

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

<http://uni-leipzig.de/~zuber/teaching/ws12/r-kurs/>

1 Aufgabe: Apply-Funktion verwenden

- Erstellen Sie eine Matrix A mit 4 Zeilen und 25 Spalten, die die Werte 1 bis 100 enthält. Analog dazu erstellen Sie eine Matrix B mit 25 Zeilen und 4 Spalten, die die Werte 1 bis 100 enthält.
- Berechnen Sie mittels dem `apply()`-Befehl den Mittelwert und die Varianz für jede Zeile von A bzw. B.
- Berechnen Sie mittels dem `apply()`-Befehl den Mittelwert und die Varianz für jede Spalte von A bzw. B.
- Standardisieren ist eine häufige Transformation von Daten; dafür wird der Mittelwert von der entsprechenden Zeile oder Spalte abgezogen und durch die entsprechende Standardabweichung geteilt. Somit besitzen die Daten einen Mittelwert von 0 und eine Standardabweichung von 1.

Standardisieren Sie die Spalten der Matrix A. Abschließend überprüfen Sie, ob die Spalten richtig standardisiert wurden.

2 Aufgabe: Datensatz OECD

Der Datensatz OECD enthält Variablen (Stand 2009), die das Wohlergehen von Kindern in den Mitgliedsstaaten messen sollen. Abgefragt wurde:

- Einkommen: das durchschnittliche Einkommen der Eltern in US Dollar
- Armut: der Anteil [in Prozent] an Kindern in einem armen Elternhaus
- Bildung: der Anteil [in Prozent] an Kindern, die ohne eine Grundausrüstung (Bücher, Schreibtisch, Computer, Internet) für Bildung auskommen
- WenigRaum: der Anteil [in Prozent] an Kindern, die auf zu wenig Raum wohnen
- Umwelt: der Anteil [in Prozent] an Kindern, die unter schlechten Umweltbedingungen leben
- Lesen: mittlerer PISA-Score zur Lesefähigkeit
- Geburtsgewicht: der Anteil [in Prozent] an Kindern, die bei der Geburt weniger als 2.5kg wiegen
- Säuglsterblichkeit: 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 [in Prozent] an 11, 13 und 15-jährigen Jugendlichen, die sich regelmäßig bewegen
 - Rauchen: der Anteil [in Prozent] an 15 jährigen Jugendlichen, die mindestens einmal die Woche rauchen
 - Alkohol: der Anteil [in Prozent] an 13-15 jährigen Jugendlichen, die mindestens zweimal betrunken waren
 - Bullying: der Anteil [in Prozent] an Kindern, die angeben in der Schule bedroht zu werden
 - Schule: der Anteil [in Prozent] an Kindern, die angeben die Schule zu mögen
-
- (a) Lesen Sie den Datensatz *oecd* mit der Funktion `data<-read.csv(file="Daten/oecd", header=TRUE)` oder `data<-read.csv(file="http://www.uni-leipzig.de/~zuber/teaching/ws12/r-kurs/praxis/oecd", header=TRUE)` ein und überprüfen Sie die Dimension der Daten.
- (b) Berechnen Sie die Mittelwerte und Varianzen der einzelnen Variablen mit einem 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 werden 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.