Université Rennes 2 — Année 2019-2020
Master 1 — Madison Giacofci/Pierre Gloaguen

Probabilités et simulations sous R

Nous allons à présent étudier la gestion de l'aléatoire et des probabilités avec le logiciel R. Ouvrez une session R avec Rstudio et placez-vous dans votre répertoire personnel.

1 Lois de probabilités usuelles

Des fonctions dédiées permettent d'accéder aux densités, fonction de répartition et fonctions quantiles des lois de probabilités usuelles. Par ailleurs, il est aussi possible de simuler des échantillons suivant ces lois. Ces fonctions sont construites suivant le format suivant :

— un préfixe spécifiant la fonctionnalité : `d` pour la densité, `p` pour la fonction de répartition, `q` pour la fonction quantile et `r` pour la simulation sous la loi considérée.

— un suffixe spécifiant la loi à laquelle on s'intéresse. Les principaux suffixes sont donnés dans le tableau ci-dessous.

loi	suffixe	paramètres	arguments
binomiale	`binom`	nb essais proba succès	`size` `prob`
géométrique	`geom`	proba succès	`prob`
normale	`norm`	moyenne écart-type	`mean` `sd`
uniforme	`unif`	borne inf borne sup	`min` `max`
χ^2	`chisq`	ddl	`df`
student	`t`	ddl	`df`
fisher	`f`	ddl 1 ddl 2	`df1` `df2`

La liste présentée dans le tableau est loin d'être exhaustive. L'ensemble des lois usuelle présentes dans la version de base du logiciel peut être consultée dans l'aide de `Distributions`.

Prise en main : Créer une fonction qui simule une variable aléatoire de loi Bernoulli de paramètre p en utilisant `runif()`.

1.1 Lois discrètes

1. Reprenez la fonction `nombre.mystère` créée dans le TD précédent (Section 1.2.2, question 2). Transformez-la pour faire des tirages entiers compris entre 1 et 10 (au lieu de 1 et 100). Choisissez un entier entre 1 et 10 et appliquez votre fonction 1000 fois de suite. Gardez le résultat dans un vecteur.
2. Représentez les résultats obtenus sous forme de fréquence relative d'apparition (utilisez la fonction `table` couplée à la fonction `plot`). Tracer sur le même graphe en rouge les probabilités théoriques attendues (on utilisera l'argument `type="h"` de la fonction `plot`).

1.2 Lois continues

1. Représentez sur un même graphique la densité et la fonction de répartition de la loi normale centrée réduite (on prendra l'intervalle $[-4, 4]$).
2. Calculez le quantile d'ordre 0.95.
3. Ajoutez un axe horizontal en 0. Á l'aide de la fonction `segments`, ajouter un segment vertical allant de l'axe des abscisses à la fonction de densité au niveau du quantile d'ordre 0.95. A quoi correspond l'aire sous la courbe à gauche du segment ?
4. Ajoutez sur le même graphique les densités des lois de student à 5 et 30 degrés de liberté en utilisant des couleurs différentes.

2 Simulations sous R

Nous avons vu qu'on peut simuler des échantillons sous les lois de probabilités usuelles grâce au préfixe `r` associée au suffixe de la loi voulue. Une autre fonction utile est donnée par la fonction `sample` permettant de faire un tirage de taille `n` dans un vecteur `x` donné, avec ou sans remise et selon les probabilités données par `prob`.
L'ensemble de ces fonctions est basée sur les techniques de génération de nombres aléatoires. Ce problème est assez complexe étant donné qu'un ordinateur effectue uniquement des opérations déterministes. L'objectif de la génération de nombres aléatoires est alors de produire des séquences de nombres se rapprochant au maximum d'une séquence purement aléatoire. De ce fait, à l'appel d'une fonction de simulation, une *graine aléatoire* est fixée par R en fonction de la date actuelle (temps machine) et du processus en cours. Les résultats obtenus sont donc différents à chaque appel. Si l'on souhaite au contraire que la simulation effectuée donne les mêmes résultats, pour des raisons de reproductibilité, il est possible pour l'utilisateur de fixer la graine aléatoire à l'aide de la commande `set.seed(n)` où `n` est un entier choisi par l'utilisateur.

2.1 Simulation d'un vecteur de loi exponentielle et calcul de vraisemblance

1. Simulez un vecteur de taille 100 suivant une loi exponentielle de paramètre $\theta = 2$ et stockez-le dans un vecteur `ech`. La dénsité de probabilité d'une loi exponentielle ($Exp(\theta)$) est :

$$f_\theta(x) = \theta e^{-\theta x} \mathbf{1}_{[0,+\infty)}(x)$$

2. Calculez la fonction de log-vraisemblance de cet échantillon pour des valeurs de θ comprises entre 0.1 et 4. Tracez-la sur un graphique.
   Rappel : Pour un échantillon $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, issus de tirages indépendants de loi $Exp(\theta)$ la log vraisemblance est donnée par :

$$l(\theta|\mathbf{x}) = \sum_{i=1}^{n} \ln\left(f_\theta(x_i)\right).$$

3. Quel est l'expression de l'estimateur du maximum de vraisemblance pour l'estimateur θ ? Calculez-le et tracez une ligne verticale rouge pour le représenter sur le graphique. Commentez.

2.2 Simulation - loi normale "simple"

1. Fixer une graine aléatoire. Simulez un vecteur de taille 10 suivant une loi normale de moyenne 2 et de variance 3 (ATTENTION aux paramètres de la loi normale sous R).
2. Appliquez le même code sur l'ordinateur de votre voisin. Commentez.

3. Simulez un échantillon de taille 1 sous la loi normale de moyenne et de variance :

$$\mu = \begin{bmatrix} 0 \\ 100 \\ 1000 \end{bmatrix} \quad \text{et} \quad \Sigma = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 100 \end{bmatrix}$$

2.3 Simulation - loi normale multivariée

Lorsque l'on souhaite simuler un vecteur de loi gaussienne, nous avons vu qu'il est facile de le faire à l'aide de la fonction `rnorm` dés lors que ses composantes sont indépendantes. Lorsque ce n'est pas le cas, on peut utiliser le résultat suivant :

Proposition. (Cholesky) Si Σ est une matrice symétrique définie positive, alors il existe une matrice triangulaire inférieure L telle que $\Sigma = LL^t$.

Conséquence. Si $X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ avec des composantes indépendantes et telles que $X_i \sim \mathcal{N}(0,1)$, alors $Y = \mu + LX$ suit une loi normale de moyenne $\mu = (\mu_1, \ldots, \mu_n)$ et de matrice de variance covariance Σ.

1. À l'aide de la fonction `chol`, calculez la décomposition de Cholesky de la matrice suivante

$$\Sigma = \begin{bmatrix} 0.1 & 0.2 & 0.2 \\ 0.2 & 10 & 0.2 \\ 0.2 & 0.2 & 100 \end{bmatrix}.$$

2. Simulez une réalisation d'une loi normale multivariée centrée et de variance-covariance Σ.

3 Illustrations des théorèmes limites

3.1 Loi des grands nombres

1. Rappelez la loi des grands nombres.
2. Simulez un échantillon de taille 1000 $(x_1, \ldots, x_{1000})$ issu d'une loi de Bernoulli de paramètre $p = 0.4$ que vous stockerez dans un vecteur.
3. Calculez les moyennes successives $m_\ell = \frac{1}{\ell}\sum_{i=1}^{\ell} x_i$ pour ℓ allant de 1 à 1000. Tracez m_ℓ en fonction de ℓ et ajoutez la droite d'équation $y = 0.4$. Commentez.

3.2 Théorème central limite

1. Rappelez le théorème central limite.
2. Rappelez le lien entre loi de Bernoulli et loi binomiale.
3. Soit $p = 0.5$ et $n = 10$. Simulez $N = 1000$ réalisations $(s_1, \ldots, s_N)$ d'une loi binomiale de paramètres n et p. Stockez dans un vecteur `Z10` les quantités

$$\frac{s_i - n \times p}{\sqrt{n \times p \times (1-p)}}$$

4. Faire de même avec $n = 50$ et $n = 500$ pour créer les vecteurs `Z50` et `Z500`.
5. Sur une même fenêtre graphique, représentez sur des graphes différents les histogrammes des variables `Z10`, `Z50` et `Z500`. Vous superposerez à chaque fois la densité de la loi normale centrée réduite.