représentatif de personnes de dire si elles sont « mécontentes », « moyennement satisfaites », « assez satisfaites », « très satisfaites », etc., de l'augmentation récente des impôts, de la météo, du recul de

[9] Voir http://www.doingbusiness.org/economyrankings/ .
[10] La méthodologie est clairement exposée ici :
http://www.doingbusiness.org/documents/DB10Easeofdoingbusinessrankmethod.pdf

l'âge de la retraite, du pouvoir arbitraire des préposés aux différents guichets administratifs et péri administratif, de la cupidité des dentistes et des avocats, de l'incurie des garagistes, etc. Chacun répond, ou ne répond pas, et l'on calcule ensuite des pourcentages en fonction des réponses obtenues. On refait le sondage un mois plus tard, éventuellement avec les mêmes personnes, et les résultats ont changé. Ici, la méthodologie est encore plus discutable, car on oblige les personnes questionnées à formuler leur opinion à travers une grille de jugement qui n'est probablement pas la leur (si on leur avait demandé de s'exprimer librement, les réponses auraient pu varier de l'interjection polie à une longue vaticination, en passant par l'expression d'envies meurtrières, mais il est probable qu'aucune des personnes interrogées n'a dans la tête la grille à travers laquelle on lui demande pourtant de formuler sa réponse).

## 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 »)[11].

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**.

[11] 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 du caractère 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 change généralement selon la population et/ou le temps. Mais ceci ne vaut que si le nombre de valeurs possibles n'est **pas trop important**, **dénombrable** et **fini** (sinon il faut procéder à un regroupement par classes de valeurs).

## B - Distribution par valeurs ou par modalités

Suivant que le caractère est quantitatif ou qualitatif, 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 caractère qualitatif pour effectuer un regroupement par modalités. Au 1er janvier 2009, 16 sur les 35 pays sont dans la modalité « 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 du caractère quantitatif « 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 $km^2$ ? 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]

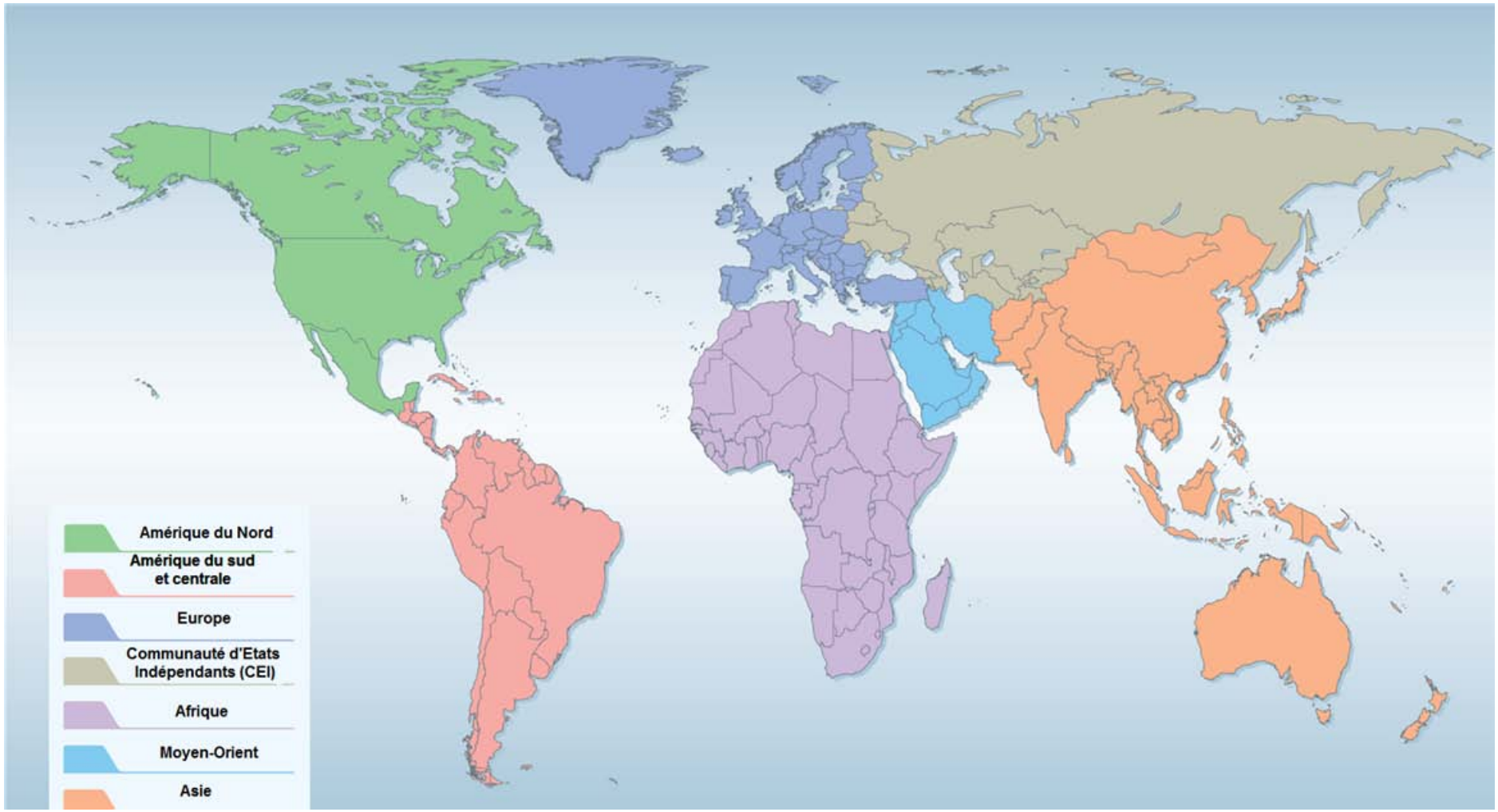
Transformation de la série simple en distribution par classes de valeurs

| 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 |

Les classes sont d'amplitudes inégales

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

## 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, 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.

## 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 se limite généralement à l'étude des populations qui comprennent un nombre fini d'unités, d'où l'analogie que l'on peut faire entre population et base de données, analogie d'autant plus pertinente que toute étude statistique commence par la constitution d'une base de données qui constitue, selon le cas, la population à étudier, ou bien un échantillon représentatif de celle-ci.

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, on divise une population en sous-ensembles au moyen de certains **caractères** ou **critères** (ou **dimensions** ou encore **caractéristiques**).

Rein n'interdit d'étudier un échantillon comme on étudie une population, en y découpant des sous-ensembles au moyen de certains caractères. Et c'est d'ailleurs ce que l'on fait lorsque la population que l'on souhaite étudier est trop importante pour être recensée de façon exhaustive et rapide. L'étude de l'échantillon se substitue alors à celle de la population, et ce sont les règles de constitution de l'échantillon, jointes aux lois de probabilité, qui permettent ensuite d'extrapoler les résultats obtenus à l'ensemble de la population.

On distingue deux sortes de caractères : les **caractères quantitatifs** et les **caractères qualitatifs**.

Les caractères quantitatifs, aussi appelés **variables**, ont des attributs. Ces attributs sont des **valeurs numériques**. Lorsque l'on peut compter ces valeurs au sein d'un intervalle, c'est que le caractère quantitatif est **discret**. Lorsque l'on ne peut pas compter ces valeurs, mais seulement les mesurer dans la population étudiée, c'est que le caractère quantitatif est **continu**.

Les caractères qualitatifs ont également des attributs appelés modalités. Lorsque l'on peut classer les modalités d'un caractère, celui-ci est dit **qualitatif ordinal**. Quand on ne peut pas ou qu'on ne veut pas[12] classer les modalités d'un caractère, celui-ci est dit **qualitatif nominal**.

Les modalités ou les valeurs des caractères 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 caractère est quantitatif) ou **distribuées par modalités** (lorsque le caractère n'est pas quantitatif). 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) :**

TD1 2008 : Sujets - Corrigés
TD1 2009 : Sujets - Corrigés

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

[12] Certains classements n'ont pas d'intérêt : par exemple, le classement alphabétique des modalités d'un caractère ne fait pas de celui-ci un caractère ordinal. D'autres classements impliquent des jugements de valeurs politiquement incorrects (par exemple, si j'étudie la couleur dans une population de fleurs et que je classe les couleurs en fonction de mes préférences, ce classement ne fait pas de la couleur un caractère qualitatif ordinal).