Chapitre 6 : Statistiques

Plan du cours

A. Statistique descriptive
B. Exercices
C. Réflexions

A. Statistique descriptive

1. Introduction à la statistique descriptive

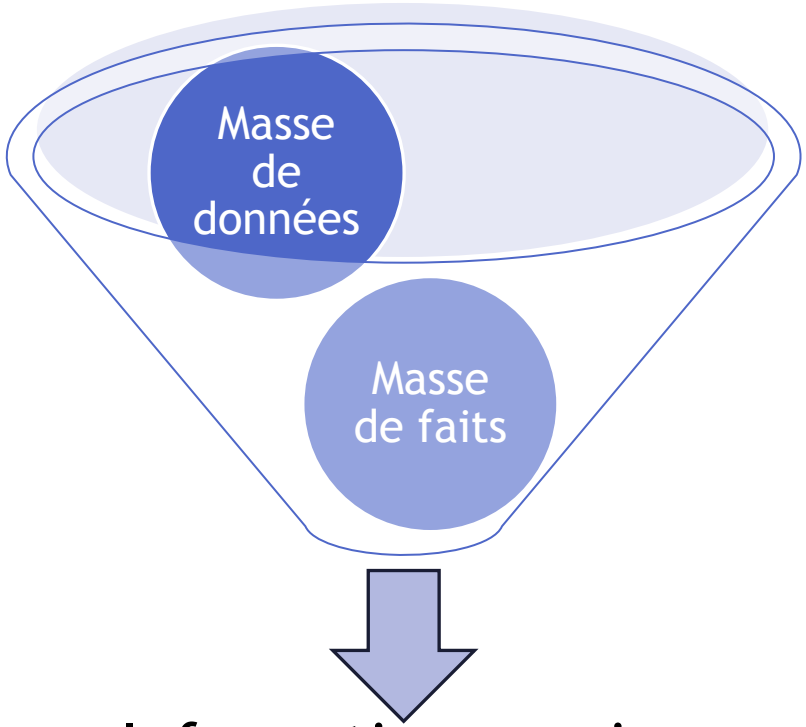
2. Définitions, classification variables & notations

3. Analyse descriptive d'une variable

1. Introduction à la statistique descriptive
2. Définitions, classification variables & notations
3. Analyse descriptive d'une variable

1. Introduction

Statistique descriptive



Information concise facilement utilisable

Statistiques démographiques
- Pyramide des âges
- Répartition hommes / femmes

…

Statistiques financières
- Indices des cours des actions
- Volume des transactions

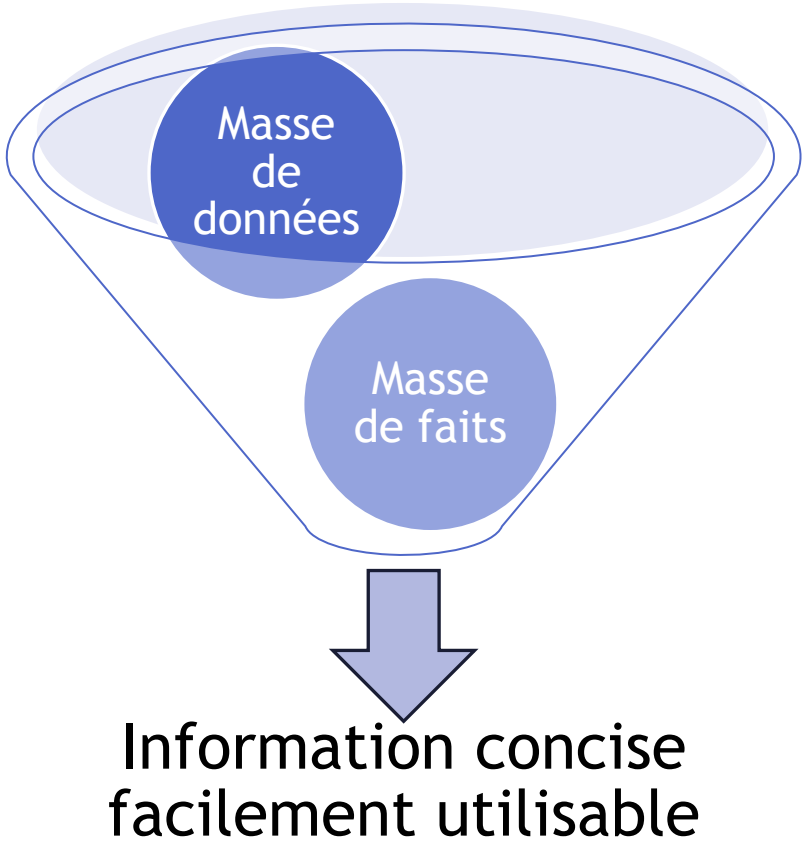
…

Statistiques économiques
- Prix à la consommation
- Indice du pouvoir d'achat

…

Entreprises
- Prévisions des ventes…

Statistique descriptive



- Synthétiser un grand ensemble de données;
- Les représenter graphiquement;
- Les analyser;
- Les présenter à d'autres.

Objectif

Le but de la statistique descriptive est de :

Représenter et résumer

utilement, objectivement et clairement

les informations disponibles dans un grand ensemble de données

sous la forme de tableaux, de graphiques et/ou de mesures numériques.

1. Introduction à la statistique descriptive
2. Définitions, classification variables & notations
3. Analyse descriptive d'une variable

2. Définitions, classification variables & notations

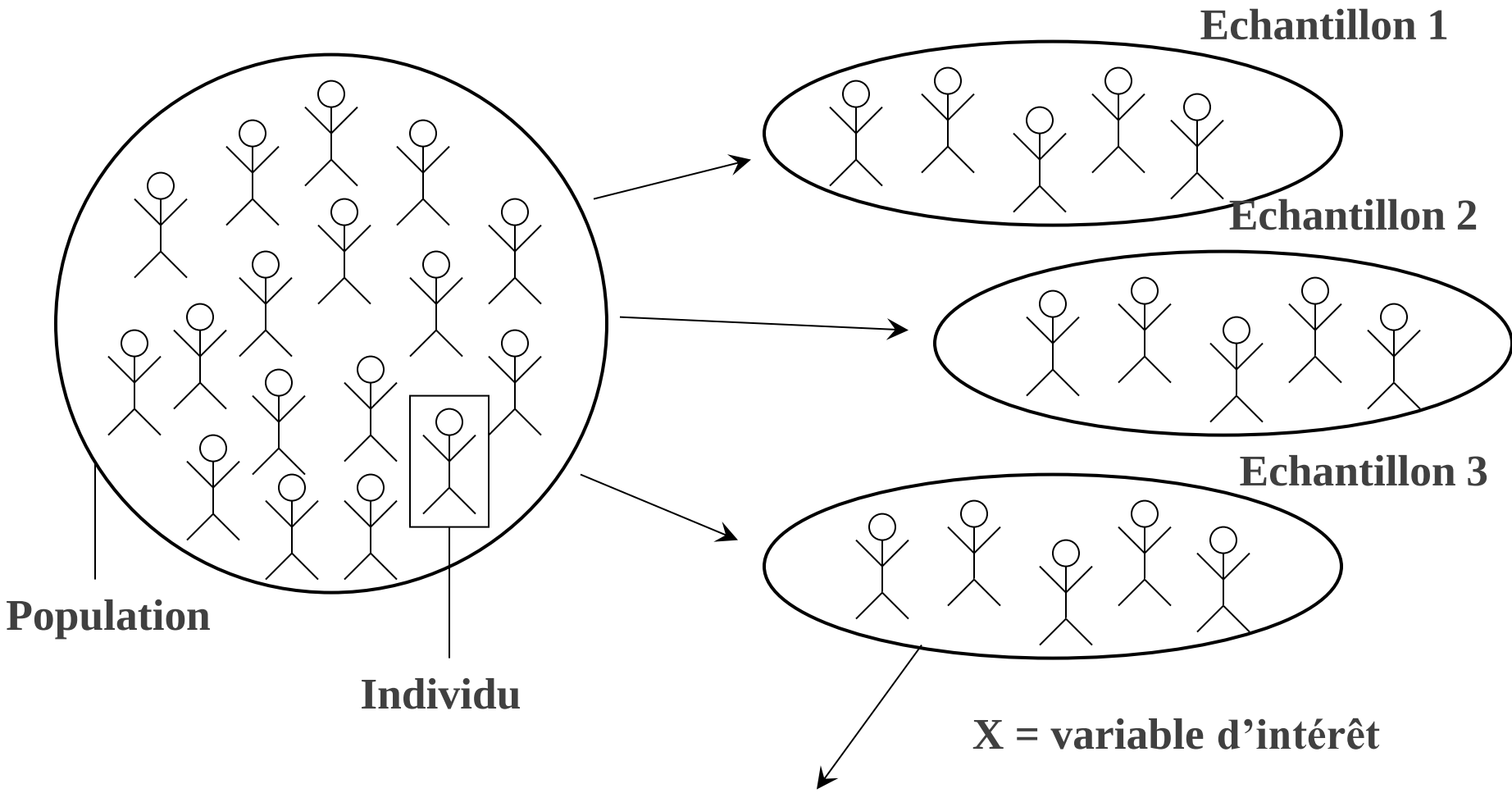
Définitions (1)

- **Population** : ensemble de toutes les personnes ou tous les faits sur lesquels porte l'étude
- **Individu** : chaque élément d'une population
- **Echantillon** : sous-ensemble de la population sur lequel on effectue l'étude

Définitions (2)

- **Variable statistique** : qualité, attribut ou caractéristique de la population à laquelle on s'intéresse
- **Observation** : valeur d'une variable pour un individu donné
- **Série statistique** : ensemble d'observations d'une variable sur un ensemble d'individus

Résumé



B. Frank, diapositive du cours de statistiques en 1BIM

Exemple

A l'aube des élections de mai 2014 en Belgique, la RTBF a effectué un sondage afin de connaître les intentions de vote en Wallonie.

Population :

Individu :

Echantillon :

Variable statistique :

Observation :

Série statistique :

Exemple

A l'aube des élections de mai 2014 en Belgique, la RTBF a effectué un sondage afin de connaître les intentions de vote en Wallonie.

Population : tous les électeurs wallons

Individu : chaque électeur wallon

Echantillon : 1000 électeurs wallons tirés au hasard

Variable statistique : l'intention de vote (parti)

Observation : l'intention de vote de Mr Dupont est: PS

Série statistique : PS, Ecolo, MR, MR, PS, MR, cdH, …

1. Introduction à la statistique descriptive
2. Définitions, classification variables & notations
3. Analyse descriptive d'une variable

Classification des variables

Type de variables

Variable **quantitative** si les valeurs qu'elle prend sont des nombres;

Variable **qualitative** sinon.

Variable qualitative (1)

Variable qui prend un nombre **fini** de valeurs possibles à caractère **qualitatif**

Exemple: Etat civil (célibataire, marié, divorcé, veuf, cohabitant)

Les valeurs possibles sont appelées **modalités** ou **niveaux** (parfois encore catégories).

Echelle de mesure :

- Ordinale s'il existe une relation d'ordre entre les catégories
- Nominale, sinon.

Variable qualitative (2)

Variable : Etat civil

- Modalités : célibataire, marié, divorcé, veuf, cohabitant
- ➔ Variable qualitative nominale

Variable : appréciation des services téléphoniques de Belgacom

- Modalités : très mauvais, mauvais, satisfaisant, bon, très bon
- ➔ Variable qualitative ordinale

Variables quantitatives(1)

Variable qui représente une **quantité** pouvant prendre un nombre fini ou infini de valeurs **numériques**.

Variable **discrète** : prend un nombre fini ou infini dénombrable de valeurs. Ces valeurs sont souvent des entiers.

Variable **continue** : prend un nombre infini non dénombrable de valeurs. Elle peut prendre n'importe laquelle des valeurs contenues dans un intervalle donné de nombres réels.

Variables quantitatives(2)

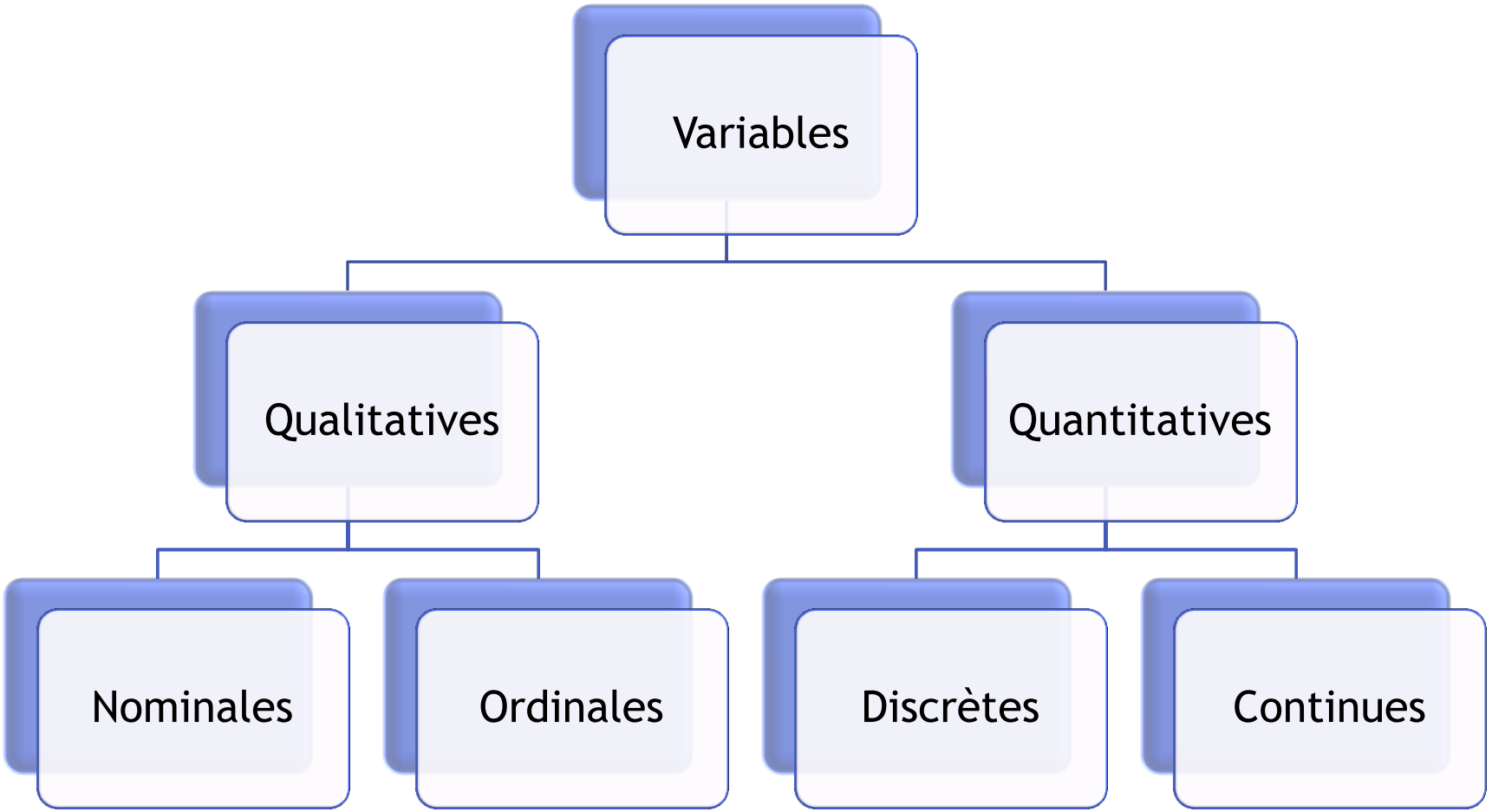
Exemples de variables quantitatives discrètes :

- Nombre d'enfants par famille
- Nombre de voitures par ménage
- Nombre d'étudiants en informatique en Europe

Exemples de variables quantitatives continues:

- Mesure du poids des enfants
- Mesure de la vitesse d'une voiture
- % de filles dans la filière informatique en Europe

Résumé



(ex: état civil) **(appréciation)** **(comptage)** **(appareil de mesure)**

Notations

Notations (1)

- X, Y, Z… : noms symboliques pour désigner une variable (on dira : soit X la variable poids…)
- N : nombre d'observations d'une série statistique, effectif total
- X_i : observation ou valeur de la variable X prise par l'individu i (i= 1,… N)
- x_i : niveau ou modalité d'une variable qualitative ou quantitative discrète
- $(X_1, X_2, X_3, \ldots X_N)$: série statistique pour la variable X

Notations (2)

Σ : signe de sommation

$$\sum_{i=1}^{N} X_i = X_1 + X_2 + X_3 + \ldots + X_N$$

1. Introduction à la statistique descriptive
2. Définitions, classification variables & notations
3. Analyse descriptive d'une variable

3. Analyse descriptive d'une variable

Analyse descriptive

- **Objectif** : résumer les caractéristiques d'**une** série statistique graphiquement ou par des chiffres
- Pour variables qualitatives :
  - Tableau de fréquences
- Pour variables quantitatives :
  - Tableau de fréquences
  - Tendance centrale
  - Dispersion
  - Distribution de la série

3.1. Analyse des variables qualitatives

Fréquence / effectif

- Soit une série statistique X : (X_1, X_2, … X_N)

- Fréquence ou effectif d'une modalité (x_i) : nombre d'individus de la population ayant cette modalité. Notation : n_i.

$$\sum_i n_i = N \text{ (effectif total)}$$

- Fréquence relative d'une modalité : effectif de la modalité divisé par l'effectif total. $f_i = n_i/N$. (entre 0 et 1)
  - Peut être exprimée en Pourcentage : fréquence relative * 100

Tableau de fréquences

Tableau regroupant, pour chaque modalité ou niveau d'une variable, l'effectif et la fréquence relative de la modalité.

Exemple : soit la variable X, groupe sanguin

Modalités : A, B, AB, O

Série statistique : (A, A, O, B, A, AB, O, …)

Exemple: tableau de fréquences pour le groupe sanguin

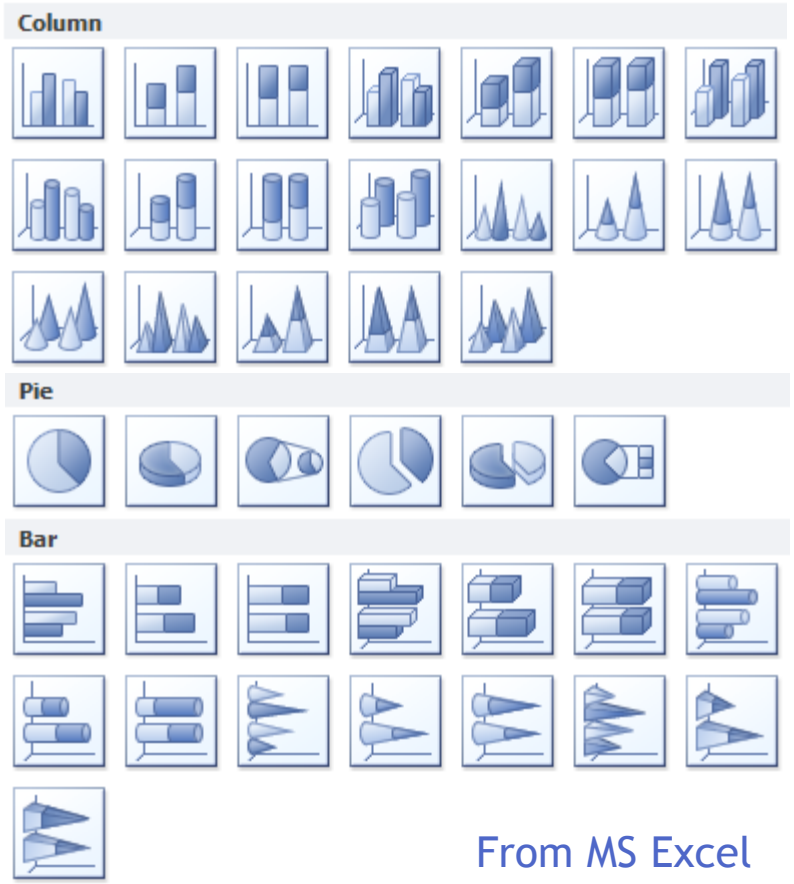
Groupe sanguin x_i	Nombre d'individus n_i	Fréquence relative $f_i=n_i/N$	Pourcentage d'individus
A	451	0,45	45%
B	79	0,08	8%
AB	32	0,03	3%
O	438	0,44	44%
	N=1000	1,00	100

Population : les belges

Echantillon : 1000 individus prélevés au hasard dans les listes des donneurs de la croix rouge

Représentation graphique

- Diagramme en colonnes (verticales ou horizontales)
- Diagramme à secteurs



From MS Excel

Exemple

«L'alimentation de l'étudiant en première année d'enseignement supérieur et vivant en kot»
A. Pierson

Représentez le fait de fumer auprès des étudiants de sexe masculin interrogés.

Codage :

Col 2	M	Masculin
	F	Féminin
Col 7	0	Non fumeur
	1	1 à 10 cig / jour
	2	11 à 20 cig / jour
	3	Plus de 20 cig / jour

Contexte : population, **taille** de l'échantillon, variable(s) statistique(s) étudiées

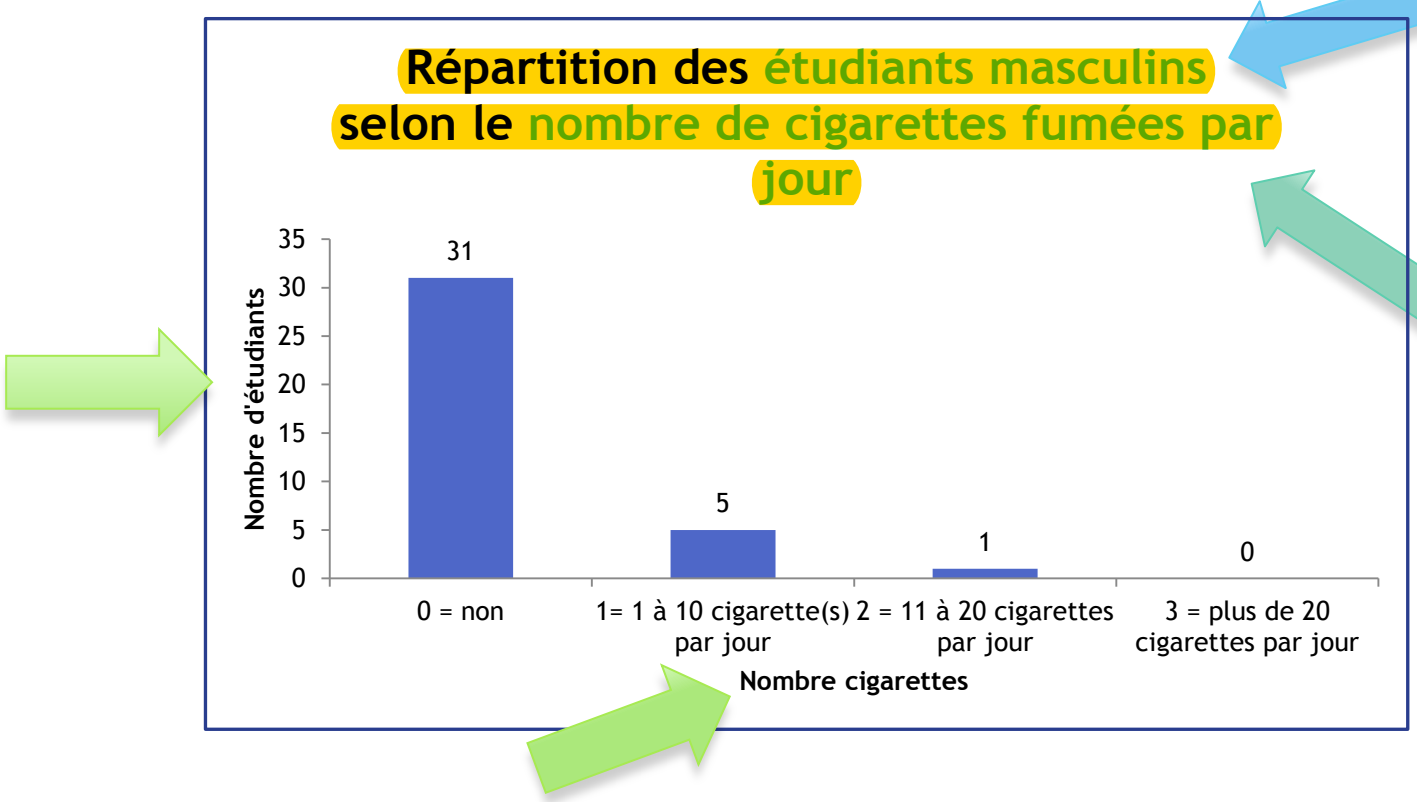
- L'enquête s'est déroulée auprès de la population étudiante en première année d'enseignement supérieur et vivant en kot.
- La population étudiée inclut les deux sexes et répond à une enquête concernant notamment la consommation de cigarettes.
- L'analyse porte sur 100 étudiants dont 37 de sexe masculin.

Tableau de distribution de fréquences

Répartition des étudiants masculins en fonction du nombre de cigarettes fumées par jour

Nombre de cigarettes	Nombre d'étudiants	Fréquence relative d'ét.	Pourcentage d'étudiants
Non fumeur	31	0,83	83%
1 à 10	5	0,14	14%
11 à 20	1	0,03	3%
Plus de 20	0	0,00	0%
Total	37	1,00	100%

Titre qui évoque le contenu du graphique :
+ titres sur les axes



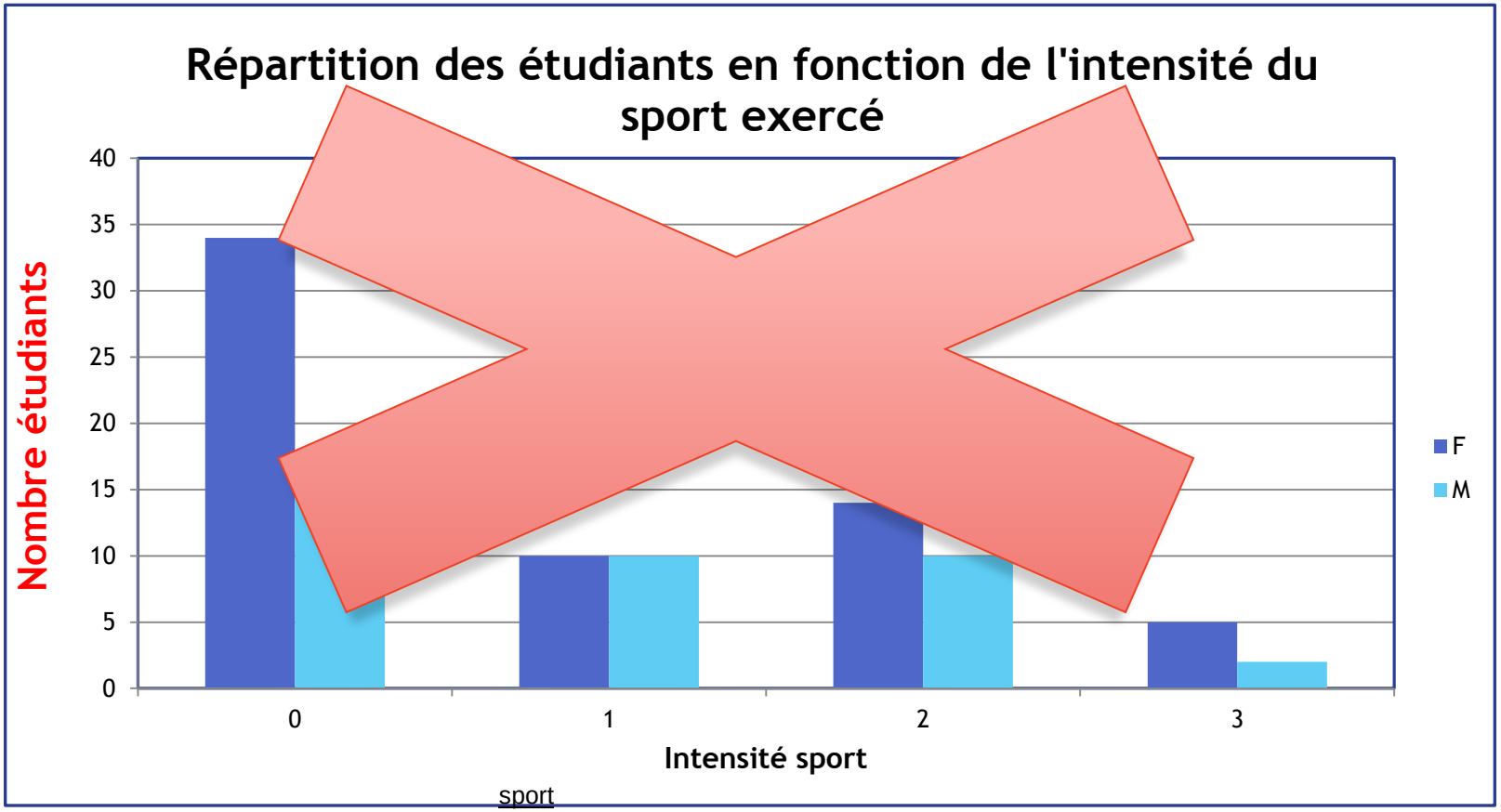
Individu (unité statistique)

Variable (d'intérêt)

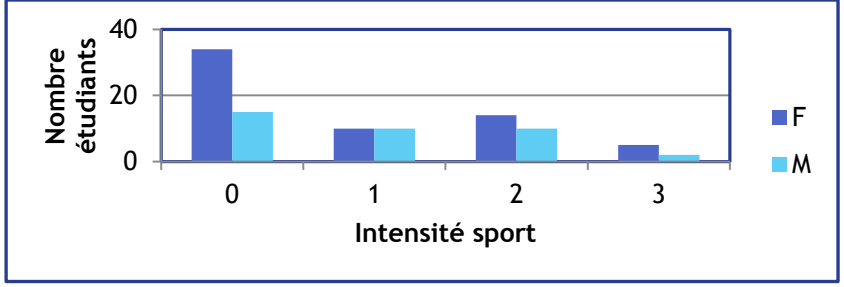
Présentation d’un graphique

Conseils

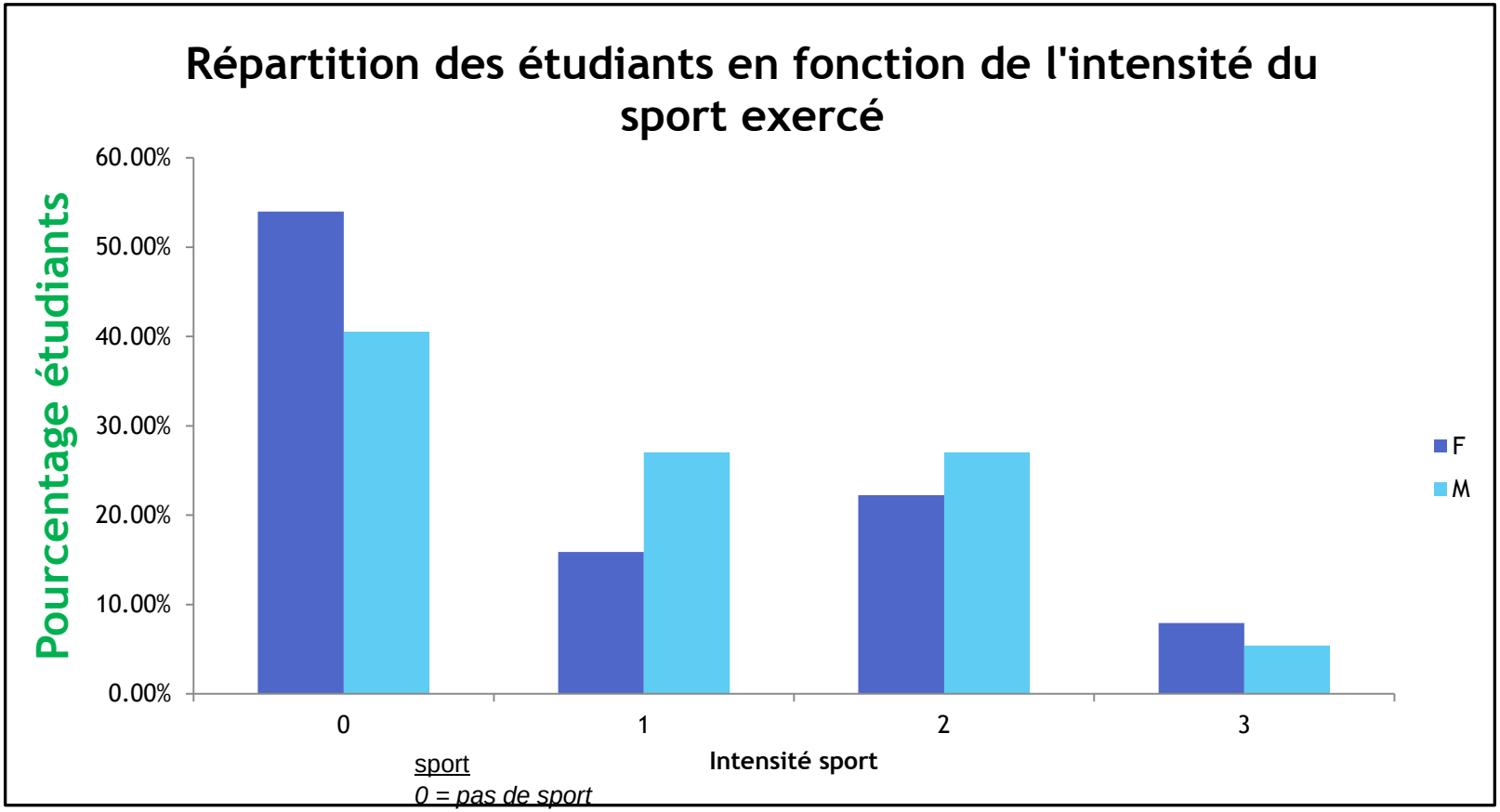
Attention on a interrogé beaucoup plus d'étudiantes que d'étudiants !



sport
0 = pas de sport
1 = faible
2 = modéré
3 = intense

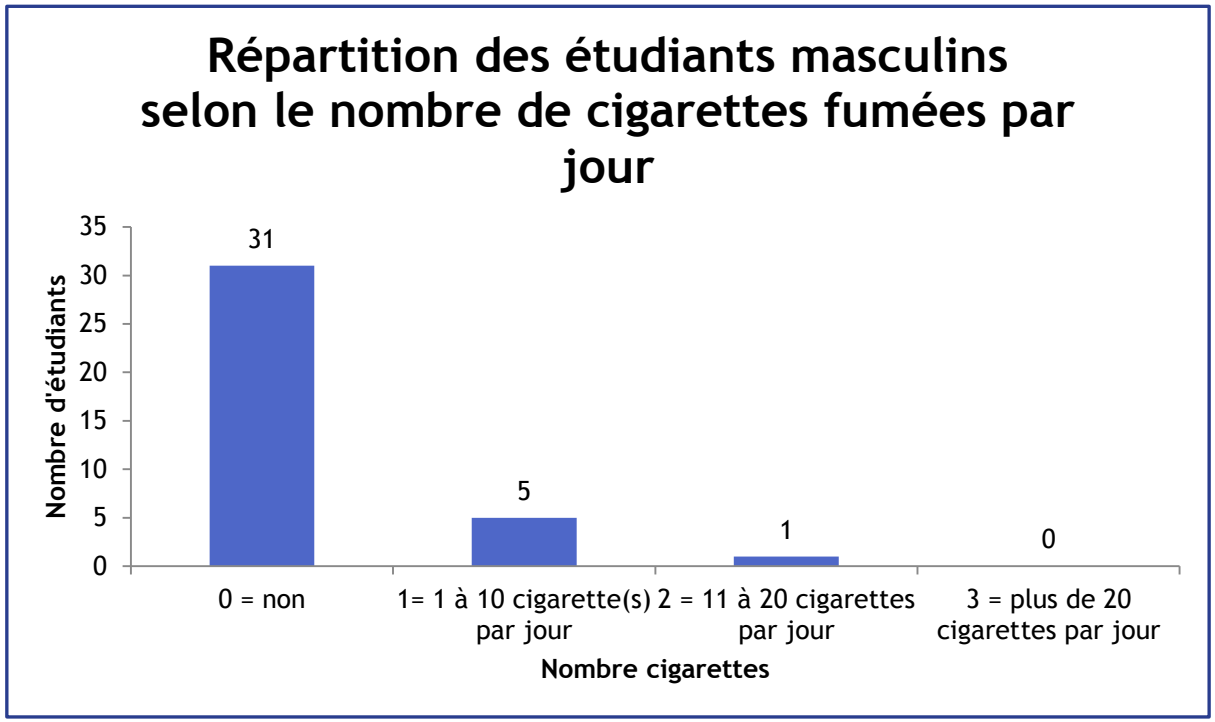


Les pourcentages sont indispensables pour les comparaisons !

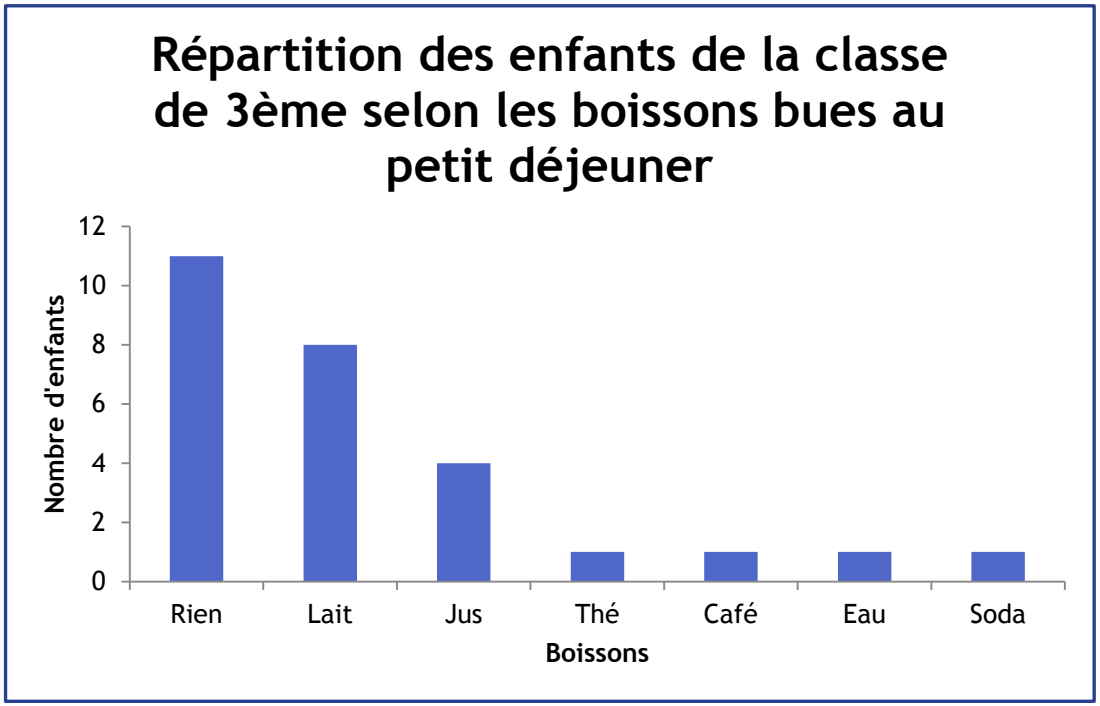


sport
0 = pas de sport
1 = faible
2 = modéré
3 = intense

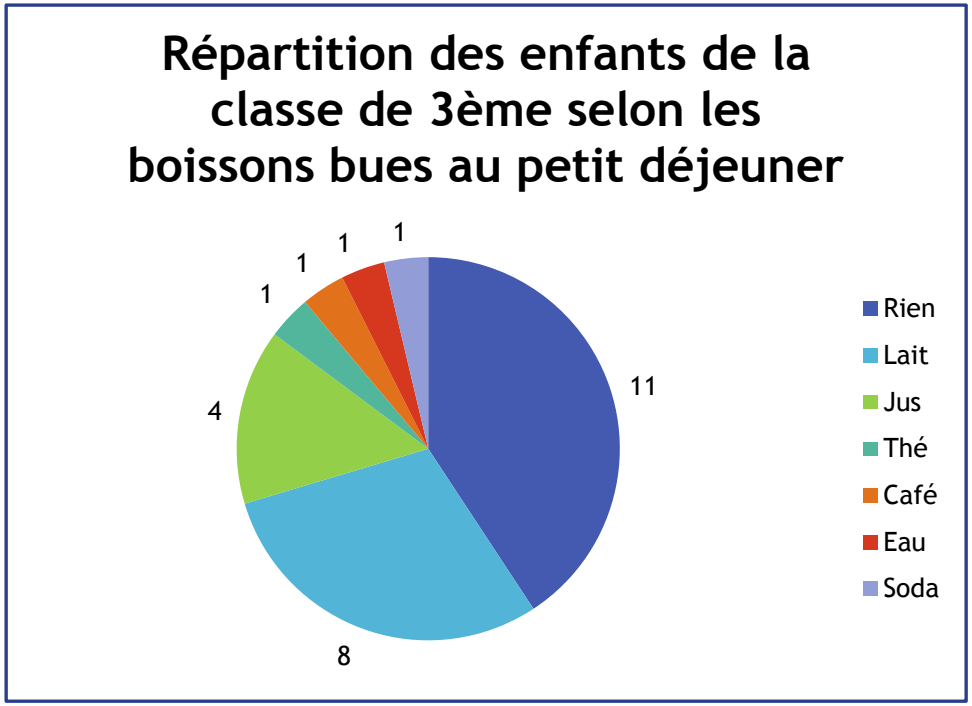
Si variable « ordonnée », on préfère souvent le diagramme à colonnes au diagramme à secteurs. L'ordre apparaît sur le diagramme à colonnes.



Si variable « non ordonnée » et nombre de modalités élevé, on préfère le diagramme à colonnes où celles-ci apparaissent triées par ordre décroissant.



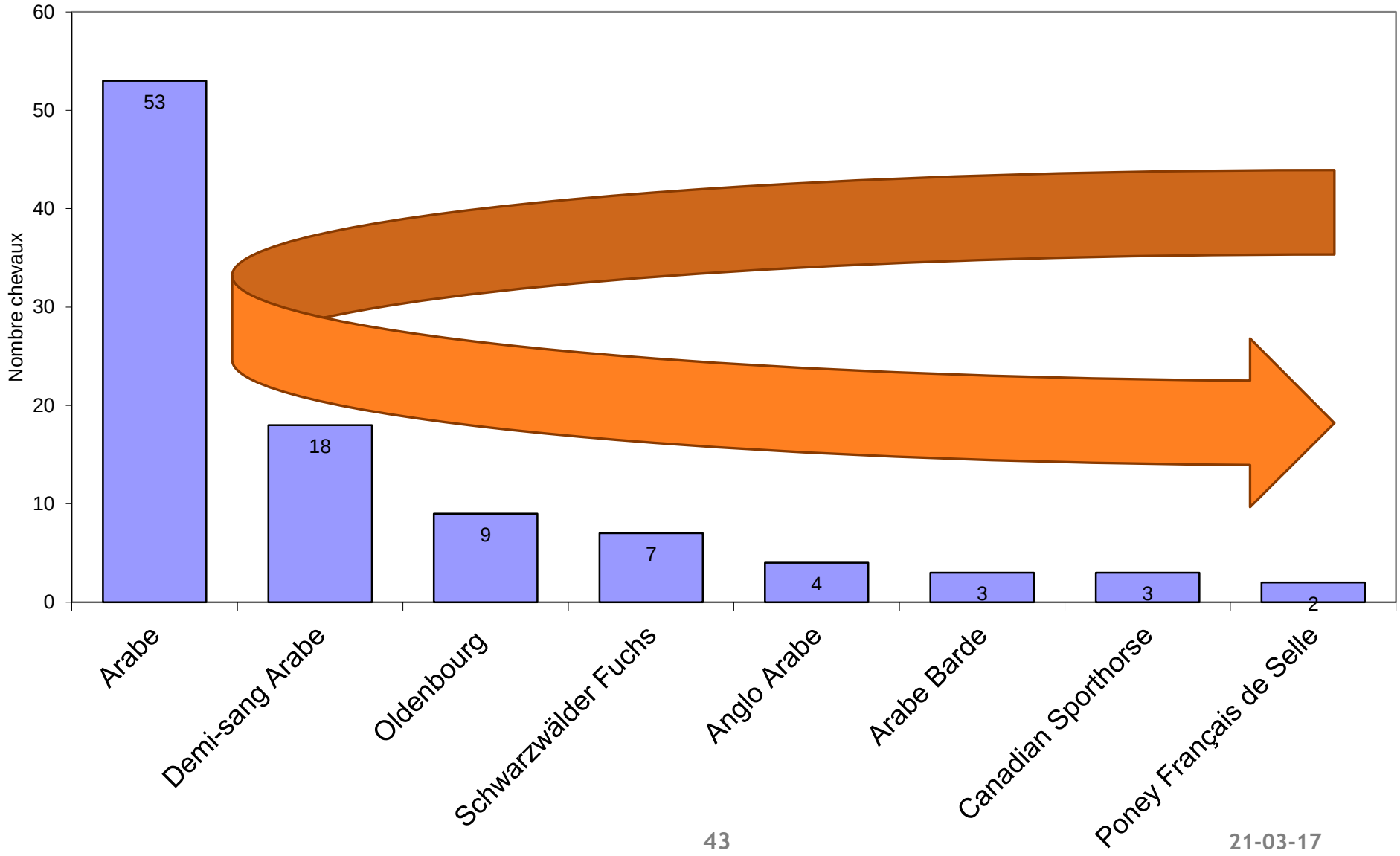
Il faut éviter un diagramme à secteurs avec trop de secteurs, ils deviennent rapidement illisibles



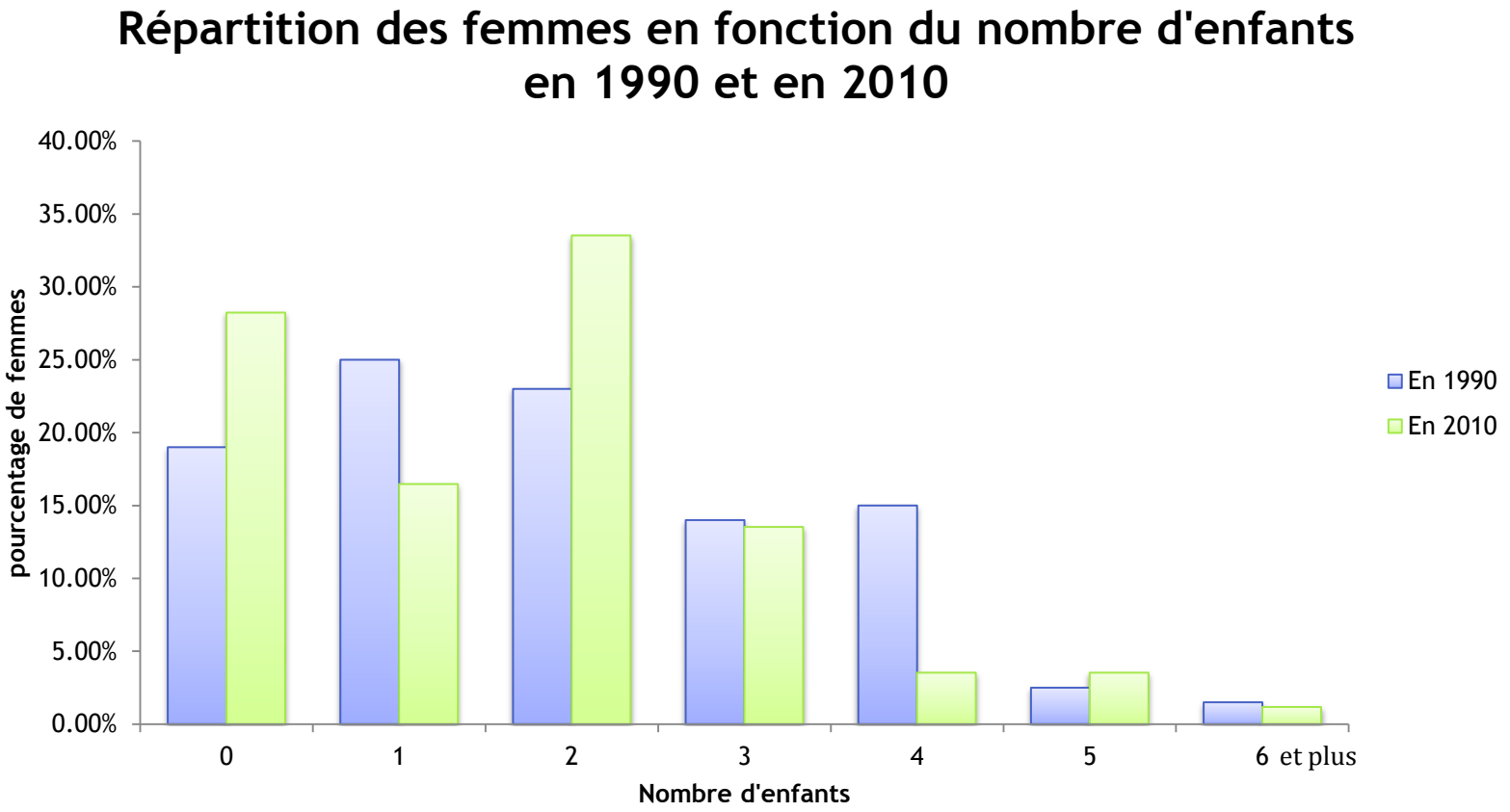
QUESTIONS ?

Illustrations

Répartition des chevaux suivant leur race



Répartition du nombre d'enfants par femme



1. Introduction à la statistique descriptive
2. Définitions, classification variables & notations
3. Analyse descriptive d'une variable

3.2. Analyse des variables quantitatives

Analyse des variables quantitatives

1. Variable quantitative discrète comportant peu de modalités
2. Variable quantitative discrète ou continue

Variables discrètes comportant peu de valeurs possibles

- Variable discrète comportant un nombre **fini** et **faible** de valeurs ➔ représentée comme une variable qualitative :
  - Tableau de fréquences
  - Diagramme en colonnes (en barres)
  - +
  - Fréquence cumulée
  - Fréquence relative cumulée
  - Fonction de répartition de la variable

Relevé de températures (°C): 7, 8, 7, 8, 10, 10, 7, 8, 8, 8, 8, 10, 8, 10, 7, 8, 10, 10, 8, 8

20 observations

Tableau de fréquences

x_j	n_j	f_j	N_j	F_j
7	4	0.2	4	0.2
8	10	0.5	14	0.7
10	6	0.3	20	1

$N=20$ $\sum_{j=1}^{3} f_j=1$

B. Frank, diapositive du cours de statistiques en 1BIM

x_j	n_j	f_j	N_j	F_j
7	4	0.2	4	0.2
8	10	0.5	14	0.7
10	6	0.3	20	1

Fréquence cumulée

- Fréquence cumulée ou effectif cumulé pour la $j^{ème}$ modalité : notation : N_j.

N_j : effectif cumulé pour le niveau "j"

$$N_j = n_1 + n_2 + \ldots + n_j$$

$$N_j = \sum_{i=1}^{j} n_i$$

- → nombre d'observations plus petites ou égales à xj dans la série statistique.

x_j	n_j	f_j	N_j	F_j
7	4	0.2	4	0.2
8	10	0.5	14	0.7
10	6	0.3	20	1

Fréquence relative cumulée

- Fréquence relative cumulée pour la $j^{ème}$ modalité : notation : F_i.

F_j : fréq. relative cumulée pour le niveau "j"

$$F_j = f_1 + f_2 + \ldots + f_j$$

$$F_j = \sum_{i=1}^{j} f_i$$

- ➔ fréquence cumulée divisée par l'effectif total : $F_i = N_i/N$

Variables discrètes et continues

- Pour variables quantitatives :
  - Tableau de fréquences (& histogramme)
  - Tendance et position
  - Dispersion
  - Distribution de la série

Regroupement des données par classe

Les nuisances sonores

Au-delà des 80 décibels, attention

Voici les résultats des 25 vols enregistrés au dessus de Bruxelles la nuit-j :

90	319	26	94	75	25	61,65	66,86	63
142	60,93	68,45	77,87	119	51,63	75,33	183	67
136	98	42	75	131	43	43		

Regroupement en classes (intervalles)

- 2 observations dans la classe [0,40[
- 14 observations dans la classe [40;80[

Nombre de classes entre 5 et 10, voire 15.

Nuisances sonores

Classes	Effectif	Fréquence relative	Effectif cumulé	Fréquence relative cumulée
[0-40[	2	0,08	2	0,08
[40-80[	14	0,56	16	0,64
[80-120[	4	0,16	20	0,8
[120-160[	3	0,12	23	0,92
[160-200[	1	0,04	24	0,96
[200-240[	0	0	24	0,96
[240-280[	0	0	24	0,96
[280-320[	1	0,04	25	1,00
Total	25	1,00		

Définir les classes

- Pas de règle stricte, entre 5 et 10, voire 15 au maximum
  - Prendre, par exemple, un chiffre proche de $\sqrt{N}$

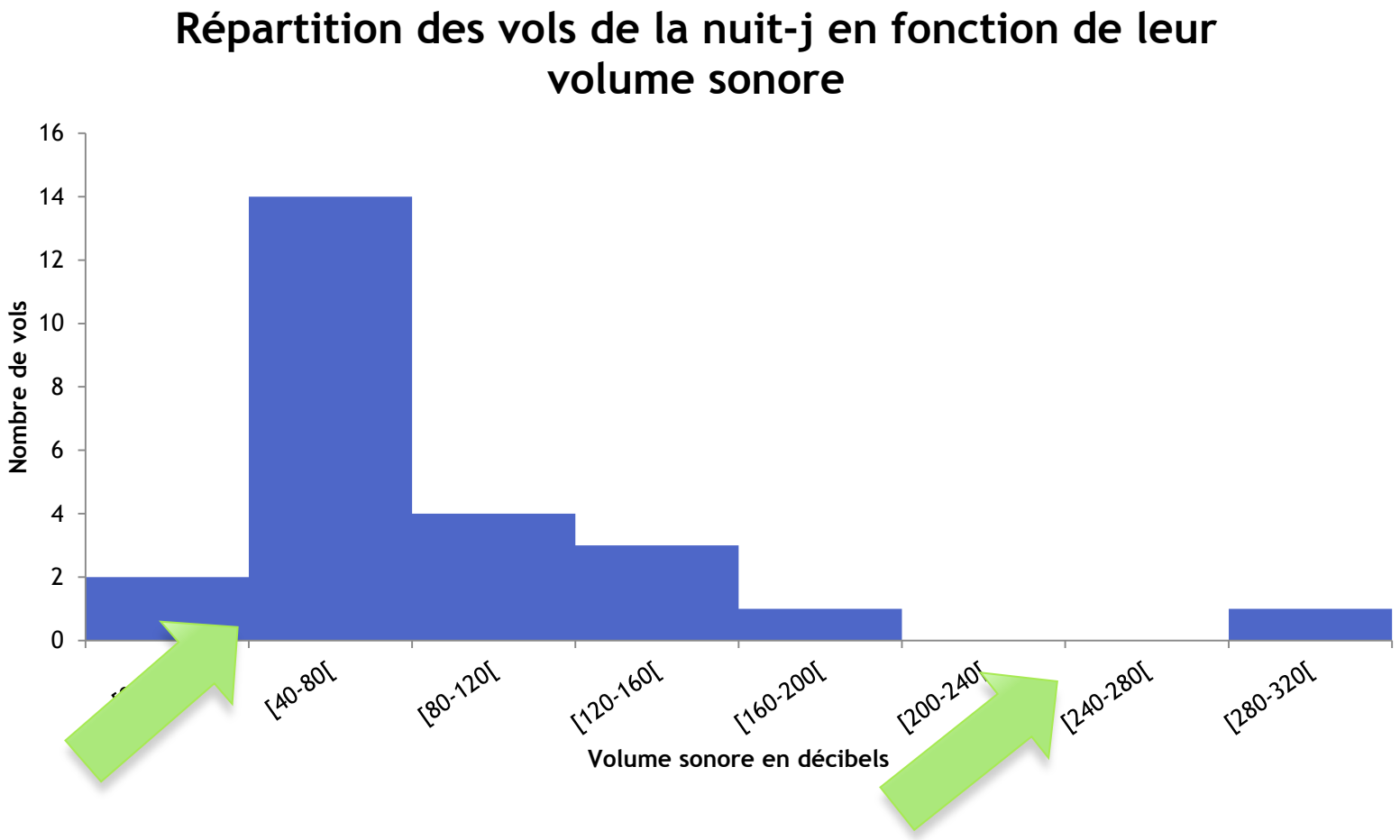
- Limites de classes $\neq$ valeurs observées
- Importance des valeurs « seuil » (80 décibels à partir desquels on dépasse la norme)

Représentations graphiques

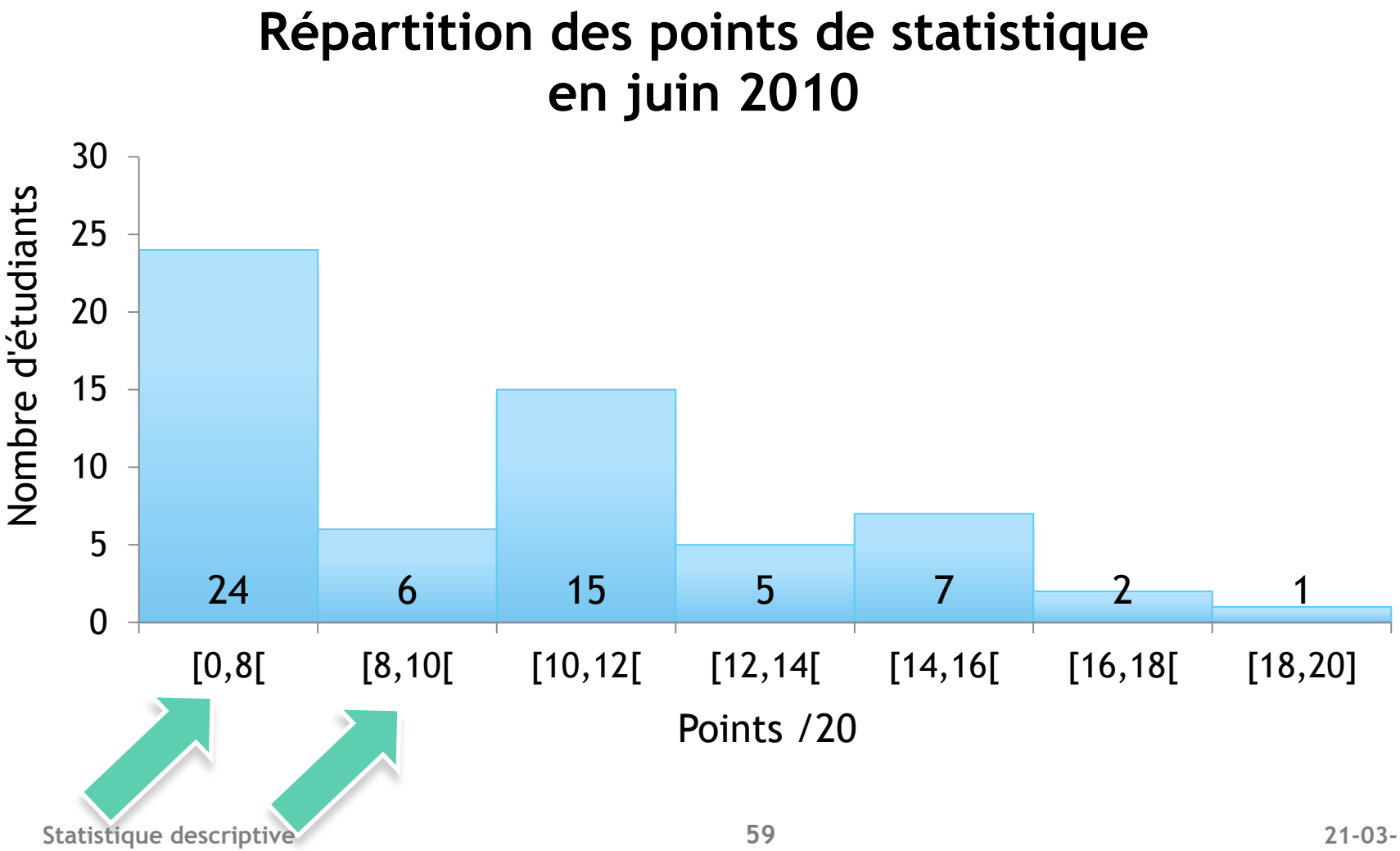
- **Histogramme**
  - pour les effectifs et fréquences relatives

- **Fonction de répartition**
  - pour les fréquences cumulées

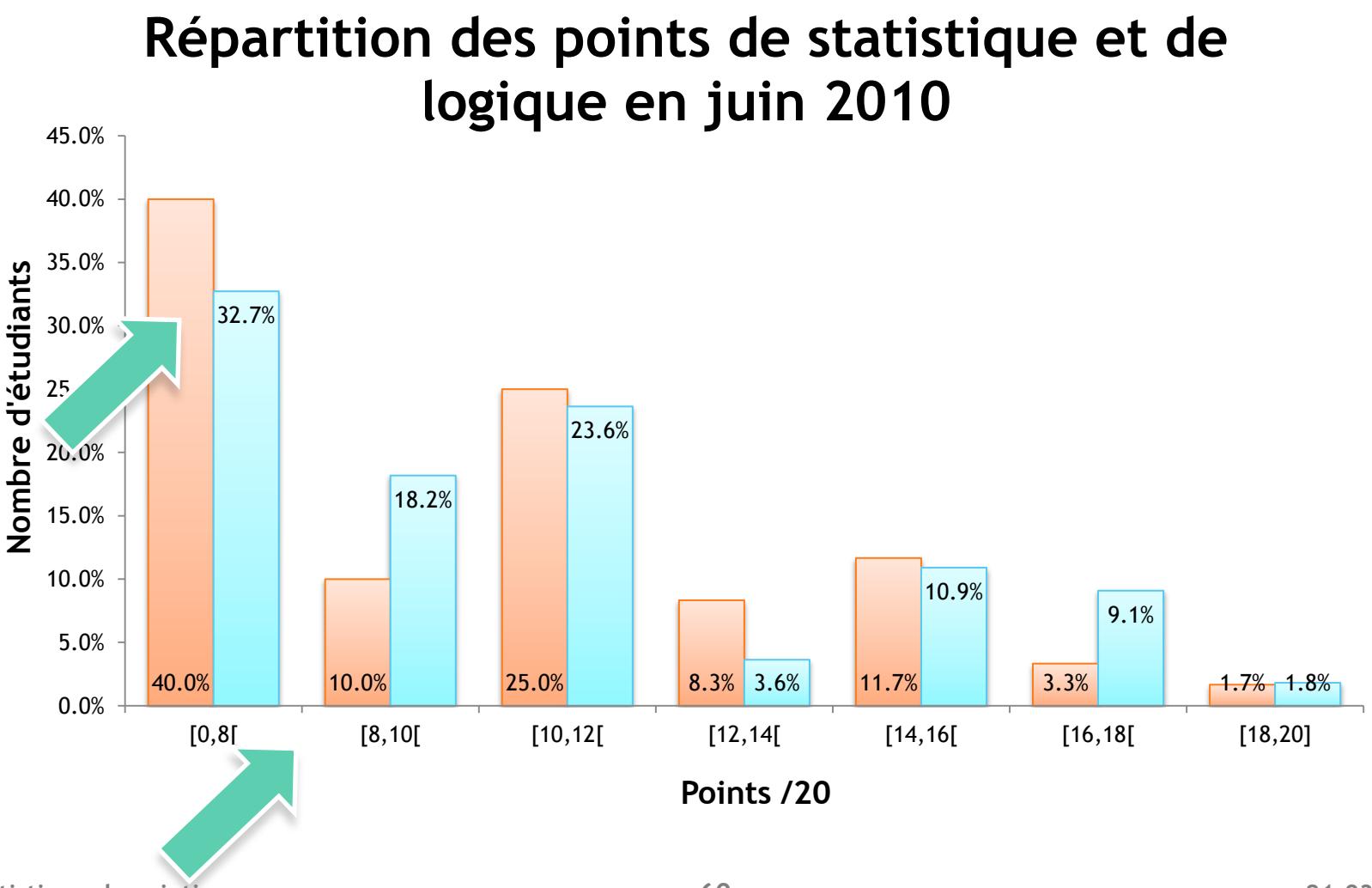
Histogramme



Exemple 1

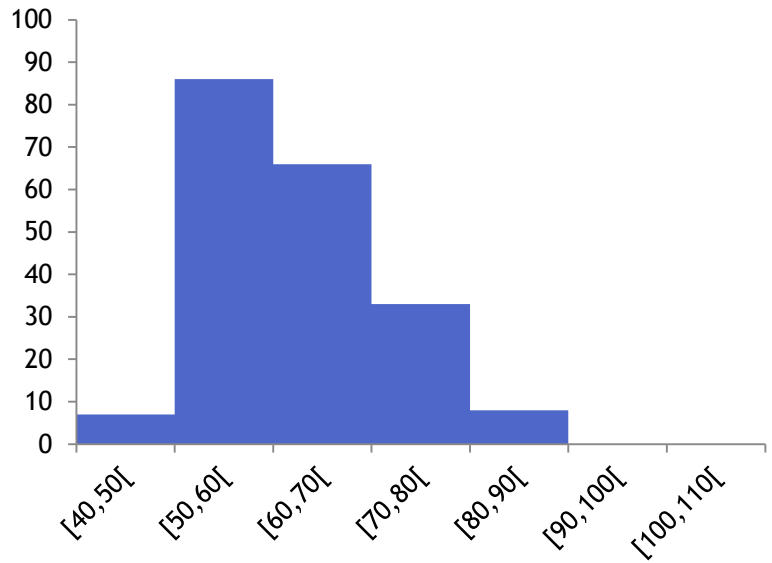

Répartition des points de statistique en juin 2010
Nombre d'étudiants
30
25
20
15
10
5
0
24
6
15
5
7
2
1
[0,8[
[8,10[
[10,12[
[12,14[
[14,16[
[16,18[
[18,20]
Points /20

Exemple 2

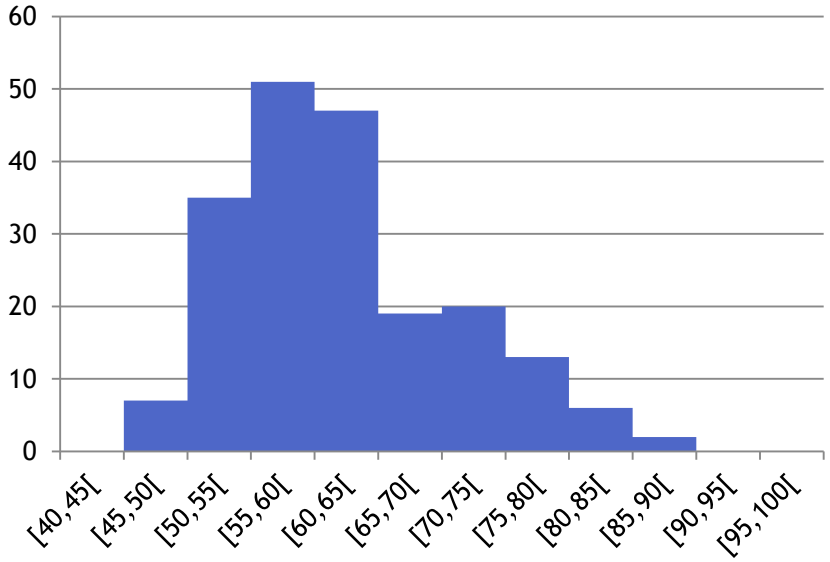

Répartition des points de statistique et de logique en juin 2010
Nombre d'étudiants
45.0%
40.0%
35.0%
30.0%
25
20.0%
15.0%
10.0%
5.0%
0.0%
40.0%
32.7%
10.0%
18.2%
25.0%
23.6%
8.3%
3.6%
11.7%
10.9%
3.3%
9.1%
1.7%
1.8%
[0,8[
[8,10[
[10,12[
[12,14[
[14,16[
[16,18[
[18,20]
Points /20

Histogramme

200 observations – 5 classes



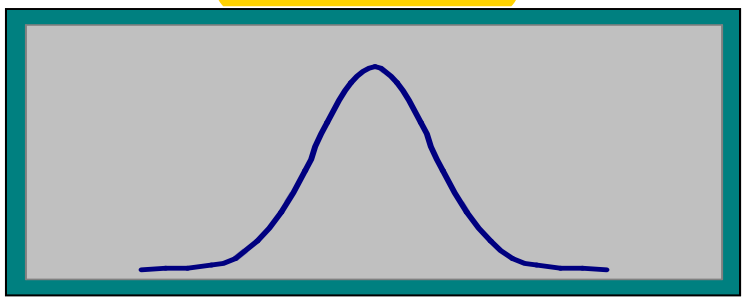
10 classes



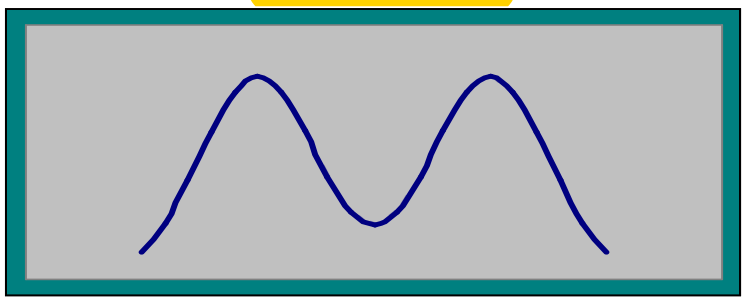
Histogramme

Distribution de fréquence de la variable :

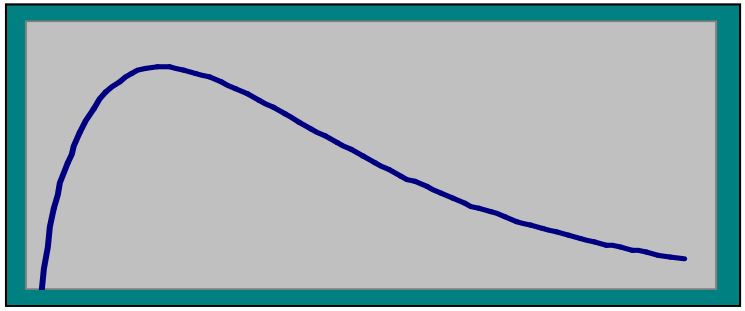
unimodale



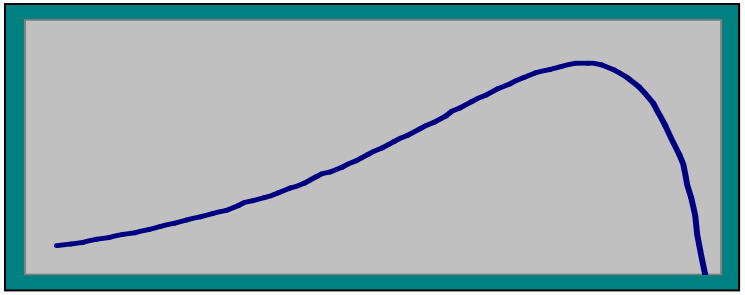
bimodale



dissymétrique à droite



dissymétrique à gauche



Résumer une série statistique quantitative

⇩

valeurs numériques

→ **Tendance centrale : mode, médiane, moyenne**

→ **Quantiles : minimum, maximum,**
premier quartile, troisième quartile
q_p $(0 \leq p \leq 1)$

→ **Variabilité (dispersion) : étendue, écart interquartile,**
variance/écart-type,
coefficient de variation (CV)

Mesures de tendance centrale

- Moyenne
- Médiane
- Mode – classe modale

Moyenne

- Moyenne d'une série stat. : la plus connue MAIS

Séries					Moy
8	9	10	11	12	10
8	9	10	11	20	11,6
0	9	10	11	12	8,4
0	9	10	11	20	10

Moyenne influençable par des valeurs extrêmes d'un même côté

Quelque chose de plus stable

Médiane

- Médiane : valeur « milieu » de la série ordonnée
  - 50% des observations ≤ médiane
- Si N est impair, (N+1)/2ième donnée

4	5	8	9	11	11	24

- Si N est pair, point milieu entre observation N/2ième et observation (N/2+1)ième

2	4	5	8	9	11	11	24

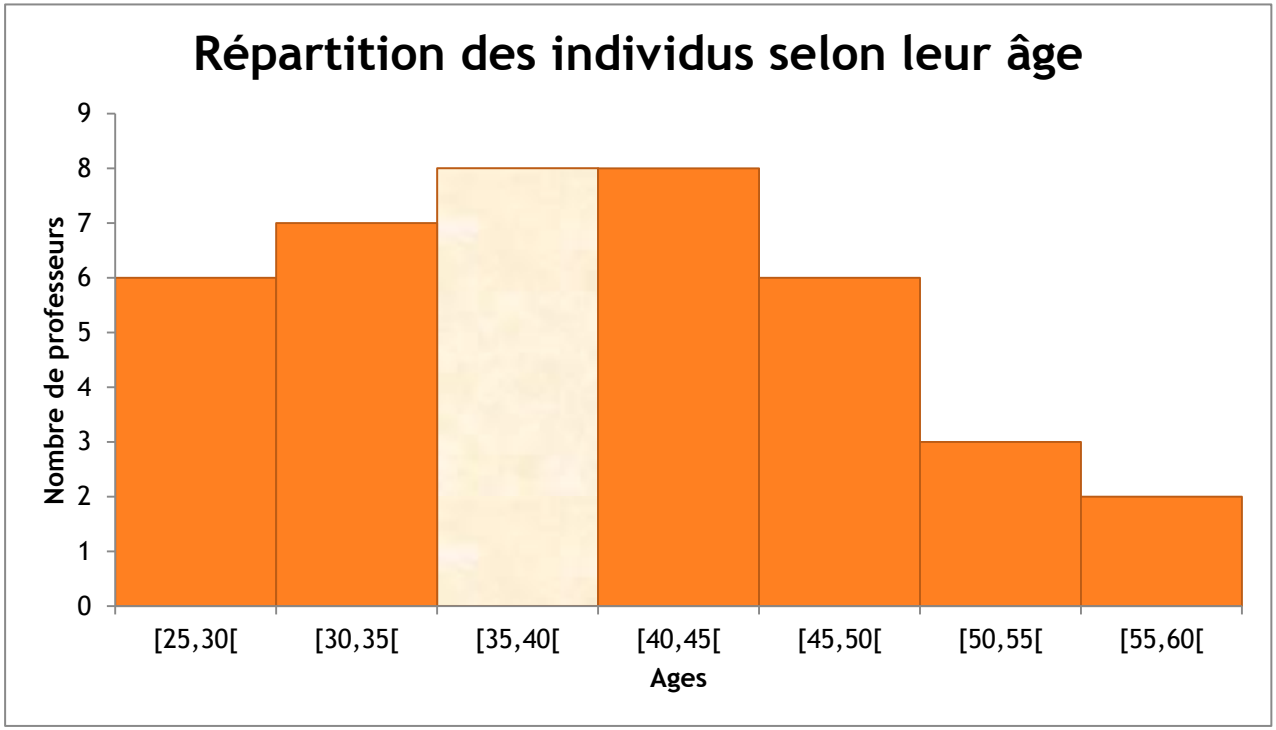
8,5

Médiane

Si variable regroupée en classes (continue),

Déterminer la classe]a;b] qui contient la médiane

Min	25
Max	59
Etendue	34
Median	39



Mode

- Mode d'une série stat. : valeur qui est observée le plus fréquemment
  - Variable discrète : modalité qui apparaît le plus souvent (avec effectif maximum)
  - Variable continue : classe modale : classe la plus élevée de l'histogramme (effectif de classe maximum)

Calcul moyenne

Moyenne ($\overline{X}$) : somme des observations divisée par le nombre total d'observations

$(X_1, X_2, ..., X_N)$ ⟹ $\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$

20 relevé de températures (°C):
7, 8, 7, 8, 10, 10, 7, 8, 8, 8, 8, 10, 8,
10, 7, 8, 10, 10, 8, 8

$$\overline{X} = (7+8+7+8+10+...+8)/20 = 8{,}4\,°C$$

Calcul moyenne

Tableau de fréquences ⇒ $\overline{X} = \frac{1}{N} \sum_{j=1}^{k} n_j x_j$

où x_j = niveau (cas discret) ou centre de classe (cas continu)

x_j	n_j	f_j
7	4	0.2
8	10	0.5
10	6	0.3

$$\overline{X} = (7*4+8*10+10*6)/20 = 8{,}4\,°C$$

Tendance centrale : comparaison

Mode : **effectif maximum**

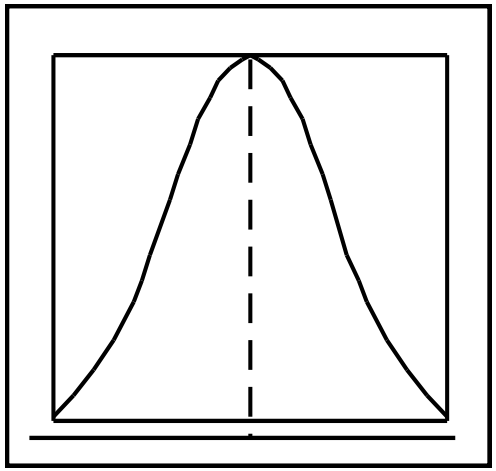
Médiane : **50% des observations,**
pas d'influence de données aberrantes

Moyenne : **toutes les observations interviennent,**
influence de données aberrantes

Les trois ensemble :
indication sur la forme de la distribution
(graphe des fréquences relatives)

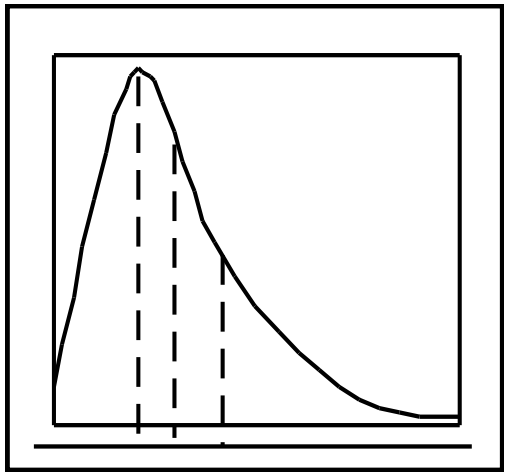
Comparaison des mesures de tendance centrale

symétrique



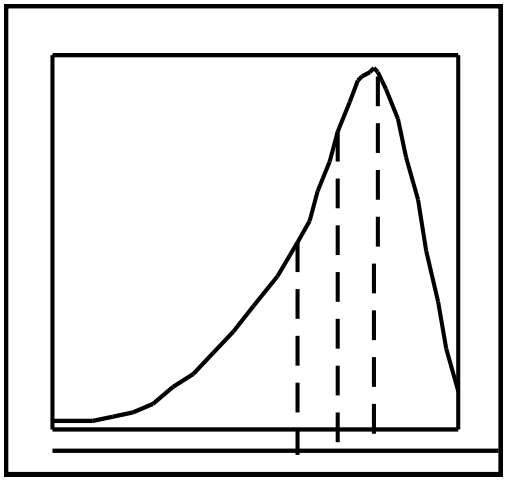
mode = moyenne = médiane

dissymétrique à droite



mode médiane moyenne

dissymétrique à gauche



moyenne médiane mode

A. Dupont, diapositive du cours de statistiques en Info2 - 2011

Autres mesures de tendance

Quantiles

Quantile q_p $(0 \leq p \leq 1)$:

valeur de la variable telle que (100*p)% des observations sont inférieures à cette valeur

q_0 : minimum (aucune valeur inférieure)

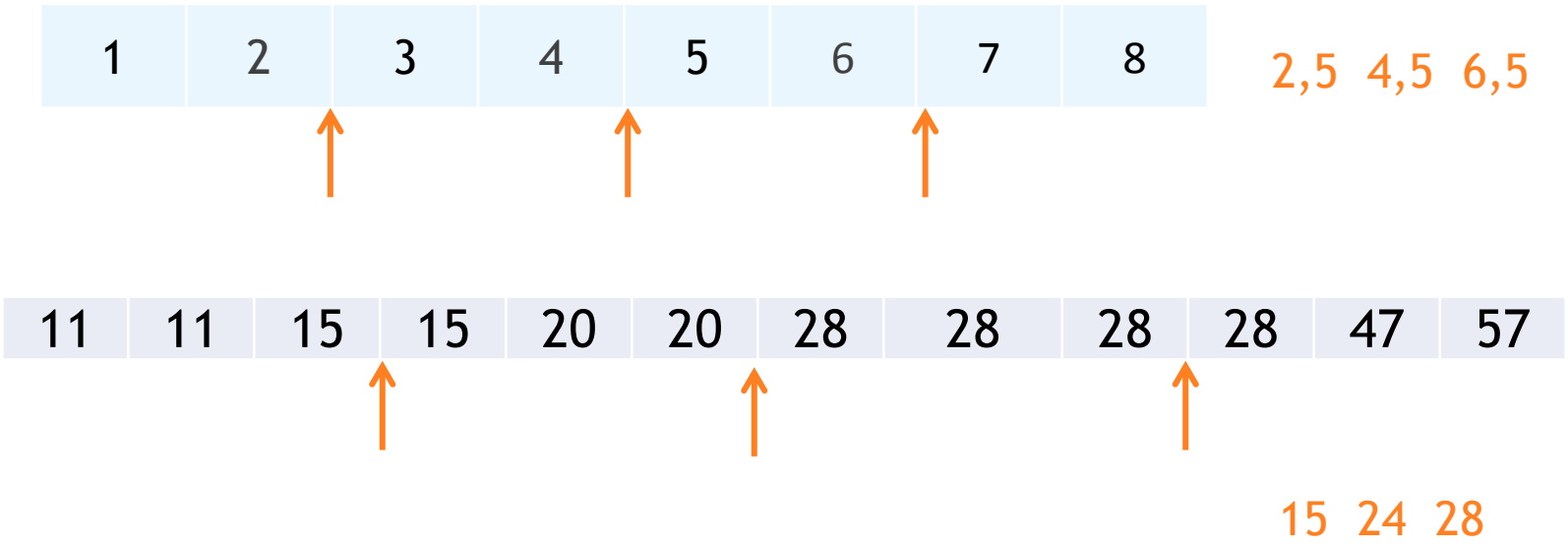
$q_{0.25}$: premier quartile (25% valeurs inférieures)

$q_{0.50}$: médiane (50% valeurs inférieures)

$q_{0.75}$: troisième quartile (75% valeurs inférieures)

q_1 : maximum (toutes les valeurs sont inférieures)

Quartiles



Quartiles

Relevé de températures (°C):

7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 10, 10, 10, 10, 10, 10

Tableau de fréquences

x_j	n_j	f_j	N_j	F_j	
7	4	0.2	4	0.2	
8	10	0.5	14	0.7	$q_1 = q_{0,25} = 8$ ←
10	6	0.3	20	1	$q_2 = q_{0,50} = 8$ ←
					$q_3 = q_{0,75} = 10$

$N=20$ $\quad \sum_{j=1}^{3} f_j = 1$

B. Frank, cours de statistiques en 1BIM

Quartiles

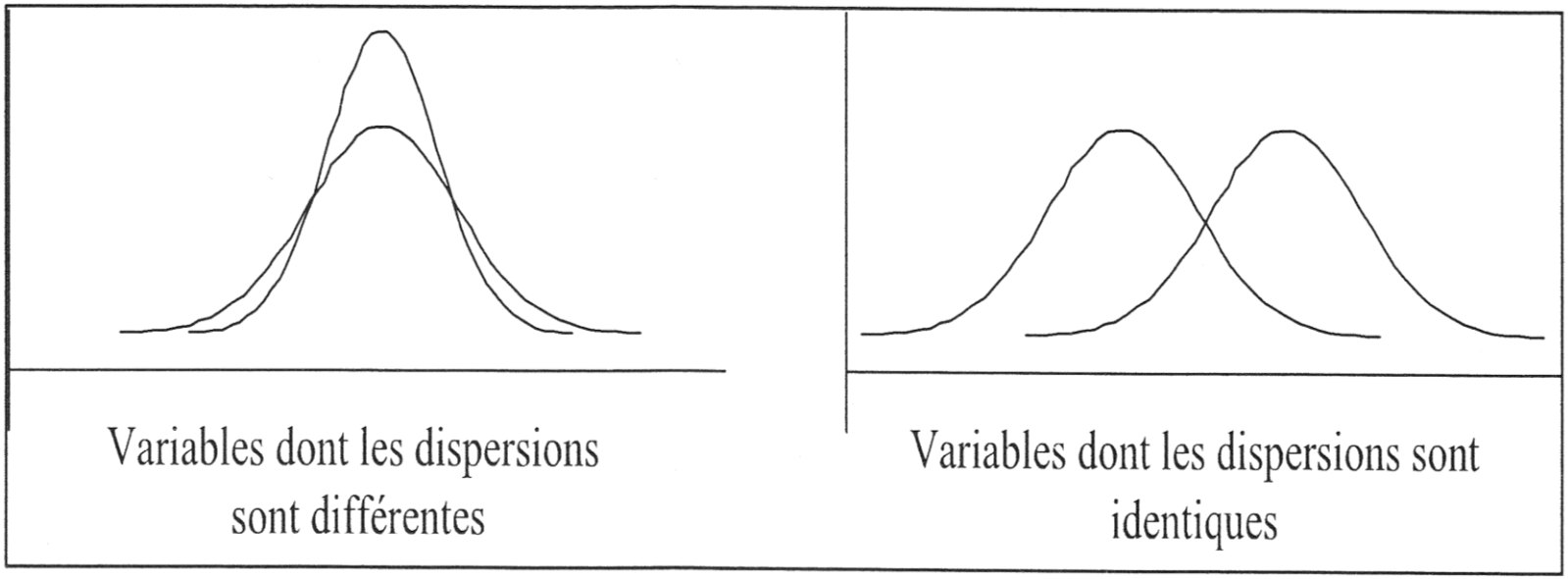
Si variable regroupée en classes (continue),

- Déterminer la classe]a;b] qui contient le quartile

Mesures de dispersion

But

Mesurer la dispersion ou variabilité d'une série statistique autour de sa tendance centrale



Etendue

- Etendue d'une série stat. : différence entre la plus grande et la plus petite valeur de la variable

Maximum - Minimum

  - Utilise très peu d'informations
  - Est sensible aux valeurs aberrantes
  - Manque de « stabilité » : dépend uniquement de deux valeurs extrêmes (même s'il y a 1000 valeurs connues)

Ecart interquartile

Ecart interquartile : $q_{0.75} - q_{0.25}$

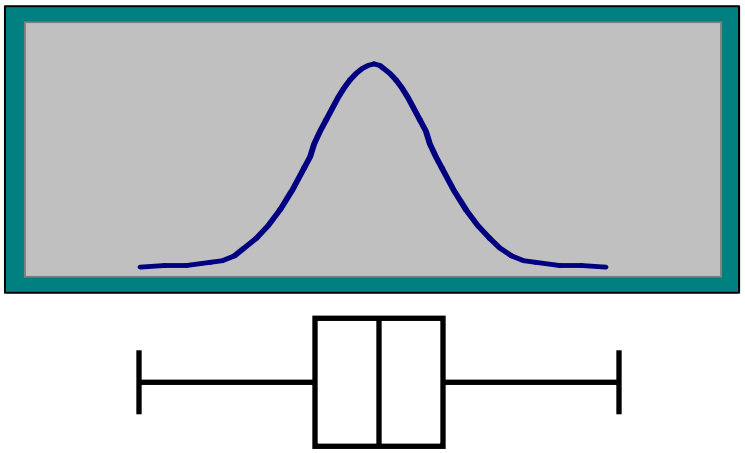
- Distance entre le 1er et le 3ième quartiles : longueur de l'intervalle contenant 50% des observations, les plus centrées
- Etendue des 50% des données « milieu » de la série
- Pas influencé par les valeurs extrêmes de la série

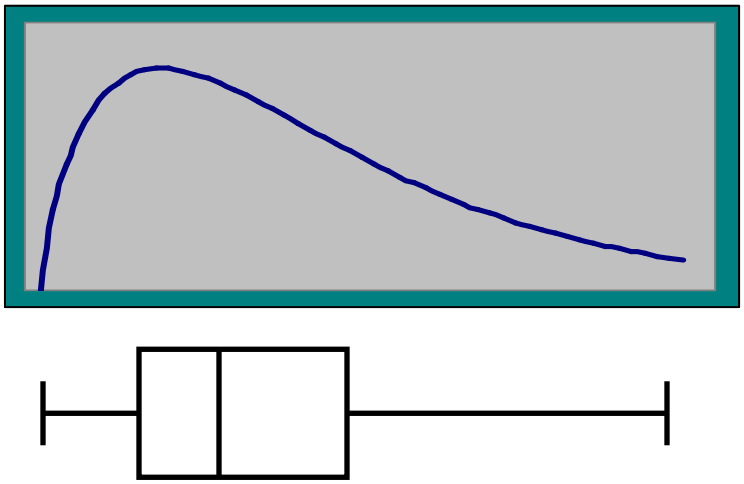
Box-plot

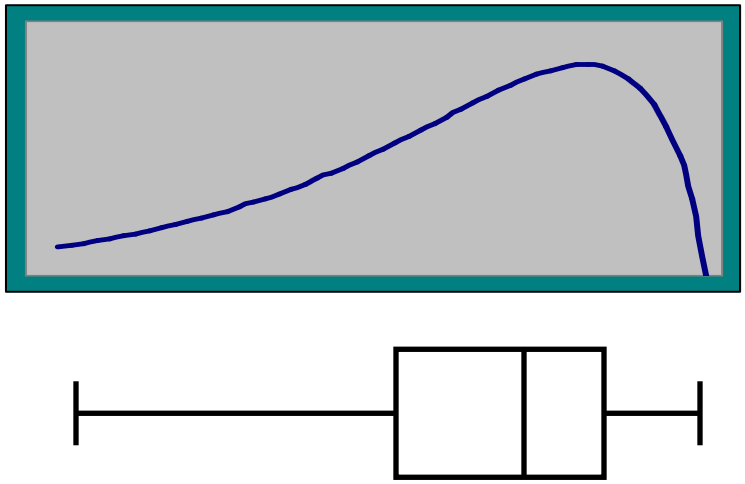
- **Graphique reprenant 5 quantiles, qui résume la structure des observations**

- **Indique si la distribution des observations est :**
  - **en forme de "cloche"**
  - **dissymétrique à droite / gauche**
  - **uniforme**
  - …

Forme de la distribution







B. Frank, diapositive du cours de statistiques en 1BIM

Box-plot

Excel Box and Whisker Diagrams (Box Plots) | Peltier Tech Blog | Excel Charts

Bonne alternative à l'histogramme quand le nombre de données est faible; en effet,

- Est unique
- Ne dépend pas d'un choix arbitraire de classes

Variance

Mesure de la dispersion par la moyenne des carrés des écarts entre les observations et la moyenne.

Pour la calculer :

- **1. Ecart à la moyenne :** $\left(X_i - \overline{X}\right)^2$
- **2. Prendre en compte toutes les observations :** $\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$
- **3. Prendre la moyenne**

$$V_X^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 = \frac{1}{N}\sum_{j=1}^{k} n_j\left(x_j - \overline{X}\right)^2$$

Variance et écart-type

➡ **Pour un échantillon, formule de la variance :**

$$S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 = \frac{1}{N-1}\sum_{j=1}^{N} n_j\left(x_j - \overline{X}\right)^2$$

Ecart-type :

➡ **Variance** ⇨ **unités au carré!!!**

Unités de base ⇨ **écart-type** population : $V_X = \sqrt{V_X^2}$

échantillon : $S_X = \sqrt{S_X^2}$

Coefficient de variation

- **Variabilité relative ne dépendant pas des unités**

$$CV_X = \frac{S_X}{\overline{X}}$$

- **Comparer la variabilité de plusieurs séries avec des unités différentes ou des ordres de grandeur différents !**

QUESTIONS ?