Biostatistiques Descriptives

Dr Marc CUGGIA

PCEM 1 – Année 2006/2007

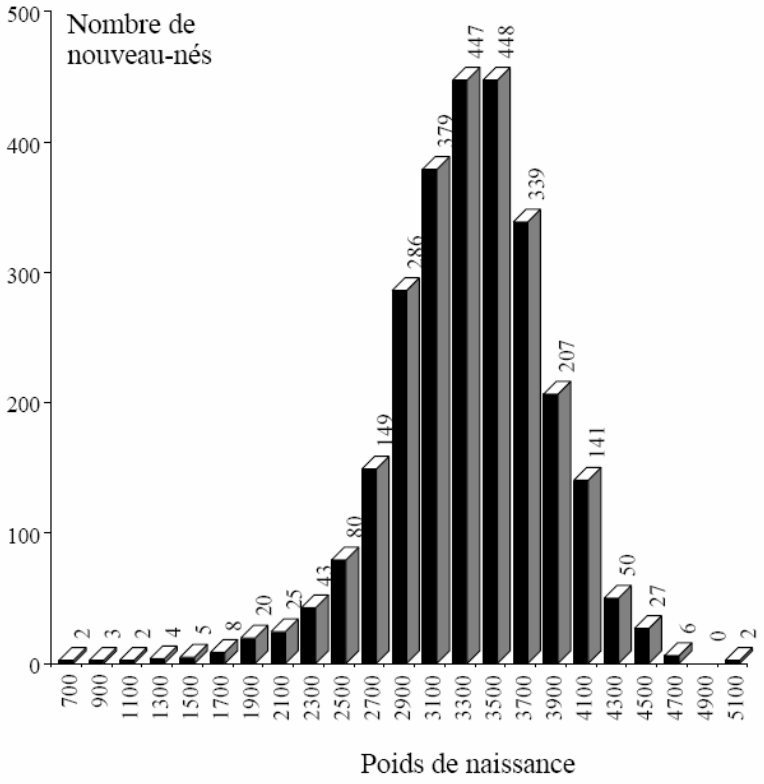
http://www.med.univ-rennes1.fr

Laboratoire d'Informatique Médicale

La variabilité est la règle dans les sciences de la vie

Exemple 1

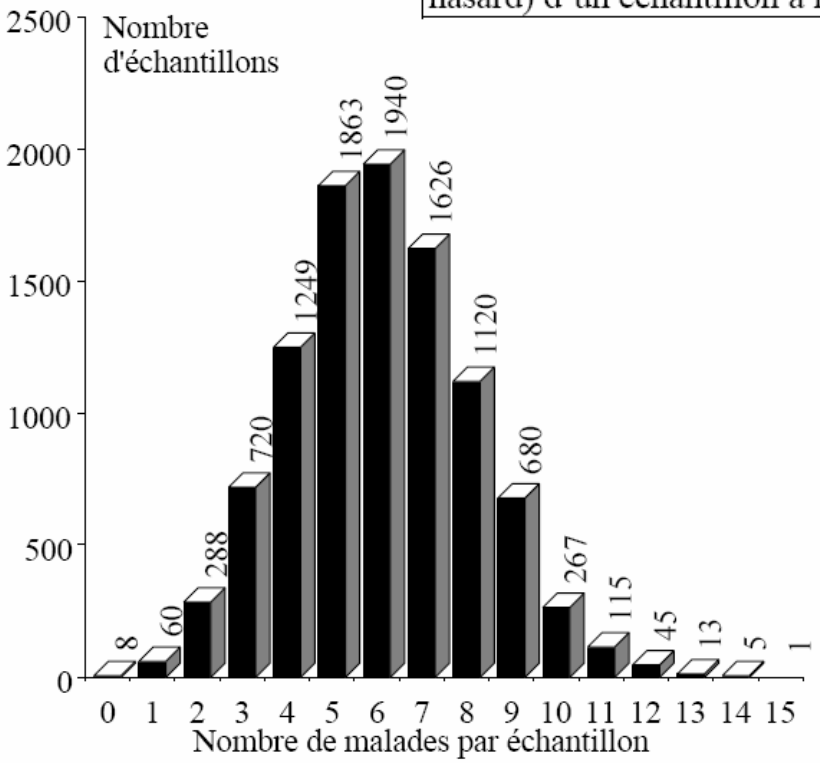
Répartition des poids de naissance de 2673 nouveau-nés



Exemple 2

Nombres de malades observés sur 10 000 échantillons de 20 sujets tirés d'une population où le pourcentage vrai de malades est 30%

Fluctuations d'échantillonnage : les observations varient (au hasard) d'un échantillon à l'autre



Conséquences des fluctuations d'échantillonnage

(1)

- On ne peut pas donner une seule valeur pour une variable telle que le poids de naissance

-> il faut des indices pour résumer les observations

moyenne, variance

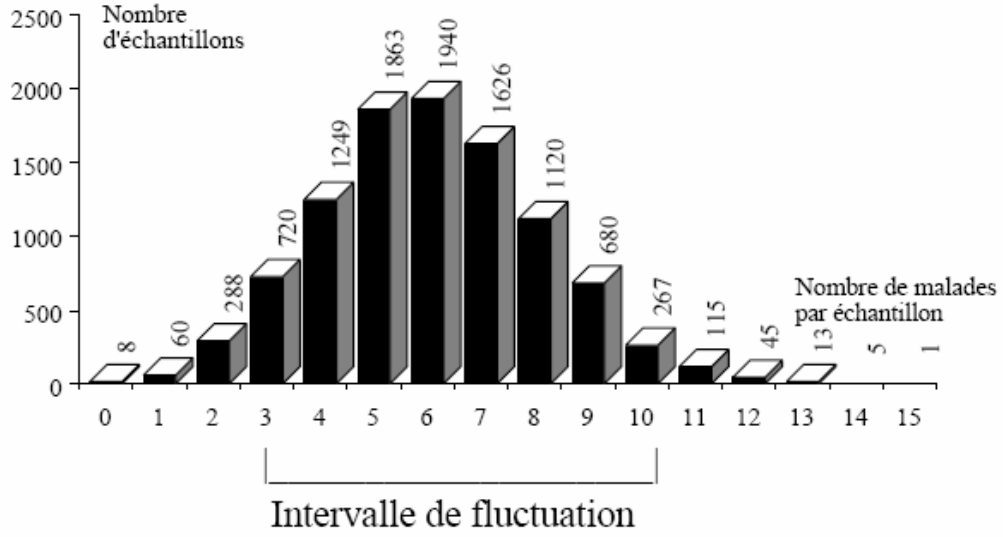
Conséquences des fluctuations d'échantillonnage

(2)

- Les conclusions qu'on peut tirer concernant un échantillon sont sujettes à erreur

Le pourcentage de malades dans un échantillon de 20 sujets est compris entre 15% et 50%
… mais seulement pour 95% des échantillons.

-> Intervalle de fluctuation



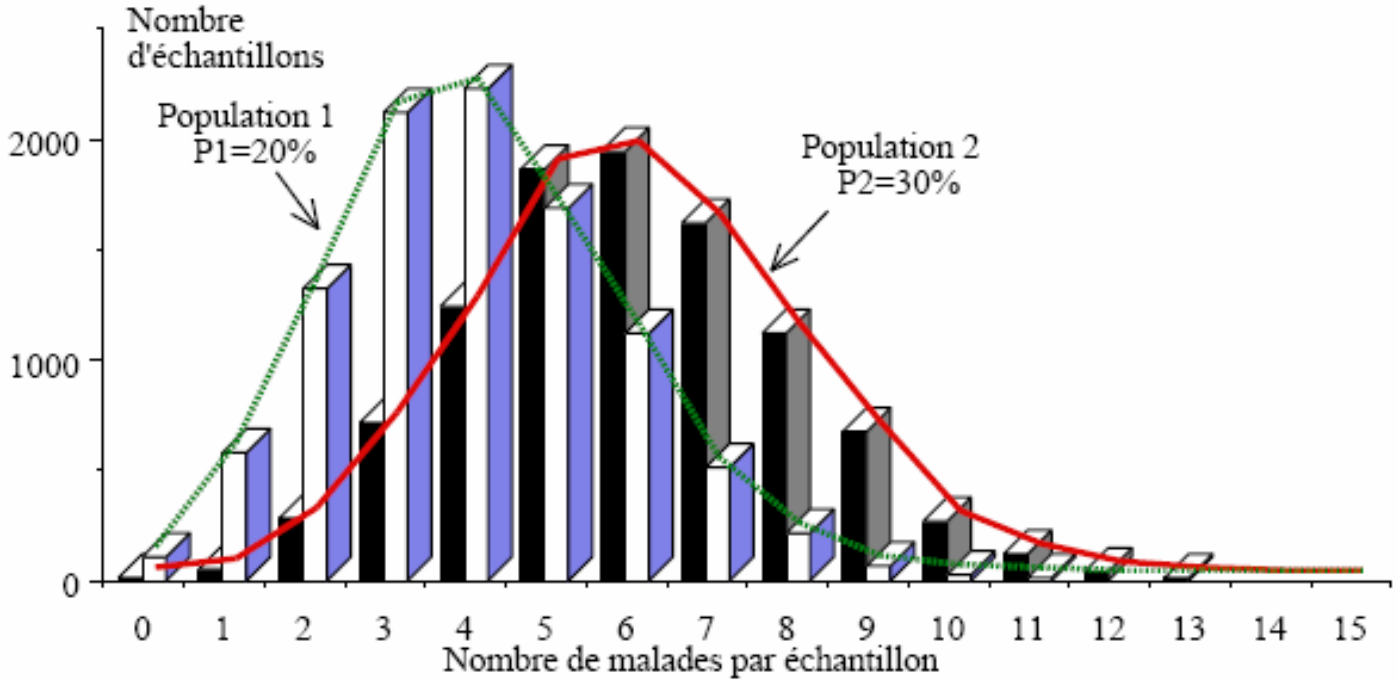
• A partir d'un échantillon, on ne doit pas donner une estimation unique d'un pourcentage (ou d'une moyenne), mais un intervalle

-> Intervalle de confiance

Conséquences des fluctuations d'échantillonnage (3)

- La comparaison de pourcentages (ou de moyennes) observés nécessite des précautions

-> Tests statistiques


Nombre
d'échantillons
Population 1
P1=20%
Population 2
P2=30%
2000
1000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Nombre de malades par échantillon

Les méthodes statistiques permettent de prendre en compte la variabilité individuelle et les fluctuations d'échantillonnage.

Le raisonnement se fait au niveau de groupes de sujets.
La constitution de ces groupes conduit souvent à simplifier la réalité.

Définitions

- La population cible est l'ensemble de tous les objets que l'on étudie.
- Un individu ou une unité statistique est un objet de cette population.
- Un échantillon est une partie choisie d'une population.
- Le nombre d'objets composant une population ou un échantillon est appelé sa taille ou effectif.
- Caractère (variable) : caractéristique ou propriété susceptible d'être possédée ou non par les individus de la population étudiée (ex : taille, glycémie, rythme cardiaque, etc..)
- Modalité : valeur que peut prendre un caractère (on peut les ordonner)

- **Effectif : nombre total « N » d'individus de la population ou de l'échantillon**
  - **n.b. : si n_i = nombre d'individus correspondant à la modalité x_i,**
  - **alors** $N = \sum n_i$
- **Fréquence d'un caractère :**
  - **nombre d'individus possédant le caractère normalisé à l'effectif total**
  - $f_i = n_i / N$

Définitions

- Lorsque l'on veut connaître certaines caractéristiques d'une population, on dit qu'on enquête sur la population.
  Une enquête peut être réalisée auprès de toute la population ou sur un échantillon.
  - Un recensement est une enquête réalisée auprès de toute la population.
  - Un sondage est une enquête réalisée sur un échantillon.

Exemples

- **Étude portant sur la consommation de tabac chez les français**
  - **la population est l'ensemble des français et la caractéristique est la consommation de tabac**

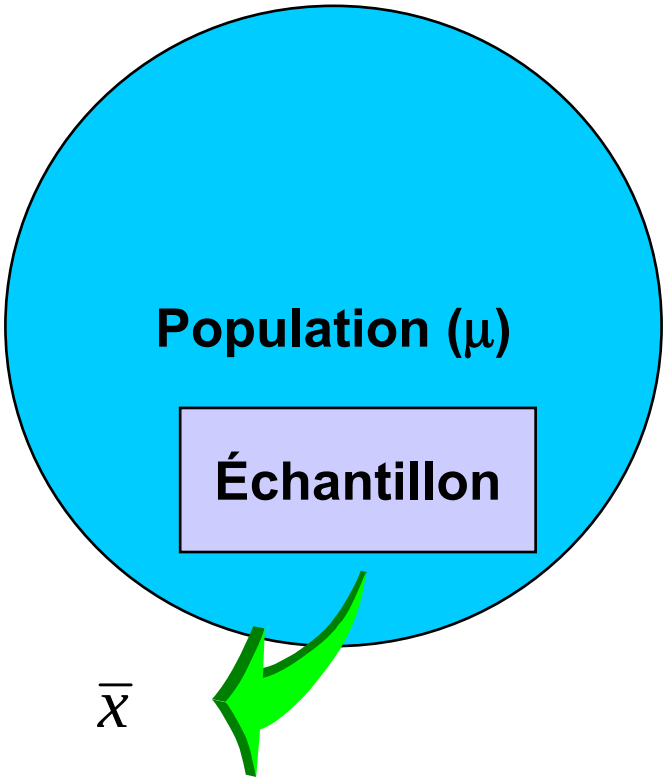
- **Étude portant sur la durée des ampoules électriques produites dans l'usine X.**
  - **La population est constituée des ampoules électriques produites à l'usine X et la caractéristique étudiée est la durée des ampoules.**

- **Une compagnie pharmaceutique veut vérifier un nouveau vaccin contre une certaine maladie.**
  - **On administre ce produit à 50 patients atteints de la maladie.**
  - **La population est formée de tous les gens atteints de la maladie, l'échantillon est formé des 50 patients à qui on a administré le médicament et la caractéristique étudiée est la réponse au médicament.**

- **Les coûts élevés et les délais trop longs, reliés à un recensement, sont les principales raisons qui nous amènent à utiliser un sondage puisque la taille d'un échantillon est beaucoup plus petite que celle de la population.**

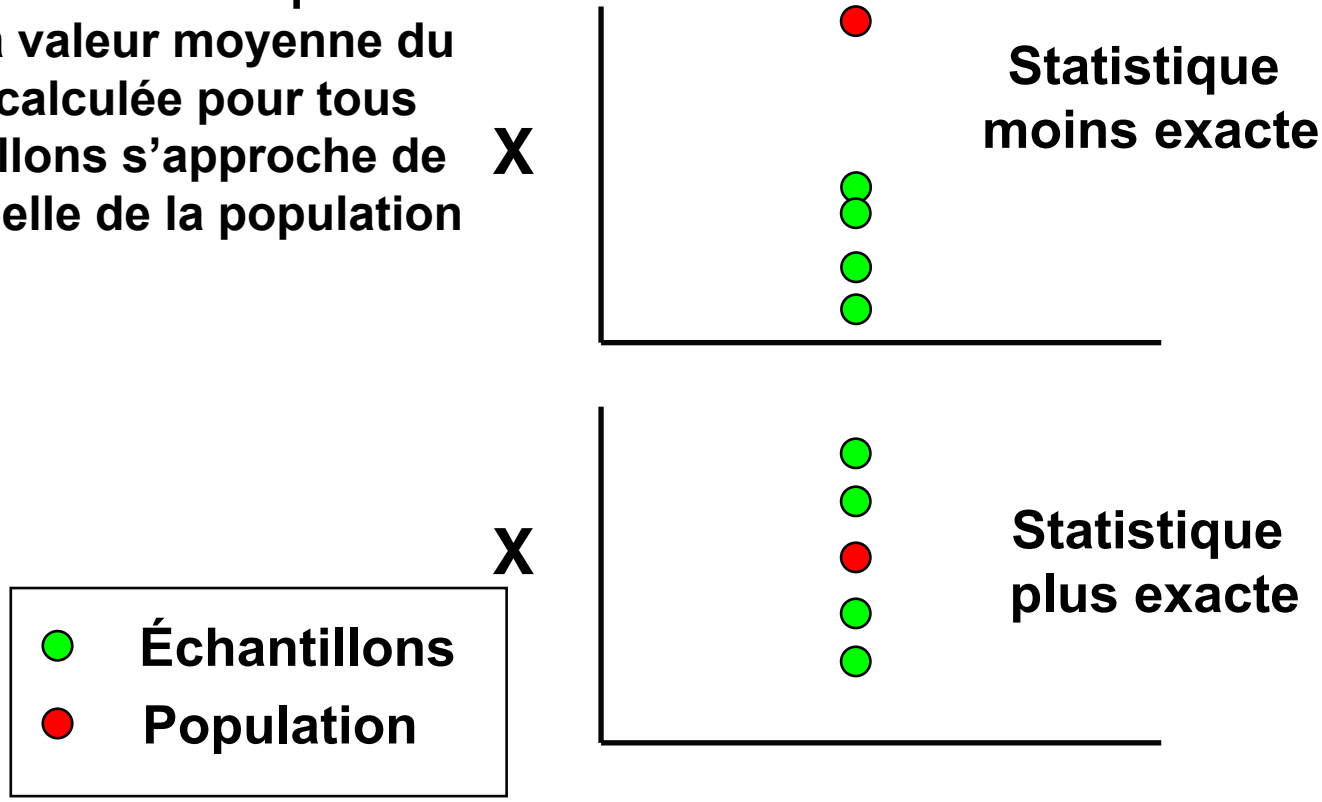
Terminologie

- ***Paramètre ou indicateur*** **: définit une population**
- ***statistique :*** **estimés des paramètres d'une population**
- **par exemple:**
- **la moyenne de la population (μ)**
  - **versus**
- **la moyenne d'un échantillon ($\bar{x}$)**



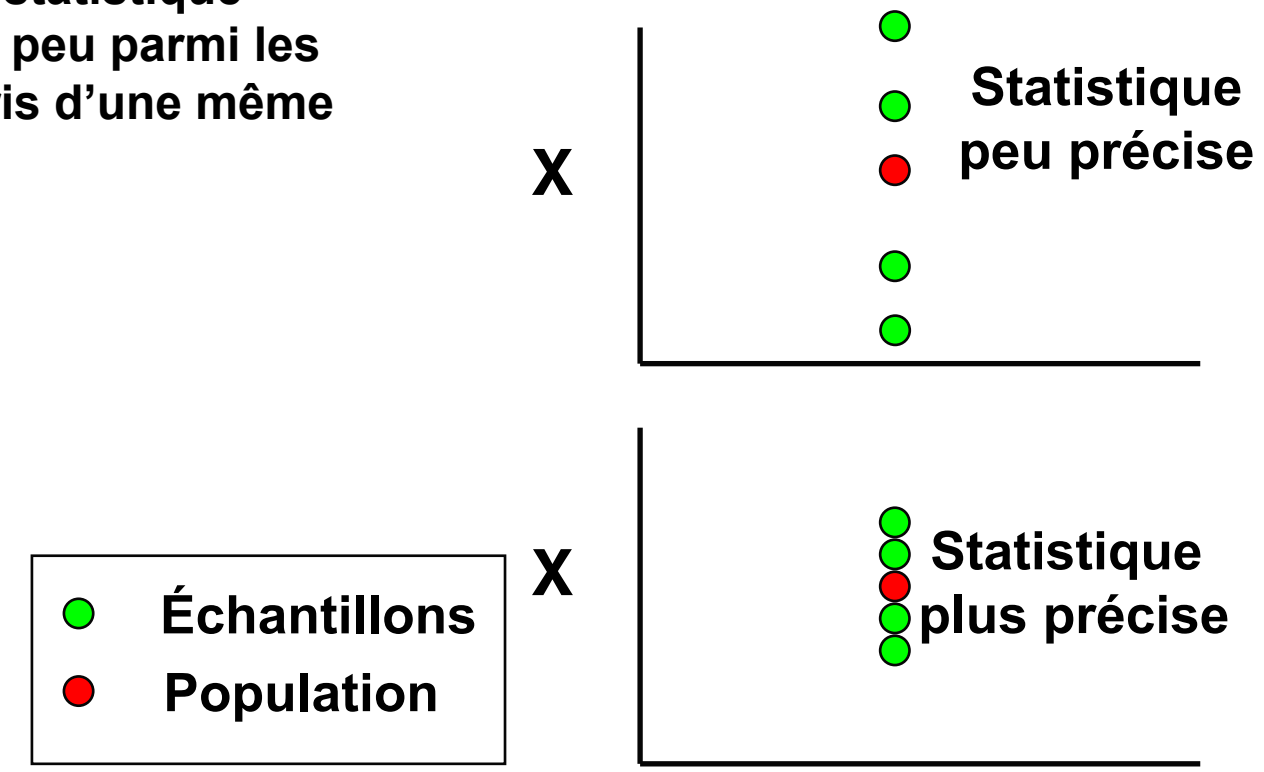
Propriétés d'une statistique

- ***Exactitude*: une statistique est exacte si la valeur moyenne du paramètre calculée pour tous les échantillons s'approche de la valeur réelle de la population**



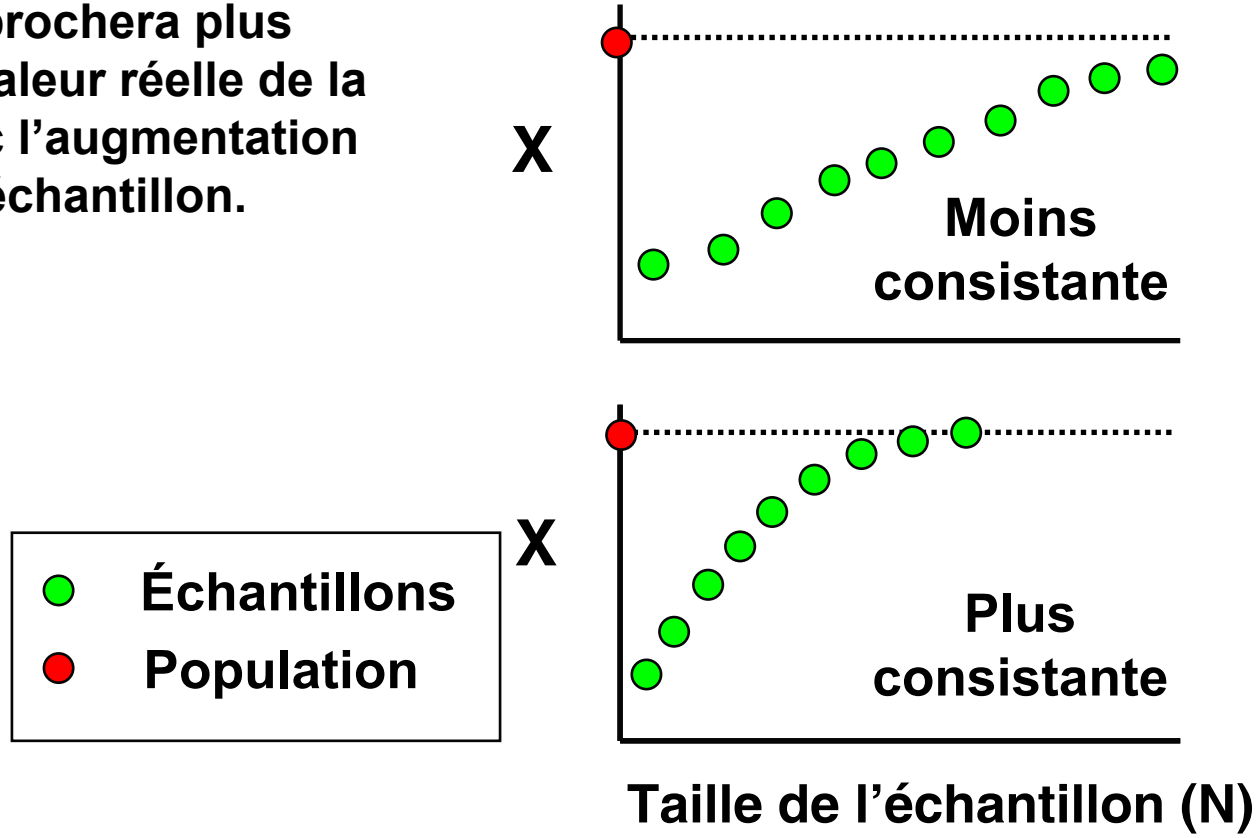
Propriétés d'une statistique

- ***Précision*: une statistique précise variera peu parmi les échantillons pris d'une même population**



Propriétés d'une statistique

- ***Consistance***: une statistique consistante approchera plus rapidement la valeur réelle de la population avec l'augmentation de la taille de l'échantillon.



Variables

- **Définition :**
- **Caractéristique ou facteur susceptible de prendre une valeur différente selon les individus (ou les unités statistiques) étudiées**
  - **Couleur de cheveux**
  - **La taille**
  - **La durée d'incubation d'une maladie**
- **Différents types de variables**
  - **Quantitatives**
  - **Qualitatives**
  - **Temporelles**

Variables qualitatives

- Non mesurables
- Leurs valeurs sont des qualités réparties en classes
- On dénombre les effectifs appartenant à chacune des classes
- 3 types
  - Variables qualitatives ordinales
  - Variables qualitatives nominales
  - Variables qualitatives binaires

- **Variables ordinales**
  - **Classes pouvant être ordonnées selon une échelle de valeur**
    - Niveau d'étude : primaire, secondaire,supérieur
    - Score de Glasgow : 1 à 15
    - Complication d'une maladie : Modérée, Moyenne, Sévère
  - **➔ pas de manipulation arithmétique**
  - **Peu être considérées comme variables semi-quantitatives**

Variables qualitatives

- **Variables qualitatives nominales**
  - **Variables dont les classes ne peuvent être hiérarchisées**
  - **Elles sont nommées mais pas ordonnées**
  - **L'ordre de présentation est arbitraire**
    - Groupe sanguin — A B O AB
    - État civil — Célibataire,marié,divorcée
    - Accident — Voie Publique, sport, jeux

- **Variables binaires** 
  - **Cas particulier de variable nominales**
  - **Prennent 2 valeurs**
  - **Dichotomique, booléenes, bernouillies**
    - Etat de santé ➔ malade, sain
    - Survie ➔ Vivant,décédé

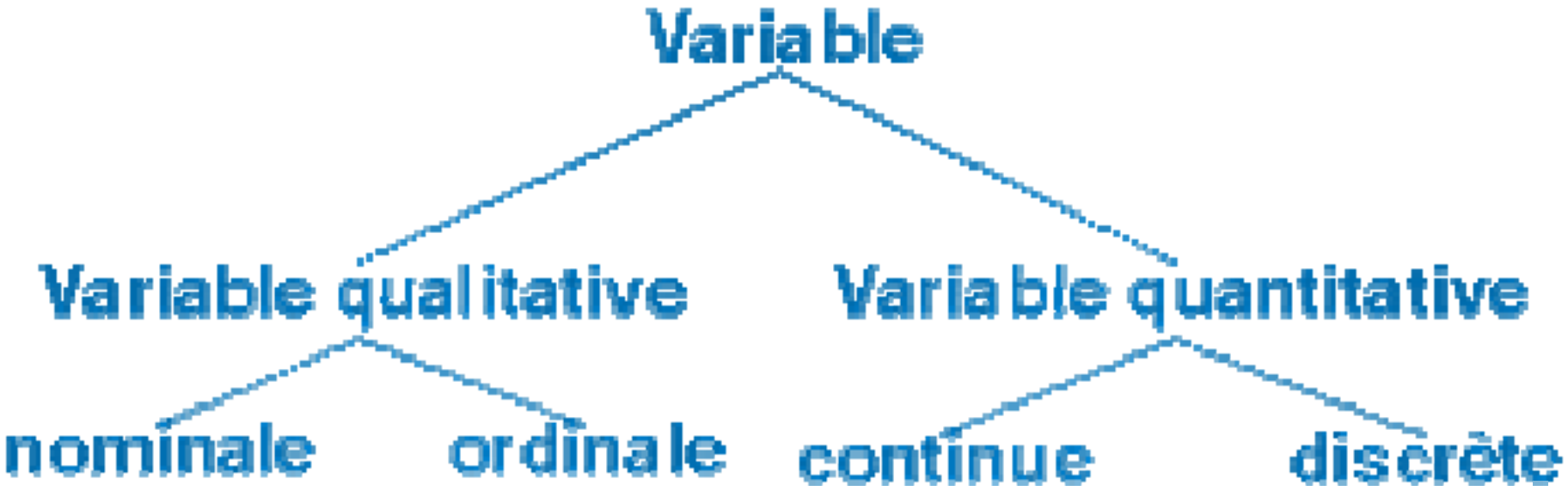
Variables quantitatives

- **Caractérisées par des valeurs numériques**
  - **Exploitable arithmétiquement**

- **Variables quantitatives continues**
  - **Prennent n'importe quelles valeurs numériques dans l'intervalle d'observation**
  - **Appartient à l'ensemble des réels : toutes les valeurs sont possibles**
    - Poids 56,3 kg
    - Taille 1,72 m
    - Cholestérol 2,22 g/l
  - **Attention au nombre de décimale**
  - **Très utilisées en médecine**
  - **La précision est limitée par l'instrument de mesure**
  - **En fait variable pas vraiment continues : saut d'intervalles**
    - TA : 12,5/82

- **Variables quantitatives discrètes**
  - **Variables numériques discontinues.**
  - **En général valeurs entières**
  - **Souvent ⇔ à un dénombrement**
    - Rechute d'une maladie 3 rechute par an
    - Rappel de vaccin 4 injections
    - Dentition 32 dents

- **Variables temporelles**
  - **Variables quantitatives particulières utilisant les unités de temps**


Variable
Variable qualitative
Variable quantitative
nominale
ordinale
continue
discrète

Variable continue à Variable discrètes DISCRETISATION

- On peut passer facilement d'une variable quantitative à une variable qualitative.
- On utilise une échelle de dépendance.
- Ex : grade en fonction de la taille
  - [0-5] - grade 1
  - [6-10] - grade 2
  - [11-20] - grade 3

- Perte d'information car on suppose que des individus différents on le même comportement dans un intervalle donné.

- **Division en classes ou en intervalles**
- **Les classes sont contiguës et ne se chevauchent pas.**

Exemple : distribution des poids chez 100 femmes

Classes	Effectifs	Fréquences	%
X	F	f = F/n	100.f
40-44	5	0,05	5,0
45-49	12	0,12	12,0
50-54	31	0,31	31,0
55-59	31	0,31	31,0
60-64	16	0,16	16,0
65-69	3	0,03	3,0
70-74	2	0,02	2,0
Total	n=100	1,00	100,0

Préciser le domaine de classe

Mesures limites	Limites réelles	Points médians	Effectifs
40-44	39,5-44,5	42	5
45-49	44,5-49,5	47	12
50-54	49,5-54,5	52	31
55-59	54,5-59,5	57	31
60-64	59,5-64,5	62	16
65-69	64,5-69,5	67	3
70-74	69,5-74,5	72	2
		Total:	100,0

Dans le cas suivant, déterminez :

. Quelles sont les variables étudiées ?
. Quelle est leur nature : qualitative (nominale ou ordinale) ou quantitative (discrète ou continue)?
. Quelle échelle a été utilisée : nominale, ordinale, d'intervalle ou de rapport)?

Conditions météorologiques dans certaines villes canadiennes.
Jeudi 13 février 1997 14h30.

Villes	**températures** (°C)	**vitesse des vents** (km/h)
Edmonton	-4	9
Calgary	-4	0
Fredericton	-11	22
Halifax	-3	33
Montréal	-17	13
Québec	-19	28
Vancouver	6	13
Toronto	-10	15

Conditions météorologiques dans certaines villes canadiennes.
Jeudi 13 février 1997 14h30.

Villes	**températures** (°C)	**vitesse des vents** (km/h)
Edmonton	-4	9
Calgary	-4	0
Fredericton	-11	22
Halifax	-3	33
Montréal	-17	13
Québec	-19	28
Vancouver	6	13
Toronto	-10	15

Les variables	**La nature des variables**	**L'échelle utilisée**
Villes canadiennes	Qualitative nominale	Échelle nominale
Températures	Quantitative continue	Échelle d'intervalle
Vitesse des vents	Quantitative continue	Échelle de rapport

Organisation des données

- **Objectif : décrire l'ensemble des données recueillies de façon synthétique**
  - **Tri de données**
  - **Regroupement en classe**
    - Discrétisation d'une variable continue en variable discrète
      - Valeurs d'un test biologique : Titrage avec seuil : test positif ou négatif
    - Transformation d'une variable quantitative discrète en variable qualitative ordinale
      - Poids : Maigre – Normal - Obèse
- **Construction d'une échelle de classification en divisant la série en classes**
- **Définition des bornes entre lesquelles on compte les individus**
  - **Perte d'information**
- **Choix des bornes (ex)**
  - **Par amplitude**
  - **Par fréquence**
  - **Par convenance**
- **Créer des groupes exclusifs (en bornant correctement les intervalles)**

Effectif et fréquence

$$f_i = \frac{n_i}{N}$$

- **i ⇔ modalité ou classe**
  - Femme, homme
  - Malade, sain
- **ni ⇔ effectif de la classe**
  - Ex : pop 1000, homme 450, femme 550
  - Freq homme, femme
- **N : Le nombre total**

Effectifs et fréquences cumulées

- Utilisées lorsque une variable est ordonnée
- Ajout à l'effectif d'une classe le total des effectifs des classes inférieures
- Fréquence cumulées = effectif cumulé / total de la série

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

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	

Distribution

- Constituée par l'ensemble des effectifs réparties dans les classes étudiées
- Pour étudier une distribution, on examine les fréquences des effectifs dans toutes les classes
- En statistique on regarde si une distribution OBSERVEE ressemble à une distribution THEORIQUE
- Si c'est le cas, on peut utiliser toutes les propriétés mathématique du modèle théorique pour étudier la distribution observée.

Description des données

- 3 procédées pour décrire un ensemble de données statistique ou un distribution
  - Les tableaux
  - Les diagrammes
  - Le calcul de paramètre ou indicateurs
- Tableaux brut de données

	v1	groupe	age	sexe	durée_ma	forme	mms	edss1	edss2	mif1
1	andré	Témoin	47	Homme	8	Secondairement progressif	28	4,5	.	116
2	armand	Témoin	50	Femme	11	Secondairement progressif	29	4,5	.	115
3	baron mc	Témoin	39	Femme	8	progressif	29	4,5	.	118
4	cardona	Témoin	54	Homme	17	Secondairement progressif	30	5,5	.	116
5	carra	Témoin	46	Homme	4	Secondairement progressif	28	4,5	.	114
6	cerveaux	Témoin	54	Femme	13	progressif	27	5,5	.	112
7	geoffroy	Témoin	39	Homme	14	Secondairement progressif	29	4,5	.	122
8	guillo	Témoin	60	Femme	7	progressif	28	4,5	.	114
9	james	Témoin	50	Femme	15	Secondairement progressif	29	3,5	.	121
10	lautredou	Témoin	58	Femme	13	progressif	26	3,5	.	120
11	le lan	Témoin	46	Homme	15	Secondairement progressif	30	4,5	.	120
12	maheo	Témoin	54	Homme	13	Secondairement progressif	29	4,5	.	112
13	martin c	Témoin	61	Femme	11	progressif	27	4,5	.	116
14	philippe	Témoin	60	Homme	11	progressif	25	4,5	.	109
15	saudrais	Témoin	41	Femme	23	Secondairement progressif	27	5,0	.	115

- Individus en ligne les variables en colonnes
- Attention CNIL

Tableaux de fréquences

Sexe

		Fréquence	Pour cent	Pourcentage valide	Pourcentage cumulé
Valide	Femme	266	33,1	33,1	33,1
	Homme	538	66,9	66,9	100,0
	Total	804	100,0	100,0	

Combinaison de variables dans un tableau

		Sexe			
		Femme		Homme	
		Count	Column Total N %	Count	Column Total N %
disc 'age (Banded)' (Banded)	0-25	4	1,5%	3	,6%
	25-50	50	18,8%	100	18,6%
	50-75	169	63,5%	360	66,9%
	>75	43	16,2%	75	13,9%
	Total	266	100,0%	538	100,0%

Pas plus de 2 variables par tableau

- **Données manquantes**
  - **Tenter de récupérer le max de données manquantes**
  - **Effectuer une double saisie par 2 opérateurs différents**
  - **Prévoir un code spécial pour les données manquantes ou aberrantes**
  - **Prévoir une règle de décision sur les données manquantes. Les représenter dans les tableaux**

Stade T tumoral

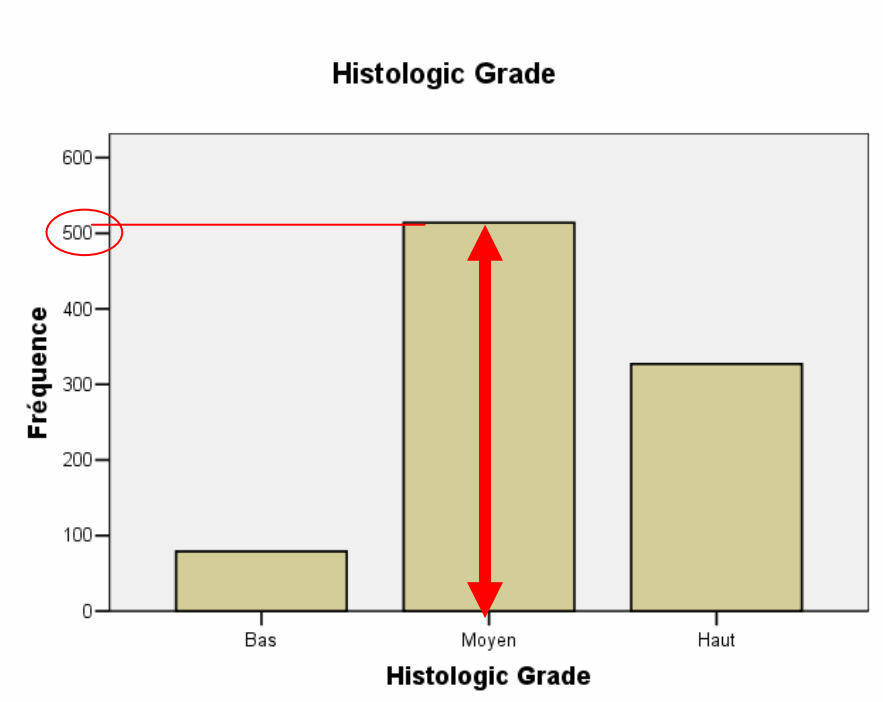
		Fréquence	Pour cent	Pourcentage valide	Pourcentage cumulé
Valide	1,00	283	35,2	35,5	35,5
	2,00	115	14,3	14,4	49,9
	3,00	383	47,6	48,0	97,9
	4,00	17	2,1	2,1	100,0
	Total	798	99,3	100,0	
Manquante	Système manquant	6	,7		
Total		804	100,0		

Les graphiques

- Les tableau représentent les données exactes
- Les graphique font ressortir une vision synthétique
- Conseils:
  - Pas de 3 d ni de camembert
  - Pas de superposition de graphe
  - Pas de colorisation abusive
  - Simple
  - Légendé (titre,axes,unités)
  - Honnête

Diagramme en barre

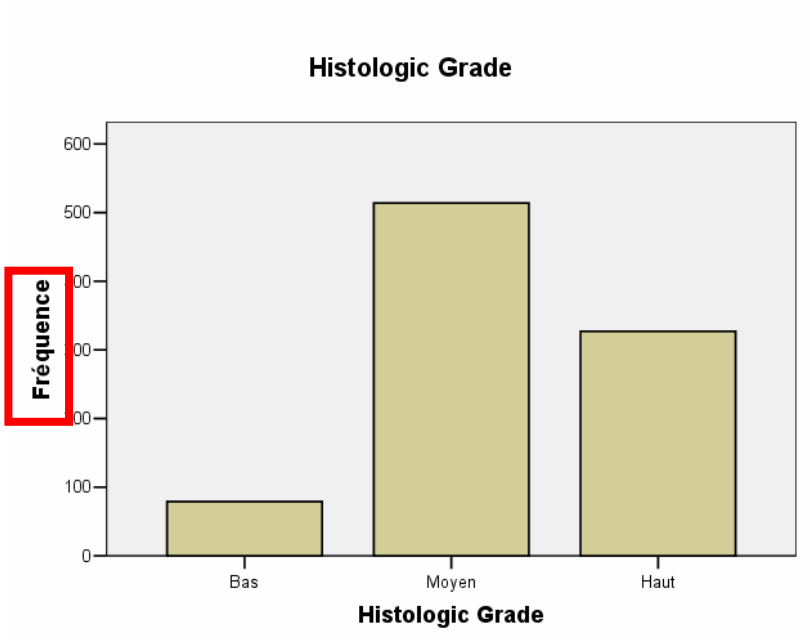
- Utilisé pour représenter une variable qualitative nominale ou ordinale
- Hauteur de chaque colonne = nombre de sujet dans la catégorie correspondante

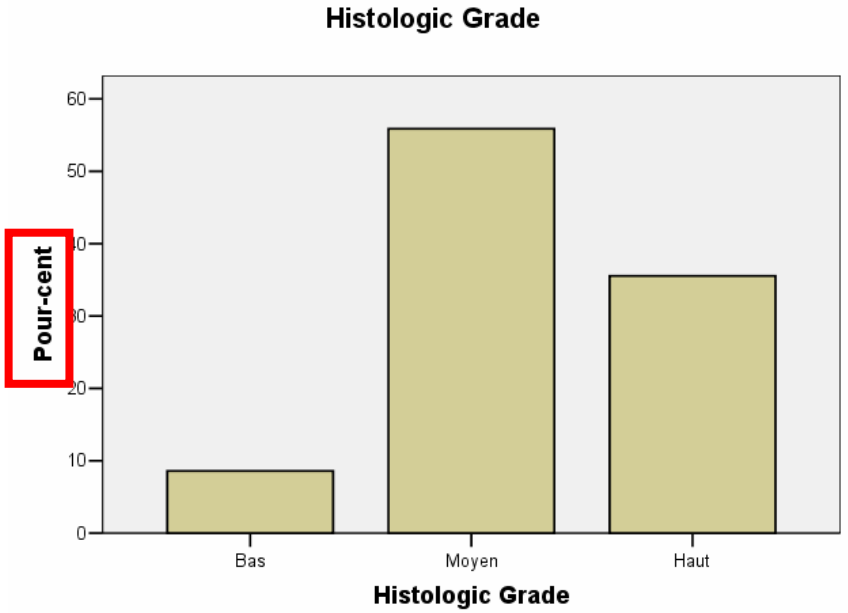


http://www.med.univ-rennes1.fr

Diagramme en barre

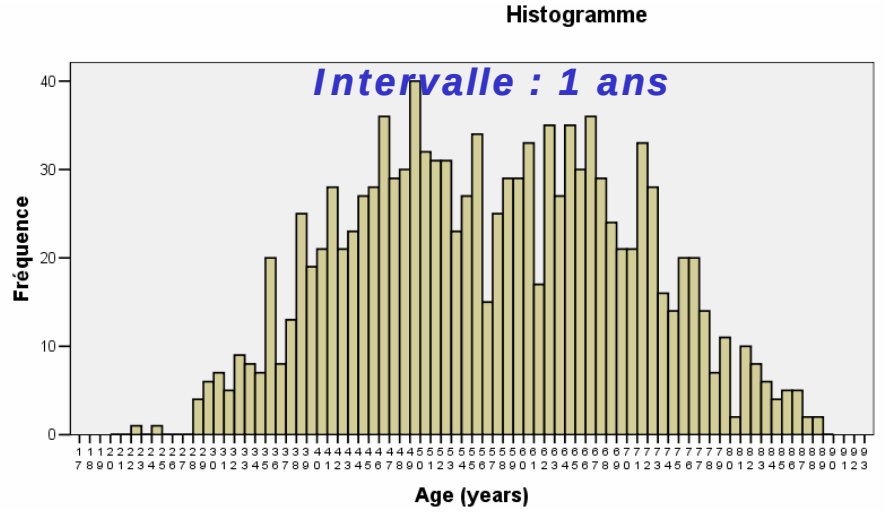
- Si on divise chaque hauteur par le nombre de sujet total de la population ou de l'échantillon, on conserve la même allure
- La hauteur ⇔ proportion de sujet dans la catégorie
- L'histogramme représente alors graphiquement l'ensemble des probabilités des différentes catégories ou classe de la variable.

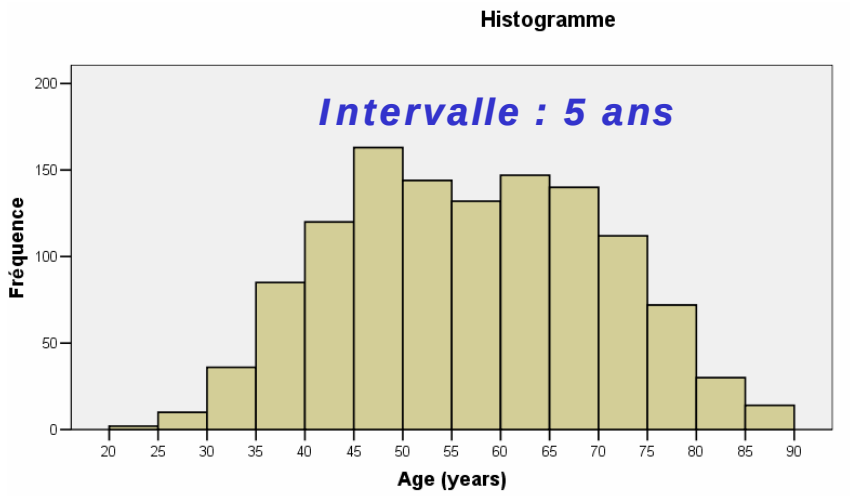


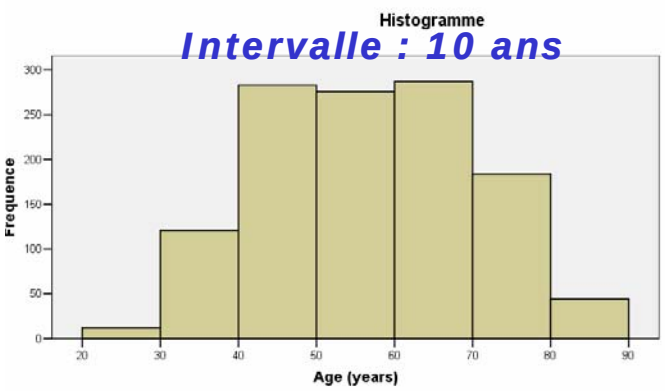


Histogramme

- **Pour les variables quantitatives**
  - **Il faut le plus souvent regrouper en classe**







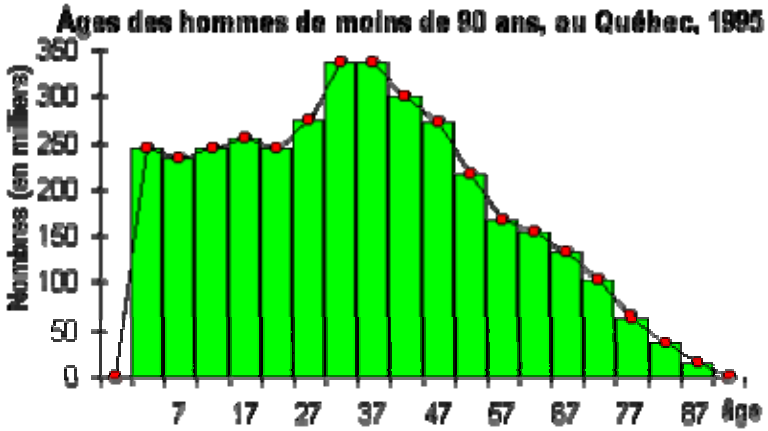
Laboratoire d'Informatique Médicale

Comment choisir les classes

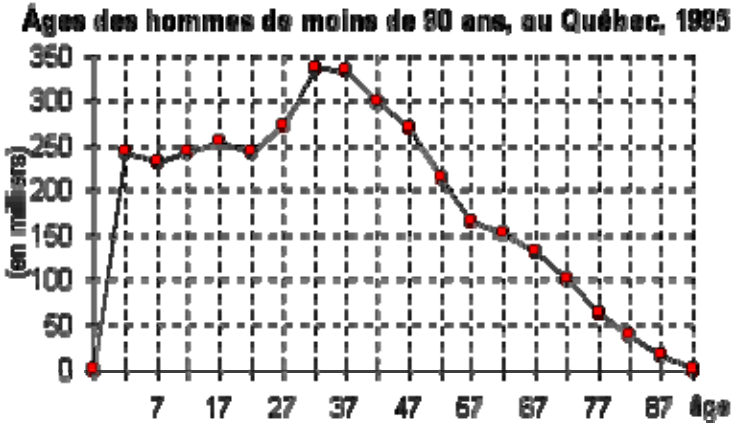
- En général : constituer des classes de largeurs égales en nombre assez grand pour représenter la répartition des sujets
- Mais pas trop pour qu'il y est suffisamment de sujet dans les classes.
- Plus le nombre de classe est grand, plus l'histogramme se rapproche d'une courbe continue

Polygonne de fréquence

- **Pour les variables quantitatives continues**

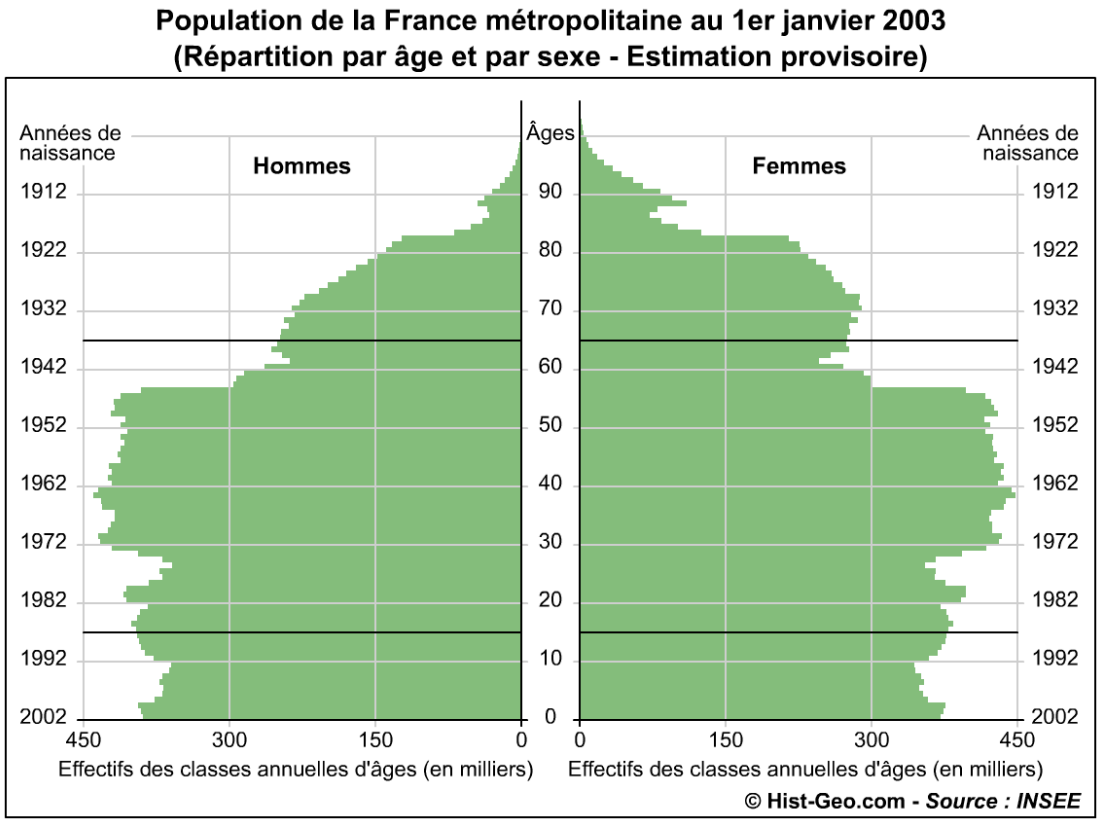


L'avantage de cette représentation est qu'on peut avoir plusieurs polygones des fréquences dans une même fenêtre. Cela fait mieux ressortir les comparaisons lorsque les variables sont nombreuses.



Pyramide des ages

- Utilisée pour montrer la distribution par age et par sexe d'une pop.
- Utilisé en démographie



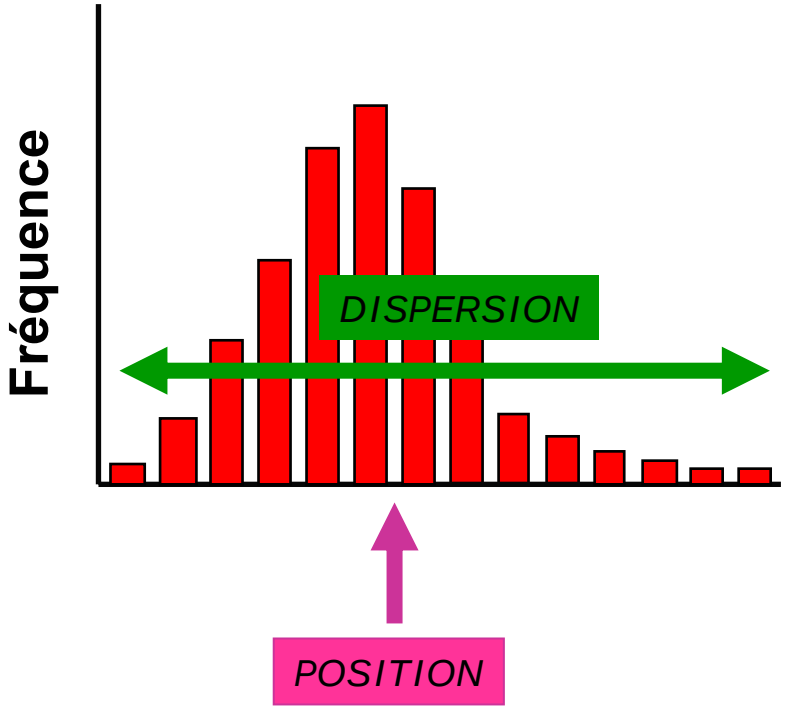
Mesures en statistiques

http://www.med.univ-rennes1.fr

Laboratoire d'Informatique Médicale

Problème général

- **Pour décrire les données, on peut**
  - **Établir des tableaux**
  - **Regrouper les données dans des classes**
  - **Dessiner des diagrammes**
- **Pour résumer les données afin de les exprimer ou les comparer**
  - **On calcule des paramètres (ou indicateurs)**
    - De POSITION
    - De DISPERSION



Paramètres

- **2 types :**
  - Paramètres de POSITION
    - Médiane
    - Quartiles, déciles, percentiles
    - Mode
    - Moyenne
    - Fréquences relatives
  - Paramètres de Dispersion
    - Extrêmes (Minimum, Maximum)
    - Entendue (Range)
    - Intervalle interquartile
    - Variance
    - Écart type
    - Coefficient de variation

Mesures en statistiques

- **Médiane**
  - **Est la valeur qui partage la série des individus en 2 groupes d'effectifs égaux.**
  - **La moitié des sujets présentent une valeur inférieure à la médiane. L'autre moitié une valeur supérieure à la médiane.**
  - **Calcul : nécessite de classer les sujets par ordre de valeur croissant.**
  - **Si la série est impaire, la médiane = valeur observée chez le sujet médian**
  - **Si la série est paire, médiane = moyenne des valeurs qui séparent en 2 la série**

- **Exemple : Calculez la médiane des deux échantillon suivants : 5 4 4 5 6 8 8 0 1**

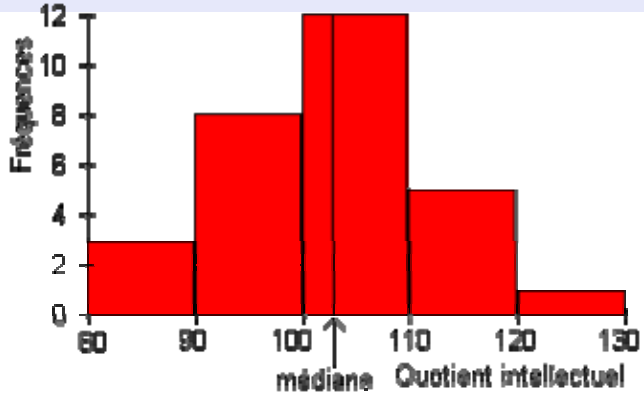
- **On ordonne les valeurs**
  - 0 1 4 4 5 5 6 8 8
  - Série impaire
- **On cherche la valeur séparant 50% des effectifs supérieurs et inférieurs**
  - 0 1 4 4 **5** 5 6 8 8
  - La médiane est 5

- **12 5 7 18 11 15 9 5**
- **On ordonne les valeurs**
  - 5 5 7 9 11 12 15 18
- **Nombre pair de valeurs : on cherche la moyenne des 2 valeurs séparant 50% des effectifs**
  - 5 5 7 **9** **11** 12 15 18
  - (9+11)/2 = 10
- **La médiane est 10**

Cas où l'on ne dispose que d'un tableau de fréquence

Distribution des quotients intellectuels

Quotient intellectuel	Fréquences	Fréquences relatives (en %)	Fréquences relatives cumulées (en %)
80 - 90 exclu	3	10,3	10,3
90 - 100 exclu	8	27,6	38
100 - 110 exclu	12	41,4	79
110 - 120 exclu	5	17,2	97
120 - 130 exclu	1	3,5	100
Total	29	100	

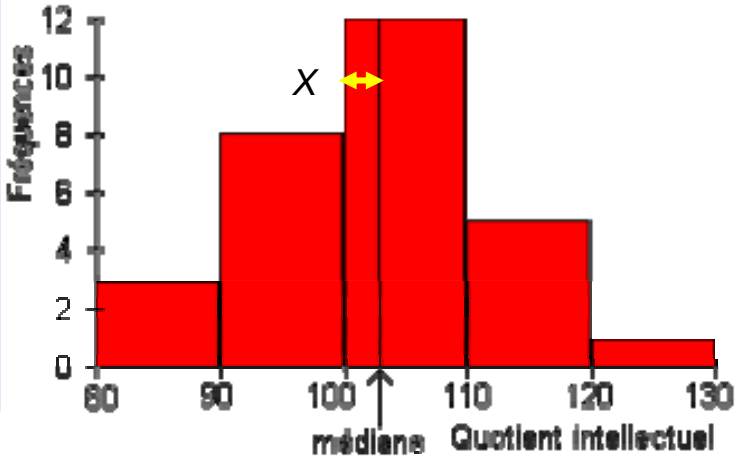


http://www.med.univ-rennes1.fr

Cas où l'on ne dispose que d'un tableau de fréquence

Distribution des quotients intellectuels

Quotient intellectuel	Fréquences	Fréquences relatives (en %)	Fréquences relatives cumulées (en %)
80 - 90 exclu	3	10,3	10,3
90 - 100 exclu	8	27,6	38
100 - 110 exclu	12	41,4	79
110 - 120 exclu	5	17,2	97
120 - 130 exclu	1	3,5	100
Total	29	100	

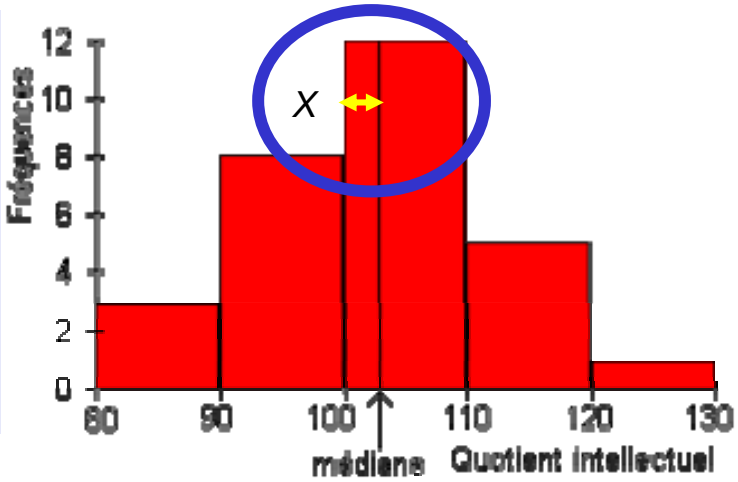


- La **classe médiane** est la classe où est située la médiane.

Cas où l'on ne dispose que d'un tableau de fréquence

Distribution des quotients intellectuels

Quotient intellectuel	Fréquences	Fréquences relatives (en %)	Fréquences relatives cumulées (en %)
80 - 90 exclu	3	10,3	10,3
90 - 100 exclu	8	27,6	38
100 - 110 exclu	12	41,4	79
110 - 120 exclu	5	17,2	97
120 - 130 exclu	1	3,5	100
Total	29	100	



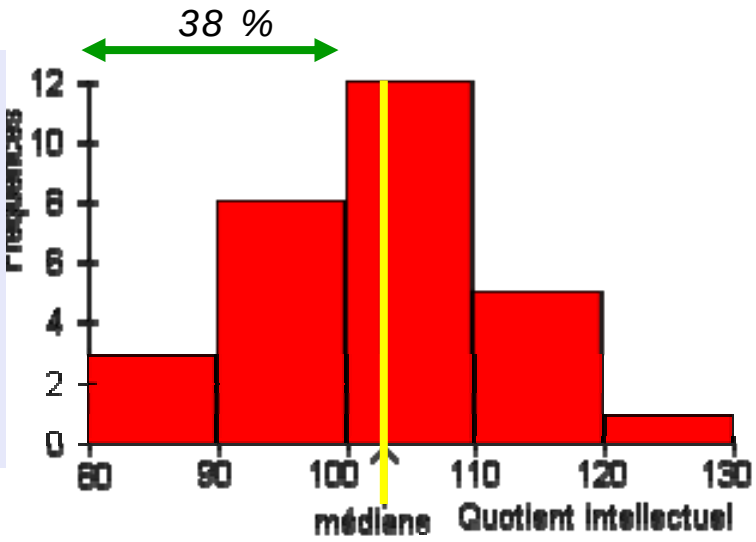
- On cherche la classe pour laquelle les fréquences cumulées
  - avant celle-ci sont plus petites ou égales à 50%
  - et après celle-ci plus grandes ou égales à 50%.
- Avant la classe 100 - 110, il y a 38% de données et après, on en a accumulé 79%.
- Donc la classe médiane est la classe 100 - 110.

- **Médiane = Borne inférieure de la classe médiane + longueur X**
- Calculer X ?

Cas où l'on ne dispose que d'un tableau de fréquence

Distribution des quotients intellectuels

Quotient intellectuel	Fréquences	Fréquences relatives (en %)	Fréquences relatives cumulées (en %)
80 - 90 exclu	3	10,3	10,3
90 - 100 exclu	8	27,6	38
100 - 110 exclu	12	41,4	79
110 - 120 exclu	5	17,2	97
120 - 130 exclu	1	3,5	100
Total	29	100	



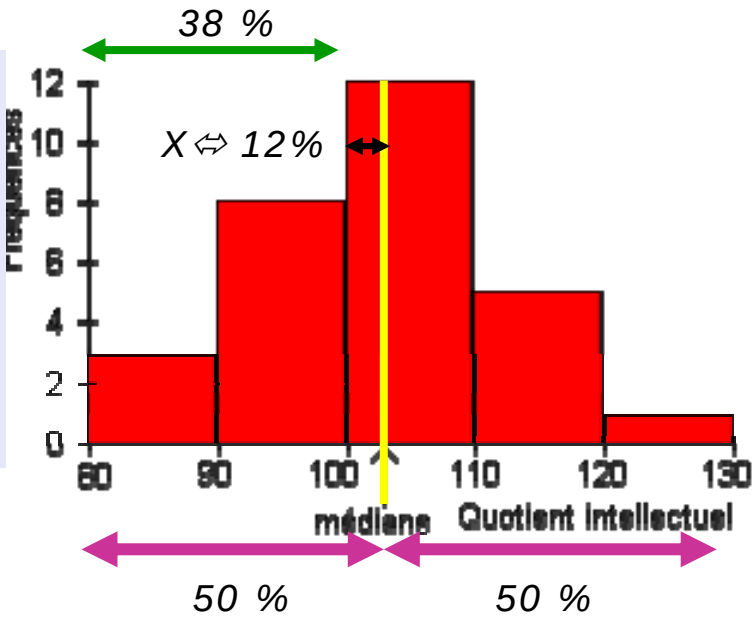
Calculer X

- 38 % des valeurs sont inférieurs à un QI de 100

Cas où l'on ne dispose que d'un tableau de fréquence

Distribution des quotients intellectuels

Quotient intellectuel	Fréquences	Fréquences relatives (en %)	Fréquences relatives cumulées (en %)
80 - 90 exclu	3	10,3	10,3
90 - 100 exclu	8	27,6	38
100 - 110 exclu	12	41,4	79
110 - 120 exclu	5	17,2	97
120 - 130 exclu	1	3,5	100
Total	29	100	



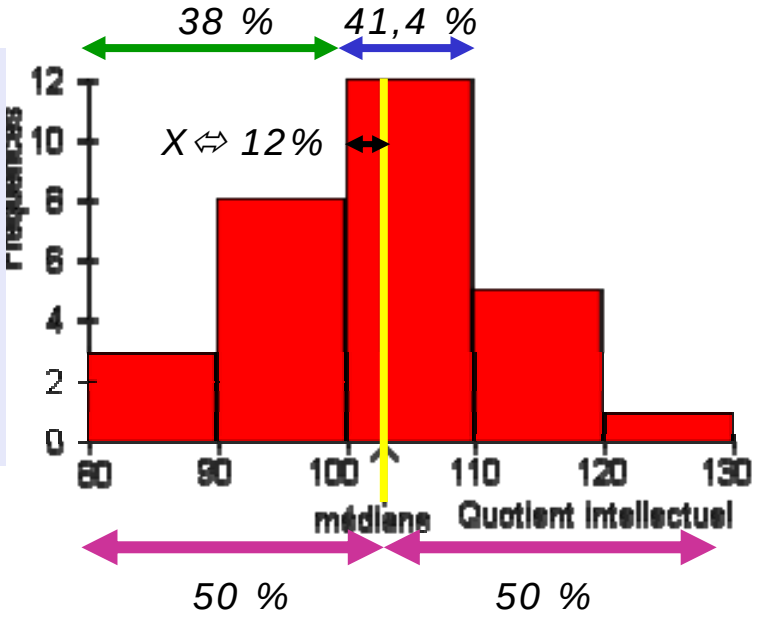
Calculer X

- 38 % des valeurs sont inférieurs à à un QI de 100
- La médiane sépare 50% des valeurs.
- La longueur X manquante ⇔ à 12% des données

Cas où l'on ne dispose que d'un tableau de fréquence

Distribution des quotients intellectuels

Quotient intellectuel	Fréquences	Fréquences relatives (en %)	Fréquences relatives cumulées (en %)
80 - 90 exclu	3	10,3	10,3
90 - 100 exclu	8	27,6	38
100 - 110 exclu	12	41,4	79
110 - 120 exclu	5	17,2	97
120 - 130 exclu	1	3,5	100
Total	29	100	



Calculer X

- 38 % des valeurs sont inférieurs à à un QI de 100
- La médiane sépare 50% des valeurs. La longueur X manquante ⇔ à 12% des données
- Or, La classe 100 -110
  - est de longueur 10 (110 - 100)
  - et contient 41,4% des données.

- À quelle longueur X correspond 12% des données?

Distribution des quotients intellectuels

Quotient intellectuel	Fréquences	Fréquences relatives (en %)	Fréquences relatives cumulées (en %)
80 - 90 exclu	3	10,3	10,3
90 - 100 exclu	8	27,6	38
100 - 110 exclu	12	41,4	79
110 - 120 exclu	5	17,2	97
120 - 130 exclu	1	3,5	100
Total	29	100	

- Il faut utiliser une règle de 3 pour faire le calcul :

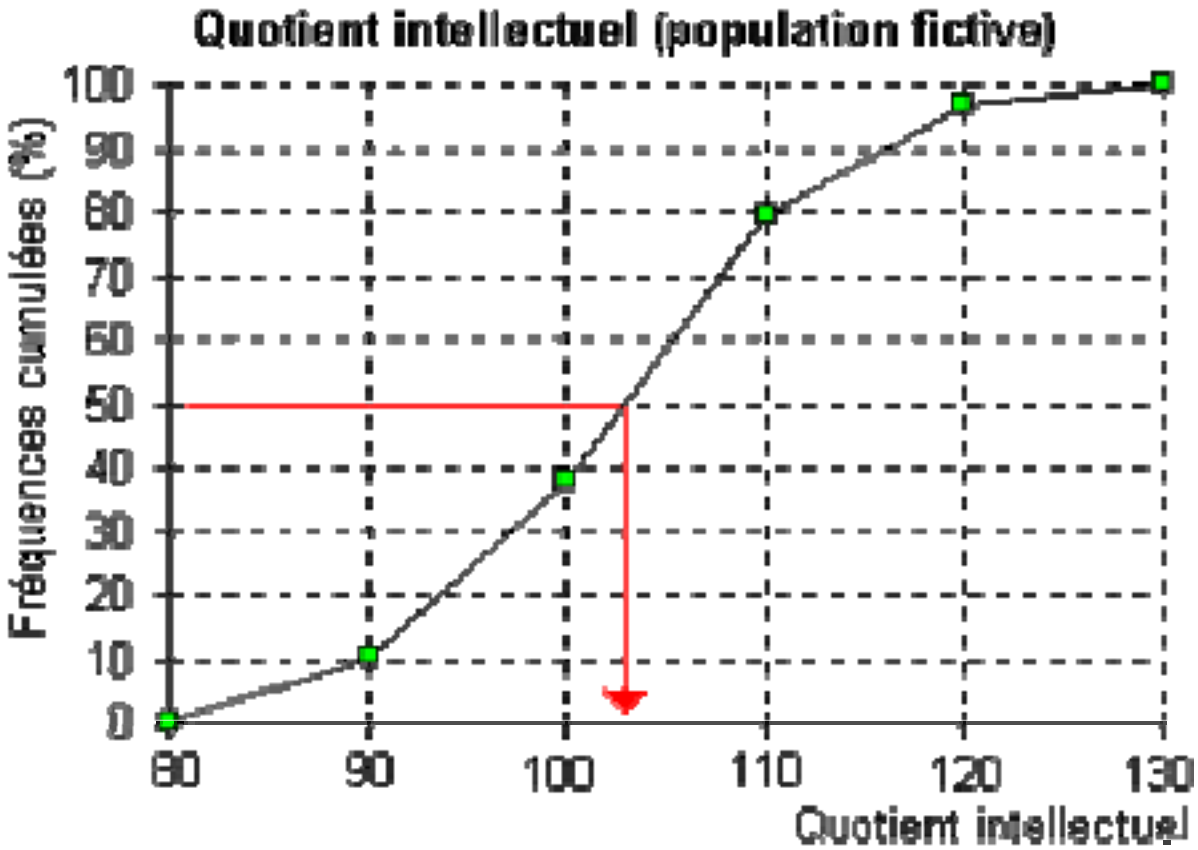
Longueur *%*
10 ➔ *41,4%*
X ➔ *12 %*

$X = 12 \times 10 / 41,4$

$X = 2,9$

Médiane = classe médiane + X
Médiane = 100 + 2,9 = 102,9

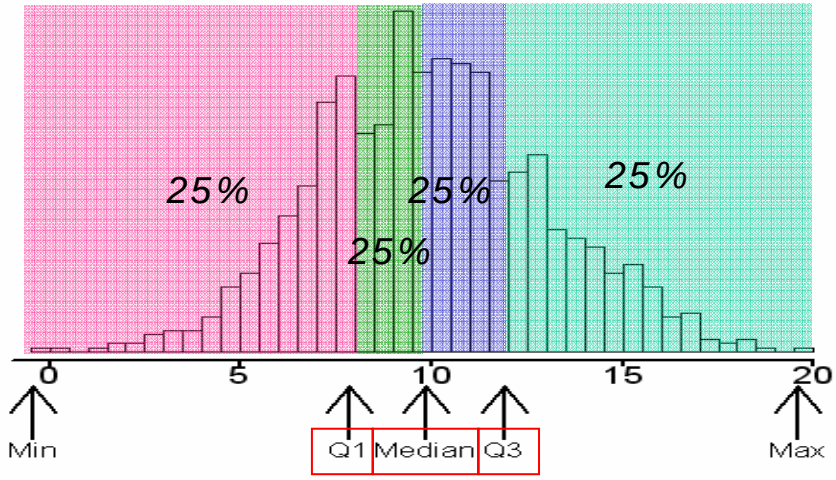
Détermination graphique de la médiane



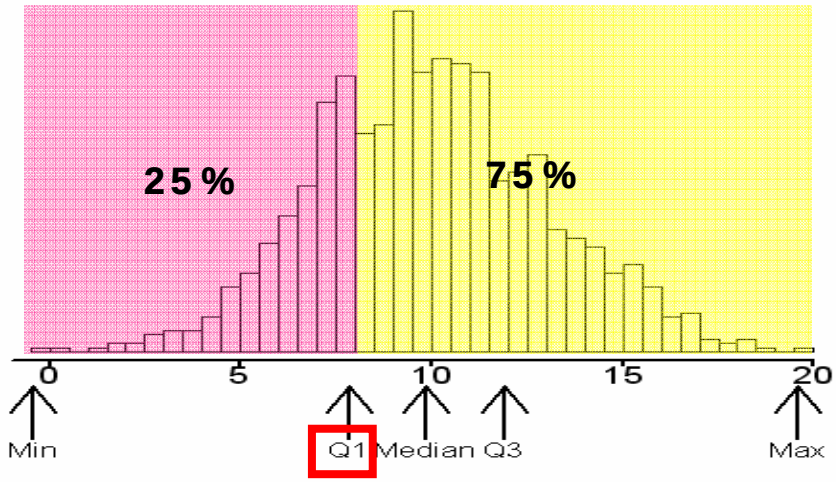
http://www.med.univ-rennes1.fr

Quartiles

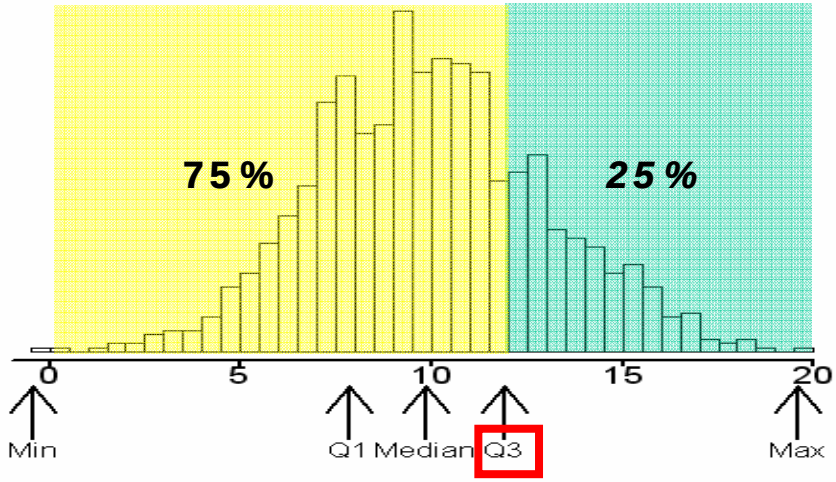
- Sont les 3 valeurs qui partagent la distribution en 4



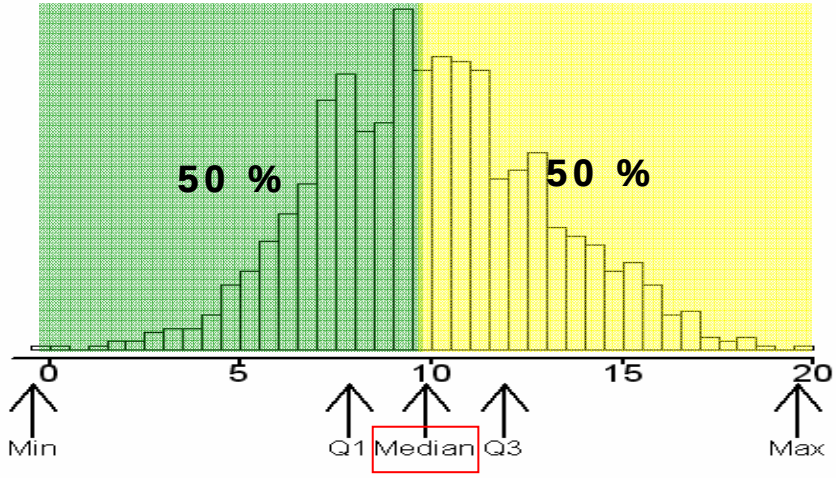
- **1er quartile : sépare 25% des valeurs les plus faibles et 75% des valeurs les plus élevés**



- **3 ème quartile : sépare 75% des valeurs les plus faibles et 25% des valeurs les plus élevés**



- Le deuxième quartile sépare 50 % des valeurs les plus faible de 50% des valeurs les plus élevées
- 2ème quartile ⇦ Médiane !



Quartiles

Q1 ⇨ effectif cumulé croissant N/4

Q2 ⇨ à Me

Q3 ⇨ 3N/4

Classes	Effectifs	Fréquences	%
X	F	f = F/n	100.f
[40-45[	5	0,05	5,0
[45-50[	12	0,12	12,0
[50-55[	31	0,31	31,0
[55-60[	31	0,31	31,0
[60-65[	16	0,16	16,0
[65-70[	3	0,03	3,0
[70-75[	2	0,02	2,0
Total	n=100	1,00	100,0

Classes	Effectifs	Fréquences	%
X	F	f = F/n	100.f
[40-45[	5	0,05	5,0
[45-50[	12	0,12	12,0
[50-55[	31	0,31	31,0
[55-60[	31	0,31	31,0
[60-65[	16	0,16	16,0
[65-70[	3	0,03	3,0
[70-75[	2	0,02	2,0
Total	n=100	1,00	100,0

- **N/4 = 100/4 = 25**
- **Classe qui contient 1er quartile est celle immédiatement au dessus des 25% inférieurs cumulés**
- **ICI c'est la classe [50-55]**

Classes	Effectifs	Fréquences	%
X	F	f = F/n	100.f
[40-45[	5	0,05	5,0
[45-50[	12	0,12	12,0
[50-55[	31	0,31	31,0
[55-60[	31	0,31	31,0
[60-65[	16	0,16	16,0
[65-70[	3	0,03	3,0
[70-75[	2	0,02	2,0
Total	n=100	1,00	100,0

- **Q1= Borne inférieur de la classe Q1 + X**
- **X?**
- **Régle de 3 :Longueur de classe %**

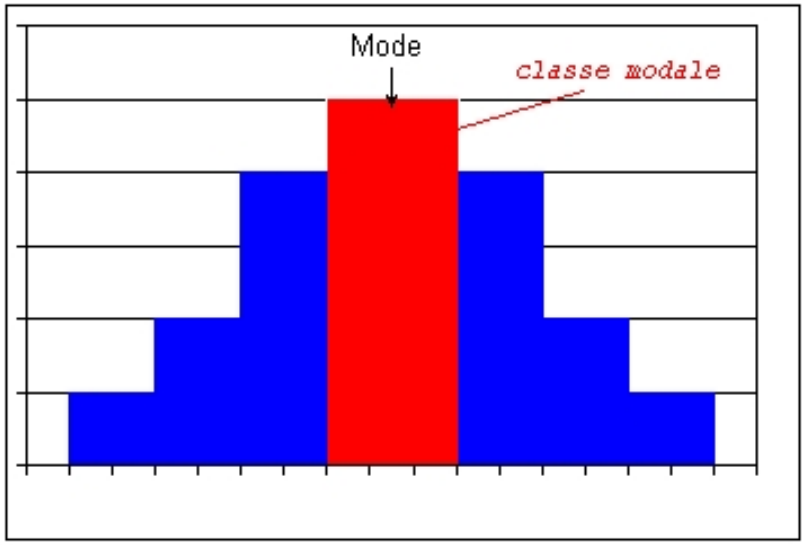
5	➔	31%
X	➔	(25%-17%)

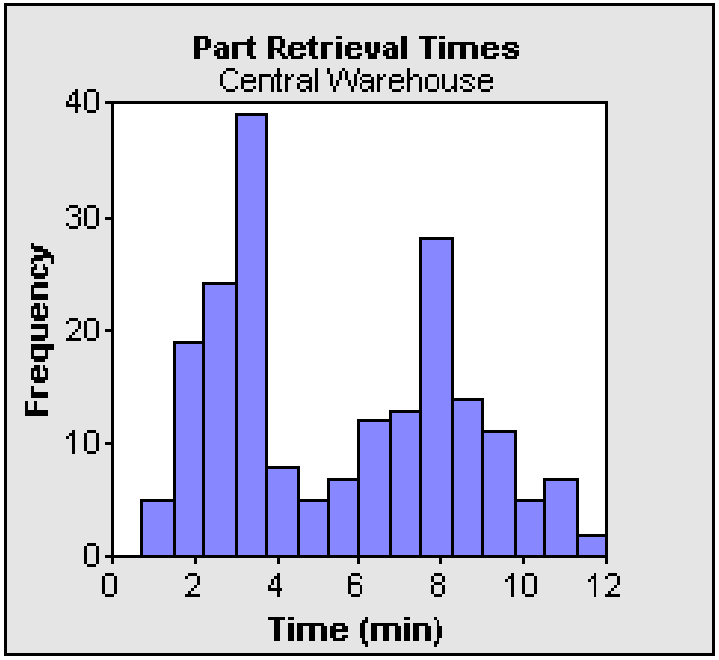
- **D'où Q1 = 50 + (25-17).5/31= 51,29**

- **Déciles**
  - **Idem : 9 valeurs qui partagent la distribution en 10 groupes de tailles égales.**
- **Percentiles**
  - **Sont les valeurs qui partagent la distribution en 100 groupes de tailles égales**
  - **Le percentile 10% ⇔ au 1er décile**
  - **Le percentile 25 % ⇔ au 1er quartile**
  - **Le percentile 50 % ⇔ à la médiane**

MODE

- **Modes**
- **Dans une distribution comportant de nombreuses données, le mode est la valeur qui revient le plus souvent**



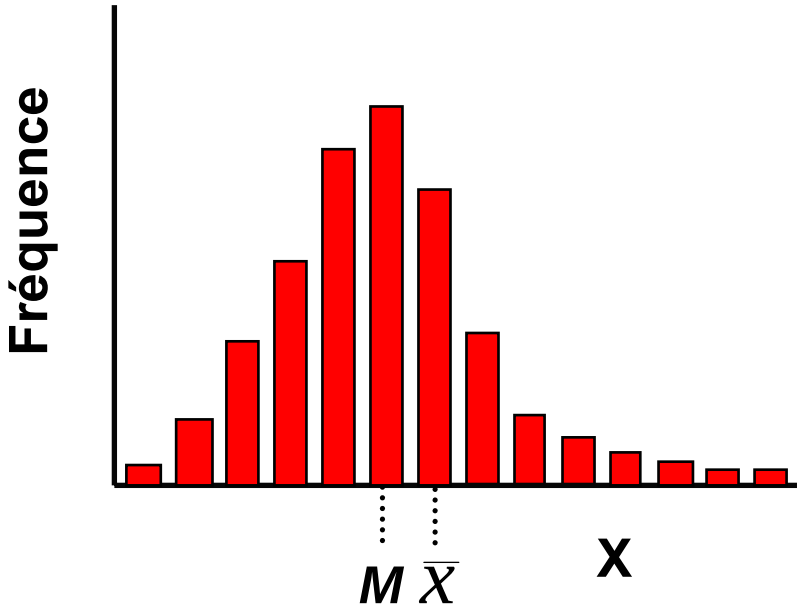


- **But uniquement descriptif**

Moyenne

- **Moyenne**
- **Indicateur de tendance centrale servant à résumer une série de données d'une variable quantitative**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Autre formule :

$$m = \frac{\sum_{i=1}^{n} fiXi}{\sum_{i=1}^{n} fi}$$

Age(Xi)	fi
2	1
5	3
6	4
8	2
somme:10	

m=(1x2+5x3+6x4+8x2)/10

m = 57/10 = 5,7 ans

- **La somme des écarts à la moyenne = 0**

$$\sum_{i=1}^{n}(Xi - m) = 0$$

Age(Xi)	fi	m=5,7
2	1	
5	3	
6	4	
8	2	
	somme : 10	

(2-5,7)+3x(5-5,7)+4x(6-5,7)+2x(8-5,7)
-3,7 – 3x0,7 + 4x0,3 + 2x2,3
-3,7 – 2,1 + 1,2 + 4,6
-5,8 + 5,8 = 0


Laboratoire d'Informatique Médicale
HTTP://WWW.MED.UNIV-RENNES1.FR
UNIVERSITE DE RENNES 1
FACULTÉ DE MÉDECINE

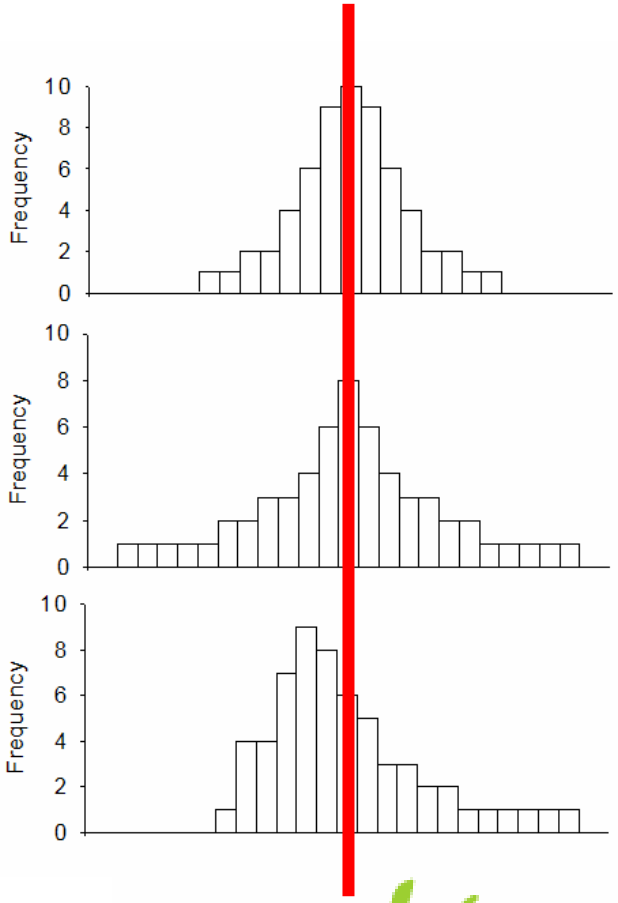
Dispersion

Dispersion

- Paramètre centraux ne résument pas complètement une distribution.

- Les paramètre mesurant la dispersion
  - Min Max
  - Étendue (range)
  - Espace interquartile (entre 1 et 3ème)

  - VARIANCE
  - ECART TYPE



Dispersion

- **Min Max :**
  - **Très sensible aux valeurs extrêmes**
  - **Permet de détecter les erreurs**
- **Étendue : Valeur Max – Valeur min**
- **Espace interquartiles**
  - **Qi = Q3 – Q1**
  - **contient 50% des valeurs de la série**

Variance

- **Variance :**
- **Caractériser l'écart de l'ensemble des valeurs**

- **Pour une valeur x_i, l'écart par rapport à la moyenne est :**

$$\Delta = (x - \mu)$$

- **les écarts étant de signe + ou -, on considère le carré des écarts**

$$(x - \mu)^2$$

- **Est la moyenne de la somme des carrés des écarts à la moyenne**

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

- **σ^2 = variance de la population (N)**

Variance

- **Variance d'un échantillon :**

$$s^2 = \frac{\sum (x - m_x)^2}{n - 1}$$

- **Si on considère une population ➔ on calcule σ^2**
- **Si on considère un échantillon ➔ on calcule S^2**

Exemple **: poids d'une population de 100 femmes**

moyenne = 54,9 kg

X	F	(X - 54,9)	$(X- 54,9)^2$	$F.(X- 54,9)^2$
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**

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

d'où $\sigma^2 = 3809 / 100$

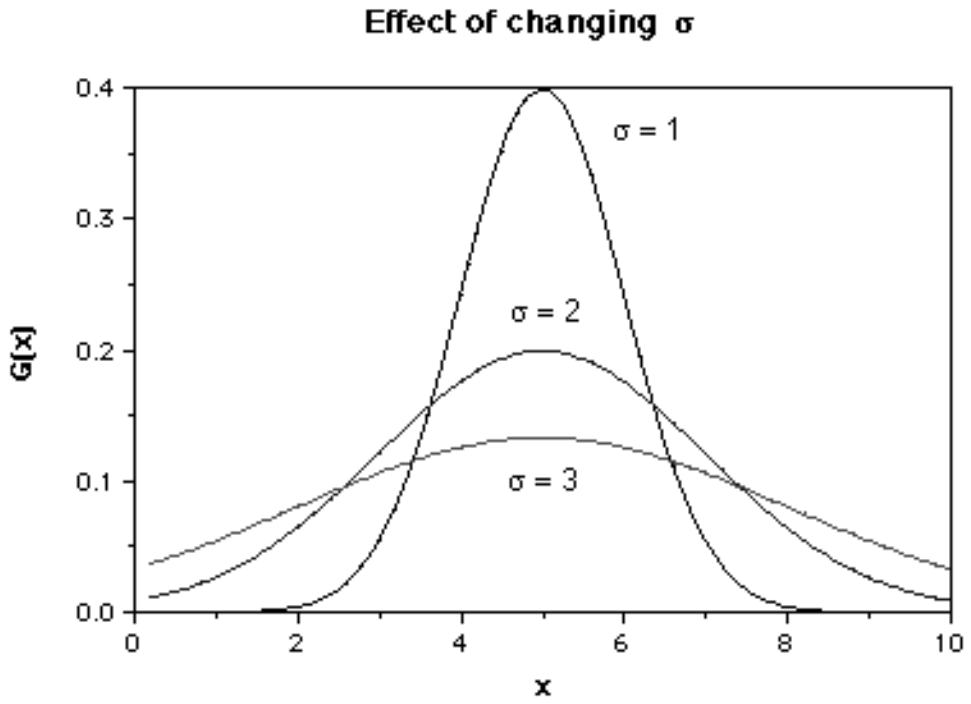
$\sigma^2 = 38,09$

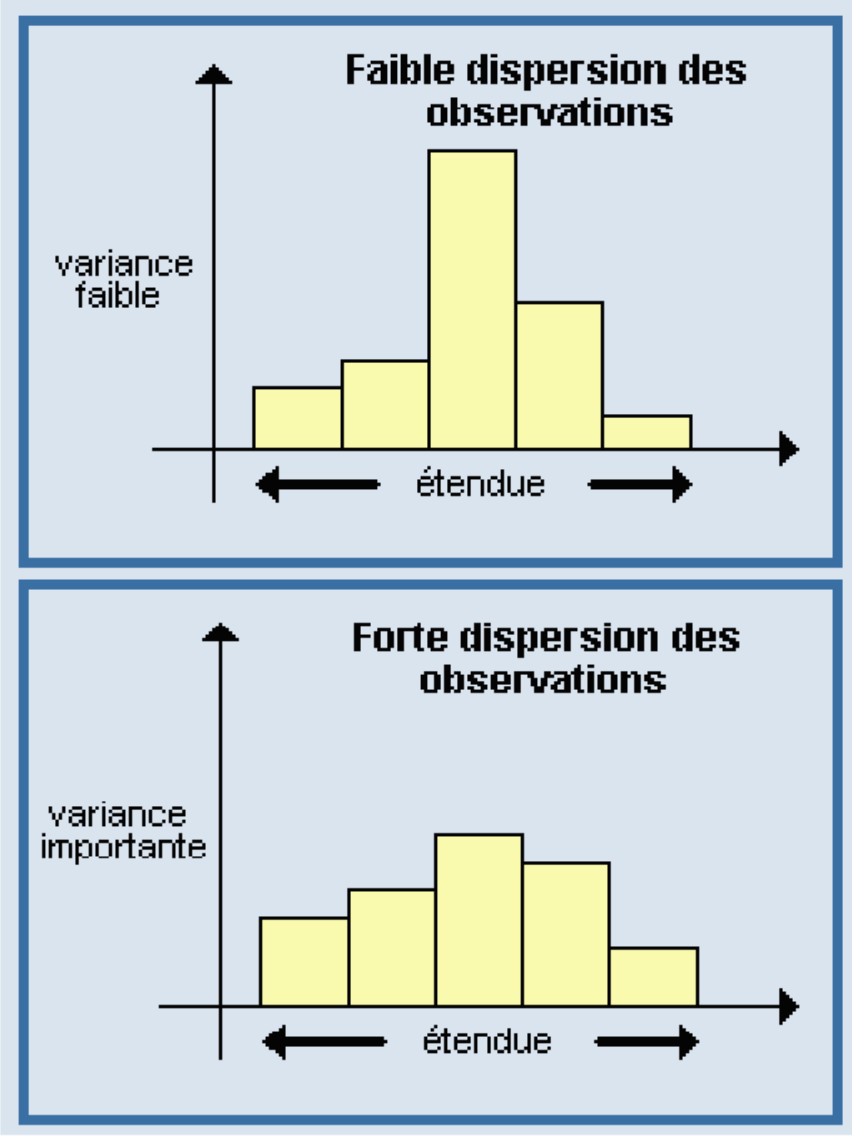
- **Autre formule de la variance :**

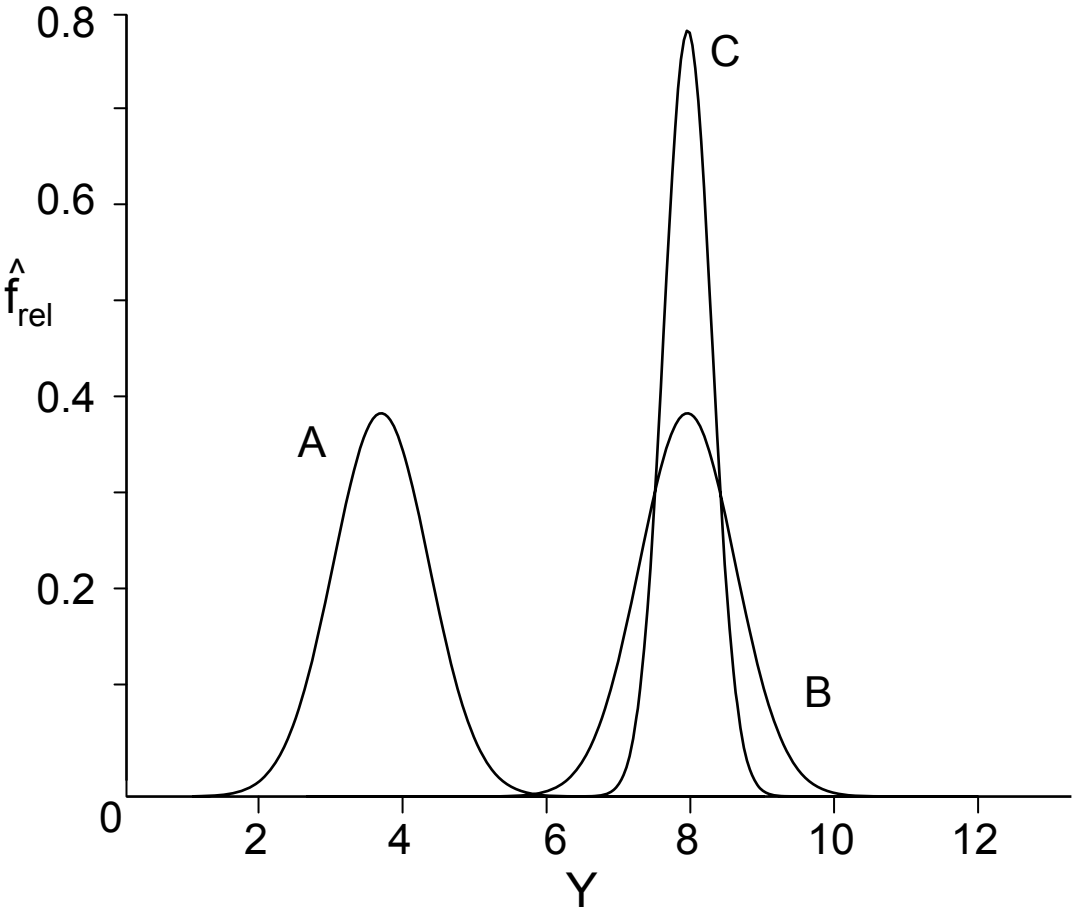
$$s^2 = \frac{\sum (X_i^2) - \frac{\sum (Xi)^2}{n}}{n-1}$$

- **Écart type :**
  - **D'une population** $\sigma = \sqrt{\sigma^2}$
  - **D'un échantillon** $s = \sqrt{s^2}$
  - **Écart type = même grandeur que la moyenne.**

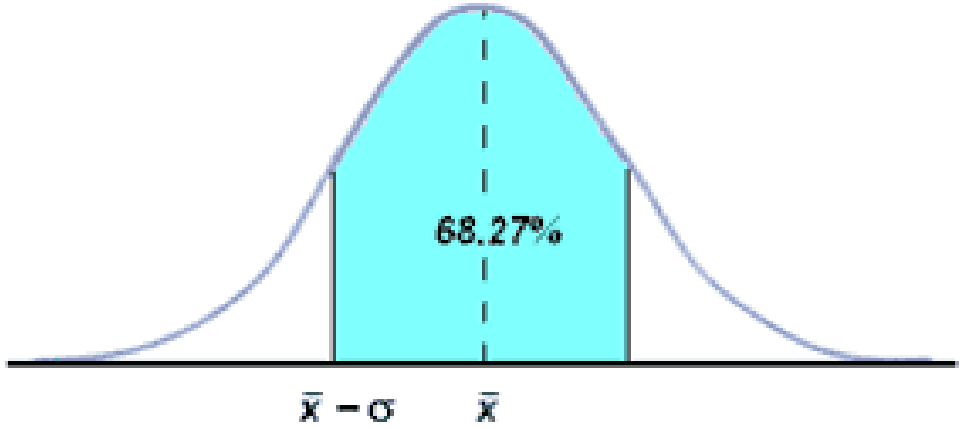
m±s

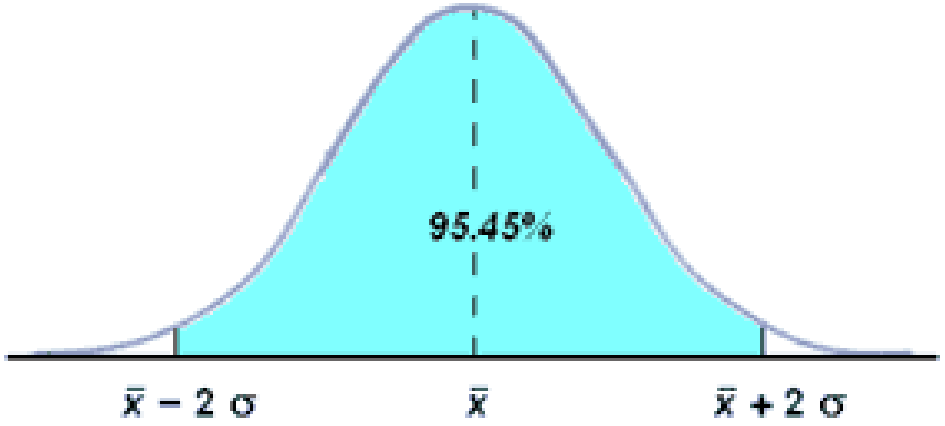

Effect of changing σ
G(x)
x
σ = 1
σ = 2
σ = 3
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10


Faible dispersion des observations
variance faible
étendue
Forte dispersion des observations
variance importante
étendue



- **Des changements pour les valeurs de la moyenne et la variance entraînent des changements dans la forme et la position de la distribution normale.**
- **A. $\mu = 4$, $\sigma = 1$**
- **B. $\mu = 8$, $\sigma = 1$**
- **C. $\mu = 8$, $\sigma = 0.5$**


68.27%
$\bar{x} - \sigma$
$\bar{x}$


95.45%
$\bar{x} - 2\sigma$
$\bar{x}$
$\bar{x} + 2\sigma$

Variable qualitative à 2 classes : Pour la population

- **Proportion d'une modalité K**
- **multipliée par 100 pour l'exprimer en pourcentage.**
- **N : taille de la population**

$$P = \frac{K}{N}$$

- **Variance : Produit de la proportion par son complément à 1**

$$\sigma^2 = P(1-P)$$

- **Ecart type : racine carrée de la variance**

$$\sigma = \sqrt{P(1-P)}$$

Variable qualitative à 2 classes : Pour un échantillon

- **Proportion d'une modalité (k) .**
- **multipliée par 100 pour l'exprimer en pourcentage.**
- **n=taille de l'échantillon**

$$p = \frac{k}{n}$$

- **Variance : Produit de la proportion par son complément à 1**

$$s^2 = p(1-p)$$

- **Ecart type : racine carrée de la variance**

$$s = \sqrt{s^2} = \sqrt{p(1-p)}$$

Exemple

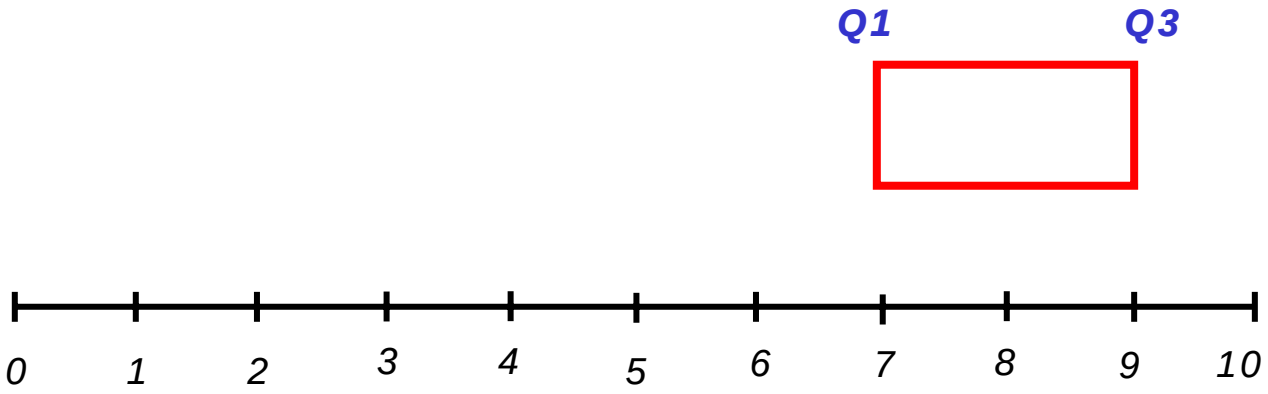
- **On considère un échantillon de 60 sujets, il y a 20 malades, les autres sont sains.**
- **Calculer**
  - **les proportions**
    - de malades
    - de non malades
  - **La variance S^2**
  - **L'écart type S**
- **$P_{malades}$=20/60 = 0,33**
- **$P_{non\ malades} = (1-P_{malades}) = 1- 0,33 = 0,67$**
- **$S^2=P_{malades}(1-P_{malades})= 0,33x0,67=0,221$**
- **S=0,47**

BOXPLOT ou Boîte à moustache

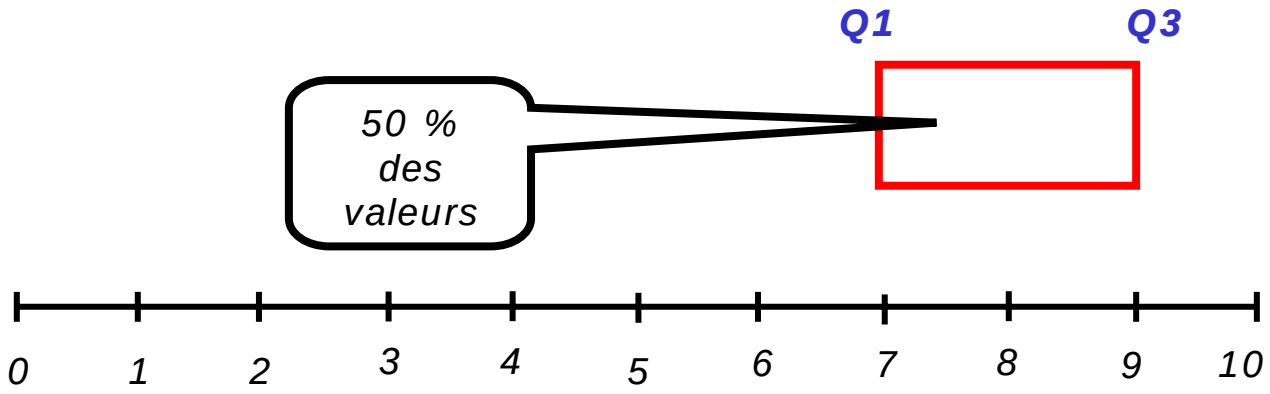
http://www.med.univ-rennes1.fr

Laboratoire d'Informatique Médicale

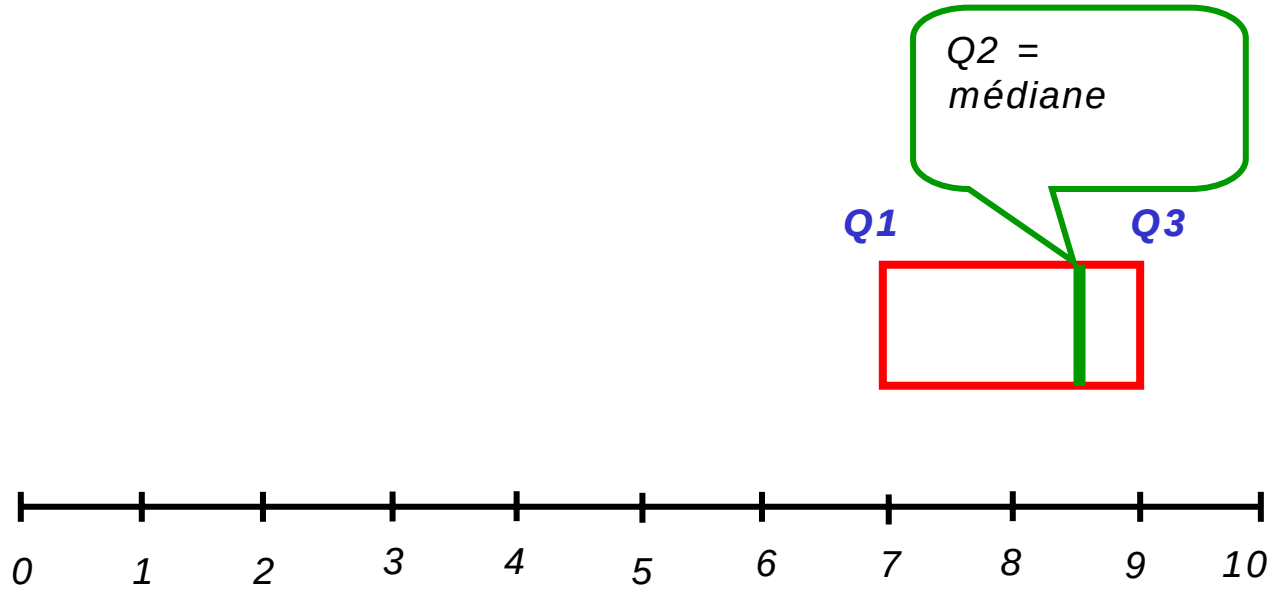
- **Représentation simple mais puissante d'un échantillon de données constituée**
  - **D'un rectangle (box) orienté selon un système de coordonnés**
  - **L'échelle de l'axe est celle des données**
  - **Les limites inférieures et supérieures correspondent au respectivement au 1er et 3ème QUARTILE**



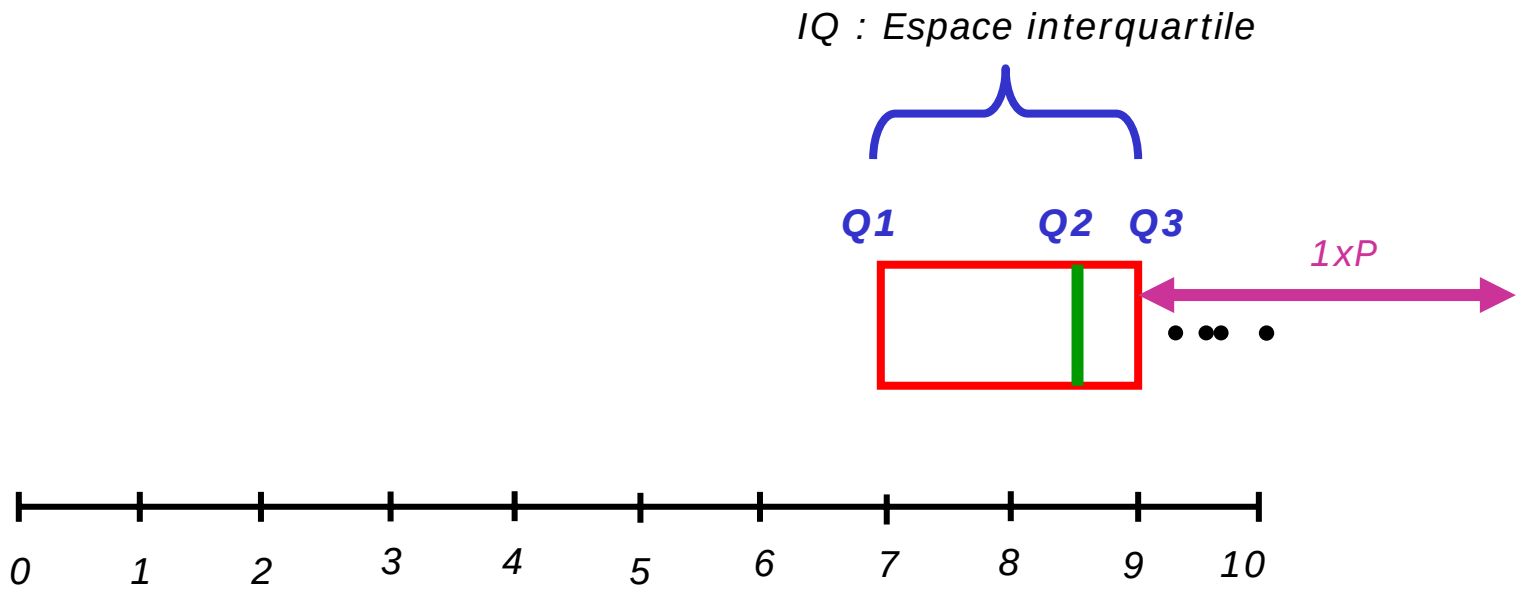
- ➔ **ainsi la boîte contient 50 % des valeurs.**



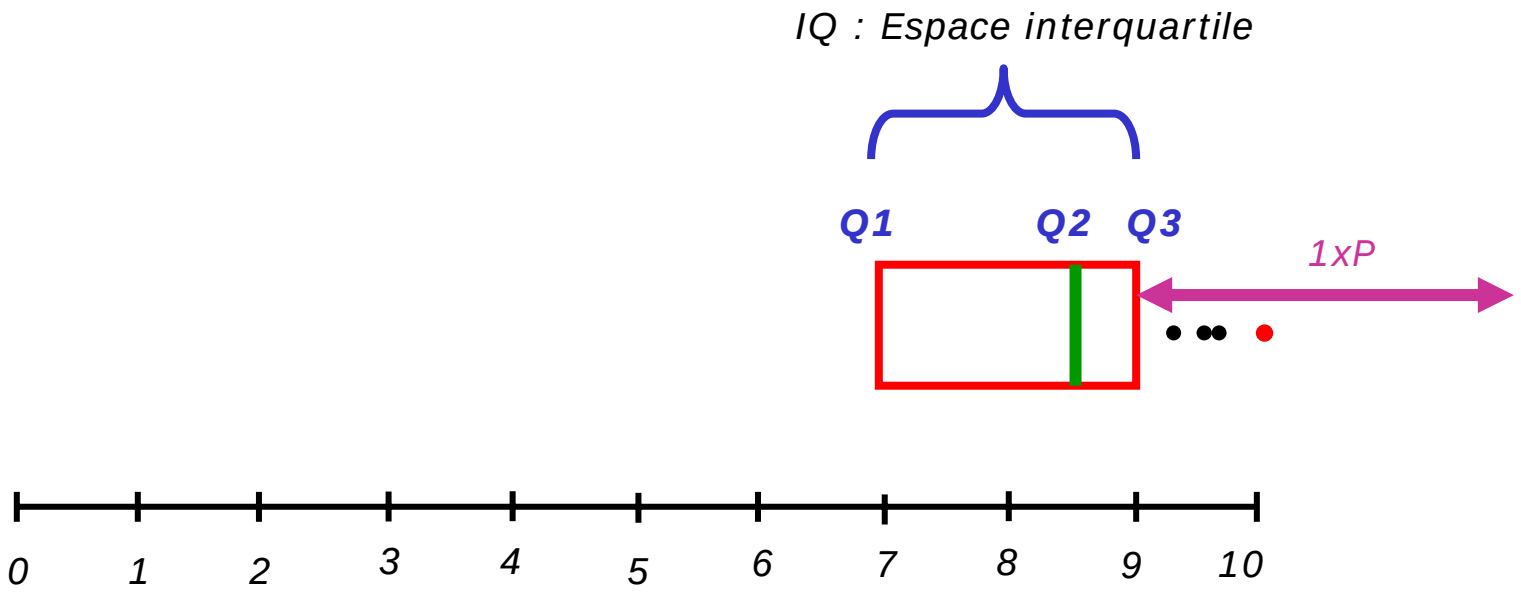
- **Le rectangle est partagé en 2 par un trait horizontal au niveau de la médiane**



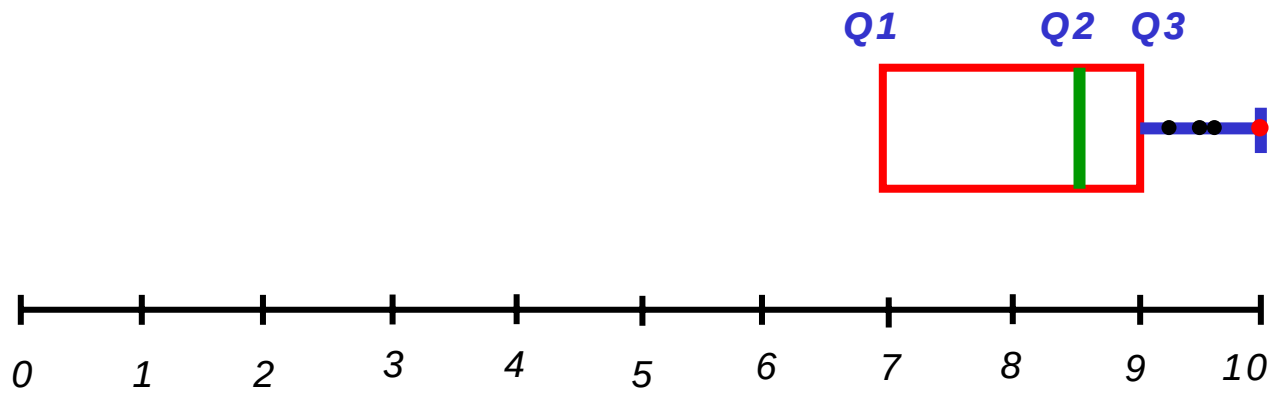
- On définit un pas P=1,5xIQ
- On considère les données située entre le sommet + 1 P



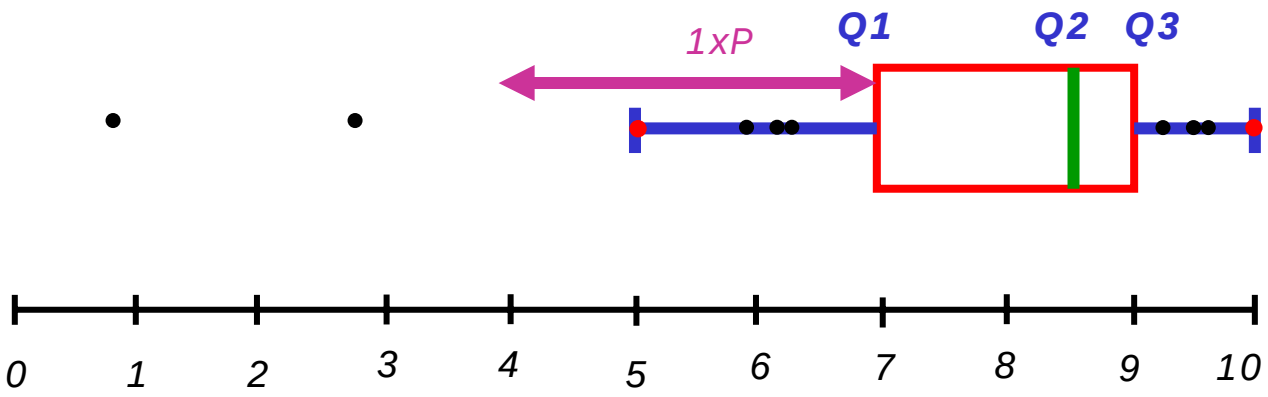
- On définit un pas P=1,5xIQ
- On considère les données située entre le sommet + 1 P



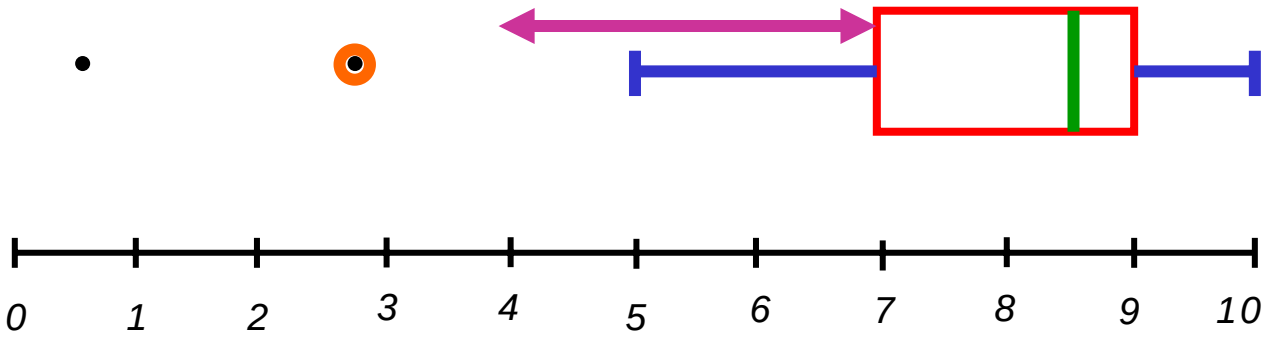
- **Un trait s'etend du milieu du sommet jusqu'à la limite supérieure**



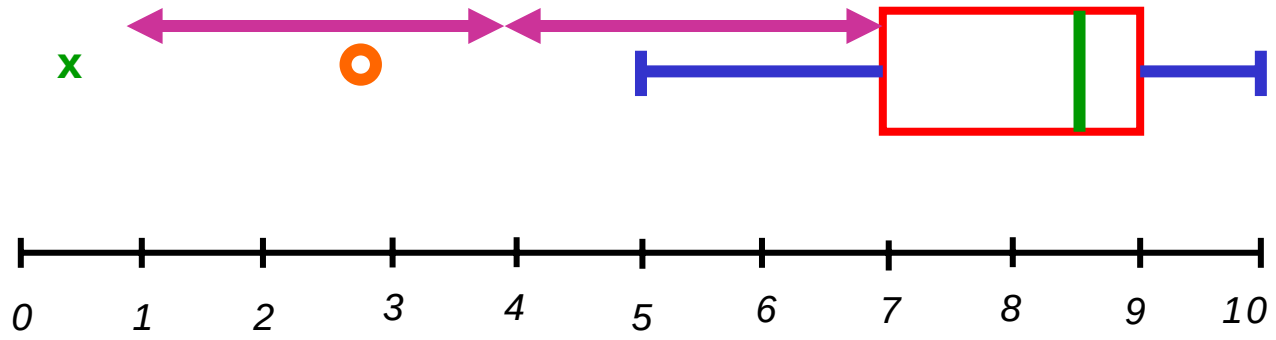
- **De manière symétrique on trouve la limite inférieure.**



- **Les observations les plus éloignées qui dépassent les limites sont marquées individuellement « O » pour outliers**

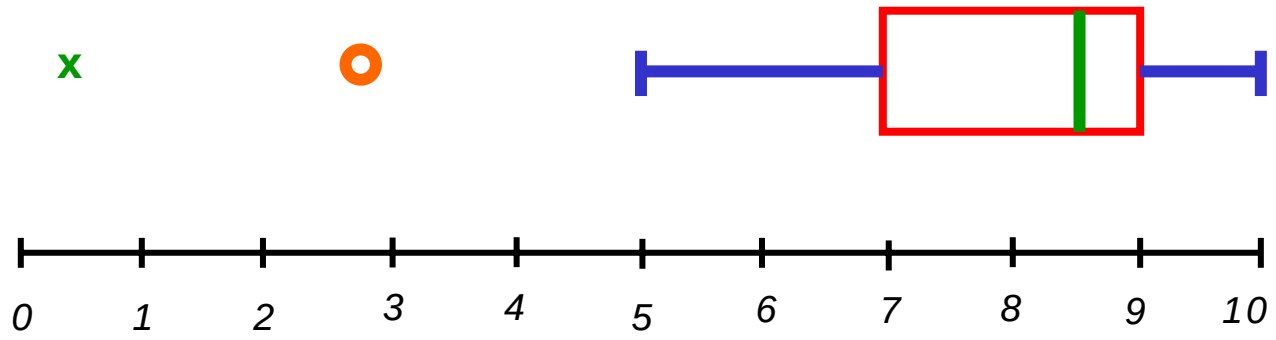


- **Les observations les plus éloignées qui dépassent les limites sont marquées individuellement « O » pour outliers**
- **Cellent qui dépasse de 2 pas sont considérées comme extrême et sont notés E ou x**

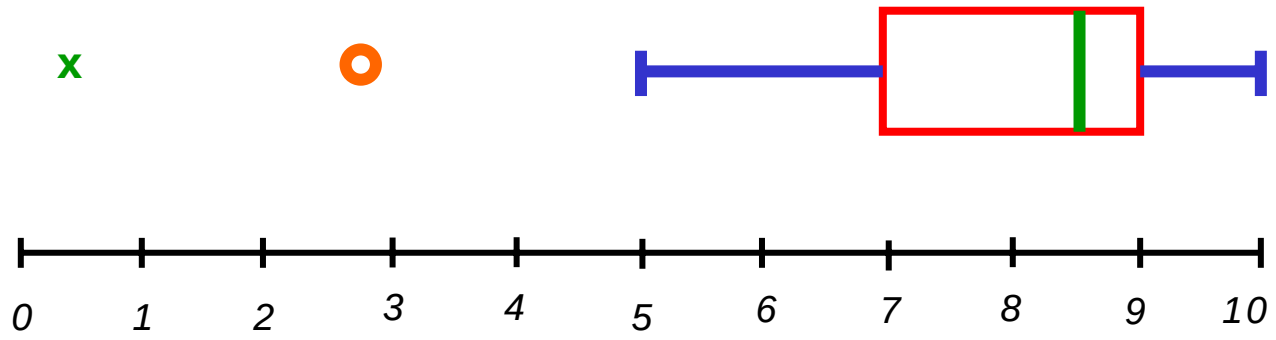


Intêret d'une boxplot

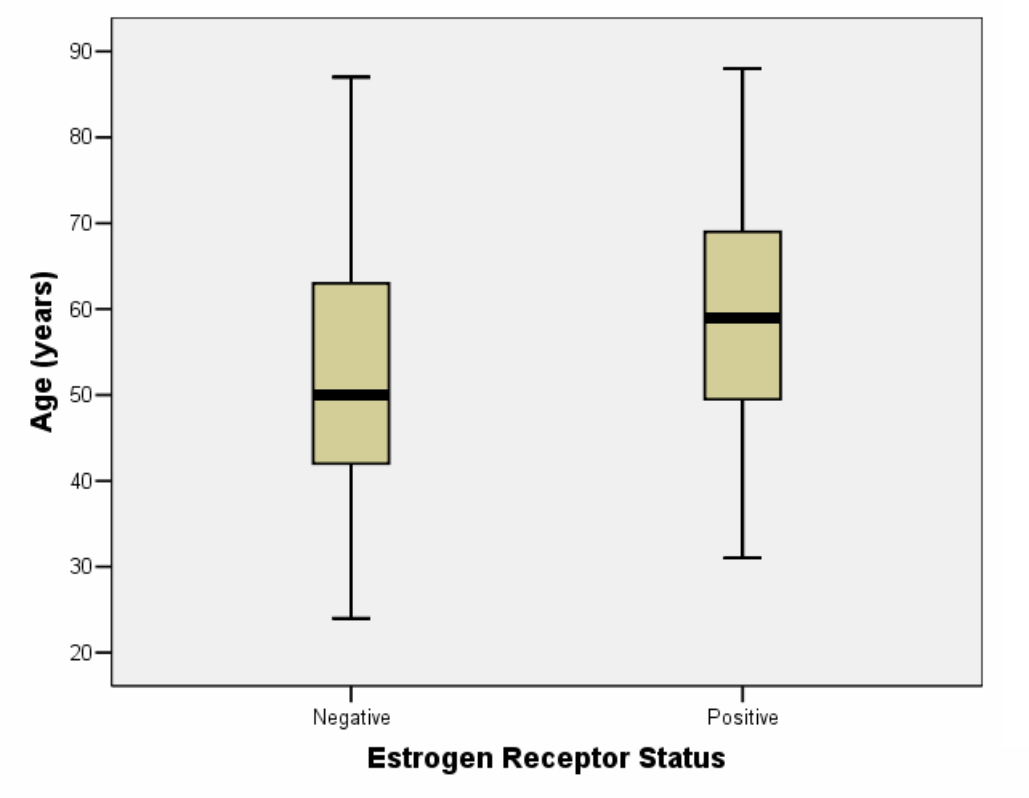
- Cinq synthèses numériques (mediane, quartiles, limites) sont représentées de façon à visualiser les informations essentielles (position, dispersion, asymétrie) de l'échantillon.
- La position est celle du box, en particulier.
- La dispersion est visualisée par la longueur du box ainsi que par écart entre les limites.
- La position du trait transversal dans le box et la différence entre les moustaches nous renseignent sur le degré d'asymétrie.



- Enfin, la fréquence et la position des outliers indiquent si l'échantillon est particulièrement étalé
- Les outliers sont souvent très intéressantes (cas exceptionnels, erreurs de mesure ou de codage, etc.).



- Plusieurs échantillons peuvent être représentés simultanément et comparés par des box-plots les uns à côté des autres.



Références

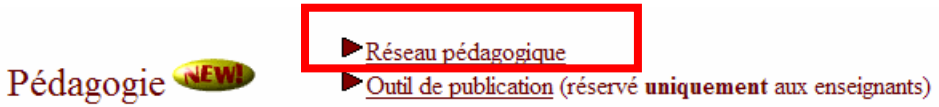
- Statistiques Epidémiologique. T Ancelle (Maloine)
- Méthode Statistiques Médecine –Biologie. Jean Bouyer (ESTEM)

Accéder au réseau pédagogique

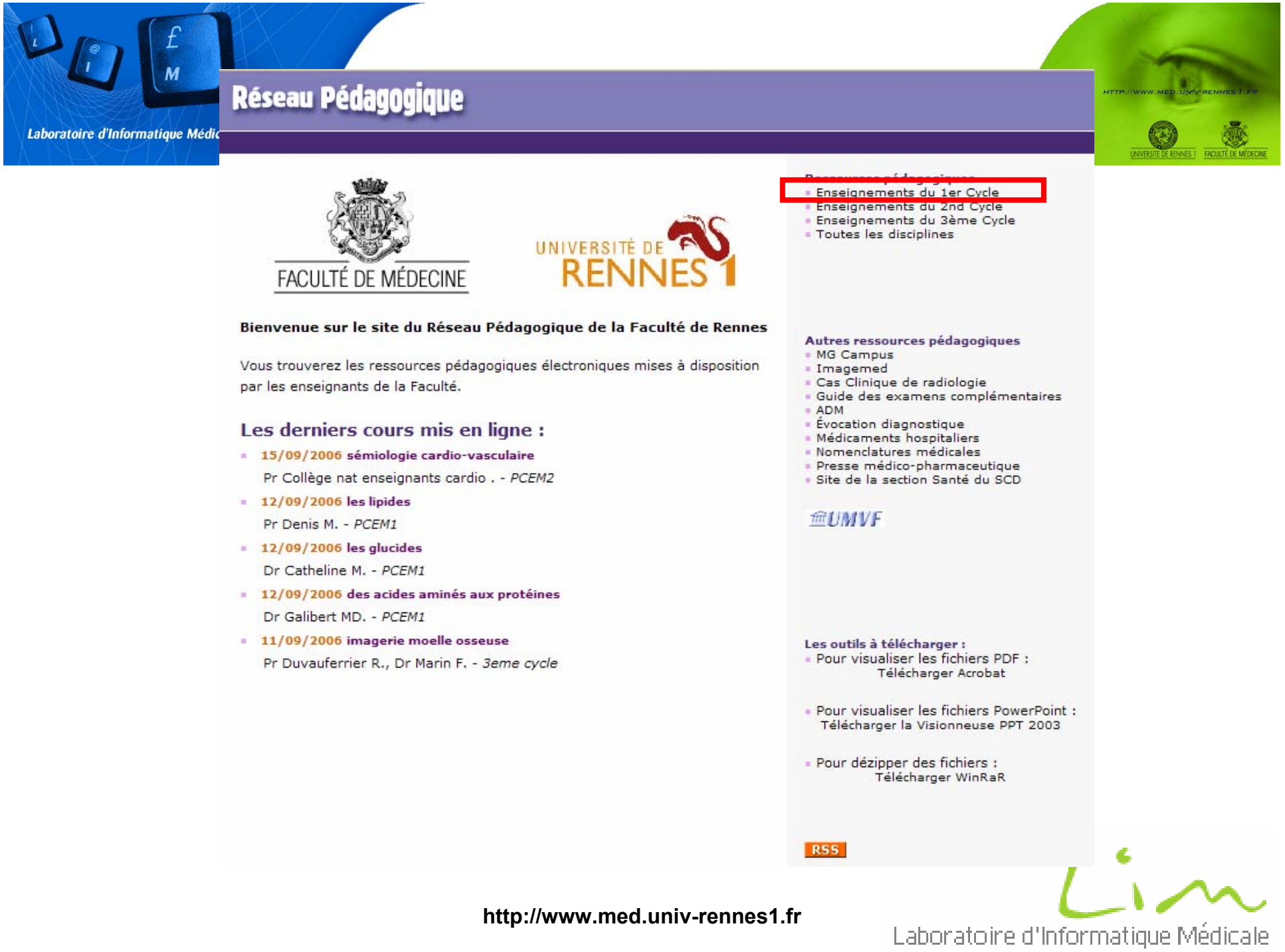
http://www.med.univ-rennes1.fr

Laboratoire d'Informatique Médicale

http://www.fac.med.univ-rennes1.fr

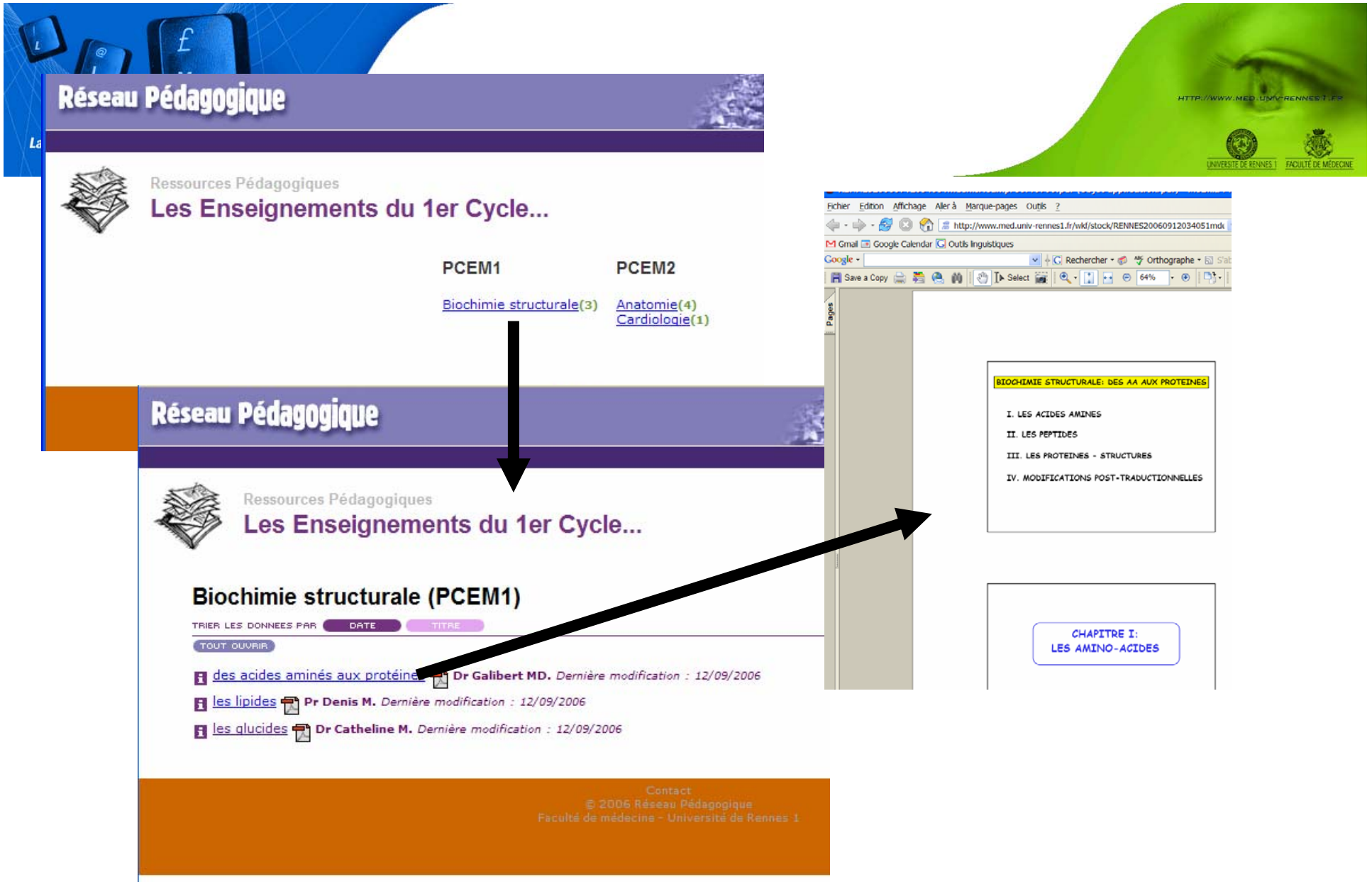


http://www.med.univ-rennes1.fr



http://www.med.univ-rennes1.fr

Laboratoire d'Informatique Médicale



http://www.med.univ-rennes1.fr

Laboratoire d'Informatique Médicale