Multivariate Verfahren

MSc Psychologie & MSc Klinische Psychologie und Psychotherapie
Wintersemester 2024/2025

Joram Soch

(5) Multivariate Deskriptivstatistik

Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen

Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen

Datenanalyseszenarien

UV	AV	Datenanalysemethoden
univariat	univariat	
multivariat	univariat	
univariat	multivariat	
multivariat	multivariat	

Datenanalyseszenarien

UV	AV
x_1	y_1
x_{11}	y_{11}
x_{12}	y_{12}
x_{13}	y_{13}
$\vdots$	$\vdots$
$\vdots$	$\vdots$
$\vdots$	$\vdots$
x_{1n}	y_{1n}

Korrelation, einfache lineare Regression, T-Tests, ANOVA

UV	AV
x_1	y_1
x_{11}	y_{11}
x_{12}	y_{12}
x_{13}	y_{13}
$\vdots$	$\vdots$
$\vdots$	$\vdots$
$\vdots$	$\vdots$
x_{1n}	y_{1n}

Datenanalyseszenarien

UV			AV
x_1	$\cdots$	x_m	y_1
x_{11}	$\cdots$	x_{m1}	y_{11}
x_{12}	$\cdots$	x_{m2}	y_{12}
x_{13}	$\cdots$	x_{m3}	y_{13}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n}	$\cdots$	x_{mn}	y_{1n}

multiple Korrelation, multiple Regression, Allgemeines Lineares Modell

UV			AV
x_1	$\cdots$	x_m	y_1
x_{11}	$\cdots$	x_{m1}	y_{11}
x_{12}	$\cdots$	x_{m2}	y_{12}
x_{13}	$\cdots$	x_{m3}	y_{13}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n}	$\cdots$	x_{mn}	y_{1n}

Datenanalyseszenarien

UV	AV		
x_1	y_1	$\cdots$	y_m
x_{11}	y_{12}	$\cdots$	y_{m1}
x_{12}	y_{13}	$\cdots$	y_{m2}
x_{13}	y_{14}	$\cdots$	y_{m3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n}	y_{1n}	$\cdots$	y_{mn}

T^2-Tests, einfaktorielle multivariate Varianzanalyse (MANOVA)

UV	AV		
x_1	y_1	$\cdots$	y_m
x_{11}	y_{12}	$\cdots$	y_{m1}
x_{12}	y_{13}	$\cdots$	y_{m2}
x_{13}	y_{14}	$\cdots$	y_{m3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n}	y_{1n}	$\cdots$	y_{mn}

Datenanalyseszenarien

UV			AV		
x_1	$\cdots$	x_{m_x}	y_1	$\cdots$	y_{m_y}
x_{11}	$\cdots$	$x_{m_x 1}$	y_{11}	$\cdots$	$y_{m_y 1}$
x_{12}	$\cdots$	$x_{m_x 2}$	y_{12}	$\cdots$	$y_{m_y 2}$
x_{13}	$\cdots$	$x_{m_x 3}$	y_{13}	$\cdots$	$y_{m_y 3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n}	$\cdots$	$x_{m_x n}$	y_{1n}	$\cdots$	$y_{m_y n}$

Kanonische Korrelationsanalyse, multivariates Allgemeines Lineares Modell

UV			AV		
x_1	$\cdots$	x_{m_x}	y_1	$\cdots$	y_{m_y}
x_{11}	$\cdots$	$x_{m_x 1}$	y_{11}	$\cdots$	$y_{m_y 1}$
x_{12}	$\cdots$	$x_{m_x 2}$	y_{12}	$\cdots$	$y_{m_y 2}$
x_{13}	$\cdots$	$x_{m_x 3}$	y_{13}	$\cdots$	$y_{m_y 3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{1n}	$\cdots$	$x_{m_x n}$	y_{1n}	$\cdots$	$y_{m_y n}$

Datenanalyseszenarien

UV	AV	Datenanalysemethoden
univariat	univariat	Korrelation, einfache lineare Regression, T-Tests, ANOVA
multivariat	univariat	multiple Korrelation, multiple Regression, Allgemeines Lineares Modell
univariat	multivariat	T^2-Tests, einfaktorielle multivariate Varianzanalyse (MANOVA)
multivariat	multivariat	Kanonische Korrelation, multivariates Allgemeines Lineares Modell

Multivariate Generalisierungen bekannter Frequentistischer Verfahren

(Einstichproben-T^2-Tests als Generalisierung von Einstichproben-T-Tests)

- Inferenz für ein bis zwei Gruppen mit multivariaten Daten

Multivariate Varianzanalyse als Generalisierung der einfaktoriellen Varianzanalyse

- Inferenz für drei oder mehr Gruppen mit multivariaten Daten

Kanonische Korrelationsanalyse als Generalisierung der Berechnung von Korrelationen

- Zusammenhangsmaß für multivariate unabhängige und abhängige Variablen

Zur Wiederholung univariater Frequentistischer Verfahren bietet sich an:

- Einheiten (8)–(11) von Wahrscheinlichkeitstheorie und Frequentistische Inferenz
- Einheiten (9)–(14) von Allgemeines Lineares Modell

Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen

Definition (Multivariate Deskriptivstatistiken)

$y_1, ..., y_n$ seien m-dimensionale Zufallsvektoren.

- Das *Stichprobenmittel* der $y_1, ..., y_n$ ist definiert als der m-dimensionale Vektor

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

- Die *Stichprobenkovarianzmatrix* der $y_1, ..., y_n$ ist definiert als die $m \times m$-dimensionale Matrix

$$C := \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}}. \tag{2}$$

- Die *Stichprobenkorrelationsmatrix* der $y_1, ..., y_n$ definiert als die $m \times m$-dimensionale Matrix

$$R := \left(\frac{(C)_{ij}}{\sqrt{(C)_{ii}}\sqrt{(C)_{jj}}} \right)_{1 \leq i,j \leq m}. \tag{3}$$

Bemerkungen

- Bei unabhängig und identisch verteilten $y_1, ..., y_n$ ist $\bar{y}$ ein unverzerrter Schätzer von $\mathbb{E}(y_i), i = 1, ..., n$.
- Bei unabhängig und identisch verteilten $y_1, ..., y_n$ ist C ein unverzerrter Schätzer von $\mathbb{C}(y_i), i = 1, ..., n$.
- Wir bezeichnen hier mit y_i sowohl den Zufallsvektor, über dessen Wahrscheinlichkeitsverteilung wir Aussagen machen können (vormals v_i), als auch konkrete Realisierungen dieses Zufallsvektors, die einen gegebenen Datensatz darstellen $(y_1, ..., y_n)$.

Theorem (Datenmatrix und multivariate Deskriptivstatistiken)

$$Y := \begin{pmatrix} y_1 & \cdots & y_n \end{pmatrix} \tag{4}$$

sei eine $m \times n$ *Datenmatrix*, die durch die spaltenweise Konkatenation der m-dimensionalen Zufallvektoren $y_1, \ldots, y_n$ gegeben ist. Dann ergeben sich

- für das Stichprobenmittel

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

- für die Stichprobenkovarianzmatrix

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

- und mit

$$D := \mathrm{diag} \left(\frac{1}{\sqrt{(C)_{11}}}, \ldots, \frac{1}{\sqrt{(C)_{mm}}} \right) \tag{7}$$

für die Stichprobenkorrelationsmatrix

$$R = DCD \tag{8}$$

Bemerkungen

- Das Theorem erlaubt eine mathematisch konzise Darstellung von $\bar{y}, C$ und R.
- Das Theorem erlaubt eine programmatisch effiziente Berechung von $\bar{y}, C$ und R.

Beweis

Die Darstellung des Stichprobenmittels ergibt sich nach durch

$$\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n} y_{i1} \\ \vdots \\ \sum_{i=1}^{n} y_{im} \end{pmatrix} = \frac{1}{n}\left(\begin{pmatrix} y_{11} & \cdots & y_{n1} \\ \vdots & \ddots & \vdots \\ y_{1m} & \cdots & y_{nm} \end{pmatrix}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right) = \frac{1}{n} Y 1_n. \tag{9}$$

Hinsichtlich der Darstellung der Stichprobenkovarianzmatrix halten wir zunächst fest, dass gilt:

$$\begin{aligned} C &:= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}} \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i y_i^{\mathrm{T}} - y_i \bar{y}^{\mathrm{T}} - \bar{y} y_i^{\mathrm{T}} + \bar{y}\bar{y}^{\mathrm{T}}\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i y_i^{\mathrm{T}} - \sum_{i=1}^{n} y_i \bar{y}^{\mathrm{T}} - \sum_{i=1}^{n} \bar{y} y_i^{\mathrm{T}} + \sum_{i=1}^{n} \bar{y}\bar{y}^{\mathrm{T}}\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i y_i^{\mathrm{T}} - n\bar{y}\bar{y}^{\mathrm{T}} - n\bar{y}\bar{y}^{\mathrm{T}} + n\bar{y}\bar{y}^{\mathrm{T}}\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i y_i^{\mathrm{T}} - n\bar{y}\bar{y}^{\mathrm{T}}\right). \end{aligned} \tag{10}$$

Beweis (fortgeführt)

Mit $1_n 1_n^{\mathrm{T}} = 1_{nn}$ ergibt sich dann weiterhin

$$\begin{aligned}
Y\left(I_n - \frac{1}{n}1_{nn}\right)Y^{\mathrm{T}} &= \left(YI_n - \frac{1}{n}Y1_{nn}\right)Y^{\mathrm{T}} \\
&= YY^{\mathrm{T}} - \frac{1}{n}Y1_{nn}Y^{\mathrm{T}} \\
&= \begin{pmatrix} y_1 & \cdots & y_n \end{pmatrix}\begin{pmatrix} y_1^{\mathrm{T}} \\ \vdots \\ y_n^{\mathrm{T}} \end{pmatrix} - \frac{1}{n}Y1_n 1_n^{\mathrm{T}}Y^{\mathrm{T}} \\
&= \sum_{i=1}^{n} y_i y_i^{\mathrm{T}} - n\left(\frac{1}{n}Y1_n\right)\left(\frac{1}{n}1_n^{\mathrm{T}}Y^{\mathrm{T}}\right) \qquad (11) \\
&= \sum_{i=1}^{n} y_i y_i^{\mathrm{T}} - n\bar{y}\bar{y}^{\mathrm{T}} \\
&= \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}} \\
&= (n-1)\,C \qquad \Leftrightarrow \qquad C = \frac{1}{n-1}Y\left(I_n - \frac{1}{n}1_{nn}\right)Y^{\mathrm{T}}
\end{aligned}$$

Schließlich ergibt sich für die Korrelationsmatrix für ein beliebiges Indexpaar i, j mit $1 \leq i, j \leq m$, dass

$$R_{ij} = \frac{(C)_{ij}}{\sqrt{(C)_{ii}}\sqrt{(C)_{jj}}} = \frac{1}{\sqrt{(C)_{ii}}}(C)_{ij}\frac{1}{\sqrt{(C)_{jj}}} = (DCD)_{ij}. \qquad (12)$$

□

Definition (Mahalanobis-Distanz)

y_1 sei ein Zufallsvektor, eine Realisation eines Zufallsvektors, ein multivariater Erwartungswert oder ein multivariates Stichprobenmittel, y_2 sei ein Zufallsvektor, eine Realisation eines Zufallsvektors, ein multivariater Erwartungswert oder ein multivariates Stichprobenmittel und C sei eine Kovarianzmatrix oder eine Stichprobenkovarianzmatrix. Dann heißt

$$D = (y_1 - y_2)^{\mathrm{T}} C^{-1} (y_1 - y_2) \tag{13}$$

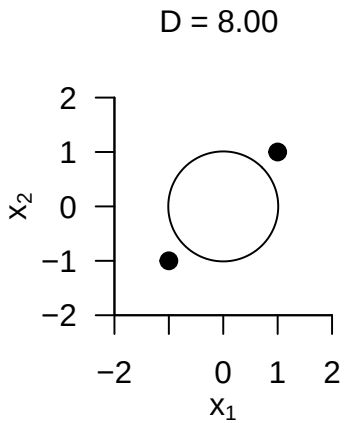
Mahalanobis-Distanz von y_1 und y_2 hinsichtlich C.

Bemerkungen

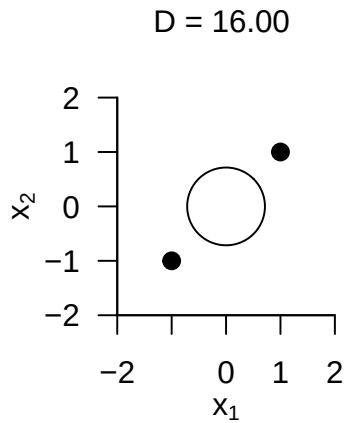
- Eine Mahalanobis-Distanz ist eine Kovarianzmatrix-normalisierte quadrierte Euklidische Distanz.
- Ähnliche Maße in der univariaten Statistik sind die z-Transformation $z = \frac{y-\mu}{\sigma}$ und Cohen's $d = \frac{\bar{y}_1 - \bar{y}_2}{s_{12}}$.
- Ähnlich wie bei z-Werten wird bei der Mahalanobis Distanz in "Einheiten von Kovarianzen" gemessen.
- Stark variante Komponenten von y_1 und y_2 tragen weniger zur Distanz bei.
- Stark kovariante Komponenten von y_1 und y_2 tragen weniger zur Distanz bei.

Mahalanobis-Distanzen als Funktion von Komponentenvarianzen

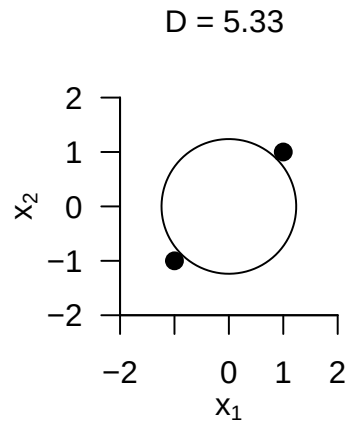
$$\Sigma := \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}$$



$$\Sigma := \begin{pmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{pmatrix}$$

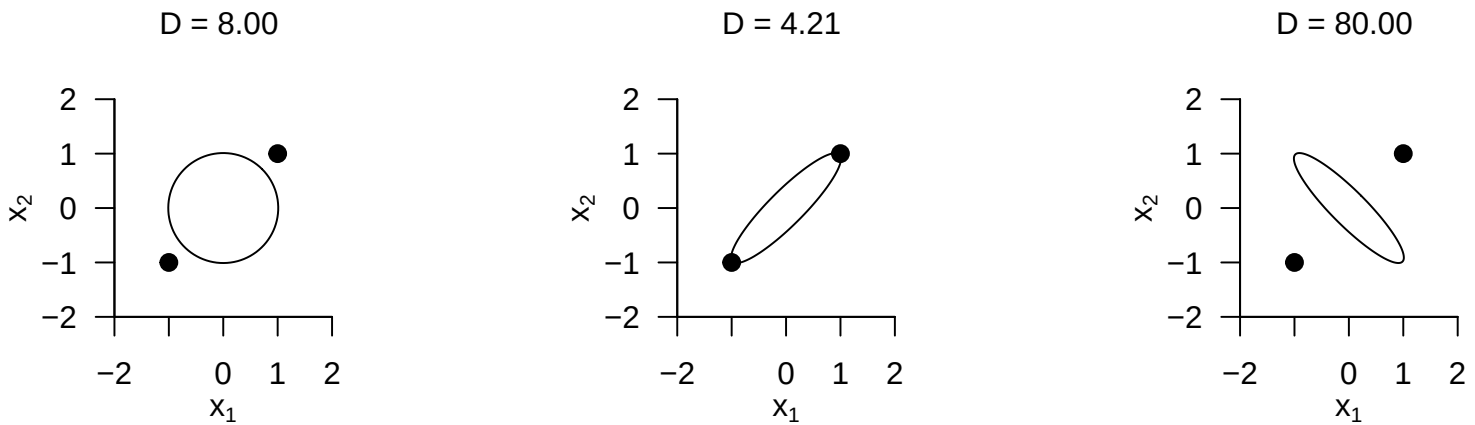


$$\Sigma := \begin{pmatrix} 1.5 & 0.0 \\ 0.0 & 1.5 \end{pmatrix}$$



Mahalanobis-Distanzen als Funktion von Komponentenkovarianzen

$$\Sigma := \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix} \qquad \Sigma := \begin{pmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{pmatrix} \qquad \Sigma := \begin{pmatrix} 1.0 & -0.9 \\ -0.9 & 1.0 \end{pmatrix}$$

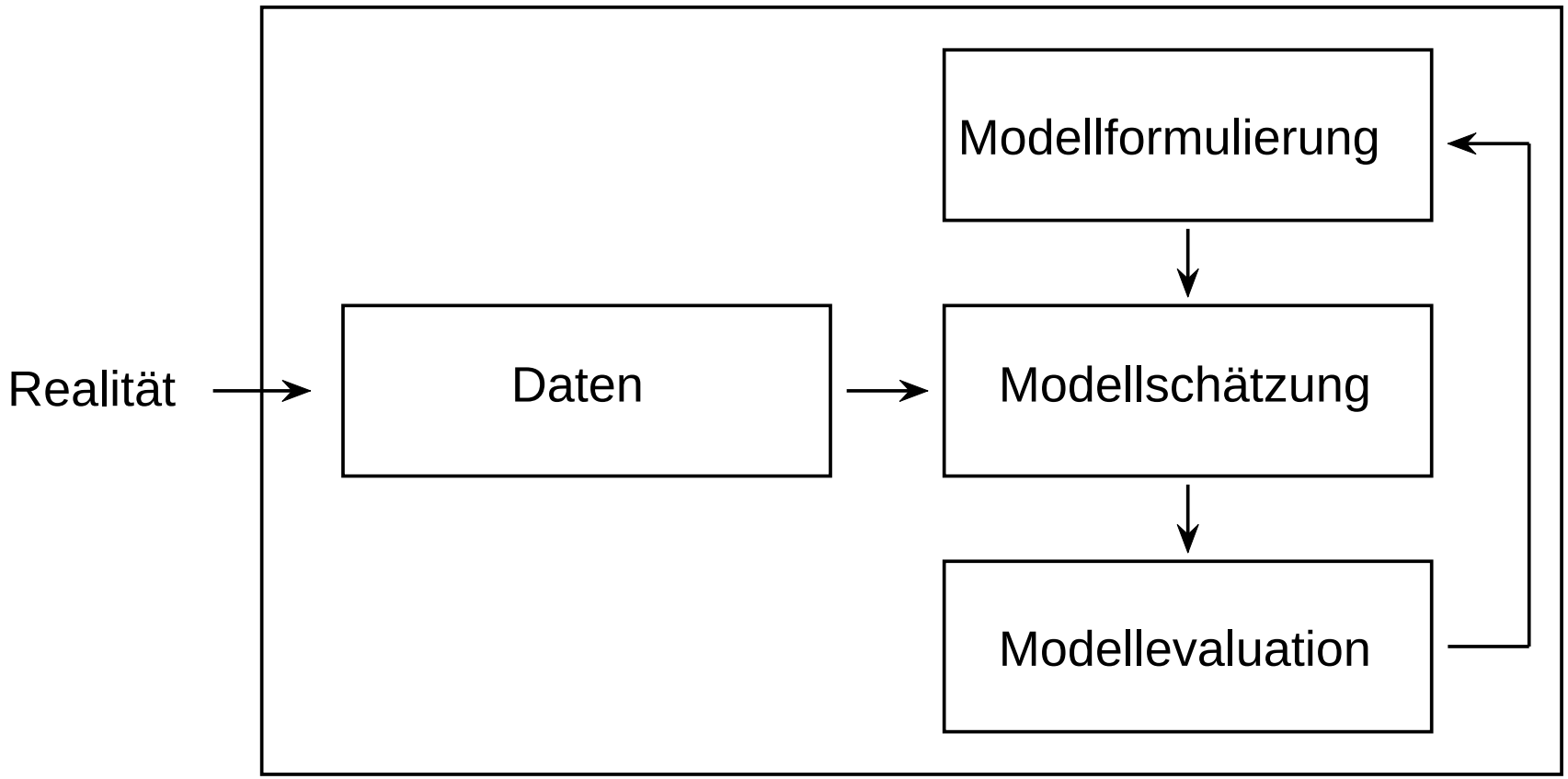


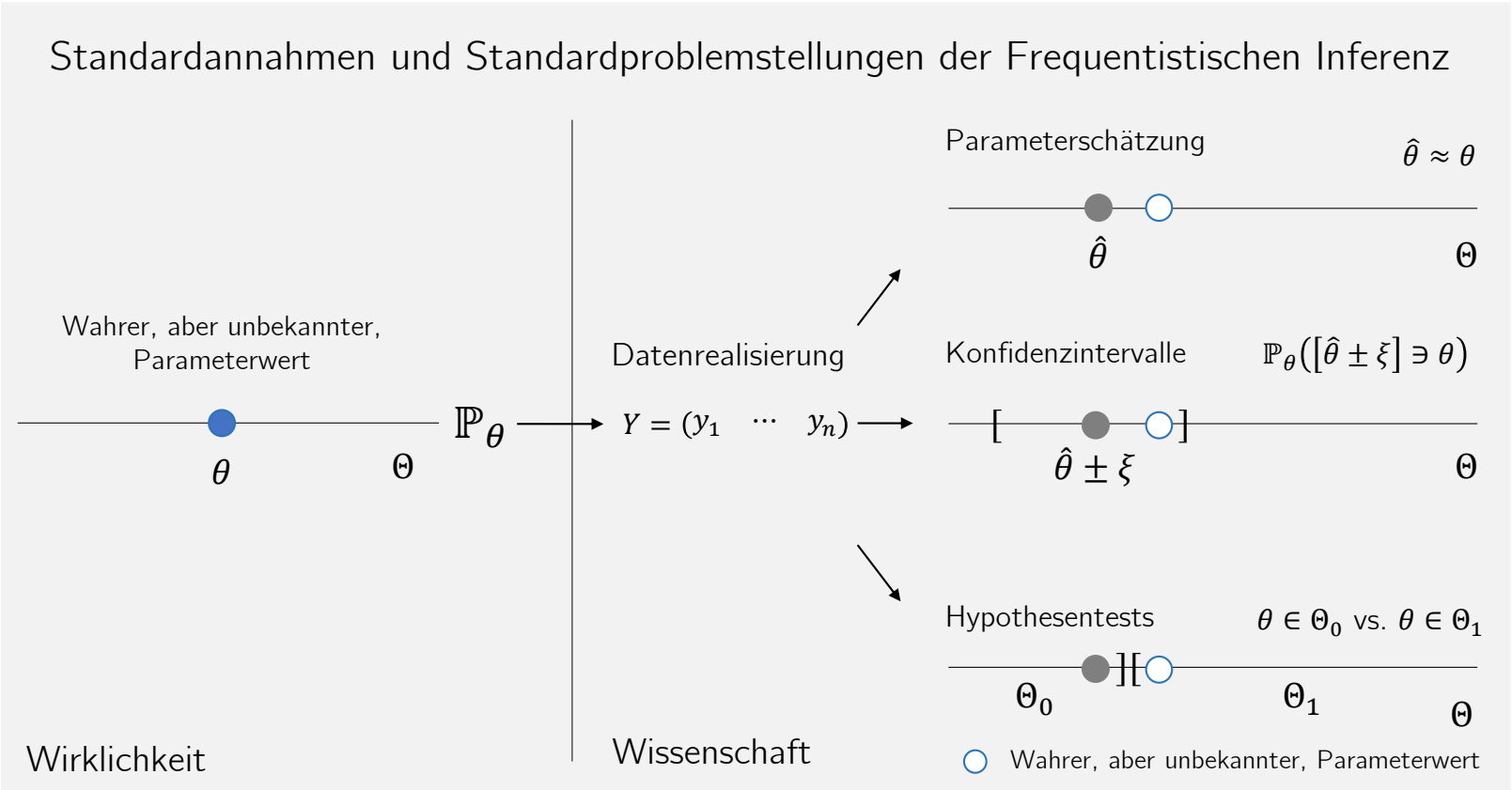
Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen


Realität
Daten
Modellformulierung
Modellschätzung
Modellevaluation


Standardannahmen und Standardproblemstellungen der Frequentistischen Inferenz
Wahrer, aber unbekannter, Parameterwert
θ
Θ
$\mathbb{P}_\theta$
Datenrealisierung
$Y = (y_1 \quad \cdots \quad y_n)$
Parameterschätzung
$\hat{\theta} \approx \theta$
$\hat{\theta}$
Θ
Konfidenzintervalle
$\mathbb{P}_\theta([\hat{\theta} \pm \xi] \ni \theta)$
$\hat{\theta} \pm \xi$
Θ
Hypothesentests
$\theta \in \Theta_0$ vs. $\theta \in \Theta_1$
Θ_0
Θ_1
Θ
Wirklichkeit
Wissenschaft
Wahrer, aber unbekannter, Parameterwert

Standardannahmen Frequentistischer Inferenz

- $\mathcal{M}$ sei ein Frequentistisches Inferenzmodell mit $y_1, ..., y_n \sim p_\theta$. Es wird angenommen, dass eine konkrete Datenmatrix $Y \in \mathbb{R}^{m \times n}$ eine der möglichen Realisierungen von $\begin{pmatrix} y_1 & \cdots & y_n \end{pmatrix}$ ist.
- Aus Frequentistischer Sicht kann man eine Studie unendlich oft wiederholen und zu jedem Datensatz Schätzer oder Statistiken auswerten, z.B. das Stichprobenmittel:

$$\text{Datensatz (1)} : Y^{(1)} = \begin{pmatrix} y_1^{(1)} & \cdots & y_n^{(1)} \end{pmatrix} \text{ mit } \bar{y}^{(1)} = \frac{1}{n} \sum_{i=1}^{n} y_i^{(1)}$$

$$\text{Datensatz (2)} : Y^{(2)} = \begin{pmatrix} y_1^{(2)} & \cdots & y_n^{(2)} \end{pmatrix} \text{ mit } \bar{y}^{(2)} = \frac{1}{n} \sum_{i=1}^{n} y_i^{(2)}$$

$$\text{Datensatz (3)} : Y^{(3)} = \begin{pmatrix} y_1^{(3)} & \cdots & y_n^{(3)} \end{pmatrix} \text{ mit } \bar{y}^{(3)} = \frac{1}{n} \sum_{i=1}^{n} y_i^{(3)}$$

$$\text{Datensatz (4)} : Y^{(4)} = \begin{pmatrix} y_1^{(4)} & \cdots & y_n^{(4)} \end{pmatrix} \text{ mit } \bar{y}^{(4)} = \frac{1}{n} \sum_{i=1}^{n} y_i^{(4)}$$

$$\text{Datensatz (5)} : Y^{(5)} = ...$$

- Um die Qualität statistischer Methoden zu beurteilen, betrachtet die Frequentistische Statistik deshalb die Wahrscheinlichkeitsverteilungen von Schätzern und Statistiken unter Annahme von $y_1, ..., y_n \sim p_\theta$. Was zum Beispiel ist die Verteilung der $\bar{y}^{(1)}$, $\bar{y}^{(2)}$, $\bar{y}^{(3)}$, $\bar{y}_n^{(4)}$, ..., also die Verteilung der Zufallsvariable $\bar{y}$?
- Wenn eine statistische Methode im Sinne der Frequentistischen Standardannahmen "gut" ist, dann heißt das also, dass sie bei häufiger Anwendung "im Mittel gut" ist. Im Einzelfall, also im Normalfall nur eines vorliegenden Datensatzes, kann sie auch "schlecht" sein.

Standardprobleme Frequentistischer Inferenz

(1) Parameterschätzung

Ziel der Parameterschätzung ist es, einen möglichst guten Tipp für wahre, aber unbekannte, Parameterwerte oder Funktionen dieser abzugeben, typischerweise mithilfe von Daten.

(2) Konfidenzintervalle

Ziel der Bestimmung von Konfidenzintervallen ist es, basierend auf der angenommenen Verteilung der Daten eine quantitative Aussage über die mit Schätzwerten assoziierte Unsicherheit zu treffen.

(3) Hypothesentests

Ziel des Hypothesentestens ist es, basierend auf der angenommenen Verteilung der Daten in einer möglichst zuverlässigen Form zu entscheiden, ob ein wahrer, aber unbekannter Parameterwert in einer von zwei sich gegenseitig ausschließenden Untermengen des Parameterraumes liegt.

Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen

Selbstkontrollfragen

1. Erläutern Sie die Begriffe der unabhängigen und abhängigen Variablen.
2. Erläutern Sie die Unterschiede zwischen Datenanalyseszenarien im Hinblick darauf, ob ihre unabhängigen und abhängigen Variablen univariater oder multivariater Natur sind. Nennen Sie jeweils ein Beispiel.
3. Geben Sie die Definition des multivariaten Stichprobenmittels wieder.
4. Geben Sie die Definition des Stichprobenkovarianzmatrix wieder.
5. Geben Sie die Definition des Stichprobenkorrelationsmatrix wieder.
6. Erläutern Sie die Berechnung des Stichprobenmittels aus der Datenmatrix mithilfe eines Matrixprodukts.
7. Geben Sie die Definition der Mahalanobis-Distanz wieder.
8. Wie verändert sich die Mahalanobis-Distanz in Abhängigkeit von den Diagonaleinträgen der Kovarianzmatrix bei gleichbleibenden Erwartungswerten?
9. Wie verändert sich die Mahalanobis-Distanz in Abhängigkeit von den Nicht-Diagonaleinträgen der Kovarianzmatrix bei gleichbleibenden Erwartungswerten?
10. Erläutern Sie die Standardprobleme der Frequentistischen Inferenz.