Übungen zur Vorlesung

Datenanalyse - Dr. Terveer, Vogt, Pohl

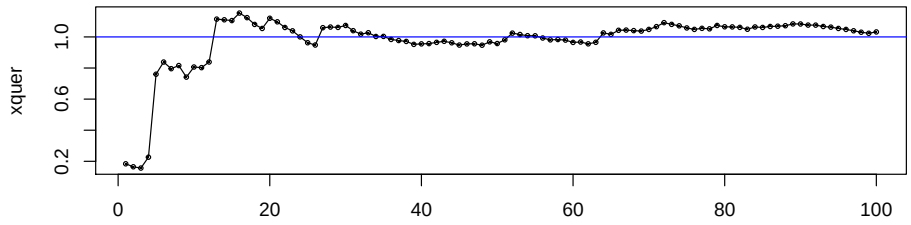
Sommersemester 2022 Blatt 1 19.04.2022

Aufgabe 1 (Klausur Datenanalyse und Simulation, Sommersemester 2015) $X, X_1, X_2, \ldots, X_n$ seien positive stochastisch unabhängige Zufallsvariablen mit einer stetigen Verteilung und der folgenden Wahrscheinlichkeitsdichte $f_X(x) = \lambda^2 x e^{-\lambda x} \mathbb{1}_{]0;\infty[}(x)$. Dabei sei $\lambda > 0$ unbekannt.

Bestimmen Sie eine Wahrscheinlichkeitsdichte von $Y = \frac{1}{X}$.

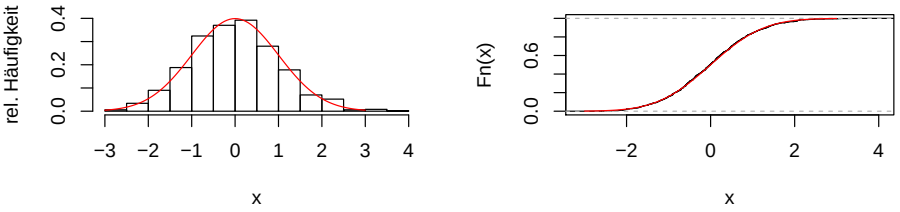
Aufgabe 2 Erzeugen Sie mit R die unten stehenden Grafiken und interpretieren Sie sie. Initialisieren Sie dazu jeweils den Zufallszahlengenerator von R mit `set.seed(0)` und informieren Sie sich über den Befehl `stats::rexp`

a) Dargestellt sind fortlaufend gebildete Mittelwerte $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$, $n = 1, \ldots, 100$ von Realisierungen zu 100 u.i.v. $Exp(1)$-verteilten Zufallsvariablen $X_1, \ldots, X_{100}$:



Hinweis: Suchen Sie nach einer R-Funktion, welche die kumulierte Summe berechnet.

b) Dargestellt sind links das Histogramm und rechts die empirische Verteilungsfunktion einer u.i.v.-Stichprobe $Y_1, \ldots, Y_{1000}$ mit $Y_i = \sqrt{100}(\bar{X}_i - 1)$, wobei $\bar{X}_i$ Mittelwert von u.i.v. Zufallsvariablen $X_{i1}, \ldots, X_{i100}$ zur $Exp(1)$-Verteilung ist:



Rot eingezeichnet sind auch Dichte (links) und Verteilungsfunktion (rechts) der Standardnormalverteilung.

Aufgabe 3 (R – Wiederholung) Nachdem Sie im letzten Semester bereits einen ersten Einblick in die Programmiersprache R erhalten haben, werden Sie in diesem Semester Ihre Kenntnisse vertiefen. Um Ihr Wissen wieder etwas aufzufrischen, steht im Learnweb wie gewohnt ein `covid_19_daily_ reports_04-12-2022`-Datensatz der Johns-Hopkins Univsersität[1] zu den aktuellen Corona-Fallzahlen bereit.

a) Lesen Sie den Datensatz in R ein und schauen Sie sich ihn an.

b) Filtern Sie den Datensatz nach Beobachtungen aus Deutschland. Clustern Sie diesen Datensatz nach dem Ward-Verfahren, indem Sie die Distanzmatrix berechnen und das Ergebnis in einem Dendrogramm darstellen. Welche Anzahl an Clustern würden Sie empfehlen?

[1] https://github.com/CSSEGISandData/COVID-19

c) Schreiben Sie eine Funktion (oder nutzen Sie Ihre Funktion aus dem letzten Semester) um in dem ungefilterten Datensatz die Fallzahlen pro Land zu aggregieren.

d) Sortieren Sie den in c) aggregierten Datensatz aufsteigend nach bestätigten Fallzahlen und plotten Sie diese. Tragen Sie den Median sowie den Mittelwert als farbige Linien ein und fügen Sie außerdem eine passende Legende hinzu. Nutzen Sie Ihr Wissen aus DuW (Außenseiter, Schiefe etc.) um den Plot zu interpretieren.

e) Erzeugen Sie ein Streudiagramm der Todes- sowie bestätigten Fallzahlen. Was war noch mal ein Korrelationsmaß? Wählen Sie ein geeignetes Maß und berechnen Sie es für die Todes- und bestätigten Fallzahlen.