Universität Bremen

FB3, A. Mändle WS 2019/20, 28.10.2019

Einführung in R
3. Aufgabenblatt

1. Schreiben Sie Ihre R-Befehle in ein R-Skript.
2. Geben Sie die Lösungen zu den Hausaufgaben in einem R-Skript ab.

Präsenzaufgabe 1

Laden Sie den Datensatz `airquality` aus dem Paket `datasets`.

a) Bereinigen Sie den Datensatz um fehlende Werte und legen Sie ihn als Variable `air3` im Speicher ab. Hinweis: `na.omit()` oder `complete.cases()`

b) Bereinigen Sie den Datensatz um fehlende Werte der 1. Spalte und speichern Sie ihn als Variable `air4` (Hinweis: `!is.na()`).

c) Bereinigen Sie den Datensatz um fehlende Werte der 1. und 2. Spalte und speichern Sie ihn als Variable `air5`.

d) Wann war der Wind (bzw. die Temperatur, Ozonkonzentration, Sonneneinstrahlung) am stärksten?

Präsenzaufgabe 2

Schreiben Sie eine Funktion, welche die m-te und n-te (m und n beliebige natürliche Zahlen) Wurzel aus den Werten eines Vektors (mit positiven Einträgen) berechnet und angibt.

Präsenzaufgabe 3

a) Schreiben Sie eine Funktion, welche den ersten Eintrag eines Vektors `x` durch `NA` ersetzt, falls `x[1]` negativ oder gleich dem Wert 999 ist. Hinweis: Lesen Sie in der Hilfe zu `?ifelse`.

b) Schreiben Sie eine Funktion, welche alle Werte, die negativ oder gleich 999 sind, durch `NA` ersetzt.

Präsenzaufgabe 4

Die Merkmale `Ozone` und `Solar.R` des Datensatzes `airquality` (im Paket `datasets`) haben Einträge mit `NA`.

a) Fügen Sie dem Datensatz eine neue Variable `Ozone1` hinzu, bei welcher die `NA`-Einträ ge in der `Ozone`-Spalte durch den Wert -99 ersetzt wurden.
Benutzen Sie den `ifelse`-Befehl!

b) Definieren Sie eine ähnliche Spalte für das Merkmal `Solar.R`.
*Benutzen Sie aber diesmal **nicht** den `ifelse`-Befehl.*

Präsenzaufgabe 5

Erzeugen Sie mithilfe der Funktion `sample()` einen Vektor, welcher die Werte $-3, -2, \ldots, 3$ und `NA` in zufälliger Reihenfolge enthält. Bilden sie eine neue Variable $y = f(x)$, die wie folgt definiert ist:

$$y = f(x) = \begin{cases} x^2, & \text{falls} \quad x \geq 0 \\ \frac{1}{x^2}, & \text{falls} \quad x < 0 \\ 0, & \text{falls} \quad x = NA \end{cases}$$

Benutzen Sie den `ifelse`-Befehl!

Hausaufgabe 1 (9 Punkte)

Die Größe einer Grundgesamtheit (in cm) sei normalverteilt mit den Parametern $\mu = 170$ und $\sigma = 10$. Verschaffen Sie sich eine Stichprobe vom Umfang $n = 2000$ aus dieser Population mittels folgendem Befehl:
`pop1<-rnorm(2000, 170, 10)`.

a. Runden Sie die Größe auf ganze Zahlen mithilfe des Befehls `round()` (d.h. `pop1` mit den gerundeten Werten überschreiben!).

b. Führen Sie den Befehl `which(pop1 > 190)` aus und interpretieren Sie das Ergebnis.

c. Welche Subgruppe wird durch `pop1[pop1 > 190]` erzeugt?

d. Bilden Sie 2 Subgruppen: 1) *Größe* < 155 und 2) *Größe* > 185.

e. Interpretieren Sie die Ergebnisse:
`rev(pop1); unique(pop1); duplicated(pop1);`
`pop1[duplicated(pop1)]; pop1[!duplicated(pop1)]`.

f. Simulieren Sie eine weitere Stichprobe `pop2` vom Umfang $n = 1000$ aus einer Grundgesamtheit, deren Größe $N(\mu = 172, \sigma = 12)$-verteilt ist.

g. Runden Sie die Größe in der 2. Stichprobe auf ganze Zahlen.

h. Ordnen Sie 1000 Individuen aus der 1. Stichproben der Zweiten **zufällig** zu. Definieren Sie dazu eine Matrix `X` mit zwei Spalten, deren 1. bzw. 2. Spalte die Individuen der 1. bzw. 2. Stichproben beinhaltet.
Hinweis: verwenden Sie `sample(1000,pop1,replace=FALSE)` zur Randomisierung.

i. Wir betrachten jede Zeile dieser Matrix als ein Paar, wobei wir die Individuen der 1. bzw. 2. Stichproben als Mann bzw. Frau bezeichnen. Bei wie vielen Paaren ist der Mann kleiner als die Frau?

Zusatzaufgabe (1 P): *entspricht der von Ihnen berechneten Wert in i) dem erwarteten Wert, den man aus der Wahrscheinlichkeitsrechnung kennt? Begründen Sie Ihre Antwort!*

Hausaufgabe 2 (6 (+2) Punkte)

a. Definieren Sie eine Funktion, die angibt, ob es in einer Spalte eines Datensatzes `NA`-Werte gibt. (Datensatz wird als Argument der Funktion übergeben!)

b. Definieren Sie eine Funktion, die angibt, ob es in einer Zeile eines Datensatzes `NA`-Werte gibt.

c. Wenden Sie diese Funktionen an den Spalten bzw. Zeilen des `airquality`-Datensatzes an.

d. **Zusatzaufgabe:** Definieren Sie eine Funktion, die für einen Datensatz angibt, in welchen Spalten es `NA`-Werte gibt.

e. **Zusatzaufgabe:** Das Gleiche wie in d., aber dieses Mal soll die Funktion angeben, in welchen Zeilen es `NA`-Werte gibt.

Hausaufgabe 3 (2+3 Punkte)

Die Variable `Wind` im Datensatz `airquality` gibt die Windsgeschwindigkeit in *mph (mile per hour)* an. Dies ist eine metrisch-skalierte Variable.

a. Bilden Sie aus dieser Variable eine neue, ordinal-skalierte Variable `wind`, wie folgt:
wind=`"schwach"`, *falls* $Wind \leq 12$,
wind=`"stark"`, *sonst*
und fügen Sie diese Variable dem Datensatz `airquality` hinzu.

b. Definieren Sie jetzt das ordinal-skalierte Merkmal *wind* als:
wind=`"schwach"`, *falls* $Wind \leq 4$,
wind=`"mild"`, *falls* $4 < Wind \leq 18$
wind=`"stark"`, *sonst.*

Benutzen Sie den `ifelse`-Befehl!
Hinweise: `wind` kann in einem Zwischenschritt zunächst als Zeichenkette (`character`) erzeugt werden. Mit der Funktion `factor(x=?, levels=?, ordered=?)` kann dann daraus ein ordinalskaliertes Merkmal erzeugt werden.

Abgabe der Lösungen: bis ***Dienstag 06.11.2018,***

an maendle@uni-bremen.de oder über das Stud.IP.