Stochastik

Serie 1

Aufgabe 1.1

Der Datensatz der OECD enthält Messgrössen, die das Wohlergehen von Kindern in den Mitgliedsstaaten ermitteln sollen. Im Jahr 2009 wurde abgefragt:

- Einkommen (Average disposable income): das durchschnittliche Einkommen der Eltern [in tausend US Dollar pro Kind].
- Armut (Children in poor homes): der Anteil [immer in Prozent] an Kindern in einem armen Elternhaus.
- Bildung (Educational Deprivation): der Anteil an Kindern, die ohne Grundausstattung (Bücher, Schreibtisch, Computer, Internet) für Bildung auskommen müssen.
- Wenig Raum (Overcrowding): der Anteil an Kindern, die auf zu wenig Raum wohnen.
- Umwelt (Poor environmental conditions): der Anteil an Kindern, die unter schlechten Umweltbedingungen leben.
- Lesen (Average mean literacy score): mittlerer PISA-Score zur Lesefähigkeit.
- Geburtsgewicht (Low birth weight): der Anteil an Kindern, die bei der Geburt weniger als 2.5 kg wiegen.
- Säuglingssterblichkeit (Infant mortality): Säuglingssterblichkeit (< 1 Jahr) [x in Tausend].
- Sterblichkeit (Mortality rates): Sterblichkeit (< 20 Jahre) [x in 100 000].
- Selbstmord (Suicide rates): Selbstmord von Jugendlichen im Alter von 15 bis 19 [x in 100 000].
- Bewegung (Physical activity): der Anteil an 11, 13 und 15 jährigen Jugendlichen, die sich regelmässig bewegen.
- Rauchen (Smoking): der Anteil an 15 jährigen Jugendlichen, die mindestens einmal die Woche rauchen.
- Alkohol (Drunkenness): der Anteil an 13-15 jährigen Jugendlichen, die mindestens zweimal betrunken waren.

- Bullying (Bullying): der Anteil an Kindern, die angeben, in der Schule bedroht zu werden.
- Schule (Liking school): der Anteil an Kindern, die angeben die Schule zu mögen.

a) Lesen Sie den Datensatz `child.txt` mit der Funktion

```
data <- read.table(file = "./child.txt", header = TRUE,
    sep = ",")
```

ein und überprüfen Sie die Dimension der Daten mit der Funktion

```
dim(...)
```

b) Bestimmen Sie den Mittelwert und Median der einzelnen Variablen mit der R-Funktion

```
summary(...)
```

c) Überprüfen Sie, ob die Niederlande in der Länderliste des Datensatzes auftaucht. Gibt es auch einen Eintrag für China? Die Spaltennamen ermitteln Sie mit der R-Funktion

```
colnames(...)
```

Finden Sie heraus, wie Sie die Zeilennamen ermitteln können.

d) Welche Werte für Bullying haben die Länder in der fünften bis zehnten Zeile? Um welche Länder handelt es sich?

e) In welchen fünf Ländern waren die meisten Jugendlichen mindestens zweimal betrunken? Wie hoch ist der maximale Prozentsatz? Benützen Sie den R-Befehl

```
order(..., na.last = ...)
```

f) In welchem Land ist die Säuglingssterblichkeit am geringsten? Wie hoch ist sie in diesem Land? Benützen Sie den R-Befehl

```
which.min(...)
```

g) In welchen Ländern ist der Prozentsatz an Jugendlichen, die sich regelmässig bewegen, kleiner als der Durchschnitt? Benützen Sie die R-Befehle

```
mean(..., na.rm = ...)
which(...)
```

h) Erstellen Sie einen neuen Datensatz, der aufsteigend nach dem Einkommen geordnet ist. Speichern Sie diesen in einer neuen `.txt` Datei. Benützen Sie den R-Befehl

```
write.table(..., col.names = ..., row.names = ..., file = ...)
```

Aufgabe 1.2

Das Dataframe `d.fuel` enthält die Daten verschiedener Fahrzeuge aus einer amerikanischen Untersuchung der 80er-Jahre. Jede Zeile (row) enthält die Daten eines Fahrzeuges (ein Fahrzeug entspricht einer Beobachtung).

a) Lesen Sie die auf Ilias abgelegte Datei `d.fuel.dat` ein mit dem folgenden R-Befehl[1]:

```
d.fuel <- read.table(file = "./Daten/d.fuel.dat",
    header = T, sep = ",")
```

Das Argument `sep=","` braucht es, weil die Kolonnen im File `d.fuel.dat` durch Kommata getrennt sind. Im File `d.fuel.dat` wurden die Zeilen durchnummeriert und daher steht in der ersten Spalte die Nummer der Zeile. Die Spalten (columns) enthalten die folgenden Variablen:

weight:	Gewicht in Pounds (1 Pound = 0.453 59 kg)
mpg:	Reichweite in Miles Per Gallon (1 gallon = 3.789 l; 1 mile = 1.6093 km)
type:	Autotyp

b) Wählen Sie nur die fünfte Zeile des Dataframe `d.fuel` aus. Welche Werte stehen in der fünften Zeile?

c) Wählen Sie nun die erste bis fünfte Beobachtung des Datensatzes aus. So lässt sich übrigens bei einem unbekannten Datensatz ein schneller Überblick über die Art des Dataframe gewinnen.

d) Zeigen Sie gleichzeitig die 1. bis 3. und die 57. bis 60. Beobachtung des Datensatzes an.

e) Berechnen Sie den Mittelwert der Reichweiten aller Autos in Miles/Gallon.

R-Hinweis:

```
mean(...)
```

f) Berechnen Sie den Mittelwert der Reichweite der Autos 7 bis 22.

g) Erzeugen Sie einen neuen Vektor `t.kml`, der alle Reichweiten in km/l, und einen Vektor `t.kg`, der alle Gewichte in kg enthält.

[1] Alternativ können Sie in `RStudio` die Datei auch mit dem Menü `ImportDataset` einlesen.

h) Berechnen Sie den Mittelwert der Reichweiten in km/l und denjenigen der Fahrzeuggewichte in kg.

Aufgabe 1.3

Bei der Ermittlung der landwirtschaftlichen Nutzfläche von Bauernhöfen in einem Bezirk ergaben sich folgende Werte (in ha):

```
2.1,2.4,2.8,3.1,4.2,4.9,5.1,6.0,6.4,7.3,10.8,12.5,13.0,13.7,14.8,17.6,19.6,
23.0,25.0,35.2,39.6
```

a) Berechnen Sie die Summen $\sum x_i$ und $\sum x_i^2$.

R-Hinweis: Benützen Sie die Funktion

```
sum(...)
```

b) Berechnen Sie den Mittelwert und die Standardabweichung (ohne die in R implementierten Funktionen, sondern aufgrund der Definition der Grössen).

R-Hinweis: Benützen Sie die Funktion

```
length(...)
```

(ohne die in R implementierten Funktionen, sondern aufgrund der Definition der Grössen).

c) Bestimmen Sie den Median (ohne die in R implementierte Funktion, sondern aufgrund der Definition der Grössen).

R-Hinweis: Benützen Sie die Funktionen

```
sort(...)
round(...)
```

d) Bestimmen Sie das 25 % und das 75 % Quantil (ohne die in R implementierte Funktion, sondern aufgrund der Definition der Grössen).

e) Bestimmen Sie nun Mittelwert, Standardabweichung, Median und das 75 % Quantil mit den R-Funktionen

```
mean(...)
sd(...)
median(...)
quantile(...)
```

f) Überprüfen Sie aufgrund des Datenvektors mit den landwirtschaftlichen Nutzflächen, dass das arithmetische Mittel der **standardisierten** Variablen

$$z_i = \frac{x_i - \bar{x}}{s_x} \quad \text{mit} \quad i = 1, \ldots, n$$

gleich null und die empirische Standardabweichung von z_i gleich 1 ist.

Kurzlösungen vereinzelter Aufgaben

A 1.3:

a) 269.1 , 5729.27

b) 12.81, 114 , 10.68

c) 10.8

d) 4.9, 17.6