Multivariate Verfahren

MSc Psychologie | MSc Klinische Psychologie und Psychotherapie

WiSe 2023/24

Prof. Dr. Dirk Ostwald

(4) Deskription und Inferenz

Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen

Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen

Datenanalyseszenarien

UV	AV	Datenanalysemethoden
Univariat	Univariat	Korrelation, Einfache Regression,T-Tests
Multivariat	Univariat	Multiple Korrelation, Multiple Regression, Allgemeines Lineares Modell
Univariat	Multivariat	Einstichproben-T^2-Tests, Einfaktorielle multivariate Varianzanalyse
Multivariat	Multivariat	Kanonische Korrelation, Multivariates Allgemeines Lineares Modell

Korrelation, Einfache Regression, T-Tests

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}

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}

Einstichproben-T^2-Tests, Einfaktorielle multivariate Varianzanalyse

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}

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}	$\cdots$	$y_{m_y n}$

Multivariate Generalisierungen bekannter Frequentistischer Verfahren WiSe 23/24

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

- Inferenz für ein bis zwei Gruppen multivariater Daten

Einfaktorielle multivariate Varianzanalyse als Generalisierung der einfaktoriellen Varianzanalyse

- Inferenz für drei oder mehr Gruppen multivariater Daten

Kanonische Korrelationsanalyse als Generalisierung der Korrelation

- Zusammenhangsmaß für multivariate unabhängige und abhängige Variablen

Zur Revision univariater Frequentistischer Verfahren

- Wahrscheinlichkeitstheorie und Frequentistische Inferenz 2022/23
- Allgemeines Lineares Modell 2023

Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen

Definition (Multivariate Deskriptivstatistiken)

$\upsilon_1, ..., \upsilon_n$ seien m-dimensionale Zufallsvektoren.

- Das *Stichprobenmittel* der $\upsilon_1, ..., \upsilon_n$ ist definiert als der m-dimensionale Vektor

$$\bar{\upsilon} := \frac{1}{n} \sum_{i=1}^{n} \upsilon_i. \tag{1}$$

- Die *Stichprobenkovarianzmatrix* der $\upsilon_1, ..., \upsilon_n$ ist definiert als die $m \times m$-dimensionale Matrix

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

- Die *Stichprobenkorrelationsmatrix* der $\upsilon_1, ..., \upsilon_n$ 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 $\upsilon_1, ..., \upsilon_n$ ist $\bar{\upsilon}$ ein unverzerrter Schätzer von $\mathbb{E}(\upsilon_i), i = 1, ..., n$.
- Bei unabhängig und identisch verteilten $\upsilon_1, ..., \upsilon_n$ ist C ein unverzerrter Schätzer von $\mathbb{C}(\upsilon_i), i = 1, ..., n$.

Theorem (Datenmatrix und multivariate Deskriptivstatistiken)

$$\Upsilon := \begin{pmatrix} \upsilon_1 & \cdots & \upsilon_n \end{pmatrix} \tag{4}$$

sei eine $m \times n$ *Datenmatrix*, die durch die spaltenweise Konkatenation von m-dimensionaler Zufallvektoren $\upsilon_1, \ldots, \upsilon_n$ gegeben sei. Dann ergeben sich

- für das Stichprobenmittel

$$\bar{\upsilon} = \frac{1}{n} \Upsilon 1_n, \tag{5}$$

- für die Stichprobenkovarianzmatrix

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

- und mit

$$D := \text{diag} \left(\sqrt{(C)_{ii}}^{-1}, i = 1, \ldots, m \right) \tag{7}$$

für die Stichprobenkorrelationsmatrix

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

Bemerkungen

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

Beweis

Die Darstellung des Stichprobenmittels ergibt sich aus

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

Die Darstellung der Korrelationsmatrix ergibt sich für ein beliebiges Indexpaar i, j mit $1 \leq i, j \leq m$ aus

$$(R)_{ij} := \frac{(C)_{ij}}{\sqrt{(C)_{ii}}\sqrt{(C)_{jj}}} = \frac{1}{\sqrt{(C)_{ii}}}(C)_{ij}\frac{1}{\sqrt{(C)_{jj}}} = (DCD)_{ij}. \tag{10}$$

Die Darstellung der Stichprobenkovarianzmatrix schließlich ergibt sich mit $1_n 1_n^T = 1_{nn}$ aus

$$
\begin{aligned}
C &:= \frac{1}{n-1}\sum_{i=1}^{n}(\upsilon_i - \bar{\upsilon})(\upsilon_i - \bar{\upsilon})^T \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\upsilon_i\upsilon_i^T - \upsilon_i\bar{\upsilon}^T - \bar{\upsilon}\upsilon_i^T + \bar{\upsilon}\bar{\upsilon}^T\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\upsilon_i\upsilon_i^T - \sum_{i=1}^{n}\upsilon_i\bar{\upsilon}^T - \sum_{i=1}^{n}\bar{\upsilon}\upsilon_i^T + \sum_{i=1}^{n}\bar{\upsilon}\bar{\upsilon}^T\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\upsilon_i\upsilon_i^T - n\bar{\upsilon}\bar{\upsilon}^T - n\bar{\upsilon}\bar{\upsilon}^T + n\bar{\upsilon}\bar{\upsilon}^T\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\upsilon_i\upsilon_i^T - n\bar{\upsilon}\bar{\upsilon}^T\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\upsilon_i\upsilon_i^T - n\left(\frac{1}{n}\Upsilon 1_n\right)\left(\frac{1}{n}1_n^T\Upsilon^T\right)\right) \\
&= \frac{1}{n-1}\left(\begin{pmatrix}\upsilon_1 & \cdots & \upsilon_n\end{pmatrix}\begin{pmatrix}\upsilon_1^T \\ \vdots \\ \upsilon_n^T\end{pmatrix} - \frac{1}{n}\Upsilon 1_n 1_n^T\Upsilon^T\right) \\
&= \frac{1}{n-1}\left(\Upsilon\Upsilon^T - \frac{1}{n}\Upsilon 1_{nn}\Upsilon^T\right) \\
&= \frac{1}{n-1}\left(\left(\Upsilon I_n - \frac{1}{n}\Upsilon 1_{nn}\right)\Upsilon^T\right) \\
&= \frac{1}{n-1}\left(\Upsilon\left(I_n - \frac{1}{n}1_{nn}\right)\Upsilon^T\right)
\end{aligned}
\tag{11}
$$

Definition (Mahalanobis Distanz)

ξ_1 sei ein Zufallsvektor, eine Realisation eines Zufallsvektors, ein multivariater Erwartungswert oder ein multivariates Stichprobenmittel, ξ_2 sei ein Zufallsvektor, eine Realisation eines Zufallsvektors, ein multivariater Erwartungswert oder ein multivariates Stichprobenmittel und Ξ sei eine Kovarianzmatrix oder eine Stichprobenkovarianzmatrix. Dann heißt

$$D = (\xi_1 - \xi_2)^T \, \Xi^{-1} \, (\xi_1 - \xi_2) \tag{12}$$

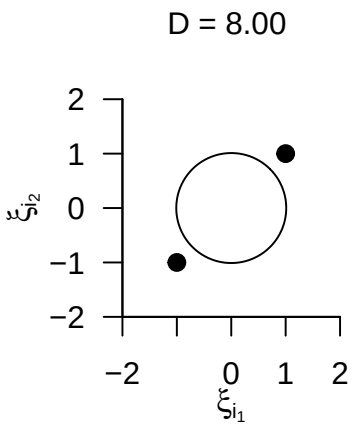
Mahalanobis Distanz von ξ_1 und ξ_2 hinsichtlich Ξ.

Bemerkungen

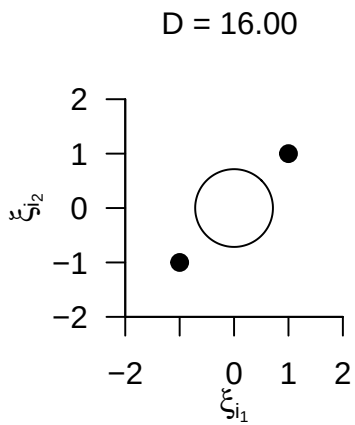
- Eine Mahalanobis Distanz ist eine Kovarianzmatrix-normalisierte quadrierte Euklidische Distanz.
- Ähnliche Maße in der univariaten Statistik sind die z-Transformation $z = \frac{y-\mu}{\sigma}$ und Cohen's $d = \frac{\bar{v}_1 - \bar{v}_2}{s_{12}}$.
- Ähnlich wie bei z-Werten wird bei der Mahalanobis Distanz in "Einheiten von Kovarianzen" gemessen.
- Stark variante Komponenten von ξ_1 und ξ_2 tragen weniger zur Distanz bei.
- Stark kovariante Komponenten von ξ_1 und ξ_2 tragen weniger zur Distanz bei.

Mahalanobis Distanzen als Funktion von Komponentenvarianzen

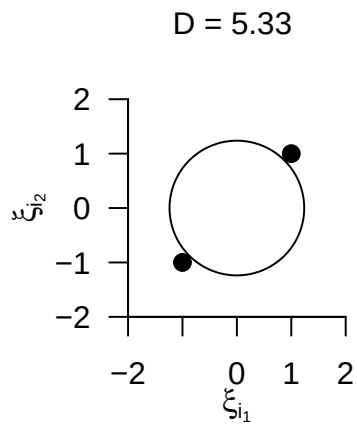
$$\Sigma := \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}$$



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

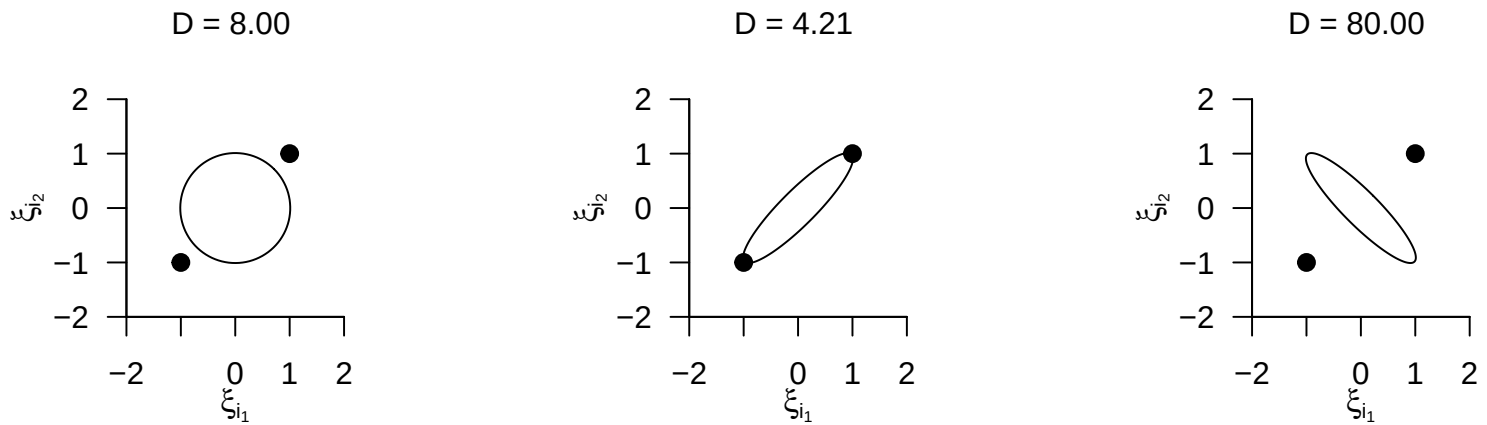


$$\Sigma := \begin{pmatrix} 1.5 & 0.0 \\ 0.0 & 1.5 \end{pmatrix}$$



Mahalanobis Distanzen als Funktion von Komponentenkovarianzen

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

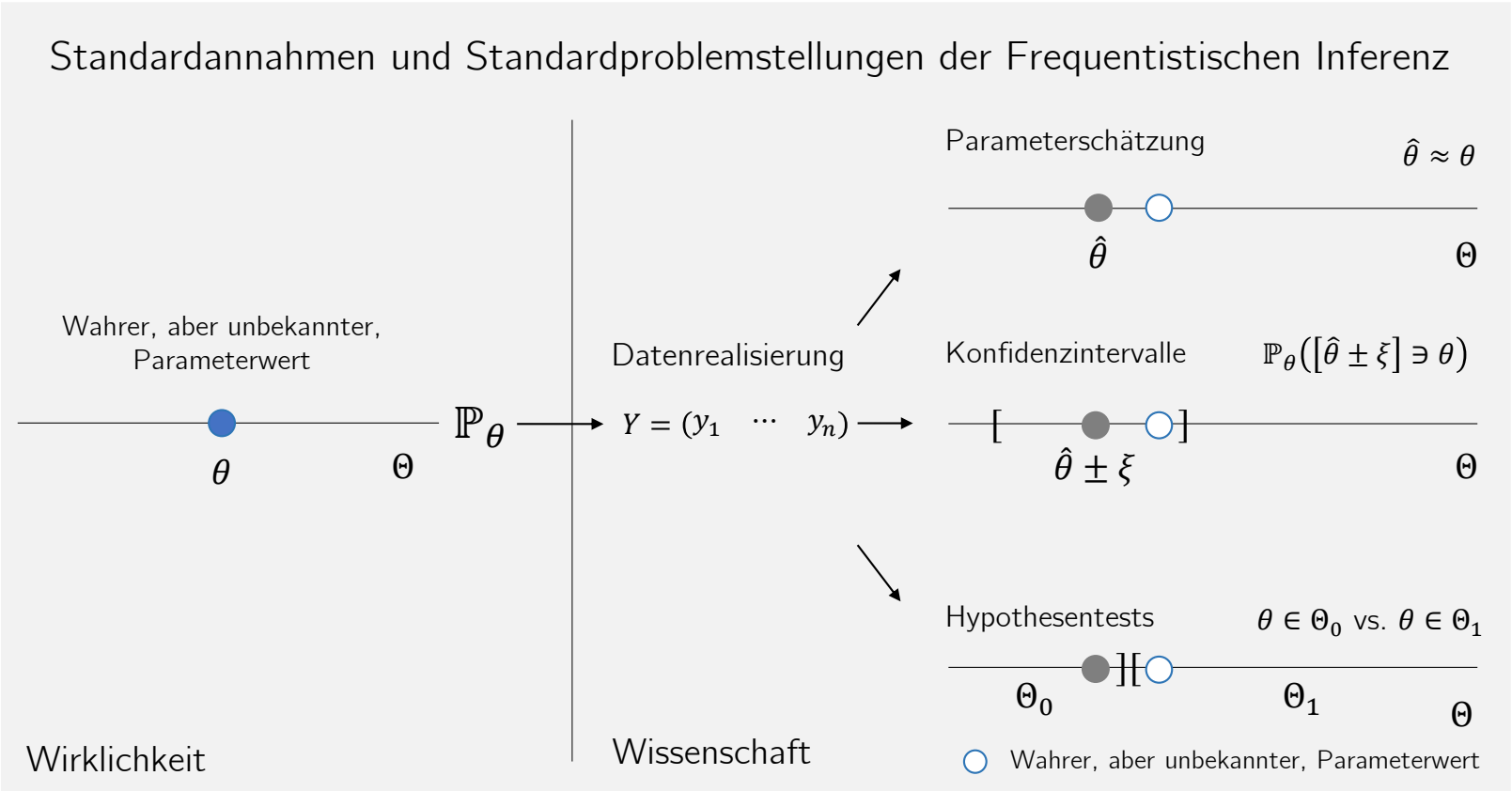


Datenanalyseszenarien

Multivariate Deskriptivstatistiken

Grundlagen Frequentistischer Inferenz

Selbstkontrollfragen


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 $\upsilon_1, ..., \upsilon_n \sim p_\theta$. Es wird angenommen, dass eine konkrete Datenmatrix $Y \in \mathbb{R}^{m \times n}$ eine der möglichen Realisierungen von $\Upsilon = \begin{pmatrix} \upsilon_1 & \cdots & \upsilon_n \end{pmatrix}$ 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 $\upsilon_1, ..., \upsilon_n \sim p_\theta$. 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 $\bar{\upsilon}$?
- 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.

Standardproblemstellungen Frequentistischer Inferenz

(1) Parameterschätzung

Ziel der Parameterschätzung ist es, einen möglichst guten Tipp für wahre, aber unbekannte, Parameterwerte oder Funktionen dieser abzugeben, typischerweise mithilfe der 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 vier prinzipielle Datenanalyseszenarien anhand der Dimensionalität ihrer UV und AV.
2. Nennen Sie Beispiele für die in den vier Datenanalyseszenarien häufig eingesetzten Datenanalyseverfahren.
3. Geben Sie die Definition des Stichprobenmittels wieder.
4. Geben Sie die Definition der Stichprobenkovarianzmatrix wieder.
5. Geben Sie die Definition des Stichprobenkorrelationsmatrix wieder.
6. Geben Sie das Theorem zu Datenmatrix und Stichprobenstatistiken wieder.
7. Geben Sie die Definition einer Mahalanobis-Distanz wieder.
8. Erläutern Sie die intuitive Bedeutung einer Mahalanobis-Distanz.