# Analyse zum Datensatz "Umfrage"

*Martin Hofbauer, Friedrich Leisch*

Der Vorschlag zum Thema der vorliegenden Umfragedaten wurde im Zuge der Lehrveranstaltung "Vertiefung in statistische Methoden der Wildtierforschung" an der Universität für Bodenkultur Wien im Wintersemester 2018/19 von den Studierenden Anja Stefanie Manninger, Vanessa Schüller und Felix Schiestl eingereicht. Der vorliegende Datensatz ist das Ergebnis dieser Umfrage, die von Studenten/Studentinnen der Lehrveranstaltung durchgeführt wurde und sich mit den Gründen für das Fernbleiben von Vorlesungen ohne Anwesenheitspflicht befasst.

*Disclaimer: Dieser Datensatz entstand im Zuge einer Lehrveranstaltung und hatte zum Ziel die Handhabung und Analyse eines selbst erhobenen Datensatzes zu üben. Dieser Datensatz ist in keinster Weise als repräsentativ anzusehen, die Ergebnisse und der Inhalt dieses Dokuments eignen sich somit nicht zur Vewendung in wissenschaftlichen Arbeiten.*

Zunächst lesen wir den vorliegenden Datensatz als csv-file ein. Mit dem hier angeführten Befehl funktioniert das nur, wenn sich die Datei direkt im aktuellen Arbeitsverzeichnis ("Working Directory") befindet. Zum Festlegen des Arbeitsverzeichnisses kann man in R den Befehl `setwd()` verwenden.

```
R> fern <- read.csv(file="VSM_2018W_HUE1_full_data.csv", sep=";")
R> summary(fern)
```

```
 f01a.immer.da f02a.lerne.besser f02b.lva.ueberschneidung
 j   : 46      Min.   :1.000     Min.   :1.00
 n   :416      1st Qu.:2.000     1st Qu.:2.00
 NA's:  5      Median :2.000     Median :2.00
               Mean   :2.144     Mean   :2.47
               3rd Qu.:3.000     3rd Qu.:3.00
               Max.   :4.000     Max.   :4.00
               NA's   :49        NA's   :50
 f02c.privater.konflikt f02d.lehrmethode  f02e.morgen     f02f.mittag
 Min.   :1.000          Min.   :1.000    Min.   :1.000   Min.   :1.000
 1st Qu.:2.000          1st Qu.:2.000    1st Qu.:2.000   1st Qu.:3.000
 Median :2.000          Median :2.000    Median :3.000   Median :4.000
 Mean   :2.511          Mean   :2.242    Mean   :2.667   Mean   :3.289
 3rd Qu.:3.000          3rd Qu.:3.000    3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :5.000          Max.   :5.000    Max.   :5.000   Max.   :5.000
 NA's   :56             NA's   :53       NA's   :50      NA's   :52
   f02g.abend      f02h.beruf      f02i.raum       f03.alter
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
 Median :4.000   Median :3.000   Median :3.000   Median :2.000
 Mean   :3.318   Mean   :2.821   Mean   :3.061   Mean   :2.115
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :4.000
 NA's   :58      NA's   :54      NA's   :56      NA's   :6
 f04.geschlecht  f05a.studium                          f05b.studium.anders
 m   :214       ubrm   :245   Agrar- und Ern<e4>hrungswirtschaft:  4
 w   :244       ktww   : 56   Biologie                          :  2
 NA's:  9       aw     : 49   internationale entwicklung        :  2
                lap    : 33   bwl                               :  1
                fw     : 20   BWL                               :  1
```

```
               (Other): 54   (Other)                          :  9
               NA's   : 10   NA's                             :448
      f06a.wohnform f06b.wohnform.anders  f07.semester    f08a.herkunft
 wg           :245   Eltern:  1          Min.   : 1.000   w      :122
 eltern       : 58   NA's  :466          1st Qu.: 3.000   n      : 93
 partner      : 57                       Median : 5.000   o      : 67
 einzel       : 52                       Mean   : 4.911   anders : 38
 studentenheim: 39                       3rd Qu.: 7.000   stmk   : 31
 (Other)      : 11                       Max.   :14.000   (Other): 98
 NA's         :  5                       NA's   :5        NA's   : 18
  f08b.herkunft.anders
 Deutschland: 16
 deutschland:  6
 suedtirol  :  4
 Suedtirol  :  4
 s<fc>dtirol:  3
 (Other)    : 19
 NA's       :415
```

```
R> dim(fern)
```

```
[1] 467  19
```

```
R> colnames(fern)
```

```
 [1] "f01a.immer.da"            "f02a.lerne.besser"
 [3] "f02b.lva.ueberschneidung" "f02c.privater.konflikt"
 [5] "f02d.lehrmethode"         "f02e.morgen"
 [7] "f02f.mittag"              "f02g.abend"
 [9] "f02h.beruf"               "f02i.raum"
[11] "f03.alter"                "f04.geschlecht"
[13] "f05a.studium"             "f05b.studium.anders"
[15] "f06a.wohnform"            "f06b.wohnform.anders"
[17] "f07.semester"             "f08a.herkunft"
[19] "f08b.herkunft.anders"
```

Wie aus der Zusammenfassung ersichtlich ist, handelt es sich um 467 Beobachtungen in 19 Variablen. Wir interessieren uns für die Beobachtungen, die Angaben in Frage 2 enthalten. Dazu erstellen wir einen neuen Datensatz `fern2`, der lediglich die Beobachtungen enthält, die Antworten in Frage 2 (also Werte $\neq$ `NA`) enthalten.

```
R> ok <- complete.cases(fern[,2:10])
R> fern2 <- fern[ok,]
R> dim(fern2)
```

```
[1] 390  19
```

```
R> summary(fern2)
```

```
 f01a.immer.da f02a.lerne.besser f02b.lva.ueberschneidung
 j:  2         Min.   :1.000     Min.   :1.000
```

```
n:388         1st Qu.:2.000     1st Qu.:2.000
              Median :2.000     Median :2.000
              Mean   :2.141     Mean   :2.477
              3rd Qu.:3.000     3rd Qu.:3.000
              Max.   :4.000     Max.   :4.000

f02c.privater.konflikt f02d.lehrmethode  f02e.morgen     f02f.mittag
Min.   :1.000          Min.   :1.000    Min.   :1.000   Min.   :1.000
1st Qu.:2.000          1st Qu.:2.000    1st Qu.:2.000   1st Qu.:3.000
Median :2.000          Median :2.000    Median :3.000   Median :4.000
Mean   :2.515          Mean   :2.228    Mean   :2.685   Mean   :3.285
3rd Qu.:3.000          3rd Qu.:3.000    3rd Qu.:4.000   3rd Qu.:4.000
Max.   :5.000          Max.   :5.000    Max.   :5.000   Max.   :5.000

  f02g.abend      f02h.beruf      f02i.raum       f03.alter
Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
Median :4.000   Median :3.000   Median :3.000   Median :2.000
Mean   :3.364   Mean   :2.859   Mean   :3.077   Mean   :2.141
3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000
Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :4.000
                                                NA's   :1
f04.geschlecht  f05a.studium                          f05b.studium.anders
m   :182        ubrm   :215   Agrar- und Ern<e4>hrungswirtschaft:  4
w   :205        ktww   : 46   internationale entwicklung        :  2
NA's:  3        aw     : 35   Biologie                         :  1
                lap    : 28   BWL                              :  1
                fw     : 18   Geographie                       :  1
                (Other): 45   (Other)                          :  7
                NA's   :  3   NA's                             :374
      f06a.wohnform f06b.wohnform.anders  f07.semester    f08a.herkunft
wg           :217   Eltern:  0           Min.   : 1.000   w      :103
partner      : 49   NA's  :390           1st Qu.: 3.000   n      : 81
einzel       : 42                        Median : 5.000   o      : 58
eltern       : 40                        Mean   : 5.067   anders : 29
studentenheim: 33                        3rd Qu.: 7.000   stmk   : 26
einzeln      :  3                        Max.   :10.000   (Other): 83
(Other)      :  6                                         NA's   : 10
 f08b.herkunft.anders
Deutschland: 10
deutschland:  6
Suedtirol  :  4
suedtirol  :  3
Italien    :  2
(Other)    : 15
NA's       :350
```

Eigentlich sollten sich in dem neuen Datensatz nun keine Beobachtungen mehr finden, die bei Frage 1 "Ja" gewählt haben. Dies ist vermutlich auf einen Eingabefehler bei Frage 1 für diese zwei Beobachtungen zurückzuführen. Wir belassen die Beobachtungen im Datensatz, da die benötigten Antworten vorhanden sind. Für uns sind lediglich die Antworten auf Frage 2 von Interesse, dies sind im Datensatz die Werte der Spalten 2 bis 10. Diese fassen wir zur leichteren Handhabung im Vektor `f2` zusammen. Einen ersten Überblick über die Daten verschaffen wir uns, indem wir die Mittelwerte (in aufsteigender Reihenfolge sortiert) und Balkendiagramme der Antworten auf einzelne Fragen betrachten. Dabei muss vor allem die Codierung der

Antwortmöglichkeiten beachtet werden. Im vorliegenden Fall sind die Antwortmöglichkeiten von 1 (“trifft völlig zu”) bis 4 (“trifft gar nicht zu”) codiert. Die Antwortmöglichkeit “keine Angabe” (codiert als NA) ist im vorliegenden Fall nicht möglich, da Beobachtungen mit fehlenden Antworten bereits entfernt wurden.

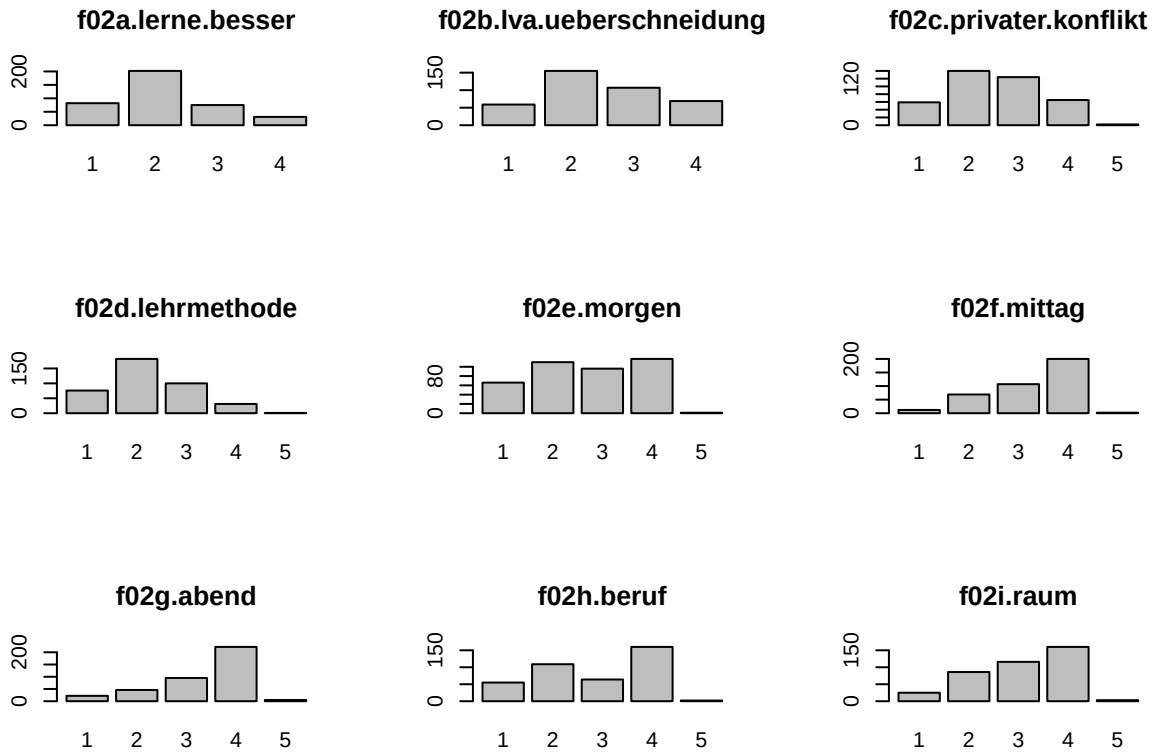
```
R> f2 <- colnames(fern2)[2:10]
R> f2
```

```
[1] "f02a.lerne.besser"         "f02b.lva.ueberschneidung"
[3] "f02c.privater.konflikt"    "f02d.lehrmethode"
[5] "f02e.morgen"               "f02f.mittag"
[7] "f02g.abend"                "f02h.beruf"
[9] "f02i.raum"
```

```
R> sort(colMeans(fern2[,f2]))
```

```
      f02a.lerne.besser         f02d.lehrmethode f02b.lva.ueberschneidung
               2.141026                 2.228205                 2.476923
 f02c.privater.konflikt              f02e.morgen               f02h.beruf
               2.515385                 2.684615                 2.858974
              f02i.raum              f02f.mittag               f02g.abend
               3.076923                 3.284615                 3.364103
```

```
R> par(mfrow=c(3,3))
R> for(n in f2) barplot(table(fern2[,n]), main=n)
```



```
R> par(mfrow=c(1,1))
```

Hohe Werte des arithmetischen Mittels zeigen hier eine Tendenz zur Ablehnung der beschriebenen Begründung. Es ist zu beachten, dass ein Wert von 2.5 keinerlei Tendenz, also im Schnitt keiner Zu- oder Ablehnung entspricht. Die Betrachtung der Balkendiagramme liefert ähnliche Erkenntnisse über die tendenzielle Zu- oder Ablehnung einzelner Gründe, bietet darüber hinaus jedoch auch einen ersten Überblick über die Streuung der Antworten.

Zur weiteren Analyse der Daten führen wir nun eine Hauptkomponentenanalyse (PCA - "Principal Component Analysis") durch.

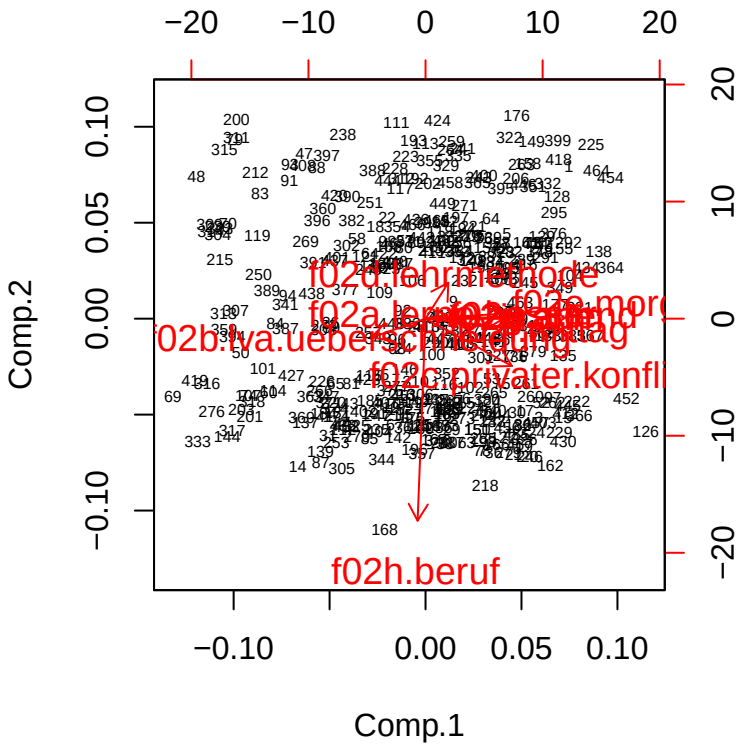
```
R> fern.pca <- princomp(fern2[,f2])
R> summary(fern.pca)
```

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
Standard deviation     1.3131537 1.1461051 1.0282911 0.9273230 0.9078637
Proportion of Variance 0.2107188 0.1605170 0.1292124 0.1050834 0.1007195
Cumulative Proportion  0.2107188 0.3712358 0.5004483 0.6055317 0.7062512
                           Comp.6    Comp.7     Comp.8     Comp.9
Standard deviation     0.83878441 0.8147907 0.74532519 0.69345378
Proportion of Variance 0.08597514 0.0811268 0.06788343 0.05876344
Cumulative Proportion  0.79222632 0.8733531 0.94123656 1.00000000
```

In dieser ersten Zusammenfassung der Hauptkomponentenanalyse können wir ablesen welcher Anteil der Varianz durch die verschiedenen Hauptkomponenten abgedeckt wird. Beispielsweise wird durch die ersten zwei Hauptkomponenten bereits 37% der Varianz im Datensatz abgedeckt.

Als nächstes wollen wir analysieren, was die ersten beiden Hauptkomponenten hauptsächlich repräsentieren. Dazu betrachten wir einen Plot des Datensatzes, wobei auf den Achsen die beiden Hauptkomponenten der Beobachtung aufgetragen werden. R bietet hierfür die Funktion `biplot()`, der Parameter `cex` ist für die Schriftgröße der Datenbeschriftungen im Plot verantwortlich.

```
R> biplot(fern.pca, cex=c(0.5, 1.2))
```




Wir erkennen, dass die Variable `f02h.beruf` fast ausschließlich zur zweiten Hauptkomponente `Comp.2` zuzuordnen ist, da sie hauptsächlich in vertikale Richtung zeigt. Die erste Hauptkomponente `Comp.1` scheint hingegen aus den Variablen `f02b.lva.ueberschneidung`, `f02c.privater.konflikt`, `f02e.morgen`, `f02f.mittag`, `f02g.abend` und `f02i.raum` zu bestehen. Diese Vermutung können wir auch durch Betrachtung der "loadings" bestätigen:

```
R> fern.pca$loadings[,1:2]
```

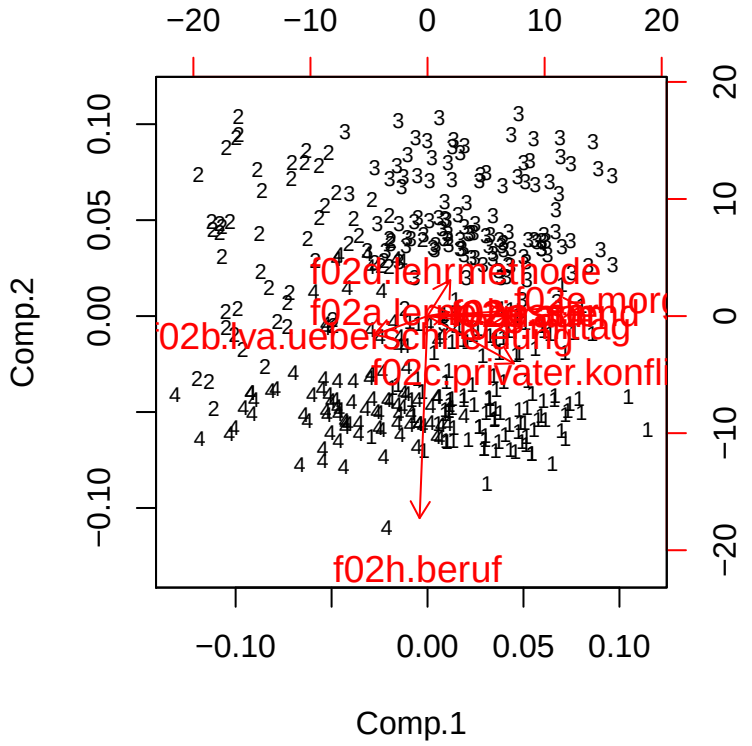
```
                          Comp.1       Comp.2
f02a.lerne.besser      0.08911636  0.017517495
f02b.lva.ueberschneidung -0.21925863 -0.087933899
f02c.privater.konflikt    0.35554946 -0.223484499
f02d.lehrmethode          0.09368021  0.169370295
f02e.morgen               0.62938821  0.046302650
f02f.mittag               0.38590265 -0.050449732
f02g.abend                0.39141504 -0.001161632
f02h.beruf               -0.03267392 -0.953198086
f02i.raum                 0.33086074  0.007217828
```

Wir erkennen vergleichsweise hohe Werte (betragsmäßig) in der ersten bzw. zweiten Hauptkomponente für die oben erwähnten Variablen. Insgesamt könnte man die erste Hauptkomponente in etwa als "Fernbleiben aus zeitlichen Gründen" und die zweite als "Fernbleiben aus beruflichen Gründen" zusammenfassen.

Als nächstes wollen wir den Datensatz in 4 Gruppen ("Cluster") unterteilen. Um Gruppen zu erhalten, in denen die Daten möglichst "nah" (im Sinne des quadrierten Abstandes) zum Gruppenmittel liegen, verwenden wir die Funktion `kmeans()`. Da es sich um eine stochastische Funktion handelt, setzen wir davor

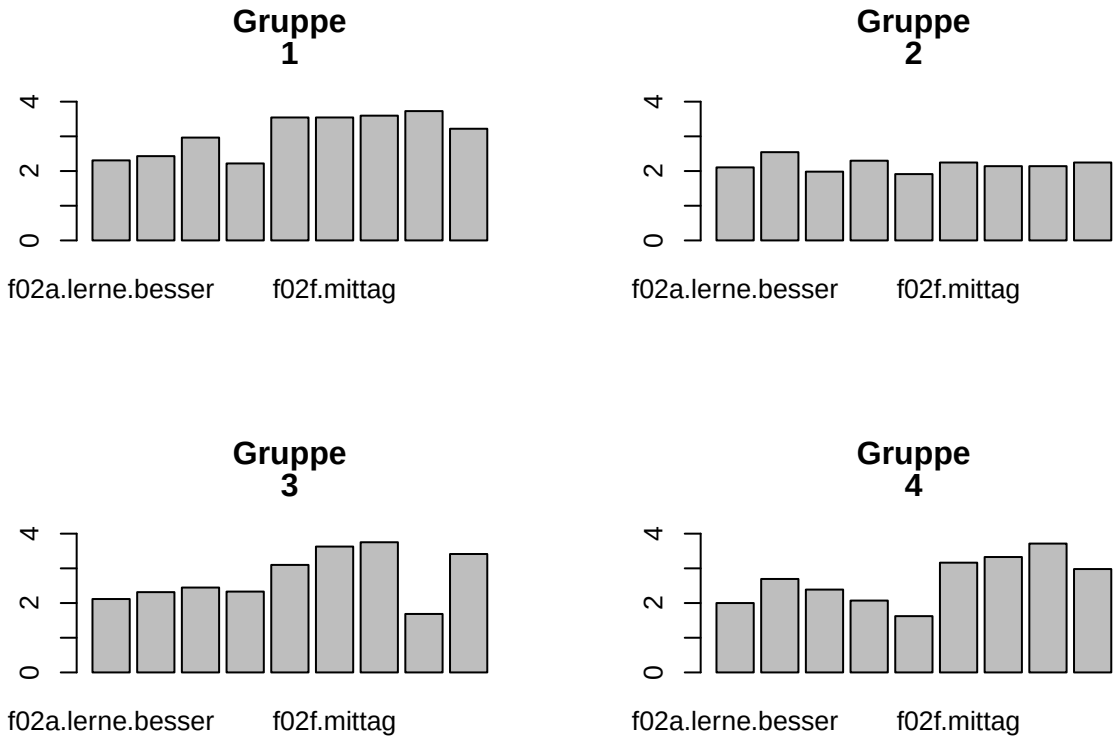
den “seed” neu damit unsere Analyse leicht reproduzierbar bleibt. Anschließend betrachten wir wieder den Biplot der PCA, beschriften die Beobachtungen jedoch nicht nach ihrer Nummerierung, sondern nach ihrer Gruppenzugehörigkeit.

```
R> set.seed(1234)
R> fern.k4 <- kmeans(fern2[,f2], 4, iter.max=100, nstart=10)
R> biplot(fern.pca, xlabs=fern.k4$cluster, cex=c(0.7,1.2))
```




Hier zeigt sich ein typisches Bild bei der Analyse von Umfragedaten: die vier Quadranten des Koordinatensystems bilden in etwa die 4 Gruppen. Die Gruppeneinteilung erfolgte im Groben also nach den beiden ersten Hauptkomponenten. Diese Eigenschaft lässt sich auch in den Balkendiagrammen der vier Gruppen erkennen:

```
R> par(mfrow=c(2,2))
R> for(n in 1:4) barplot(fern.k4$center[n,], ylim=c(0,4), main=c("Gruppe", n))
```



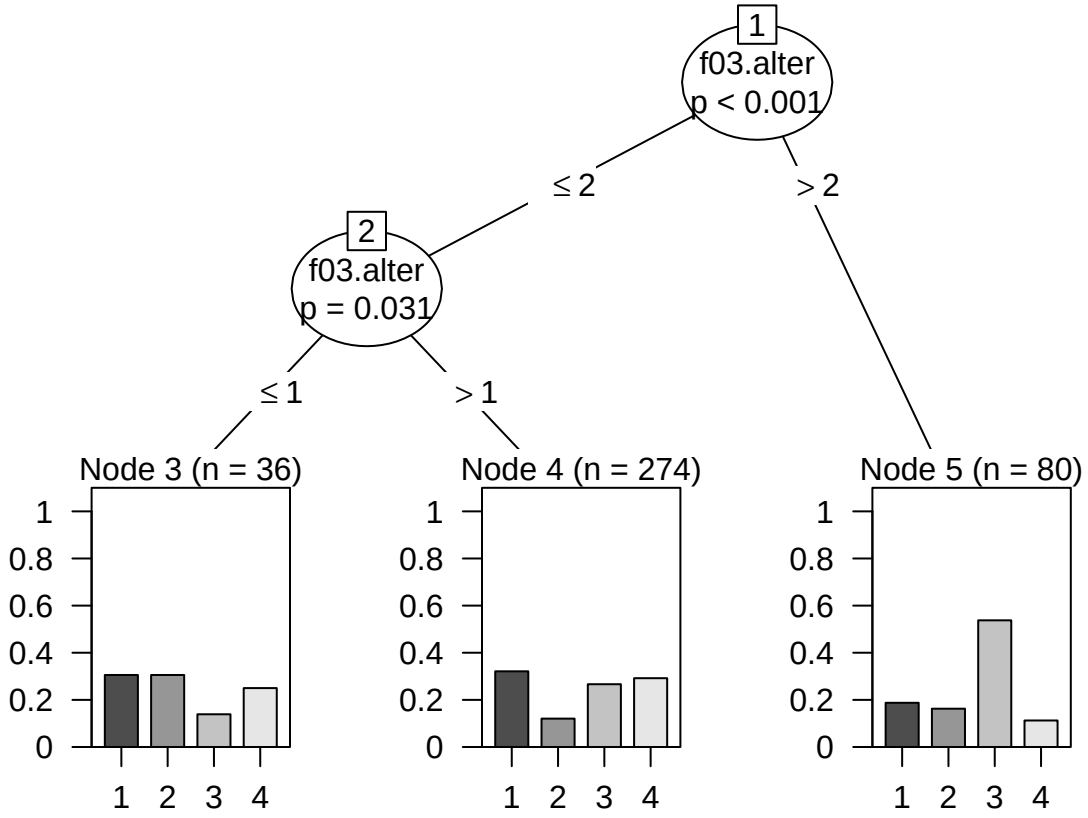

```
R> par(mfrow=c(1,1))
```

So zeigt sich beispielsweise in Gruppe 1 und Gruppe 4 eine Tendenz zu hohen Werten in der achten (vorletzten) Variable `f02h.beruf`. Im Vergleich dazu ist dieser Wert für die anderen beiden Gruppen, vor allem für Gruppe 3, wesentlich niedriger.

Es bleibt nun noch zu klären "wer" in welche Gruppe fällt, also welche Eigenschaften (Alter & Geschlecht) Befragte haben, die in die jeweiligen Gruppen fallen. Wir konzentrieren uns dabei auf die Unterscheidung nach Alter und Geschlecht. Dazu kommt die Funktion `ctree()` aus dem Paket `party` zum Einsatz, welches vorher geladen werden muss.

```
R> library(party)
```

```
R> fern.baum <- ctree(as.factor(fern.k4$cluster)~f03.alter+f04.geschlecht, data=fern2)
R> plot(fern.baum)
```




Wir erkennen, dass der Geschlechterunterschied für keine der Gruppen signifikant zu sein scheint, da diese Kategorisierung im Baumdiagramm nicht auftaucht. Die Unterscheidung nach dem Alter erfolgt in zwei Stufen. Wir erinnern uns hierfür noch einmal an die Codierung der Variable `f03.alter`:

Table 1: Codierung von Frage 3: Alter

| Alter | Codierung |
|---|---|
| <20 | 1 |
| 20 - 25 | 2 |
| 26 - 30 | 3 |
| Älter als 30 | 4 |
| keine Angabe | NA |

Die erste Unterteilung im Baumdiagramm entspricht also einer Teilung in Personen mit einem Alter $\leq 25$ und $> 25$. Die über 25 jährigen Personen fallen dabei in einen eigenen Knoten, der hauptsächlich aus Beobachtungen aus Gruppe 3 besteht. Weiter oben haben wir bereits analysiert, dass Befragte in Gruppe 3 einen auffällig niedrigen Wert der Variable `f02h.beruf` aufweisen. Dabei ist wieder zu beachten, dass niedrige Werte in dieser Variable einer hohen Zustimmung zur Aussage "Berufstätigkeit hindert mich am Besuch der Lehrveranstaltung" entsprechen. Das scheint plausibel, da ältere Studierende vermutlich eher berufstätig sind und dadurch auch die Gründe für das Fernbleiben von Vorlesungen gegeben sind. Für unter 25 jährige Personen sieht das Baumdiagramm eine weitere (feinere) Unterteilung in unter und über 20 Jährige vor. Hier ist wieder zu beobachten, dass Berufstätigkeit als Begründung für das Fernbleiben von Vorlesungen weiter an Wichtigkeit verliert je jünger der/die Befragte ist.