Tutorial 2

Verteilungsschätzung: empirische Verteilungsfunktionen, Histogramme, Histogramme mit dynamischen Klassengrenzen, Kerndichteschätzer

Auf diesem Übungsblatt beschäftigen wir uns mit Daten des Old Faithful Geysirs in Kalifornien. Daten zur Wartezeit bis zur nächsten Eruption sowie zur Dauer der Eruption finden Sie in dem Datensatz `faithful`, welcher direkt in R verfügbar ist.

Die Wartezeit ist in der Spalte `faithful$waiting` enthalten und die Eruptionsdauer in `faithful$eruptions`. Wir beschäftigen uns hier mit der Eruptionsdauer (in Minuten):

```
eruptionen <- faithful$eruptions
```

Aufgabe 1: empirische Verteilungsfunktionen und Histogramme

a) Erstellen Sie mit `plot(ecdf(eruptionen))` die empirische Verteilungsfunktion der insgesamt 272 Beobachtungen. Identifizieren Sie (grob) anhand des Plots das 25%-, 50%- und 75%-Quantil. Überprüfen Sie anschließend Ihre Werte mit dem `summary(...)` Befehl.

b) Betrachten Sie nun ein Histogramm der Daten:

```
hist(eruptionen, breaks = seq(1.5, 5.5, by = 0.5), prob = TRUE)
```

Über das Argument breaks wird hierbei festgelegt, dass das Histogramm im Bereich zwischen 1.5 und 5.5 gezeichnet wird mit einer Klassenbreite von 0.5. Die Option `prob = TRUE` bewirkt, dass die Gesamtfläche des Histogramms gleich 1 ist.

Alternativ kann mit dem Argument length die Anzahl an Klassen festgesetzt werden. Für bspw. 8 Klassen bräuchten wir 9 Intervallgrenzen:

```
hist(eruptionen, breaks = seq(1.5, 5.5, length = 8 + 1), prob = TRUE)
```

Wie lässt sich die erhaltene Form der Verteilung erklären? Kennen Sie eine Verteilung, mit der man diese Daten sinnvoll modellieren könnte?

c) Erstellen Sie Histogramme mit verschiedenen Klassenzahlen (z.B., 2, 10, 50, 100) und erinnern Sie sich daran, inwiefern dies den Bias-Varianz-Trade-off illustriert. Wie viele Klassen würden Sie für die Eruptionsdauer wählen?

Aufgabe 2: Histogramme mit dynamischen Klassengrenzen

Diese Aufgabe befasst sich mit der Modifikation des Histogramms, bei dem die Klassengrenzen dynamisch sind (siehe Slide 22 der 2. Vorlesung). Bestimmen Sie für die Eruptionsdauer aus Aufgabe 1 den Wert des entsprechenden Schätzers $\widehat{f}(x)$ an der Stelle $x = 3.75$ für $b = 0.5$.

Zusatzaufgabe (falls Sie gerne programmieren): Schreiben Sie eine `for`-Schleife, welche die Berechnung für

$$x = 1.5, 1.51, 1.52, ..., 5.48, 5.49, 5.5$$

durchführt. Plotten Sie anschließend die so erhaltenen Werte gegen x, um die gesamte geschätzte Funktion zu erhalten. Sie könnten anschließend auch noch den Wert von b variieren.

Aufgabe 3: Kerndichteschätzer

Mit den folgenden Befehlen erstellen Sie ein Histogramm der 272 Beobachtungen und fügen anschließend dem gleichen Plot einen Kerndichteschätzer hinzu:

```
hist(eruptionen, breaks = 15, prob = TRUE, main = "")
lines(density(eruptionen))
```

a) Welche beiden wesentlichen Vorteile hat der Kerndichteschätzer gegenüber dem Histogramm?
b) Nutzen Sie die Hilfefunktion (`?density`), um herauszufinden, welche Kernfunktion und welche Bandweite von `density(...)` per default zur Kerndichteschätzung genutzt werden.