

# Stochastik

## Musterlösungen zu Serie 1

### Lösung 1.1

- a) Die Datei `child.txt` lesen wir mit folgendem Befehl ein

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

Die Dimension ermitteln wir dann mit

```
dim(data)  
  
## [1] 30 21
```

Der Datensatz enthält also 30 Zeilen und 21 Spalten.

- b) Eine Zusammenfassung lässt sich mit R folgendermassen erhalten:

```
summary(data)  
  
## Average.disposable.income Children.in.poor.homes  
## Min. : 3.839 Min. : 2.740  
## 1st Qu.:16.618 1st Qu.: 8.901  
## Median :21.107 Median :11.659  
## Mean :18.848 Mean :12.372  
## 3rd Qu.:22.643 3rd Qu.:16.092  
## Max. :34.242 Max. :24.590  
##  
## Educational.Deprivation Overcrowding  
## Min. : 0.400 Min. :10.33  
## 1st Qu.: 1.000 1st Qu.:17.06  
## Median : 1.500 Median :21.57  
## Mean : 2.673 Mean :31.95  
## 3rd Qu.: 2.200 3rd Qu.:44.39  
## Max. :13.700 Max. :73.96  
## NA's :4  
## Poor.environmental.conditions  
## Min. :10.50  
## 1st Qu.:20.15  
## Median :25.49  
## Mean :25.22  
## 3rd Qu.:30.24  
## Max. :38.71
```

```

## NA's :6
## Average.mean.literacy.score Literacy.inequality
## Min. :408.7 Min. :1.475
## 1st Qu.:482.8 1st Qu.:1.623
## Median :501.3 Median :1.683
## Mean :496.3 Mean :1.665
## 3rd Qu.:512.8 3rd Qu.:1.719
## Max. :552.7 Max. :1.756
##
## Youth.NEET.rate Low.birth.weight Infant.mortality
## Min. : 1.700 Min. : 3.900 Min. : 2.300
## 1st Qu.: 4.550 1st Qu.: 5.150 1st Qu.: 3.525
## Median : 6.200 Median : 6.750 Median : 4.200
## Mean : 7.378 Mean : 6.643 Mean : 5.447
## 3rd Qu.: 8.400 3rd Qu.: 7.500 3rd Qu.: 5.250
## Max. :37.700 Max. :11.300 Max. :23.600
## NA's :3
## Breastfeeding.rates Vaccination.rates..pertussis.
## Min. :41.00 Min. :78.00
## 1st Qu.:79.00 1st Qu.:91.00
## Median :91.00 Median :95.80
## Mean :86.03 Mean :93.78
## 3rd Qu.:96.00 3rd Qu.:97.80
## Max. :99.00 Max. :99.80
## NA's :1 NA's :1
## Vaccination.rates.measles. Physical.activity
## Min. :74.00 Min. :13.10
## 1st Qu.:88.00 1st Qu.:15.80
## Median :94.00 Median :19.30
## Mean :91.52 Mean :20.13
## 3rd Qu.:96.30 3rd Qu.:21.80
## Max. :99.80 Max. :42.10
## NA's :1 NA's :4
## Mortality.rates Suicide.rates Smoking
## Min. :14.84 Min. : 1.263 Min. : 8.10
## 1st Qu.:21.17 1st Qu.: 5.037 1st Qu.:14.62
## Median :23.15 Median : 6.785 Median :16.60
## Mean :24.60 Mean : 6.856 Mean :16.51
## 3rd Qu.:25.75 3rd Qu.: 8.864 3rd Qu.:19.50
## Max. :50.23 Max. :15.950 Max. :27.10
## NA's :1 NA's :1 NA's :6
## Drunkenness Teenage.births Bullying
## Min. :10.00 Min. : 3.70 Min. : 4.200
## 1st Qu.:11.35 1st Qu.: 7.05 1st Qu.: 7.975

```

```
## Median :14.55   Median :10.60   Median : 9.650
## Mean   :15.22   Mean     :15.50   Mean    :10.979
## 3rd Qu.:17.93   3rd Qu.:17.80   3rd Qu.:13.825
## Max.    :24.80   Max.     :65.80   Max.    :25.300
## NA's    :6                      NA's     :6
## Liking.school
## Min.     :11.70
## 1st Qu.:21.40
## Median  :25.60
## Mean    :27.17
## 3rd Qu.:34.90
## Max.    :57.40
## NA's    :5
```

c) Wir wollen die Zeilenamen unseres Datensatzes ermitteln

```
rownames(data)

## [1] "Australia"      "Austria"
## [3] "Belgium"        "Canada"
## [5] "Czech Republic" "Denmark"
## [7] "Finland"        "France"
## [9] "Germany"        "Greece"
## [11] "Hungary"        "Iceland"
## [13] "Ireland"        "Italy\t"
## [15] "Japan\t"        "Korea\t"
## [17] "Luxembourg"     "Mexico"
## [19] "Netherlands"   "New Zealand"
## [21] "Norway"        "Poland"
## [23] "Portugal"      "Slovak Republic"
## [25] "Spain\t"       "Sweden"
## [27] "Switzerland"   "Turkey"
## [29] "United Kingdom" "United States\t"
```

Die Niederlande ist also im Datensatz enthalten, China hingegen nicht.

d) Die Werte für Bullying der Länder in der fünften bis zehnten Zeile lauten:

```
data[5:10, "Bullying"]

## [1] 5.5 8.0 8.0 13.6 13.9 22.0

rownames(data)[5:10]

## [1] "Czech Republic" "Denmark"
## [3] "Finland"        "France"
```

```
## [5] "Germany" "Greece"
```

e) Die Variablen heissen

```
colnames(data)

## [1] "Average.disposable.income"
## [2] "Children.in.poor.homes"
## [3] "Educational.Deprivation"
## [4] "Overcrowding"
## [5] "Poor.environmental.conditions"
## [6] "Average.mean.literacy.score"
## [7] "Literacy.inequality"
## [8] "Youth.NEET.rate"
## [9] "Low.birth.weight"
## [10] "Infant.mortality"
## [11] "Breastfeeding.rates"
## [12] "Vaccination.rates..pertussis."
## [13] "Vaccination.rates.measles."
## [14] "Physical.activity"
## [15] "Mortality.rates"
## [16] "Suicide.rates"
## [17] "Smoking"
## [18] "Drunkenness"
## [19] "Teenage.births"
## [20] "Bullying"
## [21] "Liking.school"
```

Um zu ermitteln, in welchen 5 Ländern die meisten Jugendlichen mindestens zweimal betrunken sind, benutzen wir

```
order(...)
```

um die Zeilen in aufsteigender Reihenfolge zu ermitteln

```
order(data[, "Drunkenness"], na.last = F)[26:30]

## [1] 4 22 29 7 6
```

Dabei stellt das Argument `na.last=F` die Zeilen mit den nicht vorhandenen Datenpunkten am Anfang des geordneten Datenvektors. Die Ländernamen lauten dann

```
rownames(data[order(data[, "Drunkenness"], na.last = F)[26:30],
])
```

```
## [1] "Canada"          "Poland"
## [3] "United Kingdom"  "Finland"
## [5] "Denmark"
```

In Dänemark sind die meisten Jugendlichen mindestens zweimal betrunken, nämlich

```
data["Denmark", "Drunkenness"]

## [1] 24.8
```

Prozent der dänischen Jugendlichen.

- f) Die Zeile, in der der Wert mit der kleinsten Säuglingssterblichkeit steht, lautet

```
which.min(data[, "Mortality.rates"])

## [1] 17
```

Das betreffende Land ist

```
rownames(data[which.min(data[, "Mortality.rates"]),
               ])

## [1] "Luxembourg"
```

- g) Der Mittelwert der Anzahl an Jugendlichen, die sich regelmässig bewegen, lautet

```
mean(data[, "Physical.activity"], na.rm = T)

## [1] 20.13462
```

Also in folgenden Ländern ist die Anzahl an Jugendlichen, die sich regelmässig bewegen, kleiner als im OECD Durchschnitt

```
mean.physical.activity <- mean(data[, "Physical.activity"],
                               na.rm = T)
which(data[, "Physical.activity"] < mean.physical.activity)

## [1] 2 3 8 9 10 11 14 17 18 21 22 23 26 27 28
## [16] 29

rownames(data[which(data[, "Physical.activity"] <
                    mean.physical.activity), ])

## [1] "Austria"          "Belgium"
## [3] "France"           "Germany"
```

```
## [5] "Greece"      "Hungary"
## [7] "Italy\t"     "Luxembourg"
## [9] "Mexico"      "Norway"
## [11] "Poland"      "Portugal"
## [13] "Sweden"      "Switzerland"
## [15] "Turkey"      "United Kingdom"
```

h) `order(data[, "Average.disposable.income"])`

```
## [1] 23 28 18 24 22 11 5 25 14 10 20 8 9 26 1
## [16] 3 16 7 2 12 13 15 29 6 27 19 4 21 30 17
```

```
write.table(data[order(data[, "Average.disposable.income"]),
], file = "income_ordered.txt", col.names = TRUE,
row.names = TRUE)
```

Wir können überprüfen, ob die Datei abgespeichert wurde

```
data <- read.table(file = "income_ordered.txt", header = TRUE)
rownames(data)
```

```
## [1] "Portugal"      "Turkey"
## [3] "Mexico"        "Slovak Republic"
## [5] "Poland"        "Hungary"
## [7] "Czech Republic" "Spain\t"
## [9] "Italy\t"       "Greece"
## [11] "New Zealand"   "France"
## [13] "Germany"       "Sweden"
## [15] "Australia"     "Belgium"
## [17] "Korea\t"       "Finland"
## [19] "Austria"       "Iceland"
## [21] "Ireland"       "Japan\t"
## [23] "United Kingdom" "Denmark"
## [25] "Switzerland"   "Netherlands"
## [27] "Canada"        "Norway"
## [29] "United States\t" "Luxembourg"
```

## Lösung 1.2

- Siehe Aufgabenstellung.
- Um die Daten in Tabellenform zu sehen, tippt man den Namen des Objektes ein

```
d.fuel
```

##		X	weight	mpg	type
##	1	1	2560	33	Small
##	2	2	2345	33	Small
##	3	3	1845	37	Small
##	4	4	2260	32	Small
##	5	5	2440	32	Small
##	6	6	2285	26	Small
##	7	7	2275	33	Small
##	8	8	2350	28	Small
##	9	9	2295	25	Small
##	10	10	1900	34	Small
##	11	11	2390	29	Small
##	12	12	2075	35	Small
##	13	13	2330	26	Small
##	14	14	3320	20	Sporty
##	15	15	2885	27	Sporty
##	16	16	3310	19	Sporty
##	17	17	2695	30	Sporty
##	18	18	2170	33	Sporty
##	19	19	2710	27	Sporty
##	20	20	2775	24	Sporty
##	21	21	2840	26	Sporty
##	22	22	2485	28	Sporty
##	23	23	2670	27	Compact
##	24	24	2640	23	Compact
##	25	25	2655	26	Compact
##	26	26	3065	25	Compact
##	27	27	2750	24	Compact
##	28	28	2920	26	Compact
##	29	29	2780	24	Compact
##	30	30	2745	25	Compact
##	31	31	3110	21	Compact
##	32	32	2920	21	Compact
##	33	33	2645	23	Compact
##	34	34	2575	24	Compact
##	35	35	2935	23	Compact
##	36	36	2920	27	Compact
##	37	37	2985	23	Compact
##	38	38	3265	20	Medium

```
## 39 39    2880  21  Medium
## 40 40    2975  22  Medium
## 41 41    3450  22  Medium
## 42 42    3145  22  Medium
## 43 43    3190  22  Medium
## 44 44    3610  23  Medium
## 45 45    2885  23  Medium
## 46 46    3480  21  Medium
## 47 47    3200  22  Medium
## 48 48    2765  21  Medium
## 49 49    3220  21  Medium
## 50 50    3480  23  Medium
## 51 51    3325  23   Large
## 52 52    3855  18   Large
## 53 53    3850  20   Large
## 54 54    3195  18    Van
## 55 55    3735  18    Van
## 56 56    3665  18    Van
## 57 57    3735  19    Van
## 58 58    3415  20    Van
## 59 59    3185  20    Van
## 60 60    3690  19    Van
```

c) Auswählen der fünften Beobachtung:

```
d.fuel[5, ]

##      X weight mpg  type
## 5  5    2440  32 Small
```

d) Auswählen der 1. bis 5. Beobachtung:

```
d.fuel[1:5, ]

##      X weight mpg  type
## 1  1    2560  33 Small
## 2  2    2345  33 Small
## 3  3    1845  37 Small
## 4  4    2260  32 Small
## 5  5    2440  32 Small
```

Alternativ kann man sich eine Übersicht verschaffen mit Hilfe der R-Funktion `head(...)`



```
head(d.fuel)

##      X weight mpg  type
## 1  1    2560  33 Small
## 2  2    2345  33 Small
## 3  3    1845  37 Small
## 4  4    2260  32 Small
## 5  5    2440  32 Small
## 6  6    2285  26 Small
```

e) Auswählen der 1. bis 3. und 57. bis 60. Beobachtung:

```
d.fuel[c(1:3, 57:60), ]

##      X weight mpg  type
## 1  1    2560  33 Small
## 2  2    2345  33 Small
## 3  3    1845  37 Small
## 57 57    3735  19  Van
## 58 58    3415  20  Van
## 59 59    3185  20  Van
## 60 60    3690  19  Van
```

f) Die Werte der Reichweiten stehen in der dritten Spalte, die mpg heisst. Zur Berechnung des Mittelwertes gibt es verschiedene Möglichkeiten, welche sich in der Art der Datenselektion unterscheiden:

```
mean(d.fuel[, 3])

## [1] 24.58333

mean(d.fuel[, "mpg"])

## [1] 24.58333

mean(d.fuel$mpg)

## [1] 24.58333
```

g) Auch hier gibt es wieder verschiedene Möglichkeiten. Eine davon ist:

```
mean(d.fuel[7:22, "mpg"])

## [1] 27.75
```

h) Umrechnung der Miles Per Gallon in Kilometer pro Liter und der Pounds in Kilogramm:

```
t.kml <- d.fuel[, "mpg"] * 1.6093/3.789
t.kg <- d.fuel[, "weight"] * 0.45359
```

## Lösung 1.3

```
x <- c(2.1, 2.4, 2.8, 3.1, 4.2, 4.9, 5.1, 6, 6.4, 7.3, 10.8,
      12.5, 13, 13.7, 14.8, 17.6, 19.6, 23, 25, 35.2, 39.6)
```

a)  $\sum x_i =$

```
sum(x)
```

```
## [1] 269.1
```

$\sum x_i^2 =$

```
sum(x^2)
```

```
## [1] 5729.27
```

b) Mittelwert:  $\frac{1}{n} \sum x_i =$

```
n <- length(x)
mean.x <- 1/n * sum(x)
mean.x
```

```
## [1] 12.81429
```

Standardabweichung:

```
var.x <- 1/(n - 1) * sum((x - mean.x)^2)
var.x
```

```
## [1] 114.0473
```

```
sqrt(var.x)
```

```
## [1] 10.67929
```

c) Median:

```
x.sorted <- sort(x)
0.5 * length(x)
```

```
## [1] 10.5
```

```
k <- round(0.5 * length(x) + 0.5)
k
```

```
## [1] 11
```

```
x.sorted[k]
```

```
## [1] 10.8
```

d) Das 25 % berechnet sich zu

```
0.25 * length(x)
```

```
## [1] 5.25
```

```
k <- round(0.25 * length(x) + 0.5)
```

```
k
```

```
## [1] 6
```

```
x.sorted[k]
```

```
## [1] 4.9
```

und das 75 % Quantil ergibt

```
0.75 * length(x)
```

```
## [1] 15.75
```

```
k <- round(0.75 * length(x) + 0.5)
```

```
k
```

```
## [1] 16
```

```
x.sorted[k]
```

```
## [1] 17.6
```

e)

```
mean(x)
```

```
## [1] 12.81429
```

```
sd(x)
```

```
## [1] 10.67929
```

```
median(x)
```

```
## [1] 10.8
```

```
quantile(x, 0.75)
```

```
## 75%
```

```
## 17.6
```

f) Wir stellen aufgrund unseres Beispieldatensatzes für den arithmetischen Mit-

telwert und die Standardabweichung fest, dass

```
z <- (x - mean(x)) / sd(x)
mean(z)

## [1] -2.614476e-17

sd(z)

## [1] 1
```

Im Allgemeinen ergibt sich für den arithmetischen Mittelwert von  $z_i$

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n z_i &= \frac{1}{n} \sum_{i=1}^n \frac{x_i - \bar{x}}{s_x} \\&= \frac{1}{n \cdot s_x} \sum_{i=1}^n (x_i - \bar{x}) \\&= \frac{1}{n \cdot s_x} \sum_{i=1}^n \left( \frac{n}{n} \cdot x_i - \bar{x} \right) \\&= \frac{1}{n \cdot s_x} \left( \frac{n}{n} \cdot \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} \right) \\&= \frac{1}{n \cdot s_x} (n\bar{x} - n\bar{x}) \\&= 0.\end{aligned}$$

Für die Varianz von  $z_i$  erhalten wir

$$\begin{aligned}\frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2 &= \frac{1}{n-1} \sum_{i=1}^n (z_i - 0)^2 \\&= \frac{1}{(n-1)} \sum_{i=1}^n \frac{x_i - \bar{x}}{s_x}^2 \\&= \frac{1}{s_x^2} \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2 \\&= \frac{1}{s_x^2} s_x^2 \\&= 1.\end{aligned}$$