Projet Obe Maghreb

Ecole thématique gestion et analyse de données
20 au 29 avril 2010

Qualité des données dans les enquêtes épidémiologiques

Pierre Traissac
UMR 204 « Prévention des malnutritions et pathologies associées »
IRD, Montpellier, France

Objet

- Assurance qualité des données pour une enquête épidémiologique
- Concept « Erreur d'enquête totale »
- Approche non complètement formalisée / normalisée (e.g. HACCP)
- Présentation : principes généraux
- Détails techniques : documents de référence / mise en pratique

Erreur totale d'une enquête (« total survey error »)

Question
(e.g. de nutrition)

Monde réel
(e.g. adultes Tunisiens)

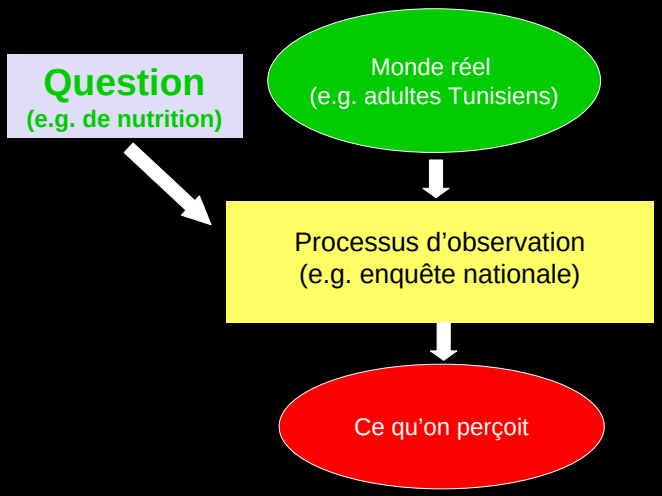
Erreur totale d'une enquête (« total survey error »)

Question
(e.g. de nutrition)

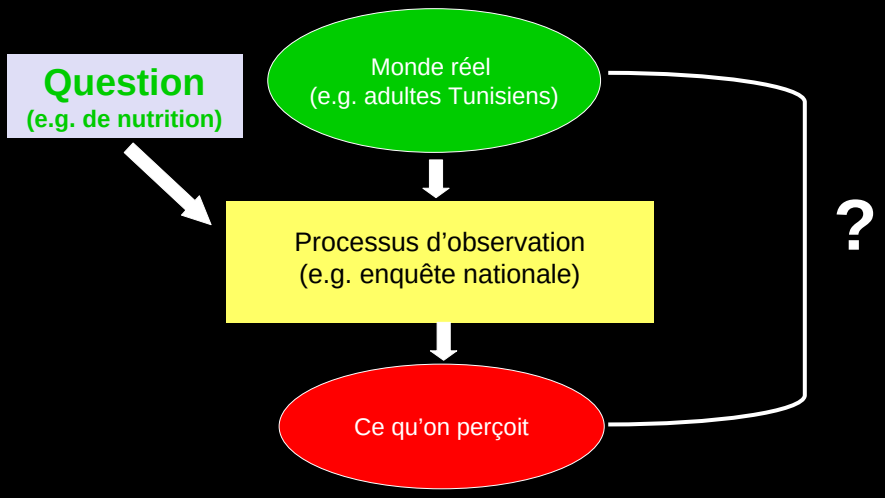
Monde réel
(e.g. adultes Tunisiens)

Processus d'observation
(e.g. enquête nationale)

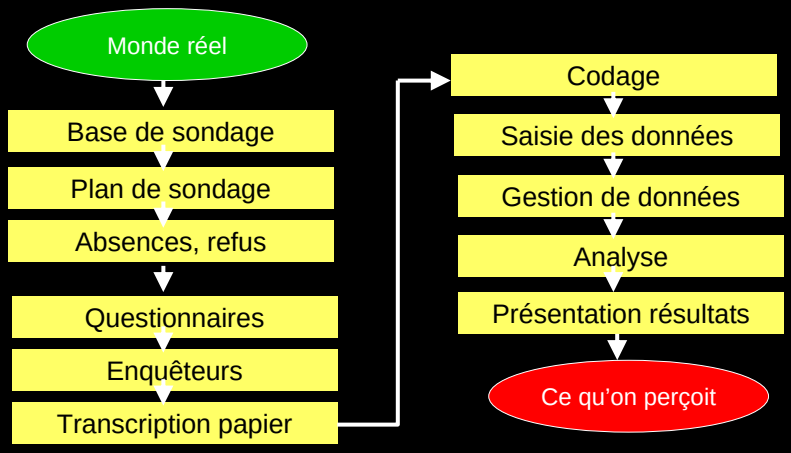
Erreur totale d'une enquête (« total survey error »)



Erreur totale d'une enquête (« total survey error »)



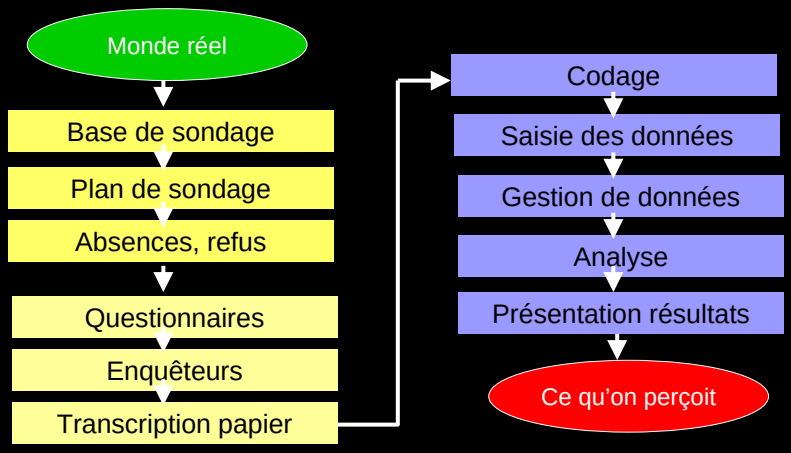
Erreur totale d'une enquête (« total survey error »)



Erreur totale d'une enquête (« total survey error »)

- Erreurs lors de la collecte de l'information sur le terrain (« response errors »)
- Erreurs dans les traitements ultérieurs (« processing errors »)

Erreur totale d'une enquête (« total survey error »)



Erreurs lors de la collecte de l'information (« response errors »)

- **Erreurs liées à la sélection des sujets**
  - population cible v.s. population source
  - variance d'échantillonnage (vs. recensement)
  - plan de sondage théorique (base de sondage et plan de sondage)
  - mise en œuvre terrain (sélection des individus, absences, refus)
- **Erreurs de mesure**
  - qualités intrinsèques des instruments de mesure et questionnaires (précision, validité, reproductibilité)
  - mise en œuvre pratique (administration du questionnaire, prise des mesures, transcription)

Erreurs dans la chaîne de traitement (« processing errors »)

- Transcription des mesures/observations/réponses
- Codage (le cas échéant)
- Saisie (lecture, frappe), validation, apurement
- Gestion des données
  - Sélections, fusion de tableaux de données
  - Calcul de nouvelles variables, indices, recodages
- Analyses (choix des méthodes, mise en oeuvre)
- Préparation de tableaux / graphiques
- Documents écrits / supports visuels

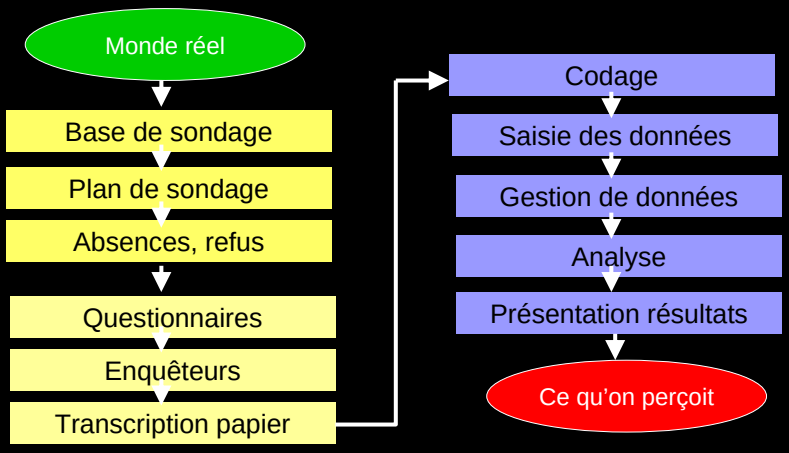
Erreurs dans la chaîne de traitement (« processing errors »)

- Transcription des mesures/observations/réponses
- Codage (le cas échéant)
- Saisie (lecture, frappe), validation, apurement
- Gestion des données
  - Sélections, fusion de tableaux de données
  - Calcul de nouvelles variables, indices, recodages
- Analyses (choix des méthodes, mise en oeuvre)
- Préparation de tableaux / graphiques
- Documents écrits / supports visuels

Principe de base n°1

« Une chaîne est aussi solide que le plus faible de ses maillons»

Erreur totale d'une enquête (« total survey error »)



Principe de base n°1

« Une chaîne est aussi solide que le plus faible de ses maillons»

- ne pas négliger certaines étapes a priori moins « nobles » : aspects « techniques » (e.g. mesures, codage, saisie, recodages) vs. « conceptuels » (e.g. échantillonnage, questionnaire, analyse)
- planification, organisation, écriture de guides opératoires, formation des opérateurs, tests pour TOUTES les étapes : sélection des ménages, administration du questionnaire, mesures, codage, saisie, apuration, gestion de données, analyse

Principe de base n°2

« Confiance n'exclut pas contrôle»

- De son propre travail, de celui des autres
- Procédures de contrôle, détection et correction d'erreurs
- Toutes les étapes de la chaîne de collecte / traitement
- Aussi pour gestion et analyse de données (cf exercices)

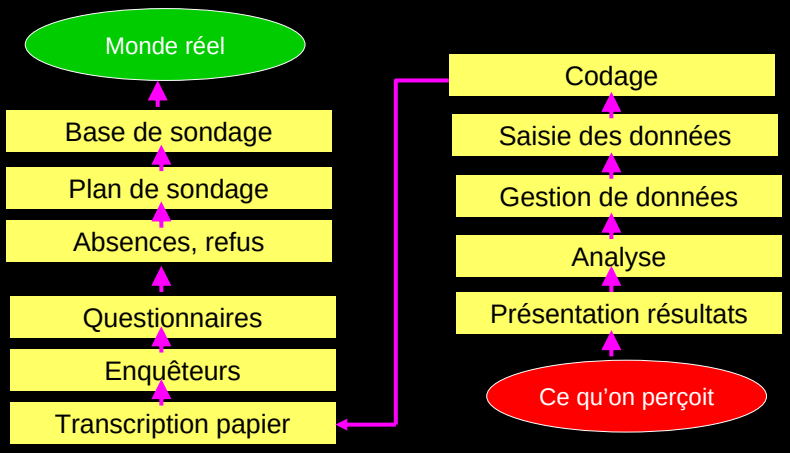
Principe de base n°3

« Traçabilité »

- But : pouvoir à tout moment remonter la chaîne de collecte/traitement jusqu'à la « donnée de base » (pour vérification, détection et correction erreurs)

Principe de base n°3

« Traçabilité »



Principe de base n°3

« Traçabilité »

- But : pouvoir à tout moment remonter la chaîne de collecte/traitement jusqu'à la « donnée de base » (pour vérification, détection et correction erreurs)
- Moyen : documentation écrite de tous les processus, étapes de traitement, difficultés éventuelles de mise en œuvre etc.

Traçabilité : documentation écrite

- Eléments concernant la méthodologie (base de sondage, échantillonnage, mesures, questionnaire)
- Modes opératoires : guides pour enquêteurs (choix des individus, mesures), pour opérateurs de saisie.
- « Journal de bord » des activités de terrain et/ou de laboratoire et de bureau (problèmes rencontrés, modes de résolution)

Traçabilité : documentation écrite

- Description précise des processus d'accès aux données (saisie, vérification, apurement)
- Archivage des questionnaires « papier »
- Description précise des processus de traitement (e.g. QFCA ➡ score de diversité alimentaire DQI, HEI …)
- Dictionnaires de variables
- Logiciels : programmes et non mode interactif

Traçabilité : documentation écrite

- Description précise des processus d'accès aux données (saisie, vérification, apurement)
- Archivage des questionnaires « papier »
- Dictionnaires de variables
- Logiciels : programmes et non mode interactif
- Description précise des processus de traitement (e.g. QFCA ➡ ingéré en divers nutriments)

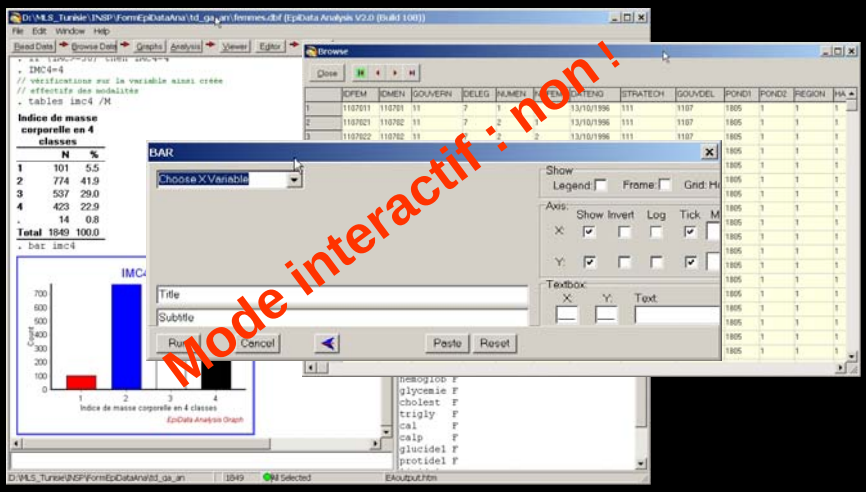
- Dictionnaire de variables

Ordre dans le fichier	nom	Contenu	format	Unités / codes	n	Pgm de création
1	nusaiad	Numéro séquentiel de saisie adulte	Numérique 4.	Sans objet	250	adultes.qes
2	idadu	Identifiant unique d'adulte	Numérique 4.	Sans objet	250	adultes.qes adultes.chk
3	idmen	Identifiant de ménage	Numérique 4.	Sans objet	250	adultes.qes adultes.chk
4	numad	Numéro adulte dans ménage	Numérique 2.	Sans objet	250	adultes.qes
5	datenq	Date d'enquête	Date (dmy 10)	jj/mm/aaaa	250	adultes.qes
6	sexe	Sexe de la personne	Numérique 1.	1: masculin 2: féminin	250	adultes.qes
7	age	Age de la personne	Numérique 3.	années révolues	245	adultes.qes
8	statmat	Statut matrimonial	Numérique 1.	1: célibataire 2: marié 3: …..	240	adultes.qes
9	gross	Grossesse visible	Numérique 1.	1:oui 2:non	145	adultes.qes
10	tailled	Taille debout	Numérique 5.1	cm	238	adultes.qes

Traçabilité : documentation écrite

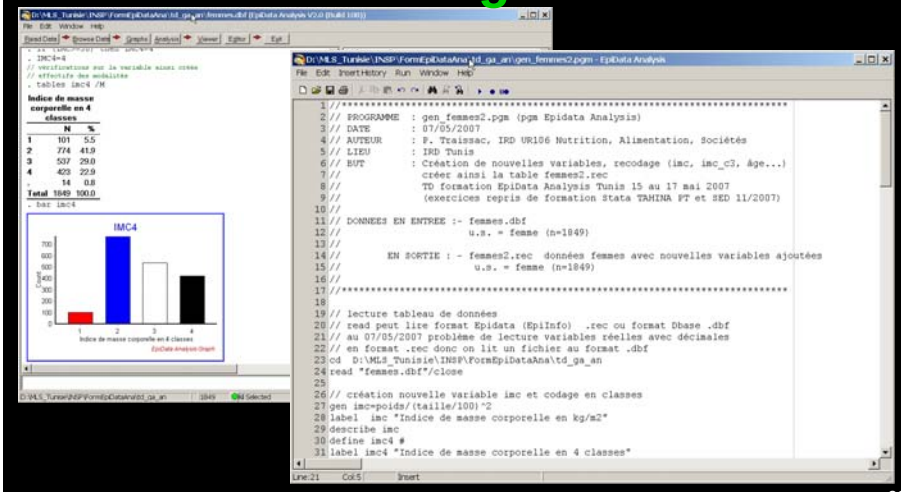
- Description précise des processus d'accès aux données (saisie, vérification, apurement)
- Archivage des questionnaires « papier »
- Dictionnaires de variables
- Logiciels : programmes et non mode interactif
- Description précise des processus de traitement (e.g. QFCA ➡ ingéré en divers nutriments)

Traçabilité : gestion et analyse



Traçabilité : gestion et analyse

Programmes : oui !



```
//*****************************************************************************
// PROGRAMME  : gen_femmes2.pgm (pgm Epidata Analysis)
// DATE       : 07/05/2007
// AUTEUR     : P. Traissac, IRD UR106 Nutrition, Alimentation, Sociétés
// LIEU       : IRD Tunis
// BUT        : Création de nouvelles variables, recodage (imc, imc_c3, âge...)
//              créer ainsi la table femmes2.rec
//              TD formation EpiData Analysis Tunis 15 au 17 mai 2007
//              (exercices repris de formation Stata TAHINA PT et SED 11/2007)
//
// DONNEES EN ENTREE :- femmes.dbf
//                      u.s. = femme (n=1849)
//
//        EN SORTIE : - femmes2.rec  données femmes avec nouvelles variables ajoutées
//                      u.s. = femme (n=1849)
//
//*****************************************************************************

// lecture tableau de données
// read peut lire format Epidata (EpiInfo)  .rec ou format Dbase .dbf
// au 07/05/2007 problème de lecture variables réelles avec décimales
// en format .rec donc on lit un fichier au format .dbf
cd  D:\MLS_Tunisie\INSP\FormEpiDataAna\td_ga_an
read "femmes.dbf"/close

// création nouvelle variable imc et codage en classes
gen imc=poids/(taille/100)^2
label  imc "Indice de masse corporelle en kg/m2"
describe imc
define imc4 #
label imc4 "Indice de masse corporelle en 4 classes"
```

Traçabilité : gestion et analyse

- **Entête des programmes (e.g. Epidata Analysis)**

```
//*************************************************************************************************
// Nom du programme : gen_femmes2.pgm
// Type de programme : Epidata Analysis
// Auteur : PT, IRD, UR 106 Nutrition, Alimentation, Sociétés
// Date : 08/05/2007
// Lieu : IRD, Montpellier – IRD, Tunis
// But :   Calculs et recodages données femmes en préalable à analyse
//           facteurs de risque de l'obésité
//
// Données en entrée :
//        femmes.rec (u.s. : femme adulte, n=1849)
//
// Données en sortie :  femmes2.rec (n=1763)
//
// Remarques diverses : pgm repris de exercices Stata SED et PT
//*************************************************************************************************
À partir d'ici les instructions de programmation proprement dites
```


Traçabilité : gestion et analyse

- **Commentaires dans les programmes**

```
// tableau d'entrée issu de la double saisie
read "femmes.rec "   /close
//          création nouvelle variable imc
gen imc=poids/(taille/100)^2
label  imc "Indice de masse corporelle en kg/m2 "
//          imc en 4 classes (bornes OMS : maigreur, surpoids, obésité)
//          cf rapport « Physical status : use and interpretation of anthropometry »
//           OMS 1995 (page 357)
define imc4 #
label imc4 "Indice de masse corporelle en 4 classes"
//           à ce stade elle n'a que des valeurs manquantes
recode imc to imc4 lo-18.4999=1 18.50-24.9999=2 25.0-29.9999=3 30-hi=4
//          on peut assigner des labels aux codes ainsi créés
labelvalue imc4 /1="maigre" /2="normal" /3="surpoids" /4="obèse"
//          codage obesité (>=30 kg/m2)
define obesite #
recode imc to obesite lo-29.999=0 30-hi=1
//           sauvegarde nouvelles variables dans table femmes2
savedata "femmes2.rec"   /replace
```

Traçabilité : documentation écrite

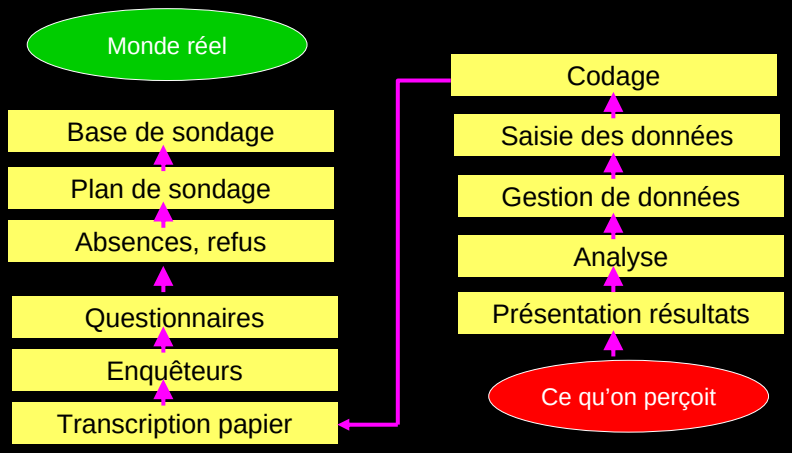
- Description précise des processus d'accès aux données (saisie, vérification, apurement)
- Archivage des questionnaires « papier »
- Dictionnaires de variables
- Logiciels : programmes et non mode interactif
- Description précise des processus de traitement (e.g. QFCA ➡ ingéré en divers nutriments)

Traçabilité : gestion et analyse

- **Documenter les opérations sur les fichiers de données**

Données en entrée	Programme	En sortie	Commentaires
menages.rec n=1763, p= 39	gen_menages2.pgm (pgm Epidata Ana.)	menages2.rec n=1763, p= 45	Calcul score biens Terciles par milieu …..
femmes.rec n=1849, p=36 menages2.rec n=1763, p= 39	gen_femmes2.pgm (pgm Epidata Ana.)	femmes2.rec n=1849, p=43 fem2men.rec n=1849, p=90	Calcul imc, âge, recodages …. Fusion données ménages
fem2men.rec n=1849, p=90	ana1_fem.pgm (pgm Epidata Ana.)	ana1_fem.log (fichier résultats)	Analyse bivariée facteurs de risque obésité

Principe de base n°3
« Traçabilité »



Un mot sur la sécurité des données

- Tous types :
  - protocoles, questionnaires, fichiers de données, programmes, résultats d'analyses,documents, rapports, courriels, images, …
- Protection accès non autorisés (serrures , mots de passe, cryptage)
- Protections anti-virus (internet +++)
- Protection contre pertes accidentelles de données :
  - sauvegardes régulières (CD-ROM, DVD, bandes, disque amovible)
  - doublage et mise en sécurité des supports de sauvergarde
- Archivage questionnaires papier (forme papier ou informatique)
- Motiver les personnes

Conclusion

- Erreur d'enquête totale
- Une chaîne est aussi solide que le plus faible de ses maillons
- Confiance n'exclut pas contrôle
- **Traçabilité (documentation ++)**
- Formation, motivation des personnes
- Sécurité des données

Quelques références

- Adelf-Aeema-Aderest-Epiter (2003). *Recommendations. Déontologie et bonnes pratiques en épidémiologie*. ADELF, AEEMA, ADEREST, EPITER**:** 31.
- Bennett S., Myatt M. et al. (2001). *Data management for surveys and trials. A practical primer using EpiData*. The EpiData Association.
- Biemer P., Lyberg L. (2003). *Introduction to survey quality*. Wiley & Sons.
- Juul S. (2001). *Take good care of your data*. Aarhus, Department of Epidemiology and Social Medicine, University of Aarhus**:** 56.
- Traissac P. (2003). *Quelques éléments de bonnes pratiques pour la saisie, gestion et analyse des données. Projet TAHINA*. INCO Med – ICA3-CT-2002-10011
- United Nations Statistics Division (2003). *Household Surveys in Developing and Transition Countries: design, implementation and analysis*. http://unstats.un.org/unsd/HHsurveys/index.htm