TISD – FICHE 2

Rappels et compléments de probabilités (avec R)

Adrien Hardy, adrien.hardy@math.univ-lille1.fr

Exercice 1 (Loi des grands nombres)

1. Rappeler l'énoncé de la loi des grands nombres.

2. Illustrer sur **R** ce théorème en utilisant un échantillon $x_1, \dots, x_n$ de loi de Bernoulli. Pour cela, faire varier n et donner une représentation graphique en fonction de n ; on pourra utiliser la commande `cumsum`.

3. Même question avec une loi de Cauchy. Commenter.

Bonus : Quelle est la loi de $\frac{1}{n}\sum_{j=1}^{n} X_j$ *quand* $X_1, \dots, X_n$ *est un échantillon de loi de Cauchy ?*

Exercice 2 (Variables exponentielles)

1. Rappeler [1] la définition d'une variable de loi $\mathcal{E}(\lambda)$ exponentielle de paramètre $\lambda > 0$, puis calculer à la main son espérance, sa fonction de répartition et sa fonction quantile.

2. Représenter sur un même graphique les densités de cette loi pour $\lambda = 0.5$ et $\lambda = 2$. Même chose pour les fonctions de répartition, puis pour les fonctions quantile.

3. (Propriété d'absence de mémoire) [2] Si $X \sim \mathcal{E}(\lambda)$ et si $0 < s < t$, prouvez (à la main) que $\mathbb{P}(X > t + s \,|\, X > s) = \mathbb{P}(X > t)$. Proposez une expérience numérique pour vérifier cette propriété, en prenant des approximations discrètes de $\mathbb{P}(X > t + s \,|\, X > s)$ et $\mathbb{P}(X > t)$ adaptées.

NB : Les lois exponentielles sont typiquement utilisées pour modéliser des durées de vie de phénomènes "sans usure". Par exemple, le temps écoulé avant qu'un composant électronique ne tombe en panne, ou la durée de vie d'un atome radioactif (loi de Rutherford et Soddy).

Exercice 3 (Fonctions de répartition)

Dans cet exercice, on notera F_n la fonction de répartition empirique associée à des observations $x_1, \dots, x_n$ i.i.d de loi gaussienne standard $\mathcal{N}(0, 1)$.

1. Et si vous ne l'avez jamais vue, faites un tour sur Wikipedia !
2. On rappelle que la probabilité conditionnelle $\mathbb{P}(A|B)$ d'un évènement A sachant un évènement B est définie par $\mathbb{P}(A \cap B)/\mathbb{P}(B)$.

1. Pour $n = 100$, tracer F_n en utilisant la commande `sort()` pour ordonner les observations. Superposer la fonction de répartition théorique F sur le même graphique. Recommencer en utilisant cette fois la commande `ecdf()` pour appeler la fonction de répartition empirique.

2. On définit $I = \{-4, -3.99, -3.98, \ldots, 4\}$ et on considère la quantité

$$D(n) = \max_{u \in I} \big|F_n(u) - F(u)\big|.$$

C'est une approximation discrète de

$$\sup_{u \in \mathbb{R}} \big|F_n(u) - F(u)\big|.$$

En effet, calculer $\mathbb{P}(|X| > 4)$ avec **R** pour $X \sim \mathcal{N}(0, 1)$ et interpréter le résultat.
Que savez-vous de $D(n)$ quand $n \to \infty$?

3. Programmer une fonction qui retourne $D(n)$ et représenter graphiquement $D(n)$ comme fonction de n avec $n = 100, 200, 300, \ldots, 20\ 000$.

Aide : Pour vectoriser une fonction que l'on a créée, on utilise `Vectorize`.

Bonus : On s'intéresse à la vitesse à laquelle $D(n)$ converge quand $n \to \infty$: Chercher empiriquement $\alpha > 0$ tel que $n^\alpha D(n)$ a une limite non triviale quand $n \to \infty$ (c'est-à-dire ne tend ni vers zéro ni vers l'infini). Il existe effectivement un objet limite, qui est à la base du test de Kolmogorov-Smirnov. Cf. suite du cours

Exercice 4 (Moyenne VS Médiane)

Calculer la moyenne et la médiane (`median`) de $8, 5, 2, 9, 8, 3, 6, 4, 100$, puis de la même série de chiffres sans le 100, et interpréter la différence.

Exercice 5 (Boîtes à moustaches)

On utilise le jeu de données `chickwts` disponible sous **R**. *Questions-préliminaires-que-vous-devrez-vous-poser-à-chaque-fois-que-vous-découvrez-un-jeu-de-données-à-partir-de-maintenant* : Que représente-il ? (On pourra utiliser l'aide.) Comment sont structurées ces données : data frame ? dimensions ? nom des variables ?

1. Après avoir tapé `meal <- chickwts$feed` et `wts <- chickwts$weight` par commodité, que représente G_1 défini par `G1 <- wts[meal=='horsebean']` ? Donner la boîte à moustaches de ce groupe (`boxplot`).

2. Donner une représentation simultanée des boîtes à moustaches des six groupes de régimes alimentaires. Interprétez ces graphiques : Si vous deviez choisir le régime alimentaire qui maximise le poids des poulets, quel serait votre choix ?

3. Au fait, pourquoi les "moustaches" font intervenir ce facteur 1.5 ? Réponse de l'inventeur (John W. Tukey) : "*Because* 1 *is too small and* 2 *is too large*". Pour comprendre, créer une fonction de $M(t)$ qui la probabilité qu'une variable $X \sim \mathcal{N}(0, 1)$ tombe entre la $t-$moustache basse et la $t-$moustache haute définies respectivement par $Q_1 - t(Q_3 - Q_1)$ et $Q_3 + t(Q_3 - Q_1)$, où Q_1 et Q_3 sont le premier et troisième quartile théoriques de X. Donner $M(1)$, $M(1.5)$ et $M(2)$, puis conclure.

Exercice 6 (Lois Gamma)

La loi $\Gamma(k, \theta)$ de paramètre de forme $k > 0$ et de taux $\theta > 0$ (ou d'échelle $1/\theta$) est la loi continue sur $[0, +\infty]$ de densité

$$f(x) = \frac{\theta^k}{\Gamma(k)}\, x^{k-1} \mathrm{e}^{-x\theta} \mathbf{1}_{[0,+\infty[}(x),$$

où $\Gamma(k)$ est la fonction Gamma.

1. Observer qu'une loi exponentielle $\mathcal{E}(\lambda)$ est une loi Gamma dont on précisera les paramètres.

2 Représenter sur le même graphique la densité d'une loi Gamma pour

$$(k, \theta) \in \{(0.5, 1), (1, 1), (2, 1), (2, 2), (2, 3)\}.$$

Propriété : On admettra que si X et Y sont deux variables indépendantes de lois respectives $\Gamma(k_1, \theta)$ et $\Gamma(k_2, \theta)$, alors $X + Y$ suit une loi $\Gamma(k_1 + k_2, \theta)$.

3. Si X, Y, Z sont trois variables i.i.d $\mathcal{E}(2)$, quelle est la loi de $X + Y + Z$? Pour le vérifier empiriquement, simuler un vecteur x de 1000 réalisations de la somme de 3 variables aléatoires i.i.d $\mathcal{E}(2)$. Donner une représentation graphique approximée de la densité théorique f de $X + Y + Z$, et tracer sur le même graphique f ainsi que les densités théorique des lois $\mathcal{E}(2)$ et $\Gamma(10, 2)$. Ensuite, faire un Q-Q plot pour comparer la distribution de x avec f, mais aussi avec $\mathcal{E}(2)$ et $\Gamma(10, 2)$.

Aide : Pour faire un QQ-plot de deux échantillons empiriques x, y, on utilise `qqplot(x,y)`. Pour comparer un échantillon x à une loi théorique "loi", on utilise `qqPlot(x, "loi", "paramètres")`, où `qqPlot` (attention à la majuscule) provient de la bibliothèque `car` qu'il faudra installer : Pour ce faire aller sous R studio dans "Packages" puis "Install" et cherchez `car`. Une fois installée, pour utiliser cette bibliothèque il suffira d'éxectuer `library(car)`.

Bonus : Montrer cette propriété des lois Gamma.

Exercice 7 (Elections et sondages)

À la veille du second tour des élections présidentielles, on cherche à estimer la proportion $p \in [0, 1]$ de la population qui a l'intention de voter pour le candidat A (et donc une proportion $1 - p$ a l'intention de voter pour le candidat B). On interroge n personnes, choisies de façon indépendante dans la population, et on suppose qu'elles répondent honnêtement. De chaque répondant $j \in \{1, \ldots, n\}$, on obtient une donnée x_j qui vaut 1 si il compte voter pour le candidat A, et 0 si il compte voter pour le candidat B. On fera la modélisation mathématique suivante : On suppose que $x_1, \ldots, x_n$ sont des réalisations d'un échantillon $X_1, \ldots, X_n$ d'une variable de loi de Bernoulli $\mathcal{B}(1, p)$. Le paramètre p est inconnu, et nous voulons l'estimer.

1. On considère la moyenne empirique $\overline{X}_n := \frac{1}{n} \sum_{j=1}^{n} X_j$. Calculer son espérance, sa variance et donner sa limite lorsque $n \to +\infty$.

2. Rappeler l'inégalité de Tchebychev, puis montrer que pour tout $t > 0$,

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| \geqslant t\right) \leqslant \frac{1}{4t^2 n}. \tag{1}$$

(a) Combien faut-il interroger de personnes pour que la probabilité que "l'écart entre l'estimateur $\overline{X}_n$ et l'inconnue p est supérieur à 1%" soit inférieure à 5% ?

(b) Télécharger sur Moodle le jeu de données "Intention_Vote.txt" (c'est un data frame) et estimer le paramètre p de ce jeu de données avec une probabilité de se tromper inférieure à 5%, et donnant la précision de l'estimation.

Remarque : L'inégalité de Tchebychev n'est pas l'inégalité la plus puissante à utiliser dans ce contexte ; on préfèrera utiliser l'inégalité d'Hoeffding, cf. cours.