ACTIVITES DE STATISTIQUE DECLINEES DE LA SIXIEME A LA TERMINALE : UN EXEMPLE AUTOUR DE LA CLIMATOLOGIE

Philippe GARAT, Florent GIROD, Damien JACQUEMOUD, Frédérique LETUE
Groupe Probabilités et Statistique – IREM de Grenoble

Résumé – En première partie d'exposé, nous présentons des exemples d'activités de Statistique proposés cette année à des élèves de collège et lycée, sur le thème de la climatologie (autour de données de la station météorologique de Besançon, Doubs). Nous montrons comment peuvent être déclinées diverses problématiques de statistique, selon le niveau des élèves et selon les éléments contenus dans les nouveaux programmes de mathématiques du lycée. Le but de ces activités de statistique est de faire comprendre la notion de fluctuation d'échantillonnage sur des exemples concrets, et d'expliquer (sans théorisation formelle) le raisonnement qui est à la base des tests d'hypothèses. En deuxième partie de l'atelier, nous exposons quelques considérations plus théoriques sur la manière dont les logiciels statistiques et les calculatrices opèrent pour calculer les fractiles empiriques d'une série statistique. En effet, plusieurs algorithmes existent, trouvant leur justification dans la théorie de l'estimation. Ce thème peut être éventuellement traité dans le cadre d'une APe.

PARTIE I. Exemples d'activités de Statistique proposées à des élèves de collège et lycée

Le groupe de travail de Probabilités et Statistique de l'IREM de Grenoble a fait le choix de travailler cette année sur la mise en place d'activités autour de données réelles, pour les élèves allant de la classe de 6ème à la celle de Terminale. L'activité présentée ici concerne les classes de 3ème et de 1ère ES, sur le thème de la météorologie.

1. Une activité en classe de Troisième

1.1. Les notions de statistiques vues en classe de 3ème

Les notions de statistiques ont été abordées en trois temps dans cette classe, dans le cadre d'une progression de type spiralée.

Dans un premier temps, les notions de médiane, d'étendue ont été vues au mois d'octobre 2011. La médiane et la moyenne en particulier ont été confrontées, mettant en évidence les spécificités de chacune de ces notions.

Dans un second temps, la notion de quartiles a été abordée. Cette activité a permis, en réinvestissant la notion de médiane, de montrer l'intérêt des quartiles pour étudier la répartition d'un échantillon de valeurs. Ce travail a eu lieu au mois de février 2012. Il a consisté dans tout d'abord à répondre à un sondage permettant le calcul de l'Indice de Masse Corporelle (IMC) par tous les élèves de 3ème du collège (figure 1).

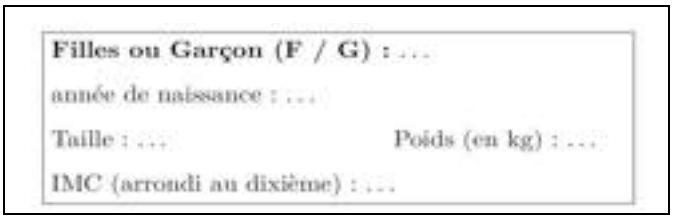

Figure 1 – Questionnaire adressé aux élèves de 3ème du collège

Ensuite, les élèves ont complété une fiche de travail en classe, cette fiche étant associée aux courbes de corpulence présentée par l'INPES (voir annexe 2).

Enfin, le dernier temps consacré aux apprentissages liés aux notions de statistiques a été consacré à un problème ouvert, laissant le choix aux élèves de la méthode la plus appropriée. C'est ce travail, réalisé au mois d'avril 2012 qui est présenté ici.

1.2. Une problématique

Une problématique est donnée en classe : « le climat de ces dernières années est-il significativement différent de celui des années précédentes ? ».

Un travail de réflexion a lieu en classe entière, pour s'approprier le sujet, aller plus loin que les clichés, et mettre en place une démarche.

Des données sont fournies : les températures minimales journalières de la ville de Besançon (données existantes depuis le 1er janvier 1890). Elles sont disponibles, comme de très nombreuses autres données sur le site ECA&D (European Climate Assessment and Dataset) (voir annexe 1).

Même après avoir ramené ces données à celles concernant les normales saisonnières (de 1971 à 2000), elles restent trop nombreuses si elles ne sont pas traitées. Un choix a été fait : faire la moyenne mensuelle des températures minimales journalières. Un tableau (donné sous tableur : voir ci-dessous) reprend toutes ces valeurs : c'est à partir de ce tableau que les élèves vont pouvoir donner des éléments de réponses par rapport à la question posée au départ.

1.3. Travail avec les données informatiques

Après un travail en groupe pour échanger et mettre en place une stratégie, la classe s'est rendue en salle informatique où était fourni le tableur cité précédemment.

Ce travail s'est prolongé à la maison. Un point a été fait quelques séances suivantes où des groupes ont présenté leur travail. Par la suite, le travail s'est terminé en tant que devoir à la maison. Le travail rendu était un texte, incluant des graphiques ou des tableaux.

1.4. Travail des élèves

Beaucoup d'élèves sont restés un peu déconcertés par cette activité. Un manque d'aisance par rapport au tableur, un manque d'habitude de traiter des problèmes ouverts, peu d'initiatives. Il a fallu relancer les groupes durant la séance en salle informatique, mais aussi par la présentation en classe de certains travaux pour échanger sur les techniques et les idées, sur la pertinence des choix faits pour traiter la question. Environ

la moitié des groupes a utilisé la moyenne pour donner des résultats du type suivant (figure 2).

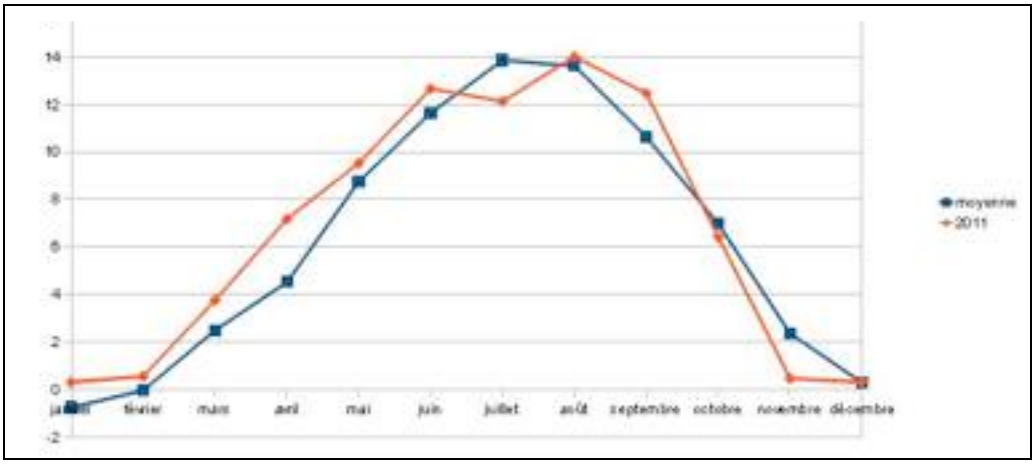

Figure 2 – Environ la moitié des groupes a utilisé la moyenne

Le texte accompagnant ce type de graphique était assez pauvre, dans la mesure où le seul fait d'être au-dessus ou en dessous de la moyenne pouvait être constaté. Un groupe a présenté ces résultats sous la forme d'un tableau :

2) Traitement numérique :

a) On a pris les quartiles et la médiane de la température moyenne minimale entre les années 1970 et 2000. Puis on a classé les années de 2001 à 2012 dans un tableau ci-dessous en fonction des quartiles.

b) **Quartiles**

	Janv.	Fév.	Mars	Avril	Mai	Juin	Juillet	Août	Sept.	Oct.	Nov.	Déc.
2001	4eme	4eme	4eme	3eme	4eme	2eme	4eme	2eme	2eme	2eme	2eme	3eme
2002	1er	4eme	4eme	3eme	4eme	4eme	3eme	3eme	2eme	3eme	4eme	4eme
2003	1er	1er	4eme	4eme	4eme	4eme	4eme	4eme	2eme	1er	4eme	3eme
2004	2eme	2eme	2eme	4eme	2eme	4eme	3eme	4eme	4eme	4eme	3eme	2eme
2005	1er	1er	2eme	4eme	3eme	4eme	4eme	1er	4eme	4eme	2eme	2eme
2006	1er	2eme	2eme	4eme	4eme	4eme	4eme	2eme	4eme	4eme	4eme	3eme
2007	4eme	4eme	2eme	4eme	4eme	4eme	3eme	2eme	1er	2eme	2eme	2eme
2008	4eme	4eme	2eme	4eme	4eme	4eme	3eme	2eme	1er	1er	4eme	3eme
2009	1er	2eme	3eme	4eme	4eme	4eme	4eme	4eme	4eme	2eme	4eme	3eme
2010	2eme	3eme	2eme	4eme	3eme	4eme	4eme	2eme	1er	2eme	4eme	1er
2011	3eme	3eme	4eme	4eme	3eme	4eme	1er	3eme	4eme	2eme	1er	3eme
2012	2eme	2eme										

3) Conclusion :

Le climat de ces dernières années est révélateur du changement climatique car l'hiver, les températures moyennes minimales sont un peu en dessous (les trous dans la couche d'ozone peuvent en être la cause) des normales saisonnières mais en été, elles sont très souvent dans le quatrième quartile (36/55 entre mars et juillet).

Figure 3 – Les résultats présentés par un groupe d'élèves

Ce groupe a mis en évidence qu'entre mars et juillet, la proportion de mois faisant partie du quatrième quartile était très élevée, beaucoup plus que le quart attendu.

On peut critiquer le choix de la période (mars à juillet) qui est un parti pris. Le fait de faire cette comparaison est tout de même une démarche intéressante : c'est un premier pas vers la comparaison d'une fréquence à une proportion attendue. De là se sont engagées des discussions en classe sur le thème de la fluctuation : est-ce « normal » que la fréquence observée ne soit pas exactement égale à un quart ? A partir de quelle valeur pourrait-on considérer que la fréquence est « anormale » ? Tout en restant qualitatif, ces questionnements permettent de préparer aux notions vues en classe de Seconde sur la fluctuation d'échantillonnage.

Un groupe a mis en place les courbes suivantes :

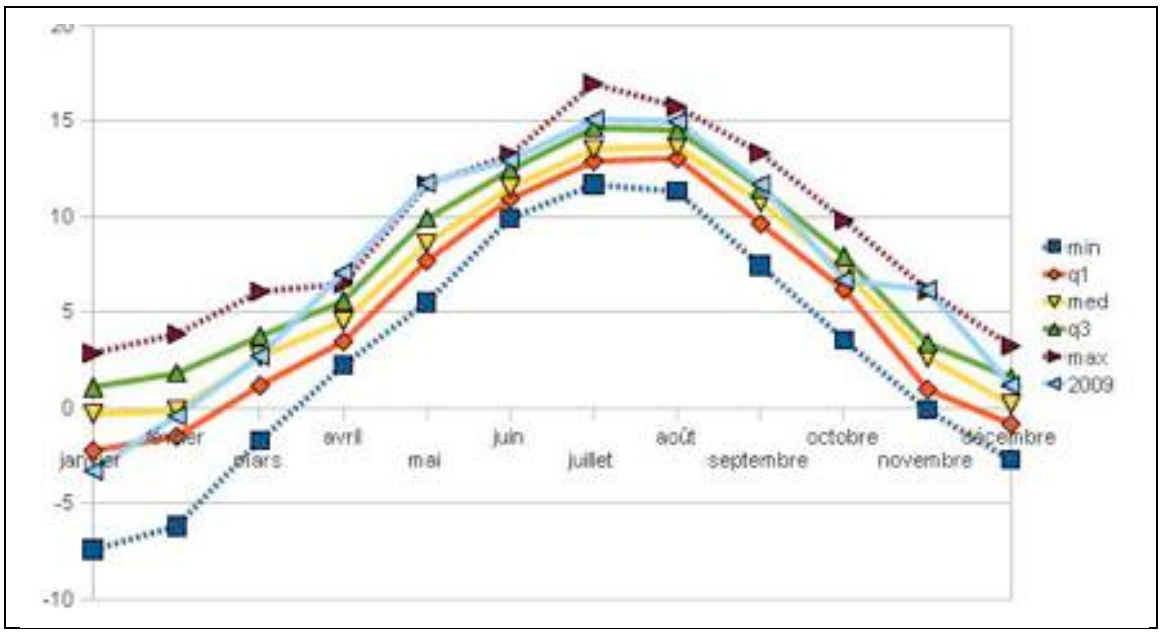

Figure 4 – Production d'un autre groupe d'élèves

Les élèves ont en quelque sorte créer les courbes de l'INPES concernant l'IMC, en mettant en place les courbes « minimum », « premier quartile », « médiane », « troisième quartile » et « maximum » en partant des valeurs des normales saisonnières.

Ils ont pu ensuite situer chaque année dans ces courbes, en qualifiant les mois de « froid », « plutôt froid », « plutôt chaud » et « chaud ».

1.5. Quelques remarques

La plupart des élèves ont le réflexe de calculer des moyennes quand on leur demande de traiter des données numériques. Le cadre scolaire avec l'importance des moyennes des notes, le fait de rencontrer cette notion assez tôt dans la scolarité n'y sont sans doute pas étrangers.

Le fait de mettre les élèves face à de nombreuses données rend pour ainsi dire obligatoire l'utilisation de l'outil informatique. Il faut s'organiser pour que les élèves puissent utiliser ces données, en classe et à la maison. Des clés USB et un site Internet ont été utilisés. Par ailleurs, quelques problèmes techniques (compatibilité OpenOffice / Excel) ont gêné quelques groupes.

Ces difficultés techniques nous ont amenés dans notre établissement à la réflexion suivante : pourquoi ne pas demander dans la liste de matériel de rentrée, d'installer

OpenOffice (et d'autres logiciels, notamment en géométrie dynamique) sur les ordinateurs à la maison, pour que les élèves puissent poursuivre ce type de travail sans difficultés techniques ? Il va de soi que l'établissement doit être en mesure de mettre à disposition des postes pour les élèves qui ne seraient pas équipés à la maison.

2. Une activité en classe de 1ère ES

2.1. Progression sur l'année

Les notions de statistiques descriptives (moyenne, écart-type, médiane et quartiles, diagramme en boîte) ont été vues en octobre 2011. La loi binomiale a été abordée en janvier 2012. La notion de fluctuation d'échantillonnage a été vue ensuite. Les calculs d'intervalle de fluctuation au seuil de 95% ont été vus en mars 2012.

2.2. Activité

La même problématique qu'en classe de 3ème a été présentée en classe de 1ère ES, sur le thème de la météo. Le même type de discussion a été mené. Les élèves ont été plus guidés, pour construire des diagrammes en boîtes relatifs aux moyennes mensuelles des températures minimales journalières. Ces courbes ont rappelé à certains les courbes présentes dans le carnet de santé.

Une deuxième partie du travail a consisté à mettre en place le modèle suivant :

- On compte le nombre de jours dans un mois, où la température minimale est inférieure à 0°C. On considère le fait que cette température soit au dessus ou en dessous de zéro respectivement comme « échec » et « succès » d'une épreuve de Bernoulli.
- Les normales saisonnières (données de 1971 à 2000) donnent le nombre de jours moyens, par mois, où la température minimale est inférieure à 0°C. Par exemple, il est de 16,6 pour le mois de Janvier (toujours pour la ville de Besançon).
- Il s'agira de construire un intervalle de fluctuation, par mois, concernant le nombre de jours où la température minimale est inférieure à 0°C.

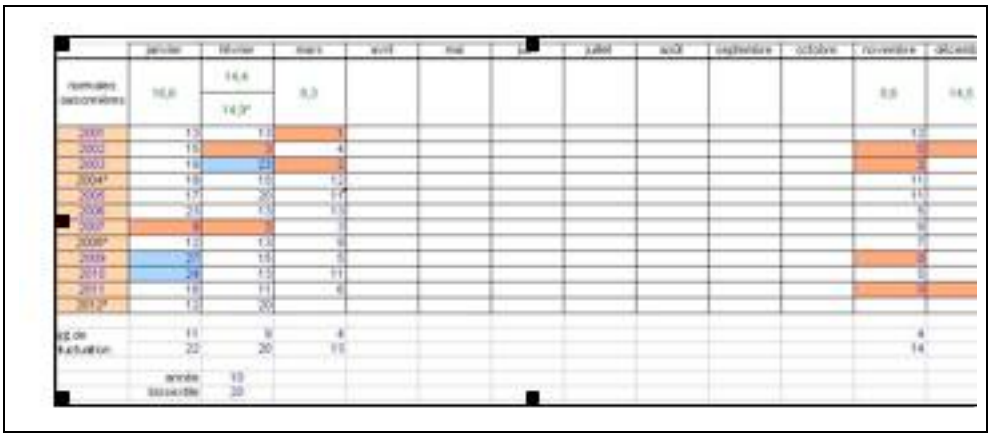

Tableau 1 – Ce tableau résume le travail décrit ci-dessus

Par exemple, pour le mois de janvier, la loi de Bernoulli de paramètres p =16,6/31 et n =31 donne un intervalle de fluctuation au seuil de 95% égal à [11 ; 22]. Cela veut dire que pour ce modèle, un mois de janvier conforme aux normales saisonnières comptera

entre 11 et 22 jours où la température minimale aura été inférieure à 0°C. Si la température minimale a été moins de 11 fois inférieure à 0°C, on considérera ce mois comme « chaud » (sur fond gris foncé dans le tableau 1), et si la température minimale a été plus de 22 fois inférieure à 0°C, on considérera ce mois comme « froid » (sur fond gris clair, dans le tableau 1). Le travail a été fait pour les mois d'hiver : janvier, février (deux calculs selon que l'année est bissextile ou pas), mars, novembre et décembre. Cela représente 6 déterminations d'intervalles de fluctuation.

2.3. Travail des élèves

Ce travail, démarré en salle informatique, s'est poursuivi à la maison sous la forme d'un devoir à la maison. Certains élèves ont joué le jeu : des allers-retours de leur travail par courriel a permis de l'enrichir pour arriver à un résultat très satisfaisant. D'autres ont vu un autre intérêt d'un travail informatique en enregistrant le travail d'un camarade sous leur nom (c'est encore plus rapide que de recopier un devoir à la maison à la main !).

Mis à part ce problème de « copie », cette activité a permis aux élèves de déterminer de nombreux intervalles de fluctuation hors d'un cadre classique. Ils ont été évalués plus classiquement par la suite, cette évaluation ne laissant pas de doute sur ceux qui avaient les recherches par eux-mêmes et ceux qui avaient simplement repris le travail.

Ce travail part d'un certain nombre d'hypothèses qui ont conduit à mettre en place un modèle : une loi binomiale pour étudier la conformité d'un mois par rapport aux normales saisonnières. Ce modèle a ses limites. En effet, l'indépendance des expériences de Bernoulli (ici, le fait que la température d'un jour est indépendante de celle de la veille) est discutable. Il faut bien mettre en évidence que cela fait partie de la démarche : mettre en place un modèle en ayant conscience de ses limites.

3. Quelques réflexions par rapport à l'arrivée des probabilités statistiques dans l'enseignement secondaire

3.1. Transformations de l'enseignement et contraintes pour l'enseignant

Il nous semble qu'il est important de problématiser les notions présentées : présenter les quartiles et les déterminer sans avoir vu leur intérêt sera très démotivant. S'il est important de réaliser des exercices d'entraînement sur plusieurs séries, il est indispensable de montrer la pertinence de ces notions. La recherche d'activités de ce type n'est pas aisée. Les manuels ne sont pas toujours très riches à ce niveau.

Par ailleurs, les probabilités-statistiques, par l'importance qu'elles ont prise, peuvent permettre de renforcer la formation du citoyen en ce sens que de nombreux documents que l'ont peut voir dans les journaux, à la télévision ou sur Internet sont critiquables voire faux. Prenons l'exemple de ce graphique présenté sur France 2 lors d'une émission consacrée aux élections présidentielles :

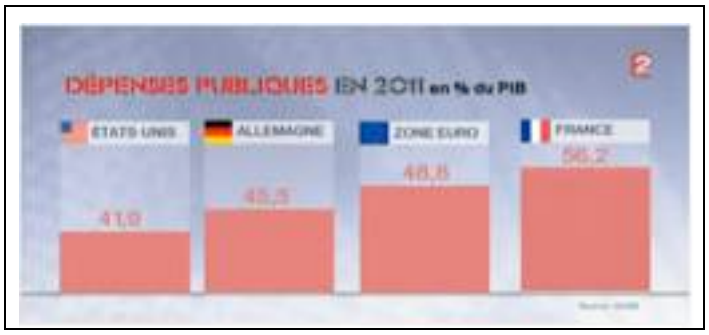

Figure 5 – Graphique extrait de l'émission « Des Paroles et des Actes » (2012)

Le type d'activités présentées nous paraît donc primordial en terme de problématisation des notions vues dans le cadre du cours, ces activités permettent également d'aiguiser l'esprit argumentatif et critique des élèves.

Cependant, une double contrainte de temps s'impose :

- pour l'enseignant : ces activités demandent beaucoup de temps en terme de préparation ;
- pour l'enseignement : ces activités, si on veut les mener jusqu'au bout, demandent beaucoup de temps.

Il faut donc faire des choix et cerner quelques thèmes sur l'année scolaire.

3.2. Répercussions pour les élèves et conséquences en terme d'évaluation

Ces activités ont exigé une continuité du travail : en classe, en salle informatique, à la maison, voire pour certains avec l'utilisation du courriel, des échanges avec l'enseignant. Beaucoup d'élèves ne sont pas habitués à ce type de démarche et ne conçoivent le travail que dans le cadre scolaire.

Les thèmes abordés, changeant de cadre, permettent un réinvestissement de notions vues en classe, permettent un transfert. Cette démarche n'est pas non plus évidente et demande à être travaillée régulièrement. Elle demande en particulier une prise d'initiative.

Enfin, évaluer ce type de travail n'est pas évident si l'on veut en donner une évaluation chiffrée. Comment valoriser la démarche, l'investissement, même si le résultat n'est pas probant ? Comment s'assurer que le travail a bien été réalisé par les élèves eux-mêmes ? Est-il indispensable de noter ce type de travail ? Si on ne le fait pas, ne court-on pas le risque d'un laisser-aller de la part de certains ? Dans les expériences réalisées, les travaux ont été évalués sous la forme de devoirs à la maison, c'est-à-dire notés avec un coefficient très faible.

PARTIE II. Calculs de fractiles empiriques à l'aide des logiciels statistiques et calculatrices

1. Introduction : importances des fractiles dans la décision statistique

Les fractiles d'une distribution permettent de construire aisément des intervalles de fluctuation pour une variable étudiée, et de façon corollaire des régions critiques, des seuils critiques. On utilise les fractiles de la loi binomiale dans le test exact d'une

proportion. Le rôle de la médiane est essentiel en tant que paramètre de localisation. De nombreux ouvrages en Génie Civil sont dimensionnés par rapport aux valeurs centennales de variables climatiques (hauteur d'eau, neige, force du vent …). Un autre exemple en santé publique : analyse de sang…

Les fractiles dans l'enseignement secondaire :

- les fractiles usuels, médiane et quartiles, sont enseignés dès la classe de Seconde ;
- le diagramme en boîte est enseigné en Première ; il est parfois préconisé d'utiliser les déciles D1 et D9 pour tronquer les « moustaches » du diagramme ;

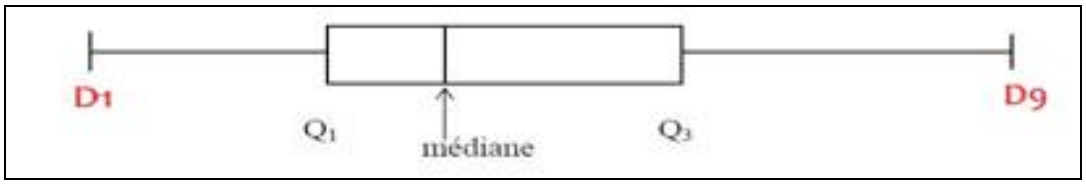

Figure 6 – Une boîte à moustaches dont les extrémités sont les déciles

- En Terminale S, le seuil $u_{0.05}$ est en réalité le fractile de probabilité 97.5 % ;
- des applications statistiques autour des fractiles sont possibles en économie, SVT, environnement, etc ...
- Le calcul des fractiles peut donner lieu à des exercices en algorithmique.

2. Fractiles d'une loi de probabilité : cas d'une loi discrète

2.1. Définition : fractiles d'une loi discrète

Soient $a_1 < a_2 < \ldots < a_k < \ldots$ l'ensemble des modalités d'une variable discrète X et soient $p_1, p_2, \ldots, p_k, \ldots$ les probabilités associées.

On appelle fractile d'ordre p de X (et on note x_p) la plus petite modalité a_j telle que la probabilité d'être inférieure ou égale à a_j soit au moins égal à p, c'est-à-dire :

$$\text{Prob}[X < a_j] < p \quad \text{et} \quad \text{Prob}[X \leq a_j] \geq p \qquad (1)$$

Soit encore :

$$p_1 + p_2 + \ldots + p_{j-1} < p \quad \text{et} \quad p_1 + p_2 + \ldots + p_{j-1} + p_j \geq p.$$

Considérant $F(x)$ la fonction de répartition de X, on peut aussi définir le fractile d'ordre p de X comme étant le plus petit élément de l'ensemble antécédent par F de $[\, p\, ;\, 1\,]$:
$x_p = \min \{\, x \in \text{IR} / F(x) \geq p \,\}$

2.2. Exemple

Considérons l'expérience aléatoire consistant à tirer *avec remise* 30 boules dans une urne contenant 1/3 de boules blanches. Cette expérience est régie par la loi binomiale B(30 ; 1/3).

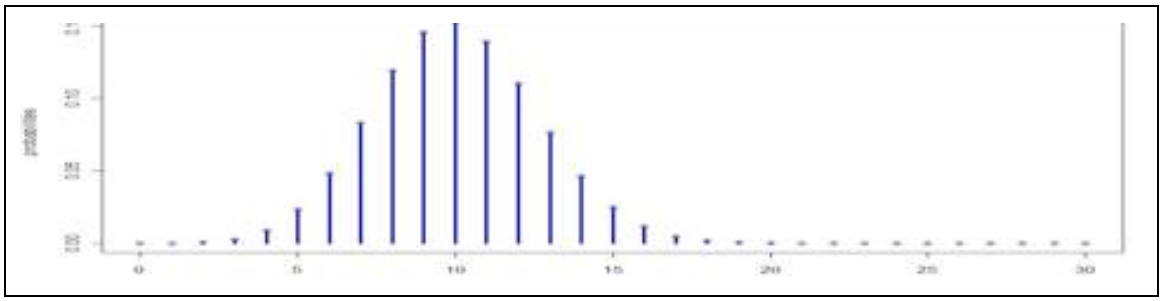

Figure 7 – Loi binomiale B(30 ; 1/3)

Le tableau des probabilités cumulées donne :

a	0	1	2	3	4	5	6	7	8	9	10
Prob[$X \leq a$]	*0.000*	*0.000*	*0.001*	*0.003*	*0.012*	*0.035*	*0.084*	*0.167*	*0.286*	*0.432*	*0.585*
a	11	12	13	14	15	16	17	18	19	20	21
Prob[$X \leq a$]	*0.724*	*0.834*	*0.910*	*0.957*	*0.981*	*0.993*	*0.998*	*0.999*	*1*	*1*	*1*
a	22	23	24	25	26	27	28	29	30		
Prob[$X \leq a$]	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*1*		

Tableau 2 – Le tableau des probabilités cumulées

Nous en déduisons par exemple les *fractiles* de la loi de X suivants :

Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
5	6	7	8	10	12	13	14	15

Tableau 3 – Les fractiles de la loi de X

Il est à noter que, par construction, la probabilité que X soit inférieur (strictement) à son fractile d'ordre p est au plus égale à p, et que la probabilité que X dépasse (strictement) son fractile d'ordre p est au plus égale à $(1 - p)$. Cette propriété sera très utile en statistique inférentielle pour la recherche de seuil critique. Nous l'écrivons :

$$\text{Prob}[X < x_p\,] < p \quad \text{et} \quad \text{Prob}[X > x_p\,] < 1 - p.$$

Cela implique en particulier :

$$\text{Prob}[X < x_{p1} \text{ ou } X > x_{1\text{-}p2}\,] < p_1 + p_2.$$

On illustre *sur l'exemple* :

$$\text{Prob}[X < Q_{0.05} \text{ ou } X > Q_{0.95}\,] < 10\ \%.$$

Le calcul exact donne :

$$\text{Prob}[X < 6 \text{ ou } X > 15\,] = 0.035 + 0.019 = 5.4\ \%.$$

2.3. Fractiles empiriques d'une série statistique : cas d'une variable statistique discrète

Au-delà de la définition des fractiles théoriques d'une variable statistique X, il se pose le problème d'estimer ces fractiles à partir d'un échantillon $(x_1, x_2, \ldots, x_n)$ lorsque la loi de X est inconnue. Nous appellerons fractiles empiriques de tels fractiles.

Le cas discret ne pose *a priori* pas de problème particulier : il suffit de remplacer dans la définition (1) les probabilités théoriques p_1, p_2, …, p_k … par les fréquences empiriques f_1, f_2, …, f_k, … obtenues lors de l'analyse fréquentielle (tri-à-plat) de l'échantillon (méthode 1).

En pratique, deux autres méthodes sont parfois proposées selon les logiciels utilisés :

1) trier l'échantillon $x_1, x_2, \ldots, x_n$ par ordre croissant des valeurs ;
2) associer à chaque valeur x_j la fréquence cumulée $F(x_j) = j / n$;
rechercher les deux valeurs d'indice $j - 1$ et j dont les fréquences cumulées encadrent la probabilité p, c'est-à-dire : $(j - 1) / n < p \leq j / n$;
3) calculer le fractile d'ordre p selon l'une des 3 méthodes :
   - méthode 1 : $Q(p) = x_j$;
   - méthode 2 : $Q(p) = x_j$ si $p < j / n$ et $Q(p) = (x_j + x_{j+1}) / 2$ si $p = j / n$;
   - méthode 3 : $Q(p) = x_{j-1}$ ou x_j selon que p est plus près de $(j - 1)$ ou de j ; lorsque $p = j / n$, on choisit x_{j-1} ou x_j correspondant à un indice pair.

Graphiquement :

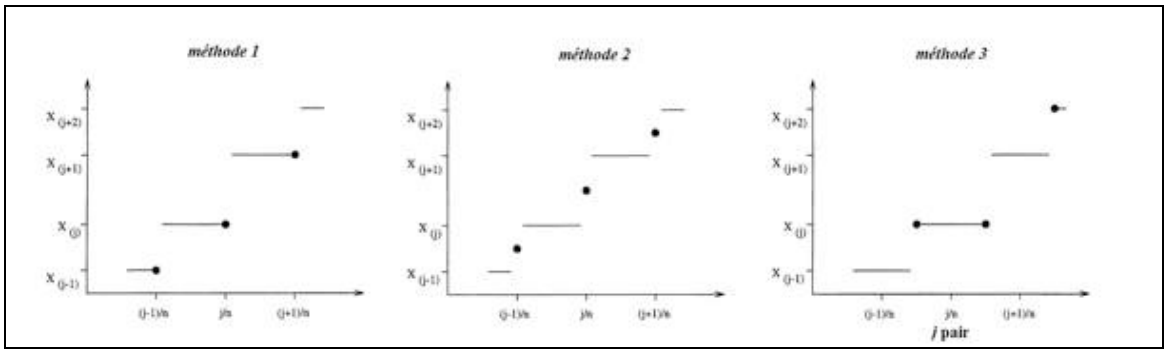

Figure 8 – Fonctions Quantile

Remarques :

- la définition « lycée » de la médiane correspond à la méthode 2 ;
- la définition « lycée » des quartiles Q1 et Q3 correspond à la méthode 1.

3. Fractiles d'une loi de probabilité : cas d'une loi absolument continue

3.1. Définition des fractiles d'une loi absolument continue (loi à densité)

Soit X une variable aléatoire absolument continue admettant une densité de probabilité $f(x)$ et une fonction de répartition $F(x)$ sur IR. On appelle fractile d'ordre p la valeur x_p de X telle que la probabilité d'être inférieure ou égal à x_p vaut exactement p ; c'est-à-dire :

$$\text{Prob}[X \leq x_p] = p \qquad (2)$$

Soit :

$$F(x_p) = p.$$

L'existence et l'unicité de x_p est garantie par le caractère continu et croissant de la fonction de répartition $F(x)$. Le fractile x_p est l'antécédent de p par la fonction $F(x)$.

$$x_p = F^{-1}(p) \qquad (3)$$

3.2. Exemple de loi continue

Considérons la loi exponentielle « unilatérale » sur IR^+ d'espérance unité, c'est-à-dire la loi définie par la densité :

$$f(x) = \exp(-x), \quad x \geq 0.$$

La fonction de répartition associée à cette loi est :

$$F(x) = 1 - \exp(-x), \quad x \geq 0.$$

Le fractile x_p d'ordre p s'obtient en résolvant en x :

$$F(x) = p.$$

On obtient :

$$x_p = -\ln(1-p).$$

On en déduit par exemple les fractiles de la loi de X suivants :

Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
0,0253	0,0513	0,1054	0,2877	0,6931	1,3863	2,3026	2,9957	3,6889

Tableau 4 – Les fractiles de la loi de X

3.3. Fractiles empiriques d'une série statistique : cas d'une variable statistique continu

Le cas continu est plus difficile que le cas discret, et de nombreux algorithmes existent, donnant des résultats ayant des propriétés mathématiques sensiblement différentes.

Par exemple, le tableur Excel adopte la méthode suivante :

- trier l'échantillon $x_1, x_2, \ldots, x_n$ par ordre croissant des valeurs ;
- associer à chaque valeur x_i la fréquence cumulée corrigée :

$$F_n^*(x_j) = (j-1)/(n-1) ;$$

- rechercher les deux valeurs x_{j-1} et x_j dont les fréquences cumulées encadrent la probabilité p, c'est-à-dire

$$F_n^*(x_{j-1}) < p \leq F_n^*(x_j) ;$$

- calculer le fractile d'ordre p à l'aide de la formule d'interpolation linéaire :

$$\tilde{Q}_p = [\,(F_n^*(x_j) - p) * x_{j-1} + (p - F_n^*(x_{j-1})) * x_j\,] / (F_n^*(x_j) - F_n^*(x_{j-1})).$$

Cette méthode consiste donc à approcher la fonction de répartition théorique de X par une fonction continue et linéaire par morceaux, notée $F_n^*(x)$ et à rechercher l'antécédent de p par $F_n^*(x)$.

Compte tenu de la définition de $F_n^*(x_j) = (j-1)/(n-1)$, on voit que cette méthode fournit une valeur de médiane empirique égale à : $x_{[(n+1)/2]}$ si n est impair, et $(x_{[n/2]} + x_{[(n+1)/2]})/2$ si n est pair. Cette méthode est une alternative à la démarche « naturelle » consistant à interpoler linéairement les fréquences cumulées non corrigées $F_n(x_j) = j/n$ ou à interpoler linéairement les milieux des paliers de la fonction de répartition empirique, c'est-à-dire :

$$F_n(x_j) = (j - 0.5)/n$$

3.4. Cinq méthodes disponibles

Diverses méthodes de calcul se différencient sur la manière de corriger les fréquences cumulées $F_n(x_j)$:

- méthode 4 : $F_n(x_j) = j / n$;
- méthode 5 : $F_n(x_j) = (j - 0.5) / n$;
- méthode 6 : $F_n(x_j) = j /(n + 1)$;
- méthode 7 (utilisée par *EXCEL* (et *R* par défaut) : $F_n(x_j) = (j - 1) / (n - 1)$;
- méthode 8 : $F_n(x_j) = (j - 1/3) / (n + 1/3)$.

On justifie les méthodes 4 à 8 en remarquant que :

$F(X(j))$ suit la loi Bêta $(j , n - j + 1)$.

Par conséquent : $E [F(X(j))] = j / (n + 1)$ et Mode $[F(X(j))] = (j - 1) / (n - 1)$.

De manière approximative : Médiane $[F(X(j))] = (j - 1/3) / (n + 1/3)$.

4. Expérimentations et Simulations

On souhaite savoir laquelle des cinq méthodes donne de meilleurs résultats pour estimer les fractiles d'une loi exponentielle à partir d'un échantillon de taille $n = 20$ ou d'un échantillon de taille $n = 50$. Des simulations Monte-Carlo (nombre de répétitions = 10.000) avec le logiciel R donnent les résultats suivants :

n = 20	Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
Méthode 1	*141*	*73*	*54*	*34*	*24*	*21*	*21*	*23*	*26*
Méthode 4	*142*	*72*	*53*	*34*	*25*	*21*	*21*	*24*	*25*
Méthode 5	*142*	*82*	*56*	*35*	*25*	*21*	*21*	*23*	*26*
Méthode 6	*142*	*71*	*52*	*35*	*25*	*22*	*25*	*31*	*26*
Méthode 7	*209*	*116*	*68*	*37*	*25*	*21*	*21*	*23*	*24*
Méthode 8	*142*	*74*	*53*	*35*	*25*	*21*	*22*	*25*	*26*

n= 50	Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
Méthode 1	*89*	*52*	*34*	*22*	*16*	*13*	*15*	*15*	*17*
Méthode 4	*60*	*45*	*34*	*22*	*16*	*13*	*14*	*15*	*17*
Méthode 5	*74*	*52*	*35*	*22*	*16*	*14*	*14*	*15*	*17*
Méthode 6	*60*	*45*	*34*	*22*	*16*	*14*	*15*	*17*	*24*
Méthode 7	*96*	*58*	*39*	*23*	*16*	*13*	*14*	*15*	*17*
Méthode 8	*67*	*49*	*35*	*22*	*16*	*14*	*14*	*16*	*19*

Tableau 5 – Erreur relative absolue moyenne par rapport à la bonne valeur des fractiles (en %)

n = 20	Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
Méthode 1	0.02	0.00	0.00	-0.01	-0.02	-0.07	-0.22	-0.42	-0.11
Méthode 4	0.02	0.00	0.00	-0.01	-0.02	-0.07	-0.22	-0.42	-0.61
Méthode 5	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.08	-0.11
Méthode 6	0.02	0.00	0.00	0.01	0.03	0.08	0.23	0.53	-0.11
Méthode 7	0.05	0.05	0.05	0.04	0.03	-0.02	-0.17	-0.37	-0.59
Méthode 8	0.02	0.02	0.02	0.02	0.03	0.04	0.10	0.23	-0.11

n = 50	Q0.025	Q0.05	Q0.10	Q0.25	Q0.50	Q0.75	Q0.90	Q0.95	Q0.975
Méthode 1	0.01	0.01	0.00	0.01	-0.01	0.01	-0.09	0.00	-0.20
Méthode 4	0.00	0.00	0.00	0.00	-0.01	-0.03	-0.09	-0.17	-0.32
Méthode 5	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.05
Méthode 6	0.00	0.00	0.00	0.00	0.01	0.03	0.09	0.22	0.53
Méthode 7	0.02	0.02	0.02	0.02	0.01	-0.01	-0.07	-0.15	-0.31
Méthode 8	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.07	0.21

Tableau 6 – Biais d'estimation

Ces résultats de simulations montrent au final que les diverses méthodes de calcul donnent des résultats sensiblement équivalents du point de vue de la précision d'estimation, à l'exception de certains fractiles extrêmes de la loi exponentielle compte tenu de la taille d'échantillon relativement faible.

ANNEXE 1

MOYENNES MENSUELLES DES TEMPERATURES MINIMALES DE 1971 A 2011 POUR LA VILLE DE BESANÇON (SOURCE : ECA&D)

moyenne mensuelle des températures minimales de 1971 à 2011 pour la ville de Besançon (toutes les températures sont données en °C)

	janvier	février	mars	avril	mai	juin	juillet	août	septembre	octobre	novembre	décembre
1971	-3	-0,77	-1,69	0,04	9,6	10,75	13,81	13,72	9,21	6,31	1,02	-0,54
1972	-1,06	1,86	2,63	4,29	7,89	9,91	12,76	12,03	7,43	4,61	2,19	-2,14
1973	-3,07	-1,49	0,12	2,25	9,14	12,09	13,46	14,92	11,03	5,2	1,66	-0,84
1974	2,39	2,18	4,21	3,57	7,35	10,34	11,91	13,79	10,57	3,56	3,53	3,11
1975	2,31	-0,06	1,9	4,56	7,69	10,72	13,26	14,5	12,41	4,86	2,8	-2,26
1976	-0,04	-0,02	0,55	3,74	8,7	12,42	14,75	11,77	9,59	7,98	3,23	-1,78
1977	1,1	3,86	4,68	3,21	7,86	11,23	12,89	12,2	8,49	7,94	3,14	0,42
1978	-0,6	-0,24	3,84	3,32	8,46	10,43	12,73	11,31	8,98	5,54	0,1	1,17
1979	-4,21	0,7	3,75	3,29	7,57	12,25	12,6	11,73	10,16	6,15	2,22	1,57
1980	-1,91	1,9	2,81	3,57	7,46	10,93	11,68	13,34	10,73	5,41	0,46	-2,74
1981	-2,71	-2,66	5,88	5,4	8,32	11,11	13	12,86	11,49	7,06	1,71	0,2
1982	0,44	-0,17	1,61	3,19	8,37	13,23	15	13,36	12,1	7,45	3,68	1,45
1983	0,89	-1,95	2,29	5,04	7,18	12,44	16,95	14,29	11,18	6,8	1,42	-0,39
1984	0,41	-0,77	-0,12	3,48	6,67	10,69	12,45	13,22	10,46	7,58	4,57	-0,04
1985	-7,42	-1,91	1,17	5,19	8,54	10,84	14,45	12,49	10,93	6,53	0,16	1,58
1986	0,1	-6,17	1,23	4,04	10,39	12,41	13,33	13,04	10,58	7,86	3,54	0,18
1987	-5,75	-0,19	0,05	5,6	6,6	11,46	14,64	13,78	13,05	8,9	3,29	0,24
1988	2,87	0,25	2,46	5,16	10,59	11,57	13,14	14,02	10,9	8,57	0,93	1,96
1989	-1	0,45	4,16	4,83	9,93	11,38	14,31	13,57	10,61	7,24	0,38	-0,23
1990	-1,24	3,76	3,34	4,53	10,14	11,64	13,96	14,11	9,52	8,87	3,21	-1,33
1991	-0,28	-1,44	5,03	3,34	6,51	11,31	14,92	14,66	12,77	6,19	2,63	-1,86
1992	-2,89	-0,47	3,08	4,94	10,19	12,56	14,64	16,33	10,76	6	4,26	1,21
1993	1,56	-2,09	1,36	6,49	9,87	12,72	13,35	13,12	10,4	6,54	-0,11	2,73
1994	1,25	1,18	6,08	4,2	9,88	12,6	16,01	16,41	11,57	7,47	6,16	3,08
1995	-0,56	3,81	1,15	5,74	8,52	11,1	16,11	14,65	9,24	9,8	2,63	-0,48
1996	-0,13	-1,68	0,68	4,83	8,88	12,49	13,37	13,5	7,79	7,06	2,91	-1,06
1997	-2,23	1,81	4,29	3,45	9,68	12,35	13,58	15,75	10,96	6,5	3,39	1,75
1998	1,11	0,09	2,74	5,62	9,72	12,49	14,18	13,17	11,22	7,66	0,01	-0,29
1999	1,26	-0,62	3,52	5,76	11,66	11,28	15,32	14,55	13,32	7,42	0,98	0,79
2000	-0,36	1,86	2,91	5,97	10,43	12,85	12,63	14,52	11,66	8,23	4,69	3,24
2001	1,62	1,2	5,62	4,63	11,47	11,1	14,74	15,14	9,43	10,32	1,14	-1,32
2002	-0,42	4,06	3,9	4,78	8,34	14,24	13,5	14,42	10,49	7,35	5,86	3,59
2003	-0,89	-2,57	3,54	5,56	10,44	17,12	15,34	17,54	10,31	5,02	4,06	0,73
2004	0,22	-0,42	1,55	5,79	7,69	12,47	13,79	14,94	11,67	9,5	2,84	-0,35
2005	-1,02	-1,86	2,12	6,32	9,34	13,6	14,84	12,1	12,14	9,18	2,08	-0,84
2006	-2,16	-0,51	2,36	5,46	10,3	13,17	16,45	13,04	14,07	10,72	4,67	1,1
2007	3,7	3,24	2,5	8,04	11,47	14,41	14,13	13,57	9,18	8,68	2,27	-0,74
2008	1,53	1,21	2,55	5,53	11,37	13,6	14,25	13,39	9,32	5,92	3,97	0,22
2009	-3,27	-0,42	2,71	7,07	11,76	12,54	15,07	14,97	11,68	8,68	6,17	1,19
2010	-2,61	0,44	1,82	5,95	9,11	13,65	15,9	13,43	9,44	6,21	4,17	-1,58
2011	0,31	0,55	3,76	7,18	9,53	12,87	12,14	14,02	12,48	6,43	0,47	0,3
2012	0,37	-0,49										

ANNEXE 2

COURBES DE CORPULENCE (POUR LES GARÇONS) ISSUE DE L'INPES