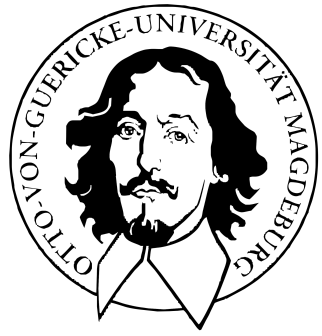


Tutorium

Wahrscheinlichketstheorie und Frequentistische Inferenz

BSc Psychologie WiSe 2022/23

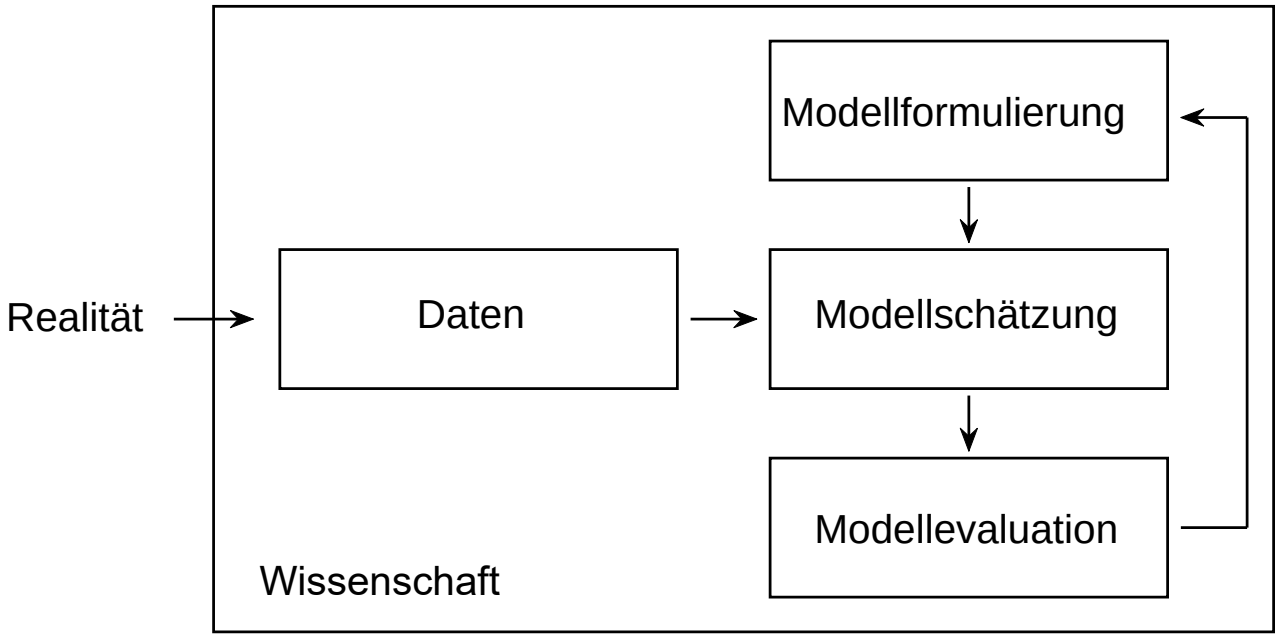
Belinda Fleischmann

(9) Grundbegriffe Frequentistischer Inferenz

Selbstkontrollfragen

1. Definieren und erläutern Sie den Begriff des parametrischen statistischen Modells.
2. Definieren und erläutern Sie den Begriff eine parametrischen statistischen Produktmodells.
3. Erläutern Sie den Unterschied zwischen univariaten und multivariaten statistischen Modellen.
4. Formulieren und erläutern Sie das Normalverteilungsmodell.
5. Formulieren und erläutern Sie das Bernoullimodell.
6. Definieren und erläutern Sie den Begriff der Statistik.
7. Definieren und erläutern Sie den Begriff des Schätzers.
8. Nennen und erläutern Sie die Standardprobleme der frequentistischen Inferenz.
9. Erläutern Sie die Standardannahmen der frequentistischen Inferenz.

Modellbasierte Datenwissenschaft



Frequentistische Inferenz

Modellformulierung	$\Rightarrow$ Statistische Modelle
Modellschätzung	$\Rightarrow$ Parameterschätzung und Konfidenzintervalle
Modellevaluation	$\Rightarrow$ Hypothesentests (cf. Allgemeines Lineares Modell)

1. Definieren und erläutern Sie den Begriff des parametrischen statistischen Modells.

Definition (Statistisches Modell)

Ein *statistisches Modell* ist ein Tripel

$$\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\}) \tag{1}$$

bestehend aus einem *Datenraum* $\mathcal{Y}$, einer σ-Algebra $\mathcal{A}$ auf $\mathcal{Y}$ und einer mindestens zweielementigen Menge $\{\mathbb{P}_\theta | \theta \in \Theta\}$ von Wahrscheinlichkeitsmaßen auf $(\mathcal{Y}, \mathcal{A})$, die durch $\theta \in \Theta$ indiziert sind. Wenn $\Theta \subset \mathbb{R}^k$ ist, heißt ein statistisches Modell auch *parametrisches* statistisches Modell und Θ heißt *Parameterraum* des statistischen Modells.

- Der Datenraum $\mathcal{Y}$ enthält alle Werte, die ein Zufallsvektor y, welcher den Vorgang der Datenbeobachtung beschreibt, annehmen kann. Im Kontext statistischer Modelle wird y auch *Daten*, *Beobachtung*, *Messung* oder *Stichprobe* genannt. Eine Realisierung von y, also konkret vorliegende Datenwerte $\tilde{y} \in \mathcal{Y}$, werden *Datensatz*, *Beobachtungswert*, *Messwert* oder *Stichprobenwert* genannt.
- $\{\mathbb{P}_\theta | \theta \in \Theta\}$ ist eine mindestens zweielementige Menge von Wahrscheinlichkeitsmaßen auf $(\mathcal{Y}, \mathcal{A})$, wobei jedes Element $\mathbb{P}_\theta$ eine Abbildung $\mathbb{P}_\theta : \mathcal{A} \to [0, 1]$ ist (vgl. Def. Wahrscheinlichkeitsraum und Wahrscheinlichkeitsmaß Einheit (2)).

SKF 1. *Das parametrischen statistischen Modell* (fortgeführt)

- Der Parameterraum Θ ist die Menge aller möglichen wahren, aber unbekannten, Parameterwerte. Der wahre, aber unbekannten, Parameterwert θ^* bleibt auch nach der statistischen Analyse unbekannt. In der mathematischen Analyse von Inferenzmethoden betrachtet man alle möglichen wahren, aber unbekannten, Parameterwerte, schreibt also einfach $\{\mathbb{P}_\theta | \theta \in \Theta\}$.
- In einem konkreten Datenanalyseproblem nimmt man an, dass die beobachteten Werte $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)$ von $y = (y_1, \ldots, y_n)$ durch θ^* generiert wurde, wobei θ^* hier den *wahren, aber unbekannten, Parameterwert* bezeichnet.
- Im Gegensatz zum Wahrscheinlichkeitsraummodell betrachtet man bei statistische Modellen zwei oder mehr Wahrscheinlichkeitsmaße, die die Verteilung von y mutmaßlich bestimmen. Das jeweils zugrundeliegende Wahrscheinlichkeitsmaß ist mit $\theta \in \Theta$ indiziert,

2. Definieren und erläutern Sie den Begriff eine parametrischen statistischen Produktmodells.

- Für ein statistisches Modell $\mathcal{M}_0 := (\mathcal{Y}_0, \mathcal{A}_0, \{\mathbb{P}_\theta^0 | \theta \in \Theta\})$ heißt das statistische Modell $\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\})$, für das $\mathcal{Y}$ das n-fache kartesische Produkt von $\mathcal{Y}_0$ mit sich selbst, $\mathcal{A}$ die entsprechende Produkt-σ-Algebra ist, und $\{\mathbb{P}_\theta | \theta \in \Theta\}$ die entsprechende Menge an Produktmaßen ist, das zu $\mathcal{M}_0$ gehörige *Produktmodell*.
- Produktmodelle modellieren die n-fache unabhängige Wiederholung eines Zufallsvorgangs. Der entsprechende Zufallsvektor $y := (y_1, ..., y_n)$ entspricht dann einer Menge von n unabhängigen Zufallsvariablen/vektoren.

3. Erläutern Sie den Unterschied zwischen univariaten und multivariaten statistischen Modellen.

Wir sprechen von einem *univariaten statistischen Modell*, wenn für ein Produktmodell die Menge $\mathcal{Y}_0$ eindimensional ist und von einem *multivariaten statistischen Modell*, wenn für ein Produktmodell die Menge $\mathcal{Y}_0$ mehrdimensional ist.

4. Formulieren und erläutern Sie das Normalverteilungsmodell.

Definition (Normalverteilungsmodell)

Das univariate parametrische Produktmodell

$$\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\}) \tag{2}$$

mit

$$\mathcal{Y} := \mathbb{R}^n, \mathcal{A} := \mathcal{B}(\mathbb{R}^n), \theta := (\mu, \sigma^2), \Theta := \mathbb{R} \times \mathbb{R}_{>0}, \tag{3}$$

also

$$\{\mathbb{P}_\theta | \theta \in \Theta\} := \left\{ \prod_{i=1}^{n} N(\mu, \sigma^2) | (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \right\}, \tag{4}$$

und damit

$$y_1, \ldots, y_n \sim N(\mu, \sigma^2) \text{ mit } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \tag{5}$$

heißt *Normalverteilungsmodell*.

- Das Normalverteilungsmodell ist ein univariates parametrisches statistischen Modell, genauer gesagt ein univariates parametrisches Produktmodell.
- Der Datenraum ist das n-fache kartesische Produkt der reellen Zahlen mit sich selbst. (vgl. Vorkurs (1) Def. Die Menge $\mathbb{R}^n$)

SKF 4. *Das Normalverteilungsmodell* (fortgeführt)

- Die σ-Algebra $\mathcal{A}$ ist definiert als die *Borelsche* σ-Algebra $\mathcal{B}(\mathbb{R}^n)$
- θ ist ein Tupel bestehend aus zwei Parametern μ und σ^2 und $\Theta := \mathbb{R} \times \mathbb{R}_{>0}$ ist die Menge aller möglichen wahren, aber unbekannten, Parameterwerte für $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}_{>0}$.
- $\{\mathbb{P}_\theta | \theta \in \Theta\}$ ist die Menge aller Wahrscheinlickeitsmaße, auf $(\mathcal{Y}, \mathcal{A})$. Jedes Element $\mathbb{P}_\theta$ dieser Menge ist das Wahrscheinlichkeitsmaß für einen möglichen wahren, aber unbekannten Parameterwertepaar $\theta := (\mu, \sigma^2)$, und ist definiert als das n-fache Produkt einer univariaten Normalverteilung mit Erwartungswertparameter μ und Varianzparameter σ^2.
- Es wird angenommen, dass jede der n Zufallsvariablen $y_1, \ldots, y_n$ normalverteilt ist mit Erwartungswertparameter μ und Varianzparameter σ^2.

5. Formulieren und erläutern Sie das Bernoullimodell.

Definition (Bernoullimodell)

Das univariate parametrische Produktmodell

$$\mathcal{M} := (\mathcal{Y}, \mathcal{A}, \{\mathbb{P}_\theta | \theta \in \Theta\}) \tag{6}$$

mit

$$\mathcal{Y} := \{0,1\}^n, \mathcal{A} := \mathcal{P}\left(\{0,1\}^n\right), \theta := \mu, \Theta :=]0,1[, \tag{7}$$

also

$$\{\mathbb{P}_\theta | \theta \in \Theta\} := \left\{ \prod_{i=1}^{n} \text{Bern}(\mu) | \mu \in]0,1[\right\}, \tag{8}$$

und damit

$$y_1, \ldots, y_n \sim \text{Bern}(\mu) \text{ mit } \mu \in]0,1[, \tag{9}$$

heißt *Bernoullimodell*.

Erläuterung analog zu SKF 4

6. Definieren und erläutern Sie den Begriff der Statistik.

Definition (Statistik)

$\mathcal{M}$ sei ein statistisches Modell und $(\Sigma, \mathcal{S})$ sei ein Messraum. Dann wird eine Zufallsvariable der Form

$$S : \mathcal{Y} \to \Sigma \tag{10}$$

Statistik genannt.

Bemerkungen

- Daten und Statistiken werden durch Zufallsvariablen modelliert. Statistiken modellieren dabei von Datenwissenschaftler:innen konstruierte Funktionen, die bestenfalls datenbasierte Information liefern, aus der sich Schlüsse über die latenten datengenerierenden Prozesse ziehen lassen.

7. Definieren und erläutern Sie den Begriff des Schätzers.

Definition (Schätzer)

$\mathcal{M}$ sei ein statistisches Modell, $(\Sigma, \mathcal{S})$ sei ein Messraum und $\tau : \Theta \to \Sigma$ sei eine Abbildung, die jedem $\theta \in \Theta$ eine Kenngröße $\tau(\theta) \in \Sigma$ zuordnet. Dann heißt eine Statistik

$$\hat{\tau} : \mathcal{Y} \to \Sigma \tag{11}$$

ein *Schätzer* für τ.

8. Nennen und erläutern Sie die Standardprobleme der frequentistischen Inferenz.

Mithilfe statistischer Modelle behandelt die Frequentistische Inferenz folgende Standardprobleme:

(1) Parameterschätzung

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

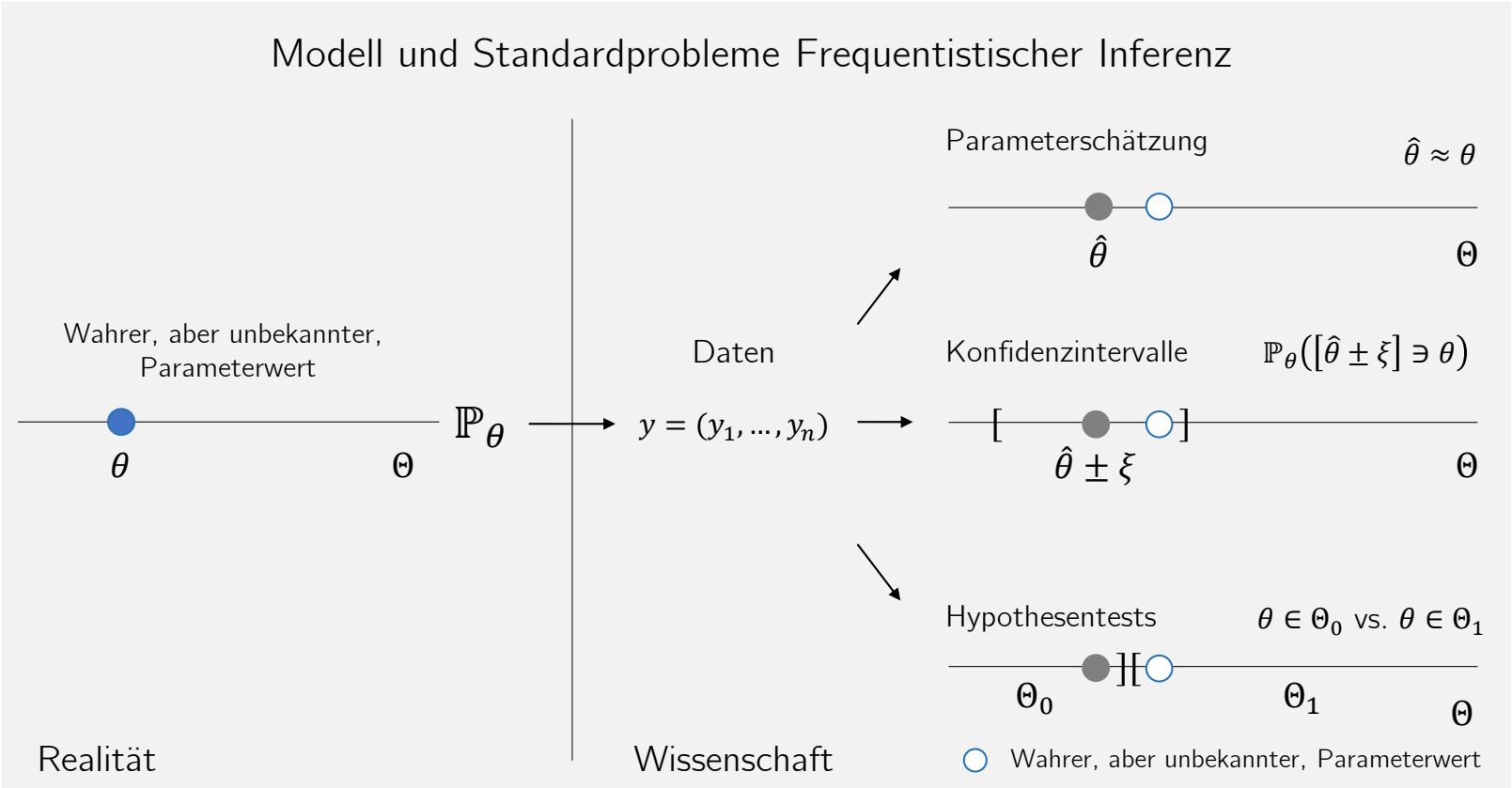
(2) Konfidenzintervalle

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.

SKF 8. *Standardprobleme der frequentistischen Inferenz* (fortgeführt)



9. Erläutern Sie die Standardannahmen der frequentistischen Inferenz.

$\mathcal{M}$ sei ein statistisches Modell mit $y_1, \ldots, y_n \sim p_\theta$.

Es wird angenommen, dass ein konkreter Datensatz eine der möglichen Realisierungen von $y_1, \ldots, y_n \sim p_\theta$ ist.

Aus Frequentistischer Sicht kann man eine Studie unendlich oft wiederholen und zu jedem Datensatz Schätzer oder Statistiken auswerten, z.B. das Stichprobenmittel:

$$\text{Datensatz (1)} : \tilde{y}^{(1)} = \left(\tilde{y}_1^{(1)}, \tilde{y}_2^{(1)}, \ldots, \tilde{y}_n^{(1)}\right) \text{ mit } \bar{y}_n^{(1)} = \frac{1}{n}\sum_{i=1}^n \tilde{y}_i^{(1)}$$

$$\text{Datensatz (2)} : \tilde{y}^{(2)} = \left(\tilde{y}_1^{(2)}, \tilde{y}_2^{(2)}, \ldots, \tilde{y}_n^{(2)}\right) \text{ mit } \bar{y}_n^{(2)} = \frac{1}{n}\sum_{i=1}^n \tilde{y}_i^{(2)}$$

$$\text{Datensatz (3)} : \tilde{y}^{(3)} = \left(\tilde{y}_1^{(3)}, \tilde{y}_2^{(3)}, \ldots, \tilde{y}_n^{(3)}\right) \text{ mit } \bar{y}_n^{(3)} = \frac{1}{n}\sum_{i=1}^n \tilde{y}_i^{(3)}$$

$$\text{Datensatz (4)} : \tilde{y}^{(4)} = \left(\tilde{y}_1^{(4)}, \tilde{y}_2^{(4)}, \ldots, \tilde{y}_n^{(4)}\right) \text{ mit } \bar{y}_n^{(4)} = \frac{1}{n}\sum_{i=1}^n \tilde{y}_i^{(4)}$$

$$\text{Datensatz (5)} : \tilde{y}^{(5)} = \ldots$$

Um die Qualität statistischer Methoden zu beurteilen betrachtet die Frequentistische Statistik deshalb die Wahrscheinlichkeitsverteilungen von Schätzern und Statistiken unter Annahme von $y_1, \ldots, y_n \sim p_\theta$. Was zum Beispiel ist die Verteilung von $\bar{y}_n^{(1)}$, $\bar{y}_n^{(2)}$, $\bar{y}_n^{(3)}$, $\bar{y}_n^{(4)}$, ... also die Verteilung der Zufallsvariable $\bar{y}_n$?

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.