



# 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 <sup>2</sup> -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
$x_1$	...	$x_m$	$y_1$
$x_{11}$	...	$x_{m1}$	$y_{11}$
$x_{12}$	...	$x_{m2}$	$y_{12}$
$x_{13}$	...	$x_{m3}$	$y_{13}$
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
$x_{1n}$	...	$x_{mn}$	$y_{1n}$

## Einstichproben- $T^2$ -Tests, Einfaktorielle multivariate Varianzanalyse

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

## Kanonische Korrelationsanalyse

### Multivariates Allgemeines Lineares Modell

UV			AV		
$x_1$	...	$x_{m_x}$	$y_1$	...	$y_{m_y}$
$x_{11}$	...	$x_{m_x 1}$	$y_{11}$	...	$y_{m_y 1}$
$x_{12}$	...	$x_{m_x 2}$	$y_{12}$	...	$y_{m_y 2}$
$x_{13}$	...	$x_{m_x 3}$	$y_{13}$	...	$y_{m_y 3}$
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
$x_{1n}$	...	$x_{m_x n}$	$y_{1n}$	...	$y_{m_y n}$

## 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)

$v_1, \dots, v_n$  seien  $m$ -dimensionale Zufallsvektoren.

- Das *Stichprobenmittel* der  $v_1, \dots, v_n$  ist definiert als der  $m$ -dimensionale Vektor

$$\bar{v} := \frac{1}{n} \sum_{i=1}^n v_i. \quad (1)$$

- Die *Stichprobenkovarianzmatrix* der  $v_1, \dots, v_n$  ist definiert als die  $m \times m$ -dimensionale Matrix

$$C := \frac{1}{n-1} \sum_{i=1}^n (v_i - \bar{v})(v_i - \bar{v})^T. \quad (2)$$

- Die *Stichprobenkorrelationsmatrix* der  $v_1, \dots, v_n$  definiert als die  $m \times m$ -dimensionale Matrix

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

### Bemerkungen

- Bei unabhängig und identisch verteilten  $v_1, \dots, v_n$  ist  $\bar{v}$  ein unverzerrter Schätzer von  $\mathbb{E}(v_i)$ ,  $i = 1, \dots, n$ .
- Bei unabhängig und identisch verteilten  $v_1, \dots, v_n$  ist  $C$  ein unverzerrter Schätzer von  $\mathbb{C}(v_i)$ ,  $i = 1, \dots, n$ .

## Theorem (Datenmatrix und multivariate Deskriptivstatistiken)

$$\Upsilon := (v_1 \quad \dots \quad v_n) \quad (4)$$

sei eine  $m \times n$  *Datenmatrix*, die durch die spaltenweise Konkatenation von  $m$ -dimensionaler Zufallvektoren  $v_1, \dots, v_n$  gegeben sei. Dann ergeben sich

- für das Stichprobenmittel

$$\bar{v} = \frac{1}{n} \Upsilon \mathbf{1}_n, \quad (5)$$

- für die Stichprobenkovarianzmatrix

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

- und mit

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

für die Stichprobenkorrelationsmatrix

$$R = D C D \quad (8)$$

### Bemerkungen

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

## Beweis

Die Darstellung des Stichprobenmittels ergibt sich nach durch

$$\bar{v} := \frac{1}{n} \sum_{i=1}^n v_i = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n v_{i1} \\ \vdots \\ \sum_{i=1}^n v_{im} \end{pmatrix} = \frac{1}{n} \left( \begin{pmatrix} v_{11} & \cdots & v_{n1} \\ \vdots & \ddots & \vdots \\ v_{1m} & \cdots & v_{nm} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right) = \frac{1}{n} \Upsilon \mathbf{1}_n. \quad (9)$$

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

$$\begin{aligned} C &:= \frac{1}{n-1} \sum_{i=1}^n (v_i - \bar{v})(v_i - \bar{v})^T \\ &= \frac{1}{n-1} \sum_{i=1}^n (v_i v_i^T - v_i \bar{v}^T - \bar{v} v_i^T + \bar{v} \bar{v}^T) \\ &= \frac{1}{n-1} \left( \sum_{i=1}^n v_i v_i^T - \sum_{i=1}^n v_i \bar{v}^T - \sum_{i=1}^n \bar{v} v_i^T + \sum_{i=1}^n \bar{v} \bar{v}^T \right) \\ &= \frac{1}{n-1} \left( \sum_{i=1}^n v_i v_i^T - n \bar{v} \bar{v}^T - n \bar{v} \bar{v}^T + n \bar{v} \bar{v}^T \right) \\ &= \frac{1}{n-1} \left( \sum_{i=1}^n v_i v_i^T - n \bar{v} \bar{v}^T \right) \end{aligned} \quad (10)$$

# Multivariate Deskriptivstatistiken

## Beweis (fortgeführt)

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

$$\begin{aligned} \Upsilon \left( I_n - \frac{1}{n} \mathbf{1}_{nn} \right) \Upsilon^T &= \left( \Upsilon I_n - \frac{1}{n} \Upsilon \mathbf{1}_{nn} \right) \Upsilon^T \\ &= \Upsilon \Upsilon^T - \frac{1}{n} \Upsilon \mathbf{1}_{nn} \Upsilon^T \\ &= \begin{pmatrix} v_1 & \dots & v_n \end{pmatrix} \begin{pmatrix} v_1^T \\ \vdots \\ v_n^T \end{pmatrix} - \frac{1}{n} \Upsilon \mathbf{1}_n \mathbf{1}_n^T \Upsilon^T \\ &= \sum_{i=1}^n v_i v_i^T - n \left( \frac{1}{n} \Upsilon \mathbf{1}_n \right) \left( \frac{1}{n} \mathbf{1}_n^T \Upsilon^T \right) \\ &= \sum_{i=1}^n v_i v_i^T - n \bar{v} \bar{v}^T \\ &= \frac{1}{n-1} \sum_{i=1}^n (v_i - \bar{v})(v_i - \bar{v})^T \\ &= C. \end{aligned} \tag{11}$$

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

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

## Definition (Mahalanobis Distanz)

$\xi_1$  sei ein Zufallsvektor, eine Realisation eines Zufallsvektors, ein multivariater Erwartungswert oder ein multivariates Stichprobenmittel,  $\xi_2$  sei ein Zufallsvektor, eine Realisation eines Zufallsvektors, ein multivariater Erwartungswert oder ein multivariates Stichprobenmittel und  $\Xi$  sei eine Kovarianzmatrix oder eine Stichprobenkovarianzmatrix. Dann heißt

$$D = (\xi_1 - \xi_2)^T \Xi^{-1} (\xi_1 - \xi_2) \quad (13)$$

*Mahalanobis Distanz* von  $\xi_1$  und  $\xi_2$  hinsichtlich  $\Xi$ .

### Bemerkungen

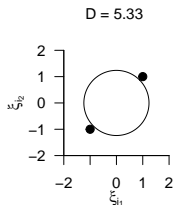
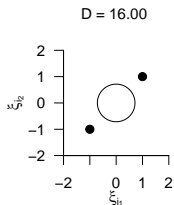
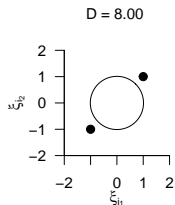
- Eine Mahalanobis Distanz ist eine Kovarianzmatrix-normalisierte quadrierte Euklidische Distanz.
- Ähnliche Maße in der univariaten Statistik sind die  $z$ -Transformation  $z = \frac{y-\mu}{\sigma}$  und Cohen's  $d = \frac{\bar{v}_1 - \bar{v}_2}{s_{12}}$ .
- Ähnlich wie bei  $z$ -Werten wird bei der Mahalanobis Distanz in "Einheiten von Kovarianzen" gemessen.
- Stark variante Komponenten von  $\xi_1$  und  $\xi_2$  tragen weniger zur Distanz bei.
- Stark kovariante Komponenten von  $\xi_1$  und  $\xi_2$  tragen weniger zur Distanz bei.

## Mahalanobis Distanzen als Funktion von Komponentenvarianzen

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

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

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

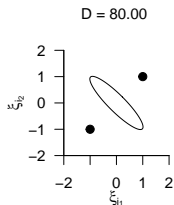
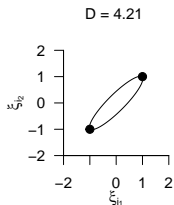
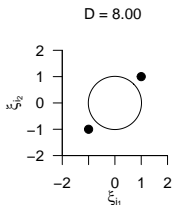


## Mahalanobis Distanzen als Funktion von Komponentenkovarianzen

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

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

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



---

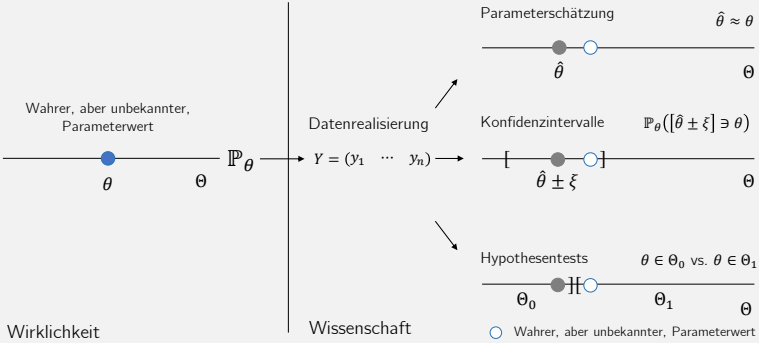
Datenanalyseszenarien

Multivariate Deskriptivstatistiken

**Grundlagen Frequentistischer Inferenz**

Selbstkontrollfragen

## Standardannahmen und Standardproblemstellungen der Frequentistischen Inferenz



## Standardannahmen Frequentistischer Inferenz

- $\mathcal{M}$  sei ein Frequentistisches Inferenzmodell mit  $v_1, \dots, v_n \sim p_\theta$ . Es wird angenommen, dass eine konkrete Datenmatrix  $Y \in \mathbb{R}^{m \times n}$  eine der möglichen Realisierungen von  $Y = (v_1 \quad \dots \quad v_n)$  ist.
- Aus Frequentistischer Sicht kann man eine Studie unendlich oft wiederholen und zu jedem Datensatz Schätzer oder Statistiken auswerten, z.B. das Stichprobenmittel:

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

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

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

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

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

- Um die Qualität statistischer Methoden zu beurteilen betrachtet die Frequentistische Statistik deshalb die Wahrscheinlichkeitsverteilungen von Schätzern und Statistiken unter Annahme von  $v_1, \dots, v_n \sim p_\theta$ . Was zum Beispiel ist die Verteilung der  $\bar{y}^{(1)}, \bar{y}^{(2)}, \bar{y}^{(3)}, \bar{y}^{(4)}, \dots$  also die Verteilung der Zufallsvariable  $\bar{v}$ ?
- 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

---