École nationale vétérinaire d'Alfort

Année 2 – Semestre S3
2023/2024

UC 0213
Communication et réglementation dans la profession vétérinaire (CoRVet)

Utilisation d'Excel et du site Internet BiostaTGV pour réaliser quelques statistiques de base

Auteur : Pr Loïc Desquilbet
Version : Juin 2023

Table des matières

I. Avant de faire des stat', une petite intro !

A. Fichier de données qui va servir de base aux exemples

Ce tutoriel va utiliser les données du fichier Excel intitulé « Pour tuto Excel stat descriptives.xlsx », présent sur la page Internet suivante, section « Fichiers Excel » : http://eve.vet-alfort.fr/course/view.php?id=353&section=5. Vous pouvez télécharger ce fichier, et suivre ce tutoriel en faisant de votre côté sur votre ordinateur ce que je vous montre ici. Ce fichier de données contient quelques caractéristiques individuelles, en colonnes, de 50 chiens (de la ligne 2 à la ligne 51). Voici les 20 premières lignes du fichier de données :

	A	B	C	D	E	F
1	Num_animal	Sterile	Femelle	Race_cl	Glycemie	Insuffisance_renale
2	1	1	1	3	1.2	0
3	2	1	0	5	0.82	0
4	3	0	0	2		
5	4	1	0	1	1.2	0
6	5	1	0	5	1.14	0
7	6	1	0	5	1.17	1
8	7	1	1	5	1.81	0
9	8	1	0	3		
10	9	0	0	5	0.86	0
11	10	1	0	5	1.21	1
12	11	0	0	5	0.34	1
13	12	0	1	2	0.95	
14	13	0	1	5	1.66	1
15	14	0	0	1	1.11	0
16	15	0	1	5	1.16	0
17	16	0	1	1	1.03	1
18	17	1	0	3	1.13	1
19	18	0	1	1	1.42	0
20	19	1	1	5		0

B. Version d'Excel utilisée

La version d'Excel que j'ai utilisée dans ce tutoriel est la version d'Office 2010. Il est possible qu'en fonction des versions que vous utilisez, il y ait quelques petites différences… J'espère qu'elles ne compromettront pas trop la compréhension de ce tutoriel, si vous suivez pas à pas ce tutoriel avec votre ordinateur et le fichier Excel que vous auriez téléchargé !

C. Quelques définitions

Dans la suite de ce tutoriel, je vais utiliser quelques termes assez spécifiques. Je préfère donc les définir, pour éviter certaines confusions ou incompréhensions…

Variable : j'ai écrit ci-dessus le mot « colonne », mais en fait, en langage stat', on parle de « variable ». (Le fichier de données ci-dessus contient 7 variables, nommées « Num_animal », « Sterile », …, et « Insuffisance_renale ».) Les variables sont de deux types : numérique & alphanumérique. Une variable numérique ne contient absolument que des chiffres, tandis qu'une variable alphanumérique peut contenir des caractères autres que des chiffres (tout caractère autre que le séparateur de décimal, « 0 », « 1 », …, ou « 9 » est considéré comme un caractère alphanumérique). Dans le fichier de données de l'exemple, les 7 variables sont numériques.

Croisement : on parle de « croisement » de deux variables quand on veut savoir comment deux variables sont associées. Par exemple, si je croise deux variables binaires telles que le sexe M/F et la stérilisation S/E[1], cela va produire un tableau à 4 cases suivant :

		Statut de stérilisation	
		Stérilisé	Entier
Sexe de l'animal	Mâle	A	B
	Femelle	C	D

Où A, B, C, et D sont les nombres d'animaux correspondant à leurs caractéristiques *croisées*.

Variable qualitative : une variable qualitative est une variable qui comprend trois classes ou plus. (Si elle n'en contient que deux, on parle de « variable binaire ».) Ces classes peuvent ou non être ordonnées. La variable « Race » dont les classes seraient « Bulldog français », « Boxer », « Caniche », et « autre race » serait une variable qualitative dont les classes ne sont pas ordonnées (on parle alors de « variable qualitative nominale »). La variable « Frequence de repas par jour », dont les classes seraient « 1 fois par jour », « 2 fois par jour », « 3 fois par jour ou plus » serait une variable qualitative dont les classes sont ordonnées (on parle de « variable qualitative ordinale »).

Variable quantitative : une variable quantitative est une variable représentant une mesure ou une quantification (avec ou sans chiffre après la virgule), prenant potentiellement un nombre important (> 5) de valeurs différentes. Par exemple, la variable *taille_elevage* représentant le nombre de vaches dans un élevage est une variable quantitative. En revanche, un score sur une échelle de 1 à 4 ne prenant que des valeurs entières est une variable davantage qualitative ordinale que quantitative. (Mais il faut avouer que la frontière est ténue, et il n'y a pas de règles strictes permettant de distinguer une variable qualitative ordinale d'une variable quantitative).

D. Vérifier la normalité d'une variable quantitative

On dit qu'une variable quantitative « suit une loi normale » si la distribution de ses valeurs suit *à peu près* une loi normale. Pour le vérifier, il faut dresser un histogramme à partir des données de l'échantillon. Pour cela, vous pouvez utiliser un fichier Excel disponible ici[2]. Un site internet permettant de dresser un histogramme se trouve à l'adresse suivante[3]. Attention lorsque vous utilisez ce site internet : si vos données sont des nombres avec virgules, quand vous collerez vos données dans le champ, vous devrez remplacer les « , » par des « . ». Par ailleurs, pour savoir quel est le nombre optimal de barres d'histogramme, je vous recommande d'utiliser la formule de Brooks-Carruthers suivante : valeur entière de $\{5.\mathrm{Log}_{10}(n)\}$ où n est le nombre de valeurs dont on cherche à savoir si elles suivent une distribution normale. Par exemple, si vous voulez vérifier la normalité à partir de 38 valeurs, le nombre optimal de barres de l'histogramme sera : valeur entière de $\{5.\mathrm{Log}_{10}(38)\} = 7$ barres. Si la distribution ne peut pas être considérée comme normale, ou s'il y a trop peu de valeurs prises par la variable *a priori* quantitative pour dresser un histogramme, nous partirons du principe que la distribution ne suit pas une loi normale.

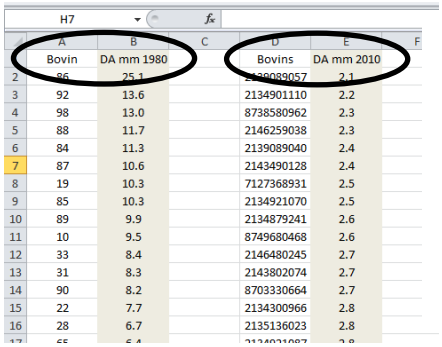
[1] S/E = stérilisé/entier
[2] http://eve.vet-alfort.fr/course/view.php?id=353&section=5
[3] http://www.socscistatistics.com/descriptive/histograms/

E. Quelques règles de base pour travailler correctement avec un fichier Excel

Pour pouvoir travailler proprement avec un fichier de données Excel, voici les recommandations que je peux faire :

- Chaque animal doit se présenter en ligne (il y a donc autant de lignes que d'animaux – sans compter la première ligne correspondant au nom des colonnes), et les caractéristiques individuelles (sexe, âge, race, …) doivent se présenter en colonnes.

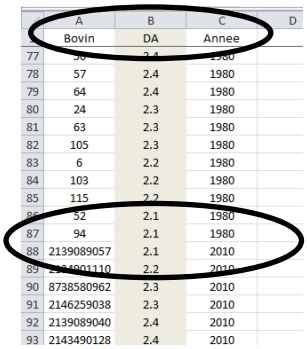
  Il arrive parfois que l'on reçoive un fichier de données où les animaux sont sur deux ou plusieurs *colonnes* (cf. ci-dessous) :



	A	B	C	D	E	F
	Bovin	DA mm 1980		Bovins	DA mm 2010	
2	86	25.1		2139089057	2.1	
3	92	13.6		2134901110	2.2	
4	98	13.0		8738580962	2.3	
5	88	11.7		2146259038	2.3	
6	84	11.3		2139089040	2.4	
7	87	10.6		2143490128	2.4	
8	19	10.3		7127368931	2.5	
9	85	10.3		2134921070	2.5	
10	89	9.9		2134879241	2.6	
11	10	9.5		8749680468	2.6	
12	33	8.4		2146480245	2.7	
13	31	8.3		2143802074	2.7	
14	90	8.2		8703330664	2.7	
15	22	7.7		2134300966	2.8	
16	28	6.7		2135136023	2.8	

Dans le tableau ci-contre, les données de DA de bovins sont séparées en deux colonnes pour 1980 et 2010. Par exemple, le bovin n°87 a une valeur de DA de 10,6, et il a été évalué en 1980. Ce fichier de données n'est pas analysable facilement en utilisant notamment les tableaux croisés dynamiques d'Excel, ou d'autres logiciels de statistique. Il peut en revanche l'être pour certaines analyses. Il ne faut donc pas supprimer cette feuille mais la dupliquer avant de la remettre en forme.

  La « remettre en forme », c'est placer tous les bovins les uns sous les autres, fusionner les deux variables « DA mm 1980 » et « DA mm 2010 » en une seule nouvelle variable que je vais nommer « DA », et enfin créer une 3ème variable que je vais appeler « Annee » qui vaudra soit « 1980 » soit « 2010 » :



	A	B	C	D
	Bovin	DA	Annee	
77	56	2.4	1980	
78	57	2.4	1980	
79	64	2.4	1980	
80	24	2.3	1980	
81	63	2.3	1980	
82	105	2.3	1980	
83	6	2.2	1980	
84	103	2.2	1980	
85	115	2.2	1980	
86	52	2.1	1980	
87	94	2.1	1980	
88	2139089057	2.1	2010	
89	2134901110	2.2	2010	
90	8738580962	2.3	2010	
91	2146259038	2.3	2010	
92	2139089040	2.4	2010	
93	2143490128	2.4	2010	

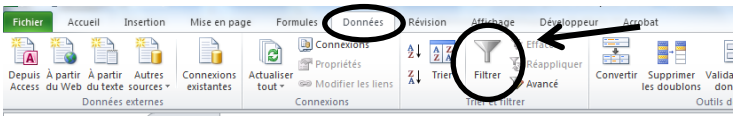
Les bovins évalués en 2010 ont été (arbitrairement) placés sous les bovins évalués en 1980.

- La 1ère ligne doit comporter le nom de la variable, si possible concis. Si vous êtes sûr(e) de ne travailler que sur Excel, vous pouvez laisser des espaces et des caractères spéciaux (accents, …) dans le nom des variables. Mais si vous devez à un moment exporter votre fichier de données vers un logiciel de statistique, alors le nom des variables ne doit comporter ni espace, ni caractère spécial (incluant donc les accents).
- Il ne doit pas y avoir de ligne totalement vide entre deux lignes d'animaux.
- Une donnée manquante doit être une cellule vide, dans Excel.
- Lorsqu'une variable est censée être numérique, ne pas mettre de caractères dans une des cellules. Cela a l'air évident, mais par exemple, pour une variable biologique (numérique, donc), il est fréquent de voir écrit « < X » (X étant un seuil de détection, par exemple). Or, le signe « < » est un signe alphanumérique qui posera des problèmes quand il faudra faire des statistiques sur cette variable biologique ! De même, si une donnée numérique manque, il ne faut pas écrire « NSP » ou « ? », pour les mêmes raisons. Au besoin, vous pouvez créer une colonne supplémentaire de commentaires, où vous pourrez y écrire tout ce que vous voudrez ! ☺

F. La fonction « Filtre » d'Excel

La fonction « Filtre » s'utilise pour remplir deux objectifs : (a) pour sélectionner des lignes selon les valeurs d'une variable qualitative, (b) pour avoir un aperçu rapide du contenu de l'ensemble d'une colonne (pour vérifier par exemple qu'il n'y a pas de valeurs aberrantes ou de signes alphanumériques qui se promènent par mégarde dans la colonne).

Pour sélectionner des individus selon les valeurs d'une variable qualitative, cliquez sur la 1ère ligne de votre fichier (la ligne qui contient le nom des variables), dans n'importe quelle colonne non vide (ci-dessous, entre les colonnes A et G), et allez dans le menu « Données » puis cliquez sur « Filtrer ».

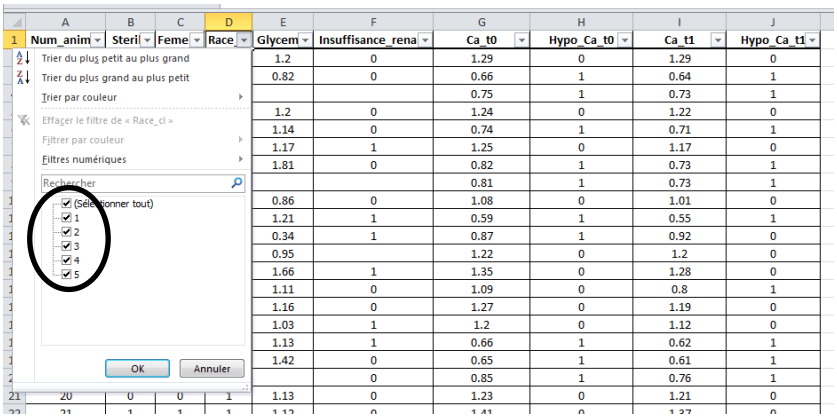


Vous obtenez ainsi des petites flèches sur le bord droit de chaque cellule sur la première ligne :

	Num_anim	Steri	Feme	Race	Glycem	Insuffisance_rena
2	1	1	1	3	1.2	0
3	2	1	0	5	0.82	0

Chaque flèche est un accès à une liste déroulante qui contient tout ce que contient la colonne ! C'est très important car cette commande permet notamment de repérer des choses un peu aberrantes pour des variables numériques, telles qu'un « ? » ou un « NSP »…

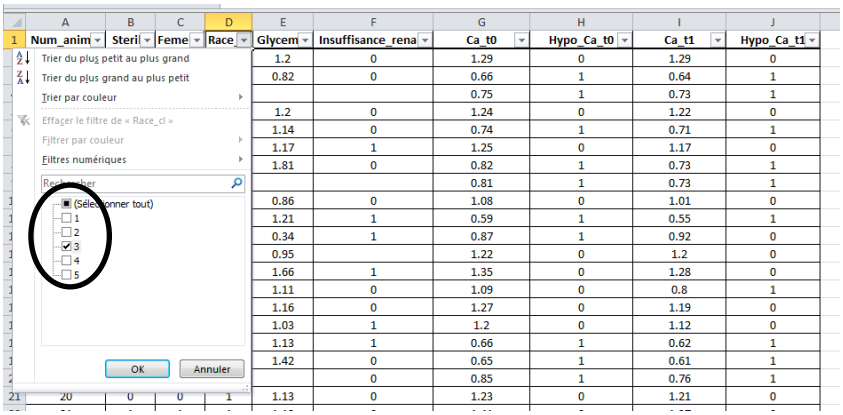
En cliquant sur la petite flèche pour la variable « Race_cl », on obtient ceci :



	Num_anim	Steri	Feme	Race	Glycem	Insuffisance_rena	Ca_t0	Hypo_Ca_t0	Ca_t1	Hypo_Ca_t1
					1.2	0	1.29	0	1.29	0
					0.82	0	0.66	1	0.64	1
							0.75	1	0.73	1
					1.2	0	1.24	0	1.22	0
					1.14	0	0.74	1	0.71	1
					1.17	1	1.25	0	1.17	0
					1.81	0	0.82	1	0.73	1
							0.81	1	0.73	1
					0.86	0	1.08	0	1.01	0
					1.21	1	0.59	1	0.55	1
					0.34	1	0.87	1	0.92	0
					0.95		1.22	0	1.2	0
					1.66	1	1.35	0	1.28	0
					1.11	0	1.09	0	0.8	1
					1.16	0	1.27	0	1.19	0
					1.03	1	1.2	0	1.12	0
					1.13	1	0.66	1	0.62	1
					1.42	0	0.65	1	0.61	1
						0	0.85	1	0.76	1
21	20	0	0	1	1.13	0	1.23	0	1.21	0

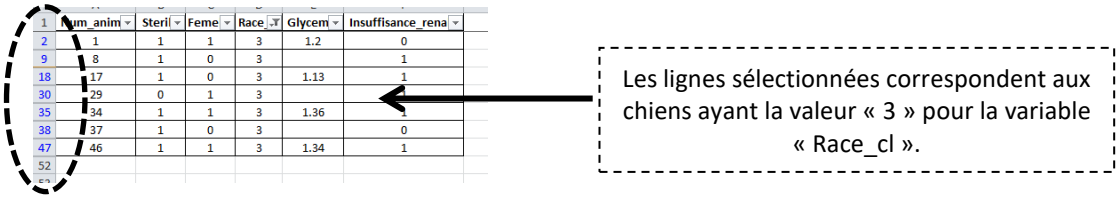
Première information que la figure ci-dessus fournit : les seules choses qui sont contenues dans la colonne « Race_cl » sont « 1 », « 2 », « 3 », « 4 », et « 5 ». Par défaut, toutes les cases sont cochées, cela signifie que toutes les lignes ont été par défaut sélectionnées pour apparaître à l'écran.

Supposons que l'on ne veuille sélectionner que les chiens de valeur « 3 » pour cette variable « Race_cl », il faut alors cocher « 3 » dans la liste déroulante :



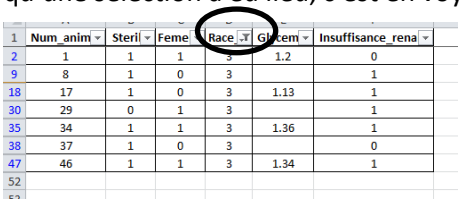
	Num_anim	Steri	Feme	Race	Glycem	Insuffisance_rena	Ca_t0	Hypo_Ca_t0	Ca_t1	Hypo_Ca_t1
					1.2	0	1.29	0	1.29	0
					0.82	0	0.66	1	0.64	1
							0.75	1	0.73	1
					1.2	0	1.24	0	1.22	0
					1.14	0	0.74	1	0.71	1
					1.17	1	1.25	0	1.17	0
					1.81	0	0.82	1	0.73	1
							0.81	1	0.73	1
					0.86	0	1.08	0	1.01	0
					1.21	1	0.59	1	0.55	1
					0.34	1	0.87	1	0.92	0
					0.95		1.22	0	1.2	0
					1.66	1	1.35	0	1.28	0
					1.11	0	1.09	0	0.8	1
					1.16	0	1.27	0	1.19	0
					1.03	1	1.2	0	1.12	0
					1.13	1	0.66	1	0.62	1
					1.42	0	0.65	1	0.61	1
						0	0.85	1	0.76	1
21	20	0	0	1	1.13	0	1.23	0	1.21	0

En cliquant ensuite sur « Ok », voici ce que l'on obtient :



	Num_anim	Steri	Feme	Race_cl	Glycem	Insuffisance_rena
2	1	1	1	3	1.2	0
9	8	1	0	3		1
18	17	1	0	3	1.13	1
30	29	0	1	3		1
35	34	1	1	3	1.36	1
38	37	1	0	3		0
47	46	1	1	3	1.34	1
52						

Les numéros de ligne sont désormais en bleu (numéros entourés en pointillés ci-dessus). C'est très important car c'est une des seules façons de voir qu'une sélection a été effectuée, et donc que vous ne travaillez pas sur la totalité du fichier de données ! Si vous n'y prenez pas gare, vous pourriez faire des stat' seulement sur une partie des animaux, alors que vous pensiez les faire sur la totalité, tout simplement parce que vous auriez oublié de retirer votre sélection ! Pour retirer la sélection, il faut cliquer à nouveau sur la flèche, et cliquer sur « (Sélectionner tout) ». Une autre façon de repérer qu'une sélection a eu lieu, c'est en voyant que la flèche initiale s'est transformée en entonnoir :



	Num_anim	Steri	Feme	Race_cl	Glycem	Insuffisance_rena
2	1	1	1	3	1.2	0
9	8	1	0	3		1
18	17	1	0	3	1.13	1
30	29	0	1	3		1
35	34	1	1	3	1.36	1
38	37	1	0	3		0
47	46	1	1	3	1.34	1
52						

G. Les tableaux croisés dynamiques

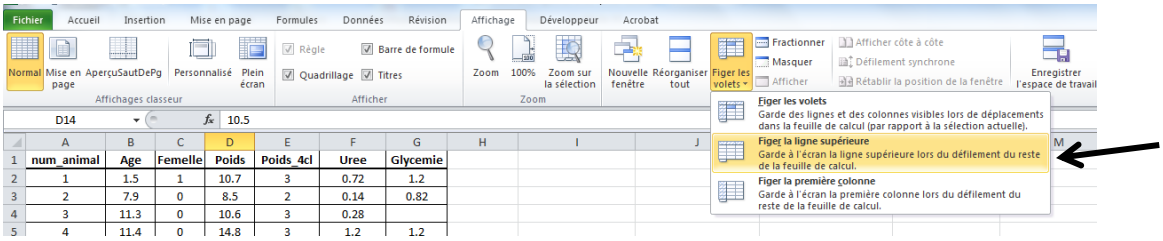
Excel a intégré depuis longtemps une procédure pour traiter les données d'un tableau très (très) puissante : le tableau croisé dynamique (TCD). Ce tutoriel va complètement s'en inspirer : certaines fonctions d'Excel pourraient permettre d'obtenir les mêmes résultats que ceux fournis par les TCD, mais de façon très souvent beaucoup plus fastidieuse. Je vais cependant présenter l'utilisation des fonctions d'Excel lorsque les TCD ne permettent pas d'obtenir l'information souhaitée. Pour comprendre l'utilisation des TCD dans ce tutoriel, il faut avoir visionné au moins un tutoriel vidéo que vous pouvez trouver sur YouTube. En voici un que j'ai sélectionné (il en existe certainement des centaines d'autres !) :

https://www.youtube.com/watch?v=E5shJb7Zndk (visionnez les premières 7min45')

Ce que je vais vous présenter des TCD dans ce tutoriel ne représente que très peu la puissance de cet outil. Je vous invite à visionner plusieurs tutoriels si vous souhaitez en savoir davantage sur les TCD…

H. Figer les « volets » dans Excel

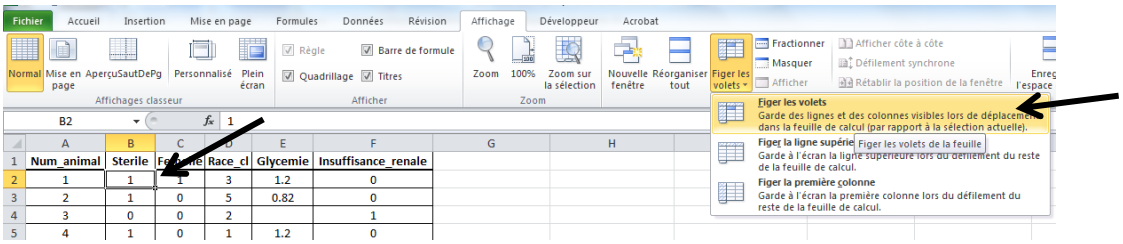
Une astuce dans Excel qui peut se révéler bien pratique est la suivante. Quand vous êtes devant votre fichier de données, et que vous descendez le chariot vertical, vous ne voyez plus apparaître le nom des variables, ce qui peut être gênant ! Pour pallier ce léger problème, placez-vous sur le fichier de données de telle façon à voir la première ligne contenant le nom des variables. Cliquez ensuite sur « Figer la ligne supérieure », dans « Figer les volets », lui-même situé dans l'onglet « Affichage » :



	num_animal	Age	Femelle	Poids	Poids_4cl	Uree	Glycemie
2	1	1.5	1	10.7	3	0.72	1.2
3	2	7.9	0	8.5	2	0.14	0.82
4	3	11.3	0	10.6	3	0.28	
5	4	11.4	0	14.8	3	1.2	1.2

Ainsi, quand vous vous promènerez verticalement dans votre fichier de données, la première ligne (contenant le nom des variables) sera toujours apparente.

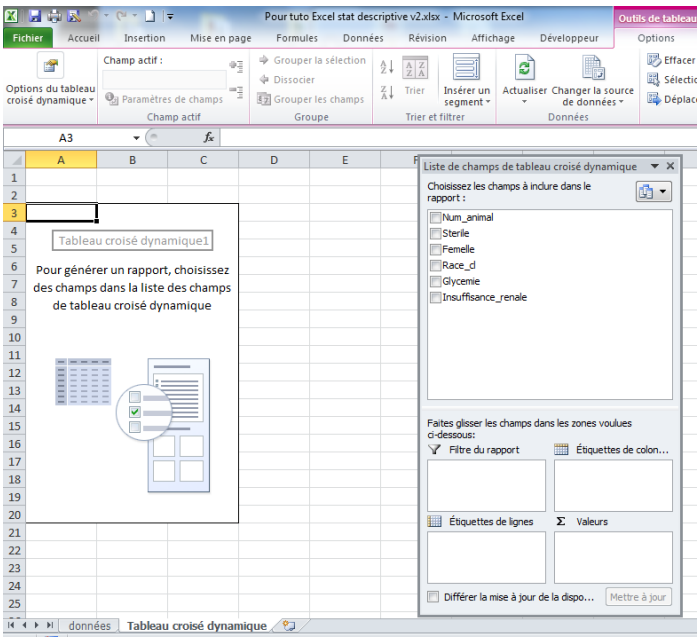
Si vous voulez figer la première ligne *et* la première colonne, vous devez sélectionner la cellule B2, et cliquez sur « Figer les volets » :



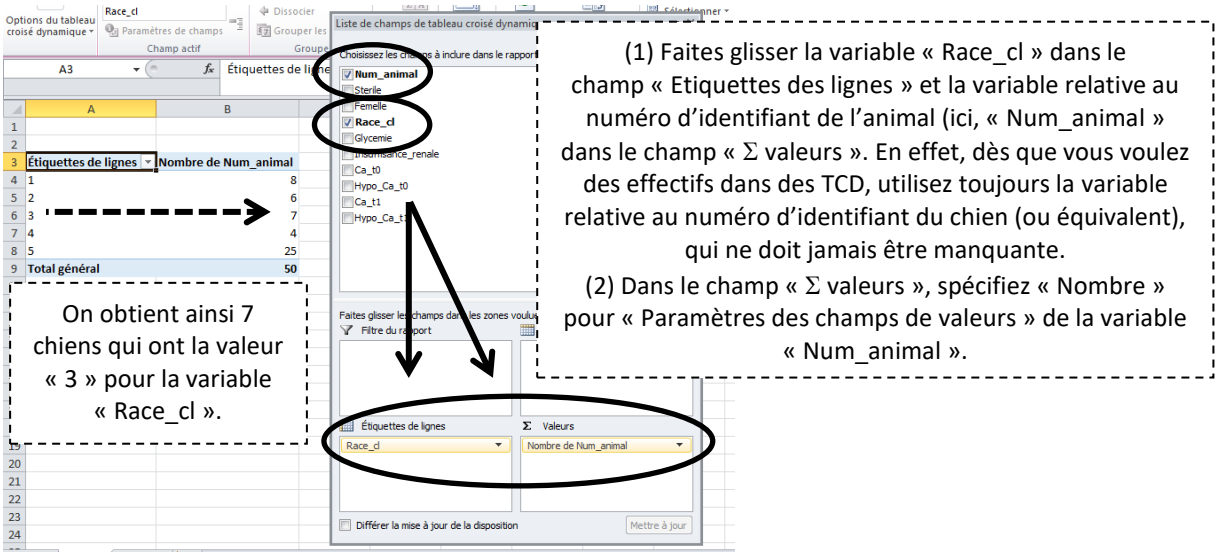
II. Statistiques de base sur une seule variable

A. Description des variables qualitatives (en classes, ou binaire)

Vous devez tout d'abord insérer un TCD (cf. https://www.youtube.com/watch?v=E5shJb7Zndk). Une fois que ce TCD est inséré, voici ce que vous obtenez :



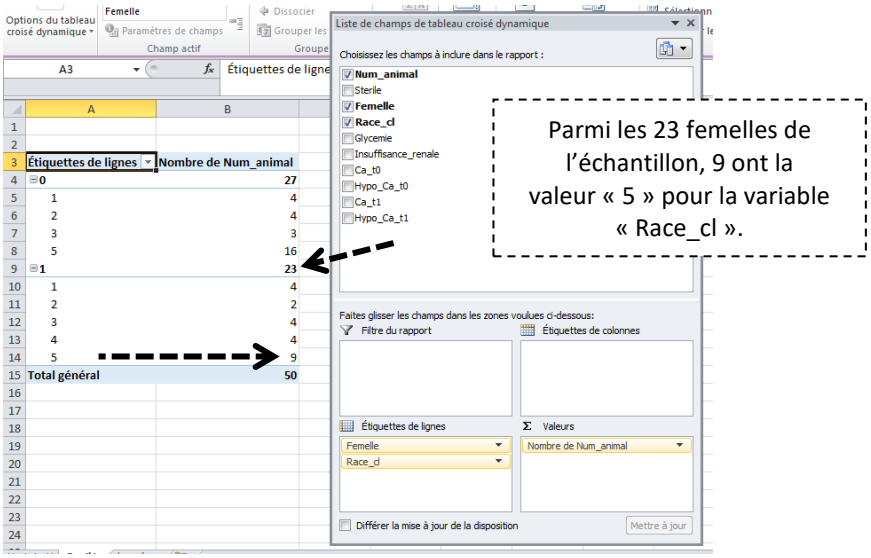
Supposons que nous souhaitions connaître le nombre de chiens de race « 3 » (valeur = « 3 » pour la variable « Race_cl ») dans le fichier de données. Voici comment procéder :



Supposons ensuite que nous souhaitions connaître le nombre de femelles parmi les chiens dont la valeur pour la variable « Race_cl » est égale à « 5 ».

Pour cela, commencez par faire le « nettoyage » du TCD, c'est-à-dire vider ses champs. Pour cela, faites glisser la variable « Race_cl » du champ « Etiquettes de lignes » vers le champ du haut où sont listées toutes les variables. Faites de même avec « Nombre de Num_animal ». Ensuite, faites apparaître les deux variables « Femelle » et « Race_cl » dans le champ « Etiquettes de lignes », avec « Femelle » au-dessus de « Race_cl », puis faites glisser la variable « Nom_animal » dans le champ « Σ valeurs », en spécifiant « Nombre ».

Voici ce que l'on obtient :



B. Description des variables quantitatives

1. *Introduction*

Pour décrire une variable quantitative, de nombreux indicateurs existent. Les principaux sont les suivants (ceux dont je vais parler dans ce tutoriel) : le minimum, le 1er quartile (le 25ème percentile[4]), la médiane, la moyenne, le 3ème quartile (le 75ème percentile), le maximum, et la Standard Deviation (SD)[5].

2. *Les fonctions d'Excel & utilisation de certaines fonctions*

Les fonctions d'Excel pour les indicateurs cités ci-dessus sont présentées dans le tableau ci-dessous.

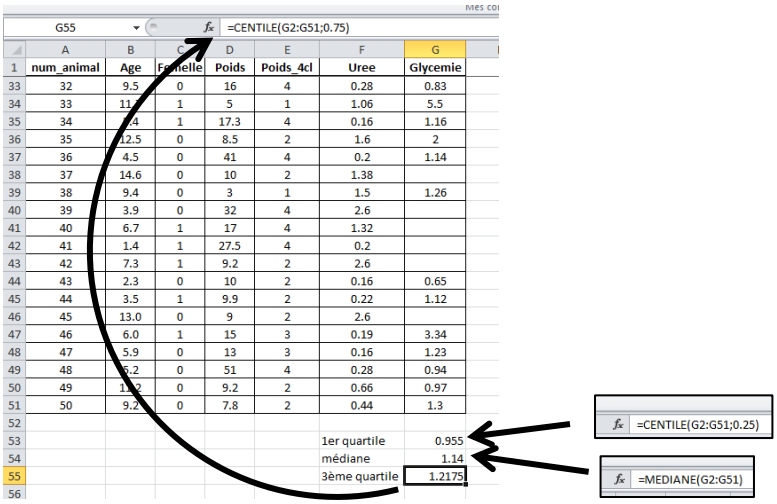
Indicateur	**Fonction dans Excel**
Minimum[a]	MIN(*plage*)
1er quartile	CENTILE(*plage*;0,25)
Médiane	CENTILE(*plage*;0,50) ou bien MEDIANE(*plage*)
3ème quartile	CENTILE(*plage*;0,75)
Maximum[a]	MAX(*plage*)
Moyenne[a]	MOYENNE(*plage*)
Standard Deviation (SD)[a]	ECARTYPE(*plage*)
Variance dans l'échantillon[a]	VAR(*plage*)
Nombre de valeurs non manquantes[a]	NBVAL(*plage*)

[a] Les TCD permettent de fournir ces indicateurs

Comme vous pouvez le voir dans le tableau ci-dessus, on ne peut pas utiliser les TCD pour obtenir des médianes ou des quartiles. Pour obtenir la médiane, le 1er quartile et le 3ème quartile de la glycémie parmi les 50 chiens, voici les trois formules qu'il faut taper :

[4] Le 25ème percentile, ou 1er quartile, est la valeur de la variable telle que 25% des individus ont une valeur au maximum égale à cette valeur. Par exemple, si le 1er quartile de la taille des femmes en France est de 1m62, cela signifie que 25% des femmes en France mesure 1m62 ou moins, et, de façon équivalente, 75% mesurent plus d'1m62.

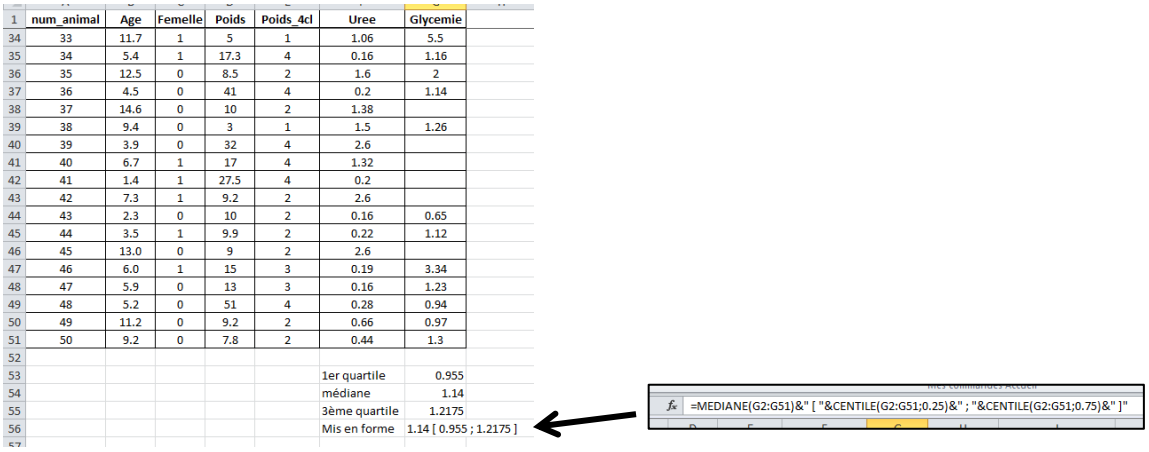
[5] Ecart-type dans l'échantillon, à ne pas confondre avec la Standard Error (SE), qui est l'écart-type d'une estimation



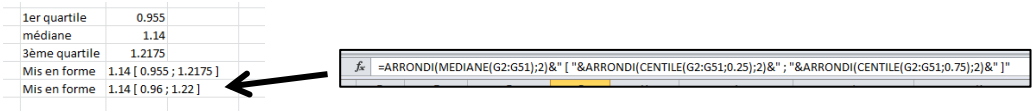
Cette présentation des résultats n'est pas extraordinairement esthétique… Je vous propose une syntaxe qui permet d'indiquer les informations dans une seule cellule, de telle façon à ensuite la copier et la coller dans un document Word : « médiane [1er quartile ; 3ème quartile] », en arrondissant à deux chiffres après la virgule, ce qui donnera ici : « 1,14 [0,96 ; 1,22] ». Pour cela, il faut utiliser les guillemets (double) et la commande « & ». Voici l'ensemble de ce qu'il faut taper (sans encore faire d'arrondi) :

« =MEDIANE(*plage*)&" ["&CENTILE(*plage*;0,25)&" ; "&CENTILE(*plage*;0,75)&"] " »

Voici ce que l'on obtient :



La fonction « ARRONDI(*valeur*;*X*) » permet d'arrondir à X chiffres après la virgule les nombres. Cela donne la formule et le résultat suivants, en souhaitant arrondir à deux chiffres après la virgule :

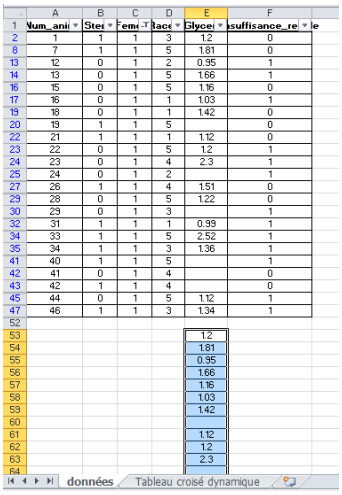


Si l'on souhaitait connaître la médiane de la glycémie seulement chez les femelles, il faut effectuer la démarche (fastidieuse, mais obligatoire car les TCD ne permettent pas d'obtenir des médianes et des quartiles) présentée ci-dessous.

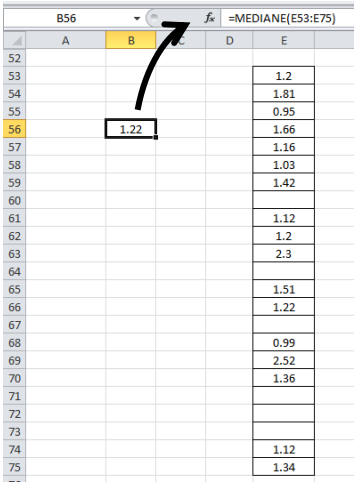
(1) Utilisez les filtres et sélectionnez les chiens ayant une valeur égale à « 1 » pour la variable « Femelle ».

Num_anim	Steril	Feme	Race	Glycem	Insuffisance_rena
1	1	1	3	1.2	0
7	1	1	5	1.81	0
12	0	1	2	0.95	1
13	0	1	5	1.66	1
15	0	1	5	1.16	0
16	0	1	1	1.03	1
18	0	1	1	1.42	0
19	1	1	5		0
21	1	1	1	1.12	0
22	0	1	5	1.2	1
23	0	1	4	2.3	1
24	0	1	2		1
26	1	1	4	1.51	0
28	0	1	5	1.22	0
29	0	1	3		1
31	1	1	1	0.99	1
33	1	1	5	2.52	1
34	1	1	3	1.36	1
40	1	1	5		1
41	0	1	4		0
42	1	1	4		0
44	0	1	5	1.12	1
46	1	1	3	1.34	1

(2) Sélectionnez la plage des valeurs de la colonne « Glycemie » et collez-la *sous* le tableau.



(3) Dans une cellule de votre choix (cellule B56 ci-dessous), tapez la formule « =MEDIANE(*plage*) » en spécifiant la plage que vous venez de copier-coller (ici, la plage E53:E75).

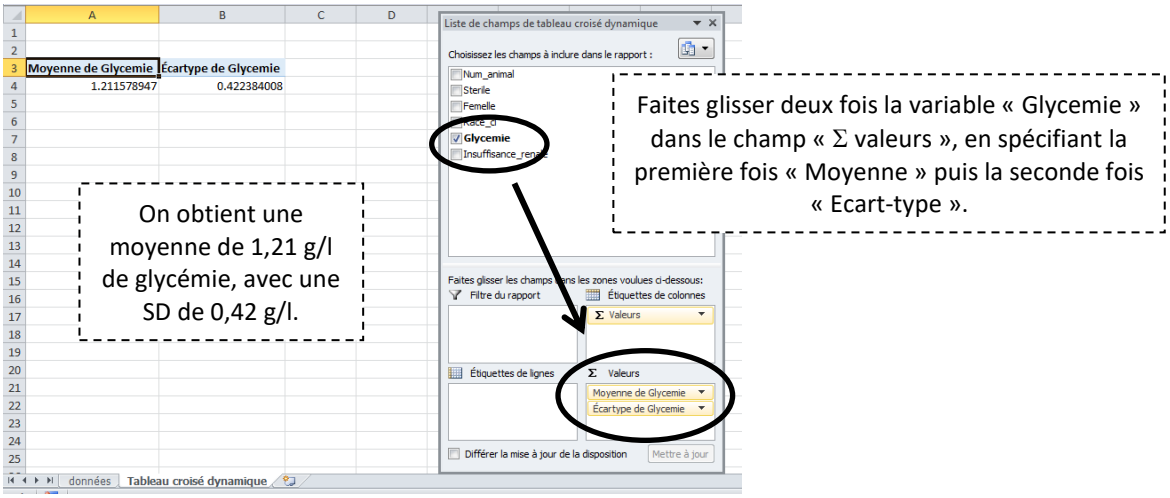


La médiane de la glycémie parmi les femelles de l'échantillon est égale à 1,22 g/l.

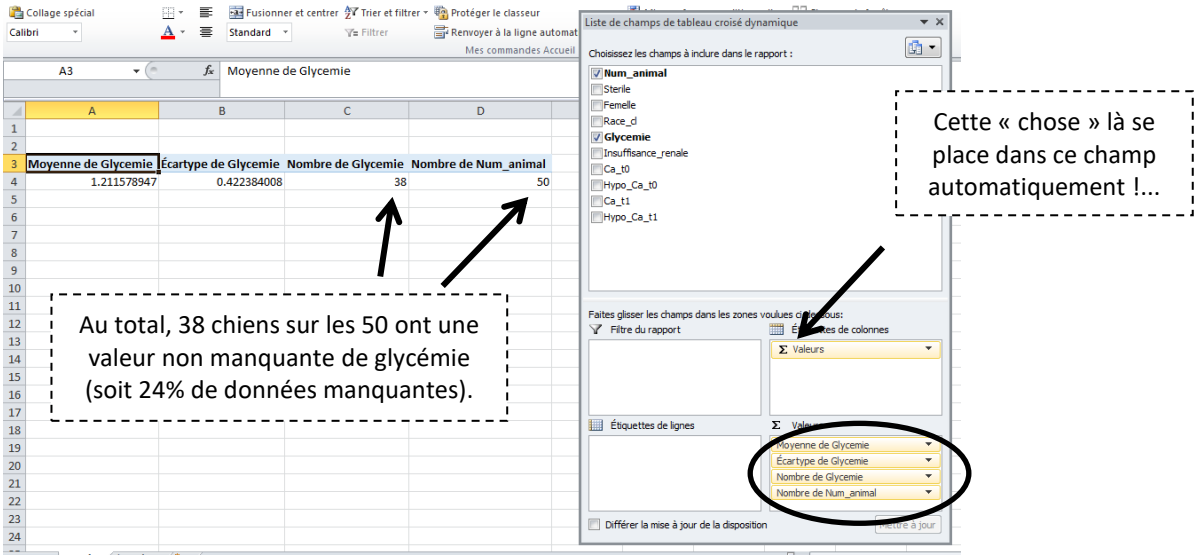
Vous pouvez remarquer qu'il y a quelques données manquantes sur la variable « Glycemie ». Excel ne les prend pas en compte dans le calcul de la médiane. Donc, cette valeur de 1,22 ne porte que sur les chiens qui ont une valeur de glycémie.

3. *Utilisation des tableaux croisés dynamiques*

On utilisera les TCD dès que l'on souhaitera fournir les indicateurs suivants : minimum, maximum, moyenne, et SD. Supposons que nous souhaitions fournir la moyenne et la SD de la variable « Glycemie » (« écart-type de glycémie » ci-dessous). Rien de plus simple avec les TCD ! Regardez…



Il peut être important de savoir sur combien de valeurs cette moyenne et cette SD ont été calculées. Pour cela, il faut faire glisser la variable « Glycemie » dans le champ « Σ valeurs » en spécifiant « Nombre ». En faisant glisser la variable « Num_animal » dans le champ « Σ valeurs » en spécifiant « Nombre », cela permet de mettre en évidence la proportion de données manquantes sur la variable « Glycemie ». Voici ce que l'on obtient :



III. Associations statistiques entre deux variables (données indépendantes)

A. Introduction

Les données sont dites « indépendantes » si elles proviennent d'individus indépendants. Des individus peuvent être considérés comme indépendants si la valeur d'une variable pour un individu donné ne dépend *a priori* pas de la valeur de cette même variable pour un autre individu de l'échantillon. Par exemple, si l'on sélectionne dans l'échantillon deux vaches d'un même élevage laitier, il y a des chances pour que ces deux vaches issues du même élevage aient des valeurs de moyennes de la production laitière plus proches que celles de deux vaches issues de deux élevages différents. Les deux vaches

d'un même élevage ne peuvent donc pas être considérées comme indépendantes du point de vue de la production laitière. Dans tout ce qui suit dans cette partie III, je ferai l'hypothèse que les individus sont indépendants, donc que l'on travaillera sur des données indépendantes. La partie IV de ce tutoriel traite une situation courante de non indépendance des données.

Le tableau ci-dessous présente les tests statistiques usuels lorsque les individus sont indépendants.

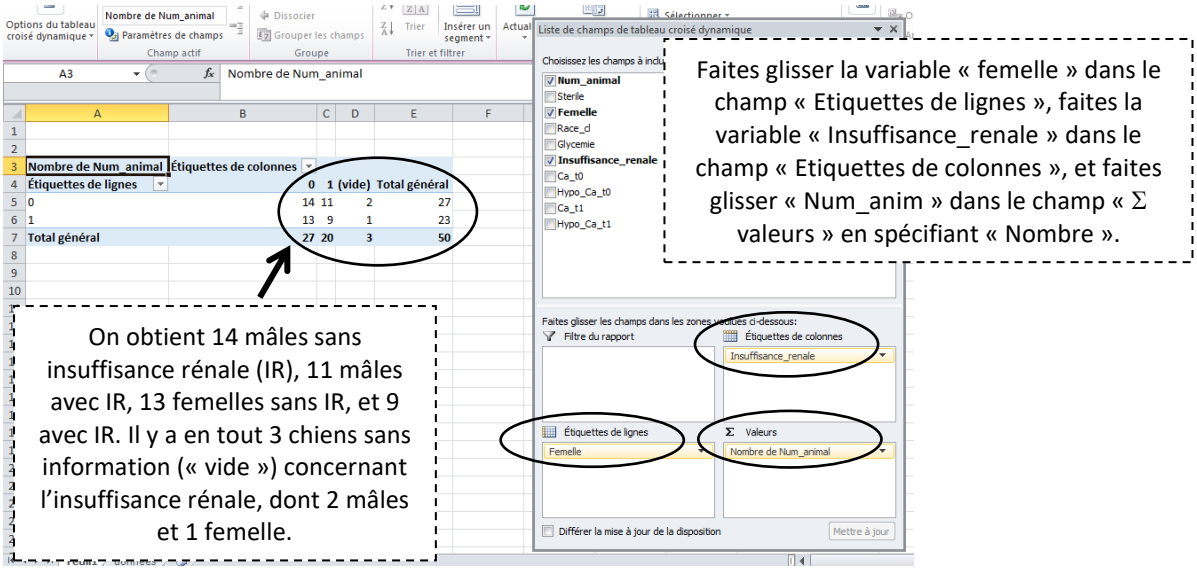
<table>
<tr><th>Tester l'association entre…</th><th>Nom du test statistique</th><th>Que fait le test ? (Indicateurs comparés)</th></tr>
<tr><td>2 variables binaires</td><td rowspan="2">Chi-2, Fisher*</td><td>Compare 2 %</td></tr>
<tr><td>1 variable binaire x 1 variable qualitative</td><td>Compare ≥ 3 %</td></tr>
<tr><td>2 variables qualitatives</td><td colspan="2">Résultats ininterprétables ⇒ il faut transformer une des deux variables en une variable binaire</td></tr>
<tr><td rowspan="2">1 variable binaire x 1 variable quantitative</td><td>Student</td><td>Compare 2 moyennes</td></tr>
<tr><td>Mann-Whitney**</td><td>Compare 2 médianes*</td></tr>
<tr><td rowspan="2">1 variable qualitative x 1 variable quantitative</td><td>ANOVA</td><td>Compare ≥ 3 moyennes</td></tr>
<tr><td>Kruskal-Wallis**</td><td>Compare ≥ 3 médianes*</td></tr>
<tr><td rowspan="2">2 variables quantitatives</td><td>Coefficient de corrélation de Pearson</td><td rowspan="2">Calcule un coefficient de corrélation (pas d'indicateurs comparés)</td></tr>
<tr><td>Coefficient de corrélation de Spearman***</td></tr>
</table>

* Test de Fisher à utiliser à la place du test du Chi-2 si au moins un des effectifs attendus est < 5
** Test « non paramétrique », à utiliser si la distribution de la variable quantitative ne suit pas une loi normale
*** Coefficient de corrélation « non paramétrique », à utiliser si la distribution d'au moins l'une des deux variables quantitatives ne suit pas une loi normale

B. Croisement de deux variables qualitatives (en classes ou binaires)

Supposons par exemple que nous souhaitions savoir s'il existe une association statistique entre le sexe du chien (variable « Femelle ») et la présence d'insuffisance rénale (IR ; variable « Insuffisance_renale »). Ces deux variables étant des variables qualitatives (elles sont même binaires toutes les deux), il faut comparer des pourcentages avec le test du Chi-2 ou de Fisher. Pour répondre à la question, trois étapes sont nécessaires.

1) La 1ère étape consiste à obtenir les effectifs issus du croisement entre les deux variables. Pour cela, utilisons les TCD. Après avoir nettoyé votre TCD, voici ce que vous devez obtenir :



2) La 2ème étape consiste à calculer les pourcentages qui vont être comparés puis testés. On veut comparer le % d'IR parmi les 25 mâles avec une information sur l'insuffisance rénale au % d'IR parmi les 22 femelles avec une information sur l'insuffisance rénale. Ces % sont respectivement de 11/25=44% et 9/22=41%[6]. Le test du Chi-2 que nous allons effectuer va tester la différence entre le % d'IR parmi les mâles (44%) et le % d'IR parmi les femelles (41%).

3) La 3ème étape consiste effectuer le test statistique sur le site Internet BiostaTGV (http://marne.u707.jussieu.fr/biostatgv/?module=tests) et de cliquer sur l'un des 4 « Chi-2 » (la page qui va apparaître sera identique quel que soit le lien sur lequel vous cliquez) :

[6] La somme de ces deux pourcentages ne doit bien évidemment pas faire 100% !!

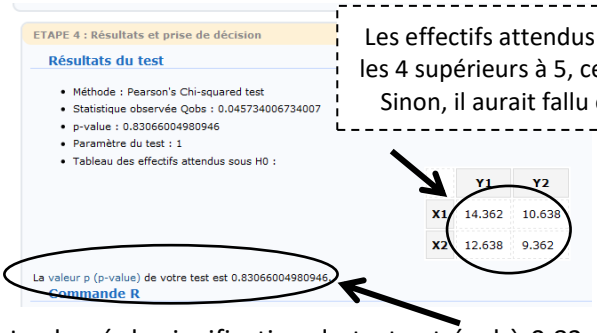
La fenêtre ci-dessous apparaît :

Dans la mesure où les variables « Femelle » et « Insuffisance_renale » sont toutes les deux binaires, vous devez taper le chiffre « 2 » pour le nombre de modalités, puis cliquer sur « Envoyer ».

Vous devez ensuite remplir le tableau à 4 cases qui sont les effectifs fournis par le TCD précédemment.

Vous devez enfin cliquer sur « Faire le test ».

Vous obtenez alors les résultats suivants :



Les effectifs attendus sont calculés, ce qui permet de vérifier qu'ils sont tous les 4 supérieurs à 5, ce qui est une des conditions de validité du test du Chi-2. Sinon, il aurait fallu effectuer le test exact de Fisher, accessible sur le site.

Le degré de signification du test est égal à 0,83, supérieur au risque d'erreur α égal à 0,05, donc les deux % d'IR comparés (44% et 41%, respectivement parmi les mâles et parmi les femelles) ne sont pas significativement différents.

C. Croisement d'une variable qualitative avec une variable quantitative

1. *Remarques introductives*

Ce type de croisement entraîne différents tests statistiques selon la nature de la variable qualitative (binaire ou en classes) et selon la distribution de la variable quantitative.

Voici en résumé les 4 tests statistiques en fonction des différentes situations rencontrées :

		Distribution de la variable quantitative	
		Normale	Non normale
Type de la variable qualitative	Binaire	Comparaison de 2 moyennes à l'aide du test de Student	Comparaison de 2 médianes à l'aide du test de Mann-Whitney
	≥ 3 classes	Comparaison de 3 moyennes ou plus à l'aide du test de l'ANOVA[7]	Comparaison de 3 médianes ou plus à l'aide du test de Kruskal-Wallis

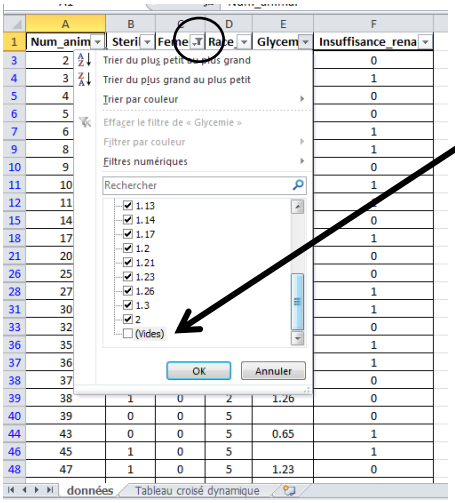
[7] ANOVA = ANalyse Of VAriance (analyse de la variance en français)

Je ne vais présenter dans ce tutoriel que les tests de Student et de Mann-Whitney. Mais l'utilisation du test de l'ANOVA rejoint celle du test de Student avec l'utilisation des TCD et du site Internet BiostaTGV, et le test du Kruskal-Wallis rejoint celle du test de Mann-Whitney avec l'utilisation des filtres dans Excel et du site Internet BiostaTGV.

2. *Comparaison de deux moyennes avec le test de Student*

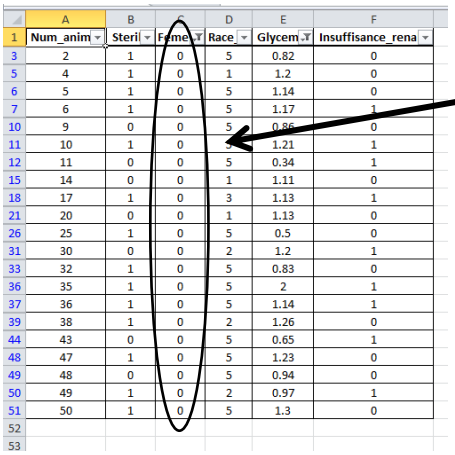
Nous allons supposer tout d'abord que la variable « Glycemie » suit une loi normale. Supposons maintenant que nous souhaitions comparer la moyenne de la glycémie parmi les chiens mâles à la moyenne de la glycémie parmi les chiens femelles. Pour répondre à la question, trois étapes sont nécessaires.

1) La 1ère étape est d'obtenir les valeurs des moyennes et des variances (dans les échantillons) dans chacun des deux groupes. Les TCD pourraient être utilisés, mais je préfère vous faire utiliser la fonction filtre d'Excel (de plus, cette méthode est identique à celle que je présenterai pour comparer les médianes). Il faut commencer par créer deux colonnes de données sur la glycémie : une colonne des valeurs de la glycémie chez les mâles seulement (en excluant donc les mâles sans valeur de glycémie), et une colonne des valeurs de la glycémie pour les femelles seulement. Pour cela, nous allons utiliser la fonction FILTRE d'Excel.



Après avoir filtré les valeurs de la variable « Femelles » en ne sélectionnant que les valeurs « 0 » (pour sélectionner les mâles), vous cliquez sur le filtre de la variable « Glycemie » en décochant la case « (Vide) » pour retirer de la sélection les données manquantes de glycémie.

Voici ce que vous obtenez :



	A	B	C	D	E	F
1	Num_anim	Steri	Feme	Race	Glycem	Insuffisance_rena
3	2	1	0	5	0.82	0
5	4	1	0	1	1.2	0
6	5	1	0	5	1.14	0
7	6	1	0	5	1.17	1
10	9	0	0	5	0.86	0
11	10	1	0	5	1.21	1
12	11	0	0	5	0.34	1
15	14	0	0	1	1.11	0
18	17	1	0	3	1.13	1
21	20	0	0	1	1.13	0
26	25	1	0	5	0.5	0
31	30	0	0	2	1.2	1
33	32	1	0	5	0.83	0
36	35	1	0	5	2	1
37	36	1	0	5	1.14	1
39	38	1	0	2	1.26	0
44	43	0	0	5	0.65	1
48	47	1	0	5	1.23	0
49	48	0	0	5	0.94	0
50	49	1	0	2	0.97	1
51	50	1	0	5	1.3	0
52						
53						

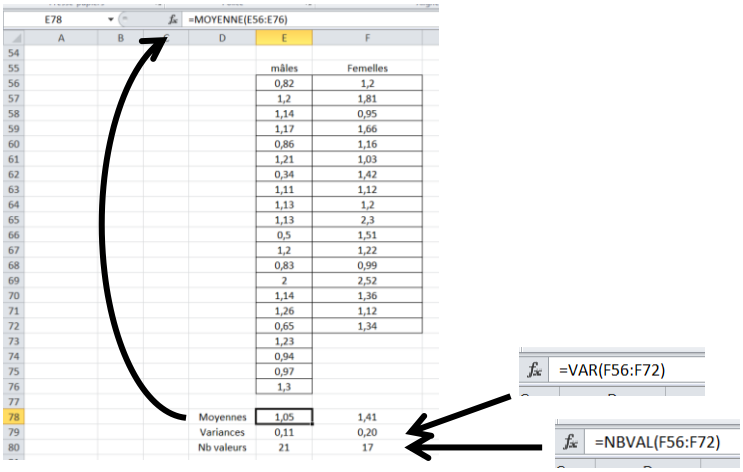
Nous n'avons bien que des chiens mâles sélectionnés.

Ensuite, vous copiez la plage des valeurs et vous les collez *sous* l'ensemble du tableau de données, ou dans un autre onglet, mais surtout pas *à côté* (à droite par exemple) du tableau de données (en effet, si vous faites cela, et comme les lignes sont déjà sélectionnées avec le filtre, vous allez les coller avec des cellules vides entre les lignes). Et vous faites exactement la même procédure pour les chiens

femelles. Pour ne pas vous mélanger les pinceaux, n'hésitez pas à écrire qui sont les mâles et qui sont les femelles, au-dessus des plages que vous venez de coller :

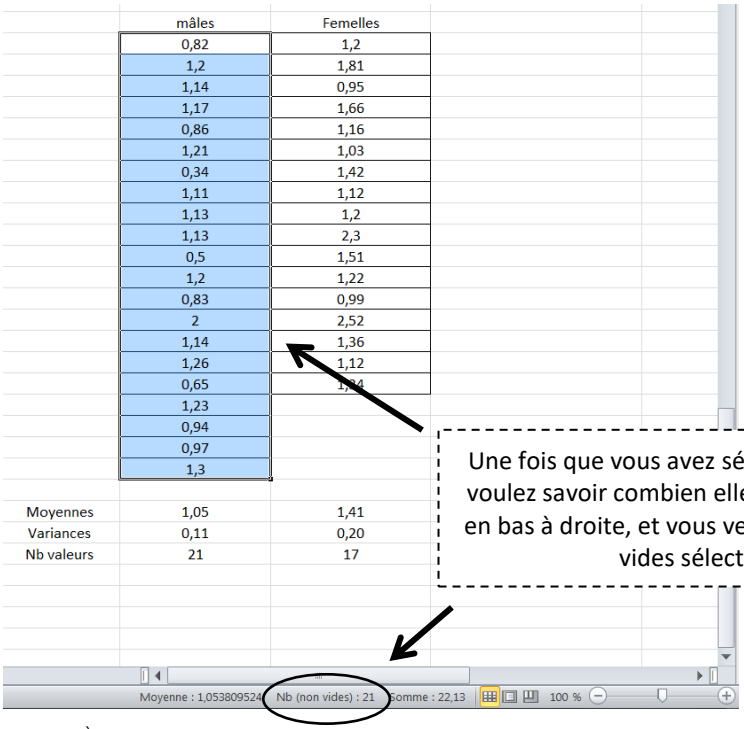
mâles	femelles
0.82	1.2
1.2	1.81
1.14	0.95
1.17	1.66
0.86	1.16
1.21	1.03
0.34	1.42
1.11	1.12
1.13	1.2
1.13	2.3
0.5	1.51
1.2	1.22
0.83	0.99
2	2.52
1.14	1.36
1.26	1.12
0.65	1.34
1.23	
0.94	
0.97	
1.3	

2) La 2ème étape consiste à calculer la moyenne, la variance, et le nombre de valeurs de glycémie qui ont été utilisées pour calculer chacune des deux moyennes. Comme vous pouvez le voir ci-dessous, les moyennes de glycémie sont de 1,05 g/L et 1,41 g/L respectivement chez les mâles (n=21) et les femelles (n=17). Vous pouvez remarquer que le rapport des variances 0,20/0,11 est compris entre 1/3 et 3, donc elles peuvent être considérées comme voisines. Ainsi, le test de Student pour séries non appariées classique sera applicable (quand les variances ne peuvent pas être considérées comme voisines, il faut utiliser un test de Student approché, tel que le test de Welch, que le site Internet BiostaTGV permet de faire !). Le test de Student que nous allons effectuer va donc tester la différence entre la moyenne de la glycémie chez les mâles (1,05 g/l) et la moyenne de la glycémie chez les femelles (1,41 g/l).



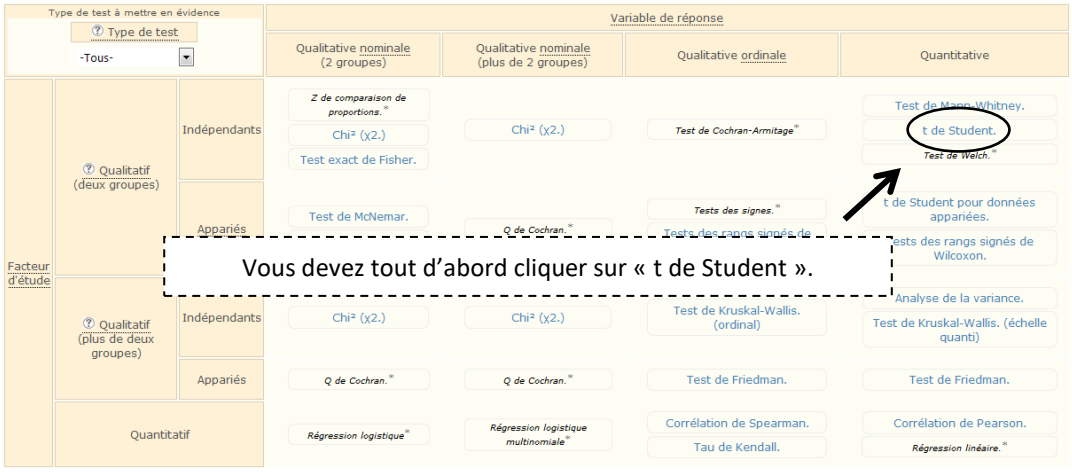
	mâles	Femelles
	0,82	1,2
	1,2	1,81
	1,14	0,95
	1,17	1,66
	0,86	1,16
	1,21	1,03
	0,34	1,42
	1,11	1,12
	1,13	1,2
	1,13	2,3
	0,5	1,51
	1,2	1,22
	0,83	0,99
	2	2,52
	1,14	1,36
	1,26	1,12
	0,65	1,34
	1,23	
	0,94	
	0,97	
	1,3	
Moyennes	1,05	1,41
Variances	0,11	0,20
Nb valeurs	21	17

Remarquons qu'il y a un autre moyen pour savoir combien de valeurs contient une colonne :



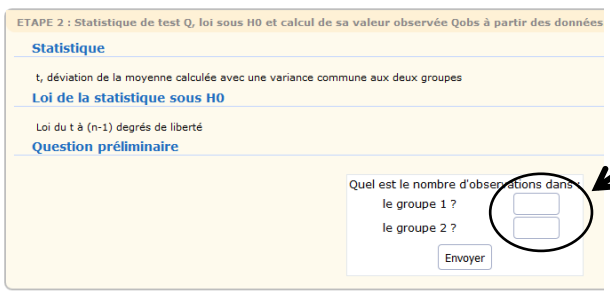
Une fois que vous avez sélectionné une plage dont vous voulez savoir combien elle contient de valeurs, regardez en bas à droite, et vous verrez le nombre de cellules non vides sélectionnées (ici, 21).

3) La 3ème étape se déroule sur le site Internet BiostaTGV. Le principe est de copier chacune des deux colonnes ci-dessus et de les coller dans les cases prévues par le site Internet. Ensuite, il suffit de cliquer sur un bouton pour effectuer le test statistique de Student.



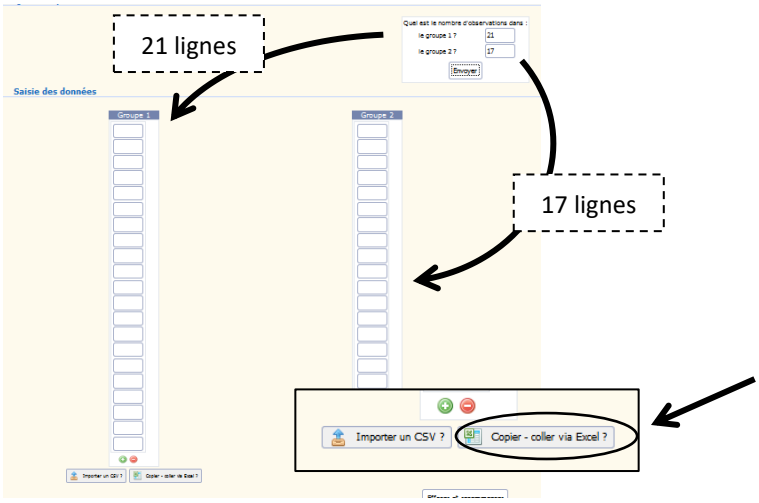
Vous devez tout d'abord cliquer sur « t de Student ».

Vous obtenez la fenêtre suivante :

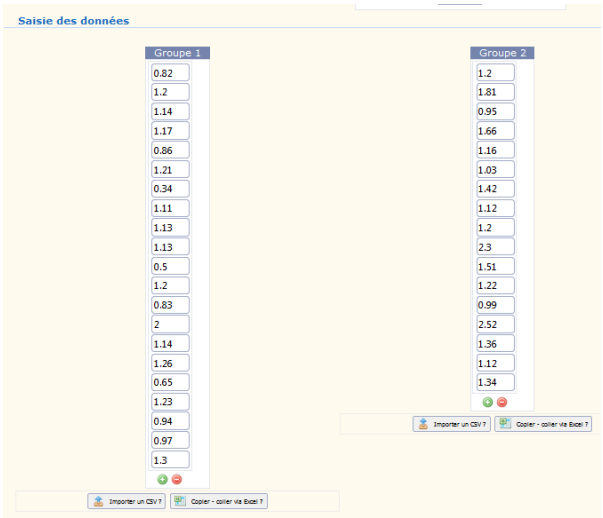


Vous devez taper le nombre de lignes que vous allez copier dans chacune des deux colonnes (les « groupes 1 et 2 » sur le site). Vous avez cette information grâce à la formule NBVAL que vous aviez tapée ci-dessus : 21 et 17 respectivement pour les mâles et les femelles.

Vous entrez les nombres d'observations dans chacun des deux groupes (21 pour les mâles et 17 pour les femelles), et vous obtenez ceci :

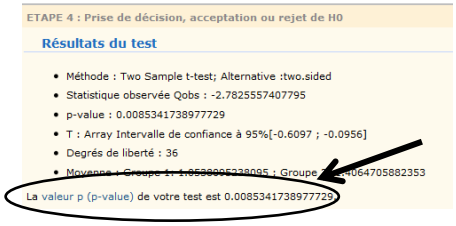


Vous copiez ensuite chaque colonne d'Excel, vous retournez sur le site, vous cliquez ensuite sur « Copier-coller via Excel ? », et vous collez (CTRL+V) ce que vous venez de copier dans le champ, puis vous cliquez enfin sur « Générer ». Voici *in fine* ce que vous obtenez :



Juste en-dessous, vous avez la partie « Options du test » :

Cochez ici cette case car nous avons vu que les variances pouvaient être considérées comme voisines. Laissez la puce « Bilatérale » cochée, et cliquez sur « Faire le test ».



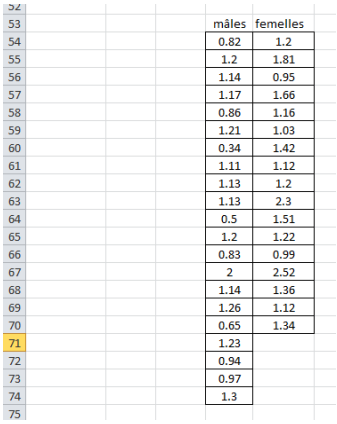
Le degré de signification du test est égal à 0,009, inférieur au risque d'erreur α égal à 0,05, donc les deux moyennes comparées (1,05 g/l et 1,41 g/l, respectivement pour les mâles et les femelles) sont significativement différentes.

3. *Comparaison de deux médianes avec le test de Mann-Whitney*

Nous allons maintenant supposer que la variable « Glycemie » ne suit pas une loi normale. Les moyennes de la glycémie entre différents groupes ne peuvent désormais plus être comparées. Il faut comparer des médianes de glycémie. Supposons donc que nous souhaitions comparer la médiane de la glycémie parmi les chiens mâles à la médiane de la glycémie parmi les chiens femelles. Pour répondre à la question, trois étapes sont nécessaires.

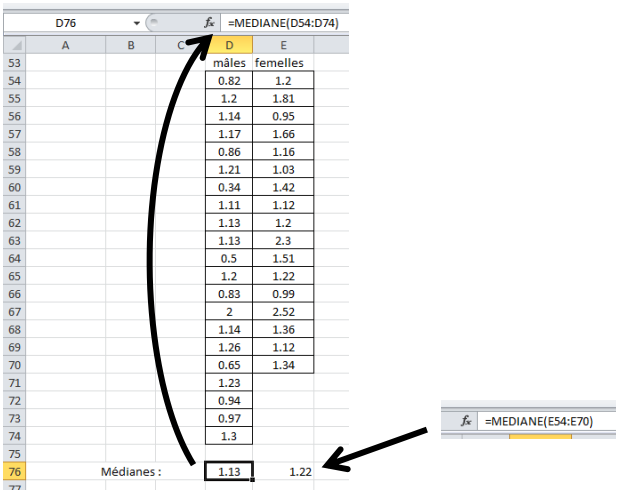
1) La démarche est identique à celle précédente pour la comparaison de moyennes. Il faudra faire apparaître les médianes avec la formule Excel MEDIANE(*plage*). Mais il faut commencer par utiliser la fonction filtre, copier les cellules de la glycémie chez les mâles, puis chez les femelles. Pour cela, il suffit de faire exactement ce que vous avez fait dans la 1ère étape pour la comparaison de deux moyennes ! (Je vous laisse relire cette étape.)

Voici ce que vous obtenez :

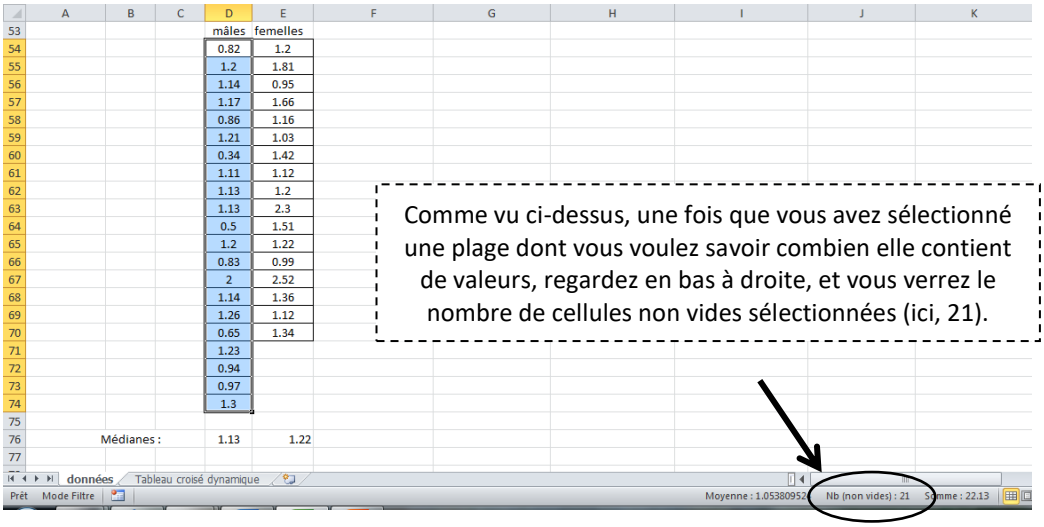


	mâles	femelles
54	0.82	1.2
55	1.2	1.81
56	1.14	0.95
57	1.17	1.66
58	0.86	1.16
59	1.21	1.03
60	0.34	1.42
61	1.11	1.12
62	1.13	1.2
63	1.13	2.3
64	0.5	1.51
65	1.2	1.22
66	0.83	0.99
67	2	2.52
68	1.14	1.36
69	1.26	1.12
70	0.65	1.34
71	1.23	
72	0.94	
73	0.97	
74	1.3	

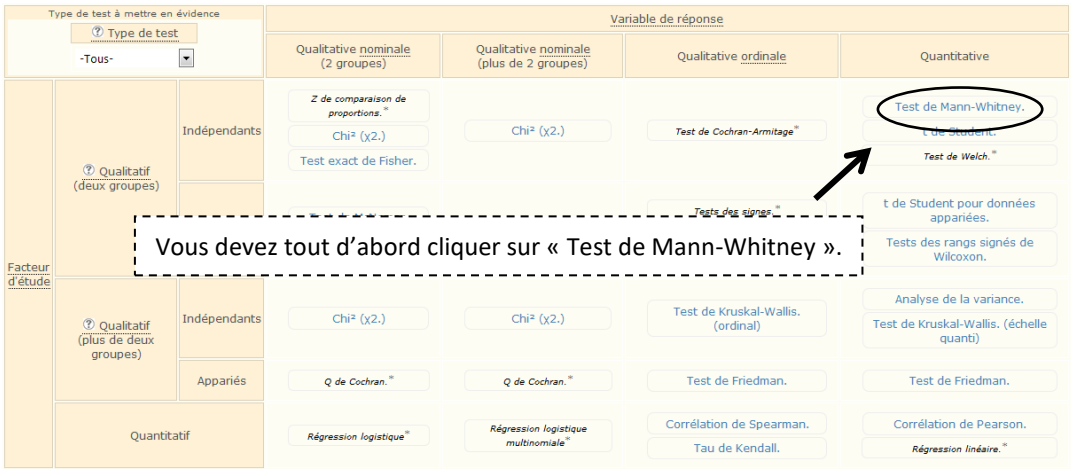
2) Sous chacune des deux colonnes, tapez la formule de la médiane en spécifiant les plages des mâles et des femelles respectives :



Le test statistique de Mann-Whitney va donc tester la différence entre la médiane de la glycémie parmi les mâles (1,13 g/l) et la médiane de la glycémie parmi les femelles (1,22 g/l). Ici, je n'ai pas remis la formule NBVAL pour connaître le nombre de valeurs de la glycémie :

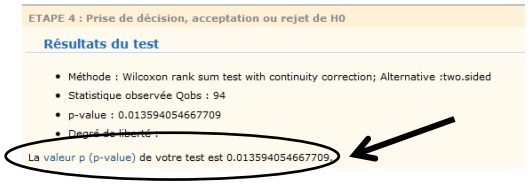


3) La 3ème étape se déroule sur le site Internet BiostaTGV. Le principe est exactement le même que pour la comparaison de moyennes avec le test de Student : copier chacune des deux colonnes de glycémie et les coller dans les cases prévues par le site Internet. Ensuite, il suffit de cliquer sur un bouton pour effectuer le test statistique de Mann-Whitney.



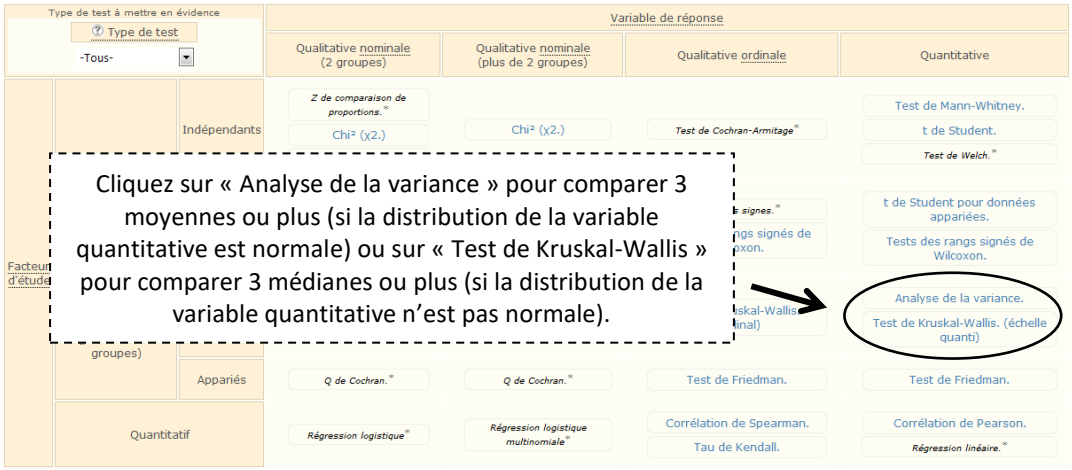
Je ne vais vous remontrer comment faire pour copier-coller les colonnes d'Excel vers le site, n'est-ce pas ?! Au besoin, relisez cette démarche ci-dessus ;-)…

Une fois que vous avez collé (généré) vos colonnes (21 lignes pour les mâles et 17 lignes pour les femelles), vous cliquez sur « Faire le test » et voici ce que vous obtenez :



Le degré de signification du test est égal à 0,01, inférieur au risque d'erreur α égal à 0,05, donc les deux médianes comparées (1,13 g/l et 1,22 g/l, respectivement pour les mâles et les femelles) sont significativement différentes.

Maintenant, pour comparer plusieurs moyennes ou plusieurs médianes, il faut reprendre les démarches ci-dessus, et cliquer sur les liens ci-dessous :



D. Croisement de deux variables quantitatives

L'objectif est de montrer que deux variables quantitatives sont associées entre elles. Pour cela, il faut calculer un coefficient de corrélation. Si les deux variables quantitatives suivent une loi normale, le coefficient de corrélation qu'il faut calculer est celui de Pearson ; si ce n'est pas le cas, il faut calculer le coefficient de corrélation de Spearman (coefficient de corrélation dit « non paramétrique »).

Pour illustrer le calcul du coefficient de corrélation, nous allons calculer le coefficient de corrélation entre la concentration en glycémie (variable « Glycemie » dans le fichier de données) et celle en calcium (variable « Ca_t0 » dans le fichier de données). Nous ferons l'hypothèse que ces deux variables suivent une loi normale. Ainsi, je vais présenter le calcul du coefficient de corrélation de Pearson avec BiostaTGV, mais la procédure est identique s'il fallait calculer un coefficient de corrélation de Spearman.

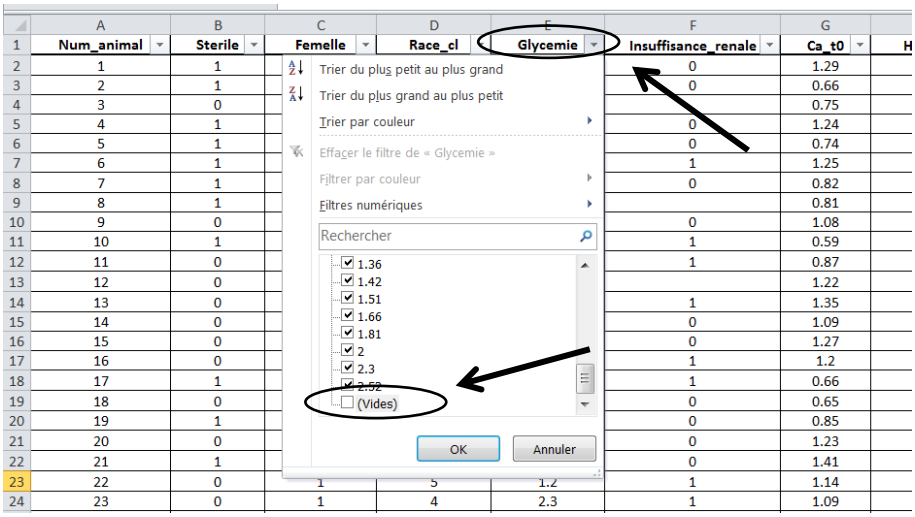
Pour calculer un coefficient de corrélation, il faut 3 étapes.

1) Ne sélectionner que les individus pour lesquels les données sont présentes pour les deux variables, et donc exclure ceux pour lesquels les données manquent pour au moins une des deux variables. Regardez ci-dessous,

	A	B	C	D	E	F	G
1	Num_animal	Sterile	Femelle	Race_cl	Glycemie	Insuffisance_renale	Ca_t0
2	1	1	1	3	1.2	0	1.29
3	2	1	0	5	0.82	0	0.66
4	3	0	0	2		0	0.75
5	4	1	0	1	1.2	0	1.24
6	5	1	0	5	1.14	0	0.74
7	6	1	0	5	1.13	1	1.25
8	7						0.82
9	8						0.81
10	9						1.08
11	10						0.59
12	11						0.87
13	12						1.22
14	13						1.35
15	14						1.09
16	15						1.27
17	16	0	1	1	1.03	1	1.2
18	17	1	0	3	1.13	1	0.66

Le chien n°3 a une donnée manquante pour la glycémie, alors que ce n'est pas le cas pour le calcium. Par conséquent, le chien n°3 ne doit pas faire partie du calcul du coefficient de corrélation

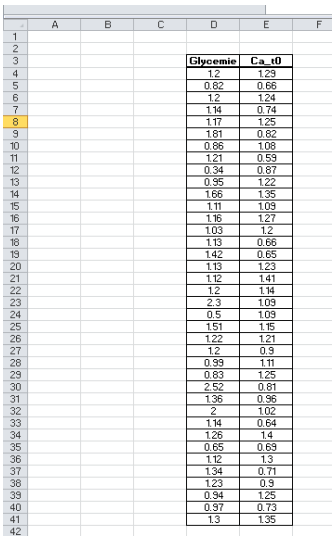
Pour sélectionner les individus dont la donnée ne manque pour aucune des deux variables dont on souhaite calculer le coefficient de corrélation, il faut utiliser les filtres en décochant pour les deux variables la case « (Vides) ». On commence par la première des deux variables :



	A	B	C	D	E	F	G
1	Num_animal	Sterile	Femelle	Race_cl	Glycemie	Insuffisance_renale	Ca_t0
2	1	1				0	1.29
3	2	1				0	0.66
4	3	0					0.75
5	4	1				0	1.24
6	5	1				0	0.74
7	6	1				1	1.25
8	7	1				0	0.82
9	8	1					0.81
10	9	0				0	1.08
11	10	1				1	0.59
12	11	0				1	0.87
13	12	0					1.22
14	13	0				1	1.35
15	14	0				0	1.09
16	15	0				0	1.27
17	16	0				1	1.2
18	17	1				1	0.66
19	18	0				0	0.65
20	19	1				0	0.85
21	20	0				0	1.23
22	21	1				0	1.41
23	22	0	1	5	1.2	1	1.14
24	23	0	1	4	2.3	1	1.09

Puis on fait de même pour la seconde si c'est nécessaire (ici, il y avait des données manquantes pour la glycémie, mais pas pour le calcium). Il est bien entendu possible qu'il n'y ait aucune donnée manquante pour des deux variables. Auquel cas, cette première étape n'est pas pertinente.

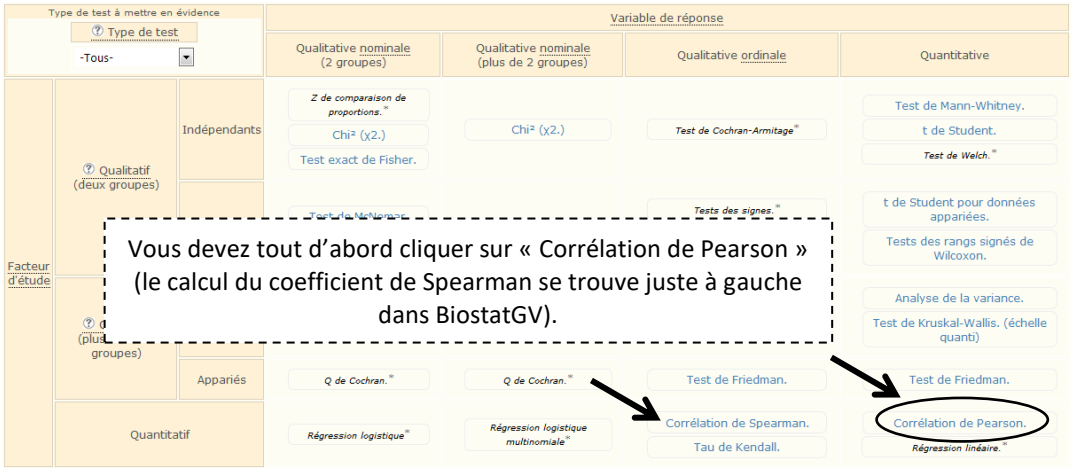
2) Une fois que les individus à exclure (à cause de données manquantes) l'ont été, il faut sélectionner les deux plages de valeurs, les copier, et les coller dans un autre onglet. Voici ce que l'on obtient :



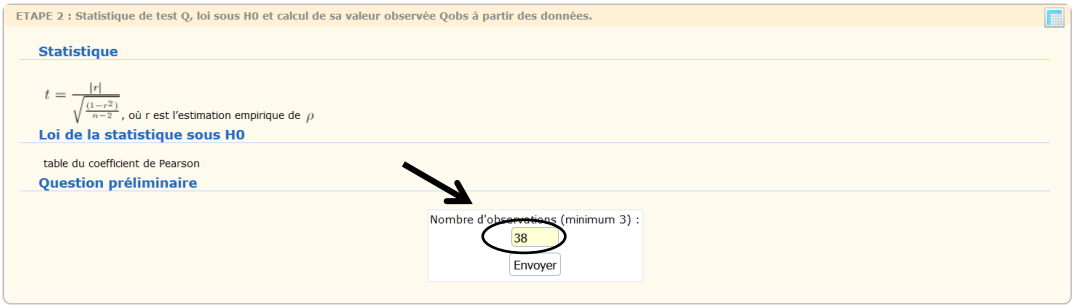
Glycemie	Ca_t0
1.2	1.29
0.82	0.66
1.2	1.24
1.14	0.74
1.17	1.25
1.81	0.82
0.86	1.08
1.21	0.59
0.34	0.87
0.95	1.22
1.66	1.35
1.11	1.09
1.16	1.27
1.03	1.2
1.13	0.66
1.42	0.65
1.13	1.23
1.12	1.41
1.2	1.14
2.3	1.09
0.5	1.09
1.51	1.15
1.22	1.21
1.2	0.9
0.99	1.11
0.83	1.25
2.52	0.81
1.36	0.96
2	1.02
1.14	0.64
1.26	1.4
0.65	0.63
1.12	1.3
1.34	0.71
1.23	0.9
0.94	1.25
0.97	0.73
1.3	1.35

Si l'on sélectionne l'ensemble des données de l'une ou l'autre variable, et en regardant en bas à droite de l'écran, on obtient le nombre de cellules non vides (cf. partie III.C.2 ci-dessus). Ici, ce nombre est de 38. Ainsi, 38 chiens ont des données pour la glycémie *et* pour le calcium.

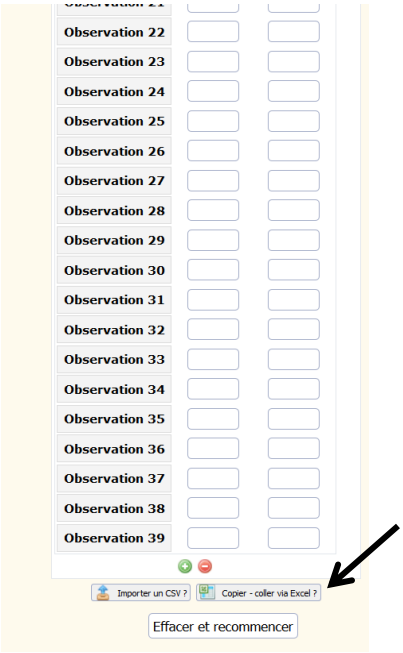
3) Il faut ensuite aller sur BiostaTGV, et sélectionner le calcul du coefficient de corrélation de Pearson :



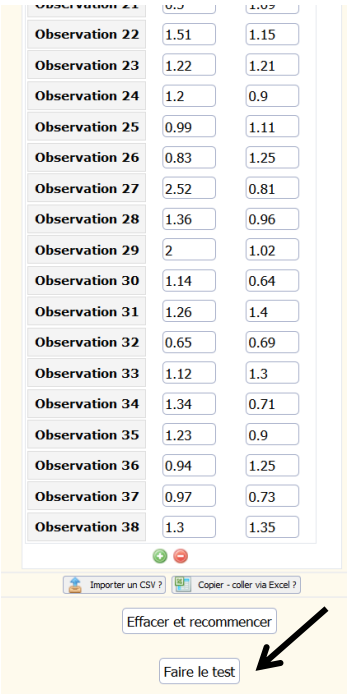
Ensuite, il faut entrer le nombre d'individus à partir desquels le coefficient de corrélation va être calculé (ici, 38 chiens) :



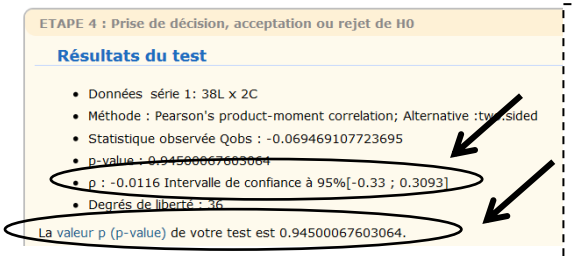
Ensuite, il faut copier la plage de valeurs pour la glycémie et le calcium et la coller dans BiostaTGV en utilisant « copier-coller via Excel » :



On obtient cela :

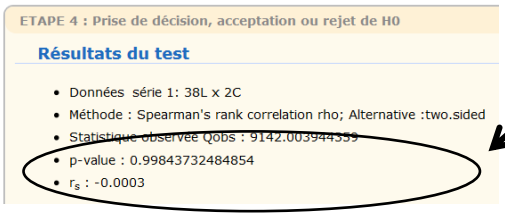


On clique ensuite sur « Faire le test », et on obtient cela :



BiostatGV fournit la valeur du coefficient de corrélation de Pearson (ρ) qui est de -0,01, assorti d'un degré de signification p = 0,94. Par conséquent, les variables « glycemie » et « Ca_t0 » ne sont pas significativement corrélées (car p > 0,05), avec une valeur du coefficient de corrélation très proche de zéro.

Si vous aviez souhaité calculer un coefficient de corrélation de Spearman, vous auriez obtenu le résultat suivant :



Le coefficient de corrélation de Spearman (r_S) vaut -0,0003, avec un degré de signification p = 1,00 (en arrondissant à deux chiffres après la virgule).

IV. Associations statistiques sur séries appariées

A. Introduction et présentation des données

On parle de « séries appariées » quand on travaille sur des (en général, deux) séries de valeurs qui proviennent d'un même individu, et ce, pour N individus. Il existe de nombreux exemples de séries appariées[8]. L'exemple que je vais utiliser dans cette partie est issu d'une situation très courante en médecine (vétérinaire) : on regarde l'évolution d'un paramètre quantitatif ou binaire suite à une intervention (initiation d'un traitement, ou autre), entre un temps t_0 (juste avant l'intervention) et un temps t_1. Les N animaux sont évalués à t_0, et ils le sont à nouveau à t_1. Il y a donc deux séries de valeurs par animal. Ces valeurs ne peuvent pas être considérées comme indépendantes, puisqu'elles sont issues d'un même animal à chaque fois ! Les tests statistiques classiques vus dans la partie III ne sont donc pas applicables.

Cette partie IV va utiliser le même fichier Excel « Pour tuto Excel stat descriptives.xlsx » qui contient les variables (colonnes) suivantes : Ca_t0, Hypo_Ca_t0, Ca_t1, Hypo_Ca_t1, représentant respectivement le taux de calcium à t_0, la présence d'une hypocalcémie à t_0 (en 0/1), le taux de calcium à t_1, et la présence d'une hypocalcémie à t_1 (en 0/1), où t_0 est la date d'initiation d'un traitement dont on craint qu'il réduise le taux de calcium circulant et t_1 la date un mois après t_0. Voici les 16 premières lignes du fichier de données :

	A	B	C	D	E	F	G	H	I	J
1	Num_animal	Sterile	Femelle	Race_cl	Glycemie	Insuffisance_renale	Ca_t0	Hypo_Ca_t0	Ca_t1	Hypo_Ca_t1
2	1	1	1	3	1.2	0	1.29	0	1.29	0
3	2	1	0	5	0.82	0	0.66	1	0.64	1
4	3	0	0	2			0.75	1	0.73	1
5	4	1	0	1	1.2	0	1.24	0	1.22	0
6	5	1	0	5	1.14	0	0.74	1	0.71	1
7	6	1	0	5	1.17	1	1.25	0	1.17	0
8	7	1	1	5	1.81	0	0.82	1	0.73	1
9	8	1	0	3			0.81	1	0.73	1
10	9	0	0	5	0.86	0	1.08	0	1.01	0
11	10	1	0	5	1.21	1	0.59	1	0.55	1
12	11	0	0	5	0.34	1	0.87	1	0.92	0
13	12	0	1	2	0.95		1.22	0	1.2	0
14	13	0	1	5	1.66	1	1.35	0	1.28	0
15	14	0	0	1	1.11	0	1.09	0	0.8	1
16	15	0	1	5	1.16	0	1.27	0	1.19	0

B. Séries appariées sur un paramètre binaire

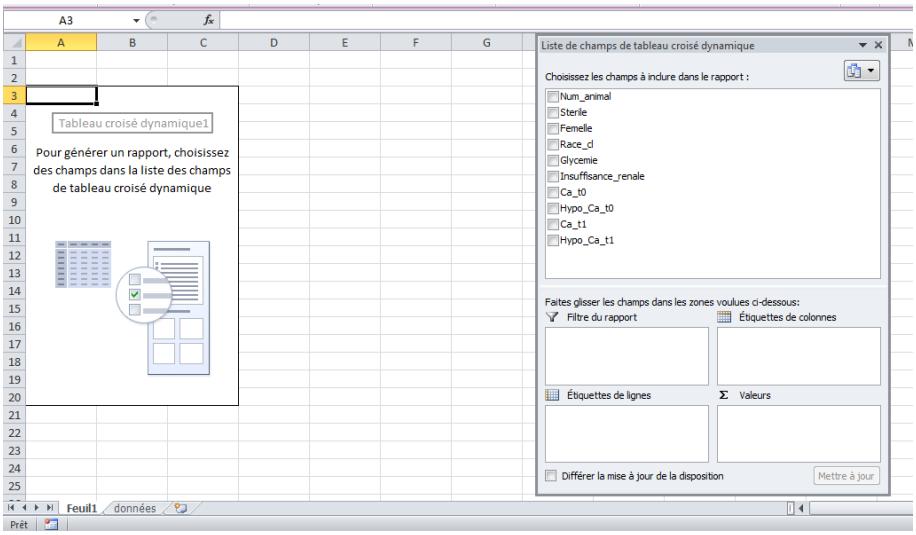
Dans cette partie, nous allons nous concentrer sur l'hypocalcémie qui est un paramètre binaire, puisqu'il est soit présent (les variables *Hypo_Ca_t0* et *Hypo_Ca_t1* valent alors « 1 »), soit absent (les variables *Hypo_Ca_t0* et *Hypo_Ca_t1* valent alors « 0 »).

La question est de savoir s'il existe une évolution significative de la présence d'hypocalcémie après un mois de traitement. On souhaite donc comparer le % de chiens hypocalcémiques à t_0 au % de chiens hypocalcémiques à t_1. Il est indispensable pour cela d'utiliser le test statistique de comparaison de % qui prend en compte le fait que l'on travaille sur les mêmes chiens, vus à deux temps différents. C'est le test statistique de McNemar[9] qu'il faut utiliser. Pour cela, 2 étapes sont nécessaires.

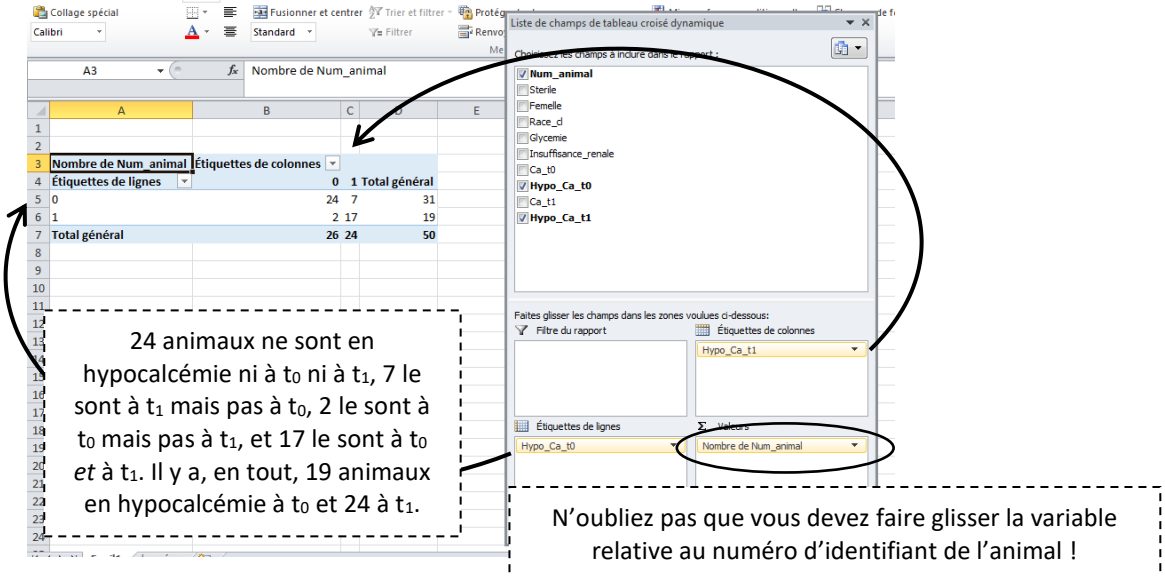
1) Il faut tout d'abord commencer par calculer ces % avec les TCD. Vous devez insérer un TCD (cf. https://www.youtube.com/watch?v=E5shJb7Zndk). Une fois que ce TCD est inséré, voici ce que vous obtenez :

[8] https://en.wikipedia.org/wiki/Paired_data

[9] Cf. ici : https://en.wikipedia.org/wiki/McNemar's_test



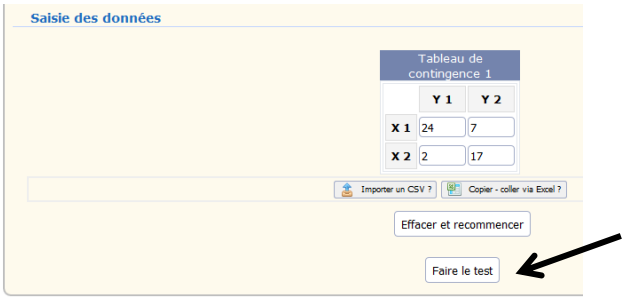
Vous devez ensuite croiser les deux variables *Hypo_Ca_t0* et *Hypo_Ca_t1* (cf. partie III.B ci-dessus). Voici ce que vous obtenez :

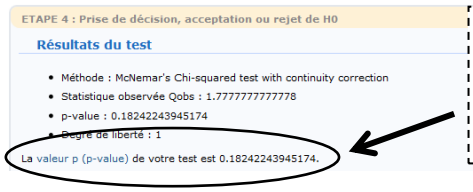


La question est de savoir s'il existe une évolution significative du % d'animaux en hypocalcémie entre t_0 et t_1. Le % d'animaux en hypocalcémie à t_0 est de 19/50, soit 38%, et celui à t_1 est de 24/50, soit 48%. La question est donc de savoir si 48% est significativement différent de 38%.

2) Il faut ensuite remplir le tableau à 4 cases des effectifs ci-dessus sur le site de BiostaTGV. Pour cela, vous devez cliquer sur « Test de McNemar » (cf. ci-dessous).

Il faut ensuite remplir le tableau à 4 cases qu'avait fourni le TCD, puis cliquer sur « Faire le test » :





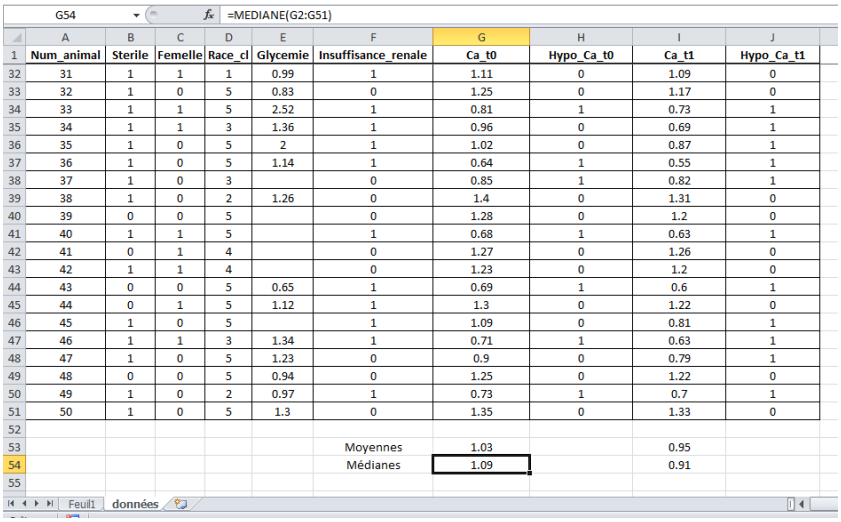
Le degré de signification du test est égal à 0,18, supérieur au risque d'erreur α égal à 0,05, donc les deux % comparés (38% et 48%, respectivement les % d'hypocalcémie à t_0 et à t_1) ne sont pas significativement différents.

C. Séries appariées sur un paramètre quantitatif

Dans cette partie, nous allons nous concentrer sur la calcémie qui est un paramètre quantitatif, exprimé ici en mmol/l. La question est ici de savoir si la moyenne (ou la médiane, dans le cas où la calcémie ne suivait pas une loi normale) de la calcémie à t_0 est significativement différente de la moyenne (ou la médiane) de la calcémie à t_1. Il est indispensable pour cela d'utiliser le test statistique de comparaison de moyennes (ou de médianes) qui prend en compte le fait que l'on travaille sur les mêmes chiens, vus à deux temps différents. Pour comparer deux moyennes provenant de séries appariées, il faut utiliser le test de Student pour séries appariées, et pour comparer deux médianes provenant de séries appariées, il faut utiliser le test de Wilcoxon pour séries appariées. Ces deux tests statistiques sont proposés par le site Internet BiostaTGV.

Pour cela, 3 étapes sont nécessaires.

1) Utiliser les formules « =MOYENNE(*plage*) » ou « =MEDIANE(*plage*) » dans chacune des deux colonnes (celle pour t_0 et celle pour t_1) pour obtenir respectivement les moyennes et médianes de la calcémie à t_0 et t_1. Voici ce que l'on obtient :



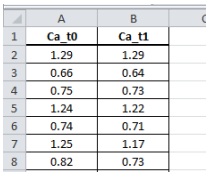
G54 : =MEDIANE(G2:G51)

	A	B	C	D	E	F	G	H	I	J
1	Num_animal	Sterile	Femelle	Race_cl	Glycemie	Insuffisance_renale	Ca_t0	Hypo_Ca_t0	Ca_t1	Hypo_Ca_t1
32	31	1	1	1	0.99	1	1.11	0	1.09	0
33	32	1	0	5	0.83	0	1.25	0	1.17	0
34	33	1	1	5	2.52	1	0.81	1	0.73	1
35	34	1	1	3	1.36	1	0.96	0	0.69	1
36	35	1	0	5	2	1	1.02	0	0.87	1
37	36	1	0	5	1.14	1	0.64	1	0.55	1
38	37	1	0	3		0	0.85	1	0.82	1
39	38	1	0	2	1.26	0	1.4	0	1.31	0
40	39	0	0	5		0	1.28	0	1.2	0
41	40	1	1	5		1	0.68	1	0.63	1
42	41	0	1	4		0	1.27	0	1.26	0
43	42	1	1	4		0	1.23	0	1.2	0
44	43	0	0	5	0.65	1	0.69	1	0.6	1
45	44	0	1	5	1.12	1	1.3	0	1.22	0
46	45	1	0	5		1	1.09	0	0.81	1
47	46	1	1	3	1.34	1	0.71	1	0.63	1
48	47	1	0	5	1.23	0	0.9	0	0.79	1
49	48	0	0	5	0.94	0	1.25	0	1.22	0
50	49	1	0	2	0.97	1	0.73	1	0.7	1
51	50	1	0	5	1.3	0	1.35	0	1.33	0
52										
53						Moyennes	1.03		0.95	
54						Médianes	1.09		0.91	
55										

Feuil1 | données

La moyenne de la calcémie est de 1,03 mmol/l à t_0, et de 0,95 mmol/l à t_1 ; la médiane de la calcémie est de 1,09 mmol/l à t_0 et de 0,91 à t_1.

2) Il faut ensuite copier-coller chacune de vos deux colonnes sur le site Internet, comme vous l'avez fait dans la 3ème étape de la partie III.C.2. Sauf que dans le cas des séries appariées, le site Internet requière le fait que les deux colonnes soient adjacentes dans Excel, ce qui n'est pas le cas dans le fichier de données ☹. Vous devez donc, dans une autre feuille de calcul Excel, coller les deux colonnes de la calcémie (celle à t_0 et celle à t_1) l'une à côté de l'autre (cf. ci-dessous en vous montrant les 8 premières lignes parmi les 50), pour ensuite faire un copier-coller qui plait à BiostaTGV !...



	A	B
1	Ca_t0	Ca_t1
2	1.29	1.29
3	0.66	0.64
4	0.75	0.73
5	1.24	1.22
6	0.74	0.71
7	1.25	1.17
8	0.82	0.73

Maintenant, vous êtes prêt(e) à faire les tests dans BiostaTGV.

3) Il faut enfin aller sur le site Internet BiostaTGV pour effectuer les tests de Student et de Wilcoxon pour séries appariées (le test de Wilcoxon pour séries appariées est le test des rangs signés de Wilcoxon sur le site Internet).

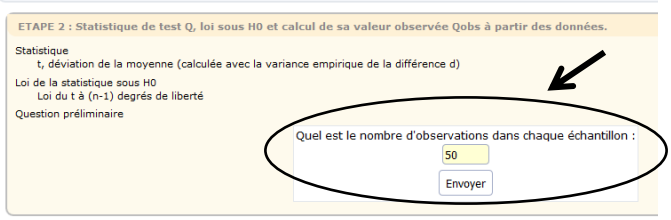
Type de test à mettre en évidence : Type de test : -Tous-

Variable de réponse

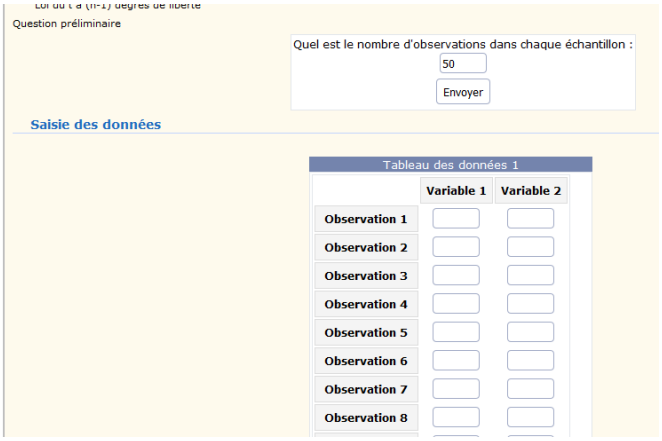
			Qualitative nominale (2 groupes)	Qualitative nominale (plus de 2 groupes)	Qualitative ordinale	Quantitative
Facteur d'étude	Qualitatif (deux groupes)	Indépendants	*Z de comparaison de proportions.**	Chi² (χ2.)	*Test de Cochran-Armitage**	Test de Mann-Whitney.
			Chi² (χ2.)			t de Student.
			Test exact de Fisher.			*Test de Welch.**
		Appariés	Test de McNemar.	*Q de Cochran.**	*Tests des signes.**	t de Student pour données appariées.
			Test exact de Fisher.		Tests des rangs signés de Wilcoxon.	Tests des rangs signés de Wilcoxon.
	Qualitatif (plus de deux groupes)	Indépendants	Chi² (χ2.)	Chi² (χ2.)	Test de Kruskal-Wallis. (ordinal)	Analyse de la variance.
						Test de Kruskal-Wallis. (échelle quanti)
		Appariés	*Q de Cochran.**	*Q de Cochran.**	Test de Friedman.	Test de Friedman.
	Quantitatif		*Régression logistique**	*Régression logistique multinomiale**	Corrélation de Spearman.	Corrélation de Pearson.
					Tau de Kendall.	*Régression linéaire.**

Je vais présenter la démarche pour le test de Student pour séries appariées, en faisant l'hypothèse que la calcémie suit une loi normale, mais la démarche est identique pour effectuer le test de Wilcoxon pour séries appariées, dans le cas où la calcémie ne suivait pas une loi normale.

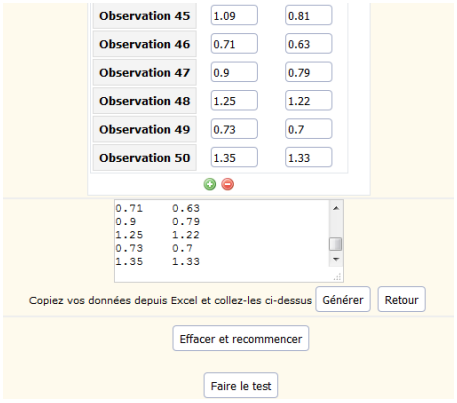
Une fois que vous avez cliqué sur « t de Student pour données appariées », vous devez d'abord fournir le nombre de lignes (c'est-à-dire, le nombre de chiens qui ont été mesurés à t0 et à t1[10]). Ici, tous les chiens de l'échantillon (n=50) ont été évalués pour la calcémie à t_0 et à t_1.



Voici ce que l'on obtient :

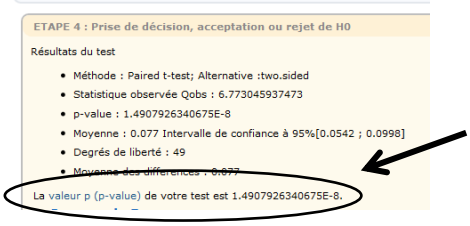


Vous devez maintenant copier-coller les 50 lignes de vos deux colonnes en même temps à partir d'Excel, en cliquant sur « Copier-coller via Excel » puis sur « Générer » :



[10] Attention, il ne faut fournir que les données complètes, c'est-à-dire celles ne manquant ni à t_0, ni à t_1. S'il y a des données manquantes à t_0 et/ou à t_1, vous devez constituer une nouvelle feuille Excel ne contenant que les individus pour lesquels les données ne manquent ni à t_0 ni à t_1.

Vous cliquez sur « Faire le test », et vous obtenez :



Le site fournit la « moyenne des différences », qui vaut 0,077 (0,08 en arrondissant) : c'est la différence entre la moyenne à t_0 (1,03) et la moyenne à t_1 (0,95). Le degré de signification du test est égal à $1{,}5 \times 10^{-8}$, inférieur au risque d'erreur α égal à 0,05, donc les deux moyennes comparées (1,03 et 0,95, respectivement pour la moyenne de la calcémie à t_0 et à t_1) sont significativement différentes.

V. Petit mot de conclusion

N'hésitez pas à me suggérer des ajouts dans ce tutoriel !... (Cela dit, ce tutoriel doit rester un tutoriel pour statistiques *de base*… !) Vos suggestions ou quelques commentaires que ce soit peuvent m'être envoyés à l'adresse suivante : loic.desquilbet@vet-alfort.fr