EAS 2023 - Inferenzstatistik

Prof. Dr. Ansgar Steland

2021

Agenda

- Einstieg in die Statistische Inferenz
- Parametrische Modelle
- Schätzverfahren

Wozu schließende Statistik?

Beispiel

Sind User der Spielekonsole __ zufrieden?

- *Umfrage unter* $n = 500$ *zufällig ausgewählten registrierten Usern.*
- $k = 400$ *sind mit ihrer Konsole zufrieden.*

Sind diese Zahlen belastbar*?*

1. *Ist der Anteil von* $k/n = 80\%$ *zufriedenen Nutzern eine gute Schätzung des wahren Anteils in der Grundgesamtheit?*
2. *Wie stark streut das Stichprobenergebnis? Wie sicher ist die Schätzung?*
3. *Wie kann objektiv nachgewiesen werden, dass der wahre Anteil zufriedener User zumindest höher als (z. B.) 75% ist?*

Wege zur Lösung...

1. Finde geeignetes Verteilungsmodell für die Daten.
   Hier: $\mathrm{Bin}(n, p)$, p unbekannt.
2. Wie kann man p (optimal?) aus den Daten schätzen?
3. Wie kann man die Hypothese $p > 0.75$ nachweisen?

Wahrscheinlichkeitsrechnung:

- Liefert Regeln, wie man mit Wahrscheinlichkeiten und Verteilungen rechnet.
- Gegeben: Stochastisches Modell $X \sim F$.
  Oft: $F = F_\vartheta$ (parametrisiert durch ϑ).
- F wird (gedanklich) als bekannt/gegeben angenommen.
- Bsp: $X \sim N(\mu, \sigma^2) \Rightarrow P(X \leq 2) = \Phi((2-\mu)/\sigma)$.
  Liefert eine Formel, die von $\vartheta = (\mu, \sigma^2)$ abhängt.
  Einsetzen spezieller Werte, z.B. $\vartheta = (4, 2)$, liefert eine konkrete Zahl.

Schließende Statistik:

- Gegeben: Verrauschte (zufallsbehaftete) Daten $X_1, \dots, X_n \sim F_\vartheta$.
- Gesucht: Das Modell F_ϑ, also ϑ.
- Ziel: Schließe aus den Daten auf das zugrunde liegende Modell.
- Relevant Schritte:
  1. Gute Modellklasse für die Daten finden. — Modellierung
  2. Schätzen des Modells aus den Daten. — Schätzen
  3. Testen: Gilt $\vartheta \in \Theta_0$ oder $\vartheta \in \Theta_1$? — Testen
  4. Untersuche, ob das Modell die Daten gut erklärt. — Modellvalidierung

Grundbegriffe: Stichprobe

Stichprobe

$X_1, \ldots, X_n$ heißt **Stichprobe** vom **Stichprobenumfang** n, wenn

$$X_1, \ldots, X_n : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$$

Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ sind.
Zufallsvektor $\mathbf{X} = (X_1, \ldots, X_n)$ nimmt Werte im **Stichprobenraum**

$$\mathcal{X} = \{\mathbf{X}(\omega) : \omega \in \Omega\} \subset \mathbb{R}^n$$

an. Realisierungen: Vektoren $(x_1, \ldots, x_n) \in \mathcal{X}$.

Hinweis

In der Statistik interessiert i.d.R. der zugrunde liegende W-Raum $(\Omega, \mathcal{A}, P)$ nicht, sondern lediglich der Stichprobenraum $\mathcal{X}$ und die Verteilung $P_{\mathbf{X}}$ von $\mathbf{X} = (X_1, \ldots, X_n)'$ hierauf!

Grundbegriffe: Verteilungsmodell

Verteilungsmodell

Eine Menge $\mathcal{P}$ von (möglichen) Verteilungen auf $\mathbb{R}^n$ (für die Stichprobe $(X_1, \dots, X_n)$) heißt **Verteilungsmodell**.

$\mathcal{P}$ heißt **parametrisches Verteilungsmodell**, falls

$$\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$$

für eine Menge $\Theta \subset \mathbb{R}^k$ von Parametervektoren.

Θ: **Parameterraum**.
D.h.: Es gibt eine Bijektion $\mathcal{P} \leftrightarrow \Theta$.

Ein Verteilungsmodell, das nicht durch einen endlichdimensionalen Parameter parametrisiert werden kann, heißt **nichtparametrisches Verteilungsmodell**.

Beispiele:

Beispiel

Parametrische Verteilungsmodelle:

1). $\mathcal{P} = \{Bin(n, p) : p \in [0, 1]\}$ *für ein festes* n.
Parameter: $\vartheta = p \in \Theta = [0, 1]$.

2). $\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, 0 < \sigma^2 < \infty\}$. *Parameter:*
$\vartheta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$.

3). *Sei* $Y = g_{net}(\mathbf{X})$ *mit* $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{I}_p)$, $\boldsymbol{\mu} \in \mathbb{R}^p$, $p \in \mathbb{N}$
$y = g_{net}(\mathbf{x})$ *Deep Learner mit Netzparametern* $\mathbf{w} \in \mathbf{W} \subset \mathbb{R}^q$, $q \in \mathbb{N}$.
Bezeichne $G_{(\boldsymbol{\mu},\mathbf{w})}(y)$ *die Verteilungsfunktion von* Y *bei Input* $\mathbf{X}$.
$\mathcal{P} = \{G_{(\boldsymbol{\mu},\mathbf{w})} : \boldsymbol{\mu} \in \mathbb{R}^p, \mathbf{w} \in \mathbf{W}\}$ *Menge möglicher Verteilungen für* Y.
Parameter: $\boldsymbol{\vartheta} = (\boldsymbol{\mu}, \boldsymbol{\theta}) \in \Theta = \mathbb{R} \times (0, \infty)$.

Nichtparametrische Verteilungsmodelle:

4). $\mathcal{P} = \{F : \mathbb{R} \to [0, 1] : F \text{ ist Verteilungsfunktion}\}$

5). $\mathcal{P} = \{f : \mathbb{R} \to \mathbb{R}^+ : f \text{ ist stetige Dichtefunktion}\}$

Statistik, Schätzfunktion, Schätzer

Statistik,...

Sei $X_1, \ldots, X_n$ eine Stichprobe (und o.E. $\mathcal{X} = \mathbb{R}^n$).
Eine Abbildung

$$T : \mathbb{R}^n \to \mathbb{R}^d$$

mit $d \in \mathbb{N}$ (oft: $d = 1$) heißt **Statistik**.

Bildet T in den Parameterraum ab, d.h.

$$T : \mathbb{R}^n \to \Theta,$$

dann heißt T **Schätzfunktion** oder kürzer **Schätzer** (für ϑ).

Allgemein: Schätzung von Funktionen $g(\vartheta)$ von ϑ durch Statistiken $T : \mathbb{R}^n \to \Gamma$ mit $\Gamma = g(\Theta) = \{g(\vartheta) | \vartheta \in \Theta\}$.

Beispiele

Beispiel: Seien $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$, und

$$T_1(X_1, \dots, X_n) = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \ ,$$

$$T_2(X_1, \dots, X_n) = S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

$T_1(X_1, \dots, X_n)$ bildet in den Parameterraum $\Theta_1 = \mathbb{R}$ für μ ab und ist daher eine Schätzfunktion für μ.
$T_1(X_1, \dots, X_n)$ bildet in den Parameterraum $\Theta_2 = (0, \infty)$ von σ^2 ab und ist daher eine Schätzfunktion für σ^2.
Standard-Notation: Ist $T : \mathbb{R}^n \to \Theta$ ein Schätzer für ϑ, dann schreibt man

$$\widehat{\vartheta} = T(X_1, \dots, X_n)$$

zu schreiben. Analog: $\widehat{F}_n(x)$ ist Schätzer für $F(x)$, etc.

Empirische Verteilungsfunktion

Allgemeinstes nichtparametrisches Modell:

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} F(x)$$

mit beliebiger Verteilungsfunktion

$$F(x) = P(X_1 \leq x), \qquad x \in \mathbb{R}.$$

Frage:

1. Wie kann man $F(x)$ ohne zusätzliche Annahmen schätzen?
2. Wie kann man einen solchen Schätzer $\widehat{F}(x)$ rechtfertigen?

Empirische Verteilungsfunktion

Empirische Verteilungsfunktion

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(X_i), \qquad x \in \mathbb{R}.$$

Hierbei: $\mathbf{1}_{(-\infty,x]}(X_i) = \mathbf{1}(X_i \leq x)$.

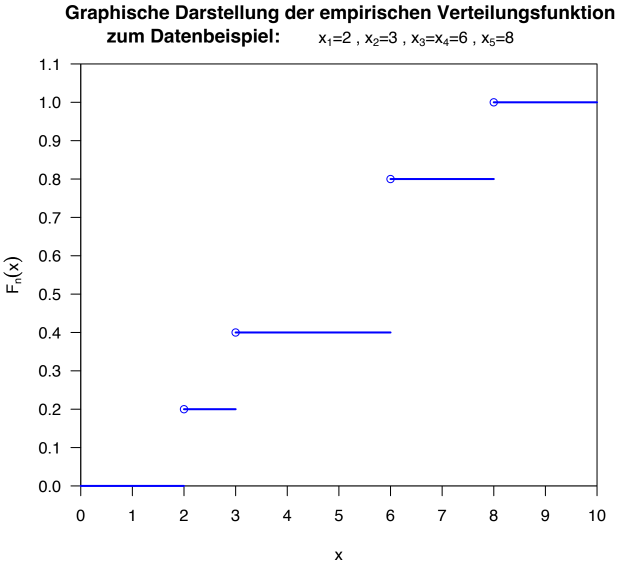
$\widehat{F}_n(x)$: Anteil der Beobachtungen, die kleiner oder gleich x sind.

1. Die Anzahl $n\widehat{F}_n(x)$ der Beobachtungen $\leq x$ ist binomialverteilt mit Parametern n und $p(x) = E(\mathbf{1}(X_i \leq x)) = F(x)$.
2. Daher folgt:

$$E(\widehat{F}_n(x)) = P(X_i \leq x) = F(x), \quad \mathrm{Var}(\widehat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}.$$

3. Nach dem Hauptsatz der Statistik konvergiert $\widehat{F}_n(x)$ mit Wahrscheinlichkeit 1 gegen $F(x)$ (gleichmäßig in x).

Empirische Verteilungsfunktion



Empirische Verteilungsfunktion

Sehr viele Statistiken leiten sich von der empirischen Verteilungsfunktion ab, z.B.:

- Arithmetisches Mittel $\overline{X}_n$.
- Stichprobenvarianz S^2.
- Empirisches Quantil.

(da die Funktion $\widehat{F}(x)$ die geordnete Stichprobe kodiert).

Dichteschätzung

Nichtparametrisches Verteilungsmodell:

$$X_1, \dots, X_n \sim f(x)$$

mit einer Dichtefunktion $f(x)$.
Mögliche Schätzer:

- Histogramm-Schätzer (schätzt eine Vergröberung der Dichte).
- Kerndichteschätzer $\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K(\frac{x-X_i}{h})$, $K : \mathbb{R} \to [0,1]$, $h > 0$ Bandbreite. $K(z) = \frac{1}{2}\mathbf{1}_{[-1,1]}(z)$ liefert das gleitende Histogramm ($\widehat{f}_n(x)$: Anteil der Beob. in $[x-h, x+h]$).

Das Likelihood-Prinzip

Wichtiges Schätzprinzip der parametrischen Statistik.
Motivation:
Information:

1. Ein Restaurant hat zwei Köche A und B.
2. Koch A versalzt die Suppe mit Wkeit 0.1.
3. Koch B versalzt die Suppe mit Wkeit 0.3.

Sie gehen ins Restaurant und essen eine Suppe. Die Suppe ist versalzen.
Wer war der Koch?

Formalisierung

Wir beobachten $x \in \{0, 1\}$. (1: Suppe versalzen, 0: nicht versalzen).
Parameter: $\vartheta \in \Theta = \{A, B\}$ (der wahre Koch).
Statistisches Problem: Schätze ϑ bei Vorliegen der Beobachtung x.
Jeder Koch erzeugt eine W-Verteilung auf $\mathcal{X} = \{0, 1\}$.

$\diagdown\ p_\vartheta(x)$ / $\vartheta\ \diagdown$	Beobachtung 0	1	Summe
A	0.9	0.1	1.0
B	0.7	0.3	1.0

Lösungsheuristik: ϑ umso plausibler, größer $p_\vartheta(x)$ ist.

Likelihood-Funktion

Likelihood-Funktion

Sei $p_\vartheta(x)$ eine Zähldichte (in $x \in \mathcal{X}$) und $\vartheta \in \Theta$ ein Parameter.
Für eine gegebene (feste) Beobachtung $x \in \mathcal{X}$ heißt die Funktion

$$L(\vartheta|x) = p_\vartheta(x), \qquad \vartheta \in \Theta,$$

Likelihood-Funktion.

Likelihood-Prinzip

Ein Verteilungsmodell ist bei gegebenen Daten plausibel, wenn es die Daten mit hoher Wahrscheinlichkeit erzeugt. Entscheide Dich für das plausibelste Verteilungsmodell!

Situation 1:
Diskreter Parameterraum $\Theta = \{\vartheta_1, \ldots, \vartheta_L\}$.
Diskreter Stichprobenraum $\mathcal{X} = \{x_1, \ldots, x_K\}$.

	x_1	$\ldots$	x_K	Summe
ϑ_1	$p_{\vartheta_1}(x_1)$	$\ldots$	$p_{\vartheta_1}(x_K)$	1
ϑ_2	$p_{\vartheta_2}(x_1)$	$\ldots$	$p_{\vartheta_2}(x_K)$	1
$\vdots$	$\vdots$		$\vdots$	
ϑ_L	$p_{\vartheta_L}(x_1)$	$\ldots$	$p_{\vartheta_L}(x_K)$	1

Algorithmus: Bestimme Spaltenmaximum für gegebene Beobachtung $x \in \{x_1, \ldots, x_K\}$.

Verallgemeinerung

Situation 2: (Standardfall bei diskreten Beobachtungen)
Parameterraum $\Theta \subset \mathbb{R}$ Intervall oder ganz $\mathbb{R}$
Diskreter Stichprobenraum: $\mathcal{X} = \{x_1, x_2, \dots\}$.
Keine Tabellendarstellungen mehr. Zeit für eine formale Definition:

Maximum-Likelihood-Schätzer

$p_\vartheta(x)$ sei Zähldichte (in $x \in \mathcal{X}$).
$\vartheta \in \Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}$ wie oben.
Dann heißt $\widehat{\vartheta} = \widehat{\vartheta}(x) \in \Theta$ **Maximum-Likelihood-Schätzer (ML-Schätzer)**, wenn für festes x gilt:

$$p_{\widehat{\vartheta}}(x) \geq p_\vartheta(x) \qquad \text{für alle } \vartheta \in \Theta.$$

(Falls Maximum nicht eindeutig, so wähle eines aus).
Hierdurch ist eine Funktion $\widehat{\vartheta} : \mathcal{X} \to \Theta$ definiert.

Maximum-Likelihood

- Also: Maximiere $(\vartheta, x) \mapsto p_\vartheta(x)$ für festes x in der Variablen $\vartheta \in \Theta$.
- Typischerweise ist $p_\vartheta(x)$ differenzierbar in ϑ.
- Wende bekannte Methoden zur Maximierung an.

Likelihood für Dichten

Problem: Was tun bei stetigen Variablen: $X \sim f_X(x)$?
Für alle $x \in \mathbb{R}$ gilt:

$$P(X = x) = 0$$

Wie kann man jetzt eine Likelihood-Funktion definieren?

Likelihood für Dichten

Idee:

1. Beobachtung x sei fest.
2. Vergröbere die Information 'x beobachtet' zu: 'ungefähr x beobachtet':
$$\{x\} \mapsto [x - dx, x + dx].$$
dx 'infinitesimal' klein.
3. Jetzt ist die Likelihood wie oben definiert:
$$L(\vartheta | [x - dx, x + dx]) = \int_{x-dx}^{x+dx} f_\vartheta(s)\, ds \approx f_\vartheta(x) \cdot (2dx).$$
4. Die rechte Seite wird maximiert, wenn $\vartheta \mapsto f_\vartheta(x)$ maximiert wird.

Likelihood für Dichten

Likelihood für Dichten

$f_\vartheta(x)$ eine Dichtefunktion (in x), $\vartheta \in \Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}$.
Für festes x heißt die Funktion

$$L(\vartheta|x) = f_\vartheta(x), \qquad \vartheta \in \Theta,$$

Likelihood-Funktion. $\widehat{\vartheta} \in \Theta$ heißt **Maximum-Likelihood-Schätzer**, wenn bei festem x gilt:

$$f_{\widehat{\vartheta}}(x) \geq f_\vartheta(x), \quad \text{für alle } \vartheta \in \Theta.$$

Likelihood einer Stichprobe

Kompakt: $X \sim f_\vartheta(x)$, f_ϑ eine Zähldichte oder Dichtefunktion. Dann ist

$$L_\vartheta(x) = f_\vartheta(x)$$

Sei nun speziell $\mathbf{X} = (X_1, \ldots, X_n)'$ mit

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$$

Dann ist die gemeinsame (Zähl-) Dichte die Produkt-(Zähl-) Dichte.
Also:

$$L_\vartheta(\mathbf{x}) = f_\vartheta(x_1) \cdots f_\vartheta(x_n)$$

(Gilt für Zähldichten und Dichtefunktionen).

Beispiele

Beispiele...

Computerexperiment: Likelihood

```
# Likelihood der B(n,p)-Verteilung
# p: (Vektor der) Erfolgswahrscheinlichkeit(en)
# y: beobachtete Anzahl der Erfolge
# n: Stichprobenumfang

likeli = function( p, n, y ) {
  choose(n,y) * p^y * (1-p)^(n-y)
}

# Bsp: n = 10, y = 7 Erfolge

n = 10; y = 7
pp = seq( 0, 1, len=100 )
L = likeli( pp, n, y )
plot( pp, L, type="l", lwd=2, col="blue" )

# ML-Schä‰zer (numerisch im Intervall [0,1] mit max. Fehler 1e-10)

optimize( likeli, c(0,1), maximum=TRUE, tol=1e-10, n=10, y=7 )
```

Computerexperiment: Likelihood