Universität Bremen

FB3, A. Mändle WS 2019/20, 04.11.2019

Einführung in R
4. Aufgabenblatt

Präsenzaufgabe 1

In Hausaufgabe 1e) von Blatt 3 haben wir die Anwendung der Befehle `unique(x)`, `duplicated(x)`, `x[!duplicated(x)]` auf einen Vektor `x` gesehen. Jetzt wollen wir die gleichen Fragen beantworten für den Fall, dass `x` ein Datensatz ist. Laden Sie den Datensatz `airquality` aus dem Paket `datasets`.

a) Interpretieren Sie den Befehl
`AirDouble <- rbind(airquality, airquality[sample(1:nrow(airquality), 10),])`.

b) Welche Subgruppen bildet man mit
`unique(AirDouble); AirDouble[!duplicated(AirDouble),]`?

c) Welche Subgruppen bildet man mit
`airquality[!duplicated(airquality[,5]),]` bzw.
`airquality[!duplicated(airquality[,6]),]`?

Präsenzaufgabe 2

Der Datensatz *gewicht.txt* (Datei im *StudIP* verfügbar) enthält das Geburtsgewicht eines Neugeborenen (bwt), das Rauchverhalten der Mutter (smoke, 0≙Nichtraucher, 1≙Raucher, 9≙fehlender Wert) und die Identifikationsnummer der Mutter (PatID). Laden Sie diesen Datensatz in R und bearbeiten Sie die folgenden Aufgaben:

a) Ersetzen Sie in der Spalte `smoke` Werte, die gleich 9 sind, durch `NA`.

b) Berechnen Sie *Minimum*, *Maximum*, *Mittelwert*, *Standardabweichung* und *Median* des Geburtsgewichts gruppiert nach dem Rauchverhalten der Mutter.

c) Der Datensatz enthält auch die Daten von Müttern mit mehreren Kindern. Gehen Sie davon aus, dass die Daten nach dem Datum geordnet sind. Bilden Sie eine Untergruppe `first_value`, die angibt, ob es sich um Daten für die jeweils erste Geburt der Mütter handelt.

d) Bilden Sie eine Untergruppe `last_value`, die angibt, ob es sich um Daten für die jeweils letzte Geburt der Mütter handelt.

e) Bilden Sie eine Untergruppe `non_repeated`, welche die Daten der Mütter danach gruppiert, ob sie nur ein Kind geboren haben.

f) Bearbeiten Sie die Teilaufgabe b) auch für die in c) bis e) erstellten Subgruppen.

Präsenzaufgabe 3

Speichern Sie den Datensatz *bilirubin.txt* (Datei im *StudIP* verfügbar) auf Ihrem Rechner und laden Sie ihn in R.

a. Berechnen Sie *Minimum*, *Maximum*, *Mittelwert*, *Standardabweichung* und *Median* vom Bilirubin-Wert und vom Alter der Patienten jeweils gruppiert nach dem Geschlecht.

b. Wieviele Patienten sind mindestens 70 Jahre alt?

c. Wieviele Frauen in diesem Datensatz sind jünger als 30 Jahre und haben Bilirubin-Werte kleiner als 1.1?

d. Wieviel Prozent der unter 30-jährigen Männer bzw. Frauen haben Bilirubin-Werte zwischen 0.4 und 1.2 ([0.4, 1.2])?

Hausaufgabe 1 (4 Punkte)

Der bekannte Iris-Datensatz von Fisher bzw. Anderson enthält Messungen der Größen von Kelch und Blütenblättern dreier verschiedener Arten von Schwertlilien.

a. Laden Sie den Datensatz `iris` aus dem Paket `datasets`

b. Rufen Sie mit dem entsprechenden R-Befehl die Hilfe zum Datensatz auf und lesen Sie die Informationen zu diesem Datensatz.

c. Berechnen Sie die Mittelwerte der 4 metrisch-skalierten Merkmale des Datensatzes gruppiert nach dem qualitativen Merkmal `Species`. Benutzen Sie dazu die Funktion `tapply()`.

d. Lösen Sie Teilaufgabe c) diesmal mit `aggregate()`

Hausaufgabe 2 (12 Punkte)

Im Folgenden benötigten Sie den Datensatz *heartatk4R.txt* (verfügbar unter `http://statland.org/R/R/heartatk4R.txt`).

a. Laden Sie den Datensatz direkt unter Angabe der URL als Dateiname in R ein. Weitere Informationen zum Datensatz finden Sie hier: `http://statland.org/R/R/heartvar.txt`

b. Berechnen Sie den Mittelwert des Alters der Patienten gruppiert nach *jeweils* `SEX`, `DIAGNOSIS`, `DRG` und `DIED`.

c. Die Krankenhausaufenthaltskosten für jeden Patienten sind in der Spalte `CHARGES` angegeben. Berechnen Sie den Mittelwert der Kosten gruppiert nach *jeweils* `SEX`, `DIAGNOSIS`, `DRG` und `DIED`. Kommentieren Sie knapp wo die größten Kosten entstehen!

d. Die Dauer des Krankenhausaufenthalts für jeden Patienten ist in der Spalte `LOS` angegeben. Berechnen Sie den Mittelwert von `LOS` gruppiert *gleichzeitig* nach `SEX`, `DIAGNOSIS`, `DRG` und `DIED`.

   Hinweis: Das zweite Argument in `tapply(X,INDEX,FUN)` bzw. `aggregate(x,by,FUN)` kann auch eine Liste mehrerer Gruppierungen sein: Bsp.: `tapply(X=mydata, INDEX=list(var1, var2),FUN=mean)`.

Hausaufgabe 3 (4 Punkte)

Im folgenden benötigen Sie den Datensatz *birthwt* aus dem R-Paket *MASS*.

a. Laden Sie den genannten Datensatz und rufen Sie die Hilfe zum Datensatz auf.

b. Vergleichen Sie die Mittelwerte vom Geburtsgewicht der Neugeborenen gruppiert nach dem Rauchverhalten ihrer Mutter.

c. Vergleichen Sie die Mittelwerte vom Geburtsgewicht der Neugeborenen gruppiert nach dem Bluthochdruck (Hypertonie) ihrer Mutter.

d. Benutzen Sie die folgenden, neuen Befehle und versuchen Sie diese zu interpretieren. Konsultieren Sie dazu die Hilfe, falls nötig.
```
boxplot(birthwt$bwt ~ birthwt$smoke)
t.test(birthwt$bwt ~ birthwt$smoke)
boxplot(birthwt$bwt ~ birthwt$ht)
t.test(birthwt$bwt ~ birthwt$ht)
```

Zusatzaufgabe (3 Punkte)

Nutzen Sie für diese Aufgabe weiterhin den `birthwt`-Datensatz.

a. Wandeln Sie die Variablen `smoke` und `ht` in Faktoren um.

b. Berechnen Sie die Mittelwerte vom Geburtsgewicht der Neugeborenen gruppiert nach dem Rauchverhalten **und** Bluthochdruck (Hypertonie) ihrer Mutter, indem Sie für das Argument `INDEX` die beiden Variablen in einer Liste übergeben.

c. Veranschaulichen Sie das Ergebnis mit dem Befehl:
`interaction.plot(birthwt$smoke, birthwt$ht, birthwt$bwt)`.

d. Eine Alternative zur Veranschaulichung bietet die Funktion `bwplot()` aus dem Paket `lattice` an. Diese Funktion mit den Angaben `bwplot(y ~ x1 | x2)` zeichnet gruppierte Boxplots einer numerischen Variable `y` gruppiert nach zwei qualitativen Variablen `x1` und `x2`. Veranschaulichen das obige Ergebnis mit Hilfe dieser Funktion.

Abgabe der Lösungen: bis ***Montag 11.11.2019,***

maendle@uni-bremen.de