

## Chapitre 1 : Statistique univariée : encodage de tableaux de données brutes, calculs de statistiques descriptives et de tableaux de fréquences

### 1. Encodage de données brutes

Dans Excel, pas de structure préétablie pour encoder les données. Mais avec quelques bons réflexes, on peut éviter de se perdre dans des fichiers mal structurés et difficiles à lire !

#### Règles de base :

**Une ligne par individu (ou « enregistrement »)**

**Une colonne par variable (ou « champ »)**

Avantage :

- Facilite l'export vers d'autres logiciels (SPSS, par exemple)
- Facilite la collaboration (car convention fréquemment utilisée)

À titre d'exemple, ouvrez le classeur *R1-9.xlsx*, feuille *Données brutes*. Vous allez encoder correctement les données suivantes, relatives à 18 sujets.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Score ( $X_i$ )	6	9	4	13	14	9	8	12	11	9	12	18	14	10	19	8	15	11
Groupe ( $Y_i$ )	1	2	2	2	1	2	2	2	1	2	1	2	2	1	1	2	1	2
Age ( $Z_i$ )	17	18	22	16	14	15	20	23	17	18	19	19	19	24	22	16	21	26
Sexe ( $W_i$ )	1	1	1	1	1	2	2	1	1	2	1	2	2	1	1	2	2	1

#### Avant de commencer, quelques règles, de sorte à augmenter la lisibilité :

**Règle 1 :** créez une première ligne (= ligne 1 du tableur) = nom des variables.

**Intérêt :** permet de discriminer les colonnes.



**Règle 2 :** créez une première variable  $i$  (=colonne A du tableur) = numéro des individus statistiques.

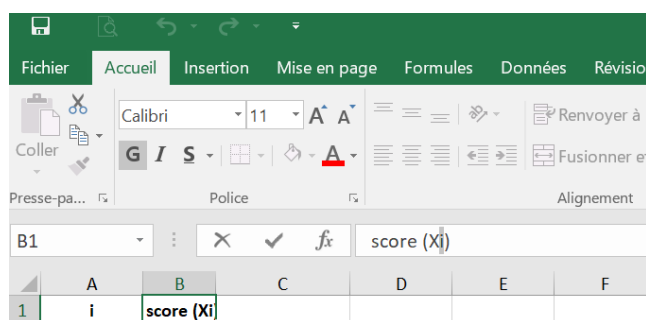
**Intérêt :** si au cours de vos analyses, vous êtes amenés à modifier l'ordre de présentations des données, cette première colonne pourra toujours servir à retrouver l'ordre initial.

## Encoder correctement les noms de variable :

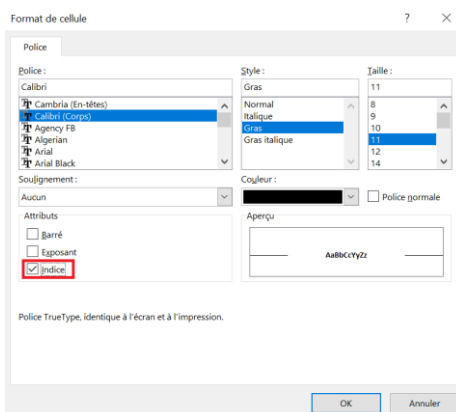
Pour rappel, une variable se symbolise par une lettre majuscule, associée à un indice « i » qui représente l'individu étudié.

Commencez par écrire le nom complet d'une variable dans la cellule adéquate, soit score (Xi) dans la cellule B1.

Ensuite, dans la barre de formule, mettez en surbrillance la partie que vous voulez transformer en indice (soit la lettre « i »). Pour ce faire, vous pouvez maintenir le bouton gauche de la souris enfoncé le temps de mettre en surbrillance ce que vous souhaitez, ou alternativement, placer le curseur entre la lettre majuscule et le « i », dans la barre de formule, enfoncer la touche « Shift »  de votre clavier, et simultanément, la flèche vers la droite .



Dans l'onglet d'affichage « accueil », cliquez sur la flèche en bas à droite de la zone « Police » du ruban pour ouvrir l'interface des paramètres de police, et cochez « indice » dans les attributs. Pour finir, cliquez sur OK.



Reproduisez la même démarche pour encoder correctement le nom des 3 autres variables.

Résultat final attendu :

l11					
	A	B	C	D	E
1	i	score (X <sub>i</sub> )	groupe (Y <sub>i</sub> )	âge (Z <sub>i</sub> )	sexe (W <sub>i</sub> )
2	1	6	1	17	1
3	2	9	2	18	1
4	3	4	2	22	1
5	4	13	2	16	1
6	5	14	1	14	1
7	6	9	2	15	2
8	7	8	2	20	2
9	8	12	2	23	1
10	9	11	1	17	1
11	10	9	2	18	2
12	11	12	1	19	1
13	12	18	2	19	2
14	13	14	2	19	2
15	14	10	1	24	1
16	15	19	1	22	1
17	16	8	2	16	2
18	17	15	1	21	2
19	18	11	2	26	1

l11					
	A	B	C	D	E
1	i	score (X <sub>i</sub> )	groupe (Y <sub>i</sub> )	âge (Z <sub>i</sub> )	sexe (W <sub>i</sub> )
2	1	6	1	17	1
3	2	9	2	18	1
4	3	4	2	22	1
5	4	13	2	16	1
6	5	14	1	14	1
7	6	9	2	15	2
8	7	8	2	20	2
9	8	12	2	23	1
10	9	11	1	17	1
11	10	9	2	18	2
12	11	12	1	19	1
13	12	18	2	19	2
14	13	14	2	19	2
15	14	10	1	24	1
16	15	19	1	22	1
17	16	8	2	16	2
18	17	15	1	21	2
19	18	11	2	26	1

→ 4 variables

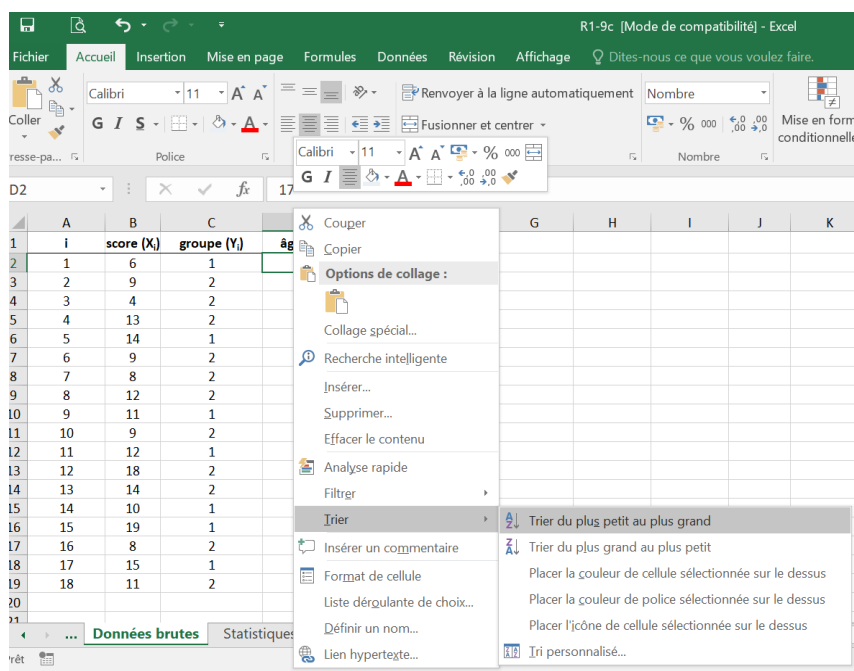
18 sujets

## Intérêt de la colonne « i » ?

Nous allons trier les observations par ordre croissant d'âge. Pour ce faire, sélectionnez l'une des cellules de la colonne « âge » (par exemple D2).

	A	B	C	D	E
1	i	score ( $X_i$ )	groupe ( $Y_i$ )	âge ( $Z_i$ )	sexe ( $W_i$ )
2	5	14	1	14	1
3	6	9	2	15	2
4	4	13	2	16	1
5	16	8	2	16	2
6	1	6	1	17	1
7	9	11	1	17	1
8	2	9	2	18	1
9	10	9	2	18	2
10	11	12	1	19	1
11	12	18	2	19	2
12	13	14	2	19	2
13	7	8	2	20	2
14	17	15	1	21	2
15	3	4	2	22	1
16	15	19	1	22	1
17	8	12	2	23	1
18	14	10	1	24	1
19	18	11	2	26	1

Ensuite, faites un clic droit sur votre souris et choisissez l'option « Trier du plus petit au plus grand ».



Le tri s'est étendu à l'ensemble de la base de données, ce qui est impératif, dans la mesure où toutes les informations qui figurent sur une ligne correspondent au même individu statistique.

Résultat :

I12					
	A	B	C	D	E
1	i	score ( $X_i$ )	groupe ( $Y_i$ )	âge ( $Z_i$ )	sexe ( $W_i$ )
2	5	14	1	14	1
3	6	9	2	15	2
4	4	13	2	16	1
5	16	8	2	16	2
6	1	6	1	17	1
7	9	11	1	17	1
8	2	9	2	18	1
9	10	9	2	18	2
10	11	12	1	19	1
11	12	18	2	19	2
12	13	14	2	19	2
13	7	8	2	20	2
14	17	15	1	21	2
15	3	4	2	22	1
16	15	19	1	22	1
17	8	12	2	23	1
18	14	10	1	24	1
19	18	11	2	26	1

Le numéro de chaque ligne (zone grisée qui apparaît juste avant la colonne A) n'a pas changé. Par contre, puisque le tri a été étendu à l'ensemble de la base de données, l'ordre d'apparition des chiffres dans la colonne « i », lui, s'est modifié !

Grâce à cela, on est capable de déduire que le sujet étant le plus jeune était le 5<sup>ème</sup> à être encodé dans la base de données.  $Z_5$  = le minimum de la variable âge = 14.

**Remarque :** les variables  $\text{groupe}(Y_i)$  et  $\text{sexe}(W_i)$  sont des variables nominales. Pourtant, les modalités de la variable semblent être chiffrées. Rappelez-vous le cours théorique : il arrive souvent que l'on attribue des chiffres comme nom d'étiquette aux modalités. Pour voir à quelle catégorie/modalité se rapporte chacune des valeurs numériques pour ces deux variables, regardez la feuille *Légendes* du fichier.

## 2. Calcul de statistiques descriptives, au départ de données brutes

Si peu de données/variables, les paramètres peuvent être calculés dans le feuillet dédié aux données brutes.

Si beaucoup de données/variables, il est préférable de calculer et encoder les statistiques descriptives sur une feuille à part :

- Créer une première variable dans la colonne A, que l'on nomme *Indicateur statistique*.
- Indiquer le nom de la variable représentée dans la première ligne de chaque colonne
- **ATTENTION :** il est très important de respecter scrupuleusement l'ordre d'apparition des variables. Cela permet d'étendre facilement les formules.

**Exemple** : ouvrez le classeur *R1-9.xlsx*, feuille *Statistiques descriptives*.

**Rappel** : pour tous les calculs ci-dessous, il vous sera demandé d'introduire dans la feuille *Statistiques descriptives* une formule faisant appel à des séries de données figurant dans la feuille *Données brutes*. Pensez à votre cours d'introduction à Excel pour savoir comment vous y prendre.

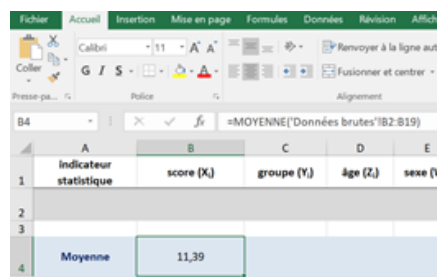
### Indicateurs de tendance centrale

#### La moyenne

La **Moyenne** est un indicateur de tendance centrale qui consiste à additionner toutes les valeurs d'une variable, et à diviser le tout par le nombre de termes additionnés.

**Fonction** : **=MOYENNE()**. L'argument à introduire entre parenthèses sera la plage de la série de données, soit l'ensemble des valeurs à considérer pour calculer la moyenne (par exemple, la série de données D2:D19 de la feuille *Données brutes* contient les données des 18 sujets relatives à la variable  $\hat{A}ge(Z_i)$ ).

Pour éviter des erreurs, un bon réflexe est d'étendre les formules lorsque c'est possible. En B4, introduisez la formule adéquate pour calculer la moyenne de la variable  $Score(X_i)$ . Affichez deux décimales.



Étirez ensuite la formule dans les cellules C4:E4.

Avez-vous obtenu le résultat ci-dessous ?

17						
	A	B	C	D	E	F
1	Indicateur statistique	score (X <sub>i</sub> )	groupe (Y <sub>i</sub> )	âge (Z <sub>i</sub> )	sexe (W <sub>i</sub> )	
2	Type de variable					
4	Moyenne	11,22	1,61	19,22	1,39	Indice tendanc

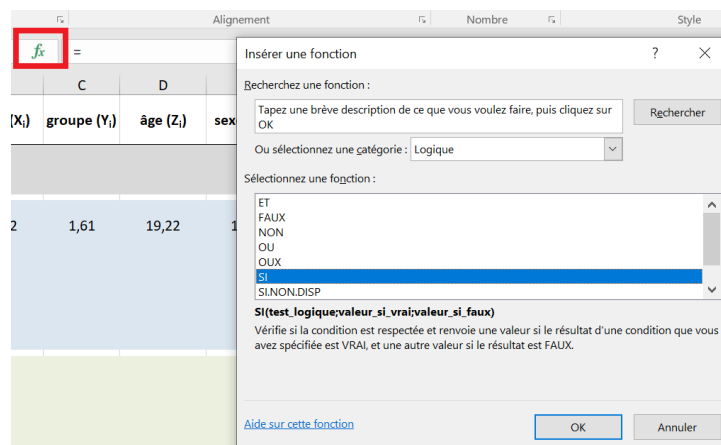
On y est presque, mais un détail devrait vous choquer...À quoi correspond le score 1,39 pour la variable Sexe(W<sub>i</sub>) ? et le 1,61 pour la variable Groupe(Y<sub>i</sub>) ?

Souvenez-vous : le calcul de la moyenne n'est pas du tout pertinent pour des variables de type nominal (ni pour n'importe quelle variable de type qualitatif)!

Nous allons corriger cela. Dans un premier temps, indiquez pour chaque variable son type. A niveau de la ligne 2, écrivez pour chaque variable, soit « qualitatif », soit « quantitatif ». Ensuite, modifiez la formule que vous avez introduite en B4 et faites-en sorte que :

- Si la variable est de type quantitatif, la moyenne est calculée
- Si la variable est de type qualitatif, une barre « / » est imprimée.

**Prérequis :** les fonctions conditionnelles (syllabus 7, niveau 2 de votre cours d'introduction à Excel). Un trou de mémoire ? N'hésitez pas à utiliser l'aide disponible dans Excel. Cliquez sur le symbole « insérer une fonction », et faites une recherche dans les fonctions logiques. Vous y trouverez le nécessaire.



## La médiane

La **Médiane** est un indicateur plus robuste que la moyenne (=moins influencé par les valeurs extrêmes). Elle consiste à calculer une valeur telle qu'il y ait autant d'observations qui lui soient inférieures ou égales qu'il n'y a d'observations qui lui soient supérieures ou égales (on découpe la distribution en 2 parts contenant le même nombre de sujets).

Fonction : **=MEDIANE()** », où l'argument à introduire entre parenthèses est la plage de la série de données.

Créez-en B5 une formule qui permet de calculer la médiane de la variable *Score(X<sub>i</sub>)*, en utilisant les fonctions conditionnelles :

- Si la variable est de type quantitatif, la médiane est calculée
- Si la variable est de type qualitatif, une barre « / » est imprimée.

Affichez une décimale et étirez la formule dans les cellules C5:E5.

## Le mode

Contrairement à la moyenne et à la médiane, le **Mode** est un indicateur de tendance centrale qui convient tant pour décrire les variables qualitatives que les variables quantitatives.

**Remarque 1** : Bien que le mode soit assez simple à calculer dans Excel lorsque les modalités d'une variable sont chiffrées, c'est plus compliqué lorsque les modalités sont non chiffrées (si vous vous demandiez pourquoi nous avons choisi des valeurs chiffrées pour décrire le sexe et le groupe, voici une partie de la réponse).

**Remarque 2** : il existe la fonction suivante dans Excel : **=MODE()**, où l'argument à introduire entre parenthèses est la plage de la série de données. Bien que cette fonction soit toujours disponible dans les versions actuelles d'Excel, elle est amenée à disparaître. Il est préférable de lui préférer deux nouvelles fonctions plus précises ayant été ajoutée ultérieurement :

- **=MODE.SIMPLE()**
- **=MODE.MULTIPLE()**

L'option **=MODE.SIMPLE()** retourne systématiquement une seule valeur, soit celle associée à la plus grande fréquence. Elle est adéquate uniquement lorsque la distribution est unimodale (ce qui est le cas pour toutes les variables du fichier R1-9.xlsx).

Fonction : **=MODE.SIMPLE()**, où l'argument à introduire entre parenthèses est la plage de la série de données.

Créez-en B6 une formule qui permet de calculer le mode de la variable  $Score(X_i)$ , en utilisant les fonctions conditionnelles :

- Si la variable est de type quantitatif, la médiane est calculée
- Si la variable est de type qualitatif, une barre « / » est imprimée.

N'affichez aucune décimale et étirez la formule dans les cellules C6:E6.

A20						
	A	B	C	D	E	F
1	indicateur statistique	score ( $X_i$ )	groupe ( $Y_i$ )	âge ( $Z_i$ )	sexe ( $W_i$ )	
2	Type de variable	quantitatif	qualitatif	quantitatif	qualitatif	
3						
4	Moyenne	11,39	/	19,22	/	Indicateurs de tendance centrale
5	Médiane	11	/	19	/	
6	Mode	11	mono-parental	19	Homme	



À présent, répondez aux questions suivantes :

- Quel type de famille est le plus représenté dans l'échantillon : les familles monoparentales, ou les familles biparentales ?
- Quel sexe est le plus représenté dans l'échantillon : les hommes ou les femmes ?

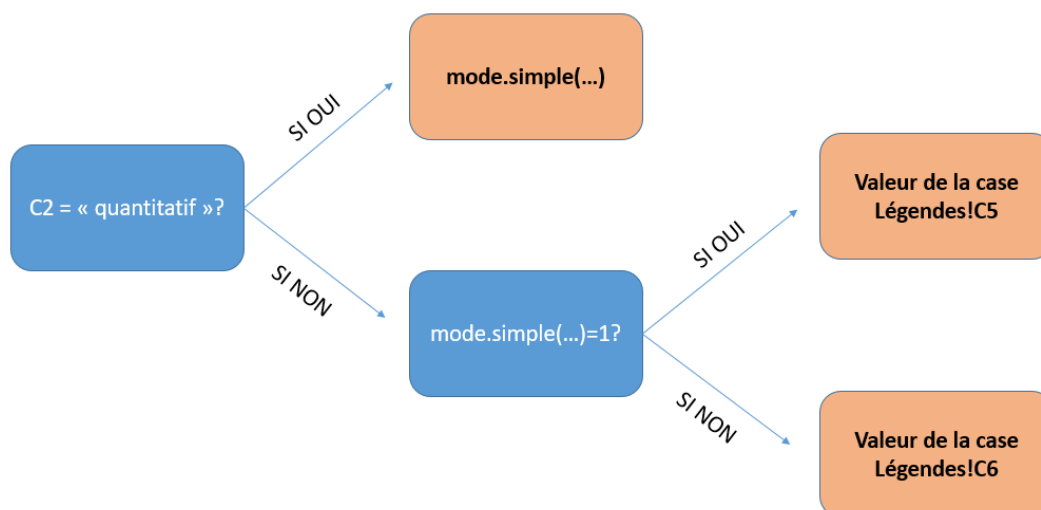
La réponse figure dans la feuille *Légende* (vous y trouverez ce que représente le « 2 » dans la variable *Groupe*(Y<sub>i</sub>), et le « 1 » dans la variable *Sexe*(W<sub>i</sub>)).

Imaginez que vous ayez quelques minutes pour analyser un grand tableau, constitué de plein de variables, et que vous deviez chaque fois switcher de la feuille *Légendes* vers la feuille *Statistique descriptives*. Peu pratique n'est-ce pas ?

Nous allons modifier la formule, de sorte que lorsque la variable est de type qualitatif, ce ne soit plus un code chiffré qui apparaisse, mais bien le nom de la catégorie représentée par ce chiffre.

**Astuce** : à nouveau, utiliser les fonctions conditionnelles, mais cette fois, en imbriquant deux conditions. Pensez à placer les « \$ » correctement, pour pouvoir étirer les formules.

Besoin d'un coup de pouce ? Commencez par modifier la formule de la cellule C6 (vous la copiez-collerez ensuite dans les cellules B6, D6 et E6). Cet arbre de décision devrait vous aider :



Résultat attendu :

	A	B	C	D	E	F
1	indicateur statistique	score ( $X_i$ )	groupe ( $Y_i$ )	âge ( $Z_i$ )	sexe ( $W_i$ )	
2	Type de variable	quantitatif	qualitatif	quantitatif	qualitatif	
3						
4	Moyenne	11,39	/	19,22	/	Indicateurs de tendance centrale
5	Médiane	11	/	19	/	
6	Mode	11	mono-parental	19	Homme	

L'option **=MODE.MULTIPLE()** retourne une seule valeur si la distribution est unimodale. Elle retourne par contre plusieurs valeurs si la distribution est bimodale. Vous l'aurez compris, l'option **=MODE.MULTIPLE()** est plus pertinente que l'option **=MODE.SIMPLE()**, puisqu'elle peut être utilisée tout le temps. Cependant, elle est plus compliquée à utiliser.

**Pourquoi plus compliquée ?** Parce qu'il s'agit d'une **formule matricielle**. Ce type de formule sera développé dans un autre chapitre.

### Indicateurs de dispersion

L'**Étendue** des données consiste à calculer l'écart entre les valeurs maximales et minimales. La calculer implique donc d'utiliser deux fonctions : celles du minimum et celle du maximum.

#### Fonctions :

- **=MIN()**, où l'argument à introduire entre parenthèses est la plage de la série de données.
- **=MAX()**, où l'argument à introduire entre parenthèses est la plage de la série de données.

Créez-en B21 une formule qui permet de calculer l'étendue de la variable *Score( $X_i$ )*, en utilisant les fonctions conditionnelles :

- Si la variable est de type quantitatif, la médiane est calculée
- Si la variable est de type qualitatif, une barre « / » est imprimée.

N'affichez aucune décimale et étirez la formule dans les cellules C21:E21.

L'**Écart interquartile** est l'écart entre le troisième quartile et le premier quartile (il correspond à la boîte centrale de la boîte à moustache).

**Rappel :**

- Le troisième quartile est égal au 75<sup>ème</sup> percentile
- Le premier quartile est égal au 25<sup>ème</sup> percentile

Dans Excel, vous pourrez utiliser soit la fonction quartile, soit la fonction percentile.

Fonctions :

- **=CENTILE()**, où les arguments à introduire entre parenthèses sont, d'une part, la plage de la série de données et d'autre part, une proportion (soit k, la proportion des valeurs de la série de données qui sont inférieures ou égales à la valeur recherchée).
  - k = 0,25 → premier quartile
  - k = 0,5 → médiane
  - k = 0,75 → troisième quartile
- **=QUARTILE()**, où les arguments à introduire entre parenthèses sont, d'une part, la plage de la série de données et d'autre part, le numéro du quartile :
  - 1 → premier quartile ;
  - 2 → second quartile ;
  - 3 → troisième quartile ;

Créez-en B22 une formule qui permet de calculer l'écart interquartile de la variable *Score*( $X_i$ ), en utilisant les fonctions conditionnelles :

- Si la variable est de type quantitatif, la médiane est calculée
- Si la variable est de type qualitatif, une barre « / » est imprimée.

Affichez deux décimales et étirez la formule dans les cellules C22:E22.

L'**Écart moyen absolu** (ou **EMA**) est une mesure de dispersion très utile, bien que peu utilisée en pratique. Elle consiste à calculer la moyenne des écarts absolus des observations d'une série de données par rapport à la moyenne arithmétique de cette série.

Fonction : **=ECART.MOYEN()**, où l'argument à introduire entre parenthèses est la plage de la série de données.

Créez-en B23 une formule qui permet de calculer l'EMA de la variable  $Score(X_i)$ , en utilisant les fonctions conditionnelles :

- Si la variable est de type quantitatif, la médiane est calculée
- Si la variable est de type qualitatif, une barre « / » est imprimée.

Affichez deux décimales et étirez la formule dans les cellules C23:E23.

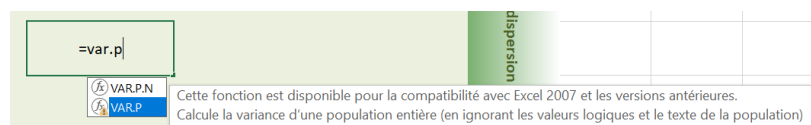
La **Variance** est une alternative à l'EMA beaucoup plus fréquemment utilisée. Elle consiste à calculer la moyenne des écarts au carré des observations d'une série de données par rapport à la moyenne arithmétique de cette série.

Il faut distinguer deux cas, lorsqu'on veut calculer la variance dans Excel :

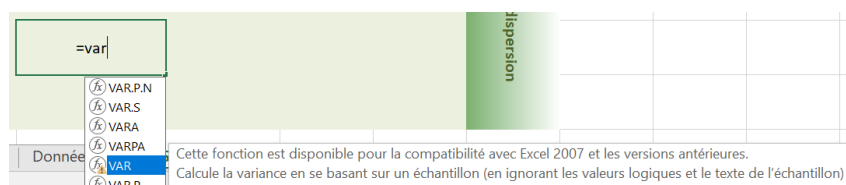
- 1) Celui où l'on veut simplement décrire notre échantillon (fonction **=VAR.P()** = statistiques descriptives).
- 2) Celui où l'on veut estimer ce que serait la variance d'une population plus large au départ de laquelle notre échantillon est récolté (fonction **=VAR()** = statistiques inférentielles).

Faites **très** attention ici. L'aide Excel pourrait vous sembler contre-intuitive et pourrait vous induire en erreur !

Voici l'explication associée à la formule **=VAR.P()** :



Et voici celle associée à la formule **=VAR()** :



Cela vous semble-t-il contradictoire avec mon explication plus haut ?

Si **non**, l'explication ci-dessous ne vous intéressera pas. Si **oui**, voici ce qu'il en est :

Comme expliqué en introduction de ce cours, il arrive souvent que l'on veuille décrire une population sur base d'un échantillon (par exemple, parce qu'il n'est pas possible de récolter les données de la population entière). Cependant, la formule de la variance que l'on vous a apprise dans ce cours est telle que bien souvent, elle aura pour effet de légèrement sous-estimer la variance de la population, c'est pourquoi il convient d'effectuer une correction.

Lorsque dans Excel, vous précisez vouloir utiliser la formule « en se basant sur un échantillon », il sera automatiquement compris que vous voulez estimer le paramètre d'une population plus large.

Lorsqu'au contraire vous précisez vouloir utiliser la formule de la variance d'une population entière, il est compris qu'il n'est pas nécessaire d'effectuer de correction, puisque le but n'est pas d'estimer la variance d'un ensemble contenant plus de sujets que récoltés.

Autrement dit, « leurrer » Excel, en affirmant que notre échantillon est en fait la population est le moyen de calculer la formule de variance sans correction, et c'est ce que j'attends de vous.

Fonction : **=VAR.P()**, où l'argument à introduire entre parenthèses est la plage de la série de données.

Créez-en B24 une formule qui permet de calculer la variance de la variable  $Score(X_i)$ , en utilisant les fonctions conditionnelles :

- Si la variable est de type quantitatif, la médiane est calculée
- Si la variable est de type qualitatif, une barre « / » est imprimée.

Affichez deux décimales et étirez la formule dans les cellules C24:E24.

Enfin, l'**écart-type** est la racine-carré de la variance. Il existe deux manières de la calculer :

- Soit vous utilisez la fonction proposée par Excel (à nouveau, il convient de choisir l'option qui permet de calculer l'écart-type d'une population entière ; cf. explication ci-dessus).
- Soit vous calculez la racine-carré de la valeur introduire à la ligne 24.

Fonctions :

- **=ECARTYPE.P()**, où l'argument à introduire entre parenthèses est la plage de la série de données.
- **OU =RACINE()**, où l'argument à introduire entre parenthèses est la cellule contenant la valeur dont on veut prendre la racine carrée.

Créez-en B25 une formule qui permet de calculer l'écart-type de la variable  $Score(X_i)$ , en utilisant les fonctions conditionnelles :

- Si la variable est de type quantitatif, la médiane est calculée
- Si la variable est de type qualitatif, une barre « / » est imprimée.

Affichez deux décimales et étirez la formule dans les cellules C25:E25.

### Mesures d'asymétrie et d'aplatissement

La mesure d'**asymétrie** sert entre autres à décider si la moyenne sera ou non un résumé pertinent de notre distribution.

Fonction : **=COEFFICIENT.ASYMETRIE.P()**, où l'argument à introduire entre parenthèses est la plage de la série de données.

Pour l'explication relative au « .P », voir notes de la variance et de l'écart-type.

Créez-en B27 une formule qui permet de calculer le coefficient G1 de la variable  $Score(X_i)$ , en utilisant les fonctions conditionnelles :

- Si la variable est de type quantitatif, la médiane est calculée
- Si la variable est de type qualitatif, une barre « / » est imprimée.

Affichez deux décimales et étirez la formule dans les cellules C27:E27.

Question : parmi toutes les variables quantitatives, laquelle présente la plus forte asymétrie ? S'agit-il d'une asymétrie positive, ou négative ?

La mesure d'**Aplatissement** est un indicateur de forme de la distribution, essentiellement lié à la densité des extrémités des distributions (plus le kurtosis est élevé, plus forte est la densité des extrémités).

Fonction : **=KURTOSIS()**, où l'argument à introduire entre parenthèses est la plage de la série de données.

Cette équation n'est pas identique à celle introduite dans le cours théorique. Le principe reste cependant identique : si le kurtosis vaut 0, alors la distribution est similaire à la distribution normale (en termes de densité). Si le kurtosis est supérieur à 0, cela signifie que la distribution est plus pointue, avec des extrémités plus lourdes que la normale et enfin, si le kurtosis est inférieur à 0, cela signifie que la distribution est plus aplatie, avec des extrémités moins denses que la normale.

Créez-en B28 une formule qui permet de calculer le coefficient G2 de la variable  $Score(X_i)$ , en utilisant les fonctions conditionnelles :

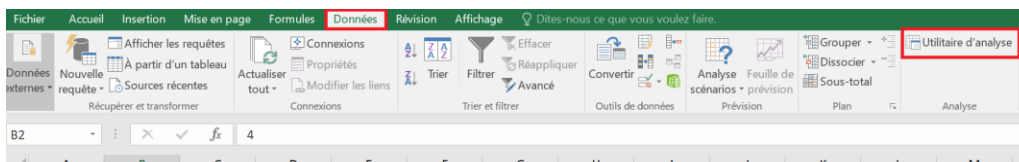
- Si la variable est de type quantitatif, la médiane est calculée
- Si la variable est de type qualitatif, une barre « / » est imprimée.

Affichez deux décimales et étirez la formule dans les cellules C28:E28.

## Calculer toutes les statistiques descriptives usuelles d'une variable, en une seule fois

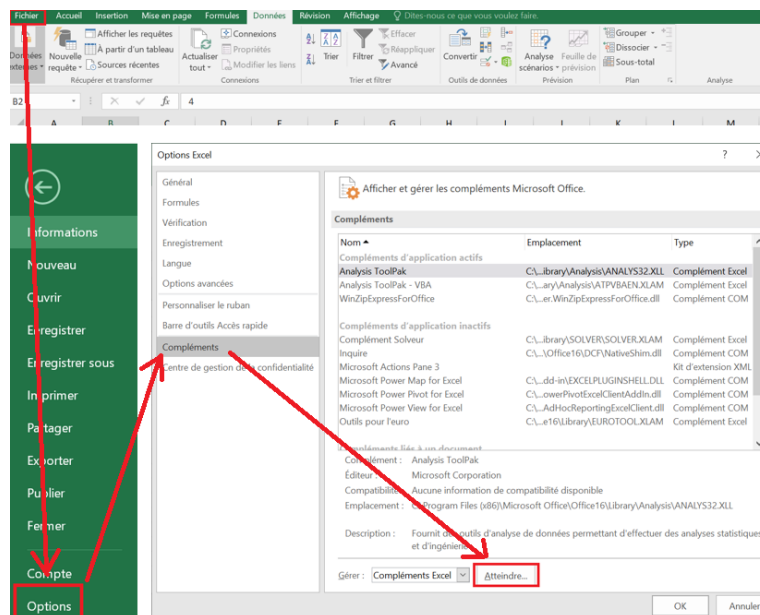
Excel propose un utilitaire d'analyse de données, qui permet entre autres de calculer automatiquement une série de paramètres descriptifs pour une ou plusieurs variables. Cet utilitaire présente cependant quelques limites, que nous exposerons après l'avoir utilisé.

Rendez-vous dans l'onglet « Données », groupe « Analyse », « Utilitaire d'analyse ».

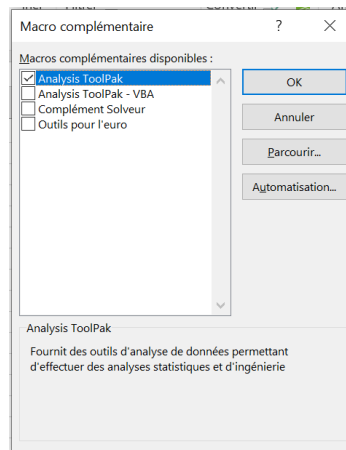


Il est possible que vous ne trouviez pas cette option. Si c'est le cas, cela signifie qu'elle n'est pas activée. Voici les étapes à suivre pour l'activer :

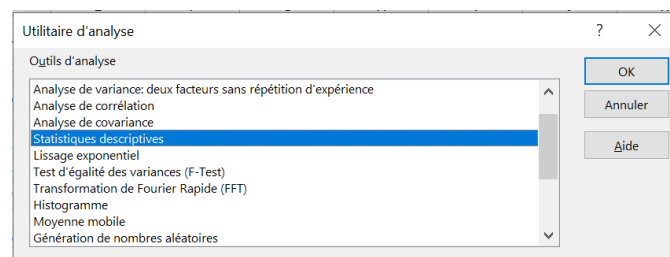
Rendez-vous dans l'onglet « Fichier », et cliquez sur « Options ». Dans l'interface qui s'ouvre, cliquez sur « Compléments ». En bas de la nouvelle interface, cliquez sur « Atteindre ».



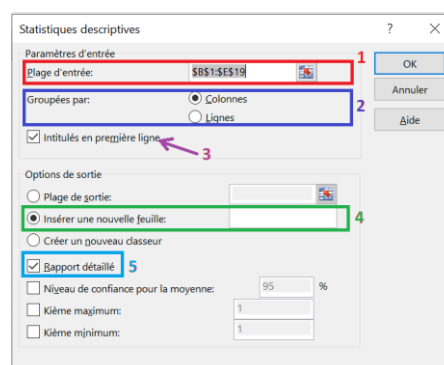
Dans les macros complémentaires disponibles, cochez « Analysis ToolPak », et cliquez sur OK.



À présent, l'utilitaire d'analyse apparaît dans l'onglet « Données. Cliquez dessus. Cochez l'option « Statistiques descriptives », puis cliquez sur OK.



- 1) Dans l'interface ouvrant, commencez par spécifier la plage d'entrée. Vous devez sélectionner toute la base de données, à l'exception de la variable *i*, soit la plage de données **B1:E19** (pensez au chapitre « Comment faire des sélections ? » de votre cours d'introduction à Excel, syllabus 1, niveau 1, p 15).
- 2) Spécifiez que les variables ont été encodées par « Colonnes » (cf. règle de base de l'encodage des données).
- 3) Cochez l'option « Intitulés en première ligne »
- 4) Précisez que vous voulez insérer les résultats dans une nouvelle feuille
- 5) Demandez un rapport détaillé



Cet outil présente de nombreuses limites, du moins dans l'usage que nous voulons en faire :



1. Il n'est pas dynamique : si les données de la feuille *Données brutes* sont modifiées, le compte rendu ne sera pas mis à jour.
2. Il ne fait pas de distinction entre les variables qualitatives et quantitatives. Lorsque les observations d'une variable sont des chiffres, des paramètres tels que la moyenne pourront toujours être calculés, même si ceux-ci ne sont pas intrinsèquement numériques (cf. variables qualitatives).
3. Il calcule par défaut la variance et l'écart-type corrigé (cf. Remarque plus haut, dans la section sur la variance). Idem pour le kurtosis.
4. Il donne une réponse chiffrée pour le mode, et non le nom de l'étiquette qui s'y rapporte.
5. Il ne fournit pas tous les paramètres désirés (il manque l'écart interquartile, et l'EMA).

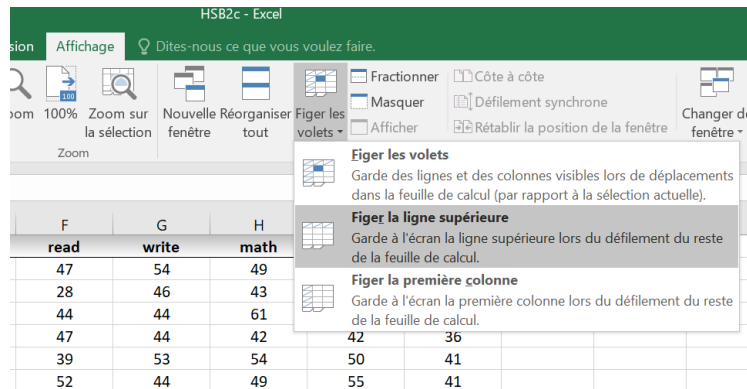
### 3. Les tableaux de fréquence

De même que pour les statistiques descriptives, il est préférable de créer les tableaux de fréquence sur une feuille à part. Ouvrez le classeur *HSB2.xlsx*. Avant toute chose, constatez que ce tableau comporte un nombre important d'observations (200 lignes de données), et que vous devrez calculer des informations en dessous de ce tableau. Descendez vers le bas du tableau, jusqu'à voir les cellules « Nombre de modalités » et « N ».

	A	B	C	D	E	F	G	H	I	J
193	191	F	privé	technique	Wallonie	47	52	43	48	61
194	192	M	privé	technique	Wallonie	65	67	63	66	71
195	193	F	privé	technique	Wallonie	44	49	48	39	51
196	194	F	privé	technique	Wallonie	63	63	69	61	61
197	196	M	privé	technique	Wallonie	44	38	49	39	46
198	197	M	privé	technique	Wallonie	50	42	50	36	61
199	198	F	privé	technique	Wallonie	47	61	51	63	31
200	199	M	privé	technique	Wallonie	52	59	50	61	61
201	200	M	privé	technique	Wallonie	68	54	75	66	66
202										
203	Nombre de modalités:			3	3					22
204	N:									

Le tableau étant conséquent, les noms des titres ne sont plus visibles (ce qui peut être embêtant). Pour résoudre ce problème, nous allons figer la première ligne du tableau, de sorte à ce qu'elle reste visible en permanence.

Dans l'onglet affichage, cliquez sur « Figer les volets », et ensuite, sur « Figer la ligne supérieure ».



À présent, nous allons déterminer le tableau de fréquence de la variable  $Programme(T_i)$ . Commencez par trier vos données par ordre alphabétique en fonction de la variable  $Programme(T_i)$ . Cela vous facilitera la création du tableau.

- 1) D'abord, il faut déterminer le nombre de modalités de la variable. Comme cela implique d'utiliser une formule de type matriciel et que ces formules font l'objet du chapitre 2, la réponse vous est donnée dans la cellule D203 de la feuille *Données brutes*. Plus tard, vous devrez être apte à déterminer cette réponse vous-mêmes.
- 2) Ensuite, il faut déterminer le nombre total d'observations (soit le nombre de cellules non vides du champ de la variable  $Programme(T_i)$ ).

Fonction : **=NBVAL()**, où l'argument à introduire entre parenthèses est la plage de la série de données.

Créez-en D204 une formule qui permet de calculer le nombre de cellules non vides du champ de la variable  $Programme(T_i)$ .

Vous disposez à présent de deux valeurs qui serviront à vérifier qu'il n'y ait pas d'erreur dans votre tableau de fréquence.

Dans la feuille *Tableau de fréquences*, créez un tableau (au départ de la cellule A3), en respectant les règles suivantes :

**Une ligne par modalité de la variable étudiée**

**Une colonne par information**

De plus,

- Créez une première ligne (= ligne 1 du tableau) = nom des colonnes
- Créez une première variable  $j$  (=colonne A du tableau) = numéro de la modalité statistique étudiée.
- Créez une ligne juste en dessous du tableau = TOTAL

	A	B	C	D
1	Tableau de fréquences de la variable "Programme (T <sub>i</sub> )"			
2				
3	j	t <sub>j</sub>	n <sub>j</sub>	f <sub>j</sub>
4	1			
5	2			
6	3			
7	TOTAL			
8				

**Rappel** : pourquoi la première colonne se nomme-t-elle « t<sub>j</sub> » ? (t minuscule et indice j ?)

**Rappel 2** : pourquoi n'y a-t-il pas de colonnes prévues pour N<sub>j</sub> et F<sub>j</sub> ?

Commencez à remplir le tableau :

1. Complétez la colonne t<sub>j</sub>. Respectez scrupuleusement le nom de chaque modalité.
2. Complétez ensuite la colonne n<sub>j</sub>.

**Fonction** : **=NB.SI()**, où l'argument à introduire entre parenthèses est la plage de la série de données.

**Attention** : comme toujours, assurez-vous de générer une fonction qui puisse être étirée de C4 jusqu'à C6. N'oubliez pas les \$.

➔ Complétez la colonne C7, en utilisant la fonction **=SOMME()**. Trouvez-vous la même réponse que dans la cellule D204 ? Si oui, c'est bon signe !

3. Complétez la colonne f<sub>j</sub>. Vous pouvez reproduire la formule de la colonne C, et diviser la formule par le n total, ou tout simplement, diviser le résultat obtenu dans la colonne C par le n total.

**Conseil** : ne divisez pas directement par la valeur de n, mais par une des cellules qui contient la valeur de n (par exemple, la cellule C7). De la sorte, si les données doivent être modifiées dans la feuille *Données brutes* et que le n change, les fréquences seront correctement ajustées.

➔ Complétez la colonne D7, en utilisant la fonction **=SOMME()**. Trouvez-vous « 1 » comme réponse ? Si oui, c'est que tout va bien !

Entraînez-vous en réalisant le tableau de fréquence de la variable *Région(U<sub>i</sub>)*, au départ de la cellule A11.

Une fois que vous avez terminé, créez un tableau de fréquence pour la variable  $Soc(Z_i)$ , soit les notes au cours de Social, au départ de la cellule A19.

Contrairement aux variables  $Programme(T_i)$  et  $Région(U_i)$ , le score au cours de Social est une variable de type quantitatif. Dès lors, il faudra ajouter les colonnes  $N_j$  et  $F_j$ .

**De plus, ATTENTION** : Souvenez-vous que lorsqu'on travaille avec des variables **quantitatives**, il faut respecter un intervalle égal entre chaque ligne ! Pour vous aider, commencez par calculer le minimum et le maximum de la variable  $Soc(Z_i)$  (cf. chapitre 1). La ligne sous celle des noms de variables représentera le score minimum. Créez ensuite des lignes, en dessous, dont la valeur avancera par pas de 1, jusqu'à atteindre le score maximum.

**Résultat attendu :**

	A	B	C	D	E	F
18	Tableau de fréquences de la variable "Soc (Z <sub>i</sub> )"					
19	<b>i</b>	<b>z<sub>i</sub></b>	<b>n<sub>i</sub></b>	<b>f<sub>i</sub></b>	<b>N<sub>i</sub></b>	<b>F<sub>i</sub></b>
20	1	26				
21	2	27				
22	3	28				
23	4	29				
24	5	30				
25	6	31				
26	7	32				
27	8	33				
28	9	34				
29	10	35				
30	11	36				
31	12	37				
32	13	38				
33	14	39				
34	15	40				
35	16	41				
36	17	42				
37	18	43				
38	19	44				
39	20	45				
40	21	46				
41	22	47				
42	23	48				
43	24	49				
44	25	50				
45	26	51				
46	27	52				
47	28	53				
48	29	54				
49	30	55				
50	31	56				
51	32	57				
52	33	58				
53	34	59				
54	35	60				
55	36	61				
56	37	62				
57	38	63				
58	39	64				
59	40	65				
60	41	66				
61	42	67				
62	43	68				
63	44	69				
64	45	70				
65	46	71				
66	TOTAL					

Complétez ensuite les colonnes  $n_j$  et  $f_j$ , exactement de la même manière que pour les variables qualitatives. Vous pouvez aisément vérifier vos réponse en calculant le total, et en vérifiant que :

- Le total de la colonne C est identique à la valeur de la cellule 'Données brutes'!J204
- Le total de la colonne D est égal à 1.

Il reste à compléter les colonnes E et F.

- Pour compléter la colonne  $N_j$ , il faut trouver une formule de telle sorte qu'à la ligne associée à la première modalité, la valeur soit identique à celle de la colonne  $n_j$  et qu'à toutes les autres lignes, la valeur soit égale à celle de la colonne  $n_j$  + la référence de la cellule au-dessus.

**Astuce** : utiliser une formule conditionnelle, où la condition est liée à la valeur de  $j$ .

- ➔ La dernière valeur de fréquence cumulée est-elle identique au total de la colonne  $n_j$  ? Si oui, vous êtes sur la bonne voie.

4. Pour compléter la colonne  $F_j$ , il faut trouver une formule de telle sorte qu'à la première ligne, la valeur soit identique à celle de la colonne  $f_j$  et qu'à toutes les autres lignes, la valeur soit égale à celle de la colonne  $f_j$  + la référence de la cellule au-dessus.

**Astuce** : si vous placez judicieusement des \$ dans la formule que vous aviez introduite dans la colonne E ( $N_j$ ), vous pourrez l'étendre à la colonne F ( $F_j$ ).

- ➔ La dernière valeur de fréquence cumulée vaut-elle 1 ? Si oui, vous êtes à présent capable d'appliquer une méthode qui fonctionne pour tout type de variable (qualitatif ET quantitatif).

Résultat attendu pour la variable  $soc(Z_i)$  :

Presse-pa...	Police	Alignement				
F76						
	A	B	C	D	E	F
17						
18	Tableau de fréquences de la variable "Soc"					
19	j	x <sub>j</sub>	n <sub>j</sub>	f <sub>j</sub>	N <sub>j</sub>	F <sub>j</sub>
20	1	26	3	0,015	3	0,015
21	2	27	0	0	3	0,015
22	3	28	0	0	3	0,015
23	4	29	0	0	3	0,015
24	5	30	0	0	3	0,015
25	6	31	8	0,04	11	0,055
26	7	32	1	0,005	12	0,06
27	8	33	1	0,005	13	0,065
28	9	34	0	0	13	0,065
29	10	35	0	0	13	0,065
30	11	36	8	0,04	21	0,105
31	12	37	1	0,005	22	0,11
32	13	38	0	0	22	0,11
33	14	39	1	0,005	23	0,115

#### 4. La fonction =SOMMEPROD()

La fonction =SOMMEPROD() est très utile, pour calculer certains paramètres statistiques au départ d'un tableau de fréquences. Illustrons ceci à travers le calcul de la moyenne.

##### La moyenne

Dans le cadre du cours théorique, vous avez appris qu'il était possible de calculer cet indicateur de tendance centrale de trois manières distinctes :

$$\text{Au départ des données brutes: } \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$\text{Au départ des fréquences absolues: } \bar{X} = \frac{\sum_{j=1}^k x_j \times n_j}{n}$$

$$\text{Au départ des fréquences relatives: } \bar{X} = \sum_{j=1}^k x_j \times f_j$$

Les deux formules au départ de fréquences ont un point commun : dans les deux cas, pour chaque ligne du tableau de fréquences, il faut multiplier la valeur de la modalité par la fréquence, et ensuite faire la somme du résultat obtenu à chaque ligne.

C'est typiquement le genre d'actions que permet de réaliser la fonction =SOMMEPROD().

Fonction =SOMMEPROD(), où les arguments à introduire entre parenthèses sont d'une part, la plage de la série de données constituant le premier terme de la multiplication (matrice 1 ;  $x_j$ ) et d'autre part, la plage de la série de données constituant le deuxième terme de la multiplication (matrice 2 ; soit  $n_j$ , ou  $f_j$ , suivant la formule considérée).

Nous allons nous entraîner. Ouvrez la feuille SOMMEPROD du fichier HSB2.xlsx. Il vous est demandé de calculer la moyenne de la variable  $soc(Z_i)$  de trois manières différentes :

- Sur base des fréquences absolues (cellule B4)
- Sur base des fréquences relatives (cellule C4)
- Sur base des données brutes (cellule D4).

Si vous obtenez la même réponse dans les trois cellules, a priori, vous avez compris le principe !

## EMA:

De même que pour la moyenne, il est possible de calculer l'EMA sur base des fréquences absolues et des fréquences relatives. Le principe de la pondération est rigoureusement identique. Si vous avez bien compris ce principe, vous n'aurez aucun mal à accepter les formules suivantes:

$$\text{Calcul au départ de la fréquence absolue: } EMA_X = \frac{\sum_{j=1}^k |x_j - \bar{X}| \times n_j}{n}$$

$$\text{Calcul au départ de la fréquence relative: } EMA_X = \sum_{j=1}^k |x_j - \bar{X}| \times f_j$$

Essayez de compléter les cellules B7 et C7. Complétez ensuite la cellule D7, afin de vérifier l'exactitude de vos réponses.

## Variance :

Cette fois encore, suivant le même principe, il est possible de calculer la variance sur base des formules suivantes:

$$\text{Calcul au départ de la fréquence absolue: } S_X^2 = \frac{\sum_{j=1}^k (x_j - \bar{X})^2 \times n_j}{n}$$

$$\text{Calcul au départ de la fréquence relative: } S_X^2 = \sum_{j=1}^k (x_j - \bar{X})^2 \times f_j$$

Essayez de compléter les cellules B8 et C8. Complétez ensuite la cellule D8, afin de vérifier l'exactitude de vos réponses.

## 5. Exercices récapitulatifs

1. Ouvrez le fichier *Eau.minerale.xlsx*.

Le fichier contient un tableau qui récapitule la composition chimique d'un certain nombre d'eaux minérales classées par pays (cf. Tomassone, Dervin & Masson 1993 BIOMÉTRIE, Modélisation des phénomènes biologiques, Masson, Paris, 553 p.). Les variables sont les suivantes :

- Nom = la dénomination d'origine de la source
- Pays = le pays d'origine
- Cl = la composition en chlore
- Ca = la composition en calcium
- Mg = la composition en manganèse
- Ni = la composition en nitrate
- SO4 = la composition en sulfate
- HCO3 = la composition en gaz carbonique

1.a. Dans la feuille *Statistiques descriptives*, commencez par préciser pour chaque variable si elle est de type « qualitatif » ou « quantitatif ».

1.b. Dans la feuille *Statistiques descriptives*, déterminez une formule, que vous pourrez étirez dans les cellules B4:H4 telle que :

- Si une variable est de type quantitatif, la valeur retournée soit la valeur au-delà de laquelle les données seront jugées extrêmement hautes
- Si une variable est de type qualitatif, la réponse soit : « / »

**Pour rappel** : on considère qu'une valeur est extrêmement haute si elle se situe à  $Q3 + (1.5 * \text{écart interquartile})$ .

1.c. Déterminez une formule, que vous pourrez étirer dans les cellules B5:H5, qui retournera, pour chaque variable, le nombre de valeurs de la base de données qui sont supérieures à la limite supérieure déterminée au point 1.b.

**Astuce** : utiliser du connecteur « & ».

1.d. Dans la feuille *Statistiques descriptives*, déterminez une formule, que vous pourrez étirez dans les cellules B6:H6 telle que :

- Si une variable est de type quantitatif, la valeur retournée soit la valeur en-deçà de laquelle les données seront jugées extrêmement basses
- Si une variable est de type qualitatif, la réponse soit : « / »

**Pour rappel** : on considère qu'une valeur est extrêmement basse si elle se situe à  $Q1 - (1.5 * \text{écart interquartile})$ .

1.e. Déterminez une formule, que vous pourrez étirer dans les cellules B7:H7, qui retournera, pour chaque variable, le nombre de valeurs de la base de données qui sont inférieures à la limite inférieure déterminée au point 1.d.

**Astuce** : utiliser du connecteur « & ».

2. Ouvrez le fichier *Entraînement.tennis.xlsx*.

2.a. Dans la cellule B33 de la feuille *Données brutes*, calculez le mode de la série, en utilisant les labels de la feuille *Légendes* (la réponse attendue n'est donc pas un chiffre).

2.b. Dans la feuille *Tableau de fréquences*, construisez un tableau de fréquences complet de la série.