Vorbereitung

Bitte führen Sie zur Vorbereitung folgende Schritte aus:

1. Starten Sie RStudio.
2. Löschen Sie den Workspace.
3. Setzen Sie das Arbeitsverzeichnis: Session » Set Working Directory » Choose Directory.
4. Öffnen Sie ein R-Skript und laden Sie den Datensatz `erstis_neu`.
5. Nachdem Sie die Aufgaben bearbeitet haben, speichern Sie das Skript unter einem geeigneten Namen ab.

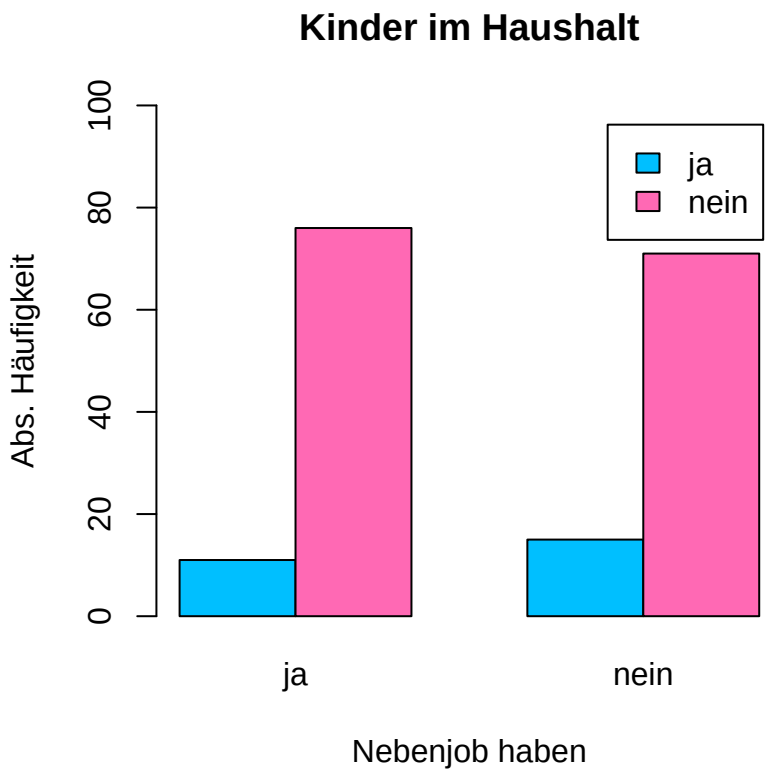
Aufgabe 1

Erstellen Sie ein gruppiertes Säulendiagramm, um die Häufigkeit von befragten Personen mit und ohne Kindern (`kinder`) für die Personen mit und ohne Nebenjob (`job`) zu vergleichen.

(i) Ändern Sie die Farbgebung, sodass die gruppierten Säulen gut unterscheidbar sind. Fügen Sie eine Legende hinzu, sodass die Farbkodierung nachvollziehbar ist.
(ii) Ergänzen Sie einen Titel für Ihre Grafik und eine Beschriftung der x-Achse und der y-Achse.
(iii) Erweitern Sie die Höhe der Grafik, so dass Platz für die Legende ist.
(iv) Speichern Sie das Balkendiagramm in Ihrem Ordner.

Lösung

```
mytab <- table(erstis$kinder, erstis$job)
barplot(mytab, beside = TRUE,
        col = c("deepskyblue", "hotpink"),
        legend = TRUE,
        ylim = c(0, 100),
        main = "Kinder im Haushalt", ylab = "Abs. Häufigkeit", xlab = "Nebenjob haben")
```



- Führen Sie die Grafik einmal mit und einmal ohne das `ylim`-Argument aus. Ohne das Argument überlappen die erstellte Legende und die pinke Säule. Um dem entgegenzuwirken, erweitern wir das angezeigte Intervall auf der y-Achse mit `ylim`.

(v) Erstellen Sie eine Häufigkeitstabelle mit Randsummen. Wie viele Fälle aus dem Datensatz `erstis_neu` wurden ausgeschlossen?

Lösung

```
addmargins(mytab)
```

```
      ja nein Sum
ja    11   15  26
nein  76   71 147
Sum   87   86 173
```

Insgesamt beinhaltet die Tabelle 173 Fälle, d.h. 18 Personen aus dem Datensatz `erstis_neu` wurden aufgrund fehlender Werte ausgeschlossen.

Aufgabe 2

Welche der folgenden Eigenschaften treffen auf den Korrelationskoeffizienten nach Pearson zu?

☐ Bei einer Korrelation von 1 kann man von einer kausalen Beziehung sprechen.
Falsch. Man kann von Korrelation nicht sofort auf Kausalität schließen. Korrelationen können auch nicht-kausale Zusammenhänge charakterisieren.

☒ Für die Berechnung müssen beide Variablen intervallskaliert sein.
Richtig. Beide Variablen müssen mindestens intervallskaliert sein (= metrisch)

- ☒ Der Korrelationskoeffizient ist invariant gegenüber linearen Transformationen.
- ☐ Wenn eine Variable mit sich selbst korreliert wird, ergibt der Korrelationskoeffizient 0.
  Falsch. Er ergibt 1.
- ☐ Der Korrelationskoeffizient nimmt Werte zwischen $-\infty$ und ∞.
  Falsch. Das gilt für die Kovarianz. Der Korrelationskoeffizient kann nur Werte zwischen -1 und 1 annehmen.

Aufgabe 3

Überprüfen Sie, ob es einen Zusammenhang zwischen zufriedener Stimmung(`stim1`) und munterer Stimmung `stim10` gibt. Welches Skalenniveau haben die beiden Variablen? Erstellen Sie dafür zunächst eine Häufigkeitstabelle und dann ein geeignetes Zusammenhangsmaß.

Lösung

```
tabelle <- table(erstis$stim1, erstis$stim10)
tabelle
```

```
   1  2  3  4  5
1  3  0  2  0  0
2  0  5  8  4  0
3  2 16 21 15  0
4  2 16 32 42  4
5  0  0  3 11  2
```

```
library(Hmisc)
rcorr.cens(erstis$stim1, erstis$stim10, outx = TRUE)[2]
```

```
      Dxy
0.4600378
```

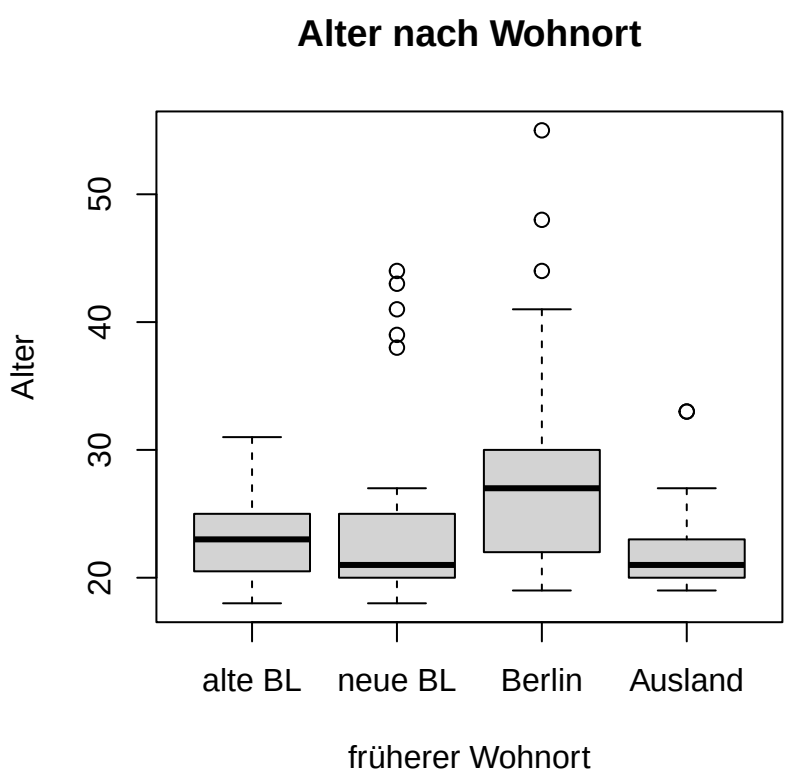
Es handelt sich um zwei kategoriale ordinalskalierte Variablen. Der *gamma*-Koeffizient beträgt 0.46. Es liegen also mehr konkordante Paare im Vergleich zu diskordanten Paare vor, was einen positiven Zusammenhang beschreibt. Personen, deren Stimmung eher zufrieden ist, sind tendenziell auch eher munter.

Aufgabe 4

Sie möchten sich das Alter (`alter`) abhängig vom früheren Wohnort (`wohnort.alt`) anschauen. (i) Visualisieren Sie dies in einem Boxplot-Diagramm. Fügen Sie dem Diagramm einen Titel hinzu und benennen Sie die Achsen adäquat.

Lösung

```
boxplot(alter ~ wohnort.alt, data = erstis,
        main = "Alter nach Wohnort",
        ylab = "Alter",
        xlab = "früherer Wohnort")
```



(ii) Lassen Sie sich die Deskriptivstatistik vom Alter abhängig vom früheren Wohnort ausgeben. Lesen Sie aus der Ausgabe ab, wie groß der Mittelwert und die Standardabweichung für Personen aus Berlin und aus dem Ausland sind. Finden Sie die entsprechenden Mittelwerte auch dargestellt im oberen Boxplot?

```
library(DescTools)
describeBy(erstis$alter, erstis$wohnort.alt)
```

```
 Descriptive statistics by group
group: alte BL
   vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 32 23.22 3.61     23      23 2.97  18  31    13 0.43    -0.82 0.64
------------------------------------------------------------
group: neue BL
   vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 26 24.85 8.38     21   23.73 2.97  18  44    26 1.32     0.07 1.64
------------------------------------------------------------
group: Berlin
   vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 86 27.33 7.05     27   26.53 6.67  19  55    36 1.21     1.94 0.76
------------------------------------------------------------
group: Ausland
   vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 18 22.61 4.31     21   22.19 2.22  19  33    14 1.47     0.91 1.02
```

Der Mittelwert für Personen mit früherem Wohnort Berlin beträgt 27.33 (SD = 7.05). Der Mittelwert für Personen mit früherem Wohnort im Ausland beträgt 22.61 (SD = 4.31). Die Mittelwerte sind nicht im

Boxplot dargestellt, stattdessen finden wir die Mediane.