



Wahrscheinlichkeitstheorie und Frequentistische Inferenz

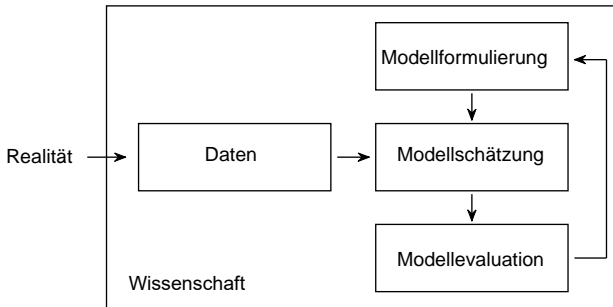
BSc Psychologie WiSe 2022/23

Prof. Dr. Dirk Ostwald

(9) Grundbegriffe Frequentistischer Inferenz

Datum	Einheit	Thema
13.10.2022	Einführung	(1) Einführung
20.10.2022	Wahrscheinlichkeitstheorie	(2) Wahrscheinlichkeitsräume
27.10.2022	Wahrscheinlichkeitstheorie	(3) Elementare Wahrscheinlichkeiten
03.11.2022	Wahrscheinlichkeitstheorie	(4) Zufallsvariablen I
10.11.2022	Wahrscheinlichkeitstheorie	(4) Zufallsvariablen II
17.11.2022	Wahrscheinlichkeitstheorie	(5) Multivariate Verteilungen
24.11.2022	Wahrscheinlichkeitstheorie	(6) Erwartungswert und Kovarianz
01.12.2022	Wahrscheinlichkeitstheorie	(7) Ungleichungen und Grenzwerte
08.12.2022	Wahrscheinlichkeitstheorie	(8) Transformationen der Normalverteilung
15.12.2022	Frequentistische Inferenz	(9) Grundbegriffe Frequentistischer Inferenz
	Weihnachtspause	
05.01.2023	Frequentistische Inferenz	(10) Parameterschätzung
12.01.2023	Frequentistische Inferenz	(11) Konfidenzintervalle
19.01.2023	Frequentistische Inferenz	(12) Hypothesentests I
26.01.2023	Frequentistische Inferenz	(12) Hypothesentests II
02.02.2023	Klausur	G44-H6, 16:00 - 17:00 Uhr
Jul 2023	Klausurwiederholungstermin	

Modellbasierte Datenwissenschaft



Frequentistische Inferenz

- Modellformulierung ⇒ Statistische Modelle
- Modellschätzung ⇒ Parameterschätzung und Konfidenzintervalle
- Modellevaluation ⇒ Hypothesentests (cf. Allgemeines Lineares Modell)

Statistische Modelle

Statistiken und Schätzer

Standardprobleme Frequentistischer Inferenz

Selbstkontrollfragen

Statistische Modelle

Statistiken und Schätzer

Standardprobleme Frequentistischer Inferenz

Selbstkontrollfragen

Definition (Statistisches Modell)

Ein *statistisches Modell* ist ein Tripel

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

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.

Bemerkungen

- Für Erwartungswerte und (Ko)Varianzen bezüglich \mathbb{P}_θ schreiben wir $\mathbb{E}_\theta, \mathbb{V}_\theta, \mathbb{C}_\theta$.
- Ein statistisches Modell \mathcal{M} heißt ein *diskretes Modell*, wenn \mathcal{Y} diskret ist und jedes \mathbb{P}_θ eine WMF p_θ besitzt, ein statistisches Modell \mathcal{M} heißt ein *stetiges Modell*, wenn $\mathcal{Y} \subset \mathbb{R}^n$ ist und jedes \mathbb{P}_θ eine WDF p_θ besitzt.
- 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*.
- Wenn für ein Produktmodell die Menge \mathcal{Y}_0 eindimensional ist, also z.B. $\mathcal{Y}_0 = \mathbb{R}$ gilt, spricht man von einem *univariaten statistischen Modell*. Wenn für ein Produktmodell die Menge \mathcal{Y}_0 mehrdimensional ist, also z.B. $\mathcal{Y}_0 = \mathbb{R}^m, m > 1$ ist, spricht man von einem *multivariaten statistischen Modell*.

Bemerkungen (fortgeführt)

- Der Vorgang der Datenbeobachtung wird durch einen Zufallsvektor v , der Werte in \mathcal{Y} annimmt, beschrieben. Im Kontext statistischer Modelle nennt man diesen Zufallsvektor *Daten*, *Beobachtung*, *Messung* oder *Stichprobe*. Eine Realisierung von v , also konkret vorliegende Datenwerte $y \in \mathcal{Y}$, werden *Datensatz*, *Beobachtungswert*, *Messwert* oder *Stichprobenwert* genannt.
- Produktmodelle modellieren die n -fache unabhängige Wiederholung eines Zufallsvorgangs. Der entsprechende Zufallsvektor $v := (v_1, \dots, v_n)$ entspricht dann einer Menge von n unabhängigen Zufallsvariablen.
- Im Gegensatz zum Wahrscheinlichkeitsraummodell betrachtet man bei statistische Modellen zwei oder mehr Wahrscheinlichkeitsmaße, die die Verteilung von v mutmaßlich bestimmen. Das jeweils zugrundeliegende Wahrscheinlichkeitsmaß ist mit $\theta \in \Theta$ indiziert,
- In einem konkreten Datenanalyseproblem nimmt man an, dass die beobachteten Werte $y = (y_1, \dots, y_n)$ von $v = (v_1, \dots, v_n)$ durch θ^* generiert wurde, wobei θ^* hier den *wahren, aber unbekanntem, Parameterwert* bezeichnet. Der wahre, aber unbekanntem, Parameterwert θ^* bleibt auch nach der statistischen Analyse unbekannt. In der mathematischen Analyse von Inferenzmethoden betrachtet man alle möglichen wahren, aber unbekanntem, Parameterwerte, schreibt also einfach $\{\mathbb{P}_\theta | \theta \in \Theta\}$.

Definition (Normalverteilungsmodell)

Das univariate parametrische Produktmodell

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

mit

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

also

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

und damit

$$v_1, \dots, v_n \sim N(\mu, \sigma^2) \text{ mit } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0} \quad (5)$$

heißt *Normalverteilungsmodell*.

Bemerkungen

- Das Normalverteilungsmodell ist Grundlage vieler populärer statistischen Verfahren.
- Diese Verfahren werden im Allgemeinen Linearen Modell integrativ betrachtet.
- Das Modell ist grundlegend durch normalverteilte Fehlerterme motiviert.

Definition (Bernoullimodell)

Das univariate parametrische Produktmodell

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

mit

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

also

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

und damit

$$v_1, \dots, v_n \sim \text{Bern}(\mu) \text{ mit } \mu \in]0, 1[, \quad (9)$$

heißt *Bernoullimodell*.

Bemerkung

- Das Bernoullimodell spielt in der statistischen Anwendung eine eher untergeordnete Rolle.

Statistische Modelle

Statistiken und Schätzer

Standardprobleme Frequentistischer Inferenz

Selbstkontrollfragen

Definition (Statistik)

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

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

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.

Beispiele (Statistiken)

\mathcal{M} sei das Normalverteilungsmodell. Dann sind zum Beispiel folgende Zufallsvariablen Statistiken, was wir hier explizit durch die Notation deutlich machen wollen, was aber oft zur Vereinfachung der Notation implizit (aber trotzdem wichtig) bleibt:

- Das *Stichprobenmittel*

$$\bar{y}_n : \mathbb{R}^n \rightarrow \mathbb{R}, y \mapsto \bar{y}_n(y) := \frac{1}{n} \sum_{i=1}^n y_i, \quad (11)$$

- Die *Stichprobenvarianz*

$$s_n^2 : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}, y \mapsto s_n^2(y) := \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y}_n(y))^2, \quad (12)$$

- Die *Stichprobenstandardabweichung*

$$s_n : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}, y \mapsto s_n(y) := \sqrt{s_n^2(y)}, \quad (13)$$

Definition (Schätzer)

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

$$\hat{\tau} : \mathcal{Y} \rightarrow \Sigma \quad (14)$$

ein *Schätzer* für τ .

Bemerkungen

- Typische Beispiele für τ sind
 - $\tau(\theta) := \theta$ für die Schätzung von θ ,
 - $\tau(\theta) := \theta_i$ mit $\theta \in \mathbb{R}^d$, $d > 1$ für die Schätzung einer Komponente von θ ,
 - $\tau(\theta) := \mathbb{E}_\theta(y_1)$ für die Schätzung des Erwartungswert,
 - $\tau(\theta) := \mathbb{V}_\theta(y_1)$ für die Schätzung der Varianz.
- Für $\hat{\tau}$ bei $\tau(\theta) := \theta$ schreibt man üblicherweise $\hat{\theta}$.
- Schätzer nehmen Zahlwerte in Σ an und heißen deshalb auch *Punktschätzer*.
- Nicht jeder Schätzer ist ein guter Schätzer, man definiert deshalb *Schätzgütekriterien*.
- Gütekriterien für Schätzer sind der Inhalt von Vorlesungseinheit (10) Parameterschätzung.

Beispiele (Schätzer)

\mathcal{M} sei das Normalverteilungsmodell.

- Dann ist zum Beispiel das Stichprobenmittel $\bar{y}_n : \mathbb{R}^n \rightarrow \mathbb{R}$ ein Schätzer für

$$\tau : \mathbb{R} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}, (\mu, \sigma^2) \mapsto \tau(\mu, \sigma^2) := \mu. \quad (15)$$

Ebenso ist \bar{y}_n ein Schätzer für

$$\tau : \mathbb{R} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}, (\mu, \sigma^2) \mapsto \tau(\mu, \sigma^2) := \mathbb{E}_{\mu, \sigma^2}(y_1). \quad (16)$$

- Weiterhin ist die konstante Funktion

$$\hat{\tau} : \mathbb{R}^n \rightarrow \mathbb{R}, y \mapsto \hat{\tau}(y) := 42 \quad (17)$$

ein Schätzer für

$$\tau : \mathbb{R} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}, (\mu, \sigma^2) \mapsto \tau(\mu, \sigma^2) := \sigma^2. \quad (18)$$

Dass eine Funktion $\hat{\tau} : \mathcal{Y} \rightarrow \Sigma$ ein Schätzer ist, heißt nicht, dass sie ein guter Schätzer ist!

Statistische Modelle

Statistiken und Schätzer

Standardprobleme Frequentistischer Inferenz

Selbstkontrollfragen

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.

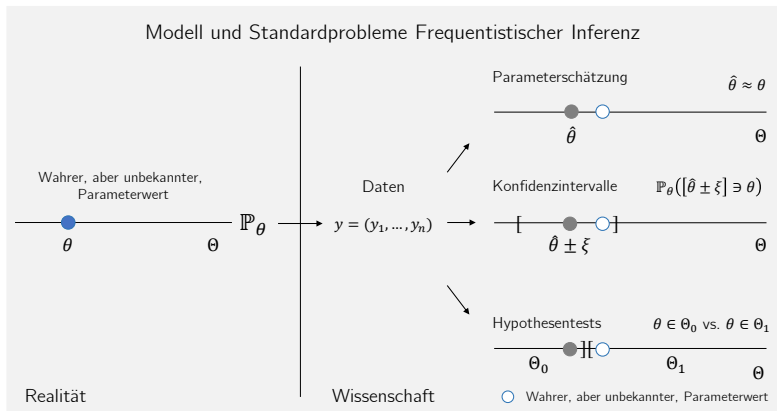
(2) Konfidenzintervalle

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.

Standardprobleme Frequentistischer Inferenz



Standardprobleme Frequentistischer Inferenz

Standardannahmen Frequentistischer Inferenz

\mathcal{M} sei ein statistisches Modell mit $v = (v_1, \dots, v_n) \sim p_\theta$. Es wird angenommen, dass ein konkreter Datensatz $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ eine der möglichen Realisierungen von $v = (v_1, \dots, v_n) \sim p_\theta$ ist.

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

$$\text{Datensatz (1)} : y^{(1)} = \left(y_1^{(1)}, y_2^{(1)}, \dots, y_n^{(1)} \right) \text{ mit } \bar{y}_n^{(1)} = \frac{1}{n} \sum_{i=1}^n y_i^{(1)}$$

$$\text{Datensatz (2)} : y^{(2)} = \left(y_1^{(2)}, y_2^{(2)}, \dots, y_n^{(2)} \right) \text{ mit } \bar{y}_n^{(2)} = \frac{1}{n} \sum_{i=1}^n y_i^{(2)}$$

$$\text{Datensatz (3)} : y^{(3)} = \left(y_1^{(3)}, y_2^{(3)}, \dots, y_n^{(3)} \right) \text{ mit } \bar{y}_n^{(3)} = \frac{1}{n} \sum_{i=1}^n y_i^{(3)}$$

$$\text{Datensatz (4)} : y^{(4)} = \left(y_1^{(4)}, y_2^{(4)}, \dots, y_n^{(4)} \right) \text{ mit } \bar{y}_n^{(4)} = \frac{1}{n} \sum_{i=1}^n y_i^{(4)}$$

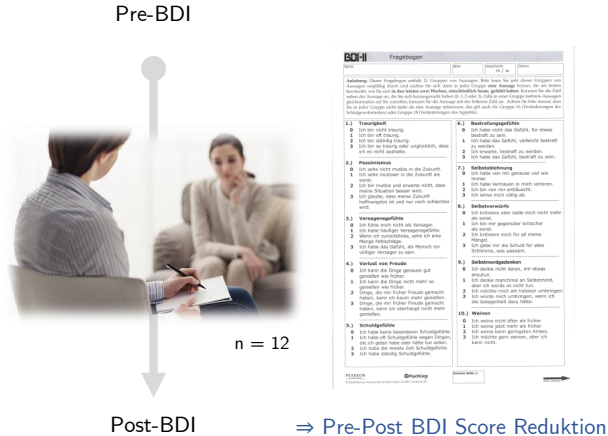
$$\text{Datensatz (5)} : y^{(5)} = \dots$$

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

Standardprobleme Frequentistischer Inferenz

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression



Standardprobleme Frequentistischer Inferenz

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

```
fname = file.path(getwd(), "9_Grundbegriffe_Frequentistischer_Inferenz.csv")  
D     = read.table(fname, sep = ",", header = T)
```

i	BDI.Reduktion
1	-1
2	3
3	-2
4	9
5	3
6	-2
7	4
8	5
9	5
10	1
11	9
12	4

Standardprobleme Frequentistischer Inferenz

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

Für die Pre-Post BDI Score Reduktion y_i der i ten von n Patient:innen legen wir das Modell

$$v_i = \mu + \varepsilon_i \text{ mit } \varepsilon_i \sim N(0, \sigma^2) \text{ u.i.v. für } i = 1, \dots, n \quad (19)$$

zugrunde. Dabei wird die Pre-Post BDI Reduktion y_i der i ten Patient:in also mithilfe einer über die Gruppe von Patient:innen identischen Pre-Post BDI Score Reduktion $\mu \in \mathbb{R}$ und einer Patient:innen-spezifischen normalverteilten Pre-Post BDI Score Reduktionsabweichung ε_i erklärt

Wie unten gezeigt ist dieses Modell äquivalent zum oben eingeführten Normalverteilungsmodell

$$v = (v_1, \dots, v_n) \sim N(\mu, \sigma^2). \quad (20)$$

Die Standardprobleme der Frequentistischen Inferenz führen in diesem Szenario auf folgende Fragen:

- (1) Was sind sinnvolle Tipps für die wahren, aber unbekanntenen, Parameterwerte μ und σ^2 ?
- (2) Wie hoch ist im Frequentistischen Sinn die mit diesen Tipps assoziierte Unsicherheit?
- (3) Entscheiden wir uns sinnvollerweise für die Hypothese, dass gilt $\mu \neq 0$?

Beispiel | Evidenzbasierte Evaluation von Psychotherapie bei Depression

Die Äquivalenz beider Modellformen folgt direkt aus der Transformation normalverteilter Zufallsvariablen durch linear-affine Funktionen. Speziell gilt im vorliegenden Fall für $\varepsilon_i \sim N(0, \sigma^2)$ u.i.v., dass

$$v_i = f(\varepsilon_i) \text{ mit } f : \mathbb{R} \rightarrow \mathbb{R}, \varepsilon_i \mapsto f(\varepsilon_i) := \varepsilon_i + \mu. \quad (21)$$

Dann gilt

$$\begin{aligned} p_{v_i}(y_i) &= \frac{1}{|1|} p_{\varepsilon_i} \left(\frac{y_i - \mu}{1} \right) \\ &= N(y_i - \mu; 0, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2\sigma^2} (y_i - \mu - 0)^2 \right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2\sigma^2} (y_i - \mu)^2 \right) \\ &= N(y_i; \mu, \sigma^2), \end{aligned} \quad (22)$$

also $v_i \sim N(\mu, \sigma^2)$ u.i.v. und damit $v_1, \dots, v_n \sim N(\mu, \sigma^2)$.

Statistische Modelle

Statistiken und Schätzer

Standardprobleme Frequentistischer Inferenz

Selbstkontrollfragen

Selbstkontrollfragen

1. Definieren und erläutern Sie den Begriff des parametrischen statistischen Modells.
2. Definieren und erläutern Sie den Begriff eines 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 Bernoulli-Modell.
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.