Université de TOURS - L1 Gestion

Cours Mathématiques et Statistiques Appliquées à la Gestion

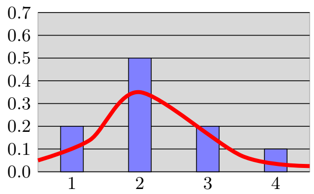
Corrigé du TD n° 8

Distributions et statistiques

Automne 2021

Le cours viendra après les TD clore le travail des étudiants. Exceptionnellement, les chargés de TD développeront les savoirs qui sont repris dans l'encadré, et qui correspondent à la première partie du chapitre V, c'est-à-dire la définition d'une distribution, discrète ou continue, l'usage de statistiques pour illustrer/représenter ces distributions, et les critères de dominance stochastique du 1er et du 2nd ordre.

On appelle distribution d'une variable la description de l'ensemble des occurences possibles de cette variable et de la fréquence d'apparition de ces occurences. On distingue les distributions discrètes dans lesquels les évènements sont finis, chacun associés avec une probabilité, des distributions continues, où ce sont les probabilités des intervalles qui sont définies.



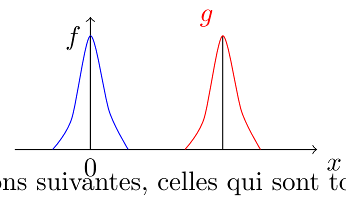
Dans le diagramme ci-dessus les fréquences sont reportées sur l'axe vertical, en tuyau d'orgue une distribution discrète, en rouge, une distribution continue.

Les statistiques de position, le *mode*, la *médiane* et la *moyenne* donnent une idée de l'ordre de grandeur de la distribution considérée. le mode est la valeur qui correspond à l'effectif le plus élevé, la médiane, est la valeur qui partage la distribution en deux sous-ensemble d'effectifs égaux. La comparaison de ces 3 caractéristiques de tendance centrale permet de juger de la symétrie de la distribution. Lorsqu'une distribution est telle que le mode est inférieur à la médiane (elle-même inférieure à la moyenne), la distribution est dite *étalée vers la droite* et vice versa.

On dit que la distribution $\mathcal{A}$ domine stochastiquement la distribution $\mathcal{B}$ au premier ordre, si elle rémunère plus que $\mathcal{B}$ dans tous les états de la nature. On dit que la domination est du second ordre, si $\mathcal{B}$ est un spread à moyenne constante de la distribution $\mathcal{A}$.
On dit que $\mathcal{B}$ est un *spread* de $\mathcal{A}$ à moyenne constante si dans la distribution $\mathcal{B}$, on a déplacé du poids du centre de la distribution vers les queues de la distribution.

1 Deux distributions suite à un transfert d'argent

Soit les deux distributions f et g représentées ci-dessous resp. en bleu et en rouge, vérifiant la propriété (0) :



$$\exists\ a > 0\ /\ g(x) = f(x-a) \qquad (0)$$

Dire des affirmations suivantes, celles qui sont toujours Vraies **(V)**, toujours Fausses **(F)** ou, Parfois vraies ou fausses **(P)** ?

1) L'espérance de la distribution de f est toujours supérieure à l'espérance de la distribution de g . . . . **V** ☐ **F** ☐ **P**

L'espérance de la distribution de f est toujours inférieure à l'espérance de la distribution de g . . . . . **V** ☐ **F** ■

FAUX. Les réalisations de f sont toutes sans ambiguïté inférieures au réalisations de g. Aussi, la moyenne de f est inférieure à la moyenne de g.

2) 0 est l'espérance de f . . . . **V** ☐ **F** ☐ **P**

0 est l'espérance de f . . . . **V** ☐ **F** ☐

PARFOIS. Si la distribution était symétrique, tel que c'est suggéré par le graphique, la moyenne serait zéro. Mais le graphique n'est pas une information suffisante pour dire que c'est toujours vrai.

3) 0 est le mode de f . . . . **V** ☐ **F** ☐ **P**

0 est le mode de f . . . . **V** ■ **F** ☐

VRAI. Zéro est le point où la distribution f est la plus haute.

4) 0 est la médiane de f . . . . **V** ☐ **F** ☐ **P**

0 est la médiane de f . . . . **V** ☐ **F** ☐

PARFOIS. Si la distribution était symétrique, tel que c'est suggéré par le graphique, la médiane serait zéro. Mais le graphique n'est pas une information suffisante pour dire que c'est toujours vrai.

5) a est l'espérance de g. . . . . **V** ☐ **F** ☐ **P**

a est l'espérance de g . . . . **V** ■ **F** ☐

OUI. le graphique suggère qu'il existe une translation entre les deux distributions, c'est ce qu'exprime l'équation

(0). Cela veut dire que le poids de la distribution f est translaté point par point d'une distance a à la distribution g. Si on appelle X la réalisation de f alors, la variable corrélée $Y = X + a$ a pour distribution g. La moyenne de g est la moyenne de f plus a : $Eg = Ef + a$.

6) La variance de f égale la variance de g. . . . . **V** ☐ **F** ☐ **P**

La variance de f égale la variance de g. . . . . **V** ■ **F** ☐

OUI. Si on poursuit l'argument précédent, Si on appelle X la réalisation de f alors, la variable corrélée $Y = X + a$ a pour distribution g. La moyenne de g est la moyenne de f plus a : $Eg = Ef + a$. Ainsi, la distribution de la distance à la moyenne dans la distribution g est la distribution f. Ainsi $VAR(f) = VAR(g)$.

7) La distribution g domine stochastiquement la distribution f au premier ordre. . . . . **V** ☐ **F** ☐ **P**

La distribution g domine stochastiquement la distribution f au premier ordre. . . . . **V** ■ **F** ☐

OUI. L'argument simple est de dire que toute réalisation de g est supérieur à n'importe quelle réalisation de f. Bien entendu, la translation analysée ci-dessus permet de développer un autre argument : pour chaque état de la nature, à X on associe $X + a$, la réalisation de g est plus élevée que la réalisation de f.

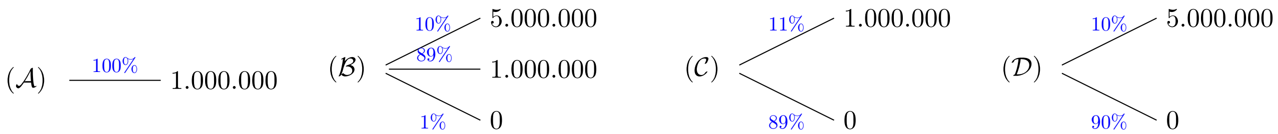
8) La distribution g domine stochastiquement la distribution f au second ordre. . . . . **V** ☐ **F** ☐ **P**

La distribution g domine stochastiquement la distribution f au second ordre. . . . . **V** ☐ **F** ■

NON. f n'est pas un spread de g. Car sinon, il faudrait qu'il y ait dans la distribution de g du poids sur des réalisations négatives.

2 Dominance stochastique de premier et de second ordre

On considère dans cet exercice les quatre distributions $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ et $\mathcal{D}$ suivantes :



1) Dire s'il existe entre ces distributions -prises deux à deux- des liens de dominance stochastique de premier ordre

A - B : Pas de dominance au 1er ordre, car l'unique gain de A n'est pas supérieur à tous les gains qu'on peut obtenir avec B, ou n'est pas inférieur à tous les gains qu'on peut obtenir avec B.

A - C : Dominance de 1er ordre de A qui paye toujours plus ou égal que C.

A - D : Pas de dominance stochastique de premier ordre, comme dans le cas A - B.

B - C : B domine C au premier ordre il suffit d'écrire les deux distributions comme des arbres à 4 branches.

B - D : B domine D au premier ordre, dont l'état le plus bas se traduit par une meilleure conjoncture dans la distribution B

C - D : Il n'y a pas de dominance stochastique de premier ordre de la D, car même si l'état haut rémunère mieux (5M $>$ 1M) il rémunère mieux moins souvent (10% $<$ 11%).

2) Dire s'il existe entre ces distributions -prises deux à deux- des liens de dominance stochastique de second ordre

A - B : Pas de dominance au 1er ordre, car l'unique gain de A n'est pas supérieur à tous les gains qu'on peut obtenir avec B, ou n'est pas inférieur à tous les gains qu'on peut obtenir avec B. Pour la dominance de second ordre, il faut déjà voir l'espérance de B. $E[B] = \frac{89}{100} 1.000.000 + \frac{10}{100} 5.000.000 = 1.390.000$. B est un spread de A, mais à moyenne plus élevée. Donc, on ne peut pas dire que A domine B au sens de la dominance stochastique de second ordre ;

A - C : Dominance de 1er ordre de A qui paye toujours plus ou égal que C.

A - D : D est un spread de A. Par ailleurs $E[D] = 500.000$, la moyenne de D est plus faible que A. Donc D est dominé par A au sens de la dominance stochastique de 2nd ordre.

B - C : B domine C au premier ordre il suffit d'écrire les deux distributions comme des arbres à 4 branches.

B - D : B domine D au premier ordre, dont l'état le plus bas se traduit par une meilleure conjoncture dans la distribution B

C - D : Il n'y a pas de dominance stochastique de premier ordre de la D, car même si l'état haut rémunère mieux (5M > 1M) il rémunère mieux moins souvent (10% < 11%). Cependant, D est un spread de C. On sait déjà que $E[D] = 500.000$**. Par ailleurs** $E[C] = 110.000$**. Il n'y a donc pas de dominance stochastique au second ordre.**

3) Dire intuitivement pourquoi tous les agents rationnels préfèrent une distribution $\mathcal{A}$ à la distribution $\mathcal{C}$ quand $\mathcal{A}$ domine stochastiquement $\mathcal{C}$ au premier ordre.

Dans le cas de distributions discrètes, quand on compare deux distributions, on définit des états de la nature, associés à une probabilité, sur lesquels les réalisations des deux distributions sont constantes. Si on peut définir un système d'états de la nature, tel que une distribution paye toujours plus que l'autre distribution, alors, si les agents ont des préférences sur les distributions, ils choisiront toujours celle qui paye plus, quel que soit l'état de la nature qui se réalise.

4) Justifier, lorsque l'on considère une distribution $\mathcal{L}$ dans laquelle il y a de l'alea avec la distribution $\mathcal{S}$ qui consiste à donner de manière sûre la moyenne de $\mathcal{L}$, que $\mathcal{L}$ est un spread à moyenne constante de $\mathcal{S}$.

Le premier critère d'une moyenne constante est vérifié, puisque la moyenne de $\mathcal{S}$ **est par définition la moyenne de** $\mathcal{L}$**.**

Le second critère de la nouvelle répartition du poids de la distribution, en partant de $\mathcal{S}$**, en enlevant du poids centre, qui dans ce cas est uniquement un singleton, et en le répartissant de manière égale (puisque la moyenne est identique) vers la gauche et vers la droite est aussi vérifié.**

En d'autres termes, cette question permet de souligner que toute distribution est un spread à moyenne constante de la distribution dégénérée qui propose sa moyenne.

5) Dire les raisons pour lesquelles un agent pourrait ne pas aimer les spread à moyenne constante.

On peut comprendre intuitivement un spread à moyenne constante comme augmentant le risque d'une distribution. Quand on dit qu'un agent n'aime pas le risque cela pourrait en particulier se traduire par le fait de ne pas aimer les spread à moyenne constante.

3 Distribution des Rendements de deux titres financiers

Un investisseur aimerait savoir lequel des deux titres A et B est le plus risqué. En gestion de portefeuille, le risque d'un titre étant souvent caractérisé par la variabilité de son rendement, il doit comparer la variance des rendements des deux titres. À cette fin, il tire un échantillon aléatoire de 8 rendements semestriels pour chacun des deux titres et il obtient comme

Rendements trimestriels								
Titre A	4,2 %	-0,6 %	2,3 %	5,1 %	4,3 %	4,9 %	7,4 %	3,8 %
Titre B	2,5 %	2,2 %	2,8 %	3,0 %	2,8 %	2,6 %	2,8 %	3,2 %

1) Calculer pour chacun de ces titres le rendement moyen

Pour le titre A, on a

$$E[R_A] = \frac{4,2 - 0,6 + 2,3 + 5,1 + 4,3 + 4,9 + 7,4 + 3,8}{8} = 3,925\%$$

Pour le titre B, on a

$$E[R_B] = \frac{2,5 + 2,2 + 2,8 + 3,0 + 2,8 + 2,6 + 2,8 + 3,2}{8} = 2,7375\%$$

2) Calculer pour chacun de ces titres la variance des rendements

On fait le tableau de la distance à la moyenne au carré pour les deux titres :

Rendements trimestriels								
$(R_i^A - ER^A)^2$	**0,51**	**16,72**	**1,41**	**2,60**	**0,66**	**1,99**	**15,30**	**0,10**
$(R_i^B - ER^B)^2$	**0,00**	**0,05**	**0,13**	**0,32**	**0,13**	**0,03**	**0,13**	**0,59**

Pour le titre A, on a

$$VAR[A] = \frac{0,51 + 16,72 + 1,41 + 2,60 + 0,66 + 1,99 + 15,30 + 0,10}{8} \approx 4,91(\%)^2 = 0,0491\%$$

Pour le titre B, on a

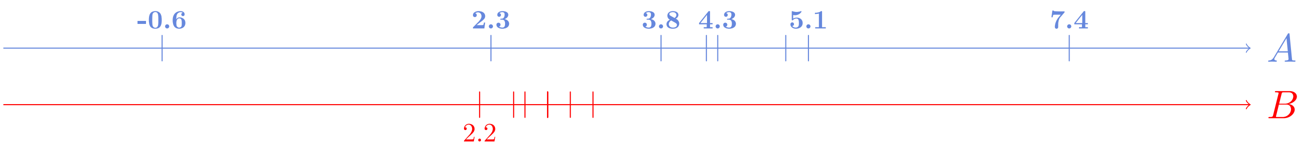
$$VAR[B] = \frac{0,00 + 0,05 + 0,13 + 0,32 + 0,13 + 0,03 + 0,13 + 0,59}{8} = 0,1725(\%)^2 = 0,001725\%$$

3) Dire sur quel critère on pourrait préférer le titre A et sur quel critère on pourrait préférer le titre B

Le titre A a un meilleur espérance de rendement, mais c'est au prix d'une variance beaucoup plus grande.

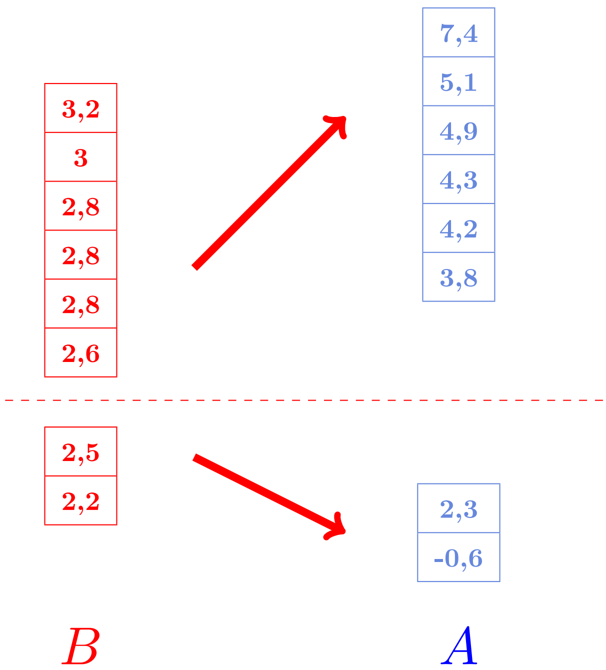
4) Est-ce que le titre B est un spread du titre A ?

Commençons par mettre en regard les différentes réalisations de A et b

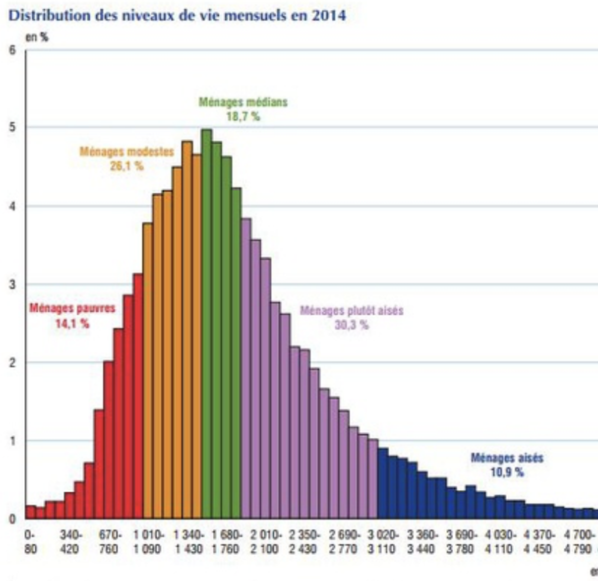


Cette première représentation laisse penser que A est un spread de B ; en effet, mis à part 2.2, toutes les réalisations de B sont comprises entre 2.3 et 3.8.

On est dans un cas assez aisé dans lequel chaque point vaut le même poids. En partant des réalisations de B, on fait un tableau qui envoie ces réalisations vers les réalisations plus hautes et plus basses de A. Il apparaît alors que A est un spread de B.



4 Ménages français en 2014



Ci-contre la distribution des revenus des ménages français en 2014, en tuyau d'orgue. Sur l'axe horizontal, chaque tranche de revenu est de 80, sur l'axe vertical le pourcentage de chacune des tranches.

1) Expliquer l'appellation ménages médian pour les tranches de revenu représentées en vert.

2) Donner une approximation des trois quartiles $Q1$, $Q2$ et $Q3$.

3) Développer votre intuition, l'espérance est-elle ci-contre inférieure à la médiane ? Étayer votre propos par une recherche internet.

Réponse à la question 1

La probabilité des ménages dont les revenus sont supérieurs aux tranches vertes est de $30,3 + 10,9 = 41,2\%$. La probabilité des ménages dont les revenus sont inérieurs aux tranches vertes est de $14,1 + 26,1 = 40,2\%$. Il y a à peu près autant de poids réparti de part et d'autre des ménages dont les revenus sont dans les tranches vertes, ce qui justifie l'appellation de tranches de revenu médianes.

Réponse à la question 2

Pour $Q1$, il faut les premiers 25% de la population. Celà comprend les ménages pauvres (14,1%) plus 10,9% des « premiers » ménages modeste. Ces ménages modestes sont répartis en 6 tranches de probabilité respectives et approximatives de 3,8%, 4,15%, 4,2%, 4,5%, 4,8% et 4,65% (la somme étant de 26,1%). Il y a donc les deux premières tranches de poids agrégé 7,95% et moins que les trois premières tranches de poids agrégé 12,15%. On doit donc prendre, pour trouver ce premier quartile les deux premières tranches des ménages modestes plus une fraction de la troisième tranche, dont le poids soit exactement $10,9 - 7,95 = 2,95\%$. Or cette troisième tranche est de $4,2\%$, on veut donc en prendre la fraction $\dfrac{2,95\%}{4,2\%} = 0,70 = 70\%$.

Les revenus de cette troisième tranche de revenus modeste, cad de cette 15 ième tranche de revenu sont compris entre $15 * 80$ et $16 * 80$, cad inclus dans l'intervalle $[1200, 1280]$.

Sous l'hypothèse que les revenus de cette troisième tranche se répartissent de manière uniforme, on peut dire que les 70% premiers sont compris entre 1200 et $1200 + 0,70 * 80 = 1256$

$$Q1 = 1256$$

Pour $Q2$, il faut les premiers 50% de la population. On sait d'après la première question que la médiane sera dans les ménages médian. Pour aller jusqu'à $Q2$, on comprend les ménages pauvres (14,1%), les ménages modestes 26,1% plus 9,8% des « premiers » ménages médians. Ces ménages modestes sont répartis en 4 tranches de probabilité respectives et approximatives de 4,95%, 4,75%, 4,65% et 4,35% (la somme étant de 18,7%). Il faut donc donc les deux premières tranches de poids agrégé 9,7% et 0,1% de la troisième tranche de poids 4,65%. On doit donc prendre, pour trouver la médiane les deux premières tranches des ménages médians plus une fraction de la troisième tranche, de poids 0,1%. Or cette troisième tranche est de $4,65\%$, on veut donc en prendre la fraction $\dfrac{0,1\%}{4,65\%} = 2,15\%$.

Les revenus de cette troisième tranche de revenus médians, cad de cette 21 ième tranche de revenu sont compris entre $20 * 80$ et $21 * 80$, cad inclus dans l'intervalle $[1600, 1680]$.

Sous l'hypothèse que les revenus de cette troisième tranche se répartissent de manière uniforme, on peut dire

que les 2,15% premiers sont compris entre 1200 et $1200 + 0,0215 * 80 = 1201,7$

$$Q2 = 1202$$

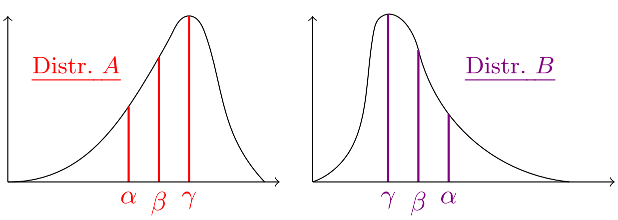
Pour $Q3$, il faut les premiers 75% de la population. Celà comprend les ménages pauvres (14,1%), les ménage modestes (26,1%), les ménages médians (18,7%) plus 16,1% des « premiers » ménages plutôt aisés. Ces ménages plutôt aisés sont répartis en 14 tranches de probabilité respectives et approximatives de 3,8%, 3,5%, 3,3%, 2,7%, 2,6%, 2,25%, 2,2, ... (la somme étant de 30,3%). Il y a donc les cinq premières tranches de poids agrégé 15,9% et moins que les six premières tranches de poids agrégé 18,15%. On doit donc prendre, pour trouver ce premier quartile les cinq premières tranches des ménages plutôt aisés plus une fraction de la sixième tranche, dont le poids soit exactement $16,1 - 15,9 = 0,2\%$. Or cette sixème tranche est de $2,25\%$, on veut donc en prendre la fraction $\dfrac{0,2\%}{2,25\%} = 8,88\%$.
Les revenus de cette sixième tranche de revenus modeste, cad de cette 28 ième tranche de revenu sont compris entre $27 * 80$ et $28 * 80$, cad inclus dans l'intervalle $[2160, 2240]$.
Sous l'hypothèse que les revenus de cette sixième tranche se répartissent de manière uniforme, on peut dire que les 8,8% premiers sont compris entre 2160 et $80 * 27,088 = 2\ 167,04$

$$Q3 = 2\ 167$$

Réponse à la question 3

La distribution est étalée quand il s'agit des hauts revenus, ce qui est un signe de concentration sur les petits revenus (comme dans la distribution B de l'exercice ci-après). La médiane se situe au-dessus du mode, donc dans les tranches de revenus des ménages plutôt aisés. La moyenne est influencée par les revenus aisés et devrait être supérieur à la médiane.
Pour l'application numérique, on sait que $Q2 \approx 1.202$. Sur internet, on trouve que l'insee publie une étude, le salaire moyen est de 2.225 euros nets des prélèvements à la source..
par exemple sur le site : https ://www.insee.fr/fr/statistiques/2121609

5 Espérance, Mode, Médiane



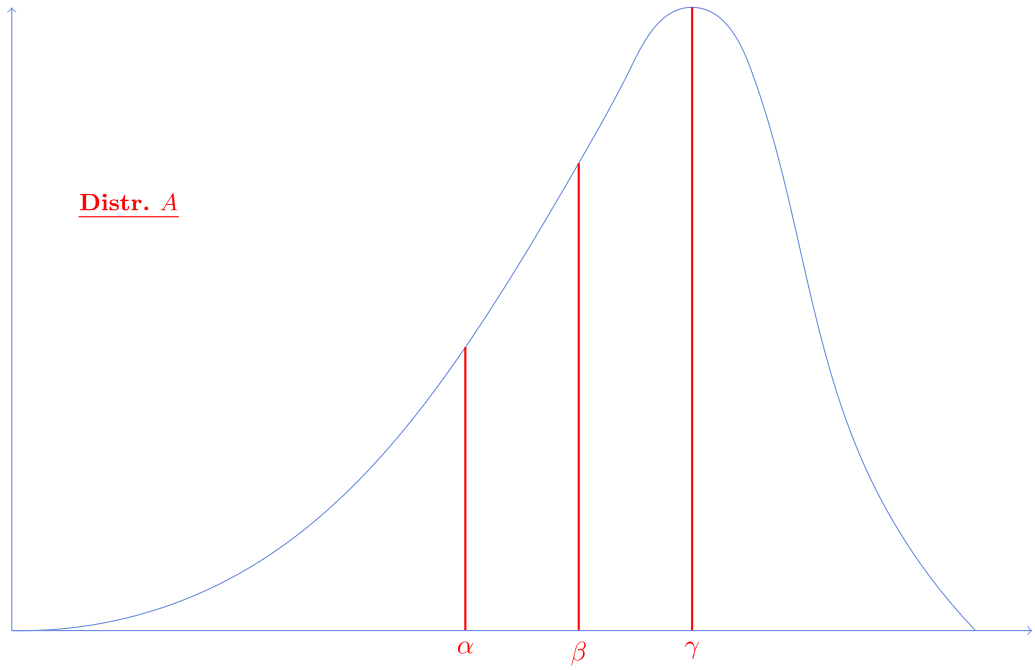
Soit les deux distributions unimodales ci-contre où l'on a représenté l'espérance, le mode, la médiane.

1) Indiquer (intuitivement) en A et B si espérance $<$ médiane

2) Avec Tableur trouver $x_1 < x_2 < x_3 < x_4 < x_5 < x_6 < x_7$ tels qu'associés aux probas respectives $\frac{3}{20}; \frac{1}{20}; \frac{5}{20}; \frac{2}{20}; \frac{3}{20}; \frac{4}{20}; \frac{2}{20}$ la distribution correspondante vérifie Espérance<mode<médiane.

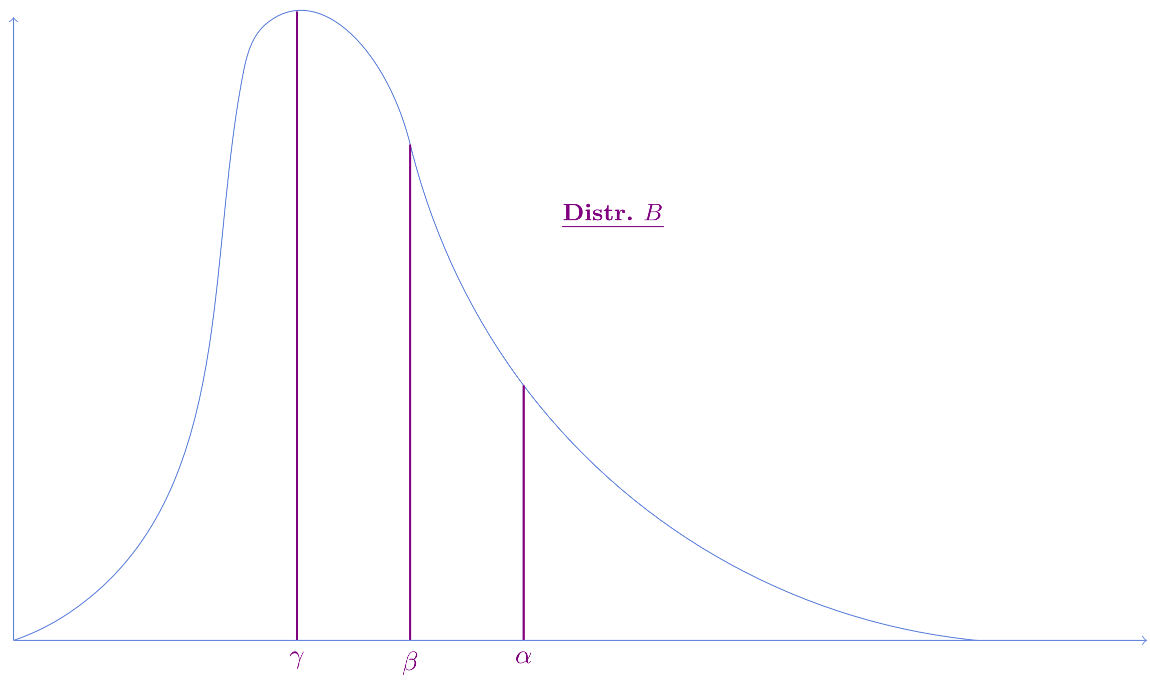
Réponse à la question 1

Si on regarde la distribution A, le fait qu'elle soit asymétrique à gauche indique que le poids est plutôt à gauche du mode. Ainsi, la médiane, sans vraiment d'ambiguité est à gauche du mode. Nommons les deux coordonnées susceptibles d'être la moyenne et la médiane



À la faveur de l'agrandissement, il semble que α ne peut pas être la médiane, et donc que β serait la médiane. Donc, par défaut, α serait la moyenne pour cette distribution.
La distribution A est une distribution asymétrique étalée vers la gauche (ou oblique à droite) ; elle est telle que la moyenne est inférieure à la médiane, ces deux caractéristiques étant elles-mêmes inférieures au mode.

Si on regarde la distribution A, le fait qu'elle soit asymétrique à gauche indique que le poids est plutôt à gauche du mode. Ainsi, la médiane, sans vraiment d'ambiguité est à gauche du mode. Nommons les deux coordonnées susceptibles d'être la moyenne et la médiane



À la faveur de l'agrandissement, il semble que α ne peut pas être la médiane, et donc que β serait la médiane. Donc, par défaut, α serait la moyenne pour cette distribution.
La distribution B est une distribution asymétrique étalée vers la droite (ou oblique à gauche) ; elle est telle que la moyenne est supérieure à la médiane, ces deux caractéristiques étant elles-mêmes supérieures au mode.
Réponse à la question 2

Si on prend des distributions bi-modales, on a plus de lattitude pour fabriquer un peu ce que l'on veut. Prenons l'arbre de probabilité suivant avec $x_1 < x_2 < x_3 < x_4 < x_5 < x_6$

Proba	**Réalisation**
$2/20$	x_7
$4/20$	x_6
$3/20$	x_5
$2/20$	x_4
$5/20$	x_3
$1/20$	x_2
$3/20$	x_1

Par construction, la médiane est x_4. La distribution est bi-modale, mais, par contruction, le mode est $x_3 < x_6$. Construisons une distributions, dont la moyenne est inférieure à x_3. On utilise Un tableur, en essayant de mettre des petites valeurs sur x_1 et x_2 et très peu de croissance à partir de x_3. On a par exemple la distribution suivante

i	1	2	3	4	5	6	7		
pi	0,15	0,05	0,25	0,10	0,15	0,20	0,10		
xi	0	0,05	0,101	0,102	0,103	0,104	0,105		
pi*xi	0	0,0025	0,02525	0,0102	0,01545	0,0208	0,0105	moyenne	0,0847
Mediane				x4=0,102					
Mode			x3=0,101						
Moyenne		x2 < E=0,0847 < x3							

FIN du corrigé du TD 8