Prof. Dr. Evgeny Spodarev
Dipl.-Math. Stefan Roth

SS 2015
28. April 2015

Stochastik I - Übungsblatt 3

Abgabe: Dienstag, 5. Mai vor Beginn der Übung.

Hinweise zu den R-Aufgaben:

- Den Namen beider abgebenden Studenten auf jedes Blatt der Ausgabe drucken!
- Immer Quelltext und Ausgabe zusammen abgeben (nicht auf getrennten Blättern). Bei Aufgaben mit Grafikausgabe Quelltext und Plots abgeben.

Aufgabe 1 (5 + 1 + 6 Punkte)

Die Verteilung der 2 Datensätze in den Spalten von `qq-analyse.data` soll bestimmt werden. Dazu verwenden wir QQ-Plots und beschränken uns auf die Exponentialverteilung, Normalverteilung und Gleichverteilung.

(a) Es seien $X, X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit absolutstetiger Verteilung. Zeige, dass

$$P(X \leq X_{(k)}) = \frac{k}{n+1},$$

$1 \leq k \leq n$.

(b) Motiviere mit (a) die im Vorlesungsskript vorgeschlagene praktische Variante der QQ-Plots.

(c) Erstelle für die 2 Datensätze QQ-Plots mit den 3 Verteilungen und entscheide jeweils, welche am besten passt und zu welcher Verteilungsfamilie die Daten folglich vermutlich gehören. Hierfür kannst Du Quantilfunktionen von **R** (z.B. `qnorm()`) verwenden, erstelle den Plot aber selbst und beschrifte ihn mit der jeweiligen Verteilung. Zeichne auch mit `abline(lm(...))` eine angepasste Gerade in das Schaubild.

Aufgabe 2 (4 + 6 Punkte)

Das Ziel dieser Aufgabe ist es, einen Kern-Dichte-Schätzer zu implementieren und auf den Datensatz `iris.data`, der auf der Homepage bereitsteht, anzuwenden.

(a) Schreibe eine Funktion in **R**, die einen Vektor `x` mit den Auswertungspunkten, einen Vektor `daten` mit Datenpunkten, die Bandbreite `h` und eine Kernfunktion `K` als Eingangsdaten hat und den Wert des Kerndichteschätzers an den Stellen in `x` zurückgibt.

(b) Plotte das Ergebnis der Funktion aus (a) angewandt auf den Epanechnikov-Kern, die Kelchblattbreite (Spalte `KelBr` in `iris.data`) sowie den Bandbreiten 0.1, 0.3, 0.5 und 1 in ein gemeinsames Schaubild. Verwende dafür unterschiedliche Linientypen (Parameter `lty`) für die Bandbreiten und beschrifte die Grafik entsprechend. Welche Bandbreite scheint am besten zu passen? Begründe Deine Entscheidung.

Aufgabe 3 (4 + 5 + 3 Punkte)

Bei einer physikalischen Reaktion sei bekannt, dass diese nach Beginn des Experiments zufällig nach T Sekunden eintritt, wobei $T \sim \mathsf{U}(t, t+1)$ für ein $t > 0$. Dabei hängt t von dem Versuchsaufbau ab. Um t zu bestimmen, werden mehrere (identische, aber unabhängige) Versuche durchgeführt, bis die gewünschte Genauigkeit erreicht ist. In dieser Aufgabe soll die Verteilung der Anzahl der Versuche (Bezeichnung N) untersucht werden, die nötig sind, um t bis auf $0,1$s genau zu bestimmen.

(a) Zeige, dass die Verteilung von N nicht von t abhängt. Wie viele Versuche benötigt man im Mittel bis die gewünschte Genauigkeit erreicht wird?

(b) Simuliere so lange unabhängige auf (t_0, t_0+1) gleichverteilte Zufallsvariablen für ein $t_0 > 0$, bis die gewünschte Genauigkeit erreicht wird. Führe das Experiment 100 Mal durch und speichere jeweils die benötigte Anzahl an Versuchen $N_1, \ldots, N_{100}$.

(c) Plotte ein Histogramm der in (b) erzeugten Daten $N_1, \ldots, N_{100}$ gemeinsam mit der Zähldichte von N.

Aufgabe 4 (4 Punkte)

Zeige die folgende Darstellung des Bravais-Pearson-Korrelationskoeffizienten ρ_{xy} im Falle binärer Daten $(x_1, y_1), \ldots, (x_n, y_n)$, d.h. falls $x_i, y_i \in \{0, 1\}$:

$$\rho_{xy} = \frac{h_{00}h_{11} - h_{01}h_{10}}{\sqrt{h_{0\cdot} \cdot h_{1\cdot} \cdot h_{\cdot 0} \cdot h_{\cdot 1}}},$$

wobei

$$\begin{aligned}
h_{00} &= \#\{(x_i, y_i);\ x_i = y_i = 0\} \\
h_{11} &= \#\{(x_i, y_i);\ x_i = y_i = 1\} \\
h_{01} &= \#\{(x_i, y_i);\ x_i = 0,\ y_i = 1\} \\
h_{10} &= \#\{(x_i, y_i);\ x_i = 1,\ y_i = 0\} \\
h_{0\cdot} &= h_{00} + h_{01} \\
h_{\cdot 0} &= h_{00} + h_{10} \\
h_{1\cdot} &= h_{10} + h_{11} \\
h_{\cdot 1} &= h_{01} + h_{11}.
\end{aligned}$$

Aufgabe 5 (6 Punkte)

In dieser Aufgabe soll der Einfluss (der Varianz) des Störterms und der Steigung der Geraden auf das Bestimmtheitsmaß bei der linearen Regression in einer Simulationsstudie untersucht werden. Dafür gehen wir davon aus, dass als Messpunkte (x-Werte) die Punkte $0.1,\ 0.2,\ \ldots,\ 3.9,\ 4$ zur Verfügung stehen. Die zu rekonstruierenden Funktionen haben die Form $y = ax + b$, wobei stets $b = 5$ gilt und a den Wert $0.1,\ 1$ oder 2 hat. Ferner soll an den Messstellen ein additiver Fehler dazukommen in Form von unabhängig und identisch $\mathsf{N}(0, \sigma^2)$-verteilten Zufallsvariablen, wobei $\sigma^2 \in \{0.1, 1, 10\}$. Führe die Studie durch, indem du für die 3 Varianzen je einen Satz Zufallsvariablen simulierst und die gestörten y-Werte für die verschiedenen Werte von a errechnest. Berechne dann jeweils für die 9 sich ergebenden Regressionsgeraden das Bestimmtheitsmaß und lege die Werte in einen `data.frame` ab. Gib diesen als Ergebnis aus und interpretiere die errechneten Werte.