Organiser son tableau de données

Philippe MICHEL

7 avril 2022

Un tableur comme Excel® est un outil de comptabilité analytique & n'a jamais été conçu pour stocker des données. Néanmoins ces logiciels sont puissants, versatiles, très pratiques, semblent simples d'emploi & permettent, sous certaines conditions, de stocker des tableaux simples. Un peu de réflexion & de travail préparatoire vous feront gagner beaucoup de temps.

Dans tous les cas on ne travaille jamais seul, le data-manager ou le statisticien [1] peuvent à tout moment vous conseiller.

Pour mémoire le tableau qui arrive au statisticien donc quand tloutes les saisies & queries sont terminées doit être complètement anonymisé.

1. *data-scientist* pour faire chic

Un seul tableau

Vos résultats doivent rentrer dans un seul tableau. Dans certains cas c'est impossible [2]. Dans ce cas c'est un tableau par onglet (avec un nom clair pour l'onglet). Par exemple, un travail sur la prise en charge d'une pathologie dans les services d'un hôpital. Vous n'allez pas ressaisir les caractéristiques du service pour chaque patient. Vous aurez donc un tableau Patients (1) avec une variable *Service* qui permettra de rattacher ce patient aux données du service contenus dans un autre tableau nommé Service (2).

2. À discuter avec le statisticien

ID	service	age	taille	...
P01	Gastro	85	158	...
P02	Neuro	65	187	...
P03	Réa	55	178	...
...	...	...	...	...

Table 1 : tableau Patient

ID	service	nb_lit	dms	...
S01	Gastro	12	5.8	...
S02	Neuro	16	6.3	...
S03	Réa	18	6.7	...
...	...	...	...	...

Table 2 : tableau Service

Un tableau c'est un rectangle plein

Par convention les cas (patients par ex.) sont en ligne & les variables (âge, poids...) en colonne donc un grand rectangle.

Ce tableau ne contient que les données, aucun calcul. Pas de moyenne en bas de colonnes. Si vous voulez faire des calculs, résumés etc. copiez l'onglet & faites tout ce que vous voulez sur la copie.

No Patient	**Age**	**Taille**	**...**
P01	45	148	...
P02	84	178	...
P03	48	**NA**	...
...	...	...	...

Données manquantes

Malheureusement vous aurez des données manquantes. La case ne doit pas être vide pour autant. Il faut noter un code arbitraire pour coder ces données. Habituellement on utilise NA pour *Not Available* mais l'important est de toujours utiliser le même code.

des Variables

Nom des variables

Le nom de la variable (ou de l'onglet) va être utiliser ensuite dans du code. Par exemple pour calculer la moyenne de l'âge :

Il donc évident qu'il doit s'agir d'un nom simple & court, pas la question posé au patient sur le questionnaire. Ce nom ne doit pas comporter d'accent, d'espace ou de caractère bizarre.

Question	Nom inutilisable	Nom correct
Quel est votre âge ?	Âge	Age
Sexe du patient	Sexe patient	Sexe
Pression artérielle systolique	PA systolique	PA_sys

Titres des variables

Chaque variable (donc chaque colonne) doit avoir un titre & un seul. Donc **UNE** & une seule ligne de titre.

Tableau inutilisable

Ce tableau, bien que parfaitement clair pour un être humain, est inutilisable. Quel est le nom de la variable en colonne 2 ?

ID	H1			H2		
	PAM	FC	SpO2	PAM	FC	SpO2
P1	123	98	98	145	111	97
P2	145	88	97	154	121	99
P3	125	78	98	145	98	98

TABLE 3 : tableau inutilisable : deux lignes pour les titres

Première solution – format court

ID	PA_H1	FC_H1	SpO2_H1	PA_H2	FC_H2	SpO2_H2
P1	123	98	98	145	111	97
P2	145	88	97	154	121	99
P3	125	78	98	145	98	98

TABLE 4 : Tableau correct - format court

Solution simple. La comparaison des PA entre H1 & H2 (test de student) va s'écrire :

FIGURE 1 : Test de Student – Données en format court

Deuxième solution – format long

Exactement les mêmes données, seule la présentation change :

ID	heure	PA	FC	SpO2
P1	H1	P1	123	98
P1	H2	145	111	97
P2	H1	145	88	97
P2	H2	154	121	99
P3	H1	125	78	98
P3	H2	145	98	98

TABLE 5 : Tableau correct - format long

Moins instinctif mais souvent plus pratique. Le même test de Student va alors s'écrire :

FIGURE 2 : Test de Student – Données en format long

Du codage des variables

Beaucoup de personnes, souvent beaucoup plus jeunes que moi, sont persuadées qu'un ordinateur ne peut gérer que des chiffres. Grande nouvelle, c'est faux ! Et votre tableau sera plus clair avec des intulés en clair (Homme, Femme) qu'avec des chiffres (1, 0). Même les scores ou échelles doivent être notés en texte. Les coder en numérique revient à dire que passer de 1 à 2 est ausi grave que de passer de 6 à 7. Vous en être sûr ?

Il reste à éviter les erreurs de saisie. On voit souvent dans la même colonne des oui , Oui , OUI & surtout oui avec une espace [3]. On peut facilement éviter ça & gagner du temps sur la saisie en utilisant l'outil Validation de données dans Excell®. Ça vous évitera aussi des âges de 548 ans & la saisie sera beaucoup plus rapide.

3. Oui, on ne voit rien, c'est pour ça que c'est traite.

Des nombres

UNE VARIABLE NUMÉRIQUE ne doit contenir que des nombres ! Et > 5 ou 5 mmol/L ne sont pas des nombres ! L'unité de mesure n' pas à être présente sur le tableau final, ni dans le titre, ni dans les données. Utilisez la validation de données aussi pour vos variables numériques. C'est un peu de travail au début mais un gain de temps & de sécurité ensuite.

Date & Heure

LES DATES & HEURES sont une source d'erreur constante. Il exite au moins trois maière d'écrire les dates d'usage courant (tableau6). La solution est de ne pas laisser Excell® choisir & d'entrer les dates en format texte ou de faire très attention au codage. En particulierdeux colonnes peuvent être codées différement sans que vous l'ayez demandé. De toute façon il faudra convertir les dates pour le logiciel de statistique donc choisissez une norme, n'importe laquelle, mais n'en changez pas en route.

De plus certains systèmes compte le temps en seconde depuis le 01/01/1900, d'autres depuis le 01/01/1904...

TABLE 6 : Cosages usuels pour les dates

France	12/08/2022 12 août 2022 12/08/22
USA	08-12-22
ISO 8601	22-08-12

Comment préparer son tableau

VOUS COMPRENEZ BIEN qu'on improvise pas un tableau de données. La bonne solution, un peu lourde mais si pratique, est d'écrire en premier un tableau de ses variables (*Code Book*) avec toutes les informations sur le modèle :

des Sauvegardes !

UN ORDINATEUR n'est qu'une machine & est sujet à la panne. Un disque dur est une mécanique fragile (qu'il soit mécanique ou SSD). Une tasse de café peut se renverser. Donc on fait des sauvegardes ! Sur plusieurs supports [4] en local & à distance. Il existe des solutions simples de sauvegarde (TimeMachinel® sur MacOSl® ou rsync sur Linux par ex.).

4. une cle USB n'est pas un support de sauvegarde

Table 7 : Tableau des variables

	nom	abrégé	type	valeurs
1	Âge du patient	age	entier	18 –110
2	Date d'admission	dateadm	date	dd/mm/yyyy
3	Sexe	sexe	facteur	F / M
4	BMI	bmi	entier	10 à 60 Kg/m^2
5	Score IGS II à l'admission	igsadm	entier	6 à 150
6	Oxygénothérapie	oxy	entier	0 à 20 L/min
7	Ventilation invasive	vi	facteur	oui/non
8	Sédation	sedation	facteur	oui/non
9	Curarisation	curarisation	facteur	oui/non/nsp
...				

Un fichier non sauvegardé n'existe pas

Remerciements

UNE GRANDE PARTIE de ces conseils est très bien expliquée sur le blog de Claire DELLA-VEDOVA dont je me suis largement inspiré.