Université Paris Dauphine — Année 2008-2009
Département MIDO
L3 - Statistique Mathématique

Projet de statistique inférentielle

Les étudiants travaillent par groupes de 1 ou 2. La rédaction d'un rapport écrit est obligatoire. Une soutenance orale peut étre exigée.
Les étudiants doivent utiliser le logiciel R[1]

Récupération des données : Chaque étudiant ou binôme, reçoit en TD un numéro de projet. Il doit récupérer le fichier correspondant. Ce fichier contient un vecteur R de 1000 réalisations indépendantes d'une variable aléatoire X de loi inconnue.

1 Objectifs

On fournit à chaque groupe d'étudiants un jeu de données i.e. un vecteur $(x_1, \ldots, x_{1000}) \in \mathbb{R}^{1000}$. Ce vecteur est la réalisation d'un 1000-échantillon $(X_1, \ldots, X_{1000})$ autrement-dit les X_i sont indépendants et identiquement distribués selon une **loi de probabilité inconnue.** On note f la densité de probabilité des X_i et F sa fonction de répartition.

BUT du PROJET : on cherche à retrouver la distribution de probabilité des X_i à partir des observations $(x_1, \ldots, x_{1000})$

CADRE de TRAVAIL : On suppose tout d'abord que la loi de probabilité de X appartient à un ensemble des lois de probabilités classiques dont la liste est rappelée ci-dessous.

[1] voir ci-besoin introduction à R : http ://www.ceremade.dauphine.fr/%7Exian/Noise/R.pdf

Liste des densités classiques

1. La loi continue uniforme sur l'intervalle $[a, b]$, avec $a < b$.
2. La loi exponentielle de paramétres x_0 et b (avec x_0 et $b > 0$) dont une version de la densité $f_{\mathcal{E}xp}$ est
$$f_{\mathcal{E}xp}(x; a, b) = b \exp(-b(x - x_0)) 1\!\!1_{[x_0,\infty[}(x).$$
3. La loi **normale** d'espérance m et de variance σ^2.
4. La loi log**-normale** généralisée de paramétres x_0 ,m et σ^2. Une variable aléatoire X suit cette loi si elle peut s'écrire de la forme $X = x_0 + \exp(m + \sigma U)$ où U suit la loi normale $N(0, 1)$.
5. La loi de **Cauchy avec changement d'origine et d'échelle.** X suit cette loi si elle peut s'écrire $X = b + aY$ (avec $a > 0$) où Y suit la loi de Cauchy usuelle.
6. La loi de **Weibull** de paramètres α et θ dont la fonction de répartition est donnée par
$$F_W(x; \theta, \alpha) = 1 - exp\,(-\theta x^\alpha)\, 1\!\!1_{[0,+\infty[}(x)$$
où les paramétres et sont strictement positifs.
7. La **loi logistique** avec changement d'origine et d'échelle. X suit cette loi si elle peut s'écrire $X = b + aY$ (avec $a > 0$), où Y suit la loi logistique usuelle, de densité
$$f_{log}(x) = \frac{e^{-x}}{(1 + e^{-x})^2}$$

Ainsi, dans un premier temps, on suppose que la densité cherchée f peut-être écrite sous la forme
$$f(x) = f_{\mathcal{T}}(x : \theta)$$
où $\mathcal{T}$ est le type de loi de probabilité (uniforme, normale, Cauchy, Weibull, logistique) et $\theta \in \mathbb{R}^d$ est son paramètre.

Question 0 : Pour chaque densité "classique", choisir arbitrairement un jeu de paramètres θ choisi et tracer la densité correspondante (on tracera les courbes sur un même graphe).

Question 1 : Tracer un histogramme des données ainsi que leur fonction de répartition. Parmi la liste de lois proposées, pouvez-vous en éliminer ? Lesquelles ? Pour quelles raisons ?

2 Estimation de la densité inconnue f

Toutes les lois proposées sont dites *paramétriques* i.e. on peut les écrire sous la forme $x \mapsto f_{\mathcal{T}}(x : \theta)$ où θ est un paramètres. Ex pour la gaussienne : $\theta = (\mu, \sigma^2)$, $f_{gauss}(x; \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(x - \mu)^2)$

Pour un type de densité fixé (ex : gaussienne, exponentielle,...), dépendant d'un paramètre θ estimé par $\widehat{\theta}$, on approche naturellement la densité par $x \mapsto f(x;\widehat{\theta})$.

Question 2 : Pour chacune des lois que vous n'avez pas éliminées, estimer ses paramètres. Superposer à l'histogramme tracé précédemment, les courbes des densités obtenues. Commentez

Indications pour l'estimation

Chaque fois que cela est possible, on utilisera la méthode des moments et celle du maximum de vraisemblance pour déterminer des estimateurs. Lorsque les estimateurs obtenus par ces deux méthodes sont différents, on comparera leurs qualités respectives. Il peut arriver que la méthode des moments ne soit pas applicable, par exemple lorsque l'espérance mathématique n'existe pas (voir exemple de la distribution de Cauchy, ci-dessous). Dans ce cas, on pourra utiliser d'autres indicateurs tels que la médiane (on identifie la médiane empirique avec la médiane théorique), l'écart-interquartile ou encore la probabilité d'appartenir à un intervalle donné. On comparera ensuite les estimateurs obtenus. On fera notamment une étude du biais, de la variance et de la vitesse de convergence

Exemple dans le cas de la loi de Cauchy

Supposons que la loi présumée est la loi de Cauchy avec changement d'origine b et changement d'échelle donné par $a > 0$. On est donc en présence d'une v.a. X de la forme $X = aY + b$, où Y est une loi de Cauchy ordinaire. On souhaite déterminer des estimateurs de a et b. La méthode du maximum de vraisemblance est en principe applicable, mais nécessite la résolution numérique d'un systéme de deux équations non linéaires. Quant à la méthode des moments, on ne peut l'utiliser car les moments théoriques n'existent pas. Dans ce cas, il est opportun de se servir des quantiles. L'identification de la médiane empirique et de la médiane théorique, égale à b, fournit immédiatement une estimation de ce paramètre. Pour estimer le paramétre d'échelle a, on peut identifier l'écart interquartile.

Question 3 : Nous nous sommes placés dans un cadre paramétrique. Estimer la densité par une méthode du noyau. Commentez.

3 Méthode graphique de la droite de Henry

On cherche dans cette partie à retrouver la loi des X_i en utilisant la fonction de répartition et à développer un critère graphique. Cette méthode repose sur la droite de Henry dont on rappelle le principe ci-dessous.

Rappels sur la droite de Henry

Soient $X_1, ..., X_n$ des variables indépendantes de même loi. Soit F la fonction de répartition de cette loi. On appelle échantillon ordonné le vecteur $(X_{(1)}, ..., X_{(n)})$ obtenu en classant les observations par ordre croissant : $X_{(1)} = \min(X_1, \ldots, X_n)$ et $X_{(n)} = \max(X_1, \ldots, X_n)$.

Soit $Z_k = F(X_{(k)})$, on peut montrer que

$$\mathbb{E}(Z_k) = \frac{k}{n+1}, \qquad \mathbb{V}(Z_k) = \frac{k(n-k+1)}{(n+2)(n+1)^2}.$$

On remarque ainsi que la variance de Z_k tend relativement vite (à la même vitesse que $1/n^2$) vers 0. Par conséquent, pour n grand, Z_k est donc très proche de $\dfrac{k}{n+1}$.

Si l'on porte sur un graphique les points de coordonnées $(X_{(k)}, F^{-1}(k/n+1))$, ces points seront approximativement alignés sur le première bissectrice. Aussi, si l'on fait subir une transformation affine à l'échantillon ordonné $Y_k = a + bX_{(k)}$, les points $(Y_k, F^{-1}(k/n+1))$ seront également alignés mais pas forcément le long de la bissectrice.
La méthode consiste donc tout d'abord à ordonner les observations $(x_1, ..., x_n)$ par ordre croissant. Soit $(y_1, ..., y_n)$ le résultat de cette opération. Si l'on trouve une fonction de répartition F pour laquelle les points $(y_k, F^{-1}(k/n+1))$ sont approximativement alignés, on peut suspecter que la vraie loi appartient à la famille de F.

Question 4 : Pour chaque distribution "classique", construire la droite de Henry (on donnera les détails de la construction). Tracer toutes les droites obtenues et commenter les résultats.

Indications

Pour les lois uniformes, normales et exponentielles..., on peut vérifier qu'un changement de variable affine ramène aux lois : uniforme sur l'intervalle $[0, 1]$, normale centrée réduite, exponentielle de paramètre 0 et 1. Pour la loi log-normale, on peut vérifier qu'un changement de variable log-linéaire ($Y = \log(X)$) ramène à la loi normale. Dans ce dernier cas, on travaillera donc sur les logarithmes des données pour tracer la droite. Pour la loi de Weibull, vous devez dans un premier temps estimer les paramètres et ensuite tracer la droite d'Henry.

4 Tests d'hypothèse

Les critères précédents sont graphiques donc subjectifs. Pour construire un critère objectif, on propose de construire des tests d'hypothèses. Plus précisément, soit $\mathcal{T}$ une famille fixée

et et θ_0 un paramètre fixé, on cherche à tester :

$$\mathcal{H}_0 : f = f_{\mathcal{T}}(.;\theta_0) \quad \text{versus} \quad \mathcal{H}_1 : f \neq f_{\mathcal{T}}(.;\theta_0)$$

Question 5 : Appliquer le test de Kolmogorov-Smirnov, celui-ci est implémenté dans la fonction `ks.test` de R. Utilisez cette fonction pour tester pour les distributions que vous n'avez pas éliminées auparavant. Conclusion ?

5 Comparaison à un autre jeu de données

Considérons maintenant le deuxième jeu de données fourni (commun à tous les étudiants). On cherche à savoir si votre jeu de données et ce nouveau jeu de données sont issus de la même loi de probabilité.
On cherche à faire un test de l'hypothèse :

$$\mathcal{H}_0 : (X_1 \dots X_{1000}) \stackrel{\mathcal{L}}{=} (Z_1 \dots Z_{1000}) \quad \text{versus} \quad \mathcal{H}_1 : (X_1 \dots X_{1000}) \stackrel{\mathcal{L}}{\neq} (Z_1 \dots Z_{1000})$$

Question 7 : Tracer l'histogramme de $z_1, \dots, z_{1000}$. Qu'en pensez vous ?

Question 8 : Utilisez la fonction `ks.test` pour effectuer le test précédent. On justifiera l'utilisation de cette fonction et on commentera la sortie du test. Commentez.