Mise en œuvre d'études en statistique descriptive : méthode et outils

Dr Jean-Charles DUFOUR

jean-charles.dufour@univ-amu.fr

SESSTIM (Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale) UMR 912

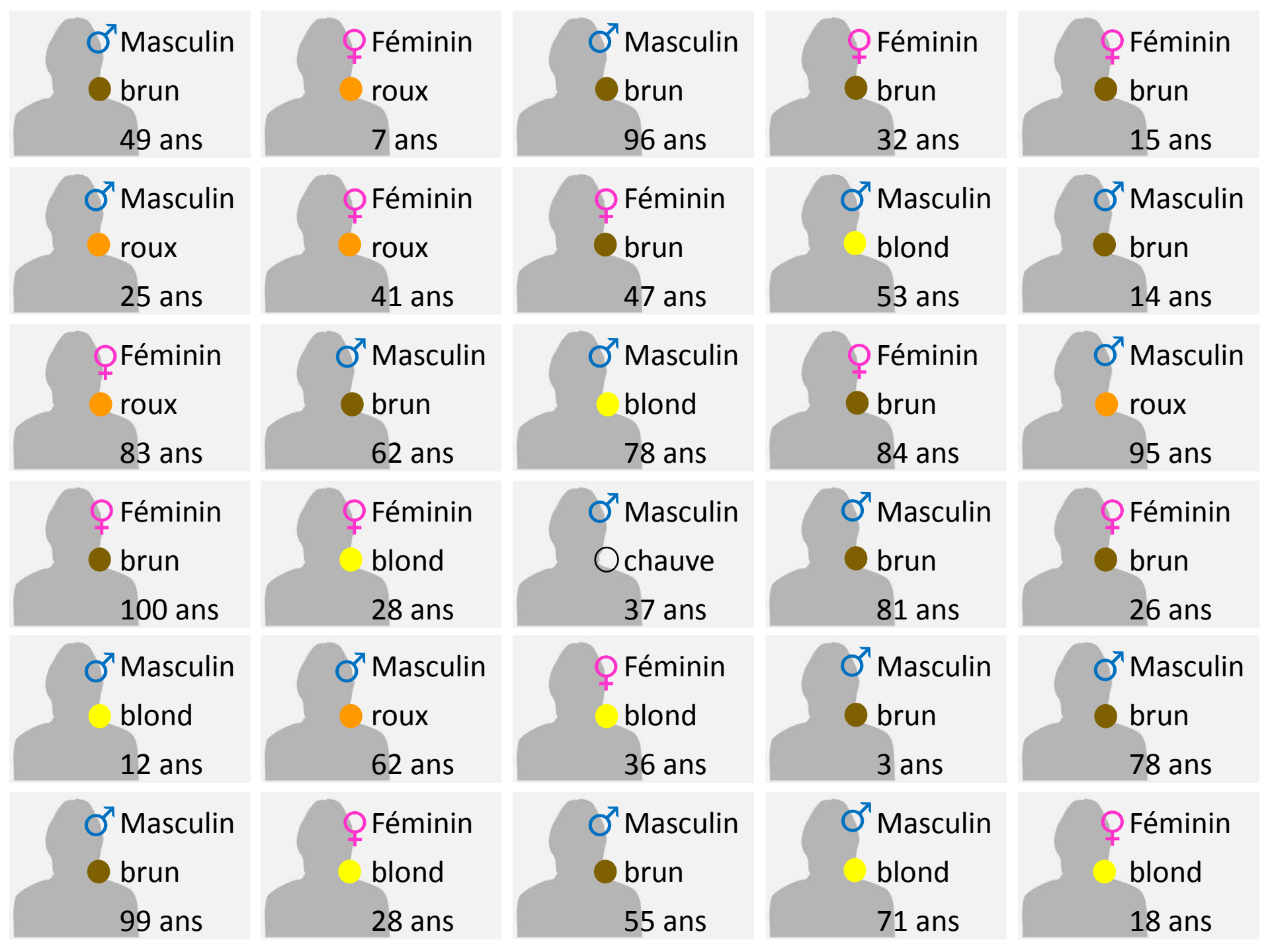
Plan du cours

- Nos pauvres cerveaux ont-ils vraiment besoin des stats ?
  - Qu'est ce qu'une étude descriptive
  - Notion d'individus

- Admettons que oui : comment s'y prendre ?
  - Les étapes de la mise en œuvre d'une étude descriptive
  - Passage à l'acte : les grands principes à respecter pour construire un questionnaire et une base de données

Pourquoi faire une étude descriptive

Que pouvez-vous dire à propos de ces individus


Masculin
brun
49 ans
Féminin
roux
7 ans
Masculin
brun
96 ans
Féminin
brun
32 ans
Féminin
brun
15 ans
Masculin
roux
25 ans
Féminin
roux
41 ans
Féminin
brun
47 ans
Masculin
blond
53 ans
Masculin
brun
14 ans
Féminin
roux
83 ans
Masculin
brun
62 ans
Masculin
blond
78 ans
Féminin
brun
84 ans
Masculin
roux
95 ans
Féminin
brun
100 ans
Féminin
blond
28 ans
Masculin
chauve
37 ans
Masculin
brun
81 ans
Féminin
brun
26 ans
Masculin
blond
12 ans
Masculin
roux
62 ans
Féminin
blond
36 ans
Masculin
brun
3 ans
Masculin
brun
78 ans
Masculin
brun
99 ans
Féminin
blond
28 ans
Masculin
brun
55 ans
Masculin
blond
71 ans
Féminin
blond
18 ans

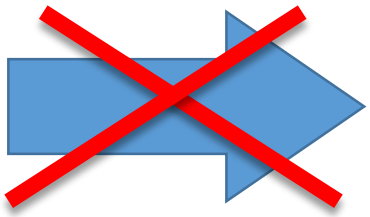
Que pouvez-vous dire à propos de ces individus statistiques

Qu'est-ce qu'une étude descriptive ?

- C'est une étude qui **décrit** certaines **caractéristiques** d'un **groupe** d'**individus**
- Décrire : c'est en fait résumer par des grandeurs statistiques (*moyennes, médianes, modes, écarts-types, …*) ou représenter par des graphiques (*histogrammes, nuages de points, …*) des données disponibles pour chacun des individus
- Paradoxe ? : la description du groupe passe d'abord par le recueil d'informations à un niveau individuel

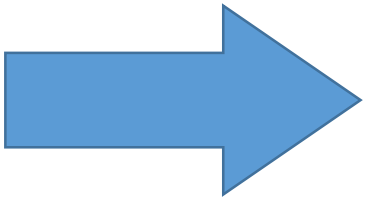
Qu'est-ce qu'une étude descriptive ?

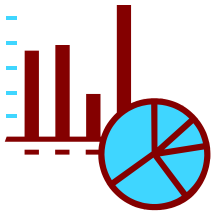
Ce que notre cerveau ne sait pas faire seul

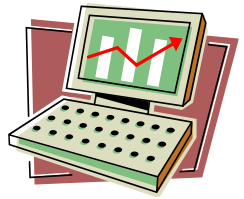


… les statistiques descriptives peuvent l'y aider

♂ Masculin brun 49 ans	♀ Féminin roux 7 ans	♂ Masculin brun 96 ans	♀ Féminin brun 32 ans	♀ Féminin brun 15 ans
♂ Masculin roux 25 ans	♀ Féminin roux 41 ans	♀ Féminin brun 47 ans	♂ Masculin blond 53 ans	♂ Masculin brun 14 ans
♀ Féminin roux 83 ans	♂ Masculin brun 62 ans	♂ Masculin blond 78 ans	♀ Féminin brun 84 ans	♂ Masculin roux 95 ans
♀ Féminin brun 100 ans	♀ Féminin blond 28 ans	♂ Masculin chauve 37 ans	♂ Masculin brun 81 ans	♀ Féminin brun 26 ans
♂ Masculin blond 12 ans	♂ Masculin roux 62 ans	♀ Féminin blond 36 ans	♂ Masculin brun 3 ans	♂ Masculin brun 78 ans
♂ Masculin brun 99 ans	♀ Féminin blond 28 ans	♂ Masculin brun 55 ans	♂ Masculin blond 71 ans	♀ Féminin blond 18 ans







Qu'est-ce que n'est pas une étude descriptive ?

- La statistique descriptive se limite à analyser des observations : elle à pour but de décrire ce qui est, et non de tester des hypothèses

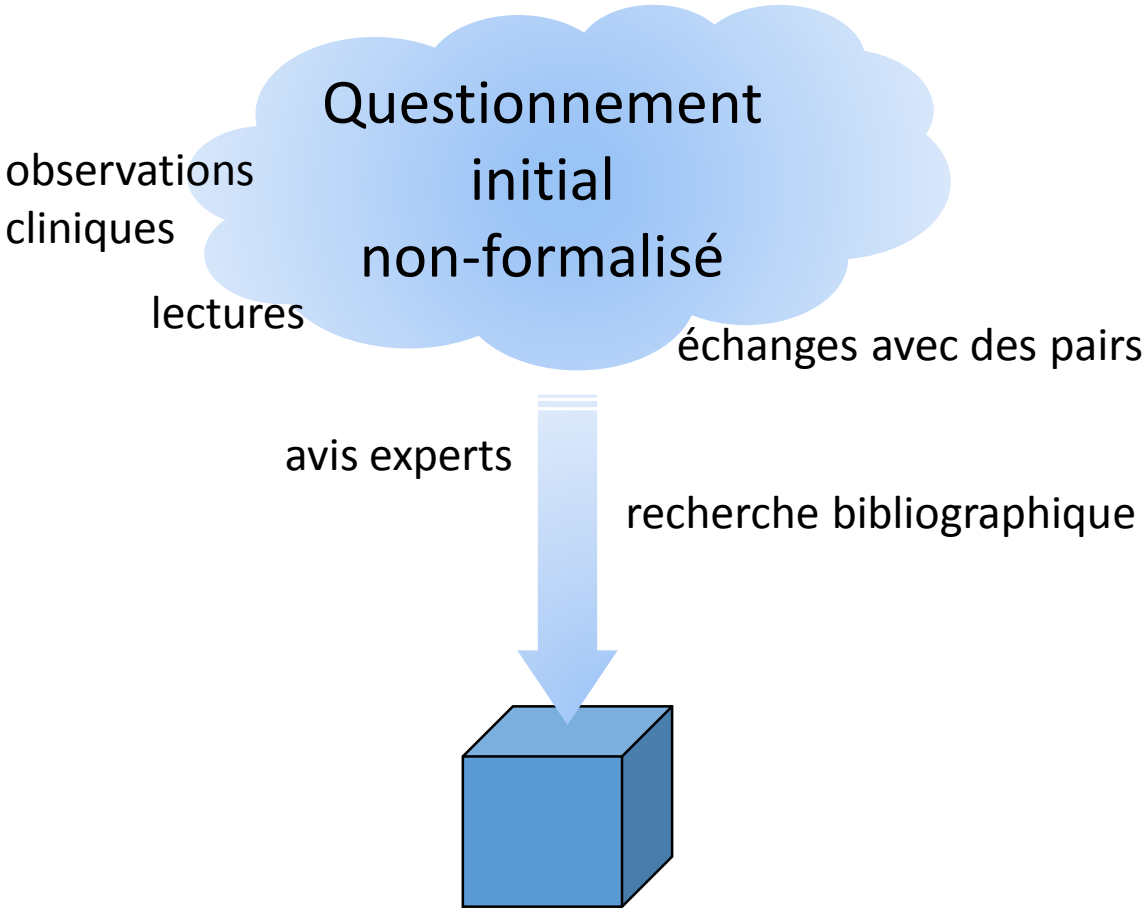
- La statistique inférentielle permet de déduire (avec une certaine marge d'erreur) des caractéristiques d'une population en partant de données décrites seulement sur un échantillon de cette population

Mise en œuvre d'une étude descriptive

Plusieurs étapes

1) Formulation du problème
2) Choix d'une méthodologie et élaboration du design de l'étude
3) Collecte les données
4) Préparation et analyse des données
5) Rédaction du rapport d'étude

1) Formulation du problème



Question scientifique formalisée :

Problème(s) à décrire, hypothèse(s) à vérifier

2) Méthodologie et design de l'étude

- **Définition de la population à décrire**
  - Echantillon *vs* Toute la Population
  - Critères d'inclusion / critères d'exclusion
  - Caractéristiques des individus statistiques (détermination des variables)

- **Type d'étude descriptive :**
  - Itération :
    - *Ponctuelle (ou transversale)* : recueil unique
    - *Longitudinale* : recueils répétés
  - Distance temporelle avec l'événement observé :
    - *Prospective*
    - *Rétrospective*

3) Collecte des données : sources des données

- **Données primaires**
  observations, entretiens individuels, sondages
  → structurer correctement les données colligées
  (questionnaires *ad hoc* et base de données)

- **Données secondaires** préexistantes
  (ex: dossiers de soins, PMSI, INSEE, ….)
  → bien connaitre le contexte et le mode de recueil initial

3) Collecte des données : modalité de recueil

- Enquêtes et interrogatoires ciblés
  - Par téléphone
  - En face-à-face
  - Par correspondance
  - Par internet
- Observation systématique
  - Par le sujet lui-même ou par un observateur externe
  - Observation automatisée / mécanique
  - Analyse de traces (numériques par exemple)

3) Collecte des données : points clés

- Étape à coupler, si possible, avec la construction correcte de la base de données
- Choix techniques et organisationnels pratiques ++
- Objectif majeur = conserver et organiser de manière cohérente les données colligées

4) Préparation et analyse des données

- La préparation est plus ou moins importante/facile en fonction de la qualité de la base de données
- A minima :
  - quelques regroupement de données (calculs sommes, moyennes,…)
  - quelques vérifications (ex: erreurs de saisies, changements d'unités, …)
- Peut tourner au cauchemar :

  - base mal conçue, mal structurée
  - variables mal pensées
  - saisie non contrôlées/non-guidées

4) Préparation et **analyse des données**

- Analyse descriptives fait appel aux méthodes vues dans les cours précédents :
  - Regroupements, calculs de grandeurs statistiques (moyennes, variances, écarts-types, quantiles, …)
  - Tableaux de contingences et de présentation des données
  - Représentations graphiques (histogramme, diagramme en secteurs, nuages de points,…)

5) Rédaction du rapport d'étude

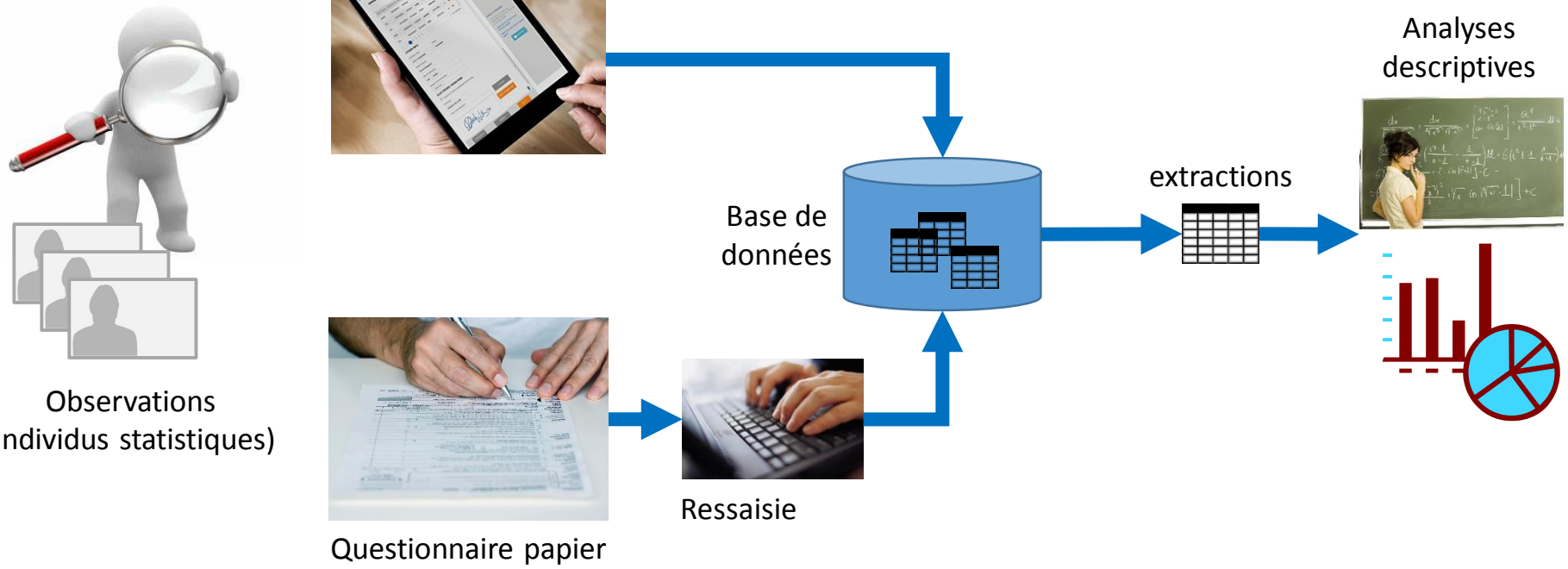
Sort du cadre de ce cours …

…mais doit exposer clairement les étapes 1 à 4 vues précédemment !

Comment colliger les données…

…et constituer la base de données

Schéma classique (données primaires / questionnaire ad hoc)



Conception du formulaire

- C'est un sujet en soi : seules quelques notions sont énoncées ici !

- Organisation du questionnaire
  - Présentation du questionnaire et de l'étude
  - **Questions introductives et qualifiantes**
  - **Questions spécifiques**
  - **Questions d'identification**
  - Remerciements !

Conception du formulaire

- **Règles de base pour la formulation des questions**
  1. Traiter un point unique par question
  2. Formulations précises et concises
  3. Utiliser un langage simple et adapté
  4. Rester neutre dans la formulation
  5. Éviter les exemples et les généralisations
  6. Éviter les questions reposant sur la mémoire
  7. Utiliser des questions filtres

Conception du formulaire

- 4 types d'objectif pour les questions :
  - questions d'introduction
  - **questions destinées à collecter de l'information**
  - questions de vérification
  - questions destinées à masquer l'objectif de l'étude

Conception du formulaire

- 2 grands types de structure pour les questions :
  - questions ouvertes (ou non structurées)

Pourquoi utilisez-vous principalement vous internet ?

  - questions fermées (ou structurées) – à privilégier pour les études descriptives

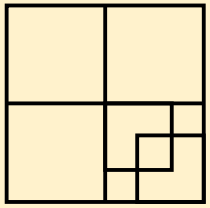
Choix multiples / réponse unique

Quel est votre sexe ?

- ○ Masculin
- ○ Féminin

Pas de choix imposé mais réponse unique (numérique)

Combien de carrés voyez-vous ?



Conception du formulaire

- **questions fermées (ou structurées)** – à privilégier pour les études descriptives

Choix multiples / réponses multiples

Quels sports pratiquez vous ?

- ☐ Judo
- ☐ Tennis
- ☐ Football
- ☐ Ski
- ☐ Yoga
- ☐ Danse
- ☐ Pétanque
- ☐ Sieste (en compétition)

Choix multiples / réponse unique ordonnée
(ici échelle de Likert)

Ce cours est génial :

- ○ Pas du tout d'accord
- ○ Pas d'accord
- ○ Ni en désaccord ni d'accord
- ○ D'accord
- ○ Tout à fait d'accord

Conception du formulaire

- **Toujours faire un pré-test du questionnaire (sur un petit échantillon représentatif) afin de détecter :**
  - **des questions mal rédigées**
  - **des modalités de réponses inadaptées ou ambiguës**
  - **des problèmes de questions filtres**
  - **des réponses incohérentes entre elles**
  - **…**

- **Faire également un traitement statistique du pré-test :**
  - **détection des erreurs de saisies potentielles**
  - **détection variables trop souvent manquantes/ incomplètes**
  - **détection variables ayant une trop faible variance**
  - **vérification de la faisabilité des analyses envisagées**
  - **…**

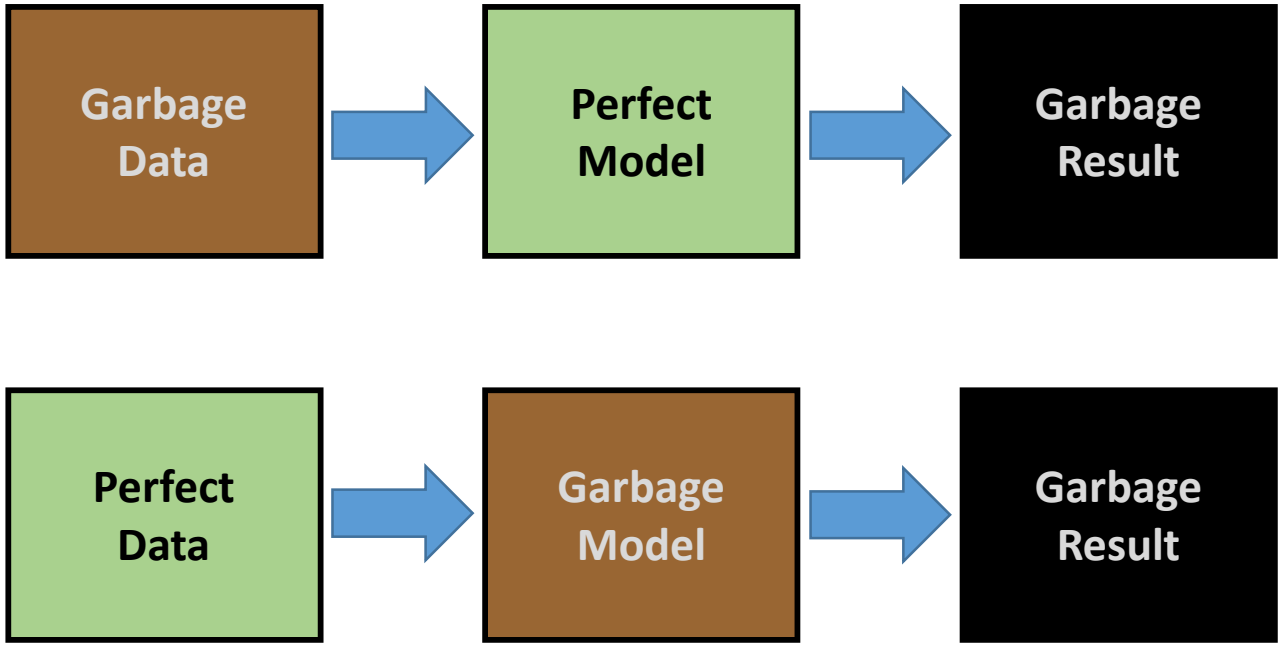
Constitution de la base de données

Base de données et statistiques

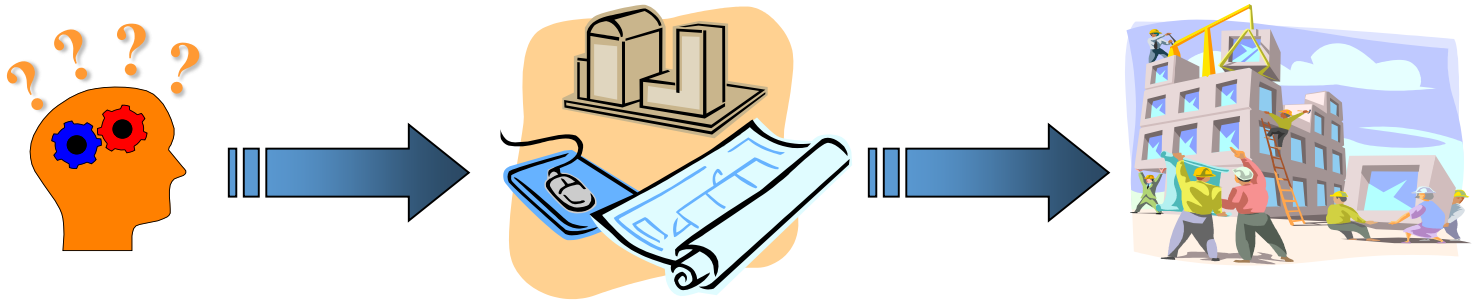
- Constitution et gestion de la base doivent faire partie du processus de l'étude

- Des données médiocres et/ou médiocrement gérées :
  - Compromettent les analyses
  - Conclusions erronées
  - Gaspillage (temps, argent, travail, …)

Paradigme « Garbage In – Garbage Out »

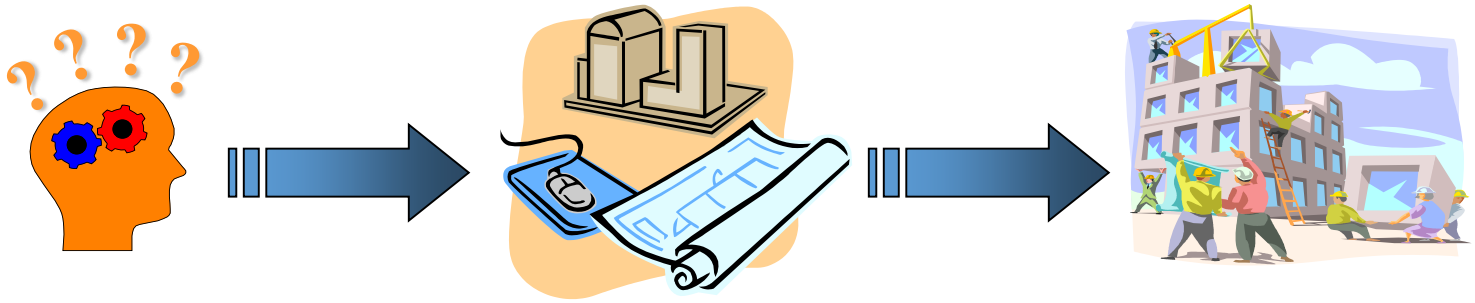


Avant de se lancer…se questionner !



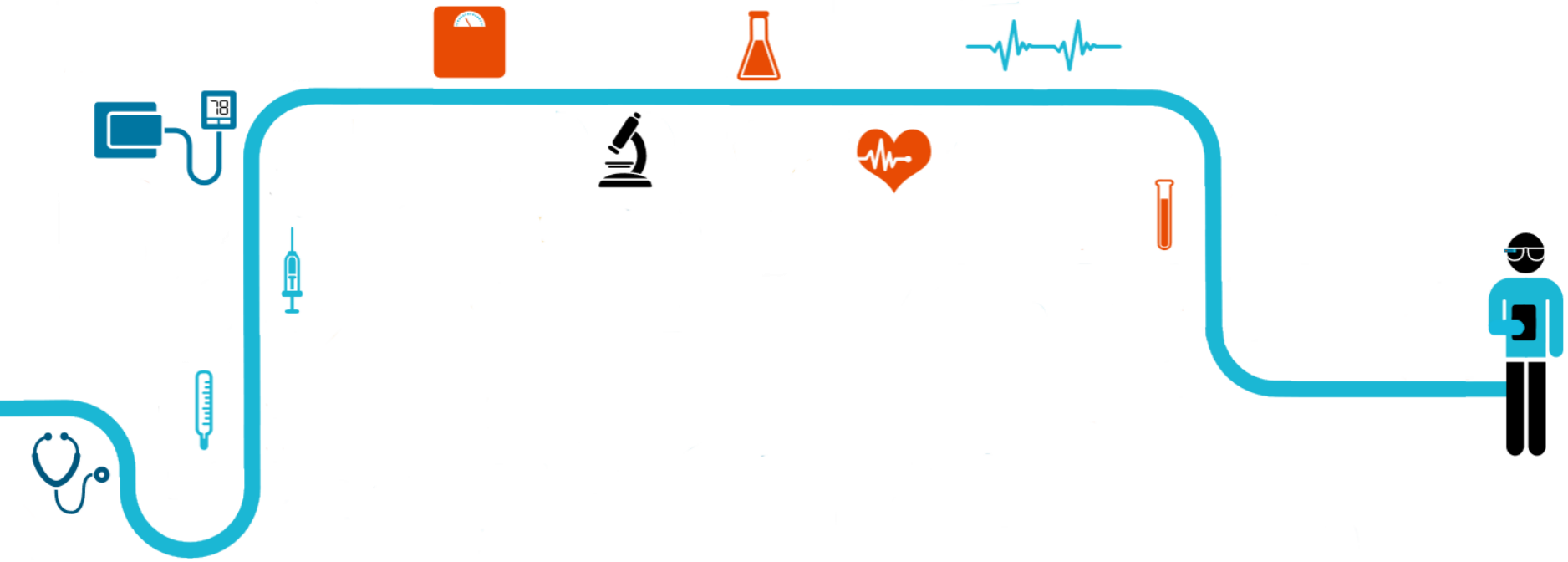
- Quelles sont les **finalités** ?
  - Objectif **principal**
  - Objectifs **secondaires**
- Quels sont les **critères de jugement** et comment les **mesurer** ?

Avant de se lancer…se questionner !



- Quelles sont les « **types objets** » et quels sont les **éléments d'informations** qui les caractérisent ?
- Parmi ces éléments quels sont ceux **à prendre en compte** (et ceux à oublier) ?

Résister à la tentation de tout collecter !



Mieux vaut des variables bien choisies, préalablement identifiées par rapport à votre objectif et dont vous pourrez contrôler la qualité !

Base de données : principes essentiels

- Informations organisées dans 1 ou plusieurs **tables**
- Table est composée de **lignes** et de **colonnes**
  - Ligne = **enregistrement**, observation
  - Colonne = **variable**, attribut, champ, élément d'information

Table « Patient »

Nom	Prenom	DateNaiss	Sexe
Rogne	Olive	18/05/1922	M
Dézieux	Jacqueline	26/02/1942	F
Greux	Nadine	06/09/1967	F

Valeur de la variable « Sexe » pour le 1er enregistrement = **une donnée**

1 Enregistrement = 1 Ligne

Base de données : principes essentiels

- Chaque enregistrement doit avoir un **identifiant unique** (numéro, code alphanumérique, …) :
  - Permet de désigner facilement et de manière certaine un et un seul enregistrement
  - Permet de faciliter l'anonymisation

IDPatient	Nom	Prénom	DateNaiss	Sexe
P1	Rogne	Olive	18/05/1922	M
P2	Dezieux	Jacqueline	26/02/1942	F
P3	Greux	Nadine	06/09/1967	F

Base de données : principes essentiels

- En théorie, autant de tables que de « types d'objets »

Table « Patient »

IDPatient	Nom	Prenom	DateNaiss	Sexe
P1	Rogne	Olive	18/05/1922	M
P2	Dézieux	Jacqueline	26/02/1942	F
P3	Greux	Nadine	06/09/1967	F

Table « Bilan »

IDBilan	DateBilan	HemoGb	Creat	TP
B1	18/07/2014	12,8	194	60
B2	12/08/2014	12,5	106	58
B3	17/09/2014	11,9	83	65
B4	22/10/2014	12,3	98	62

Base de données : principes essentiels

- On utilise les identifiants lorsqu'il faut lier (faire références à) des enregistrements

Table « Patient »

IDPatient	Nom	Prenom	DateNaiss	Sexe
P1	Rogne	Olive	18/05/1922	M
P2	Dézieux	Jacqueline	26/02/1942	F
P3	Greux	Nadine	06/09/1967	F

Table « Bilan »

IDPatient	IDBilan	DateBilan	HemoGb	Creat	TP
P3	B1	18/07/2014	12,8	194	60
P3	B2	12/08/2014	12,5	106	58
P3	B3	17/09/2014	11,9	83	65
P1	B4	22/10/2014	12,3	98	62

Nommer correctement les variables

- Pour ne pas vivre un cauchemar lors de l'analyse

  - Donnez un nom simple, court et descriptif/explicite aux variables ~~VAR1, VAR2, VAR3,…~~
  - Pas de variables portant le même nom
  - Evitez les caractères spéciaux (&, $, %, *,…) et les espaces

Valeurs des variables

- **Une variable doit avoir une valeur « atomique » (non décomposable)**
  - **Une variable ne doit pas contenir une liste/suite de valeurs**

IDPatient	Nom	Prenom	Traitement
P1	Rogne	Olive	**3 Doliprane 1000 mg, 3 Bactrim, 3 Spasfon Lyoc 160 mg**
P2	Dézieux	Jacqueline	**1 Valium 10 mg**
P3	Greux	Nadine	**3 Spasfon Lyoc 160 mg, 3 Clamoxyl 500 mg**

Valeurs des variables

- **Uniformisez les valeurs et évitez les ambiguïtés**

IDPatient	Nom	Prenom	DiagPrincipal
P1	Rogne	Olive	**Infarctus mésentérique**
P2	Dézieux	Jacqueline	**HTA**
P3	Greux	Nadine	**Hypertension**
P4	Patamob	Adhemar	**Hypertension artérielle**
P5	Cive	Jean	**Infarctus**
P6	Zepoher	Agathe	**Hypertension art.**

Valeurs des variables

- **Uniformisez les valeurs et évitez les ambiguïtés**

IDPatient	Nom	Prenom	DiagPrincipal
P1	Rogne	Olive	**Infarctus mésentérique**
P2	Dézieux	Jacqueline	**HTA**
P3	Greux	Nadine	**HTA**
P4	Patamob	Adhemar	**HTA**
P5	Cive	Jean	**Infarctus du myocarde**
P6	Zepoher	Agathe	**HTA**

Valeurs des variables

- **Ne réinventez pas la roue !**
  Utilisez autant que possible des référentiels reconnus

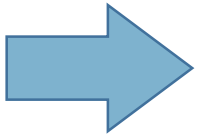
IDPatient	Nom	Prenom	DiagPrincipal
P1	Rogne	Olive	**K55.9**
P2	Dézieux	Jacqueline	**I10**
P3	Greux	Nadine	**I10**
P4	Patamob	Adhémar	**I10**
P5	Cive	Jean	**I21.9**
P6	Zepoher	Agathe	**I10**

Code	Libellé
I10	hypertension essentielle (primitive)
I21.9	infarctus aigu du myocarde
K55.9	infarctus mésentérique aiguë

Valeurs des variables

- Pour les variables qualitatives ordinales :
  il peut être utilise d'opter pour des codes numériques ou alphanumériques
  - Exemple :

TypeDouleur
Douleur forte
Douleur modérée
Douleur absente
Douleur insupportable
Douleur faible
Douleur minime
Douleur intense



TypeDouleur
D4
D3
D0
D6
D2
D1
D5

Valeurs des variables

- Une variables doit toujours avoir des valeurs du même type et au même format
  - Ne mentionnez pas les unités dans la valeur
  - Utilisez toujours la même unité
  - Utilisez toujours le même séparateur décimal

IDBilan	DateBilan	HemoGb
B1	**25-06-2014**	**13,6**
B2	**18 fev 2014**	**12,8 g/dL**
B3	**12/08/2014**	**7,5 mmol/l**
B4	**17 septembre 2014**	**11.9 g/dL**
B5	**22/10/2014**	**9.3 mmol/l**
B6	**2014-03-23**	**NR**

Valeurs des variables

- Attention aux variables contenant le résultat d'un calcul ou d'une agrégation

Exemples :

- « Poids » et « Taille »  « IMC »
- « Date de Naissance » et « Date événement »  « Age »
- « Dose unitaire » et « Nombre de prises »  « Dose cumulée »

Dans tout les cas…

- Documentez votre base
  - Faite un listing des différentes variables avec leurs significations, leurs unités, les codages utilisés,…
- Faite des vérifications régulières
  - Dénombrez et passez en revue les différentes valeurs de vos variables qualitatives pour repérer les codes erronés/mal saisis
  - Tracer des histogrammes des variables quantitatives pour repérer les valeurs aberrantes ou non-numériques

Quels logiciels ?

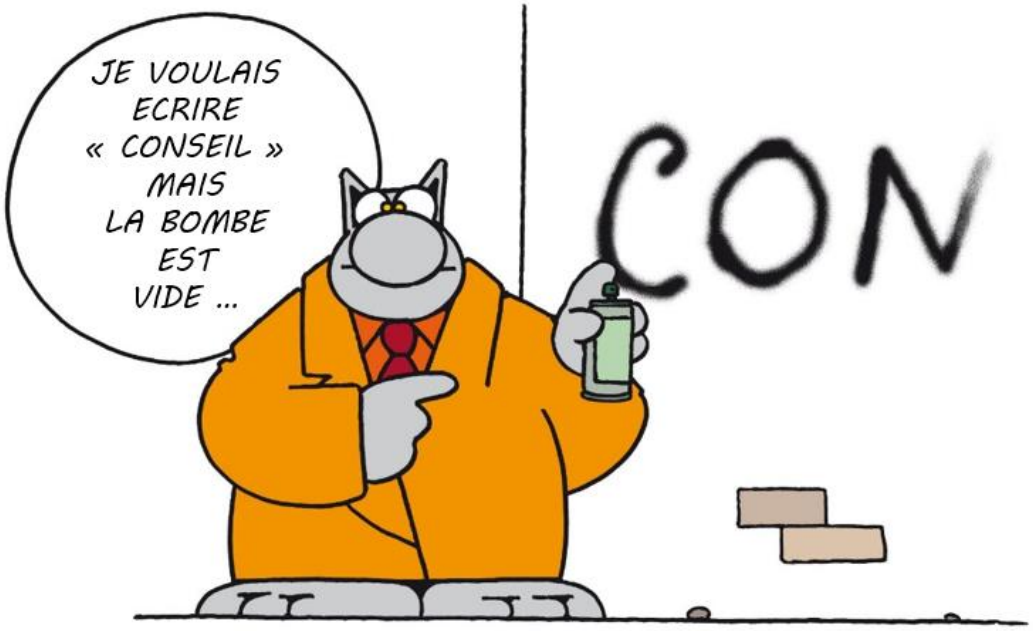
- Tableurs (MS Excel / LibreOffice Calc)
  - Pratique pour l'analyse descriptive des données
  - Pas adapté à la saisie/conservation si beaucoup de données
  - Pas indiqué s'il y a de nombreuses tables

- Outils en ligne (Google Forms, SurveyMonkey, …)
  - Facilité d'utilisation
  - Bien adapté à la création de questionnaires multi-supports
  - Attention à la confidentialité/propriété des données

Quels logiciels ?

- SGBD ‘grand public’ (MS Access / LibreOffice Base)
  - Prise en main plus complexe
  - Plutôt pour les bases « mono-utilisateur » (< 2Go de données)
- Outils ‘pro’
  - Compétences techniques requises

Attention aux imprévus techniques



Attention aux imprévus techniques

Si ce n'est pas sauvegardé
à 3 endroits différents
ce n'est pas sauvegardé !

Exercice

Vous souhaitez mettre en place une étude pour apprécier certains troubles cognitifs de votre patientèle

1. **Proposer des critères d'inclusions / exclusions dans l'étude**
2. **Donnez 3 variables que vous allez observer**
   - **Une quantitative**
   - **Une ordinale**
   - **Une non-ordinale**
3. **Etablissez le questionnaire pour ces 3 variables**
4. **Sous quelle forme allez-vous les enregistrer ?**

Exercice : une correction possible

- Critères d'inclusions :
  - Tout patient vu en consultation entre le 01/01/2014 et le 30/06/2014
  - Age > 70 ans
- Critères d'exclusions :
  - Résidant dans un EHPAD
  - Antécédents psychiatriques connus

Exercice : une correction possible

1 - Nombre de mots retenus (quantitatif discret)

L'examinateur lit cette liste de mots « musée, sauterelle, passoire, camion, limonade », et demande au sujet de les répéter

Combien de mots sont retenus parfaitement ?

☐ (0 à 5)

IDPatient	Sexe	Age	NbMotsRet
P1	M	75	5
P2	F	79	3
P3	F	81	1

Exercice : une correction possible

2 - Oublis de rendez-vous (ordinal)

Oubliez-vous des rendez-vous ou des engagements ?

Jamais	Rarement	Parfois	Souvent	Très souvent
○	○	○	○	○
(0)	(1)	(2)	(3)	(4)

IDPatient	Sexe	Age	NbMotsRet	OublisRDV
P1	M	75	5	0
P2	F	79	3	1
P3	F	81	1	3

Exercice : une correction possible

2 - Médicament pris (non-ordinal)

Médicament pris par le/la patient(e) ?

- ☐ Antidépresseur
- ☐ Anxiolytique
- ☐ Anti-inflammatoire
- ☐ Antihypertenseur
- ☐ Antalgique

Remarque :
Il peut aussi être intéressant de coder Oui = 1 et Non = 0

IDPatient	Sexe	Age	NbMotsRet	OublisRDV	Anti DEP	Anti ANX	Anti INF	Anti HTA	Anti ALG
P1	M	75	5	0	Oui	Non	Oui	Non	Non
P2	F	79	3	1	Non	Non	Non	Non	Non
P3	F	81	1	3	Non	Oui	Non	Oui	Oui

Merci pour votre attention

Dr Jean-Charles DUFOUR

jean-charles.dufour@univ-amu.fr

SESSTIM (Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale) UMR 912

Faculté de Médecine
Aix-Marseille Université

Principales sources d'inspiration

- Supports de cours de Vincent Jalby (http://www.unilim.fr/pages_perso/vincent.jalby/m1aes/cours.html)
- Illustrations diapos 4 et 9 issues du projet «the face of tomorrow » (Sydney) via http://rue89.nouvelobs.com