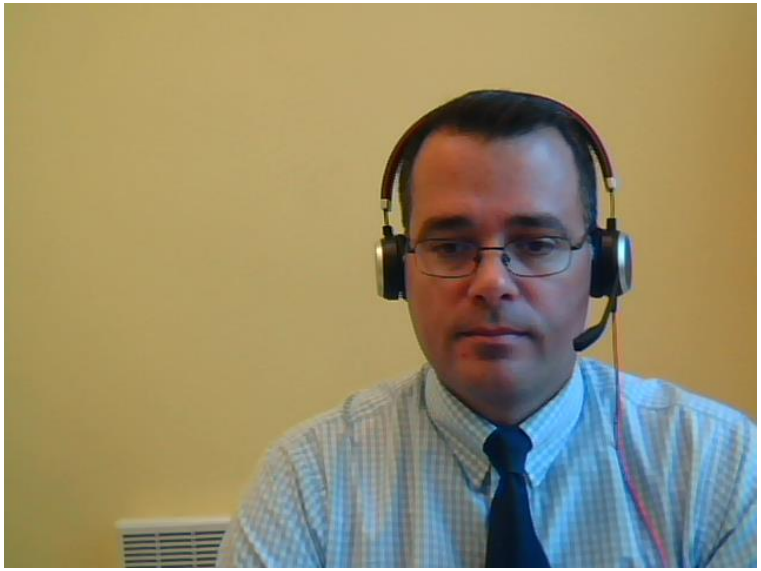
Imputation de données manquantes en réutilisation de données

I. Données manquantes
II. Imputation en recherche sur les personnes
III. Imputation en recherche sur des données



Données manquantes

- Prosaïquement, « case vide » dans un tableau
- Codage
  - Logiciels de statistique : NA (*not available*)
  - Excel : laisser vide
  - Eviter les codages type 999 ou -1…
- Notion relative, dépend également du formatage

Représentation « informaticiens »

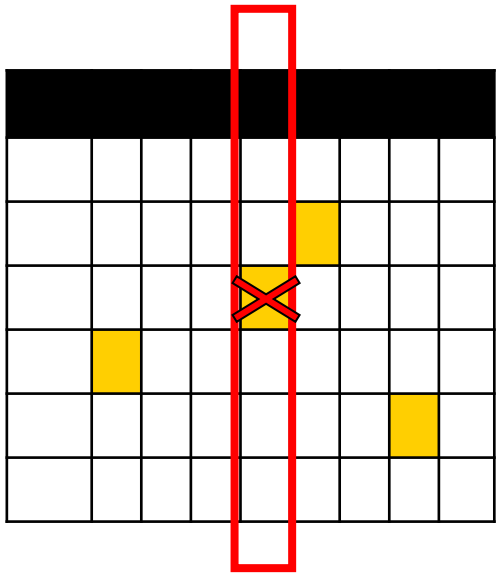
Personne	Date	Valeur
Josette	Lundi	4.3
Josette	Jeudi	3.8
Lucien	Lundi	3.4

Représentation « épidémiologistes »

Personne	Lundi	Jeudi
Josette	4.3	3.8
Lucien	3.4	NA

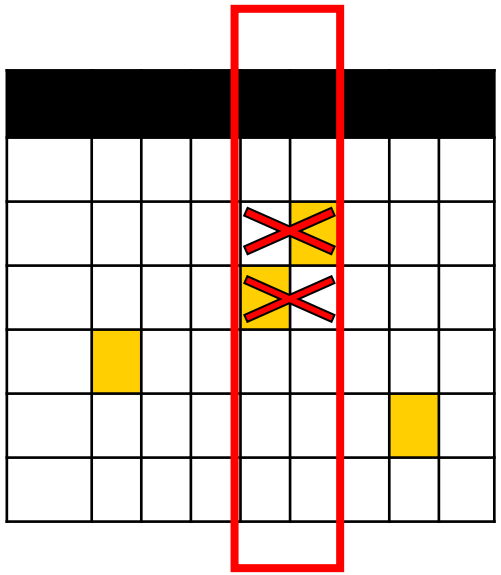
Traitements habituels : analyses univariées

- Analyse d'une seule colonne à la fois
- Si proportion acceptable (<10%) :
  - analyse sans les NA (exclusion automatique)
  - +/- signaler la proportion de valeurs manquantes
- Si proportion inacceptable : éviter d'analyser…
- Ex :
  - L'âge moyen est de 84,5 ans (DS=14,8 ; 1,2% de valeurs manquantes)

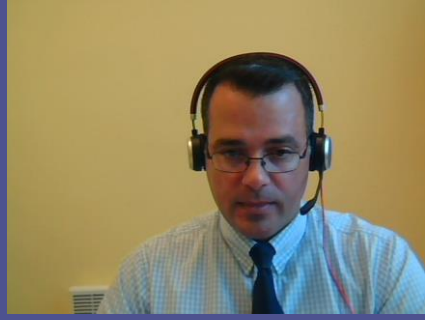


Traitements habituels : analyses bivariées

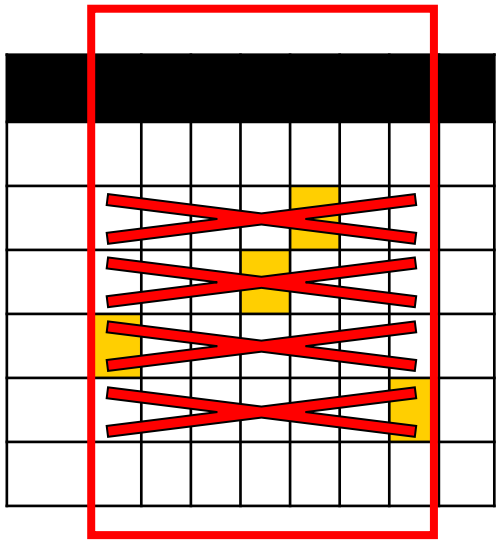
- Analyse de deux colonnes à la fois
- Si proportion de NA acceptable :
  - analyse des « cas complets » (exclusion automatique)
  - +/- signaler la proportion de valeurs manquantes
- Si proportion inacceptable : éviter d'analyser…

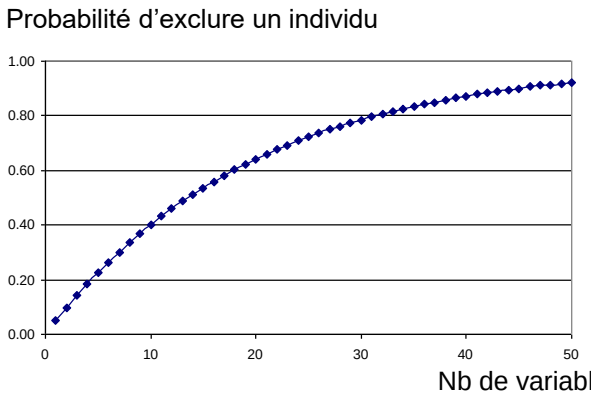


Traitements habituels : analyses multivariées

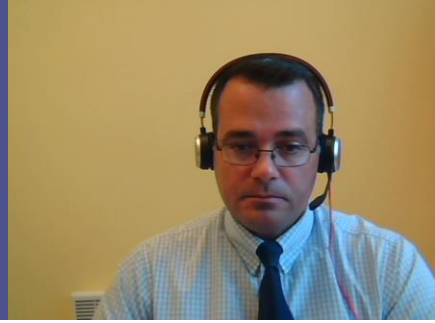


- Analyse de plus de 2 colonnes à la fois
- Problème :
  - Analyse en cas complet => un « gruyère » de données !
  - Les sujets restants sont rares, et peut-être peu représentatifs
  - Plus on incorpore de colonnes, plus on exclut d'individus !
- Autre solution parfois nécessaire : imputer les données manquantes





En recherche traditionnelle « sur les personnes »



- Protocole de recueil de données défini à l'avance
- Financement de personnel dédié au recueil (ARC = attachés de recherche clinique)
- Hypothèse indispensable (est souvent raisonnable dans ce cadre) :
  - les données sont manquantes par hasard (*missing completely at random*, *MCAR*)
  - Le fait qu'une valeur soit manquante ne donne aucune information sur sa valeur réelle
- Alors on peut
  - imputer les données manquantes par des procédés mathématiques (ex : NIPALS…)
  - (utile essentiellement pour les analyses multivariées)

En recherche « sur les données » (*data reuse*)

- Recherches réalisées sur des données existantes, recueillies en routine
  - Dossiers patient (électronique ou papier)
  - Bases nationales des assurances santé
  - etc.
- Hypothèse de données manquantes par hasard :
  - Elle est totalement délirante !
  - => illustration à venir
- Imputation
  - Ne pas utiliser les méthodes mathématiques usuelles
  - => solutions valides et simples à venir

Exemple très réaliste

- On recherche une fracture du fémur à l'aide d'une radio de la cuisse dans 1815 séjours pris au hasard
- Données réelles : variable « Fracture du fémur »
  - 1 (radio faite, fracture présente) : 2
  - 0 (radio faite, fracture absente) : 3
  - *NA* (radio pas faite) : 1810
- Interprétation :
  - 40% de valeurs 1
  - Mais 99,7% de valeurs manquantes
  - En vrai, la proportion de fractures parmi les NA est non pas de 40% mais de 0,55%
  - Parce que les examens ne sont réalisés qu'en cas de présomption de positivité ! Motifs évidents : éthiques, économiques et de disponibilité
- Solutions :
  - Imputation « classique » de données manquantes : absolument inacceptable
  - Imputation par la valeur « 0 » : tout à fait acceptable dans ce contexte
    - 0 (fracture absente) : 1810+3 = 1813 (au lieu de 1803, effectif réel)
    - 1 (fracture présente) : 2 (au lieu de 12, effectif réel)
  - Imputation par une valeur qualitative « non mesuré » : très pertinent également, selon…
    - Fracture « absente » : 3
    - Fracture « présente » : 2
    - Fracture « non mesuré » : 1810

Pour information, données source

- L'exemple précédent est ultraréaliste
- Base nationale du PMSI, une année, France entière :
  - Code CCAM : NBQK001 Radiographie de la cuisse
  - Code CIM10 : S72 fracture du fémur

		fracture fémur		
		non	oui	
radio cuisse	non	17 770 581	95 583	*17 866 164*
	oui	28 605	20 666	*49 271*
		17 799 186	*116 249*	*17 915 435*

- Réduction à 1/10000 :

		fracture fémur		
		non	oui	
radio cuisse	non	1 800	10	*1 810*
	oui	3	2	*5*
		1 803	*12*	*1 815*

Merci de votre attention !

Objectif Thèse

http://objectifthese.org
contact@objectifthese.org