MISE EN FORME D'UN FICHIER DE DONNEES
Les données doivent être saisies sur un tableur (exemple avec Excel)

1. Format du fichier

- Une colonne=une variable
- Une ligne=un sujet
- Saisir les données sur la même feuille pour pouvoir comparer vos données

2. Identification des sujets

- Une ligne de données par sujet
- Un identifiant unique (par exemple un N° spécifique au sujet)

3. Présentation des données selon leur type

- **Données continues** : un nombre par sujet. Exemple :

N° Sujet	*Taille*
1	178
2	180
3	160

- **Données qualitatives** : Utilisation de codage plutôt que du texte pour les variables qualitatives
  - lister toutes les réponses possibles et coder en 1/2/3… Ex : codage de la sévérité

N° Sujet	*Sévérité de l'évènement indésirable*
1	1
2	3
3	2

Sur une feuille annexe, préciser la correspondance du codage. Ex :
1 : très sévère
2 : sévère
3 : peu sévère

  - Cas particuliers des données binaires : codage en 0/1 : 1 = OUI 0 = NON

- Si vous avez recueilli des données quantitatives approximatives, codez-les aussi. Exemple :

N° Sujet	*Cigarettes/jour*		*Cigarettes/jour (variable codée)*
1	5		1
2	>20	➔	3
3	10 à 15		2

1 : [0-9[
2 : [10-19[
3 : ≥ 20 cig. /jour

- Données qualitatives non exclusives les unes des autres : les transformer en plusieurs variables binaires. Exemple : codage des complications (1=oui/0=non)

N° Sujet	*Complication infectieuse*	*Complication mécanique*	*Autre complication*
1	1	1	0
2	0	0	0
3	1	0	1

Interprétations :

- Le sujet 1 a deux complications (infectieuse et mécanique)
- Le sujet 2 n'a aucune complication
- Le sujet 3 a deux complications (infectieuse et autre complication)

- **Données censurées** : étudier le **délai de survenue d'un événement** (ex : décès). Il faut définir :
  - **date d'origine** correspondant à un événement "commun" à tous les sujets. (ex : date origine = date diagnostic, date début d'un traitement). *N.B. : c'est l'événement qui est commun à tous les sujets et non pas la date (i.e. la date de diagnostic, par exemple, varie d'un sujet à l'autre…)*
  - **date de point**, la date à laquelle on "fait le point". (Ex : date de point = 1er janvier 2000➔ recueil des informations antérieures au 1er janvier 2000 dont le statut des sujets (vivant/décédé au 1er janvier 2000). Toute information postérieure à la date de point ne sera pas prise en compte dans l'analyse (*ex : si un sujet décède le 5 janvier 2000, et que la date de point est le 1er janvier 2000, il sera pris en compte dans l'analyse, avec le statut "vivant" au 1er janvier 2000*).

    Trois colonnes de données sont donc nécessaires :
    - date d'origine (ex : date de diagnostic)
    - date de l'événement, s'il a eu lieu
    - date de dernière nouvelle pour les sujets chez qui l'événement n'a pas eu lieu (on dit alors que ces sujets sont censurés). Pour cette date de dernière nouvelle soit le sujet, à la date de dernière nouvelle est :
      - toujours suivi à la date de point et l'événement n'a pas eu lieu ➔ date de dernière nouvelle = date de point
      - soit le sujet est perdu de vue ➔ date de dernière nouvelle = dernière date à laquelle le sujet a été vu

N.B. (+++) *: lorsqu'on s'intéresse à un événement qui n'est pas le décès, comme la survenue d'une rechute, pour un sujet chez qui la rechute a été observée, la date de l'événement est la date de survenue de la rechute. Pour un tel patient, on ne tiendra nullement compte de ce qui s'est passé après la rechute même si le patient a été revu au cours d'un suivi post rechute.*

Exemple (date de point le 01/01/2000)

N° Sujet	*Date d'origine*	*Date de l'évènement*	*Date de dernière nouvelle*
1	17/02/1985	18/03/1988	
2	01/06/1988		01/05/1990
3	12/09/1986	19/03/1988	01/01/2000

- Sujet 1 : évènement le 18/03/1988
- Sujet 2 : perdu de vue et vivant au 01/05/1990 = date dernière nouvelle
- Sujet 3 : vivant au 01/01/2000 (= date de point) donc censuré

4. Règles générales pour la saisie des données

- 1ère ligne = nom des variables ➔ 2nde ligne= début des données
  - nom simplifié, sans espace, descriptif, unique
- 1ère colonne de gauche = identification des sujets
- Homogénéité des données correspondant à une variable - respect des unités

Exemple d'une situation problématique :

Problème d'identification des sujets : doublon

N° Sujet	*Taille*
1	178
2	1.80
2	180
3	160

Données problématiques :
- non respect de l'unité
- lettre O saisie à la place du zéro!

- Si une donnée est manquante**, laisser la case vide** et n'utiliser aucun caractère particulier tel NC, 9, 0, etc… Exemple :

N° Sujet	*Type de traitement*
1	1
2	
3	2

Donnée manquante

- Donner des dates plutôt que des durées : un recueil de date est plus aisé à réaliser qu'un calcul de durée (fait facilement sur ordinateur). Ex :
  - il est préférable de recueillir la date de naissance et d'inclusion plutôt que l'âge
  - ne pas recueillir le terme de la grossesse à l'accouchement, mais la date d'accouchement et la date présumée du début de grossesse

  Format conseillé : JJ/MM/AAAA (ex : 12/01/1998). Conventions :
  - si le jour est inconnu ➔ mettre le 15 du mois
  - si le mois est inconnu ➔ mettre le mois de juin (06)

- Ne pas saisir de commentaires qui seraient associés à une donnée non sûre, ou à vérifier. Ne pas utiliser de point d'interrogation ni d'astérisque. Ex de donnée inexploitable :

N° Sujet	*Date de diagnostic*
1	08/05/1992
2	10/12/1993
3	19/09/1994 (diagnostic suspecté lors de la visite du 23/02/1994)

Donnée inexploitable car format date non respecté (*les commentaires ne peuvent être exploités dans l'analyse*)

5. Contrôle de la qualité des données

Des contrôles simples peuvent être effectués afin de détecter des données aberrantes dues à des erreurs de recueil, de saisie. Exemple sous Excel, ne pas hésiter à utiliser les fonctions proposées telles que MIN, MAX, NB, NB.SI …