Randolf Altmeyer

BZQ II: Stochastikpraktikum
Wintersemester 2016
Humboldt-Universität zu Berlin

Projektaufgaben Block 1

1 Infinite-Monkey-Theorem (5P)

Laut dem Infinite-Monkey-Theorem (siehe `https://de.wikipedia.org/wiki/Infinite-Monkey-Theorem`) wird ein Affe, der unendlich lange zufällig auf einer Schreibmaschine tippt, fast sicher jede beliebige Zeichenkette schreiben.

1. Formuliere die Aussage des Theorems mathematisch präzise und beweise sie mit Hilfe des Lemmas von Borel-Cantelli.

2. Schreibe eine R-Funktion, die solange zufällig Zeichen $a, b, \ldots, y, z$ auswählt bis eine vorgegebene Zeichenkette komplett erschienen ist. Werte die Funktion mehrmals aus und diskutiere die durchschnittliche Anzahl von Zeichen bis die Zeichenkette vollständig erscheint.

Hinweise: Verwende die Befehle `letters` und **`sample`**. In R sind Strings nicht wie in anderen Programmiersprachen als Vektoren von einzelnen Zeichen implementiert. Allerdings kann man mit **`strsplit`** aus einem String einen solchen Vektor machen:

```
> strsplit("test","")[[1]]
[1] "t" "e" "s" "t"
```

2 Monte-Carlo-Approximationen (5P)

1. Finde eine Näherung für π über eine Monte-Carlo-Approximation von $\mathbb{E}\left[\mathbf{1}_S\left(X\right)\right]$, wobei $S = \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1\}$ und $X \sim Unif[-1, 1]^2$.

2. Approximiere die Wahrscheinlichkeit $\mathbb{P}(X > 20)$ für $X \sim N(0, 1)$ über eine Monte-Carlo-Approximation. Was ist das Problem? Schreibe die Wahrscheinlichkeit $\mathbb{P}(X > 20)$ über eine Variablensubstitution als Erwartungswert $\mathbb{E}[g(U)]$ für $U \sim Unif(0, 1/20)$ und eine integrierbare Funktion g. Approximiere $\mathbb{E}[g(U)]$ durch Monte-Carlo-Approximation und vergleiche mit der ersten Methode.

3. Berechne $\int_0^1 h_i(x)dx$ für $h_1(x) = (\cos(50x) + \sin(20x))^2$ und für eine beliebige andere „unglatte“ Funktion h_2. Verwende zuerst die `integrate`-Methode von R und dann eine Monte-Carlo-Approximation. Vergleiche die Ergebnisse.

4. Zeige (durch eine Monte-Carlo-Simulation), dass der Erwartungswert der Fläche eines zufälligen Dreiecks, wobei die drei Eckpunkte zufällig gleichverteilt auf dem Einheitsquadrat sind, etwa 11/144 ist. Benutze hierfür, dass der Flächeninhalt eines Dreiecks mit Eckpunkten (x_1, y_1), (x_2, y_2), (x_3, y_3) gegeben ist durch $\frac{1}{2}|\text{Det}A|$, wobei

$$A = \begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{pmatrix}.$$

Hinweise: Zu jeder Monte-Carlo-Approximation gehört eine Fehlerauswertung. Implementiere und plotte dafür die jeweilige Monte-Carlo-Methode in Abhängigkeit von der Samplegröße n zusammen mit Konfidenzbändern in einem Plot. Verwende den Befehl **cumsum**. Für die Konfidenzbänder schätze entweder die Varianz für die jeweilige Samplegröße durch eine weitere Monte-Carlo-Simulation der Länge m oder verwende den **quantile**-Befehl, z.B. mit Hilfe des folgenden Codes:

```
> mEst <- apply(mData, 2, cumsum)/(1:n)
> mUpperLowerBounds <- apply(mEst, 1, quantile, c(.025, .975))
> plot(mUpperLowerBounds[1,])
> lines(mUpperLowerBounds[2,])
```

mData ist dabei eine $n \times m$-dimensionale Matrix.

3 Eine (naive) Datenanalyse (5P)

Betrachte den Datensatz „faithful“, der mit R ausgeliefert wird. Lies zuerst die enstprechende Hilfe (mit `?faithful`). Analysiere die Ausbruchsdauern und Wartezeiten auf den nächsten Ausbruch in den Teilpopulationen mit kurzen Ausbruchsdauern (≤ 3) und langen Ausbruchsdauern (> 3). Vergleiche die jeweiligen Verteilungen mit Hilfe von Histogrammen, Boxplots und QQ-Plots. Gibt es einen Zusammenhang zwischen Ausbruchsdauern und Wartezeiten (erkläre mit Hilfe eines Scatterplots)? Wie lange muss man im Schnitt auf den nächsten Ausbruch warten (verwende z.B. **summary**) und wie stark streut diese Schätzung?

Hinweise: Recherchiere die Begriffe Histogramm, Boxplot, Scatterplots und QQ-Plots (z.B. auf Wikipedia). Erkläre wozu man sie verwendet und welchen Informationsgehalt sie für die Datenanalyse haben.

4 Normalverteilung (5P)

Vergleiche die folgenden Methoden zur Erzeugung von Zufallszahlen $Z \sim N(0, 1)$. Verwende dazu Histogramme, QQ-Norm-Plots und Kolmogorov-Smirnov-Tests. Diskutiere auch das Verhalten für Werte außerhalb von $[-2, 2]$.

1. Erzeuge Zufallszahlen $U_1, U_2 \overset{iid}{\sim} U([0, 1])$ und setze $Z = Z_1 = \sqrt{-2 \log U_1} \cos(2\pi U_2)$, $Z_2 = \sqrt{-2 \log U_1} \sin(2\pi U_2)$ (Box-Muller-Methode).

2. Erzeuge Zufallszahlen $U_1, \ldots, U_m \overset{iid}{\sim} U([-1/2, 1/2])$ für ein $m \in \mathbb{N}$ und setze $Z = \sum_{i=1}^m U_i$. Welche Wahl von m führt zu guten Ergebnissen?

3. Verwende den Accept-Reject-Algorithmus mit der Dichte der Laplace-Verteilung

$$g_\alpha(x) = \frac{\alpha}{2} \exp(-\alpha |x|), \quad \alpha > 0,$$

als Kandidat. Warum sollte man $\alpha = 1$ wählen? Wieviele Zufallsvariablen bezüglich g_1 müssen im Schnitt erzeugt werden, um eine Zufallsvariable bezüglich $N(0,1)$ zu erhalten?

Hinweis: Wende die Inversionsmethode an, um Zufallszahlen bezüglich g_α zu erzeugen.

Hinweise zur Abgabe:

Alle Dateien (pdf der Auswertung, R-Code für jede einzelne Aufgabe in eigener Datei) „gepackt“ (z.B. mit zip) mit dem Namen *UE1_Student1_Student2* bis zum 15.11. per Email an *altmeyrx@math.hu-berlin.de* schicken.