

Bernd Klaus (bernd.klaus@imise.uni-leipzig.de)
Verena Zuber (verena.zuber@imise.uni-leipzig.de)

<http://uni-leipzig.de/~zuber/teaching/ws09/r-kurs/>

Den folgenden Aufgaben liegt der Datensatz OECD zu Grunde, er enthält Variablen (Stand 2009), die das Wohlergehen von Kindern in den Mitgliedsstaaten messen sollen. Abgefragt wurde:

- Einkommen: das durchschnittliche Einkommen der Eltern [in tausend US Dollar pro Kind]
- Armut: der Anteil [immer in Prozent] an Kindern in einem armen Elternhaus
- Bildung: der Anteil an Kindern, die ohne eine Grundausstattung (Bücher, Schreibtisch, Computer, Internet) für Bildung auskommen
- WenigRaum: der Anteil an Kindern, die auf zu wenig Raum wohnen
- Umwelt: der Anteil an Kindern, die unter schlechten Umweltbedingungen leben
- Lesen: mittlerer PISA-Score zur Lesefähigkeit
- Geburtsgewicht: der Anteil an Kindern, die bei der Geburt weniger als 2.5kg wiegen
- Säuglingssterblichkeit: Säuglingssterblichkeit (<1 Jahr) [x in Tausend]
- Sterblichkeit: Sterblichkeit (<20 Jahre) [x in 100 000]
- Selbstmord: Selbstmord von Jugendlichen im Alter von 15 bis 19 [x in 100 000]
- Bewegung: der Anteil an 11, 13 und 15 jährigen Jugendlichen, die sich regelmäßig bewegen
- Rauchen: der Anteil an 15 jährigen Jugendlichen, die mindestens einmal die Woche rauchen
- Alkohol: der Anteil an 13-15 jährigen Jugendlichen, die mindestens zweimal betrunken waren
- Bullying: der Anteil an Kindern, die angeben in der Schule bedroht zu werden
- Schule: der Anteil an Kindern, die angeben die Schule zu mögen
- Geo: Faktor, der die geographische Lage eines Landes beschreibt, dabei steht "E" für Europa und "R" für den Rest der Welt.

1 Aufgabe: Datensatz OECD

- (a) Lesen Sie den Datensatz *oecdM* mit der Funktion `data<-read.csv(file="Daten/oecdM", header=TRUE)` ein und überprüfen Sie die Dimension der Daten.
- (b) Berechnen Sie die Mittelwerte und Varianzen der einzelnen Variablen mit dem geeigneten `apply` Befehl.
- (c) Überprüfen Sie, ob die Niederlande im der Länderliste des des Datensatzes auftaucht. Gibt es auch einen Eintrag für China? (Benutzen sie die R-Hilfe, um herauszufinden wie man auf die Ländernamen zugreifen kann.)

- (d) In welchem Land waren die meisten Jugendlichen mindestens zweimal betrunken? Wie hoch ist der maximale Prozentsatz?
- (e) In welchem Land ist die Säuglingssterblichkeit am geringsten? Wie hoch ist sie in diesem Land?
- (f) In welchen Ländern ist der Prozentsatz an Jugendlichen, die sich regelmäßig bewegen, kleiner als der Durchschnitt?
- (g) In welchen Ländern fühlen sich besonders viele Kinder in der Schule bedroht? Als Indikator für "besonders viel" soll ein Bullying Wert gelten, der mindestens eine Standardabweichung (standard deviation) vom Mittelwert aller Ländern entfernt ist.
- (h) Erstellen Sie einen neuen Datensatz, der aufsteigend nach dem Einkommen geordnet ist. Speichern Sie diesen in einer neuen *.csv* Datei.
- (i) Wieviele Länder im Datensatz *oecdM* gehören zu Europa, wieviele zum Rest der Welt? Stellen Sie das Ergebnis in einem Kuchendiagramm dar und verwenden sie dazu die Farben grün ("green") und blau ("blue").
- (j) Visualisieren sie die Variable *Lesen*, getrennt nach dem Faktor Geo, in einem vertikalen Stripchart.

2 Aufgabe: Quantile und Plots

- (a) Erstellen Sie einen Boxplot für die Variable "Bildung". Was fällt Ihnen auf?
- (b) Untermauern Sie die Beobachtung aus Aufgabe (a) durch Berechnung einiger Quantile mit Hilfe der Funktion `quantile()`.
- (c) Stellen Sie zudem die aufsteigend geordneten Werte der Variable *Bildung* mit Hilfe der Funktion `plot()` als Kurve dar.
- (d) Begründen Sie anhand Ihrer Beobachtungen, dass das 75% Quantil der Daten einen guten Trennpunkt zwischen Ländern mit "guter" und "schlechter" Grundausrüstung für Bildung darstellt.
- (e) Bestimmen Sie eine Liste mit Ländern, in denen es an der Grundausrüstung für Bildung besonders fehlt.

3 Aufgabe: Mittlere absolute Abweichung

Die **mittlere absolute Abweichung**, kurz auch als **MAD** bezeichnet, ergibt sich für Daten $X = \{x_1, x_2, \dots, x_n\}$ als

$$\text{MAD} = \text{median}(|x_i - x_{0.5}|)$$

Sie ist ein robustes Maß für die Standardabweichung.

- (a) Ermitteln Sie für alle Variablen des Datensatzes die MAD und die Standardabweichung.
- (b) Bestimmen Sie jene zwei Variablen, die die größten Unterschiede zwischen MAD und Standardabweichung aufweisen.
- (c) Versuchen Sie mit Hilfe von graphischen Methoden eine Erklärung zu finden.