Mise en forme d'une base de données sur tableur (ex : Excel®)

Les exemples détaillés sont dans le fichier excel « exemple_tableur.xlsx »

Organisation générale

- Une ligne = un sujet (un patient ou une venue, une analyse… selon l'étude)
- Une colonne = une variable
- Nom des variables : en 1e ligne vous devez indiquer le nom de vos variables, sans caractère spécial (accent, cédille, ponctuation…) et sans espace : utiliser l'underscore (tiret du 8 _). Ces noms de variables doivent être courts et explicites.
- Ne précisez pas l'unité au sein du nom de la variable (celle-ci doit être mentionnée dans la légende, cf plus bas)
- Ne pas fusionner des cellules

A éviter	Préférez plutôt
Date n°1, Date N°2	date_entree, date_sortie
Sévérité de l'évènement indésirable	severite_ei
Date du diagnostic, sévérité du diagnostic	diagnostic_date, diagnostic_severite

Légende : pour chacune de vos variables, fournir dans un document annexe (ou autre feuille Excel) les éléments suivants :

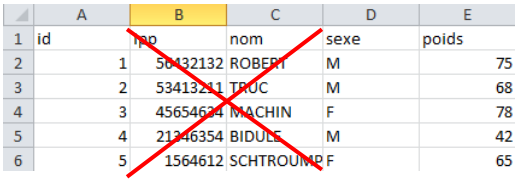
- Nom de la variable
- Description
- Type de variable (numérique, date, heure, date et heure)
- Valeurs possibles.
- Unité (si numérique)
- Préciser si la variable est obligatoire ou non

Identification des sujets

- En première colonne, mettez toujours un identifiant du sujet propre à l'étude Il peut s'agir par exemple d'un nombre allant de 1 à N par exemple Classiquement on nomme cette variable id

	A	B	C	D	E
1	id	age	sexe	poids	taille
2	1	32	M	75	175
3	2	45	M	68	179
4	3	78	F	78	166
5	4	15	M	42	170
6	5	75	F	65	171

- Toute information directement identifiante doit être éliminée de votre base de données
Ex : IPP, IEP, nom, prénom, date de naissance (remplacez par l'âge)



	A	B	C	D	E
1	id	ipp	nom	sexe	poids
2	1	56432132	ROBERT	M	75
3	2	53413211	TRUC	M	68
4	3	45654634	MACHIN	F	78
5	4	21346354	BIDULE	M	42
6	5	1564612	SCHTROUMP	F	65

- Si vous devez absolument conserver ces informations, créez un 2e fichier associant ces données nominatives aux id de vos sujets. ➔ Conservez cette table de correspondance en un unique exemplaire, au sein de votre établissement, sur un poste avec accès sécurisé.

Table de correspondance

	A	B	C
1	id	ipp	nom
2	1	56432132	ROBERT
3	2	53413211	TRUC
4	3	45654634	MACHIN
5	4	21346354	BIDULE
6	5	1564612	SCHTROUMP

Base de données

	A	B	C	D	E
1	id	age	sexe	poids	taille
2	1	32	M	75	175
3	2	45	M	68	179
4	3	78	F	78	166
5	4	15	M	42	170
6	5	75	F	65	171

Valeurs manquantes

Il existe 2 situations où une valeur peut être manquante dans un tableur :

- Vous n'avez pas encore rempli la case en question
- La valeur n'est définitivement pas disponible : c'est une valeur en soit

En laissant une case vide lorsque vous savez que celle-ci ne sera jamais remplie (donnée inexistante, information définitivement perdue, etc), vous prenez le risque au prochain passage de vous demander s'il s'agit d'un oubli ou d'une donnée inexistante.

Pour éviter cela, renseignez NA (not available) ou NC (non connu) lorsque la donnée n'est pas disponible. Faite un choix de convention (NA/NC/autre) et conservez-le tout au long de votre fichier.

Variables quantitatives

Renseignez simplement la valeur numérique dans votre champ. Attention aux valeurs décimales, pensez à utiliser une virgule (ou un point si Excel est en anglais). Si Excel ne reconnait pas une valeur numérique, celle-ci apparait sur la gauche de la cellule.

E
poids
12,4
12.4
78
42
65

Ici 12,4 est écrit avec un point en 2e ligne : Excel ne reconnait pas la valeur comme étant numérique.

Seules des éventuelles valeurs manquantes (NC/NA) devraient apparaitre sur la gauche du champ pour une variable numérique.

Dates et heures

A l'instar des valeurs numériques, une date et une heure correctement reconnues par Excel sont placées à droite du champ. Veillez à ce que cela soit toujours le cas, sans quoi l'analyse statistique sur ces champs sera impossible.

- Pour renseigner une date, le format attendu par Excel (version française) est : JJ/MM/AAAA (ex : 31/12/2020).
- Pour une heure, le format attendu est : HH:MM (ex : 23:59). Les secondes sont optionnelles : HH:MM:SS (ex : 23:59:59).

Dates et heures reconnues (à droite du champ)

11/07/1989
11:37
11:59:59

Dates et heures non reconnues (à gauche du champ)

11 07 1989
11h37
11-59-59

- Vous pouvez également renseigner une date et une heure à la fois en suivant ce format : JJ/MM/AAAA HH:MM:SS (ex : 31/12/2020 23:59:59)

Variables qualitatives

Pour toute variable qualitative, renseignez de préférence un code numérique plutôt que la valeur textuelle. Vous limitez ainsi le risque de fautes de frappe. La correspondance entre le code numérique et la valeur textuelle est à renseigner dans la légende.

Exemple : variable binaire

C
sexe
homme
femme
Femme
Homme
M
F

A éviter
Risque de variation au fil du temps : homme = Homme = M

C
sexe_masculin
1
0
0
1
1
0

A préférer
Ici on fait le choix d'une valeur (homme)
Et on renseigne de façon binaire

Exemple : variable nominale

A éviter

C
service_entree
Urgences
Dermatologie
uro
urgence
Urologie
dermato

Préférez ceci

C
service_entree
1
2
3
1
3
2

Avec une légende
(sur deuxième feuille)

D	E
code	valeur
1	Urgences
2	Dermatologie
3	Urologie

Exception : les variables de texte libre

Si vous devez collecter du texte libre (commentaire, conclusion, etc), il n'est évidemment pas possible de faire l'usage de codes-valeurs. Rédigez simplement le texte dans le champ. Notez cependant que ce type de champ ne peut pas être exploité en analyse statistique sans traitement

complexe (recherche de mots-clés par exemple). N'utilisez ce type de champ qu'en cas de nécessité absolue.

Variables de survie

L'étude de survie avant un événement (maladie, décès, rémission, etc) nécessite :

- Une date d'origine : entrée du sujet dans l'étude (ex : date de diagnostic)
- Une variable binaire précisant si l'événement étudié a eu lieu : oui/non
- Une date de dernières nouvelles :
    - Si l'événement étudié a eu lieu : date de l'événement (car les données a posteriori sont censurées, elles ne nous intéressent pas)
    - Si l'événement n'a pas eu lieu :
        - Le patient est toujours suivi : on renseigne la date de point (date à laquelle le recueil est fait)
        - Le patient est perdu de vue : on renseigne la date de dernières nouvelles

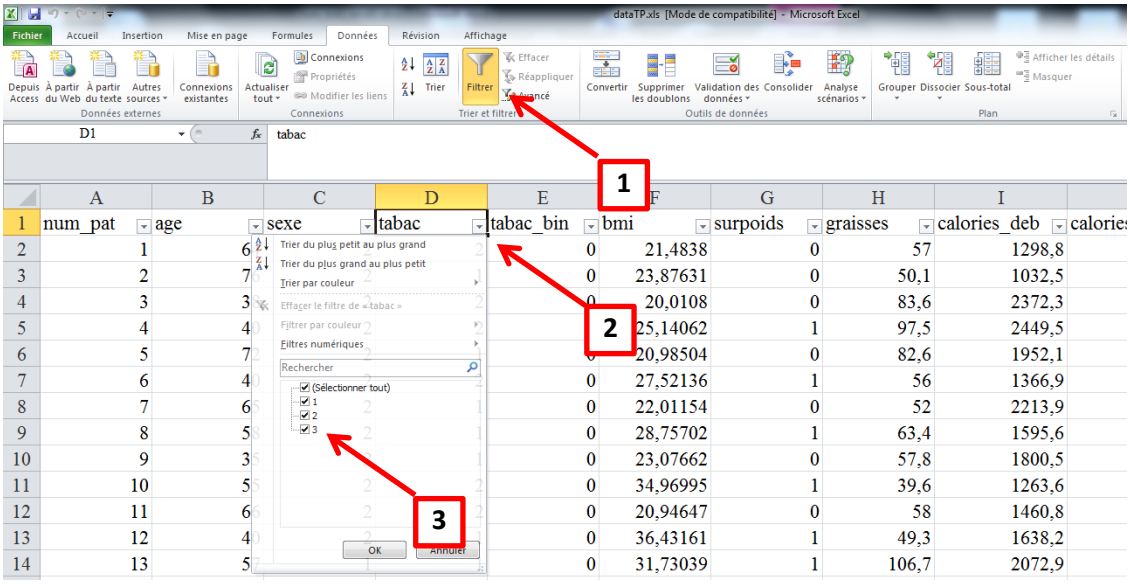
⇨ *Cf fichier excel*

Commentaires

Ne faites pas mention de commentaires dans les différents champs prévus pour vos variables car cela ne permettrait plus d'exploiter vos données (valeur non reconnue). A la place, prévoyez si besoin une dernière colonne « commentaire » dans laquelle vous inscrirez toutes les remarques propres au sujet concerné.

Vérification de la qualité de la saisie

Une fois la base constituée, intégrer des filtres sur votre ligne de noms de variables (1). Ceci vous permettra, en cliquant sur la flèche qui apparaît en bas à droite chaque nom de variable (2), de vérifier qu'il n'y a pas de valeurs aberrantes ou manquantes (3).



	num_pat	age	sexe	tabac	tabac_bin	bmi	surpoids	graisses	calories_deb	calorie
2	1	6			0	21,4838	0	57	1298,8	
3	2	7			0	23,87631	0	50,1	1032,5	
4	3	3			0	20,0108	0	83,6	2372,3	
5	4	4			0	25,14062	1	97,5	2449,5	
6	5	7			0	20,98504	0	82,6	1952,1	
7	6	4			0	27,52136	1	56	1366,9	
8	7	6			0	22,01154	0	52	2213,9	
9	8	5			0	28,75702	1	63,4	1595,6	
10	9	3			0	23,07662	0	57,8	1800,5	
11	10	5			0	34,96995	1	39,6	1263,6	
12	11	6			0	20,94647	0	58	1460,8	
13	12	4			0	36,43161	1	49,3	1638,2	
14	13	5			0	31,73039	1	106,7	2072,9	