TP 1 : Simulation, visualisation de données avec R

Ce TP n°1 a pour but la maîtrise des outils classiques de simulation probabiliste et de réprésentation de données.

Exercice 0 : Reprise en main de la console : calculs rapides Recopier dans la console chacun de ces calculs, puis l'éxécuter (en appuyant sur Entrée). Assurez-vous de bien comprendre le résultat renvoyé par R.

`3+1.2, 2-7.8, 2/5, 2*5.5, sqrt(100), log(2), exp(1), log(exp(3)), log10(10000), sin(pi/2), cos(pi)`

La commande `?` devant une fonction permet d'obtenir de l'aide. Par exemple, pour demander de l'aide sur la fonction `exp` on écrira:

```
?exp
```

R permet aussi de définir des tableaux à l'aide de `c(...)`. Définir un tableau `tab` contenant 6 nombres réels de votre choix. Ensuite, que renvoient les commandes suivantes ?

`tab[2], tab[-2], tab[2:5], tab[c(2,5)], tab+0.5, tab/2, tab*5, exp(tab)`

Pour la suite du TP, y a quatre commandes simples à retenir avec R pour chaque loi usuelle : `rmaloi` permet de simuler des variables selon maloi, `dmaloi` calcule la densité de maloi, `pmaloi` pour la fonction de répartition, et `qmaloi` pour les quantiles de maloi.

Exercice 1 : Loi normale, commandes classiques pour les lois continues

1. Simulez une réalisation d'un échantillon de $n = 100$ variables independantes et de même loi normale de moyenne $\mu = -2$ et de variance 9. On utilisera donc la fonction `rnorm` qui prend en paramètres n, `mean` (la moyenne) et `sd` (l'écart-type).

Calculer la moyenne et la variance de la réalisation.

Afficher la représentation de la réalisation sous la forme d'une boîte à moustaches (fonction `boxplot`)

2. Tracer l'histogramme du jeu de données ainsi obtenu (fonction `hist` avec option `freq=F`), et réfléchir à comment régler le paramètre `breaks`. Puis à l'aide de la fonction `curve` (avec option `add=T,col="red"`), superposer à l'histogramme la densité théorique de la loi en question.

3. Répéter les deux premières questions avec $n = 1000$ puis $n = 10000$. Qu'observe-t-on?

4. Pour une réalisation $(x_1, \ldots, x_n)$ d'un échantillon, la fonction de répartition empirique F_{emp} est définie par

$$\forall y \in \mathbb{R},\ F_{emp}(y) = \frac{1}{n}\sharp\left\{1 \leq i \leq n,\, x_i \leq y\right\}.$$

A l'aide des fonctions `plot` et `ecdf`, représenter la fonction de répartition empirique de la réalisation. Superposer au graphique la fonction de répartition (théorique) de la loi normale étudiée. Qu'observe-t-on ?

Exercice 2 : Lançons des fléchettes ! On dispose d'un carré de côté 2. On place l'origine O du repère au centre du carré $[-1,1] \times [-1,1]$. On considère tout d'abord l'expérience aléatoire suivante : on lance au hasard des fléchettes dans le carré, avec une abscisse et une ordonnée indépendantes et toutes deux de loi uniforme sur $[-1,1]$.

1. Simuler $n = 100$ lancers indépendants de fléchettes. Les représenter graphiquement sur le carré qui sert de cible.

2. On repère la qualité d'un lancer à la distance D de la flèche au centre de la cible. Tracer l'histogramme des réalisations de D.

Quelle est la moyenne de la réalisation ? Et la moyenne théorique (espérance) de D ?

3. Refaire l'histogramme de D pour $n = 10^7$. (Indication : là encore, attention à bien calibrer `breaks`).

Quelle semble être la loi de D sachant $D \leq 1$? Interprétez.

4. Proposez deux méthodes pour estimer $\mathbb{P}(D \leq 1)$: l'une à l'aide de simulations, l'autre à l'aide d'un calcul.

5. Maintenant arrive un autre lanceur de fléchettes qui semble bien plus aguerri, et que l'abscisse x et l'ordonnée y du lancer sont deux variables indépendantes, gaussiennes centrées de variance $1/2$. Calculer la moyenne et représenter l'histogramme de réalisations indépendantes de D sous ce nouveau modèle. Commentez.

Selon vous, ce joueur est-il vraiment meilleur ?

Exercice 3 : Promenons-nous sur $\mathbb{Z}$... Un promeneur habite dans un monde à une seule dimension. Il décide d'aller se balader pour prendre l'air entre deux confinements. On suppose que ces déplacements ne peuvent se faire que sur l'axe des entiers relatifs ($\mathbb{Z}$) et on les modélise comme suit: Au temps $t = 0$, il est chez lui, au point 0, et à chaque pas de temps, il se déplace indépendemment vers la droite (de $+1$ donc) avec une probabilité p, ou bien vers la gauche (de -1) avec une probabilité $q = 1 - p$. La donnée successive de sa position en fonction du temps est appelée une marche aléatoire.

Une façon simple de simuler une telle marche aléatoire du temps $t = 0$ au temps $t = T$ est la suivante : il suffit de simuler un vecteur de taille T de variables $X_1, \ldots, X_T$ indépendantes et de même loi, satisfaisant

$$\mathbb{P}(X_1 = -1) = 1 - p, \quad \mathbb{P}(X_1 = +1) = p.$$

Ensuite, il suffit de réaliser que la position de notre promeneur au temps t, notée S_t, est donnée par

$$S_t := \sum_{s=1}^{t} X_s.$$

1. Simuler à l'aide de R une marche aléatoire de taille $T = 30$, avec $p = 0.65$. Stocker cette marche dans un vecteur *marche*. (Indication : pour simuler les X_i on pourra utiliser des variables i.i.d. de loi uniformes sur [0,1], puis regarder si elles sont inférieures à p...)

2. Refaire une simulation pour $T = 200$, et représenter graphiquement la marche aléatoire S_t en fonction du temps t: faites cela plusieurs fois. Le promeneur rentrera-t-il chez lui un jour ? Interprétez.

3. Quelle valeur de p semble être la plus judicieuse pour espérer rentrer que le promeneur rentre un jour chez lui ? Refaire plusieurs simulations avec la nouvelle valeur de p, pour $T = 2000$, et représenter graphiquement la marche aléatoire S_t en fonction du temps t. Qu'observe-t-on cette fois-ci ? Interprétez.

Dans toute la suite on prend $p = 1/2$. On s'intéresse au premier temps T_0 de retour en 0, c'est-à-dire au premier instant où le promeneur revient chez lui. Mathématiquement,

$$T_0 := \inf \{n \in \mathbb{N}^*,\ S_n = 0\},$$

avec la convention $T_0 = \infty$ si $\forall n \in \mathbb{N}^*, S_n \neq 0$. On peut en fait montrer que presque sûrement $T_0 < \infty$.

4. Compléter le code ci-dessous, qui définit la fonction `T0()` (sans argument) qui simule et renvoie le temps de retour en 0 pour une marche du promeneur de paramètre $p = 1/2$.

```
T0 <- function(){
t = 0
position = 0
while (F) {# <-- A compléter
# t=  <-- A compléter
position <- position + 2*(runif(1)<0.5)-1}
t
}
```

5. A l'aide de la fonction `T0`, simuler $m = 1000$ valeurs de $T0$. Calculer son espérance (empirique), sa variance (empirique). Représenter son histogramme. A votre avis, que vaut l'espérance (théorique) de T_0 ?