Projektpraktikum II (Stochastik)
Wintersemester term 2018/19
Humboldt-Universität zu Berlin

Sebastian Holtz

Projektaufgaben Block 1

1. 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 (kurz) wozu man sie verwendet und welchen Informationsgehalt sie für die Datenanalyse haben.

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

   (a) 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).

   (b) 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?

   (c) Verwende den Accept-Reject-Algorithmus mit der Dichte der Laplace-Verteilung zum Parameter $\alpha > 0$,

   $$g_\alpha(x) = \frac{\alpha}{2}\exp(-\alpha|x|), \quad x \in \mathbb{R},$$

   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 die Zufallszahlen bezüglich g_α zu erzeugen. Plotte für $n = 100$ auf diese Art erzeugte Zufallszahlen $Y_1, \ldots, Y_n \sim g_\alpha$ die empirische sowie die theoretische Verteilungsfunktion und zeichne in denselben Graphen Monte-Carlo-Bänder für $m = 100$ Iterationen.

3. 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.

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

   (b) 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"
```

4. (a) Finde eine Näherung für π über eine Monte-Carlo-Approximation von $\mathbb{E}\left[\mathbb{1}_S(X)\right]$, wobei $S = \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1\}$ und $X \sim U[-1,1]^2$. Berechne mit dieser Methode π approximativ für $n = 10^k,\ k = 1, \ldots, 6$.

   (b) Berechne das Integral $\int_0^\infty x^2 \exp(-x)dx$ analytisch und dann numerisch mittels der Funktion `integrate`.

   (c) Berechne das Integral in 4b näherungsweise mit der Monte-Carlo-Integration als empirisches zweites Moment. Zeige, dass $\int_0^\infty x^2 \exp(-x)dx = 2\int_0^\infty x \exp(-x)dx$ gilt und nutze dies, um das Integral alternativ durch $2n^{-1}\sum_{i=1}^n X_i$ zu berechnen. Vergleiche beide Ansätze für $n = 1000,\ 10\,000,\ 100\,000$ generierte Zufallszahlen.

   (d) Schreibe eine Funktion, die in $N = 50$ Iterationen die beiden Methoden aus 4c für ein Sample vom Umfang $n = 10\,000$ durchführt und als Rückgabe die empirische Varianz für beide Methoden ausgibt. Berechne die (theoretische) Varianz der Methoden exakt und vergleiche mit den empirischen Werten.

Lösungen in einer zip-Datei per Mail bis zum 14. November, 23:59 Uhr, abgeben.