Introduction à la statistique
UE : Bases en statistiques - L1 MIASHS
Université de Bordeaux

Jérémie Bigot

Institut de Mathématiques de Bordeaux (IMB)
Bureau : 114 (Site de La Victoire) et 213 (Site de Talence)

e-mail : jeremie.bigot@u-bordeaux.fr

Annnée 2016 - 2017

Comment définir la Statistique ?

Comment définir la Statistique ?

- Une production de l'esprit humain à partir de l'observation ?
- La statistique, c'est calculer des moyennes à partir d'un tableau de chiffres...
- La statistique est-elle une discipline des mathématiques liée au probabilités ?

Comment définir la Statistique ?

- Une production de l'esprit humain à partir de l'observation ?
- La statistique, c'est calculer des moyennes à partir d'un tableau de chiffres...
- La statistique est-elle une discipline des mathématiques liée au probabilités ?
- La statistique peut-elle se définir à partir de son utilisation dans un métier ?

  Le statisticien a pour profession la mise au point et l'utilisation d'outils statistiques... S'agit-il d'un métier à part entière ?

Statisticien, un métier pour l'avenir ?

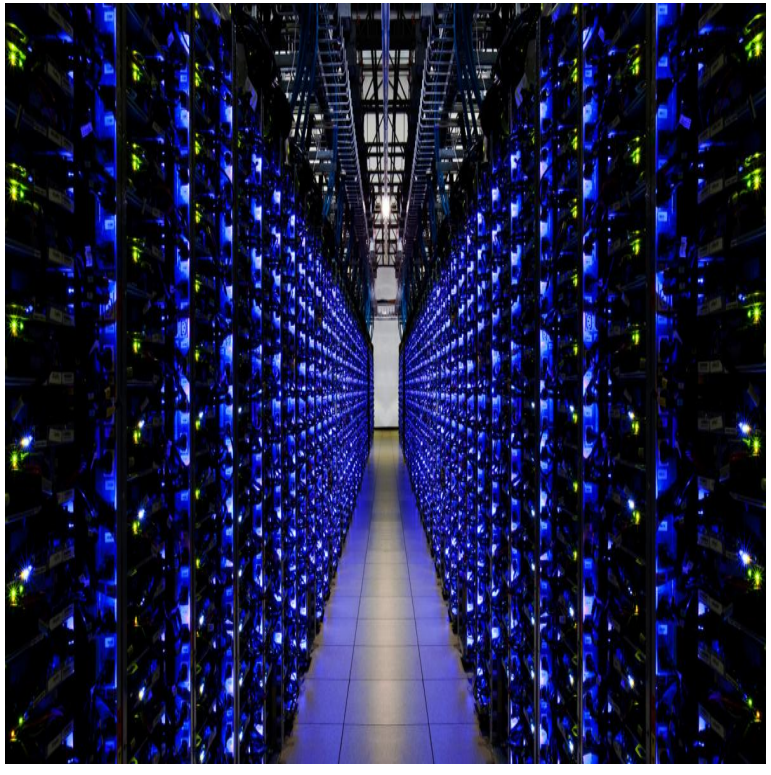
En 2009, Hal Varian, chef économiste de Google, déclarait :

Being a statistician is the sexiest job of the 21st century !

Le métier de statisticien sera le plus sexy du 21ème siècle !

D'où vient cette affirmation qui semble indiquer le fait que notre société a besoin de la statistique ?

Que représentent ces photos ?



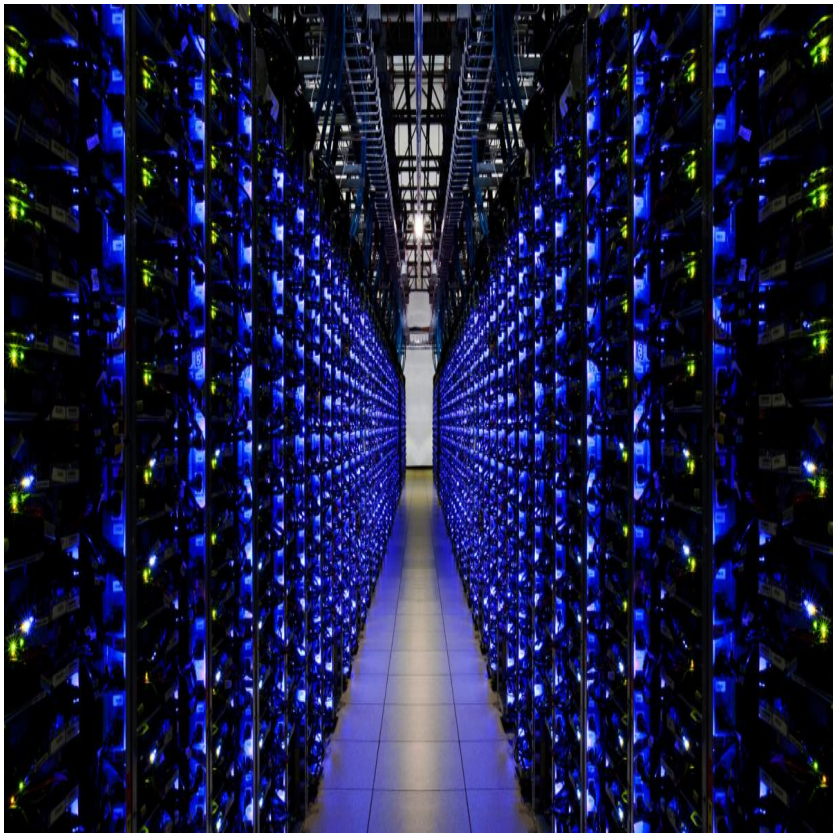
Que représentent ces photos ?

©T.EK/CASCADE MEDIA/NCC

Vue aérienne d'un des datacenters de Facebook implanté en Suède

Source : `https://lejournal.cnrs.fr/`

Que représentent ces photos ?



©GOOGLE

Vue intérieure d'un des datacenters de Google

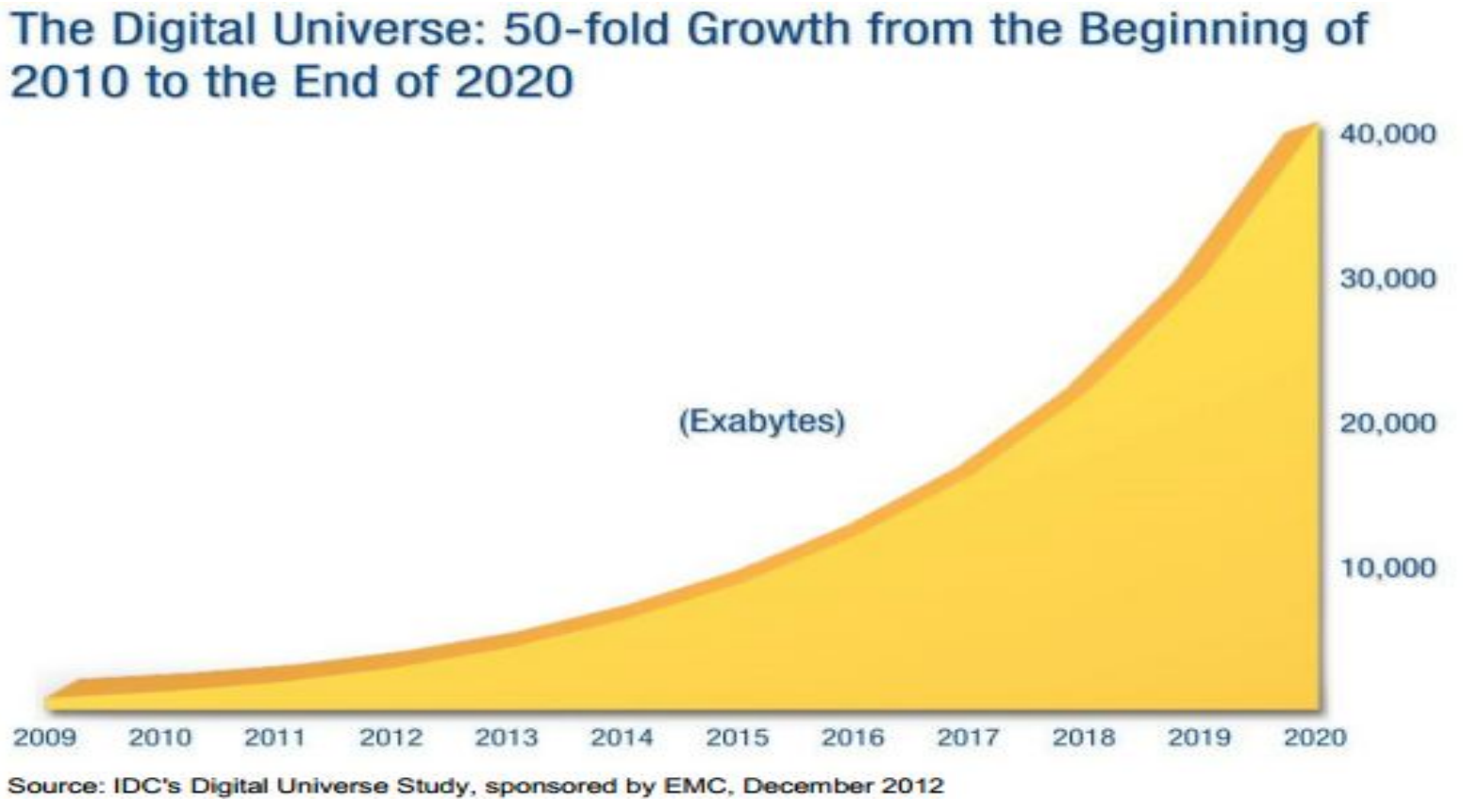
Source : `https://lejournal.cnrs.fr/`

Un volume croissant de données dans le monde

- Chaque année, l'équivalent de 912 500 000 000 000 000 000 octets d'informations sont publiées dans le monde soit 912,5 exaoctets (10^{18} octets) par an selon une estimation faite vers mi-2012 par IBM.
- De 2005 à 2020, la masse de données digitales va passer de 130 exabytes à 40,000 exabytes, soit 40 trillions de gigabytes (ou plus de 5.200 gigabytes pour chaque être humain en 2020).

Source : `http://www.planetoscope.com`
Site de statistiques mondiales en temps réel

Le siècle du Big Data ?



Question : comment traiter cette masse de données et en extraire de l'information ?

Réponse : besoin de la Statistique !

1 Un rapide panorama de la Statistique

2 Notion de variable quantitative

3 Notion de variable qualitative

4 Les enjeux actuels de la statistique

5 Organisation de l'UE et ressources sur la statistique

Tentative de définition

Le terme **"La Statistique"** (en Anglais **statistics**) désigne l'ensemble des sciences et méthodes d'analyse de **données** (en Anglais **data**) relatives à un groupe **d'individus ou d'unités**.

La Statistique comprend

- le recueil / collecte/ enregistrement des données
- **le traitement et l'interprétation des données collectées**
- **la présentation (représentations graphiques) de l'information issue de ces données**
- l'aide à la décision

Différentes composantes

La statistique est un domaine des mathématiques qui possède deux grandes tendances :

- **la statistique mathématique** : une composante théorique qui s'appuie sur la théorie des probabilités
- **la statistique appliquée** : une composante liée à l'analyse de données qui est utilisée dans de nombreux domaines de l'activité humaine

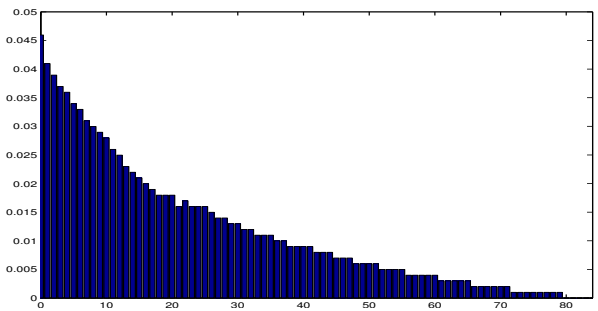
Domaines d'application de la statistique

Démographie : le recensement permet, à un instant donné, de connaître la composition d'une population

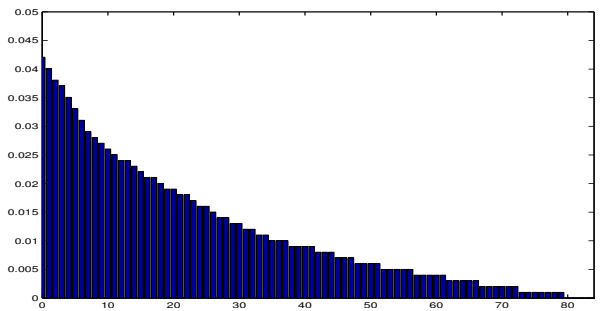
Domaines d'application de la statistique

Démographie : répartition par âge des individus dans un pays donné pour l'année 2000 (âge entre 0 and 84).

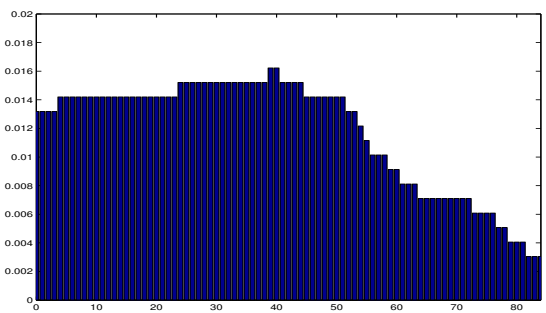
IPC. 2000. International Data Base. International Programs Center. US Census Bureau. URL : http ://www.census.gov/ipc/www/idb/



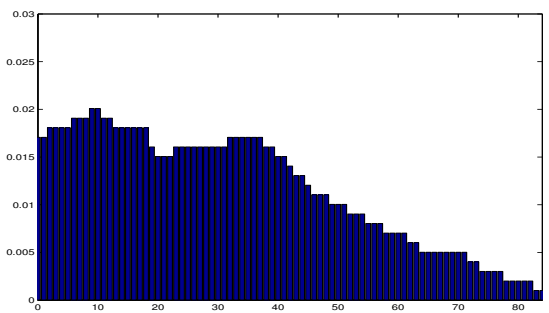
(a) Afghanistan



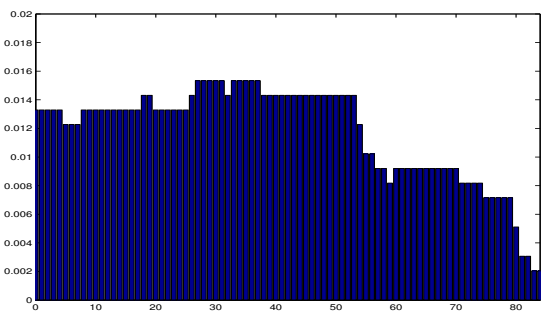
(b) Angola



(c) Australie



(d) Chili



(e) France

Données disponibles pour 217 pays sur plusieurs années.

Domaines d'application de la statistique

Sciences économiques et sociales, et économétrie : étude de l'évolution d'un secteur économique ou bien d'un groupe d'individus - Travaux de l'INSEE (Institut national de la statistique et des études économiques, créé en 1946)

Assurance et finance : utilisation de la statistique pour le calcul de risque ou bien de prix.

Finance et économétrie : étude et prévision/estimation de l'évolution de l'indice S&P 500.

Domaines d'application de la statistique

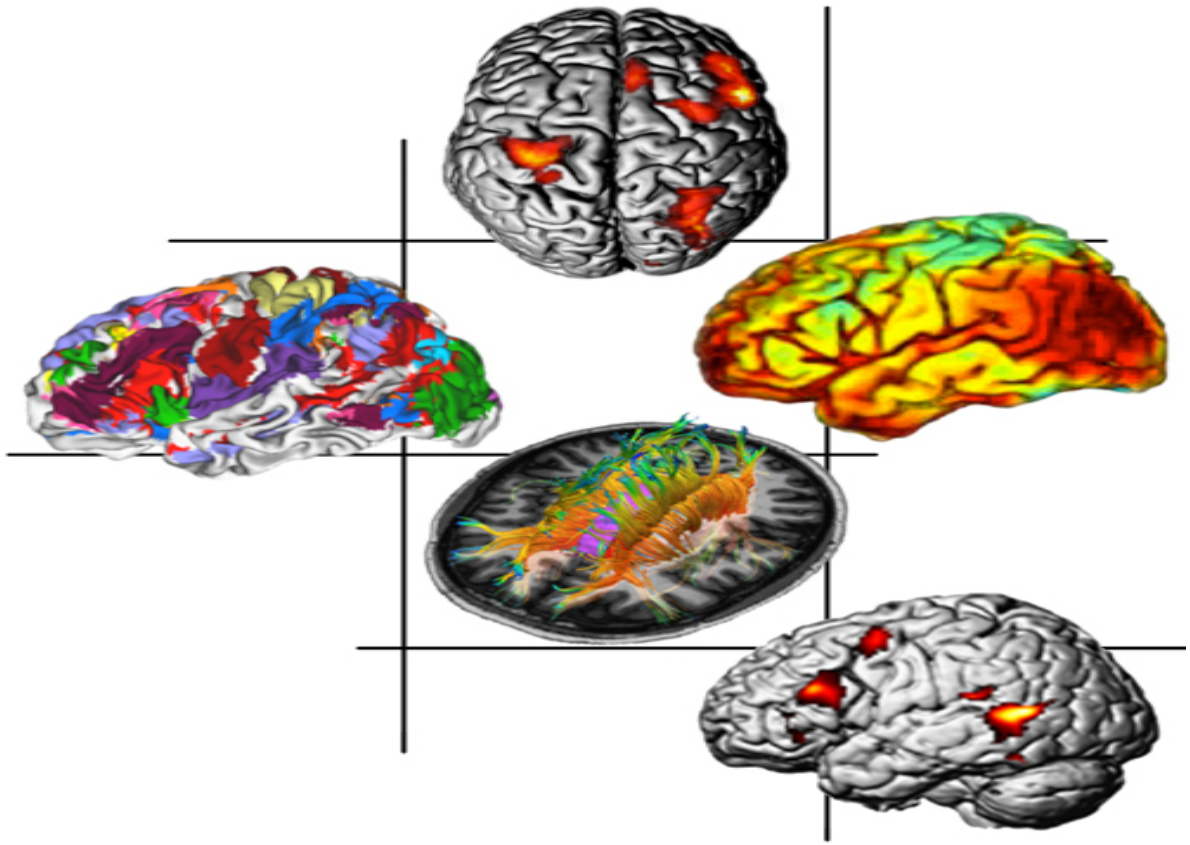
Sondage d'opinion : outil très présent dans le marketing, les médias, et le monde politique.

Domaines d'application de la statistique

Médecine, biologie, psychologie, sciences cognitives : étude de l'évolution de maladies, validation d'un nouveau traitement, bio-informatique, génomique, étude du fonctionnement cérébral.

Domaines d'application de la statistique

Sciences cognitives : images 3D d'activation du cerveau



Source : Logo du Groupe d'Imagerie Neurofonctionnelle (GIN, UMR 5296), Bordeaux - `http://www.gin.cnrs.fr`

Domaines d'application de la statistique

Météorologie : prévision du temps et de l'évolution du climat.

Et bien d'autres encore !

Quelques notions de base

- Echantillon et Population
- Statistique descriptive
- Statistique inférentielle

Echantillon et Population

- **Population** : ensemble d'unités ou d'individus étudiés par le statisticien
  Exemple : la population des individus vivant en France.
- **Echantillon** : ensemble issu de la population à la disposition du statisticien
- Si l'échantillon = population toute entière : statistique descriptive
- A partir d'un échantillon, on souhaite en déduire des conclusions sur la population tout entière : statistique inférentielle.

Statistique descriptive

- **Définition** : description d'un ensemble de données
- **La statistique descriptive, c'est de la communication !**

Statistique inférentielle

- **Définition** : la statistique inférentielle, c'est la science qui permet de tirer des conclusions sur une population à partir d'un échantillon issu de cette population.

- **Attention** : il existe toujours un biais de l'échantillon par rapport à la population qu'il convient d'essayer de corriger !

Un exemple

Une étude portant sur 200 personnes indique que 50 d'entre elles portent des lunettes :

- 25 % des personnes de l'échantillon portent des lunettes : il s'agit de **statistique descriptive**
- on en conclut que le % de personnes qui portent des lunettes dans l'ensemble de la population est compris entre 23 et 27 % : il s'agit de **statistique inférentielle**

Objectifs de la statistique descriptive

- Pour elle-même, c'est à dire la description d'un échantillon = population (**ce cours**)
- Apprendre à ordonner, présenter de façon claire, simplifier et abstraire de l'information à partir d'un ensemble de données - **Importance des représentations graphiques !**
- Etape indispensable avant une modélisation statistique de la loi de l'échantillon et de faire de l'inférence sur la population (**plus tard dans votre formation**)

Etude d'un échantillon

- Que ce soit dans un but descriptif ou inférentiel, la statistique passe par **l'étude de données** d'un échantillon

- Les **données statistiques** se présentent sous la forme d'**individus** ou d'**unités** pour lesquels sont mesurés un certain nombre de **caractères** appelés plus généralement **variables**

Exemple : Etude de 4 individus notés **1, 2, 3 et 4** pour lesquels on observe les caractères (ou variables) **Sexe, Nombre de frères et soeurs, Taille (cm)**

Etude d'un échantillon

- Ces données se présentent généralement sous la forme d'un tableau avec en **lignes les individus ou unités** et en **colonnes les variables**

Exemple : tableau de données à 4 individus et 3 variables :

	Sexe	Nombre de frères et soeurs	Taille(cm)
1	M	1	143.2
2	M	2	149.6
3	F	0	144.4
4	F	2	146.8

Quelques notations

Soit deux **caractères statistiques** ou **variables** notés X, Y.

Un **échantillon** pour ces deux variables peut se représenter par la donnée de deux suites de nombres :

$$\{x_1, x_2, \ldots, x_n\}, \text{ et } \{y_1, y_2, \ldots, y_n\},$$

où :

- n est le nombre d'observations (ou encore la **taille de l'échantillon**)
- $\forall i,\ 1 \leq i \leq n$, x_i représente la valeur de la variable X pour l'individu i,
- $\forall i,\ 1 \leq i \leq n$, y_i représente la valeur de la variable Y pour l'individu i.

Les données présentées sous cette forme sont appelées les **données brutes**.

Quelques notations

Soit deux **caractères statistiques** ou **variables** notés X, Y.

Un **échantillon** pour ces deux variables peut se représenter par la donnée de deux suites de nombres :

$$\{x_1, x_2, \ldots, x_n\}, \text{ et } \{y_1, y_2, \ldots, y_n\}.$$

On peut également représenter ces données sous la forme d'un tableau :

Individu	X	Y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

Différents cadres d'étude statistique

Pour un tableau de données à n individus et k variables, on peut étudier les données de plusieurs façons :

- **1 variable à la fois** $\rightarrow$ statistique descriptive à 1 variable, étude d'un caractère statistique
- **2 variables à la fois** $\rightarrow$ statistique descriptive à 2 variables, étude **simultanée** de 2 caractères statistiques
- **+ de 2 variables à la fois** $\rightarrow$ statistique exploratoire, analyse des données (méthodes plus lourdes et plus complexes, calculs intensifs).

Un peu de vocabulaire

L'ensemble des individus / sujets / unités constitue un **échantillon** (ou encore une **série statistique**), formant ainsi un sous-ensemble d'un groupe (beaucoup) plus grand appelé **population**.

Les caractères statistiques ou variables peuvent être de plusieurs natures. On distingue en particulier :

- Les variables **quantitatives**
- Les variables **qualitatives**

1 Un rapide panorama de la Statistique

2 Notion de variable quantitative

3 Notion de variable qualitative

4 Les enjeux actuels de la statistique

5 Organisation de l'UE et ressources sur la statistique

Définition

Les **variables quantitatives** sont celles que l'on peut mesurer et pour lesquelles l'addition a un sens. On distingue :

- Les **variables quantitatives discrètes** : variables prenant leurs valeurs dans un ensemble fini ou dénombrable (en général $\mathbb{N}$).

  Exemples : âge d'une personne (en années), nombre d'enfants dans une famille.

- Les **variables quantitatives continues** : variables qui peuvent prendre **n'importe quelle valeur** dans un intervalle de $\mathbb{R}$, c'est à dire mesurées avec une très grande précision.

  Exemples : taille d'un individu (en cm), poids (en g), temps d'une réaction chimique (en micro-secondes).

Exemple de variables quantitatives

Exemple : tableau de données à 4 individus et 3 variables :

	Sexe	**Nombre de frères et soeurs**	**Taille(cm)**
1	M	1	143.2
2	M	2	149.6
3	F	0	144.4
4	F	2	146.8

Nombre de frères et soeurs : variable quantitative **discrète**

Taille(cm) : variable quantitative **continue**

1 Un rapide panorama de la Statistique

2 Notion de variable quantitative

3 Notion de variable qualitative

4 Les enjeux actuels de la statistique

5 Organisation de l'UE et ressources sur la statistique

Définition

Les **variables qualitatives** sont celles qui sont... non-quantitatives !

Exemples : nationalité d'une personne, profession, degré de satisfaction, stade de maladie, couleur de cheveux, vote, origine...

Les valeurs possibles d'une variable qualitative sont appelées les **modalités**.

Exemples de modalités :

- nationalité d'une personne : "Français", "Britannique"
- profession : "Plombier", "Médecin"

Exemple de variable qualitative

Exemple : tableau de données à 4 individus et 3 variables :

	Sexe	**Nombre de frères et soeurs**	**Taille(cm)**
1	M	1	143.2
2	M	2	149.6
3	F	0	144.4
4	F	2	146.8

Sexe : variable qualitative

Modalités : M (Masculin) et F (Féminin)

1 Un rapide panorama de la Statistique

2 Notion de variable quantitative

3 Notion de variable qualitative

4 Les enjeux actuels de la statistique

5 Organisation de l'UE et ressources sur la statistique

Quel rôle pour la statistique au 21ème siècle ?

Les nouvelles technologies d'acquisition de données conduisent à des échantillons de taille de plus en plus importante où le :

- nombre n d'individus / unités est très grand
- nombre k de variables est très grand

L'analyse statistique de ce type de données posent des défis à la fois d'ordre

- informatique, numérique (stockage/lecture des données)
- mais également d'ordre méthodologique : les outils usuels de la statistique (développé lors du 20ème siècle) ne sont plus adaptés, et il s'agit d'en inventer de nouveaux.

Quel rôle pour la statistique au 21ème siècle ?

La Statistique est donc actuellement une discipline extrêmement vivante et en plein essor !

- Nombreux travaux de recherche - de nouvelles méthodologies statistiques à mettre en place pour de nouveaux types des données.
- Opportunités d'emploi en hausse

Le défi lancé par Netflix aux statisticiens

Qu'est-ce que Netflix ?

Le défi lancé par Netfilx aux statisticiens

Netflix est une entreprise américaine qui offre un service de location de films et séries TV sur Internet. Elle est principalement localisée aux Etats-Unis, Canada, Amérique du Sud, et dans quelques pays d'Europe.

Après la location d'un film, il est généralement demandé à l'utilisateur de donner son avis sur le film au travers d'une note (par exemple entre 1 et 5).

Il s'agit d'une demande très classique de la part de nombreuses entreprises (par exemple Amazon)....

Quel est l'intérêt pour Netflix de recueillir l'avis de ses clients sur les films qu'ils louent ?

Le défi lancé par Netfilx aux statisticiens

Clients					
M. Untel	1	?	2	5	?
Mme Dupont	1	3	4	?	2
Mme Henry	5	?	3	?	?
M. Georges	5	5	2	5	?

Est-il possible de prédire la note que donnerait un client à un film qu'il n'a pas encore vu ?

Le défi lancé par Netfilx aux statisticiens

Entre 2006 et 2009, Netflix a proposé un défi visant à améliorer les performances de son propre algorithme de construction d'un système de recommandation de films à partir des opinions et évaluations de leurs clients sur les films qu'ils ont déjà vus (principe des méthodes dites de **filtrage collaboratif**).

Le défi lancé par Netfilx aux statisticiens

Les données fournies par Netflix : un tableau contenant

- $n = 480189$ **clients** (lignes) - nombre d'individus
- $k = 17770$ **films** (colonnes) - nombre de variables
- 100 480 507 résultats de votes

Compétition ouverte à tous ! But : mettre au point un algorithme capable de prédire les notes manquantes. Cet algorithme est évalué sur un ensemble test dont seul Netflix connait les vraies notes ! Il s'agit un problème de **Statistique Inférentielle**.

En juin 2007 plus de 20 000 équipes s'étaient enregistrées pour la compétition originaires de plus de 150 pays, et 2 000 équipes ont proposé plus de 13 000 possibilités de prédiction (source Wikipedia) !

Le défi lancé par Netfilx aux statisticiens

Les données fournies par Netflix : un tableau contenant

- $n = 480189$ **clients** (lignes) - nombre d'individus
- $k = 17770$ **films** (colonnes) - nombre de variables
- 100 480 507 résultats de votes

Un peu de **Statistique Descriptive** sur ce jeu de données :

- pourcentage de films notés : $\frac{100480507}{480189 \times 17770} \approx \frac{100.10^6}{8.10^9} = \approx 1.18\%$
- le nombre de films notés en moyenne par client est de 200
- le nombre moyen de votes par film est de 5000
- certains films sont notés moins de 3 fois
- un client a noté plus de 17 000 films !

Le défi lancé par Netfilx aux statisticiens

Les données fournies par Netflix : un tableau contenant

- $n = 480189$ **clients** (lignes) - nombre d'individus
- $k = 17770$ **films** (colonnes) - nombre de variables
- 100 480 507 résultats de votes

En 2009, le défi a finalement été remporté par l'équipe "BellKor's Pragmatic Chaos" qui a fourni le taux d'erreur de prédiction le plus faible (grand prix de 1 million de dollars le 21/09/2009).

1 Un rapide panorama de la Statistique

2 Notion de variable quantitative

3 Notion de variable qualitative

4 Les enjeux actuels de la statistique

5 Organisation de l'UE et ressources sur la statistique

Répartition cours/TD de l'UE

- 5 amphis de 2h
  - Notions de données, échantillon, variables
  - Statistique descriptive à une ou deux variables
- Notes de cours à compléter :
  - Statistique descriptive (34 pages)

  Où les trouver ?

  `https://sites.google.com/site/webpagejbigot/enseignements`
- 9 séances de TD de 2h
  - Analyse de données avec Excel (séances sur machine)
  - Exercices applicatifs
- 1 DS **- 1/3 note finale**
- 1 examen terminal écrit **- 2/3 note finale**

Ressources documentaires sur la statistique

- Société française de statistique (SFDS) :

SOCIETE FRANÇAISE
DE STATISTIQUE

`http://www.sfds.asso.fr`

- Brochure sur les métiers de la statistique (ONISEP / SFDS)

`http://www.sfds.asso.fr/images/zoom-statistique-2011.pdf`

Références bibliographiques

- Livres : G. Saporta, Probabilités, Analyse des Données et Statistique, 3e édition révisée, Technip, 2011.
- WikiStat : `http://wikistat.fr`