Programmieren mit Statistischer Software **9. Block**
Grün **SS 2017**

9 Graphik

Aufgabe 1:

Laden Sie den Datensatz `CO2`.

- Visualisieren Sie die Daten geeignet.
- Bestimmen Sie den Mittelwert und die Standardabweichung der Variable `uptake` für jeden Teildatensatz, wenn Sie die Daten bzgl. der Faktoren `Type` und `Treatment` aufsplitten.

Aufgabe 2:

Laden Sie den Datensatz `Titanic`. Dieser enthält die Anzahl der Passagiere auf der Titanic aufgeteilt nach Klasse, Alter, Geschlecht und Überlebensstatus.

- Bestimmen Sie die Randtabelle nur für die Variablen Geschlecht und Überlebensstatus. Berechnen Sie daraus die relativen Überlebenshäufigkeiten von Frauen und Männern und visualisieren sie diese Randtabelle geeignet.
- Bestimmen Sie weiters die relativen Überlebenshäufigkeiten von Frauen und Männern separat für die Klassen und visualisieren Sie diese.

Aufgabe 3:

Laden Sie den Census Income Datensatz, der im UCI Machine Learning Repository unter `http://archive.ics.uci.edu/ml/datasets/Census+Income` zur Verfügung steht, herunter und lesen Sie ihn nach R ein.

- Achten Sie darauf `NA`s richtig einzulesen. Bei welchen Variablen kommen `NA`s vor und wieviele sind es jeweils?
- Recodieren Sie die Variable über den Familienstand (`marital-status`) so, dass alle Kategorien, die mit `Married` beginnen zu einer Kategorie zusammengefasst werden. Erstellen Sie einen parallelen Boxplot von Alter bzgl. des Familienstands.

Hinweis: Auf der Hilfseite von `level` wird gezeigt, wie man Kategorien einer nominellen Variable zusammenfassen kann.

Aufgabe 4:

Laden Sie den Datensatz `road` aus dem Paket **MASS**

- Wie viele Beobachtungen und Variablen hat der Datensatz?
- Versuchen Sie graphisch zu analysieren, ob das Risiko eines tödlichen Straßenunfalls von der Populationsdichte, der Länge der Landstraßen, der Temperatur und vom durchschnittlichen Verbrauch pro Fahrer abhängt.

- Gibt es auffällige Beobachtungen? Entfernen Sie diese gegebenenfalls und erzeugen Sie die analogen Graphiken, wo diese nicht mehr im Datensatz enthalten sind.

Aufgabe 5:

Laden Sie den Datensatz Student Performance vom UCI Machine Learning Repository unter

`http://archive.ics.uci.edu/ml/datasets/Student+Performance`

herunter und lesen Sie den Teil über die Leistungen in Mathematik in R ein.

- Wie viele Beobachtungen und Variablen hat der Datensatz?
- Bestimmen Sie erst geeignete Maßzahlen für den Zusammenhang für die Variablen `G1`, `G2` und `G3`.
- Wandeln Sie die Variable `G3` in eine Variable `G3ord` um, indem Sie die Beobachtungen geeignet in gute, mittlere und schlechte Schüler einteilen.
- Bestimmen Sie nun geeignete Maßzahlen für den Zusammenhang zwischen `G3ord` und den Attributen 1–30. Wo ist der Zusammenhang am stärksten?

Hinweis: Bzgl. geeigneter Maßzahlen siehe C. Duller, Einführung in die Statistik mit Excel und SPSS.