Présentation des données pour une analyse statistique

Ce document décrit les points essentiels à vérifier avant d'analyser des données par un logiciel statistique.

Sommaire

Dr Caroline Tournoux-Facon et Alexandre Rollet - 20091

I. Règles à respecter lors de la constitution d'une base de données

A. Repérer l'unité statistique d'analyse

Il s'agit de l'unité élémentaire de l'étude, le plus souvent le patient, repérée par un numéro d'identification unique et possédant un certain nombre d'attributs ou variables la décrivant.

B. Donner un numéro identifiant unique pour chaque patient

Il est saisi dans la première colonne de la table
Il permet d'anonymiser les données, de remonter aux données sources pour vérification
Cet identifiant peut être de la forme « 3 premières lettres du nom » et « 3 premières lettres du prénom » (ex : JEADUP pour JEAN DUPONT)

C. Mettre les données sur une seule page du tableur dans un format rectangulaire

- Chaque ligne correspond à un sujet (= « unité statistique », « observation »), il peut y avoir plusieurs lignes pour un sujet en cas de données répétées (par exemple si le patient a été vu lors de plusieurs consultations).
- Chaque colonne correspond à une seule variable (= « valeur de l'attribut considéré de l'unité statistique ».
- L'intersection de chaque ligne et colonne doit contenir la valeur unique de la variable pour le sujet considéré.

Lors de la constitution de la base de données, il faut raisonner en terme de sujets et de variables et non pas en terme de présentation des résultats. Exemple : si un groupe de sujet a eu le traitement A et l'autre le traitement B, il doit simplement y avoir une variable (colonne nommée « GROUPE ») qui contient A ou B pour chaque sujet.

D. Donner un nom simple aux variables

- La première ligne de la base de données (en-tête de colonne) doit contenir les noms de chaque variable. Il faut essayer d'être assez descriptif sans que le nom soit trop long, éviter les noms comme VAR1, VAR2 …
- Vérifier de ne pas avoir deux fois le même nom de variables, chaque colonne doit avoir un en-tête unique.
- Certains logiciels statistiques étant assez contraignants :
    - Le nom de variable ne doit pas dépasser 8 caractères
    - Il est possible d'utiliser des lettres (plutôt majuscules) et des chiffres mais pas de caractères accentués, de caractères spéciaux (&, $, %, -) ou d'espace (utiliser l'underscore « _ »).

- Le premier caractère doit être alphabétique.

- Faire un **listing** des variables en annexe (sur WORD ou sur la feuille 2 d'EXCEL) avec la signification des variables, leurs unités et le codage des réponses.

E. Coder convenablement les variables qualitatives

- Il faut donner un nom unique à chaque catégorie de la variable qualitative.
- Les codes alphabétiques sont plus informatifs et plus faciles à mémoriser alors que les codes numériques peuvent être plus pratiques et permettre d'imposer un ordre de classement.
- Il sera toujours possible, au moment de l'analyse, de regrouper les catégories.

F. Saisir les variables quantitatives avec soin

Les variables quantitatives ne doivent être que numériques : les cellules de la colonne ne doivent comporter aucun texte, en particulier l'unité de mesure ne doit pas être saisie avec la valeur (dans Excel, la valeur doit s'aligner à droite de la cellule)

- Ne pas saisir des >, < ou ?
- Si la valeur n'est pas connue, il faut laisser la case vide
- Attention aux O (lettre) et 0 (chiffre), I (lettre) et 1 (chiffre)
- Attention au caractère séparateur décimal (soit virgule soit point)
- Attention à la précision : elle doit être toujours la même (même nombre de décimales) pour une même variable
- Etre constant dans le format de saisie d'une date (le transfert d'une date pose souvent problème)

G. Vérifier et revérifier les données avant toute analyse

- Calculer les fréquences des catégories de chaque variable qualitative pour repérer les codes inconnus ou mal saisis.
- Tracer l'histogramme des variables quantitatives pour repérer les données aberrantes ou non-numériques.
- Revérifier les données, les noms des variables. Le temps gagné sur la correction des erreurs et la mise en forme des données sera investi dans une meilleure analyse et explication des résultats

H. Enregistrer le fichier

Le nom du fichier doit comporter le sujet du travail, les initiales de l'auteur ainsi que la date de modification ou le numéro de version

II. Consignes pour l'élaboration du plan d'analyse

- Indiquer dans le listing des variables celles qui ne feront pas l'objet d'analyse

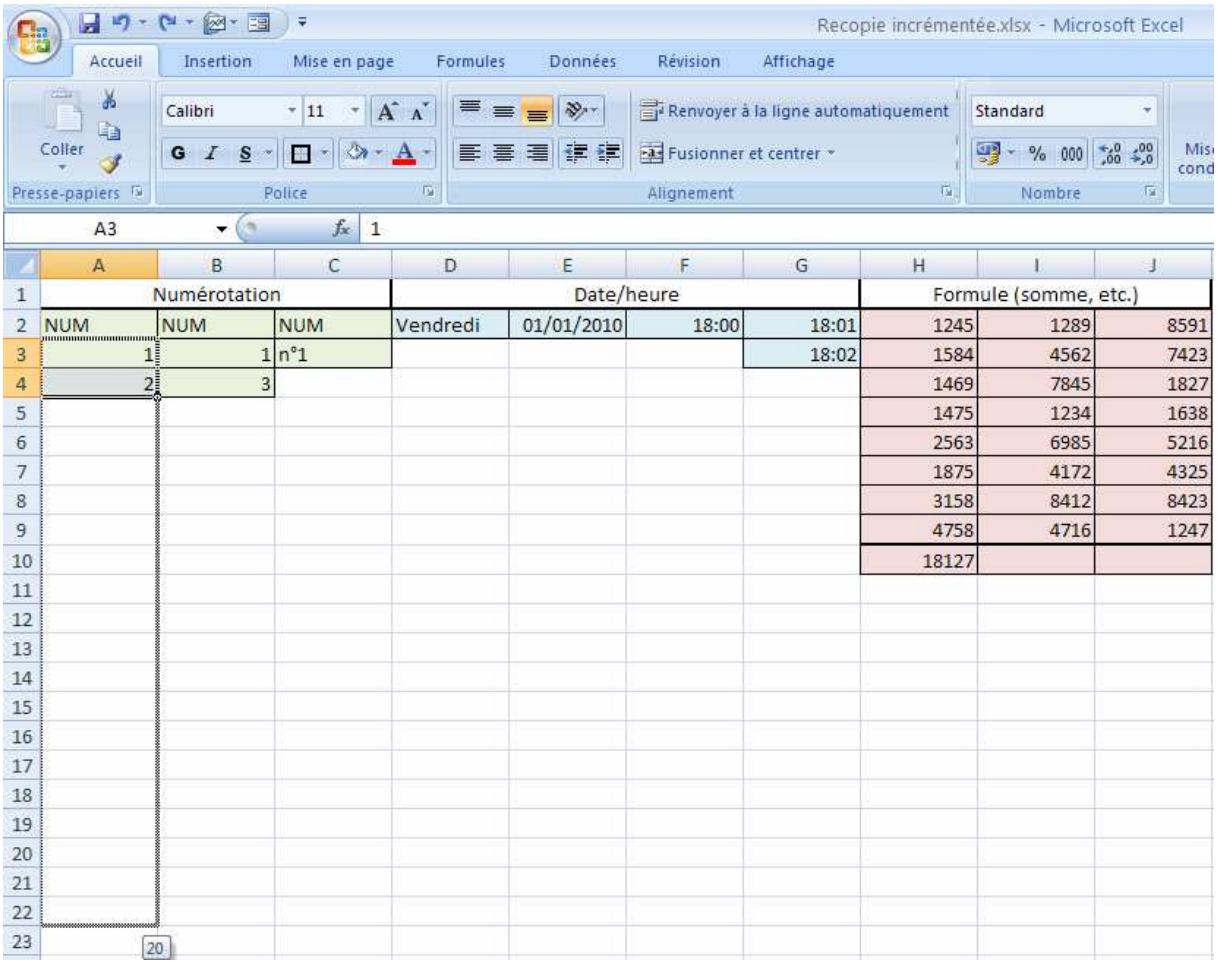
- Formaliser le plan d'analyse pour les analyses comparatives : à quelles questions doit-on répondre? quelles variables doit-on comparer? Cibler les demandes sur des analyses nécessaires, utiles en termes d'interprétation, de compréhension et/ou de comparaison d'après des données cliniques consensuelles et d'après la littérature.

III. Quelques fonctions utiles d'Excel

A. Recopie incrémentée :

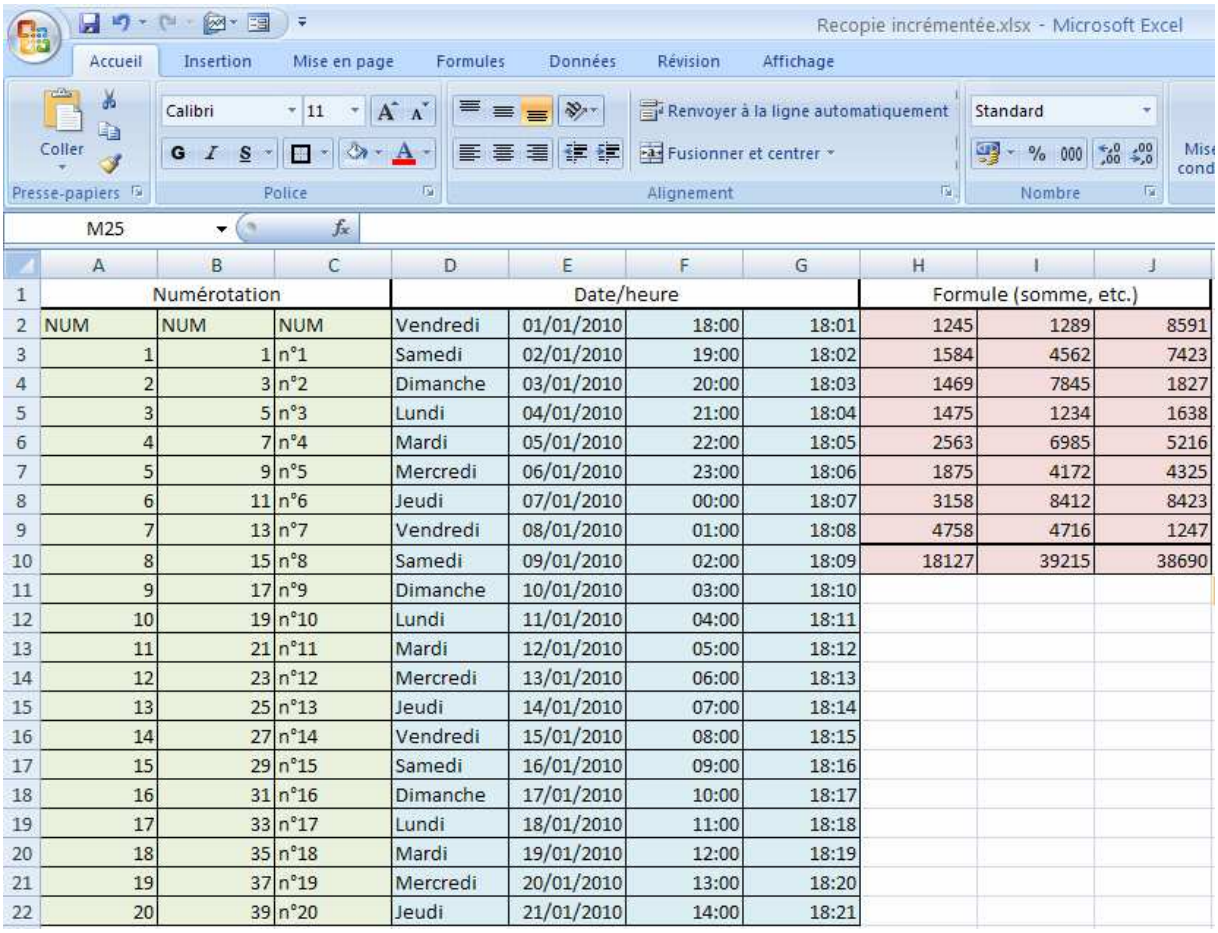
Cette fonction permet un gain de temps considérable, en évitant de répéter x fois les mêmes opérations (numéroter de 1 à 30, appliquer une même formule à plusieurs colonnes…)

- (Exemple recopie incrémentée.xlsx) Il faut sélectionner la ou les cellules souhaitées (si on veut incrémenter un intervalle), placer le curseur dans le coin en bas à droite de la sélection puis cliquer sans relâcher et déplacer jusqu'à la cellule voulue.



	A	B	C	D	E	F	G	H	I	J
1	Numérotation			Date/heure				Formule (somme, etc.)		
2	NUM	NUM	NUM	Vendredi	01/01/2010	18:00	18:01	1245	1289	8591
3	1	1	n°1				18:02	1584	4562	7423
4	2	3						1469	7845	1827
5								1475	1234	1638
6								2563	6985	5216
7								1875	4172	4325
8								3158	8412	8423
9								4758	4716	1247
10								18127		
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										

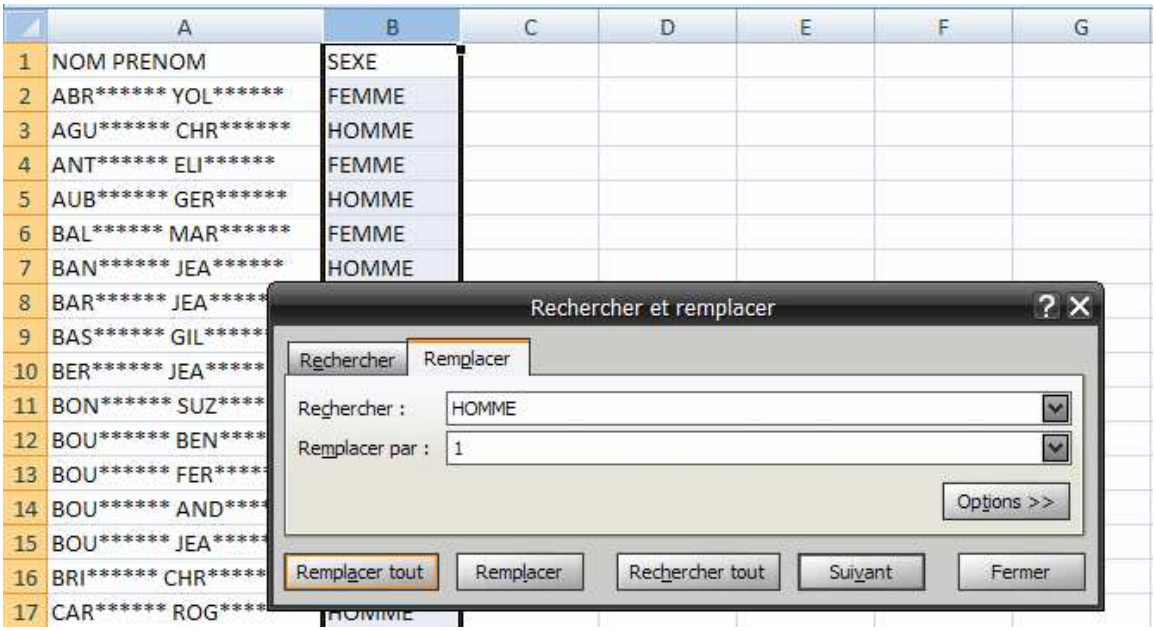
- La recopie incrémentée fonctionne avec différents formats (numérique, date, heure, formule…)



	A	B	C	D	E	F	G	H	I	J
1	Numérotation			Date/heure				Formule (somme, etc.)		
2	NUM	NUM	NUM	Vendredi	01/01/2010	18:00	18:01	1245	1289	8591
3	1	1	n°1	Samedi	02/01/2010	19:00	18:02	1584	4562	7423
4	2	3	n°2	Dimanche	03/01/2010	20:00	18:03	1469	7845	1827
5	3	5	n°3	Lundi	04/01/2010	21:00	18:04	1475	1234	1638
6	4	7	n°4	Mardi	05/01/2010	22:00	18:05	2563	6985	5216
7	5	9	n°5	Mercredi	06/01/2010	23:00	18:06	1875	4172	4325
8	6	11	n°6	Jeudi	07/01/2010	00:00	18:07	3158	8412	8423
9	7	13	n°7	Vendredi	08/01/2010	01:00	18:08	4758	4716	1247
10	8	15	n°8	Samedi	09/01/2010	02:00	18:09	18127	39215	38690
11	9	17	n°9	Dimanche	10/01/2010	03:00	18:10			
12	10	19	n°10	Lundi	11/01/2010	04:00	18:11			
13	11	21	n°11	Mardi	12/01/2010	05:00	18:12			
14	12	23	n°12	Mercredi	13/01/2010	06:00	18:13			
15	13	25	n°13	Jeudi	14/01/2010	07:00	18:14			
16	14	27	n°14	Vendredi	15/01/2010	08:00	18:15			
17	15	29	n°15	Samedi	16/01/2010	09:00	18:16			
18	16	31	n°16	Dimanche	17/01/2010	10:00	18:17			
19	17	33	n°17	Lundi	18/01/2010	11:00	18:18			
20	18	35	n°18	Mardi	19/01/2010	12:00	18:19			
21	19	37	n°19	Mercredi	20/01/2010	13:00	18:20			
22	20	39	n°20	Jeudi	21/01/2010	14:00	18:21			

B. Rechercher et remplacer (Ctrl+H) :

Pratique en cas de recodage de variables : après avoir sélectionné les données, faire Ctrl+H ouvre une boîte de dialogue (par exemple rechercher « oui » et remplacer par « 1 » puis cliquer sur « remplacer tout »)



C. Collage spécial :

1. Coller valeurs :

Les données transmises doivent être vierges de toute formule, si ce n'est pas le cas :

- Sélectionner toutes les données, faire copier
- Sur une cellule vide d'une autre feuille de travail, faire : clic droit, « collage spécial », sélectionner : « coller valeurs »

2. Transposer :

Si les données ont été rentrées avec les patients en colonnes et les variables en lignes:

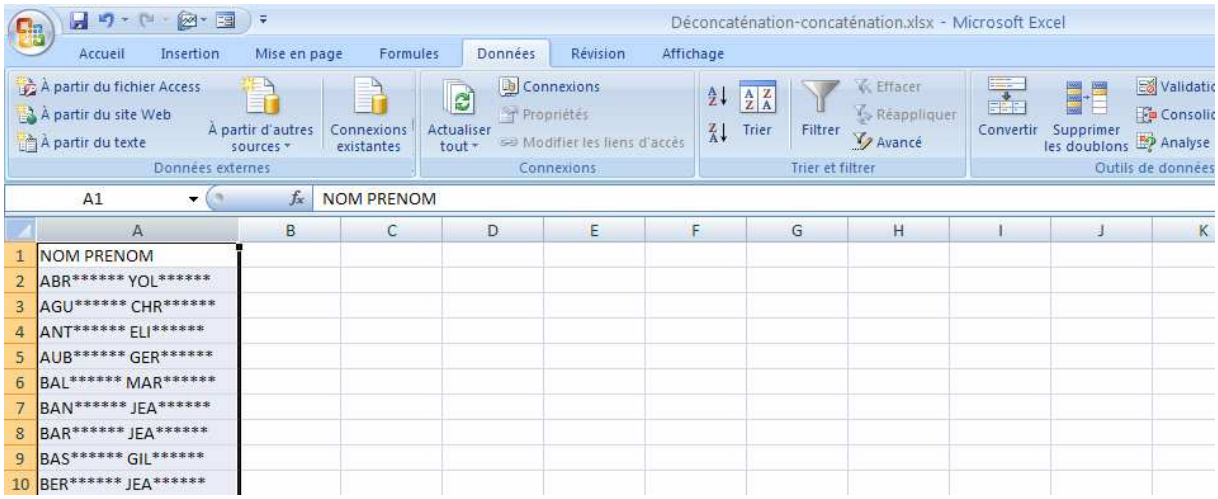
- Sélectionner toutes les données, faire copier
- Sur une cellule vide d'une autre feuille de travail, faire : clic droit, « collage spécial », sélectionner : coller « valeurs » et « transposé »

D. Concaténation et déconcaténation

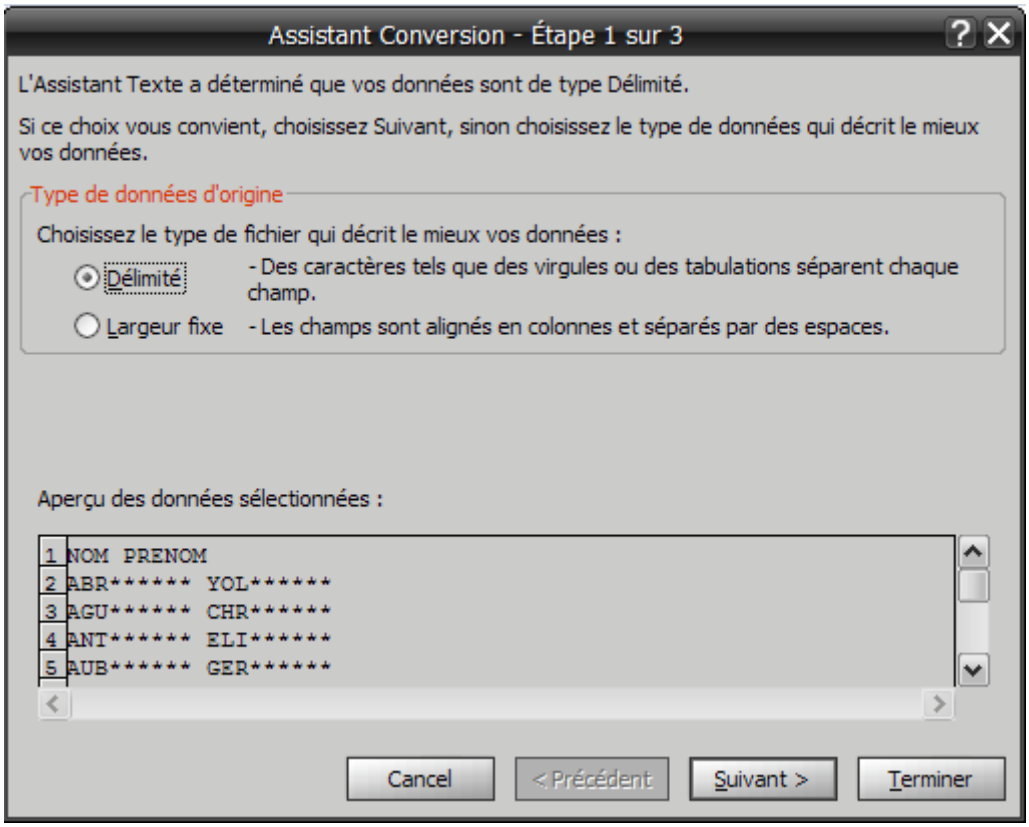
Obtenir une colonne contenant les 3 premières lettres du nom et les 3 premières lettres du prénom séparées par un espace en vue d'une anonymisation des données

1. Déconcaténation (séparer noms et prénoms)

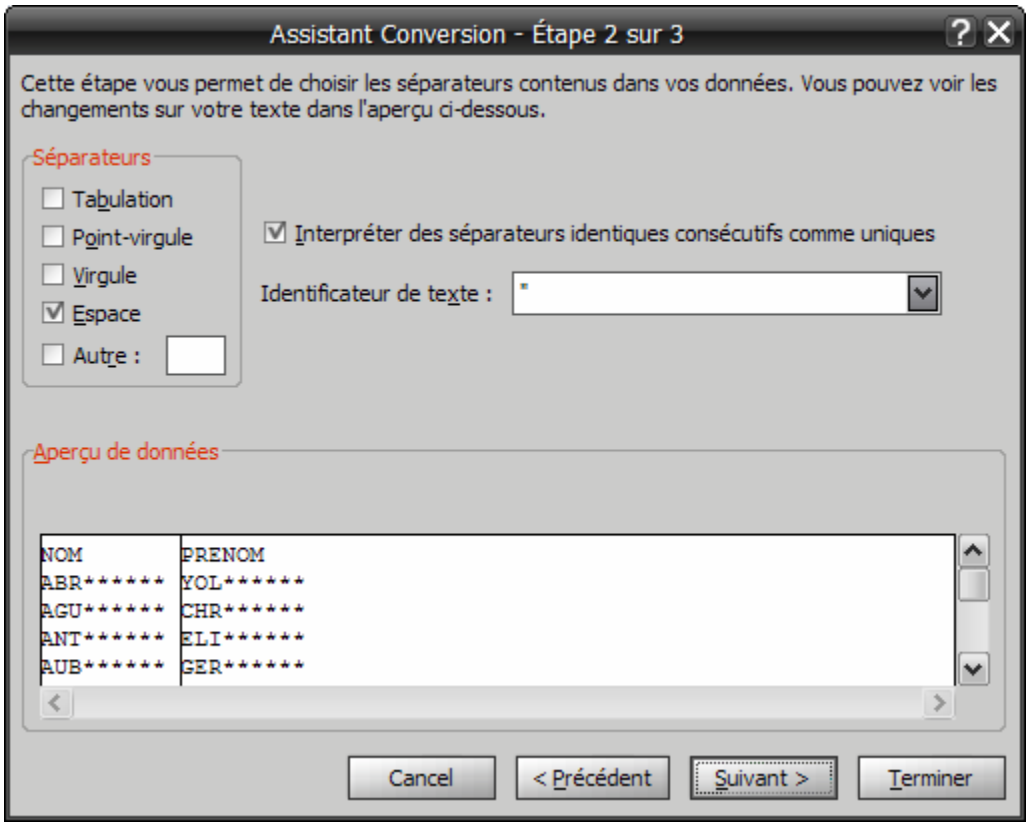
- Sélectionner les données correspondantes aux noms et prénoms (Exemple déconcaténation-concaténation.xlsx)



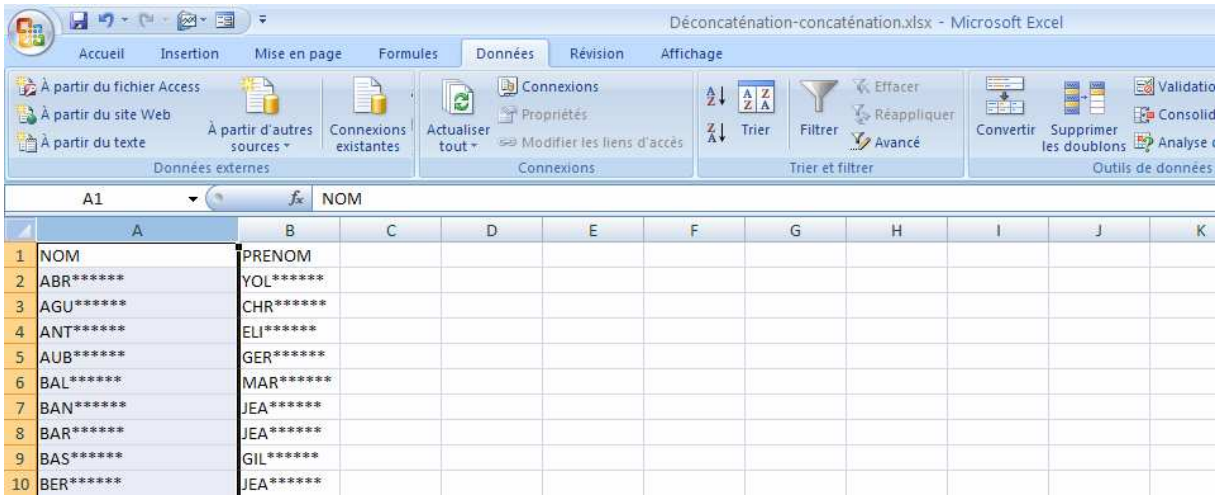
- Utiliser la fonction « convertir » du cadre « outils de données » situé sur l'onglet « données »
- Sélectionner « délimité » pour le type de données d'origine, puis cliquer sur « suivant » :



❖ Sélectionner « espace » pour le séparateur, puis cliquer sur « suivant » :

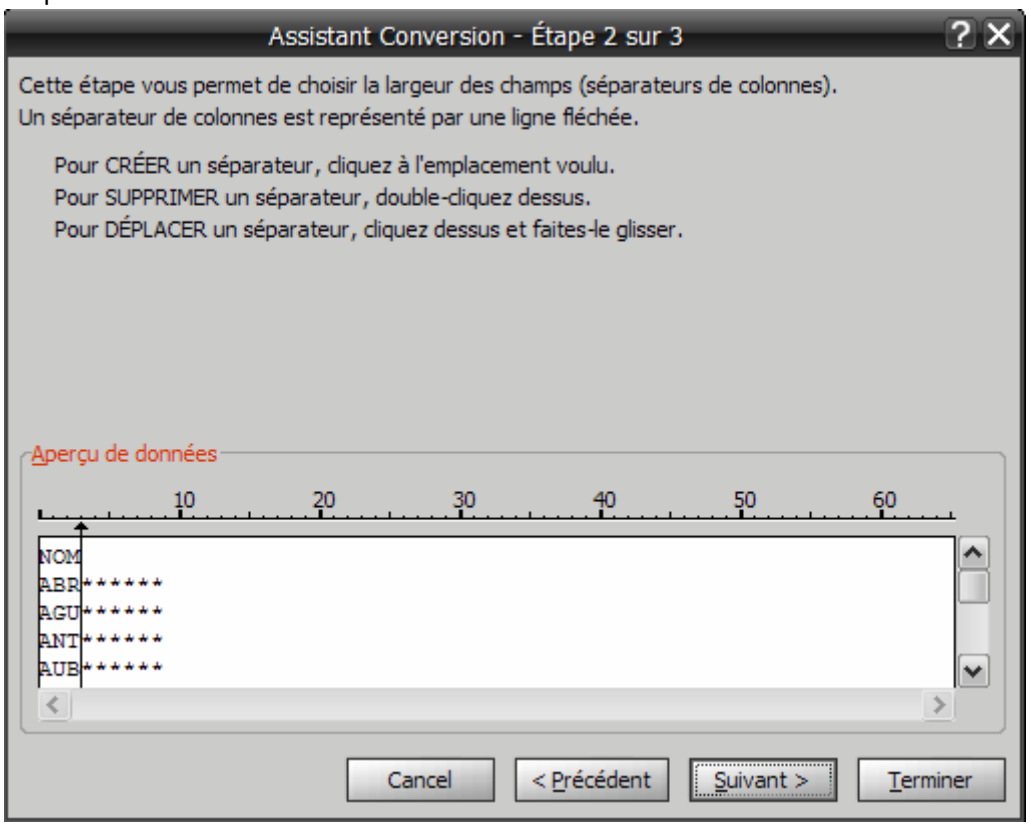


- Indiquer la cellule de destination (prévoir des colonnes vierges pour ne pas écraser les données : obtention de 3 colonnes si noms et prénoms composés sans tiret), puis faire « terminer » :

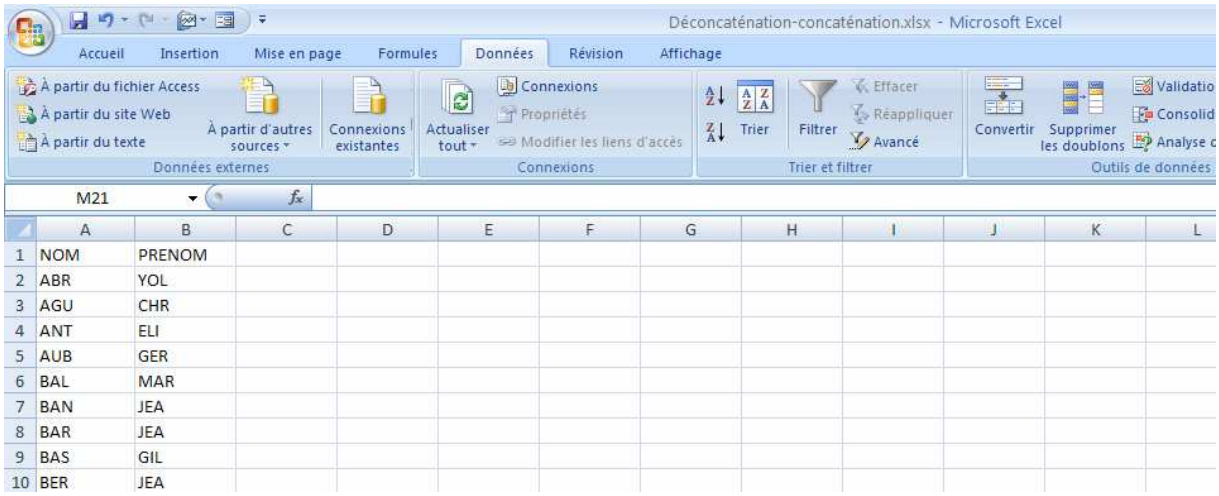


2. Tronquer les noms et les prénoms (garder les 3 premiers caractères)

- Penser à insérer une colonne entre la colonne NOM et la colonne PRENOM
- Sélectionner les données correspondantes aux noms
- Utiliser la fonction « convertir » du cadre « outils de données » situé sur l'onglet « données »
- Sélectionner « largeur fixe » pour le type de données d'origine, puis cliquer sur « suivant »
- Dans « aperçu de données », créer un séparateur après les 3 premiers caractères, puis cliquer sur « terminer » :



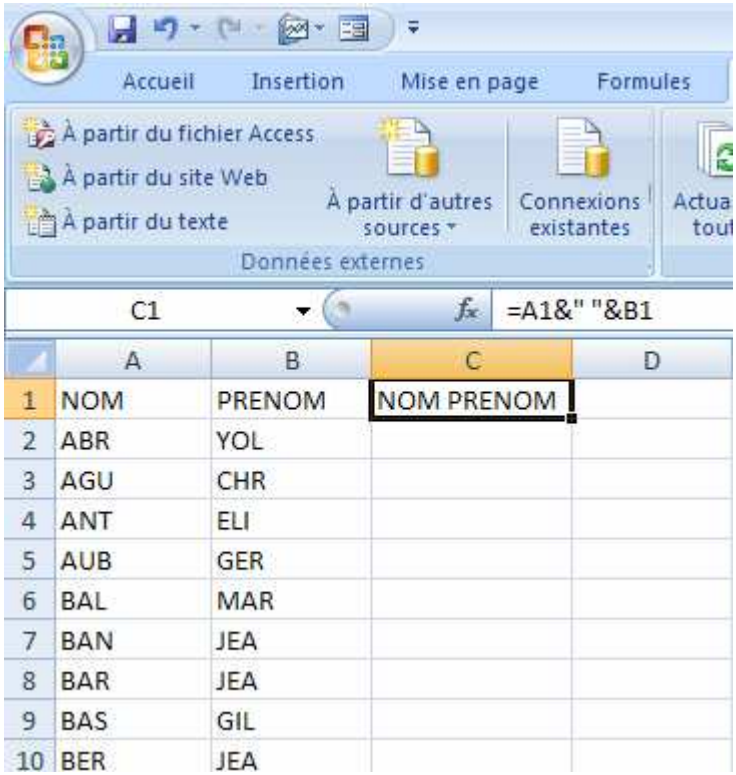
- Supprimer la colonne de la partie tronquée
- Faire la même chose pour la colonne PRENOM :



	A	B	C	D	E	F	G	H	I	J	K	L
1	NOM	PRENOM										
2	ABR	YOL										
3	AGU	CHR										
4	ANT	ELI										
5	AUB	GER										
6	BAL	MAR										
7	BAN	JEA										
8	BAR	JEA										
9	BAS	GIL										
10	BER	JEA										

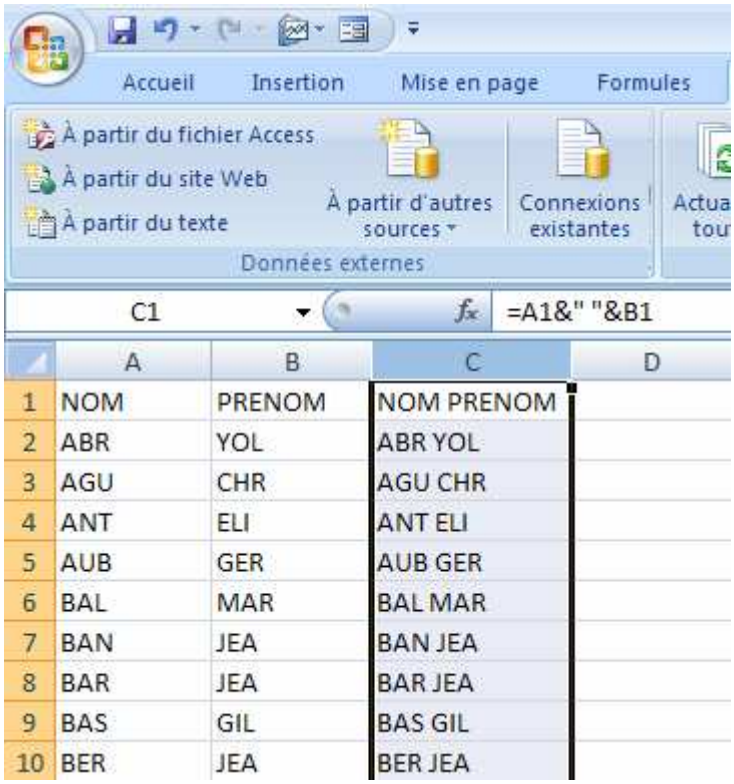
3. Concaténation (rassembler les noms et prénoms tronqués dans une même colonne)

- Dans la cellule C1, taper « =A1&'' ''&B1 » ce qui signifie : réunir les contenus des cellules A1 et B1 dans la cellule C1 en les séparant par un espace (taper seulement =A1&B1 pour réunir les contenus sans espace. Valider avec Entrée :



	A	B	C	D
1	NOM	PRENOM	NOM PRENOM	
2	ABR	YOL		
3	AGU	CHR		
4	ANT	ELI		
5	AUB	GER		
6	BAL	MAR		
7	BAN	JEA		
8	BAR	JEA		
9	BAS	GIL		
10	BER	JEA		

- Pour terminer, utiliser la recopie incrémentée pour assembler les autres noms et prénoms tronqués :



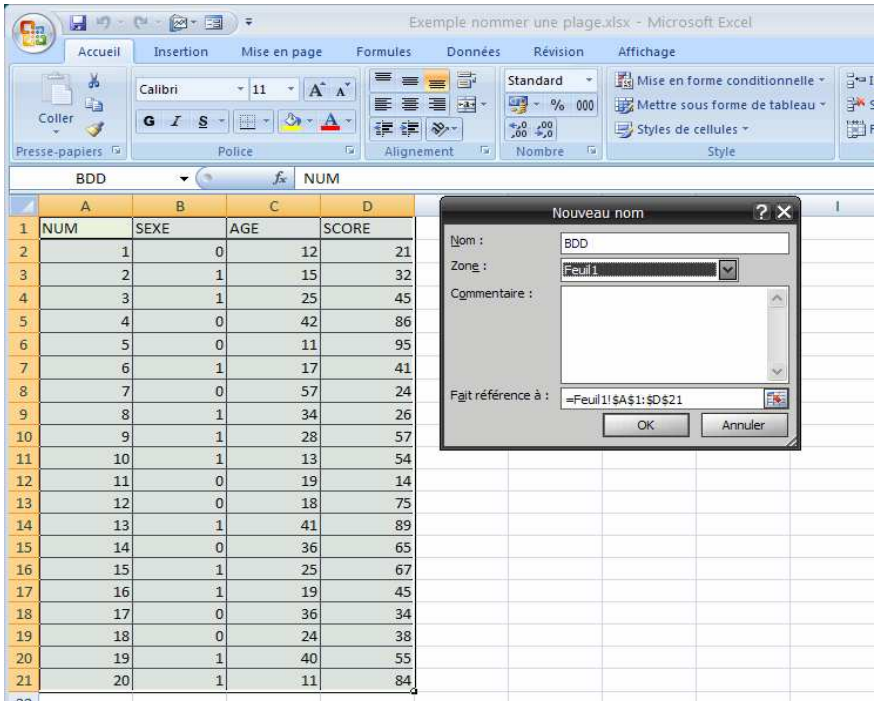
	A	B	C	D
1	NOM	PRENOM	NOM PRENOM	
2	ABR	YOL	ABR YOL	
3	AGU	CHR	AGU CHR	
4	ANT	ELI	ANT ELI	
5	AUB	GER	AUB GER	
6	BAL	MAR	BAL MAR	
7	BAN	JEA	BAN JEA	
8	BAR	JEA	BAR JEA	
9	BAS	GIL	BAS GIL	
10	BER	JEA	BER JEA	

E. Savoir nommer une plage et utiliser une fonction (NB, MOYENNE, SOMME…) sur une variable en prenant en compte seulement les patients répondant à certains critères

Nommer une plage permet d'assigner un nom à un groupe de cellule et de l'utiliser avec des fonctions

1. Nommer la base de donnée

- Sélectionner un groupe de cellule, faire clic droit et choisir « nommer une plage »



	A	B	C	D
1	NUM	SEXE	AGE	SCORE
2	1	0	12	21
3	2	1	15	32
4	3	1	25	45
5	4	0	42	86
6	5	0	11	95
7	6	1	17	41
8	7	0	57	24
9	8	1	34	26
10	9	1	28	57
11	10	1	13	54
12	11	0	19	14
13	12	0	18	75
14	13	1	41	89
15	14	0	36	65
16	15	1	25	67
17	16	1	19	45
18	17	0	36	34
19	18	0	24	38
20	19	1	40	55
21	20	1	11	84

- Dans la boîte de dialogue assigner un nom à la plage, ici « BDD », faire OK

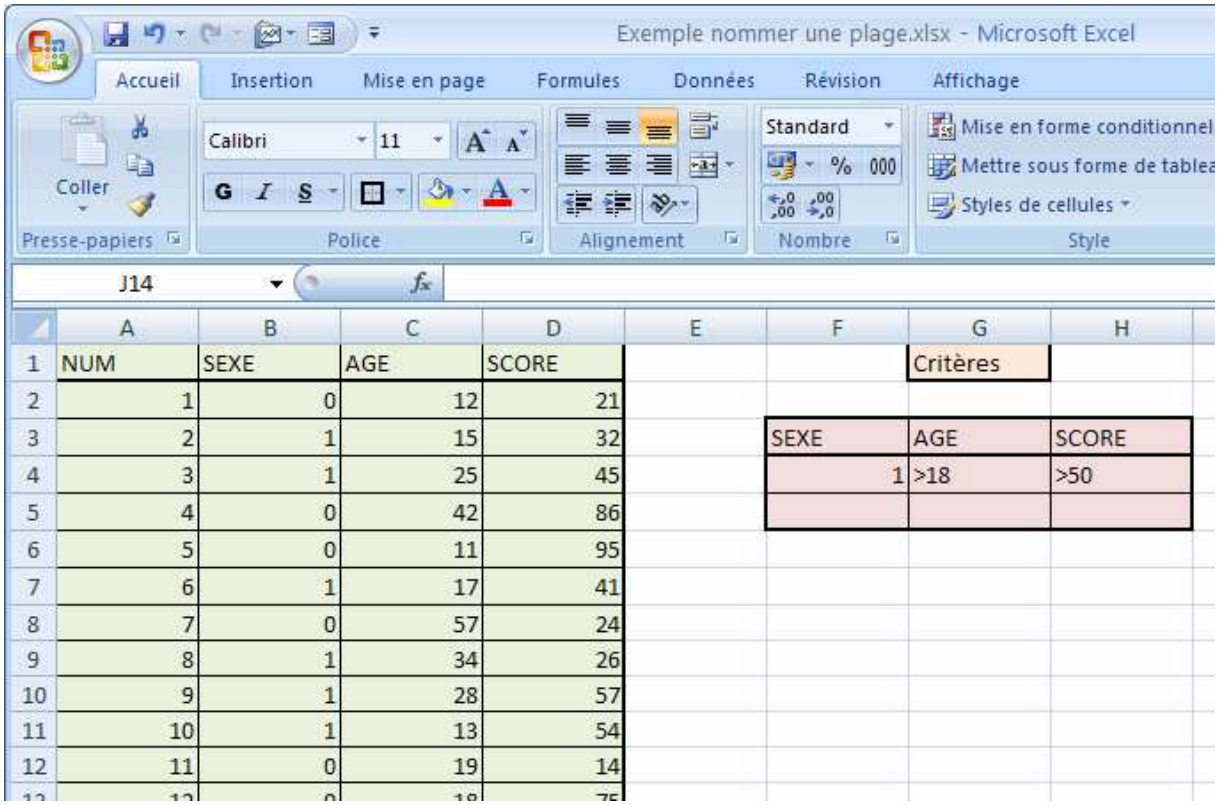
2. Nommer une colonne

- De même sélectionner les données d'une colonne sans l'en-tête, clic droit, nommer une plage
- Si un en-tête est présent en première ligne, Excel va le choisir automatiquement comme nom de la plage
- Assigner un nom aux colonnes NUM, SEXE, AGE et SCORE

3. Utilisation à travers un exemple : la fonction BDNB

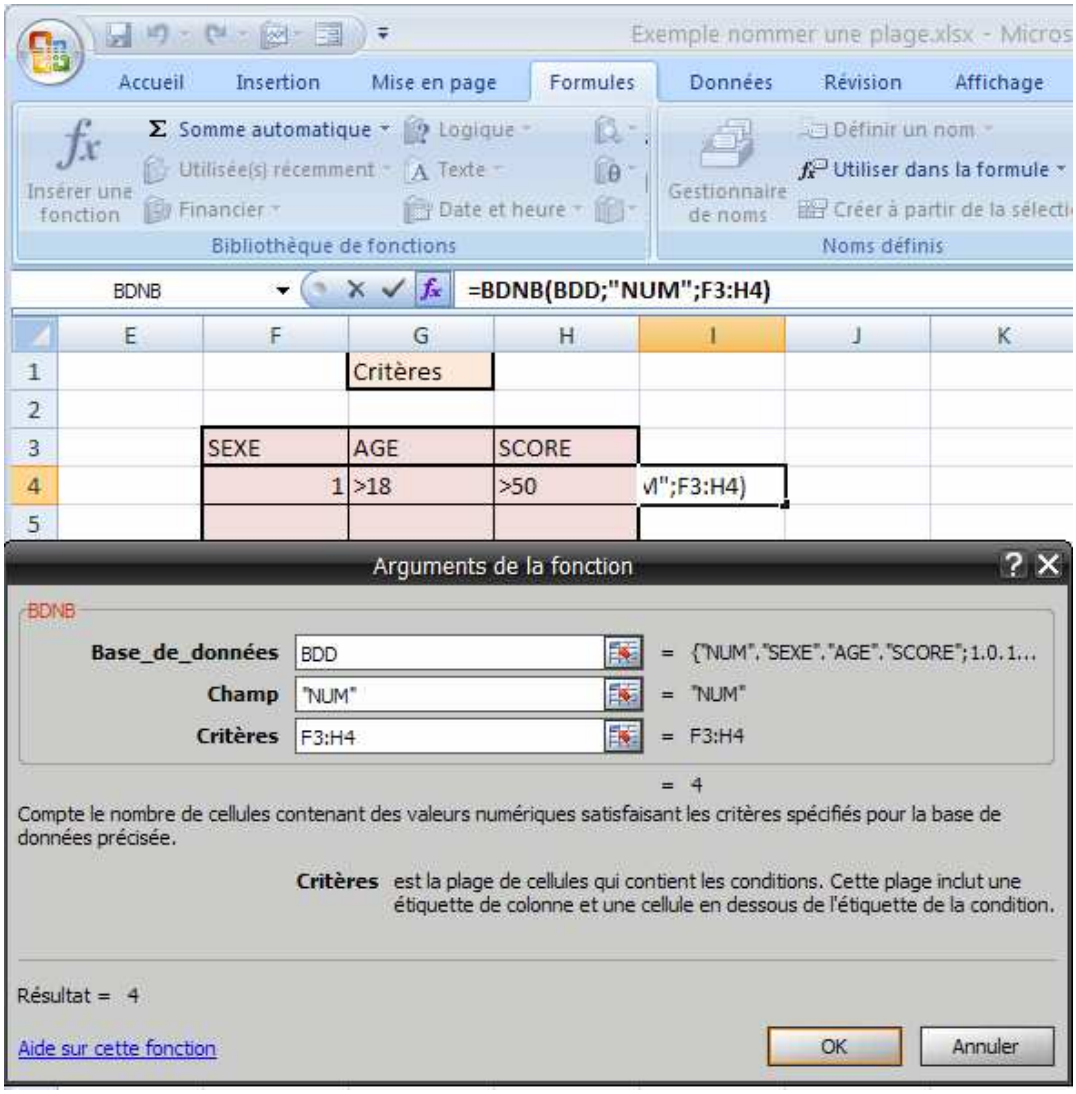
Cette fonction est utile pour calculer un effectif répondant à plusieurs critères dans une base de données

- Après avoir nommé la base de données, et les différentes colonnes, créer un tableau avec les mêmes en-têtes que la base de données
- Indiquer les critères auxquels doit répondre le calcul d'effectif, par exemple pour le calcul du nombre d'hommes majeurs avec un score supérieur à 50: « 1 » sous SEXE, « >18 » sous AGE, et >50 sous score



	A	B	C	D	E	F	G	H
1	NUM	SEXE	AGE	SCORE			Critères	
2	1	0	12	21				
3	2	1	15	32		SEXE	AGE	SCORE
4	3	1	25	45		1	>18	>50
5	4	0	42	86				
6	5	0	11	95				
7	6	1	17	41				
8	7	0	57	24				
9	8	1	34	26				
10	9	1	28	57				
11	10	1	13	54				
12	11	0	19	14				

- Sélectionner la cellule où doit apparaître le résultat
- Onglet « Formules », Insérer une fonction, catégorie « tous », choisir « BDNB »
- Dans la boîte de dialogue, taper BDD pour « Base_de_données », ''NUM'' **entre guillemets** pour « Champ », puis sélectionner les cellules contenant les critères accompagnés des en-têtes (de F3 à H4), faire OK : le résultat (4) s'inscrit dans la cellule voulue



- Les fonctions BDMIN, BDMAX, BDMOYENNE, BDECARTYPE, BDSOMME, etc. s'utilisent de la même manière