Université de Toulouse 2015-2016
L3 SID

TP1 : Introduction à R et illustrations de théorèmes classiques

Lors de ce premier TP, nous allons nous familiariser avec le logiciel R. Nous verrons d'abord comment simuler des lois et manipuler leurs fonction de répartition et densité. Ensuite, nous illustrerons des théorèmes classiques de probabilité sur des exemples.

1 Introduction à la simulation avec R

Cette première partie s'appuiera sur l'aide du logiciel (en anglais), que l'on peut trouver dans la rubrique "aide", "manuels" et "an introduction to R". Nous nous concentrerons sur la section 8 de ce manuel : Probability distributions.

1.1 Réalisations de variables aléatoires

Pour générer des réalisations de variables aléatoires, on utilise la commande "r", que l'on accole au nom de la loi que l'on souhaite simuler. Taper par exemple dans la console "rnorm" : la liste des paramètres est alors détaillée. Simuler une variables aléatoire gaussienne X de loi normale centrée réduite. Comment simuler un vecteur de 3 réalisations indépendantes de loi normale de moyenne 2 et de variance 4 ? Simuler ensuite un vecteur de réalisations indépendantes de loi béta de paramètres 1 et 2, puis de loi exponentielle de paramètres 1, et enfin de loi binômiale pour $n = 5$ et $p = 0.5$.

1.2 Fonction de densité

La commande est "d". Taper par exemple dans la console "dnorm" pour comprendre le fonctionnement de cette commande. Combien vaut la densité de la loi normale centrée réduite en 0 ? Vérifier avec la commande. Essayer la même chose pour la loi binomiale. Que se passe-t-il si l'on demande la fonction de densité en 0.3 ? Pourquoi ?

L'intérêt de cette commande réside surtout dans le fait qu'elle permet de tracer facilement des fonctions de densité. Aller voir pour cela la fonction "plot" dans l'aide (section "Graphical procedure"). Commencer par créer un vecteur d'abscisses par la commande "$x < --10:10$". Tracer ensuite la fonction de densité de la loi normale centrée réduite en ces points. Relier ces points par des segments. Pour avoir plus de précision, il faut bien entendu mettre plus de points dans le vecteur d'abscisses. Pour cela, on peut utiliser la commande "$x < -seq(-5, 5, by = .2)$". Il existe plusieurs paramètres réglables pour avoir des courbes de différents design.

Avec le logiciel R il est aussi possible de tracer des histogrammes à partir d'un échantillon. Aller voir l'aide sur la fonction "$hist$". Tracer alors un histogramme d'un vecteur de 10000 réalisations indépendantes d'une loi normale centrée réduite.

Dans le paragraphe de l'aide sur les fonctions de densité, il est résumé comment tracer un histogramme et superposer la densité réelle (en remplaçant "plot" par "lines"). Superposer alors la densité de la loi normale standard à l'histogramme précédemment tracé.

1.3 Fonction de répartition

Cette fois-ci la commande est "p". Aller voir dans l'aide. Que vaut la fonction de répartition de la loi normal centrée réduite en 0 ? Vérifier avec la commande. Tracer grâce à cette même aide, la fonction de répartition de la loi normale standard avec le vecteur d'abscisse de votre choix. Attention au type de la courbe, il faut une allure cohérente avec une fonction de répartition.

Ce logiciel nous permet aussi de tracer des fonctions de répartitions empiriques. Créer un vecteur de 100 réalisations indépendantes de la loi de votre choix. Comment pourrait-on construire une fonction de répartition empirique ? Aller regarder l'aide sur la fonction "*ecdf*". Tracer alors la fonction de répartition empirique de votre vecteur (faites varier la taille du vecteur), puis superposer la fonction de répartition.

2 Illustrations de théorèmes

On considère une variable X à valeurs dans $\{0, 1, 3\}$, distribuée comme suit : $P(X = 0) = 0.5$, $P(X = 1) = 0.25$, $P(X = 3) = 0.25$.

1. Proposer une façon de simuler X avec
   Suggestion : utiliser 2 variables de Bernoulli de paramètre 1/2.
2. On considère $X_1, X_2, \ldots$ une suite infinie de variables indépendantes de même loi que X. Soit $\overline{X}_n$ la moyenne empirique des X_i pour $i \in \{1, \ldots, n\}$.
   (a) Simuler la suite des X_i pour $i \in \{1, \ldots, 10000\}$.
   (b) Produire un graphique représentant l'évolution de $\overline{X}_n$ pour n variant de 1 à 10000.
   Que constate-on ? Pouvait-on s'y attendre ?
3. On désire maintenant approfondir comment $\overline{X}_{500}$ varie autour de sa valeur moyenne.
   (a) Proposer une transformation affine de $\overline{X}_{500}$, de la forme $Y = a * (\overline{X}_{500} + b)$, qui suive approximativement la loi $\mathcal{N}(0, 1)$.
   (b) Simuler avec 10000 réalisations indépendantes de la variable Y. Nous les noterons $Y_1, \ldots, Y_{10000}$.
   (c) Confirmer l'approximation gaussienne en réalisant un histogramme des valeurs prises par les Y_j pour $j \in \{1, \ldots, 10000\}$, sans oublier de tracer la densité gaussienne correctement renormalisée. Commenter l'écart entre l'histogramme et la densité gaussienne.
4. Recommencer la question 2 avec la loi de Cauchy. Que se passe-t-il ? Pourquoi ?
5. Simuler 100 fois une suite de $X_i^{(j)}$ pour $i \in \{1, \ldots, 20\}$. Pour chaque suite, construire l'intervalle de confiance à 95% pour $\overline{X}^{(j)}$. Compter le nombre d'intervalles de confiance contenant $\mathbb{E}(X)$. Conclure.