Aufgabenblatt
Abgabe: bis Montag, 07.12.2009; 10:00 Uhr

Diese Zettel sollen Sie mit dem Statistikprogramm R vertraut machen. Zunächst werden die grundlegende Bedienung und wichtige Befehle erklärt. Im Anschluss daran finden sich Aufgaben.
Durch das regelmäßige Bearbeiten dieser Aufgaben kann ein Teilnahmeschein erworben werden.
Alle Aufgaben sind Programmieraufgaben. Bitte kopieren Sie die eingegebenen Befehle und die Ausgaben von R in eine Textdatei (Kommentare können hinter der Raute `#` eingegeben werden), und senden Sie diese zur Abgabe per e-Mail an
`mentemeier@uni-muenster.de`
Eine schriftliche Abgabe ist nicht erforderlich.

Das Statistikprogramm R ist frei verfügbar, und kann von `http://www.r-project.org/` heruntergeladen werden. Zum Starten in der Uni siehe gesonderte Anleitung.

Theorie

Monte-Carlo-Methoden

Das starke Gesetz der großen Zahlen besagt, dass für P-fast jede Realisierung $(x_n)_{n\geq 1}$ einer Folge $(X_n)_{n\geq 1}$ von unabhängig, identisch verteilte Zufallsgrößen und jede Funktion h mit $\mathrm{E}\,|h(X_1)| < \infty$ gilt

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(x_i) = \mathrm{E}\, h(X_1).$$

Diese Tatsache kann beispielsweise zur näherungsweisen Berechnung komplizierter Integrale benutzt werden: Im einfachsten Fall Integrale über kompakte Intervalle, z.B. ist

$$\int_0^1 \sin\left(\frac{1}{x}\right) dx = \mathrm{E} \sin\left(\frac{1}{X}\right),$$

wenn X eine auf $[0,1]$ gleichverteilte Zufallsgröße ist. Zur Monte-Carlo-Integration erzeugt man nun (in R mittels `runif(n,a,b)`) genügend viele gleichverteilte Zufallszahlen, und berechnet mittels

$$\frac{1}{n} \sum_{i=1}^{n} \sin\left(\frac{1}{x_i}\right)$$

einen Näherungswert für das Integral. Im allgemeinen Fall sei X eine Zufallsgröße mit (auf dem Integrationsbereich positiver) Lebesgue-Dichte $f(x)$, dann ist

$$\int_I g(x)\ dx = \int_I \left(\frac{g(x)}{f(x)}\right) f(x)\ dx = \mathrm{E}\left(\frac{g(X)}{f(X)}\right).$$

Hier muss die Existenz von $\mathrm{E}\left|\frac{g(X)}{f(X)}\right|$ vorausgesetzt werden (meist zeigt man die Beschränktheit des Quotienten), und man berechnet als Näherungswert für das Integral dementsprechend

$$\frac{1}{n}\sum_{i=1}^{n}\frac{g(x_i)}{f(x_i)}.$$

Auch Flächeninhalte lassen sich mittels Monte-Carlo-Methoden bestimmen, dies geschieht durch mehrdimensionale Monte-Carlo-Integration über die entsprechende Indikatorfunktion.

Empirische Verteilungsfunktion

Eine wichtige Fragestellung in der Statistik ist, welche Wahrscheinlichkeitsverteilung einen gegebenen Datensatz am besten modelliert. Dies kann aufgrund theoretischer Überlegungen (physikalische Gesetze, Gedächtnislosigkeit, Unabhängigkeit) geschehen; aber auch die grafische Darstellung des Datensatzes mag Indizien geben.
Ein Hilfsmittel ist die sog. empirische Verteilungsfunktion, für einen (reellwertigen) Datensatz $(x_1, \ldots, x_n)$ definiert durch

$$\mathrm{ecdf}(t) = \frac{1}{n}\sum_{i=1} n1_{(-\infty,t]}(x_i).$$

Nach dem Satz von Glivenko-Cantelli nähert sich die empirische Verteilungsfunktion einer zugrundeliegenden Verteilungsfunktion mit wachsendem n an. In `R` liefert `ecdf(x)` die emp. Vtlg.fkt. eines Datensatzes `x`.

Brownsche Bewegung

Eine Standard-Brownsche Bewegung ist ein stochastischer Prozess $(X_t)_{t\geq 0}$ in stetiger Zeit mit P-fast sicher stetigen Pfaden, „stochastisch unabhängigen Zuwächsen“, und der Eigenschaft, dass $X_t \sim N(0,t)$ für jeden Zeitpunkt t.

Verteilungen in R, Erzeugen von Zufallszahlen

Zu fast allen gängigen Verteilungen gibt existieren in R vier Funktionen, unterschieden durch die Präfixe `d,p,q` und `r`, gefolgt vom Namen der Verteilung, z.B. `norm` für die Normalverteilung. Die Eingabe von `pnorm(x,`μ`,`σ`)` liefert den Wert der Verteilungsfunktion von $N(\mu,\sigma^2)$ an der Stelle x. Das Präfix `d` steht für die Lebesgue- bzw. Zähldichte der jeweiligen Verteilung, `qnorm(p,`μ`,`σ`)` liefert das `p`-Quantil einer $N(\mu,\sigma^2)$-Verteilung. Mittels `rnorm(n,`μ`,`σ`)` wird ein Vektor mit `n` Realisierungen einer $N(\mu,\sigma^2)$-verteilten Zufallsgröße erzeugt. Eine Übersicht über implementierte Verteilungen findet sich auf der Praktikumshomepage unter „Wichtige R-Befehle“.

Plots

Die Funktion `plot()` kann mit verschiedenen Daten arbeiten, sind x und y bspw. zwei gleichlange Vektoren, so liefert `plot(x,y)` ein Streudiagramm dieser beiden Vektoren, wohingegen `plot(ecdf(x))` die emp. Vtlg.-funktion ausgibt. Wie bei Histogrammen können mittels der Parameter `xlab, ylab` die Achsenbeschriftungen verändert werden. Der Befehl `par(mfrow=c(3,2))` sorgt z.B. dafür, dass in einem Grafikfenster 3 Zeilen und 2 Spalten für Plots zur Verfügung stehen, bevor ein neues Fenster geöffnet wird. Der *Par*ameter `las` kann die Werte 0 bis 3 annehmen, und steuert die Ausrichtung der Achsenbeschriftung.
Grundsätzlich erzeugt R jeden Plot neu, will man nachträglich noch Grafikelemente einfügen, so helfen Befehle wie `lines()` oder `points()`, die als Argument 2 Vektoren mit x- bzw. y-Koordinaten (Funktionswerten) erwarten.

Aufgaben

Aufgabe 1. (5 Punkte)
Zunächst eine kurze Wiederholung: Erzeugen Sie einen Vektor mit 1000 Realisierungen einer Standard-Normalverteilung, und lassen Sie ein Histogramm plotten. Experimentieren Sie mit den Unterteilungspunkten, so dass das Histogramm der Gaußschen Glockenkurve ähnlich wird. Gibt es im `hist`-Befehl eine Option, um relative Häufigkeiten statt absoluter Häufigkeiten plotten zu lassen?

Nun zu Monte-Carlo-Methoden.

(a) Berechnen Sie mittels Monte-Carlo-Integration (n=1 000 000) einen Näherungswert für $\int_0^1 \sin\left(\frac{1}{x}\right) dx$.

(b) Berechnen Sie mittels Monte-Carlo-Integration (n=1 000 000) einen Näherungswert für

$$\int_0^1 \sin(x) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Beachten Sie die Integralgrenzen!

(c) Der Flächeninhalt eines Viertel-Einheitskreises ist $\frac{\pi}{4}$. Nutzen Sie mehrdimensionale Monte-Carlo-Integration (n=1 000 000), um einen Näherungswert für $\frac{\pi}{4}$ zu bestimmen. **Tipp**: Wikipedia-Eintrag zu „Monte-Carlo-Integration“.

Aufgabe 2. (3 Punkte)
Erzeugen Sie vier Bilder in einer Graphik, die jeweils (unterschiedliche) Approximationen von Pfaden einer zweidimensionalen (Standard-)Brownschen Bewegung darstellen. Lassen Sie dazu viermal die Werte einer zweidimensionalen Brownschen Bewegung in den Punkten k/n mit $n = 10.000$ und $k = 1, \ldots, n$ von R randomisieren und interpolieren Sie dazwischen linear. Beschriften Sie die Achsen mit $B_t^{(1)}$ bzw. $B_t^{(2)}$, wobei Sie wie üblich die Beschriftung der Ordinate orthogonal zur Ordinate vornehmen. Überschreiben Sie die Bilder mit Pfad 1 bis Pfad 4.

Aufgabe 3. (8 Punkte)
Auf der Praktikums-Homepage finden Sie die Datei `gamma.txt`. Sie enthält 50 Zufallszahlen, die gemäß einer Gamma-Verteilung mit unbekanntem Parametern erstellt wurden.

(a) Lesen Sie die Datei ein und plotten Sie die Empirische Verteilungsfunktion.

(b) Bestimmen Sie den Maximum-Likelihood-Schätzer für den Datensatz. Benutzen Sie dabei die Funktion `optim`, um das Maximum der Log-Likelihood-Funktion näherungsweise zu bestimmen. Plotten Sie anschließend die Verteilungsfunktionsfunktion der Gamma-Verteilung mit diesen Parametern in das bestehende Bild.

(c) Bestimmen Sie den Momenten-Methode-Schätzer für den Datensatz und plotten Sie die Verteilungsfunktionsfunktion der Gamma-Verteilung mit diesen Parametern in das bestehende Bild.

(d) Nehmen Sie nun an, die Daten kämen von einer Weibull-Verteilung mit unbekannten Parametern. Bestimmen Sie nun den Maximum-Likelihood-Schätzer unter dieser Annahme und plotten Sie die Verteilungsfunktionsfunktion der Weibull-Verteilung mit diesen Parametern in das bestehende Bild. Fügen Sie anschließend eine Legende in das Bild ein.

Aufgabe 4. (4 Punkte)
Erzeugen Sie Zufallsvektoren x und y mit jeweils 1000 $N(1,1)$ bzw. $Exp(1)$-verteilten Einträgen.

(a) Zeichnen Sie die empirischen, und tatsächlichen Verteilungsfunktionen jeweils in die selbe Grafik.

(b) Plotten Sie die Dichtefunktionen beider Verteilungen (**Tipp:** Nutzen Sie die Funktion `curve()`).

(c) Erstellen Sie mit dem Befehl `box.plot()` Box-Plots der Vektoren x bzw. y, vergleichen Sie diese und denken Sie an die oben gezeichneten Dichtefunktionen. Was fällt Ihnen auf?