

Moyennes, mode, médiane, quartiles

On ne va pas travailler ici avec des séries chronologiques (les valeurs prises par ce qui nous intéresse à différentes époques), mais avec des distributions. Une distribution est un ensemble de valeurs prises par un caractère chez différents individus. On prend ici « individus » au sens statistique, qui est très général. Cela peut être les étudiants d'un TD (on peut étudier la distribution de leurs notes), mais aussi des pays (on peut étudier la distribution de leurs richesses), des entreprises (on peut étudier la distribution de leurs effectifs d'employés)...

Ce qui va nous intéresser, c'est de résumer par quelques indicateurs les grandes caractéristiques d'une distribution. On peut se demander si elle est plus ou moins hétérogène (forts écarts de notes dans le TD, forts écarts de richesses entre pays...). Cela nécessite des calculs un peu compliqués que l'on verra sur la fiche « variance et écart-type ». Mais on peut d'abord plus simplement caractériser la distribution par un nombre représentatif : mode, médiane ou moyenne, qui va nous donner une idée, très simplifiée, de l'ordre de grandeur des valeurs présentes dans la série.

Par exemple, vous connaissez déjà bien la moyenne la plus simple, que l'on appelle moyenne arithmétique (la somme de toutes les valeurs divisée par le nombre d'individus). Si la moyenne des notes du TD est de 10, cela ne me dit pas grand chose sur la note de chacun et sur l'hétérogénéité ; mais si je compare avec des matières où la note est de 6 ou de 14, cela me donne déjà une indication d'ensemble sur le niveau des étudiants du TD dans la matière considérée. Il existe d'autres indicateurs que la moyenne arithmétique, qui remplissent à peu près la même fonction, avec quelques nuances.

La moyenne arithmétique

Elle est très facile à calculer en Excel. Il suffit de taper

`=MOYENNE(ici sélectionner le bloc de cellules comprenant les valeurs du caractère étudié)`

Par exemple dans la situation suivante :

Nom de l'étudiant	Note
Aaaa	12
Bbbb	9
Cccc	13
Dddd	7

Pour faire la moyenne des notes, il faudra sélectionner les quatre cellules qui contiennent ces notes, dans la deuxième colonne.

Il peut arriver que certaines valeurs soient manquantes au sein du bloc de cellules. Par exemple un étudiant absent n'a pas de note : dans la case qui lui correspond, il va y avoir l'indication « absent », ou bien rien du tout. L'avantage d'Excel est qu'il ne tient compte dans le calcul que des nombres (il ignore les mots et il ne remplace pas les cellules vides par un zéro). On peut donc sélectionner le bloc de cellules qui nous intéresse, même s'il comprend des « blancs ».

La moyenne pondérée

On l'a déjà vue à propos des indices synthétiques : c'est par exemple la moyenne de vos notes au bac, pondérée par les coefficients de chaque matière.

Elle ne s'automatise pas en Excel. Il faut taper toute la formule :

`=(valeur1*coeff1+valeur2*coeff2+valeur3*coeff3+...)/(coeff1+coeff2+coeff3+...)`

Ou bien on peut faire cela de façon un peu plus rusée, si on a de nombreux termes :

- premièrement, remplir une nouvelle colonne (ou ligne) avec les résultats de valeur1*coeff1 ; puis valeur2*coeff2 ; puis valeur3*coeff3, etc.
- deuxièmement, faire la somme des valeurs de cette colonne, à l'aide de la fonction =SOMME ou du bouton correspondant (qui porte le signe Σ).
- troisièmement, faire la somme des coefficients, de tête ou avec le même bouton.
- quatrièmement, faire la division.

Contrairement aux apparences, cela va plus vite... Exemple :

Note	Coeff	Nouvelle colonne
5	1	=(cellule contenant « 5 » * cellule contenant « 1 ») résultat : 5 (égal à valeur1*coeff1)
13	3	Recopier la formule ci-dessus vers le bas résultat : 39
8	4	Recopier la formule ci-dessus vers le bas résultat : 32
12	3	Recopier la formule ci-dessus vers le bas résultat : 36
	=SOMME(<i>les quatre cellules ci-dessus</i>) résultat : 11 (somme des coeff)	=SOMME(<i>les quatre cellules ci-dessus</i>) résultat : 112 (somme des valeur*coeff)

Pour connaître la moyenne pondérée, il ne reste plus qu'à diviser 112 par 11 (soit une moyenne légèrement supérieure à 10).

A noter :

Si on vous donne non pas les notes obtenues par chaque étudiant, l'un derrière l'autre, mais un tableau d'effectifs, comme ceci :

Note	Effectif des étudiants ayant obtenu cette note
5	1
8	3
10	4
12	3

On peut calculer la moyenne du TD exactement comme une moyenne pondérée, les effectifs servant de coefficients. Donc on peut utiliser la méthode ci-dessus.

Le mode

Le mode est simplement la valeur la plus fréquente dans une distribution, celle qui caractérise le plus grand nombre d'individus. Il peut y avoir plusieurs modes.

Exemples :

Note	3	6	9	10	11	14	16	18
Effectif	1	3	5	6	5	3	1	1

Le mode est de 10 : c'est la note la plus fréquente, celle qui a été obtenue par le plus grand effectif d'étudiants.

Note	3	6	9	10	11	14	16	18
Effectif	1	3	5	5	4	3	1	1

Ici, il y a deux modes : 9 et 10.

Si vous avez une longue liste de valeurs, individu par individu (par exemple les notes de tous les étudiants de l'université, en face de leurs noms), il peut être intéressant de faire rechercher le mode automatiquement par Excel. La formule est tout simplement :

=MODE(*bloc de cellules concerné*)

La médiane

La moyenne a un gros défaut : elle est très sensible aux valeurs extrêmes. Par exemple, si une seule personne dans le TD obtient 0 ou 20, alors que les autres ont des notes autour de 10, elle va fortement peser sur la moyenne. On aura alors l'impression que le TD est « plutôt faible » ou « plutôt fort » dans la matière, à cause de cette seule personne. La médiane est un indicateur qui permet au contraire de ne pas faire attention aux valeurs extrêmes. De ce fait, on l'utilise en complément de la moyenne.

La médiane est la valeur du caractère (par exemple la note, le PNB, le nombre d'ouvriers...) telle qu'on ait la moitié des individus en dessous de cette valeur et la moitié au-dessus.

Exemple :

Note	1	8	9	10	11	12
Effectif	1	3	5	6	5	3

La moyenne de ce TD est de 9,6. Si l'étudiant qui a obtenu 3 quitte le TD, la moyenne des autres monte à 10.

L'effectif total du TD est de 23 personnes. Donc, si l'on range les notes par ordre croissant, la 12^e personne sera « au milieu » : 11 personnes auront eu une moins bonne note et 11 personnes une meilleure note. En l'occurrence, en regardant le tableau, on voit que la 12^e personne a obtenu la note de 10. La médiane du TD est donc de 10.

Si l'étudiant qui a obtenu 3 quitte le TD, l'effectif passe à 22, nombre pair. La médiane va donc se trouver entre la note de la 11^e personne et celle de la 12^e personne. Comme la 11^e et la 12^e personne ont obtenu la note de 10, la médiane est toujours de 10. L'existence de quelques individus se situant en position « extrême » n'a en effet pas d'influence sur elle.

A noter : si on a un effectif pair, la médiane va se trouver entre les valeurs caractérisant deux individus. Si elles sont égales, comme ci-dessus, le cas est simple. Si elles sont différentes, par convention, la médiane est égale à la moyenne arithmétique de ces deux valeurs.

Exemple :

Note	8	9	10	11	12
Effectif	3	8	5	5	1

Ici, on a de nouveau un effectif de 22 personnes. La médiane se situe entre la 11^e et la 12^e. Mais la 11^e personne a obtenu 9 et la 12^e personne a obtenu 10. Par convention, la médiane est de 9,5.

De toute façon, en Excel, tout est plus simple : si vous disposez d'un bloc de cellules avec les valeurs caractérisant chaque individu, il suffit, comme pour la moyenne, de taper :

=MEDIANE(*sélectionner ce bloc de cellules*)

Là encore, Excel ne tient pas compte des valeurs manquantes.

Mode, moyenne et médiane

Il peut être intéressant de connaître ces trois indicateurs à la fois, car, ensemble, ils permettent déjà d'obtenir certaines précisions sur une distribution. En particulier, comme la moyenne est « tirée » vers les valeurs extrêmes, on sait que :

- si la moyenne est beaucoup plus basse que la médiane, quelques individus ont des valeurs de caractère beaucoup plus basses que l'ensemble des autres.
- si la moyenne est beaucoup plus haute que la médiane, quelques individus ont des valeurs de caractère beaucoup plus hautes que l'ensemble des autres.

Exemple :

Dans une entreprise, la moyenne mensuelle des salaires est de 9 300 F. Mais le mode est de 6 000 F et la médiane est de 7 000 F. Comment est-ce possible ? On a en fait la distribution suivante, très inégale :

Salaire	6000	7000	8000	60000
Effectif	10	5	5	1

Les quartiles

On peut généraliser le principe de la médiane. Au lieu de couper la distribution en deux et de regarder qui est « à la frontière », on peut la couper en 4, en 10, ou en n'importe quel nombre. On parle alors de quartiles, de déciles, de centiles (si on coupe en 100)... les quartiles étant les plus utilisés.

Le premier quartile va donc être la valeur telle que 25% (le quart) des individus soient « en dessous » et 75% « au-dessus ». Le deuxième quartile, qui coupe en deux moitiés, n'est autre que la médiane : c'est son autre nom. Le troisième quartile est la valeur telle que 75% des individus soient « en dessous » et 25% « au-dessus ».

Excel permet encore très simplement d'obtenir les quartiles si l'on connaît les valeurs pour chaque individus : le premier quartile Q1 s'obtient par :

=QUARTILE(*bloc de cellules comprenant les valeurs du caractère étudié*;1)

Le deuxième quartile est la médiane ; le troisième, Q3, s'obtient par :

=QUARTILE(*bloc de cellules comprenant les valeurs du caractère étudié*;3)

On appelle « écart interquartile » l'écart entre le premier et le troisième quartile, soit $Q3 - Q1$. Il donne une première idée de l'hétérogénéité de la distribution.

Exemple :

Note	6	9	11	13
Effectif	8	15	12	5

Ici l'effectif est de 40, donc on peut repérer facilement les quartiles :

$Q1 = 9$, $Q2 = \text{médiane} = 9$, $Q3 = 11$

En effet, 10 personnes ont moins de 9 et 30 ont plus ; 20 personnes ont moins de 9 et 20 ont plus ; 30 personnes ont moins de 11 et 10 ont plus.

Si on compare maintenant avec la distribution suivante :

Note	4	7	9	11	13	15	18
Effectif	6	5	10	6	2	7	4

$Q1 = 7$, $Q2 = 9$, $Q3 = 15$. Avec la même médiane, on a un écart interquartile de 8 points, contre 2 points dans le premier cas, qui signale l'hétérogénéité de la distribution.