Université de Lille 1 – U.F.R. de Mathématiques
M1 MAS – Ingénierie Mathématique – Semestre 1, 2017

TISD [1] – DM 1

Travail à faire en binôme. Un rapport et un script R par binôme doivent être déposés sur Moodle avant le vendredi 6 octobre 23h55. Le nom des étudiants doit apparaître en commentaire au début du script. Le rapport doit être au format PDF (travaillez avec R Markdown, ou Latex, ou OpenOffice, ou Word, etc, puis exportez le fichier en format PDF non éditable). Vous y détaillerez les réponses aux questions, les résultats graphiques, et y ferez part de vos commentaires. Il n'est pas nécessaire d'y inclure les programmes. La clarté et la présentation des rapports & scripts seront appréciés dans la note.

Problème 1

On souhaite étudier la répartition des salaires annuels moyens pour l'année 2016 au sein des différents pays de la zone Euro.

A.1. Pour ce faire, fouiller le site des données statistiques de l'OCDE [2] et exporter les salaires moyens annuels en 2016, en euros et de ces pays seulement, au format CSV. Importer ce fichier CSV sous R via File/Import Dataset ; on ne gardera que les variables COUNTRY et Value.

A.2. Donner une représentation graphique de ces salaires en labellisant chaque point par le nom du pays en question ; on pourra utiliser `text`.

A.3. On veut calculer l'indice de Gini et visualiser la courbe de Lorenz de ce jeu de données.

(a) **(Théorie)** Etant donné $x_1, \ldots, x_n$ des nombres réels positives, donner une formule pour l'indice de Gini associé comme une fonction simple des données, c'est à dire facilement implémentable sous R. Partez de la formule du cours et détaillez les étapes du calcul.

(b) Ecrire une fonction qui, à des données $x = (x_1, \ldots, x_n)$ de valeurs positives, renvoie le graphe de Lorenz associé ainsi que la valeur de l'indice de Gini.

(c) Donner le graphe de Lorenz et l'indice de Gini du jeu de données de l'OCDE.

A.4. Rédigez une courte synthèse de l'étude de ces données où vous présenterez votre conclusion, comme si vous vous adressiez à un public non-mathématicien.

B. Chercher sur internet un autre jeu de données de votre choix qui contient au moins 10 variables représentant des salaires ou assimilés, puis effectuer une analyse similaire à la précédente. Donnez le lien internet de ces données brutes et précisez les transformations préalables que vous avez appliquées si ça a été le cas.

---

1. Responsable : Adrien Hardy. Laboratoire Paul Painlevé, Université des Sciences et Technologies de Lille, Bâtiment M3, Bureau 306. Email : adrien.hardy@math.univ-lille1.fr
2. http ://stats.oecd.org/

Problème 2

Le but ici est d'observer et de comprendre numériquement le théorème central limit (TCL). C'est l'occasion de vous remettre à jour avec la notion de convergence en loi (aussi appelée convergence en distribution) si ce n'est pas déjà fait.

1. (Théorie) Pour tout $n \geq 1$, on considère $X_1, \ldots, X_n$ un échantillon i.i.d d'espérance $\mu \neq 0$ et de variance $\sigma^2 > 0$ finies, ainsi que la moyenne empirique associée

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Quelle est la limite de $\overline{X}_n$ quand $n \to \infty$ et en quel sens ? On considère ensuite la variable

$$\mathtt{E}^{(n)} := \frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu).$$

Après avoir remarqué que l'on peut écrire,

$$\overline{X}_n = \mu + \frac{\sigma}{\sqrt{n}} \mathtt{E}^{(n)},$$

que représente cette variable $\mathtt{E}^{(n)}$? Quelle est sa limite quand $n \to \infty$ et en quel sens ?

2. Pour illustrer ce dernier résultat, choisissez une loi pour X_1, autre qu'une loi normale $\mathcal{N}(\mu, \sigma^2)$. Pour n fixé, simuler $m = 2000$ réalisations $\mathtt{e}_1^{(n)} \ldots, \mathtt{e}_m^{(n)}$ indépendantes de la variable $\mathtt{E}^{(n)}$, puis dessiner l'histogramme, avec l'option `breaks=80`, de ces réalisations où l'on superposera la densité d'une loi $\mathcal{N}(0,1)$. Donner ces histogrammes pour $n = 10$, $n = 100$, et $n = 1000$. Examinez également les QQ-plots des échantillons $\mathtt{e}_1^{(n)} \ldots, \mathtt{e}_m^{(n)}$ par rapport à une loi théorique $\mathcal{N}(0,1)$ pour ces trois valeurs de n. Expliquez ce que ces expériences représentent et commentez vos résultats.

3. (Théorie) Pour être quantitativement plus précis, on considère la fonction de répartition (théorique !) $F^{(n)}$ de la variable $\mathtt{E}^{(n)}$. Si Φ est la fonction de répartition d'une variable $\mathcal{N}(0,1)$, que peut-on dire de $|F^{(n)}(u) - \Phi(u)|$ quand $n \to \infty$? Qu'est-ce que cela veut dire en termes de convergence de variables aléatoires ? L'inégalité de Berry–Esseen, dont on trouvera une référence sur internet ou à la bibliothèque, donne une borne quantitative sur la vitesse de convergence de $|F^{(n)}(u) - \Phi(u)|$. Pour tout n fixé, trouver un exemple de variable aléatoire pour laquelle l'inégalité de Berry–Essen ne vous donne pas plus d'information que l'inégalité triangulaire, à savoir $|F^{(n)}(u) - \Phi(u)| \leq 2$.

4. Reprendre la question 2. avec $n = 100$ et pour X_1 la variable aléatoire issue de la question précédente. Que constatez-vous sur les graphiques ? Soit $F_m^{(n)}$ la fonction de répartition empirique associé à un échantillon i.i.d de taille m de $\mathtt{E}^{(n)}$. Vers quoi tend $F_m^{(n)}$ quand $m \to \infty$ et de quelle façon ? Calculer $|F_m^{(n)}(0) - \Phi(0)|$ sur plusieurs échantillons. Interprétez ces résultats.

5. Dans de nombreux ouvrages non-mathématiques où l'on utilise cependant la statistique, on peut lire que quand $n = 20$, voir même $n = 10$, il est raisonnable de faire l'approximation que $\mathtt{E}^{(n)}$ se comporte déjà comme une variable $\mathcal{N}(0,1)$. Qu'en pensez-vous ?