Université Paris Dauphine — Année 2012-2013
Département MIDO
L3 - Statistique Mathématique

Projet de statistique mathématique. Première partie

Le projet compte pour 30% de la note finale de Statistique Mathématique.

Déroulement du projet : Les étudiants travaillent par groupes de 2 ou 3 étudiants du même groupe de TD. Un rapport écrit en version papier sur cette première partie doit être rendu lors de la semaine du 1 mars au chargé de TD. Tout retard équivaut à un 0 pour cette première partie qui compte pour 6/20 de la note finale du projet.

Les étudiants doivent utiliser le logiciel R [1]. Les principales fonctions utilisées doivent apparaître en annexe dans le rapport.

Récupération des données : Chaque binôme ou trinôme constitué d'étudiants du même groupe de TD envoie au chargé de TD les noms composants le binôme ou trinôme. En retour de mail, il reçoit :
– un numéro de binôme ou trinôme,
– des données dans un fichier .txt.
Les données envoyées sont des données réelles, représentant les volumes au bid ou à l'ask dans un carnet d'ordre d'actions observé durant une journée.

1 Objectif de la première partie

L'objectif de la première partie du projet est double :

1. Réaliser un traitement des données réelles afin de rendre l'inférence statistique possible,
2. Ce traitement ayant été effectué, identifier graphiquement la loi suivie par les données traitées parmi une famille de lois classiques.

Commençons par une présentation du contexte.

2 Présentation des données

Le carnet d'ordre d'un titre boursier est un recueil des ordres d'achat et de vente des actifs de ce titre. Le bid et l'ask (en français : l'offre et la demande) sont les termes employés sur les marchés financiers pour désigner le prix auquel les intervenants vendent (le bid) ou achètent des actifs (l'ask). Il y a transaction lorsqu'un vendeur propose, pour un titre,

1. voir ci-besoin introduction à R : http ://www.ceremade.dauphine.fr/%7Exian/Noise/R.pdf ou le site officiel http ://www.r-project.org/

un prix équivalent à celui d'un acheteur. Le carnet d'ordre liste le volume des ordres en attente au bid et à l'ask. Par exemple, considérons le carnet d'ordre

Volumes en attente à l'ask	Ask	Bid	Volumes en attente au bid
200	42	43	100
500	41	43, 2	120
110	40, 9	43, 5	120
1000	40	43, 8	25
500	39		

Supposons désormais que des vendeurs (généreux) proposent 300 actifs au prix de 41 et 500 actifs à 42,5. Les acheteurs les plus généreux obtiennent alors la transaction au prix qu'ils ont fixés à 42, un prix supérieur à celui des 300 actifs. Ils acquièrent 200 actifs. On compare ensuite le prix à l'ask de la tranche suivante avec le prix au bid. Comme ils sont égaux, la transaction à lieu pour les 300-200=100 actifs restant. Les 500 actifs à 42,5 apparaissent dans le carnet d'ordre car leur prix est supérieur à l'ask. Le carnet d'ordre devient

Volumes en attente à l'ask	Ask	Bid	Volumes en attente au bid
400	41	42, 5	500
110	40, 9	43	100
1000	40	43, 2	120
500	39	43, 5	120
		43, 8	25

3 Traitement des données

En pratique, seuls les prix et les volumes au bid et à l'ask sont observés après chaque transaction, i.e. seule la première ligne des tableaux ci dessus est connue. Les volumes observés à l'ask vous sont fournit dans un fichier TXT. C'est une réalisation d'une v.a. à valeurs entières qu'on notera $(x_t)_{1\leq t\leq n}$.

Question 1 : Importer les données réelles sous R en utilisant les fonctions "read.table" ou "read.delim" et tracer la courbe d'évolution des volumes.

Cette courbe peut présenter des anomalies. Par exemple, elle peut avoir une valeur fixe sur une certaine période. Ce comportement n'est pas cohérent avec la nature des donnée : les volumes fluctuent toujours du fait de nombreux ordres donnés par de nombreux acteurs.

Question 2 : Traiter les données en retirant des volumes $(x_t)_{1\leq t\leq n}$ les données tout comportement suspect. Justifier votre démarche.

Par abus de notation, on notera toujours $(x_t)_{1\leq t\leq n}$ la série traitée même si le nombre de réalisations est plus petit que celui d'origine du fait du traitement. Nous modélisons ces données comme étant les réalisations d'un échantillon aléatoire $(X_1, \ldots, X_n)$.

Question 3 : Discuter l'hypothèse d'indépendance entre observations grâce à la fonction acf sous R et à partir du problème de départ.

Bien que cette hypothèse d'indépendance peut se révéler peu réaliste, nous ne la remettrons pas en cause ici, le traitement de la dépendance des données dépassant le cadre du cours de L3 de statistique mathématique.

4 Inférence statistique

On suppose avoir obtenu, après traitement des données (c.f. Section précédente) n réalisations x_t d'observations X_t, des variables aléatoires indépendants et identiquement distribués selon une **loi de probabilité inconnue** P. On note f la densité de probabilité des X_i et F sa fonction de répartition.

Remarquons que les volumes traités sont issus de variables aléatoires (v.a.) X_t à valeurs entières. Toutefois, la majorité des lois classique sont des lois continues. En première approximation, on suppose donc que la loi de probabilité P a suffisamment de valeurs distinctes et que P appartient à un ensemble de lois de probabilités continues classiques dont la liste est rappelée ci-dessous.

Liste des densités classiques

1. La loi continue **uniforme** sur l'intervalle $[a, b]$, avec $a < b$.
2. La loi **exponentielle** "translatée" de paramétres $x_0 \in \mathbb{R}$ et $a > 0$ dont une version de la densité $f_{\mathcal{E}}$ est
$$f_{\mathcal{E}}(x; x_0, b) = a \exp(-a(x - x_0)) \mathbb{1}_{[x_0, \infty[}(x).$$
3. La loi **log-normale** généralisée de paramétres $x_0 \in \mathbb{R}$, $\mu \in \mathbb{R}$ et $\sigma^2 > 0$. Une variable aléatoire X suit cette loi si elle peut s'écrire de la forme $X = x_0 + \exp(\mu + \sigma N)$ où N suit la loi normale $N(0, 1)$.
4. La loi **Gamma** translatée qui peut s'écrire $X = x_0 + \gamma$ où γ suit une loi Gamma de paramètres $a > 0$ et $b > 0$ de densité :
$$f_\gamma(y) = \frac{1}{b^a \Gamma(a)} \cdot e^{-x/b} \cdot x^{a-1} \cdot \mathbf{1}_{x \in \mathbb{R}_+}.$$
La paramétrisation utilisée ici est celle du logiciel R (différente de celle utilisée en cours).
5. La loi de **Weibull** translatée où $X = x_0 + W$ avec $x_0 \in \mathbb{R}$ et W qui suit une loi de Weibull de paramètres $\alpha > 0$ et $\beta > 0$, de fonction de répartition :
$$F_W(x; \theta, \alpha) = 1 - \exp\left(-\beta x^\alpha\right) \mathbb{1}_{[0, +\infty[}(x).$$
6. La loi de **Pareto** translatée où $X = x_0 + p$ où p suit une loi de Pareto de paramètres $\alpha > 0$ et $\beta > 0$ de densité
$$f_P(x; \alpha, \theta) = \frac{\alpha \beta^\alpha}{x^{\alpha+1}} \mathbb{1}_{x > \beta}.$$
7. La distribution des valeurs extrêmes généralisées (**GEV**) (generalized extreme values en anglais) de paramètres $\mu \in \mathbb{R}$, $\sigma > 0$ et $\xi \in \mathbb{R}$ de fonction de répartition
$$G_{(\mu,\sigma,\xi)}(x) = \exp\Big(- \Big(1 + \frac{x - \mu}{\sigma} \xi\Big)_+^{-1/\xi}\Big), \qquad x \in \mathbb{R} \text{ et } a_+ = \max\{a, 0\}.$$

Question 4 : Pour chacune de ces lois "classiques" du tableau ci-dessus, décrire l'ensemble des paramètres Θ dans le modèle paramétrique (P_θ, Θ) et tracer une ou plusieurs densités correspondant à des jeux de paramètres θ choisis de manière à ce que l'on puisse distinguer toutes les "formes" de densité possibles typiques des 9 lois (utiliser le package "evir" pour les lois GEV).

Question 5 : Tracer un histogramme des données $x_1, \ldots, x_n$ (fonction "hist" sous R). Parmi la liste de lois proposées, pouvez-vous en éliminer ? Lesquelles ? Pour quelles raisons ? On pourra utiliser les notions de symétrie, support, dispersion etc...