Intro Datenanalyse mit R

Jan-Philipp Kolb

06 Mai, 2019

Streuungsmaße

In Basis R sind die wichtigsten Streuungsmaße enthalten:

- Varianz: `var()`
- Standardabweichung: `sd()`
- Minimum und Maximum: `min()` und `max()`
- Range: `range()`

```
ab <- rnorm(100)
var(ab)
```

```
## [1] 1.010516
```

```
sd(ab)
```

```
## [1] 1.005244
```

```
range(ab)
```

```
## [1] -2.795246  2.202099
```

Extremwerte

```
min(ab)
```

```
## [1] -2.795246
```

```
max(ab)
```

```
## [1] 2.202099
```

Fehlende Werte

- Sind `NA`s vorhanden muss dies der Funktion mitgeteilt werden

```
ab[10] <- NA
var(ab)
```

```
## [1] NA
```

Bei fehlenden Werten muss ein weiteres Argument mitgegeben werden:

```
var(ab,na.rm=T)
```

```
## [1] 1.015309
```

Beispieldatensatz Tourismus

Daten importieren

```
dat <- rio::import("http://offenedaten.frankfurt.de/dataset/48
```

```
anteil_international9 <- dat$`Tourismus Anteil der Gäste aus
diff_9 <-dat$`Tourismus Veränderung der Anzahl der Gäste 2009
```

Kategorien bilden

```
tourismus_kat4 <- cut(anteil_international9,4)
tourismus_kat3 <- cut(diff_9,3)
```

Häufigkeiten und gruppierte Kennwerte

- Eine Auszählung der Häufigkeiten der Merkmale einer Variable liefert `table()`
- Mit `table()` sind auch Kreuztabellierungen möglich indem zwei Variablen durch Komma getrennt werden: `table(x,y)` liefert Häufigkeiten von `y` für gegebene Ausprägungen von `x`

```
table(tourismus_kat4)
```

```
## tourismus_kat4
## (4.16,14.1]   (14.1,24]   (24,33.9] (33.9,43.8]
##          12           8           4           1
```

Tabellieren - weiteres Beispiel

```
?table
```

```
table(tourismus_kat4,tourismus_kat3)
```

```
##                 tourismus_kat3
## tourismus_kat4 (-13.2,-1.73] (-1.73,9.73] (9.73,21.2]
##    (4.16,14.1]              7             4           1
##    (14.1,24]                7             1           0
##    (24,33.9]                4             0           0
##    (33.9,43.8]              1             0           0
```

Eine weitere Tabelle

```
data(esoph)
table(esoph$agegp)
```

```
##
## 25-34 35-44 45-54 55-64 65-74   75+
##    15    15    16    16    15    11
```

Häufigkeitstabellen

- `prop.table()` liefert die relativen Häufigkeiten
- Wird die Funktion außerhalb einer `table()` Funktion geschrieben erhält man die relativen Häufigkeiten bezogen auf alle Zellen

Die Funktion ‘`prop.table()`

```
table(esoph$agegp,esoph$alcgp)
```

```
##
##           0-39g/day 40-79 80-119 120+
##    25-34          4     4       3    4
##    35-44          4     4       4    3
##    45-54          4     4       4    4
##    55-64          4     4       4    4
##    65-74          4     3       4    4
##    75+            3     4       2    2
```

Die Funktion `prop.table`

```
?prop.table
```

```
prop.table(table(esoph$agegp,esoph$alcgp),1)
```

```
##
##           0-39g/day     40-79    80-119      120+
##    25-34 0.2666667 0.2666667 0.2000000 0.2666667
##    35-44 0.2666667 0.2666667 0.2666667 0.2000000
##    45-54 0.2500000 0.2500000 0.2500000 0.2500000
##    55-64 0.2500000 0.2500000 0.2500000 0.2500000
##    65-74 0.2666667 0.2000000 0.2666667 0.2666667
##    75+   0.2727273 0.3636364 0.1818182 0.1818182
```

Die aggregate Funktion

- Mit der `aggregate()` Funktion können Kennwerte für Untergruppen erstellt werden
- `aggregate(x,by,FUN)` müssen mindestens drei Argumente übergeben werden:

```
aggregate(state.x77,by=list(state.region),mean)
```

```
##         Group.1 Population   Income Illiteracy Life Exp    
## 1     Northeast   5495.111 4570.222   1.000000 71.26444  4.
## 2         South   4208.125 4011.938   1.737500 69.70625 10.
## 3 North Central   4803.000 4611.083   0.700000 71.76667  5.
## 4          West   2915.308 4702.615   1.023077 71.23462  7.
##      Frost       Area
## 1 132.7778  18141.00
## 2  64.6250  54605.12
## 3 138.8333  62652.00
## 4 102.1538 134463.00
```

Beispieldatensatz - apply Funktion

```
ApplyDat <- cbind(1:4,runif(4),rnorm(4))
```

```
apply(ApplyDat,1,mean)
```

```
## [1] 0.2023834 0.6412275 1.2120531 1.5657344
```

```
apply(ApplyDat,2,mean)
```

```
## [1]  2.5000000  0.7767190 -0.5606702
```

Die Funktion apply

```
apply(ApplyDat,1,var)
```

```
## [1] 1.740430 2.098120 2.544774 4.548241
```

```
apply(ApplyDat,1,sd)
```

```
## [1] 1.319254 1.448489 1.595235 2.132661
```

```
apply(ApplyDat,1,range)
```

```
##           [,1]       [,2]        [,3]      [,4]
## [1,] -1.320385 -0.8828022 -0.06559754 0.0261038
## [2,]  1.000000  2.0000000  3.00000000 4.0000000
```

```
apply(ApplyDat,1,length)
```

```
## [1] 3 3 3 3
```

Argumente der Funktion apply

- Für `margin=1` die Funktion `mean` auf die Reihen angewendet,
- Für `margin=2` die Funktion `mean` auf die Spalten angewendet,
- Anstatt `mean` können auch andere Funktionen wie `var`, `sd` oder `length` verwendet werden.

Die Funktion tapply

```
ApplyDat <- data.frame(Income=rnorm(5,1400,200),
                       Sex=sample(c(1,2),5,replace=T))
```

- Auch andere Funktionen können eingesetzt werden. . . . - Auch selbst programmierte Funktionen
- Im Beispiel wird die einfachste eigene Funktion angewendet.

```
ApplyDat
```

```
##      Income Sex
## 1 1408.530   1
## 2 1411.878   2
## 3 1324.617   2
## 4 1610.984   2
## 5 1500.536   2
```

Beispiel Funktion tapply

```
tapply(ApplyDat$Income,ApplyDat$Sex,mean)
```

```
##        1        2
## 1408.530 1462.004
```

```
tapply(ApplyDat$Income,
       ApplyDat$Sex,function(x)x)
```

```
## $`1`
## [1] 1408.53
##
## $`2`
## [1] 1411.878 1324.617 1610.984 1500.536
```

Links Datenanalyse

- Die Benutzung von `apply`, `tapply`, etc. (Artikel bei R-bloggers)
- Quick-R zu deskriptiver Statistik
- Quick-R zur Funktion `aggregate`