5 Multivariate Deskriptivstatistik

5.1 Mathematische Grundlagen

Für diese und die folgenden Einheiten gehen wir davon aus, dass sich die Realisierungen $y_1, \dots, y_n \in \mathbb{R}^m$ eines Zufallsvektors $v \in \mathbb{R}^m$ zu einer $m \times n$ Matrix Y zusammenfassen lassen, die wir gewöhnlich als *Datenmatrix* bezeichnen:

$$Y = \begin{pmatrix} y_1 & \dots & y_n \end{pmatrix} \in \mathbb{R}^{m \times n} \, . \tag{1}$$

Bitte beachten Sie, dass sich dieses Format – die Anordnung von Datenpunkten in Spalten – von dem gewöhnlich in `CSV`-Dateien oder etwa `R`-Dataframes verwendeten Format – der Anordnung von Datenpunkten in Zeilen – unterscheidet. Die Ursache hierfür liegt darin, dass gemäß unserer Theorie in den Einheiten (2) *Matrizen* und (4) *Multivariate Normalverteilungen* Vektoren und Zufallsvektoren grundsätzlich Spaltenvektoren sind, sodass sie hier horizontal bzw. nebeneinander angeordnet werden.

In der Vorlesung haben wir drei *multivariate Deskriptivstatistiken* für eine derartige Sammlung von Realisierungen kennengelernt, deren Definition wir im Folgenden wiederholen.

Stichprobenmittel

$$\bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2}$$

Stichprobenkovarianzmatrix

$$C := \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}} \tag{3}$$

Stichprobenkorrelationsmatrix

$$R := \left(\frac{(C)_{ij}}{\sqrt{(C)_{ii}} \sqrt{(C)_{jj}}} \right)_{1 \leq i,j \leq m} \tag{4}$$

Darüber hinaus wurde in der Vorlesung gezeigt, dass diese Deskriptivstatistiken sich als Matrixprodukte auf Grundlage der Datenmatrix wie folgt darstellen lassen.

Stichprobenmittel

$$\bar{y} = \frac{1}{n} Y 1_n \tag{5}$$

Stichprobenkovarianzmatrix

$$C = \frac{1}{n-1} \left(Y \left(I_n - \frac{1}{n} 1_{nn} \right) Y^{\mathrm{T}} \right) \tag{6}$$

Stichprobenkorrelationsmatrix

$$R = DCD \quad \text{mit} \quad D := \operatorname{diag}\left(\frac{1}{\sqrt{(C)_{11}}}, \dots, \frac{1}{\sqrt{(C)_{mm}}} \right) \tag{7}$$

5.2 Analyse in R

Die Datei `FADE_SAME.csv` enthält den in der ersten Seminarsitzung vorgestellten Datensatz. Erklären Sie die Funktion des folgenden R-Codes:

```r
# Daten einlesen
fname = 'FADE_SAME.csv'                              # Dateiname
D     = read.csv(fname)                              # Dataframe

# Variablen extrahieren
vars  = c('novelty.FADE','novelty.SAME',             # Variablen
          'memory.FADE', 'memory.SAME')
n     = nrow(D)                                      # Anzahl Datenpunkte
m     = length(vars)                                 # Anzahl Variablen
Y     = matrix(rep(0,n*m), nrow = n)                 # zeilenweise Datenmatrix
for (j in 1:m) {
    Y[,j] = D[,vars[j]]                              # j-te Variable
}

# Datenmatrix transponieren
Y     = t(Y)                                         # spaltenweise Datenmatrix
print(dim(Y))                                        # Überprüfung Datenmatrix
```

5.3 Erste Programmieraufgabe

Berechnen Sie nun die multivariaten Deskriptivstatistiken für diesen Datensatz anhand der oben beschriebenen Formeln. Gehen Sie dazu wie folgt vor:

- Erzeugen Sie mit `diag(n)` die n-dimensionale Einheitsmatrix I_n.
- Erzeugen Sie mit `matrix()` eine $n \times n$ Einsmatrix J_n.
- Berechnen Sie das Stichprobenmittel $\bar{y}$ anhand der obigen Formel, d.h. indem Sie Y mit der ersten Spalte von J_n multiplizieren und das Produkt mit $1/n$ malnehmen.
- Berechnen Sie die Stichprobenkovarianzmatrix C anhand der obigen Formel, d.h. indem Sie das Matrixprodukt $Y\left(I_n - \frac{1}{n}J_n\right)Y^{\mathrm{T}}$ bestimmen und es mit $1/(n-1)$ multiplizieren.
- Extrahieren Sie mit `diag()` die Diagonaleinträge von C, nehmen Sie das Reziproke von deren Quadratwurzeln via `1/sqrt()` und erzeugen Sie mit `diag()` aus den resultierenden Werten wieder eine Diagonalmatrix, die wir D nennen.
- Berechnen Sie die Stichprobenkorrelationsmatrix R anhand der obigen Formel, d.h. indem Sie das Matrixprodukt DCD bestimmen.
- Geben Sie die Resultate ihrer Analyse aus. Sie sollten folgende Ergebnisse erhalten:

```
          [,1]
[1,] -1.8163838
[2,] -0.1993077
[3,] -1.1176979
[4,] -0.5112862
```

```
           [,1]        [,2]        [,3]        [,4]
[1,]  0.41849409 -0.28080692  0.01752792  0.06186873
[2,] -0.28080692  0.43955708 -0.05147251  0.08905219
[3,]  0.01752792 -0.05147251  0.33872503 -0.41351203
[4,]  0.06186873  0.08905219 -0.41351203  0.79214018
```

```
           [,1]       [,2]        [,3]       [,4]
[1,]  1.00000000 -0.6547198  0.04655453  0.1074547
[2,] -0.65471985  1.0000000 -0.13339639  0.1509162
[3,]  0.04655453 -0.1333964  1.00000000 -0.7982951
[4,]  0.10745473  0.1509162 -0.79829513  1.0000000
```

Überprüfen Sie nun die multivariaten Deskriptivstatistiken für diesen Datensatz anhand der Berechnung auf Grundlage von R-Funktionen. Gehen Sie dazu wie folgt vor:

- Berechnen Sie (i) das Stichprobenmittel $\bar{y}$ mit `rowMeans()` als Zeilenmittel von Y (alternativ: mit `colMeans()` als Spaltenmittel von Y^{T}), (ii) die Stichprobenkovarianzmatrix C mit `cov()` aus Y^{T} und (iii) die Stichprobenkorrelationsmatrix R mit `cor()` aus Y^{T}.
- Geben Sie die Resultate ihrer Analyse aus. Die Ergebnisse sollten identisch sein.

```
          [,1]
[1,] -1.8163838
[2,] -0.1993077
[3,] -1.1176979
[4,] -0.5112862
```

```
            [,1]        [,2]        [,3]        [,4]
[1,]  0.41849409 -0.28080692  0.01752792  0.06186873
[2,] -0.28080692  0.43955708 -0.05147251  0.08905219
[3,]  0.01752792 -0.05147251  0.33872503 -0.41351203
[4,]  0.06186873  0.08905219 -0.41351203  0.79214018
```

```
            [,1]       [,2]        [,3]       [,4]
[1,]  1.00000000 -0.6547198  0.04655453  0.1074547
[2,] -0.65471985  1.0000000 -0.13339639  0.1509162
[3,]  0.04655453 -0.1333964  1.00000000 -0.7982951
[4,]  0.10745473  0.1509162 -0.79829513  1.0000000
```

5.4 Analyse in R

Im zweiten Teil wollen wir dieselben Deskriptivstatistiken noch einmal anhand anderer Variablen aus dem Datensatz `FADE_SAME.csv` berechnen. Erklären Sie dazu zunächst den folgenden R-Code und die daraus resultierende Datenmatrix:

```r
# Daten einlesen
fname = 'FADE_SAME.csv'                              # Dateiname
D     = read.csv(fname)                              # Dataframe

# Variablen extrahieren
vars  = c('age', 'memory', 'MWT.B')                  # Definition Variablen
Y     = D[,vars]                                     # zeilenweise Datenmatrix
Y     = t(Y)                                         # spaltenweise Datenmatrix
print(dim(Y))                                        # Überprüfung Datenmatrix
```

5.5 Zweite Programmieraufgabe

Die Variable "MWT-B" (Mehrfachwahl-Wortschatz-Intelligenztest, Variante B) ist nicht für alle Datenpunkte vorhanden. Wir werden daher zunächst nur die Datenpunkte auswählen, für die der MWT-B gemessen wurde, bevor wir die Deskriptivstatistiken für diese Datenmatrix berechnen. Gehen Sie dazu wie folgt vor:

- Für Datenpunkte, deren MWT-B fehlt, hat die dritte Zeile der Datenmatrix den Wert "NA" (not available). Der Ausdruck `!is.na(Y[3,])` ist ein Vektor von der Länge der Datenmatrix, der `TRUE` ist, wo der MWT-B vorhanden und demzufolge nicht "NA" ist. Nutzen Sie diesen Vektor, um die Spalten der Datenmatrix `Y` zu indizieren und auf diese Weise nur die Werte mit vorhandenem MWT-B auszuwählen. Speichern Sie das Ergebnis in die Variable `Y`.
- Berechnen Sie die Anzahl der Datenpunkte n und die Anzahl der Variablen m als Anzahl der Spalten bzw. Zeilen der so neu erzeugten Datenmatrix Y. Geben Sie diese Werte aus: n sollte 258 sein.
- Ermitteln Sie die multivariaten Deskriptivstatistiken wie im Abschnitt 5.3. Es genügt hier, wenn Sie die Deskriptivstatistiken mittels der oben beschriebenen Formeln als Matrixprodukte berechnen.
- Geben Sie die Resultate ihrer Analyse aus. Sie sollten folgende Ergebnisse erhalten:

```
[1]   3 258


             [,1]
age    47.6666667
memory  0.7921368
MWT.B  28.8837209


               age       memory       MWT.B
age    421.2503243 -0.553940597 39.28404669
memory  -0.5539406  0.005936424 -0.01683189
MWT.B   39.2840467 -0.016831889 13.21599855


           [,1]        [,2]        [,3]
[1,]  1.0000000 -0.35029233  0.52649725
[2,] -0.3502923  1.00000000 -0.06009252
[3,]  0.5264972 -0.06009252  1.00000000
```

5.6 Lückentext

Füllen Sie mit den in der Übung gewonnenen Erkenntnissen den folgenden Lückentext aus und präsentieren Sie die Ergebnisse im Seminar:

Lückentext: Die Korrelation der beiden Novelty-Scores beträgt _____ und die Korrelation der beiden Memory-Scores beträgt _____. Die Korrelation der beiden FADE-Scores beträgt _____ und die Korrelation der beiden SAME-Scores beträgt _____. Die Scores sind demzufolge innerhalb eines fMRT-Kontrasts _____ (positiv/negativ) korreliert und innerhalb eines Score-Typs _____ (positiv/negativ) korreliert. In Absolutwerten sind Korrelationen der Scores innerhalb eines Kontrasts _____ (stärker/schwächer) als zwischen Kontrasten. Mit Blick auf die demographischen Variablen ist die Korrelation zwischen Personenalter und Gedächtnisleistung _____ (positiv/negativ), weil _____; und die Korrelation zwischen Personenalter und MWT-B _____ (positiv/negativ), weil _____.

5.7 Mögliche Klausurfrage

Präsentieren Sie im Seminar folgende Klausurfrage und erklären Sie die richtige Antwort:

Frage: Gegeben sei ein Datensatz $y_1, \dots, y_n$ mit $y_i \in \mathbb{R}^m$ für $i = 1, \dots, n$ sowie dessen multivariates Stichprobenmittel $\bar{y}$. Wie ist die Stichprobenkovarianzmatrix der $y_1, \dots, y_n$ definiert?

a) $C := \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}}$

b) $C := \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^{\mathrm{T}}(y_i - \bar{y})$

c) $C := \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}}$

d) $C := \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^{\mathrm{T}}(y_i - \bar{y})$

5.8 Kinderwitz

Was sagt der Pirat auf dem Bauernhof?

„Heu! ... Ah!“ :trowtnA