

# R Kurzreferenz

*Karsten Lübke*

## Vorbemerkungen

Eine Übersicht von nützlichen R Funktionen innerhalb der Datenanalyse.

Diese Kurzreferenz beschreibt einen kleinen Teil der R Funktionen, wobei größtenteils auf das Zusatzpaket **mosaic** zurückgegriffen wird. Sie basiert größtenteils auf der Vignette *Minimal R* von Randall Pruim.

Weitere Hilfe und Beispiele finden Sie, wenn Sie

```
?Befehl
```

eingeben.

- R unterscheidet zwischen Groß- und Kleinbuchstaben
- R verwendet den Punkt `.` als Dezimaltrennzeichen
- Fehlende Werte werden in R durch `NA` kodiert
- Kommentare werden mit dem Rautezeichen `#` eingeleitet; der Rest der Zeile von R dann ignoriert.
- R wendet Befehle direkt an
- R ist objektorientiert, d. h. dieselbe Funktion hat evtl. je nach Funktionsargument unterschiedliche Rückgabewerte
- Zusätzliche Funktionalität kann über Zusatzpakete hinzugeladen werden. Diese müssen ggf. zunächst installiert werden
- Mit der Pfeiltaste nach oben können Sie einen vorherigen Befehl in der Konsole wieder aufrufen
- Eine Ergebnisuweisung erfolgt über `<-`

Innerhalb von **mosaic**:

```
analysiere(y ~ x | z , data=Daten)
```

- d. h., modelliere `y` in Abhängigkeit von `x` getrennt bzw. bedingt für `z` aus dem Datensatz `Daten`. Dabei können Teile (z. B. `y` und/ oder `z`) fehlen.<sup>1</sup>

Zusatzpakete müssen vor der ersten Benutzung einmalig installiert und geladen werden:

```
install.packages("Paket") # Einmalig installieren  
library(Paket) # Laden, einmalig in jeder Sitzung
```

## Daten

Daten einlesen und Datenvorverarbeitung sind häufig der (zeitlich) aufwendigste Teil einer Datenanalyse. Da die Daten die Grundlage sind, sollte auch hier sorgfältig gearbeitet und überprüft werden.

### Daten einlesen

```
read.table() # Allgemeinen Datensatz einlesen. Achtung: Optionen anpassen  
read.csv2() # csv Datensatz einlesen (aus deutschsprachigem Excel)  
file.choose() # Datei auswählen  
meineDaten <- read.csv2(file.choose())
```

U. a. mit Hilfe des Zusatzpaketes **readxl** können Excel Dateien eingelesen werden:

```
meineDaten <- read_excel(file.choose())
```

---

<sup>1</sup>Beim Mac ist `~` die Tastenkombination `alt+n`, `|` die Tastenkombination `alt+7`

## Daten verarbeiten

```
str() # Datenstruktur
head() # Obere Zeilen
tail() # Untere Zeilen
nrow(); ncol() # Anzahl Zeilen; Spalten
rownames(); colnames() # Zeilennamen, Spaltennamen
```

## Daten transformieren

Einzelne Variablen eines Datensatzes können über `$` ausgewählt werden: `Daten$Variable`. Allgemein kann über `Daten[i,j]` die i. Zeile und j. Spalte ausgewählt werden, wobei auch mehrere oder keine Zeile(n) bzw. Spalte(n) ausgewählt werden können. Über `c()` wird ein Vektor erzeugt.

```
as.factor() # Daten als Faktoren definieren
relevel() # Faktorstufen umordnen
droplevels() # Ungenutzte Faktorstufen entfernen
recode() # Umkodierung von Werten, Paket car
as.numeric() # Faktorstufen als numerische Daten verwenden
cut() # Aufteilung numerischer Werte in Intervalle

subset() # Teilmenge der Daten auswählen
na.omit() # Zeilen mit fehlenden Werten entfernen

log() # Logarithmusfunktion
exp() # Exponentialfunktion
sqrt() # Quadratwurzelfunktion
abs() # Betragsfunktion

rowSums() # Zeilensumme
rowMeans() # Zeilenmittelwert
```

Innerhalb des Paketes `dplyr` (wird mit `mosaic` geladen) gibt es u. a. folgende Funktionen:

```
filter() # Filtert Beobachtungen eines Datensatzes
select() # Wählt Variablen eines Datensatzes aus
mutate() # Erzeugt neue Variable bzw. verändert bestehende
rename() # Benennt Variablen um
arrange() # Sortiert Beobachtungen eines Datensatzes
```

Dabei übergibt die “Pfeife” `%>%` das Ergebnis der vorhergehenden Funktion an die folgende.

## Grafische Verfahren

Vor jeder mathematisch-statistischen Analyse sollte eine explorative, grafische Analyse erfolgen. Die folgenden Befehle sind aus dem Paket `mosaic`.

```
bargraph() # Balkendiagramm
histogram() # Histogramm
bwplot() # Boxplot
xyplot() # Streudiagramm
```

Nicht aus dem Paket `mosaic` sind:

```
mosaicplot() # Mosaicplot
corrplot() # Korrelationsplot, Paket corrplot
ggpairs() # Matrixplot, Paket GGally
heatmap() # Heatmap
```

## Deskriptive Statistik

Eine gute Zusammenfassung liefert der `mosaic` Befehl:

```
favstats()
```

Ansonsten (`mosaic` angepasst):

```
tally() # Tabellierung, Häufigkeiten
mean() # Arithmetischer Mittelwert
median() # Median
sd() # Standardabweichung
var() # Varianz
IQR() # Interquartilsabstand
cov() # Kovarianz
cor() # Korrelationskoeffizient
```

## Inferenzstatistik

### Randomisierung, Simulationen

Größtenteils `mosaic`:

```
set.seed() # Zufallszahlengenerator setzen
rflip() # Münzwurf
do() # Wiederholung (Schleife)
sample() # Stichprobe ohne Zurücklegen
resample() # Stichprobe mit Zurücklegen
rnorm() # Normalverteilte Zufallszahlen
```

### Verteilungen

Innerhalb der Funktionen müssen ggf. die Parameter, d. h. `mean=`, `sd=`, bzw. `df=` angepasst werden.

```
pchisq() # Verteilungsfunktion Chi2 Verteilung
qchisq() # Quantilsfunktion Chi2 Verteilung
pnorm() # Verteilungsfunktion Normalverteilung
qnorm() # Quantilsfunktion Normalverteilung
pt() # Verteilungsfunktion t-Verteilung
qt() # Quantilsfunktion t-Verteilung
```

Analoger Aufbau für weitere Verteilungen, z. B. `_pnorm()`, `_f()`.

### Testverfahren

Einige der Testverfahren wurden von `mosaic` angepasst.

```
t.test() # t-Test
prop.test() # Binomialtest (approximativ)
xchisq.test() # Chi2-Test
aov() # Varianzanalyse
```

Der Nicht-parametrische Wilcoxon-Test `wilcox.test()` ist nicht im Paket `mosaic`, hat daher einen leicht anderen Funktionsaufruf. Einen Test auf Normalverteilung führt der Shapiro-Wilk Test durch: `shapiro.test()`.

## Multivariate Verfahren

```
lm() # Lineare Regression  
glm(), family="binomial" # Logistische Regression  
residuals() # Residuen einer Regression  
fitted() # Angepasste Werte einer Regression
```

In **mosaic** kann das Ergebnis einer solchen Regression über **makeFun()** in eine einfache mathematische Funktion überführt werden. **plotFun()** zeichnet das Ergebnis. **step()** führt eine Variablenselektion durch.

Weitere Verfahren - nicht **mosaic**:

```
hclust() # Hierarchische Clusteranalyse  
kmeans() # k-Means Clusterverfahren  
prcomp() # Hauptkomponentenanalyse (PCA)  
rpart() # Klassifikations- und Regressionsbäume, Paket rpart
```

### Versionshinweise:

Erstellt von Karsten Lübke unter der Lizenz Creative Commons Attribution-ShareAlike 3.0 Unported.

- Datum erstellt: 2017-04-28
- R Version: 3.4.0
- **mosaic** Version: 0.14.4