

RFIDEC — cours 1 : Rappels de probas/stats (1/3)

Christophe Gonzales

LIP6 – Université Paris 6, France

- Pas de contrôle continu
- Note finale d'UE en trois parties :
 - ① Examen écrit en milieu de semestre (35%)
 - ② Examen écrit en fin de semestre (35%)
 - ③ Contrôle sur machine (30%)

Introduction à la reconnaissance des formes

Reconnaissance des formes

Algorithmes \implies reconnaissance automatique de régularités,
de motifs dans un amas de données.

Utilisation de ces motifs pour des tâches (classification, etc)

Exemple :

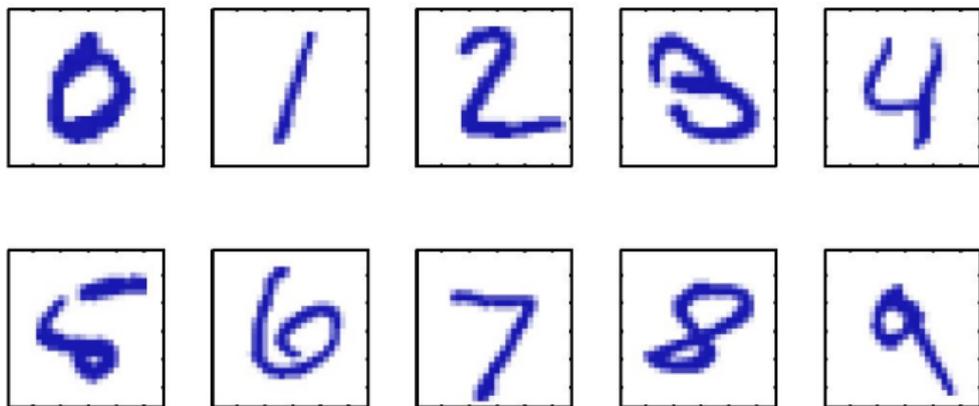
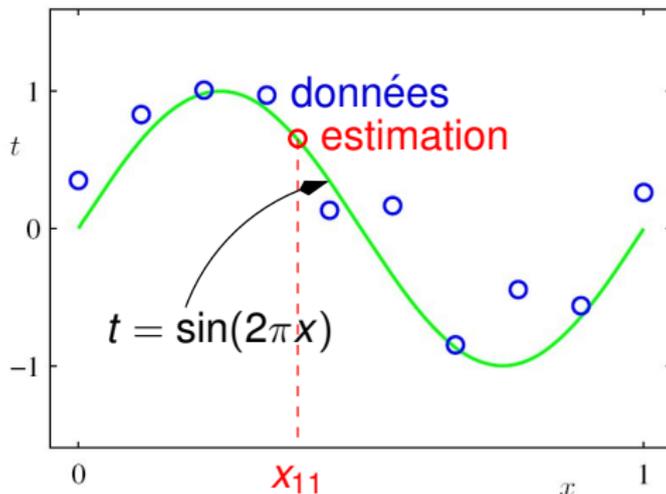


Image = données (pixels) *Classification* = reconnaître le chiffre

Exemple : problème d'ajustement (1/5)



Observations

(x_1, t_1)
⋮
 (x_{10}, t_{10})

\Rightarrow courbe $\sin(2\pi x)$ \Rightarrow estimation de t_{11}

\Rightarrow reconnaissance de la courbe verte

Exemple : problème d'ajustement (2/5)

Idée : estimer la courbe verte par un polynôme :

$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \cdots + w_Mx^M = \sum_{j=0}^M w_jx^j$$

⇒ **Problèmes** :

- déterminer les « meilleurs » coefficients w_j
 - déterminer la « meilleure » valeur de M
- « meilleur » ⇒ critère d'optimalité

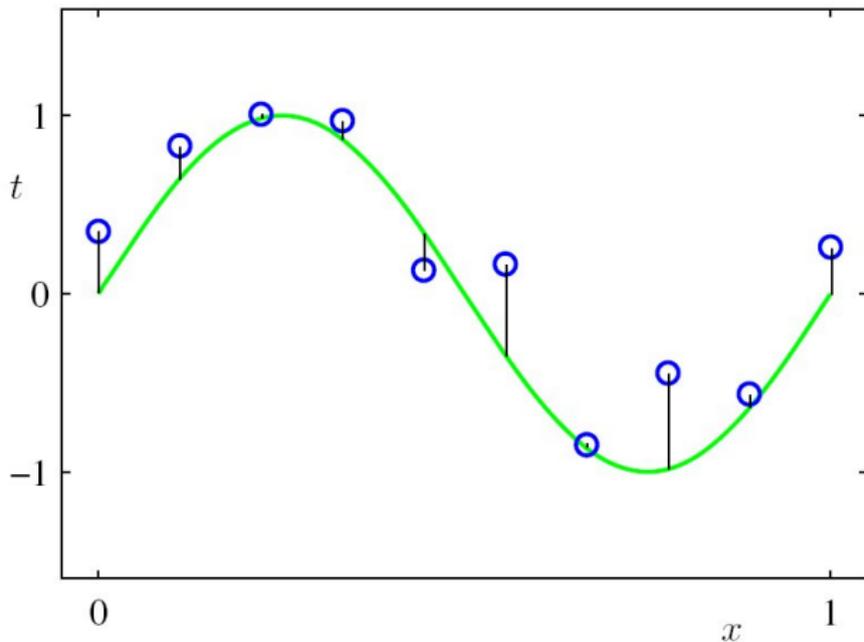
critère d'optimalité possible

Minimiser une **fonction d'erreur** mesurant l'inadéquation entre la courbe $y(x, \mathbf{w})$ et les points observés :

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{10} [y(x_i, \mathbf{w}) - t_i]^2$$

Exemple : problème d'ajustement (3/5)

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{10} [y(x_i, \mathbf{w}) - t_i]^2$$



$$\textit{Problème 1} : \mathbf{w}^* \text{ tel que } \mathbf{w}^* = \min_{\mathbf{w}} E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{10} [y(x_i, \mathbf{w}) - t_i]^2$$

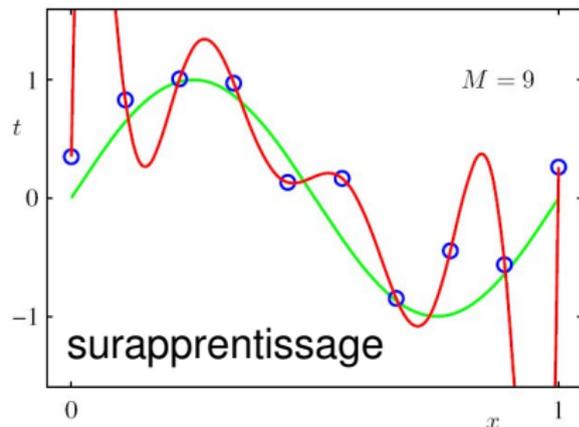
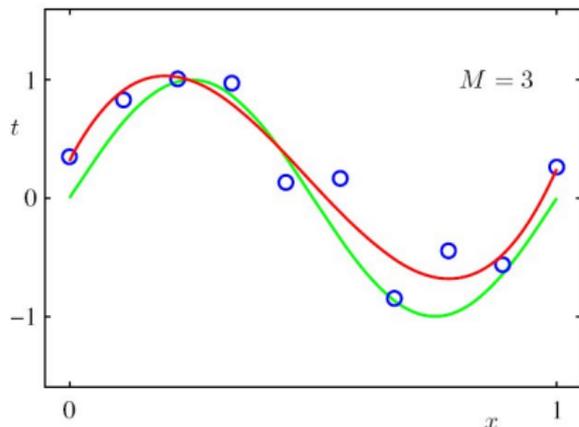
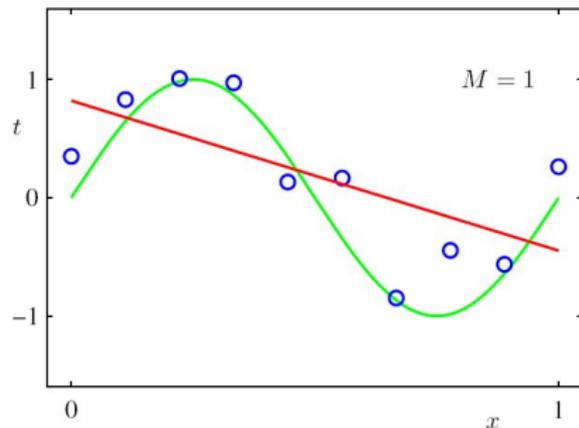
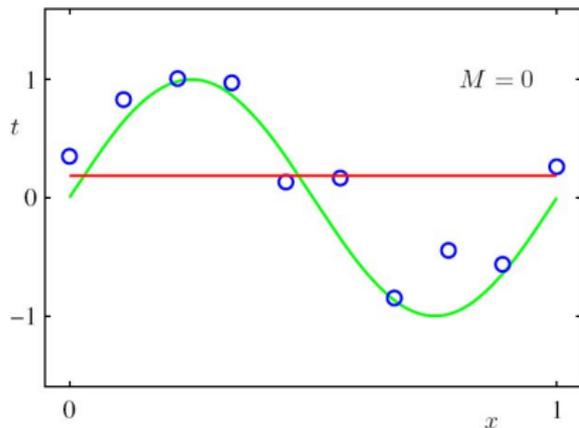
$E(\mathbf{w})$ fonction quadratique en $\mathbf{w} \implies$ 1 seule solution

$$\implies \frac{dE(\mathbf{w})}{d\mathbf{w}} = 0$$

\implies résoudre un système linéaire

Problème 2 : Choix de $M \implies$ sélection de modèle

Exemple : problème d'ajustement (5/5)



- 1 Introduction et rappels de proba/stat
- 2 Estimations ponctuelles à partir d'échantillons
- 3 Intervalles de confiance, tests d'hypothèses
- 4 Tests d'ajustement et maximum de vraisemblance
- 5 MAP et apprentissage non paramétrique
- 6 Régression linéaire
- 7 Clustering (K-means, Kohonen)
- 8 Classification bayésienne
- 9 Classification gaussienne
- 10 Classifieur linéaire : LDA, perceptron

- 1 Introduction à la statistique descriptive
- 2 vocabulaire de stat descriptive
- 3 distributions et représentations
- 4 indicateurs de « moyenne »
- 5 indicateurs de dispersion

Statistique

- recueil de données
- traitement de ces données
- l'exploitation (interprétation, prévision, etc) sort du cadre de la stat descriptive

Principe de la statistique descriptive

- des amas de données
- stat descriptive \implies synthèse, résumé d'informations

données synthétiques \implies exploitation



pas forcément besoin de probabilités

Exemple

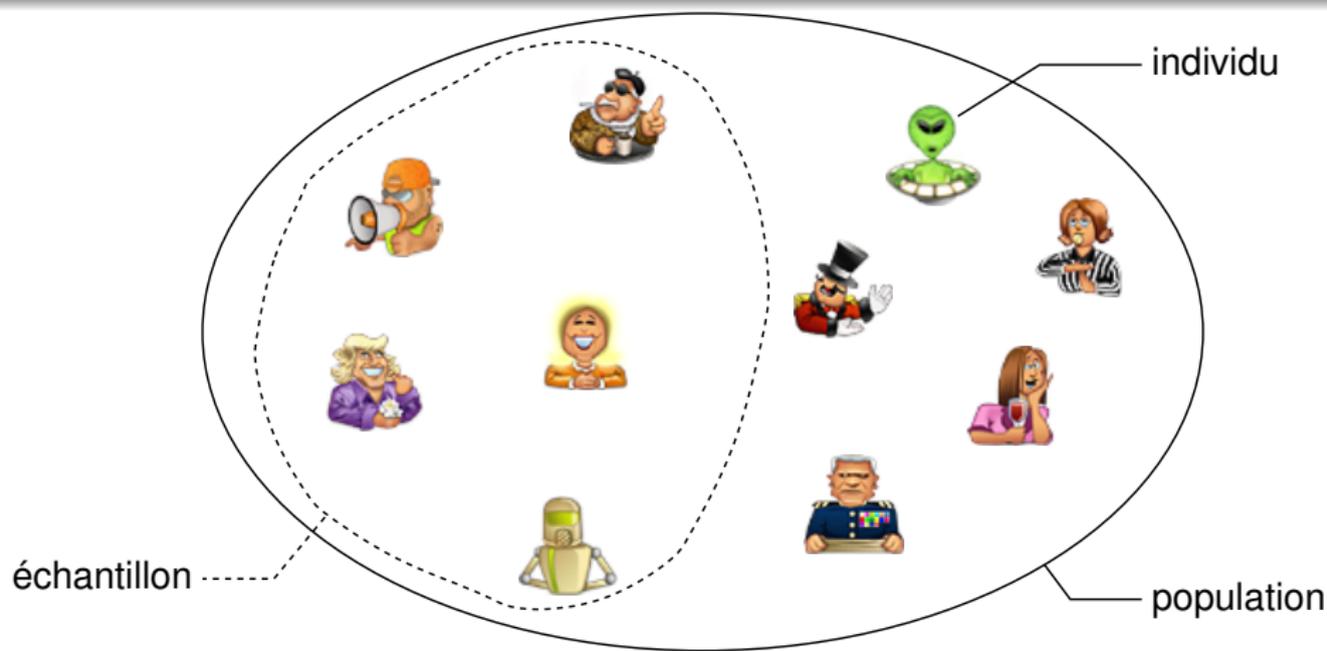
Salaires de cadres masculins (exprimés en k€)

18	10	12	12	10	14	14	14	12	12	16	18	10
16	16	12	12	14	14	14	14	12	16	14	14	16
12	10	18	12	18	18	14	14	14	12	16	14	16
10	16	16	14	14	14	12	16	16	14	14	12	14

Salaires de cadres féminins (exprimés en k€)

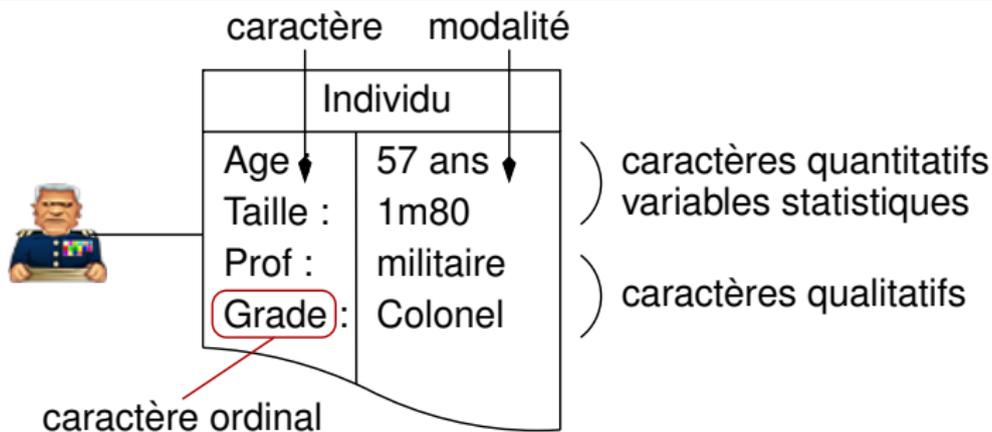
16	16	14	14	14	14	10	12	12	08	20	14	14
14	20	14	12	12	16	10	10	18	10	16	12	12
14	18	12	12	16	08	14	10	12	14	12	16	14

Question : Les hommes sont-ils mieux payés que les femmes ?



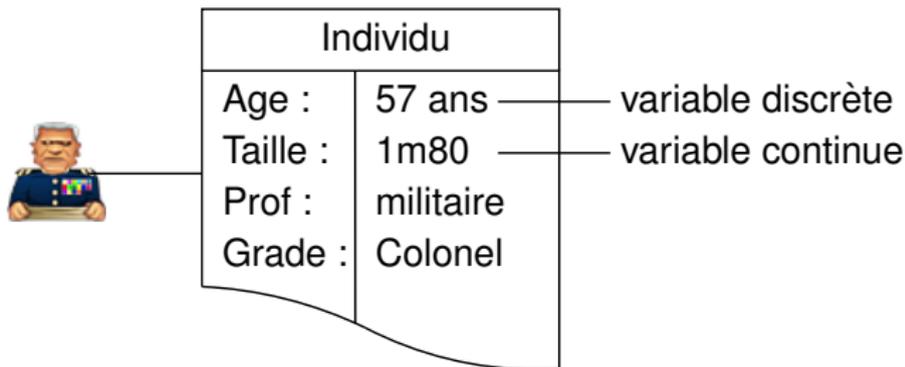
Définitions

- **population (statistique)** : ensemble des objets (ou personnes) sur lesquels porte l'étude
- **individu** : chaque élément de la population



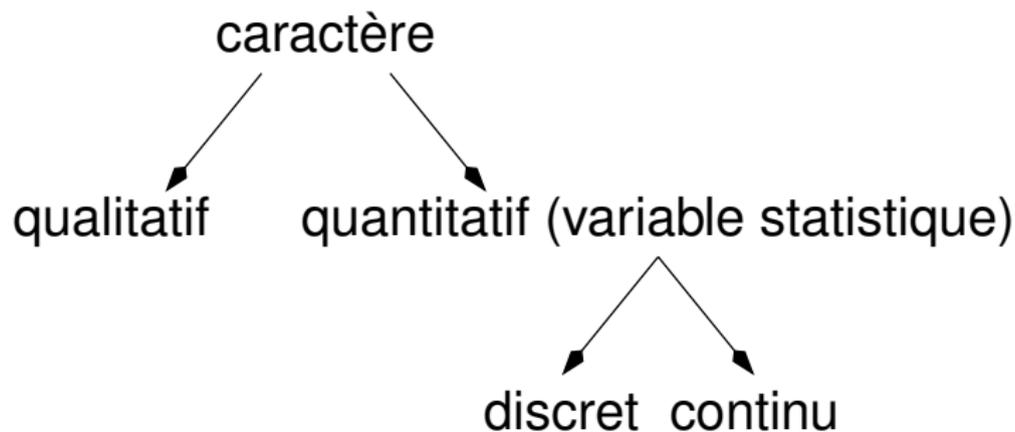
Définitions

- **Caractères** : critères d'étude de la population
- **Modalités** : les valeurs que peuvent prendre les caractères
- **Caractère quantitatif ou Variable statistique** : ensemble de modalités = des nombres + échelle mathématique
- **Caractère qualitatif ou Variable catégorielle** : caractère non quantitatif
- **Caractère ordinal** : les modalités sont ordonnées



Définitions sur les variables statistiques

- **Variable discrète** : définie sur un espace discret (par exemple des entiers)
- **Variable continue** : définie sur un continuum (toutes les valeurs numériques d'un intervalle)



Quelques définitions

- X : caractère défini sur une population de N individus
- $\{x_1, \dots, x_l\}$ modalités de X
- $N_j =$ **effectif** de x_j
= nombre d'individus pour lesquels X a pris la valeur x_j
- **fréquence ou effectif relatif** : $f_j = \frac{N_j}{N}$
- **distribution** de X : ensemble des couples $\{(x_1, f_1), (x_2, f_2), \dots\}$

Idée :

calculer la distribution des caractères sur l'amas de données
 \implies résume les informations importantes

Définition du diagramme en bâtons

- caractère X de distribution $\{(x_1, f_1), (x_2, f_2), (x_3, f_3), \dots\}$
- diagramme en bâtons = graphe dans lequel on associe à chacune des modalités x_i (représentées sur l'axe horizontal) un bâton de hauteur f_i
- représentation en colonnes : idem mais en élargissant les bâtons

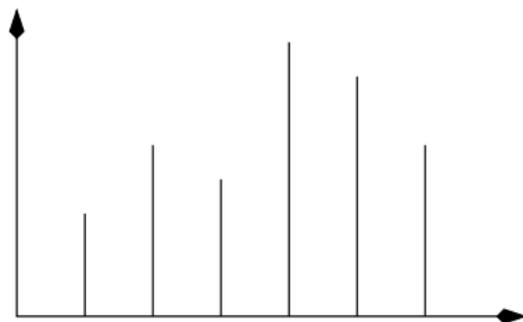
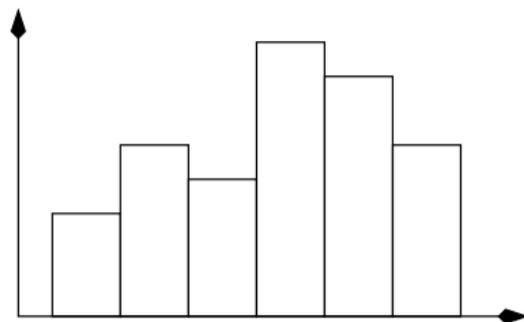


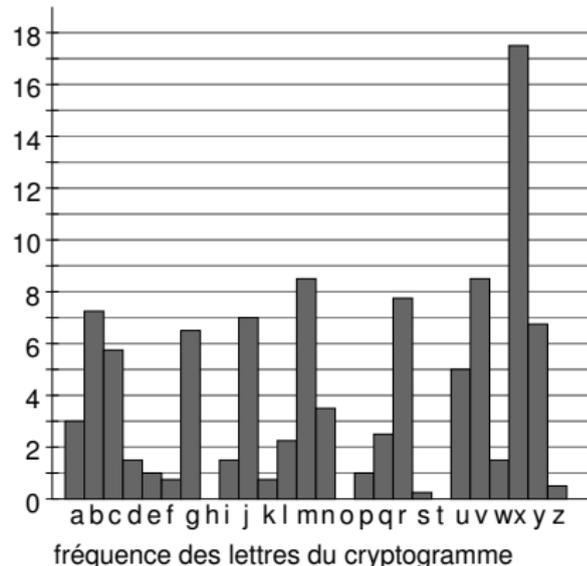
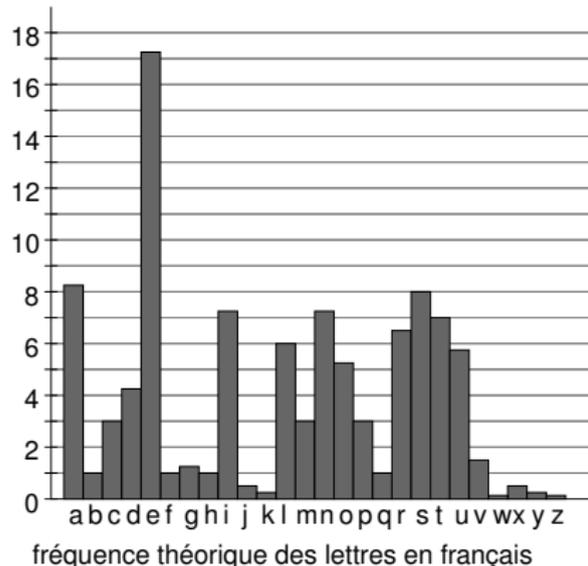
diagramme en bâtons



représentation en colonnes

Application du diagramme en bâtons

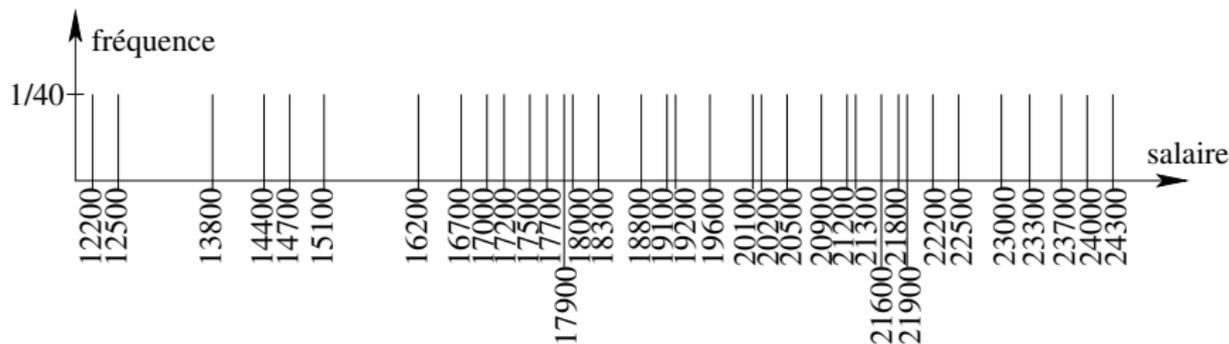
XY AXJ BYRJMYJ, MQQMUVVXYJ GXR NCBWJR N' UYX
LMBY N' PCLLX XJ BGR XAVBDBVXYJ, XY IMAX NU
AMYNXGMFVX, RUV GX QGMJVX NU LUV NU QMGMBR
VCEMG.



Problèmes des diagrammes en bâtons

Salaires des cadres masculins (en k€)

19,6	12,5	17,7	18,8	19,1	14,7	21,9	22,5	21,8
20,1	16,2	20,5	12,2	25,4	20,9	21,2	15,5	21,3
17,9	25,0	14,4	21,7	18,3	16,7	23,0	17,0	24,3
27,0	27,7	18,0	17,5	23,7	21,6	13,8	22,2	19,2
20,2	15,1	19,6	17,2	23,3	24,0			



variables continues \implies diagramme en bâtons inutilisable

Idée force : utiliser les diagrammes en bâton ssi la taille de la population \gg au nombre de modalités des caractères

sinon \implies discrétiser le caractère

Discrétisation

regrouper toutes les modalités appartenant à certains intervalles dans des **classes** de données

Caractéristiques d'une classe

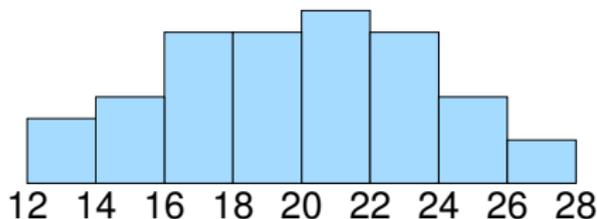
- \mathcal{C} : la classe $[a, b[$
- **centre** de la classe \mathcal{C} : $c = \frac{a+b}{2}$
- **amplitude** de la classe \mathcal{C} : $h = b - a$

Exemple de discrétisation (1/3)

Salaires des cadres masculins (en k €)								
19,6	12,5	17,7	18,8	19,1	14,7	21,9	22,5	21,8
20,1	16,2	20,5	12,2	25,4	20,9	21,2	15,5	21,3
17,9	25,0	14,4	21,7	18,3	16,7	23,0	17,0	24,3
27,0	27,7	18,0	17,5	23,7	21,6	13,8	22,2	19,2
20,2	15,1	19,6	17,2	23,3	24,0			

Classe de salaire	effectif	fréq.	Classe de salaire	effectif	fréq.
[12000, 14000[3	3/42	[20000, 22000[8	8/42
[14000, 16000[4	4/42	[22000, 24000[7	7/42
[16000, 18000[7	7/42	[24000, 26000[4	4/42
[18000, 20000[7	7/42	[26000, 28000[2	2/42

Diagramme
en colonnes :

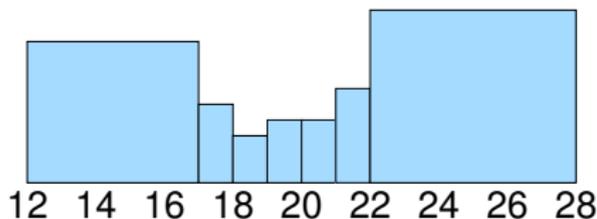


Exemple de discrétisation (2/3)

Salaires des cadres masculins (en k €)								
19,6	12,5	17,7	18,8	19,1	14,7	21,9	22,5	21,8
20,1	16,2	20,5	12,2	25,4	20,9	21,2	15,5	21,3
17,9	25,0	14,4	21,7	18,3	16,7	23,0	17,0	24,3
27,0	27,7	18,0	17,5	23,7	21,6	13,8	22,2	19,2
20,2	15,1	19,6	17,2	23,3	24,0			

Classe de salaire	effectif	fréq.	Classe de salaire	effectif	fréq.
[12000, 17000[9	9/42	[20000, 21000[4	4/42
[17000, 18000[5	5/42	[21000, 22000[6	6/42
[18000, 19000[3	3/42	[22000, 28000[11	11/42
[19000, 20000[4	4/42			

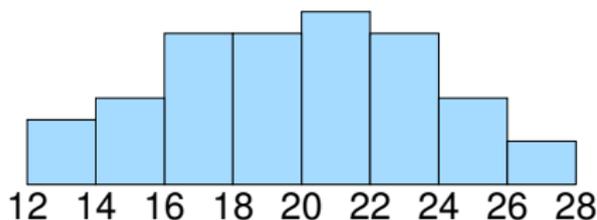
Diagramme
en colonnes :



Exemple de discrétisation (3/3)

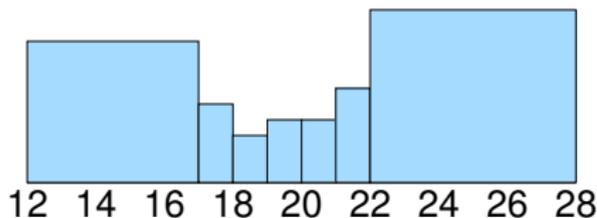
Première discrétisation :

Diagramme
en colonnes :



Deuxième discrétisation :

Diagramme
en colonnes :



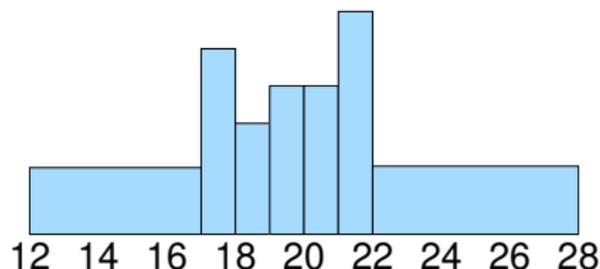
N'utiliser le diagramme en colonnes que pour des discrétisations uniformes

Définition d'un histogramme

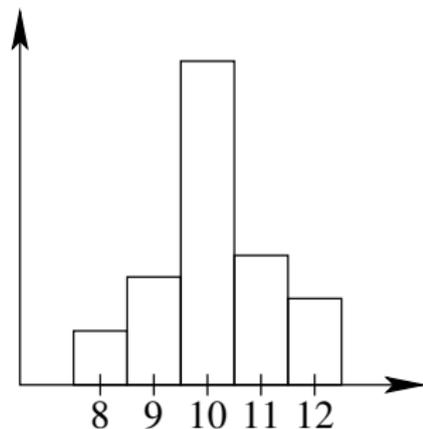
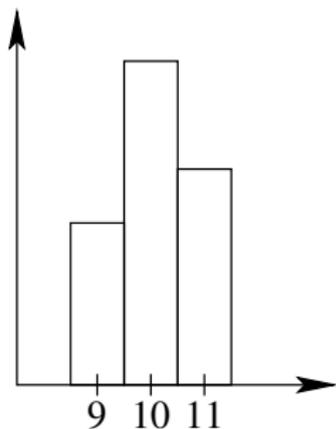
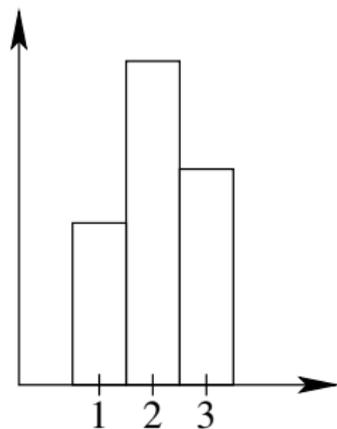
graphe dont l'axe des abscisses représente les modalités et qui associe à chaque classe de modalités un rectangle dont la base correspond aux bornes de cette classe et dont la hauteur est égale à :

$$\text{hauteur} = \frac{\text{fréquence}}{\text{base}}.$$

⇒ la fréquence d'une classe est égale à la surface de son rectangle



Synthèse d'histogrammes ?



Caractéristiques statistiques :

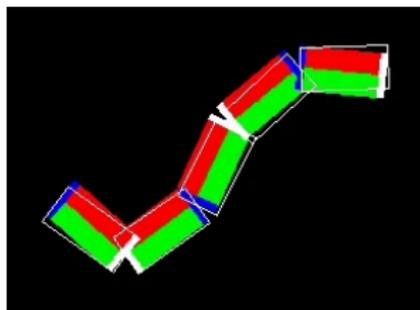
- localisation sur l'axe des $X \implies$ moyenne, médiane, mode
- étendue \implies écart-type, variance

Moyenne d'une variable statistique discrète

- X : variable statistique discrète
- modalités : $\{x_1, x_2, \dots, x_l\}$
- population de N individus
- **Moyenne de X** : $\mu_X = \sum_{i=1}^l f_i x_i = \frac{1}{N} \sum_{i=1}^l N_i x_i$

Moyenne d'une variable statistique continue

- Y : variable statistique continue
- modalités regroupées en l classes de centres $\{y_1, y_2, \dots, y_l\}$
- population de N individus
- **Moyenne de Y** : $\mu_Y = \sum_{i=1}^l f_i y_i = \frac{1}{N} \sum_{i=1}^l N_i y_i$



Propriétés de la moyenne

Propriété 1

- X : variable statistique
- $Y = aX + b$, où a et b sont des nombres réels
- $\mu_Y = a\mu_X + b$

Propriété 2

- X : variable statistique
- n sous-populations distinctes S_1, \dots, S_n de N_1, \dots, N_n individus
- μ_i la moyenne de X sur la $i^{\text{ème}}$ sous-population
- μ_X : moyenne de X sur la population $S = S_1 \cup S_2 \cup \dots \cup S_n$

$$\mu_X = \frac{\sum_{i=1}^n N_i \mu_i}{\sum_{i=1}^n N_i}.$$

Médiane d'une variable statistique discrète

- *moyenne* = valeur centrale des modalités
- *médiane* = valeur de la variable pour laquelle la moitié de la population a une modalité inférieure à cette valeur et l'autre moitié en a une supérieure

Médiane d'une variable statistique discrète

- X : variable statistique discrète, modalités : $\{x_1, \dots, x_I\}$
- population de N individus
- *médiane* de X = tout nombre δ tel que :

$$\sum_{i \in \{j: x_j < \delta\}} N_i \leq N/2 \quad \text{et} \quad \sum_{i \in \{j: x_j > \delta\}} N_i \leq N/2$$

variable statistique X						
modalité	0	1	2	3	4	5
effectif	2	3	4	5	3	2

Données triées																		
0	0	1	1	1	2	2	2	2	3	3	3	3	3	4	4	4	5	5

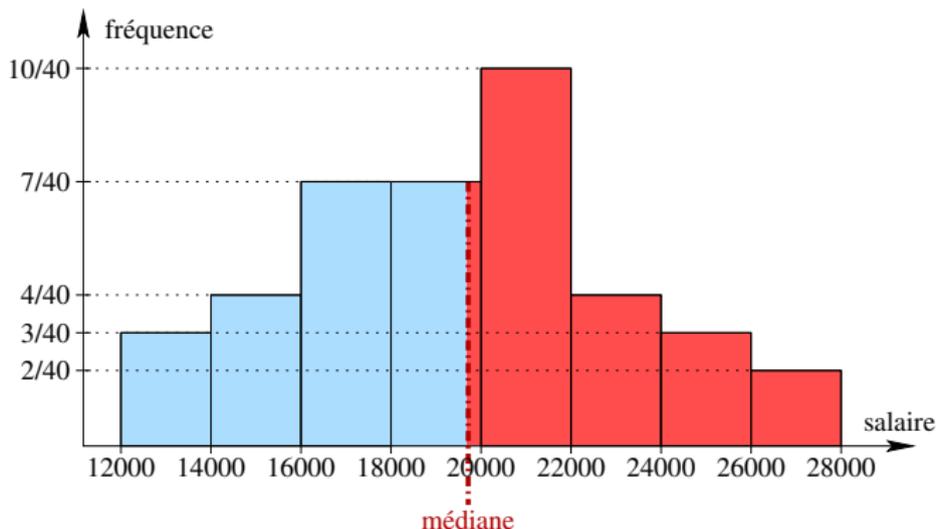
variable statistique X						
modalité	0	1	2	3	4	5
effectif	2	3	4	4	3	2

Données triées																	
0	0	1	1	1	2	2	2	2	3	3	3	3	4	4	4	5	5

Médiane d'une variable statistique continue

Médiane d'une variable statistique continue

- X : variable statistique continue
- **Médiane** = le nombre δ tel que les aires situées de part et d'autre de ce nombre dans l'histogramme représentant X sont égales



Quantile d'une variable discrète

- X : variable statistique discrète, modalités $\{x_1, \dots, x_l\}$
- population de N individus
- **quantile d'ordre α** = tout nombre δ tel que :

$$\sum_{i \in \{j: x_j < \delta\}} N_i \leq \alpha N \quad \text{et} \quad \sum_{i \in \{j: x_j > \delta\}} N_i \leq (1 - \alpha)N$$

Quantile d'une variable continue

- X : variable statistique continue
- **quantile d'ordre α** = le nombre δ tel que les aires situées de part et d'autre de ce nombre dans l'histogramme représentant X sont égales respectivement à $\alpha \times$ aire totale et $(1 - \alpha) \times$ aire totale

Principaux quantiles	
valeur de α	nom du quantile d'ordre α
$1/2$	médiane
$i/4$ ($i = 1, 2, 3$)	$i^{\text{ème}}$ quartile
$i/5$ ($i = 1, 2, 3, 4$)	$i^{\text{ème}}$ quintile
$i/10$ ($i = 1, 2, \dots, 9$)	$i^{\text{ème}}$ décile
$i/100$ ($i = 1, \dots, 99$)	$i^{\text{ème}}$ centile

Propriété

- X : variable statistique
- $Y = aX + b$, où a et b sont des nombres réels
- $\text{médiane}(Y) = a \times \text{médiane}(X) + b$

Application de la médiane

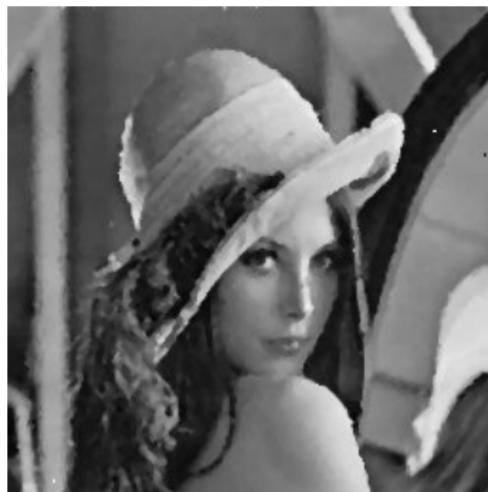
Filtre médian :

x_{11}	x_{12}	x_{13}
x_{21}	x_{22}	x_{23}
x_{31}	x_{32}	x_{33}



	M	

$M = \text{médiane des } x_{ij}$



Étendue d'une variable discrète

- X : variable statistique discrète
- **Étendue** de X = la différence entre la plus grande modalité de X et la plus petite modalité

Étendue d'une variable continue

- X : variable statistique continue
- **Étendue** de X = la différence entre la borne supérieure de la classe associée à la plus grande valeur observée et la borne inférieure de la classe associée à la plus petite valeur observée



Mesure de dispersion peu utilisée

Idée force : analyser globalement les déviations entre les valeurs prises sur un individu et la moyenne de la population

Variance d'une variable discrète

- X : variable statistique discrète, modalités : x_1, \dots, x_l
- population : N individus

- **Variance** de X :
$$\sigma_X^2 = \sum_{i=1}^l f_i (x_i - \mu_X)^2 = \frac{1}{N} \sum_{i=1}^l N_i (x_i - \mu_X)^2$$

Variance d'une variable continue

- Y : variable statistique continue
- modalités : l classes de centres $\{y_1, y_2, \dots, y_l\}$
- population : N individus

- **Variance** de Y :
$$\sigma_Y^2 = \sum_{i=1}^l f_i (y_i - \mu_Y)^2 = \frac{1}{N} \sum_{i=1}^l N_i (y_i - \mu_Y)^2$$

Calcul pratique de la variance

$$\begin{aligned}\sigma_X^2 &= \frac{1}{N} \sum_{i=1}^l N_i (x_i - \mu_X)^2 \\ &= \frac{1}{N} \left(\sum_{i=1}^l N_i x_i^2 \right) - \mu_X^2\end{aligned}$$

Propriété

- X : variable statistique
- $Y = aX + b$, où a et b sont des nombres réels
- $\sigma_Y^2 = a^2 \sigma_X^2$

Problème de la variance : unités différentes des données

X exprimé en € $\implies \sigma_X^2$ exprimé en €^2

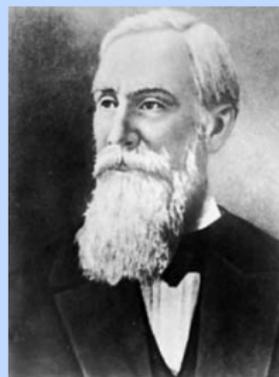
Écart-type

- X : variable statistique (discrète ou continue)
- **Écart-type** de X : σ_X , la racine carrée de la variance de X

Propriété

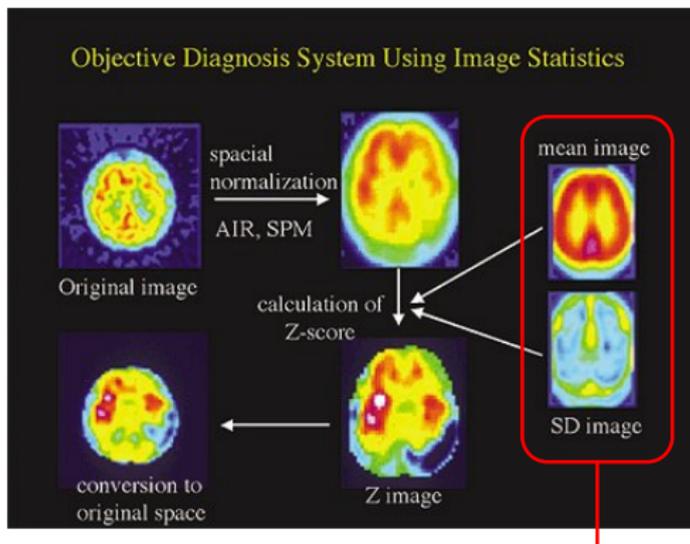
- X : variable statistique
- $Y = aX + b$, où a et b sont des nombres réels
- $\sigma_Y = |a|\sigma_X$

Théorème de Bienaymé-Tchébicheff



- X une variable statistique discrète
 - moyenne μ_X et écart-type σ_X
 - la proportion de valeurs de X observées entre $\mu_X - k\sigma_X$ et $\mu_X + k\sigma_X$ est strictement supérieure à $1 - 1/k^2$
- Propriété valable pour les variables statistiques continues :
la proportion d'individus pour lesquels la valeur de X se situe entre $\mu_X - k\sigma_X$ et $\mu_X + k\sigma_X$ est strictement supérieure à $1 - 1/k^2$
- \Rightarrow $\left\{ \begin{array}{l} \text{plus de 88\% de la population entre } \mu_X - 3\sigma_X \text{ et } \mu_X + 3\sigma_X \\ \text{plus de 96\% de la population entre } \mu_X - 5\sigma_X \text{ et } \mu_X + 5\sigma_X \end{array} \right.$

Application de l'écart-type



base de 1200 patients

- 1 normaliser l'image du cerveau du patient (SPM : Statistical Parametric Mapping)
- 2 comparer avec la base de patients
- 3 produire l'image comparée
- 4 diagnostiquer la maladie