Multivariates Schätzen und Testen

Aufgabe 1: Kovarianzmatrix

In einer Stichprobe der Größe $n = 3$ erheben Sie zwei Merkmale X_1 und X_2 und erhalten folgende Datenmatrix:

$$\mathbf{X} = \begin{pmatrix} 1 & -4 \\ 0 & 2 \\ 2 & -1 \end{pmatrix}.$$

Schätzen Sie die Kovarianzmatrix von $\boldsymbol{X} = (X_1, X_2)$.

Lösung:

Allgemein: $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \dots \\ \boldsymbol{x}_n^\top \end{pmatrix}$, $\quad n = 3, p = 2, x_i \in \mathbb{R}^p$ für $i = 1, \dots, n$.

- Mittelwertsvektor $\bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_{.1} \\ \bar{x}_{.2} \end{pmatrix}$ mit $\bar{x}_{.j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$

$$\bar{x}_{.1} = \frac{1+2}{3} = 1; \quad \bar{x}_{.2} = \frac{-4+2-1}{3} = -1 \quad \Rightarrow \bar{\boldsymbol{x}} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

- Schätzung von $Cov(\boldsymbol{x})$: $\widehat{Cov}(\boldsymbol{x}) = \boldsymbol{S}$

$$
\begin{aligned}
\boldsymbol{S} &= \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top\right) \\
&= \frac{1}{2}\left\{\begin{pmatrix}1 & -4\\ -4 & 16\end{pmatrix} + \begin{pmatrix}0 & 0\\ 0 & 4\end{pmatrix} + \begin{pmatrix}4 & -2\\ -2 & 1\end{pmatrix} - 3\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\right\} \\
&= \frac{1}{2}\begin{pmatrix}2 & -3\\ -3 & 18\end{pmatrix} = \begin{pmatrix}1 & -1.5\\ -1.5 & 9\end{pmatrix}
\end{aligned}
$$

- Alternative: Verwende Zentrierungsmatrix $\boldsymbol{H}$:

$$
\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top = \begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix} - \frac{1}{3}\begin{pmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{pmatrix}
$$

$$
\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^\top\boldsymbol{H}\boldsymbol{X} = \ldots = \begin{pmatrix}1 & -1.5\\ -1.5 & 9\end{pmatrix}
$$

Aufgabe 2: Multivariate Tests I

In einer Studie (Seal (1964), S. 106) wurde die Länge (x_1) und Breite (x_2) des Schädels bei 35 erwachsenen weiblichen Fröschen untersucht. Es ergaben sich folgende empirische Größen:

$$\bar{\boldsymbol{x}}_w = \begin{bmatrix} 22.860 \\ 24.397 \end{bmatrix}, \quad \boldsymbol{S}_w = \begin{bmatrix} 17.683 & 20.290 \\ & 24.407 \end{bmatrix}.$$

a) Testen Sie, unter der Annahme $(x_1, x_2)^T \sim N_2(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ mit

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 20 \\ 25 \end{bmatrix} \quad \boldsymbol{\Sigma}_0 = \begin{bmatrix} 20 & 0 \\ 0 & 25 \end{bmatrix}$$

die Hypothese $H_0: \boldsymbol{\mu}_w = \boldsymbol{\mu}_0$ auf einem Signifikanzniveau von $\alpha = 0.05$.

b) Testen Sie, ob $\boldsymbol{\mu}_w = \boldsymbol{\mu}_0$ bei unbekannter Kovarianzmatrix ($\alpha = 0.05$).

c) Äquivalente Messungen wurden bei 14 erwachsenen männlichen Fröschen durchgeführt, mit folgenden Ergebnissen:

$$\bar{\boldsymbol{x}}_m = \begin{bmatrix} 21.821 \\ 22.843 \end{bmatrix}, \quad \boldsymbol{S}_m = \begin{bmatrix} 18.479 & 19.095 \\ & 19.273 \end{bmatrix}.$$

Überprüfen Sie unter der Annahme identischer Kovarianzmatrizen in beiden Subpopulationen, ob $\boldsymbol{\mu}_w = \boldsymbol{\mu}_m$ gilt ($\alpha = 0.05$).

Lösung:

a) Das zu testende Hypothesenpaar lautet

$$H_0 : \boldsymbol{\mu}_w = \boldsymbol{\mu}_0 \ , \ H_1 : \boldsymbol{\mu}_w \neq \boldsymbol{\mu}_0$$

Da die Kovarianzmatrix als bekannt vorausgesetzt wird, verwenden wir folgende Teststatistik

$$T^2 = n_w(\bar{\boldsymbol{x}}_w - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}_w - \boldsymbol{\mu}_0).$$

Es gilt

$$T^2 \overset{H_0}{\sim} \chi^2(p).$$

Testentscheidung: Wir lehnen H_0 ab, falls $T^2 > \chi^2_{1-\alpha}(p)$. Wir erhalten

$$\bar{\boldsymbol{x}}_w - \boldsymbol{\mu}_0 = \begin{bmatrix} 22.860 \\ 24.397 \end{bmatrix} - \begin{bmatrix} 20 \\ 25 \end{bmatrix} = \begin{bmatrix} 2.860 \\ -0.603 \end{bmatrix}$$

sowie

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} 0.05 & 0 \\ 0 & 0.04 \end{bmatrix}$$

Der Stichprobenumfang beträgt $n_w = 35$ und $p = 2$, da wir zwei Merkmale betrachten.

Damit erhalten wir

$$\begin{aligned} T^2 &= 35 \begin{bmatrix} 2.860 & -0.603 \end{bmatrix} \begin{bmatrix} 0.05 & 0 \\ 0 & 0.04 \end{bmatrix} \begin{bmatrix} 2.860 \\ -0.603 \end{bmatrix} \\ &= 14.82. \end{aligned}$$

Die Realisation unserer Teststatistik vergleichen wir mit dem 0.95 Quantil der χ^2-Verteilung mit 2 Freiheitsgraden:

$$T^2 = 14.82 > 5.99 = \chi^2_{0.95}(2).$$

Ergebnis: Der hypothetische Erwartungswertvektor $\boldsymbol{\mu}_0$ liegt außerhalb des 0.95-Konfidenzintervall. H_0 wird daher abgelehnt.

b) Das zu testende Hypothesenpaar lautet:

$$H_0 : \boldsymbol{\mu}_w = \boldsymbol{\mu}_0 \ , \ H_1 : \boldsymbol{\mu}_w \neq \boldsymbol{\mu}_0$$

Da die wahre Kovarianzmatrix als unbekannt vorausgesetzt wird, müssen wir sie mit Hilfe der Stichprobe schätzen und folgende Teststatistik (Hotelling Einstichproben T^2-Test) verwenden

$$T^2 = \frac{n_w(n_w - p)}{p(n_w - 1)}(\bar{\boldsymbol{x}}_w - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_w - \boldsymbol{\mu}_0).$$

Es gilt

$$T^2 \overset{H_0}{\sim} F(p, n_w - p).$$

H_0 wird abgelehnt, wenn gilt:

$$T^2 > F(p, n_w - p; 1 - \alpha).$$

Wir erhalten

$$\bar{\boldsymbol{x}}_w - \boldsymbol{\mu}_0 = \begin{bmatrix} 22.860 \\ 24.397 \end{bmatrix} - \begin{bmatrix} 20 \\ 25 \end{bmatrix} = \begin{bmatrix} 2.860 \\ -0.603 \end{bmatrix}$$

sowie als Inverse von $\boldsymbol{S}_1$

$$\boldsymbol{S}_w^{-1} = \frac{1}{17.683 \cdot 24.407 - 20.290^2} \begin{bmatrix} 24.407 & -20.290 \\ -20.290 & 17.683 \end{bmatrix} = \begin{bmatrix} 1.2262 & -1.0193 \\ -1.0193 & 0.8884 \end{bmatrix}.$$

Mit $p = 2$ und $n_w = 35$ erhalten wir

$$\begin{aligned} T^2 &= \frac{35 \cdot 33}{68} \begin{bmatrix} 2.860 & -0.603 \end{bmatrix} \begin{bmatrix} 1.2262 & -1.0193 \\ -1.0193 & 0.8884 \end{bmatrix} \begin{bmatrix} 2.860 \\ -0.603 \end{bmatrix} \\ &= 235.55 \end{aligned}$$

Diese Realisation der Teststatistik vergleichen wir mit dem 0.95 Quantil der F-Verteilung mit $p = 2, n_w - p = 35 - 2 = 33$ Freiheitsgraden

$$T^2 = 235.55 > 3.284918 \approx F(2, 33; 0.95).$$

Testentscheidung: H_0 wird abgelehnt.

c) Das zu testende Hypothesenpaar lautet nun

$$H_0 : \boldsymbol{\mu}_w = \boldsymbol{\mu}_m, \;\; H_1 : \boldsymbol{\mu}_w \neq \boldsymbol{\mu}_m$$

Es handelt sich um einen Zwei-Stichprobentest. Da laut Aufgabenstellung gleiche Kovarianzmatrizen in den Subpopulationen unterstellt werden, können wir folgende Teststatistik verwenden

$$\begin{aligned} T^2 &= \frac{n_w + n_m - p - 1}{(n_w + n_m - 2) \cdot p} \cdot \frac{n_w \cdot n_m}{n_w + n_m} \cdot (\bar{\boldsymbol{x}}_w - \bar{\boldsymbol{x}}_m)^\top \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}}_w - \bar{\boldsymbol{x}}_m) \\ \text{mit} \quad \boldsymbol{S} &= \frac{1}{n_w + n_m - 2} \left[(n_w - 1)\boldsymbol{S}_w + (n_m - 1)\boldsymbol{S}_m \right] \\ \text{und} \quad \boldsymbol{S}_i &= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^T, \quad i = w, m \end{aligned}$$

Es gilt

$$T^2 \overset{H_0}{\sim} F(p, n_w + n_m - p - 1)$$

und damit lehnen wir H_0 ab, falls

$$T^2 > F(p, n_w + n_m - p - 1; 1 - \alpha).$$

Für die gepoolte Datenmatrix erhalten wir mit $n_w = 35$ und $n_m = 14$ und damit $n_w + n_m - 2 = 47$

$$\begin{aligned} \boldsymbol{S} &= \frac{1}{47} \left\{ 34 \begin{bmatrix} 17.683 & 20.290 \\ 20.290 & 24.407 \end{bmatrix} + 13 \begin{bmatrix} 18.479 & 19.095 \\ 19.095 & 19.273 \end{bmatrix} \right\} \\ &= \frac{1}{47} \left\{ \begin{bmatrix} 601.222 & 689.860 \\ 689.860 & 829.838 \end{bmatrix} + \begin{bmatrix} 240.227 & 248.235 \\ 248.235 & 250.549 \end{bmatrix} \right\} \\ &= \begin{bmatrix} 17.903 & 19.959 \\ 19.959 & 22.987 \end{bmatrix} \end{aligned}$$

Für die Inverse erhalten wir

$$\boldsymbol{S}^{-1} = \frac{1}{17.903 \cdot 22.987 - 19.959^2} \begin{bmatrix} 22.987 & -19.959 \\ -19.959 & 17.903 \end{bmatrix} = \begin{bmatrix} 1.747 & -1.517 \\ -1.517 & 1.361 \end{bmatrix}.$$

Für die Differenz der Mittelwertvektoren erhalten wir

$$\bar{\boldsymbol{x}}_w - \bar{\boldsymbol{x}}_m = \begin{bmatrix} 22.860 \\ 24.397 \end{bmatrix} - \begin{bmatrix} 21.821 \\ 22.843 \end{bmatrix} = \begin{bmatrix} 1.039 \\ 1.554 \end{bmatrix}.$$

Damit erhalten wir

$$T^2 = \frac{46}{94} \cdot \frac{35 \cdot 14}{49} \cdot 0.273 = \frac{46}{94} \cdot 2.733 = 1.337$$

als Realisation unserer Teststatistik. Diese vergleichen wir mit dem 0.95-Quantil der F-Verteilung mit $p = 2$ und $n_w + n_m - p - 1 = 46$ Freiheitsgraden. Es gilt

$$T^2 = 1.337 < 3.2 \approx F(2, 46; 0.95).$$

Testentscheidung: H_0 kann nicht verworfen werden.

Aufgabe 3: Multivariate Tests II

Für die Fußball-Bundesliga liegen für 16 verschiedene Vereine die beiden Merkmale 'Anzahl an Toren in der 1. Halbzeit' und 'Anzahl an Toren in der 2. Halbzeit' vor, jeweils für die Saisons 2013/14 ($\boldsymbol{X}^{(1)}$) und 2014/15 ($\boldsymbol{X}^{(2)}$). Mittels eines geeigneten multivariaten Tests soll überprüft werden, ob sich die beiden Merkmale zwischen den betrachteten Saisons unterscheiden.

	$\boldsymbol{x}_i^{(1)}$		$\boldsymbol{x}_i^{(2)}$		$\boldsymbol{d}_i$	
	1. HZ	2. HZ	1. HZ	2. HZ	1. HZ	2. HZ
Bayern München	37	57	29	51		
VfL Wolfsburg	26	37	34	38		
Bor. M'Gladbach	27	32	24	29		
Bor. Dortmund	32	48	22	25	10	23
Eintr. Frankfurt	18	22	23	33	-5	-11
Bayer Leverkusen	32	28	18	44	14	-16
Werder Bremen	20	22	24	26	-4	-4
Schalke 04	30	33	19	23	11	10
Mainz 05	19	33	17	28	2	5
1899 Hoffenheim	34	38	25	24	9	14
Hamburger SV	20	31	11	14	9	17
Hertha BSC	19	21	16	20	3	1
Hannover 96	22	24	15	25	7	-1
SC Freiburg	12	31	13	23	-1	8
VfB Stuttgart	26	23	16	26	10	-3
FC Augsburg	26	21	13	30	13	-9

a) Vervollständigen Sie die fehlenden Werte der Differenzenvektoren $\boldsymbol{d}_i = \boldsymbol{x}_i^{(1)} - \boldsymbol{x}_i^{(2)}$ in obiger Tabelle! Zeigen Sie außerdem, dass gilt

$$\bar{\boldsymbol{d}} = \begin{pmatrix} 5.1 \\ 2.6 \end{pmatrix}$$

b) Die geschätzte Kovarianzmatrix des obigen Differenzenvektors ist gegeben durch

$$\boldsymbol{S} = \begin{pmatrix} 45.3 & 11.3 \\ 11.3 & 108.3 \end{pmatrix}$$

Testen Sie mit einem geeigneten Test zum Signifkanzniveau von 5% und unter der Annahme von normalverteilten Variablen für beide Saisons, ob sich die erwarteten Anzahlen an Toren für Halbzeit 1 und Halbzeit 2 über die Saisons hinweg signifkant verändert haben.

c) Lesen Sie die Rohdaten `buligoals.rda` von der Übungshompage in `R` ein und vollziehen Sie den gesamten Test nach.

Lösung:

a) Tabelle

	$\boldsymbol{x}_i^{(1)}$		$\boldsymbol{x}_i^{(2)}$		$\boldsymbol{d}_i$	
	1. HZ	2. HZ	1. HZ	2. HZ	1. HZ	2. HZ
Bayern München	37	57	29	51	8	6
VfL Wolfsburg	26	37	34	38	-8	-1
Bor. M'Gladbach	27	32	24	29	3	3
Bor. Dortmund	32	48	22	25	10	23

$$\bar{d}_{EH} = \frac{1}{16}(8 - 8 + 3 + 10 - 5 + 14 - 4 + 11 + 2 + 9 + 9 + 3 + 7 - 1 + 10 + 13) = \frac{81}{16} = 5.1$$

$$\bar{d}_{ZH} = \frac{1}{16}(6 - 1 + 3 + 23 - 11 - 16 - 4 + 10 + 5 + 14 + 17 + 1 - 1 + 8 - 3 - 9) = \frac{42}{16} = 2.6$$

b) Zwei-Stichprobenfall bei verbundener Stichprobe

Hypothesen:

$$\mathrm{H}_0: \quad \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)} \quad vs. \quad \mathrm{H}_1: \quad \boldsymbol{\mu}^{(1)} \neq \boldsymbol{\mu}^{(2)}$$

bzw mit $\boldsymbol{d}_i = \boldsymbol{x}_i^{(1)} - \boldsymbol{x}_i^{(2)}$

$$\mathrm{H}_0: \quad \mathbb{E}(\boldsymbol{d}_i) = \mathbf{0} \quad vs. \quad \mathrm{H}_1: \quad \mathbb{E}(\boldsymbol{d}_i) \neq \mathbf{0}$$

Teststatistik (jeweils mit gerundeten Zwischenergebnissen)

$$\begin{aligned}
T^2 &= \frac{(n-p)n}{(n-1)p}(\bar{\boldsymbol{d}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1}(\bar{\boldsymbol{d}} - \boldsymbol{\mu}_0) \\
\boldsymbol{A} &= \begin{pmatrix} a & b \\ c & d \end{pmatrix} \rightarrow \boldsymbol{A}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \\
\boldsymbol{S}^{-1} &= \frac{1}{45.3 \cdot 108.3 - 11.3^2}\begin{pmatrix} 108.3 & -11.3 \\ -11.3 & 45.3 \end{pmatrix} = \frac{1}{4778.3}\begin{pmatrix} 108.3 & -11.3 \\ -11.3 & 45.3 \end{pmatrix} \\
T^2 &= \frac{(n-p)n}{(n-1)p}(\bar{\boldsymbol{d}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}^{-1}(\bar{\boldsymbol{d}} - \boldsymbol{\mu}_0) \\
&= \frac{(16-2)16}{(16-1)2} \cdot \frac{1}{4773.77}\begin{pmatrix} 5.1 & 2.6 \end{pmatrix}\begin{pmatrix} 108.3 & -11.3 \\ -11.3 & 45.3 \end{pmatrix}\begin{pmatrix} 5.1 \\ 2.6 \end{pmatrix} \\
&= \frac{224}{143349} \cdot 2823.4 = 4.412
\end{aligned}$$

H_0 wird abgelehnt wenn gilt

$$\begin{aligned}
T^2 &> F_{1-\alpha}(p, n-p) \\
F_{1-\alpha}(p, n-p) &= F_{0.95}(2,\, 14) = 3.74
\end{aligned}$$

H_0 wird abgelehnt, die Anzahlen an Toren unterscheiden sich zwischen des Saisons signifikant.