Cours L3 MIAGE

Analyse statistique de données quantitatives réelles

On rappelle tout d'abord qu'une variable aléatoire est dite quantitative si elle prend des valeurs numériques sur lesquelles cela a un sens d'effectuer des opérations arithmétiques (somme, moyenne, variance, ...). Les données quantitatives sont des observations des valeurs x_1,...,x_n prises par cette variable aléatoire quantitative sur les n individus de l'échantillon.

Certains exercices utilisent la librairie lycee que vous pouvez télécharger en vous connectant au site de la MIAGE, dans les répertoires MEMBRES puis Sophie Jacquot. Une ancienne version est disponible à l'URL http://transdoc.univ-orleans.fr/get?k=8tZLAIWS5bUP6l8fI23.

A. Données quantitatives discrètes

On parle de données quantitatives discrètes lorsque l'on observe qu'un petit nombre (en général moins de 20) de modalités (valeurs) dans les observations.

On peut utiliser tous les outils de statistique descriptive décrits pour les données qualitatives ordinales:

- les modalités
- les effectifs
- les effectifs cumulés
- les fréquences
- les fréquences cumulées
- le mode: obtenu à partir des effectifs ou des fréquences
- la médiane
- les quartiles
- les modalités maximales et minimales

On peut représenter graphiquement la distribution (i.e. les fréquences) sous la forme de diagrammes en bâtons, de diagrammes en barres, ou de diagrammes en secteurs. On utilise le plus souvent un diagramme en bâtons pour pouvoir représenter les distances entre les modalités.

On représente également assez souvent la courbe des fréquences cumulées. Supposons que m_1<m_2<...<m_p sont les p modalités observées rangées par ordre croissant et f_1, .., f_p les fréquences associées. On note Fc(m_1), …,Fc(m_p) les fréquences cumulées définies par Fc(m_s)=f_1+...+f_s, pour tout s=1,...,p. Elles correspondent aux proportions de données qui sont inférieures ou égales à chaque modalité m_i.
La fonction des fréquences cumulées est définie de la manière suivante:
Fc(x)= 0 si x<m_1
Fc(m_s) si m_s<=x<m_{s+1}, s=1,..., p-1
1 si x>m_p
C'est une version empirique (i.e. calculée à partir des données) de la fonction de répartition P(X<=x) de la variable aléatoire X dont les données sont des observations.

Il est également intéressant de tracer la boîte à moustache associée aux données. Ce graphique permet de visualiser comment se situe 50% des observations autour de la médiane ainsi que les modalités minimales et maximales. Plus il a une forme allongée , plus les données sont dispersées. Les données distantes de la médiane de plus de 1.5 fois l'écart interquartile sont considérés comme des valeurs extrêmes et ne sont pas prises en compte pour tracer la boîte à moustache. Ces valeurs sont représentés par des points.

Lorsque l'on considère des données quantitatives, cela a un sens de considérer d'additionner ou de soustraire les modalités. On peut donc s'intéresser à d'autres caractéristiques des données que sont:

- la moyenne: moyenne arithmétique des données
  à partir de la série brute: $\bar{x}=(x_1+...+x_n)/n$
  à partir d'une représentation condensée: $\bar{x}=(m_1*e_1+...+m_p*e_p)/n=m_1*f_1+...+m_p*f_p$
- la variance: écart carré moyen entre les données et la moyenne
  à partir de la série brute: $V=((x_1-\bar{x})^2+...+(x_n-\bar{x})^2)/n=(x_1^2+...+x_n^2)/n-(\bar{x})^2$
  à partir d'une représentation condensée
  $V=((m_1-\bar{x})^2*e_1+...+(m_p-\bar{x})^2*e_p)/n=(m_1^2*e_1+...+m_p^2*e_p)/n-(\bar{x})^2$
- l'ecart-type: c'est la racine carré de la variance Sd= $V^{1/2}$
- l'étendue: écart entre la modalité maximale et la modalité minimale.
- l'espace interquartile: écart entre le 1er et le 3eme quartile.
- L'écart moyen à la moyenne: `(|x_1-`$\bar{x}$`|+...+|x_n-`$\bar{x}$`|)/n`
- l'écart moyen à la médiane: `(|x_1-med|+...+|x_n-med|)/n`

Remarques:

1. *Pour des raisons de qualité d'estimation on utilise souvent une version modifiée de V:*
   $V_2=((x_1-\bar{x})^2+...+(x_n-\bar{x})^2)/(n-1)=(x_1^2+...+x_n^2)/(n-1)-n(\bar{x})^2/(n-1)$

   C'est cette valeur que donne R.
2. *Le mode, la médiane et la moyenne sont des caractéristiques empiriques de centralité des données.*
3. *La variance, l'écart-type, l'étendue, l'écart moyen à la moyenne, l'écart moyen à la médiane et l'espace interquartile sont quant aux eux des indicateurs de la dispersion des données. Plus leurs valeurs sont grandes et plus les données sont dispersées. La variance et l'écart-type sont les indicateurs les plus couramment utilisés.*

Afin de mieux comprendre les notions présentées plus haut faire les exercices suivants:

EXERCICE 1:

Lors d'une étude statistique portant sur le nombre d'enfants des ménages résidant dans une zone géographique particulière, on recueille les données suivantes:

Nombre d'enfants	Effectifs observés
0	43
1	58
2	49
3	25
4	13
5	8
6	4

1. Créer un vecteur correspondant à ces données.
2. Représenter la distribution par un diagramme en bâtons et déterminer la valeur du mode. Utiliser range pour retrouver les modalités maximales et minimales.
3. Donner le tableau des fréquences et des fréquences cumulées. Tracer la courbe des fréquences cumulées.
4. En déduire la valeur des quartiles. Utiliser la commande boxplot.stats pour controler vos résultats
5. Déterminer la moyenne, la variance, et l'écart-type (à la main puis avec R)

EXERCICE 2:

Dans une classe de 20 étudiants, l'âge se répartit de la manière suivante:

Age	19	20	21	22	23
Effectif	1	3	9	5	2

1. Créer un vecteur correspondant à ces données.
2. Déterminer le mode, la médiane et la moyenne de ces données (à la main puis avec R)
3.Comment ces valeurs changent-elles si un étudiant de 45 ans vient rejoindre la classe? Commenter.

EXERCICE 3:
On étudie la fréquentation d'un guichet dans une grande banque au centre de Paris dans la but d'améliorer ce service. On fait une étude sur 2000 périodes de 2 minutes choisies sur la tranche horaire de 11h à 12h et on compte le nombre de personnes x_i entrant dans l'agence pendant la période i donnée. Les données recueillies se trouvent dans la librairie lycee dans le fichier guichet.

1. Charger ces données sur R. Créer un vecteur file contenant la colonne « queue » du fichier guichet. Déterminer la moyenne, le mode et la médiane de ces données.
2. Tracer la boîte à moustache associée à ces données, et déterminer les valeurs des quartiles, de la variance et de l'écart-type.
3. Construire le diagramme en bâton correspondant à la distribution des données. Représenter également la courbe des fréquences cumulées.
4. Proposer une modélisation de la distribution du caractère étudié. Représenter sur un même graphique la distribution empirique et la modélisation.

EXERCICE 4:
Une compagnie d'assurance a effectué une étude concernant le nombre d'accidents déclarés en 1998 par ses souscripteurs de police d'assurance « tous risques ». Le résultat est donné dans le tableau ci-dessous:

Nombre k d'accident déclarés	0	1	2	3
Effectifs	14698	1466	74	3

1. Créer un vecteur correspondant à ces données
2. Quelle est la population étudiée, la taille de l'échantillon, le caractère considéré, la nature de ce caractère?
3. Par quels graphiques peut-on représenter ces données. Faites-en deux.
4. Déterminer la moyenne empirique, la variance empirique, la médiane, le premier et le troisième quartile de ces données.
5. Par quel modèle probabiliste peut-on envisager de modéliser le nombre d'accidents?

EXERCICE 5:

On étudie la population des ménages d'un pays lointain. On observe le nombre d'enfants par ménages pour un échantillon de taille 1000 de cette population. On obtient les données suivantes:

Nb enfants	1	2	3	4	5	6	7	8	10
Effectif	510	248	132	55	29	15	2	8	1

1. Tracer la distribution empirique sous forme de diagramme en bâtons.
2. Calculer la moyenne empirique de ces données.
3. On voudrait voir si il est envisageable de modéliser la distribution du nombre d'enfants par une loi géométrique. Sachant que l'espérance d'une loi géométrique de paramètre p est donnée par (1-p)/p, on propose d'estimer p par (1-mean())/mean( ). Comparer distribution empirique et modélisation.

Cours L3 MIAGE

Analyse statistique de données quantitatives continues

Par simplicité, on ne traite dans ce cours que le cas où les données sont dans **IR**. Certaines des méthodes évoquées peuvent se généraliser à des espaces de dimension supérieure.

On parle de données quantitatives continues quand les données proviennent de l'observation d'une variable aléatoire quantitative discrète qui prend un grand nombre de valeurs ou bien de l'observation d'une variable quantitative continue pouvant prendre une infinité de valeurs (e.g. toutes les valeurs contenues dans un (ou plusieurs) intervalle de **IR)**.

D'autre part, il arrive assez souvent que l'on dispose de données regroupées par classes. Dans ce cas, les observations dont on dispose ne sont pas les valeurs exactes de la caractéristique que l'on étudie mais plutôt les classes (en général ce sont des intervalles de **IR**) auxquelles elles appartiennent. Il est bien évident que l'on disposerait de plus d'information à partir des valeurs prises par la variable sur l'échantillon. Cependant, on peut avoir privilégié ce type de données lors de l'enquête statistique par commodité ou bien parce que la variable étudiée est trop sensible (il est par exemple moins gênant de demander à des individus leur tranche de salaire que leur salaire précis). Il est important de bien choisir les classes considérées lors de l'enquête statistique car elles vont conditionner la pertinence de l'étude statistique que l'on pourra faire des données.

Nous allons distinguer dans ce qui suit le cas où l'on observe réellement les valeurs prises par notre variable (données brutes) de celui où l'on observe seulement la classe à laquelle elles appartiennent (données classées).

A. Lorsque l'on dispose des données brutes

La majorité des outils décrits plus haut pour les variables quantitatives discrètes peuvent être considérés. Cependant, comme l'exercice introductif suivant permet de le constater, la nature même des données nécessite de proposer d'autres méthodes pour représenter leur distribution.

EXERCICE (INTRODUCTION):

1. Lancer R et générer 150 données provenant d'une loi N(0,1).
2. Tracer la distribution avec un diagramme en bâtons.
3. Qu'observe-t-on? Commenter.
4. Quel est le mode de la variable.

On comprend bien à partir de l'exercice précédent qu'il nous faut proposer une autre manière de caractériser la distribution (effectifs et fréquences) de nos données et de définir la valeur du mode. En effet, il est assez rare, lorsque l'on étudie des données continues, d'observer plusieurs fois la même valeur sur l'échantillon. Lorsque les données proviennent réellement de l'observation d'une variable continue, les modalités observées dans les données sont généralement toutes différentes et si certaines valeurs sont répétées plusieurs fois, cela vient souvent du fait que l'on considère des valeurs arrondies et/ou du fait que les instruments de mesure sont gradués. Ce phénomène s'explique notamment par le fait que la probabilité que deux variables continues indépendants soient égales est nulle.

La fréquence associée à la valeur x est une version empirique de la probabilité P(X=x). Les variables aléatoires discrètes prennent un nombre fini ou au plus dénombrable de valeurs (modalités m_i). Leur distribution peut être caractérisée par les probabilités associées à chacune de ces modalités $P(X=m_i)$. Cela n'est plus vrai pour les variables continues pour lesquelles on a

P(X=x)=0 pour tout x. Par conséquent, cela n'a plus d'intérêt de considérer comme on l'a fait jusqu'à présent les différentes modalités puis les effectifs et fréquences qui leur sont associés. Pour la même raison, on doit proposer une nouvelle méthode pour déterminer le mode.

On peut cependant remarquer que les notions d'effectifs ou de fréquences cumulés sont encore pertinentes pour les données quantitatives continues. Représenter la courbe des fréquences cumulées donne une version empirique de la fonction de répartition $F(x)=P(X\leq x)$ (qui est une manière pertinente de caractériser la loi d'une variable continue).

La distribution d'une variable continue peut également être caractérisée par sa densité. Il s'agit de la dérivée (notée souvent f) de la fonction de répartition F. En tout point x, la densité correspond donc à la limite du quotient $P(x-\epsilon\leq X\leq x+\epsilon)/2\epsilon$ lorsque ϵ tend vers 0. Par conséquent, elle représente la densité de données que l'on peut s'attendre à observer autour de la valeur x. Il existe différentes méthodes permettant de construire des versions empiriques de f à partir des données brutes.
Nous ne verrons dans ce cours que la version la plus simple, appelée histogramme, qui consiste à utiliser la densité de données observées sur différentes classes. Cela revient donc à approcher la densité par une fonction constante sur chacune des classes.

CONSTRUCTION DE L'HISTOGRAMME ET DU MODE

Considérons tout d'abord M classes $(b_{0,}b_1),(b_{1,}b_2),\ldots,(b_{M-1},b_M)$ disjointes et observons les effectifs e_i (nombre de données appartenant à la classe) et fréquences f_i (proportion de données appartenant à la classe) qui leur sont associés. On a l'habitude de noter $c_j=(b_{j-1}+b_j)/2$ le centre de la classe j=1,...,M. La fréquence f_j associée à la classe (b_{j-1},b_j) est une version empirique de la probabilité $P(X\in(b_{j-1},b_j))$.
La notation (b_j,b_{j+1}) utilisée dans ce cours ne suppose pas que les classes ont une forme particulière (ouvertes ou fermées à droite et à gauche). Cela peut avoir une importance en pratique . Toutefois, si les données proviennent de l'observation d'une variable continue $P(X\in(b_{j-1},b_j))$ ne change pas que la classe contienne les extrémités ou pas car P(X=x)=0.

EXERCICE (FREQUENCE, DENSITE et CLASSE MODALE):
On réalise une enquête concernant les salaires. On obtient les données suivantes (les classes pouvant par exemple correspondre à des tranches d'impôt). On peut soit aboutir à ces données classées à partir des données brutes soit avoir trouvé plus simple de demander aux personnes interrogées à quelle tranche leur salaire appartenait plutôt que leur salaire exact.

Salaires	12000-14000	14000-16000	16000-21000	21000-26000	26000-36000	36000-50000
Effectif	50	190	210	40	4	2

1. Calculer les fréquences associées à ces classes. Dans quelle classe trouve-t-on la plus grande proportion de données?
2. Calculer les densités associées à ces classes. Dans quelle classe trouve-t-on la plus grande densité de données? Commenter.
3. Comment proposez-vous d'adapter la notion de mode?

On s'intéresse souvent non pas à la proportion de valeurs appartenant à une classe que modélise la fréquence mais plutôt à la « densité » de ces observations. Il est alors important, comme on l'a vu dans l'exercice précédent de relativiser la fréquence par l'amplitude de la classe. On associe à

chaque classe (b_{j-1}, b_j) sa densité (empirique) $d_j = f_j/(b_j - b_{j-1})$. On remarque que si les classes sont de même amplitude, les densités sont proportionnelles aux fréquences.
La distribution empirique est visualisée au travers d'un histogramme (dépendant des classes utilisées) représentant la densité de chaque classe par une barre de hauteur égale à sa densité (et donc d'aire égale à la fréquence associée). L'aire totale de l'histogramme est donc de 1 (c'est la somme des fréquences).

La notion de mode est remplacée par celle de classe modale. C'est la classe pour laquelle la densité est la plus forte.

Le passage des données brutes aux données regroupées par classes correspond à une perte d'information puisque seuls les effectifs des classes sont retenus. La manière dont on choisit les classes que l'on considère joue un rôle important dans la pertinence des résultats que l'on pourra obtenir par la suite. Le choix des extrémités des classes de fait à partir des données brutes. Le nombre M de classes doit être modéré (entre 5 et 12 en général). Le découpage en classe est souvent choisi de manière à obtenir des classes de même amplitude ou des classes de même effectifs (e.g. 10% de la population dans chaque classe). Enfin, l'étude statistique de ces données regroupées (on dira plutôt classées) est ensuite basée sur l'hypothèse que les données sont réparties de manière homogène dans chacune des classes (la densité empirique y est modélisée par une constante). Si ce n'est vraisemblablement pas le cas il faut peut être choisir un découpage plus approprié.

EXERCICE:
Charger les données robinet contenues dans la librairie lycee. Créer un vecteur cons contenant les consommations .

1. Donner les quartiles, la moyenne et la variance de ces données.
2. Tracer la boîte à moustache et la courbe des fréquences cumulées.
3. Représenter la distribution par un histogramme:
   - sans préciser d'option (ce sont en fait les effectifs que trace R)
   - afficher les densités avec l'option freq
   - fixer le nombre de classes à 5 avec nclass
   - utiliser les classes données par 0,50,100,200,350,550,800,1100,1450,3000
4. D'après la forme de l'histogramme on peut penser modéliser la consommation par une loi exponentielle. L'espérance d'une loi exponentielle de paramètre c étant 1/c on propose d'approcher le paramètre c de l'exponentielle par 1/mean(cons). Tracer avec lines la densité de la loi exponentielle de paramètre 1/mean(cons). Commenter.

B. Lorsque l'on dispose de données classées

Dans certains cas on ne dispose après l'enquête statistique que de données classées. Cela peut avoir été choisi par commodité ou bien parce que la variable étudiée est trop sensible. Lorsque l'on est dans cette situation, on fait l'hypothèse que les valeurs non observées sont uniformément réparties dans chaque classe pour pouvoir réaliser notre étude statistique. On utilise les mêmes notations que dans la partie précédente pour les extrémités $b_{0,}\ldots, b_M$ et les centres $c_{1,}\ldots, c_{M-1}$.

Il est nécessaire de proposer une construction adaptée des outils et indices standards (médiane, moyenne, variance, quartile, courbe des fréquences cumulées) ne nécessitant pas d'observer les valeurs de la variable mais simplement leur classe d'appartenance. Ces généralisations ont pour point de départ l'hypothèse que les données sont uniformément réparties sur chaque classe.

Pour le calcul de la moyenne et de la variance (donc de l'écart-type), les valeurs non observées de la classe (b_j, b_{j+1}) sont remplacées par les centres des classes ce qui donne:

$$x=(c_1.e_1+...+c_M.e_M)/n=c_1.f_1+...+c_M.f_M$$

$$V=(c_1-x)^2.f_1+...+(c_M-x)^2.f_M=(c_1^2.e_1+...+c_M^2.e_M)/n-x^2=c_1^2.f_1+...+c_M^2.f_M-x^2$$

La courbe des fréquence cumulée est quant à elle remplacée par la courbe cumulative (notée ici CC) qui une version empirique de la fonction de répartition $F(x)=P(X\leq x)$. Cette courbe est linéaire par morceaux et continue. Elle est définie de manière habituelle en chaque extrémité de classe b_j comme la proportion F_j de données qui sont inférieures ou égales à b_j . Entre ces points, elle est linéaire (de pente $a_j=(F_j-F_{j-1})/(b_j-b_{j-1})$ sur chaque classe (b_{j-1}, b_j)). Ce qui correspond bien à l'hypothèse que les valeurs non observées sont uniformément réparties dans chaque classe. En effet, pour une variable uniforme sur (a,b), F(x)=(x-a)/(b-a) pour tout x dans (a,b) et pour tout x dans (b_{j-1}, b_j) , $CC(x)=F_{j-1}+a_j.(x-b_{j-1})$.

On définit ensuite la médiane comme la plus petite valeur $x_{0.5}$ telle que $CC(x_{0.5})=0.5$. Il s'agit bien d'une version empirique de la médiane théorique qui est définie pour des lois continues comme la plus petite valeur telle que F(x)=0,5 (de manière générale elle est définie de la manière suivante inf{ $x, F(x)\geq 0,5$ } mais ici F est continue et prend donc toutes les valeurs entre 0 et 1). On peut définir également les premiers et troisièmes quartiles empiriques comme les valeurs telles que CC(x)=0.25 et CC(x)=0.75 respectivement. Plus généralement, on définit pour tout ordre a le quantile empirique de niveau a comme la valeur x_a (non nécessairement unique) vérifiant $CC(x_a)=a$. C'est une version empirique du quantile de la loi de X (l'unicité du quantile de la loi théorique nécessite de vérifier certaines conditions).

On peut déterminer graphiquement la valeur du quantile empirique d'ordre a à partir du graphique représentant la courbe cumulative en prenant pour x_a l'abscisse qui correspond au point d'ordonnée a de la courbe.

Pour calculer précisément la valeur d'un quantile (empirique) dans le cas de données continues il faut suivre la méthode suivante:

- Déterminer la classe (b_{j_a-1}, b_{j_a}) dans laquelle se situe le quantile (à partir du graphique de la courbe cumulative, des valeurs des fréquences cumulées correspondant aux extrémités des classes ou des valeurs des fréquences associées aux différentes classes).
- Comme on l'a déjà expliqué plus haut sur chaque classe (b_{j-1}, b_j) , la courbe cumulative correspond à une interpolation linéaire entre la valeur F_{j-1} prise au point b_{j-1} et la valeur F_j prise au point b_j . Par conséquent elle est donnée par l'équation suivante:

  $$CC(x)=F_{j-1}+(x-b_{j-1}).(F_j-F_{j-1})/(b_j-b_{j-1})$$

  où par convention $F_0=0$. Chercher le quantile empirique correspond donc à résoudre l'équation CC(x)=a sur la classe (b_{j_a-1}, b_{j_a}) , c'est à dire à résoudre par rapport à x l'équation suivante:

  $$F_{j_a-1}+(x-b_{j_a-1}).(F_{j_a}-F_{j_a-1})/(b_{j_a}-b_{j_a-1})=a$$

  Ce qui nous donne:

  $$x_a=b_{j_a-1}+(a-F_{j_a-1}).(b_{j_a}-b_{j_a-1})/(F_{j_a}-F_{j_a-1})$$

EXERCICE (Mise en pratique sans R): Une enquête statistique a été menée sur la répartition des exploitations agricoles françaises en fonction de la SAU (Surface Agricole Utilisée) exprimée en hectares. La SAU est une variable quantitative continue que l'on observe au travers de 6 classes. Sur 1000 exploitations étudiées on observe les données classées suivantes:

SAU	Effectif	Fréquence	Densité
SAU<5	240		
$5 \leq SAU < 10$	109		
$10 \leq SAU < 20$			0.0178
$20 \leq SAU < 35$		0.203	
$35 \leq SAU < 50$		0.102	
$50 \leq SAU < 100$	168		

1. Compléter le tableau précédent.
2. Déterminer la classe modale et tracer l'histogramme.
3. Calculer la moyenne et l'écart-type
4. Tracer la courbe cumulative et donner la valeur de la médiane et des quartiles.

Quelques commandes R:
Définir le vecteur c des centres
Définir le vecteur b des extrémités des classes
Définir le vecteur e des effectifs des classes
Calculer la taille de l'échantillon
Calculer les fréquences
Obtenir les densités
Créer un vecteur v où les centres c_j apparaissent e_j fois
Tracer l'histogramme
Calculer la moyenne
Calculer la variance
Calculer l'ecart-type
Calculer les fréquences cumulées associées aux extrémités des classes
Tracer la courbe cumulative
Calculer la valeur de le quantile d'ordre a (la médiane correspond à a=0,5) si il est dans la classe (b_{j-1}, b_j)

EXERCICE
A la demande de la chambre syndicale des fabricants de produits surgelés, une enquête portant sue les dépenses mensuelles en Euros de produits surgelés chez les ménages dotés d'un frazer (3 étoiles) a été faite. Les résultats de cette enquête sont les suivants:

Classe	[0;20[	[20;40[	[40;60[	[60;80[	[80;100[	[100;120[	[120;140[	[140;160[	[160;200[	[200;300[
Centre	10	30	50	70	90	110	130	150	180	250
Effect	25	25	45	60	75	85	75	50	45	15

Partie 1 :
A.) 1.Calculer les indices de centralité et de dispersion (moyenne et écart-type) à partir. des données classées. Les comparer avec ceux obtenus à l'étape précédente.
2.Tracer l'histogramme et la courbe cumulative. Déterminer les quartiles.

B.) 1.En fait on dispose de la série brute des données dans le fichier surgeles de la librairie lycee. Créer un vecteur data contenant les valeurs de ventes dans la table surgeles.

2.Calculer les paramètres de centralité et de dispersion des données brutes. Comparer ces valeurs avec celles obtenues à l'étape A. Comparer la courbe des fréquences cumulées associée à data et celle obtenue à partir des données classées. Commenter.

Partie 2:

A.) Déterminer la classe modale par le calcul des densités. Le visualiser en traçant l'histogramme des données (utiliser les classes données dans le tableau).

B.) A partir des résultats obtenus dans la Partie 1 et de la forme de l'histogramme, proposer une modélisation de la distribution des dépenses des ménages français en ce qui concerne les produits surgelés. Comparer la modélisation considérée et l'histogramme en traçant la densité de celle-ci sur le même graphique que l'histogramme.

EXERCICE:

Pour juger un risque de défaillance financière des entreprises, on utilise le ratio R défini par le quotient de la marge brute d'autofinancement (cash flow) par la dette totale. On a calculé ce ratio pour 600 entreprises réputées saines. Les résultats sont répertoriés dans le tableau ci-desous.

Ratio R	[0.16;0.421[	[0.421;0.61[	[0.61;0.83[	[0.83;0.988[	[0.988;1.24[
Nombre d'entreprises	35	150	274	109	32

1. Déterminer la moyenne m et l'écart type empirique de cette série statistique.
2. Tracer les représentations graphiques de l'histogramme et de la courbe cumulative associés à ces donnés.
3. Déterminer graphiquement puis par le calcul la médiane et les quartiles.
4. D'après l'allure de l'histogramme et les valeurs trouvées au 1., proposer une modélisation de la distribution de R.

EXERCICE:

La durée d'atterrissage d'un avion est le temps, mesuré en secondes, qui s'écoule entre la prise en charge par la tour de contrôle jusqu'à l'immobilisation totale de l'appareil sur la piste.
Afin de faire face au flux croissant des avions se posant à Toulouse-Blagnac, une restructuration des services de la tour de contrôle visant à diminuer la durée du processus d'atterrissage est réalisée.
A la suite de la restructuration, une enquête, effectuée sur 1000 avions a donné les résultats suivants:

Classe	[60;120[	[120;140[	[140;180[	[180;200[	[200;260[
Effectif	112	176	461	157	94

1. Tracer l'histogramme et déterminer la classe modale.
2. Tracer la courbe cumulative puis déterminer la médiane et les quartiles (graphiquement puis par le calcul)
3. Calculer à la main la moyenne et l'écart-type associé à ces données. Avant la restructuration, la durée d'atterrissage avait pour moyenne 160 secondes et pour médiane 156 secondes. Commenter.
4. Proposer une modélisation de la distribution de la durée d'atterrissage. Comparer la avec l'histogramme.