Simulations de variables aléatoires

Travaux dirigés

Pierre Gloaguen

La plupart des exercices de cette feuille nécessite la confection de programme en `R`.

Afin de garder trace de vos exercices, pensez à sauvegarder le script associé, voire à répondre à l'exercice dans un fichier *Rmarkdown* (extension `.Rmd`).

L'environnement `Rstudio` est plus que vivement conseillé pour programmer en `R`.

1 Générateurs pseudos aléatoires

1.1 Génération de loi uniforme

1. À l'aide du logiciel R, programmez un générateur à congruences pour la loi uniforme. Ce générateur prendra la forme d'une fonction prenant en argument:

- Un entier `n` donnant la taille de l'échantillon voulu.
- 4 entiers `a, m, c, x0` correspondant aux paramètres du générateurs vu en cours.

2. À l'aide de cette fonction, générer un échantillon de taille 10000 pour les valeurs

- `a = 41358`
- `m = 2^31 -1`
- `c = 0`

et la graine de votre choix. Refaites la même procédure avec

- `a = 3`
- `m = 2^31 -1`
- `c = 0`

et

- `a = 101`
- `m = 2311`
- `c = 0`

Vous stockerez chacun des échantillons obtenus

3. Pour chacun des échantillons obtenus, tracez l'histogramme empirique. Quels échantillons vous semblent tirés selon une loi uniforme $U[0, 1]$? En utilisant la fonction `ks.test`, effectuez un test de Kolmogorov-Smirnoff d'adéquation pour la loi uniforme. Que concluez vous sur la qualité des 3 générateurs?

4. Pour chacun des échantillons $(u_1, \dots, u_{10000})$ obtenus, tracez u_n en fonction de u_{n-1}. Que pouvez vous conclure sur la qualité des 3 générateurs?

2 Méthode d'inversion

2.1 Loi exponentielle

On rappelle qu'une variable aléatoire X est de loi exponentielle de paramètre $\lambda > 0$ si elle a pour fonction de densité $f_X(x) = \lambda \mathrm{e}^{-\lambda x} \mathbf{1}_{x \geq 0}$

1. En utilisant la méthode d'inversion, proposez un algorithme de simulation pour une variable aléatoire exponentielle.
2. Ecrire une fonction `R` mettant en oeuvre cette algorithme. Cette fonction prendra deux paramètres en entrée:

- `n` La taille de l'échantillon;
- `lambda` Le paramètre de la loi exponentielle

Vous testerez la qualité de votre fonction sur un échantillon de taille 10000, en comparant graphiquement l'histogramme empirique obtenu à la densité de la loi exponentielle correspondante.

2.2 Loi discrète

On considère une variable aléatoire discrète X à valeurs dans l'ensemble $\{1, \ldots, K\}$, dont la loi est définie par le vecteur de probabilité $(p_1, \ldots, p_K)$, i.e.:

$$\mathbb{P}(X = k) = p_k \tag{1}$$

$$\sum_{k=1}^{K} p_k = 1 \tag{2}$$

1. Pour tout $u \in]0, 1[$, écrire l'expression de l'inverse généralisée de la fonction de répartition de X.
2. En déduire un algorithme de simulation pour toute variable aléatoire discrète dans un ensemble fini.
3. Utilisez cet algorithme de simulation pour simuler un échantillon de taille 10000 loi binomiale de paramètres $n = 10$ et $p = 0.5$ avec `R`. Vous comparerez les fréquences obtenues avec les fréquences théoriques.

3 Méthode de rejet

3.1 Simulation d'une loi de Poisson pour $\lambda < 1$

1. On utilisant le résultat de l'exercice 2.2, proposez un algorithme pour simuler une variable aléatoire de Bernouilli de paramètre $p \in]0, 1[$.
2. En déduire un algorithme pour simuler une loi géométrique de paramètre p sur $\mathbb{N}$.
3. On souhaite obtenir un échantillon d'une loi de de Poisson de paramètre $\lambda \in]0, 1[$ par méthode d'acceptation rejet. On se propose d'utiliser comme loi de proposition la loi géométrique sur $\mathbb{N}$ de paramètre $1 - \lambda$. Définir l'algorithme de rejet correspondant.
4. Quelle est la probabilité d'acceptation dans l'algorithme de rejet?
5. Faites une fonction `R` permettant de générer une loi géométrique de paramètre p. Utiliser cette fonction dans une autre fonction `R` permettant de simuler selon une loi de Poisson de paramètre $p \in]0, 1[$. Simuler ainsi un échantillon de taille 10000. Comparer la distribution obtenue à celle de la vraie loi.

3.2 Loi uniforme sur le disque unité

1. À partir d'une variable aléatoire uniforme sur $[0,1]$, proposez une transformation pour simuler une loi uniforme sur $[-1,1]$.
2. Proposer une méthode d'acceptation rejet pour simuler, à partir de deux variables aléatoires indépendantes de loi uniforme sur $[-1,1]$, une variable aléatoire uniforme sur le disque unité.
3. Quelle est la probabilité d'acceptation de l'algorithme?
4. Ecrire une fonction `R` mettant en place la génération de variables aléatoires sur le disque unité. Dans cet algorithme, gardez en mémoire le nombre d'essais nécessaire avant chaque acceptation.
5. Générer un échantillon de taille 10000. Vérifiez graphiquement que ces points sont uniformément répartis sur le disque unité. Vérifiez également que le nombre d'essais moyens avant acceptation est en adéquation avec ce qui est attendu.

3.3 Proposition optimale

La loi normale tronquée de support $[b,+\infty[$ est définie par la densité f proportionnelle, pour tout réel x, à

$$f_1(x) = \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\mathbf{1}_{x\geq b}, \quad \text{avec } \mu>0,\ \sigma>0.$$

On propose de simuler suivant la loi de densité f par une méthode de rejet.

3.3.1 Méthode naïve

1. On note Φ la fonction de répartition de la loi normale centrée réduite. Montrer que f satisfait l'inégalité suivante pour tout réel x:
$$f(x) \leq \frac{1}{\sigma\sqrt{2\pi}\Phi(\frac{\mu-b}{\sigma})}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$
2. En déduire l'algorithme du rejet. Que peut-on dire du nombre d'essais moyen avant acceptation ?

3.3.2 Une distribution instrumentale alternative

On suppose que $b>\mu$. On considère la loi exponentielle translatée de b, $\tau\mathcal{E}(\lambda,b)$, de densité

$$g_\lambda(x) = \lambda e^{-\lambda(x-b)}\mathbf{1}_{x\geq b}, \quad x\in\mathbb{R}.$$

3. Montrer pour $x\geq b$ que
$$\frac{f_1(x)}{g_\lambda(x)} \leq \begin{cases} \frac{1}{\lambda}\exp\left\{\lambda(\mu-b)+\frac{(\lambda\sigma)^2}{2}\right\} & \text{si } \mu+\lambda\sigma^2>b, \\ \frac{1}{\lambda}\exp\left\{-\frac{(b-\mu)^2}{2\sigma^2}\right\} & \text{sinon.} \end{cases}$$
4. Proposer une méthode de simulation de la loi de densité f.
5. Calculer la valeur de λ^* telle que le temps moyen de calcul de la méthode proposée soit le plus petit possible.
6. En `R`, mettre en oeuvre les deux méthodes afin de constater empiriquement les différences.

4 Méthode de transformation

4.1 Simulation de lois Gaussiennes. Algorithme de Box-Muller

Soient X et Y deux variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$

1. Montrer que si U et V sont deux variables aléatoires indépendantes de loi $\mathcal{U}[0,1]$ alors le couple
$$\left(\sqrt{-2\ln(U)}\cos(2\pi V), \sqrt{-2\ln(U)}\sin(2\pi V)\right)$$
a la même loi que le couple (X,Y).
2. Ecrire une fonction `box_muller` permettant de simuler une loi $\mathcal{N}(0,1)$ en `R`. Vous comparez l'histogramme obtenu à la vrai densité de la loi.
3. En déduire, pour tout $\mu \in \mathbb{R}^{\not 2}$ et toute matrice 2×2 symmétrique semi-définie positive Σ une méthode pour simuler une variable aléatoire $Z \sim \mathcal{N}(\mu,\Sigma)$.

5 Autour de l'accepation rejet

5.1 Acceptation rejet étendu: Cas de deux fonctions positives.

Pour cette preuve, vous pourrez mimer les étapes de la preuve dans le cas usuel, détaillée dans le poly.

On se propose de montrer que pour que simuler selon une densité par algorithme d'acceptation rejet, il n'est nécessaire de connaître la densité qu'*à la constante de normalisation près.* Cette propriété est très utile dans le cas où le calcul de la constante de normalisation est coûteux, voir impossible (typiquement en statistiques Bayesiennes).

Plus formellement, soient $\tilde{f}$ une fonction positive et g une densité de probabilité, toutes deux définies sur $\mathbb{R}^d$ telles que:

- $0 < \int_{\mathbb{R}^d} \tilde{f}(x)\mathrm{d}x < \infty$. On note respectivement $I(\tilde{f})$ cette intégrale et
$$f(x) = \frac{\tilde{f}(x)}{I(\tilde{f})}$$
la densité associée à cette fonction positive.
- Il existe $M > 0$ tel que, pour tout réel x, $\tilde{f}(x) \leq Mg(x)$.

On note
$$\alpha(x) := \frac{\tilde{f}(x)}{Mg(x)}.$$
Soit $(U_m)_{m\geq 1}$ une suite de variables aléatoires i.i.d. de loi uniforme sur $[0,1]$. Soit $(Y_m)_{m\geq 1}$ une suite de variables aléatoires indépendantes et identiquement distributée, de densité donnée par g. On note T la variable aléatoire (à valeurs dans $\mathbb{N}^*$):
$$T = \inf\{m, \text{ tel que } U_m \leq \alpha(Y_m)\}.$$
.

1. Montrer la variable aléatoire $X := Y_T$ (T-ième valeur de la suite $(Y_m)_{m\geq 1}$) a pour densité f.
2. Donnez alors la loi de la variable aléatoire T. Quelle est l'espérance de T?
3. En déduire un estimateur de $I(\tilde{f})$ par méthode de Monte Carlo, obtenu uniquement à partir de l'algorithme d'acceptation rejet défini plus tôt.
4. Grâce au théorème central limite, donnez l'expression d'un intervalle de confiance asymptotique à 95% pour $I(\tilde{f})$, ne dépendant d'aucune quantité inconnue.

5.2 Recyclage dans l'acceptation rejet

Dans cette section, on se replace dans le cadre classique de l'acceptation rejet.

On se propose d'approcher une intégrale du type: $J = \mathbb{E}_f[\varphi(X)]$ où $f($ est la densité de la variable aléatoire X sur $\mathbb{R}^d$ selon laquelle on ne sait pas simuler, et φ est une fonction intégrable par rapport à cette densité.

À partir d'une densité $g(x)$ sur $\mathbb{R}^d$ selon laquelle on sait simuler, et telle que

$$\exists M > 0, \text{ tel que } \forall x \in \mathbb{R}^d, \ f(x) \leq Mg(x)$$

on obtient, par algorithme d'acceptation-rejet (pour un tel M fixé) un échantillon de variables aléatoires *i.i.d.* $X_1, \ldots, X_n$ de loi donnée par f.

Pour obtenir cet échantillon de taille n, on a simulé $N \geq n$ variables aléatoires indépendantes $Y_1, \ldots, Y_N$ de densité g. On note $Z_1, \ldots, Z_{N-n}$ l'échantillon i.i.d. de variables aléatoires ayant été rejetées dans l'algorithme d'acceptation rejet.

5. Donner l'expression de la densité de la variable aléatoire Z_1.

6. En déduire que
$$\hat{J}_N = \frac{1}{N}\left(\sum_{i=1}^{n} \varphi(X_i) + \sum_{j=1}^{N-n} \frac{(M-1)f(Z_j)}{Mg(Z_j) - f(Z_j)}\varphi(Z_j)\right)$$
est un estimateur sans biais de J. Quelle est l'intérêt de cette méthode selon vous?

5.3 Application

On reprend l'exemple vu en cours d'introduction à la statistique bayésienne. Vous reprendrez le même modèle ainsi que les mêmes données utilisées.

7. En utilisant le même prior que celui du cours, ainsi que le même loi de proposition g, implémentez l'algorithme d'acceptation rejet pour tirer selon le posterior. Les algorithmes efficaces seront valorisés.

8. Implémentez cette méthode, et tracez les densités empiriques des posteriors obtenus. Vous donnerez également une estimation de $\mathbb{E}[\theta|y_{1:n}]$ ainsi que l'intervalle de confiance associé.

9. À partir de cette méthode, et en utilisant les questions 3 et 4, donner une estimation ainsi qu'un intervalle de confiance pour à 95% de la quantité
$$\int_{\mathbb{R}^4} \pi(\theta)L(y_{1:n}\theta)\mathrm{d}\theta$$

10. Afin d'estimer de $\mathbb{E}[\theta|y_{1:n}]$, implémenter l'estimateur $\hat{J}_N$ de la question 6 (avec le même algorithme d'acceptation rejet que pour la question 8). Donnez un intervalle de confiance asymptotique pour l'estimation obtenue.

11. Comparez les deux estimateurs et commentez.