Les données d'une enquête épidémiologique

Loïc Desquilbet

Département des **S**ciences **B**iologiques et **P**harmaceutiques

Ecole **N**ationale **V**étérinaire d'**A**lfort

– Master SEMHA –
23 septembre 2021

Plan

I. La gestion des données

II. Le recueil des données

III. La saisie des données

IV. Une fois les données saisies

I. La gestion des données

Remarques introductives

- La qualité de l'interprétation des résultats issues des données de l'enquête dépend de…

  - La qualité du protocole d'enquête (mode de recrutement, collecte de l'information, …)

  - L'exécution du protocole sur le terrain tel qu'il a été écrit

  - Analyses statistiques correctes et adaptées

  - Interprétation correctes des résultats (pas de sur-interprétation, discussion de la présence de biais, discussion de l'impact des biais s'ils sont présents)

Qualité du protocole

Une bonne qualité du protocole nécessite…

- Une analyse bibliographique poussée
- De pouvoir répondre à
  - « Qu'est-ce qui est connu ? »
  - « Que reste-t-il à montrer ? »
- Une définition de l'objectif principal répondant à une partie de « que reste-t-il à montrer ? »
- Définition de la population cible

Qualité du protocole

Une bonne qualité du protocole nécessite…

- De définir des critères explicites
  - D'inclusion
  - D'exclusion
- De décrire en détail les différents modes de recueil des données (prélèvements, fiches cliniques, questionnaires, base hospitalière, mesures de performances, …)

Structure d'un fichier de données

- 1 colonne = 1 variable
- Parfois, 1ère ligne = nom de la variable
- La ligne = l'individu
- Nom des variables
  - Nombre limité de caractères
  - Pas d'accent
  - Pas d'espace
- Valeurs numériques / alphanumériques dans les colonnes
- Données manquantes : dépend du logiciel et du type numérique / alphanumérique de la variable

  Variable numérique : case vide (Excel), « . » (SAS), …

Type de variables

- Variables numériques
  - 2 classes : variable binaire (à coder en 0/1 si ce n'est pas déjà fait – cf. cours sur les modèles)
  - 3 classes ou plus : variable qualitative
    - Nominale : la valeur des chiffres n'a pas de sens « plus/moins grand »
    - Ordinale : la valeur des chiffres a un sens « plus/moins grand »
  - Quantitative : continue (chiffre avec virgule possible) ou discrète (dénombrement par exemple)

Type de variables

- Variables de dates

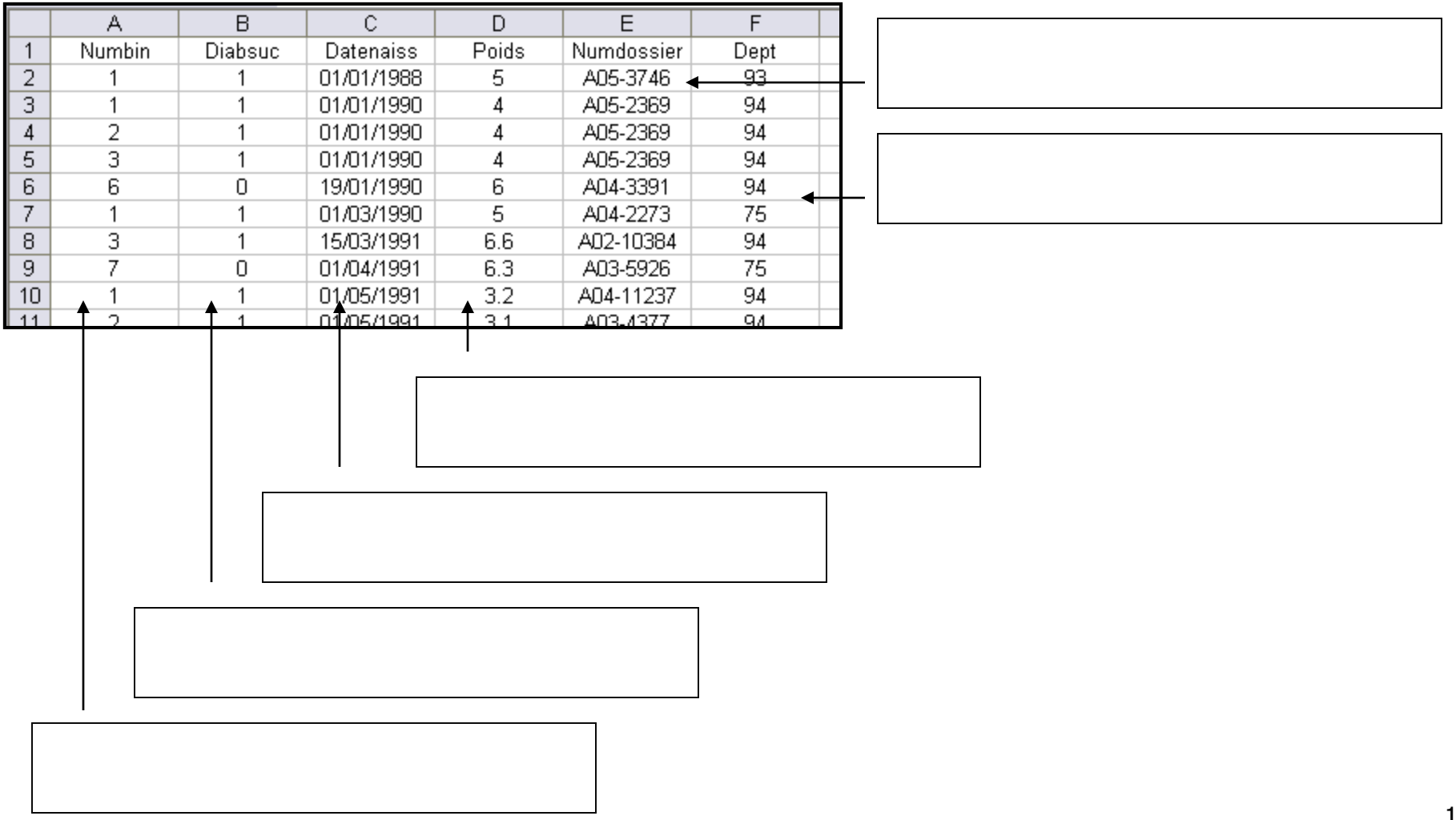
  Dans (quasiment) tous les logiciels de traitement de données, les dates sont un nombre de jours passés depuis le 01/01/1900 (Excel, …)

- Variables alphanumériques (par exemple, numéro d'identification de l'individu)

  A limiter car traitement statistique par la suite très fortement compromis !

Type de variables

Illustration



	A	B	C	D	E	F
1	Numbin	Diabsuc	Datenaiss	Poids	Numdossier	Dept
2	1	1	01/01/1988	5	A05-3746	93
3	1	1	01/01/1990	4	A05-2369	94
4	2	1	01/01/1990	4	A05-2369	94
5	3	1	01/01/1990	4	A05-2369	94
6	6	0	19/01/1990	6	A04-3391	94
7	1	1	01/03/1990	5	A04-2273	75
8	3	1	15/03/1991	6.6	A02-10384	94
9	7	0	01/04/1991	6.3	A03-5926	75
10	1	1	01/05/1991	3.2	A04-11237	94
11	2	1	01/05/1991	3.1	A03-4377	94

Type d'un fichier de données

- Fichier Excel
  - 1 fichier de données par feuille de calcul
  - Gérer correctement le nom des feuilles de calculs
  - Préférer 1 fichier Excel par fichier de données (une seule feuille de calcul utilisée par fichier de données)
- Fichier Texte $\Rightarrow$ quel séparateur de colonne ?
  - Si séparateur = tabulation $\Rightarrow$ extension : .txt
  - Si séparateur = « ; » ou « , » $\Rightarrow$ extension : .csv
  - Importation possible dans R

Type d'un fichier de données

- Fichier EpiData

  Exportable sous fichier txt, pour import dans R

- Fichiers SAS (extension : .sas7bdat), Stata (extension : .dta), …

Nom d'un fichier de données

- Doit toujours contenir la date de création du fichier de données
  - A la fin du nom
  - Selon le format AAAAMMJJ (ainsi, quand on trie alphabétiquement, on trie *de fait* chronologiquement !)

Tri alplabétique = tri chronologique	Tri alphabétique ≠ tri chronologique
Data_enq_20120915	Data_enq_03102012
Data_enq_20121003	Data_enq_15092012
Data_enq_20121017	Data_enq_17102012

II. Le recueil des données

Difficultés face aux données à recueillir / recueillies

- Les données sont déjà saisies dans un logiciel qui n'est pas un tableur (exemple, une base hospitalière) et qui ne sont pas exportables
  - Collecter l'information pertinente, localisée dans les méandres de la base de données
  - Créer le masque de saisie, puis saisie des données avec contrôles

- Les données sont déjà saisies dans un tableur (exemple, Excel)
  - Structure des données doit être : un individu / ligne, une variable / colonne
  - Vérification des données saisies nécessaire (format, données manquantes, absence de caractères alphanumériques si variable *a priori* numérique, …)

Difficultés face aux données à recueillir / recueillies

- Les données sont écrites sur une fiche papier (ou cahier d'observation)

  Créer le masque de saisie, puis saisie des données avec contrôles

- Aucune donnée n'est saisie : toutes les manips sont à effectuer
  - Quelles données collecter ???
  - Temps pour collecter les données (≈ ⇔ recruter les individus dans l'étude)
  - Recueil des données sur une fiche papier / questionnaire papier / questionnaire en ligne sur Internet
  - Créer le masque de saisie, puis saisie des données avec contrôles (inutile si questionnaire en ligne, mais attention au format des données !)

Recueil des données par questionnaire

- Tous les points-clé pour concevoir un questionnaire : cf. site sur AlforPro
- Ne pas confondre la question que l'on *se* pose, et la question que l'on va poser aux individus !
- Toujours se mettre à la place de la personne qui va répondre pour éviter les données manquantes parce que l'on a oublié une réponse possible, non proposée dans le questionnaire
- Tester le questionnaire pour évaluer sa longueur

  On ne peut pas *tout* récolter comme infos $\Rightarrow$ se limiter à celles qui permettront de répondre à l'objectif initial

Recueil des données par questionnaire

- Toujours recueillir un numéro d'identifiant unique qui devra être saisi

  (Il faut se fixer un « protocole » d'attribution de ce numéro ID unique)

- Demander des dates, plutôt que des délais (éviter tout calcul mental lors d'un remplissage de questionnaire !!)

- Pas de question ouverte* !

  Uniquement des cases à cocher ou des nombres ou dates à écrire

- Si l'étude consiste à rechercher les facteurs de risque d'une maladie, ne pas d'oublier de collecter les potentiels facteurs de confusion

* Ou alors, à vos risques et périls…

Recueil des données par questionnaire

Questions qui demandent potentiellement plusieurs réponses possibles…

La journée ou le soir quand vous êtes chez vous, votre chien dort :

- ☐ dans un endroit qui lui est réservé (panier, couverture, niche, …)
- ☐ dans un endroit où vous vous reposez aussi (canapé, lit, …)
- ☐ n'importe où

1ère solution pour créer le masque

La journée ou le soir quand vous êtes chez vous, votre chien dort :

Dans un endroit qui lui est réservé (panier, couverture, niche, …)	☐ Oui	☐ Non
Dans un endroit où vous vous reposez aussi (canapé, lit, …)	☐ Oui	☐ Non
N'importe où	☐ Oui	☐ Non

Mais parfois difficile à interpréter ensuite !

2ème solution pour éviter plusieurs réponses

La journée ou le soir quand vous êtes chez vous, votre chien dort principalement :

- ☐ dans un endroit qui lui est réservé (panier, couverture, niche, …)
- ☐ dans un endroit où vous vous reposez aussi (canapé, lit, …)
- ☐ n'importe où

Echelle de Likert

- Echelle de mesure concernant le degré d'accord ou de désaccord vis-à-vis d'une affirmation

- En général 5 choix de réponse
  - Pas du tout d'accord
  - Pas d'accord
  - Ni en désaccord ni d'accord
  - D'accord
  - Tout à fait d'accord

$\Rightarrow$ Variable qualitative ordinale

A propos des variables qualitatives ordinales

Nombre de modalités pair ou impair ?!

- Nombre pair
  - On force l'interviewé à choisir
  - Mais pas de position médiane pour les indécis ou les non-concernés ($\Rightarrow$ prévoir une modalité NSP* ou NC**)
- Nombre impair
  - On offre la possibilité de ne pas trancher et rester d'un avis médian
  - Mais la modalité du milieu peut être une « modalité refuge »

* Ne sait pas

** Non concerné

III. La saisie des données

La saisie des données

- La saisie des données : utiliser un logiciel adapté !
  - EpiData (car contrôles lors de la saisie)
  - Excel n'est pas un logiciel de saisie de données ! **A éviter…**
- Données manquantes sur fiche papier $\Rightarrow$ Faut-il recontacter la personne pour avoir des infos ?...

  Attention à ne pas chercher à collecter l'information manquante à partir de critères liés à la « maladie » que l'on étudie !
- Contrôles de la saisie indispensable
  - Au cours de la saisie (EpiData ; possible avec Excel avec les options de « Validation des données » dans « Données »)
  - Relecture de la saisie : la faire à deux de préférence

IV. Une fois les données saisies

Création de variables

- Contexte

  - On veut parfois créer de nouvelles variables à partir de variables existantes

  - Exemples

    > Création d'un « temps de survie » (cf. cours sur l'analyse de survie) = date de l'événement – date de d'inclusion dans l'étude

    > Recodage d'une variable qualitative au moment du recueil en variable binaire

    Exemple : « Rarement », « parfois », « souvent », « toujours » recodé en « Rarement » versus « Pas rarement »

Création de variables

- Contexte

  - On veut parfois créer de nouvelles variables à partir de variables existantes

  - Exemples

    > Recodage d'une variable quantitative au moment du recueil en variable qualitative ordinale

    Exemple : consommation de cigarettes initialement recueillie en nb cig / j, recodé en classes : 0,]0-10],]10-20], > 20

Création de variables

- Excel avec filtres puis formules

- Codage de la création d'une variable (par ex., dans R) permet la traçabilité de la création de la variable

- Une fois les variables créées :

  Dictionnaire des variables **indispensable** (liste des variables du fichier de données avec leur codage et la signification du codage employé)

Une fois le fichier de données obtenu

- Le mettre en lecture seule (clic droit → Propriété → Lecture seule)

- **Toute modification d'un fichier de données doit être enregistrée sous un fichier avec un nouveau nom**

- « Carnet de laboratoire »

  Carnet dans lequel toutes les modifications du fichier de données initial sont répertoriées (avec leur raison)

  → Objectivité de la justification de la modification des données (≠ basée sur la mémoire seule !)