جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

كلية العلوم والتقنيات
FACULTÉ DES SCIENCES ET TECHNIQUES

BIOSTATISTIQUE

Pr MOUILLY MUSTAPHA

Année Universitaire 2019/2020

جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

كلية العلوم والتقنيات
FACULTÉ DES SCIENCES ET TECHNIQUES

Généralités

Partie I

Probabilité & Distributions Théoriques

Partie II

Statistique Descriptive

Partie III

Statistique Inférentielle

Partie VI

Plan Expérimental/Tests Statistiques

Coronavirus COVID-19 Global Cases by Johns Hopkins CSSE

Total Confirmed
79,517

Confirmed Cases by Country/Region
- 77,150 Mainland China
- 833 South Korea
- 691 Others
- 215 Italy
- 147 Japan
- 89 Singapore
- 79 Hong Kong
- 61 Iran
- 35 Thailand
- 35 US
- 30 Taiwan
- 22 Australia
- 22 Malaysia
- 16 Germany
- 16 Vietnam
- 13 United Arab Emirates
- 13 UK

Country/Region | City, St/Prov

Last Updated at (M/D/YYYY)
2/24/2020 12:13:10 p.m.

Lancet Article: Here. Mobile Version: Here. Visualization: JHU CSSE. Automation Support: Esri Living Atlas team.
Data sources: WHO, CDC, ECDC, NHC and DXY. Read more in this blog. Contact US.
Downloadable database: GitHub: Here. Feature layer: Here.
Point level: City level - US, Canada and Australia; Province level - China; Country level - other countries.
Time Zones: lower-left corner indicator - your local time; lower-right corner plot - UTC.

Total Deaths
2,626

- 2,495 deaths **Hubei** Mainland China
- 19 deaths **Henan** Mainland China
- 12 deaths Iran
- 12 deaths **Heilongjiang** Mainland China
- 8 deaths South Korea
- 6 deaths **Anhui** Mainland China
- 6 deaths **Chongqing** Mainland China
- 6 deaths **Guangdong** Mainland China
- 6 deaths

Total Recovered
25,153

- 16,748 recovered **Hubei** Mainland China
- 940 recovered **Henan** Mainland China
- 786 recovered **Guangdong** Mainland China
- 782 recovered **Zhejiang** Mainland China
- 727 recovered **Hunan** Mainland China
- 663 recovered **Anhui** Mainland China
- 645 recovered **Jiangxi** Mainland China
- 453 recovered **Jiangsu** Mainland China
- 343 recovered

Actual | Logarithmic | Daily Increase

Généralités

Introduction

La Biostatistique : Définition et intérêt

Variabilité

Unité statistique

Population

Échantillon

Variables statistiques

Les mesures de bases

Conclusion

Généralités

Introduction

La biostatistique : Définition et intérêt

Variabilité

Unité statistique

Population

Échantillon

Variables statistiques

Les mesures de bases

Conclusion

Introduction

- Pourquoi un module de Biostatistique Licences LPS & LBVA ?

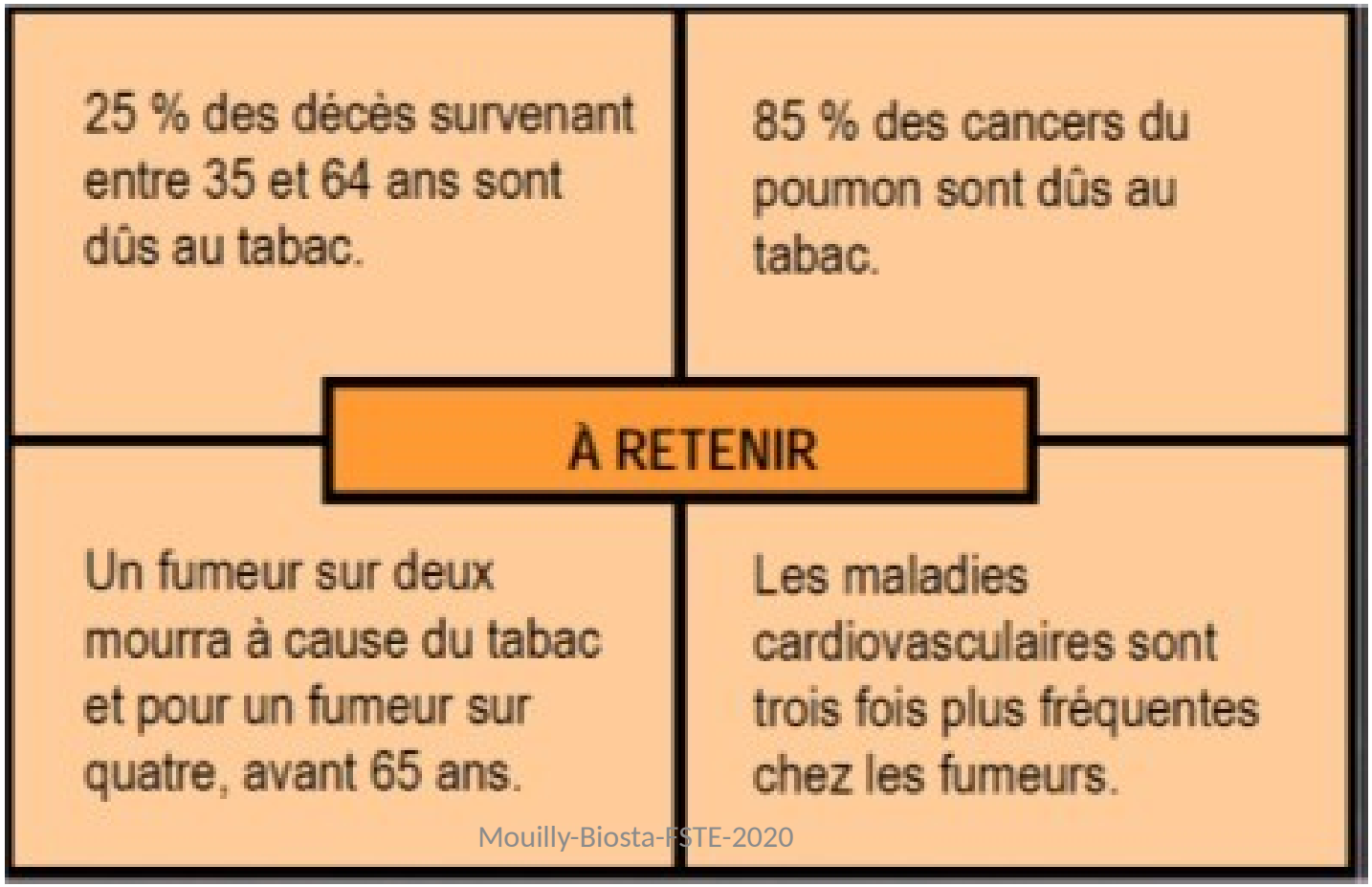
- Méthodologie statistique

- Raisonnement

- Incertitude

Introduction

- **Exemple: tabac** et **cancer** du poumon



Généralités

Introduction

La biostatistique : Définition et intérêt

Variabilité

Unité statistique

Population

Échantillon

Variables statistiques

Les mesures de bases

Conclusion

Définitions

- **Statistique** :
  - Ensemble des méthodes scientifiques consistant à **réunir**, **organiser**, **présenter** des données numériques sur une/plusieurs particularité(s) commune(s) chez un groupe de personnes ou de choses, puis à **analyser**, tirer des **conclusion** et **prendre des décisions.**

- Une statistique : un nombre calculé à partir d'**observations**.

Historique

- **Dénombrement Guerre Et Impôts**

- **Véritable début : 18^{ième} siècle**

- Première classification des causes de décès
- Calcul des probabilités (P. S. de Laplace, K. F. Gauss, S. D. Poisson)
- 1853: premier congrès

Historique (suite)

- **Première moitié du 20ième siècle**
  - Statistiques biologiques et psychologiques
    - *Biométrie et Psychométrie*
  - 1940 Recherche opérationnelle

- **Deuxième moitié du 20ième siècle**
  - Développement de l'informatique
  - Analyse des données

Domaines d'utilisation de la statistique

- Statistique officielle
- Banques – Assurances
- **Santé et Sciences de la vie**
- **Environnement (air, eau, foresterie, pêche …)**
- Sciences humaines
- Entreprises – Industrie (contrôle de qualité, études de marché, management…)
- Presse – Médias
- …

Discipline transversale

Biostatistique

- **Définition**
- Application des concepts et principes statistiques à des **données médicales**, **biologiques** et de **santé public**.

- **Exemple**
- Effet d'une plante médicinale
- le nombre de patients présentant la leishmanioses dans la région Darâa Tafilalet
- Les facteurs de risques des maladies cardiovasculaires chez la population oasienne
- …???

Biostatistique : deux branches distinctes

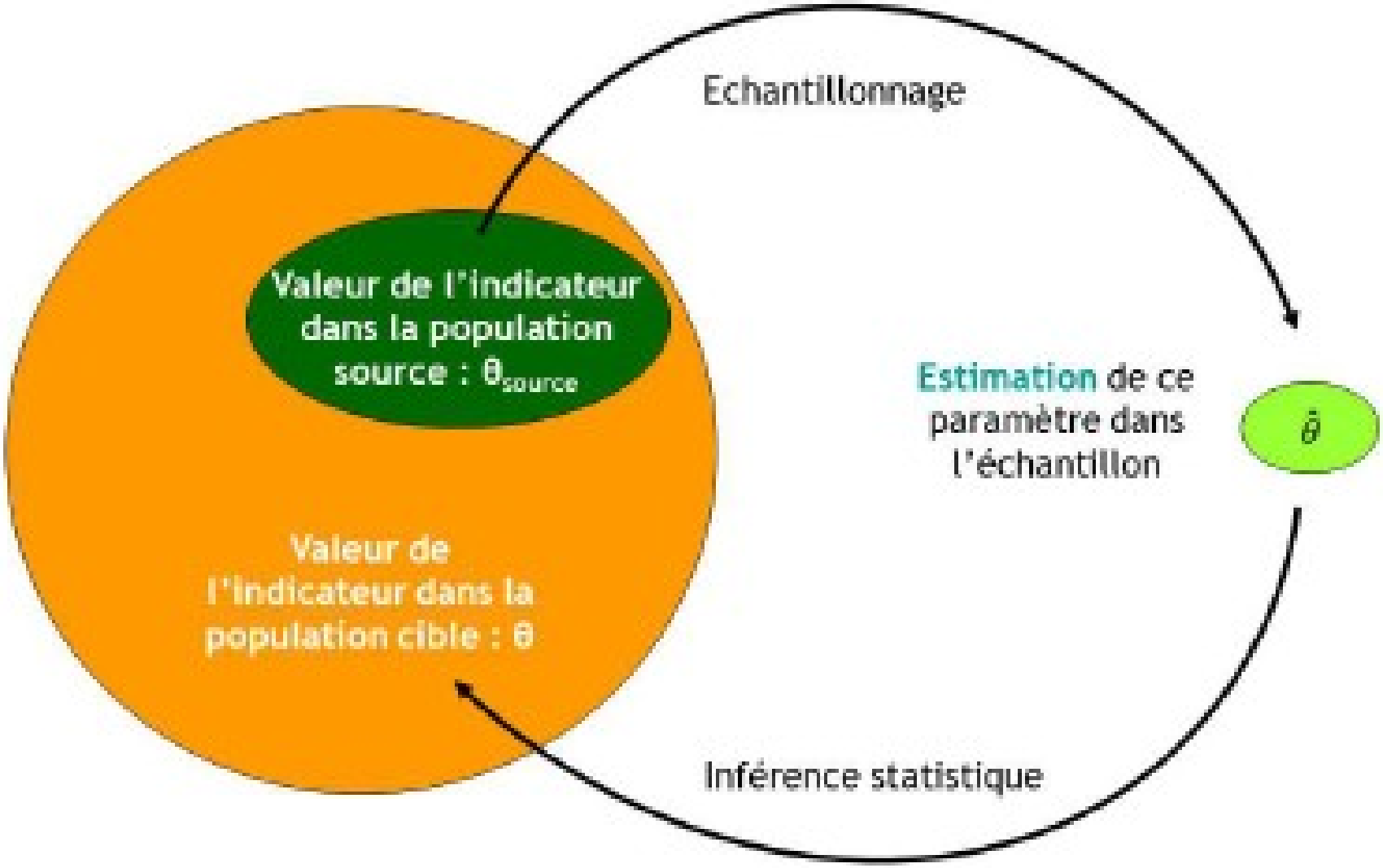
STATISTIQUE DESCRIPTIVE

- Organisation, présentation et analyse des données
- Synthèse de l'information: résumés statistiques.
- Expression des résultats : représentation graphique.
- **Étape préliminaire**

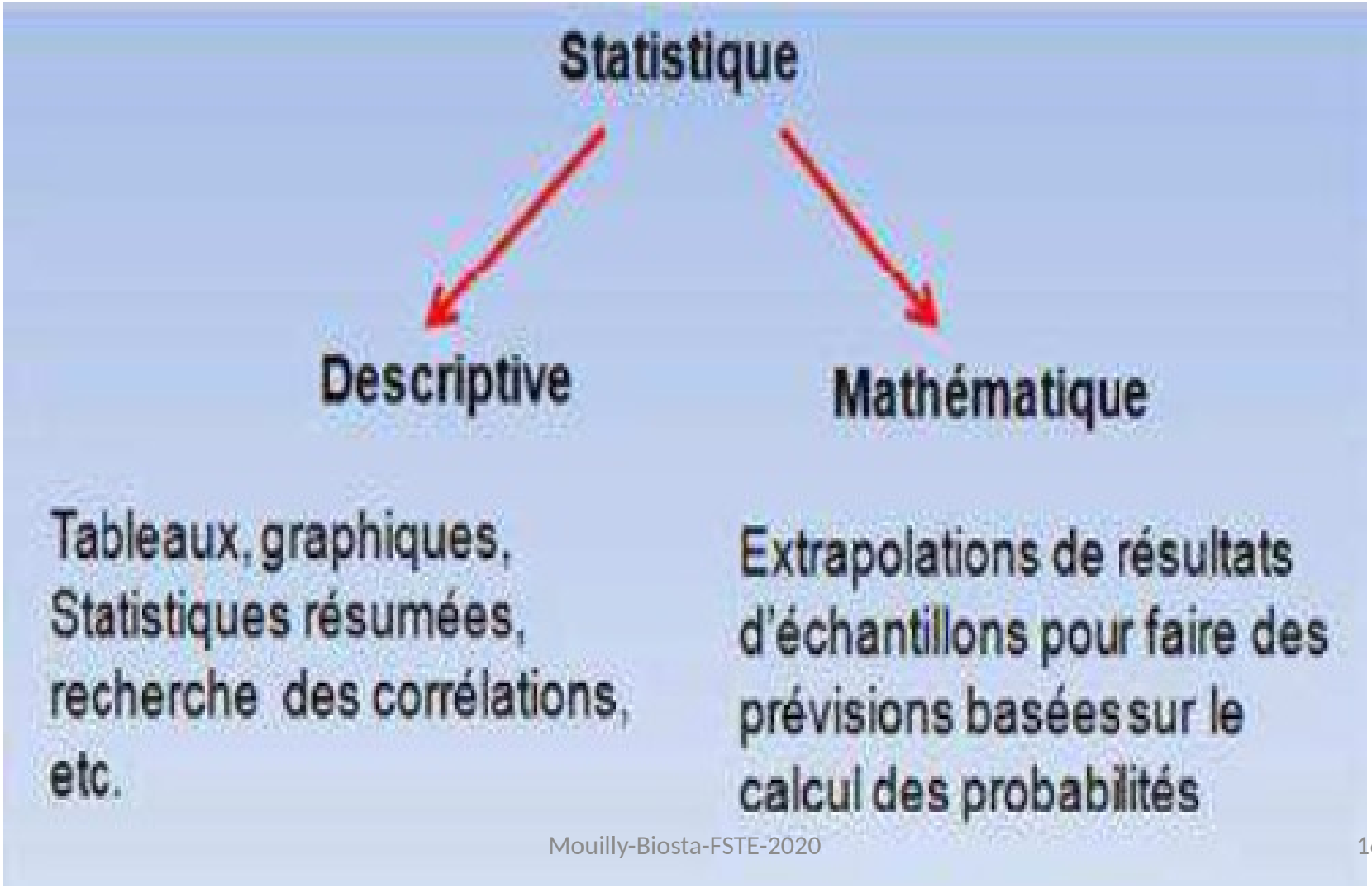
STATISTIQUE INFÉRENTIELLE

- Permet de **généraliser** à de grands ensembles d'éléments les **conclusions** tirées des résultats obtenus à partir d'un nombre réduit d'observations

•Inférer = tirer une conclusion à partir de propositions ou de faits, et de règles


Echantillonnage
Valeur de l'indicateur dans la population source : θ_{source}
Estimation de ce paramètre dans l'échantillon
$\hat{\theta}$
Valeur de l'indicateur dans la population cible : θ
Inférence statistique

Biostatistique : deux branches distinctes



A quoi sert la biostatistique?

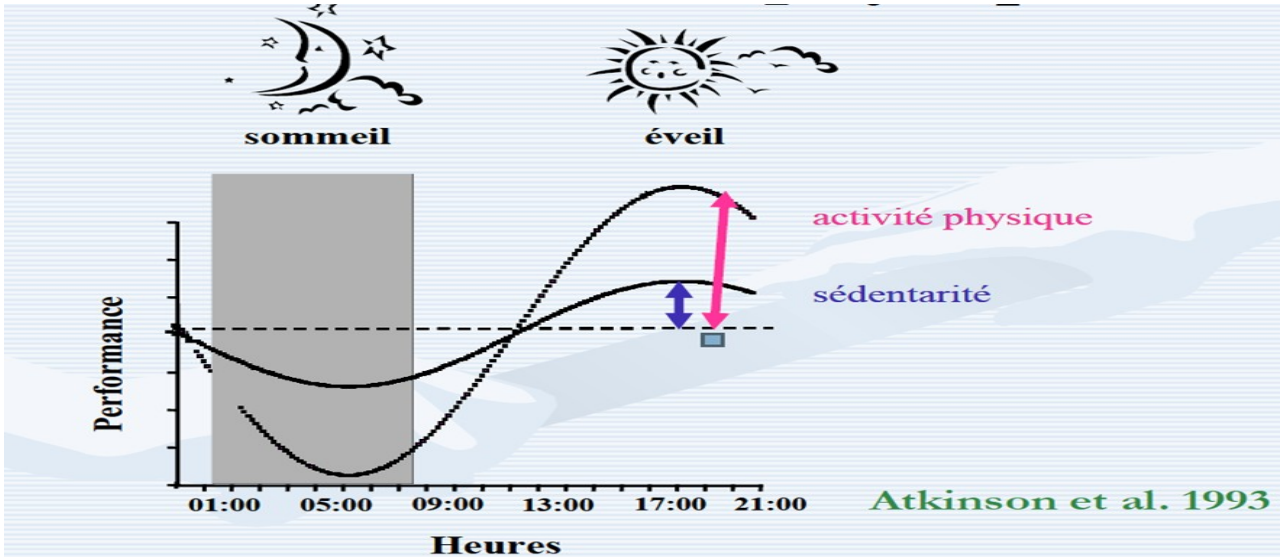
- **Gestion des observations (des données)**
  - Recueillir, présenter, analyser
- Aider à la **prise de décisions** et à la **résolution de problèmes.**

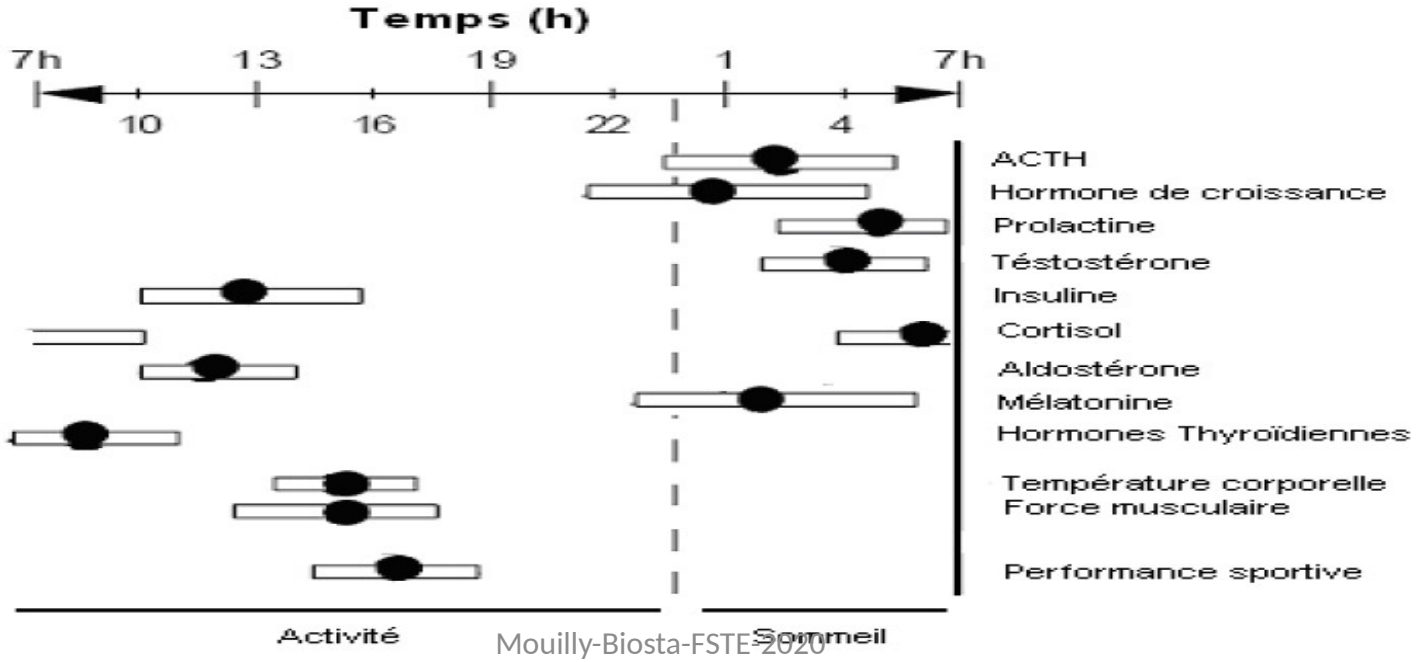
- **Comprendre et mener correctement des**
  - Expériences
  - Enquêtes
  - Travaux de recherche

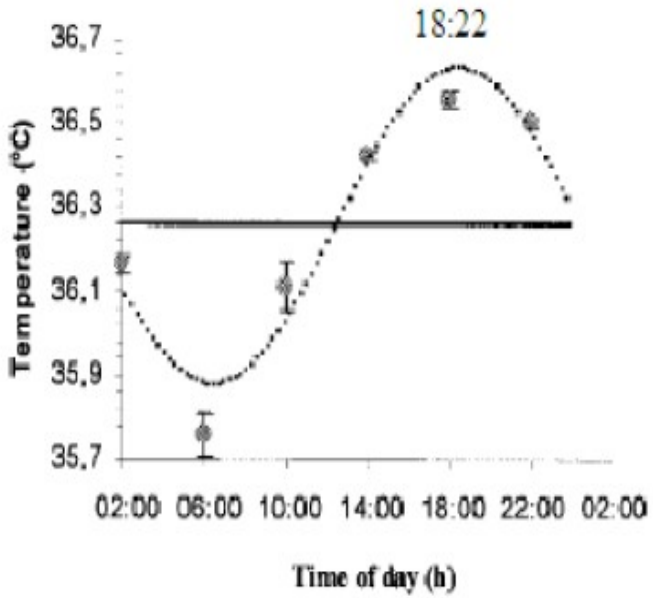
Pourquoi des statistiques en biologie?

- **Variabilité?**

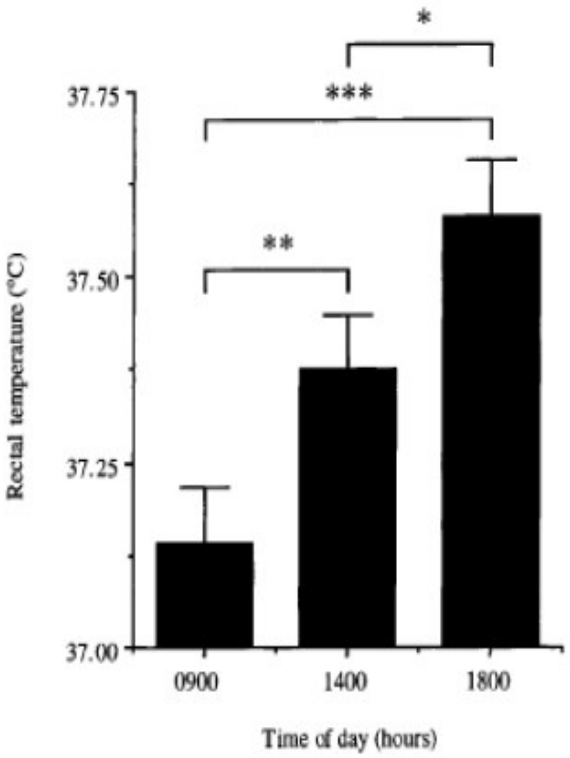
- **Échantillon?**


sommeil
éveil
activité physique
sédentarité
Performance
01:00
05:00
09:00
13:00
17:00
21:00
Heures
Atkinson et al. 1993

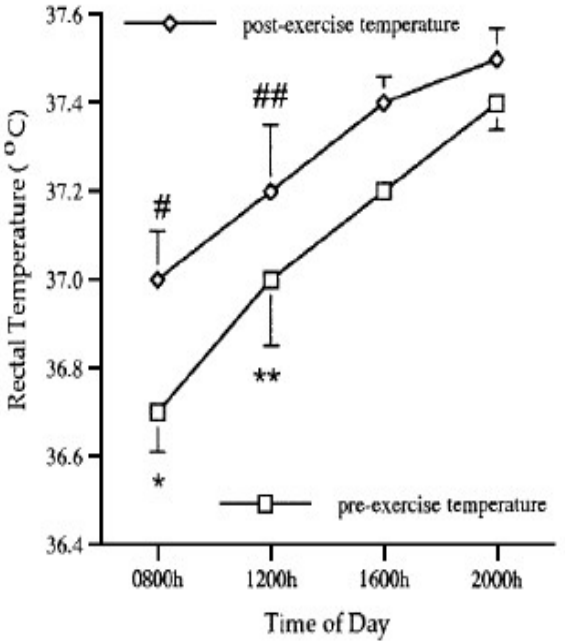

Temps (h)
7h
13
19
1
7h
10
16
22
4
ACTH
Hormone de croissance
Prolactine
Téstostérone
Insuline
Cortisol
Aldostérone
Mélatonine
Hormones Thyroïdiennes
Température corporelle
Force musculaire
Performance sportive
Activité
Sommeil



Souissi et al.(2004)



Bernard et al.(1998)



Deschenes et al. (1998)

Généralités

Introduction

La biostatistique : Définition et intérêt

Variabilité

Unité statistique

Population

Échantillon

Variables statistiques

Les mesures de bases

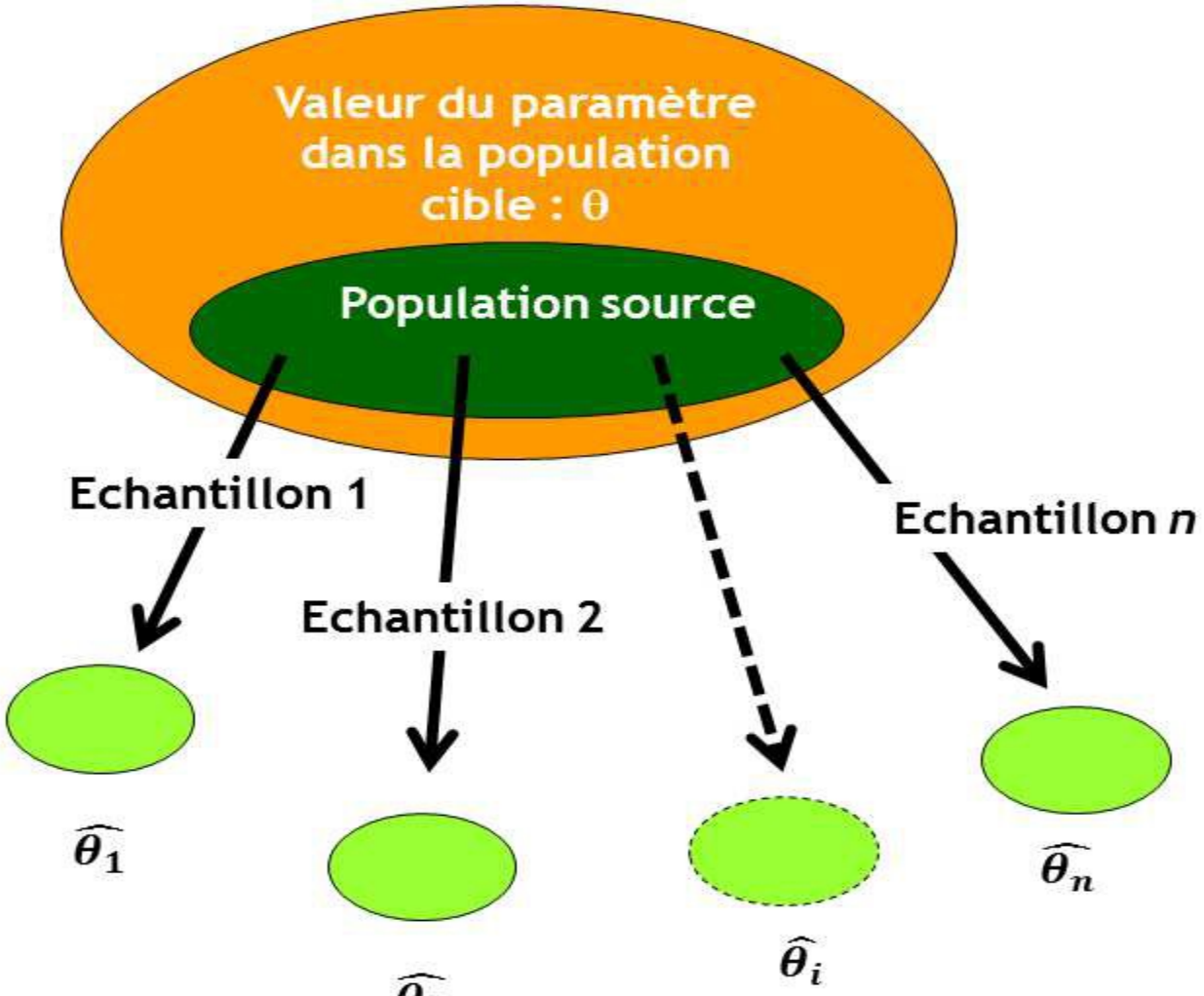
Conclusion

- Les observations obtenues pour un échantillon de taille *n=30* *figurent dans la table suivante.* ***Quel est le %....????***

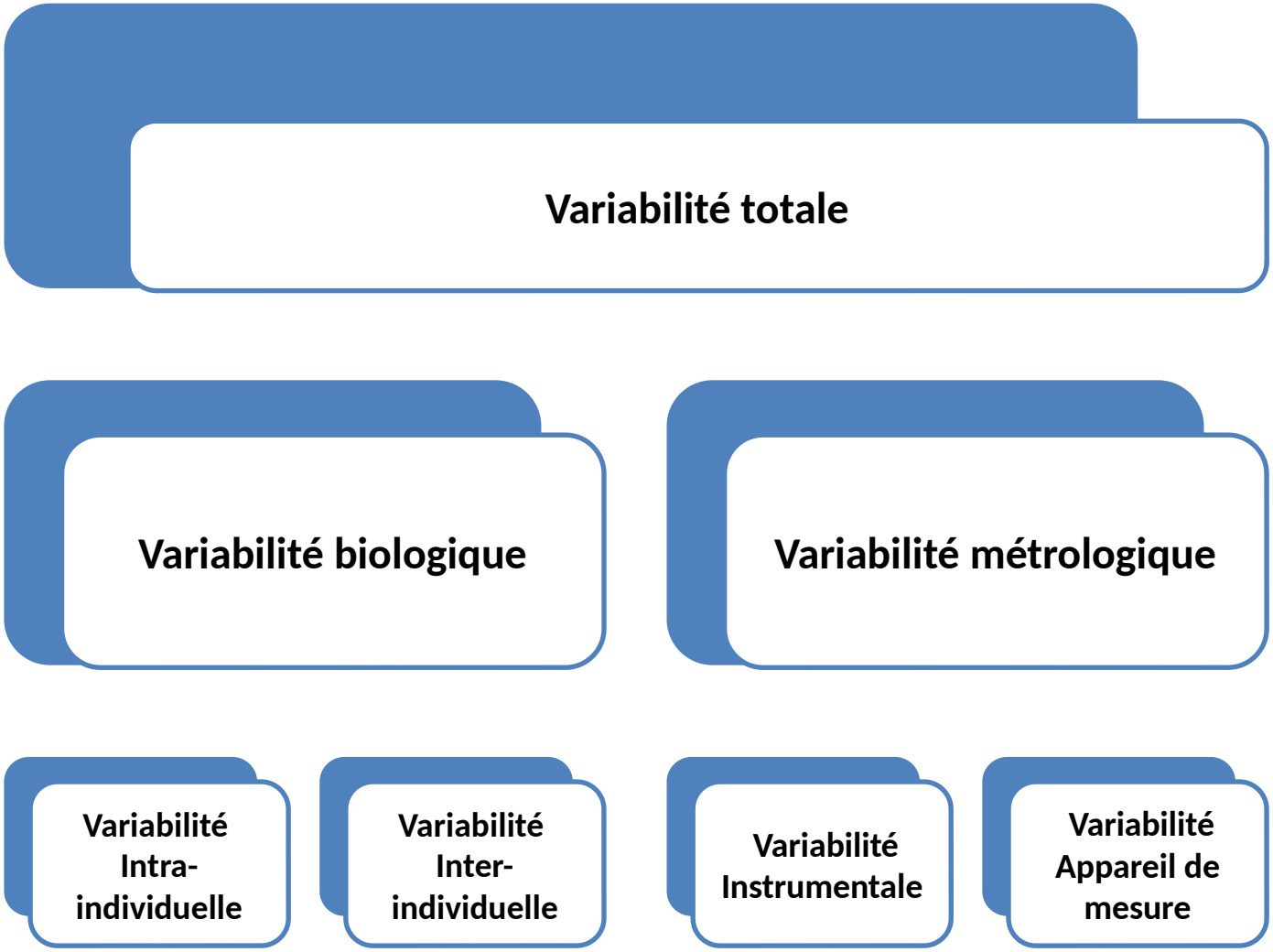
Taille	poids	sexe	frère	couleur
169	55	M	5	Marron
150	46	F	1	Noir
160	63	M	4	Noir
171	75	M	4	Noir
159	52	F	2	Marron
161	61	M	3	Marron
170	56	F	0	Noir
167	60	M	2	Bleu
164	61	M	2	Vert
163	50	M	3	Marron
164	62	F	0	Marron
150	48	F	3	Vert
151	43	M	2	Noir
167	49	F	1	Bleu

Taille	poids	sexe	frère	couleur
168	52	F	0	Noir
157	47	F	1	Noir
167	63	M	2	Marron
172	70	M	4	Noir
153	54	F	5	Marron
164	67	F	3	Noir
165	53	F	2	Marron
170	73	M	4	Bleu
160	65	M	5	Vert
161	50	M	2	Noir
162	67	F	0	Marron
158	53	F	6	Noir
157	48	F	3	Marron
164	47	F	1	Marron

Fluctuation d'échantillonnage



Variabilité



La variabilité = la règle et non l'exception

Variabilité:
Variabilité biologique

Variabilité inter-individuelle
Caractéristiques qui différent d'un individu à l'autre

Variabilité intra-individuelle
Caractéristiques évoluant dans le temps chez un même individu

La variabilité = la règle et non l'exception

Variabilité:

Variabilité métrologique

Variabilité de la mesure

Variabilité expérimentale

Variabilité appareil de mesure

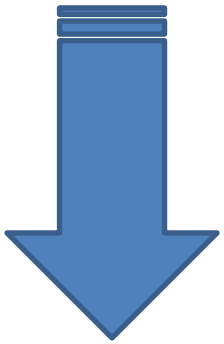
Essayer de mesurer plusieurs (100) fois la taille en mm d'un individu : vous trouverez des valeurs différentes cependant dans l'absolu un individu a une taille et une seule.

La variabilité = la règle et non l'exception

Variabilité:

Conséquences

- Il n'existe pas de « vraie valeur »
  - Valeur usuelle
- Valeur exceptionnelle, hors norme



- **Difficulté pour interpréter et utiliser les informations mesurées**

La variabilité = la règle et non l'exception

Variabilité:
Conséquences

- Caractéristique dans le domaine des sciences de la santé et sciences de la vie : **VARIABILITE**
  - Chez un même individu
  - Entre les individus
  - Entre les groupes d'individus
  - etc.

Biostatistique : Traiter les problèmes de variabilité dans les données résultantes.

La variabilité = la règle et non l'exception

Généralités

Introduction

La biostatistique : Définition et intérêt

Variabilité

Unité statistique

Population

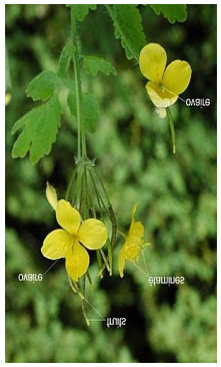
Échantillon

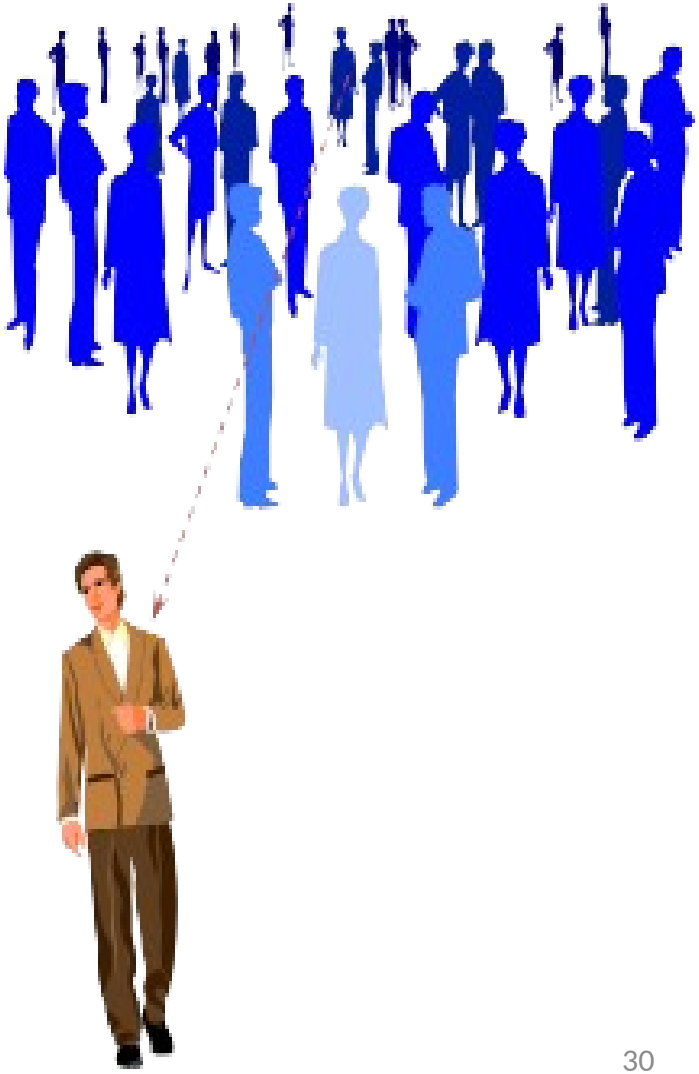
Variables statistiques

Les mesures de bases

Conclusion

Unité statistique ou individu





Unité statistique ou individu

- **Définition**
- Une unité distincte chez laquelle on peut observer une ou plusieurs caractéristiques données.
- Élément sur lequel sont effectuées des observations ou des mesures
- **Exemples**
- Individu, animal, plante, organe, cellule, champ de microscope,...

La définition de l'individu dépend des paramètres étudiés

Généralités

Introduction

La biostatistique : Définition et intérêt

Variabilité

Unité statistique

Population

Échantillon

Variables statistiques

Les mesures de bases

Conclusion

Population

- **Définition**
- Ensemble d'individus (ou unités statistiques) sur lequel on étudie une ou plusieurs caractéristiques et qui sont de même nature.
- **Exemples**
- Ensemble des vertébrés et des invertébrés dans un site
- Ensemble des étudiants inscrits à la FSTE
- ….

Population

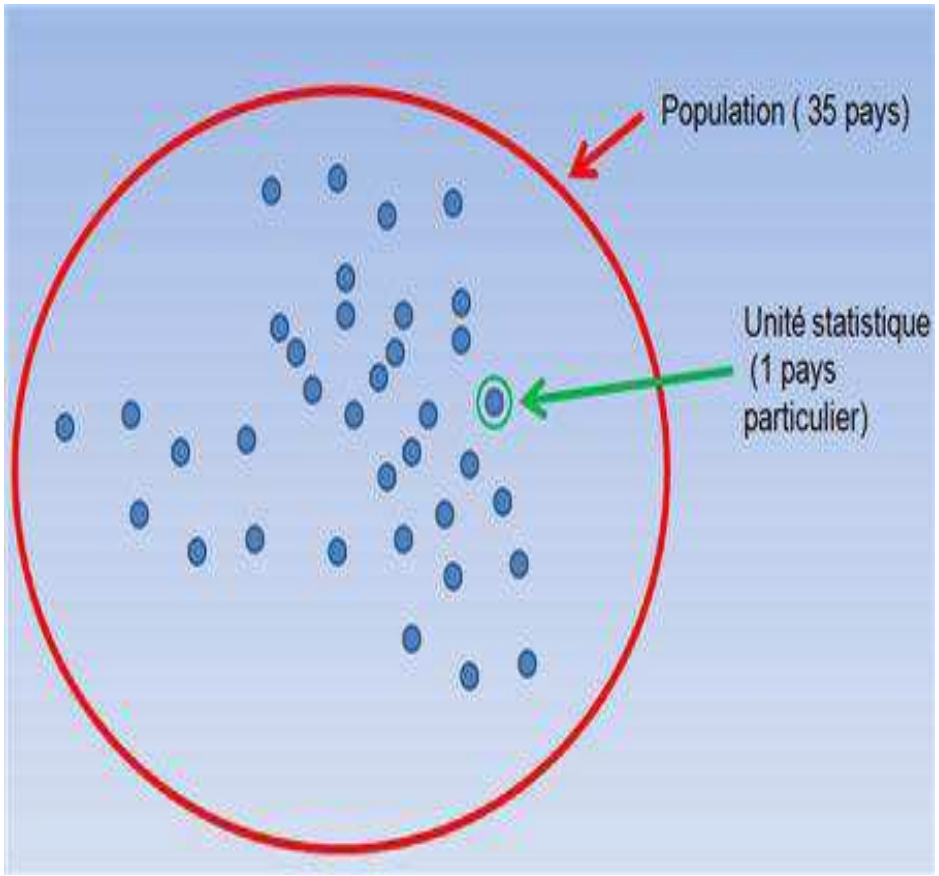
- On peut ne pas connaitre tous les individus qui composent une population.
- Une population peut être partitionnée en sous populations.

Taille de la population

- Le nombre d'individus constituant la population.
- Généralement très grande.

Notation : N

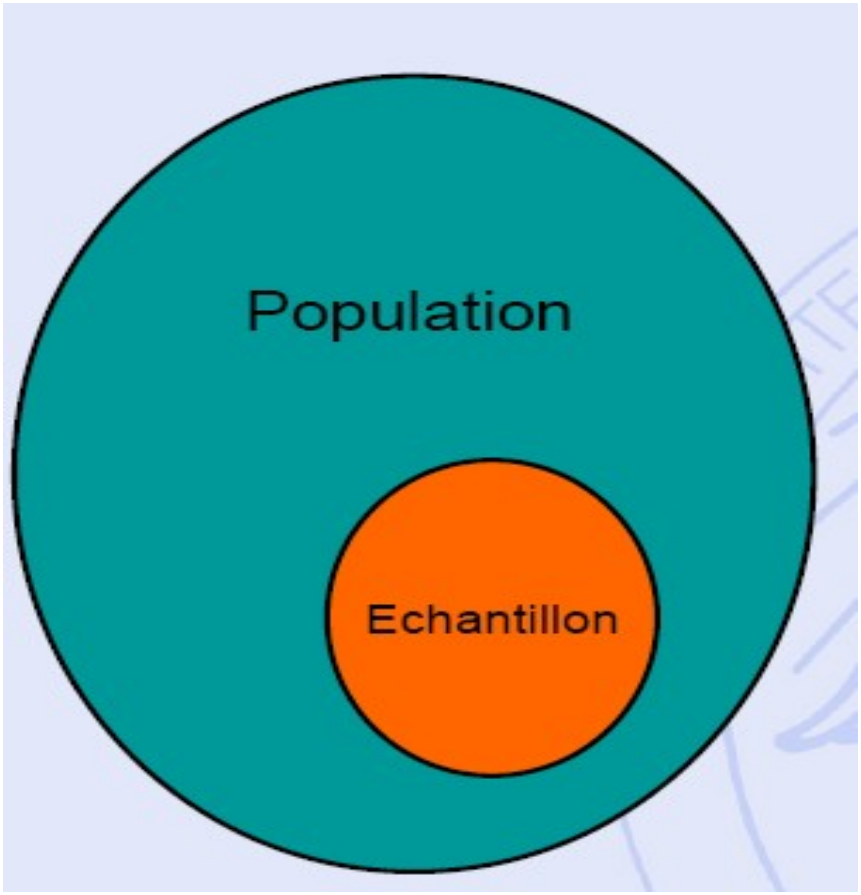
Population



- Si population **limitée** : Exhaustive –Recensement :Tous les sujets de la population sont « examinés »

- Exemple : tous les étudiants de la FSTE

Population



- Si population importante : Une partie des sujets de la population sont « examinés »

- Exemple : UN ECHANTILLON d'étudiants

Généralités

Introduction

La biostatistique : Définition et intérêt

Variabilité

Unité statistique

Population

Échantillon

Variables statistiques

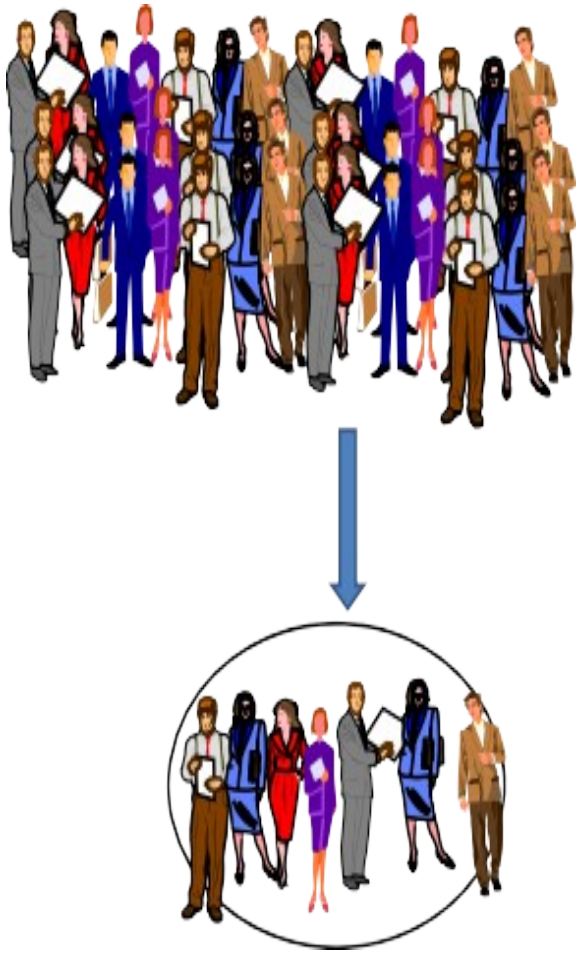
Les mesures de bases

Conclusion

Échantillon

1.Définition

- Sous ensemble d'une population et qui est de de taille finie



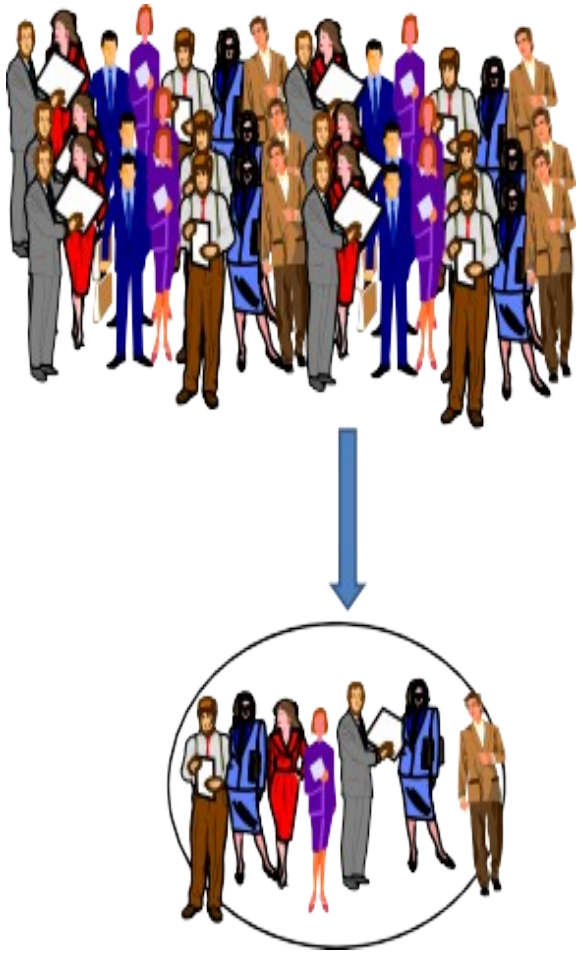
2.Exemples

- 100 NN atteints de paralysie cérébrale

- 60 étudiants de la FSTE

Échantillon

- On connait tous les individus qui composent un échantillon



- Taille de l'échantillon : le nombre d‘individu dans l'échantillon.
  - **Notation** : ***n***

Pourquoi travaille- t- on sur un échantillon ?

Échantillon:
Inférence

- **Échantillon représentatif :**
  - Échantillon qui reflète fidèlement la complexité et la composition de la population.

- **Échantillonnage aléatoire :**
  - Prélèvement au hasard, et de façon indépendante, d'un certain nombre "*n*" d'éléments à partir d'une population de "*N* " éléments.
  - Chaque élément de la population doit avoir *la* même ***probabilité*** d'être sélectionné.

Généralités

Introduction

La biostatistique : Définition et intérêt

Variabilité

Unité statistique

Population

Échantillon

Variables statistiques

Les mesures de bases

Conclusion

Variable statistique

- **Définition**
  - Caractéristique ou facteur susceptible de prendre une valeur différente selon les individus (ou les unités statistiques) étudiés;
  - Notation : X , Y , W , ... (*caractères*).

- **Exemple :**
- Le sexe des patients hospitalisés
- Le poids, la taille,
- La durée d'incubation d'une maladie

Variable : variations

Variable statistique

- Caractéristique mesurable pour toutes les unités statistiques;
- Valeurs: les mesures distinctes d'une caractéristique donnée.

–Notation : x1 , x2 , ... (**modalités**)

–**Exemple** :

- Pour la couleur des yeux : {noir ; bleu ; vert...}
- Pour le sexe : {homme ; femme}

Variable statistique

- **Valeurs possibles**
- Tous les résultats possibles *a priori* si on fait une observation d'une variable
- **Exemple**
  - Sexe : x1: Masculin, x2: Féminin
  - Groupe ABO: x1: A, x2 : B, x3: O, x4: AB

- **Valeur observée**
- Résultat *a posteriori* d'une observation d'une variable
- **Exemple** :
  - Sexe : x=x1: Masculin
  - Groupe ABO: x=x4 : AB

Variable statistique: Types

- **Différentes natures ou types :**

- **Des catégories** (Sexe, Couleur, Forme,…)
- **Des rangs** (position dans un ordre de préférence)
- **Des comptages** (nombre d'objets, nombre d'espèces, …)
- **Des mesures physiques** (température, pH, poids,…)
- Des profils, des items ….

Variable statistique: Types

- Variable qualitative
  - ***Définition***
    - Une variable statistique est qualitative si ses valeurs, ou modalités correspondent à des « qualités » :
    - Non mesurables sur une échelle
    - Présence ou absence d'une caractéristique

    (on définie des classes)

Variable qualitative nominale **Variable qualitative ordinale**

- Variable quantitative

Variable statistique:
Variable qualitative nominale

- **Définition**

– C'est une variable qualitative dont les modalités ne sont pas ordonnées.

– Si deux modalités : Dichotomique/ binaire

- **Exemples**

– Variable binaire
  - Sexe : homme ou femme
  - État de santé : malade ou sain

– Groupe sanguin
  - A, B, AB, O

– Situation familiale :
  - Célibataire, marié, divorcé, veuf

Variable statistique:
Variable qualitative ordinale

- **Définition**
- C'est une variable qualitative dont les modalités sont naturellement ordonnées *(il existe un ordre entre les classes).*

- **Exemple**
- **Niveau d'étude**
  - Primaire, secondaire, universitaire.
- **Stade de gravité d'une maladie**
  - Modéré, sévère, très sévère

Variable statistique:
Variable qualitative

- Population : ???
- Variable :???
- Unité statistique : ???
- Valeurs :???

Variable statistique:
Variable quantitative

- **Définition**

- Une variable statistique est quantitative si ses valeurs sont des nombres exprimant une **quantité**, sur lesquels les opérations arithmétiques (somme, différence…) ont un sens.

Variable quantitative continue **Variable quantitative discrète**

Variable statistique:
Variable quantitative discrète

- **Définition**
  - Une variable discrète peut prendre un nombre limité (le nombre de possibilités est fini) de valeurs isolées, généralement entières.

- **Exemples**
- Nombre d'enfants d'une famille : 0, 1, 2, 3, 4, …10

Variable statistique:
Variable quantitative continue

- **Définition**
  - Une variable continue peut prendre n'importe quelle valeur (une infinité de valeurs possibles) entre le minimum et le maximum pour une certaine échelle de mesure.
- **Exemples**
- Poids, taille, niveau de cholestérol,… .

Variable statistique:
Variable quantitative

- Population: ???
- Variable: ???
- Unité statistique: ???
- Valeurs: ???

Variable statistique:

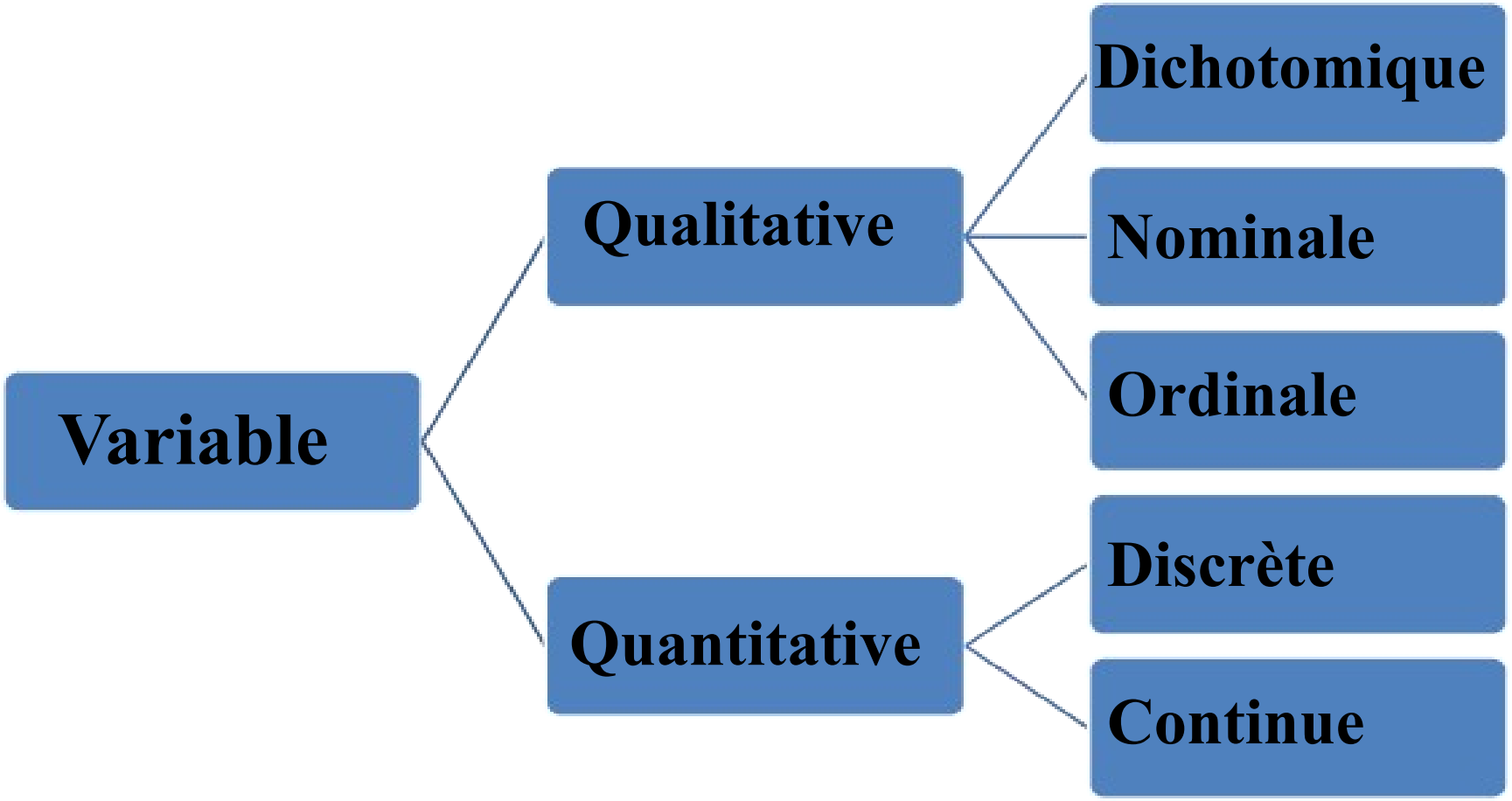
Variables continues ou discrètes ?

- On peut grouper une donnée continue ou discrète en classes de valeurs : donnée ordinale:

- Ex : âge < 20, 20-25, 25-30, >30

- Ex : nb cigarettes/j = 0, 1-10, 11-20, > 20

Variables statistiques



Généralités

Introduction

La biostatistique : Définition et intérêt

Variabilité

Unité statistique

Population

Échantillon

Variables statistiques

Les mesures de bases

Conclusion

Les mesures de base

Proportion

Ratio

Indice

Taux

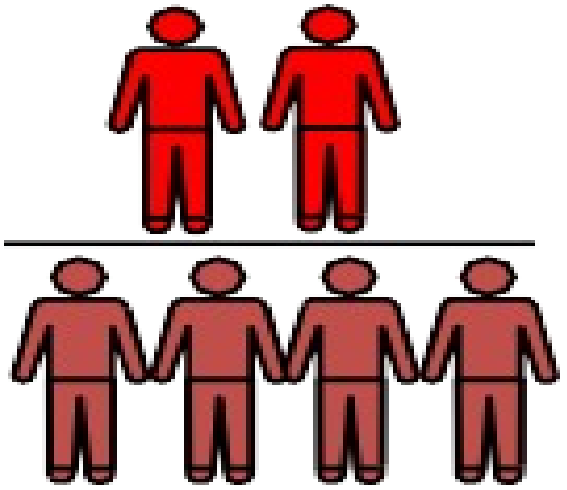
Rapport

Les mesures de base:
Rapport

- Expression de la relation qui existe entre deux quantités.
- Il est de la forme:

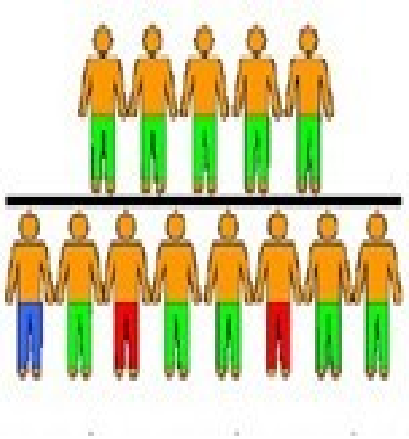
$$\frac{X}{Y} \times K$$

- **Numérateur**
- **Dénominateur**



Les mesures de base:
Proportion

- **Définition:**
- C'est le rapport où le numérateur est une part du dénominateur.
- Numérateur et dénominateur : même nature
- Expression :
  - Nombre compris entre 0 et 1
  - Pourcentage (pour mille, pour dix mille, …)



n inclus dans d

k_{2n} puissance de 10

souvent k=100

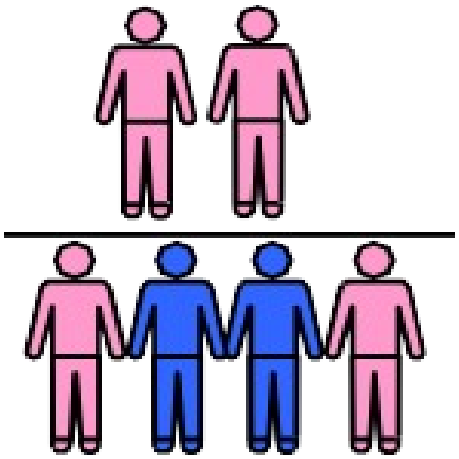
$$\text{Proportion} = \frac{n}{d} * k$$

Les mesures de base:
Proportion

- Une population de 1000 personnes dont 300 fumeurs
- ⇒la proportion des fumeurs dans ma population?

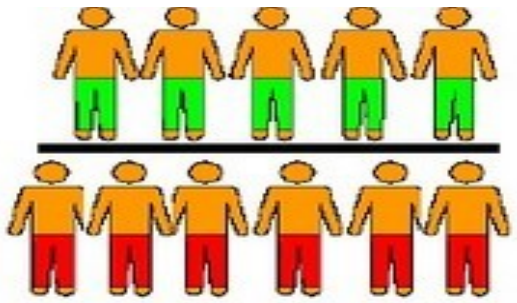
- La proportion des fumeurs = 300/1000 = 30 %

$$\frac{2}{4} = 0.5 = 50\%$$



Les mesures de base:
Ratio

- **Définition**
- C'est le rapport des effectifs de **deux classes** d'une **même variable.**
- Le numérateur n'est pas inclus dans le dénominateur.
- Numérateur et dénominateur : même nature
- Expression : nombre sans unité



$$\text{Ratio} = \frac{n}{d}$$

$$\frac{1}{4} = 0.25 \text{ Homme pour une femme}$$

Les mesures de base:
Ratio (exemples)

- Dans une faculté, nous avons 1200 étudiants dont 400 sont de sexe féminin, le sexe ratio M/F?

Le sexe ratio M/F = 800 / 400 = 2

- Une population de 1000 personnes dont 300 malades, le ratio personnes malades/ personnes non malades

Le ratio malades/ non malades=300/700 = 0,43

Cancer : ratio 10,6 hommes pour une femme ???

Les mesures de base:
Indice

•Définition:

- C'est le rapport des effectifs de **deux classes** de deux **variables différentes.**
- Le numérateur n'est pas compris dans le dénominateur.
- Numérateur et dénominateur :
  - –Nature différente.
  - –Référent à des événements différents.
  - –Exemple : **1 Kw/ personne, 10l d'eau/ Habitant**

Les mesures de base:
Taux

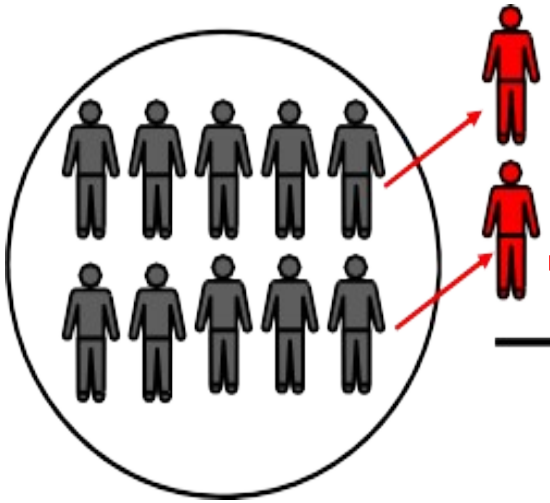
- **Définition :**
  - C'est une forme particulière de la proportion qui renferme la notion de **temps**, il exprime la vitesse de changement d'un phénomène dans le temps
  - Numérateur : individus ayant subi un événement au cours du temps
  - Dénominateur : l'ensemble des individus susceptibles de connaître l'événement pendant cette période
  - Expression : nombre de cas pour 10 ^{x} personnes - temps

Les mesures de base:
Taux

$$\text{Taux} = \frac{\text{Nb de cas survenus au cours d'une période donnée}}{\text{Effectif de la population à risque au cours de la même période}} \times 10^{x}$$

Les mesures de base:
Taux

- **Taux = probabilité de survenue d'un événement**
- **Événement = maladie ou décès**



$$\frac{x}{y} * k$$

x nb événements

y population exposée

k = puissance de 10

Les mesures de base:
Taux

- **Taux de mortalité maternelle**

$$TMM = \frac{\text{Nb de décès maternels au cours d'une période}}{\text{Nb de N}^{ces}\text{ vivantes au cours de la même période}}$$

En 2010 : TMM = 110 décès pour 100 mille naissances vivantes

Conclusion

- Biostatistique et variabilité
  - **La variabilité est la règle**
  - **La variabilité est non prévisible**
- Biostatistique : nécessité pour les biologistes

Classification des variables

Exercice 1 :

- Lors d'une étude prospective concernant des patients diabétiques du service d'endocrinologie d'un l'hôpital X, les données recueillies étaient :
  - L'âge en année,
  - Le sexe (masculin, féminin),
  - L'ancienneté du diabète (moins de 6 mois, entre 6mois et 1 an, plus d'un an),
  - Les antécédents médicaux (HTA, infection urinaire, maladie ophtalmique, autres),
  - Les antécédents de sédentarité (oui, non),
  - Le traitement antidiabétique (règles hygiéno-diététiques, antidiabétique oral, insuline),
  - L'indice de masse corporel (IMC) (normal, surpoids, obésité)
  - Et la glycémie à jeun.
- Q1 : Déterminer, sous forme d'un tableau, la nature des différentes variables recueillies.

Exercice

Variables	Modalités	Nature
L'âge en année		Quantitative continue
Le sexe	Masculin, féminin	Qualitative binaire
L'ancienneté du diabète	Moins de 6 mois, entre 6mois et 1 an, plus d'un an	Qualitative ordinale
Les antécédents médicaux	HTA, infection urinaire, maladie ophtalmique, autres	Qualitative nominale
Les antécédents de sédentarité	Oui, non	Qualitative binaire
Le traitement antidiabétique	Règles hygiéno-diététiques, antidiabétique oral, insuline	Qualitative nominale
L'indice de masse corporel (IMC)	Normal, surpoids, obésité	Qualitative ordinale
La glycémie à jeun		Quantitative continue

Probabilité & Distributions Théoriques

Les principales lois de probabilité

Introduction

- **Les probabilités visent à étudier des expériences aléatoires et à construire des modèles mathématiques pour analyser des situations impliquant l'incertitude.**
- **Une expérience est un processus qui génère des résultats. On distingue 2 types :**
  - **Une expérience aléatoire (stochastique) s'elle peut avoir plusieurs résultats possibles que l'on peut ni prévoir, ni calculer.**
  - **Une expérience déterministe : Le résultat de l'expérience est connu avec certitude avant même d'effectuer cette expérience**

Introduction

- La statistique probabiliste ou la théorie de probabilité s'intéresse à l'étude de l'aspect aléatoire des phénomènes aléatoires;

- Le but poursuivi est l'élaboration d'outils mathématiques pouvant servir à cette étude c'est **à dire construire des modèles mathématiques pour analyser des situations impliquant l'incertitude**.

Phénomènes aléatoires

- **Nous distinguons deux genres de phénomènes:**

- Ceux qui sont régis par des **lois déterminées,** On peut se rendre compte du résultat de l'expérience à l'avance et sans recours à l'expérience.

- Par exemple pour les **lois de Newton** de la pesanteur, on peut avoir une idée sur le temps de chute avec certitude (aux erreurs de mesures près) d'un corps à partir d'une distance fixée.

Phénomènes aléatoires (suite)

- Par contre ils existent d'autres phénomènes qui **n'obéissent pas à des lois déterminées**.

- Par exemple si l'expérience consiste à jeter un dé on **ne peut prédire avec certitude** les points qui apparaîtront sur la face supérieure. Cela nous conduit aux définitions :

Définitions

- **Un phénomène est dit aléatoire si on ne peut prédire avec certitude avant** l'observation du phénomène le résultat qui surviendra. Même si on répète l'expérience plusieurs fois et dans les mêmes conditions le résultat variera d'une observation à l'autre.

- **Une expérience aléatoire est le mécanisme permettant l'observation d'un** phénomène aléatoire.

Exemples de phénomènes aléatoires

- **Aléatoire**

1) **Lancer un dé ou plus non truqués.**
2) **Mesurer le taux de pollution de l'air d'une ville chaque année, à une date donnée.**
3) **Observer le niveau d'eau d'un barrage, à une date donnée.**
4) **Acheter un appartement à 500.000 Dirhams et le revendre après 3 ans. On ne peut prévoir la valeur de ces actions dans 3 ans.**

- On lance un dé, et on s'intéresse au nombre qui apparait sur la face supérieure du dé. Cette expérience est une expérience aléatoire : son **résultat**, qui s'appelle **l'évènement**, dépend du hasard.

- Les **résultats possibles** de cette expérience aléatoire s'appelle **l'univers des possibles**. Si le dé comporte 6 faces, l'univers des possibles est **Ω = {1,2,3,4,5,6}** qui s'appelle aussi **l'ensemble fondamental**

Notion de probabilité

- Cas d'un dé: la probabilité d'avoir l'une des six faces = 1/6

- Cas d'une pièce de monnaie: la probabilité d'avoir Pile = probabilité d'avoir Face = ½

- Cas d'un sac contenant deux boules noires et une blanche: P(N) = 2/3; et P(B) = 1/3

Différence entre probabilité et fréquence

- Revenons sur l'exemple du dé, quel est la fréquence d'avoir 6 la face 6 en lançant un dé 10 fois?

$$f = \frac{\text{nombre de fois où on obtient un 6}}{\text{nombre de tirages}}.$$

- Quand le nombre de tirages augmente, la fréquence de réalisation de A tend à se stabiliser autour d'un nombre limite, compris entre 0 et 1.

- Ce nombre limite signifie intuitivement la chance qu'a l'événement A de se produire lorsqu'on réalise une expérience : on l'appelle probabilité de A, et on le note P(A).

- Dans notre exemple, on a bien sûr P(A)=1/6 si le dé n'est pas pipé et P(B)=1/3.

$$P(A) = \frac{\text{Nombre de cas favorables à A}}{\text{Nombre de cas possibles}}$$

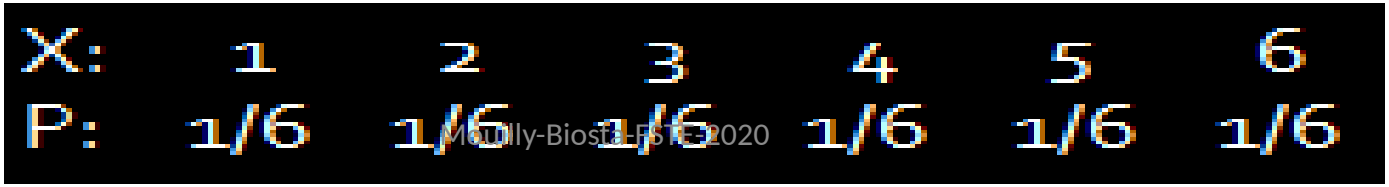
Notion de variables aléatoire et distribution de probabilités

- De manière générale, à tout évènement aléatoire on peut associer une variable aléatoire X susceptible de prendre certaines valeurs : x1 x2 . . . . . .xn correspondant aux diverses éventualités possibles;

- C'est une variable aléatoire qui peut prendre n'importe quelle valeur dans un univers statistique fini ou infini.

- L’ensemble des probabilités : p1 p2 . . . . pn associées aux valeurs prises par la variable aléatoire constitue une distribution de probabilités

- Pour le jet d’un dé on aura la distribution suivante:



Paramètre d'une distribution de probabilités

- La moyenne : symbolisée par $\bar{x}$, est la somme des produits $x_i p_i$

$$E(X) = \bar{x} = \sum_{i=1}^{n} x_i p_i$$

- Ex: Pour le jet d'un dé, la moyenne est:

$$\bar{x} = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3,5$$

- Variance et écart type d'une distribution de probabilité On peut définir un indice de dispersion

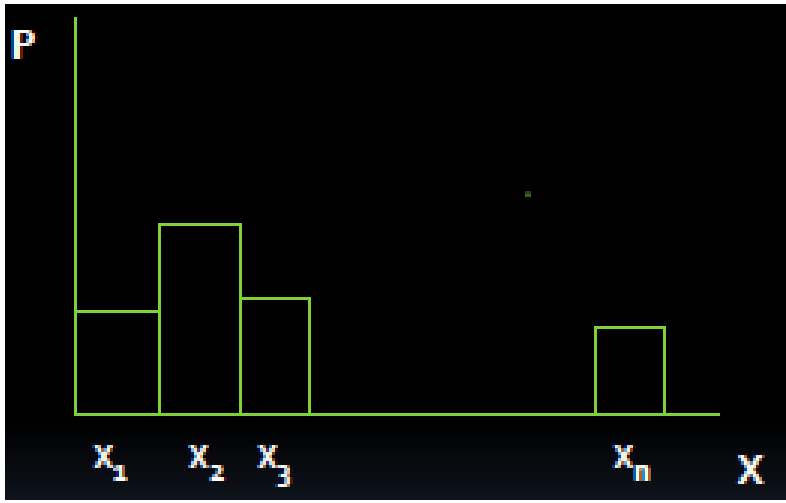
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- Dans le cas d'un dé, la variance $\sigma^2 = 3,67$

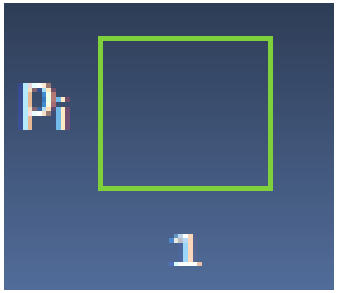
- L'écart type $\sigma = \sqrt{3,67} = 1,9$

Représentation graphique d'une distribution de probabilités

- Probabilités simples
- A chaque valeur xi correspond un rectangle i de hauteur pi et de base égale à l'unité
- Chaque rectangle a comme surface si = pi * 1
- La surface totale sous l'histogramme ST = 1 = Σpi

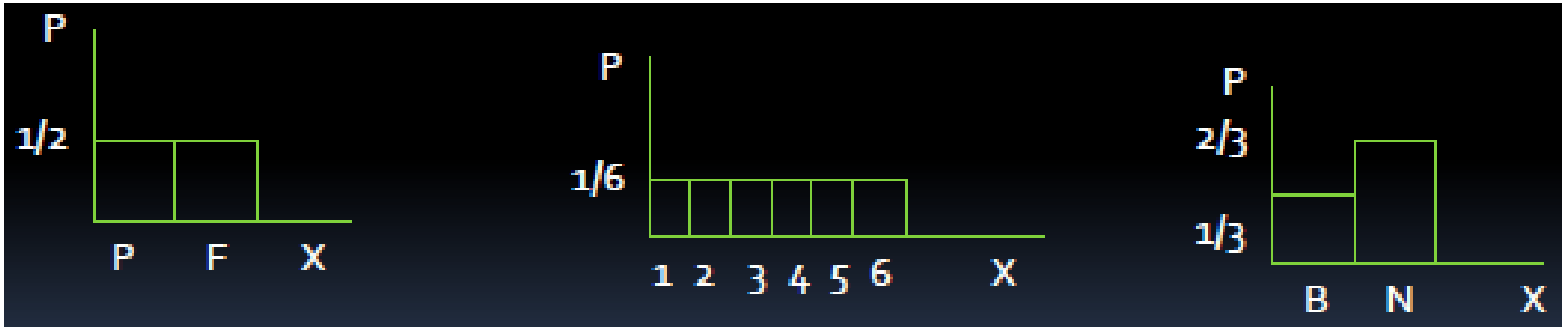


n est petit

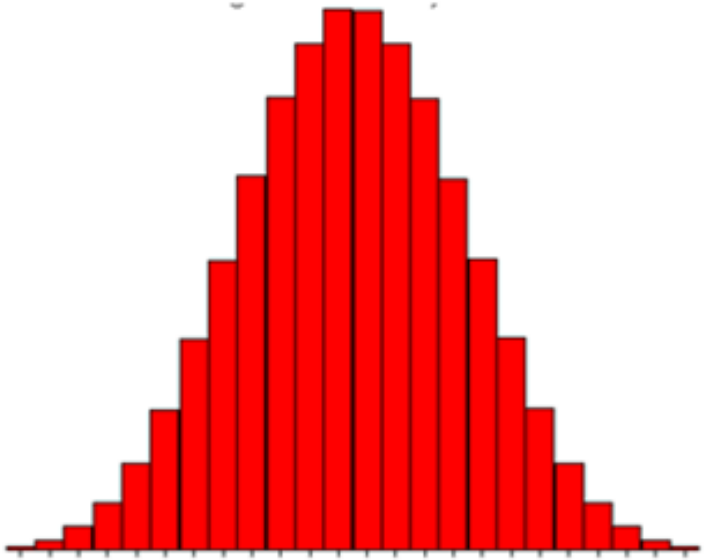


Exemples

- Jet d'une pièce de monnaie
- Lancement d'un dé
- Tirage d'une boule dans un sac contenant deux noires et une blanche

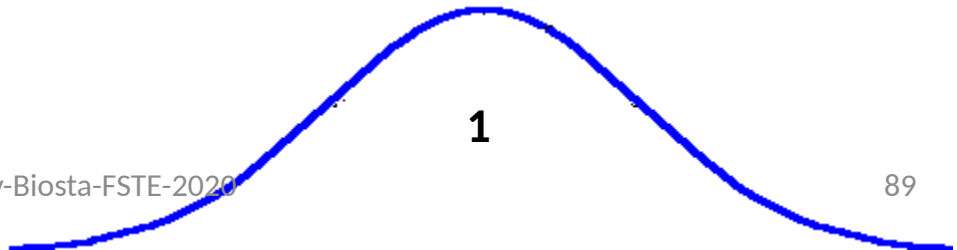


- n est grand Ou n $\to \infty$

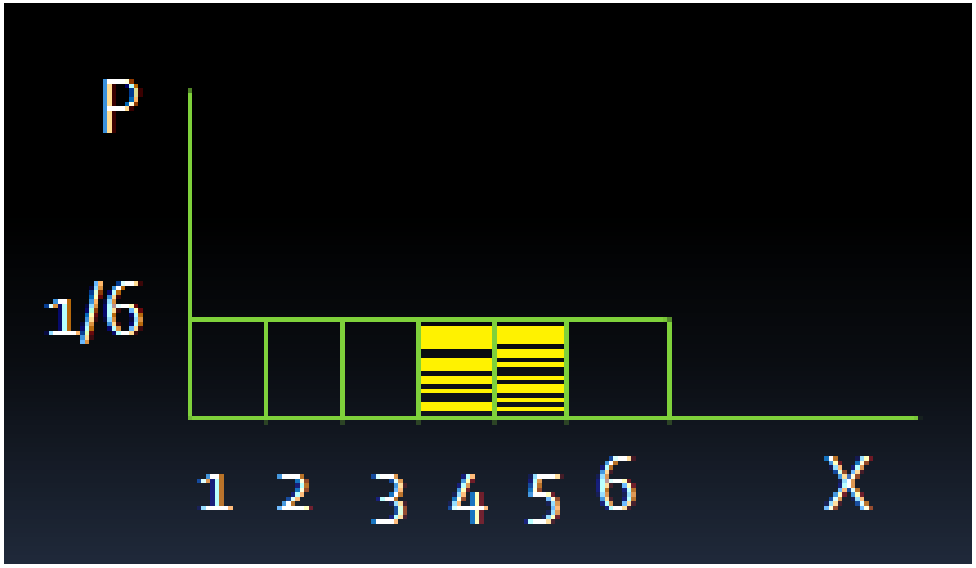


- Le nombre de rectangles augmente et leur largeur se rétrécit de plus en plus de telle manière que **les rectangles se transforment en bâtons de hauteur pi**, Si l'on joint les limites supérieures des bâtons on obtient **une courbe en cloche symétrique qui représente la loi normale**

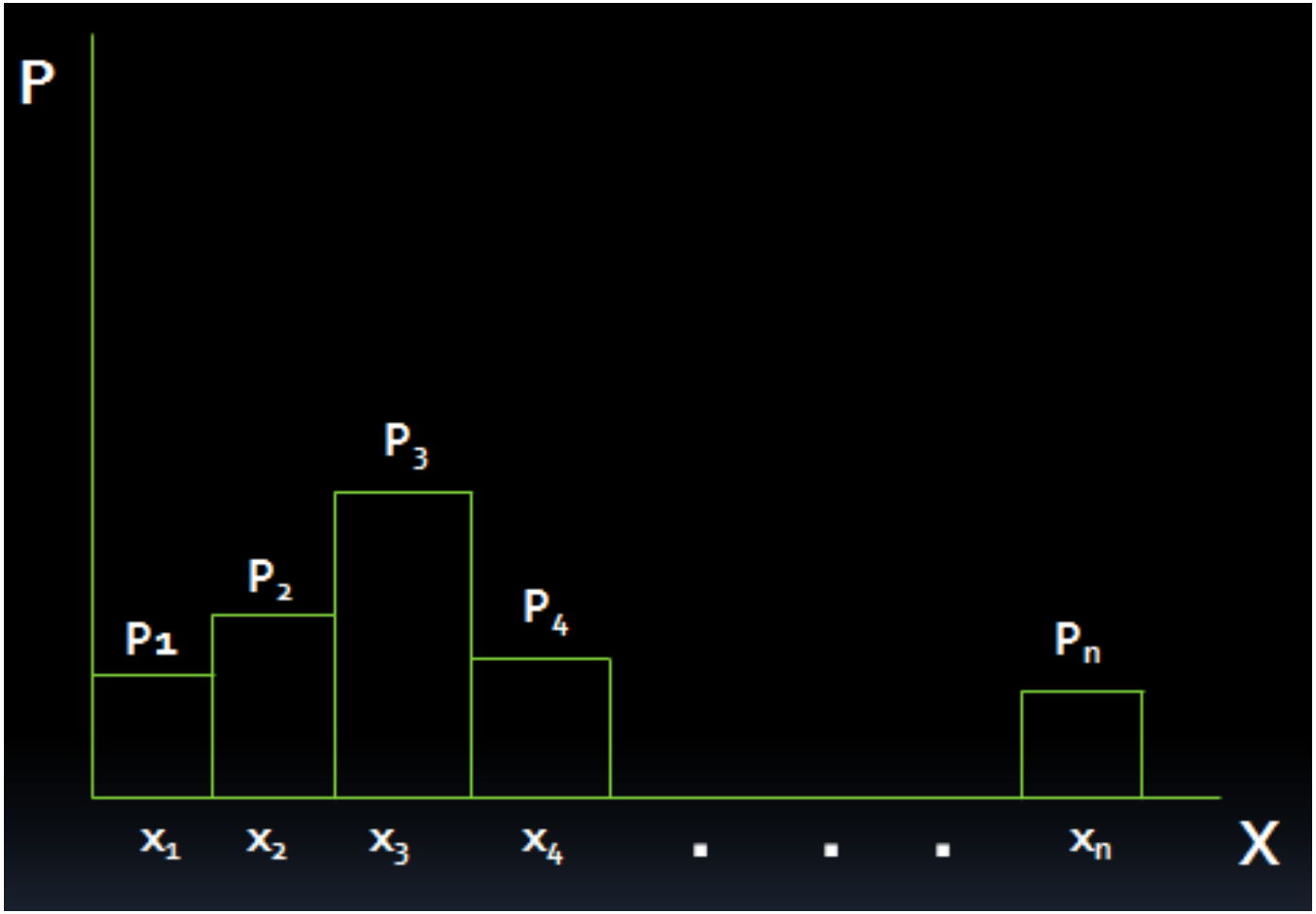
Cette courbe en cloche s'appelle la fonction de densité de probabilité symbolisée par f(x)



- Probabilités cumulées
- Notion de probabilités partielles et probabilités totales

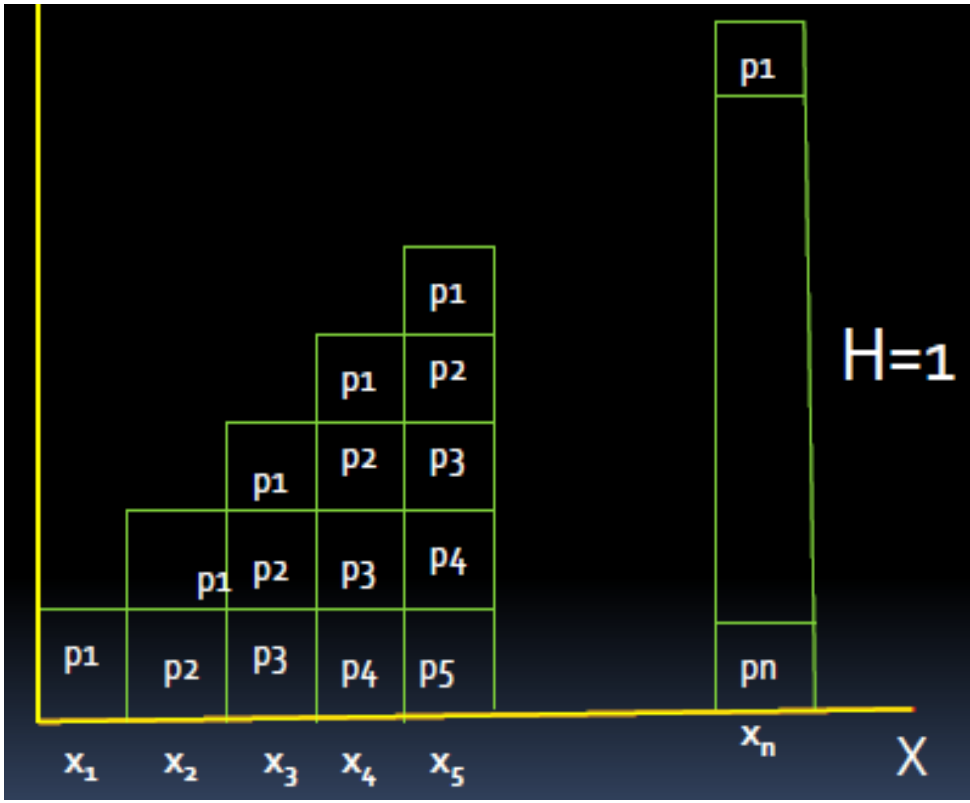


- **P(4 ou 5) = 1/6 + 1/6 = 2/6**

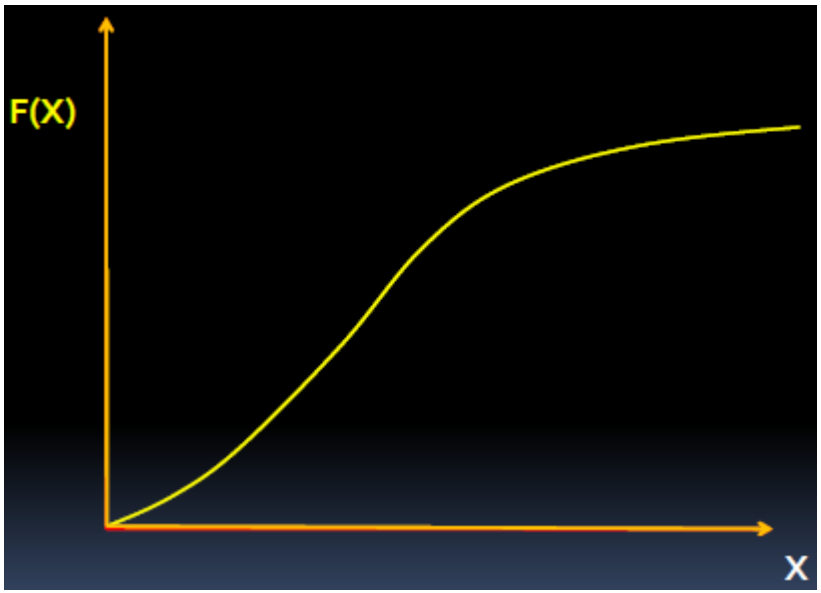

P
P1
P2
P3
P4
Pn
x1
x2
x3
x4
xn
X

- **n est petit**
- La probabilité d'avoir x1 ou x2 sera représentée par un rectangle de base x2 et hauteur égale à la somme des deux probabilités p1 et p2

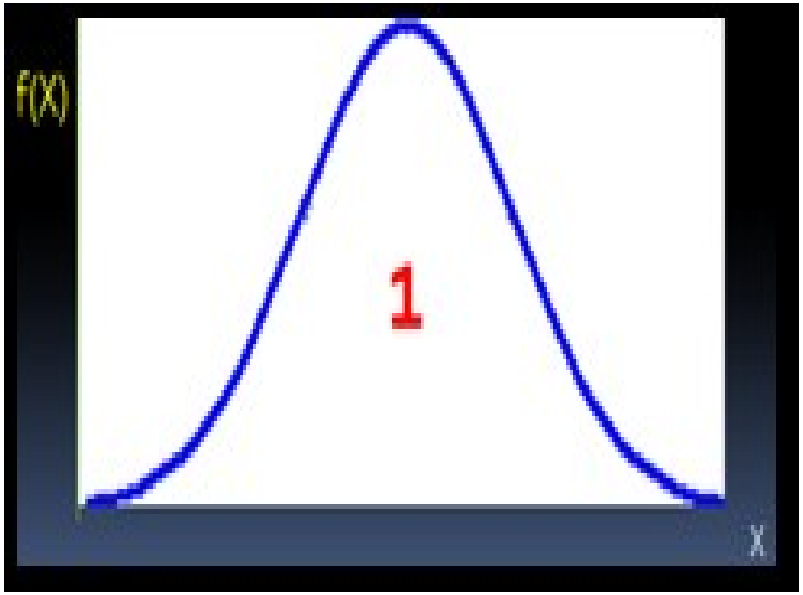
- Probabilités cumulées et n est petit ⇒ Diagramme intégral

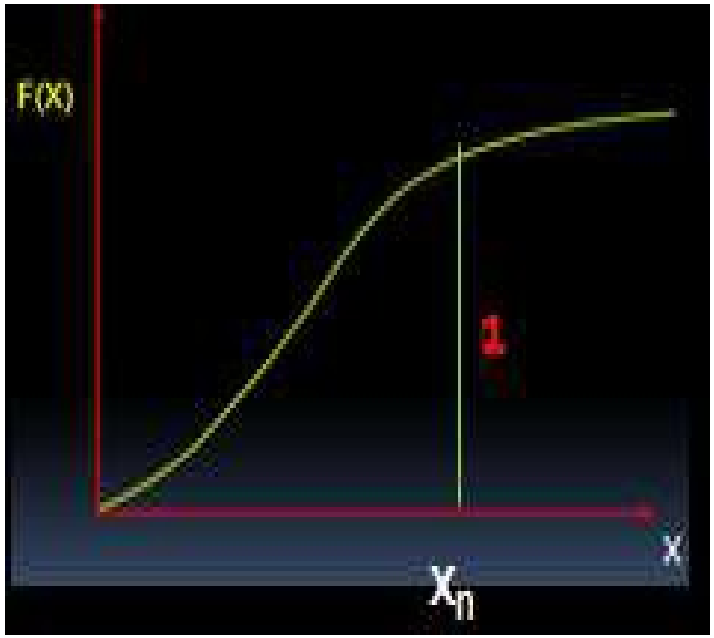


- n est grand

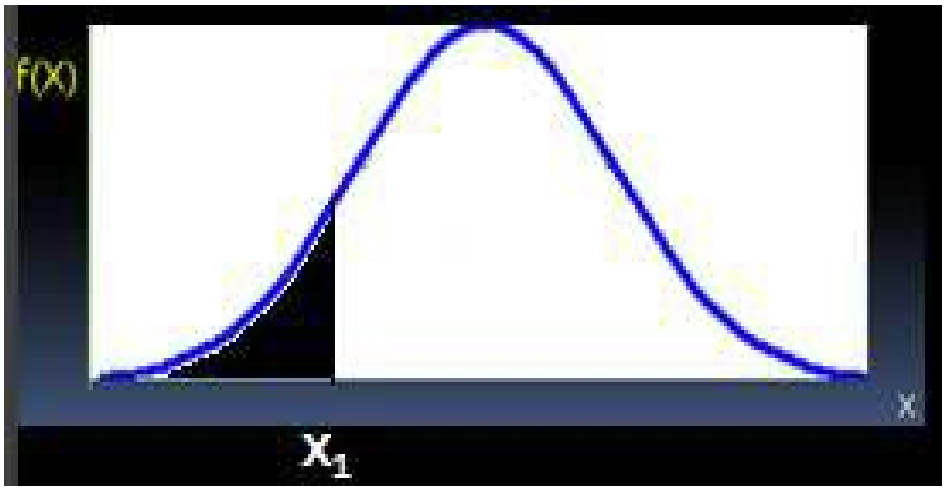


- Le nombre de rectangles augmente et leurs largeur diminue de telle manière qu'ils se transforme en bâtons. Quand on joint les limites supérieures des bâtons on obtient **la courbe cumulative** qui **s'appelle la fonction de répartition F(X).**

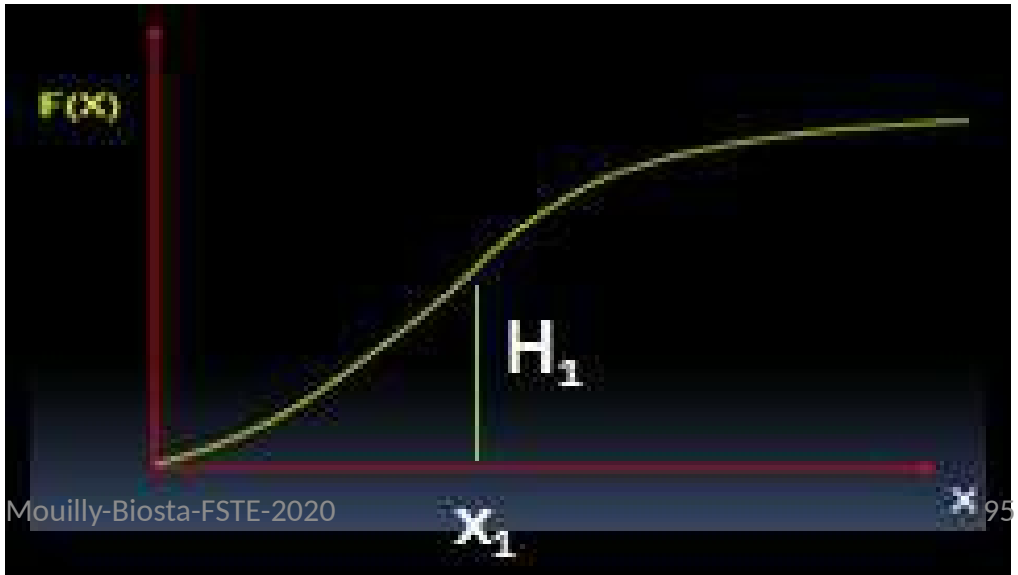


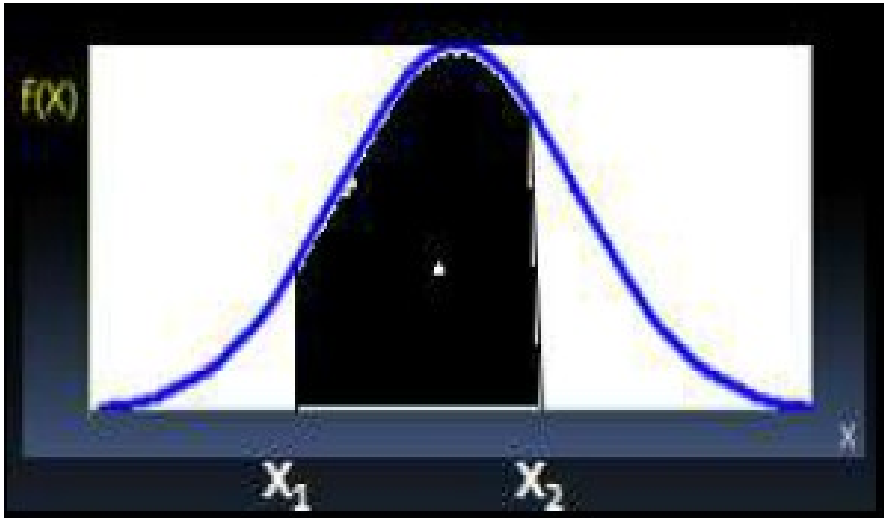


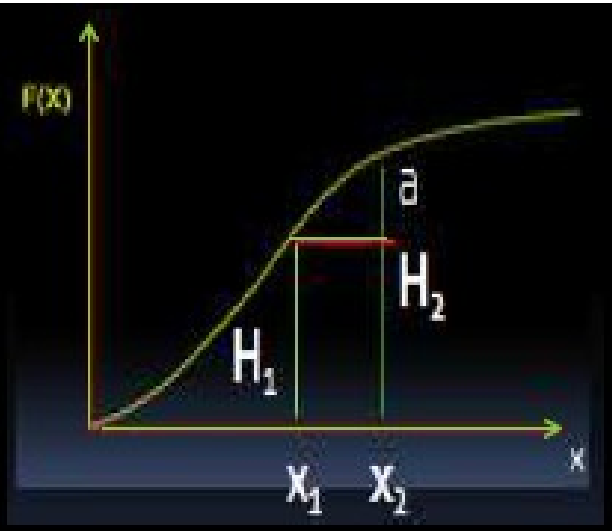
- Pt = St = Surface sous la courbe = hauteur Hn = 1



- **P (X < x1)**

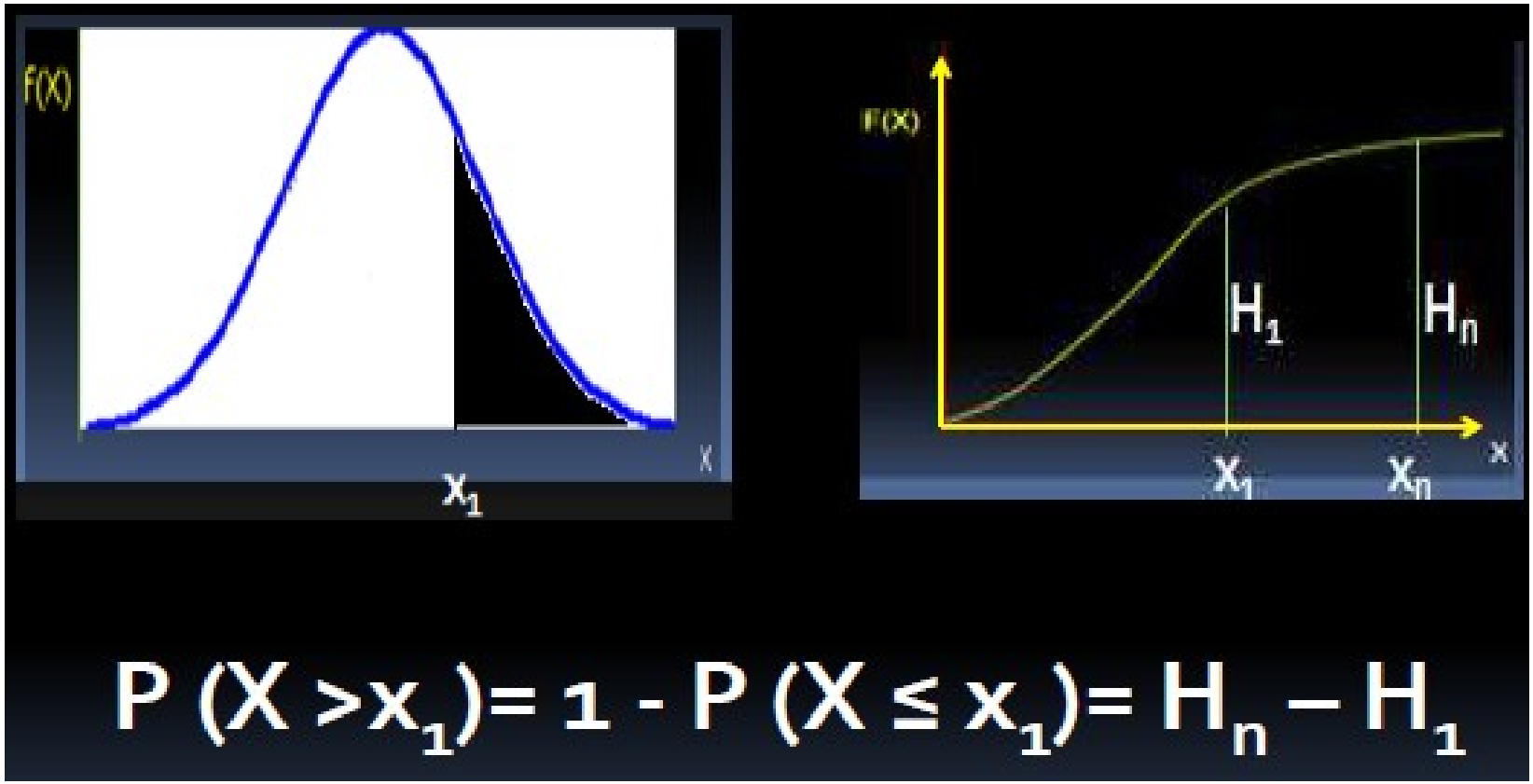






$$P(x_1 \leq X \leq x_2) = H_2 - H_1 = a$$

La formule $Pr(x_k \leq X \leq x_n) = \sum_{i=k}^{n} p_i$ est analogue à $Pr(a \leq X \leq b) = \int_a^b f(x)dx$


F(X)
X
x_1
F(X)
H_1
H_n
x
x_1
x_n
$P(X > x_1) = 1 - P(X \leq x_1) = H_n - H_1$

Concepts de Base

Variable Discontinue

- La loi Uniforme
- La loi de Bernoulli
- La loi Binomiale
- La loi de Poisson

Variable Continue

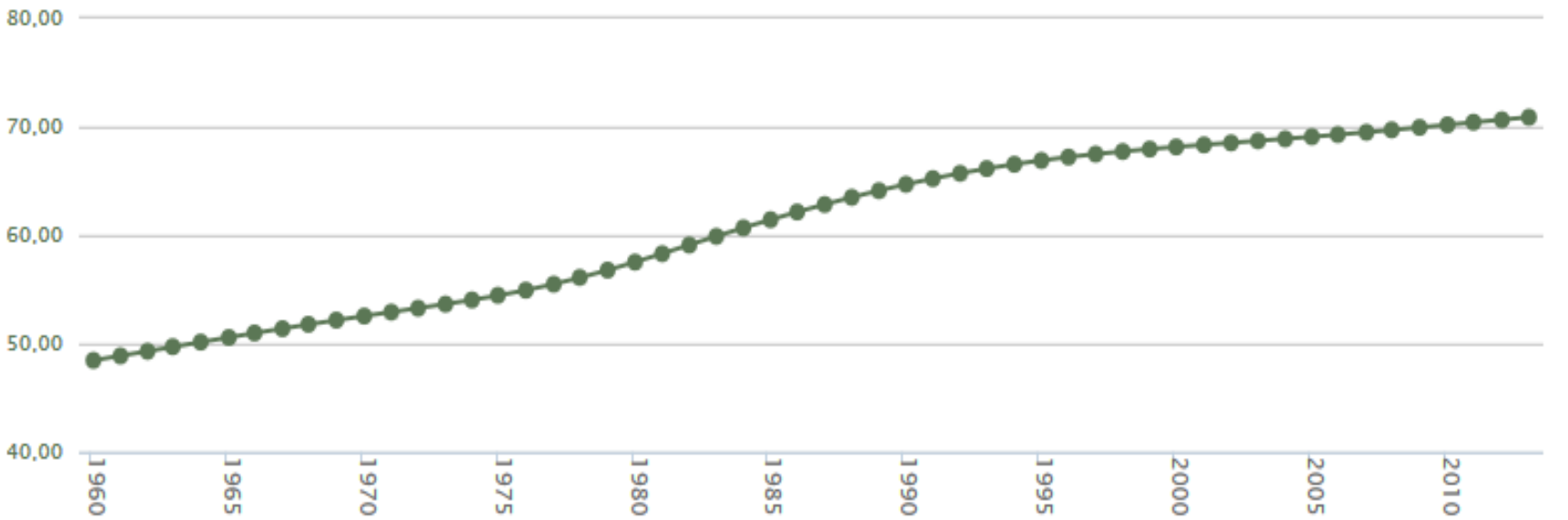
- La loi Normale
- La loi normale réduite

Paramètres principaux

- Espérance
- Variance

Exemple : Espérance de vie à la naissance (année) au Maroc

Il s'agit du nombre d'années que les personnes vivent en moyenne dans un pays donné. Cette donnée exige que les conditions (socio-médicales) prévalant à leur naissance demeurent les mêmes tout au long de leur vie.



La Banque Mondiale|Perspective Monde, Université de Sherbrooke

Loi Uniforme

Une distribution de probabilité suit une loi uniforme lorsque toutes les valeurs prises par la variable aléatoires sont équiprobables. Si *n* est le nombre de valeurs différentes prises par la variable aléatoire

$\forall K \in \Omega$ $\quad \forall_{i,}\ P(X = k) = 1/n$

Avec l'univers $\Omega = \{1.....n\}$
L'ensemble de tous les résultats (issues) possibles

Espérance $\quad E(X) = (n+1) / 2$

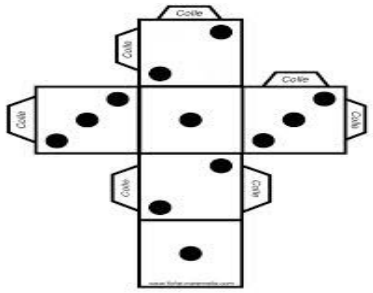
Variance $\quad V(X) = (n^2-1) / 12$

Loi Uniforme

La distribution des chiffres obtenus au lancer de dé (si ce dernier est non pipé) suit une loi uniforme dont la loi de probabilité est la suivante :

X	1	2	3	4	5	6
$P(X = k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$E(X) =$??? et $V(X) =$???



E(X) = (n+1) / 2 V(X) = $(n^2-1) / 12$

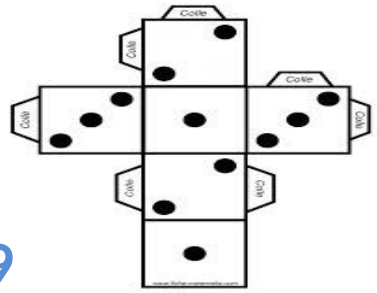
Loi Uniforme

La distribution des chiffres obtenus au lancer de dé (si ce dernier est non pipé) suit une loi uniforme dont la loi de probabilité est la suivante :

X	1	2	3	4	5	6
$P(X = k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Espérance $E(X) = (6+1) / 2 = 3.5$

Variance $V(X) = (6^2-1) / 12 = 2.9$



Loi de Birnoulli

- **Cette loi définie toute expérience aléatoire admettant 2 issues (Ω) exactement**
- **Ces 2 issues doivent être indépendantes : c.à.d. la réalisation de l'une n'influence pas la réalisation de l'autre.**
- **L'une appelée succès, noté S dont la probabilité de réalisation est p**
- **L'autre appelée échec note E = S dont la probabilité de réalisation est q = 1-p**
- **L'univers des éventualités est Ω = {0, 1}.**

Loi de Birnoulli

Dans une épreuve de Bernoulli de paramètre p, si on appelle X la variable aléatoire prenant la valeur 1 en cas de succès et 0 en cas d'échec, on dit que X est une variable de Bernoulli de paramètre p, elle suit la loi de Bernoulli de paramètre p

k	1	0
P(X = k)	p	1-p

Espérance $E(X) = p$

Variance $V(X) = p\,(1-p)$

Loi de Birnoulli

Exemples

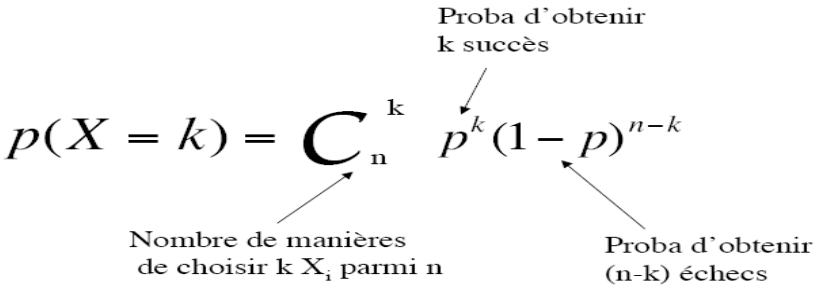
1) **un lancer de pièce de monnaie bien équilibrée est une épreuve de Bernoulli de paramètre p = ½ (le succès S étant indifféremment « obtenir PILE » ou « obtenir FACE ».**

2) **Un lancer de dé cubique bien équilibré dont les faces sont numérotées de 1 à 6, dans lequel on s'intéresse à l'apparition de S: «obtenir un 1» est une épreuve de Bernoulli de paramètre p = 1/6 et la probabilité de $\bar{S}$ est donc**

1-p = 5/6

Loi Binomiale

- **Décrite pour 1ère fois par I. NEWTON en 1676 et démontrée par J. BERNOULLI en 1713.**
- **Les épreuves doivent être répétées et indépendantes d'une même expérience de Bernoulli. Chaque expérience n'a que deux résultats possibles : succès ou échec.**
- **A cette expérience multiple on associe une variable aléatoire X qui mesure le nombre de succès obtenus dans n essais, suit la loi binomiale est définie par :**



$\frac{n!}{k!(n-k)!}$ est souvent noté $\binom{n}{k}$ ou C_n^k

Les $\binom{n}{k}$ s'appellent coefficients du binôme.

Notation : Tirage avec remise à n essai ⊟ B(n,p)

Loi de Bernoulli *versus* Loi Binomiale

- **La loi de Bernoulli est une distribution de probabilité, qui prend la valeur 1 avec la probabilité *p* et 0 avec la probabilité q =1 – p.**

- Notation : X suit la loi $B(1, p)$
- Univers : $X(\Omega) = \{0, 1\}$
- Loi : $P(X = 0) = p$
- Espérance et variance : $E(X) = p$, $Var(X) = p(1 - p)$

→ **La loi binomiale modélise un tirage avec remise parmi un ensemble de n objets (essais) de 2 types de solution, avec une probabilité de p et q**

- Notation : X suit la loi $B(n, p)$
- Univers : $X(\Omega) = \{0, 1, \cdots, n\}$
- Loi : $P(X = k) = C_n^k p^k (1 - p)^{n-k}$
- Espérance et variance : $E(X) = np$, $Var(X) = np(1 - p)$

Loi de Poisson

- **C'est une Loi introduite en 1838 par S.D. POISSON. Elle décrit le comportement du nombre d'évènements rares (ex: effets secondaires des médicaments, maladies génétiques, criminalité!….) qui se produisent dans un laps de temps fixé avec une fréquence moyenne connue et indépendamment du temps écoulé depuis l'évènement précédent.**
- **Cette une loi dérive de la Loi Binomiale. Ainsi cette dernière tend vers Poisson si p diminue et n augmente. En pratique si $p<0.05$, l'approximation est satisfaisante si $n>50$.**
- **λ (réel strictement positif) =nb moyen de ces évènements (accidents…..) par unité de temps.**

$$E(X) = \lambda$$

$$V(X) = \lambda$$

Loi de Poisson

Une v.a. X suit une loi de Poisson de paramètre réel positif λ, notée

$X \sim \mathcal{P}(\lambda)$

si elle suit:

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

$$E(X) = \lambda$$

$$V(X) = \lambda$$

Exemple : si un certain type d'évènements se produit en moyenne 4 fois par minute, pour étudier le nombre d'évènements se produisant dans un laps de temps de 10 minutes, on choisit comme modèle une loi de Poisson de paramètre $\lambda = 4 \times 10 = 40$

Loi de Poisson

- Situation

Connaissant le nombre moyen **λ** d'événements attendus pendant une période donnée, quelle est la probabilité d'observer k individus ayant subi cet événement pendant une période équivalente?

P(λ)

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Avec *e* = 2,718

Loi Normale

- **C'est une distribution théorique : une idéalisation mathématique qui ne se rencontre pas exactement dans la nature. Mais de nombreuses distributions réellement observées s'en rapprochent.**

- **C'est une Loi qui a été introduite par I. de MOIVRE en 1733. Elle est également appelée Loi gaussienne, Loi de Gauss ou Loi de Laplace-Gauss des noms de LAPLACE (1749-1827) et GAUSS (1777-1855).**

Loi Normale

- **C'est l'une des lois de probabilité les plus adaptées pour modéliser des phénomènes naturels issus de plusieurs événements aléatoires**

distribution continue

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2 / 2\sigma^2}$$

- **La densité de probabilité de cette loi est donnée par**

Où x est un réel; μ est la moyenne ; σ^2 est la variance.

- **On la note par** $N(\mu, \sigma^2)$

$$E(X) = \mu$$
$$V(X) = \sigma^2$$

- **Cette formule est classé par les 10 équations qui ont le plus changé la science!!…**

Loi Normale

Exemple 1 : Le poids des étudiants de la FP de Safi suit une loi normale de moyenne 77 kg et d 'écart type 5 kg ⊔ $N(77,25)$

Exemple 2 : Chez l'adulte normale (non diabétique), la glycémie suite une loi normale de moyenne 0.8 g/L et d 'écart type 0.3 g/L ⊔ $N\ (0.8, 0.09\ )$

Propriétés de la loi normale : La moyenne = La médiane = Le mode

La distribution normale est symétrique.

$$P(X < \mu) = P(X > \mu) = 0.5$$

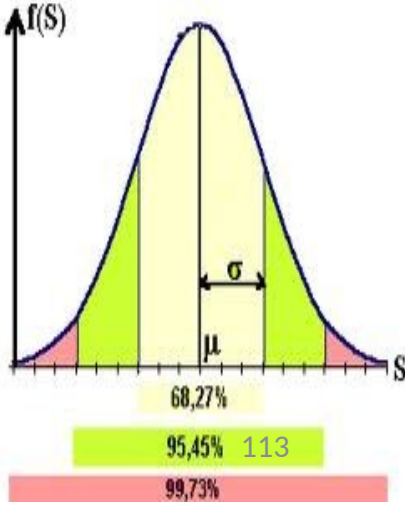
La moyenne est μ et la variance σ²

$$P(\mu - \sigma < X < \mu + \sigma) = 0.68$$

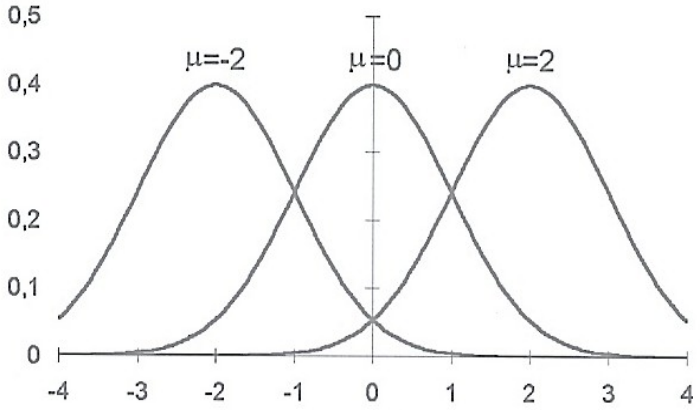
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.99$$

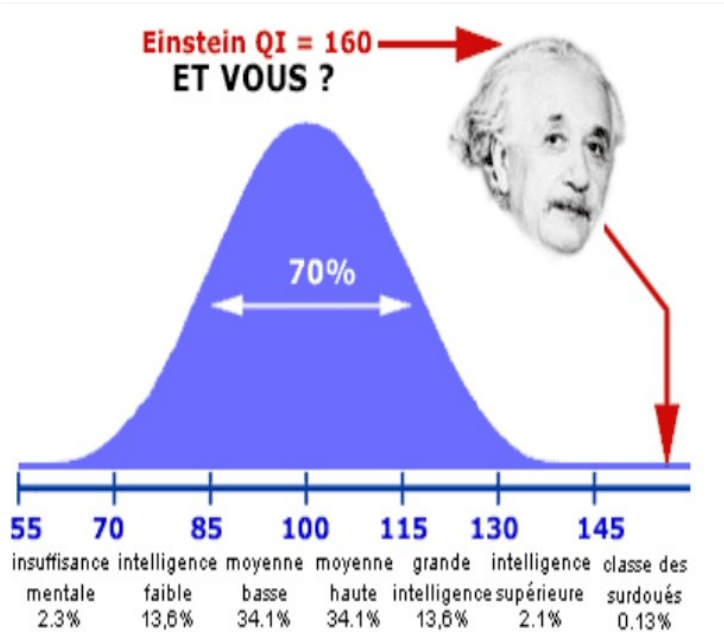
$P(a<X<b)$ **= Aire sous la courbe de la loi normale de *a* à *b***

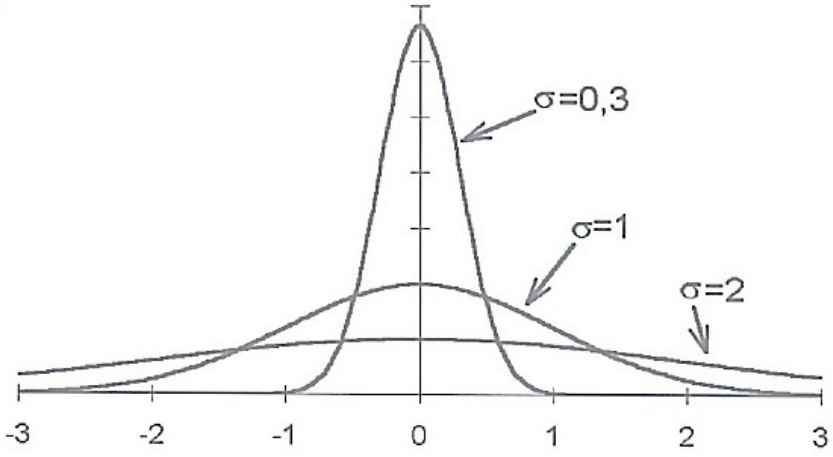


Loi Normale

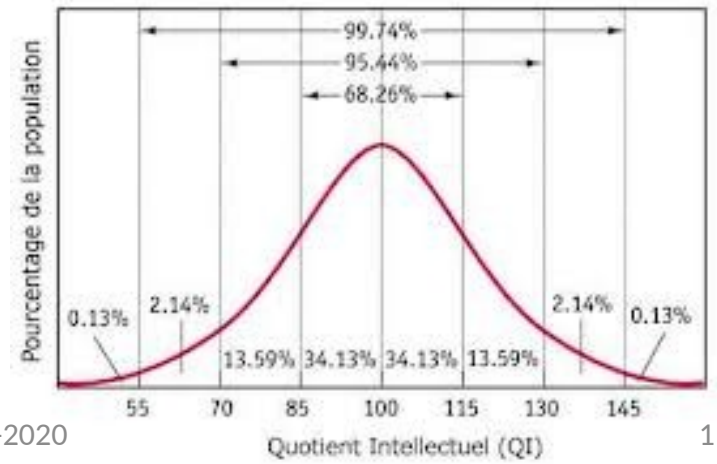


N(μ, 1) pour les valeurs de μ -2 ; 0 et 2





$N(0, \sigma^2)$ pour les valeurs de σ 0,3 ; 1 et 2

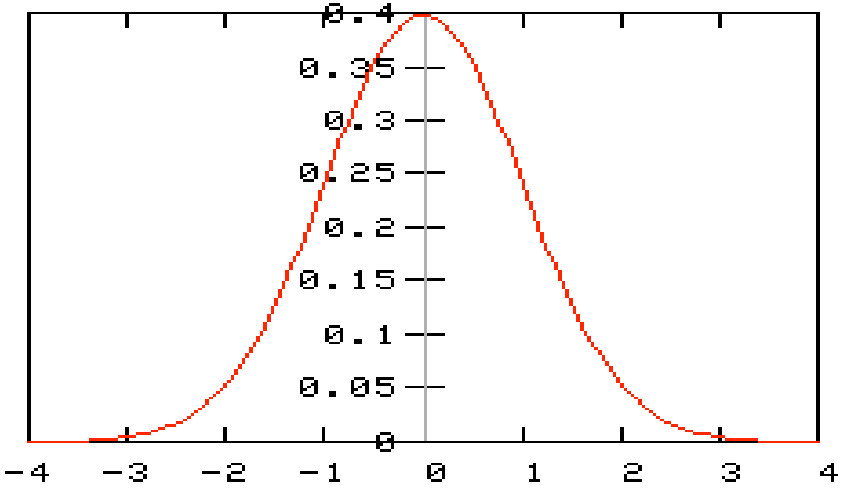


Loi Centré réduite

Loi normale centrée réduite

- On dit que la distribution est centrée si E(X)=0 et réduite si V(X)=1
- La distribution normale centrée réduite $\mathcal{N}(0;1)$ est définie par

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$



- **Les tests paramétriques ⊟ <u>distribution normale</u> & les variances des échantillons sont homogènes.**
- **Les tests non-paramétriques ⊟ Ils peuvent donc être utilisés même si les conditions de validité des tests paramétriques ne sont pas vérifiées.**

Partie II : Statistique Descriptive

Partie II : Statistique Descriptive

Avant propos

le Expérimentale

Statistique Descriptive

- 1ère étape :On collecte des données :
  - ◊ soit de manière exhaustive
  - ◊ soit par sondage
- 2ème étape : On trie les données que l'on organise en tableaux, diagrammes, etc...
- 3ème étape : On interprète les résultats : on les compare avec ceux déduits de la théorie des probabilités.

On pourra donc :

⇒ évaluer une grandeur statistique comme la moyenne ou la variance (estimateurs, intervalles de confiance).

⇒ savoir si deux populations sont comparables (tests d'hypothèses).

⇒ déterminer si deux grandeurs sont liées et de quelle façon (corrélation, ajustement analytique).

Les conclusions, toujours entachées d'un certain pourcentage d'incertitude, nous permettent alors de prendre une décision.

Partie II : Statistique Descriptive

On peut identifier deux grandes familles :

- **Des méthodes descriptives (visant à structurer et résumer l'information)**

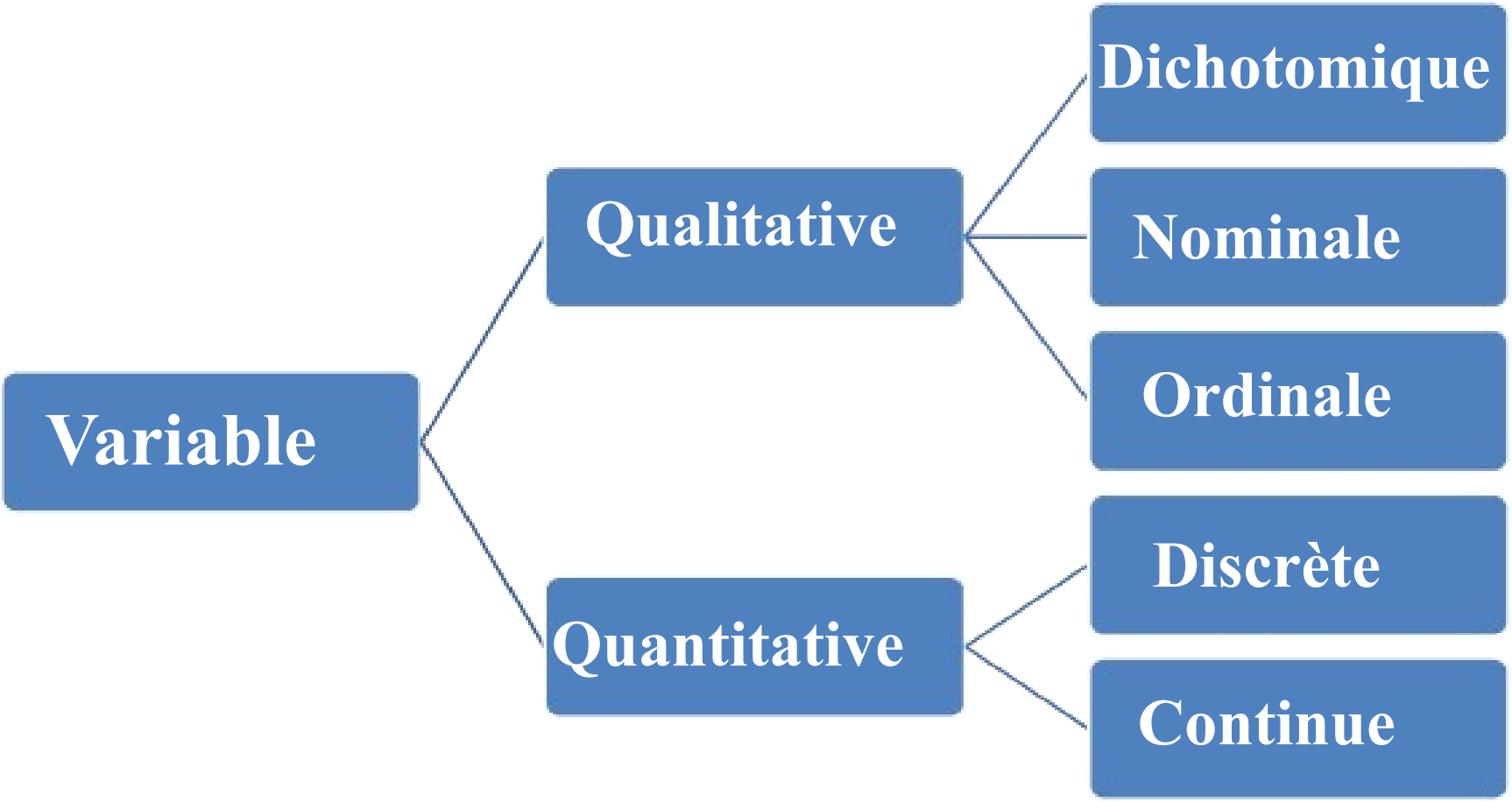
- **Des méthodes explicatives visant à expliquer une ou des variables dites « dépendantes » (variables à expliquer) par un ensemble de variables dites « indépendantes » (variables explicatives)**

Rappels: Base de données

Nom	Situation de famille	Nombre d'enfants	Age	sexe
Patient 1	Marié	2	30	M
Patient 2	Veuf	3	45	M
Patiente 3	Mariée	0	27	F
Patiente 4	Célibataire	0	32	F
Patient 5	Marié	1	39	M
….	….	….	….	….

Le nombre d'individus étant généralement grand, une telle série brute est **difficilement lisible et interprétable**. Il est indispensable de **la résumer**.

Rappels : variables



Partie II : Statistique Descriptive

- **Chapitre 1 : Statistique univariée (1 dimension)**
  1. Variable Quantitative (Discontinue & Continue)
  2. Variable Qualitative (Nominale & Ordinale)
  3. Représentation Graphique
- **Chapitre 2 : Statistique descriptive bivariée (2 dimensions)**
  1. Introduction
  2. Variables Qualitatives
     a) Tables de contingence
     b) Paramètres de corrélation : V de Cramer, CC, Coefficient Phi,…
  3. Variables quantitatives
     a) Covariance
     b) Paramètres de corrélation : R^2, Coefficient de Spearman,…
     c) Régression
  4. Variables (Qualitative & Quantitative)
  5. Représentation Graphique
- **Chapitre 3 : Statistique multivariée (x dimensions)**

Chapitre 1 : Statistique univariée (1 dimension)

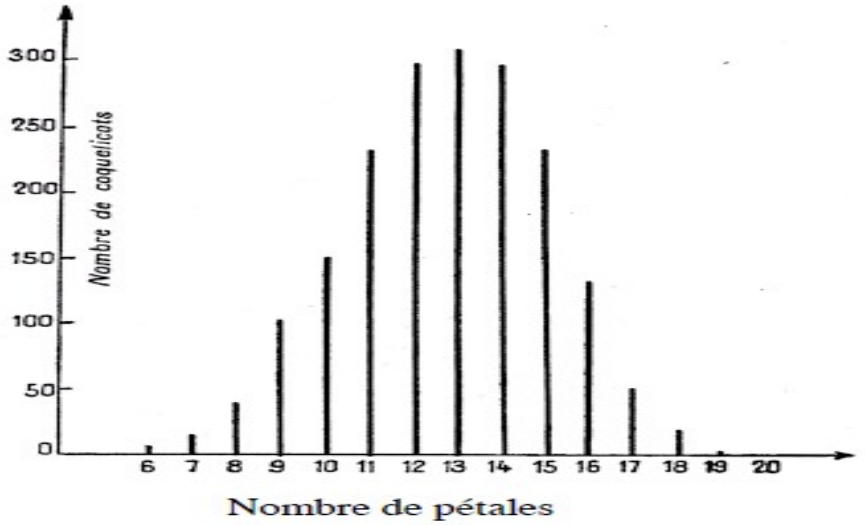
1.
    1. Distributions des fréquences
    2.
2.
3.

Chapitre 1 : Statistique univariée

Variable Quantitative Discontinue (discrète)

= **Série non groupée**

1. **Les distributions des fréquences**



En multipliant par 100 on passera de la fréquence relative au pourcentage

$$\sum n_i = N \qquad fi = \frac{n_i}{N}$$

- $n_1, n_2, \ldots, n_p$ **sont les <u>effectifs</u> (=fréquences absolues).**
- **N est l'<u>effectif total</u>**
- f_i **est la <u>fréquence</u> (relative) de xi, tel que :**

$$\sum_i f_i = 1$$

$$(0 \leq f_i \leq 1)$$

=d'apparition

NB. Statistique et Probabilités : Fréquence versus probabilité.

- Une fréquence est une proportion d'observations
- Une probabilité est la mesure d'une incertitude sur un événement

NB. Cas des fréquences (effectifs) cumulées

- x_1 est la plus petite valeur, x_p la plus grande des valeurs observées.

N_i est l'effectif cumulé c'est dire le nombre d'observations ayant des valeurs inférieures ou égales à x_i

F_i est la fréquence cumulée c'est à dire la fréquence des observations ayant des valeurs inférieures ou égales à x_i

Concernait la variable « note/20 en biostatistique » de la classe X 2019-20,

Déterminer l'effectif, la fréquence, l'effectif cumulé et la fréquence cumulée au sein de cette population ?

{ 7 16 8 15 13 14 9 14 13 13 7 13 11 8 15 11 9 11 16 11 9 15 9 8 11 }

Note	7	8	9	11	13	14	15	16
Effectif	2	3	4	5	4	2	3	2
Fréquence	0,08	0,12	0,16	0,20	0,16	0,08	0,12	0,08
Effectif cumulé	2	5	9	14	18	20	23	25
Fréquence cumulée	0,08	0,20	0,36	0,56	0,72	0,80	0,92	1

NB. Cas des fréquences (effectifs) cumulées

- x_1 est la plus petite valeur, x_p la plus grande des valeurs observées.

N_i est l'effectif cumulé c'est dire le nombre d'observations ayant des valeurs inférieures ou égales à x_i

F_i est la fréquence cumulée c'est à dire la fréquence des observations ayant des valeurs inférieures ou égales à x_i

Concernait la variable « note/20 en biostatistique » de la classe de Mastérants SAB 2019-20,

Déterminer l'effectif, la fréquence, l'effectif cumulé et la fréquence cumulée au sein de cette population ?

{ 7 16 8 15 13 14 9 14 13 13 7 13 11 8 15 11 9 11 16 11 9 15 9 8 11 }

Note	7	8	9	11	13	14	15	16
Effectif	2	3	4	5	4	2	3	2
Fréquence	0,08	0,12	0,16	0,20	0,16	0,08	0,12	0,08
Effectif cumulé	2	5	9	14	18	20	23	25
Fréquence cumulée	0,08	0,20	0,36	0,56	0,72	0,80	0,92	1

Il y a 2+3+4+5 = 14 élèves qui ont une note inférieure ou égale à 11sur 20.

0,08+0,12+0,16+0,20+0,16+0,08 = 0,80

La fréquence des élèves ayant eu une note inférieure ou égale à 14/ 20 est de 0,80 soit 80%.

Pour obtenir ce résultat on peut aussi utiliser l'effectif cumulé : 20 ÷ 25 = 0,8.

Effectifs cumulés croissants:

Nombre d'individus pour lesquels la variable est **inférieure ou égale** à x_i.
Résultat de l'addition, de proche en proche, des effectifs d'une distribution observée en commençant par le 1er.

Nbre produits financiers	Nombre de Clients	Effectifs cumulés croissants	Effectifs cumulés décroissants
0	103	103	360
1	115	218	257
2	95	313	142
3	35	348	47
4	10	358	12
5	2	360	2
Total :	360		

Effectifs cumulés décroissants:

Nombre d'individus pour lesquels la variable est **supérieure ou égale** à x_i.
Résultat de l'addition, de proche en proche, des effectifs d'une distribution observée en commençant par le dernier.

Valeurs de la variable x_i	Effectif n_i	Effectifs cumulés croissants N_i	Effectifs cumulés décroissants N'_i
x_1	n_1	$N_1 = n_1$	$N'_1 = n_k + \ldots + n_1 = n$
x_2	n_2	$N_2 = n_1 + n_2$	$N'_2 = n_k + \ldots + n_2$
x_3	n_3	$N_3 = n_1 + n_2 + n_3$	$N'_3 = n_k + \ldots + n_3$
…	…	….	….
x_{k-1}	n_{k-1}	$N_{k-1} = n_1 + \ldots + n_{k-1}$	$N'_{k-1} = n_k + n_{k-1}$
x_k	n_k	$N_k = n_1 + \ldots + n_k = n$	$N'_k = n_k$
Total :	n		

Nombre de produits financiers x_i	Nombre de clients n_i	Effectifs cumulés croissants N_i	Effectifs cumulés décroissants N'_i	Fréquences f_i	Fréquences cumulées croissantes F_i	Fréquences cumulées décroissantes F'_i
0	103	103	360	0,2861	0,2861	1
1	115	218	257	0,3194	0,6055	0,7139
2	95	313	142	0,2639	0,8694	0,3945
3	35	348	47	0,0972	0,9666	0,1306
4	10	358	12	0,0278	0,9944	0,0334
5	2	360	2	0,0056	1	0,0056
Total :	360			1		

Il y a 313 clients possédant un nombre de produits financiers inférieur ou égal à 2

Il y a 47 clients possédant un nombre de pro. fin. supérieur ou égal à 3

La proportion de clients possédant un nombre de pro. fin. inférieur ou égal à 4 est de 99,44%

La proportion de clients possédant un nombre de pro. fin. supérieur ou égal à 1 est de 71,39%

Effectif Cumulée Croissant **ECC**

→ d'une Valeur est associée à la somme des effectifs des valeurs inférieures.

Fréquence Cumulée Croissante **FCC**

→ d'une Fréquence est associée à la somme des fréquences inférieures.

Effectif Cumulé Décroissante **ECD**

→ d'une Valeur est associée à la somme des effectifs des valeurs supérieures.

Fréquence Cumulée Décroissante **FCD**

→ d'une Fréquence est associée à la somme des fréquences supérieures.

Les étudiants d'établissement universitaire sont répartis en 5 salles différentes , Déterminer les paramètres suivantes : ECC, FCC, ECD et FCD ?

Nombre d'étudiants	Salle n°1	Salle n°2	Salle n°3	Salle n°4	Salle n°5
Effectif	23	32	32	11	2

Effectif Cumulée Croissant EFC

→ d'une Valeur est associée à la somme des effectifs des valeurs inférieures.

Fréquence Cumulée Croissante FCC

→ d'une Fréquence est associée à la somme des fréquences inférieures.

Effectif Cumulée Décroissante EFD

→ d'une Valeur est associée à la somme des effectifs des valeurs supérieures.

Fréquence Cumulée Décroissante FCD

→ d'une Fréquence est associée à la somme des fréquences supérieures.

Les étudiants d'établissement universitaire sot répartis en 5 salles différentes , Déterminer les paramètres suivantes : EFC, FCC, EFD et FCD ?

Nombre d'étudiants	salle n°1	salle n°2	salle n°3	salle n°4	salle n°5
Effectif Fréquence (%)	23	32	32	11	2

N= 100

la Fréquence Cumulée croissante (FCC)

Nombre d'étudiants	Salle n°1	Salle n°2	Salle n°3	Salle n°4	Salle n°5
Fréquence (en %)	23	32	32	11	2
Fréquences cumulées croissantes (%)	23	55	87	98	100

Effectif Cumulée Croissant EFC

→ d'une Valeur est associée à la somme des effectifs des valeurs inférieures.

Fréquence Cumulée Croissante FCC

→ d'une Fréquence est associée à la somme des fréquences inférieures.

Effectif Cumulée Décroissante EFD

→ d'une Valeur est associée à la somme des effectifs des valeurs supérieures.

Fréquence Cumulée Croissant FCD

→ d'une Fréquence est associée à la somme des fréquences supérieures.

Les étudiants d'établissement universitaire sot répartis en 5 salles différentes , Déterminer les paramètres suivantes : EFC, FCC, EFD et FCD ?

Nombre d'étudiants	salle n°1	salle n°2	salle n°3	salle n°4	salle n°5
Effectif Fréquence (%)	23	32	32	11	2

N= 100

la Fréquence Cumulée Croissante (FCC)

Nombre d'étudiants	Salle n°1	Salle n°2	Salle n°3	Salle n°4	Salle n°5
Fréquence (en %)	23	32	32	11	2
Fréquences cumulées croissantes (%)	23	55	87	98	100

la Fréquence Cumulée Déroissante (FCD)

Nombre d'étudiants	Salle n°1	Salle n°2	Salle n°3	Salle n°4	Salle n°5
Fréquence (en %)	23	32	32	11	2
Fréquences cumulées croissantes (%)	23	55	87	98	100
Fréquences cumulées décroissantes(%)	100	77 ~~98~~	45 ~~87~~	13 ~~55~~	2 ~~23~~

Variable Quantitative Continue

= Série groupée

1. Les distributions des fréquences

Nécessite d'une transformation des données en classes, pour construire les intervalles, on doit au maximum respecter les règles suivantes :

- **<u>Si possible,</u> La classe doit contenir un ensemble de valeurs représentant le minimum de variation & les amplitudes des classes doivent être égaux en préférence ???**

NB1. Dans les calculs, une classe sera représentée par son centre, qui est le milieu de l'intervalle.

- **Chaque classe (sauf la dernière) contient sa borne inférieure mais pas sa borne supérieure [min~~, max~~]**
- **Une fois la classe constituée, on considère les individus répartis uniformément entre les deux bornes**

NB2. ce qui entraîne une perte d'informations par rapport aux données brutes.

Répartition des étudiants selon leurs notes

[0;4]	20
[4;6]	60
[6;8]	90
[8;10]	100
[10;12]	70
[12;14]	70
[14;16]	40
[16;20]	20

Quel est le problème?

Répartition des étudiants selon leurs notes

[0;4[	20
[4;6[	60
[6;8[	90
[8;10[	100
[10;12[	70
[12;14[	70
[14;16[	40
[16;20]	20

Taille (cm) des étudiants

Limites réelles des classes	Fréquences relatives en %	Fréquences cumulées croissantes en %		Fréquences cumulées décroissantes en %	
154,5-159,5	10	10 =	10	10 + 90 =	**100**
159,5-164,5	0	0 + 10 =	10	0 + 90 =	90
164,5-169,5	20	20 + 10 =	30	20 + 70 =	90
169,5-174,5	20	20 + 30 =	50	20 + 50 =	70
174,5-179,5	25	25 + 50 =	75	25 + 25 =	50
179,5-184,5	20	20 + 75 =	95	20 + 5 =	25
184,5-189,5	5	5 + 95 =	***100***	5 =	5
Total	***100***				

Quel est le choix de ces intervalles ????

Chapitre 1 : Statistique univariée (1 dimension)

1.
    1. Distributions
    2. **Paramètres de position, de dispersion et de forme**
2.
3.

3 Indicateurs caractérisent la distribution d'une série statistique quantitative:

Les paramètres de position :

Moyenne « arithmétique », Médiane, Mode, Fractile, Min, Max

Les paramètres de dispersion :

Variance, Écart-type, Coefficient de Variance, Étendue (Amplitude), Ecart Interquartiles, Interquartile relatif, Intervalle de Kelley, Interdécile relatif

Les paramètres de forme :

Aplatissement (Kurtosis), Asymétrie (Skewness)

3 Indicateurs caractérisent la distribution d'une série statistique quantitative:

Les paramètres de position (tendance centrale) :

Moyenne « arithmétique », Médiane, Mode, Fractile, Min, Max

Les paramètres de dispersion :

Variance, Écart-type, Coefficient de Variance, Étendue (Amplitude), Ecart Interquartiles, Interquartile relatif, Intervalle de Kelley, Interdécile relatif

Les paramètres de forme :

Aplatissement (Kurtosis), Asymétrie (Skewness)

Moyenne arithmétique

Avantages

- Se prête facilement aux calculs et tests statistiques
- Bon indicateur si distribution symétrique et dispersion faible

Inconvénients

- Sensible aux valeurs extrêmes
- Représente mal une population hétérogène (polymodale)

Variable Discontinue

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

$$\bar{x} = x_1 f_1 + x_2 f_2 + \ldots + x_n f_n = \sum_{i=1}^{n} x_i \times f_i.$$

Soit x_1, x_2, … x_i;….x_k une série statistique où chacune des valeurs élémentaire x_i est répétée n_i fois (sa fréquence étant f_i).

Variable Continue

Les données seront organisées en classes de centre ci,

- **on remplacera xi par ci**

- **Cette moyenne arithmétique est la plus ancienne méthode employée pour caractériser un ensemble de données et indiquer une tendance centrale. Elle représente le centre de gravité de la distribution.**

NB. Il existe d'autres types de moyennes :

- **Moyenne géométrique**
- **Moyenne harmonique**
- **Moyenne quadratique**

→ **Moyenne arithmétique pondérée : Il faut tenir compte des coefficients**

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n_1 + n_2 + \ldots + n_k} = \frac{1}{N} \times \sum_{i=1}^{i=k} n_i x_i$$

Matière	Coefficient	note
Physique	4	12
Chimie	4	8
Philosophie	1	3
Histoire/Géographie	1	2
Maths	10	15

Déterminer la moyenne pondérée ?

→ **Moyenne arithmétique pondérée : Il faut tenir compte des coefficients**

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n_1 + n_2 + \ldots + n_k} = \frac{1}{N} \times \sum_{i=1}^{i=k} n_i x_i$$

Matière	Coefficient	note	notes coefficientées
Physique	4	12	4x12= 48
Chimie	4	8	4x8 = 32
Philosophie	1	3	1x3 = 3
Histoire/Géographie	1	2	1x2 = 2
Maths	10	15	10 x15 = 150
Total des coefficients : 4+4+1+1+10=20			
Total des notes coefficients :			*235*

Moyenne pondérée : = 11 ,75 (235/20)

NB. Moyenne non pondérée = 8 (40/5)

Très important Technique d'échantillonnage pour éviter la surreprésentation d'un critère donné

3 Indicateurs caractérisent la distribution d'une série statistique quantitative:

Les paramètres de position (tendance centrale) :

Moyenne « arithmétique », Médiane, Mode, Fractile, Min, Max

Les paramètres de dispersion :

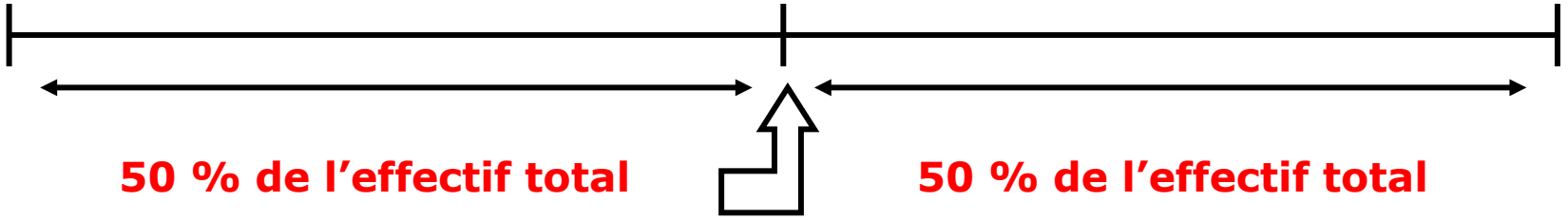
Variance, Écart-type, Coefficient de Variance, Étendue (Amplitude), Ecart Interquartiles, Interquartile relatif, Intervalle De Kelley, Interdécile relatif

Les paramètres de forme :

Aplatissement (Kurtosis), Asymétrie (Skewness)

Médiane

La médiane est la valeur de la série qui partage la distribution en 2 sous-ensembles d'égal effectif.



Effectif correspondant à la médiane de la série

Avantages

- **Moins sensible aux valeurs extrêmes que la moyenne**
- **Bon indicateur si distribution asymétrique**

Inconvénients

- **Se prête mal aux calculs statistiques**
- **Classement peut être long si les valeurs sont nombreuses**

Médiane

• Si n est impair i est un entier et la valeur médiane existe dans la série statistique.

• Si n est pair le rang i tombe entre deux valeurs, la médiane correspondra à la moyenne des 2 valeurs qui encadrent ce rang.

- si N est impair (exemple N = 7) :

$a_1\ a_2\ a_3\ a_4\ a_5\ a_6\ a_7$

3 3 4 4 6 6 7

$m = a_4$ = 4

7 3 6 4 4 6 3

- si N est pair (exemple N = 8) :

$a_1\ a_2\ a_3\ a_4\ a_5\ a_6\ a_7\ a_8$

3 3 4 4 6 6 7 8

$m = \frac{a_4 + a_5}{2}$ = 5

8 3 4 4 6 6 7 3

Moyenne - Médiane

Série de valeurs:	Moyenne	Médiane
10, 12, 18, 20, 25		
7, 12, 18, 20, 25		
10, 12, 18, 20, 45		

Moyenne - Médiane

Série de valeurs:	Moyenne	Médiane
10, 12, 18, 20, 25	17	18
7, 12, 18, 20, 25	16,4	18
10, 12, 18, 20, 45	21	18

La moyenne est sensible aux valeurs extrêmes

La médiane est insensible aux valeurs extrêmes

On fait une étude statistique sur les 50 notes attribuées par un jury à un examen, voici les résultats obtenus en classant ces notes par ordre croissant

Notes	Effectif
0	1
1	2
2	2
3	3
4	2
5	3
6	2
7	3
8	4
9	3
10	2
11	3
12	4
13	4
14	3
15	1
16	2
17	1
18	2
19	2
20	1

Déterminer la médiane de la note des étudiants ???

Utilisation des effectifs (fréquences) cumulés

Rappel : $(x_1 , x_2 , x_3 , x_4 , \ldots\ldots, x_n)$ est défini de la façon suivante:

Si $n = 2p$ est pair , M est le centre de l'intervalle $[x_p ; x_{p+1}]$

Si n est impair, M est le nombre x_p où $p = (n + 1)/2$.

n = 50 est pair, il faut donc prendre le centre de [9 ; 10]

Utilisons la colonne des effectifs cumulés pour déterminer la médiane : il y a 50 notes, la 25ème note est 9 et la 26ème : 10.

Dans le tableau il n'y a pas de valeur partageant la série statistique en deux groupe de même effectif , (l'effectif total est pair) dans ce cas l'intervalle médian est [9;10] et on prendre pour médiane **le centre de cet intervalle : 9,5**

Notes	Effectif	Effectif cumulé
0	1	1
1	2	3
2	2	5
3	3	8
4	2	10
5	3	13
6	2	15
7	3	18
8	4	22
9	3	25
10	2	27
11	3	30
12	4	34
13	4	38
14	3	41
15	1	42
16	2	44
17	1	45
18	2	47
19	2	49
20	1	50

Médiane

Variable Continue

La série est regroupée par classes : plusieurs méthodes de détermination de la médiane

Méthode d'interpolation linéaire

Utilisons la colonne des effectifs cumulés pour déterminer la médiane : il y a 50 notes, 50 % de l'effectif total c'est 25, la médiane est ici la note correspondant à l'effectif cumulé 25.

Notes	Effectifs	Effectifs cumulés
[0 ; 5[	10	10
[5 ; 8[	8	18
[8 ; 12[	12	30
[12 ; 15	11	41
[15 ; 20	9	50
	50	

D'après la colonne "effectif cumulé" :
18 personnes ont moins de 8
30 personnes ont moins de 12

Médiane

Variable Continue

La série est regroupée par classes : plusieurs méthodes de détermination de la médiane

Méthode d'interpolation linéaire

Notes	Effectifs	Effectifs cumulés
[0 ; 5[	10	10
[5 ; 8[	8	18
[8 ; 12[	12	30
[12 ; 15	11	41
[15 ; 20	9	50
	50	

Il y a 50 notes, 50 % de l'effectif total c'est 25, la médiane est ici la note correspondant à l'effectif cumulé 25.

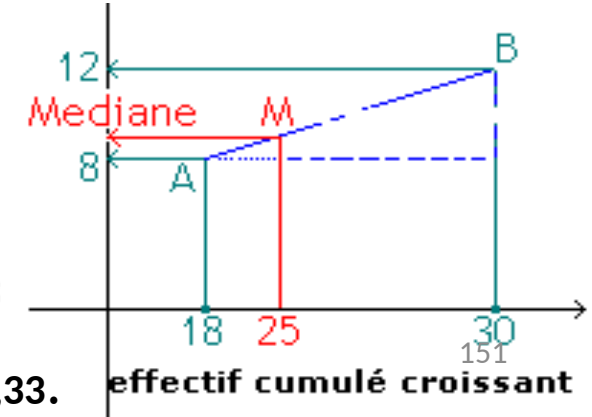
Les points A, M, B sont alignés ce qui se traduit par les droites (AM) et (AB) ont même coefficient directeur

$$\frac{Me-8}{25-18}=\frac{12-8}{30-18}$$

$$\frac{Me-8}{7}=\frac{4}{12}$$

$$Me-8=\frac{4}{12}\times 7$$

$$Me=8+\frac{4}{12}\times 7\approx 10,33$$



La médiane est environ 10,33

50 % environ des personnes ont eu moins de 10,33 et 50 % plus de 10,33.

3 Indicateurs caractérisent la distribution d'une série statistique quantitative:

Les paramètres de position :

Moyenne « arithmétique », Médiane, Mode, Fractile, Min, Max

Les paramètres de dispersion :

Variance, Écart-type, Coefficient de Variance, Étendue (Amplitude), Ecart Interquartiles, Interquartile relatif, Intervalle De Kelley, Interdécile relatif

Les paramètres de forme :

Aplatissement (Kurtosis), Asymétrie (Skewness)

Mode ou valeur dominante

Avantages

- Bon indicateur dans le cas de distributions asymétriques
- Bon indicateur de population hétérogène Insensible aux valeurs extrêmes

Inconvénients

- Se prête mal aux calculs statistiques
- Sensible aux variations d'amplitude de classes

Variable discontinue

Correspond à la valeur la plus fréquente. x_i correspondant au n_i (ou f_i) maximum. Il peut y avoir un ou plusieurs modes.

10, 9, 12, 11, 10, 8, 14 ,11 ,9 ,16 ,5 ,12 ,10 ,11 ,10 ,13

Déterminer le mode de cette population ???

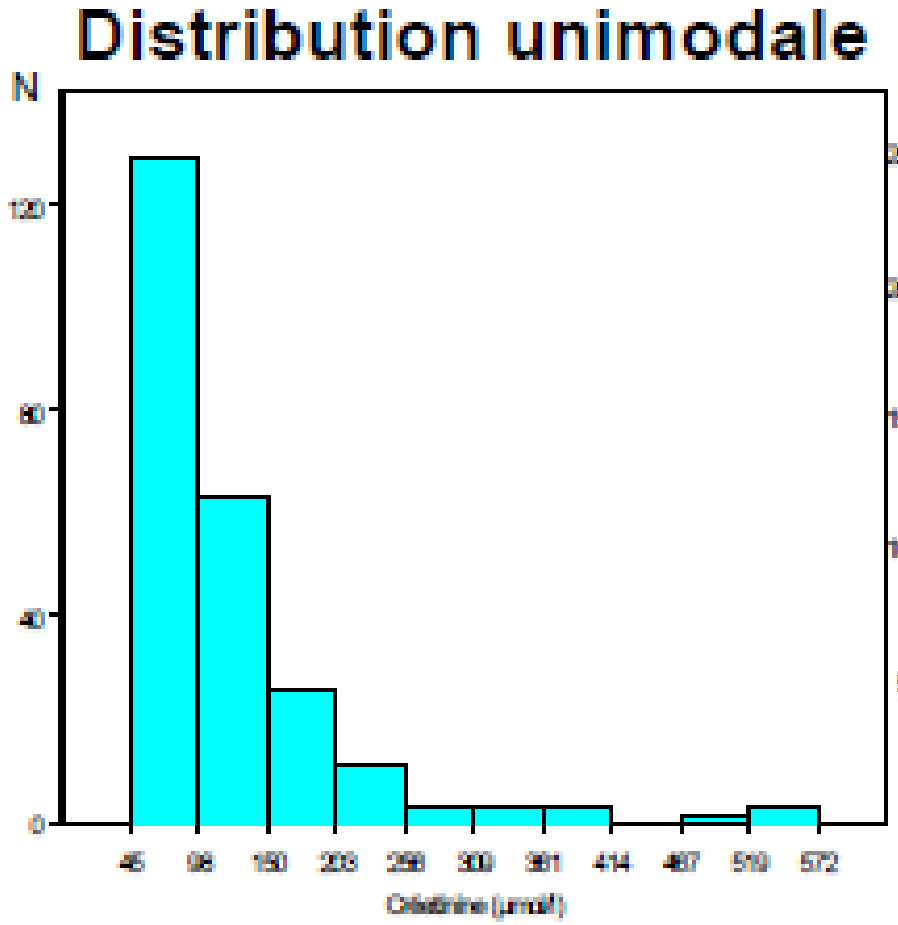
10, 9, 12, 11, 10, 8, 14 ,11 ,9 ,16 ,5 ,12 ,10 ,11 ,10 ,13

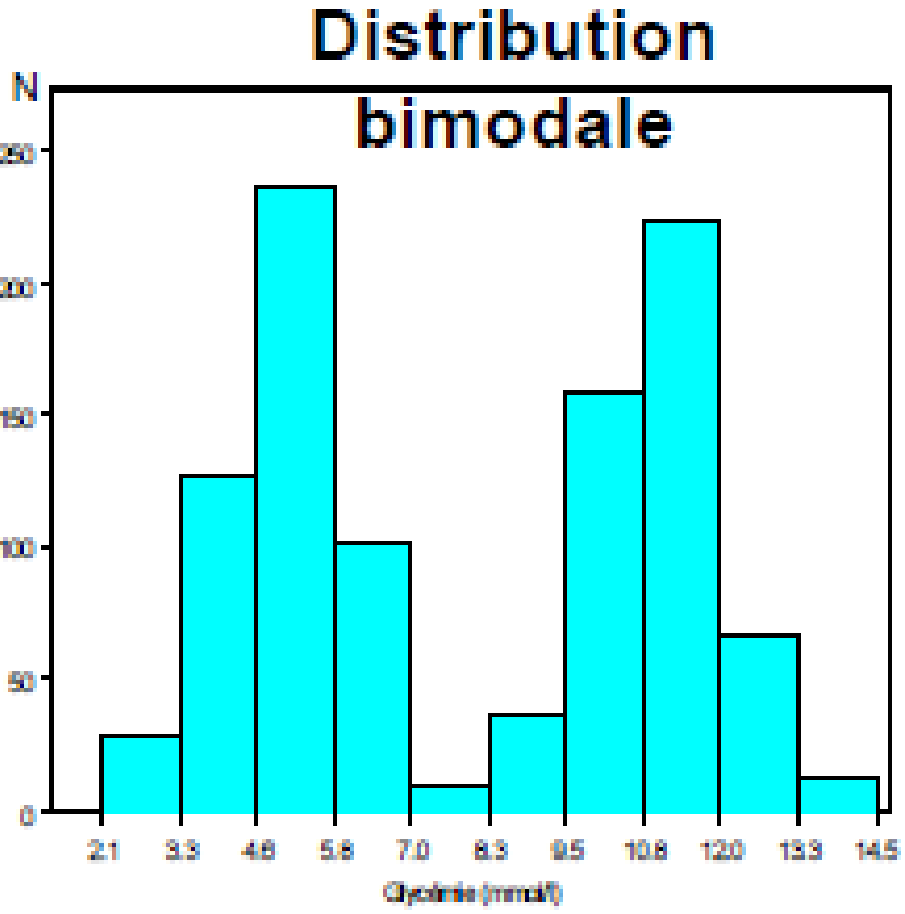
Déterminer le mode de cette population ???

notes	5	8	9	10	11	12	13	14	16
effectifs	1	1	2	4	3	2	1	1	1
effectifs cumulés	1	2	4	8	11	13	14	15	16

Moyenne = 10,69 / Médiane = 10,5 / Mode = 10

Mode


Distribution unimodale
N
120
80
40
0
45
98
150
203
256
309
361
414
467
519
572
Créatinine (µmol/l)


Distribution bimodale
N
250
200
150
100
50
0
2.1
3.3
4.6
5.8
7.0
8.3
9.5
10.8
12.0
13.3
14.5
Glycémie (mmol/l)

Classe modale

Variable continue

C'est la classe pour laquelle le quotient (effectif/amplitude) est maximal. Notons que le quotient effectif/amplitude s'appelle la densité d'effectif de la classe. Il peut exister plusieurs modes ou plusieurs classes modales.

Classes	Effectifs
[10;15[	10
[15;25[	18
[25;30[	15
[30;50[	30
[50;55[	7
Total	80

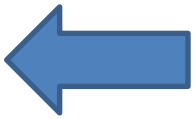
Quelle est la classe modale??

Classe modale

Variable continue

C'est la classe pour laquelle le quotient (effectif/amplitude) est maximal. Notons que le quotient effectif/amplitude s'appelle la densité d'effectif de la classe. Il peut exister plusieurs modes ou plusieurs classes modales.

Classes	Effectifs	fréquence
[10;15[	10	0,125
[15;25[	18	0,225
[25;30[	15	0,1875
[30;50[	30	0,375
[50;55[	7	0,0875
Total	80	1



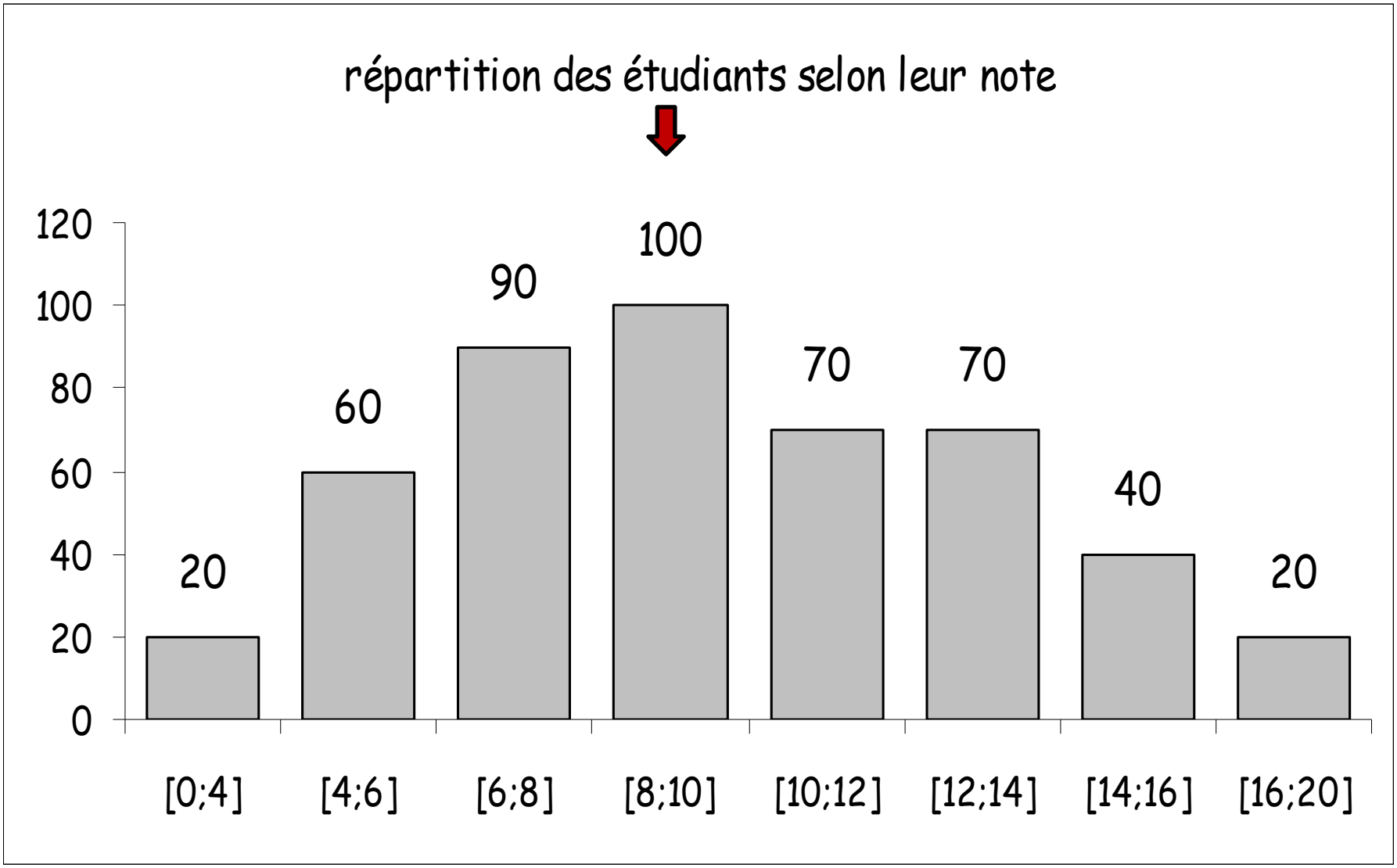
???????

Classe modale

Variable continue

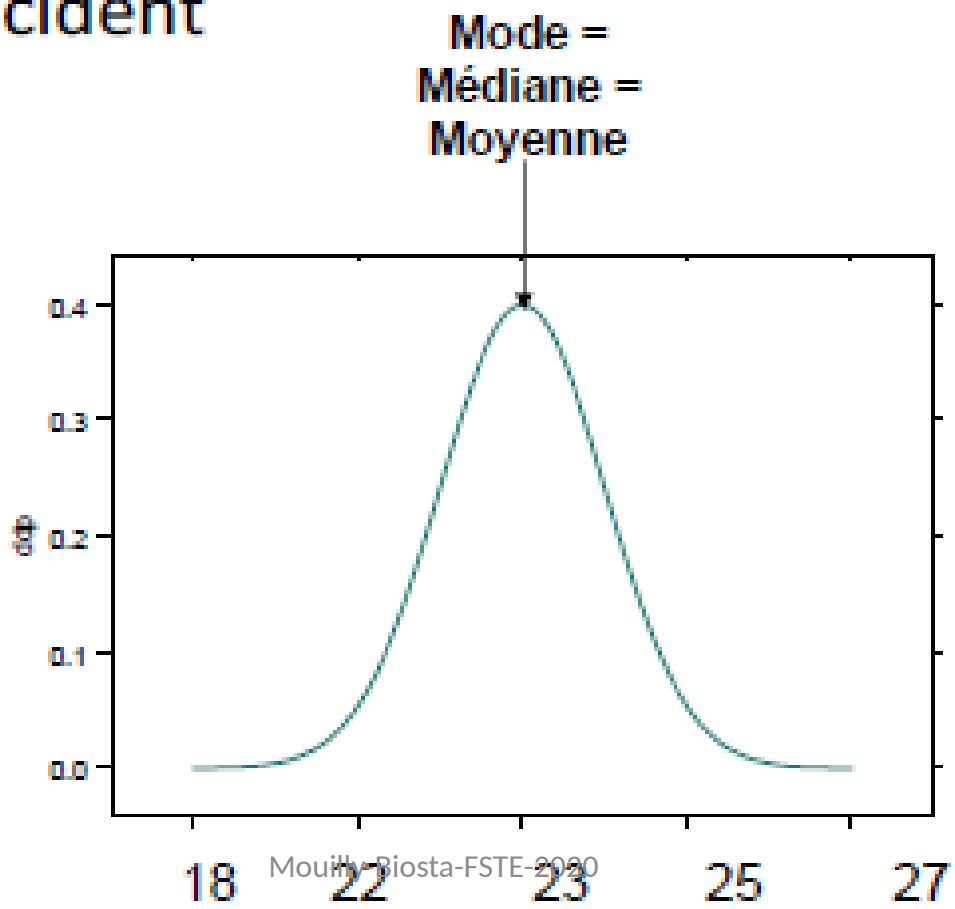
C'est la classe pour laquelle le quotient (effectif/amplitude) est maximal. Notons que le quotient effectif/amplitude s'appelle la densité d'effectif de la classe. Il peut exister plusieurs modes ou plusieurs classes modales.

Classes	Effectifs	fréquence	Amplitude	densité d'effectif
[10;15[	10	0,125	5	2
[15;25[	18	0,225	10	1,8
[25;30[	15	0,1875	5	3
[30;50[	30	0,375	20	1,5
[50;55[	7	0,0875	5	1,4
Total	80	1		


répartition des étudiants selon leur note
120
100
80
60
40
20
0
20
60
90
100
70
70
40
20
[0;4]
[4;6]
[6;8]
[8;10]
[10;12]
[12;14]
[14;16]
[16;20]

Mode, médiane, moyenne

- Si distribution unimodale, **symétrique**
  - les 3 coïncident

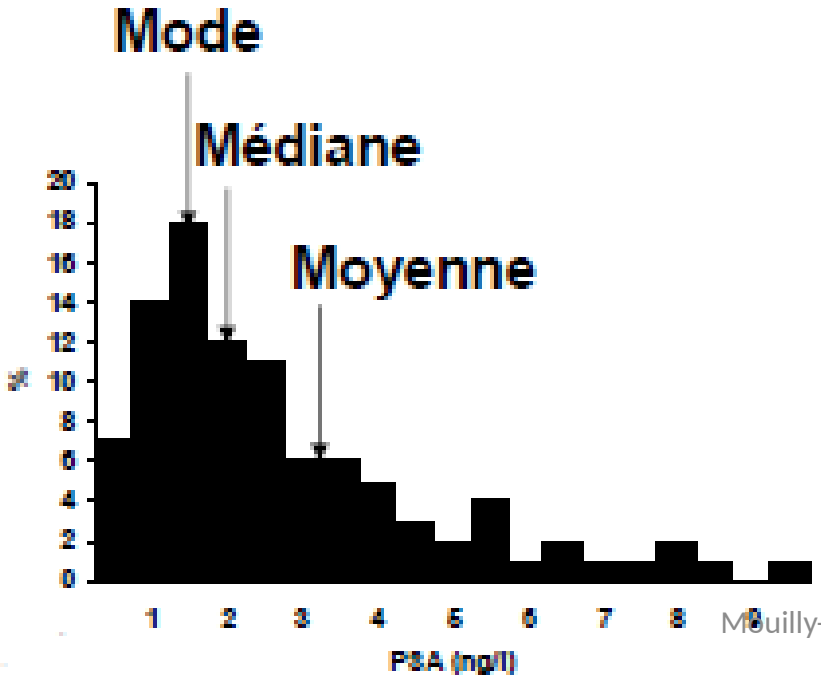


Mode, médiane, moyenne

- Si distribution asymétrique

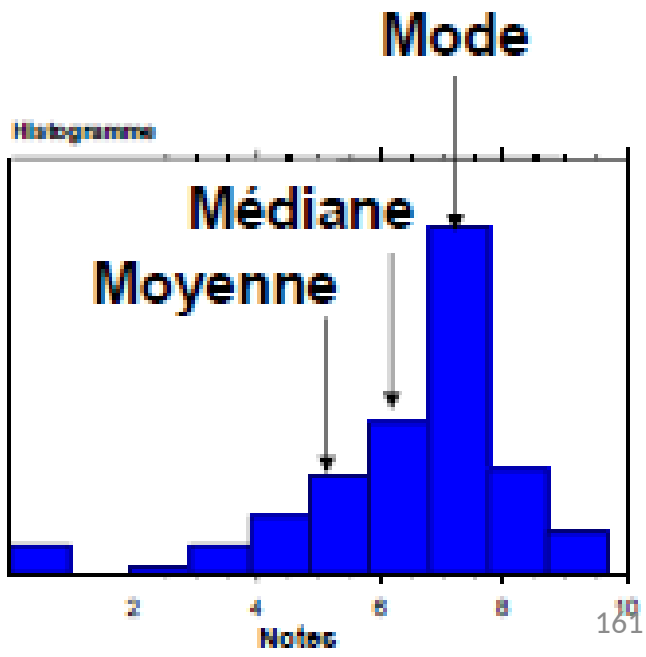
à droite

mode < médiane < moyenne



à gauche

moyenne < médiane < mode



3 Indicateurs caractérisent la distribution d'une série statistique quantitative:

Les paramètres de position :

Moyenne « arithmétique », Médiane, Mode, Fractile, Min, Max

Les paramètres de dispersion :

Variance, Écart-type, Coefficient de Variance, Étendue (Amplitude), Ecart Interquartiles, Interquartile relatif, Intervalle De Kelley, Interdécile relatif

Les paramètres de forme :

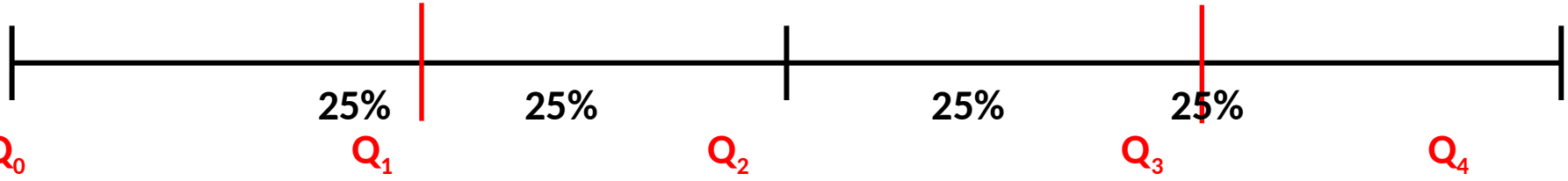
Aplatissement (Kurtosis), Asymétrie (Skewness)

Fractiles : quartiles, déciles, centiles

Quartiles

Variable discrète

- **Le 1er quartile, noté Q_1, est une valeur de la série; telle que 25 % au moins des valeurs de la série sont inférieures ou égales à Q_1; et telle que 75% au moins des valeurs de la série sont supérieures ou égales à Q_1 ⊔ $F_i = 0,25$**
- **Le 2eme, noté Q_2 = médiane ⊔ $F_i = 0,50$**
- **Le 3eme quartile, noté Q_3, est : une valeur de la série; telle que 75% au moins des valeurs de la série sont inférieures ou égales à Q_3; et telle que 25% au moins des valeurs de la série sont supérieures ou égales à Q_3 ⊔ $F_i = 0,75$**



N =	Q1	Q2	Q3
N = 4n	entre la valeur de rang n et celle de rang n+1	entre la valeur de rang 2n et celle de rang 2n+1	entre la valeur de rang 3n et celle de rang 3n+1
N = 4n + 1	entre la valeur de rang n et celle de rang n+1	la valeur de rang 2n+1	entre la valeur de rang 3n+1et celle de rang 3n+2
N = 4n + 2	la valeur de rang n+1	entre la valeur de rang 2n+1 et celle de rang 2n+2	la valeur de rang 3n+2
N = 4n + 3	la valeur de rang n+1	la valeur de rang 2n+2	la valeur de rang 3n+3

Exemple: **on fait une étude statistique sur les 50 notes attribuées par un jury à un examen, voici les résultats obtenus en classant ces notes par ordre croissant (variable discrète).**

Q_0 Q_1 Q_2 Q_3 Q_4 Q_5

Notes	Effectif	Effectif cumulé
0	1	1
1	2	3
2	2	5
3	3	8
4	2	10
5	3	13
6	2	15
7	3	18
8	4	22
9	3	25
10	2	27
11	3	30
12	4	34
13	4	38
14	3	41
15	1	42
16	2	44
17	1	45
18	2	47
19	2	49
20	1	50

N = 4n + 2	la valeur de rang n+1 Q_1	entre la valeur de rang 2n+1 et celle de rang 2n+2 Q_2	la valeur de rang 3n+2 Q_3

n=12 50=4*12 +2

Notes	Effectif	Effectif cumulé
0	1	1
1	2	3
2	2	5
3	3	8
4	2	10
5	3	13
6	2	15
7	3	18
8	4	22
9	3	25
10	2	27
11	3	30
12	4	34
13	4	38
14	3	41
15	1	42
16	2	44
17	1	45
18	2	47
19	2	49
20	1	50

N = 4n + 2	la valeur de rang n+1 Q_1	entre la valeur de rang 2n+1 et celle de rang 2n+2 Q_2	la valeur de rang 3n+2 Q_3

n=12 50=4*12 +2

Le premier quartile Q_1 = 5 (rang 13)

Le second quartile Q_2 = 9,5 (entre rang 25 et 27)

Le troisième quartile Q_3 = 13 (rang 38)

Notes	Effectif	Effectif cumulé
0	1	1
1	2	3
2	2	5
3	3	8
4	2	10
5	3	13
6	2	15
7	3	18
8	4	22
9	3	25
10	2	27
11	3	30
12	4	34
13	4	38
14	3	41
15	1	42
16	2	44
17	1	45
18	2	47
19	2	49
20	1	50

Exemple: on fait une étude statistique sur les 50 notes attribuées par un jury à un examen, voici les résultats obtenus en classant ces notes par ordre croissant (variable discrète).

Min = Q_0

Max= Q_4

Mediane= Q_2

~~Q_5~~ ????

Notes	Effectif	Effectif cumulé
0	1	1
1	2	3
2	2	5
3	3	8
4	2	10
5	3	13
6	2	15
7	3	18
8	4	22
9	3	25
10	2	27
11	3	30
12	4	34
13	4	38
14	3	41
15	1	42
16	2	44
17	1	45
18	2	47
19	2	49
20	1	50

Déciles

- **Le terme vient de dix. Les déciles notés D1, D2,….,D9 partagent la série en 10 parties d'égal effectif.**
- **Médiane est alors = D5**

Les Centiles (Percentiles)

- **Le terme vient de cent. Les paramètres C1, C2……C99 (99 centiles) d'ordre 1, 2,…99% partagent la série en 100 parties de même taille.**
- **Médiane = C50**
- **10ième percentile : x_i tel que Fi = 0,10**

NB. Pour ces fractiles c'est le même raisonnement de calcul réalisé pour les quartiles

Exemple: **on fait une étude statistique sur les 50 notes attribuées par un jury à un examen, voici les résultats obtenus en classant ces notes par ordre croissant (variable discrète).**

D_0 D_5 D_{10} ?????

C_0 C_{50} C_{100} ?????

Notes	Effectif	Effectif cumulé
0	1	1
1	2	3
2	2	5
3	3	8
4	2	10
5	3	13
6	2	15
7	3	18
8	4	22
9	3	25
10	2	27
11	3	30
12	4	34
13	4	38
14	3	41
15	1	42
16	2	44
17	1	45
18	2	47
19	2	49
20	1	50

Exemple: **on fait une étude statistique sur les 50 notes attribuées par un jury à un examen, voici les résultats obtenus en classant ces notes par ordre croissant (variable discrète).**

Min = Q_0 = D_0 = C_0

Max= Q_4 = D_{10} = C_{100}

Mediane= Q_2 = D_5 = C_{50}

Notes	Effectif	Effectif cumulé
0	1	1
1	2	3
2	2	5
3	3	8
4	2	10
5	3	13
6	2	15
7	3	18
8	4	22
9	3	25
10	2	27
11	3	30
12	4	34
13	4	38
14	3	41
15	1	42
16	2	44
17	1	45
18	2	47
19	2	49
20	1	50

Fractiles : quartiles, déciles, centiles

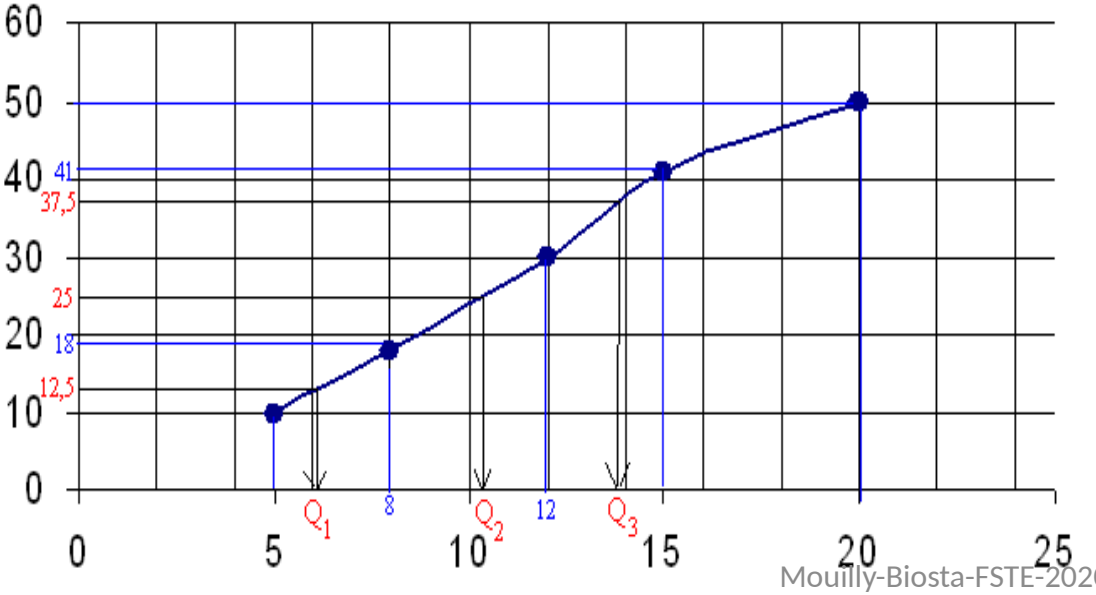
Quartiles

Variable continue

Calcule des quartiles par interpolation linéaire

Notes	Effectifs	Effectifs cumulés
[0 ; 5[	10	10
[5 ; 8[	8	18
[8 ; 12[	12	30
[12 ; 15	11	41
[15 ; 20	9	50
	50	

Construction d'un polygone des effectifs cumulés croissants :



Rappel

L'interpolation linéaire est la méthode la plus simple pour estimer la valeur prise par une fonction continue entre deux points déterminés (interpolation).

Elle consiste à utiliser pour cela la fonction affine (de la forme f(x) = a.x +b) passant par les deux points déterminés.

3 Indicateurs caractérisent la distribution d'une série statistique quantitative:

Les paramètres de position :

Moyenne « arithmétique », Médiane, Mode, Fractile, Min, Max

Les paramètres de dispersion :

Variance, Écart-type, Coefficient de Variance, Étendue (Amplitude), Ecart Interquartiles, Interquartile relatif, Intervalle de Kelley, Interdécile relatif

Les paramètres de forme :

Aplatissement (Kurtosis), Asymétrie (Skewness)

Variable Discontinue

Les distributions statistiques peuvent, tout en ayant des caractéristiques de tendance centrale voisines, être très différentes. Il est donc nécessaire de mesurer la dispersion des valeurs autour des tendances centrales.

- **Variance noté V(X) ou VAR(X) : Il mesure la dispersion de part et d'autre de la moyenne. C'est la moyenne de la somme carrés des écarts à la moyenne**

$$VAR(X) = \frac{1}{N}\sum_i (x - \mu)^2$$

Avec pondération

$$VAR(X) = \frac{1}{N}\sum_i n_i (x - \mu)^2$$

$$\sigma(X) = \sqrt{VAR(X)}$$

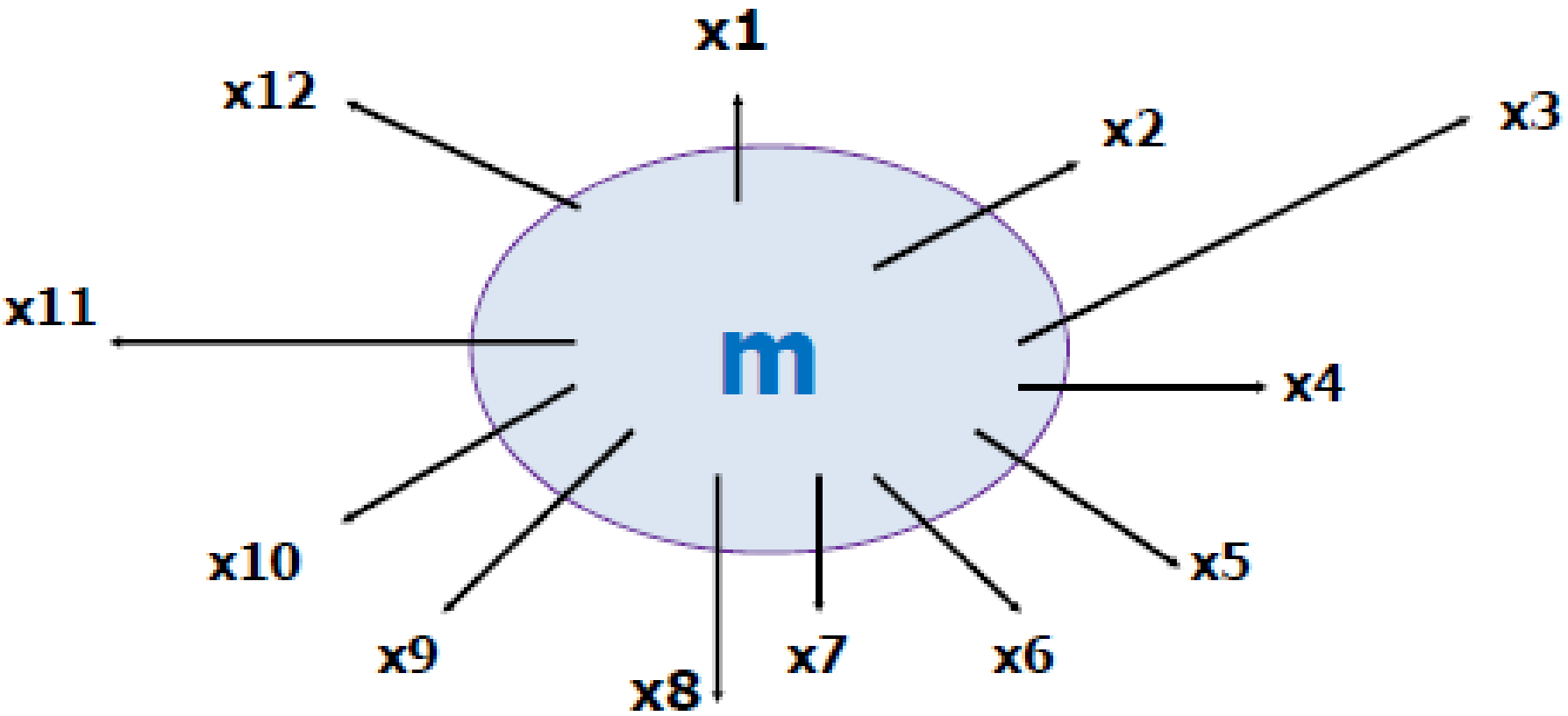
- **Écart type = Ecart moyen quadratique=Standard Déviation "SD"**

= la racine carré de la variance et donc toujours positif ou nul.

- **Plus σ est grand, plus les valeurs du caractère sont dispersées autour de la moyenne**
- **Plus σ est petit, plus les valeurs du caractère sont groupées autour de la moyenne**

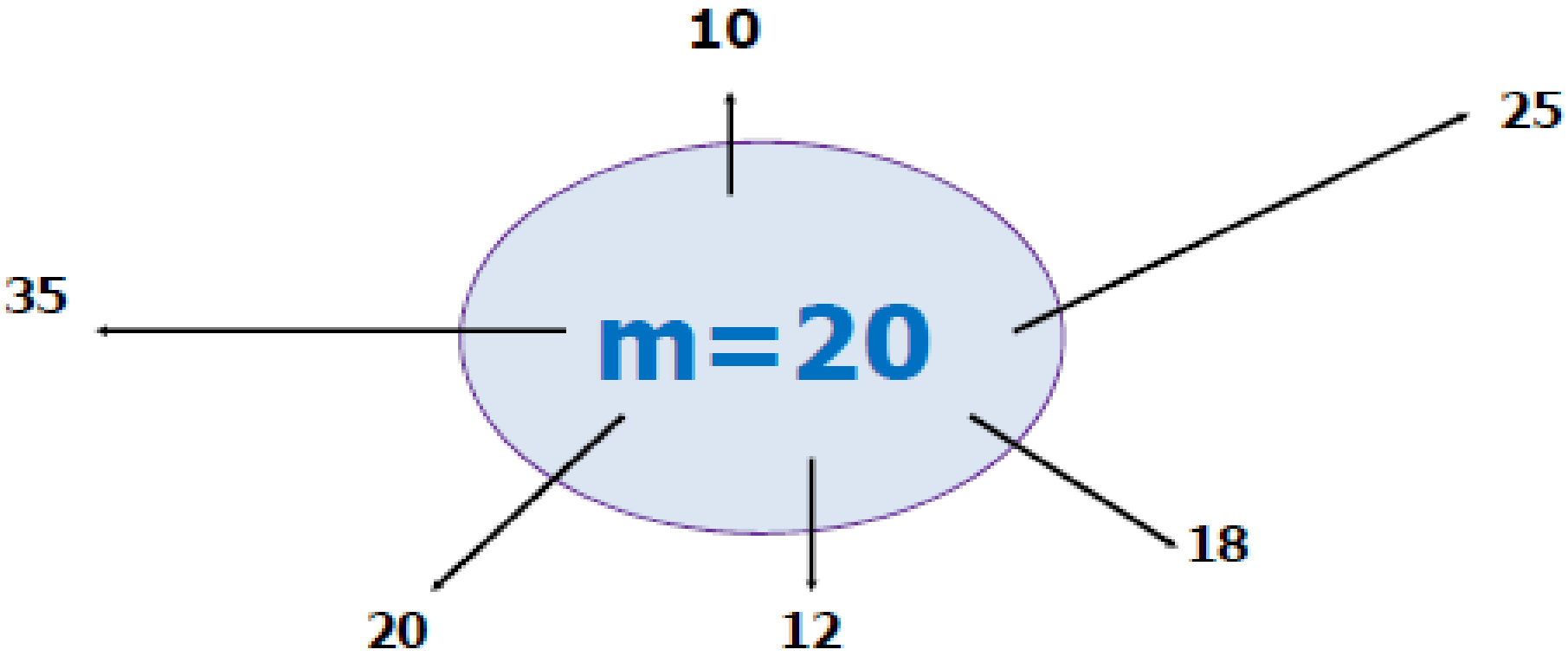
NB. Dans le cas d'un échantillon $VAR(X) = \frac{1}{n-1}\sum_i (x - \bar{x})^2$ $VAR(X) = \frac{1}{N}\sum_i n_i (x - \bar{x})^2$ $\delta(X) = \sqrt{VAR(X)}$

Écart-type



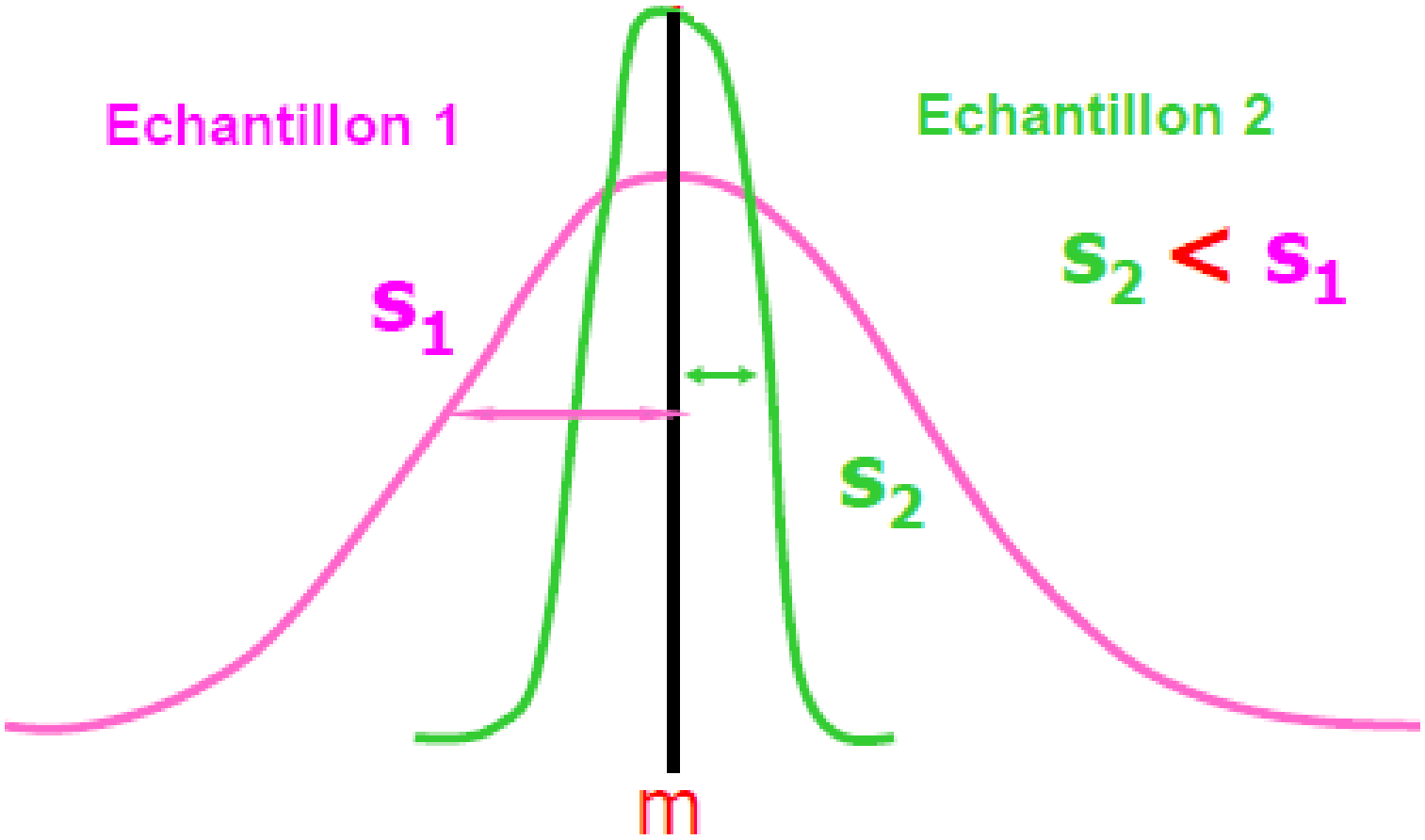
Représente l ’écart moyen des données de l’échantillon par rapport à la moyenne

Écart-type



Représente l ’écart moyen des données de l’échantillon par rapport à la moyenne

Signification probabiliste de l'écart-type



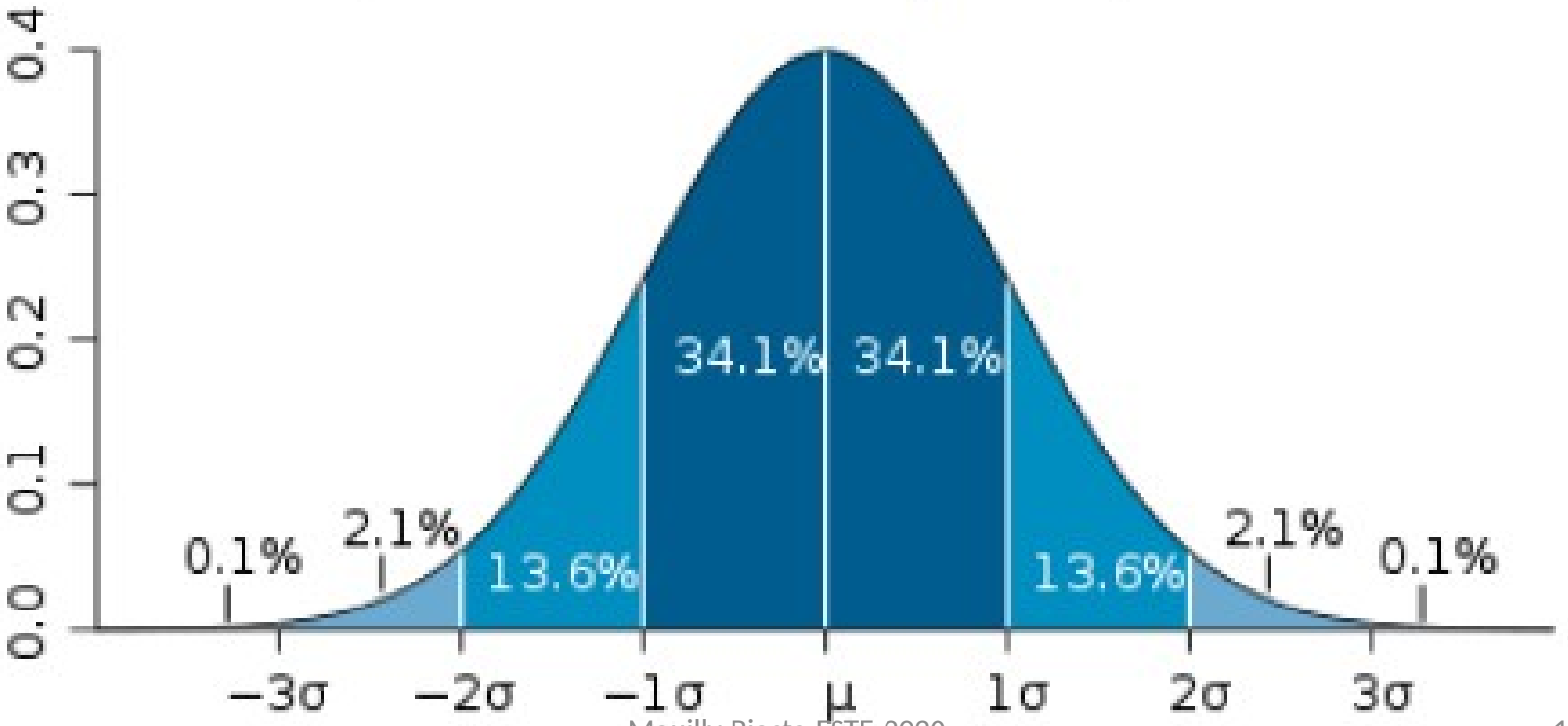
Signification probabiliste de l'écart-type

50 % des individus en-dessous de la moyenne et 50 % au-dessus

68 % des individus entre μ-1σ et μ+1σ

95 % des individus entre μ-1,96σ et μ+1,96σ

99,7 % des individus entre μ -3σ et μ+3σ



La variance
exemple

- *Calculer la variance s^2 : 10, 12, 18, 20, 25, 35*
  - *Calculer la moyenne : m=20*

Observations xi	10	12	18	20	25	35
Différence à la moyenne xi- 20						
Carré de la différence à la moyenne						

  - *Calculer la somme des carrés de la différence à la moyenne :*
  - *Diviser la somme des carrés par n -1 soit :*

La variance
exemple

- *Calculer la variance s^2 : 10, 12, 18, 20, 25, 35*
  - *Calculer la moyenne : m=20*

Observations xi	10	12	18	20	25	35
Différence à la moyenne xi- 20	-10	-8	-2	0	+5	+15
Carré de la différence à la moyenne	100	64	4	0	25	225

  - *Calculer la somme des carrés de la différence à la moyenne : 100+64+4+0+25+225=418*
  - *Diviser la somme des carrés par n -1 soit :* $\frac{418}{6-1}$ = 83,6

→Le Coefficient de Variation =Ecart type relatif, noté C.V

c'est le rapport entre l'écart type et la moyenne, il permet de comparer le taux de dispersion entre distributions même qui ne possèdent pas la même unité

$$C.V = \frac{\sigma_x}{\bar{x}} 100$$

C.V est sans unité

Plus le coefficient de variation est petit, plus la série est homogène.

D'une manière générale, la population étudiée est considérée

Homogène lorsque le CV < 15%

Dispersé lorsque CV >15%

Exemple : Récolte d'abricots dans plusieurs parcelles homogènes d'une ferme biologique

moyenne = 54,9 kg

x	n_i	(x-54,9)	$(x-54,9)^2$	$n_i (x-54,9)^2$
42	5	- 12,9	166,41	832,05
47	12	- 7,9	62,41	748,92
52	31	- 2,9	8,41	260,40
57	31	+ 2,1	4,41	136,71
62	16	+ 7,1	50,41	806,56
67	3	+ 12,1	146,41	439,23
72	2	+ 17,1	292,41	584,82
Total	**100**			**3808,82**

Ecart type simple $\mathrm{VAR(X)} = \frac{1}{N}\sum_i (x - \mu)^2$

Ecart type pondéré $\mathrm{VAR(X)} = \frac{1}{N}\sum_i n_i (x - \mu)^2$

Exemple : Récolte d'abricots dans plusieurs parcelles homogènes d'une ferme biologique

moyenne = 54,9 kg

x	n_i	(x-54,9)	$(x-54,9)^2$	$n_i (x-54,9)^2$
42	5	- 12,9	166,41	832,05
47	12	- 7,9	62,41	748,92
52	31	- 2,9	8,41	260,40
57	31	+ 2,1	4,41	136,71
62	16	+ 7,1	50,41	806,56
67	3	+ 12,1	146,41	439,23
72	2	+ 17,1	292,41	584,82
Total	100			3808,82

Ecart type simple

Ecart type pondéré

$\sigma^2 = 3808,82 / 100 = 38,09\ Kg^2$

$\sigma = 6.17\ Kg$

récolte moyenne est

$54,90 \pm 6,17$ Kg

C.V. = 6,17 /54,9=11,3%

Variable Continue

Ci : centre de la classe

$$VAR(X) = \frac{1}{n}\sum_i (ci - \bar{x})^2 \qquad \sigma(X) = \sqrt{VAR(X)}$$

Avec pondération

$$VAR(X) = \frac{1}{n}\sum_i ni(ci - \bar{x})^2$$

Exemple

Calculer la variance et l'écart type. Rappelons que $m = 35.5$ ans.

Age (années)	Centre des classes	Effectifs n_i	$\lvert x_i - m \rvert$	$(x_i - m)^2$	$(x_i - m)^2 n_i$
[20, 25[	22.5	9			
[25, 30[	27.5	27			
[30, 35[	32.5	36			
[35, 40[	37.5	45			
[40, 45[	42.5	18			
[45, 50[	47.5	9			
[50, 55[	52.5	3			
[55, 60[	57.5	3			
		$\Sigma n_i = N = 150$			

A partir de la définition :

$$V_X = \sigma^2 = \frac{\sum_i (x_i - m)^2 n_i}{\sum_i n_i}$$

On déduit :

Exemple

Calculer la variance et l'écart type. Rappelons que $m = 35.5$ *ans*.

Age (années)	Centre des classes	Effectifs n_i	$\lvert x_i - m \rvert$	$(x_i - m)^2$	$(x_i - m)^2 n_i$
[20, 25[	22.5	9	13	169	1521
[25, 30[	27.5	27	8	64	1729
[30, 35[	32.5	36	3	9	324
[35, 40[	37.5	45	2	4	180
[40, 45[	42.5	18	7	49	882
[45, 50[	47.5	9	12	144	1296
[50, 55[	52.5	3	17	189	867
[55, 60[	57.5	3	22	484	1452
		$\sum n_i = N = 150$			8250

A partir de la définition :

$$V_X = \sigma^2 = \frac{\sum_i (x_i - m)^2 n_i}{\sum_i n_i}$$

On déduit :

$$\sigma^2 = V_X = \frac{8250}{150} = 55$$

$$\sigma = \sqrt{V_X} = \sqrt{55} \approx 7.4 \text{ ans}$$

⊔ Etendu =Amplitude

- **Ecart entre la plus grande (max) et la plus petite (min) des observations**
- **Ce paramètre est totalement lié à ces 2 valeurs extrêmes et donc peu fiable. Néanmoins, il donne une première idée de la dispersion des observations.**

→ Ecart interquartile : Q3 – Q1

- **C'est un indicateur de dispersion. Il mesure l'écart entre le 3ème et 1er quartile.**
- **Cet intervalle correspond à 50% des observations. situées autour de la médiane.**

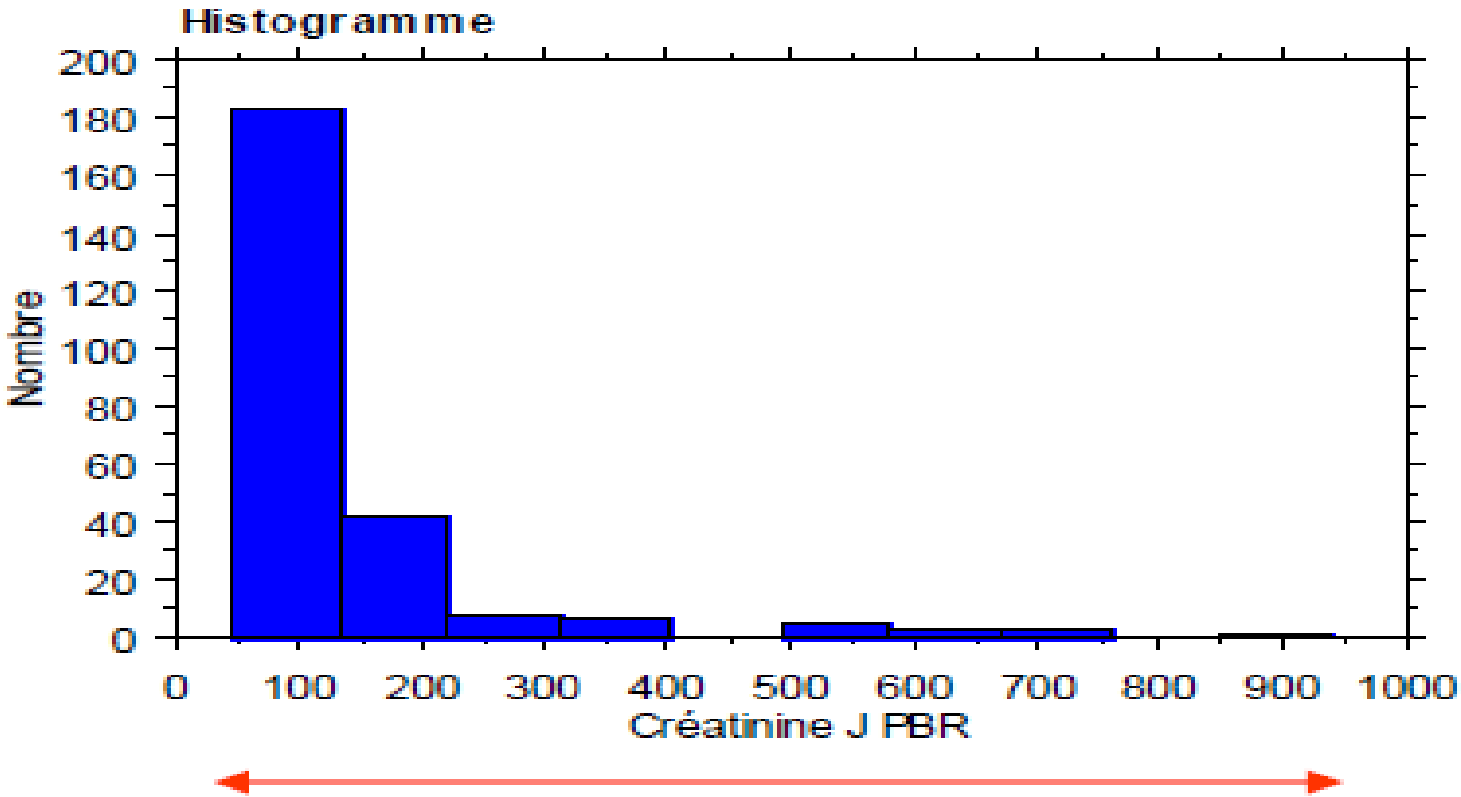
Cette mesure n'est pas sensible aux valeurs éloignées. Elle a l'avantage par rapport à celui de l'étendue d'écarter les valeurs extrêmes, souvent sans signification.

→ Interquartile relatif = Ecart Semi-interquartile : Q3 – Q1 / Q2

→ Intervale de Kelley = D9-D1 : mesure l'écart pour 80% des observations

→ Interdécile relatif = D9 – D1/ D5

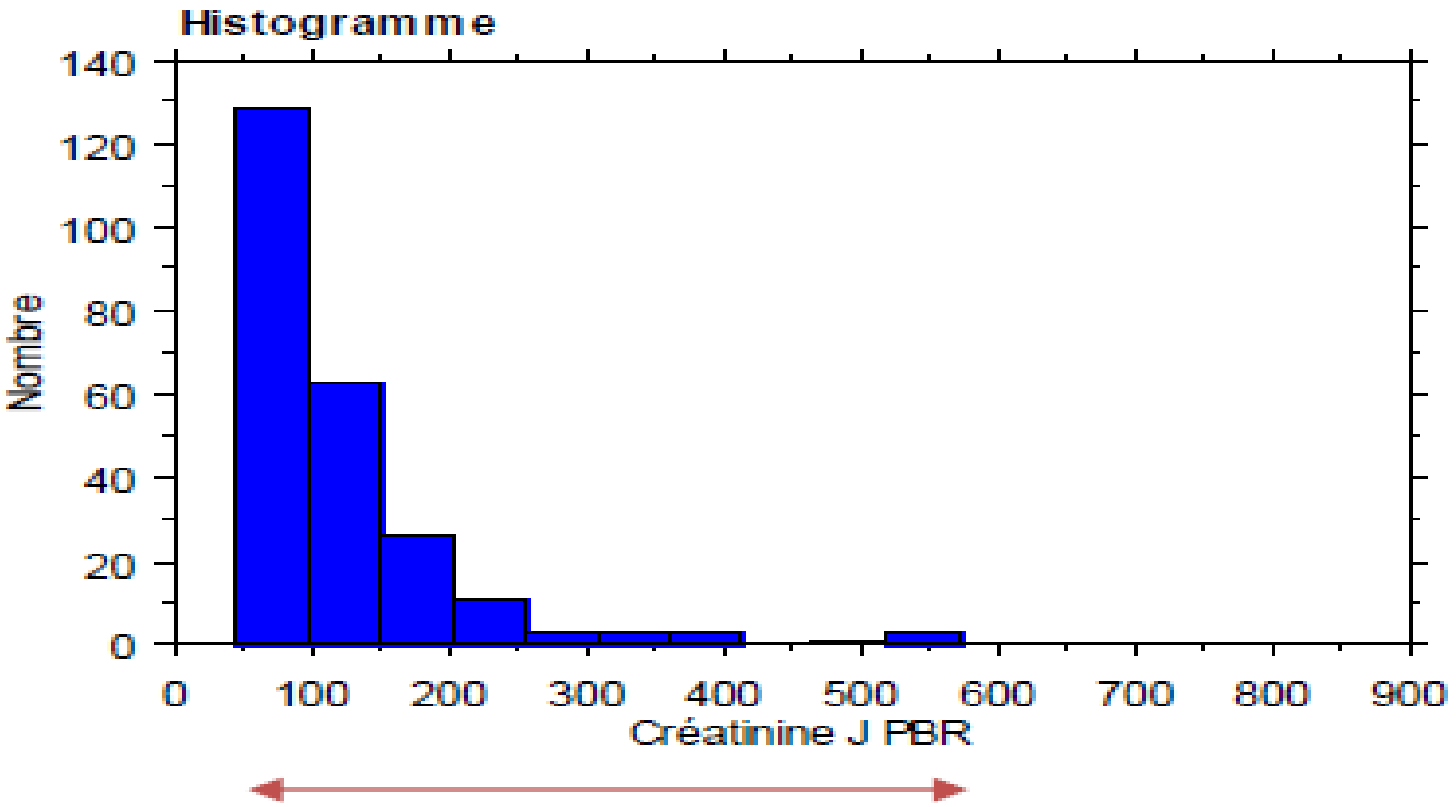
Etendu



Valeur min = 45µmol/l

Valeur max = 939 µmol/l

Etendu



Valeur min = 45µmol/l

Valeur max = 572 µmol/l

Approche Probabiliste

Variable aléatoire -> Variable Continue versus Discontinue

Voir partie 1 ▭ Distribution théorique

Variance & Esperance

$$V(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$$

$$\sigma(X) = \sqrt{\mathrm{VAR}(X)}$$

$$E(X) = \sum_{i=1}^{n} x_i p_i$$

$$V[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \sum_{i=1}^{N} p_i \cdot (x_i - \mathbb{E}[X])^2$$

$$\sigma(X) = \sqrt{VAR(X)}$$

$$\mathbb{E}[X] = \sum_{i=1}^{N} p_i \cdot x_i = \sum_{i=1}^{N} x_i \cdot \mathbb{P}[X = x_i]$$

Variable Discontinue

- **La loi Uniforme**

$$\mathbb{E}[X] = \frac{n+1}{2}, \qquad V(X) = \frac{n^2 - 1}{12}.$$

- **La loi de Bernoulli**

$$\mathbb{P}[X = 1] = p, \qquad \mathbb{P}[X = 0] = 1 - p.$$

- **La loi Binomiale**

$$\mathbb{E}[X] = np, \qquad V[X] = np(1 - p).$$

- **La loi de Poisson**

$$\mathbb{E}[X] = \lambda, \qquad V[X] = \lambda.$$

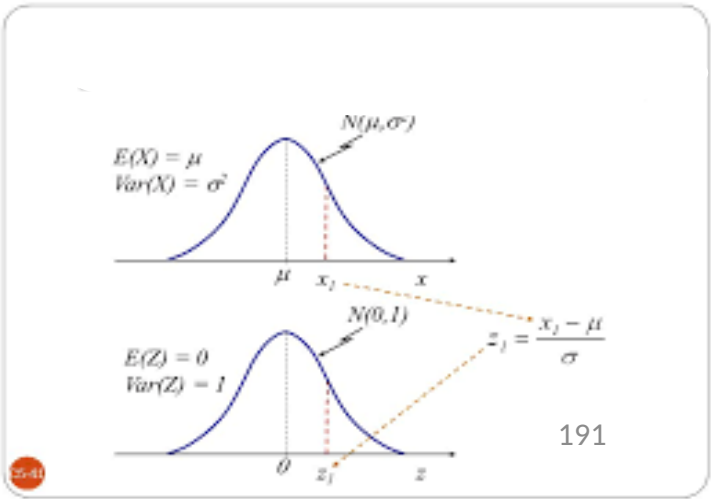
Variable Continue

- **La loi Normale**

$$E(X) = \mu$$
$$V(X) = \sigma^2$$

- **La loi normale réduite**

$$E(X) = 0$$
$$V(X) = 1$$



3 Indicateurs caractérisent la distribution d'une série statistique quantitative:

Les paramètres de position :

Moyenne « arithmétique », Médiane, Mode, Fractile, Min, Max

Les paramètres de dispersion :

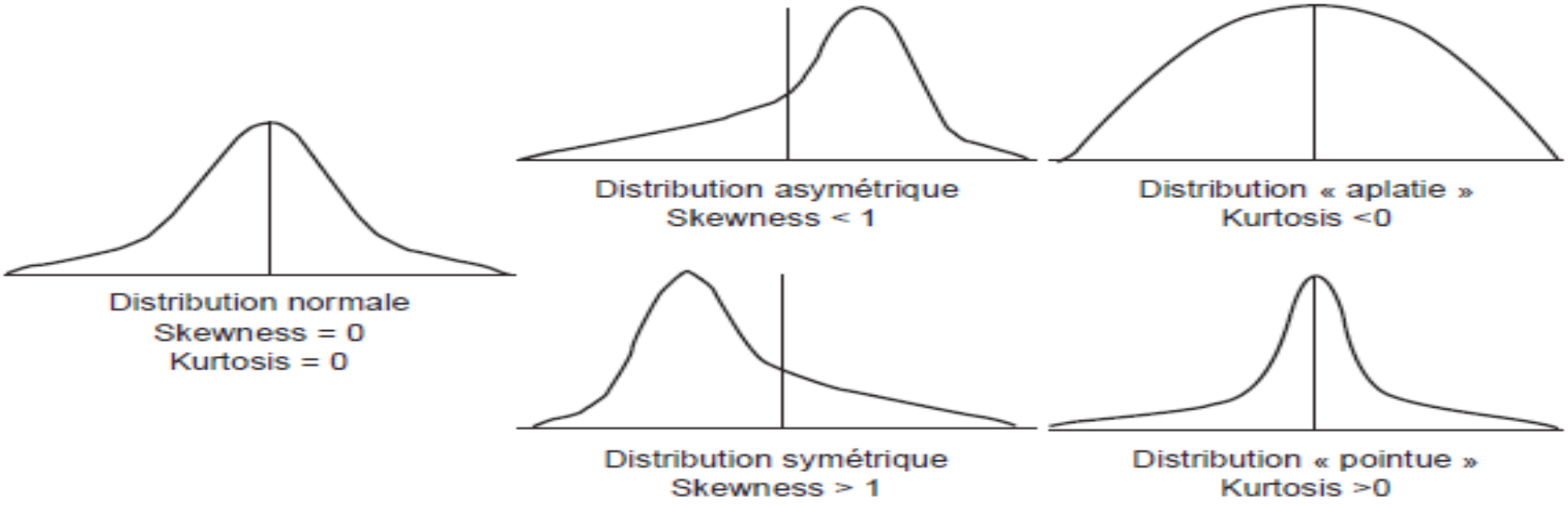
Variance, Écart-type, Coefficient de Variance, Étendue (Amplitude), Ecart Interquartiles, Interquartile relatif, Intervalle de Kelley, Interdécile relatif

Les paramètres de forme :

Aplatissement (Kurtosis), Asymétrie (Skewness)

En complément de l'étude de la position et de la dispersion, il est intéressant de repérer la forme (déjà mise en évidence graphiquement) par de mesures de son asymétrie (Skewness) et de son aplatissement (Kurtosis).

Représentations de distributions statistiques.



Coefficient d'asymétrie (skewness)

population

$$\gamma_1 = \frac{1}{N}\sum\left(\frac{x_i - \mu}{\sigma}\right)^3$$

échantillon

$$\gamma_1 = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i - m}{s}\right)^3$$

Interprétation - pour ce cours:

- Si $S_K < 0$ il y a une asymétrie négative (ou un biais à gauche) moyenne > médiane
- Si $S_K \geq 0$ il y a une asymétrie positive (ou un biais à droite) moyenne < médiane
- Si $|S_K| \leq 0{,}5$, l'asymétrie est négligeable
- Si $0{,}5 < |S_K| \leq 2$, l'asymétrie est modérée
- Si $|S_K| > 2$, l'asymétrie est prononcée

Coefficient aplatissement (kurtosis)

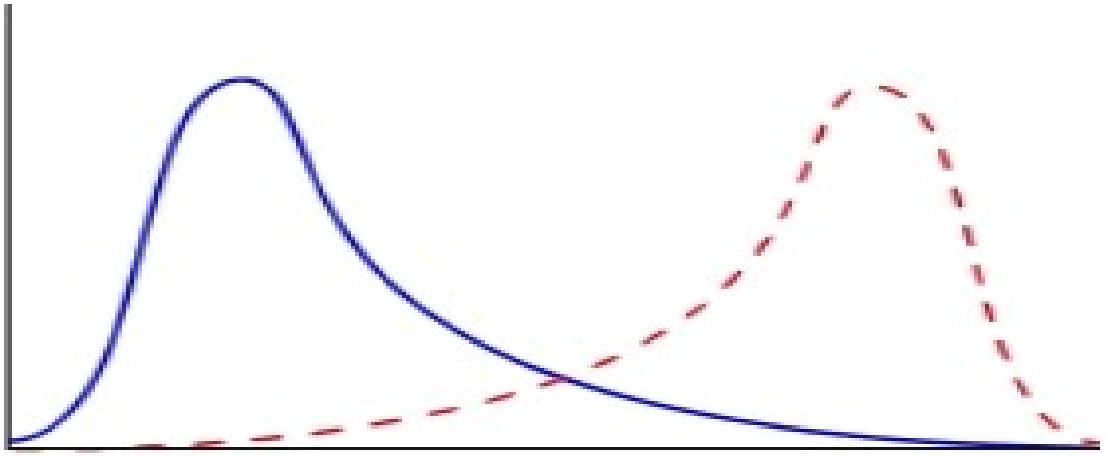
population

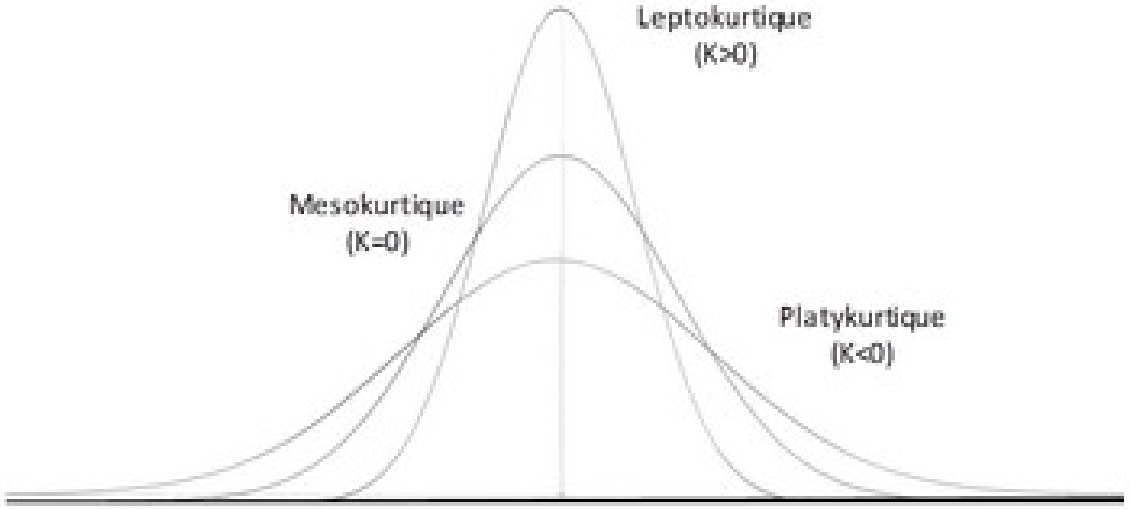
$$\gamma_2 = \frac{1}{N}\sum\left(\frac{x_i - \mu}{\sigma}\right)^4$$

échantillon

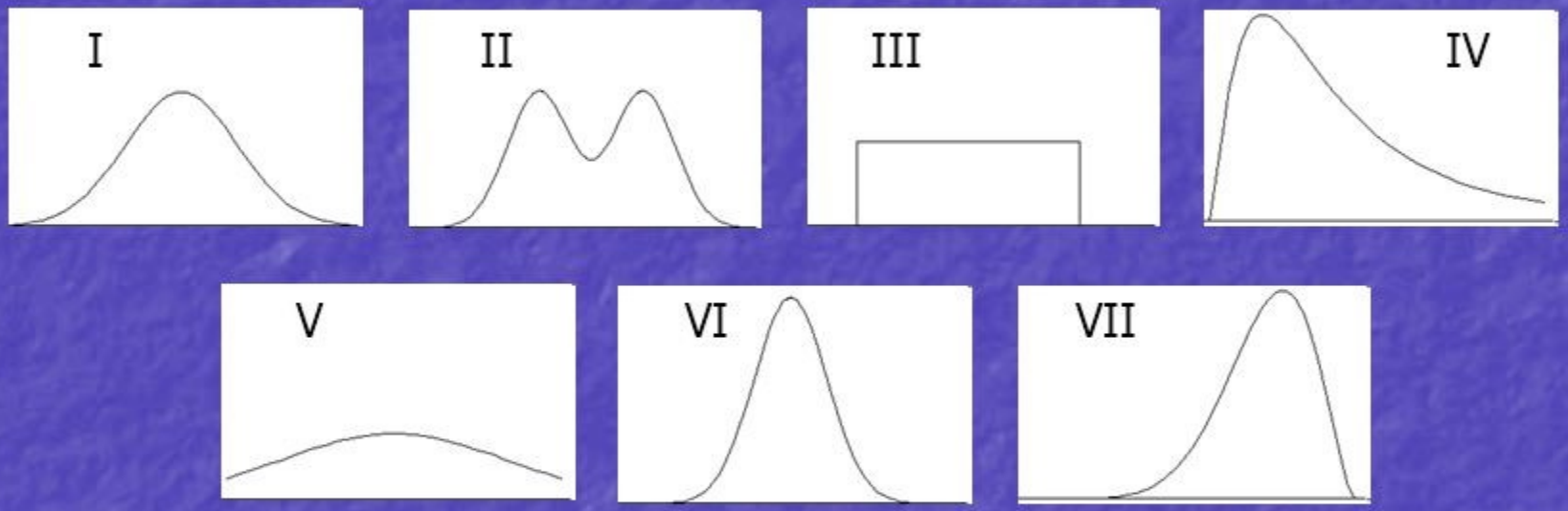
$$\gamma_2 = \frac{n(n+1)}{(n-1)(n-2(n-3)}\sum\left(\frac{x_i - m}{s}\right)^4$$

- Si $K < 0$ la courbe est aplatie (distribution **platykurtique**)
- Si $K \geq 0$ la courbe est étirée (distribution **leptokurtique**)
- Si $|K| \leq 0{,}5$ la courbe n'est ni trop aplatie, ni trop étirée (distribution **mésokurtique**)
- Si $0{,}5 < |K| \leq 2$, l'aplatissement est modéré
- Si $|K| > 2$, l'aplatissement est prononcé


skewness positif
skewness négatif


Leptokurtique
(K>0)
Mesokurtique
(K=0)
Platykurtique
(K<0)

Formes des distributions de fréquences



Modalité

- unimodale : I, IV,V, VI, VII
- bimodale : II
- Rectangulaire : III

Courbure (kurtosis)

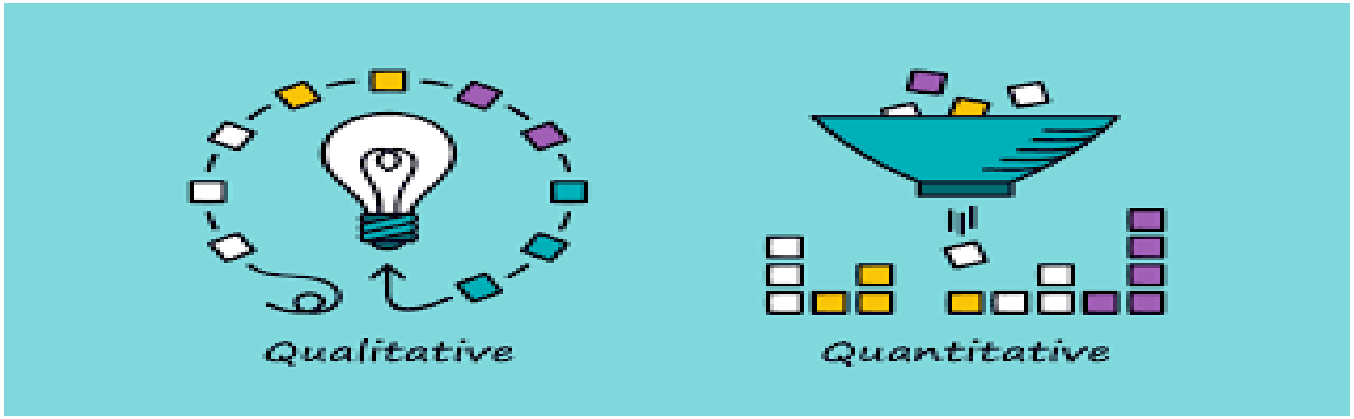
- Mesokurtique : I, II
- Platykurtique : V
- Leptokurtique : IV,VI,VII

Symétrie

- symétrique : I, II, III, V, VI
- asymétrique : IV, VII

Chapitre 1 : Statistique univariée (1 dimension)

1.
    1. Distributions des fréquences
    2. Paramètres de position, de dispersion et de forme
2. **Variable Qualitative (Nominale & Ordinale)**
3.



IMPORTANT

Variables

Quantitatives
- Discrètes
- Continues

Qualitatives
- Nominales
- Ordinales
- Binaires

MODALITÉS

Temporelles
- Date
- Heure

une variable catégorielle = Facteur

Est une variable qui prend pour valeur des *modalités*, des *catégories* ou bien des *niveaux*, par opposition aux variables quantitatives qui mesurent sur chaque individu une quantité

Valeurs numériques

Caractérisation

- **Répartition proportionnelle :**
  - ✓ **Pourcentage %**
  - ✓ **Par mille ‰**
  - ✓ **Sex ratio…**
- **Taux**
- **Risque**
- **Prévalence**
- **Incidence**
- **……**

Caractéristiques de la population

Tableau I

Caractéristiques de la population étudiée (patients infectés et non infectés).

	Total patients inclus n = 702	Patients infectés n = 91	Patients non infectés n = 611	OR (95 %)	Valeur de p
Âge moyen en jours (écart-type)	3,20 ± 5,23	3,26 ± 5,29	3,19 ± 5,22	0,55 [−1,2–1,16]	NS
Sex-ratio	1,4	1,75	1,39	1,55 [0,99–2,41]	0,052
Masculin (%)	414	58 (63,7)	356	1,44 [0,99–2,09]	
Féminin (%)	288	33 (27,3)	255	0,93 [0,86–0,99]	
Âge gestationnel (SA)					
SA + 6j	74 (10,6 %)	19 (20,8 %)	55 (9 %)		< 0,001
SA + 6j	182 (25,9 %)	36 (39,6 %)	146 (23,9)		< 0,001
≥ 37 SA	446 (63,5 %)	36 (39,6 %)	410 (67,1 %)		NS
Poids d'admission					
1000–1499 g	74 (10,6 %)	19 (20,8 %)	55 (9 %)		< 0,001
1500–1999 g	141 (21 %)	18 (19,8 %)	123 (20,1 %)		< 0,001
2000–2499 g	103 (14,7 %)	18 (19,8 %)	85 (13,9 %)		< 0,001
2500–3999 g	314 (44,8 %)	30 (32,9 %)	284 (64,5 %)		NS
> 4000 g	70 (9,9 %)	6 (6,7 %)	64 (10,5 %)		NS

Répartition des différents signes cliniques retrouvés chez les patients adultes atteint de leishmaniose viscérale

Signes cliniques	Effectif	Pourcentage (n = 12)
fièvre	12	100%
splénomégalie	7	58%
amaigrissement	7	58%
pâleur	4	33%
arthralgies	4	33%
adénopathies	1	8%
hépatomégalie	3	25%
Asthénie + anorexie	3	25%

Chapitre 1 : Statistique univariée (1 dimension)

1.
    1. Distributions des fréquences
    2. Paramètres de position, de dispersion et de forme
2. Variable Qualitative (Nominale & Ordinale)
3. **Représentation Graphique**

Représentation graphique

Définition : C'est un mode d'expression qui permet "visuellement" de saisir et de mémoriser un certain nombre d'informations (en complément de tableaux).

Cette représentation peut répondre à deux types d'objectifs:

- **Être un moyen de communication et permettre de véhiculer une information.**
- **Être un instrument de travail et permettre une vue d'ensemble synthétique du phénomène étudié , ce qui en facilite l'analyse.**

Caractérisation, selon :

- **La nature de la série représentée : chronologiques, spatiaux, quantitatifs ou qualitatifs**
- **l'échelle retenue : échelle arithmétique, logarithmique, exponentielle,…**

Représentation graphique

Recommandations

- **Bien choisir le titre**
- **Légendes claires et concises Annoter les axes (ex : l'abscisse représente l'âge en années, l'ordonnée les effectifs)**
- **Ne pas surcharger un rapport de graphiques (trop de communication tue la communication !!!)**
- **Privilégier les tableaux pour les résultats communs et mettre en évidence les résultats intéressants au moyen de graphiques**
- **Utiliser les indicateurs et les représentations graphiques adéquats !!!**

NB. Un bon graphique est souvent plus explicite qu'un long discours.

Représentation graphique: (synthèse)

Variables Qualitatives

- ✓ Diagramme circulaire = Diagramme en secteurs ou « en camembert = pie-chart »
- ✓ Diagramme en bâtons
- ✓ Diagramme en barres =à Bande =en Tuyaux d'orgues

Variables Quantitatives

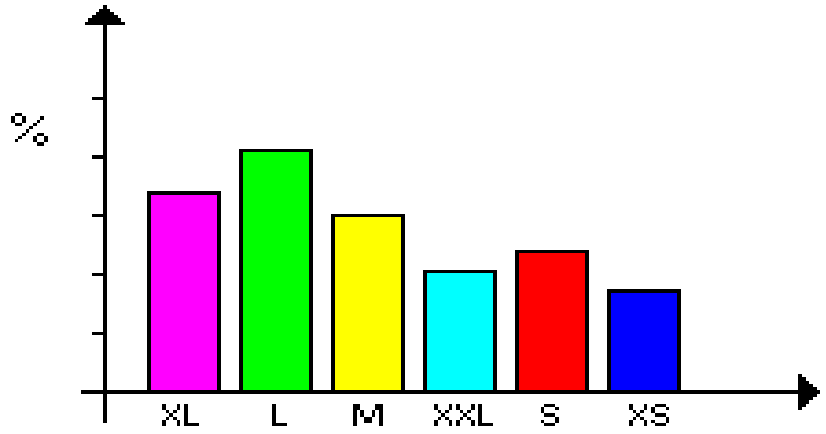
Valeurs Discontinues

- ✓ Diagramme en bâtons
- ✓ Histogramme
- ✓ Polygone des effectifs (ou fréquences)
- ✓ Courbe cumulative des effectifs (ou des fréquences)

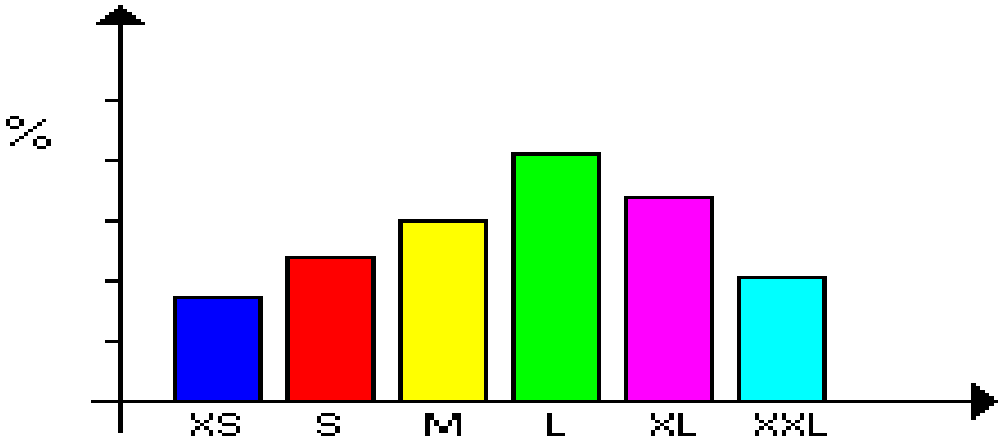
Valeurs continues

- ✓ Histogramme
- ✓ Polygone des effectifs (ou fréquences)
- ✓ Courbe cumulative des effectifs (ou des fréquences)

Représentation graphique (suite)

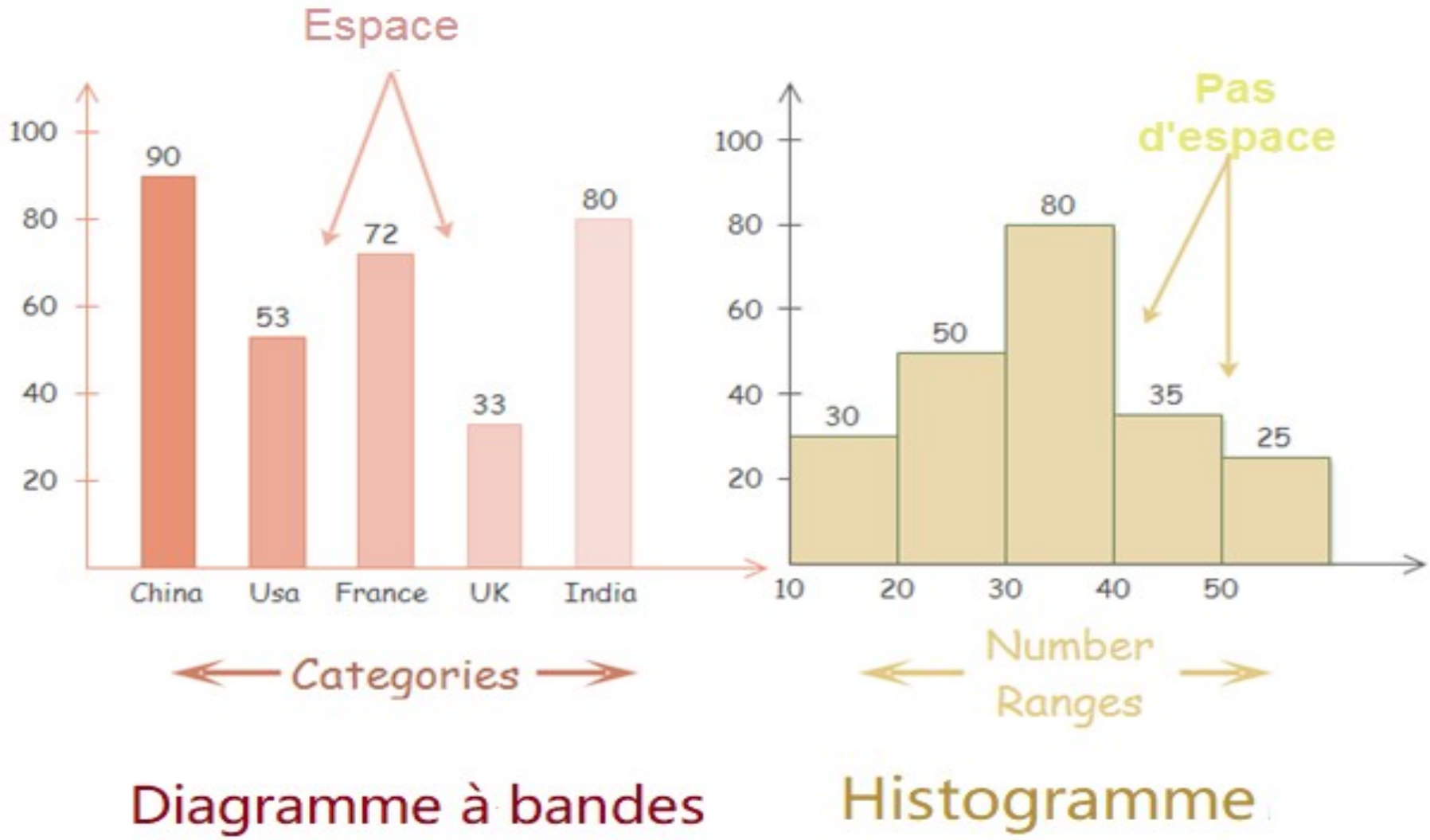


graphique incorrect



graphique correct

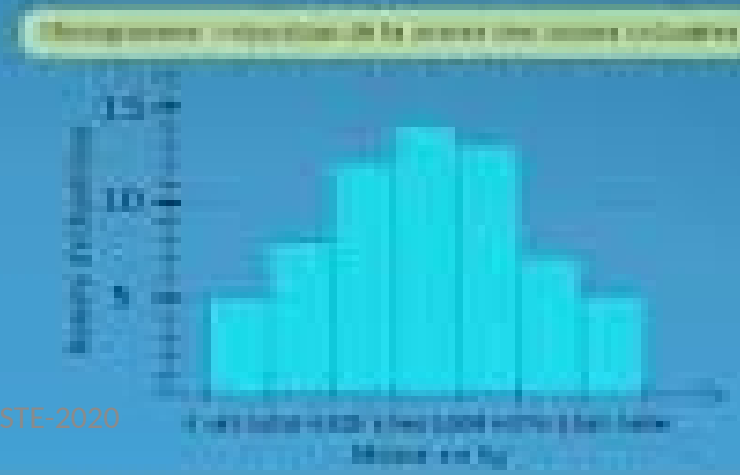
Représentation graphique (suite)



Représentation graphique (suite)

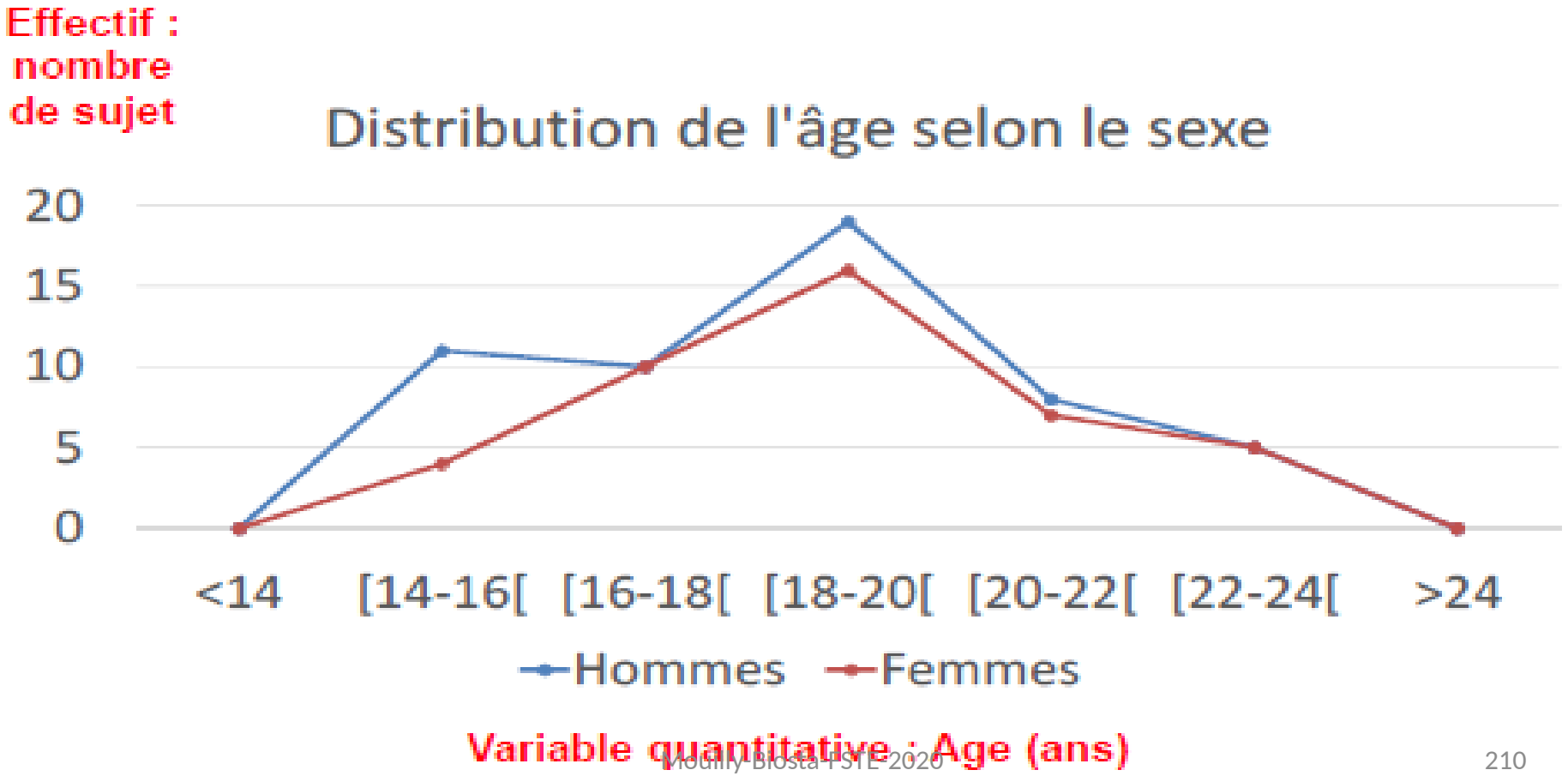
Diagrammes en bâtons …

…et histogrammes

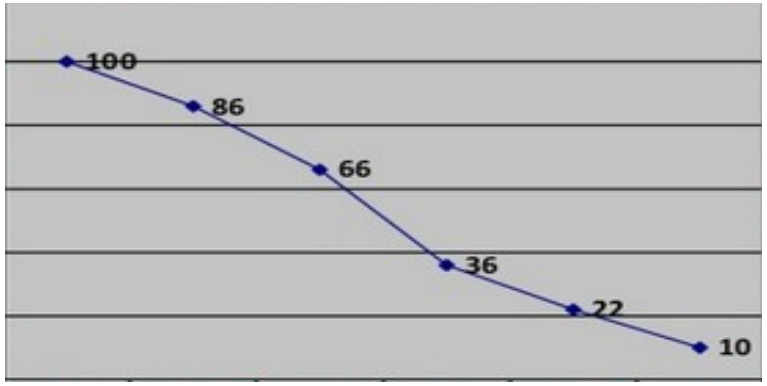


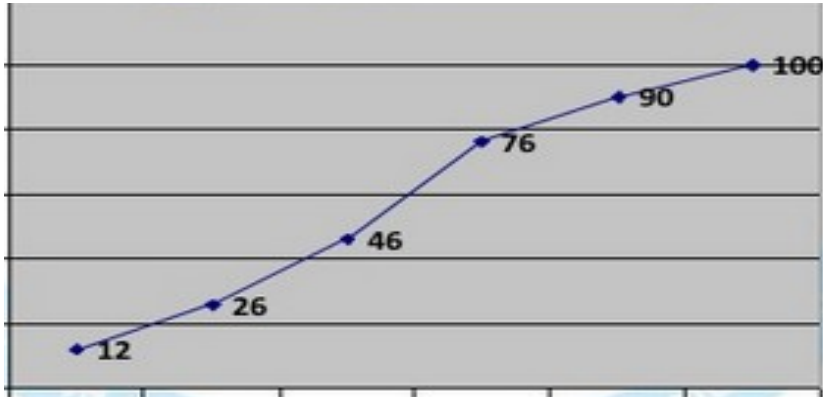
Représentation graphique (suite)

Polygone de fréquence
Exemples



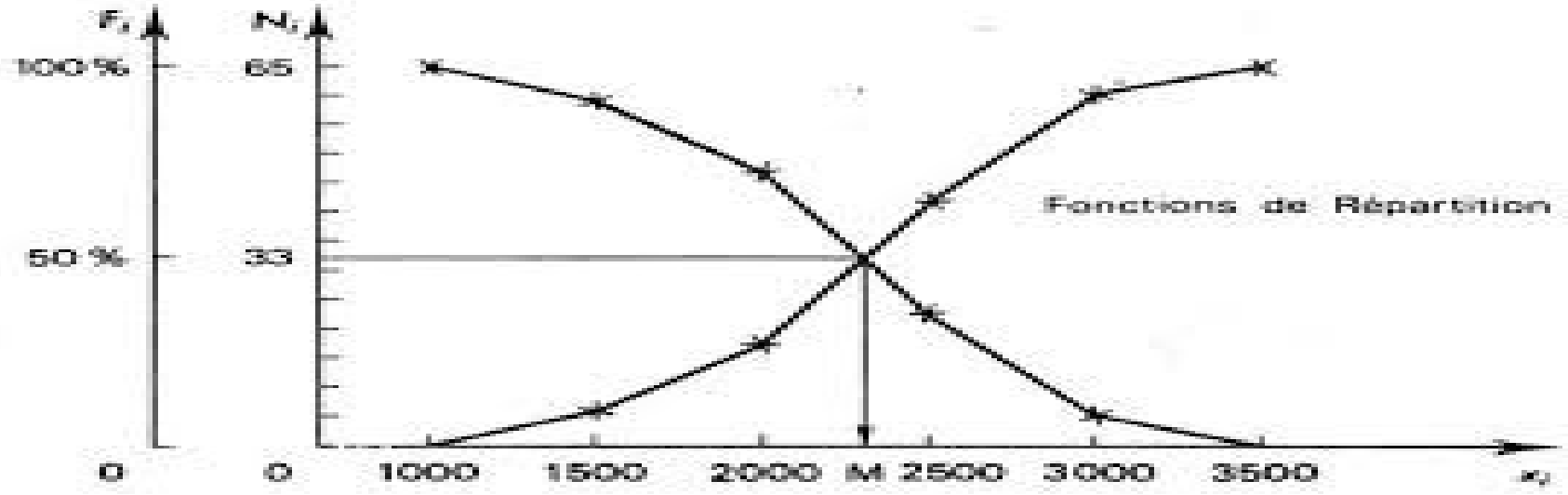
Représentation graphique (suite)





CCD

CCM

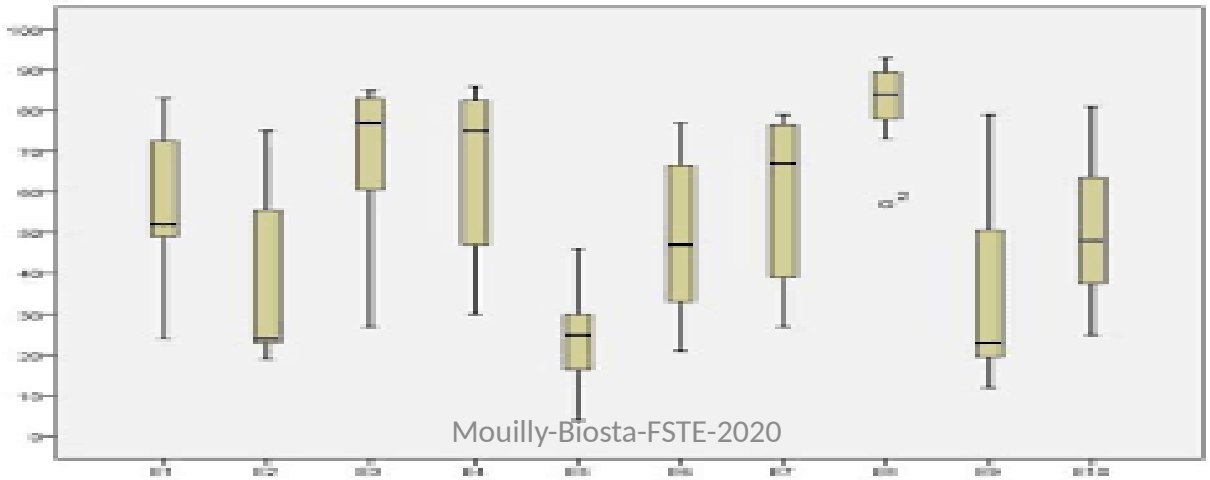


Représentation graphique (suite)

Dispersion

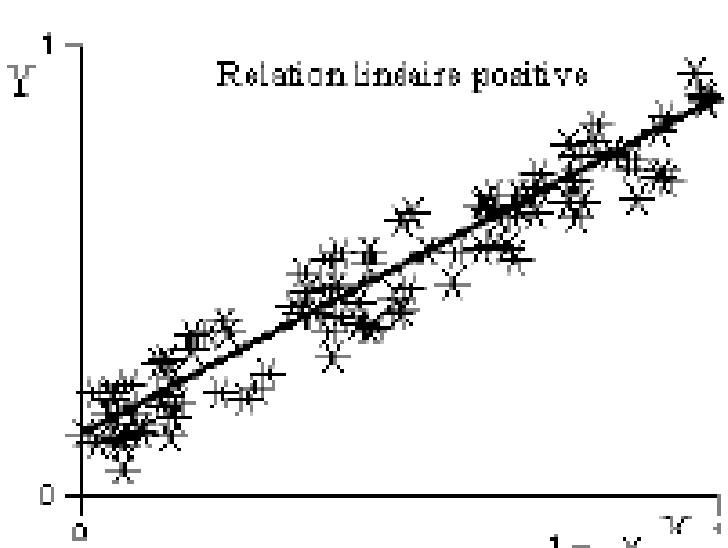
Diagramme à Moustache = Box plots

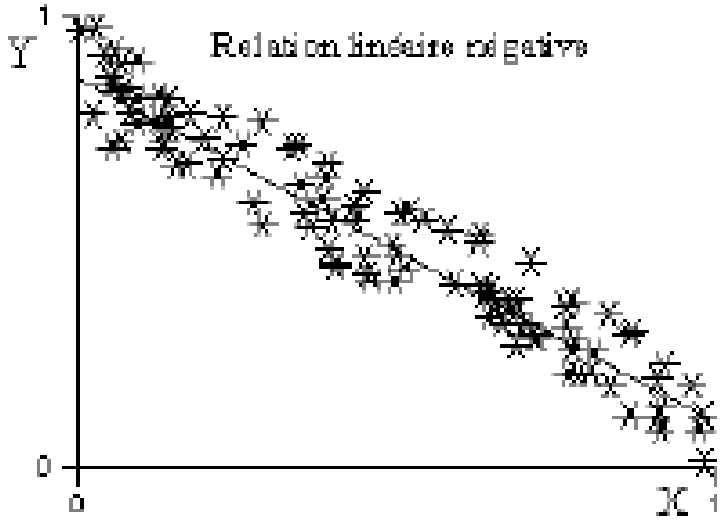
Un diagramme qui met en évidence cinq des paramètres d'une série statistique : le minimum, le premier quartile, la médiane, le troisième quartile et le maximum.



Représentation graphique (suite)

Diagramme de dispersion en nuage de point





Représentation graphique (suite)

Camembert ou (Pie Chart)

- Type de variable :
  - Variable qualitative nominale
- Permet de bien visualiser la part relative de chaque modalité.
- Camembert :
  - Cercle divisé en secteurs
  - Chaque secteur : une classe de la variable
  - La surface du secteur : proportionnelle à la fréquence
  - Nombre de secteurs : moyenne de 6

Exemple : Camembert ou (Pie Chart)

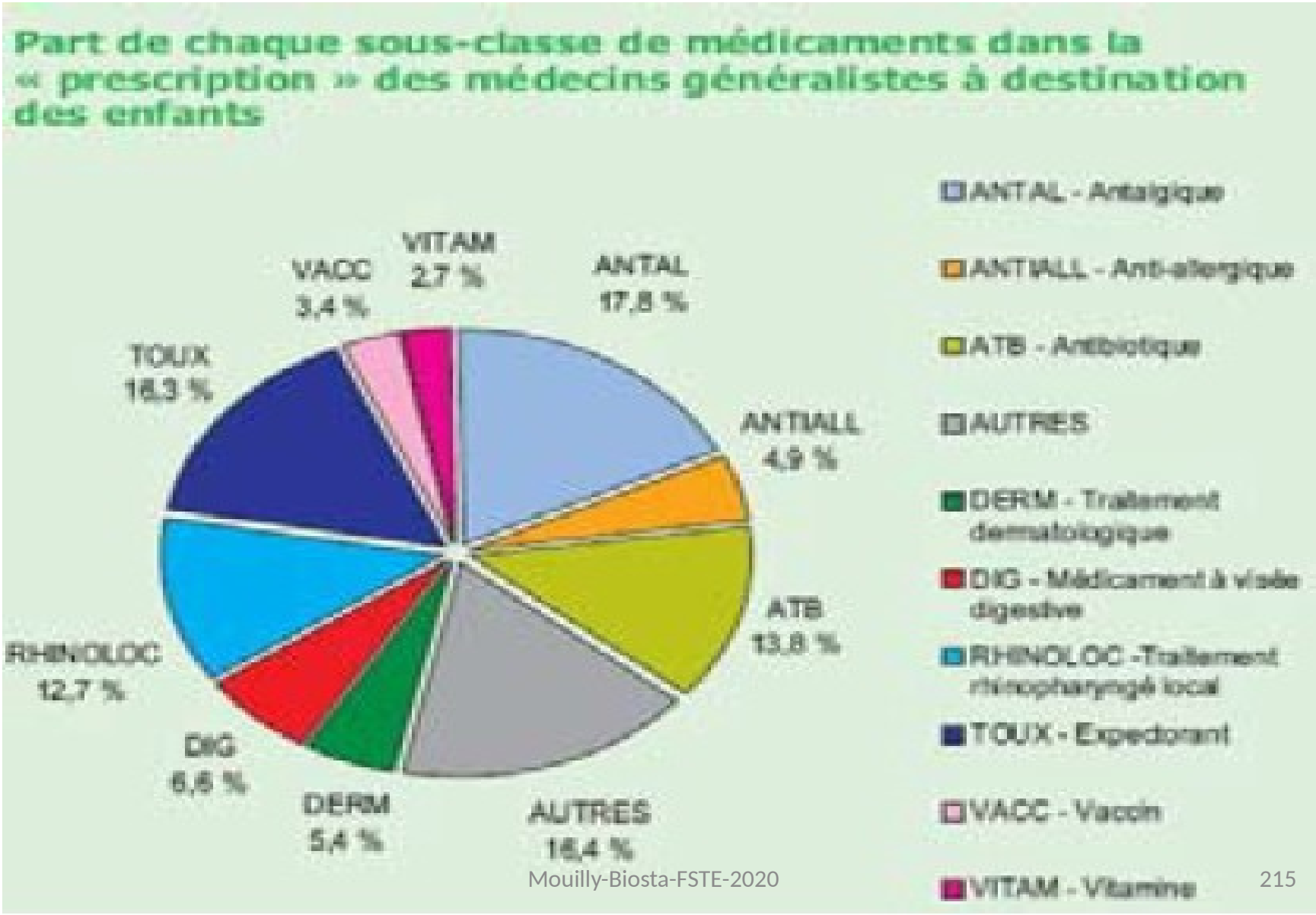


Diagramme en barres horizontales

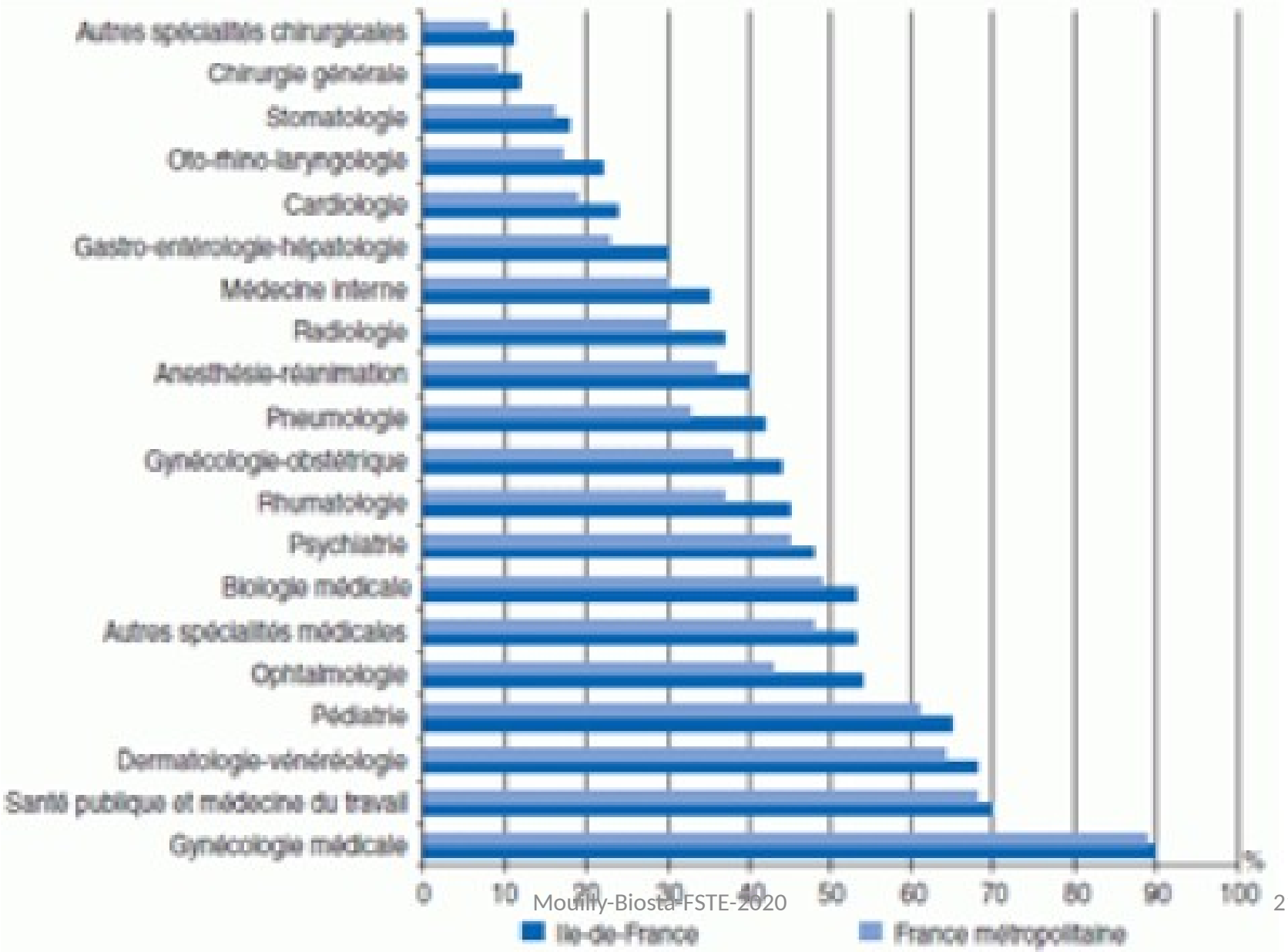
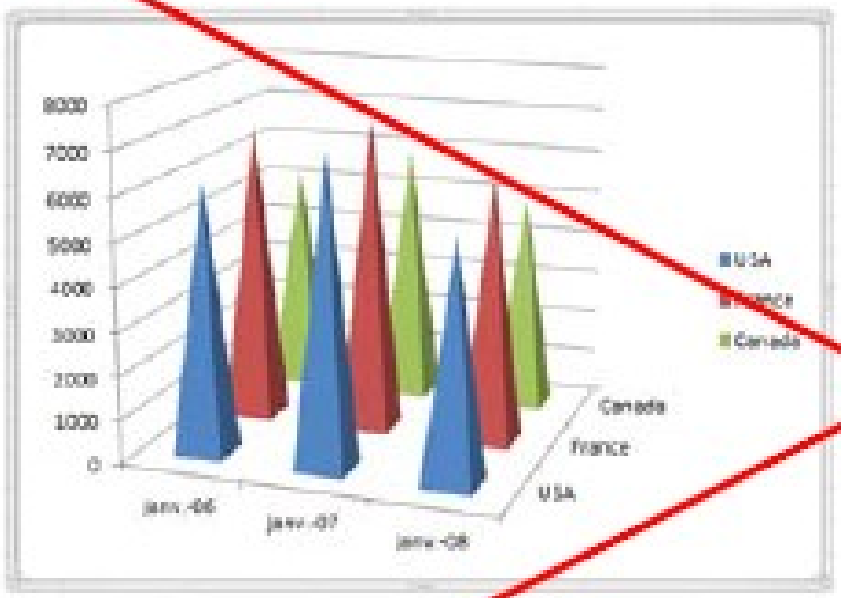
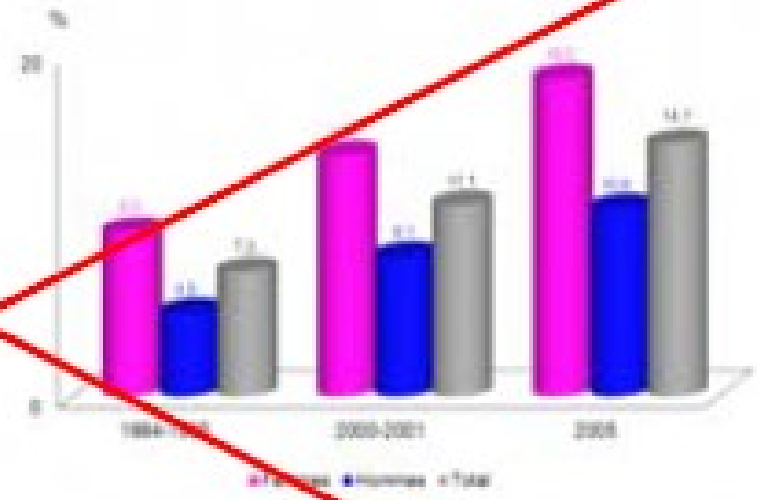


Diagramme en barres





RESUME

VARIABLE QUALITATIVE

Nominale

Effectifs ou Fréquences

Diagramme en barres

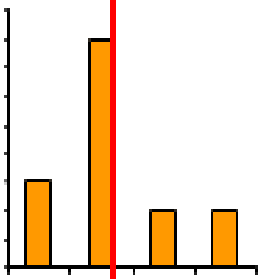
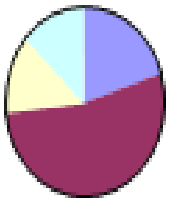


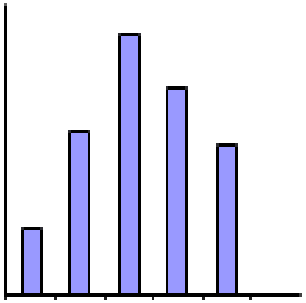
Diagramme circulaire



Ordinale

Effectifs ou Fréquences

Diagramme en barres



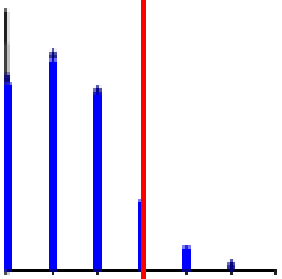
Modalités dans l'ordre

VARIABLE QUANTITATIVE

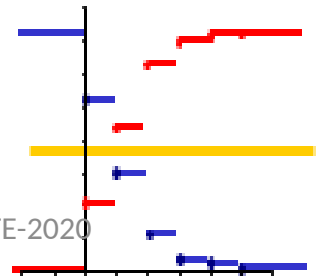
Discrète

Effectifs ou Fréquences

Diagramme en bâtons



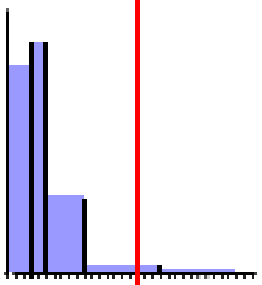
Courbes cumulatives des effectifs ou des fréquences

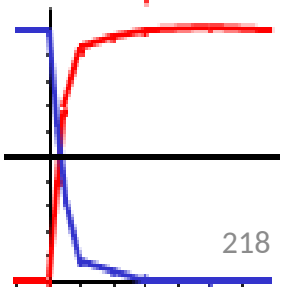


Continue

Effectifs ou Fréquences

Histogramme





Partie II : Statistique Descriptive

- **Chapitre 1 : Statistique univariée (1 dimension)**
  1. Variable Quantitative (Discontinue & Continue)
  2. Variable Qualitative (Nominale & Ordinale)
  3. Représentation Graphique
- **Chapitre 2 : Statistique descriptive bivariée (2 dimensions)**
  1. Introduction
  2. Variables Qualitatives
     a) Tables de contingence
     b) Paramètres de corrélation : V de Cramer, CC, Coefficient Phi,…
  3. Variables quantitatives
     a) Covariance
     b) Paramètres de corrélation : R^2, Coefficient de Spearman,…
     c) Régression
  4. Variables (Qualitative & Quantitative)
  5. Représentation Graphique
- **Chapitre 3 : Statistique multivariée (x dimensions)**

Ne pas confondre !!!

- **Corrélation ⟹ Concomitance, Nature et Intensité de la relation, …**
- **Régression ⟹ Modélisation : Constitution d'un model Probabiliste**
- **Effet (action) ⟹ Test d'H : H_0 ; H_1 : X à un effet significatif sur Y Seuil de 5%, 1%, …**
- **Causalité ⟹ Randomisation (aléatoire): X implique Y (Effet Cause)**

La statistique descriptive bivariée

1. Introduction

- **Etudier la relation qu'il pourrait y avoir entre deux variables distinctes X et Y;**
- **La mesure de ces deux variables sur n éléments (individus, objets, ...) donne lieu à une série statistique bivariée;**
- **Ces analyses diffèrent selon la nature des deux variables, peuvent être :**
  - **Qualitatives;**
  - **Quantitatives**
  - **L'une Quantitative et l'autre Qualitative**

2. Variables qualitatives

a)Table de contingence=Tableau a double entrée= Croisés dynamiques

Distribution marginale

Statut \ Niveau d'études	Prim	Sec	Sup	Total
Marié	2	5	5	12
Célibataire	2	3	3	8
Total	4	8	8	20

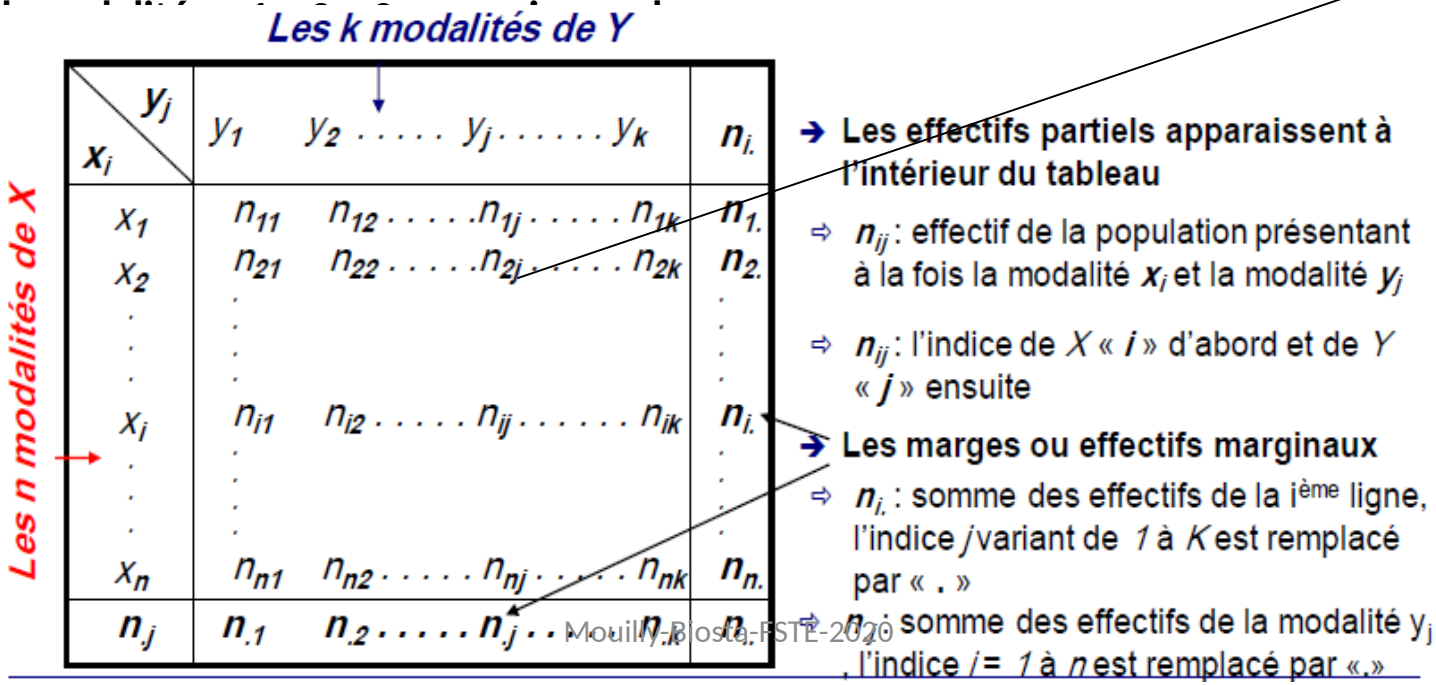
Distribution marginale

Distribution conditionnelle

Considérons n individus décrits simultanément selon deux caractères X et Y.

- **X possède n modalités : x1, x2, x3, ……., xi, …….xn**
- **Y possède**

Distribution conditionnelle



Sujet	Poids	Taille
1	70	170
2	80	180
3	65	165
4	75	175
5	90	182
6	73	170
7	60	162
8	68	165
9	83	180
….	…	…

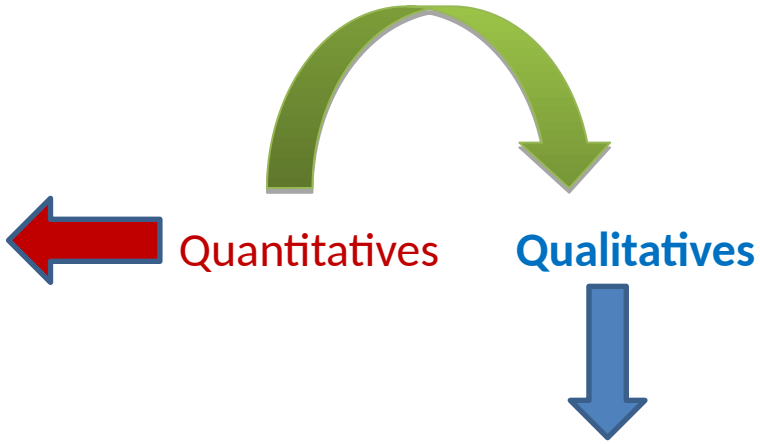


Table de contingence

	Poids									
Taille	60	65	68	70	73	75	80	83	90	Tot.
162	1									1
165		1	1							2
170				1	12					13
175						10				10
180				4		15	1		2	22
182									1	1
Tot.	1	1	1	5	12	25	1	1	3	48

Exemple

Nombre de mesures totale n

Total de chaque ligne = l_i

Total de chaque colonne = c_j

Effectif d'un cas = n_{ij}

Fréquences relatives:

n_{ij} / l_i : % en ligne

n_{ij} / c_j % en colonne

n_{ij} / n %

l_i / n

Table de contingence

	Cheveux			
Yeux	Blonds	Bruns	Autres	Tot. (l_i)
Clairs	**50**	**20**	**30**	100
Foncés	**60**	**80**	**60**	200
Tot. (c_j)	110	100	90	***300***

300 = Nombre total de mesures.

100 = Nombre d'individus ayant les yeux clairs.

110 = Nombre d'individus ayant les cheveux blonds.

50 / 300 = % d'individus ayant les cheveux blonds et les yeux clairs.

50 / 110 = % d'individus parmi les blonds ayant les yeux clairs.

50 / 100 = % d'individus parmi les yeux clairs ayant les cheveux blonds.

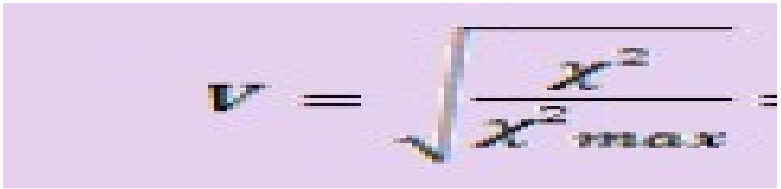
Exemple

Table de contingence à 3 variables

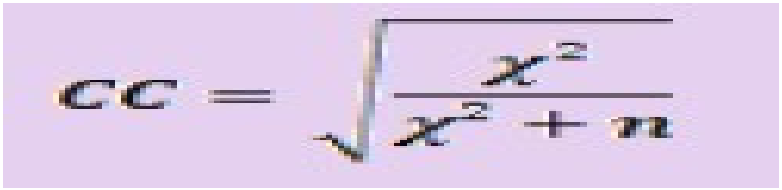
		Cancer du poumon +	Cancer du poumon -	Total
Tabagisme +	Ethylisme +	70	630	700
	Ethylisme -	30	270	300
	Total	100	900	1 000
Tabagisme -	Ethylisme +	3	297	300
	Ethylisme -	7	693	700
	Total	10	990	1 000
Total		110	1 890	2 000

b) Paramètre de corrélation

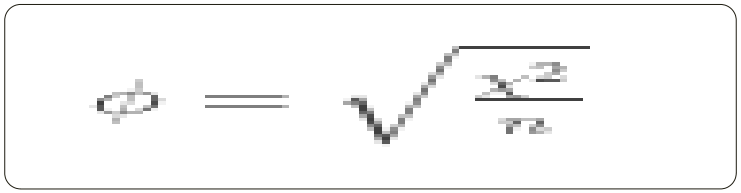
- **V de Cramer**



- **Le coefficient de contingence (CC)**



- **Le coefficient phi (de Pearson)**



- **D'autres paramètres d'association :**

T de Tschuprow,…

3. Variables quantitatives

a) Covariance

La covariance est positive si X et Y ont tendance à varier dans le même sens, et négative si elles ont tendance à varier en sens contraire

On pourra écrire la formule de covariance de la façon suivante :

$$Cov(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad \textbf{Ou} \qquad Cov(x,y) = \frac{1}{N}\sum_{i=1}^{i=p}\sum_{j=1}^{j=q} n_{ij}(x_i - \bar{x})(y_i - \bar{y})$$

$$\mathrm{cov}(x,y) = \frac{1}{n}\sum xiyi - \bar{x}\bar{y}$$

Exemple :

10 étudiants ont passé l'examen partiel et l'examen général et ont obtenu les notes suivantes :

	Partiel (X)	Général (Y)	X Y
	71	83	5893
	49	62	3038
	80	76	6080
	73	77	5621
	93	89	8277
	85	74	6290
	58	48	2784
	82	78	6396
	64	76	4864
	32	51	1632
Total	**687**	**714**	**50875**

$$\text{cov}(x, y) = \frac{1}{n} \sum xiyi - \overline{x}\,\overline{y}$$

Cov (x , y) =

La covariance est positive ou négative selon que le type de liaison entre les deux séries. On dit que les deux variables sont liées et le degré de liaison se mesure par le coefficient de corrélation.

La covariance est nulle ou presque nulle lorsqu'il y a compensation entre les deux séries, c'est à dire si les deux variables sont indépendantes

Exemple :

10 étudiants ont passé l'examen partiel et l'examen général et ont obtenu les notes suivantes :

	Partiel (X)	Général (Y)	X Y
	71	83	5893
	49	62	3038
	80	76	6080
	73	77	5621
	93	89	8277
	85	74	6290
	58	48	2784
	82	78	6396
	64	76	4864
	32	51	1632
Total	**687**	**714**	**50875**

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum xiyi - \overline{x}\,\overline{y}$$

$$\mathrm{Cov}\,(x\,,\,y) \;=\; (50875 / 10) - (687/10) \times (714/10) \;=\; 182.32$$

La covariance est positive ou négative selon que le type de liaison entre les deux séries. On dit que les deux variables sont liées et le degré de liaison se mesure par le coefficient de corrélation.

La covariance est nulle ou presque nulle lorsqu'il y a compensation entre les deux séries, c'est à dire si les deux variables sont indépendantes

3. Variables quantitatives

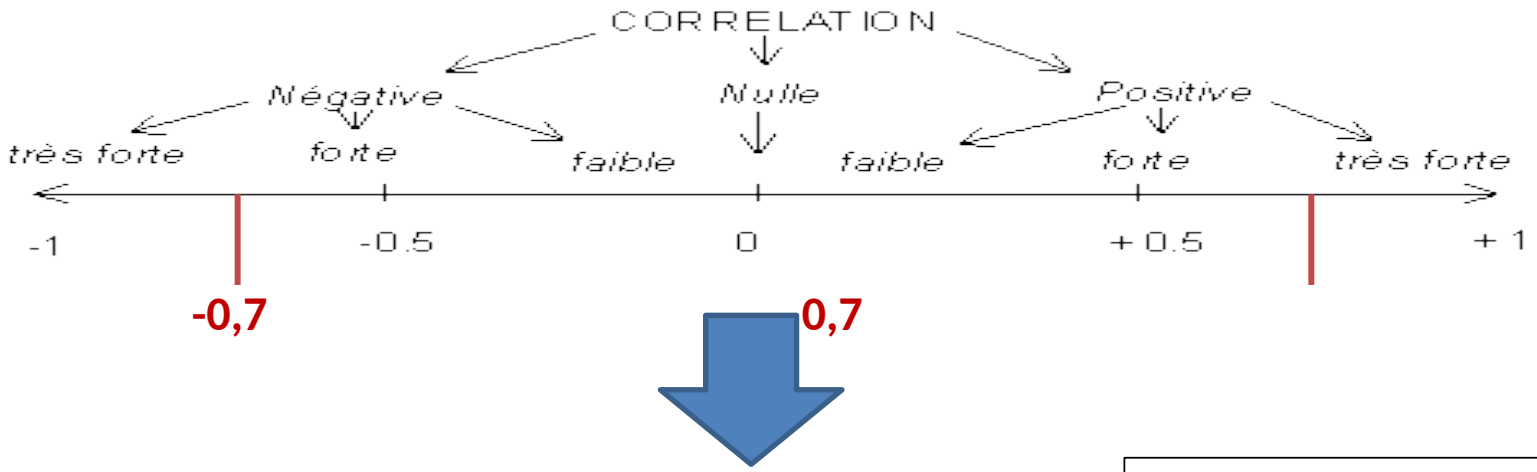
b) Paramètre de corrélation

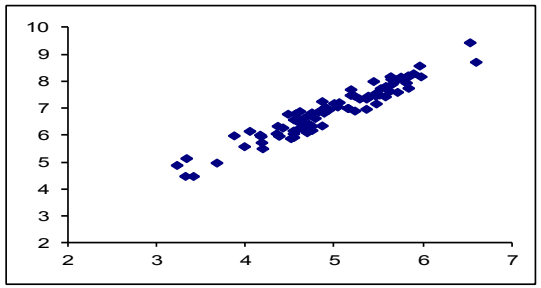
b.1. Coefficient de Corrélation linéaire Bravais-Pearson

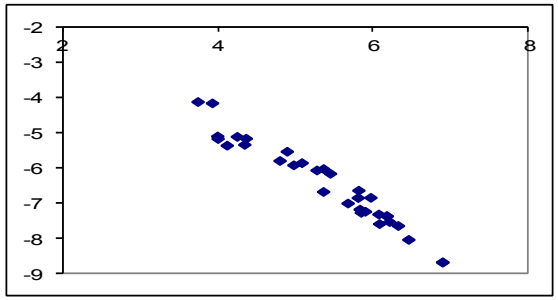
$$r = \frac{\text{covariance}(X,Y)}{\sqrt{\text{var}(X) * \text{var}(Y)}}$$

Bravais-Pearson

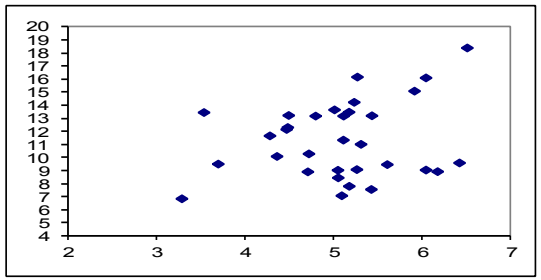
- **Le coefficient de corrélation permet de mesurer l'association linéaire entre X et Y**
- **Il reste identique si on change d'unité ou d'origine**
- **r varie de -1 à + 1 :**
  - **Corrélation positive (0 < r < 1) : relation proportionnelle**
  - **Corrélation négative (-1 < r < 0) : relation inversement proportionnelle**
- **si r = 0 pas de liaison ⊔ 2 variables indépendantes**
- **si r = 1 (ou -1) ⊔ 2 variables liées**

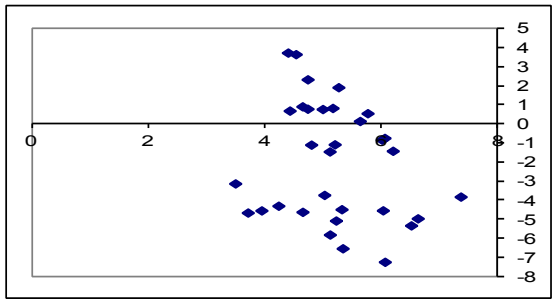




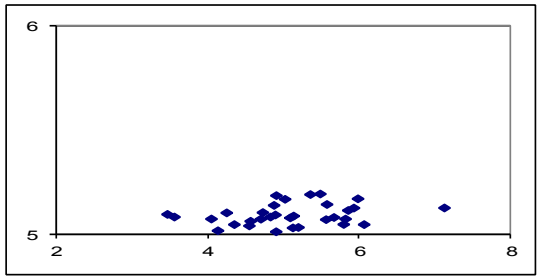


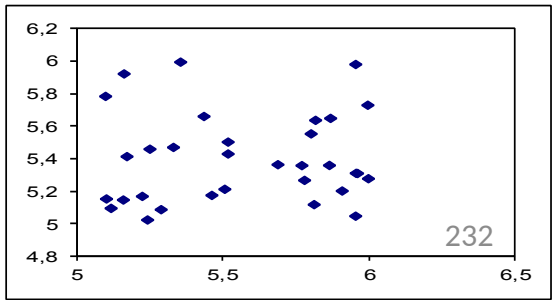
Forte corrélation





Faible corrélation





Pas de corrélation r voisin de 0

Interpréter le coefficient de Bravais-Pearson

comment ?

le placer au carré

est alors interprété

comme la proportion de la variance de Y qui est attribuable à la variance de X.

qui se résume

En résumé:

- r^2 = proportion de *liaison* entre X et Y
- $1 - r^2$ = proportion d'*aliénation* (absence de liaison entre les deux variables)

on en conclut

En conclusion:

des coefficients inférieurs à .7 sont considérés comme « peu concluants ».

exemples

r =.914

= corrélation positive et assez forte

r^2 = .84

$1 - r^2$ = .16

→ la proportion de liaison de 84% et la proportion d'aliénation est de 16%.

r =.232

= corrélation positive mais assez faible

r^2 = .05

$1 - r^2$ = .95

→ la proportion de liaison de 5% et la proportion d'aliénation est de 95%.

r =.510

= corrélation positive

r^2 = .26

$1 - r^2$ = .74

→ la proportion de liaison de 26% et la proportion d'aliénation est de 74%.

r =.725

= corrélation positive

r^2 = .53

$1 - r^2$ = .47

→ la proportion de liaison de 53% et la proportion d'aliénation est de 47%.

→ cela devient intéressant !

b.2. Coefficient de corrélation de rang de Spearman = Coefficient de Spearman.

Limites du Coefficient de Bravais-Pearson

- En principe, le coefficient de Bravais-Pearson (Pearson) n'est applicable que pour mesurer la relation entre **deux variables X et Y ayant une distribution de type gaussien** (Approche paramétriques) et ne comportant pas de valeur exceptionnelles.

- Si ces conditions ne sont pas vérifiées, l'emploi de ce coefficient peut aboutir à des **conclusions erronées** sur la présence ou l'absence d'une relation.

NB. **On notera également que l'absence d'une relation linéaire ne signifie pas l'absence de toute relation entre les deux caractères étudiés.**

b.2. Coefficient de corrélation de rang de Spearman = Coefficient de Spearman

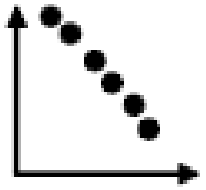
- Une approche **non paramétrique**, caractérisé par ce coefficient qui examine s'il existe une relation entre le **rang** des observations pour X et Y
- Détecter l'existence de relations monotones (croissante ou décroissante), quelle que soit leur forme précise (linéaire, exponentiel, puissance, ...)
- **Ce coefficient est très utile lorsque l'analyse du nuage de point révèle une forme curviligne dans une relation qui semble mal s'ajuster à une droite.**
- On notera également qu'il est préférable au coefficient de Pearson lorsque les distributions X et Y sont dissymétriques et/ou comportent des valeurs exceptionnelles.

Notation : ρ (rho) ou r_s

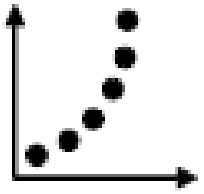
NB. Dans la formule, On n'utilise alors pas les **VALEURS** des observations dans les calculs mais leur **RANG**.

Ce coefficient varie entre -1 et +1. Son interprétation est la même que celui de Pearson, mais il permet de mettre en évidence des relations non-linéaires lorsqu'elles sont positives ou négatives.

A
corrélation positive

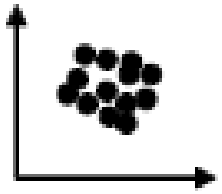


B
corrélation négative



C
corrélation positive

D
pas de corrélation, mais dépendance



E
indépendance

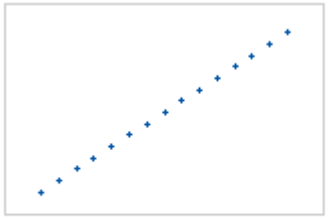
liaison :

monotone linéaire croissante

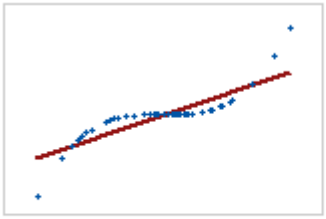
monotone linéaire décroissante

monotone non linéaire croissante

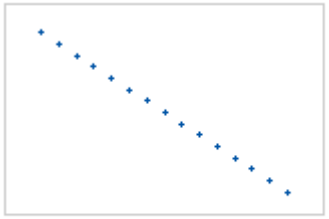
non monotone



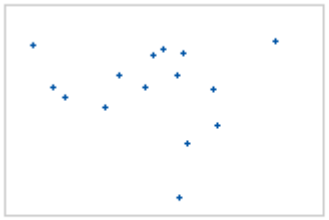
Pearson = +1, Spearman = +1



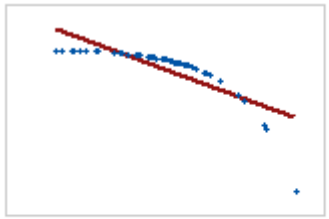
Pearson = +0,851, Spearman = +1



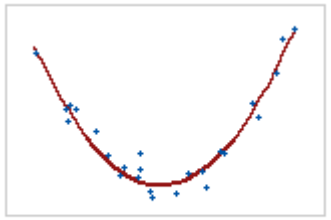
Pearson = −1, Spearman = −1



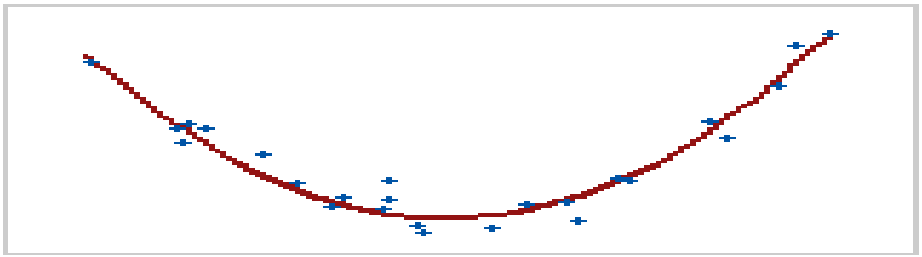
Pearson = −0,093, Spearman = −0,09



Pearson = −0,799, Spearman = −1



Coefficient de 0



Coefficient de 0

Le cas d'autres relations non linéaires

- Les coefficients de corrélation de Pearson ne mesurent que des relations linéaires
- Les coefficients de corrélation de Spearman ne mesurent que des relations monotones.

→Par conséquent, une relation significative peut exister même si les coefficients de corrélation sont de 0

→Nécessité d'examinez un nuage de points afin de déterminer la forme de la relation.

→**Il existe d'autres paramètres de corrélation des rangs : exemple le Coefficient de de Kendall, noté « tau » (τ)**

NB. Le calcul d'un coefficient de corrélation ne constitue qu'une première étape dans l'analyse de la relation entre deux caractères. Il s'agit tout au plus d'une étape exploratoire qui doit être validée par un test de la significativité de la relation et par une vérification de la validité de la relation (absence de biais).

4. Variables quantitatives

c) Régression

RAPPEL

L'analyse de corrélation ⇨ déterminer l'intensité : fort, faible...

L'analyse de régression ⇨ trouver une relation de y en fonction de x.

X et Y sont deux variables aléatoires

- **X c'est la variable explicative (indépendante) et Y c'est la variable expliquée (dépendante = réponse)**
- **X en fonction de Y a un sens (poids/taille) ⇨ taille est explicable par poids**

≠

Y en fonction de X a un autre sens (taille/poids) ⇨ poids est explicable par taille

régression linaire

y= ax + b + Erreur

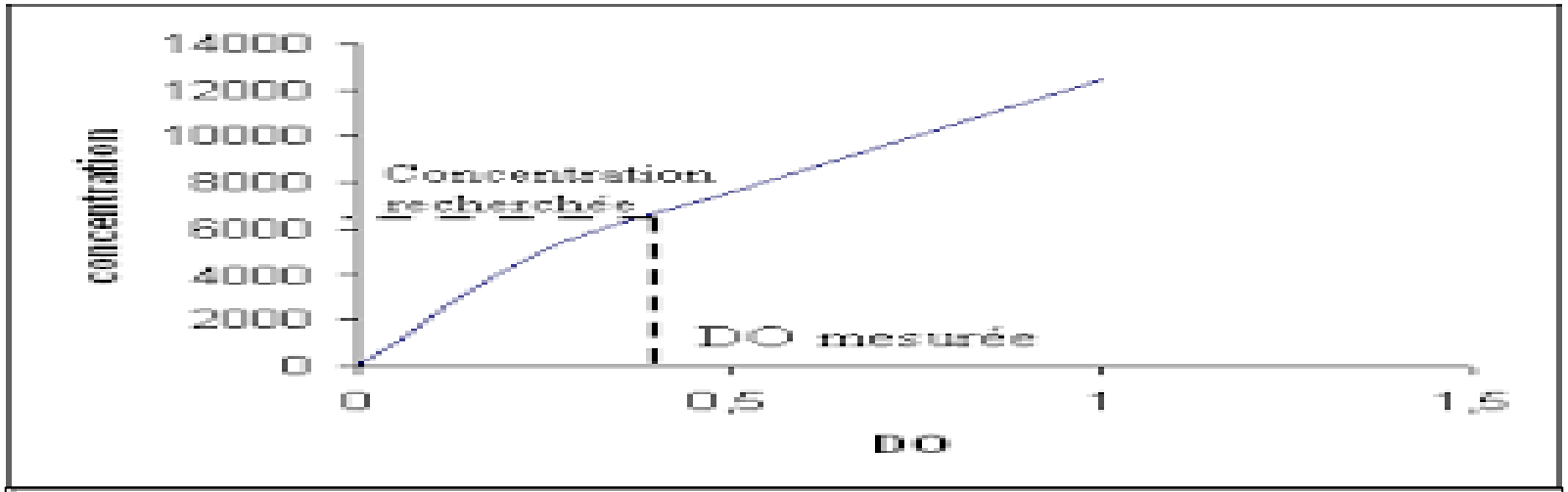
Exemples

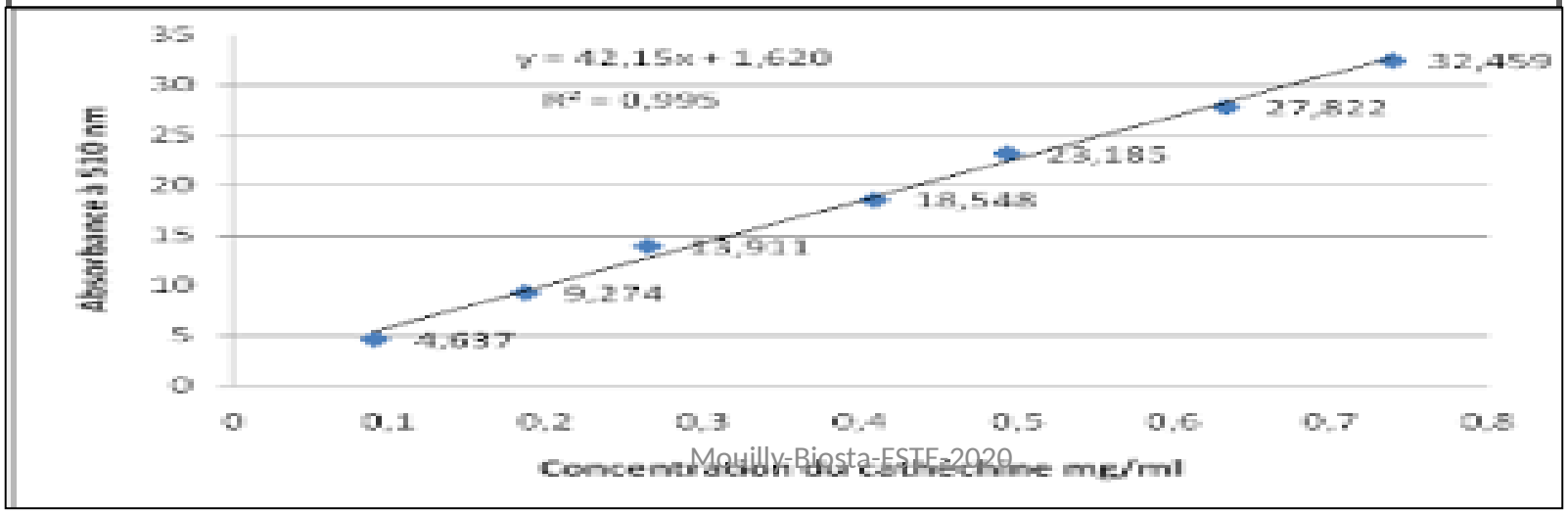
- **Consommation en eau et une population**

  ***X* = nombre d'habitants; *Y* = eau consommée**

- **Nombre d'heures passées à réviser un examen et la note obtenue**

  ***X* = heures passées à réviser ; Y = note obtenue**





4. Variables (Qualitative et Quantitative)

Pour mesurer l'association entre ces 2 types variables, on doit "orienter" d'abord la relation en décidant que l'une des variables joue le rôle de :

- ✓ **la (les) variable (s) explicative (s) (X=indépendante(s)=Covariable(s)=Exogène(s))**
- ✓ **la variable** expliquée = la variable à prédire (Y=variable dépendante=réponse=Endogène)

- Si on choisit **la variable qualitative** comme **variable explicative,** on peux faire de l'analyse de variance :
  - **ANOVA**
    - ✓ 1 facteur (**ANOVA-Univarié**) : exemple sexe et taille (quantitative)
    - ✓ 2 facteurs (**ANOVA-Bivariée**) : exemple sexe & mode de vie et taille (quantitative)
- Si à l'inverse on choisit **la variable quantitative** comme **variable explicative (=régresseur)**
  - **Régression**
    - ✓ **Linéaire** : Ajustement nécessite une fonction affine
      - **Simple : 1 régresseur** ⊐ **2 cas :**
        - ○Gouaussien ($Y = \beta_0 + \beta_1 X + \varepsilon$) ⊐ Distribution de Y doit être normale
        - ○Généralisé (Poisson, Bernoulli)
      - **Multiple** : **2 ou plusieurs régresseurs X** ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$)
    - ✓ **Polynomiale :** Ajustement nécessite une fonction polynômiale
      - **Simple : 1 régresseur** ($Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p + \varepsilon$)
      - **Multiple** : **2 ou plusieurs régresseurs X…**
      - **NB.** Dans le cadre de **la Régression logistique** ⊐ la variable Y est qualitative avec :
        - ✓ 2 modalités possibles {0, 1} ⊐ **binaire**
        - ✓ Plusieurs modalités possibles {0, 1, 2,…} ⊐ **multinomiale**

Important : **Les variables X sont exclusivement continues ou binaires.**

Régression linéaire simple

- Historiquement, le modèle linéaire a été développé par Fisher, avec applications en génétique et en agronomie

  Dans ce cours, on va traiter que le cas de la RL Normale

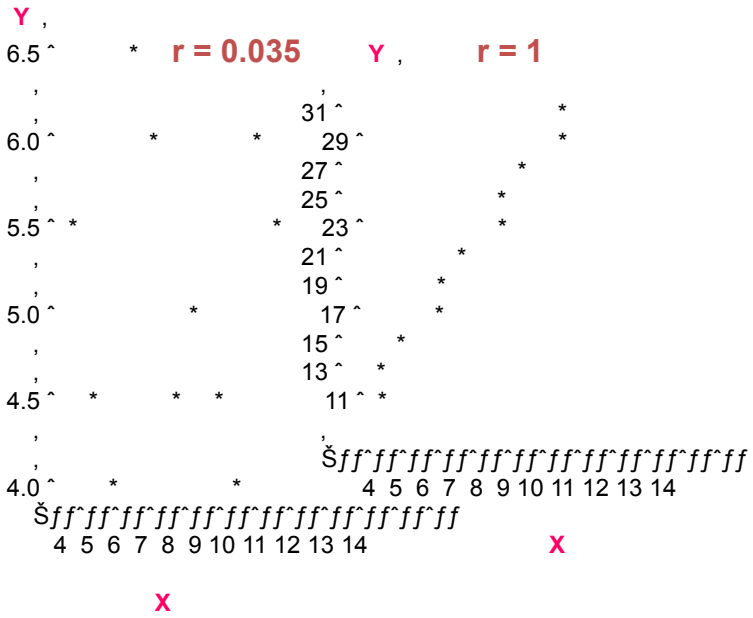
- décrire une relation linéaire entre deux variables quantitatives ou encore pour pouvoir prédire **Y** pour une valeur donnée de **X**, nous utilisons une droite de régression $Y = \beta_0 + \beta_1 X + \varepsilon$
- Puisque tout modèle statistique n'est qu'une approximation il y a toujours une erreur, notée ε dans le modèle, car le lien linéaire n'est jamais parfait.
- S 'il y avait une relation linéaire parfaite entre **Y** et **X**, le terme d'erreur serait toujours égale à 0, et toute la variabilité de **Y** serait expliquée par la variable indépendante **X**.

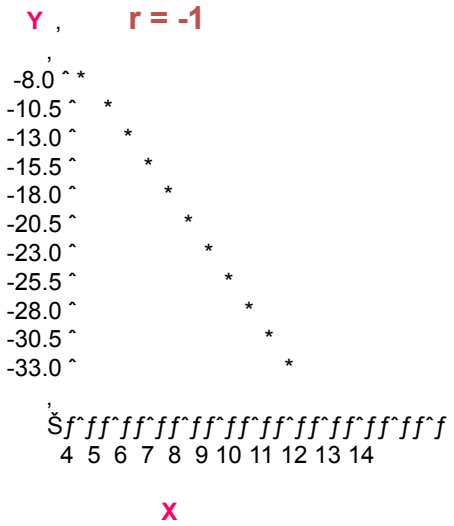
Conditions

- ✓ Variables aléatoires indépendantes
- ✓ Y suit une loi Normale
- ✓ Homoscédasticité : σ^2 doit être constante

Le coefficient de corrélation ***r*** de Pearson sert à mesurer l'intensité de la relation **linéaire** entre deux variables quantitatives.

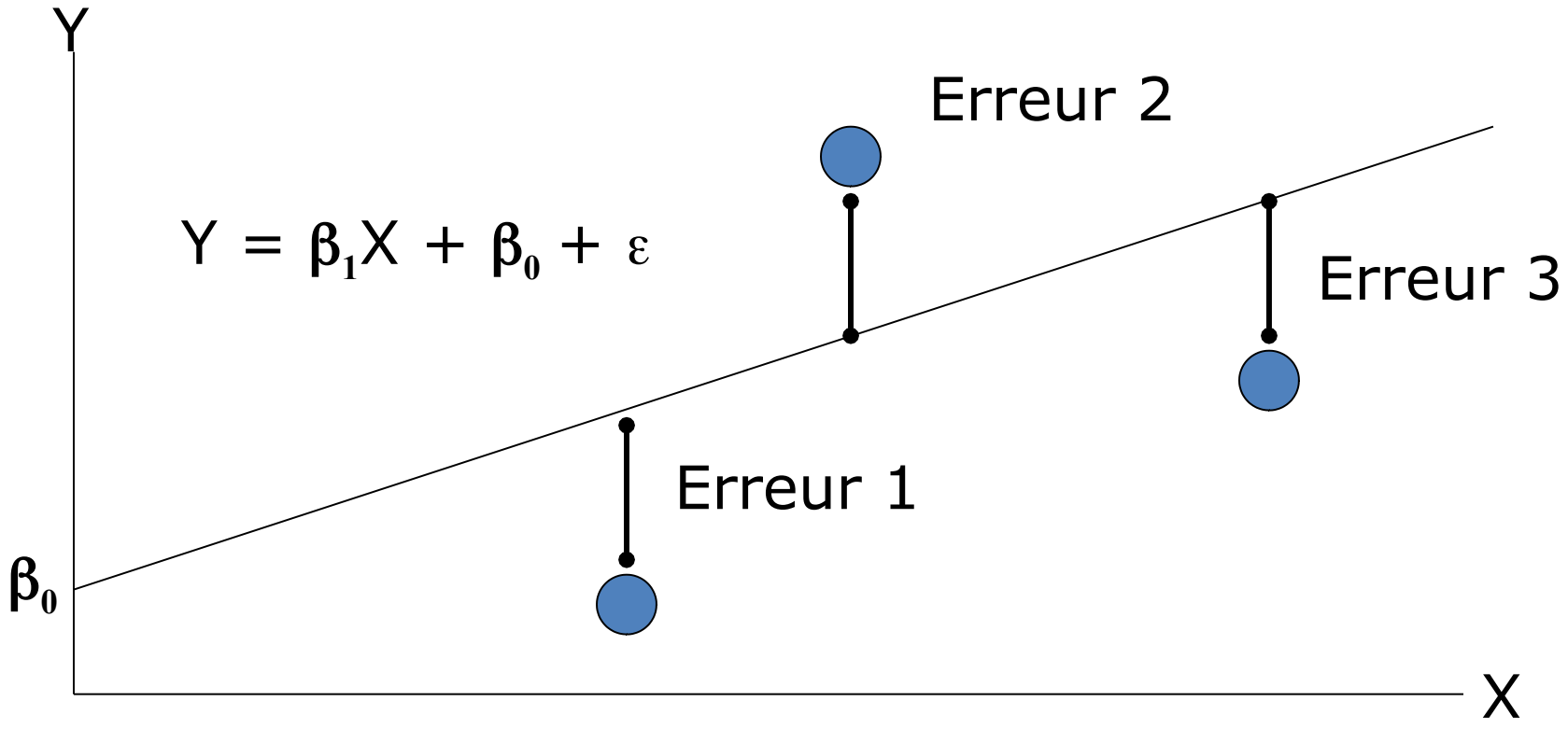
- Le coefficient de corrélation ***r*** prendra des valeurs entre -1 et 1.
- S 'il existe une relation **linéaire parfaite** entre **X** et **Y** alors ***r*** = ± 1 (***r*** =1 si **X** et **Y** varient dans le même sens et ***r*** = -1 si **X** varie dans le sens opposé à **Y**).
- Si ***r*** = 0, ceci indique qu 'il n 'y a pas de lien linéaire entre **X** et **Y**.
- Plus la valeur de ***r*** s 'éloigne de 0 pour s 'approcher de ± 1 plus l 'intensité du lien linéaire entre **X** et **Y** grandit.


Y
6.5
6.0
5.5
5.0
4.5
4.0
r = 0.035
4 5 6 7 8 9 10 11 12 13 14
X
Y
r = 1
31
29
27
25
23
21
19
17
15
13
11
4 5 6 7 8 9 10 11 12 13 14
X


Y
r = -1
-8.0
-10.5
-13.0
-15.5
-18.0
-20.5
-23.0
-25.5
-28.0
-30.5
-33.0
4 5 6 7 8 9 10 11 12 13 14
X

Méthode des moindres carrés

NB. On peut aussi estimer le modèle par maximum de vraisemblance ou encore par inférence bayésienne.



Pente ⇨ $\beta_0 = \dfrac{Cov(X,Y)}{\sigma^2(X)}$ $\quad r = \dfrac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$

- Donc, pour une valeur donnée de **X**, nous aimerions estimer **Y**.
- Ainsi, à l'aide des données de l'échantillon nous estimerons les paramètres β_0 et β_1 du modèle de régression de façon à minimiser la somme des carrés des erreurs.
- Le coefficient de corrélation au carré est appelé **coefficient de détermination et nous indique le pourcentage de la variabilité de Y expliquée par X:**

$$R^2 = 1 - (n-2)/(n-1)\{S_e / S_y\}^2,$$

où S_e est l'écart type des erreurs et S_y est l'écart type de Y.

On peut également utiliser le coefficient de détermination ajusté pour nous indiquer le pourcentage de la variabilité de **Y** expliquée par **X:**

$$R^2_{\text{ajusté}} = 1 - \{S_e/S_y\}^2 .$$

Parmi les 3 modèles précédents, lequel choisiriez vous et pourquoi?

- **Modèle 1**:
  - **Y= 16209 + 102*(X).**
  - **R^2 = 58,8%. Donc 58,8% de la variabilité de Y est expliquée par le X .**
- **Modèle 2**:
  - **Y= -347 + 22021*(X).**
  - **R^2 = 39,3%. Donc 39,3% de la variabilité de Y est expliquée par le X.**
- **Modèle 3**:
  - **Y= 32428 + 38829*(X).**
  - **R^2 = 33,9%. Donc 33,9% de la variabilité de Y est expliquée par le X.**

Régression linéaire multiple

- Il est fort possible que la variabilité de la variable dépendante **Y** soit expliquée non pas par une seule variable indépendante X mais plutôt par une combinaison linéaire de plusieurs variables indépendantes $X_1, X_2, \ldots, X_p$.
- Dans ce cas le modèle de régression multiple est donné par:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- Aussi, à l'aide des données de l'échantillon nous estimerons les paramètres $\beta_0, \beta_1, \ldots, \beta_p$ du modèle de régression de façon à minimiser la somme des carrés des erreurs.

- Le coefficient de corrélation multiple $\mathbf{R^2}$, aussi appelé **coefficient de détermination**, nous indique le **pourcentage** de la variabilité de **Y** expliquée par les variables indépendantes $\mathbf{X_1, X_2, \ldots, X_p}$.

- Lorsqu'on **ajoute** une ou plusieurs variables indépendantes dans le modèle, le coefficient $\mathbf{R^2}$ **augmente**.

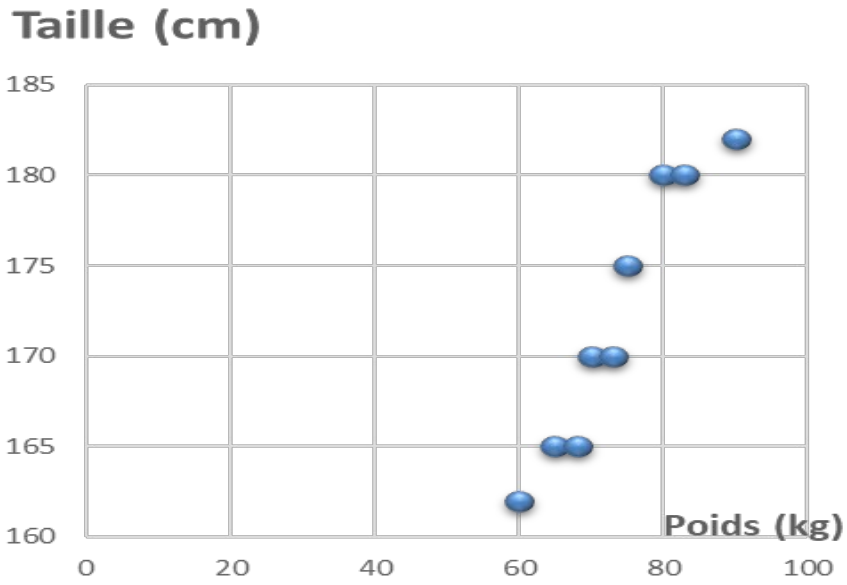
- La question est de savoir si le coefficient $\mathbf{R^2}$ augmente de façon significative.

- Notons qu'on ne peut avoir plus de variables indépendantes dans le modèle qu'il y a d 'observations dans l'échantillon (règle générale: $n \geq 5p$).

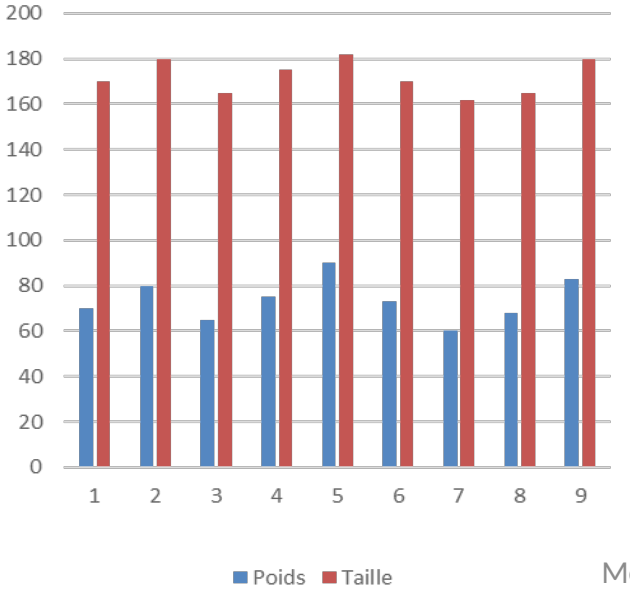
5. Représentation Graphique

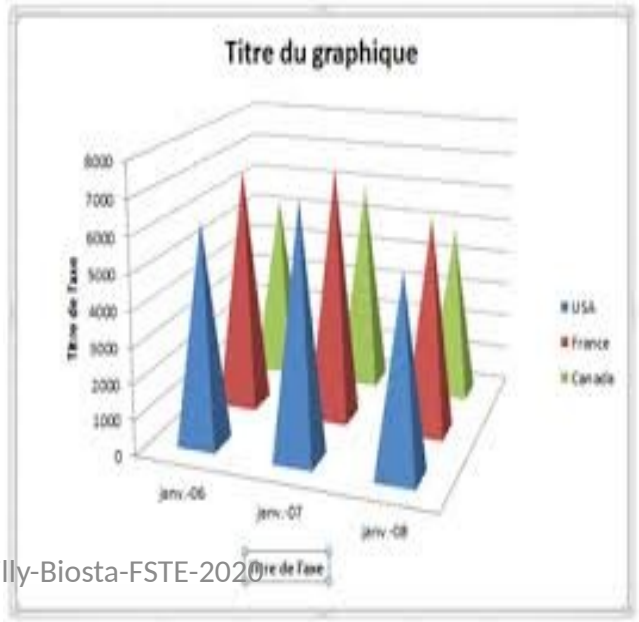
Nuage de points

Les points sont obtenus en représentant chaque couple d'observation (xi; yi) par un point dans le plan (discrète ou continue)
= diagramme de dispersion

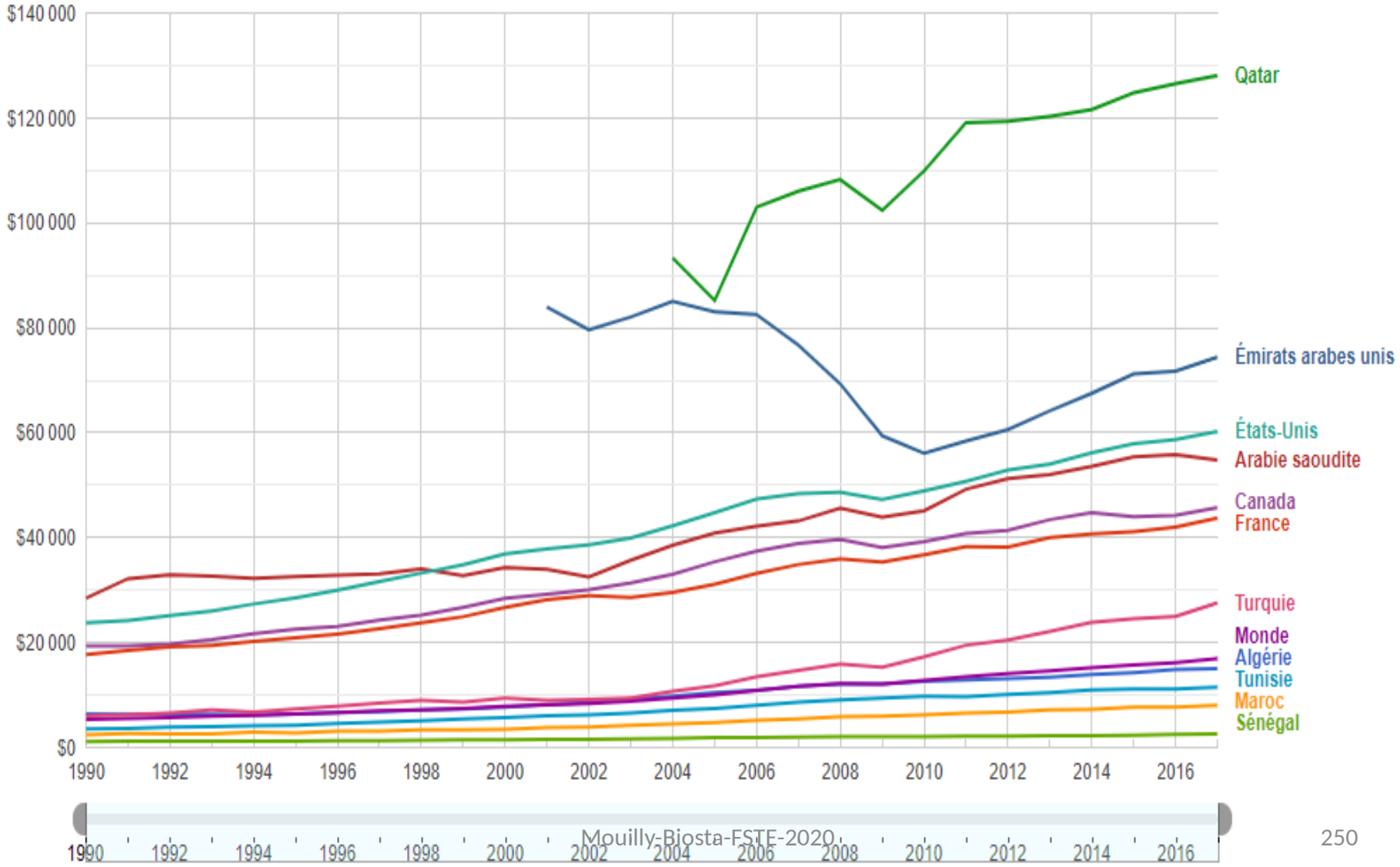


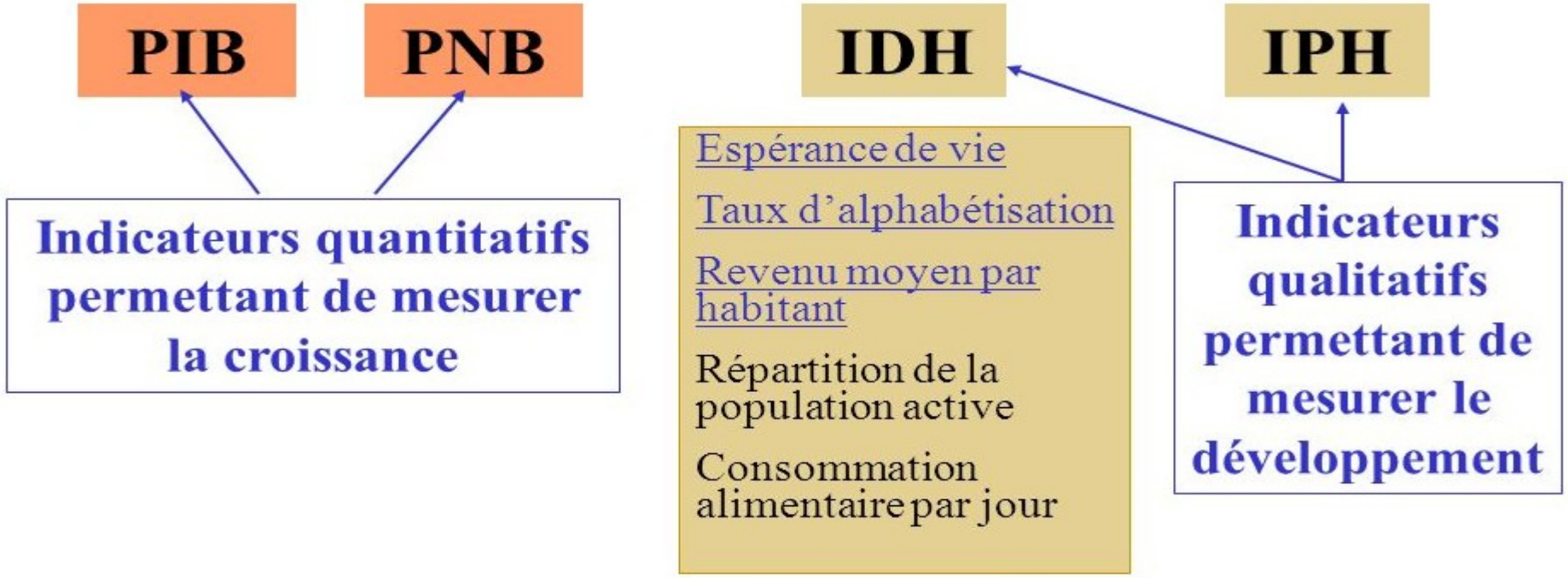
Histogramme





RNB par habitant en dollars PPA





- **PIB** = Produit Intérieur Brut par habitant. Il s'agit d'un système permettant de mesurer l'activité économique d'un pays en se basant sur le revenu moyen de ses citoyens

=Travail (W) + Capital (K) Somme des valeurs ajoutées

Limites : Les travaux domestiques, L'économie souterraine,…

- **PNB (RNB)** = Produit National Brut, Il s'agit d'un système permettant de mesurer la richesse produite au cours d'une année par l'ensemble des résidents et des ressortissants d'un pays, C'**est** la nationalité (**quelle** que soit la résidence).

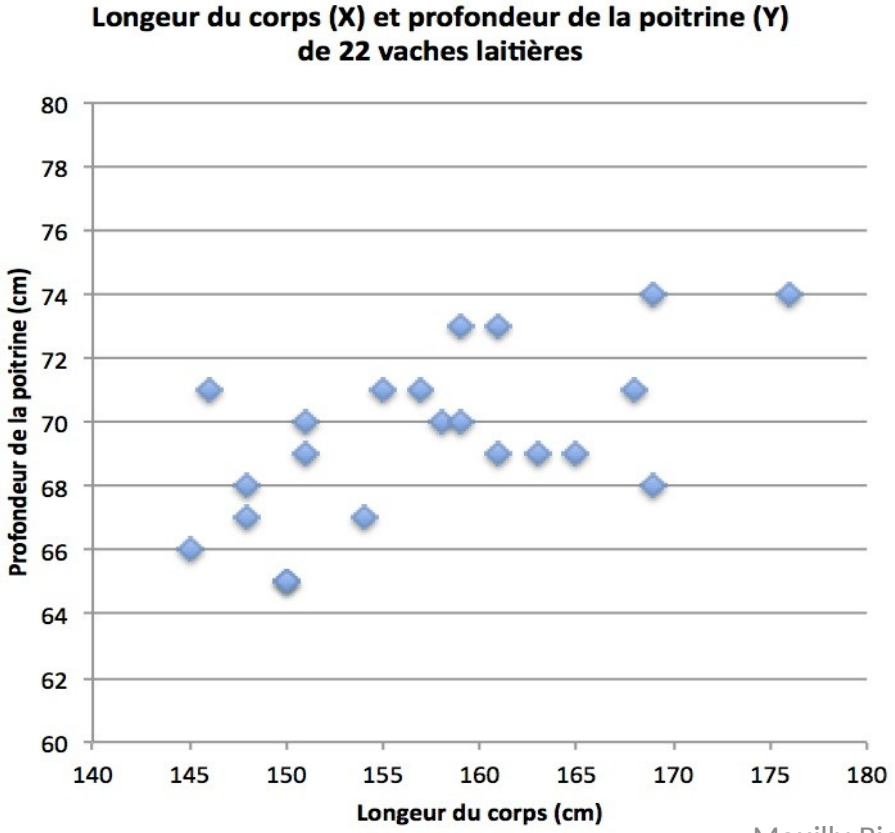
La **différence entre le PIB** et le **PNB est** le critère de délimitation géographique

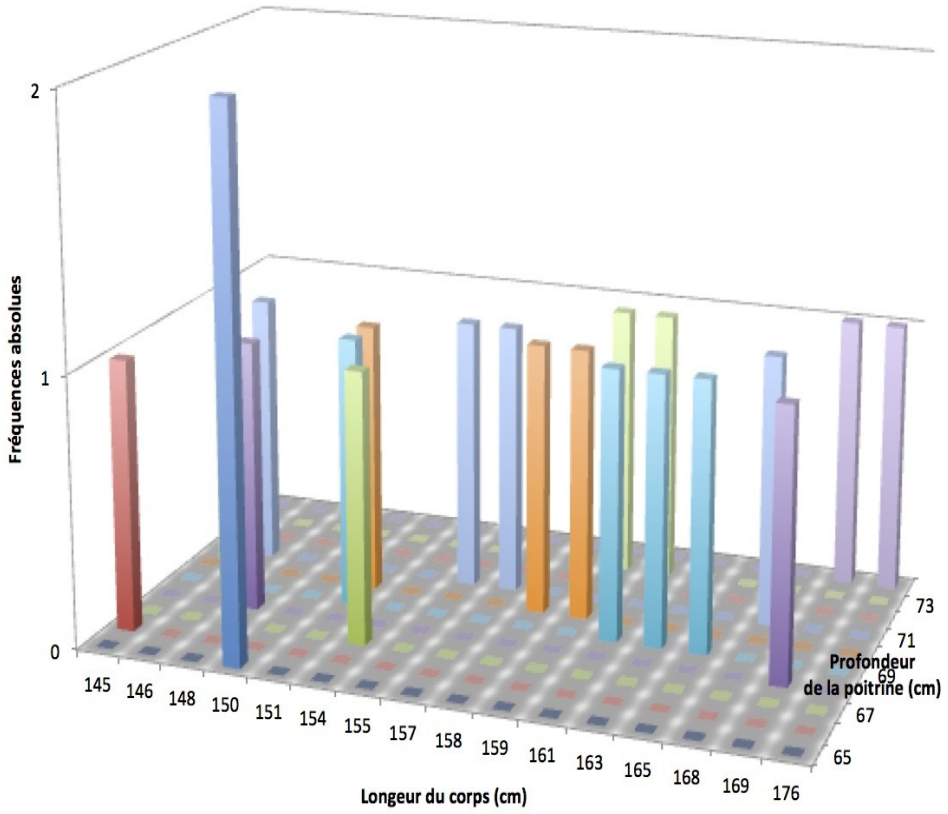
- **PPA** = Parité de Pouvoir Achat : C'est un taux de conversion monétaire qui permet d'exprimer dans une unité commune les pouvoirs d'achat des différentes monnaies.

Représentation spatiale 3D dans R^3

Il est composé de parallélépipèdes rectangles (ou prismes carrés) juxtaposés, dont chacune des bases correspond à une cellule du tableau à double entrée et dont la hauteurs est égale ou proportionnelle à la fréquence absolue ou relative de cette cellule

X	145	146	148	148	150	150	151	151	154	155	157	158	159	159	161	161	163	165	168	169	169	176
Y	66	71	67	68	65	65	69	70	67	71	71	70	70	73	69	73	69	69	71	68	74	74

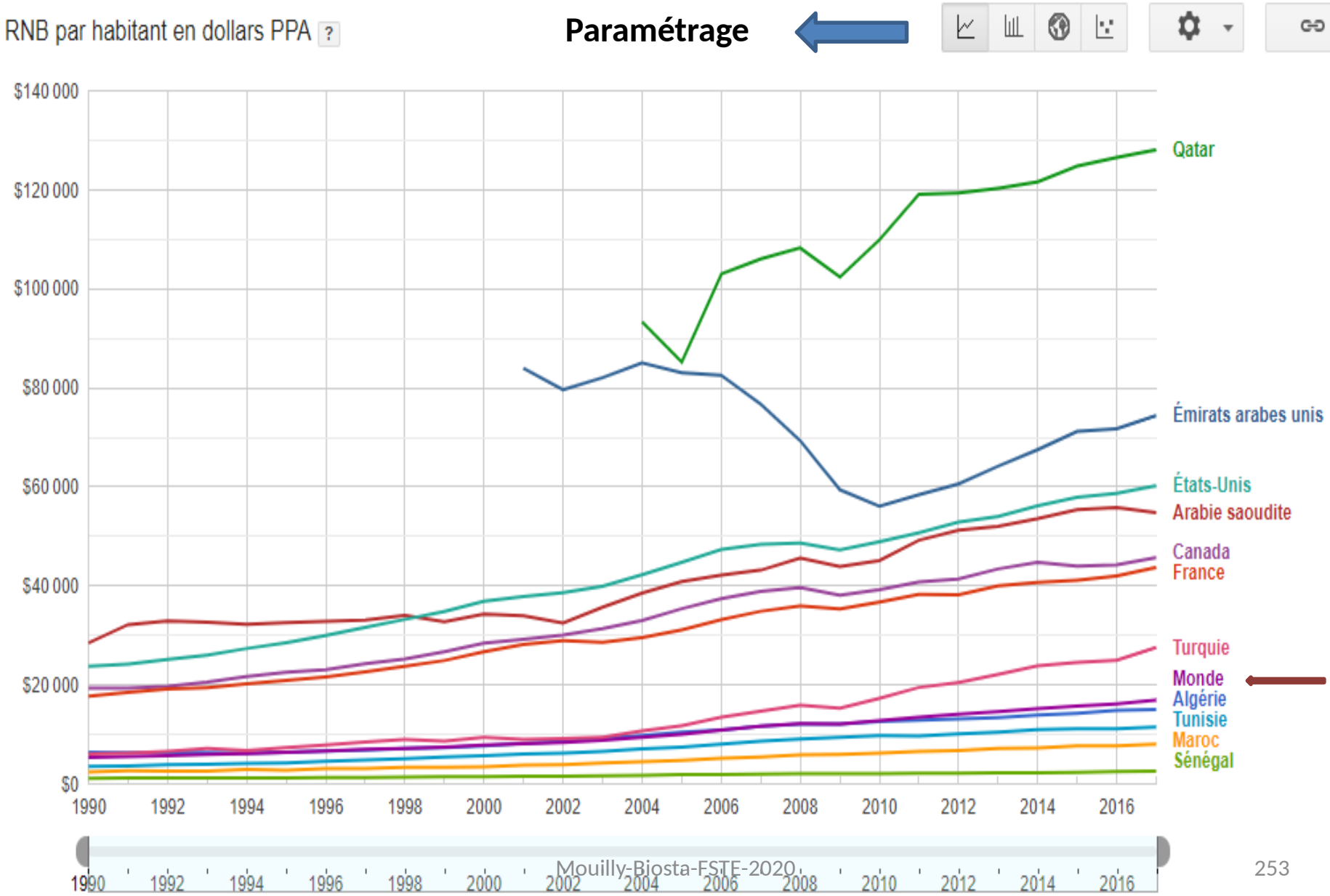




Web-Source

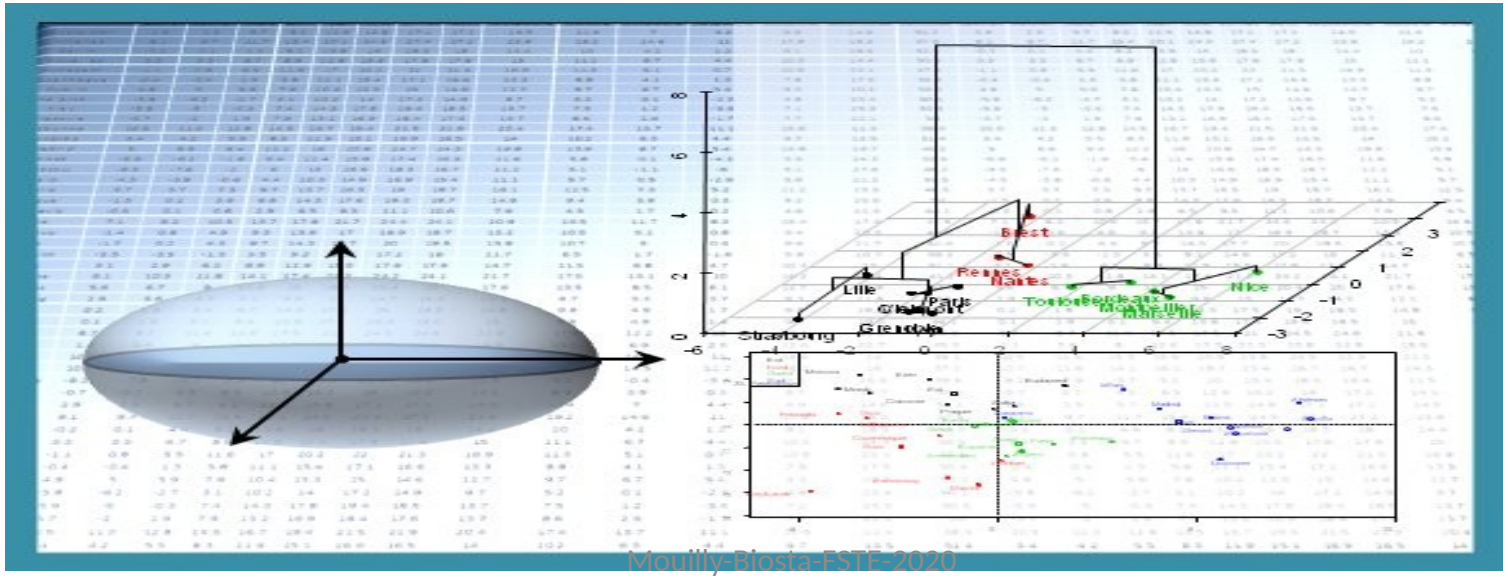
https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_&met_y=ny_gnp_pcap_pp_cd&idim=country:MAR:TUN:DZA&hl=fr&dl=fr#!ctype=l&strail=false&bcs=d&nselm=h&met_y=ny_gnp_pcap_pp_cd&scale_y=lin&ind_y=false&rdim=region&idim=country:MAR:TUN:DZA:QAT:TUR:FRA&ifdim=region&tdim=true&tstart=661219200000&tend=1513296000000&hl=fr&dl=fr&ind=false

Paramétrage



Partie II : Statistique Descriptive

- **Chapitre 1 : Statistique univariée (1 dimension)**
- **Chapitre 2 : Statistique descriptive bivariée (2 dimensions)**
- **Chapitre 3 : Statistique multivariée (x dimensions) : ACP, AFC, ACM…**



La statistique descriptive Multivariée

1. Généralités

- On désigne par statistique descriptive multidimensionnelle l'ensemble des méthodes de la statistique descriptive permettant de traiter simultanément un nombre quelconque de variables.
- Ces méthodes sont purement descriptives, c'est-a-dire qu'elles ne supposent, aucun modèle sous-jacent, de type probabiliste

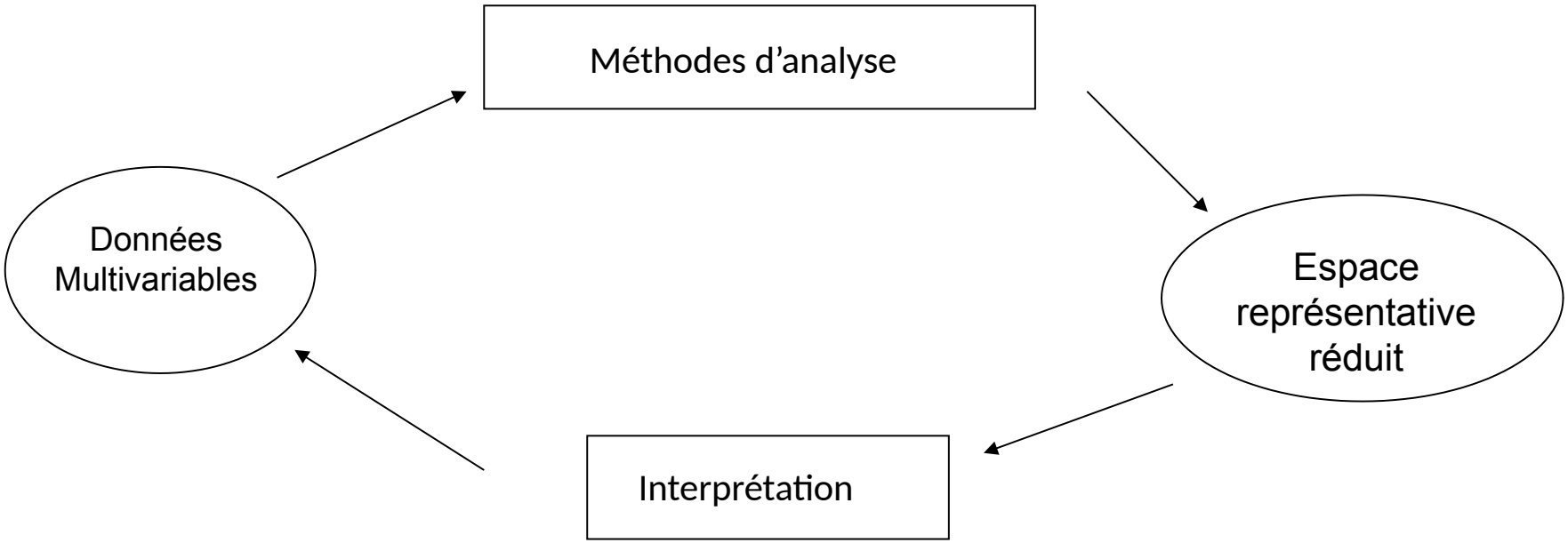
- Les méthodes les plus classiques de la statistique descriptive multidimensionnelle sont les méthodes factorielles, qui consistent a rechercher des facteurs en nombre restreint et résumant le mieux possible les données considérées.

BUT

- Synthétiser, structurer l'information contenue dans des données multidimensionnelles (I individus, K variables) **--->** q facteurs ou Composants Principales = Combinaisons linéaires des variables qui sont indépendants les uns aux autres (I <q<K)

Ces q facteurs que l'on va définir vont résumer l'information contenue dans le tableau initial et vont aussi maximiser la dispersion du nuage des observations.

Schéma global de Analyse multidimensionnelle des données

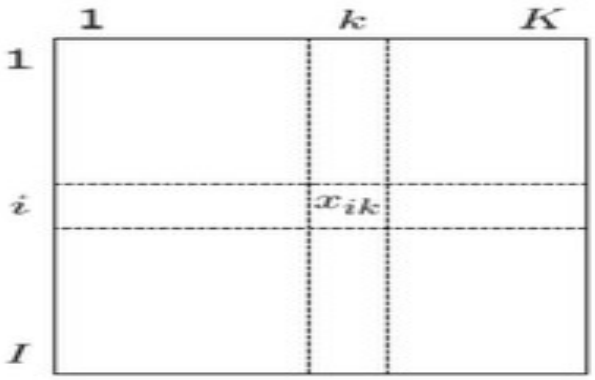


2. ACP : Analyse en Composants Principaux

Exemples d'application :

- Analyse sensorielle : note du descripteur *k* pour le produit *i*
- Ecologie : concentration du polluant *k* dans la rivière *i*
- Economie : valeur de l'indicateur *k* pour l'année *i*
- Génétique : expression du gène *k* pour le patient *i*
- Biologie : mesure *k* pour l'animal *i*
- Marketing : valeur d'indice de satisfaction *k* pour la marque *i*
- Sociologie : temps passé à l'activité *k* par les individus de la CSP *i*

L'ACP s'intéresse à des tableaux de données rectangulaires avec des individus en lignes et des variables quantitatives en colonnes



Pour la variable *k*, on note :

la moyenne : $\bar{x}_k = \frac{1}{I}\sum_{i=1}^{I} x_{ik}$

l'écart-type : $s_k = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ik} - \bar{x}_k)^2}$

Exemples : Données de Tp en France durant les 30 dernières années

- 15 individus (lignes) : villes de France
- 14 variables (colonnes) :
  - 12 températures mensuelles moyennes (sur 30 ans)
  - 2 variables géographiques (latitude, longitude)

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4	46.08	3.26

Deux façons de voir le tableau :
comme un ensemble de lignes
ou comme un ensemble de colonnes

Etude des individus

- Quand dit-on que 2 individus se ressemblent du point de vue de l'ensemble des variables ?
- Si beaucoup d'individus, peut-on faire un bilan des ressemblances ?

$\Rightarrow$ construction de groupes d'individus, partition des individus

Etude des variables

- Recherche des ressemblances entre variables
- Entre variables, on parle plutôt de liaisons
- Liaisons linéaires sont simples, très fréquentes et résument de nombreuses liaisons $\Rightarrow$ coefficient de corrélation

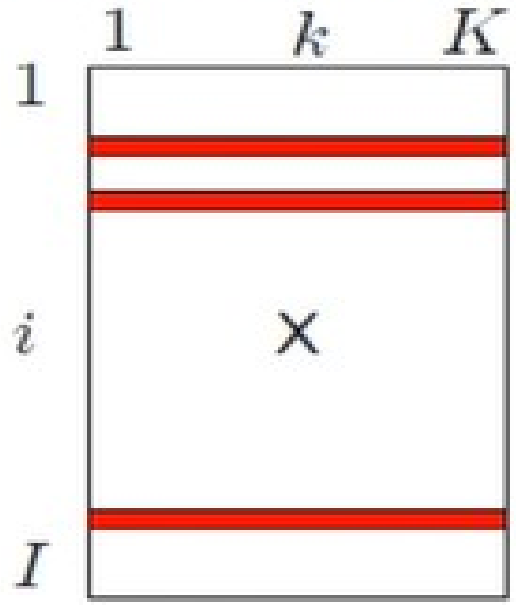
$\Rightarrow$ visualisation de la matrice des corrélations
$\Rightarrow$ recherche d'un petit nombre d'indicateurs synthétiques pour résumer beaucoup de variables

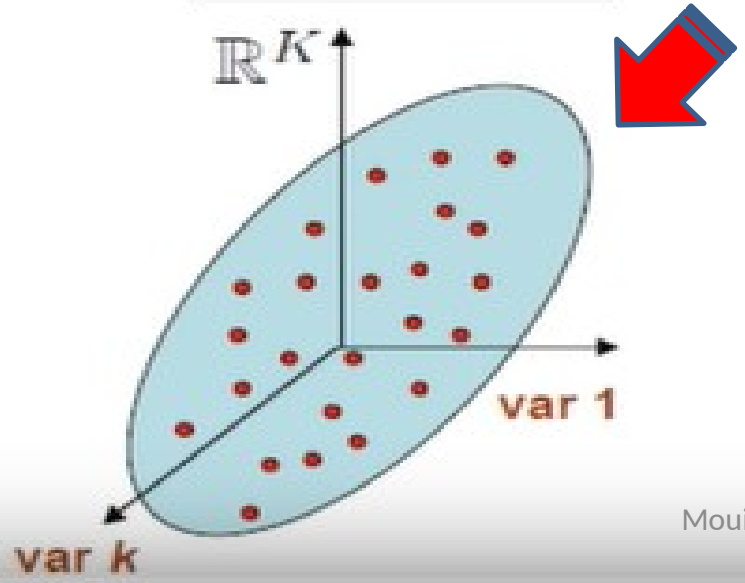
Lien entre les deux études

- Caractérisation des classes d'individus par les variables
  $\Rightarrow$ besoin de procédure automatique
- Individus spécifiques pour comprendre les liaisons entre variables
  $\Rightarrow$ utilisation d'individus extrêmes (en terme de variables : langage abstrait mais puissant, revenir aux individus pour voir les choses plus simplement)

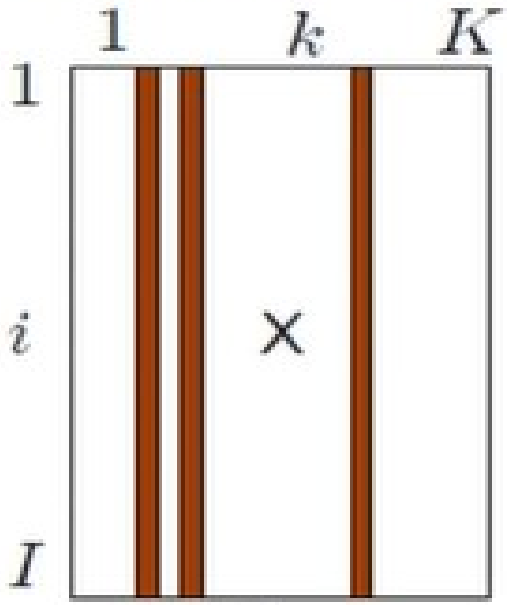
Deux nuages de points

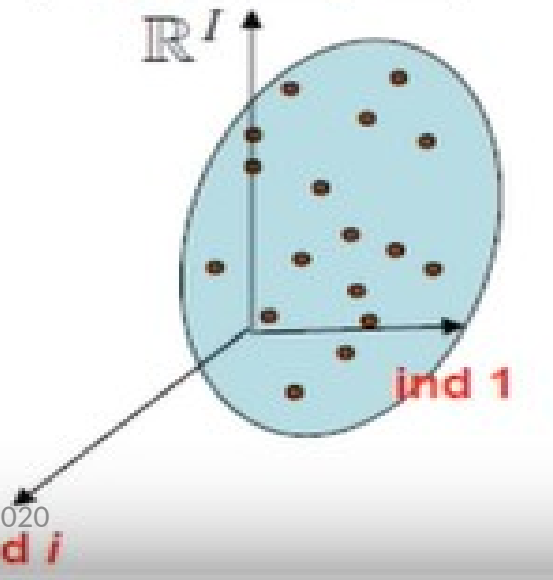
Etude des individus





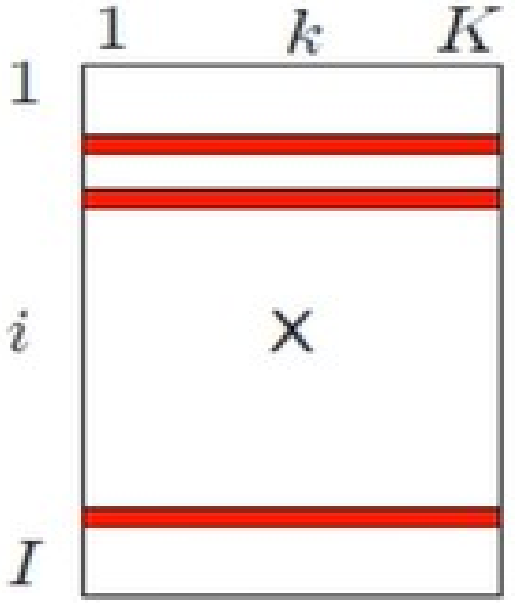
Etude des variables

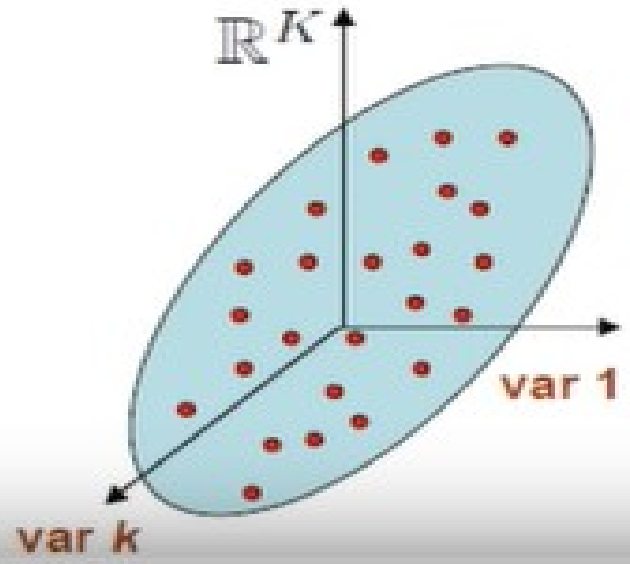




Deux nuages de points

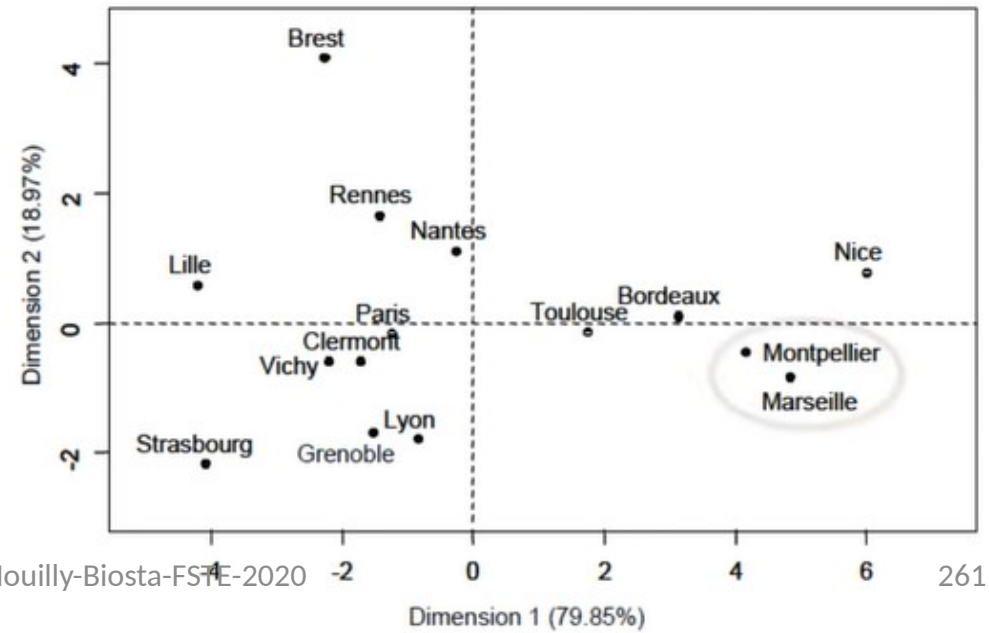
Etude des individus



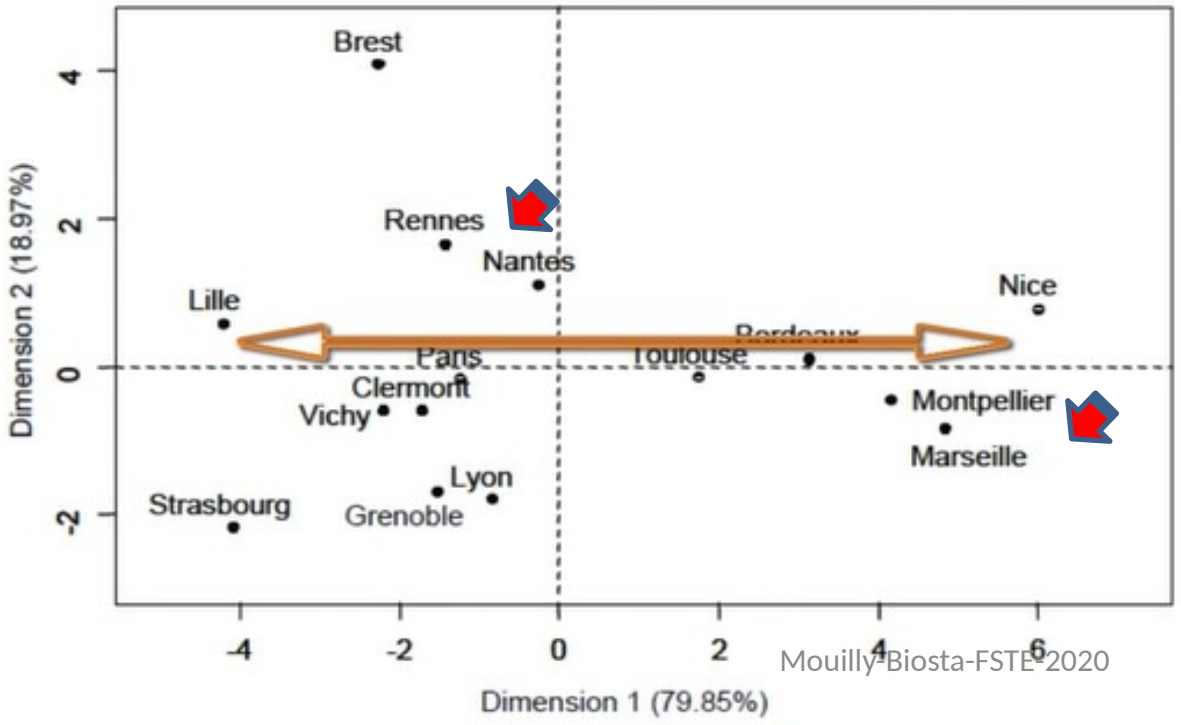


	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.1	46.08	3.26

Exemple : graphe des individus



	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.1	46.08	3.26

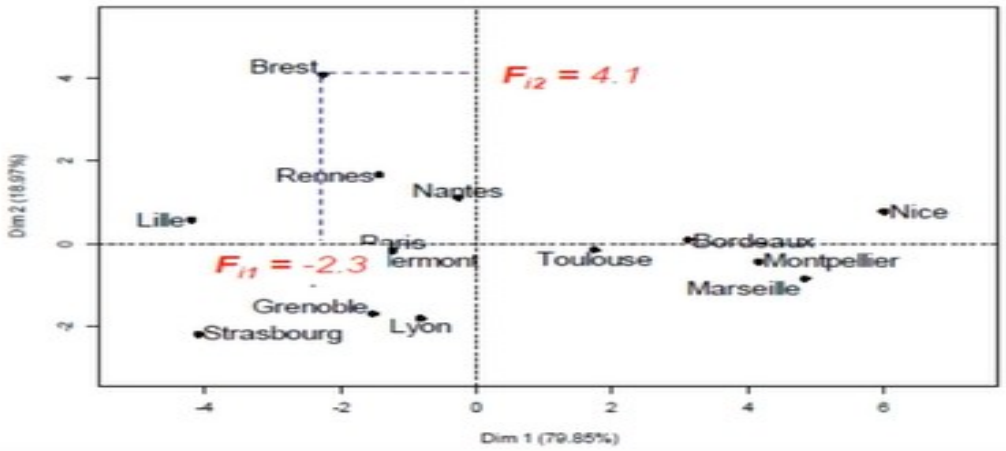


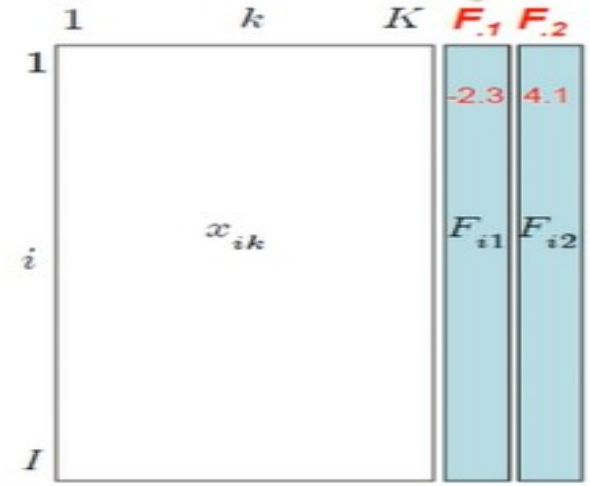
Pendant toute la période d'année, la température

- ✓ **Axe 1 : à Montpelier et à Marseille d'une part et à Rennes et Nantes sont très proches**
- ✓ **Axe 2 : à Lille et à Nice sont très opposées**

Interprétation du graphe des individus grâce aux variables

Considérons les coordonnées des individus sur les axes comme des variables



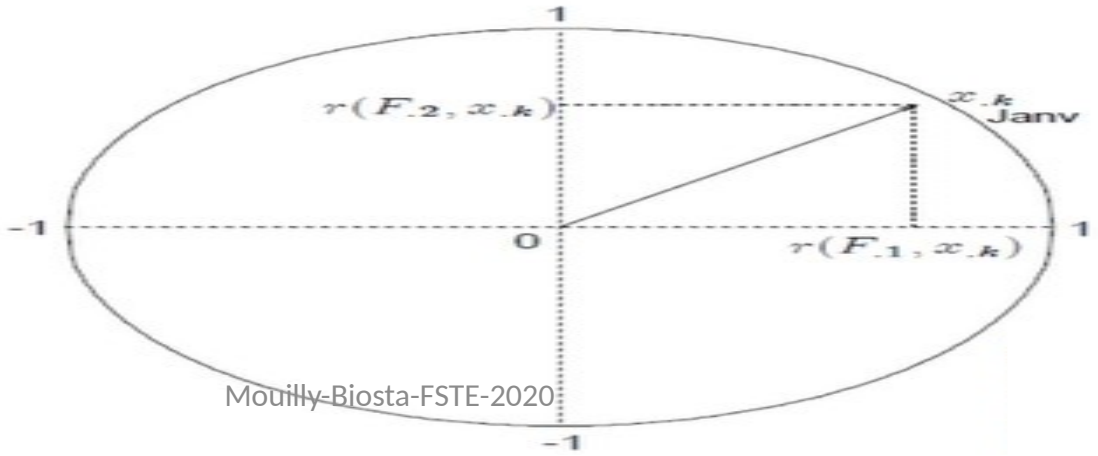


L'analyse est effectuée sur les données centrées réduites (de moyenne égale à 0 et d'écart-type éga à 1), de façon à éliminer les biais dus à des différences d'unités ou d'échelles entre les variables.

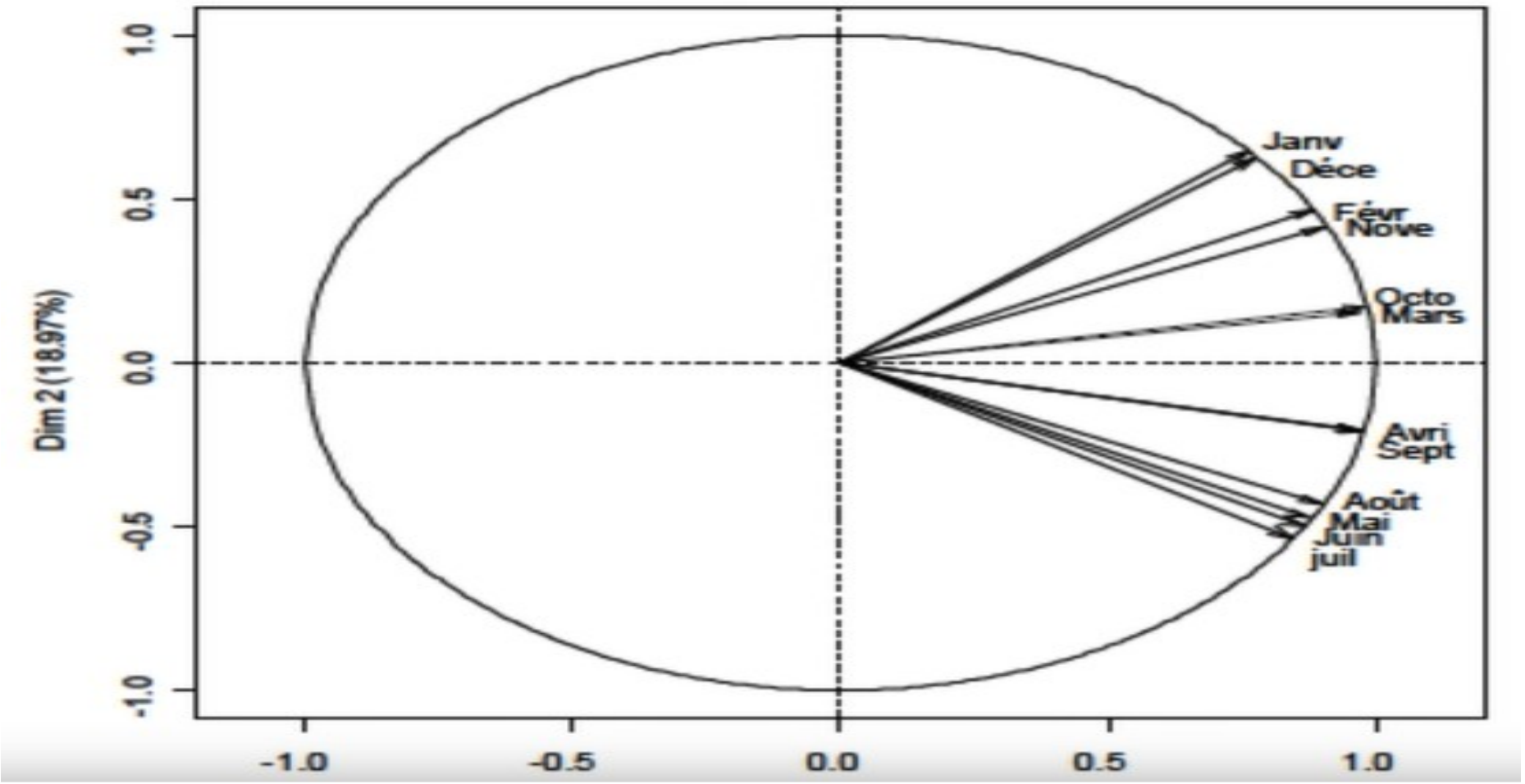
L'information totale contenue dans le nuage de points, appelée inertie totale, est égale à la dispersion des points du nuage autour du point moyen. Elle se mesure donc comme la somme des variances

- Corrélations entre la variable $x_{.k}$ et $F_{.1}$ (et $F_{.2}$)

Prenons le cas d'une seule variable : janvier



Cercle de corrélation



Graphe du Cercle de corrélation

→ **Toutes les variables étudiés sont présentées**

NB. Coefficient de détermination R^2 >70% ⇔ 2 Variables sont fortement liées

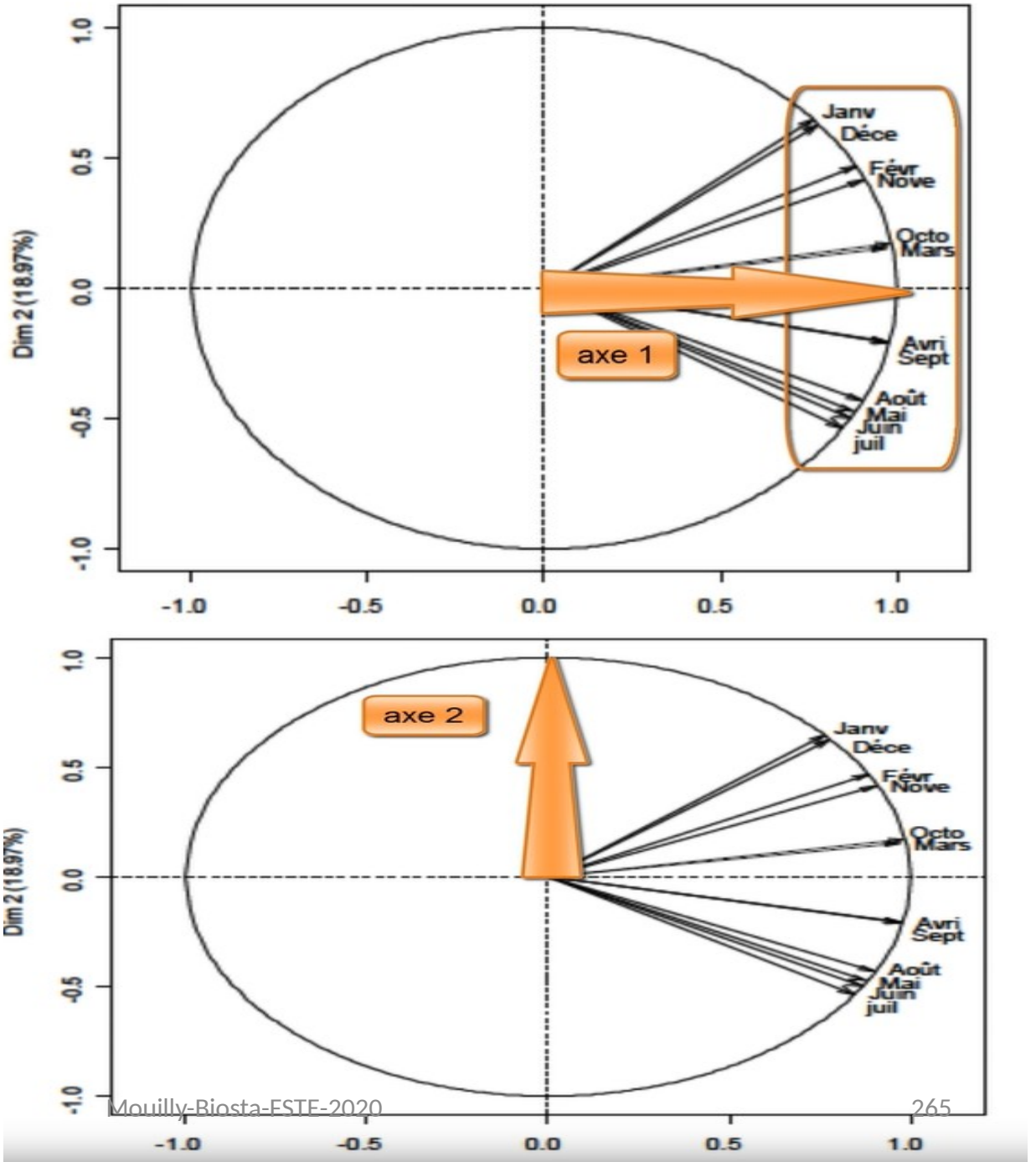
Axe 1:

important facteur de différentiation (inertie)

- **Toutes les variables avons une corrélation positive (sup à 0,6-0,7)**
- **Fortes corrélation pour oct. et mars (~1) ⊔ même comportement**

Axe 2:

Moins important que F2



2. AFC : Analyse Factorielles des Correspondances

- Ces méthodes d'analyse tout comme celles d'analyse en composantes principales s'utilisent pour décrire et hiérarchiser les relations statistiques qui peuvent exister entre des individus placés en ligne et des variables placées en colonnes dans un tableau rectangulaire de données.
- Elle s'applique à des tableaux de contingence c'est-à-dire des tableaux croisant deux variables qualitatives.
- La spécificité de l'AFC est qu'elle considère en même temps un nuage de point représentant les lignes (individus) et un autre représentant les colonnes (variables).

→ Le principe est d'étudier simultanément les lignes et les colonnes et de mettre en évidence les "correspondances", c'est-à-dire les liaisons entre ces deux ensembles.

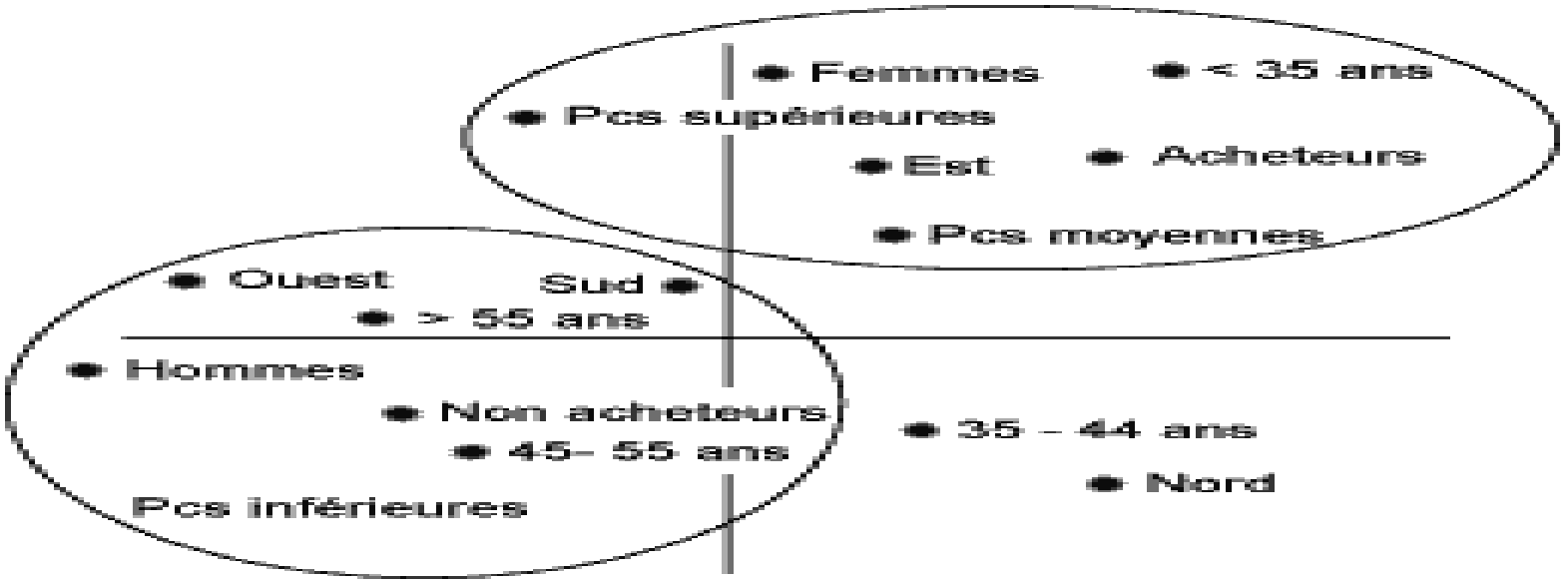
Les logiciels d'AFC fournissent donc en sortie une ou plusieurs figures de plans factoriels sur lesquels sont positionnés à la fois les individus et le variables.

→ Le but est de réduire la complexité des données tout en donnant le maximum d'informations possibles et accessibles.

Le tableau donne la répartition d'une population par classe d'âge et loisir préféré:

	Moins de 15 ans	15 à 24 ans	25 à 39 ans	40 à 60 ans	Plus de 60 ans
TV	322	114	72	135	130
Théâtre	1	17	85	92	14
Cinéma	90	220	192	87	7
Lecture	23	38	57	73	80
Restaurant	7	53	158	49	13
Night-Club	0	87	109	21	0
Concert	27	153	130	47	1

Exemple
Procéder à une typologie des acheteurs et non acheteurs d'un produit ou d'un service

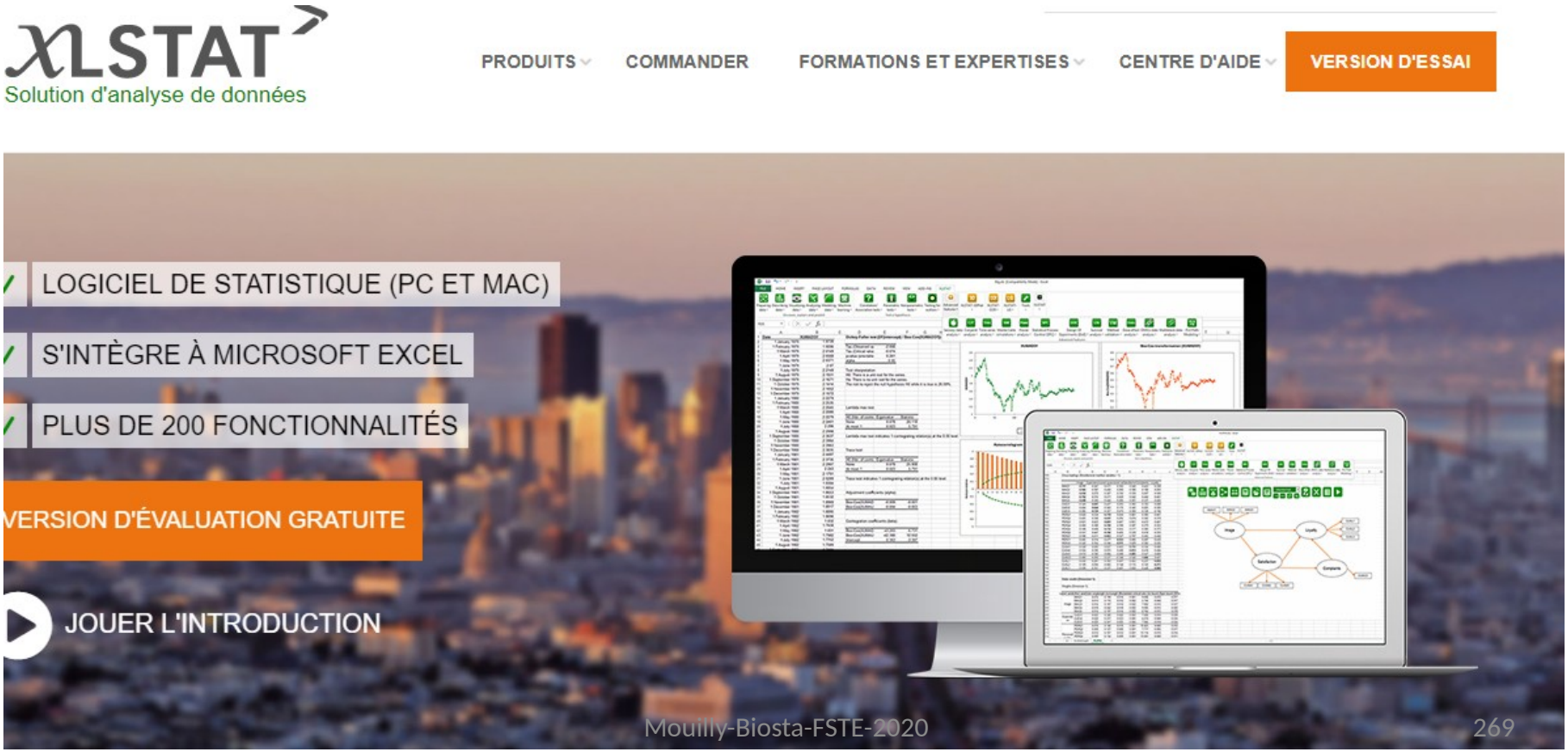


F1 : L'axe horizontal fortement corrélé à l'âge nous montre ici qu'il existe un effet d'âge important : qui est pour beaucoup dans le positionnement des points.
F2 : L'axe vertical traduit un effet catégories sociales.

- Au plus deux points sont proches sur le mapping, au plus les deux variables correspondantes sont liées.
- L'A.F.C. nous permet ainsi d'identifier les deux profils discriminants des acheteurs et des non acheteurs.

Le logiciel XLSTAT

- https://www.xlstat.com/fr/



Version d'essai de 14 jours et version gratuite

Profitez de toutes les 200 fonctionnalités de XLSTAT-Premium gratuitement pendant 14 jours, puis d'une version gratuite limitée comprenant 13 fonctionnalités. **

Prénom *

Nom de famille *

Adresse mail *

Téléphone *

Organisation *

Poste *

Pays *

Afghanistan

Vous êtes *

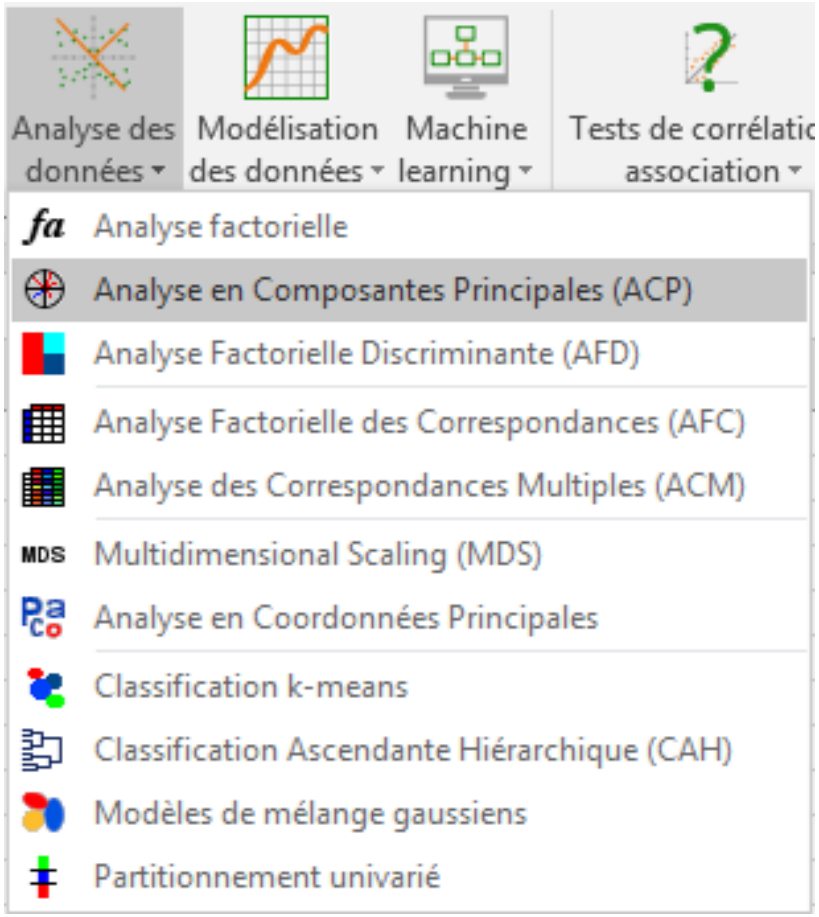
Sélectionnez...

Vista, Win7, Win8, Win10

≥ 10.8

Analyse en Composantes Principales

XLSTAT / Analyse de données / Analyse en Composantes Principales, ou cliquez sur le bouton correspondant de la barre **Analyse de données** (voir ci-dessous).



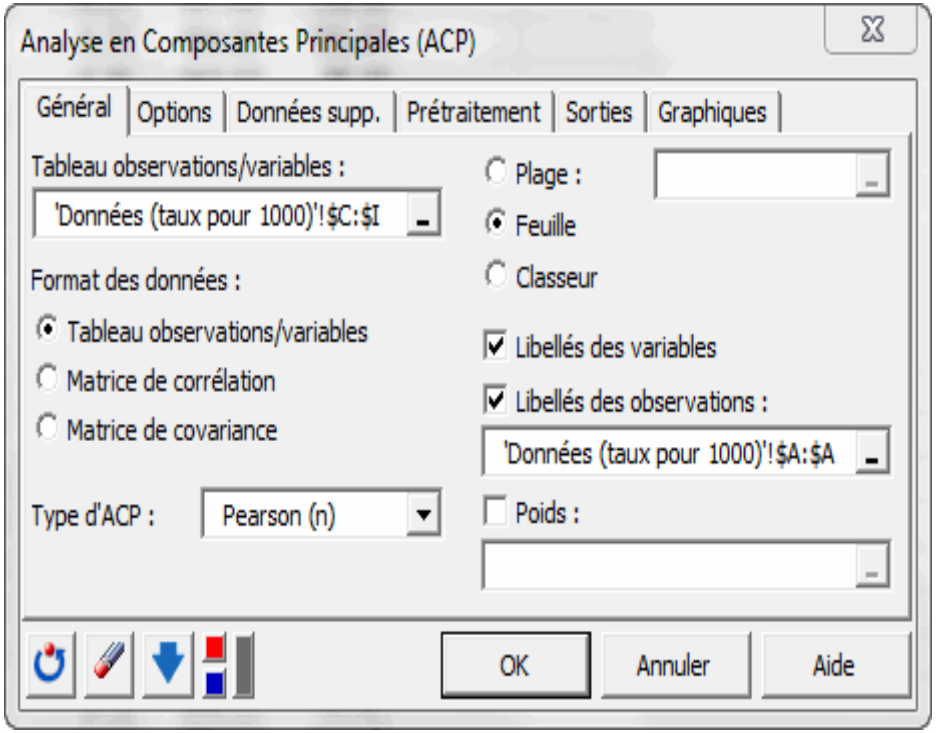
Une fois le bouton cliqué, la boîte de dialogue correspondant à l'**Analyse en composantes principales** apparaît. Vous pouvez alors sélectionner les données sur la feuille Excel.

L'option **Libellés des variables** est activée, car la première ligne de données contient le nom des variables.

Le **Format des données** choisi ici est **Observations/Variables** car c'est bien le format des données de départ.

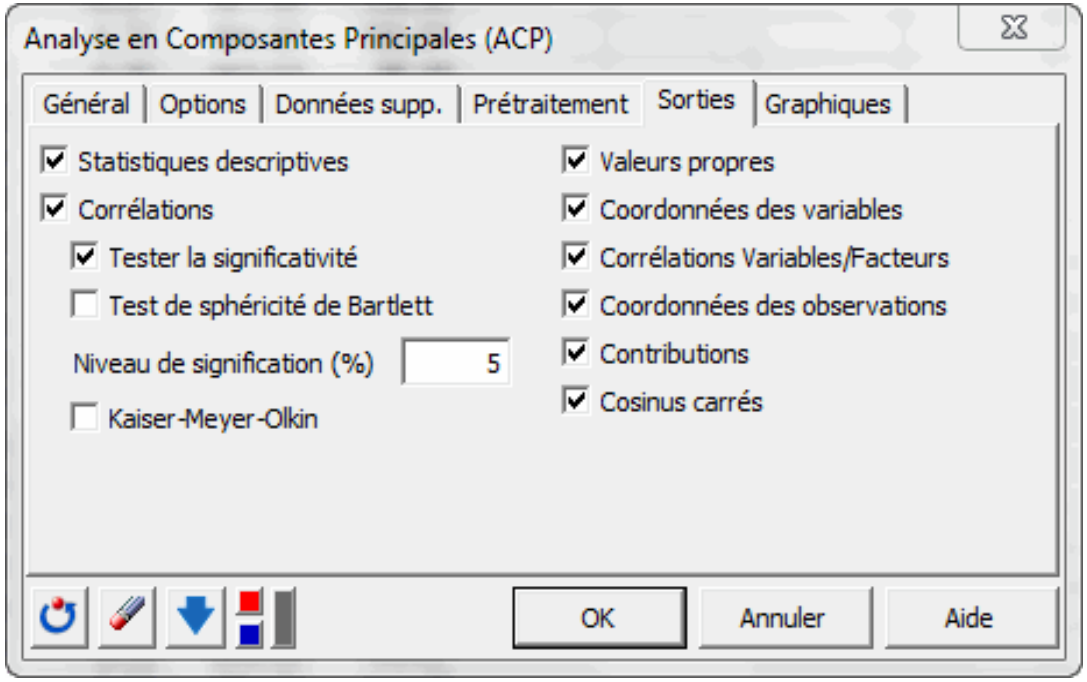
Le **Type d'ACP** choisi est **Pearson (n)**, ce qui signifie que les calculs seront basés sur une matrice composée des coefficients de corrélation de Pearson,

le coefficient de Pearson étant le coefficient de corrélation classiquement utilisé. Les matrices de covariance allouent plus de poids aux variables ayant des variances élevées.

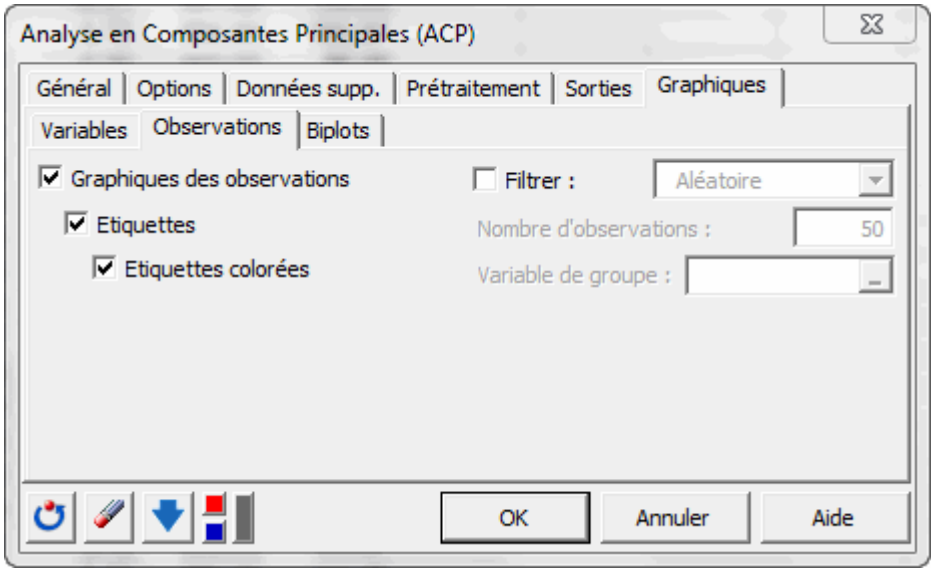


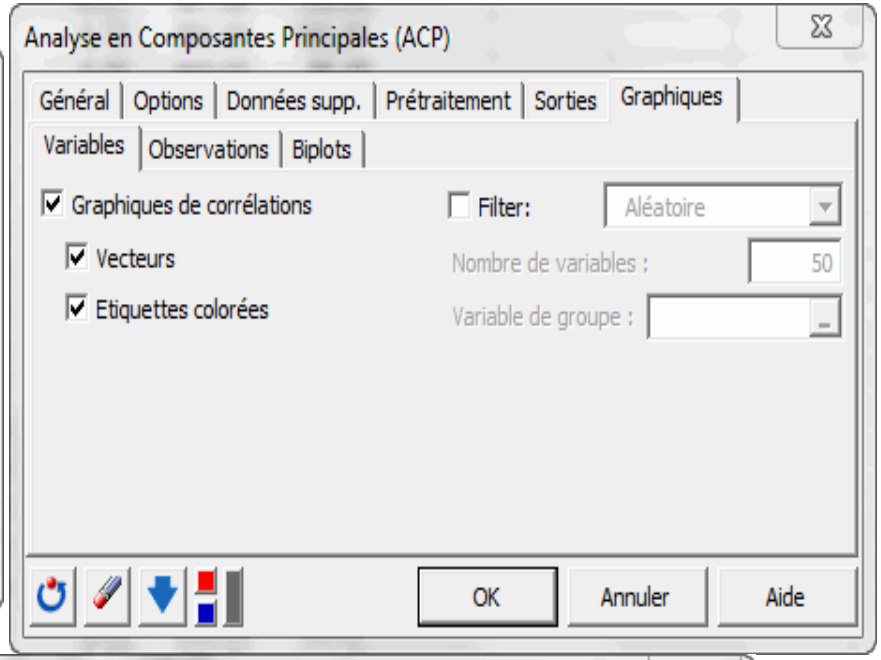
Les corrélations de Spearman peuvent être plus appropriées lorsque l'ACP est exécutée sur des variables aux distributions différentes. Les corrélations polychoriques sont adaptées aux variables ordinales.

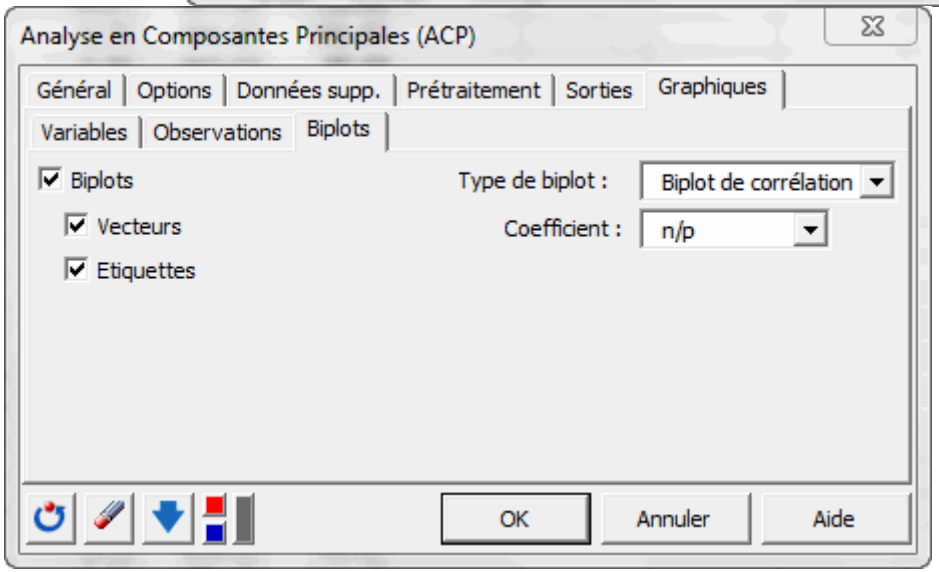
Dans l'onglet **Sorties**, on a choisi d'activer l'option **Tester la significativité** pour afficher en gras les corrélations significativement différentes de 0.



Dans l'onglet **Graphiques**, les options d'**Etiquettes** sont toutes activées afin que les libellés des variables et des observations soient bien affichés.
L'option de **filtrage des observations à afficher** est aussi désactivée afin d'afficher toutes les observations

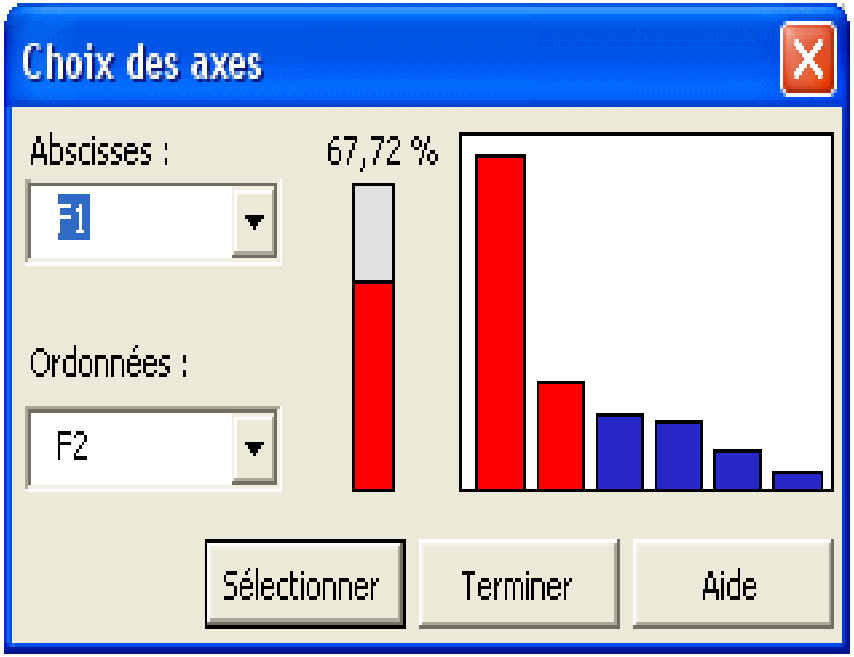


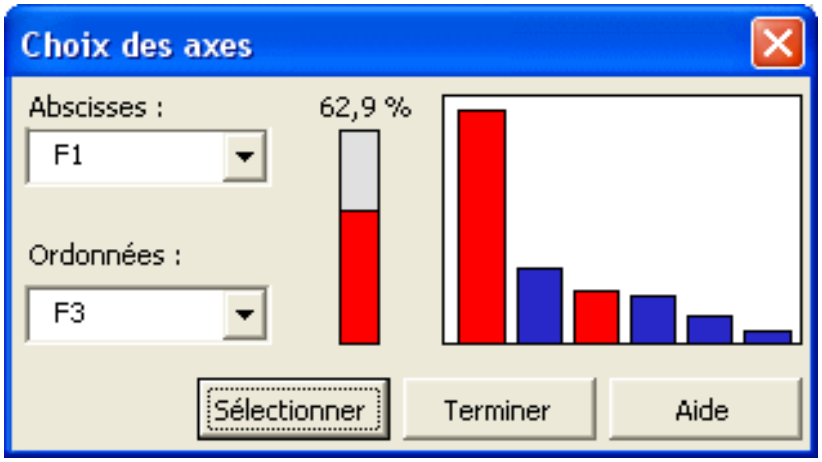


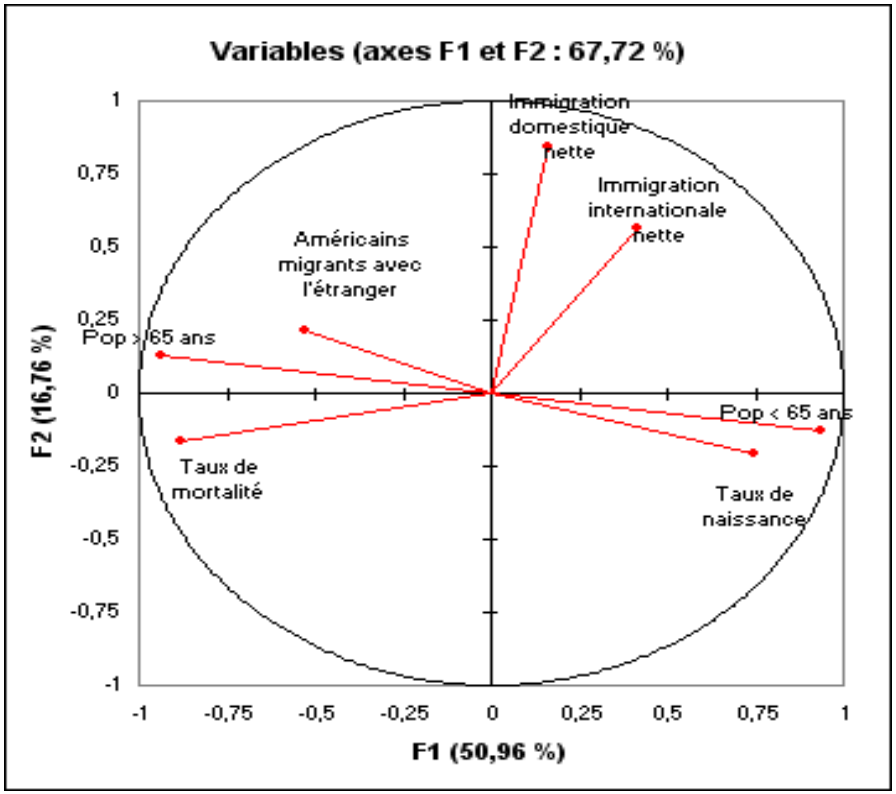


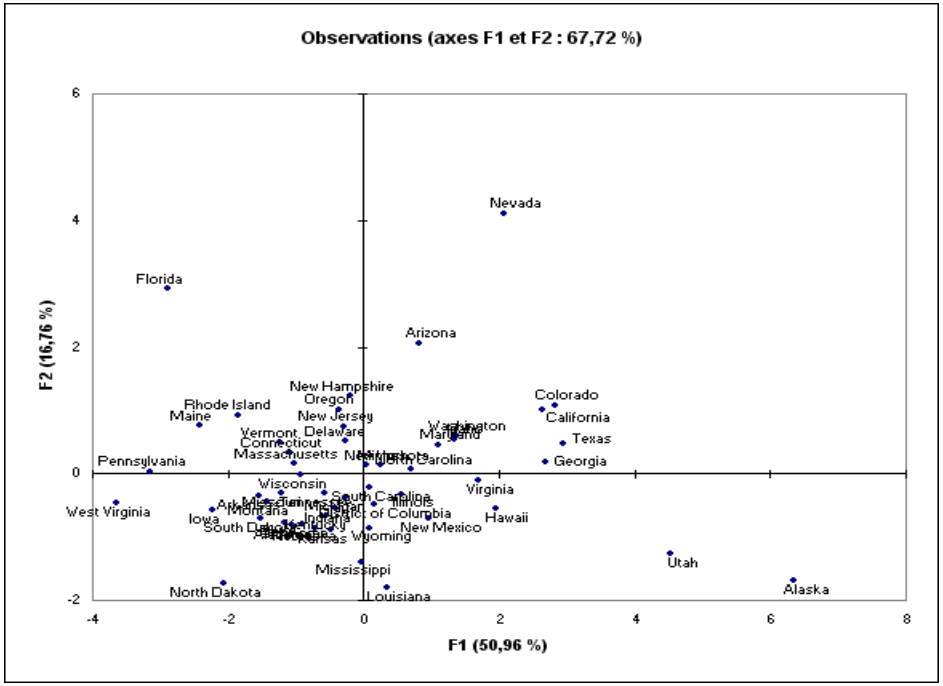
Les calculs commencent lorsque vous cliquez sur le bouton **OK**.

une nouvelle boîte vous permet de choisir les axes pour lesquels les graphiques doivent être affichés. Dans notre cas, le pourcentage de variabilité représenté sur les deux premiers axes n'est pas particulièrement élevé (67.72%)



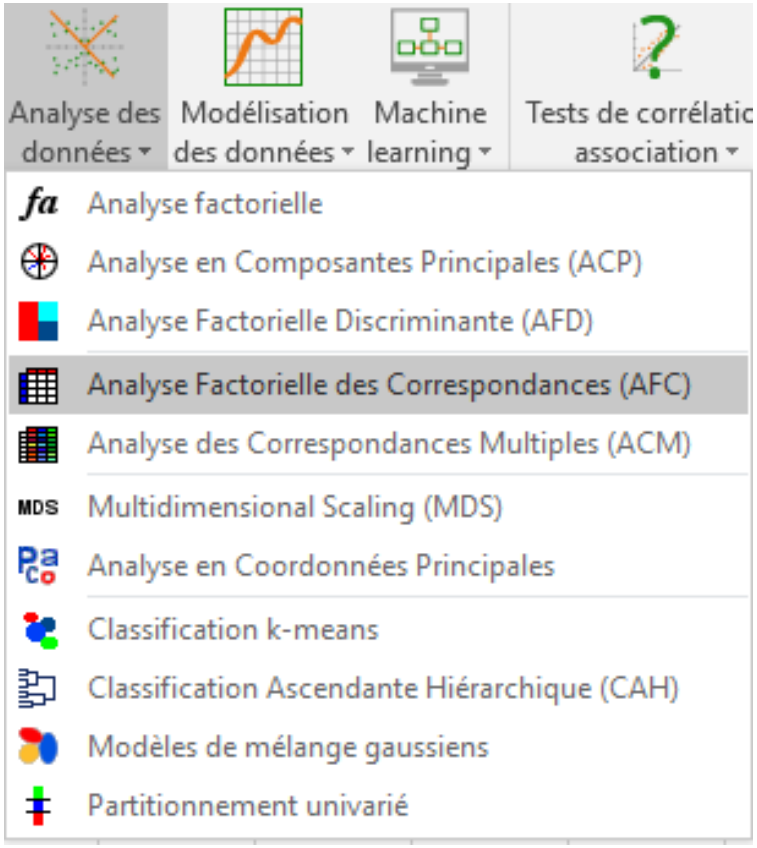



Variables (axes F1 et F2 : 67,72 %)
Immigration domestique nette
Immigration internationale nette
Américains migrants avec l'étranger
Pop > 65 ans
Pop < 65 ans
Taux de mortalité
Taux de naissance
F2 (16,76 %)
F1 (50,96 %)
1
0,75
0,5
0,25
0
-0,25
-0,5
-0,75
-1
-1
-0,75
-0,5
-0,25
0
0,25
0,5
0,75
1

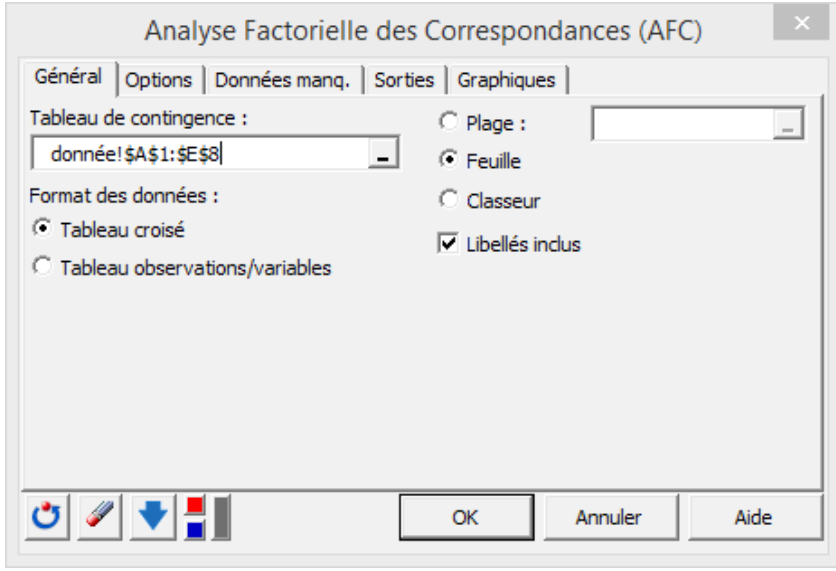

Observations (axes F1 et F2 : 67,72 %)
F2 (16,76 %)
F1 (50,96 %)
Nevada
Florida
Arizona
New Hampshire
Oregon
Rhode Island
Maine
New Jersey
Colorado
California
Vermont
Delaware
Connecticut
Massachusetts
Texas
Georgia
Pennsylvania
Wisconsin
Virginia
West Virginia
Iowa
Hawaii
New Mexico
Wyoming
Utah
Mississippi
North Dakota
Louisiana
Alaska
6
4
2
0
-2
-4
-2
0
2
4
6
8

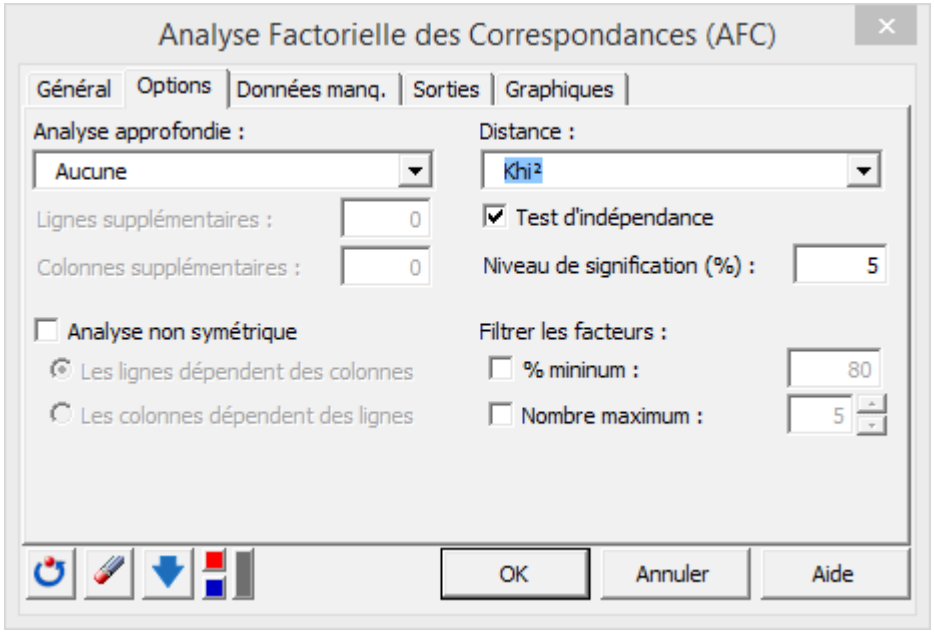
Analyse Factorielle des Correspondances AFC

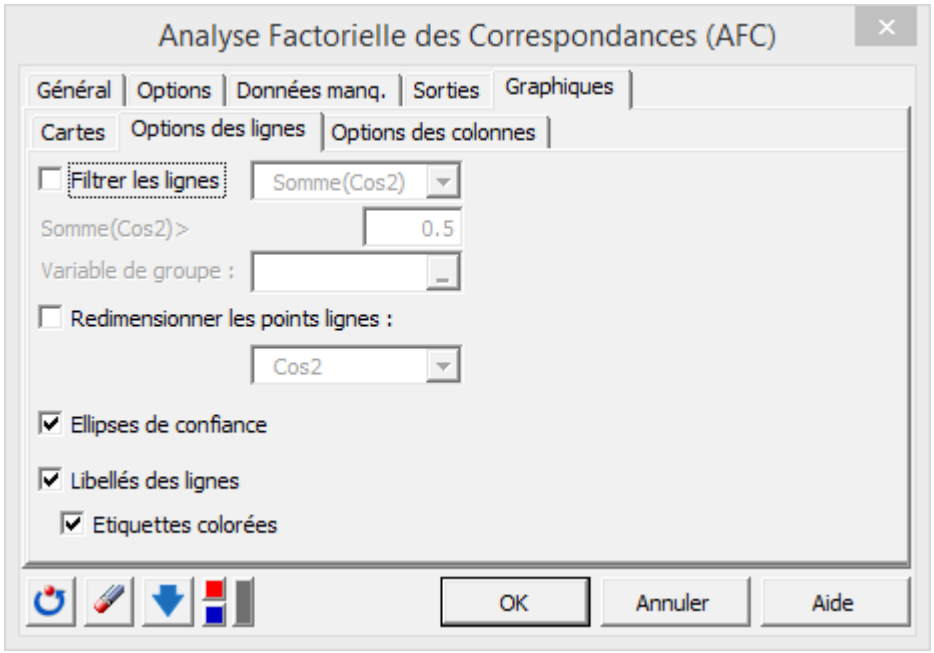
XLSTAT / Analyse de données / AFC ou cliquez sur le bouton correspondant de la barre **Analyse de données** (voir ci-dessous).

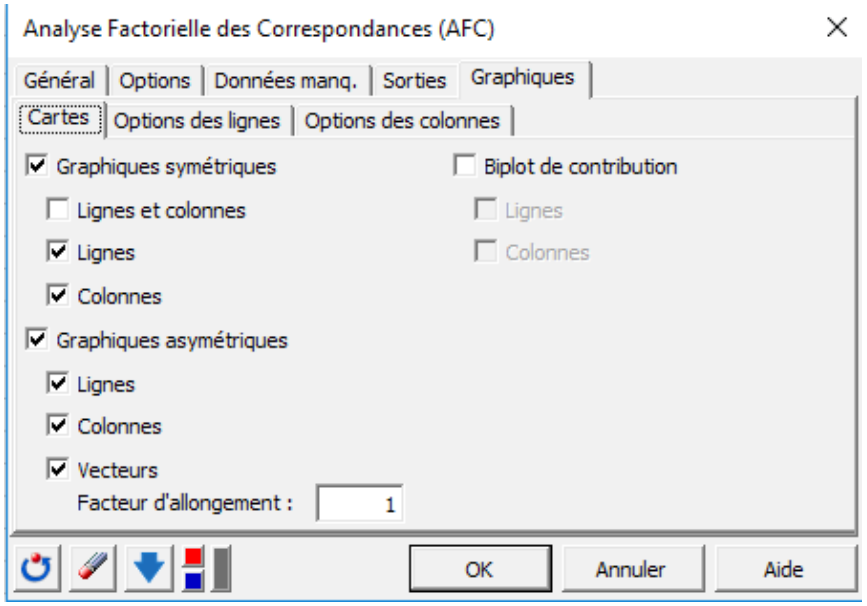


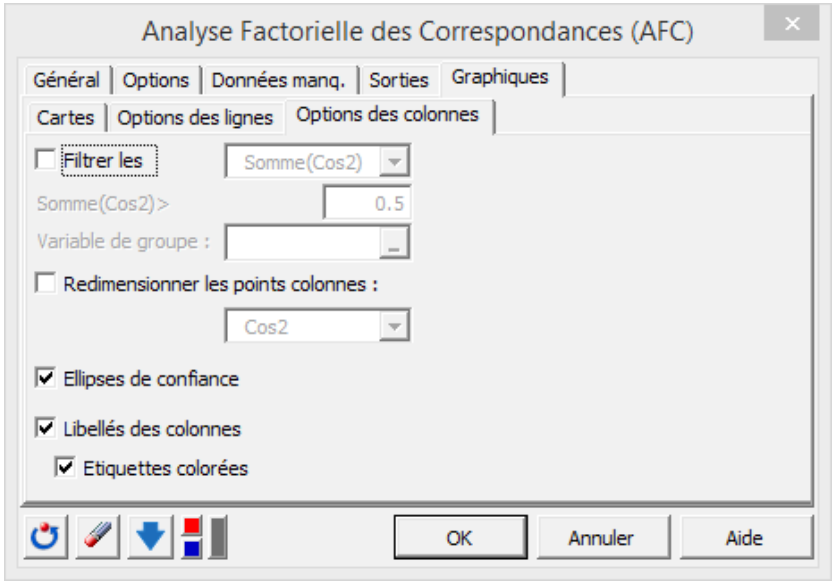
LES PARAMETRES






Analyse Factorielle des Correspondances (AFC)
Général
Options
Données manq.
Sorties
Graphiques
Cartes
Options des lignes
Options des colonnes
Filtrer les lignes
Somme(Cos2)
Somme(Cos2)>
0.5
Variable de groupe :
Redimensionner les points lignes :
Cos2
Ellipses de confiance
Libellés des lignes
Etiquettes colorées
OK
Annuler
Aide


Analyse Factorielle des Correspondances (AFC)
Général
Options
Données manq.
Sorties
Graphiques
Cartes
Options des lignes
Options des colonnes
Graphiques symétriques
Biplot de contribution
Lignes et colonnes
Lignes
Colonnes
Lignes
Colonnes
Graphiques asymétriques
Lignes
Colonnes
Vecteurs
Facteur d'allongement :
1
OK
Annuler
Aide

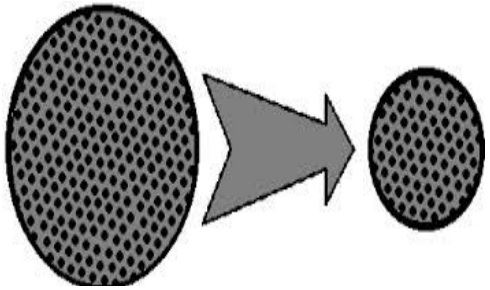

Analyse Factorielle des Correspondances (AFC)
Général
Options
Données manq.
Sorties
Graphiques
Cartes
Options des lignes
Options des colonnes
Filtrer les
Somme(Cos2)
Somme(Cos2)>
0.5
Variable de groupe :
Redimensionner les points colonnes :
Cos2
Ellipses de confiance
Libellés des colonnes
Etiquettes colorées
OK
Annuler
Aide

Partie III : Statistique Inférentielle

1. Les Principes de L'Inférence Statistique
2. Les méthodes d'échantillonnage
3. Principaux paramètres de L'Inférence Statistique
4. Les méthodes relatives aux moyennes
5. Les méthodes relatives aux proportions
6. Taille d'échantillon

1 : Les Principes de L'Inférence Statistique

- **L'inférence statistique consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon issu de cette population. Ces caractéristiques, une fois connues, reflètent avec une certaine marge d'erreur celles de la population.**



Types d'enquêtes statistiques

Enquête exhaustive

\+
Photographie de la population (t)
Mesure de la valeur vraie de Y
Création d'une base de sondage

\-
Opération coûteuse
Longue
Fastidieuse

Enquête par échantillonnage

\+
Coût réduit
Temps réduit
Ressources humaines réduites
Facilité d'exécution

\-
Choix des unités
Choix de l'estimateur

Analyse combinatoire
(Dénombrement)

Permutation	Arrangement	Combinaison
Disposition **ordonnée** de **tous** les éléments d'un ensemble.	Disposition **ordonnée** d'un **certain** nombre d'éléments d'un ensemble.	Disposition **non ordonnée** d'un **certain** nombre d'éléments d'un ensemble.

Combinaison

Simple (sans répétition) **:** est une collection de *k* objets pris simultanément parmi *n*, donc sans tenir compte de l'ordre d'apparition. Elle est dite simple si on ne peut prendre chaque objet qu'une fois au plus.

= Échantillonnage sans remise, sans ordre $\Leftrightarrow$ $C_n^k = = A_n^k / k! = n! / k!(n - k)!$

(lu « k parmi n » ou « combinaison de k parmi n »);

NB. $A_n^k = n! / (n - k)!$

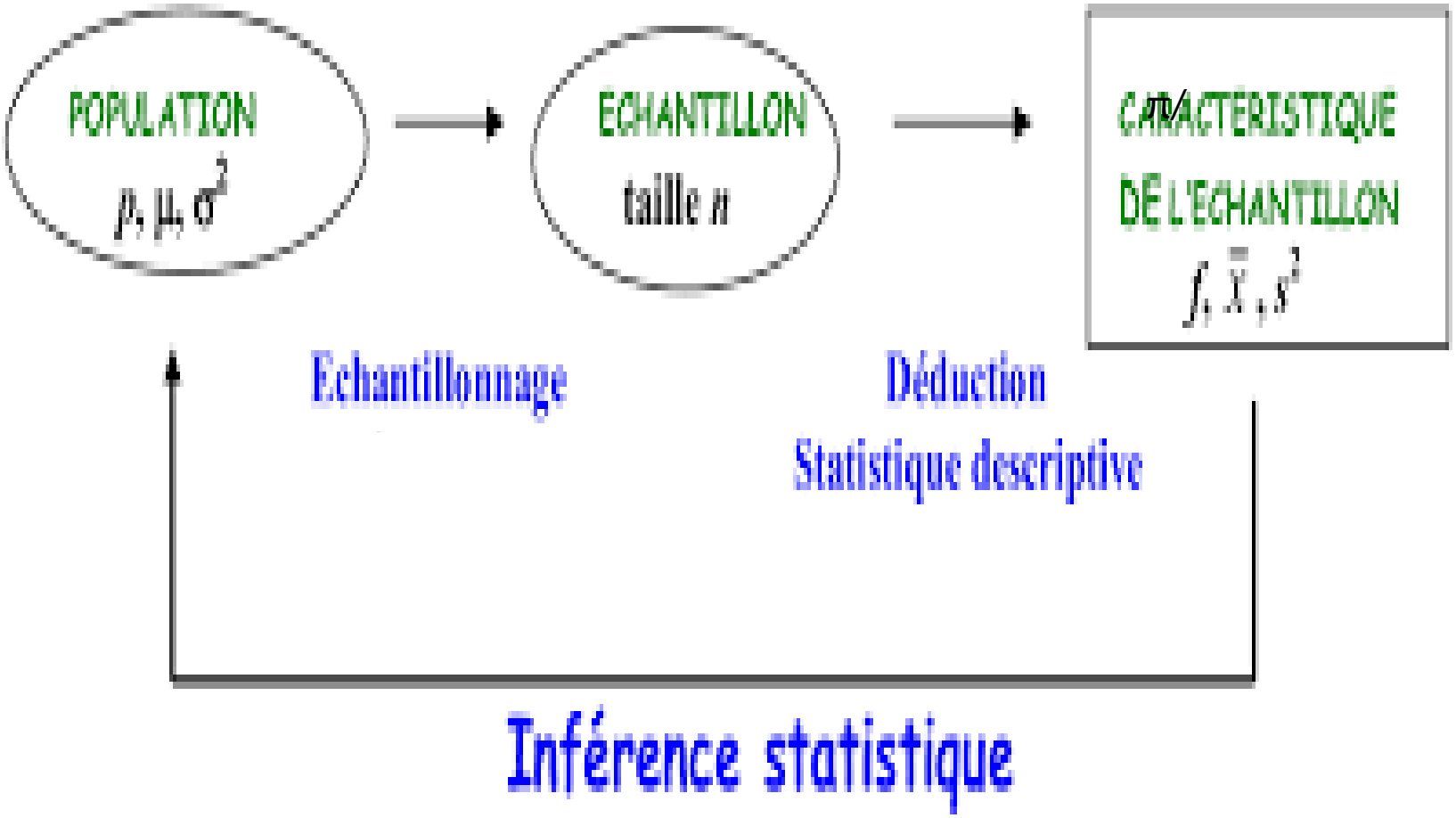
Exemples : Combinaison sans répétition de 2 (k) individus dans une population de taille 4 (n) *{1,2,3, 4}* $C_4^2 = 4! / 2!(4-2)! = 6$ (1,2) (1,3) (1,4) (2,3) (2,4) (3,4)

Avec répétition **: est une combinaison avec des répétitions permises.**

= Échantillonnage avec remise, sans ordre ➔ $(n+k-1)! / (n - 1)!k!$

Exemples : Combinaison avec répétition de 2 (k) individus dans une population de taille 4 (n) ***{1,2,3, 4}*** $=5! / 3!2! = 10$ (1,1) (1,2) (1,3) (1,4) (2,2) (2,3) (2,4) (3,3) (3,4) (4,4)

Recensement



Partie 3 : Statistique Inférentielle

1. Les Principes de L'Inférence Statistique
2. Les méthodes d'échantillonnage
3. Principaux paramètres de L'Inférence Statistique
4. Les méthodes relatives aux moyennes
5. Les méthodes relatives aux proportions
6. Taille d'échantillon

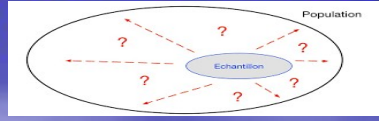
2 : Méthodes d'échantillonnage

Ces techniques d'étude dépendent de plusieurs critères :

- ✓ Contraintes financières
- ✓ Contraintes administratives
- ✓ Population finie
- ✓ Population infinie
- ✓ Population homogène
- ✓ Population échantillonnée

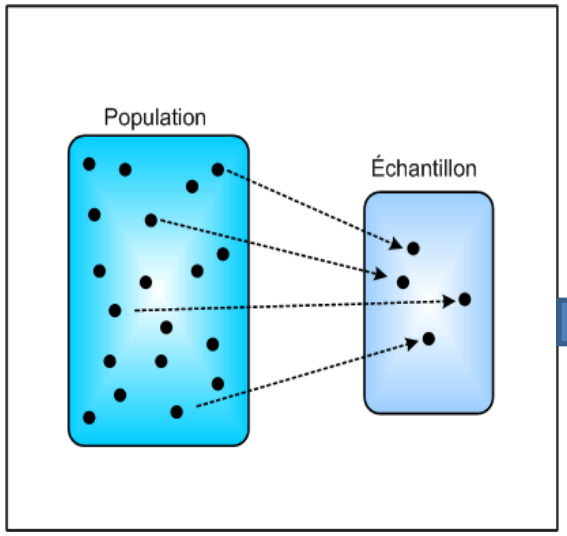
- **Méthodes probabilistes**
  - ✓ Echantillonnage aléatoire simple
  - ✓ Echantillonnage systématique (par intervalle)
  - ✓ Echantillonnage stratifié
  - ✓ Echantillonnage par grappes
  - ✓ Phases multiples

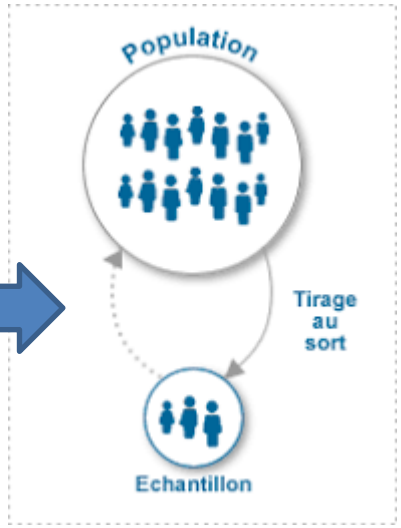
- **Méthodes non probabilistes (Raisonnées ou Empiriques)**
  - ✓ Echantillonnage par quotas
  - ✓ Echantillonnage "volontaire"
  - ✓ Echantillonnage de convenance
  - ✓ Echantillonnage selon le jugement
  - ✓ Echantillonnage de boule de neige
  - ✓ Echantillonnage structuré aléatoire

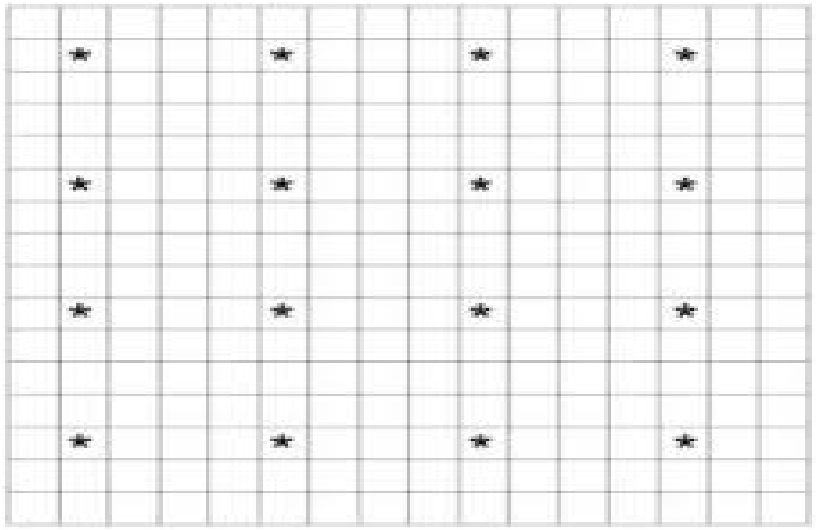


Méthodes d'échantillonnage

- Méthodes
  - Non probabilistes =Raisonnées =Empiriques
    - De convenance
    - De jugement
    - Boule de neige
    - Par quota
    - Volontaire
    - Structuré aléatoire
  - Probabilistes
    - Aléatoire simple
    - Systématique
    - Stratifié
    - Grappes
    - Phases multiples





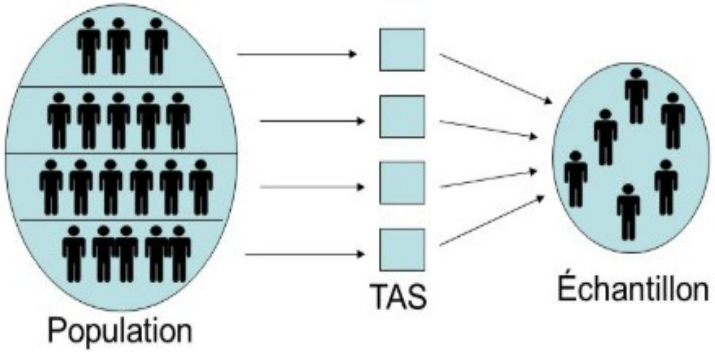


Aléatoire simple

Systématique (contrôle de qualité)

1• L'échantillon est constitué par un sondage aléatoire simple par strate :

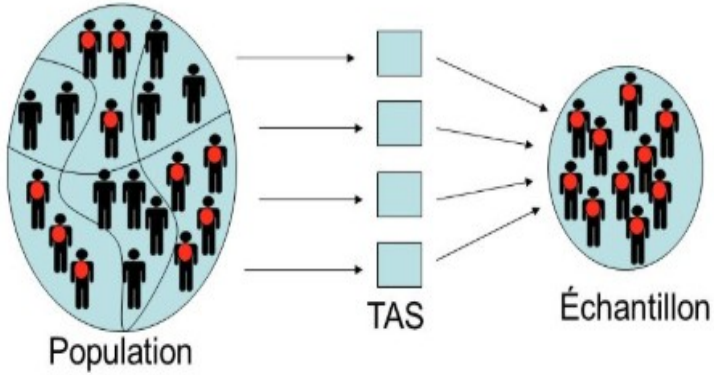
2• Tirage au sort des unités dans chaque strate



Stratifié

1 • Ce sont les grappes qui sont tirées au sort dans la population

2 • L'ensemble des sujets d'une grappe tirée au sort sera enquêté



Grapping

Partie 3 : Statistique Inférentielle

1. Les Principes de L'Inférence Statistique
2. Les méthodes d'échantillonnage
3. Principaux paramètres de L'Inférence Statistique
4. Les méthodes relatives aux moyennes
5. Les méthodes relatives aux proportions
6. Taille d'échantillon

L'Intervalle de confiance : noté IC

C'est un intervalle qui encadre une valeur réelle ou autre statistiques comme la moyenne, la médiane ou la variance, proportions,….que l'on cherche à estimer à l'aide de mesures prises par un procédé aléatoire.

NB. Ne pas confondre avec l'intervalle de fluctuation (Intervalle de pari)

- **Cet intervalle est basé sur la connaissance des paramètres de la population**

Les valeurs les plus courants sont : 95%, 99%, 99,9%….

Le Risque d'erreur : noté α

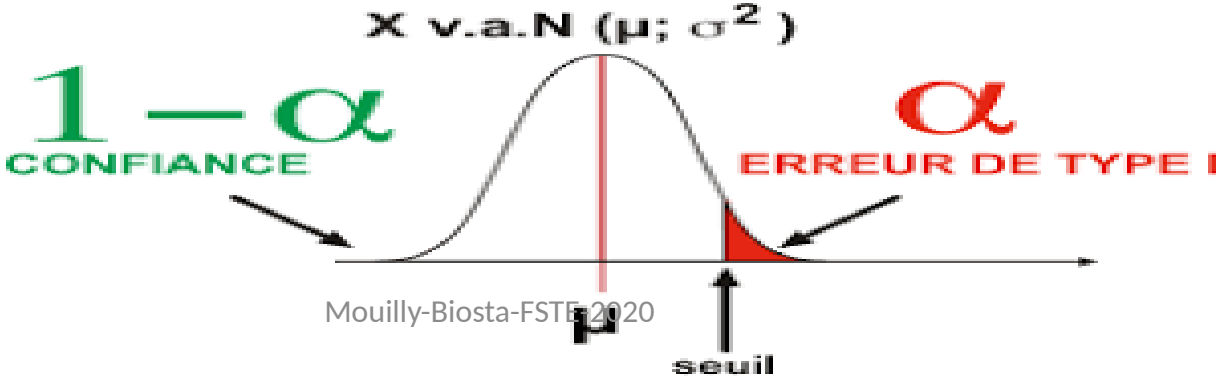
Pour définir cet intervalle de confiance, nous devons d'abord déterminer quels sont les risques d'erreurs que nous pourrions accepter.

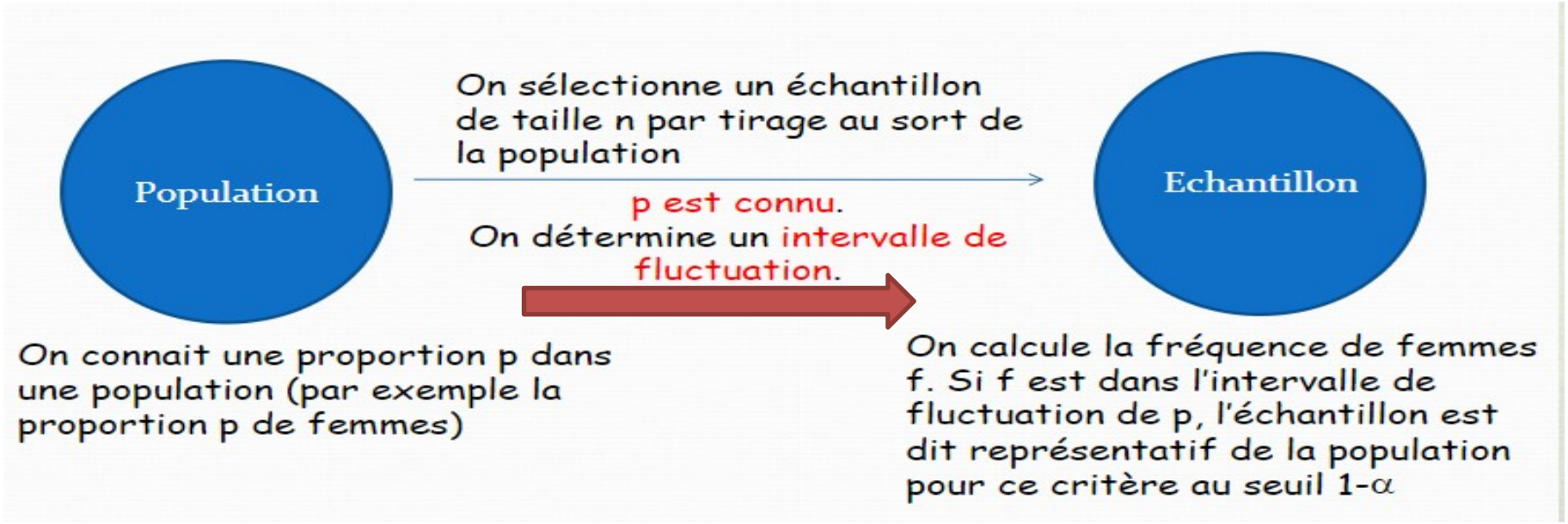
- **Classiquement ce risque d'erreur est fixé arbitrairement et les valeurs les plus courantes sont 5%, 1% ou 0,1%…**

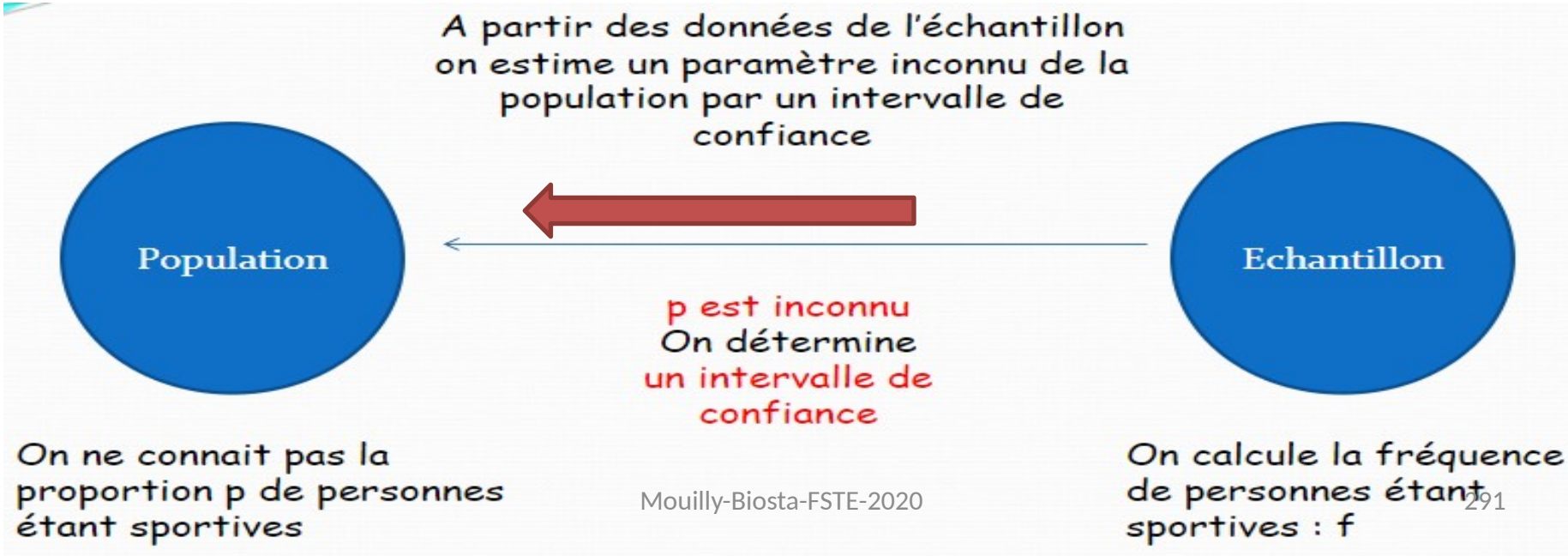
La confiance : noté $1-\alpha$

- **La confiance (1-α) étant centrée, l'erreur α se répartit de part et d'autre : α/2 à gauche, et α/2 à droite (Loi Centrée réduite)**

Les valeurs les plus courants sont : 95%, 99%, 99,9%….




Population
On sélectionne un échantillon de taille n par tirage au sort de la population
Echantillon
p est connu.
On détermine un intervalle de fluctuation.
On connait une proportion p dans une population (par exemple la proportion p de femmes)
On calcule la fréquence de femmes f. Si f est dans l'intervalle de fluctuation de p, l'échantillon est dit représentatif de la population pour ce critère au seuil 1-α


A partir des données de l'échantillon on estime un paramètre inconnu de la population par un intervalle de confiance
Population
Echantillon
p est inconnu
On détermine
un intervalle de confiance
On ne connait pas la proportion p de personnes étant sportives
On calcule la fréquence de personnes étant sportives : f

Partie 3 : Statistique Inférentielle

1. Les Principes de L'Inférence Statistique
2. Les méthodes d'échantillonnage
3. Principaux paramètres de L'Inférence Statistique
4. Les méthodes relatives aux moyennes
5. Les méthodes relatives aux proportions
6. Taille d'échantillon

3. Les méthodes relatives aux moyennes

Condition de validité : $n \geq 30$

σ^2 est connue

$$ME = z_{\frac{\alpha}{2}}.\sigma_{\bar{X}}$$

avec $\sigma_{\bar{X}}$ = erreur standard

Si remise ou sans remise et $N \geq 20n$ ou $\frac{n}{N} \leq 0{,}05$: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Si sans remise et $N < 20n$ ou $\frac{n}{N} > 0{,}05$: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

σ^2 est inconnue

$$ME = z_{\frac{\alpha}{2}}.\sigma_{\bar{X}}$$

avec $\sigma_{\bar{X}}$ = erreur standard

$$\mu = \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Erreur Standard = ES
Marge Erreur= ME
Intervalle de Confiance=IC

Si remise ou sans remise et $N \geq 20n$ ou $\frac{n}{N} \leq 0{,}05$: $\sigma_{\bar{X}} = \frac{S}{\sqrt{n}}$

Si sans remise et $N < 20n$ ou $\frac{n}{N} > 0{,}05$: $\sigma_{\bar{X}} = \frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

Si: $n < 30$

$$ME = k.\sigma_{\bar{X}}$$

avec k = inégalité de chebyshev où $k^2 = \frac{1}{\alpha}$

Si remise : $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Si sans remise : $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

$$ME = t_{\frac{\alpha}{2},dl}.\sigma_{\bar{X}} \quad avec \quad dl = n-1$$

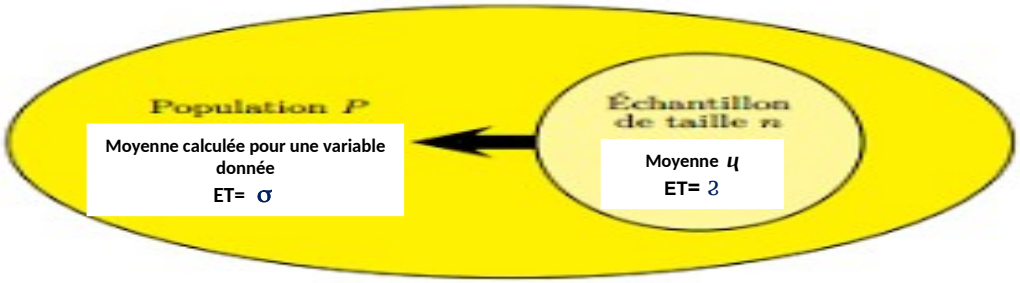
avec $\sigma_{\bar{X}}$ = erreur standard

Si remise ou sans remise et $N \geq 20n$ ou $\frac{n}{N} \leq 0{,}05$: $\sigma_{\bar{X}} = \frac{S}{\sqrt{n}}$

Si sans remise et $N < 20n$ ou $\frac{n}{N} > 0{,}05$: $\sigma_{\bar{X}} = \frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

- Fr ⊔ ES=ESM ET Ang ⊔ SE=SEM A ne pas confondre avec ET= σ = SD
- ddl désigne le nombre deVA qui ne peuvent être déterminées par une équation ou formule statistique
- $z_{a/2}$ est la valeur de z qui correspond à une surface de $\alpha/2$ sous la queue supérieure de la distribution de la loi normale centrale réduit ⊔ Coefficient Critique

α =10% ⊔ $IC_{90\%}$ ⊔ $z_{a/2} = z_{0,25=}$ 1,645 α =5% ⊔ $IC_{95\%}$ ⊔ $z_{a/2} = z_{0,25=}$ 1,960 α =1% ⊔ $IC_{99\%}$ ⊔ $z_{a/2} = z_{0,05}$ = 2,576



On veut estimer la taille moyenne des étudiantes à la FSM. Dans cette optique, nous avons tiré d'une façon aléatoire un échantillon de 40 étudiants.

La moyenne observée est de l'ordre de 169 cm avec un ET estimé à 6,16 cm.

1. Est-ce que les conditions de normalité sont remplies ?

Effectif est grand ⊔ n ≥ 30

2. Quelle est l'$IC_{95\%}$ de la taille des étudiants de la FSM?

Au risque de 5% l'$IC_{95\%}$ ⊔ ±1,91 ⊔ [167,1 cm; 170,1 cm].

Cela signifie que l'estimation de la taille moyenne « vraie » des étudiantes de cette Faculté se situe à 95% entre 167,1 & 170,1cm

3. Quelle est l'$IC_{99\%}$ la taille des étudiants de la FSM?

Au risque de 1% l'$IC_{99\%}$ ⊔ ± 2,51⊔ [166,5 cm; 171,5 cm].

Cela signifie qu'à 99%, on ai moins précis dans notre estimation, car l'amplitude de l'IC augmente, ainsi que la marge d'erreur.

Partie 3 : Statistique Inférentielle

1. Les Principes de L'Inférence Statistique
2. Les méthodes d'échantillonnage
3. Principaux paramètres de L'Inférence Statistique
4. Les méthodes relatives aux moyennes
5. Les méthodes relatives aux proportions
6. Taille d'échantillon

4. Les méthodes relatives aux Proportions

Condition de validité : $n \geq 30$ et $n^* p \geq 5$, $n * (1 - p) \geq 5$

$$ME = z_{\frac{\alpha}{2}} . \sigma_{\bar{P}}$$

avec $\sigma_{\bar{P}} =$ erreur standard

Si remise ou sans remise et $N \geq 20n$ ou $\frac{n}{N} \leq 0{,}05$: $\sigma_{\bar{P}} = \sqrt{\frac{\bar{P}(1-\bar{P})}{n}}$

Si sans remise et $N < 20n$ ou $\frac{n}{N} > 0{,}05$: $\sigma_{\bar{P}} = \sqrt{\frac{\bar{P}(1-\bar{P})}{n}} \sqrt{\frac{N-n}{N-1}}$

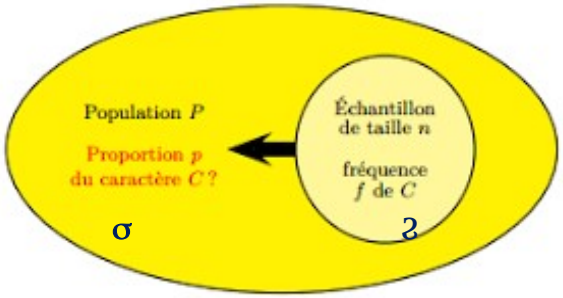
***NB.* Pour les petits échantillons loi Student ne s'applique pas**

$$\pi = P \pm z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}}$$

Erreur Standard

Marge d'Erreur

Intervalle de Confiance



Exemple : $n = 100,\ \alpha = 0{,}05,\ p = 0{,}12$

$$IC_{0,95} = \left[0{,}12 - 1{,}96\sqrt{\frac{0{,}12 \times 0{,}88}{100}}\ ;\ 0{,}12 + 1{,}96\sqrt{\frac{0{,}12 \times 0{,}88}{100}}\right] = [0{,}06\ ;\ 0{,}18]$$

Condition de validité :

n ≥ 30 ; n* p ≥ 5; n * (1 - p) ≥ 5

Dans notre cas : n ≥ 30 ; 100 x 0,12 = 12 ≥ 5 ; 100 x 0,88 = 88≥ 5 ▭ OK

Niveau de confiance C	Niveau de risque α	$z_{a/2}$
90%	10%	1,645
95%	5%	1,960
99%	1%	2,576

•En étudiant la fréquence de l'apparition de la grippe à Midelt pendant l'hiver, on ait obtenu une fréquence observée *p égale à 0,12*. Supposons que cette valeur ait été obtenue sur la base de 300 individus.

1. Est-ce que conditions de validité sont remplies ?

d'abord, on établit p qui est = 0,12;

l'effectif est bien n ≥ 30; n * p = **36** & n * (1- p) = **264** ≥ 5

2. Quelle est l'$IC_{95\%}$ de la fréquence d'apparition de cette maladie ?

Au risque de 5% l'$IC_{95\%}$= [0,08 ; 0,16]; p =0,12 ± 0,04 (exactement 0,03677286)

Cela signifie que cette valeur observée de 12 % sur les 300 individus nous fait qu'indiquer ceci : la fréquence « vraie » se situe dans la fourchette de : 8% et16 %.

Supposons que cette même valeur 12 % ait été obtenue sur la base de 100 individus. l'IC_{95}%= [0,06 ; 0,18]; p =0,12 ± 0,06

Sur la base de cette valeur 12 %, on est maintenant en mesure d'affirmer avec un risque d'erreur de 5%, que la fréquence « vraie » se situe dans le domaine 6 % & 18 % ⇨ Donc moins de précision

On cherche à estimer au niveau de la région Tafilalt le pourcentage de personnes diabétiques de type 2. Pour cela on effectue un sondage sur un échantillon représentatif de 1000 personnes. Nos résultats indiquent que 150 personnes sondés sont touchés par cette maladie.

1. Est-ce que conditions de validité sont remplies ?

d'abord, on établit p qui est = 0,15; l'effectif est bien n ≥ 30

n * p = **150** & n * (1- p) = **850** ≥ 5

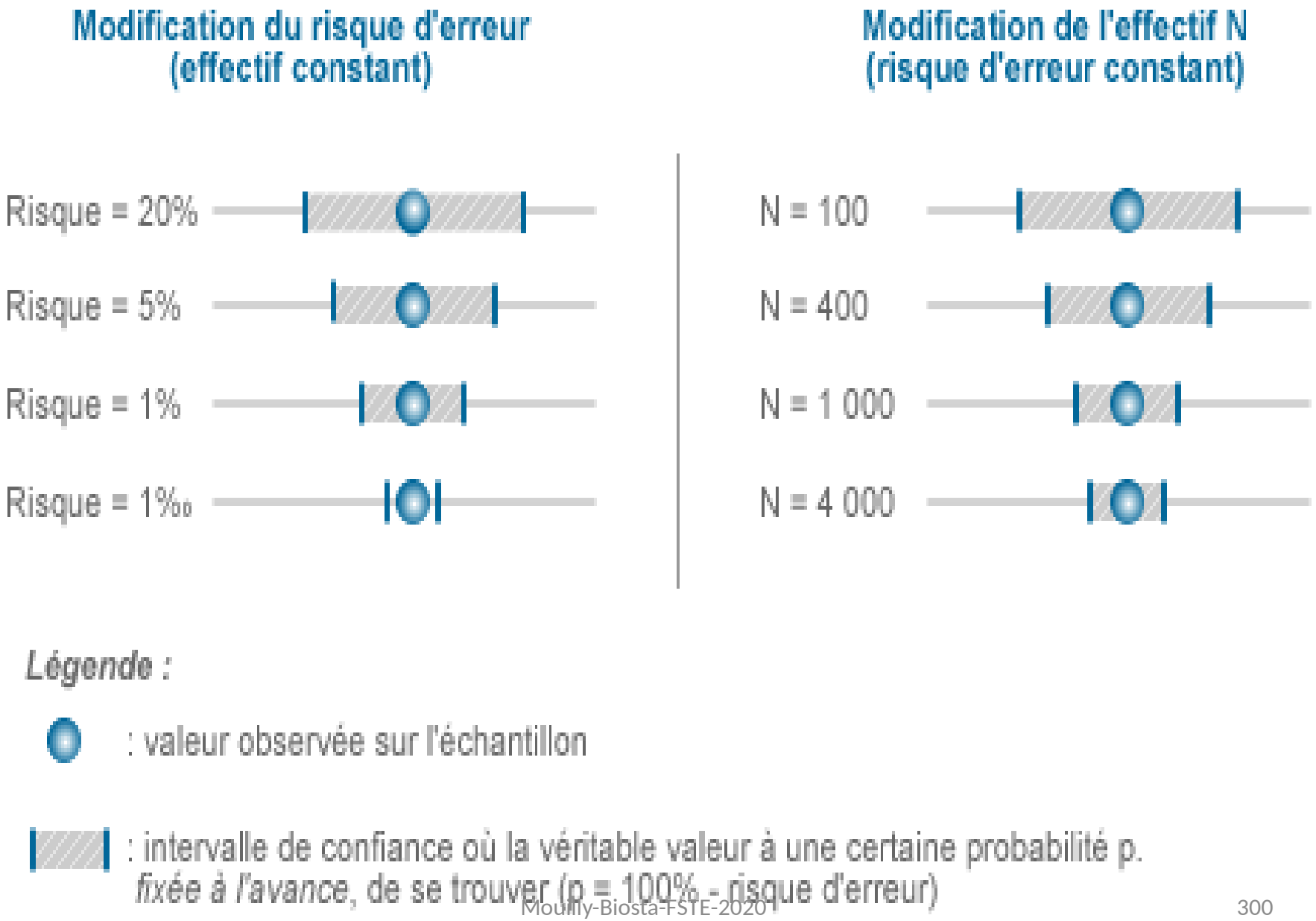
2. Quelle est l'intervalle de confiance à 95 % des individus diabétiques de type 2 lors de ce sondage ?

Au risque de 5% l'$IC_{95\%}$= [0,13; 0,17]; p =0,15 ± 0,02

3. Quelle est l'intervalle de confiance à 95 % dans le cas ou on va tripler le sondage initiale? ⇨ n = 3000

$IC_{95\%}$= [0,14; 0,16]; p =0,15 ± 0,01

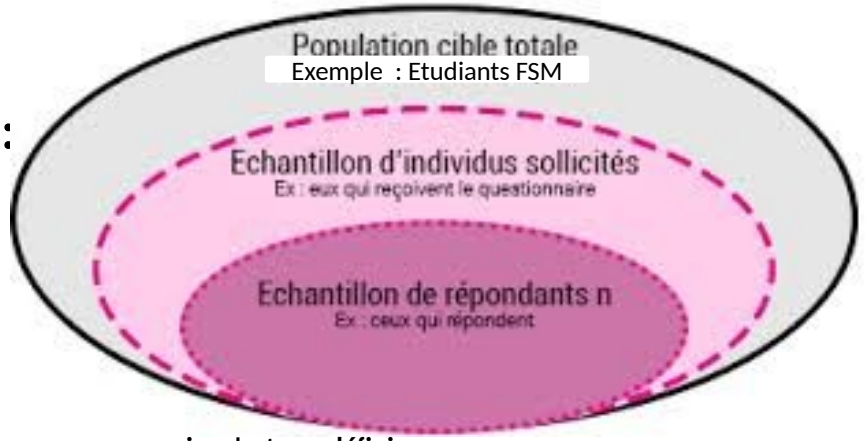
Nous avons une plus grande précision et IC & ME ont diminué


Modification du risque d'erreur
(effectif constant)
Risque = 20%
Risque = 5%
Risque = 1%
Risque = 1‰
Modification de l'effectif N
(risque d'erreur constant)
N = 100
N = 400
N = 1 000
N = 4 000
Légende :
: valeur observée sur l'échantillon
: intervalle de confiance où la véritable valeur à une certaine probabilité p.
fixée à l'avance, de se trouver (p = 100% - risque d'erreur)

Partie 3 : Statistique Inférentielle

1. Les Principes de L'Inférence Statistique
2. Les méthodes d'échantillonnage
3. Les méthodes relatives aux moyennes
4. Les méthodes relatives aux proportions
5. Taille d'échantillon

5 : Taille de l'échantillon



La Taille d'échantillon dépend de :

- **la taille de la population N**
- **la précision (seuil de confiance)**
- **la ME acceptée de l'estimation**

Les Proportions ⊟ **Si N EST CONNUE**

Dans le cas ou n est petit vs N

$$n' = \frac{N \cdot n}{N + n}$$

$$n = \frac{\frac{z^2_{\alpha/2} \times p(1-p)}{e^2}}{1 + \left(\frac{z^2_{\alpha/2} \times p(1-p)}{e^2 N}\right)}$$

si p n'est pas définie on prend par défaut 0,5 P(1-p)=0,25 ⊟ max

⊟ **Si N EST INCONNUE**

$$n = \frac{z^2_{\alpha/2} \times p(1-p)}{e^2}$$

Les moyennes ⊟ **N n'a pas d'influence**

Population

$$n = \left(\frac{Z_{\alpha/2}\, s}{e}\right)^2 \quad \left(\frac{Z_{\alpha/2}\, \sigma}{e}\right)^2$$

Echantillon

e **: la ME donnée en %** ⊟ **Les valeurs les plus courantes sont 1%, 2% et 5%**

NB. Plus que $Z_{\alpha/2}$ augmente plus que n augmente

Les Proportions

Un événement ayant une probabilité de réalisation de 40 %, en prenant un niveau de confiance de 95 % et une marge d'erreur de 5 %, la taille d'échantillon devra être :

$n = 1{,}96^2 \times 0{,}4 \times 0{,}6 / 0{,}05^2 = 368{,}79$

soit 369 individus.

N, ME et α sont fixes

N	p	α	e	n
100 000	1%	5%	3,00%	42
100 000	5%	5%	3,00%	203
100 000	10%	5%	3,00%	384
100 000	15%	5%	3,00%	543
100 000	20%	5%	3,00%	682
100 000	30%	5%	3,00%	894
100 000	40%	5%	3,00%	1022
100 000	45%	5%	3,00%	1054
100 000	50%	5%	3,00%	1064

maximum

Si N est connue et p inconnue, = 0,5 par défaut

	Confidence level = 95%			Confidence level = 99%		
	Margin of error			Margin of error		
Population size	5%	2,5%	1%	5%	2,5%	1%
100	80	94	99	87	96	99
500	217	377	475	285	421	485
1.000	278	606	906	399	727	943
10.000	370	1.332	4.899	622	2.098	6.239
100.000	383	1.513	8.762	659	2.585	14.227
500.000	384	1.532	9.423	663	2.640	16.055
1.000.000	384	1.534	9.512	663	2.647	16.317

Les Moyennes

En prenant un niveau de confiance de 99 % et une marge d'erreur de 5%, quelle est la taille nécessaire d'un échantillon d'étudiantes de la FSM pris au hasard pour faire une étude sur l'IMC. Sachant que σ=1,3 kg.m^{-1}

n = ???

$n = (1{,}96 \times 1{,}3)^2 / 0{,}05^2 = 2596{,}9216$

$$\left[\frac{Z_{\alpha/2}\,\sigma}{e}\right]^2$$

Les Moyennes

En prenant un niveau de confiance de 95 % et une marge d'erreur de 5%, quelle est la taille nécessaire d'un échantillon d'étudiantes de la FSM pris au hasard pour faire une étude sur l'IMC. Sachant que σ=1,3 kg.m^{-1}

$$\left[\frac{Z_{\alpha/2}\,\sigma}{e}\right]^2$$

n = (1,96 × 1,3)²/ 0,05² = 2596,92

Si ME est seulement 2 %
n = (1,96 × 1,3)²/ 0,02² = 16230,1

Si niveau de confiance est 99 %
n = (2.576 × 1,3)²/ 0,05² = 3501,79

جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

كلية العلوم والتقنيات
FACULTÉ DES SCIENCES ET TECHNIQUES

Partie I

Probabilité & Distributions Théoriques

Partie II

Statistique Descriptive

Partie III

Statistique Inférentielle

Partie VI

Plan Expérimental/Tests Statistiques

Partie VI : Plan Expérimental/Tests Statistiques

1. **Plan expérimental**
2. **Méthodologie d'application d'un test**
3. **Tests d'hypothèses**
4. **Niveau de signification d'un test**

1 : Plan expérimental

La méthode expérimentale est la seule procédure qui permette de d'établir un lien entre des événements internes ou externes à l'individu et le comportement de ce dernier. Elle consiste à :

- ✓Isoler les variables censées influencer le comportement étudié.
- ✓Construire une situation expérimentale contrôlée le plus rigoureusement possible.
- ✓Tester des hypothèses formulées à partir d'un contexte théorique en les soumettant aux faits de manière à les infirmer ou les confirmer..

Construire un plan d'expérience équivaut à la mise en place d'une stratégie qui va permettre

- ✓de maximiser la probabilité de détecter les effets réels des VI sur la ou les VD
- ✓minimiser la probabilité que les conclusions tirées puissent être dues à l'influence de variables non contrôlées.

- ✓**la variable explicative (X=indépendante=Covariable=Exogène)**
- ✓**la variable expliquée (Y=variable dépendante=réponse=Endogène)**

- Le scientifique est souvent amené à comprendre comment réagit un système en fonction des facteurs susceptible de le modifier en faisant appel à des méthodes expérimentales scientifiques qui consistent à tester, par des expériences répétées, la validité d'une hypothèse. Pour visualiser cette évolution, il fixe les facteurs d'entrée, mesure une réponse et essaye par la suite d'établir des relations de cause à effet entre les réponses et les facteurs.

- Le cours de plans d'expériences décrit quelles sont les expériences à réaliser et comment répartir des traitements sur les unités expérimentales dans l'objectif d'obtenir un maximum d'information sur le phénomène étudié en un minimum d'expériences. Ceci est primordial si l'objectif est un gain de temps ou de qualité. Il décrit également l'exploitation des résultats obtenus par analyses statistiques.

- Par contre il n'insiste que peu sur le choix des facteurs expérimentaux, le domaine expérimental et les soins à apporter dans la conduite des essais.

Partie VI : Plan Expérimental/Tests Statistiques

1. Plan expérimental
2. Méthodologie d'application d'un test
3. Tests d'hypothèses
4. Niveau de signification d'un test

2 : Méthodologie d'application d'un test

- Le test utilisé doit être précisé avec le résultat
- Un test pour chaque situation est défini selon :

  - Objectif de l'analyse ? : Décrire, Résumer, Inférer….
  - Type de la variable :
    - quantitatif (continue, discrète)
    - qualitatif (nominale, binomiale, ordinale, texte)
  - Nombre de variable (univariée, bivariée, multivariée)
  - Taille de la population ou échantillon petit ou grand
  - Séries dépendants (liées?) ou indépendantes
  - Séries appariées ou non appariés
  - Mesures uniques ou répétées

Partie VI : Plan Expérimental/Tests Statistiques

1. Plan expérimental
2. Méthodologie d'application d'un test
3. Tests d'hypothèses
4. Niveau de signification d'un test

hypothesis testing إختبار الفرضيات

test d'hypothèses

Hypothesentest Prueba de hipótesis

ipotesi di prova

проверки гипотезы хипотеза тест

3 : Test d'Hypothèses

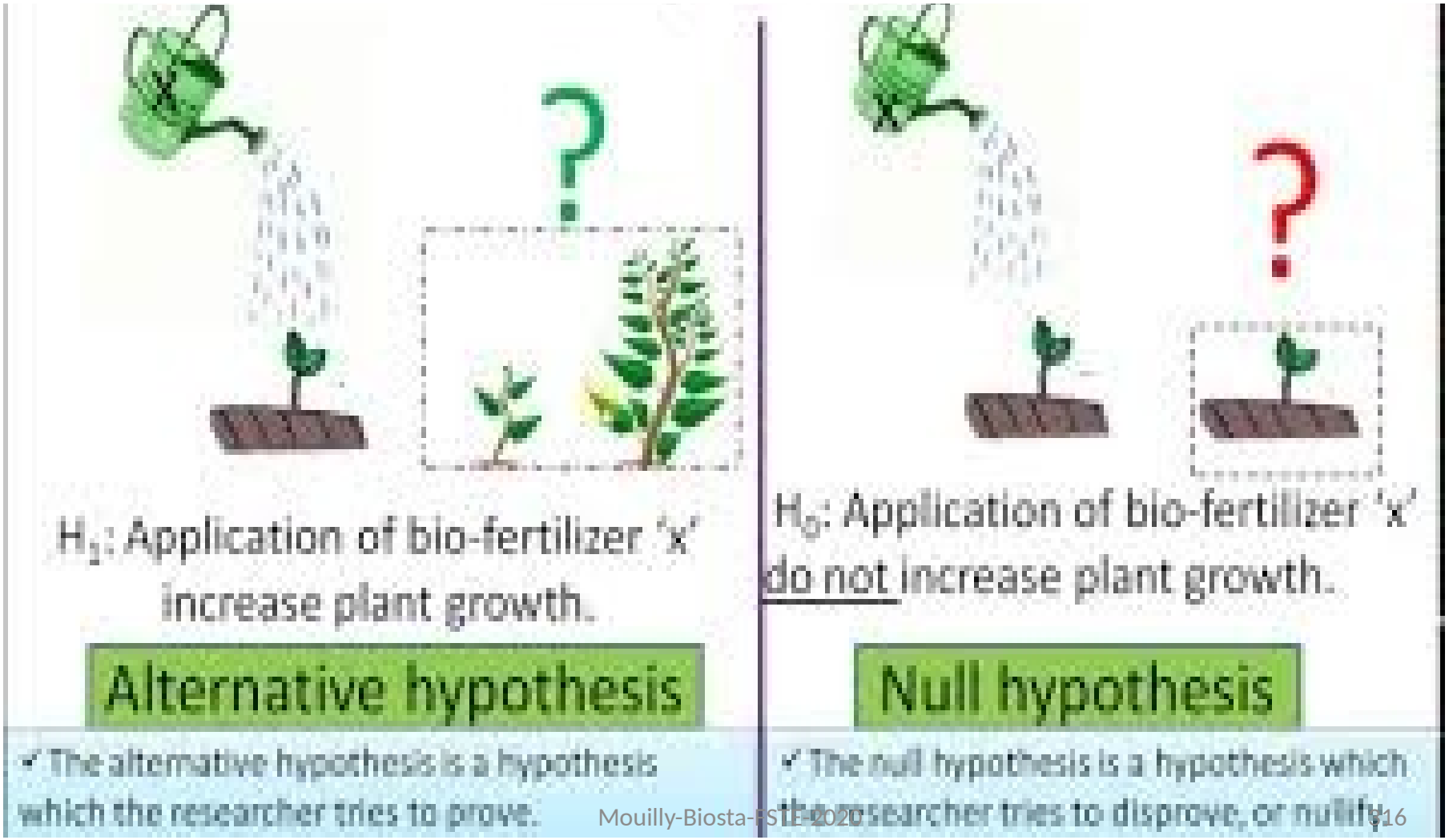
- ✓ Les hypothèses **nulle H_0** et **alternatives H_1** sont deux déclarations s'excluant mutuellement sur une population.
- ✓ Un test hypothétique utilise des données d'échantillon pour déterminer si l'hypothèse nulle peut être rejetée.
- ✓ L'hypothèse **nulle H_0** affirme qu'un paramètre de la population (la moyenne, l'écart type, taux, etc.) est égal à une valeur hypothétisée.
- ✓ L'hypothèse alternative **H_1** affirme qu'un paramètre de la population est plus petit, plus grand ou différent de la valeur hypothétisée dans l'hypothèse nulle. L'hypothèse alternative est celle que vous pensez être vraie ou que vous espérez démontrer.

Une hypothèse statistique peut-être :

- **Unilatérale signifie une seule possibilité**
- **Bilatérale signifie 2 possibilités**

Hypothèse statistique unilatérale	➧	H0 : A = B et H1 : A > B ou A < B
ou		ou
Hypothèse statistique bilatérale	➧	H0 : A = B et H1 : A ≠ B (A > B et A < B)

Exemple d'Hypothèse Unilatérale



Partie VI : Plan Expérimental/Tests Statistiques

1. **Plan expérimental**
2. **Méthodologie d'application d'un test**
3. **Tests d'hypothèses**
4. **Niveau de signification d'un test**

5 : Niveau de signification du test

Afin d'obtenir une règle de décision qui permette de choisir entre le rejet et le non rejet de l'hypothèse nulle, il faut fixer une valeur de la probabilité, qu'on appelle niveau de signification du test et qu'on désigne par **le symbole α**.

- ✓ si $p \leq \alpha$: on rejette H_0 au niveau de signification **α**
- ✓ si $p > \alpha$: on ne rejette pas H_0 au niveau de signification **α**
- ✓ La valeur généralement adoptée pour **α est 0.05**
- ✓ Fréquemment aussi, on considère trois niveaux (seuil de signification) : **0.05, 0.01 et 0.001** et, en fonction de p, on a les situations suivantes :
- ✓ si $p > 0.05$: la différence est non significative
- ✓ si $p \leq 0.05$: la différence est significative

identifiant	sex	tabagisme	age	StatutP	Santé	residence	NGI20
id1	1	oui	18	oui	m	privee	11
id2	1	oui	19	oui	m	privee	4
id3	0	oui	18	oui	m	privee	10
id4	0	oui	22	oui	b	privee	8
id5	0	oui	27	non	mo	cite	6
id6	1	oui	18	oui	m	cite	4
id7	1	oui	23	oui	m	cite	10
id8	1	oui	20	oui	mo	privee	8
id9	0	oui	22	oui	m	cite	10
id10	0	non	19	non	b	cite	11
id11	0	non			b	cite	14
id12	0	non			m	privee	14
id13	0	oui			m	cite	19
id14	0	oui			mo	cite	16
id15	0	oui			b	privee	13
id16	1	non			mo	privee	12
id17	0	non			b	privee	14
id18	0	non			b		15
id19	1	non			b	privee	10
id20	0	non			b		12
id21	0	non			b	cite	12
id22	0	non	18	non	b	cite	13
id23	0	non	19	non	b	cite	18
id24	1	non	18	non		cite	15
id25	0	non	22	non	m	cite	19
id26	0	non	27	non	m		5
id27	0			non	m	privee	13
id28	0	oui	23	oui		cite	13
id29	1	non	20	non	b		12
id30	1	non	22	non	m	cite	14
id31	1	non	19	non	m	cite	10
id32	1	non	19	non	mo	cite	12
Check in	32	32	31	31	30	28	319 32
	0	0	1	1	0	2	4 0

Principaux Tests statistiques
Champs d'application : Biologie

Tests paramétriques : postulats sur la famille de distribution dont est tiré l'échantillon

tests non paramétriques : certaines variables échappent à ces hypothèses de par leur nature (qualitatives nominales ou ordinales) : transformation en variables ordinales

=> perte d'information = de puissance
=> gain de robustesse (moins d'influence des valeurs extremes)

Principaux Tests statistiques
Champs d'application : Biologie

Tests paramétriques
et de leurs équivalents non paramétriques

• **Test paramétrique**	**Test non paramétrique**
• Test t de Student non apparié	Test de Mann et Whitney
• Test t de Student apparié	Test signé de Wilcoxon
• Analyse de variance	Test de Kruskall et Wallis
• Corrélation linéaire	Test de Spearman
• Chi² de Pearson	Chi² de Mac Nemar

Principaux Tests statistiques
Champs d'application : Biologie

- **Etude de deux variables quantitatives**
  - Comparaison de deux moyennes : Test Student apparié, non apparié

  NB. Régression linéaire simple

  Modélisation

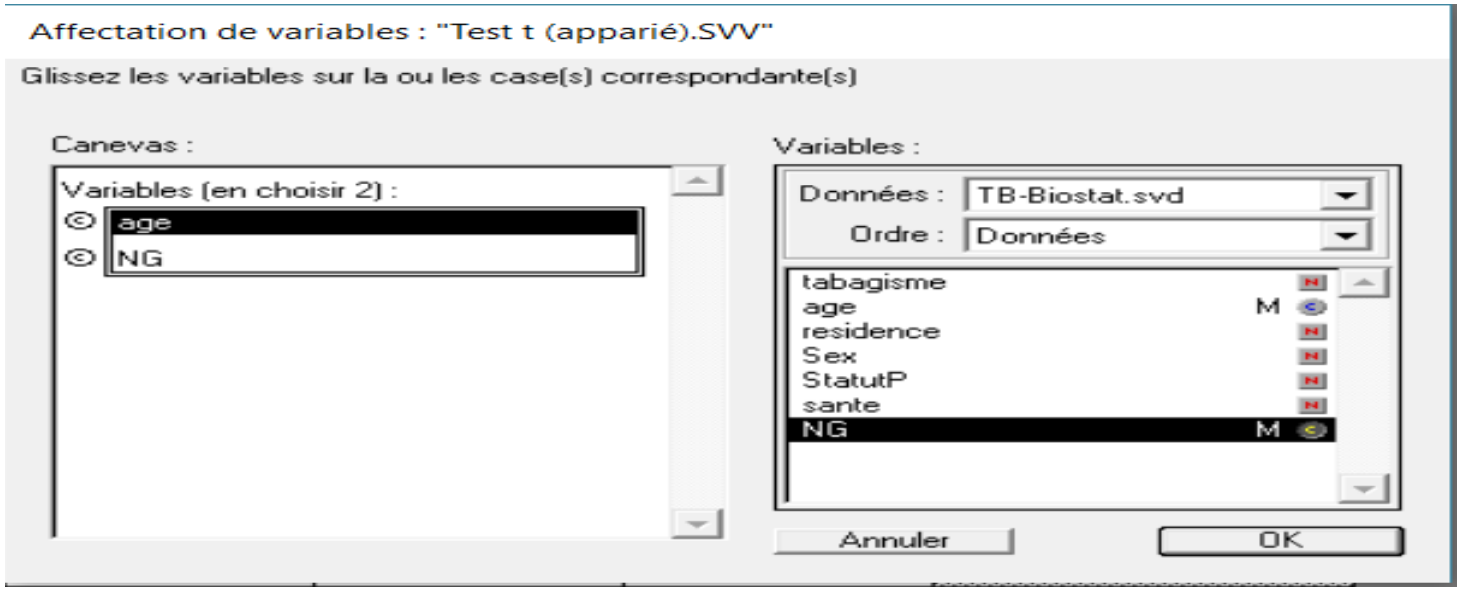
- **Etude d'une quantitative et une variable qualitative**
  - Analyse de la variance (ANOVA)-Uni-Bi-variée

- **Etude de deux variables qualitatives**
  - Test du Khi 2

TEST T de STUDENT

W.Gosset (1876-1937)

- Conditions : Normalité ou n>30



Test-t séries appariées
Ecart théorique = 0

	Ecart moyen	DDL	t	p
age, NG	9,581	30	9,818	<,0001

Principaux Tests statistiques
Champs d'application : Biologie

- **Etude de deux variables quantitatives**
  - Comparaison de deux moyennes : Test Student apparié, non apparié
  - Régression linéaire simple

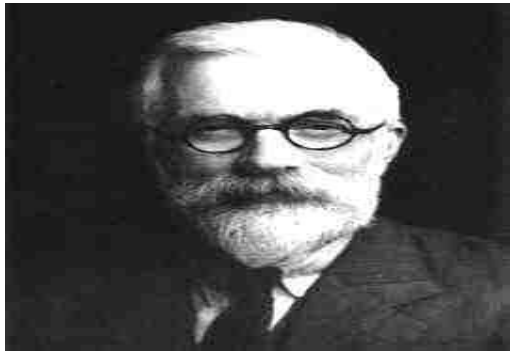
Modélisation

- **Etude d'une quantitative et une variable qualitative**
  - Analyse de la variance (ANOVA)-Uni-variée, Bi-variéee

- **Etude de deux variables qualitatives**
  - Test du Khi 2

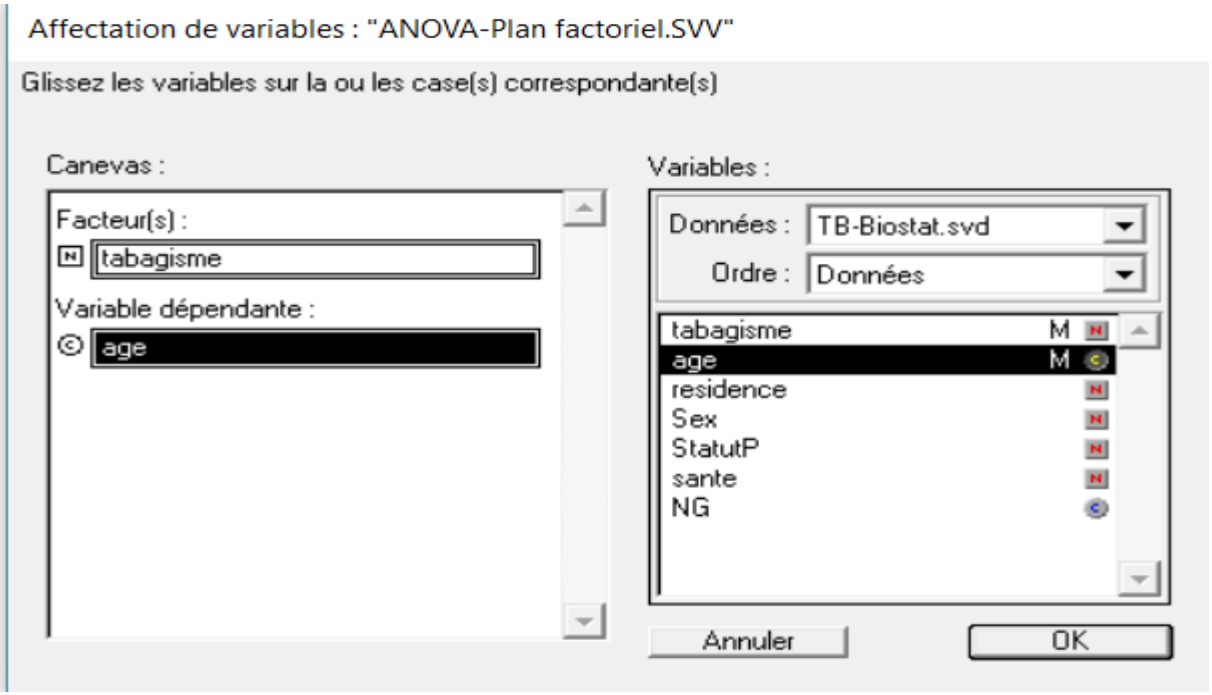
ANOVA "ANalysis Of VAriance"

R. FISHER (1890-1962)



Conditions :

- Normalité ou n>30
- Indépendance
- Homogénéité des variances = homoscédasticité



ANOVA "ANalysis of VAriance"

R. FISHER (1890-1962)

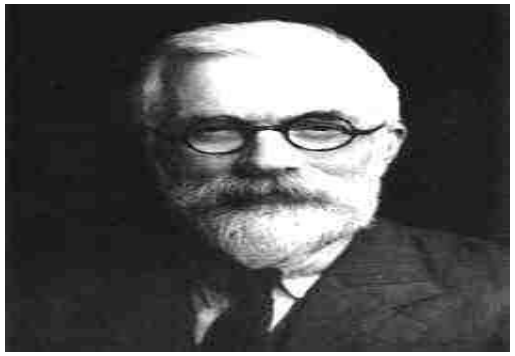


Tableau d'ANOVA pour age

	DDL	Somme des carrés	Carré moyen	Valeur de F	Valeur de p
tabagisme	1	,159	,159	,011	,9161
Résidus	27	380,668	14,099		

Modèle II estimation des composants de la variance : •
3 cas omis (manquants).

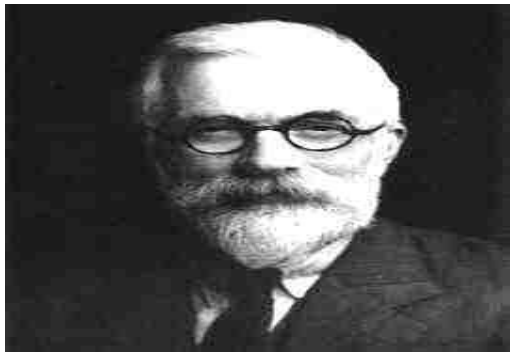
Tableau des Moy. pour age
Effets : tabagisme

	Nombre	Moy.	Dév. Std	Err. Std
non	16	21,313	4,094	1,024
oui	13	21,462	3,282	,910

3 cas omis (manquants).

ANOVA "ANalysis of VAriance"

R. FISHER (1890-1962)



Affectation de variables : "ANOVA-Plan factoriel.SVV"

Glissez les variables sur la ou les case(s) correspondante(s)

Canevas :

Facteur(s) :

tabagisme

Variable dépendante :

NG

Variables :

Données : TB-Biostat.svd

Ordre : Données

tabagisme M
age
residence
Sex
StatutP
sante
NG M

Annuler OK

ANOVA "ANalysis of VAriance"

R. FISHER (1890-1962)

Tableau d'ANOVA pour NG

	DDL	Somme des carrés	Carré moyen	Valeur de F	Valeur de p
tabagisme	1	47,978	47,978	3,847	,0599
Résidus	28	349,222	12,472		

Modèle II estimation des composants de la variance : 2,41
2 cas omis (manquants).

Tableau des Moy. pour NG
Effets : tabagisme

	Nombre	Moy.	Dév. Std	Err. Std
non	17	12,706	2,710	,657
oui	13	10,154	4,394	1,219

2 cas omis (manquants).

ANOVA "ANalysis Of VAriance"

R. FISHER (1890-1962)

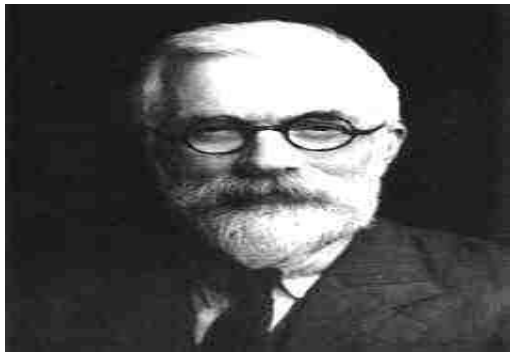


Tableau d'ANOVA pour NG

	DDL	Somme des carrés	Carré moyen	Valeur de F	Valeur de p
StatutP	1	66,452	66,452	5,812	,0230
Résidus	27	308,721	11,434		

Modèle II estimation des composants de la variance : 3,911
3 cas omis (manquants).

Tableau des Moy. pour NG
Effets : StatutP

	Nombre	Moy.	Dév. Std	Err. Std
non	17	12,824	2,157	,523
oui	12	9,750	4,615	1,332

3 cas omis (manquants).

ANOVA "ANalysis of VAriance"

Post Hoc

Le rejet de H0 signifie qu'il y a une grande probabilité qu'au moins il y a une différence entre les groupes. L'analyse Post Hoc est nécessaire pour nous indiquer ou se situe la différence entre la ou les moyens.

One-way Independent ANOVA

If $p < 0.05$ (significant, ie H_o rejected), then must do Post Hoc test (multiple pairwise comparison test)

Post Hoc Tests

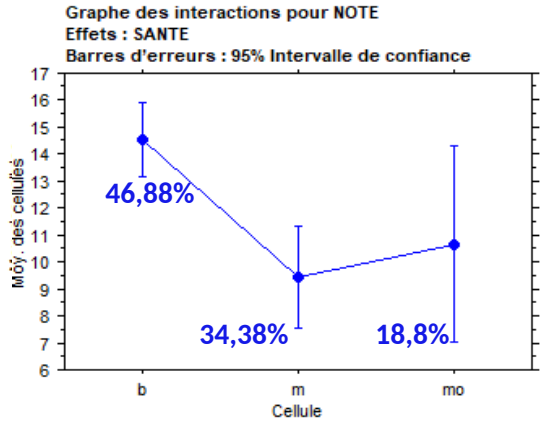
- **Tukey Test** — *If homogenous, No control*
- **Dunnette Test** — *If not homogenous, Has control*
- **Bonferroni Test** — *For repeated measure*

There also exists other types of tests : Sheffe , PLSD of Fisher,....

Exemple Post Hoc-ANOVA : Effet de l'état de santé (3 modalités : b, m, mo) sur la note

Statistiques descriptives

	NOTE
Moy.	12,063
Dév. Std	3,609
Erreur Std	,638
Nombre	32
Minimum	4,000
Maximum	19,000
Variance	13,028
Coef. Var.	,299
Etendue	15,000
Somme	386,000
Som. Carrés	5060,000
Aplat.	,459
Médiane	12,000
Interquartile	4,000
Mode	12,000
10% Moy. élaguée	12,192
DAM	2,000



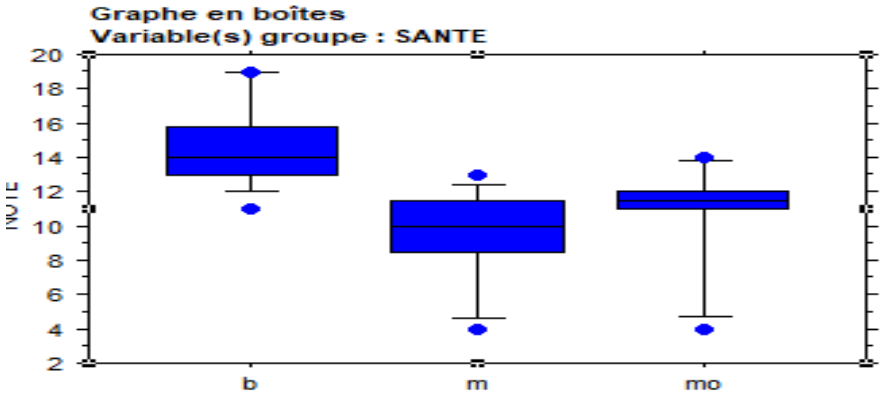


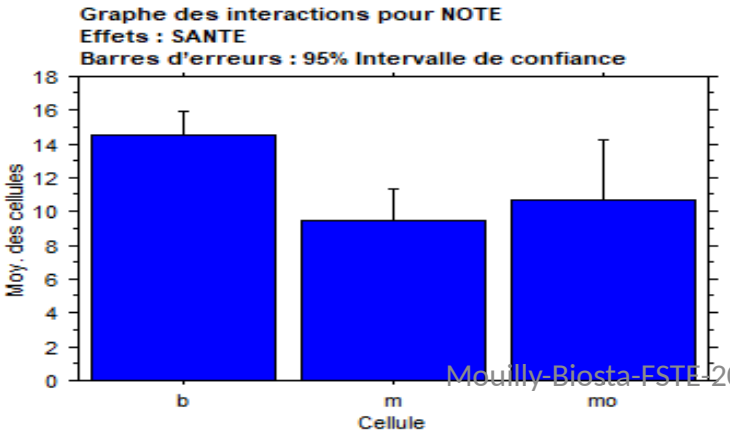
Tableau d'ANOVA pour NOTE

	DDL	Somme des carrés	Carré moyen	Valeur de F	Valeur de p
SANTE	2	178,081	89,041	11,436	,0002
Résidus	29	225,794	7,786		

Modèle II estimation des composants de la variance : 8,1

Tableau des Moy. pour NOTE
Effets : SANTE

	Nombre	Moy.	Dév. Std	Err. Std
b	15	14,533	2,503	,646
m	11	9,455	2,806	,846
mo	6	10,667	3,445	1,406



PLSD de Fisher pour NOTE
Effets : SANTE
Niveau de signif. 5 %

	Ecart moyen	Ecart critique	Valeur de p	
b, m	5,08	2,27	<,0001	S
b, mo	3,87	2,76	,0076	S
m, mo	-1,21	2,90	,3991	

Scheffe pour NOTE
Effets : SANTE
Niveau de signif. 5 %

	Ecart moyen	Ecart critique	Valeur de p	
b, m	5,079	2,857	,0004	S
b, mo	3,867	3,477	,0267	S
m, mo	-1,212	3,653	,6965	

Bonferroni/Dunn pour NOTE
Effets : SANTE
Niveau de signif. 5 %

	Ecart moyen	Ecart critique	Valeur de p	
b, m	5,079	2,814	<,0001	S
b, mo	3,867	3,425	,0076	S
m, mo	-1,212	3,598	,3991	

Principaux Tests statistiques
Champs d'application : Biologie

- **Etude de deux variables quantitatives**
  - Comparaison de deux moyennes : Test Student apparié, non apparié
  - Régression linéaire simple

    Modélisation

- **Etude d'une quantitative et une variable qualitative**
  - Analyse de la variance (ANOVA)-Univariée

- **Etude de deux variables qualitatives**
  - Test du Khi 2

χ^2 = Khi2 = Chi2 "Chi square"
K. Pearson (1857 - 1936)

Comparaison possibles

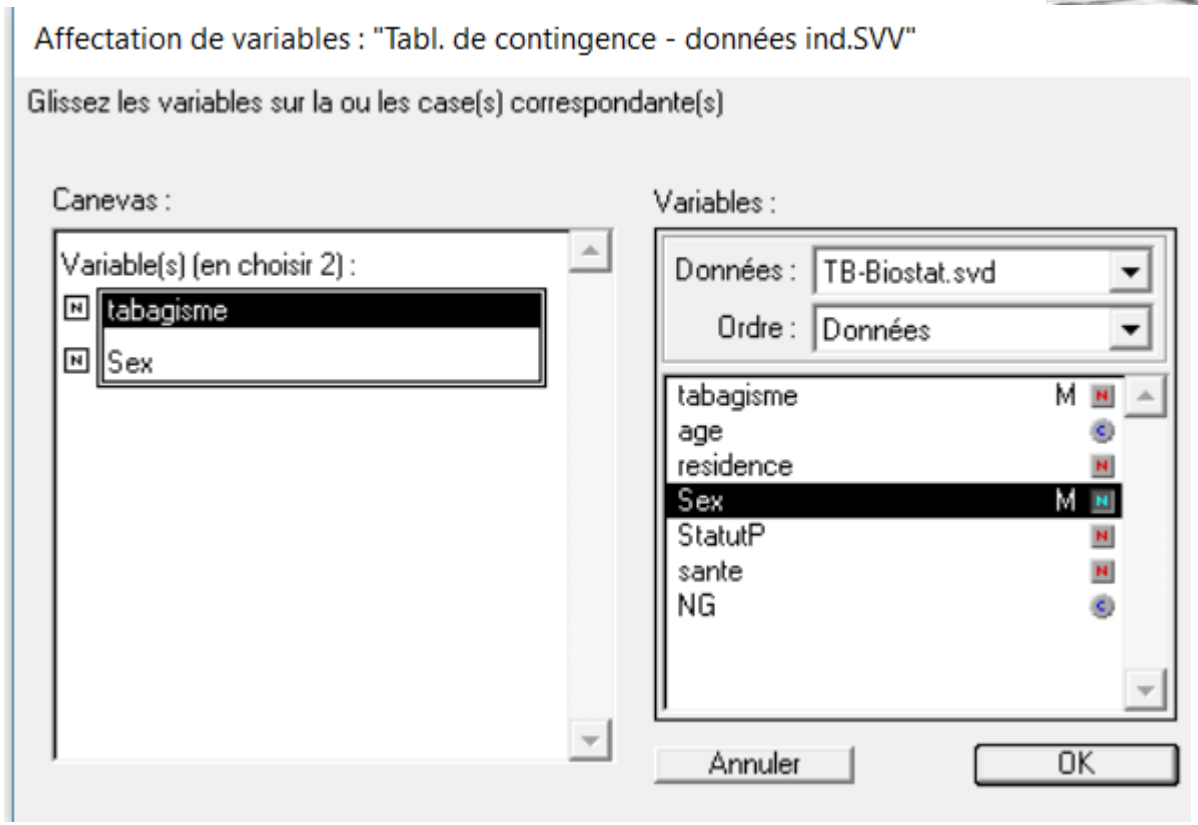
- De 2 pourcentages :
  - ✓ Comparaison d'un pourcentage observé à une valeur théorique (test d'adéquation)
  - ✓ Comparaison de 2 pourcentages observés : (test d'indépendance)

- Conditions : Tous les effectifs C_{ij} sont sup. à 5

- Cas des petits échantillons :
  - ✓ 3< l'un des effectifs C_{ij} <5_{ij} **Test corrigé Yates**
  - ✓ l'un des effectifs C_{ij} est inf à 3 **Test exact de Fisher**

χ^2 = Khi2 = Chi2 "Chi square"

K. Pearson (1857 - 1936)



χ^2 = Khi2 = Chi2 "Chi square"

K. Pearson (1857 - 1936)

Tableau "résumé" pour tabagisme, Sex

Manquants	4
DDL	1
Chi 2	,050
p (Chi 2)	,8232
G-carré	,050
p (G-carré)	,8233
Coef. de contingence	,042
Phi	,042
Chi 2 corrigé	0,000
p corrigé	>,9999
Prob. exacte de Fisher	>,9999

Fréquences observ. pour tabagisme, Sex

	f	m	Totaux
non	6	10	16
oui	5	7	12
Totaux	11	17	28