Vorbereitung

Bitte führen Sie zur Vorbereitung folgende Schritte aus:

1. Starten Sie RStudio.
2. Löschen Sie den Workspace.
3. Setzen Sie das Arbeitsverzeichnis dort, wo Sie Ihre Daten abgelegt haben: Session › Set Working Directory › Choose Directory.
4. Öffnen Sie ein R-Skript.
5. Nachdem Sie die Aufgaben bearbeitet haben, speichern Sie das Skript unter einem geeigneten Namen ab.

Aufgabe 1 - Modus & Häufigkeitstabelle

Erstellen Sie eine Häufigkeitstabelle für die Variable der Kursgruppe (`gruppe`) und bestimmen Sie den Modus, (a) indem Sie ihn aus der Tabelle ablesen und (b) indem Sie eine angemessene Grafik erstellen und den Modus daran ablesen.

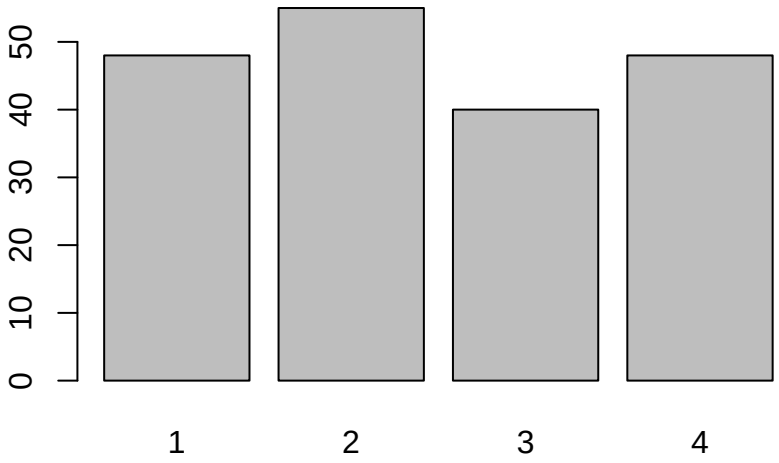
Lösung

```
table(erstis$gruppe)
```

```
 1  2  3  4
48 55 40 48
```

- Mo = `Gruppe 2`

```
barplot(table(erstis$gruppe))
```



Zusatzaufgabe: Benutzen Sie einen anderen Befehl, um sich den Modus direkt ausgeben zu lassen. (Hinweis: Installieren Sie dafür ggf. neue Pakete.)

Lösung

```
#install.packages("DescTools")
library(DescTools)
Mode(erstis$gruppe, na.rm = T)
```

```
[1] 2
attr(,"freq")
[1] 55
Levels: 1 2 3 4
```

Aufgabe 2 - Median

(i) Erstellen Sie eine Tabelle, die die prozentuale und die kumulierte prozentuale Häufigkeitsverteilung des Items `lz17` („Wenn ich mein Leben noch einmal leben könnte, würde ich kaum etwas ändern.“) gemeinsam abbildet (auf zwei Nachkommastellen gerundet). Bestimmen Sie den Median für diese Variable (a) anhand der erstellten Tabelle, (b) anhand des Befehls für Quantile, und (c) anhand des Befehls für den Median.

Lösung

Häufigkeitstabelle für das Item `lz17`.

```
(lz17abs <- table(erstis$lz17))
```

```
 1  2  3  4  5  6  7
 7 22 19 46 35 36 23
```

Prozentual und kumulierte prozentuale Häufigkeitsverteilung von lz17, auf zwei Nachkommstellen gerundet.

```
lz17pro <- 100*prop.table(lz17abs)
lz17kum <- cumsum(lz17pro)
(lz17tab <- round(cbind(lz17pro, lz17kum), digits = 2))
```

```
  lz17pro lz17kum
1    3.72    3.72
2   11.70   15.43
3   10.11   25.53
4   24.47   50.00
5   18.62   68.62
6   19.15   87.77
7   12.23  100.00
```

- Spezialfall: 50% wird in Klasse 4 genau erreicht, aber nicht überschritten.
- Vorgehen “per Hand” -> Anzahl gültiger Werte gerade oder ungerade?

```
sum(lz17abs)
```

```
[1] 188
```

- Gerade Anzahl -> wir mitteln also 4 (höchster Wert untere Hälfte) und 5 (geringster Wert obere Hälfte), um bei Md = 4,5 anzugelangen.

```
quantile(erstis$lz17, type = 5, na.rm = TRUE)
```

```
  0%  25%  50%  75% 100%
 1.0  3.0  4.5  6.0  7.0
```

- Berechnung des Medians mittels `quantile()` Befehl

```
median(erstis$lz17, na.rm = TRUE)
```

```
[1] 4.5
```

- Berechnung des Medians mittels `median()` Befehl

(ii) Eine Person sagt "Mindestens 10 % der Personen haben einen Wert angekreuzt, der kleiner oder gleich meinem Wert ist." Welchen Wert hat die Person?

Lösung

```
quantile(erstis$lz17, type = 5, na.rm = TRUE, probs = c(.1))
```

```
10%
  2
```

- Die Person hat einen Wert von 2.

(iii) Eine Person sagt "Mindestens 70 % der Personen haben einen Wert angekreuzt, der kleiner oder gleich meinem Wert ist." Welchen Wert hat die Person?

```
quantile(erstis$lz17, type = 5, na.rm = TRUE, probs = c(.7))
```

```
70%
  6
```

- Die Person hat einen Wert von 6.

Alternativ könnte man die entsprechenden Werte auch aus der Häufigkeitstabelle ablesen. Hier schauen wir, wann bei den kumulierten Häufigkeiten 10% bzw. 70% erstmalig überschritten sind.

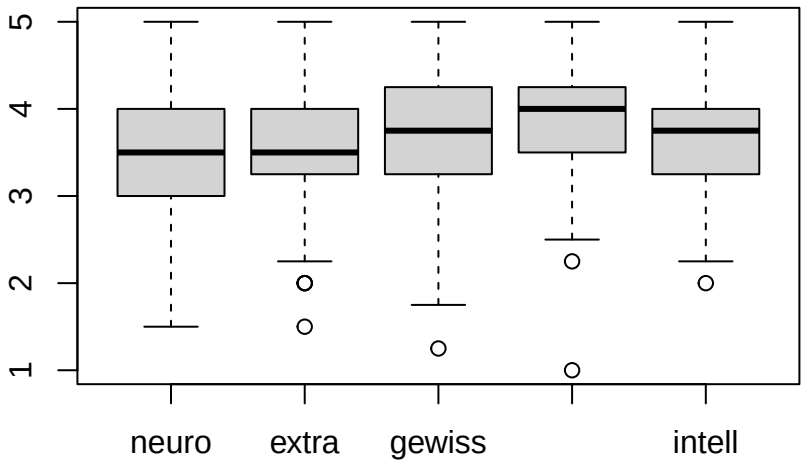
Aufgabe 3 - Median II

(i) Schauen Sie sich die fünf Persönlichkeitsdimensionen (`neuro`, `extra`, `gewiss`, `vertraeg`, `intell`) an. Beantworten Sie die Frage, für welche Skala der Median am höchsten ist. (a) graphisch, (b) mithilfe eines Befehls.

Lösung: Verträglichkeit hat den höchsten Median.

(i) graphisch

```
set <- c("neuro", "extra", "gewiss", "vertraeg", "intell")
boxplot(erstis[, set])
```



(ii)

```
median(erstis$neuro, na.rm = T)
```

```
[1] 3.5
```

```
median(erstis$extra, na.rm = T)
```

```
[1] 3.5
```

```
median(erstis$gewiss, na.rm = T)
```

```
[1] 3.75
```

```
median(erstis$vertraeg, na.rm = T)
```

```
[1] 4
```

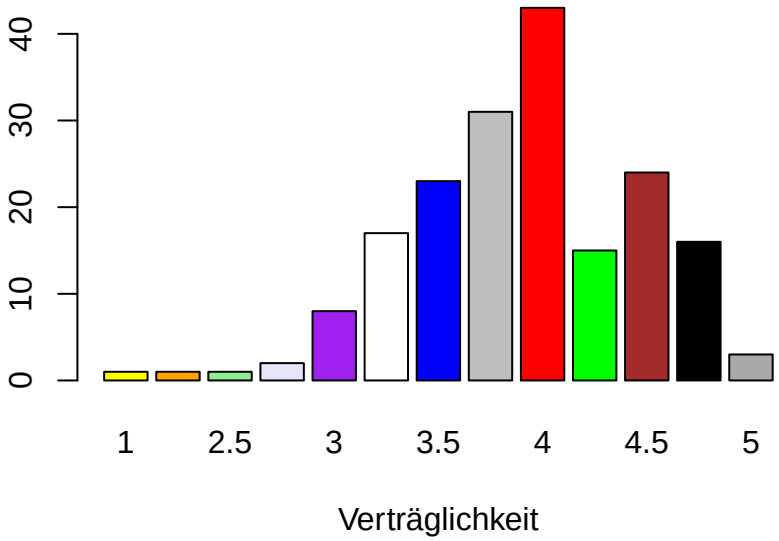
```
median(erstis$intell, na.rm = T)
```

```
[1] 3.75
```

(ii) Erstellen Sie ein Balkendiagramm für die Variable mit dem höchsten Median. Färben Sie alle Balken in unterschiedlichen Farben, dabei die Mediankategorie rot. Schreiben Sie den Namen der Variablen an die X-Achse.

Lösung

```
barplot(table(erstis$vertraeg),
        xlab = "Verträglichkeit",
        col = c("yellow","orange","lightgreen",
                "lavender","purple","white","blue","grey","red",
                "green","brown","black","darkgrey") )
```


0
10
20
30
40
1
2.5
3
3.5
4
4.5
5
Verträglichkeit

Zusatzaufgabe: Relativer Informationsgehalt

Suchen Sie aus dem Datensatz die Stimmungs-Variable (`stim1` bis `stim12`) mit den meisten fehlenden Werten. Bestimmen Sie für diese Variable den relativen Informationsgehalt H. Ist Ihrer Meinung nach der relative Informationsgehalt H bestmöglich geeignet, um die Streuung der Merkmale darzustellen?

Lösung

```
set2 <- c("stim1","stim2","stim3","stim4","stim5",
          "stim6","stim7","stim8","stim9","stim10",
          "stim11","stim12")
summary(erstis[,set2])
```

```
     stim1           stim2           stim3          stim4           stim5      
 Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.00   1st Qu.:1.000   1st Qu.:2.000  
 Median :4.000   Median :3.000   Median :3.00   Median :2.000   Median :2.000  
 Mean   :3.537   Mean   :2.777   Mean   :2.73   Mean   :1.794   Mean   :2.598  
 3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:4.00   3rd Qu.:2.000   3rd Qu.:3.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :4.000   Max.   :5.000  
 NA's   :3       NA's   :3       NA's   :2      NA's   :2       NA's   :2      
     stim6           stim7           stim8          stim9           stim10     
 Min.   :1.00   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.00   1st Qu.:2.000   1st Qu.:3.00   1st Qu.:2.000   1st Qu.:3.000  
 Median :3.00   Median :3.000   Median :4.00   Median :2.000   Median :3.000  
 Mean   :3.19   Mean   :3.026   Mean   :3.64   Mean   :2.519   Mean   :3.175  
 3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.:3.000   3rd Qu.:4.000  
 Max.   :5.00   Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000  
 NA's   :2      NA's   :2       NA's   :2      NA's   :2       NA's   :2      
     stim11          stim12     
 Min.   :1.000   Min.   :1.000  
 1st Qu.:1.000   1st Qu.:3.000  
 Median :2.000   Median :3.000  
 Mean   :1.852   Mean   :3.101  
 3rd Qu.:2.000   3rd Qu.:4.000  
 Max.   :5.000   Max.   :5.000  
 NA's   :2       NA's   :2      
```

- `stim1` und `stim2` haben beide die meisten fehlenden Werte (3). Zur Übung bestimmen wir den relativen Informationsgehalt mal für beide Variablen.

Für `stim1` beträgt H

```
h_j <- prop.table(table(erstis$stim1))
-(1/log(length(h_j))) * sum(h_j*log(h_j))
```

```
[1] 0.7611197
```

Für `stim2` beträgt H

```
h_j <- prop.table(table(erstis$stim2))
-(1/log(length(h_j))) * sum(h_j*log(h_j))
```

```
[1] 0.8535155
```

Der relative Informationsgehalt kann für die beiden Stimmungsvariablen berechnet werden. Um das Skalenniveau der Variablen bestmöglich auszunutzen, eigenen sich Quantile jedoch besser. Beim relativen Informationsgehalt erhalten wir lediglich eine Aussage, wie stark die Streuung über die verschiedenen Kategorien ist (homogenes vs. heterogenes Merkmal). Quantile berücksichtigen auch die Ordnung der Kategorien.