Multivariate Verfahren

Multivariate Schätz- und Testprobleme

Annika Hoyer

Sommersemester 2020

Schätzen und Testen - Inhalt

Schätzprobleme

Datengrundlage

Wir betrachten Zufallsvektoren $\mathbf{X}$ mit $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$ und $\boldsymbol{\Sigma} = cov(\mathbf{X})$.

Die Stichprobenvariablen $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}$ liefern die Datenmatrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_{(1)}^\top \\ \vdots \\ \mathbf{x}_{(n)}^\top \end{pmatrix} .$$

$\rightarrow$ Die Zeilen von $\mathbf{X}$ sind unabhängig. Die Spalten von $\mathbf{X}$ im Allgemeinen nicht.

Schätzer für den Erwartungswert

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{(i)} = \frac{1}{n}\begin{pmatrix}\sum_{i=1}^{n} x_{i1}\\ \vdots \\ \sum_{i=1}^{n} x_{ip}\end{pmatrix} = \begin{pmatrix}\bar{\mathbf{x}}_1\\ \vdots \\ \bar{\mathbf{x}}_p\end{pmatrix}$$

$\rightarrow$ Bestimmung des Mittelwerts für jedes Merkmal/ jede Spalte

Beispiel: PISA-Studie

Erste Spalte: Lesekompetenz, zweite Spalte: Mathematische Grundbildung, dritte Spalte: Naturwissenschaftliche Grundbildung

$$\mathbf{X} = \begin{pmatrix} 538 & 533 & 528 \\ 507 & 520 & 496 \\ 396 & 334 & 375 \\ \dots & \dots & \dots \\ 504 & 493 & 499 \end{pmatrix}$$

$$\hat{\mu} = \bar{\mathbf{x}} = \begin{pmatrix} 493.45 \\ 493.16 \\ 492.61 \end{pmatrix}$$

$\rightarrow$ Im Bereich Lesekompetenz im Durchschnitt am meisten Punkte, Leistungen im Bereich Naturwissenschaftliche Grundbildung im Mittel am schlechtesten

Eigenschaften von $\bar{\mathbf{x}}$

1. $\sum_{i=1}^{n} (\mathbf{x}_{(i)} - \bar{\mathbf{x}}) = \sum_{i=1}^{n} \mathbf{x}_{(i)} - n\bar{\mathbf{x}} = \mathbf{0}$
   $\rightarrow$ der Schätzer $\bar{\mathbf{x}}$ entspricht dem Schwerpunkt der Daten.
2. $\mathbb{E}(\bar{\mathbf{X}}) = \frac{1}{n}\mathbb{E}(\mathbf{X}_{(i)}) = \frac{1}{n} n\boldsymbol{\mu} = \boldsymbol{\mu}$
   $\rightarrow$ der Schätzer $\bar{\mathbf{x}}$ ist erwartungstreu (unverzerrt).
3. $\text{Cov}(\bar{\mathbf{X}}) = \frac{1}{n}\boldsymbol{\Sigma}$
4. $\mathbb{E}(||\bar{\mathbf{X}} - \boldsymbol{\mu}||^2) < \mathbb{E}(||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2)$ für jeden anderen linearen unverzerrten Schätzer $\hat{\boldsymbol{\mu}}$.
   $\rightarrow$ der Schätzer $\bar{\mathbf{x}}$ ist BLUE (Best Linear Unbiased Estimator).

Zentrierungsmatrix

- Voraussetzung einiger multivariater Verfahren: zentrierte Merkmale
- Vorgehen: Subtraktion des Mittelwerts $\bar{x}_j$ von jedem Wert x_{ij}: $\tilde{x}_{ij} = x_{ij} - \bar{x}_j$
- Mittelwert eines zentrierten Merkmals ist gleich 0:

$$\bar{\tilde{x}} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ij} - \frac{1}{n}\sum_{i=1}^{n}\bar{x}_j = \bar{x}_j - \frac{1}{n}n\bar{x}_j = 0$$

- Zentrierte Datenmatrix:

$$\tilde{\mathbf{X}} = \begin{pmatrix} x_{11} - \bar{x}_1 & \dots & x_{1p} - \bar{x}_p \\ \dots & \dots & \dots \\ x_{n1} - \bar{x}_1 & \dots & x_{np} - \bar{x}_p \end{pmatrix}$$

Beispiel: PISA-Studie

$$\tilde{\mathbf{X}} = \begin{pmatrix} 34.55 & 39.84 & 35.39 \\ 13.55 & 26.84 & 3.39 \\ -97.45 & -159.16 & -117.61 \\ \dots & \dots & \dots \\ 10.55 & -0.16 & 6.39 \end{pmatrix}$$

- Erkennbar, wie sich jedes Land vom Mittelwert unterscheidet
- Australien (erstes Land) liegt in allen Bereichen über dem Durchschnitt
- Brasilien (drittes Land) liegt in allen Bereichen unter dem Durchschnitt

Zentrierungsmatrix

Definition: Zentrierungsmatrix

Die Matrix $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ heißt **Zentrierungsmatrix**. Dabei ist $\mathbf{I}$ die Einheitsmatrix und $\mathbf{1}$ der Einservektor. Die Zentrierungsmatrix ist symmetrisch.
Es gilt:

$$\tilde{\mathbf{X}} = \mathbf{H}\mathbf{X}.$$

Zentrierungsmatrix - Beweis (1)

Es gilt:

$$\mathbf{HX} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\mathbf{X} = \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{X}$$

Betrachten wir zunächst $\frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{X}$. Da $\mathbf{1}$ der summierende Vektor ist, gilt

$$\mathbf{1}^\top\mathbf{X} = (\sum_{i=1}^{n} x_{i1}, ..., \sum_{i=1}^{n} x_{ip}).$$

Da $\frac{1}{n}$ ein Skalar ist, gilt

$$\frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{X} = \mathbf{1}\frac{1}{n}\mathbf{1}^\top\mathbf{X}.$$

Es gilt

$$\frac{1}{n}\mathbf{1}^\top\mathbf{X} = (\bar{x}_1, ..., \bar{x}_p).$$

Zentrierungsmatrix - Beweis (2)

Somit folgt

$$\frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{X} = \mathbf{1}\frac{1}{n}\mathbf{1}^\top\mathbf{X} = \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}(\bar{x}_1, \cdots, \bar{x}_p) = \begin{pmatrix}\bar{x}_1 & \cdots & \bar{x}_p \\ \vdots & \ddots & \vdots \\ \bar{x}_1 & \cdots & \bar{x}_p\end{pmatrix}$$

Also gilt

$$\begin{aligned}\mathbf{HX} &= \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\mathbf{X} = \begin{pmatrix}x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np}\end{pmatrix} - \begin{pmatrix}\bar{x}_1 & \cdots & \bar{x}_p \\ \vdots & \ddots & \vdots \\ \bar{x}_1 & \cdots & \bar{x}_p\end{pmatrix} \\ &= \begin{pmatrix}x_{11} - \bar{x}_1 & \cdots & x_{1p} - \bar{x}_p \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{np} - \bar{x}_p\end{pmatrix} = \tilde{\mathbf{X}}\end{aligned}$$

Erinnerung: Kovarianzmatrix

$$\text{Cov}(\mathbf{X}) = \mathbf{\Sigma} = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))^\top\right]$$

$$= \begin{pmatrix} \text{Cov}(X_1, X_1) & \dots & \text{Cov}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \dots & \text{Cov}(X_p, X_p) \end{pmatrix} \in \mathbb{R}^{p \times p}.$$

Was muss geschätzt werden?

- Hauptdiagonalenelemente/Varianzen
- Kovarianzen auf den Nebendiagonalen

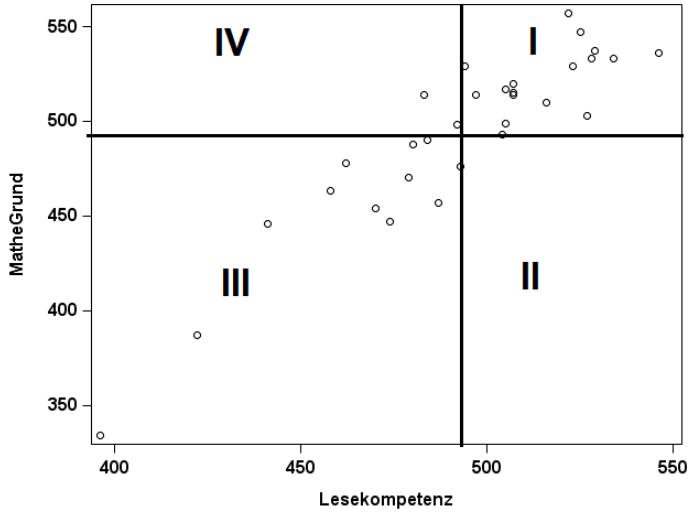
Empirische Standardabweichung

- Maß für die Streuung eines univariaten Merkmals
- Möglichkeit, die Elemente auf der Hauptdiagonalen der Kovarianzmatrix zu schätzen
- Stichprobenvarianz:

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

- Standardabweichung: $s_j = \sqrt{s_j^2}$

Beispiel: PISA-Studie



- Quadrant I: Länder mit Punktzahl in den Bereichen Lesekompetenz und Mathematische Grundbildung über dem Durchschnitt
- Quadrant III: Länder mit Punktzahl in den Bereichen Lesekompetenz und Mathematische Grundbildung unter dem Durchschnitt
- Quadrant II: Länder mit Punktzahl im Bereich Lesekompetenz über dem Durchschnitt, im Bereich Mathematische Grundbildung unter dem Durchschnitt
- Quadrant IV: Länder mit Punktzahl im Bereich Lesekompetenz unter dem Durchschnitt, im Bereich Mathematische Grundbildung über dem Durchschnitt

$\rightarrow$ Positiver Zusammenhang: meiste Beobachtungen in den Quadranten I und III

$\rightarrow$ Negativer Zusammenhang: meiste Beobachtungen in den Quadranten II und IV

$\rightarrow$ Kein Zusammenhang: gleichmäßige Verteilung der Punkte

Empirische Kovarianzmatrix

Sei j das Merkmal auf der x-Achse, k das Merkmal auf der y-Achse und x_{ij} die Ausprägung des $j-$Merkmals beim $i-$Objekt. Dann gilt:

- Quadrant I: $x_{ij} > \bar{x}_j,\ x_{ik} > \bar{x}_k \rightarrow x_{ij} - \bar{x}_j > 0,\ x_{ik} - \bar{x}_k > 0$
- Quadrant II: $x_{ij} > \bar{x}_j,\ x_{ik} < \bar{x}_k \rightarrow x_{ij} - \bar{x}_j > 0,\ x_{ik} - \bar{x}_k < 0$
- Quadrant III: $x_{ij} < \bar{x}_j,\ x_{ik} < \bar{x}_k \rightarrow x_{ij} - \bar{x}_j < 0,\ x_{ik} - \bar{x}_k < 0$
- Quadrant IV: $x_{ij} < \bar{x}_j,\ x_{ik} > \bar{x}_k \rightarrow x_{ij} - \bar{x}_j < 0,\ x_{ik} - \bar{x}_k > 0$

$\rightarrow$ Produkt $(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ im ersten und dritten Quadranten positiv, im zweiten und vierten Quadranten negativ

Empirische Kovarianzmatrix

Definition: Empirische Kovarianz

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

heißt **empirische Kovarianz** zwischen dem j-ten und k-ten Merkmal.

Es gilt:

$$s_{kk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)(x_{ik} - \bar{x}_k) = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2 = s_k^2$$

Empirische Kovarianzmatrix - Alternative Darstellung

$$\mathbf{S} = (s_{jk}) = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_{(i)} - \bar{\mathbf{x}})(\mathbf{x}_{(i)} - \bar{\mathbf{x}})^\top$$

$$= \frac{1}{n-1} \left(\sum_{i=1}^{n} \mathbf{x}_{(i)} \mathbf{x}_{(i)}^\top - n \bar{\mathbf{x}} \bar{\mathbf{x}}^\top \right)$$

Unter Verwendung der Zentrierungsmatrix:

$$\mathbf{S} = \frac{1}{n-1} (\mathbf{HX})^\top (\mathbf{HX}) = \frac{1}{n-1} \mathbf{X}^\top \mathbf{H}^\top \mathbf{HX} = \frac{1}{n-1} \mathbf{X}^\top \mathbf{HX}$$

Empirische Kovarianzmatrix

Darstellung als empirische Kovarianzmatrix bei p Merkmalen:

$$\mathbf{S} = \begin{pmatrix} s_1^2 & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_p^2 \end{pmatrix}$$

Wegen $s_{jk} = s_{kj}$ ist die empirische Kovarianzmatrix symmetrisch.

Beispiel: PISA-Studie

Für die PISA-Studie ergibt sich:

$$\mathbf{S} = \begin{pmatrix} 1109.4 & 1428.3 & 1195.6 \\ 1428.3 & 2192.9 & 1644.0 \\ 1195.6 & 1644.0 & 1419.0 \end{pmatrix}$$

- Positive Kovarianzen
- Kovarianz zwischen den Merkmalen Mathematische Grundbildung und Naturwissenschaftliche Grundbildung am größten

Eigenschaften von $\mathbf{S}$

1. $\mathbb{E}(\mathbf{S}) = \mathbf{\Sigma}$
2. Falls $\mathbf{x} \sim N_p(\mu, \mathbf{\Sigma})$, dann sind $\bar{\mathbf{x}}$ und $\mathbf{S}$ unabhängig.
3. Falls $\mathbf{x} \sim N_p(\mu, \mathbf{\Sigma})$, dann gilt

$$(n-1)\mathbf{S} = \sum_{i=1}^{n} (\mathbf{x}_{(i)} - \bar{\mathbf{x}})(\mathbf{x}_{(i)} - \bar{\mathbf{x}})^\top \sim W_p(\mathbf{\Sigma}, n-1)$$

Mehr-Stichproben-Fall

- Grundgesamtheit in g Gruppen partitioniert
- Betrachten Zufallsvektoren $\mathbf{X}_r,\ r = 1, \cdots, g$ mit $\boldsymbol{\mu}_r = \mathbb{E}(\mathbf{X}_r) = \mathbb{E}(\mathbf{X}|\, Y = r)$ und $\boldsymbol{\Sigma}_r = \mathrm{Cov}(\mathbf{X}_r)$

Beispiel: PISA-Studie

Punkte nach Gruppen, LK: Lesekompetenz, MG: Mathematische Grundbildung, NG: Naturwissenschaftliche Grundbildung

Gruppe 1				Gruppe 2				Gruppe 3			
Land	LK	MG	NG	Land	LK	MG	NG	Land	LK	MG	NG
FIN	546	536	538	AUS	528	533	528	GR	474	447	461
J	522	557	550	B	507	520	496	GB	523	529	532
FL	483	514	476	BR	396	334	375	IRL	527	503	513
L	441	446	443	DK	497	514	481	I	487	457	478
A	507	515	519	D	484	490	487	LV	458	463	460
S	516	510	512	F	505	517	500	MEX	422	387	422
CH	494	529	496	IS	507	514	496	PL	479	470	483
CZ	492	498	511	CDN	534	533	529	RUS	462	478	460
				ROK	525	547	552	E	493	476	491
				NZ	529	537	528	H	480	488	496
				N	505	499	500				
				P	470	454	459				
				USA	504	493	499				

Empirische Varianzzerlegung

- Datensituation: Realisationen $\mathbf{x}_{ri}$ der unabhängigen Zufallsvariablen $\mathbf{X}_{ri}$ für $r = 1, \cdots, g$ und $i = 1, \cdots, n_r$
- Gesamtmittel:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{r=1}^{g} \sum_{i=1}^{n_r} \mathbf{x}_{ri}$$

- Mittelwerte der Gruppen:

$$\bar{\mathbf{x}}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} \mathbf{x}_{ri}$$

Beispiel: PISA-Studie

$$\bar{\mathbf{x}} = \begin{pmatrix} 493.452 \\ 493.161 \\ 492.613 \end{pmatrix} \quad \bar{\mathbf{x}}_1 = \begin{pmatrix} 500.125 \\ 513.125 \\ 505.625 \end{pmatrix}$$

$$\bar{\mathbf{x}}_2 = \begin{pmatrix} 499.308 \\ 498.846 \\ 494.615 \end{pmatrix} \quad \bar{\mathbf{x}}_3 = \begin{pmatrix} 480.500 \\ 469.800 \\ 479.600 \end{pmatrix}$$

Empirische Varianzzerlegung

- Streuung der Gruppenmittel um das Gesamtmittel unter Berücksichtigung verschiedener Gruppengrößen:

$$\mathbf{B} = \sum_{r=1}^{g} n_r(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})^\top$$

$\rightarrow$ Zwischen-Gruppen-Streumatrix

- Streuung innerhalb der Gruppen:

$$\mathbf{W} = \sum_{r=1}^{g} \sum_{i=1}^{n_r} = (\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)(\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)^\top$$

$\rightarrow$ Inner-Gruppen-Streumatrix

- Gesamtstreuung:

$$\mathbf{T} = \sum_{r=1}^{g} \sum_{i=1}^{n_r} (\mathbf{x}_{ri} - \bar{\mathbf{x}})(\mathbf{x}_{ri} - \bar{\mathbf{x}})^\top = \mathbf{W} + \mathbf{B}$$

Beispiel: PISA-Studie

$$\mathbf{B} = \begin{pmatrix} 2479.53 & 4524.25 & 2532.51 \\ 4524.25 & 9066.03 & 5266.13 \\ 2532.51 & 5266.13 & 3100.00 \end{pmatrix}$$

$$\mathbf{W} = \begin{pmatrix} 30802.14 & 38325.49 & 33335.91 \\ 38325.49 & 56720.17 & 44054.81 \\ 33335.91 & 44054.81 & 39469.35 \end{pmatrix}$$

$$\mathbf{T} = \begin{pmatrix} 33281.67 & 42849.74 & 35868.42 \\ 42849.74 & 65786.20 & 49320.94 \\ 35868.42 & 49320.94 & 42569.35 \end{pmatrix}$$

Schätzung im Mehr-Stichprobenfall

Falls $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_g$, dann gilt

$$\boldsymbol{S} = \frac{1}{n-g}\sum_{r=1}^{g}(n_r - 1)\boldsymbol{S}_r = \frac{1}{n-g}\sum_{r=1}^{g}\sum_{i=1}^{n_r}(\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)(\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)^\top$$
$$= \frac{1}{n-g}\mathbf{W}$$

Eigenschaften:

1. Falls $\mathbf{x}_r \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma})$, so ist $\mathbf{W} \sim W_p(\boldsymbol{\Sigma}, n-g)$ und unabhängig von $\mathbf{B}$
2. Falls $\mathbf{x}_r \sim N_p(\boldsymbol{\mu}_r, \boldsymbol{\Sigma})$ und $\boldsymbol{\mu} = \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g$, so ist $\mathbf{B} \sim W_p(\boldsymbol{\Sigma}, g-1)$ und damit $\mathbf{T} \sim W_p(\boldsymbol{\Sigma}, n-1)$

Ähnlichkeits- und Distanzmaße

Ähnlichkeitskoeffizient und Distanzmaß

- Ziel: Bestimmung der Ähnlichkeit von n Objekten, bei denen p Merkmale erhoben wurden
- Ähnlichkeitskoeffizient s_{ij}, der umso größer ist, je ähnlicher sich die Objekte i und j sind
- Oftmals normiert: $0 \leq s_{ij} \leq 1$
- Alternative: Distanzmaß d_{ij}, das Unähnlichkeit zwischen dem i-ten und j-ten Objekt misst

$\rightarrow$ je größer, desto unterschiedlicher

- Wenn $0 \leq s_{ij} \leq 1$ dann $0 \leq d_{ij} \leq 1$ für $d_{ij} = 1 - s_{ij}$
- Darstellung in Distanzmatrix $\mathbf{D}$:

$$\mathbf{D} = \begin{pmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{pmatrix}$$

Beispiel: Studierendenbefragung

Befragung von Studierenden zum Alter ihrer Mutter und ihres Vaters

Alter von Mutter und Vater von 6 Studierenden

Studierender	Alter der Mutter	Alter des Vaters
1	58	60
2	61	62
3	55	59
4	59	64
5	54	54
6	52	55

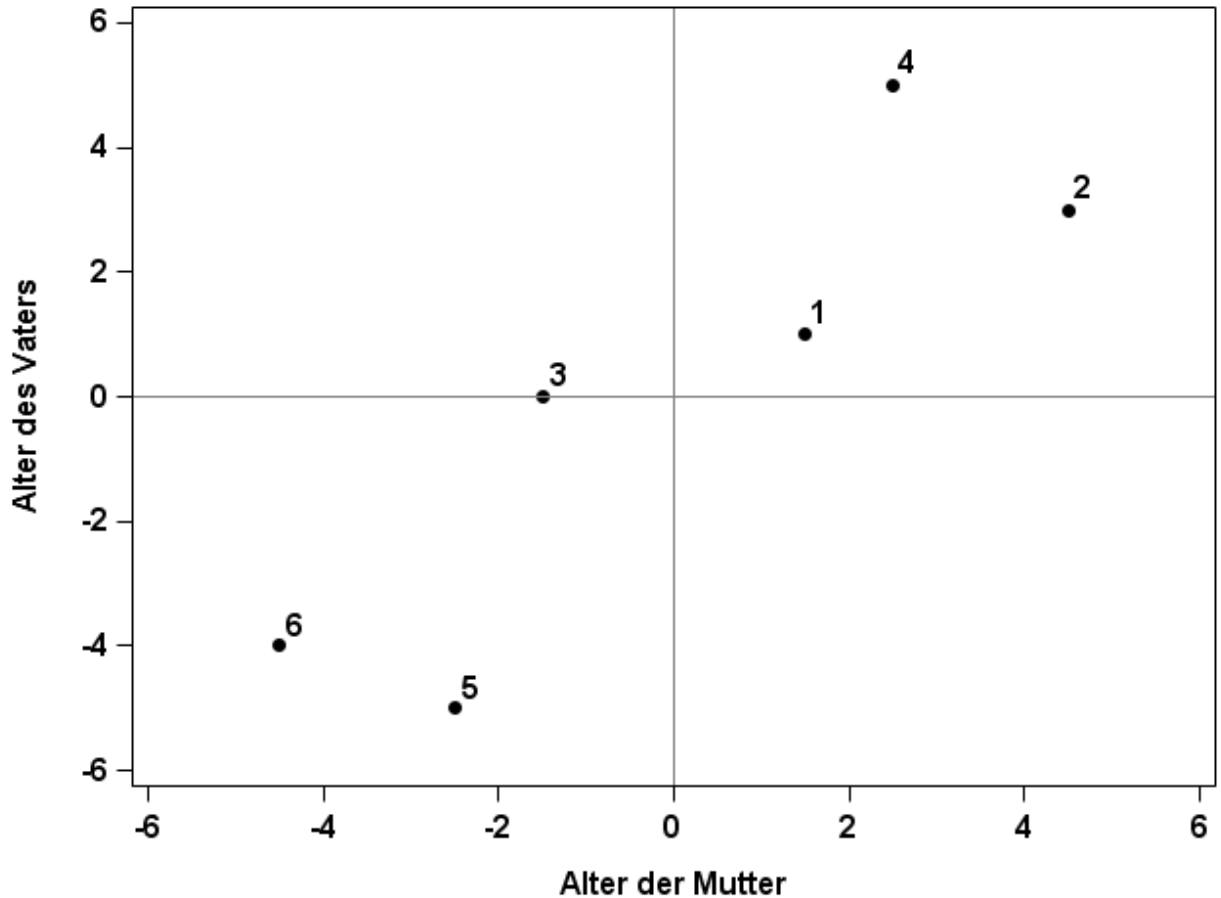
Beispiel: Studierendenbefragung

Zentrieren der Beobachtungen

Zentriertes Alter von Mutter und Vater von 6 Studierenden

Studierender	Alter der Mutter	Alter des Vaters
1	1.5	1
2	4.5	3
3	-1.5	0
4	2.5	5
5	-2.5	-5
6	-4.5	-4

Beispiel: Studierendenbefragung



Distanzmaß

- Naheliegend, als Distanz den kürzesten Abstand zwischen den Punkten zu wählen
- Seien $\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix}$ und $\mathbf{x}_j = \begin{pmatrix} x_{j1} \\ x_{j2} \end{pmatrix} \in \mathbb{R}^2$
- Kürzester Abstand nach Satz des Pythagoras (euklidische Distanz):

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$$

Beispiel: Studierendenbefragung

- Distanz zwischen den ersten beiden Studierenden:

$$d_{12} = \sqrt{(58-61)^2 + (60-62)^2} = \sqrt{13} = 3.6$$

- Distanzmatrix mit den euklidischen Distanzen:

$$\mathbf{D} = \begin{pmatrix} 0.0 & 3.6 & 3.2 & 4.1 & 7.2 & 7.8 \\ 3.6 & 0.0 & 6.7 & 2.8 & 10.6 & 11.4 \\ 3.2 & 6.7 & 0.0 & 6.4 & 5.1 & 5.0 \\ 4.1 & 2.8 & 6.4 & 0.0 & 11.2 & 11.4 \\ 7.2 & 10.6 & 5.1 & 11.2 & 0.0 & 2.2 \\ 7.8 & 11.4 & 5.0 & 11.4 & 2.2 & 0.0 \end{pmatrix}$$

Distanzmaß

- Übertragung auf höherdimensionale Räume mit

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \text{ und } \mathbf{x}_j = \begin{pmatrix} x_{j1} \\ \vdots \\ x_{jp} \end{pmatrix}$$

$\rightarrow$ $d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top(\mathbf{x}_i - \mathbf{x}_j)}$

- Bei ungleichen Varianzen der Merkmale $\rightarrow$ skalierte euklidische Distanz mit Stichprobenvarianzen $s_1^2, \cdots, s_p^2$:

$$d_{ij}^s = \sqrt{\sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{s_k^2}} = \sqrt{\sum_{k=1}^{p} \left(\frac{x_{ik}}{s_k} - \frac{x_{jk}}{s_k} \right)^2}$$

Matrixdarstellung

- Es gilt:

$$d_{ij}^s = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{V}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$$

mit

$$\mathbf{V} = \begin{pmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_p^2 \end{pmatrix}$$

als Diagionalmatrix mit positiven Hauptdiagonalelementen

$\rightarrow$

$$\mathbf{V}^{0.5} = \begin{pmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_p \end{pmatrix} \text{ und } \mathbf{V}^{-0.5} = \begin{pmatrix} \frac{1}{s_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{s_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{s_p} \end{pmatrix}$$

Matrixdarstellung

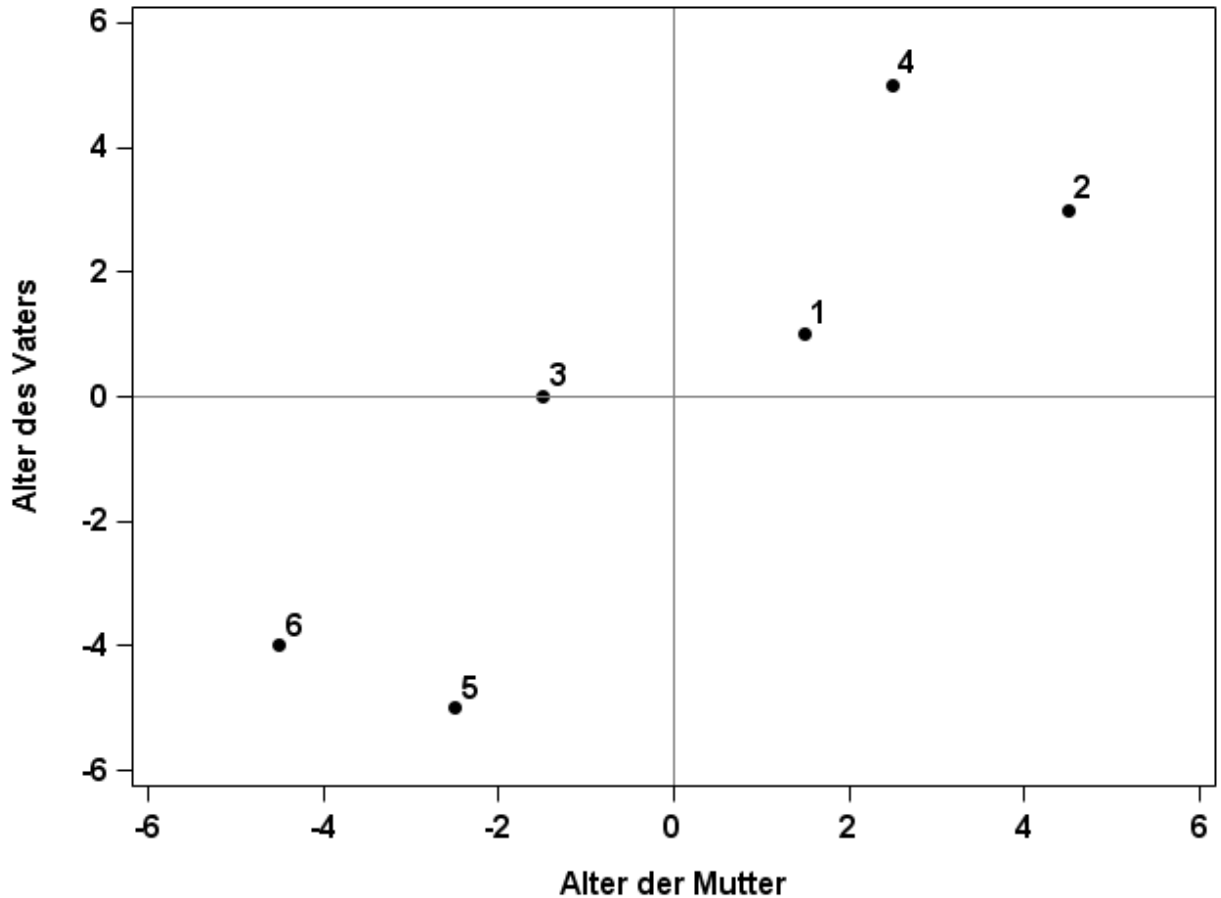
- $j-$te skalierte Beobachtung:

$$\mathbf{V}^{-0.5}\mathbf{x}_j = \begin{pmatrix} \frac{x_{j1}}{s_1} \\ \vdots \\ \frac{x_{jp}}{s_p} \end{pmatrix}$$

- Es ergibt sich:

$$\begin{aligned}
&(\mathbf{V}^{-0.5}\mathbf{x}_i - \mathbf{V}^{-0.5}\mathbf{x}_j)^\top(\mathbf{V}^{-0.5}\mathbf{x}_i - \mathbf{V}^{-0.5}\mathbf{x}_j) = \\
&(\mathbf{V}^{-0.5}(\mathbf{x}_i - \mathbf{x}_j))^\top\mathbf{V}^{-0.5}(\mathbf{x}_i - \mathbf{x}_j) = \\
&(\mathbf{x}_i - \mathbf{x}_j)^\top(\mathbf{V}^{-0.5})^\top(\mathbf{V}^{-0.5}(\mathbf{x}_i - \mathbf{x}_j)) = \\
&(\mathbf{x}_i - \mathbf{x}_j)^\top\mathbf{V}^{-0.5}(\mathbf{V}^{-0.5}(\mathbf{x}_i - \mathbf{x}_j)) = \\
&(\mathbf{x}_i - \mathbf{x}_j)^\top\mathbf{V}^{-1}(\mathbf{x}_i - \mathbf{x}_j)
\end{aligned}$$

Beispiel: Studierendenbefragung

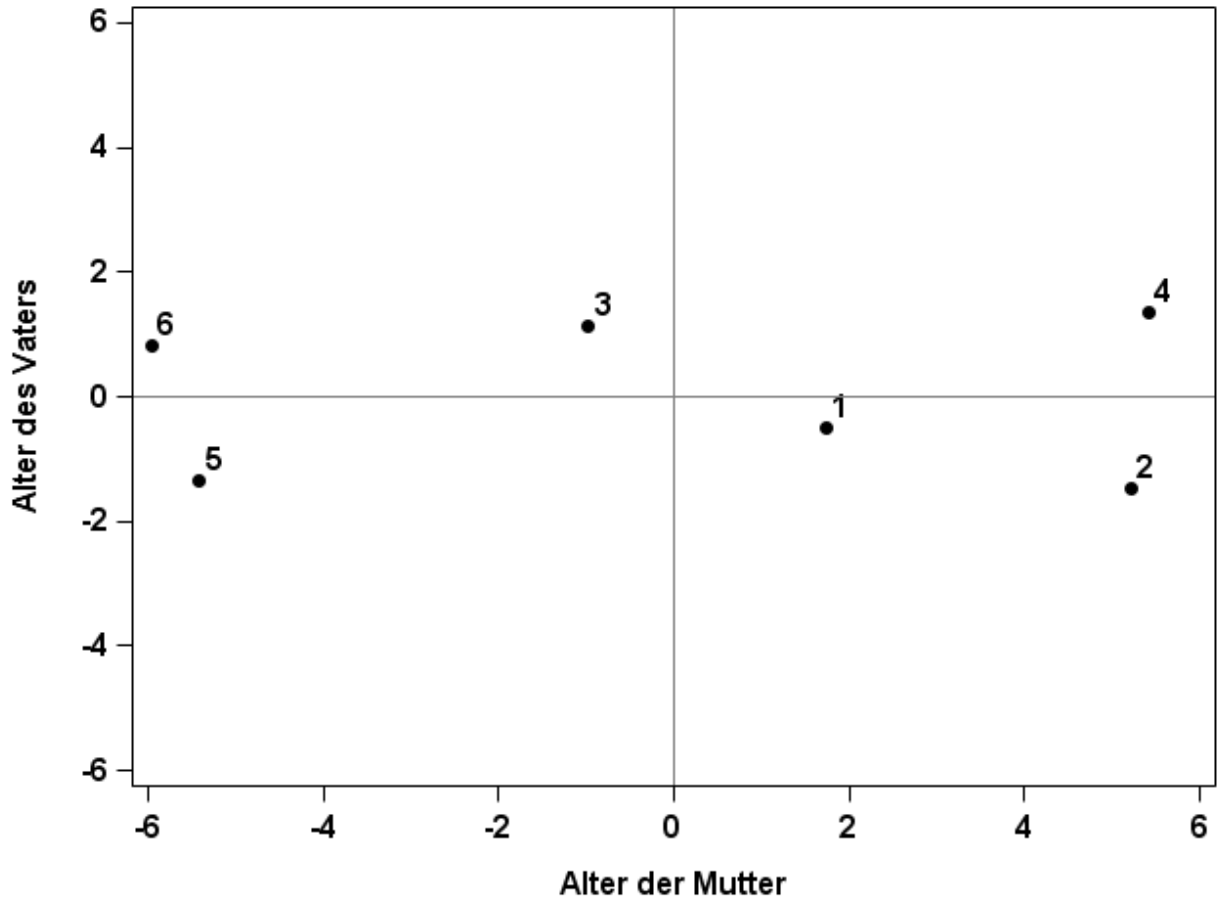


Beispiel: Studierendenbefragung

Skalierte euklidische Distanzen mit Stichprobenvarianzen $s_1^2 = 11.5$ und $s_2^2 = 15.2$:

$$\mathbf{D} = \begin{pmatrix} 0.0 & 1.0 & 0.9 & 1.1 & 1.9 & 2.2 \\ 1.0 & 0.0 & 1.9 & 0.8 & 2.9 & 3.2 \\ 0.9 & 1.9 & 0.0 & 1.7 & 1.3 & 1.4 \\ 1.1 & 0.8 & 1.7 & 0.0 & 3.0 & 3.1 \\ 1.9 & 2.9 & 1.3 & 3.0 & 0.0 & 0.6 \\ 2.2 & 3.2 & 1.4 & 3.1 & 0.6 & 0.0 \end{pmatrix}$$

Beispiel: Studierendenbefragung



Beispiel: Studierendenbefragung

Zentriertes und rotiertes Alter von Mutter und Vater von 6 Studierenden

Studierender	Alter der Mutter	Alter des Vaters
1	1.74	-0.49
2	5.21	-1.47
3	-0.98	1.14
4	5.42	1.35
5	-5.42	-1.35
6	-5.96	0.82

Beispiel: Studierendenbefragung

- Stichprobenvarianzen $s_1^2 = 25.1$ und $s_2^2 = 1.6$
- Skalierte euklidische Distanzen zwischen den zentrierten und rotierten Beobachtungen:

$$\mathbf{D} = \begin{pmatrix} 0.00 & 1.04 & 1.40 & 1.63 & 1.58 & 1.85 \\ 1.04 & 0.00 & 2.40 & 2.23 & 2.12 & 2.87 \\ 1.40 & 2.40 & 0.00 & 1.29 & 2.16 & 1.03 \\ 1.63 & 2.23 & 1.29 & 0.00 & 3.04 & 2.31 \\ 1.58 & 2.12 & 2.16 & 3.04 & 0.00 & 1.72 \\ 1.85 & 2.87 & 1.03 & 2.31 & 1.72 & 0.00 \end{pmatrix}$$

$\rightarrow$ Bei korrelierten Daten Rotation erforderlich, damit die Daten hinsichtlich der neuen Koordinatenachsen unkorreliert sind

Spektralzerlegung

- Spektralzerlegung der Kovarianzmatrix:

$$\mathbf{S} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^\top$$

- $\mathbf{T}$: orthogonale Matrix mit den normierten Eigenvektoren von $\mathbf{S}$ als Spalten

$\rightarrow$ $\mathbf{T}\mathbf{T}^\top = \mathbf{I}$ und $\mathbf{T}^\top\mathbf{T} = \mathbf{I}$

$\rightarrow$ $\mathbf{T}^\top$ ist die Rotationsmatrix

- $\mathbf{\Lambda} = diag(\lambda_1, \cdots, \lambda_p)$: Diagonalmatrix der geordneten Eigenwerte

Rotierte Beobachtungen

- Rotierte Beobachtungen $\mathbf{T}^\top \mathbf{x}_i$ sind unkorreliert
- Eigenwerte sind die Varianzen der rotierten Merkmale
- $\mathbf{\Lambda}^{-0.5}\mathbf{T}^\top \mathbf{x}_i$ sind die rotierten und skalierten Beobachtungen

$$\begin{aligned} \text{Var}(\mathbf{\Lambda}^{-0.5}\mathbf{T}^\top \mathbf{x}_i) &= \mathbf{\Lambda}^{-0.5}\mathbf{T}^\top \text{Var}(\mathbf{x}_i)(\mathbf{\Lambda}^{-0.5}\mathbf{T}^\top)^\top \\ &= \mathbf{\Lambda}^{-0.5}\mathbf{T}^\top \mathbf{S}\mathbf{T}\mathbf{\Lambda}^{-0.5} \\ &= \mathbf{\Lambda}^{-0.5}\mathbf{\Lambda}\mathbf{\Lambda}^{-0.5} \\ &= \mathbf{I} \end{aligned}$$

$\rightarrow$ unkorreliert, mit gleicher Varianz

Rotierte Beobachtungen

Bestimmung der euklidischen Distanzen zwischen den rotierten und skalierten Beobachtungen:

$$\sqrt{(\mathbf{\Lambda}^{-0.5}\mathbf{T}^\top\mathbf{x}_i - \mathbf{\Lambda}^{-0.5}\mathbf{T}^\top\mathbf{x}_j)^\top(\mathbf{\Lambda}^{-0.5}\mathbf{T}^\top\mathbf{x}_i - \mathbf{\Lambda}^{-0.5}\mathbf{T}^\top\mathbf{x}_j)} =$$

$$\sqrt{(\mathbf{\Lambda}^{-0.5}\mathbf{T}^\top(\mathbf{x}_i - \mathbf{x}_j))^\top\mathbf{\Lambda}^{-0.5}\mathbf{T}^\top(\mathbf{x}_i - \mathbf{x}_j)} =$$

$$\sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top(\mathbf{\Lambda}^{-0.5}\mathbf{T}^\top)^\top\mathbf{\Lambda}^{-0.5}\mathbf{T}^\top(\mathbf{x}_i - \mathbf{x}_j)} =$$

$$\sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top\mathbf{T}\mathbf{\Lambda}^{-0.5}\mathbf{\Lambda}^{-0.5}\mathbf{T}^\top(\mathbf{x}_i - \mathbf{x}_j)} =$$

$$\sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top\mathbf{T}\mathbf{\Lambda}^{-1}\mathbf{T}^\top(\mathbf{x}_i - \mathbf{x}_j)}$$

$\rightarrow \mathbf{T}\mathbf{\Lambda}\mathbf{T}^\top$ ist Spektralzerlegung von $\mathbf{S}^{-1}$

Mahalanobis-Distanz

Definition: Mahalanobis-Distanz

Die euklidischen Distanzen zwischen den rotierten und skalierten Beobachtungen sind gleich

$$d_{ij}^M = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$$

d_{ij}^M nennt man auch Mahalanobis-Distanz.

Beispiel: Studierendenbefragung

Bestimmung der Mahalanobis-Distanz zwischen

$$\mathbf{x}_1 = \begin{pmatrix} 58 \\ 60 \end{pmatrix} \quad \text{und} \quad \mathbf{x}_2 = \begin{pmatrix} 61 \\ 62 \end{pmatrix}.$$

Es gilt

$$\mathbf{S} = \begin{pmatrix} 11.5 & 11.6 \\ 11.6 & 15.2 \end{pmatrix} \quad \text{und} \quad \mathbf{S}^{-1} = \begin{pmatrix} 0.378 & -0.288 \\ -0.288 & 0.286 \end{pmatrix}.$$

Mit

$$\mathbf{x}_1 - \mathbf{x}_2 = \begin{pmatrix} 58 \\ 60 \end{pmatrix} - \begin{pmatrix} 61 \\ 62 \end{pmatrix} = \begin{pmatrix} -3 \\ -2 \end{pmatrix}$$

gilt

$$\begin{aligned} d_{12}^M &= \sqrt{\begin{pmatrix} -3 & -2 \end{pmatrix} \begin{pmatrix} 0.378 & -0.288 \\ -0.288 & 0.286 \end{pmatrix} \begin{pmatrix} -3 \\ -2 \end{pmatrix}} \\ &= \sqrt{\begin{pmatrix} -3 & -2 \end{pmatrix} \begin{pmatrix} -0.558 \\ 0.292 \end{pmatrix}} = \sqrt{1.09} = 1.04 \end{aligned}$$

Mahalanobis-Distanz

- Mahalanobis-Distanz oftmals verwendet, um Abstand von Punkten $\mathbf{x}_1 \cdots, \mathbf{x}_p$ zu einem Punkt $\mathbf{z}$ zu bestimmen

$\rightarrow$ $d_i^M = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})^\top \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}}$

- Im Beispiel:

$$\bar{\mathbf{x}} = \begin{pmatrix} 56.5 \\ 59.0 \end{pmatrix}$$

$\rightarrow$ $d_1^M = 0.52, d_2^M = 1.56, d_3^M = 0.92, d_4^M = 1.52, d_5^M = 1.52, d_6^M = 1.36$

Manhattan-Metrik

- Euklidische Distanz: kürzester Abstand zwischen zwei Punkten (Hypothenuse im rechtwinkligen Dreieck)
- Alternatives Distanzmaß: Summe der Längen der beiden Katheten

$\rightarrow$ kürzeste Verbindung zwischen zwei Punkten bei Betrachtung einer Stadt mit einem rechtwinkligen Straßennetz

- Manhattan-Metrik oder City-Block-Metrik

Manhattan-Metrik

Definition: Manhattan-Metrik

Die Manhattan-Metrik zwischen dem i-ten und j-ten Objekt mit den Merkmalsvektoren

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \quad \text{und} \quad \mathbf{x}_j = \begin{pmatrix} x_{j1} \\ \vdots \\ x_{jp} \end{pmatrix}$$

ist definiert als

$$d_{ij} = \sum_{k=1}^{p} \mid x_{ik} - x_{jk} \mid .$$

Beispiel: Studierendenbefragung

- Distanz zwischen den Studierenden 1 und 2:

$$d_{12} = | 58 - 61 | + | 60 - 62 | = 5$$

- Distanzmatrix

$$\mathbf{D} = \begin{pmatrix} 0 & 5 & 4 & 5 & 10 & 11 \\ 5 & 0 & 9 & 4 & 15 & 16 \\ 4 & 9 & 0 & 9 & 6 & 7 \\ 5 & 4 & 9 & 0 & 15 & 16 \\ 10 & 15 & 6 & 15 & 0 & 3 \\ 11 & 16 & 7 & 16 & 3 & 0 \end{pmatrix}$$

Manhattan-Metrik

- Skalierung der Merkmale: Dividieren durch die Spannweite R_k (Differenz aus Maximum und Minimum) des Merkmals k
- $$d_{ij} = \sum_{k=1}^{p} \frac{| x_{ik} - x_{jk} |}{R_k}$$
- Beispiel:

$$\mathbf{D} = \begin{pmatrix} 0 & 0.5 & 0.4 & 0.5 & 1 & 1.2 \\ 0.5 & 0 & 1 & 0.4 & 1.6 & 1.7 \\ 0.4 & 1 & 0 & 0.9 & 0.6 & 0.7 \\ 0.5 & 0.4 & 0.9 & 0 & 1.6 & 1.7 \\ 1 & 1.6 & 0.6 & 1.6 & 0 & 0.3 \\ 1.2 & 1.7 & 0.7 & 1.7 & 0.3 & 0 \end{pmatrix}$$

Testprobleme

Wiederholung: Einstichprobentest auf μ

- $p = 1$ und Varianz der normalverteilten Stichprobe bekannt
- Gauß-Test für $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$ mit Teststatistik

$$T = \sqrt{n}\frac{\bar{x} - \mu_0}{\sigma}$$

- T unter H_0 χ^2-verteilt mit einem Freiheitsgrad
- Ablehnung von H_0, wenn $T > \chi^2_{1;1-\alpha}$

Test für den Erwartungswert

Gegeben sind Daten $\mathbf{X} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)})^\top$, wobei $\mathbf{x}_{(i)} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Hypothesen

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

Unterscheidung:

1. Kovarianz $\boldsymbol{\Sigma}$ ist bekannt
2. Kovarianz $\boldsymbol{\Sigma}$ ist unbekannt

1. Fall: $\boldsymbol{\Sigma}$ bekannt

- Teststatistik:

$$T^2 = \left(\sqrt{n}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)\right)^\top \left(\sqrt{n}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)\right) \overset{H_0}{\sim} \chi^2(p)$$

bzw.:

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$$

$\rightarrow$ entspricht der Mahalanobis-Distanz zwischen $\bar{\mathbf{x}}$ und $\boldsymbol{\mu}_0$

- H_0 ablehnen, wenn gilt: $T^2 > \chi^2(p; 1-\alpha)$

1. Fall: $\boldsymbol{\Sigma}$ bekannt

$(1-\alpha)$-Konfidenzbereich:

$$\mathbb{P}\left\{n(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)\right\} \leq \chi^2(p;1-\alpha)$$

$\rightarrow$ entspricht einem Ellipsoid mit Mittelpunkt $\bar{\boldsymbol{x}}$

Union-Intersection-Test nach Roy (1957)

- Idee: Versuche multivariate Hypothese H_0 als Durchschnitt (Intersection) von univariaten Hypothesen zu schreiben
- H_0 lässt sich auch als Schnitt darstellen

$$H_0 = \bigcap_{\mathbf{a} \in \Gamma} H_{0,a},$$

wobei Γ eine beliebige Indexmenge ist

- Testprinzip:
  Lehne H_0 ab, falls mindestens eine $H_{0,a}$ abgelehnt wird.
  Damit ergibt sich der Ablehnbereich

$$\bigcup_{\mathbf{a} \in \Gamma} \{\text{Ablehnungsbereich für } \mathbf{a}\}$$

als Vereinigung (Union) der Ablehnungsbereiche der einzelnen Tests

Union-Intersection-Test nach Roy (1957)

- Der Test einer Nullhypothese $H_{0,a}$ entspricht einem eindimensionalen Testproblem, in diesem Fall einem einfachen Gauß-Test mit Teststatistik

$$\sqrt{n}\frac{\bar{x} - \mu_0}{\sigma}$$

- Der multivariate Test auf den Erwartungswert lässt sich sowohl als *Likelihood-Quotient-Test* als auch als *Union-Intersection-Test* ableiten.

Beispiel: Union-Intersection-Test

Sei $x_1, \cdots, x_n$ eine Zufallsvariable mit $x_i \sim \mathrm{N}(\mu, \sigma^2), \forall i$.
Hypothese: $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$
H_0 als Durchschnitt:

$$H_0 : \{\mu : \mu \leq \mu_0\} \cap \{\mu : \mu \geq \mu_0\}$$

1. Test: $H_{0L} : \mu \leq \mu_0$ gegen $H_{1L} : \mu > \mu_0$ wird verworfen, falls

$$\sqrt{n}\frac{\bar{x} - \mu_0}{\sigma} \geq t_L$$

2. Test: $H_{0L} : \mu \geq \mu_0$ gegen $H_{1L} : \mu < \mu_0$ wird verworfen, falls

$$\sqrt{n}\frac{\bar{x} - \mu_0}{\sigma} \leq t_U$$

Für $t_L = -t_U \geq 0$ ist dies der zweiseitige t-Test:

$$\sqrt{n}\frac{|\ \bar{x} - \mu_0\ |}{\sigma} \geq t_L$$

Simultane Konfidenzbereiche

Nach dem Prinzip von Bonferroni konstruiert man für alle p Komponenten $\mu_j,\ j = 1, \ldots, p$, individuelle Konfidenzintervalle zum Niveau α_j, so dass gilt

$$\sum_{j=1}^{p} \alpha_j = \alpha\,, \qquad \text{Klassiker: } \alpha_j = \frac{\alpha}{p}$$

Sei E_j: 'Ereignis, dass das j-te Intervall den Parameter μ_j enthält', dann gilt:

$$P\Big(\bigcap_{j=1,\ldots,p} E_j\Big) = 1 - P\Big(\bigcup_{j=1,\ldots,p} \bar{E}_j\Big) \geq 1 - \sum_{j=1}^{p} P(\bar{E}_j) = 1 - \sum_{j=1}^{p} \alpha_j$$

2. Fall: $\mathbf{\Sigma}$ unbekannt

- Verwendung der empirischen Kovarianzmatrix $\mathbf{S}$ als Schätzer für $\mathbf{\Sigma}$

$\rightarrow$ T^2-Teststatistik von Hotelling:

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)\,,$$

- Folgt unter $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ einer Hotelling-Verteilung
- Es gilt:

$$\frac{n-p+1}{(np}\,T^2 \overset{H_0}{\sim} F(p, n-p+1)\,.$$

H_0 ablehnen, wenn gilt: $T^2 > F(p, n-p+1; 1-\alpha)$

Beispiel: Freitag, der 13.

- Untersuchung, ob Aberglaube so stark ist, dass Auswirkungen im Alltag nachweisbar sind
- Erhebung von Verkehrsaufkommen an zwei Straßenabschnitten, sowie Noteinlieferungen aufgrund von Verkehrsunfällen für Paare von Freitagen (Scanlon, Luben, Scanlon, Singleton (1993))

Differenz: Freitag, der 6. - Freitag, der 13.

X_1	X_2	X_3
698	1104	60
1037	1889	159
1911	2416	21
2761	4382	-33
1839	321	-123

Beispiel: Freitag, der 13.

- Zu überprüfende Hypothese: Menschen verhalten sich an einem Freitag, den 13., anders als an einem anderen Freitag
- Formale Nullhypothese: Differenzen müssen bei 0 zentriert sein

$$H_0 : \mu = 0$$

Beispiel: Freitag, der 13.

- Schätzer aus der Stichprobe:

$$\bar{\mathbf{x}} = \begin{pmatrix} 1649.2 \\ 2022.4 \\ 16.8 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 655058.2 & 839692.9 & -52237.2 \\ 839692.9 & 2369662.3 & 15839.1 \\ -52237.2 & 15839.1 & 11032.2 \end{pmatrix}$$

$$\mathbf{S}^{-1} = \begin{pmatrix} 0.00001952 & -0.00000761 & 0.00010336 \\ -0.00000761 & 0.00000339 & -0.00004090 \\ 0.00010336 & -0.00004090 & 0.00063876 \end{pmatrix}$$

- Hotellings T^2-Statistik: $T^2 = 5 \cdot \bar{\mathbf{x}}^\top \mathbf{S}^{-1} \bar{\mathbf{x}} = 96.7$
- Es gilt: $(5-3+1)/(5 \cdot 3) T^2 = 19.34$

$\rightarrow$ p-Wert: $\mathbb{P}((5-3+1)/(5 \cdot 3) T^2 \geq 19.34) = 0.018$

Tests für zwei Stichproben

Gegeben sind Daten

- $\mathbf{x}^1_{(1)}, \ldots, \mathbf{x}^1_{(n_1)} \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$
- $\mathbf{x}^2_{(1)}, \ldots, \mathbf{x}^2_{(n_2)} \sim \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$

Hypothesen

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

Unterscheidung:

1. Unabhängige Stichproben
2. Verbundene Stichproben

1. Fall: Unabhängige Stichproben

- Empirische Kovarianzmatrizen:

$$\mathbf{S}_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j^\top (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j \quad (j = 1, 2)$$

- Verschmelzen zu Gesamtschätzung (analog zum t-Test für unverbundene Stichproben):

$$\boldsymbol{S}_{pl} = \frac{1}{n_1 + n_2 - 2} \{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2\}$$

$\rightarrow$ Teststatistik

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \boldsymbol{S}_{pl}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

- Zusammenhang mit F-Verteilung:

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \overset{H_0}{\sim} F(p, n_1 + n_2 - p - 1) .$$

2. Fall: Verbundene Stichproben

In diesem Fall gilt:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \Leftrightarrow \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = 0$$

$\rightarrow$ entspricht einem Ein-Stichproben-Problem für die Differenzen

$$\boldsymbol{d}_{(i)} = \mathbf{x}^1_{(i)} - \mathbf{x}^2_{(i)}, \; i = 1, \ldots, n$$

Teststatistik (mit unbekannter Kovarianz)

$$T^2 = \frac{(n-p)n}{(n-1)p} \bar{\mathbf{d}}^\top \boldsymbol{S}_d^{-1} \bar{\mathbf{d}} \overset{H_0}{\sim} F(p, n-p)$$

Multivariate Varianzanalyse (MANOVA)

Gegeben ist eine Grundgesamtheit, die in g Gruppen partitioniert ist, d.h. die Daten

$$\mathbf{x}^r_{(1)}, \ldots, \mathbf{x}^r_{(n_r)} \sim \mathrm{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}), \quad r = 1, \ldots, g$$

Hypothesen

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_g \quad \text{vs.} \quad H_1 : \exists k, \ell : \boldsymbol{\mu}_k \neq \boldsymbol{\mu}_\ell$$

Erwartungswert in Gruppe r: $\boldsymbol{\mu}_r = \left(\mu^r_1, \ldots, \mu^r_p\right)^\top$

$n = n_1 + \cdots + n_r$

MANOVA

- Gesamtstreuung = Streuung in den Gruppen + Streuung zwischen den Gruppen
- $(n-1)\mathbf{S}_{gesamt} = \mathbf{W} + \mathbf{B}$ mit

$$\begin{aligned}
\mathbf{S}_{gesamt} &= \frac{1}{n-1} \sum_{r=1}^{g} \sum_{i=1}^{n_r} (\mathbf{x}_{ir} - \bar{\mathbf{x}})^\top (\mathbf{x}_{ir} - \bar{\mathbf{x}}) \\
\mathbf{W} &= \sum_{r=1}^{g} \sum_{i=1}^{n_r} (\mathbf{x}_{ir} - \bar{\mathbf{x}}_g)^\top (\mathbf{x}_{ir} - \bar{\mathbf{x}}_g) \\
\mathbf{B} &= \sum_{r=1}^{g} n_g (\bar{\mathbf{x}}_g - \bar{\mathbf{x}})^\top (\bar{\mathbf{x}}_g - \bar{\mathbf{x}})
\end{aligned}$$

MANOVA

- Teststatistik (Wilks Lambda):

$$\Lambda = |\mathbf{I} + \mathbf{W}^{-1}\mathbf{B}|^{-1} = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{B}|}$$

- Variabilität zwischen Gruppen sehr stark $\rightarrow$ Nenner deutlich größer als Zähler

$\rightarrow$ Lehne H_0 bei kleinen Werten von Λ ab

- Verteilung unter H_0 nur für spezielle Situationen bekannt
- Falls n groß, dann Approximation von Bartlett
- Es gilt: Lehne H_0 ab, wenn

$$-\left(n - 1 - \frac{p+g}{2}\right) \ln(\Lambda) > \chi^2_{p(g-1);1-\alpha}$$

Beispiel: Anzeigen in Magazinen

- Ziel: Untersuchung, ob Anzeigen in Magazinen nach dem Bildungsstand ihrer Leser ausgerichtet werden
- Klassifizierung von 30 Magazinen entsprechend dem Bildungs-Level ihrer Leser
- Auswahl von jeweils 3 Magazinen aus dem obersten, mittleren und untersten Niveau-Drittel
- Zufallsstichproben von jeweils 6 Anzeigen aus jedem Magazin
- Variblen:
  - $X_1 =$ Anzahl der Wörter,
  - $X_2 =$ Anzahl der Sätze,
  - $X_3 =$ Anzahl der Wörter mit mindestens 3 Silben

Beispiel: Anzeigen in Magazinen

Gruppe 1			Gruppe 2			Gruppe 3		
X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
205	9	34	191	25	13	162	14	16
203	20	21	219	17	22	31	6	9
229	18	37	205	23	25	85	11	10
208	16	31	57	7	3	111	12	3
146	9	10	105	10	5	88	11	12
...	...	...	...	...	...	...	...	...

Beispiel: Anzeigen in Magazinen

- Wilks Lambda: $\Lambda = 0.7997$
- Approximation von Bartlett:

$$-(54 - 1 - \frac{3+3}{2}) \ln(0.7997) = 11.176 \overset{a}{\sim} \chi^2(6, 1-\alpha)$$

- Zugehöriger p-Wert: 0.083
- Variation innerhalb der Gruppen übertrifft Variation zwischen den Gruppen:

$$\mathbf{W} = \begin{pmatrix} 219866.8 & 13013.4 & 29125.3 \\ 13013.4 & 1326.8 & 1896.9 \\ 29125.3 & 1896.9 & 5781.9 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 10141.8 & 191.1 & 1268.0 \\ 191.1 & 6.4 & -3.8 \\ 1268.0 & -3.8 & 435.6 \end{pmatrix}$$