

Das Statistische Bundesamt veröffentlicht jedes Jahr die Gesamtstudierendenanzahl in Deutschland (<https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/BildungForschungKultur/BildungForschungKultur.html>). Die Daten können auch für die Jahre von 1975–2011 eingesehen werden, aufgeteilt nach Geschlecht und Herkunft. Weil die Datenaufbereitung im Fokus steht, sind die Daten bereits als Textdatei auf der Veranstaltungs-Homepage hinterlegt.

Aufgabe 1:

Laden Sie den Datensatz `studierende.txt` von der Veranstaltungs-Homepage (http://www.statistik.lmu.de/institut/ag/statsoz_neu/lehre/2013_SoSe/StatSoft/material/data/studierende.txt) herunter und lesen Sie ihn in R ein und bringen Sie alle Variablen in ein sinnvolles Format.

Tipp: Konvertieren Sie Anzahlen nicht gleich in Faktoren.

Tipp: Die 1000er Trennzeichen könnten Probleme machen.

Aufgabe 2:

Erstellen Sie 2 neue Variablen des Datensatzes (`alle_I` und `alle_D`), welche die Gesamtzahl der Studierenden je Jahr und Herkunft enthält.

Tipp: `?transform`

Aufgabe 3:

Erstellen Sie eine weitere Variable, die den Ausländeranteil an den Gesamt-Studierenden enthält.

Lassen Sie sich eine statistische Kurzzusammenfassung dafür ausgeben und bestimmen Sie die Jahre, in denen dieser am höchsten bzw niedrigsten war.

Tipp: Die Funktionen `which.min` und `which.max` werden Ihnen weiterhelfen.

Aufgabe 4:

Visualisieren Sie

1. die empirische Dichtefunktion des Ausländeranteils und
2. den zeitlichen Verlauf des Ausländeranteils im Verlauf der Jahre.

Tipp: Die Funktion `plot(x,y)` zeichnet einen Scatterplot für 2 gleich lange Vektoren `x` und `y`, wobei sich ein Punkt aus (x_i, y_i) ergibt.

Aufgabe 5:

Jetzt wollen wir wissen, wie hoch die jeweilige absolute Änderung der Studierendenzahl von Jahr zu Jahr ist. Erstellen Sie dazu 3 neue Variablen (nicht als Teils des Datensatzes `studierende`).

1. `diff_alle` für die Änderung der Gesamten Studierenden,
2. `diff_w` für die Änderung der weiblichen Studierenden und
3. `diff_m` für die Änderung der männlichen Studierenden.

Tipp: Sie können es auf mehrere Arten lösen!

Aufgabe 6:

Fassen Sie die in Aufgabe 5 gebildeten Variablen zu einem neuen Datensatz `aenderung` zusammen und fügen Sie diesem noch eine Variable `jahr` hinzu, die für jede Änderung das Jahr angibt mit dem das Vorjahr verglichen wird, also die Werte 1976 bis 2011 enthält.

Aufgabe 7:

Erstellen Sie je Geschlecht eine Variable im Datensatz `aenderung`, welche die Änderung in Schritten von 20.000 diskretisiert. Dabei sollen beide Variablen am Schluss die selben Levels haben!

Aufgabe 8:

Erstellen Sie die Häufigkeitsverteilung für die diskretisierten Änderungen und visualisieren Sie diese.

Tipp: Wenn Sie `barplot` eine Matrix übergeben, wobei jede Zeile eine Häufigkeitstabelle enthält, dann werden beide Häufigkeiten in der Grafik verwendet.

Tipp: Mit dem Parameter `beside = TRUE` zu `barplot` werden die beiden Häufigkeiten nebeneinander dargestellt.