Aufgabe 03

Gruppe 01

6 5 2020

Aufgabe 07 Deskriptive Statistik

```
#Workspace laden
load("data/yingtan_20_ueb3.RData")
Ca <- ljz$Ca_exch
```

a) Wenden Sie die Methode summary auf die austauschbaren Ca-Ionen an und erläutern Sie kurz die mit dieser Funktion gewonnenen Parameter.

```
#summary der austauschbaren Ca-Ionen
summary(Ca)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.772  13.633  19.491  21.926  28.718  94.311
```

Mit dem Befehl summary() werden die Quantile, der Median, der Mittelwert sowie Minimum und Maximum der Beobachtungswerte der Variable angegen. Das arithmetische Mittel der gemessenen austauschbaren Ca-Ionen ist hier 21,926 µmol/g. Der Median, bei dem die Anzahl der Werte halbiert wird, beträgt 19,491 µmol/g. Das bedeutet, dass mindestens die Hälfte kleiner oder gleich Median und mindestens die Hälfte der Daten größer oder gleich Median ist (Steland 2016). Da das arithmetische Mittel empfindlicher gegenüber Ausreißern ist (ebd.), und es größer als der Median ist, scheint es hier Ausreißer im oberen Wertebereich zu geben. Für die Quantile wird die Anzahl der Werte in vier gleich große Teile geteilt. Das erste Quantil endet bei 13,633 µmol/g, das zweite Quantil ist gleich Median, das dritte Quantil endet bei 38,718 µmol/g und das vierte Quantil ist gleich der größte Beobachtungswert, also 94,311 µmol/g.

b) Ermitteln Sie wiederum für die austauschbaren Ca-Ionen die Streuungsparameter Varianz und Standardabweichung sowie die zentralen Momente Schiefe und Kurtosis (bezogen auf NV mit kurtosis = 0).

```
var(Ca)
```

```
## [1] 123.8529
```

```
sd(Ca)
```

```
## [1] 11.12892
```

```
library(psych)
describe(Ca)
```

```
##    vars   n  mean    sd median trimmed   mad  min   max range skew kurtosis
## X1    1 335 21.93 11.13  19.49   20.98 10.22 3.77 94.31 90.54  1.5     5.55
##      se
## X1 0.61
```

```
library(moments)
skewness(Ca)
```

```
## [1] 1.510328
```

```
kurtosis(Ca)
```

```
## [1] 8.605555
```

Varianz und Standardabweichung sind Streuungsparameter (Steland 2016). Sie geben die Streuung der Werte zum arithmetischen Mittelwert (Walser 2011) oder auch die Variabilität der Werte (Hedderich and Sachs 2018) an. Die Varianz für die austauschbaren Ca-Ionen ist 123,85 µmol/g. Die Standardabweichung, also die Wurzel aus der Varianz ist 11,13 µmol/g. Das bedeutet.. Der Kurtosis-Wert ist 8,6 (5,55) (> 0), daher verläuft die Verteilung der austauschbaren Ca-Ionen im Bezug auf die Normalverteilung steilgipflig verläuft. Der Wert für die Schiefe ist 1,5 (> 0), was zeigt, dass die Verteilung zudem linkssteil und rechtsschief verläuft (Steland 2016).

Alternative Lösung mittels selbsterstellter Funktionen:

```
#Varianz
varf <- function(x) {
  mean((x - mean(x))^2)
}

varf(Ca)
```

```
## [1] 123.4832
```

```
#Standardabweichung
stabw <- function(x) {
  sqrt(mean((x-mean(x))^2))
}

stabw(Ca)
```

```
## [1] 11.1123
```

Die Differenz zwischen den Ergebnissen für die Varianz der Funktion von RStudio und der selbsterstellten Funktion ist 0,37. Das liegt daran, dass RStudio eine erwartete Fehlerkorrektur vornimmt und durch (n-1) und nicht nur durch n teilt, wobei n die Anzahl der Messwerte darstellt (Steland 2016). Folglich weicht auch die eigens berechnete Standardabweichung leicht von der automatisch berechneten ab.

```
#Schiefe
schiefe <- function(x) {
  1/(length(x) * (sqrt(var(x))^3)) * sum((x-mean(x))^3)
}

schiefe(Ca)
```

```
## [1] 1.50357
```

```
kurtosisf <- function(x) {
  1/(length(x) * (sqrt(var(x))^4)) * sum((x-mean(x))^4)-3
}

kurtosisf(Ca)
```

```
## [1] 5.554256
```

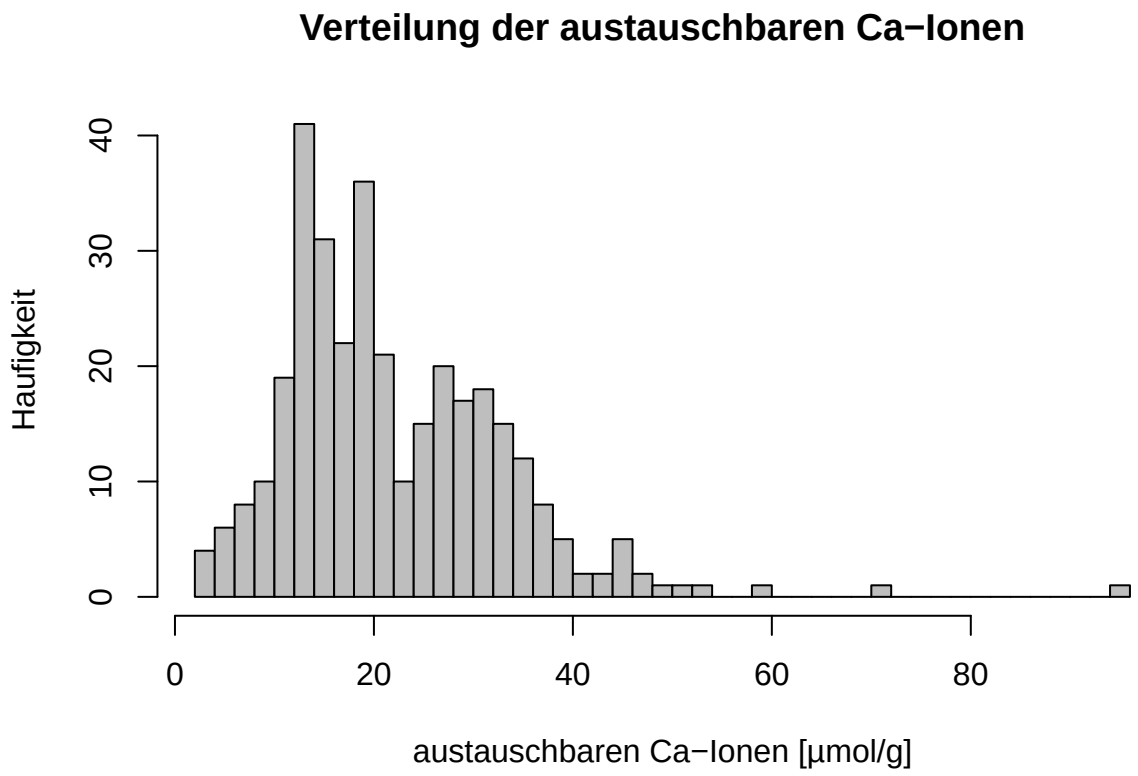
Auch für die Schiefe und die Kurtosis gibt es unterschiedliche berechnungsmöglichkeiten. Der selbst berechnete Kurtosiswert ist gleich dem Kurtosiswert aus dem Package psych, aber z.B. niedriger als der aus dem Package psych.
Der selbst berechnete Wert für die Schiefe hingegen ist gleich dem Skewness-Wert aus dem Package moments

und dem aus dem Package psych.

Aufgabe 08 Dichte-Histogramme und Box-Whisker-Plots in R

a) Erstellen Sie ein Dichte-Histogramm für die austauschbaren Ca-Ionen. Achten Sie dabei auf aussagekräftige Klassenweiten. Ändern Sie Titel und Achsenbeschriftungen sinnvoll ab und färben Sie die Balken grau ein.

```
hist(Ca,
     nclass = 50,
     main = "Verteilung der austauschbaren Ca-Ionen",
     xlab = "austauschbaren Ca-Ionen [µmol/g]",
     ylab = "Haufigkeit",
     col = "gray")
```

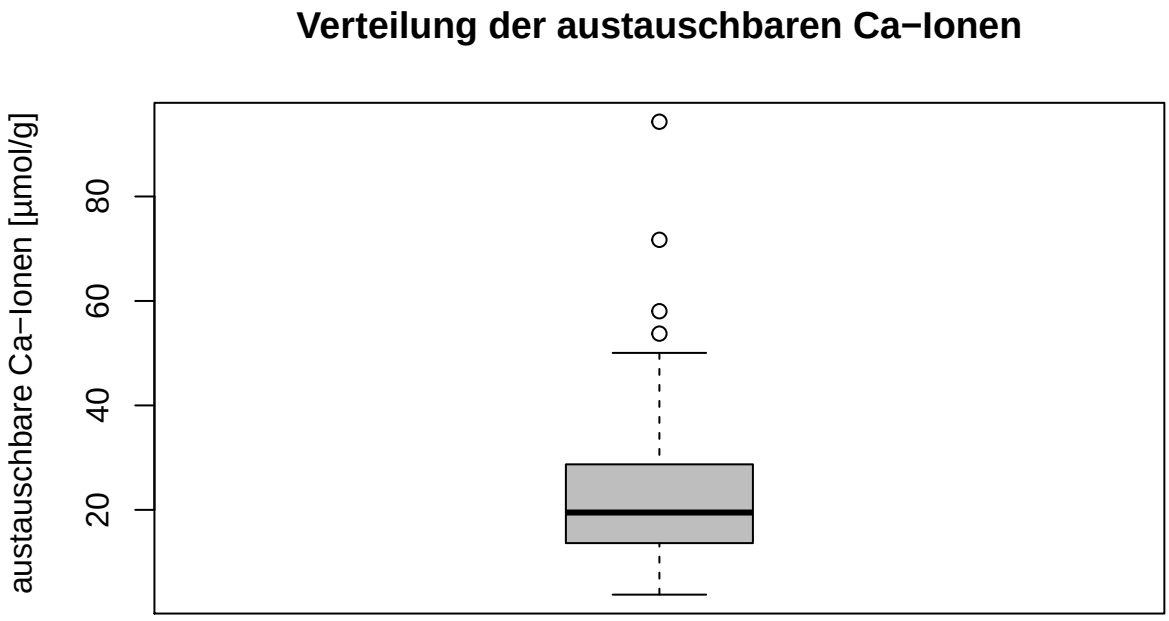


b) Um welche Verteilung handelt es sich bei der Ca-Ionenkonzentration dem visuellen Eindruck nach? Setzen Sie ihre Vermutung in Bezug zu den in Aufgabe 7b. errechneten Formparametern.

Die meisten Werte befinden sich in den ersten beiden Quantilen. Dadurch ist die Verteilung rechtsschief und linkssteil. Das spiegelt sich in dem Formparameter für die Schiefe (0,61 (1,5) > 0) wieder. Darüber hinaus verläuft die Verteilung steilgipflig. Das bedeutet, dass der Wertebereich, indem 90% der beobachteten Werte liegen kleiner ist, als bei der Normalverteilung und so der Median kleiner als der Mittelwert ist. Außerdem ist der Parameter für die Kutosis > 0 (Steland 2016).

c) Erstellen Sie nun für die austauschbaren Ca-Ionen einen Box-Whisker-Plot und untersuchen Sie das Randverhalten der Verteilung. Geben Sie mögliche Ausreißer in ihrem Protokoll an (range = 1,5). Vergeben Sie auch hier einen sinnvollen Titel sowie passende Achsenbeschriftungen. Verkleinern Sie die Balkenbreite, um eine ansprechende Grafik zu erzeugen.

```r
#Erstellen von Boxplots
boxplot(Ca,
        main = "Verteilung der austauschbaren Ca-Ionen",
        ylab = " austauschbare Ca-Ionen [µmol/g]",
        boxwex = 0.4,
        col = "grey")
```




```r
#Ausreisser aussortieren
boxplot.stats(ljz$Ca_exch)
```

```
## $stats
## [1]  3.7720 13.6330 19.4910 28.7175 50.0500
##
## $n
## [1] 335
##
## $conf
## [1] 18.18884 20.79316
##
## $out
## [1] 71.707 58.034 94.311 53.743
```

```r
x <- boxplot.stats(ljz$Ca_exch)$out
dplyr::filter(ljz, ljz$Ca_exch >= min(x))
```

```
## # A tibble: 4 x 9
##   OBJECTID SAMPLING   EAST   NORTH     C Ca_exch Mg_exch K_exch Na_exch
```

```
##      <int> <fct>     <dbl>   <dbl> <dbl>   <dbl>   <dbl>  <dbl> <chr>
## 1       86 regular  492991 3121890 0.909    71.7   10.3   2.02  0.052
## 2       91 regular  492991 3123990 1.23     58.0    7.90  0.419 0.887
## 3       99 regular  493291 3123090 0.432    94.3    7.92  4.18  0.035
## 4      197 regular  492541 3123390 1.46     53.7    6.94  1.18  0.887
```

Vor allem im oberen Wertebereich befinden sich Ausreißer. Sie liegen im 1,5-fachen Interquartelsabstand (Boxenlänge) von der Box entfernt. Die Box selbst geht vom ersten bis zum dritten Quantil und enthält die zentralen 50% der Werte (Steland 2016). Die Querstriche zeigen jeweils die Größte und kleinste Beobachtung ohne Ausreißer.
Um einen besseren Überblick über die Verteilung der austauschbaren Ca_Ionen zu erhalten, sollten die zwei höchsten Werte (94,31 µmol/g und 71,7 µmol/g) evtl. aus der Berechnung rausgelassen werden, da dort scheinbar Messfehler vorliegen (Walser 2011).

Aufgabe 09 Plotten in R

Boxplots mit den Pakten base graphics, lattice und ggplot2 für die austauschbaren Ca-Ionen für drei unterschiedliche Datensätze (regular, catA und catQ)

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

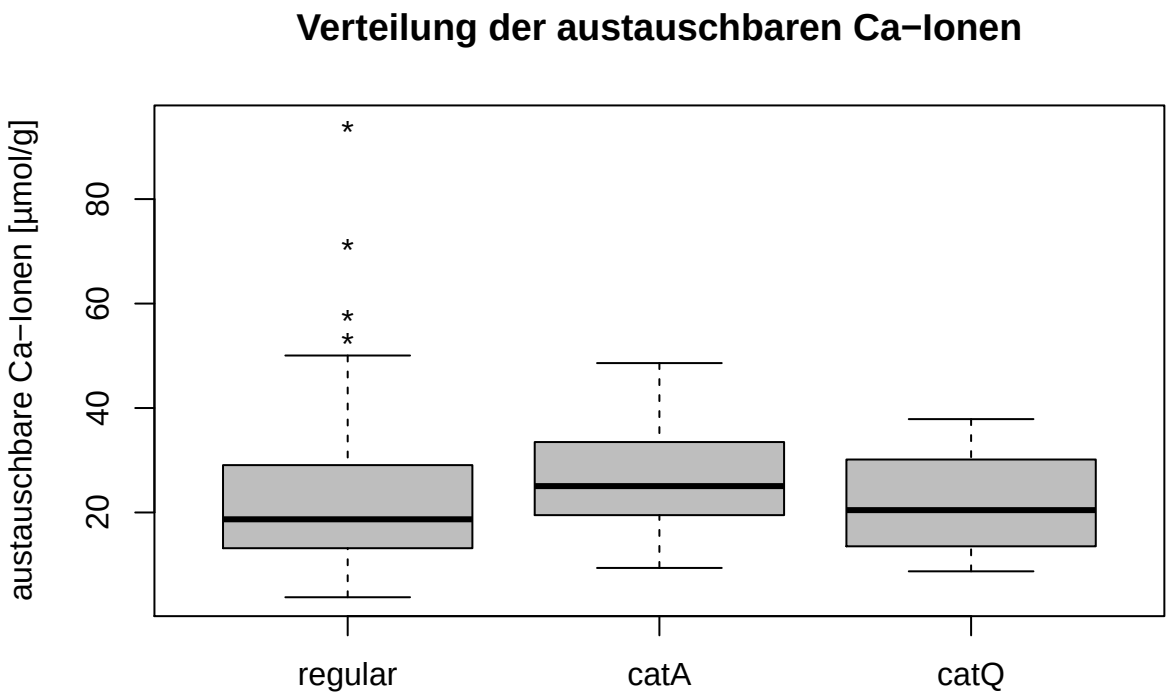
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
ljzmod <- ljz %>%
  dplyr::filter(SAMPLING == "regular" | SAMPLING == "catA" | SAMPLING == "catQ") %>%
  dplyr::select(OBJECTID, SAMPLING, Ca_exch)
```

In dem tibble ljzmod sind nun die Samples regular, catA und catQ der austauschbaren Ca-Ionen nach ihrer Object-ID in je einer Spalte sortiert.

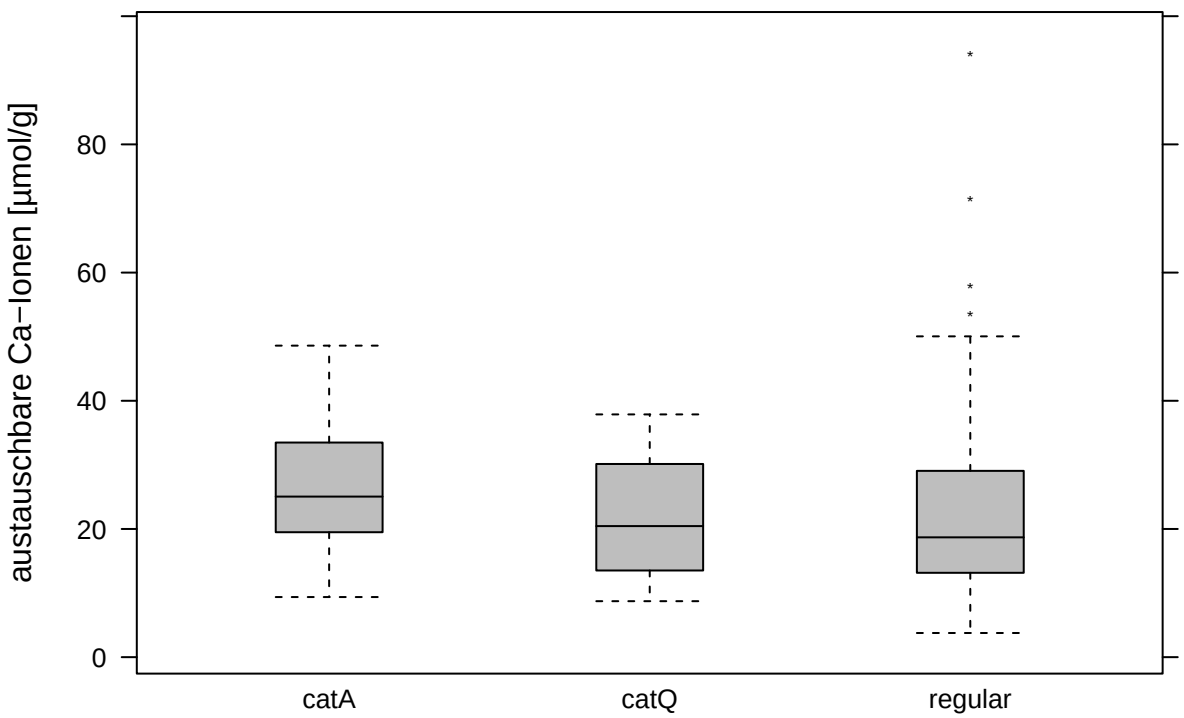
- Paket base graphics

```
#Variablen der einzelnen Samplings
regular <- ljzmod$Ca_exch[ljzmod$SAMPLING == "regular"]
catA <- ljzmod$Ca_exch[ljzmod$SAMPLING == "catA"]
catQ <- ljzmod$Ca_exch[ljzmod$SAMPLING == "catQ"]

#mit Ausreisser
boxplot(regular, catA, catQ,
        names = c("regular", "catA", "catQ"),
        ylab = "austauschbare Ca-Ionen [µmol/g]",
        main = "Verteilung der austauschbaren Ca-Ionen",
        col = "gray",
        width = c(0.4, 0.4, 0.4),
        pch = "*")
```



- Paket lattice

```
library(lattice)

#mit Ausreisser
bwplot(Ca_exch ~ SAMPLING, data = ljzmod,
       ylab = "austauschbare Ca-Ionen [µmol/g]",
       main = "Verteilung der Austauschbaren Ca-Ionen",
       box.ratio = 0.5,
       par.settings = list(box.rectangle = list(col = "black",
                           fill = "gray"),
                           box.dot = list(pch = "|"),
                           box.umbrella = list(col = "black"),
                           plot.symbol = list(col = "black",
                                              pch = "*")))
```



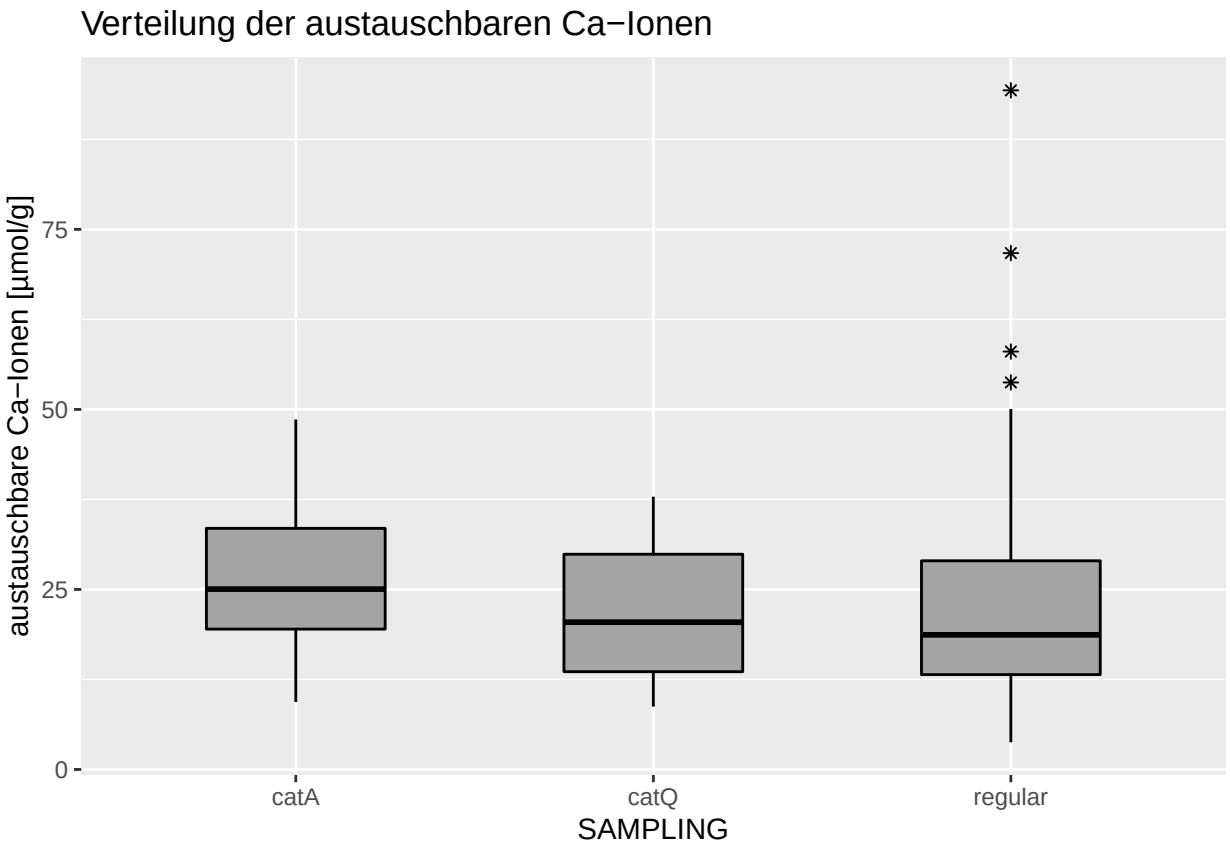
- Paket ggplot2

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
#mit Ausreisser
ggplot(ljzmod, aes(x = SAMPLING,
                   y = Ca_exch)) +
  labs(title = "Verteilung der austauschbaren Ca-Ionen",
       y = "austauschbare Ca-Ionen [µmol/g]") +
  geom_boxplot(fill = '#A4A4A4',
               color = "black",
               width = 0.5,
               outlier.shape = 8,
               orientation = TRUE)
```



Die Länge des Whiskers wird im base graphics Paket und im Paket lattice mit range beschrieben. Bei ggplot2 heißt das Argument coe. In allen Paketen ist die Default-Einstellung 1,5. Das Bedeutet, dass Werte, die den 1,5-fachen Interquartelsabstand (Boxenlänge) von der Box entfernt liegen, als Ausreißer gelten (Steland 2016).

Literatur

Hedderich, Jürgen, and Lothar Sachs. 2018. *Angewandte Statistik*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-56657-2.

Steland, Ansgar. 2016. *Basiswissen Statistik*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-49948-1.

Walser, Hans. 2011. *Statistik Für Naturwissenschaftler*. 1. Auflage. Bern, Stuttgart, Wien: Haupt Verlag.