EL 3019

I. RAPPELS : DONNEES ABERRANTES

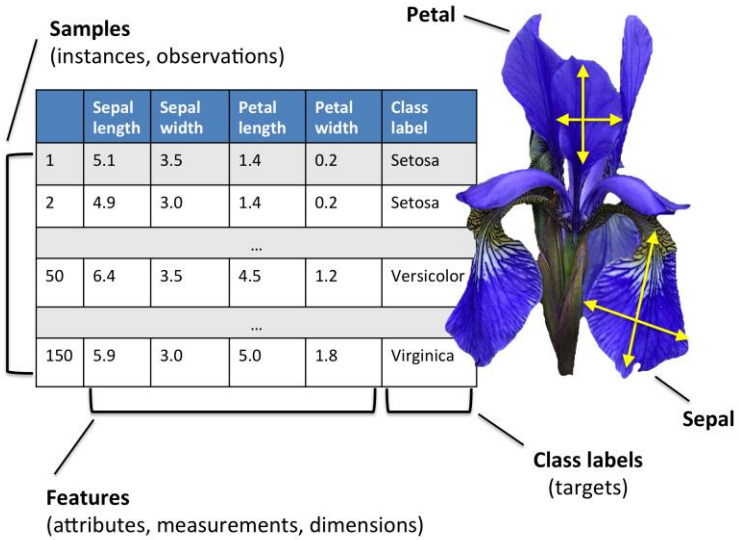
II. IMPUTATION DES DONNEES MAQUANTES

III. RECODAGE DES VARIABLES

I. Données aberrantes- Outliers

1 Rappel des dénominations sur un exemple *… qui nous servira pour les TP*

- Iris de Fisher : 150 enregistrements (individus) / 5 attributs (variables)
  - 4 attributs continus (quantitatifs)
  - 1 attribut discret (qualitatif) à 3 niveaux



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

2 Définition

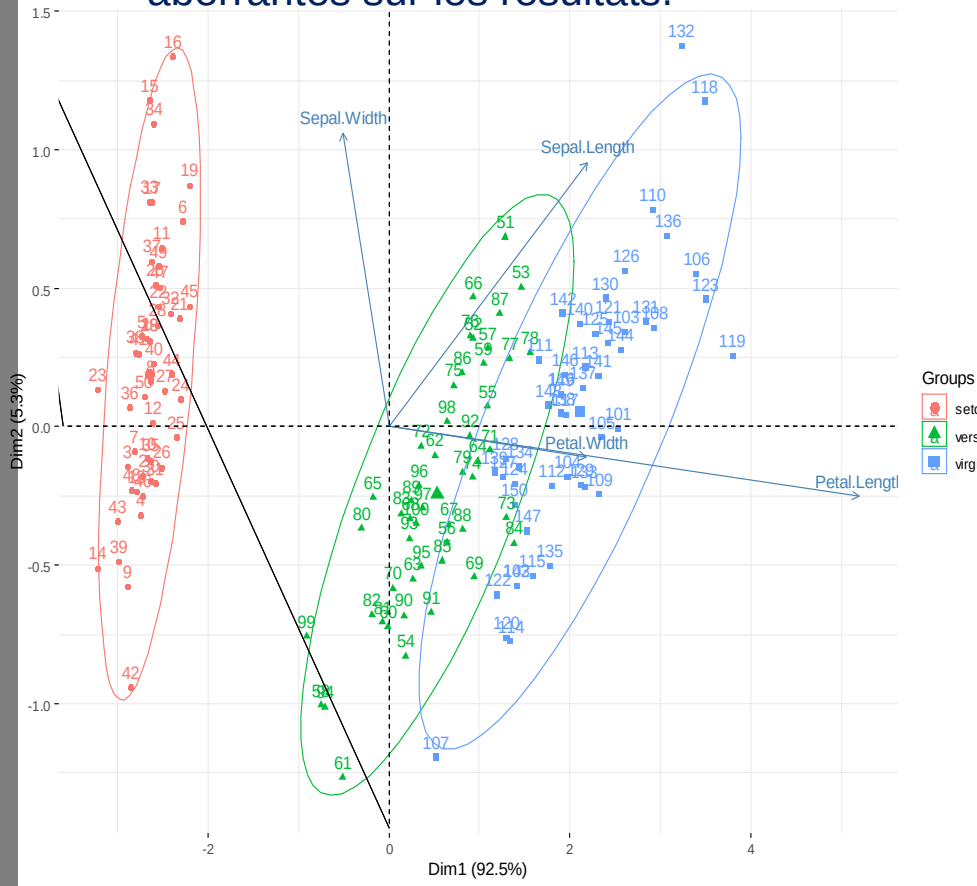
- Une donnée aberrante est une observation qui est beaucoup plus « distante » que les autres observations par rapport à la tendance centrale (la moyenne et particulièrement la médiane). Elle peut être due à la variabilité interne du phénomène observé mais le plus souvent due à une erreur expérimentale.

- Elle peut influencer très fortement les résultats d'une analyse et conduire à des interprétations erronées

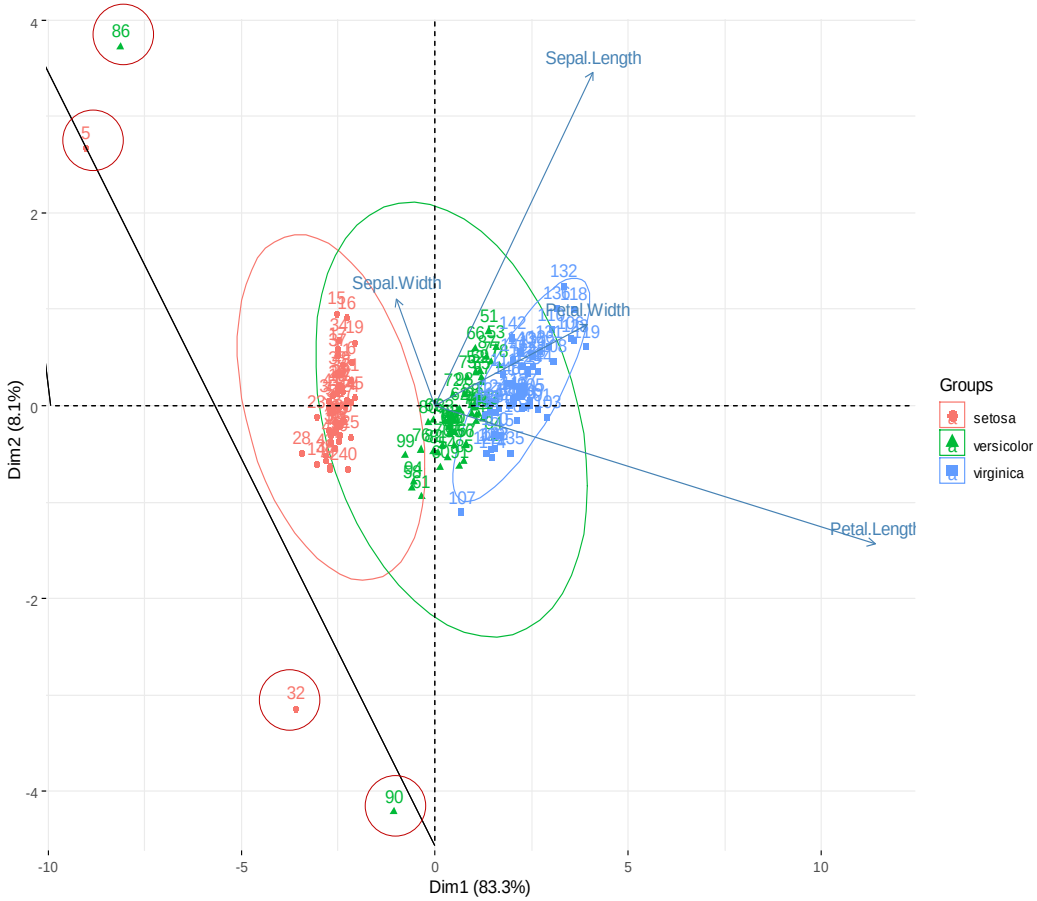
I. Données aberrantes- Outliers

3 Exemple

On désire analyser les données par une technique permettant de voir dans un même plan toutes les variables et tous les individus. Cette technique est très utilisée en fouille de données et sera enseignée l'année prochaine. Ici il s'agit de voir simplement quelle est l'influence de données aberrantes sur les résultats.



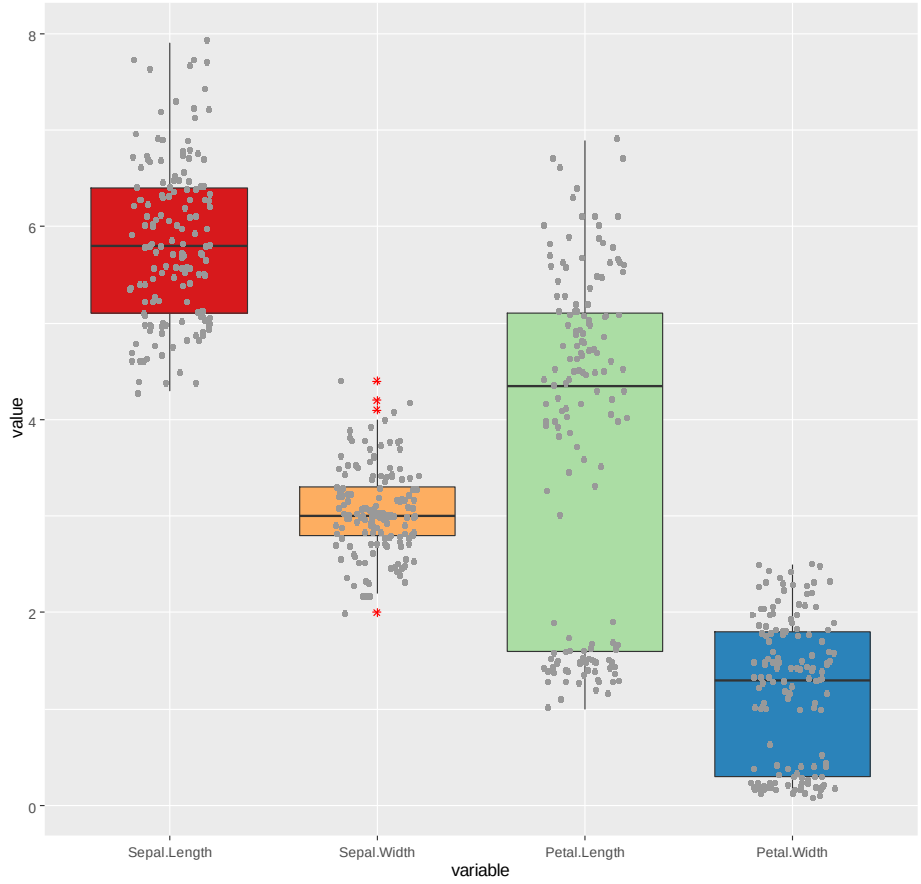
Absence de données aberrantes



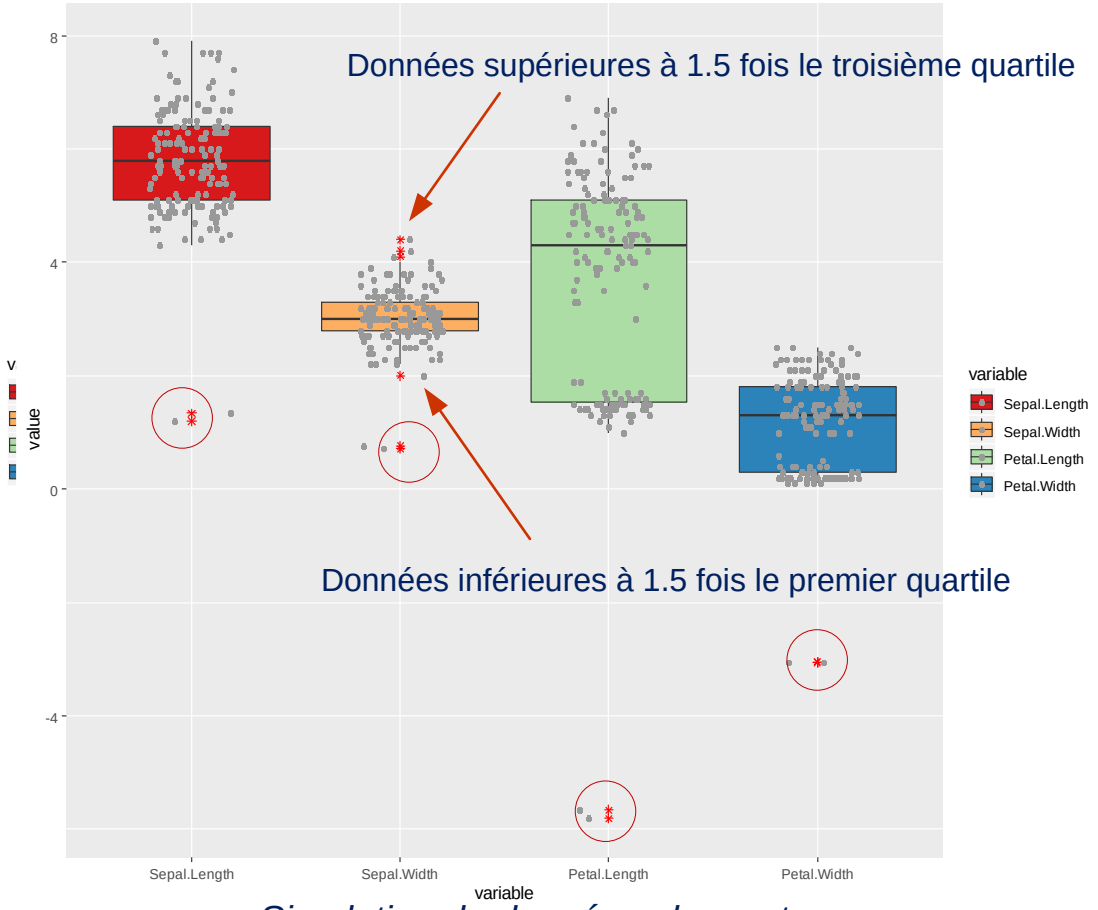
Simulation de 4 données aberrantes (sur 600 données !!)

4 Diagnostique

On calcule les quantiles (cf.cours précédents). Ces derniers peuvent être visualisés à l'aide d'une boîte à moustaches (Box plot). Ces graphiques doivent être effectuées pour chaque variable. Il s'agit d'un pré requis INDISPENSABLE en analyse de données !!!!



Absence de données aberrantes



Simulation de données aberrantes

5

Que faire lorsque l'on rencontre des données abberantes ?

- Eliminer les données

  Eliminer une donnée aberrante équivaut le plus souvent à éliminer l'enregistrement (l'individu). S'il y a de nombreux outliers dans un jeux de données, quelle sera la pertinence des données après élimination nottament en terme de représentativité et de distribution ?. Ce n'est pas la meilleure méthode

- Estimer la donnée par des méthodes analogues à celles utilisées pour les données manquantes (cf chapitre suivant)

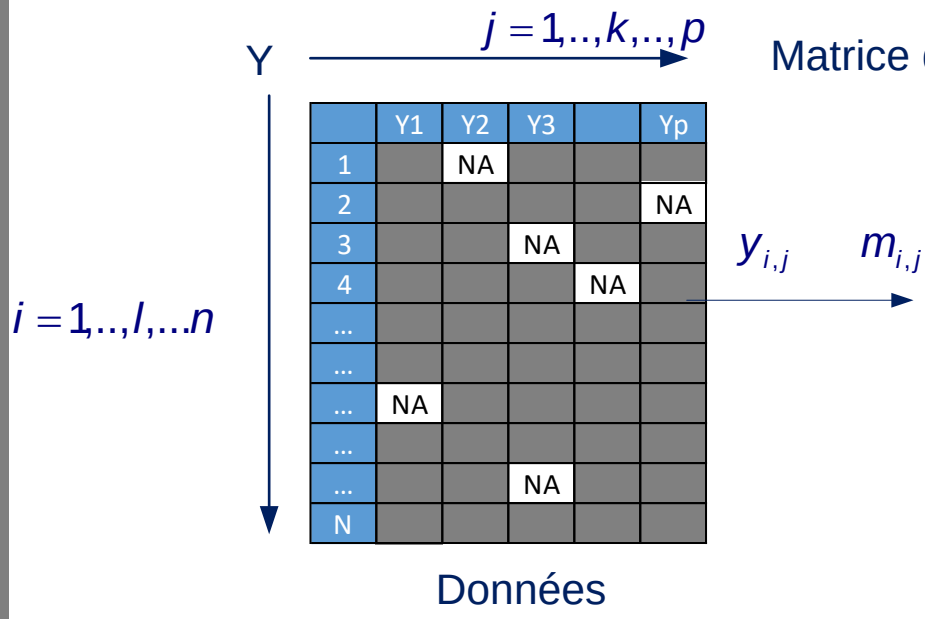
II. Imputation des données manquantes

- L'ensemble des données avec lequel on doit travailler n'est pas toujours complet. Ignorer les données manquantes peut entrainer, outre une perte d'information et de précision, un biais élevé susceptible d'invalider les modèles d'analyses
- Les données manquantes ne doivent concerner que quelques variables et que quelques observations. Si une variable manque pour toutes les observations, il s'agit d'une variable non observée (variable latente)
- La majorité des méthodes statistiques courantes ainsi que de nombreux algorithmes utilisés en machine learning supposent que toutes les observations associées à toutes les variables soient spécifiées

1 → Origine des données manquantes

- Les valeurs n'ont pas été mesurées
- Les valeurs sont mesurées mais inutilisables (**erreurs, valeurs aberrantes**)
- Les valeurs sont indisponibles ex de réponse : 'ne sait pas'
- Les valeurs sont censurées : limite de détection - qualification / quantification

II. Imputation des données manquantes



$y_{i,j}$ $m_{i,j}$

Matrice des données manquantes (M)

	X1	X2	X3		Xp
1	0	1	0	0	0
2	0	0	0	0	1
3	0	0	1	0	0
4	0	0	0	1	0
...	0	0	0	0	0
...	0	0	0	0	0
...	1	0	0	0	0
...	0	0	0	0	0
...	0	0	1	0	0
N	0	0	0	0	0

Matrice indicatrice

$$Y = (Y_{obs}, Y_{mis})$$

$$Y_{obs} = Y_{M=0}$$
$$Y_{mis} = Y_{M=1}$$

On définit Y en fonction de la matrice indicatrice

2 → Typologie des données manquantes (MD)

- Les données manquantes ne sont pas toujours issues du « pur » hasard. Little & Rubin propose une typologie (largement utilisée) qui distingue 3 catégories :

  - → Missing completely at random (MCAR)
  - → Missing at random (MAR)
  - → Missing not at random (MNAR)

II. Imputation des données manquantes

2.1 Missing Completely At Random (MCAR)

- La probabilité d'absence d'une observation est la même pour TOUTES les observations. Il s'agit donc d'une constante
- Autrement dit, la fait de ne pas avoir la valeur pour une variable Y_j est indépendante des autres variables $Y_{j \neq k}$
- L'absence des données ne dépend pas des valeurs de Y → $P(Y/M) = P(M)$ pour tout Y

→ Exemple : Y1 = age, Y2 = Sexe, Y3 glycémie
La probabilité que l'âge soit NA ne dépend ni du sexe ni des valeurs de la glycénie. Elle est la même pour tous les sujets

Remarque : si la quantité de MCAR n'est pas trop importante, ignorer les cas avec données manquantes ne biaisera pas l'analyse mais il y aura une **perte** de précision

2.2 Missing At Random (MAR)

- La probabilité d'absence d'une observation est liée à une ou plusieurs autres variables observées
- Autrement dit, le fait de ne pas avoir la valeur pour une variable Y_j est dépendante des autres variables $Y_{j \neq k}$ ***observées***
- Dans un processus MAR, l'absence de données dépend uniquement des Yobs : $P(M/Y) = p(M/Y_{obs})$

II. Imputation des données manquantes

- Exemple 1 : Y1 = age, Y2 = Sexe, Y3 = glycémie
  La probabilité que l'âge soit NA dépend su sexe et ou la glycémie (**valeurs observées**)

- Exemple 2 : On recueille la pression artérielle (PA) de plusieurs patients. La PA est suivie surtout chez les personnes âgées. Il y aura plus de données manquantes chez les personnes jeunes

2.3 Missing Not At Random **(MNAR)**

- La probabilité d'absence d'une observation dépend des valeurs non observées et n'est pas aléatoire
- La probabilité d'absence d'une observation peut dépendre de la variable en question ou dépendre d'une autre variable dont les valeurs sont non observées
- L'estimation de ces valeurs nécessitent de poser des hypothèses fortes et d'obtenir des informations complémentaires
- Dans un processus MNAR, l'absence de données dépend des Yobs **et** des Ymis

- Exemple 1 : Y1 = age, Y2 = Sexe, Y3 = glycémie
  La probabilité que l'âge soit NA dépend des valeurs manquantes pour le sexe et ou la glycémie (**valeurs non observées**)

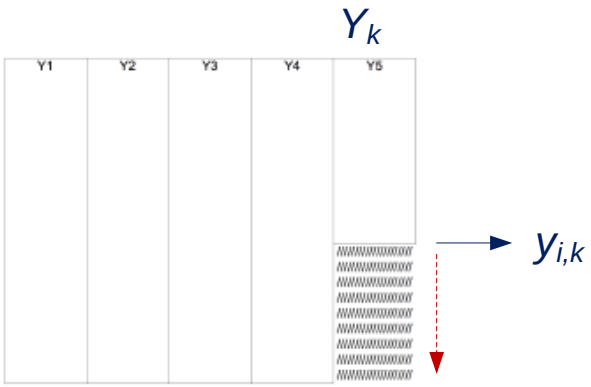
- Exemple 2 : On receuille les revenus de plusieurs personnes. Pour les personnes ayant des revenus très élevés, on a un risque accru d'avoir des données manquantes car elles ne veulent pas les fournir

II. Imputation des données manquantes

3 Répartition des données manquantes

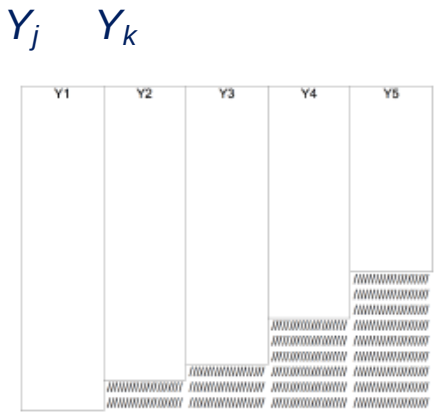
Valeurs manquantes univariées

Ne concernent qu'une seule variable Y_k pour une observation $y_{i,k}$ à partir de laquelle il n'y aura plus d'observations

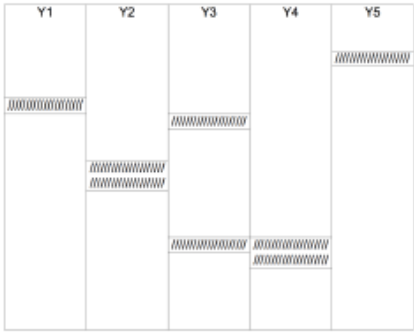


Valeurs manquantes monotones

si Y_j manquante pour un individu i implique que toutes les variables suivantes Y_k avec $k > j$ sont manquantes pour cet individu.



Valeurs manquantes arbitraires et non monotones



II. Imputation des données manquantes

3 Les méthodes d'estimation des données manquantes

3.1 Liste des principales méthodes … du moins les plus utilisées

- Délétion
  - Par liste
  - Partielle
- Imputation simples
  - Par une valeur unique représentative de la variable concernée
    - LOOCF, CMCF (complétion stationnaire)
    - Moyenne, médiane
  - Complétion par regroupement et centrage
    - Hot-Deck
    - Cold-Deck

    Algorithmes de classification automatique
  - Complétion par combinaisons linéaires
    - Régressions locales, simples, multiples, logistiques,….
    - NIPALS
    - SVD
    - ….

    … non abordées …..
- Imputation multiples
  - Equations chainées
  - Par EM algorithme

II. Imputation des données manquantes

3.2 Méthodes d'imputations par délétion

3.2.1 Délétions par liste (list wise)

- On ne considère que les individus (enregistrements) pour lesquels toutes les données sont disponibles
- La proportion d'observations complètes peut être faible même si, pour chaque variable (attribut), la probabilité qu'une donnée soit observée est grande
- Résultats généralement non biaisés pour les données MCAR mais fortement biaisés pour MAR et MNAR

Diminution de la précision et de la puissance

List wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

3.2.2 Délétions partielles : analyse des cas disponibles (avaible-case analysis)

- On effectue donc les analyses avec différentes tailles d'échantillons
- Certaines analyses ne seront pas compatibles, et certaines méthodes de DM / ML ne tolèrent pas de valeurs manquantes
- Résultats généralement non biaisés pour les données MCAR mais fortement biaisés pour MAR et MNAR

Pair wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

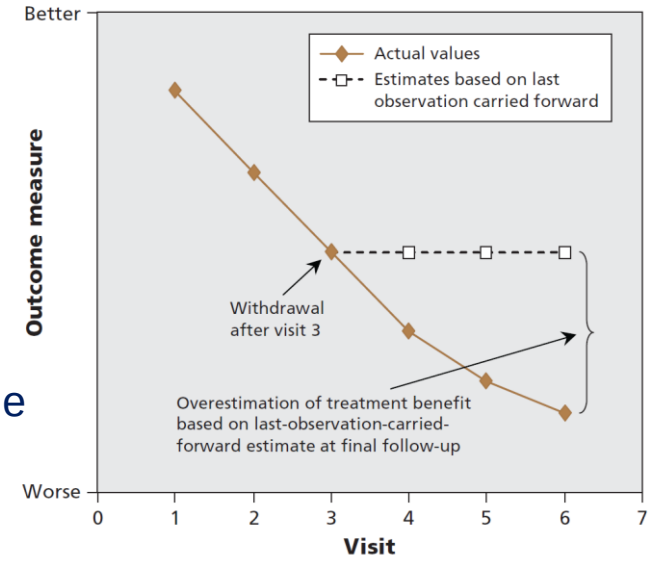
II. Imputation des données manquantes

3.3 Méthodes d'imputations simples

3.3.1 Par une valeur unique

LOCF : Last Observation Carried Forward

- On remplace la (les) donnée(s) manquante(s) par le (les) dernières valeur(s) observée(s)
- Utilisée pour des mesures répétées (essais cliniques, études de survie,…)



- Suppose que la « vraie » valeur reste inchangée depuis la dernière mesure (dépend du type d'étude, de l'évolution du critère)

CMCF : Concept Most Common Attribute Value Fitting

- On remplace la (les) donnée(s) manquante(s) par la valeur observée la plus fréquente

Remplacement par la moyenne ou la médiane

cf. transparent suivant ...

II. Imputation des données manquantes

Remplacement par la moyenne ou la médiane (1)

Person	Age	Sex	Years of Education
1	47	M	16
2	45	F	?
3	19	M	11
4	31	F	?
5	24	M	12
6	41	F	?
7	36	M	20
8	50	M	12
9	53	F	13
10	17	M	10
11	53	F	12
12	21	F	12
13	18	F	11
14	34	M	16
15	44	M	14
16	45	M	11
17	54	F	14
18	55	F	10
19	29	F	12
20	32	F	10

Remplacement par la moyenne

Person	Age	Sex	Years of Education	
1	47	M	16	
2	45	F	?	← 12.70
3	19	M	11	
4	31	F	?	← 12.70
5	24	M	12	
6	41	F	?	← 12.70
7	36	M	20	
8	50	M	12	
9	53	F	13	
10	17	M	10	
11	53	F	12	
12	21	F	12	
13	18	F	11	
14	34	M	16	
15	44	M	14	
16	45	M	11	
17	54	F	14	
18	55	F	10	
19	29	F	12	
20	32	F	10	

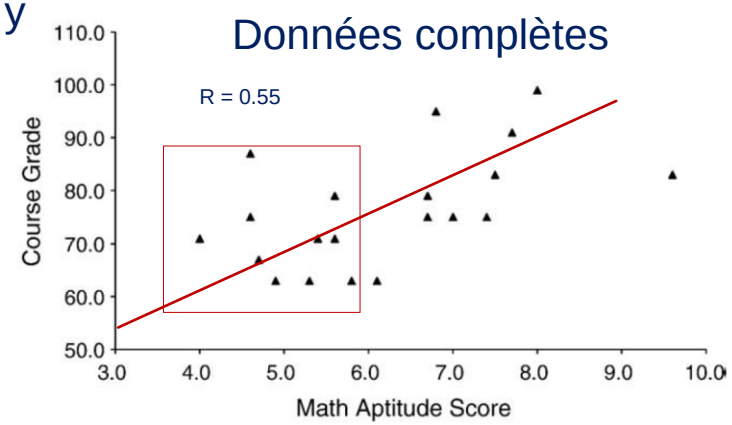
II. Imputation des données manquantes

Remplacement par la moyenne ou la médiane (2)

Méthode peu robuste et très sensible aux valeurs extrêmes

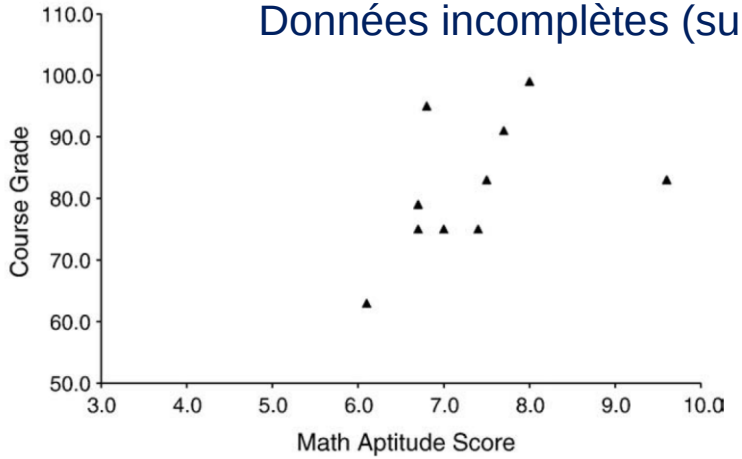
y

Données complètes



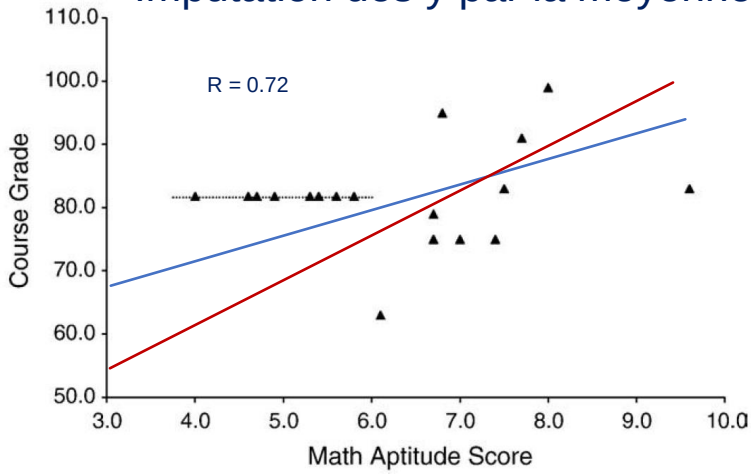
y

Données incomplètes (sur y)



Journal of School Psychology 48 (2010) 5–37

Imputation des y par la moyenne



II. Imputation des données manquantes

Remplacement par la moyenne ou la médiane (3)

- Calcul de la moyenne / médiane d'un attribut en fonction de sous groupes

Person	Age	Sex	Years of Education
1	47	M	16
2	45	F	?
3	19	M	11
4	31	F	?
5	24	M	12
6	41	F	?
7	36	M	20
8	50	M	12
9	53	F	13
10	17	M	10
11	53	F	12
12	21	F	12
13	18	F	11
14	34	M	16
15	44	M	14
16	45	M	11
17	54	F	14
18	55	F	10
19	29	F	12
20	32	F	10

← (13+12+10+14)/4 =12.25

← (11+12+10+12)/4 =11.25

Sex \ Age	<=34	>=35
M	Persons 3, 5, 10, 14	Persons 1, 7, 8, 15, 16
F	Persons 4, 12, 13, 19, 20	Persons 2, 6, 9, 11, 17, 18

Person	Age	Sex	Years of Education
1	47	M	16
2	45	F	?
3	19	M	11
4	31	F	?
5	24	M	12
6	41	F	?
7	36	M	20
8	50	M	12
9	53	F	13
10	17	M	10
11	53	F	12
12	21	F	12
13	18	F	11
14	34	M	16
15	44	M	14
16	45	M	11
17	54	F	14
18	55	F	10
19	29	F	12
20	32	F	10

← 12.25

← 11.25

← 12.25

Attention ! Nécessite expertise métier (choix des niveaux des facteurs, des attributs,…..)

II. Imputation des données manquantes

3.3 Méthodes d'imputations simples (suite)

3.3.2 Complétion par regroupement et centrage

Hot-Deck

- La valeur manquante est remplacée par une valeur observée chez un individu ayant **les mêmes caractéristiques**

Cold-Deck

- La valeur manquante est remplacée par une valeur observée chez un individu ayant **les mêmes caractéristiques mais provenant d'une autre source d'information**

« Même caractéristique » ↔ « plus proche voisin »
- métrique
- choix des variables

Méthode des k plus proches voisins

- Choix d'un entier $k : 1 \geq k \geq n$
- Calculer les distances sur les variables renseignées $d(Y_i, Y_j)$
- Retenir les k observations pour lesquelles les distances sont les plus petites $Y_{i_k}, \ldots, Y_{j_k}$
- Affecter aux valeurs manquantes la moyenne des valeurs des k voisins $Y_{miss} = \frac{1}{k}\left(Y_{i_k} + \ldots + Y_{j_k}\right)$

k plus proches voisins
- choix du k !
- métrique : distance euclidienne ou de Mahalanobis

II. Imputation des données manquantes

Imputation par la méthode des plus proches voisins (exemple)

	y1	y2	y3
1	32	*NA*	2.80
2	32	4.9	*NA*
3	40	30.0	4.38
4	10	2.8	3.21
5	6	2.7	2.73
6	20	2.8	2.81
7	32	4.6	2.88
8	32	10.9	2.90
9	32	8.0	3.28
10	30	1.6	3.20

1 *Variables renseignées (y1 et y3) en ôtant le sujet 2*

	y1	y3
1	32	2.80
3	40	4.38
4	10	3.21
5	6	2.73
6	20	2.81
7	32	2.88
8	32	2.90
9	32	3.28
10	30	3.20

$d(Y_i, Y_j)$

	Distances
1-3	8.154532
1-4	22.003820
1-5	26.000094
1-6	12.000004
1-7	0.080000
1-8	0.100000
1-9	0.480000
1-10	2.039608

Calcul de la distance (sur les variables renseignées) 2

3 *Estimation de la valeur manquante en fonction de nombre de plus proches voisin k*

	y1	y2	**y3**
1	**32**	***NA***	**2.80**
2	32	4.9	*NA*
3	40	30.0	4.38
4	10	2.8	3.21
5	6	2.7	2.73
6	20	2.8	2.81
7	**32**	**4.6**	**2.88**
8	**32**	**10.9**	**2.90**
9	**32**	**8.0**	**3.28**
10	30	1.6	3.20

k = 1: estim = 4.6

k = 2 : estim = (4.6+10.9)/2 = 7.4

k = 3 : estim = (4.6+10.9 + 8)/ 3 = 7.53

$$Y_{miss} = \frac{1}{k}\left(Y_{i_k} + \ldots + Y_{j_k}\right)$$

II. Imputation des données manquantes

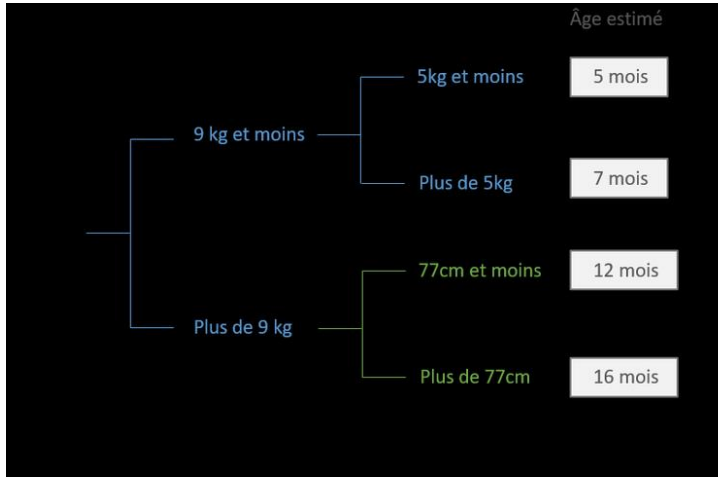
- Méthode par arbre (juste un aperçu...)
  - On remplace les données manquantes en utilisant un arbre de décision
  - Le critère de segmentation dépend du type d'arbre (entropie, gini, Khi-deux)
  - les arbres permettent d'identifier des «zones» où les observations sont homogènes.

Age	Poids	Taille
5	4,8	60
5	5,3	66
NA	6,3	67
6	7,2	67
9	8,2	66

- On procède alors à une estimation locale des paramètres (moyenne, médiane,….)

Age	Poids	Taille
9	8,2	66
15	10,5	79
5	5,3	66
6	7,2	67
16	10,1	80
17	11,2	82
NA	10,9	69
NA	6,3	67
5	4,8	60
13	10,5	76

Age	Poids	Taille
5	4,8	60
5	5,3	66
NA	6,3	67
6	7,2	67
9	8,2	66
16	10,1	80
15	10,5	79
13	10,5	76
NA	10,9	69
17	11,2	82

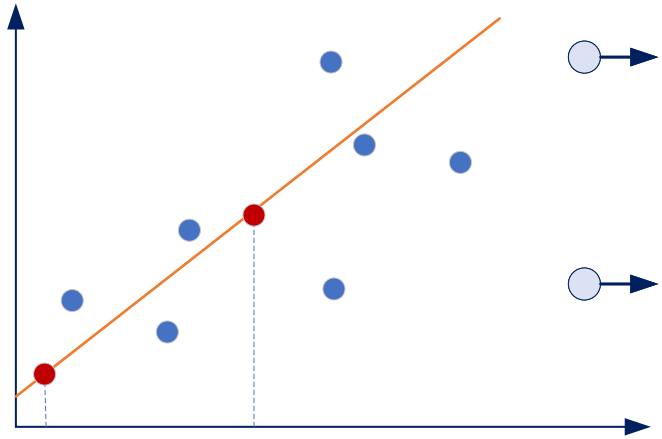


II. Imputation des données manquantes

Méthode par régression

- Remplacement d'une valeur manquante par une valeur prédite obtenue par régression

	Y1	Y2
1		NA
2		
3		
4		
...		NA
...		
...		
...		
...		
N		

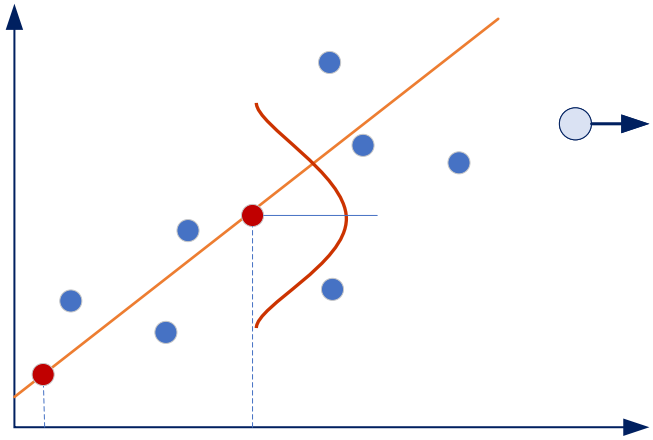


On effectue la régression sur les cas complets

$$\hat{\beta}_c = \left(Y_1^t Y_1\right)^{-1} Y_1^t Y_2 \qquad \hat{\beta}_c = (\alpha, \beta)$$

Imputation à partir du modèle de régression

$$y_{i,2} = \alpha + \beta y_{i,1}$$



Possibilité d'ajouter un aléa à la prédiction

$$y_{i,2} = \alpha + \beta y_{i,1} + \varepsilon_i \qquad \varepsilon_i \rightarrow N(y_{i,2}, \sigma^2)$$

II. Imputation des données manquantes

Méthode par régression multiple (et itérative)

	y1	y2	y3
1	32	*NA*	2.80
2	32	4.9	*NA*
3	40	30.0	4.38
4	10	2.8	3.21
5	6	2.7	2.73
6	20	2.8	2.81
7	32	4.6	2.88
8	32	10.9	2.90
9	32	8.0	3.28
10	30	1.6	3.20

	y1	y2	y3
1	32	*NA*	2.80
2	32	4.9	*NA*
3	40	30.0	4.38
4	10	2.8	3.21
5	6	2.7	2.73
6	20	2.8	2.81
7	32	4.6	2.88
8	32	10.9	2.90
9	32	8.0	3.28
10	30	1.6	3.20

1. Régression de y_2 sur les variables y_1 et y_3

$$\hat{y}_{i,2} = \hat{\alpha}_0 + \hat{\alpha}_1 y_{i,1} + \hat{\alpha}_2 y_{i,3}$$

$$\hat{y}_{1,2} = -39.9 + 0.15(32) + 13.8(2.80) = 4.09$$

$$\hat{\alpha}_0 = -39.9,\quad \hat{\alpha}_1 = 0.15\quad \hat{\alpha}_2 = 13.80$$

	y1	y2	y3
1	32	4.09	2.80
2	32	4.9	*NA*
3	40	30.0	4.38

2. Régression de y_3 sur les variables y_1 et y_2

$$\hat{y}_{i,3} = \hat{\beta}_0 + \hat{\beta}_1 y_{i,1} + \hat{\beta}_2 y_{i,2}$$

$$\hat{y}_{2,3} = -40.9 + 0.13(40) + 14.05(4.9) = 3.69$$

	y1	y2	y3
1	32	4.09	2.80
2	32	4.9	3.69
3	40	30.0	4.38

II. Imputation des données manquantes

- Méthode par régression multiple (et itérative)… *suite*

	y1	y2	y3
1	32	4.09	2.80
2	32	4.9	3.69
3	40	30.0	4.38
4	10	2.8	3.21
5	6	2.7	2.73
6	20	2.8	2.81
7	32	4.6	2.88
8	32	10.9	2.90
9	32	8.0	3.28
10	30	1.6	3.20

3. On réitère la régression de y_2 sur les variables y_1 et y_3 en prenant toutes les valeurs

$$\hat{y}_{1,2} = 3.69$$

4. On réitère la régression de y_3 sur les variables y_1 et y_2 en prenant toutes les valeurs

$$\hat{y}_{2,3} = 2.97$$

	y1	y2	y3
1	32	3.69	2.80
2	32	4.9	2.97
3	40	30.0	4.38

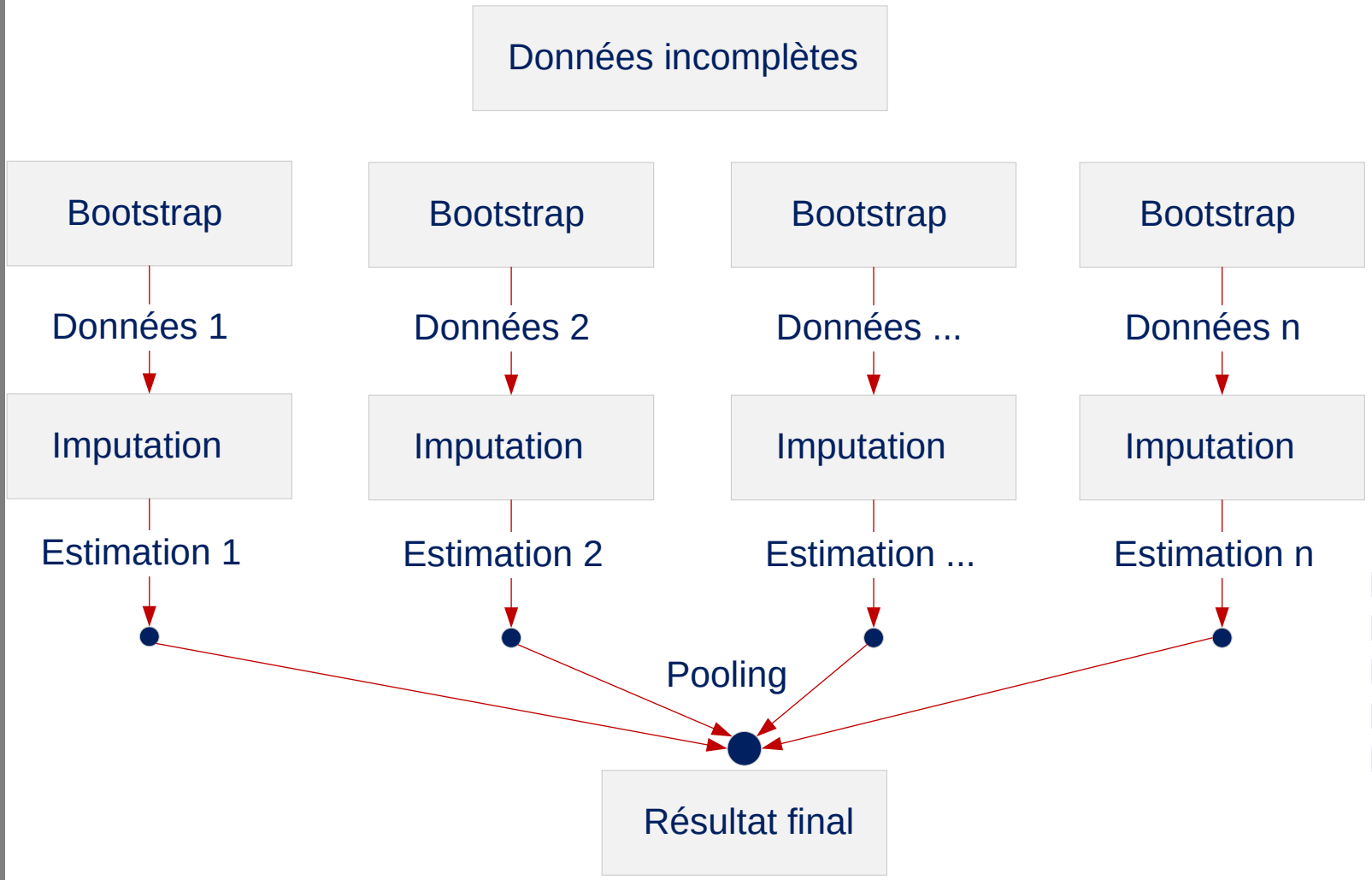
... On effectue les calculs jusqu'à ce que les estimations ne changent plus

- **Propriétés**
  - On peut utiliser différents modèles de régression, simple, multiple, polynomiale, logistique,...
  - Fournit une estimation ponctuelle correcte mais sous estime la variance
  - La « correction » de la variance peut se faire en introduisant un aléa (à la condition que l'on connaisse ou que l'on suppose a priori la distribution des erreurs

II. Imputation des données manquantes

3.4 Méthodes d'imputations multiples

Consistent à créer plusieurs valeurs possibles d'une valeur manquante (notamment par bootstrap)



EM algorithme,
Forêt aléatoire,
Régression multiple,
Modèles Bayésiens
Knn,...

Statistique

III. Recodage des variables

1 Introduction

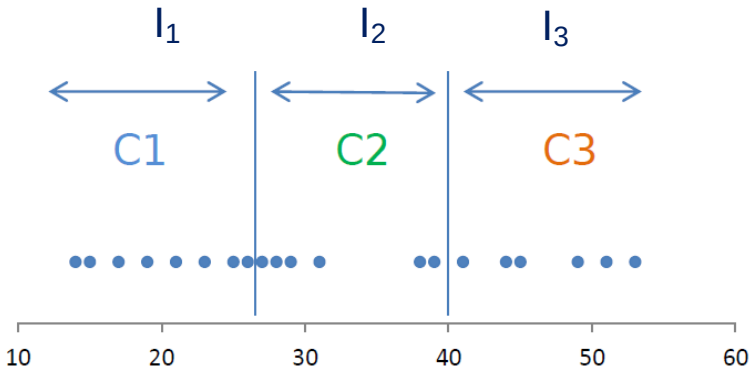
- La discrétisation consiste à transformer une variable quantitative en une variable qualitative ordinale. Elle opère par découpage de la variable en classe.
  - Comment déterminer les bornes de chaque intervalle ?
  - Comment déterminer le nombre d'intervalles ?
- Son utilisation est très fréquente en apprentissage statistique
  - Certaines méthodes supervisées ne manipulent que des descripteurs qualitatifs (exemple classification bayésienne si l'on dispose pas à priori de la loi de distribution de(s) la variable(s))
  - Les variables discrétisées sont souvent plus faciles à appréhender lors de l'interprétation
  - Certaines procédures d'apprentissage s'avèrent souvent beaucoup plus rapide. C'est la cas notamment des méthodes supervisées par arbres de décision
  - La discrétisation est une approche simple et efficace pour harmoniser les bases hétérogènes et contenant des descripteurs de nature différentes
- On distingue deux grandes approches
  - Les méthodes non supervisées
  - Les méthodes supervisées

III. Recodage des variables

2 Les méthodes non supervisées

2.1 Découpage en amplitude → Intervalles de largeur (amplitude) égale

$I_1 = I_2 = I_3$



- On fixe a priori k, le nombre de classes
- $I = \frac{\max - \min}{K}$

$b_1 = \min + I$

$b_2 = \min + 2I$

$b_n = \min + nI$

On calcule la largeur de chaque intervalle

On en déduit les k-1 bornes

Le problème est donc de déterminer K.

- Différentes méthodes se proposent d'estimer dans un premier temps un intervalle « optimal » permettant de calculer K

III. Recodage des variables

→ Estimation de l'amplitude des intervalles (K)

Appellation	Formule
Brooks-Carruthers	$5 \times \log_{10}(n)$
Huntsberger	$1 + 3.332 \times \log_{10}(n)$
Sturges	$\log2(n + 1)$
Scott	$(max - min)/(3.5 \times \sigma \times n^{-1/3})$
Freedman-Diaconis	$(max - min)/(2 \times IQ \times n^{-1/3})$

σ: écart-type
IQ : intervalle interquartiles

Les deux dernières approches exploitent plus d'informations en provenance des données

→ Avantages

- Rapidité de calcul et simplicité
- Ne modifie pas la forme de la distribution

→ Inconvénient

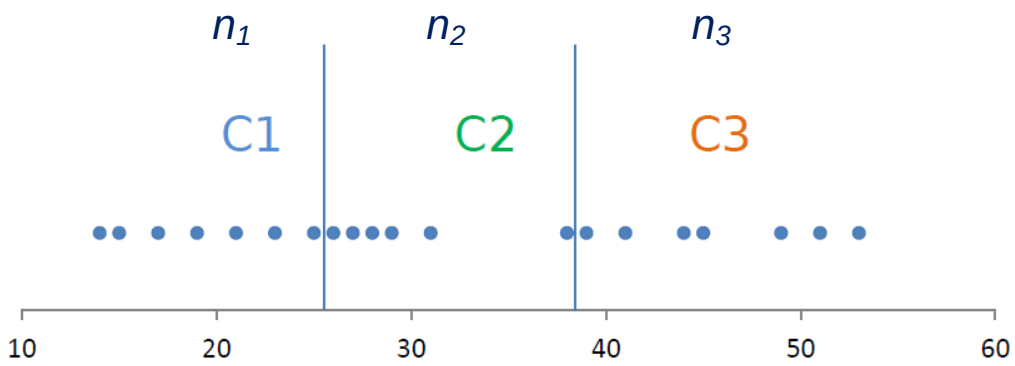
- Choix du K
- Sensibilité aux points extrêmes
- Possibilité d'avoir des intervalles avec peu d'individus (voire vides)

III. Recodage des variables

2.2 Découpage en fréquences égales

- Les amplitudes des intervalles sont différentes mais le nombre d'individus (n) dans chaque intervalle est le même

$n_1 = n_2 = n_3$



- Ce découpage égalise la distribution des données
- Les amplitudes peuvent être calculées à partir des quantiles

→ Avantages

- Rapidité de calcul et simplicité
- Intervalle avec un nombre déterminée d'individus
- Egalisation de la distribution des données (distribution uniforme)

→ Inconvénients

- Choix du K (donc des percentiles et des quantiles)
- Seuils ne tenant pas compte des proximités entre les valeurs

2.3 Moyennes emboitées

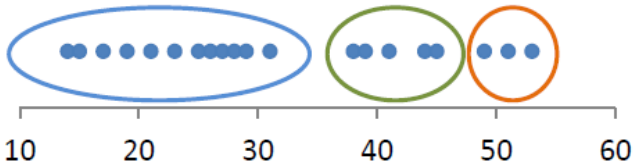
Algorithme descendant. On découpe avec la moyenne. Puis, de part et d'autre de ce premier seuil, on découpe avec les moyennes locales respectives, etc. Nombre de classes est forcément une puissance de 2.

2.4 Grandes différences relatives *(pour info)*

On trie les données de manière croissante. On repère les grands écarts entre 2 valeurs successives. On découpe si écart > seuil exprimé en % de l'écart-type des valeurs (ou en % du MAD – median absolute deviation – si l'on souhaite se prémunir du problème des valeurs aberrantes).

2.5 Méthodes descendantes

- La dispersion des données peut montrer des « regroupements » en paquets plus ou moins homogènes. On s'intéresse alors aux caractéristiques de la dispersion

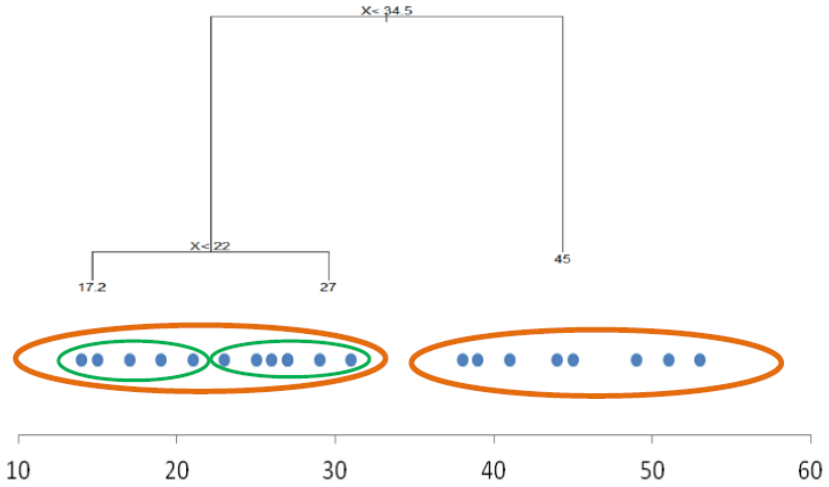


III. Recodage des variables

- Trouver la meilleure séparation binaire et continuer récursivement dans chaque sous groupe jusqu'au déclenchement d'une règles d'arrêt

  - La règle d'arrêt peut dépendre l'intervalle de classe, du nombre de classes ou des effectifs par classes

  - On peut utiliser une régression par arbre binaire (CART par exemple). Elle consiste à construire des subdivisions qui maximisent les dispersions interclasses (Attention cependant à la forte sensibilité au paramètrage de la construction de l'arbre)



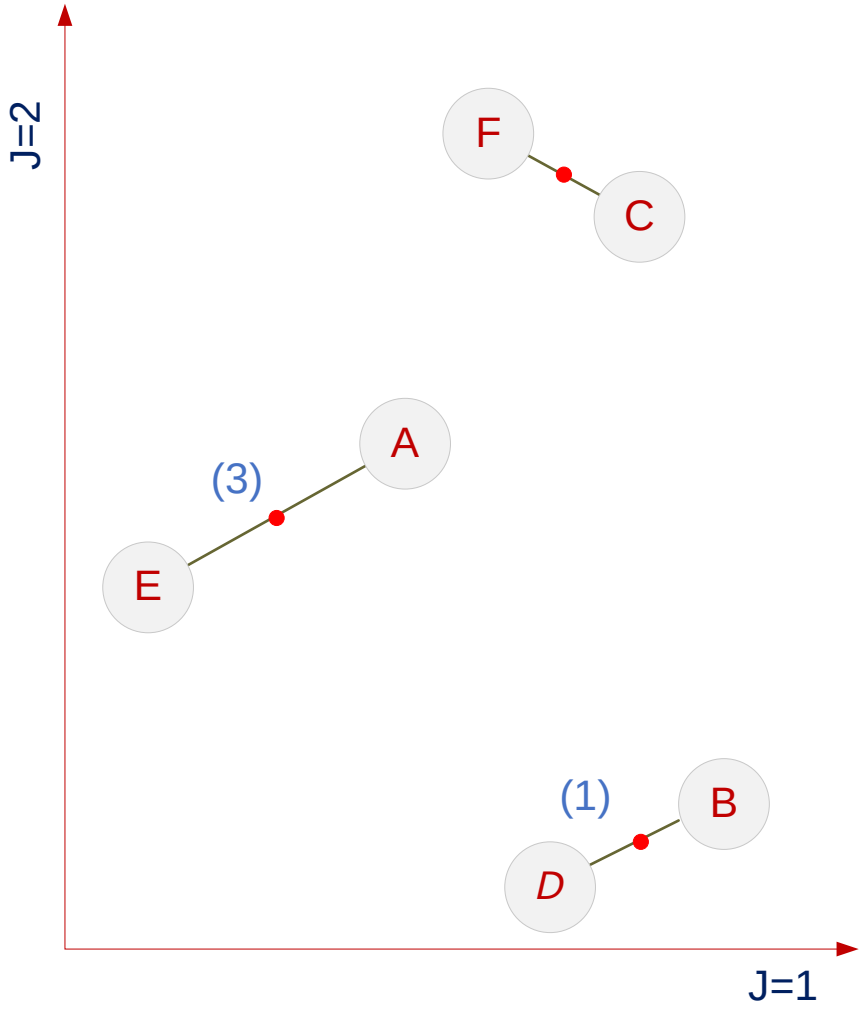
2.5 Méthodes ascendentes

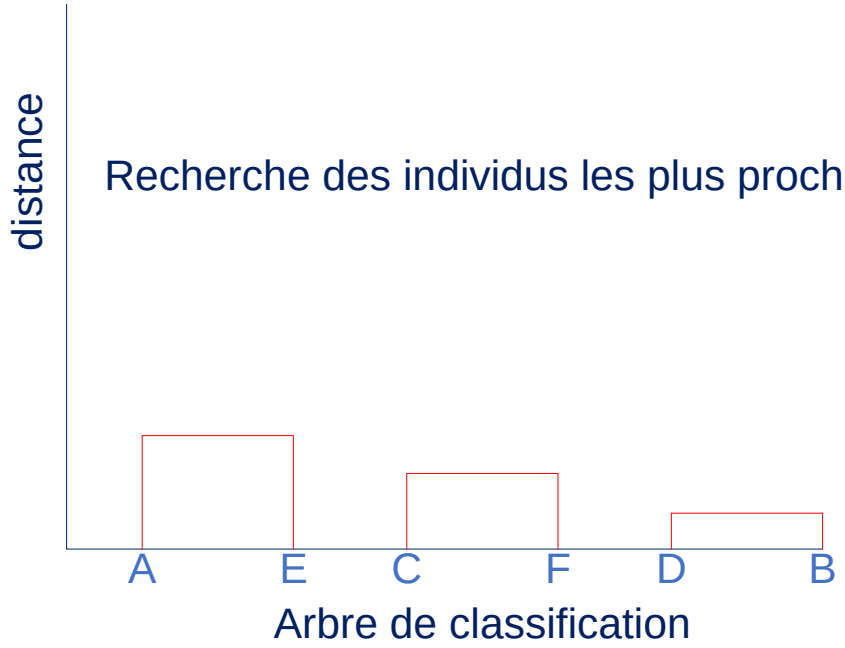
- L'objectif est le même que pour l'approche descendante, il s'agit de tenir compte de la dispersion des données au sein d'une distribution en utilisant un algorithme récursif qui effectue des regroupements

- On utilisera la classification ascendante hiérarchisée (CAH)

III. Recodage des variables

Processus

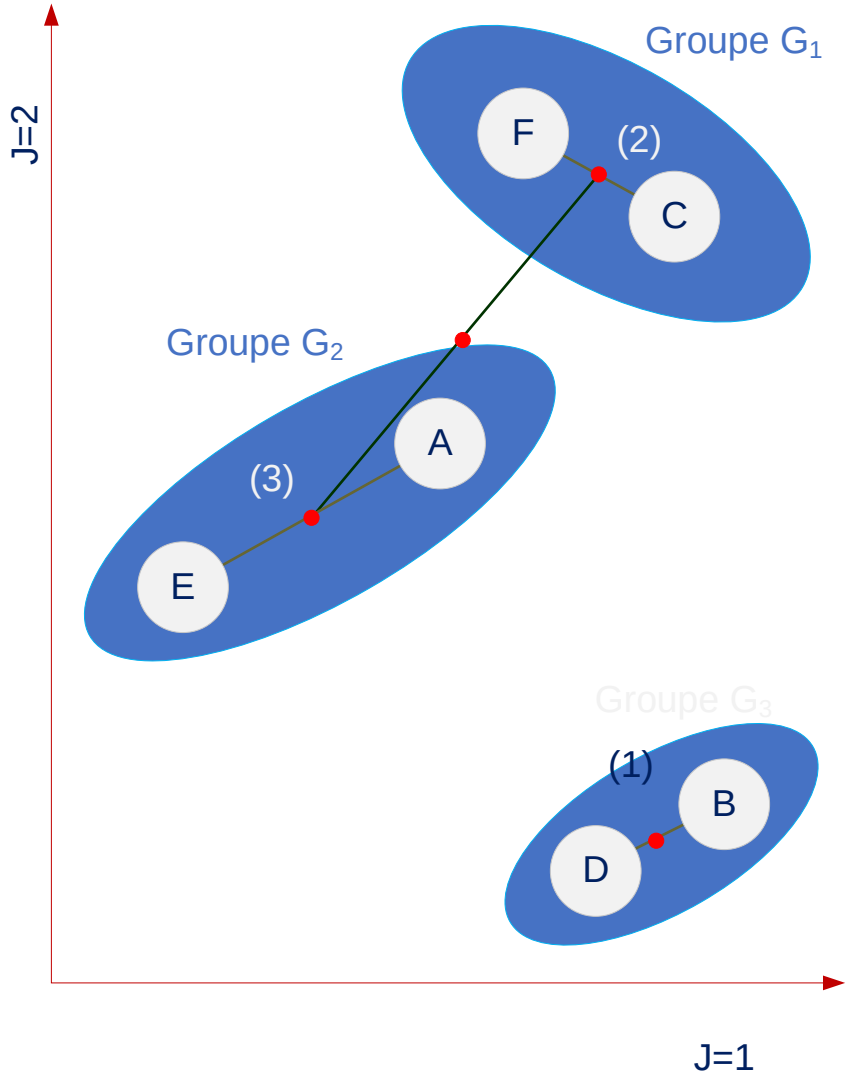


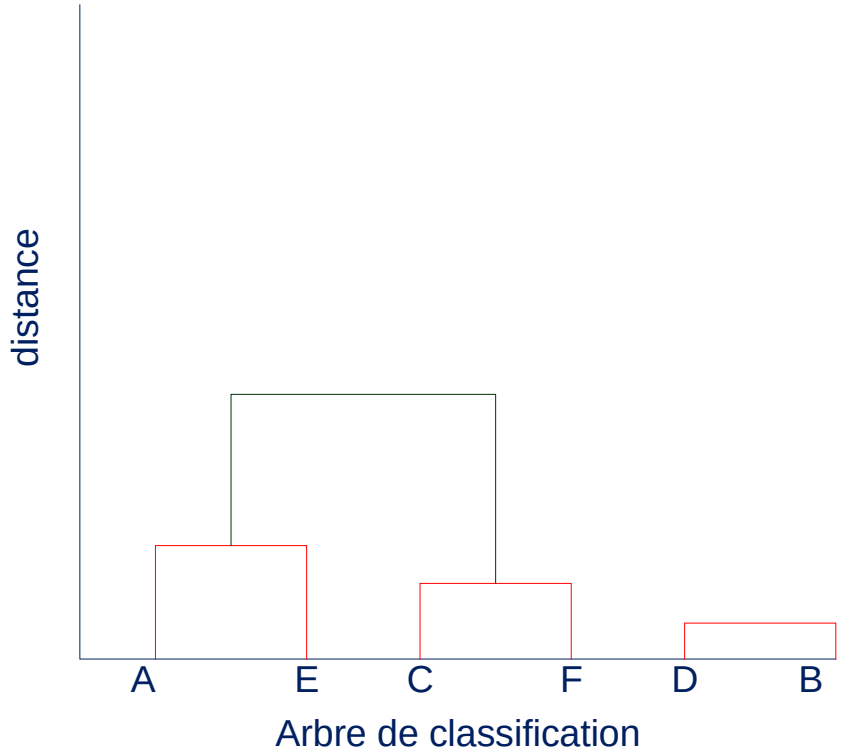


Recherche des individus les plus proches (2 à 2)

Arbre de classification

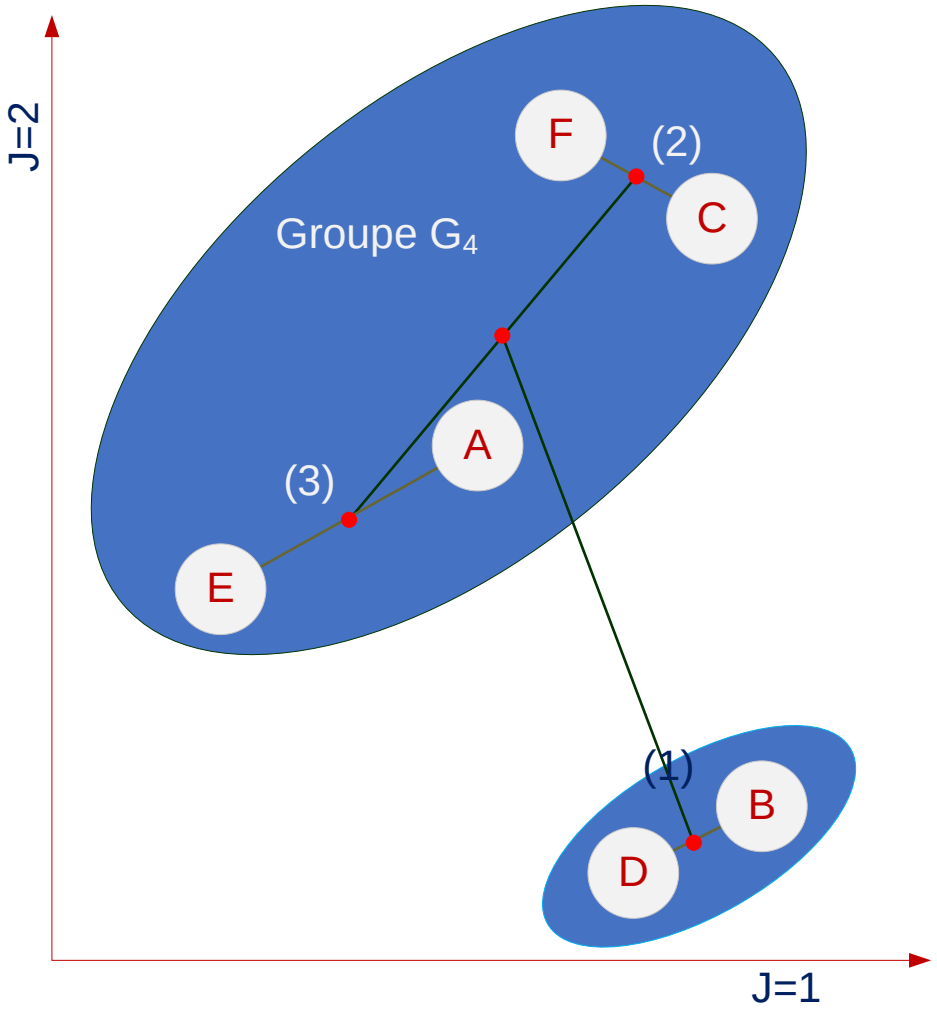
III. Recodage des variables

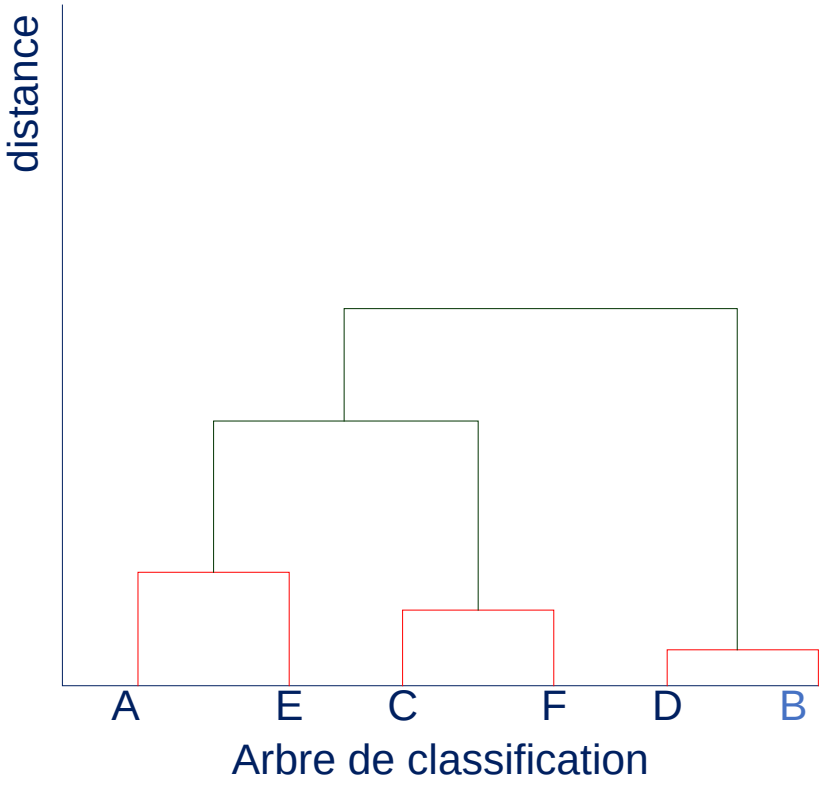




III. Recodage des variables

La hauteur des branches correspond à la distance entre les éléments regroupés





Arbre de classification

III. Recodage des variables

- Remarques sur les méthodes descendantes et ascendantes
  - On tient compte de la dispersion des données (qui peut être quantifiée)
  - Produit des classes compactes
  - Méthodes descendantes nettement plus rapide que les méthodes ascendantes