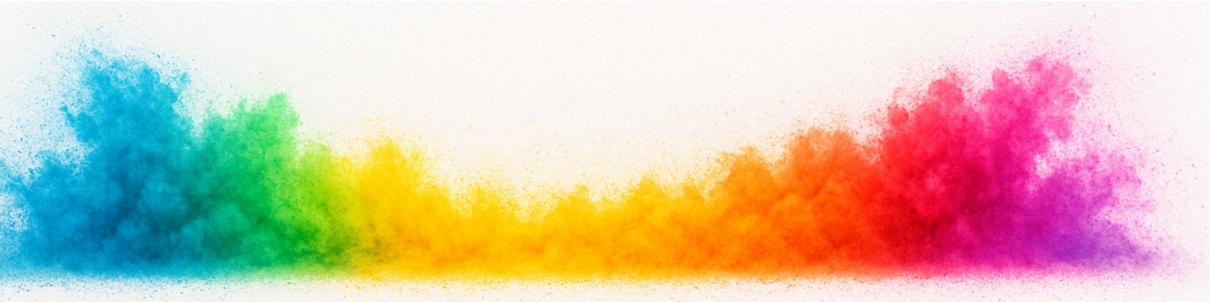
Mathematische Grundlagen von statistischen Tests und Schätzern für Psycholog:innen

Dr. Julien P. Irmer, Dr. Susanne Frick



Vorstellungsrunde

- Wer sind wir?
- Wer seid ihr?
- Was macht ihr?
- Welche Vorkenntnisse bringt ihr mit?
  - Kennt ihr Simulationsstudien?
  - Könnt ihr selbst Simulationen durchführen?
  - Könnt ihr Funktionen in `R` schreiben?


results = numeric(nsim)
seeds = sample.int(10^6, nsim)
for (i in 1:nsim) {
results[i] = simulate_model(i)
})
mean(results)
sim = replicate(8,
X <- rnorm(n)
coef(lm(X ~ 1)[[1]])
hist(sim)
https://tu-dortmund.sciebo.de/s/Ay5RTbzJ2dWo9Qi

Agenda

■ abstrakt ■ anwendbar

1. Begrüßung und Erwartungen
2. Zufallsvariablen verstehen
   - Mengen, Wahrscheinlichkeitsräume und messbare Abbildungen
3. Mittelwerte
   - Erwartungswert, Varianz und große Zahlen
4. Schätzer
   - Allgemeines und Maximum Likelihood Schätzer
   - *Mein eigener ML-Schätzer inkl. Standardfehler*
5. Machine Learning
   - Bias-Varianz-Trade-Off
   - Ausblick

Zeitplan

Zeit	Titel
9:00–9:15	Begrüßung
9:15–10:30	Grundlagen: Wie kommt der Zufall in die Mathematik?
10:30–10:45	Kaffeepause
10:45–12:15	Der Mittelwert, Asymptotiken und zentrale mathematische Sätze
12:15–13:15	Mittagspause
13:15–14:45	Schätzer am Beispiel von Maximum Likelihood
14:45–15:00	Kaffeepause
15:00–16:45	Von robusten Maximum Likelihood zu Machine Learning
16:45–17:00	Abschlussrunde

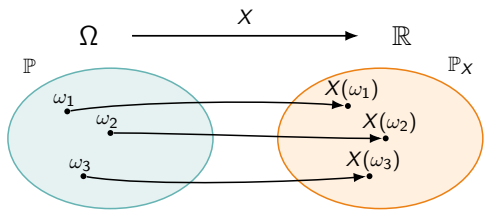
Ziel: Wie bekommen wir Zufall in die Mathematik?

Definition

Eine Zufallsvariable ist eine messbare Abbildung $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})$.

Was ist, bzw. was bedeutet

- Ω, bzw. $\mathbb{R}$
- $\mathcal{A}$, bzw. $\mathcal{B}$
- "messbare Funktion"
- $\mathbb{P}$, bzw. das von X induzierte Maß $\mathbb{P}_X$



Warum

- ist eine Zufallsvariable eine Abbildung und wozu ist das nützlich?
- wird im Zusammenhang mit Zufallsvariablen so viel "*integriert*"?
- reicht es später oft, nur mit $\mathbb{P}_X$ zu rechnen?

Ergebnismenge Ω

Definition

Die *Ergebnismenge* Ω enthält alle möglichen Ausgänge eines Zufallsexperiments.
Ein einzelnes $\omega \in \Omega$ heißt *Ergebnis*.

Endlicher Fall: $\Omega = \{\omega_1, \ldots, \omega_n\}$.
Intuition: „Liste“ aller Möglichkeiten.

Beispiel: zweimal würfeln

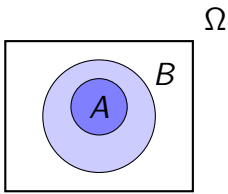
$$\Omega = \Big\{ (\boxed{\bullet}, \boxed{\bullet}), (\boxed{\bullet}, \boxed{\begin{smallmatrix}\bullet & \\ & \bullet\end{smallmatrix}}), \ldots, (\boxed{\bullet}, \boxed{\begin{smallmatrix}\bullet & \bullet \\ \bullet & \bullet \\ \bullet & \bullet\end{smallmatrix}}), (\boxed{\begin{smallmatrix}\bullet & \\ & \bullet\end{smallmatrix}}, \boxed{\bullet}), \ldots, (\boxed{\begin{smallmatrix}\bullet & \bullet \\ \bullet & \bullet \\ \bullet & \bullet\end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \bullet \\ \bullet & \bullet \\ \bullet & \bullet\end{smallmatrix}}) \Big\}$$

$\#\Omega = 36$. Ein mögliches Ergebnis ist z. B. $\omega = (\boxed{\begin{smallmatrix}\bullet & \bullet \\ & \\ \bullet & \bullet\end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \\ & \bullet \\ & \bullet\end{smallmatrix}})$.

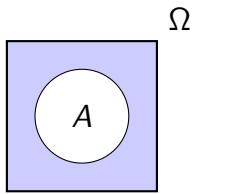
Exkurs: Mengenoperationen

Wichtige Operationen

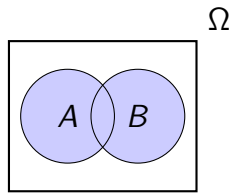
- **Teilmenge:** $A \subseteq B$ (alles in A liegt auch in B)
- **Komplement:** $A^c = \Omega \setminus A$ (alles in Ω, was nicht in A ist)
- **Vereinigung:** $A \cup B = \{\omega : \omega \in A \text{ oder } \omega \in B\}$ (mindestens eins von beiden tritt ein)
- **Schnitt:** $A \cap B = \{\omega : \omega \in A \text{ und } \omega \in B\}$ (beide Ereignisse gleichzeitig)



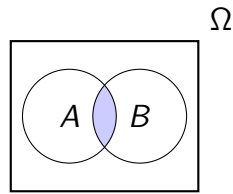
Teilmenge $A \subseteq B$



Komplement A^c



Vereinigung $A \cup B$



Schnitt $A \cap B$

Ereignisse $\mathcal{A}$

Definition

Eine σ-Algebra $\mathcal{A}$ auf Ω ist eine Menge von Teilmengen von Ω, für die gilt

- $\emptyset, \Omega \in \mathcal{A}$ ("unmögliche" und "sichere" Ergebnisse immer mit dabei)
- Wenn $A \in \mathcal{A}$, dann auch $\Omega \setminus A = A^c \in \mathcal{A}$. (auch das Gegenteil)
- Wenn $A_i \in \mathcal{A}$, dann $\bigcup_i A_i \in \mathcal{A}$. (auch "oder"-Kombinationen)

Warum nützlich?

- *Vergleichbarkeit:* für jeden Sachverhalt ist auch das Gegenteil definierbar.
- *Stabilität:* Ereignisse können logisch verknüpft werden („oder", „und").
- *Messbarkeit:* Grundlage, um Wahrscheinlichkeiten sinnvoll zu definieren.

Beispiel

Ereignis "Summe ≥ 11": $A = \left\{ (\square, \square), (\square, \square), (\square, \square) \right\} \in \mathcal{A}$

insb. $A \subseteq \Omega$.

Wahrscheinlichkeitsmaß $\mathbb{P}$

Definition

Ein *Wahrscheinlichkeitsmaß* $\mathbb{P} : \mathcal{A} \to [0, 1]$ erfüllt:

- Nichtnegativität: $\mathbb{P}(A) \geq 0$ für alle $A \in \mathcal{A}$
- σ-Additivität: Für disjunkte A_i gilt $\mathbb{P}\big(\biguplus_i A_i\big) = \sum_i \mathbb{P}(A_i)$.
- Normiertheit (sonst nur ein Maß): $\mathbb{P}(\Omega) = 1$

Beispiel: zweimal ideal würfeln

$\mathbb{P}(\{\omega\}) = 1/36$ für jedes $\omega \in \Omega$,
faire Würfel, alle gleich wahrscheinlich;
also $(\#\Omega = 36, \#A = 3)$:

$$\mathbb{P}(A) \;=\; \mathbb{P}(\text{“Summe} \geq 11\text{”}) \;=\; \tfrac{3}{36}$$

Übersetzung

Rechnen mit Mengen: $A \cup B$, $A \cap B$, A^c.
Wahrscheinlichkeiten „messen“ die Größe von Ereignissen auf der Skala 0–1.

Exkurs: Rechenregeln für Wahrscheinlichkeiten

Wichtige Regeln

- $\mathbb{P}(\emptyset) = 0$ (unmögliches Ereignis)
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ (Gegenteil)
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ (Doppeltzählen vermeiden)
- $\mathbb{P}(A \cap B) \leq \min\{\mathbb{P}(A), \mathbb{P}(B)\}$ (höchstens so groß wie das kleinere)
- **Union Bound:** $\mathbb{P}(\cup_i A_i) \leq \sum_i \mathbb{P}(A_i)$ (Obergrenze)

Beispiel: zwei Würfel

$A = \{\text{Summe} = 11\} = \{(\boxed{\cdot\cdot}, \boxed{\cdot\cdot\cdot}), (\boxed{\cdot\cdot\cdot}, \boxed{\cdot\cdot}), (\boxed{\cdot\cdot}, \boxed{\cdot\cdot\cdot})\}$, $\mathbb{P}(A) = \frac{2}{36}$

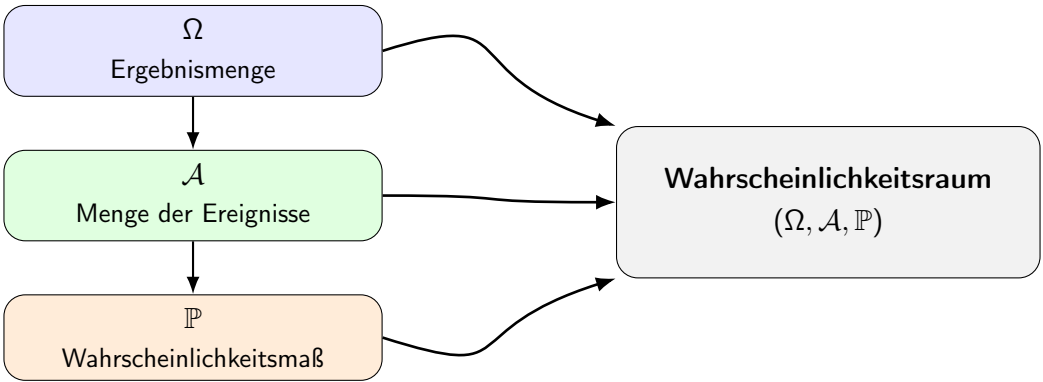
$B = \{\text{beide gleich}\} = \{(\boxed{\cdot}, \boxed{\cdot}), \ldots, (\boxed{\cdot\cdot\cdot}, \boxed{\cdot\cdot\cdot})\}$, $\mathbb{P}(B) = \frac{6}{36}$

$\mathbb{P}(A \cup B) = \frac{2}{36} + \frac{6}{36} - \frac{0}{36} = \frac{8}{36}$

Zufallsvorgang = Wahrscheinlichkeitsraum

Definition

Ein *Zufallsvorgang* wird vollständig modelliert durch den **Wahrscheinlichkeitsraum** $(\Omega, \mathcal{A}, \mathbb{P})$.



Warum brauchen wir den Wahrscheinlichkeitsraum?

Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$

- Ω: alle möglichen Ausgänge (z.B. $\Omega = \{(\boxed{\bullet}, \boxed{\bullet}), \ldots, (\boxed{\vdots}, \boxed{\vdots})\}$)
- $\mathcal{A}$: Ereignisse = Mengen von Ausgängen, über die wir reden können (z.B. „gerade Augenzahl“)
- $\mathbb{P}$: weist diesen Ereignissen Wahrscheinlichkeiten zu (z.B. $\mathbb{P}(\text{gerade Augenzahl}) = 3/6$)

Warum so abstrakt?

- Würfelaugen sind eigentlich keine Zahlen – wir modellieren ihnen zugeordnete Größen.
- Psychologische Zustände oder Gefühle sind (bevor sie gemessen wurden) auch keine Zahlen – wir können sie trotzdem in Ω einbetten.
- Durch die Abstraktion ist das Modell *allgemein und flexibel*.

Von $(\Omega, \mathcal{A}, \mathbb{P})$ zur Zufallsvariable

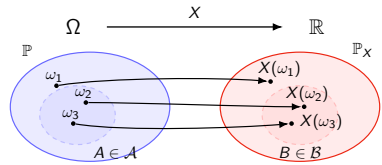
Idee: Wir wollen *beobachtbare* (numerische) Größen modellieren - wir wollen Ergebnisse in $\mathbb{R}$ mit zugehöriger σ-Algebra $\mathcal{B}$. Das leisten **Zufallsvariablen**:

$$X : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow (\mathbb{R}, \mathcal{B})$$

Definition (Messbarkeit)

Für jedes messbare Zielereignis $B \in \mathcal{B}$ gilt:

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}.$$



Induziertes Maß / Verteilung von X

Das von X induzierte Maß

$$\mathbb{P}_X(B) := \mathbb{P}\big(X^{-1}(B)\big), \quad B \in \mathcal{B}, \quad \text{heißt } \textbf{Verteilung von } X.$$

Von $(\Omega, \mathcal{A}, \mathbb{P})$ zur Zufallsvariable - Beispiele mit Würfeln

$\Omega = \{(\boxed{\bullet}, \boxed{\bullet}), \ldots, (\boxed{\vdots}, \boxed{\vdots})\}, \quad \#\Omega = 36.$

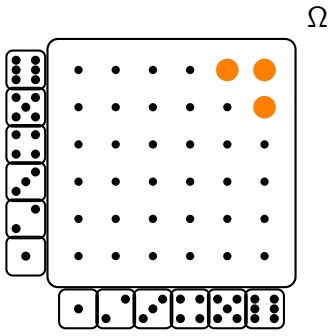
Beispiele

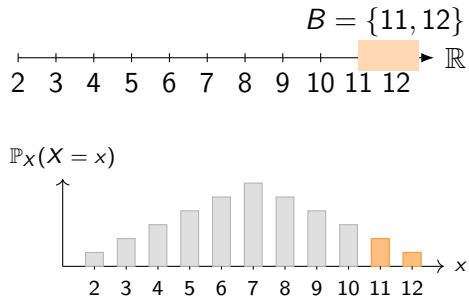
- **Summe:** $S(\boxed{\vdots}, \boxed{\vdots}) = i + j$
  $S^{-1}(\{7\}) = \{(\boxed{\bullet}, \boxed{\vdots}), (\boxed{\cdot\,\cdot}, \boxed{\cdot\,\cdot}), \ldots, (\boxed{\vdots}, \boxed{\bullet})\}$
- **Maximum:** $M(\boxed{\vdots}, \boxed{\vdots}) = \max\{i, j\}$

Alle diese Abbildungen sind messbar, da ihre Urbilder in $\mathcal{A}$ liegen.

Messbarkeit anschaulich: Urbild vs. Bild

Beispiel: Ereignis $B = \{x \in \mathbb{R} : x \geq 11\}$
$\Rightarrow$ „Augensumme beim Würfeln ist mindestens 11“.



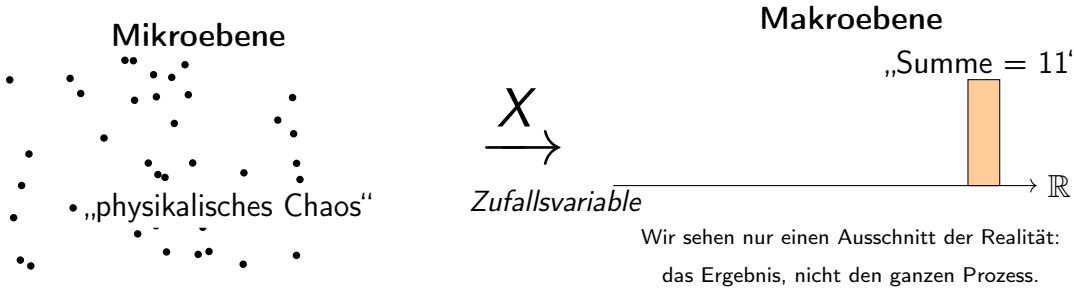


$$X^{-1}(B) = \{(\square, \square), (\square, \square), (\square, \square)\}$$

$$\mathbb{P}_X(B) = \mathbb{P}(X \geq 11) = \mathbb{P}\big(X^{-1}(B)\big)$$

Figures created with ChatGPT 5

Vom Chaos zum Modell: Warum Zufallsvariablen?



Kernaussage: Zufallsvariablen reduzieren komplexe Realität auf *beobachtbare Größen* – und machen Wahrscheinlichkeit berechenbar.

Warum braucht die Psychologie Zufallsvariablen?

- In Experimenten erfassen wir nie die „ganze Realität“ – sondern immer nur einen **Ausschnitt**:
  - Reaktionszeit
  - Skalenwerte
  - Antwort „ja / nein“
- Zufallsvariablen übersetzen *komplexe innere Prozesse* (Aufmerksamkeit, Gedächtnis, Motivation) in **messbare Größen**.
- Nur dadurch können wir diesen Größen **Wahrscheinlichkeiten und Unsicherheit** zuordnen.
- Ohne diese Abbildung gäbe es keine **statistische Inferenz** – weder Hypothesentests noch Konfidenzintervalle.
- Unsere Daten sind immer nur ein **Teil der Wirklichkeit** – wir beobachten das Ergebnis, nicht den ganzen Prozess.

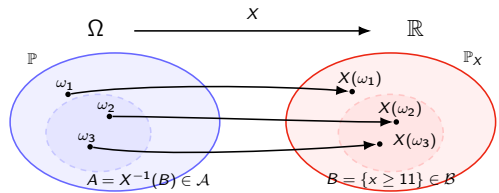
Verteilungen: von $\mathbb{P}_X$ zu F und f am Beispiel $X \geq 11$

Pushforward-Verteilung

$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B))$.

Beispiel

- X = Augensumme,
  $\Omega = \{(\boxdot, \boxdot), \ldots, (\boxplus, \boxplus)\}$.
- $B = \{x \geq 11\} = \{11, 12\}$, $A = X^{-1}(B)$.
- $\mathbb{P}_X(B) = 3/36$, mit
  $p(11) = 2/36$, $p(12) = 1/36$.
- $X^{-1}(B) =$
  $\{(\boxplus, \boxplus), (\boxplus, \boxplus), (\boxplus, \boxplus)\}$.



Darstellungen

$F_X(x) = \mathbb{P}(X \leq x)$ (Verteilungsfunktion)
$p(x) = \mathbb{P}(X = x)$, $F(x) = \sum_{t \leq x} p(t)$ (diskret)
$f(x) = F'(x)$, $\mathbb{P}(X \in B) = \int_B f(x)\, dx$ (stetig)

Anschaulich

„Dichte“ = lokale Wahrscheinlichkeit, „Verteilungsfunktion“ = aufsummierte Wahrscheinlichkeit.

Warum überall Integrale?

Darstellungen

$F_X(x) = \mathbb{P}(X \le x)$ (Verteilungsfunktion)
$p(x) = \mathbb{P}(X = x),\ F(x) = \sum_{t \le x} p(t)$ (diskret)
$f(x) = F'(x),\ \mathbb{P}(X \in B) = \int_B f(x)\,dx$ (stetig)

Erwartungswert

Für messbares $g : \mathbb{R} \to \mathbb{R}$:

$$\mathbb{E}[g(X)] = \begin{cases} \sum_x g(x)\,p(x), & \text{diskret,} \\ \int_{\mathbb{R}} g(x)\,f(x)\,dx, & \text{stetig.} \end{cases}$$

Interpretation

Wir rechnen über X nur noch auf $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$. $\to$ Darum tauchen Dichten, Verteilungsfunktionen, Likelihoods und *Integrale überall* auf.

Wie Zufallsvariablen konvergieren

Definitionen

Fast sicher (f.s.): $X_n \xrightarrow{f.s.} X$, wenn $\mathbb{P}\{\lim X_n = X\} = 1$.

In Wahrscheinlichkeit (Probability, $\mathbb{P}$): $X_n \xrightarrow{\mathbb{P}} X$, wenn $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$.

In Verteilung (Distribution, d): $X_n \xrightarrow{d} X$, wenn $F_{X_n}(x) \to F_X(x)$ (stetig in x).

Beziehungen

$$\text{f.s.} \quad \Rightarrow \quad \text{in W-keit} \quad \Rightarrow \quad \text{in Verteilung}$$

Umkehrungen im Allgemeinen falsch.

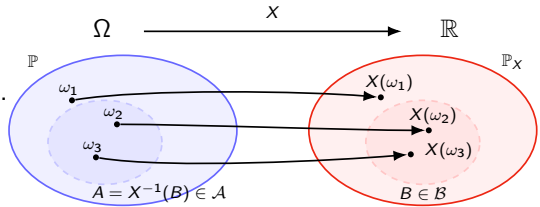
Beispiele

- In W-keit, nicht f.s.: $X_n = \mathbb{1}_{\{U_n \leq 1/n\}}$.
- In Verteilung, nicht in W-keit: $X_{2n-1} = U$, $X_{2n} = 1 - U$.

Wie kommt Zufall in die Mathematik?

Die Bausteine

- Ω: alle **theoretisch möglichen Welten**.
- $\mathcal{A}$: **Fragenkatalog** (Ereignisse), über die wir sprechen können.
- $\mathbb{P}$: **Unsicherheit als Zahl** $0-1$.
- X: **Übersetzung ins Messbare** (Beobachtbares in $\mathbb{R}$).



Einordnung: Wo dockt das an?

- **Analysis/Integration**: Erwartungswerte $\mathbb{E}[g(X)] = \int g \, d\mathbb{P}_X$, Integralrechenregeln für $\mathbb{E}[\cdot]$
- **Lineare Algebra/Geometrie**: (Ko)varianz, Projektionen, orthogonale Zerlegungen.
- **Optimierung/Numerik**: MLE/Log-Likelihood, Gradient/Hesse, Standardfehler.

Kernaussage

$(\Omega, \mathcal{A}, \mathbb{P})$ + messbares $X \Rightarrow \mathbb{P}_X$: Zufall wird *präzise* und *rechenbar* — Basis für Erwartungen, Likelihood & Inferenz.

Zeitplan

Zeit	Titel
9:00–9:15	Begrüßung
9:15–10:30	Grundlagen: Wie kommt der Zufall in die Mathematik?
10:30–10:45	Kaffeepause
10:45–12:15	Der Mittelwert, Asymptotiken und zentrale mathematische Sätze
12:15–13:15	Mittagspause
13:15–14:45	Schätzer am Beispiel von Maximum Likelihood
14:45–15:00	Kaffeepause
15:00–16:45	Von robusten Maximum Likelihood zu Machine Learning
16:45–17:00	Abschlussrunde

Figures created with ChatGPT 5

Zeitplan

Zeit	Titel
9:00–9:15	Begrüßung
9:15–10:30	Grundlagen: Wie kommt der Zufall in die Mathematik?
10:30–10:45	Kaffeepause
10:45–12:15	Der Mittelwert, Asymptotiken und zentrale mathematische Sätze
12:15–13:15	Mittagspause
13:15–14:45	Schätzer am Beispiel von Maximum Likelihood
14:45–15:00	Kaffeepause
15:00–16:45	Von robusten Maximum Likelihood zu Machine Learning
16:45–17:00	Abschlussrunde

Teaser: Was passiert bei großen Stichproben?

$$\frac{1}{\sqrt{n}} X \approx N(\mu, \sigma)$$

Zentrale Fragen

- Was ist und bedeutet ein **Erwartungswert**?
- Welche Eigenschaften hat die **Varianz**?
- Was sind Erwartungswert und Varianz des **Mittelwertes**?
- Welche gundlegenden Sätze gelten für **große Zahlen**?
  - Wie verhalten sich der Erwartungswert und die Häufigkeitsverteilung bei **steigender Stichprobengröße**?
  - Warum ist so vieles **normalverteilt**?

Mittelwert und Erwartungswert

- (Erinnerung: Mittelwert: $\frac{h_1}{n}a_1 + \frac{h_2}{n}a_2 + \ldots \frac{h_k}{n}k$) mit a_k = Wert, h_k = Häufigkeit
- Diskrete Zufallsvariable

$$\mathbb{E}[X] = p_1x_1 + p_2x_2 + \cdots + p_kx_k = \sum_{i \geq 1} p_ix_i$$
$$\equiv f(x_1)x_1 + f(x_2)x_2 + \ldots f(x_k)x_k$$

  p_k = Wahrsch.verteilung, $f(x_k)$ = Wahrsch.funktion/Dichte

- Stetige Zufallsvariable

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x)dx$$

in Worten

Der Erwartungswert ist ein Mittel, gewichtet mit den Wahrscheinlichkeiten.

Rechenregeln für Erwartungswerte

- Transformationsregel: Für $Y = g(x)$ mit einer reellen Funktion $g(x)$ gilt:

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_{i \geq 1} g(x_i) p_i = \sum_{i \geq 1} g(x_i) f(x_i)$$

  z.B. Für $Y = aX + b$ gilt $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
- Ist $f(x)$ symmetrisch um c, so ist $\mathbb{E}[X] = c$
- Summe von Zufallsvariablen gewichtet mit beliebigen Konstanten $a_1, \ldots, a_n$:

$$\mathbb{E}[a_1 X_a + \cdots + a_n X_n] = a_1 \mathbb{E}[X_1] + \cdots + a_n \mathbb{E}[X_n]$$

- Produkt unabhängiger Zufallsvariablen: $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Varianz und Verschiebungssatz

$$\begin{aligned}\mathbb{V}\mathrm{ar}[X] &= \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^2 \, f(x) dx \\ &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned}$$

Implikationen

Der Verschiebungssatz wird häufig verwendet, wenn mit Erwartungswerten gerechnet wird. (Die Herleitung erfolgt analog zur nächsten Folie.)

Varianz als Minimum der Quadratischen Abweichung

$\mathbb{E}\left[(X-c)^2\right]$ wird minimal für $c = \mathbb{E}[X]$:

$$
\begin{aligned}
&\mathbb{E}\left[(X-c)^2\right] \\
=&\mathbb{E}\left[X^2 - 2Xc + c^2\right] \\
=&\mathbb{E}[X^2] - 2c\mathbb{E}[X] + c^2 \\
=&\mathbb{E}[X^2] + \mathbb{E}[X]^2 - \mathbb{E}[X]^2 - 2c\mathbb{E}[X] + c^2 \\
=&\mathbb{E}[X^2] - \mathbb{E}[X]^2 + (c - \mathbb{E}[X])^2 \\
=&\mathbb{V}\mathrm{ar}[X] + (c - \mathbb{E}[X])^2
\end{aligned}
$$

Implikationen / in Worten

Der Erwartungswert ist diejenige Konstante, von der die quadratische Abweichung einer Zufallsvariablen am kleinsten ist. Diese Abweichung ist die Varianz.

Weitere Eigenschaften der Varianz

- Verschiebungsregel allgemeiner:

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}\left[(X-c)^2\right] - (\mathbb{E}[X]-c)^2$$

- Lineare Tranformation

$$\mathbb{V}\mathrm{ar}[aX+b] = a^2\mathbb{V}\mathrm{ar}[X]$$

- Für unabhängige Zufallsvariablen gewichtet mit beliebigen Konstanten $a_1, \ldots, a_n$:

$$\mathbb{V}\mathrm{ar}[a_1X_1 + \cdots + a_nX_n] = a_1^2\mathbb{V}\mathrm{ar}[X_1] + \cdots + a_n^2\mathbb{V}\mathrm{ar}[X_n]$$

Erwartungswert und Varianz des Mittelwertes

für unabhängig und identisch verteilte X, gilt:

Erwartungswert

$$\begin{aligned}\mathbb{E}\left[\mathrm{M}_n(X)\right] &= \frac{1}{n}\left(\mathbb{E}[X_1] + \mathbb{E}[X_2] + \ldots \mathbb{E}[X_n]\right) \\ &= \frac{1}{n} \cdot n \cdot \mathbb{E}[X] \\ &= \mathbb{E}[X]\end{aligned}$$

Varianz

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\mathrm{M}_n(X)\right] &= \frac{1}{n^2}\left(\mathbb{V}\mathrm{ar}[X_1] + \mathbb{V}\mathrm{ar}[X_2] + \cdots + \mathbb{V}\mathrm{ar}[X_n]\right) \\ &= \frac{1}{n^2} \cdot n \cdot \mathbb{V}\mathrm{ar}[X] \\ &= \frac{\mathbb{V}\mathrm{ar}[X]}{n}\end{aligned}$$

Der Erwartungswert des Mittelwertes - Beispiel

X = Summe der Augenzahl bei zweifachem Würfelwurf 

bei 4 realisierten Wiederholungen $\mathrm{M}_4(x) = \frac{1}{4} \cdot (7 + 5 + 7 + 9)$

sortiert nach Ausprägung $= \frac{1}{2} \cdot 7 + \frac{1}{4} \cdot 5 + \frac{1}{4} \cdot 9$

für n Wiederholungen $\mathrm{M}_n(x) = \frac{1}{n} \cdot (x_1 + x_2 + \cdots + x_n)$

nach k Ausprägungen "sortiert" $= \frac{h_1}{n} \cdot x_1 + \frac{h_2}{n} \cdot x_2 + \cdots + \frac{h_k}{n} \cdot x_k$

Der Erwartungswert des Mittelwertes - Beispiel

X = Summe der Augenzahlen bei zweifachem Würfelwurf

Mittelwert

$$\begin{aligned}\mathrm{M}_n(x) &= \frac{h_1}{n} \cdot x_1 + \frac{h_2}{n} \cdot x_2 + \cdots + \frac{h_k}{n} \cdot x_k \\ &= \sum_{i=1}^{k} \frac{h_i}{n} x_i\end{aligned}$$

Erwartungswert

$$\begin{aligned}\mathbb{E}[X] &= p_1 x_1 + p_2 x_2 + \cdots + p_k x_k \\ &= \sum_{i=1}^{k} p_i x_i\end{aligned}$$

Schlussfolgerung

Damit sich der Mittelwert dem Erwartungswert annähert, muss jede relative Häufigkeit $\frac{h_i}{n}$ sich der Wahrscheinlichkeit p_i annähern

Schwaches Gesetz der großen Zahl(en)

Für beliebig kleines $c > 0$ gilt

$$\mathbb{P}\left(|\mathrm{M}_n(X) - \mathbb{E}[X]| \leq c\right) \to 0 \quad \text{für} \quad n \to \infty$$

In Worten

Der Mittelwert $\mathrm{M}_n(X)$ konvergiert in Wahrscheinlichkeit gegen den Erwartungswert $\mathbb{E}[X]$.

Beweis mit Ungleichung von Tschebyscheff

Ungleichung von Tschebyscheff:

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq c\right) \leq \frac{\mathbb{V}\mathrm{ar}[X]}{c^2}$$

$$\mathbb{P}\left(|X - \mathbb{E}[X]| < c\right) \geq 1 - \frac{\mathbb{V}\mathrm{ar}[X]}{c^2}$$

Für den Mittelwert $\mathrm{M}_n(X)$ gilt:
$\mathbb{E}[\mathrm{M}(x)] = \mathbb{E}(X)$ und
Für $n \to \infty$: $\mathbb{V}\mathrm{ar}[\mathrm{M}_n(X)] = \frac{\mathbb{V}\mathrm{ar}[X]}{n} \to 0$
daher $\mathbb{P}\left(|\mathrm{M}_n(X) - \mathbb{E}[X]| \geq c\right) \to 0$

Intuition

Wenn die Varianz gegen 0 geht, wird die Abweichung zwischen Mittelwert und Erwartungswert beliebig klein.

Voraussetzung ist hier, dass $\mathbb{V}\mathrm{ar}$ und $\mathbb{E}$ endlich sind.

Beispiel: Simulation

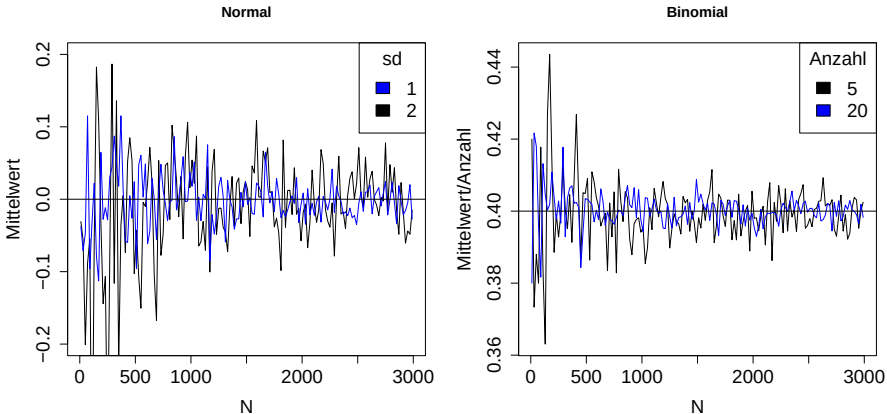


Abbildung: Mittelwert in Abhängigkeit der Stichprobengröße N

Hauptsatz der Statistik / Konvergenz in Verteilung (Satz von Glivenco-Cantelli)

X Zufallsvariable mit Verteilungsfunktion $F(x)$
Dann gilt für die Verteilungsfunktion $F_n(x)$ gebildet aus (unabhängig) und identisch wie X verteilten $X_1, \ldots, X_n$:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| \leq c\right) \to 1 \quad \text{für} \quad n \to \infty \tag{1}$$

In Worten

Die empirische Verteilungsfunktion approximiert die theoretische für $n \to \infty$.

sup = Supremum = "kleinste obere Schranke", kleinster Wert, der $\geq$ alle Werte ist

Beispiel: Simulation

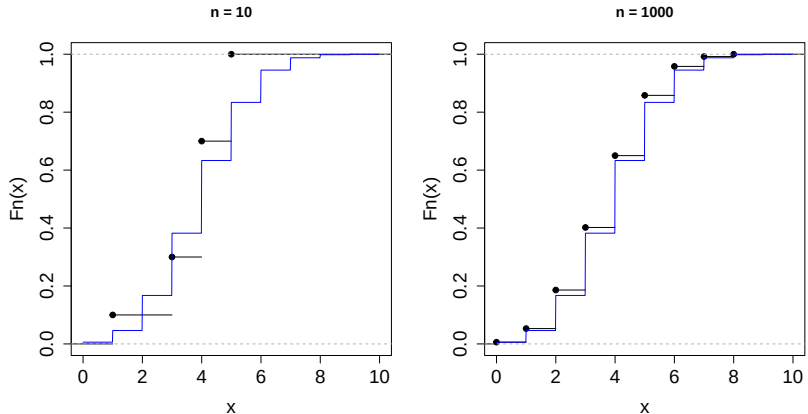


Abbildung: Empirische vs. theoretische Verteilungsfunktion, Binomialverteilung mit $p = 0.4$ und Anzahl $= 10$

Figures from pxhere.com under CC0

Aufgabe

- Testet das GGZ oder den Hauptsatz der Statistik mit anderen Verteilungen (z.B. Poisson) oder anderen Verteilungsparametern (z.B. SD = 2).
  - Beispiele für Verteilungen: Poisson, Beta, Exponential, Hypergeometrisch ...
  - Bennenung der Funktionen in R am Beispiel Normalverteilung (`norm`: Dichte `dnorm`, Verteilung `pnorm`, Quantil `qnorm`, Zufallszahl `rnorm`
- GGZ: `1_example_GGZ.R`
- Hauptsatz der Statistik: `example_glivenco_cantelli.R`

Zentraler Grenzwertsatz (ZGWS)

Für die standardisierte Summe Z_n unabhängig und identisch verteilter* X, gilt an jeder Stelle $z \in \mathbb{R}$:

$$Z_n = \frac{X_1 + X_2 + \ldots X_n - n\mu}{\sqrt{n}\sigma} = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

$$F_n(z) \to \Phi(z) \text{ für } n \to \infty \quad \text{alternativ: } \sqrt{n}\, Z_n \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\ 1)$$

Φ = Verteilungsfunktion der Standardnormalverteilung

Anders ausgedrückt

Die Verteilung einer Summe X von Zufallsvariablen nähert sich für $n \to \infty$ einer Normalverteilung an: $\frac{1}{\sqrt{n}} X \approx \mathcal{N}(\mu, \sigma)$

*Allgemeinere Varianten, die i.i.d. lockern; keine der X_i sollte die anderen deutlich dominieren (Fahrmeier et al., 2023, S.324).

Aufgabe zur Auswahl

Büchervergleich: Der zentrale Grenzwertsatz

- Psychologie: Eid et al. (2015). Statistik und Forschungsmethoden. Kap. 8.4
- Statistik: Fahrmeier et al. (2023). Statistik - Der Weg zur Datenanalyse. Kap. 7.1.2 https://link.springer.com/book/10.1007/978-3-662-67526-7
- Mathematik: Klenke (2020). Wahrscheinlichkeitstheorie. Kap. 15.5 https://link.springer.com/book/10.1007/978-3-642-36018-3

R-Übung

`2_example_glivenco_cantelli.R` für zentralen Grenzwertsatz umbauen

Figures created with ChatGPT 5

Zusammenfassung: Von der Zufallsvariable zur Verteilung

Was nützt uns das?

- Wir können beschreiben, wie eine Zufallsvariable verteilt ist (Erwartungswert und Varianz).
- Der Erwartungswert minimiert die quadratische Abweichung.
  (siehe auch Regression)
- asymptotisches Verhalten von Zufallsvariablen bestimmen: GGZ, Satz von Glivenco-Cantelli und ZGWZ.

Teaser: Ein Schätzer ist auch eine Zufallsvariable ...

Zeitplan

Zeit	Titel
9:00–9:15	Begrüßung
9:15–10:30	Grundlagen: Wie kommt der Zufall in die Mathematik?
10:30–10:45	Kaffeepause
10:45–12:15	Der Mittelwert, Asymptotiken und zentrale mathematische Sätze
12:15–13:15	Mittagspause
13:15–14:45	Schätzer am Beispiel von Maximum Likelihood
14:45–15:00	Kaffeepause
15:00–16:45	Von robusten Maximum Likelihood zu Machine Learning
16:45–17:00	Abschlussrunde

Figures created with ChatGPT 5

Zeitplan

Zeit	Titel
9:00–9:15	Begrüßung
9:15–10:30	Grundlagen: Wie kommt der Zufall in die Mathematik?
10:30–10:45	Kaffeepause
10:45–12:15	Der Mittelwert, Asymptotiken und zentrale mathematische Sätze
12:15–13:15	Mittagspause
13:15–14:45	Schätzer am Beispiel von Maximum Likelihood
14:45–15:00	Kaffeepause
15:00–16:45	Von robusten Maximum Likelihood zu Machine Learning
16:45–17:00	Abschlussrunde

Teaser: Asymptotische Normalität von Schätzern

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\Big(0,\, \mathcal{I}(\theta_0)^{-1}\Big)$$

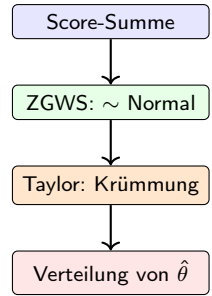
Zentrale Fragen

- Was ist ein **Modell**?
- Was ist ein **Parameter**?
- Was genau ist ein **Schätzer** $\hat{\theta}$?
- Welche **Arten** von **Schätzern** gibt es?
- Warum (und wann) gilt die **asymptotische Normalität**?
- Welche **Regularitätsannahmen** braucht man?
- Wie entstehen daraus **Standardfehler und Konfidenzintervalle**?

Werkzeugkasten für die MLE-Asymptotik

Bausteine

- **Modell & Parameter** (θ, θ_0).
- **Likelihood & Log-Likelihood**.
- **Ableitungen:** Steigung (Score), Krümmung (Hesse).
- **Fisher-Information:** Erwartung von Krümmung/Varianz.
- **Taylor-Approximation:** lokal Parabel.
- **GGZ & ZGWS:** Summen $\rightarrow$ Grenzwerte.
- **Regularität:** Glattheit, Identifizierbarkeit.



Heute auch: Binomial-Beispiel & Sandwich (robust).

Modell und Parameter

Statistisches Modell

$\mathcal{M} = \{P_\theta : \theta \in \Theta\}$. Familie von Wahrscheinlichkeitsmaßen P_θ auf $(\mathbb{R}, \mathcal{B})$.

Intuition

- Modell = Menge aller *Kandidaten-Verteilungen*
- legt fest, welche Verteilungsarten wir zulassen

Beispiel

- alle $N(\mu, \sigma^2)$ mit beliebigen $\mu \in \mathbb{R}, \sigma > 0$

Parameter

Ein **Parameter** θ ist ein Element, der eine konkrete Verteilung P_θ innerhalb des Modells auswählt. Θ ist Parameterraum.

Intuition

- Parameter = *Auswahlmechanismus* innerhalb des Modells
- bestimmt, welche Verteilung tatsächlich gilt, auch durch Θ-Wahl

Beispiel

- Spezifisches $\theta_0 = (\mu_0, \sigma_0^2)$ in $N(\mu, \sigma^2)$, z.B. $\mu_0 = 0, \sigma_0^2 = 1$ (std.)

Von Ω zum Schätzer

Zufallsvariablen

$X_i : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})$, liefert $X_i(\omega) = x_i$ für $\omega \in \Omega$.

Schätzer

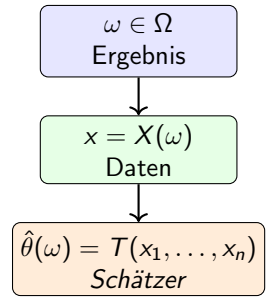
$\hat{\theta} = T(X_1, \ldots, X_n) : (\Omega, \mathcal{A}, \mathbb{P}) \to (\Theta, \mathcal{B}_\Theta)$.

- Auch $\hat{\theta}$ ist eine Zufallsvariable.

Intuition

- ω: konkreter Ausgang des Experiments
- $X(\omega) = x$: gemessene Daten
- $\hat{\theta}(\omega)$: „Prozedur“, die daraus eine Zahl macht

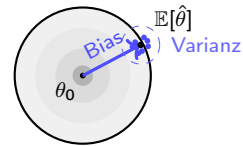
Daten $\to$ Schätzer $\to$ Zahl $\approx$ Parameter



Wie wählt man einen Schätzer?

Allgemeine Definition

Ein **Schätzer** ist eine Funktion der Daten: $\hat{\theta} = T(X_1, \ldots, X_n)$.
$\rightarrow$ *Aber: Es gibt unendlich viele mögliche Funktionen T.*



Zentrale Frage

Formale Kriterien

- **Bias / Erwartungstreue:** $\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta_0$; erwartungstreu falls $= 0$.
- **Varianz:** $\mathbb{V}\text{ar}[\hat{\theta}]$.
- **MSE:** $\text{MSE}(\hat{\theta}, \theta_0) = \mathbb{E}[(\hat{\theta} - \theta_0)^2]$.

Asymptotische Eigenschaften

- **Konsistenz:** $\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0$.
- **Asymptotische Normalität:** $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.
- **Effizienz:** kleinstmögliche Varianz unter erwartungstreuen Schätzern (Cramér–Rao).

θ_0: wahrer Populationsparameter

Methoden zur Konstruktion von Schätzern

Prinzip

Die **Eigenschaften** definieren, was wir wollen. Die **Methoden** liefern Regeln, wie wir $\hat{\theta}$ bestimmen.

Methoden

- **Method of Moments (MoM)** Setzt Stichprobenmomente = Modellmomente.
- **Least Squares (LS)** Minimiert $\sum(y_i - g(\theta))^2$.
- **Maximum Likelihood (ML)** Maximiert $\mathcal{L}(\theta) = \prod f(x_i \mid \theta)$.

Intuition

- MoM: „Passe Mittelwerte/Varianzen an“.
- LS: „Minimiere Fehlerabstand (räumlich)“.
- ML: „Wähle θ, das die beobachteten Daten am wahrscheinlichsten macht“.

Vektorwertige Ableitungen: Gradient, Jacobian und Hessian

Mathematik

Sei $f : \mathbb{R}^p \to \mathbb{R}^k$ stetig partiell differenzierbar.

Gradient (Spezialfall $k = 1$):

$$\nabla f(\theta) = \begin{bmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \\ \vdots \\ \frac{\partial f}{\partial \theta_p} \end{bmatrix} \in \mathbb{R}^p$$

Hessian (für $k = 1$):

$$H(\theta) = \begin{bmatrix} \frac{\partial^2 f}{\partial \theta_1^2} & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2 f}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 f}{\partial \theta_2^2} & \cdots & \frac{\partial^2 f}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 f}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial^2 f}{\partial \theta_p^2} \end{bmatrix} \in \mathbb{R}^{p \times p}$$

Jacobian (allgemein, $k \geq 1$):

$$J(\theta) = \begin{bmatrix} \frac{\partial f_1}{\partial \theta_1} & \frac{\partial f_1}{\partial \theta_2} & \cdots & \frac{\partial f_1}{\partial \theta_p} \\ \frac{\partial f_2}{\partial \theta_1} & \frac{\partial f_2}{\partial \theta_2} & \cdots & \frac{\partial f_2}{\partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_k}{\partial \theta_1} & \frac{\partial f_k}{\partial \theta_2} & \cdots & \frac{\partial f_k}{\partial \theta_p} \end{bmatrix} \in \mathbb{R}^{k \times p}$$

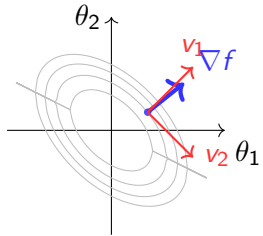
Symmetrie (Satz von Schwarz): Falls $\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}$ stetig ist, dann gilt

$$\frac{\partial^2 f}{\partial \theta_i \partial \theta_j} = \frac{\partial^2 f}{\partial \theta_j \partial \theta_i} \quad \Rightarrow \quad H(\theta) \text{ symmetrisch.}$$

Intuition & Beispiel (mit Grafik)

Intuition

- **Gradient/Jacobian**: Richtung des steilsten Anstiegs.
- **Hessian**: Krümmung; Eigenwerte λ: $\lambda > 0$ Minimum, $\lambda < 0$ Maximum, gemischt $\rightarrow$ Sattel.

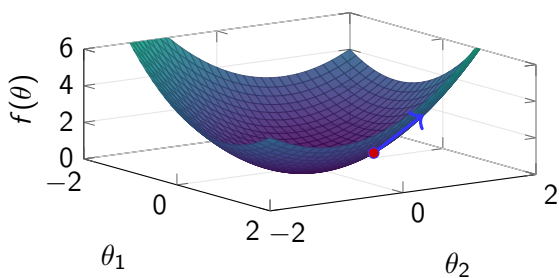


Beispiel ($p = 2$)

$f(\theta_1, \theta_2) = \theta_1^2 + \theta_1\theta_2 + \theta_2^2$.

$$\nabla f = \begin{bmatrix} 2\theta_1 + \theta_2 \\ \theta_1 + 2\theta_2 \end{bmatrix}, \quad H = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

Eigenvektoren: $v_1 = (1, 1)$, $v_2 = (1, -1)$.



Beispiel: Minimum ($\lambda_1, \lambda_2 > 0$)

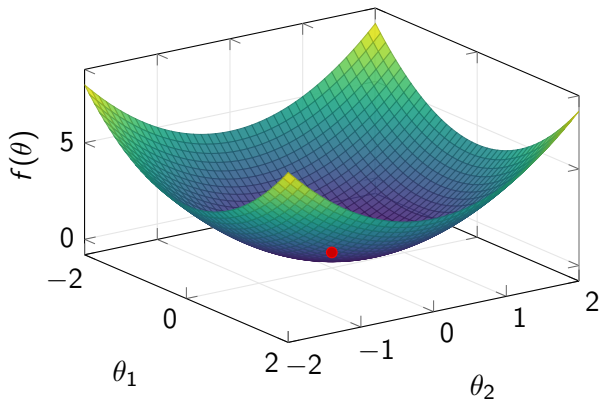
Funktion

$$f(\theta_1, \theta_2) = \theta_1^2 + \theta_2^2$$

Gradient und Hessian

$$\nabla f = \begin{bmatrix} 2\theta_1 \\ 2\theta_2 \end{bmatrix}, \qquad H = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

Eigenwerte: $2, 2 > 0 \Rightarrow$ streng konvex/negativ definit (Minimum).



Beispiel: Maximum ($\lambda_1, \lambda_2 < 0$)

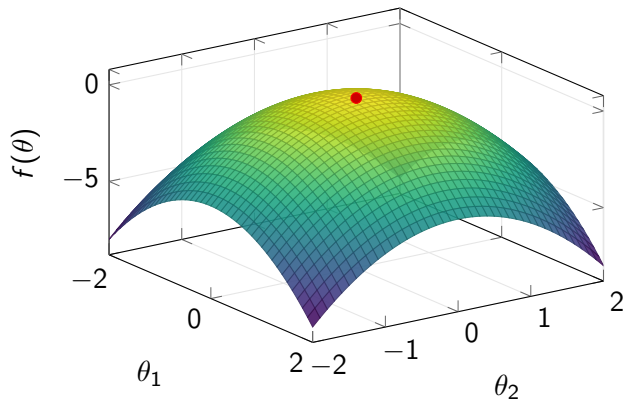
Funktion

$$f(\theta_1, \theta_2) = -\theta_1^2 - \theta_2^2$$

Gradient und Hessian

$$\nabla f = \begin{bmatrix} -2\theta_1 \\ -2\theta_2 \end{bmatrix}, \qquad H = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}.$$

Eigenwerte: $-2, -2 < 0 \Rightarrow$ streng konkav/positiv definit (Maximum).



Beispiel: Sattelpunkt (gemischte Eigenwerte)

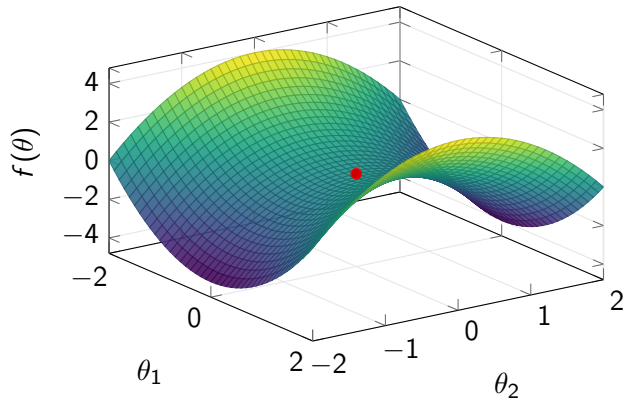
Funktion

$$f(\theta_1, \theta_2) = \theta_1^2 - \theta_2^2$$

Gradient und Hessian

$$\nabla f = \begin{bmatrix} 2\theta_1 \\ -2\theta_2 \end{bmatrix}, \qquad H = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}.$$

Eigenwerte: $+2, -2 \Rightarrow$ Sattelpunkt.



Vom Erwartungswert zur Regression

Allgemeine Idee

Parameter oft definiert als Lösung von $\theta_0 = \arg\min_\theta \ \mathbb{E}\big[(Y - m_\theta(X))^2\big]$,
z.B. $m_\theta(X) = \mu$ (Mittelwert) oder $m_\theta(X) = \beta_0 + \beta_1 X = \mathbb{E}[Y|X]$ (Regression).

Gesetz der großen Zahlen

Da $\mathbb{E}[\cdot]$ unbekannt ist: $\mathbb{E}[(Y - m)^2] \ \approx \ \frac{1}{n}\sum_{i=1}^{n}(y_i - m)^2$
$\rightarrow$ Erwartungswert wird durch Stichprobenmittel ersetzt.

Beispiele

- **Mittelwert:** $\hat{\mu} = \arg\min_\mu \frac{1}{n}\sum(y_i - \mu)^2 \ \Rightarrow \ \hat{\mu} = \bar{y}$.
- **Regression:** $\hat{\theta} = \arg\min_{\beta_0,\beta_1} \frac{1}{n}\sum(y_i - \beta_0 - \beta_1 x_i)^2$. $\Rightarrow \ \beta_1 = \frac{\widehat{\mathbb{C}\mathrm{ov}}[x,y]}{\widehat{\mathbb{V}\mathrm{ar}}[x]}$, $\beta_0 = \bar{y} - \beta_1\bar{x}$.

Ableitung $= 0 \rightarrow$ Score-Gleichungen, $\theta_0 =$ wahrer Populationsparameter.

Least Squares: Normalgleichungen

Quadratsumme und Ableitung

$$\mathcal{Q}(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Minimierung $\Leftrightarrow$ Setze Ableitungen $= 0$ (Partielle Ableitung):

$$\frac{\partial \mathcal{Q}}{\partial \beta_0} = 0, \qquad \frac{\partial \mathcal{Q}}{\partial \beta_1} = 0$$

„Score-Bedingungen“ für OLS.

Lösung (Normalgleichungen)

$$\hat{\beta}_1 = \frac{\widehat{\mathbb{C}\mathrm{ov}}[x, y]}{\widehat{\mathbb{V}\mathrm{ar}}[x]}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Least Squares: Normalgleichungen

Quadratsumme

$$\mathcal{Q}(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Ableitungen = 0 (*Score-Bedingungen*)

$$\frac{\partial \mathcal{Q}}{\partial \beta_0} = -\frac{2}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial \mathcal{Q}}{\partial \beta_1} = -\frac{2}{n} \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Normalgleichungen

$$\sum_{i=1}^{n} y_i = n\,\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2$$

Auflösen nach $\hat{\beta}_1, \hat{\beta}_0$

$$\bar{y} = \beta_0 + \beta_1 \bar{x} \quad \Rightarrow \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \beta_1 \Big(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \Big)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\widehat{\mathbb{C}\mathrm{ov}}[x, y]}{\widehat{\mathbb{V}\mathrm{ar}}[x]}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Hinweis: $\widehat{\mathbb{C}\mathrm{ov}}, \widehat{\mathbb{V}\mathrm{ar}}$ können mit $1/n$ oder $1/(n-1)$ definiert sein; $\hat{\beta}_1$ ist davon unabhängig.

Dichte, Verteilungsfunktion und Likelihood

Dichte / Verteilungsfunktion

- Zufallsvariable $X \sim F_\theta$
- **Dichte:** $f(x \mid \theta)$
- **Verteilungsfunktion:** $F(x \mid \theta) = \mathbb{P}(X \leq x|\theta)$
- **Interpretation:** Parameter θ ist fest, x ist „zufällig“.

Likelihood

- Beobachtungen $x_1, \ldots, x_n$ fix
- **Likelihood:** $\mathcal{L}(\theta) = \prod_{i=1}^n f(x_i \mid \theta)$
- **Log-Likelihood:** $\log \mathcal{L}(\theta) = \sum_{i=1}^n \log f(x_i \mid \theta)$
- **Interpretation:** Daten x_i sind fest, θ wird „gesucht“.

Anschaulich

- Dichte/Verteilungsfunktion: „Wie wahrscheinlich sind Daten, *wenn* θ gilt?“
- Likelihood: „Wie plausibel ist θ, *gegeben* die Daten?“

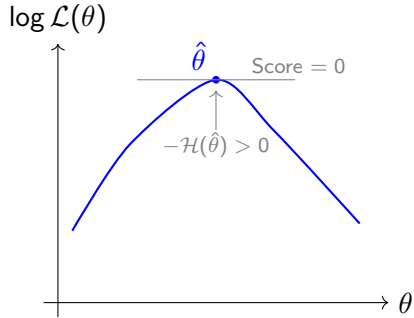
Maximum Likelihood

Definition

- **Likelihood:** $\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$
- **Log-Likelihood:** $\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta)$
- **ML-Schätzer:** $\hat{\theta} = \arg\max_{\theta \in \Theta} \log \mathcal{L}(\theta)$

Beispiele

- Bin(n, p): $\hat{p} = \frac{k}{n}$ (mit $k = \sum x_i$)



Schematische log $\mathcal{L}$: Maximum, Score=0, konkav.

Intuition

- Wähle $\hat{\theta}$, sodass die beobachteten Daten unter dem Modell am plausibelsten sind.
- Am Maximum gilt: Steigung = 0 (Score), Krümmung negativ (Hesse).

Likelihood und Erwartungswert

Sample (datenbasiert)

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta)$$

- definiert den MLE: $\hat{\theta} = \arg\max_\theta \log \mathcal{L}(\theta)$
- basiert auf den beobachteten Daten

Erwartung (theoretisch)

$$\mathbb{E}[\log f(X \mid \theta)]$$

- Grenzwert von $\frac{1}{n} \log \mathcal{L}(\theta)$ für $n \to \infty$
- Grundlage für Konsistenz und Asymptotik

Intuition

- Definition: MLE maximiert die **empirische Log-Likelihood**.
- Theorie: gute Eigenschaften, weil sie asymptotisch die **erwartete Log-Likelihood** approximiert.

Score, Hessian, Fisher-Information

Ableitungen der log-Likelihood

- **Score:** $\mathcal{S}(\theta) = \frac{\partial}{\partial \theta} \log \mathcal{L}(\theta)$ $(\nabla_\theta \log \mathcal{L}(\theta))$
- **Hesse:** $\mathcal{H}(\theta) = \frac{\partial^2}{\partial \theta^2} \log \mathcal{L}(\theta)$ $(\nabla^2_\theta \log \mathcal{L}(\theta))$
- **Fisher-Info:** $\mathcal{I}(\theta_0) = \mathbb{E}_{\theta_0}\big[-\mathcal{H}(\theta_0)\big] = \mathbb{V}\mathrm{ar}_{\theta_0}\big[\mathcal{S}(\theta_0)\big]$ ($d \times d$-Matrix)

Intuition

- Score = zufällig, Steigung der log $\mathcal{L}$. Auch pro Beobachtung bestimmbar.
- Hesse = zufällig, Krümmung der log $\mathcal{L}$.
- Fisher-Information = Erwartung/Varianz *unter dem wahren Parameter* θ_0.
- Am MLE $\hat{\theta}$: $\mathcal{S}(\hat{\theta}) = 0$, $\mathcal{H}(\hat{\theta}) < 0$, bzw. neg. def..

Beispiel: Binomialverteilung $X \sim \mathrm{Bin}(n, p)$

Ableitungen der logLikelihood

$$\log \mathcal{L}(p) = \log \binom{n}{X} + X \log p + (n - X) \log(1 - p)$$

$$\mathcal{S}(p) = \frac{X}{p} - \frac{n - X}{1 - p} \Rightarrow \hat{p} = \tfrac{X}{n}, \qquad \mathcal{H}(p) = -\tfrac{X}{p^2} - \tfrac{n-X}{(1-p)^2}$$

$$\mathcal{I}(p_0) = \mathbb{E}_{p_0}[-\mathcal{H}(p_0)] = \mathbb{V}\mathrm{ar}_{p_0}[\mathcal{S}(p_0)] = \tfrac{n}{p_0(1-p_0)}$$

Interpretation

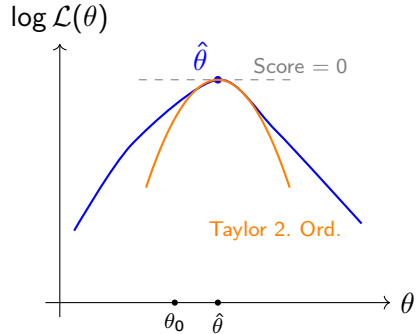
- Zufallsvariable: X = Anzahl Erfolge in n Versuchen.
- Score $= 0 \Rightarrow \hat{p} = X/n$; Hesse < 0 bestätigt Maximum.
- Fisher-Info: Erwartung/Varianz *unter wahrem* p_0.

Taylor-Entwicklung (1D) – 2. Ordnung

Um θ_0 in 2. Ordnung (mit Rest)

Für glattes $g(\theta)$ gilt:

$$g(\hat{\theta}) = \underbrace{g(\theta_0)}_{\text{Wert in } \theta_0} + \underbrace{g'(\theta_0)\,(\hat{\theta} - \theta_0)}_{\text{1. Ordnung: linear (Tangente)}} + \underbrace{\tfrac{1}{2}\, g''(\theta_0)\,(\hat{\theta} - \theta_0)^2}_{\text{2. Ordnung: Krümmung (Parabel)}} + \underbrace{\mathcal{R}_3(\hat{\theta}, \theta_0)}_{\text{Rest}}.$$



Wozu nutzt man das?

- Komplexe Funktionen lokal durch einfache Polynome ersetzbar.
- In Statistik: macht $\log \mathcal{L}$ lokal „parabolisch“.
- Äquivalent zu Linearisierung der Scores $\mathcal{S}$.

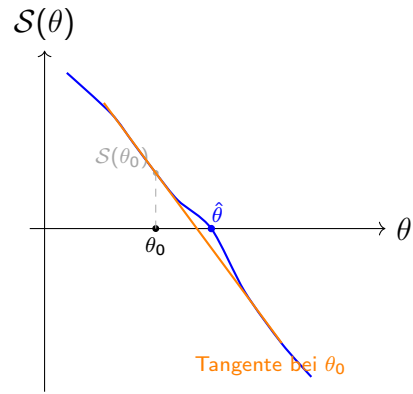
Score-Funktion (1D) – Taylor 1. Ordnung & Linearisierung

- $\mathcal{S}(\theta) = \frac{\partial}{\partial\theta} \log \mathcal{L}(\theta)$,
- $\mathcal{H}(\theta) = \frac{\partial^2}{\partial\theta^2} \log \mathcal{L}(\theta)$,
- $\mathcal{J}(\theta) = -\mathcal{H}(\theta)$.

$$\mathcal{S}(\theta) \;\approx\; \mathcal{S}(\theta_0) + \mathcal{S}'(\theta_0)(\theta - \theta_0),$$

mit $\mathcal{S}'(\theta_0) = \mathcal{H}(\theta_0) = -\mathcal{J}(\theta_0)$. Damit ist

$$\hat{\theta} \;\approx\; \theta_0 - \frac{\mathcal{S}(\theta_0)}{\mathcal{S}'(\theta_0)} \;=\; \theta_0 + \frac{\mathcal{S}(\theta_0)}{\mathcal{J}(\theta_0)}.$$



Linearisierung bei θ_0 (eine Tangente) & Newton-Formel

Warum ist der MLE asymptotisch normal?

1) Definitionen (ausgewertet bei θ_0)

$$\mathcal{S}_n(\theta_0) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i \mid \theta_0), \quad \mathcal{H}_n(\theta_0) = \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(X_i \mid \theta_0),$$

mit $\mathbb{E}[S_n(\theta_0)] = 0$, weil θ_0 im Mittel die Stelle des Maximums der Log-Likelihood ist.

2) Grenzwerte

$$\frac{1}{\sqrt{n}}\mathcal{S}_n(\theta_0) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \mathcal{I}(\theta_0)), \qquad -\frac{1}{n}\mathcal{H}_n(\theta_0) \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}[-\mathcal{H}(\theta_0)]. \qquad (= \mathcal{I}(\theta_0))$$

3) Taylor-Verknüpfung

$$0 = \mathcal{S}_n(\hat{\theta}) \approx \mathcal{S}_n(\theta_0) + \mathcal{H}_n(\theta_0)(\hat{\theta} - \theta_0).$$

$$\implies \sqrt{n}(\hat{\theta} - \theta_0) \approx -\left(\frac{1}{n}\mathcal{H}_n(\theta_0)\right)^{-1} \frac{1}{\sqrt{n}}\mathcal{S}_n(\theta_0).$$

Einschub: $\mathbb{V}\mathrm{ar}$-Rechenregeln multivariater Zufallsvariablen

Sei $X \in \mathbb{R}^p$ Zufallsvektor mit Erwartungswert $\mu = \mathbb{E}[X]$ und Kovarianzmatrix $\Sigma = \mathbb{V}\mathrm{ar}[X]$.

- **Lineare Transformation:** Für konstante Matrix A und Vektor b gilt

$$Y = AX + b \quad \Rightarrow \quad \mathbb{E}[Y] = A\mu + b, \quad \mathbb{V}\mathrm{ar}(Y) = A\Sigma A^\top.$$

- **Kovarianz mit linearer Transformation:**

$$\mathbb{C}\mathrm{ov}[AX, BX] = A\Sigma B^\top.$$

- **Varianz einer Linearkombination:**

$$\mathbb{V}\mathrm{ar}\left[a^\top X\right] = a^\top \Sigma a, \quad a \in \mathbb{R}^p.$$

Klassisch vs. robust: Fisher-Information

Klassisch (korrektes Modell)

$$\underbrace{\mathbb{V}\mathrm{ar}_{\theta_0}\big[\mathcal{S}(\theta_0)\big]}_{\mathcal{J}} \;=\; \underbrace{-\,\mathbb{E}_{\theta_0}\big[\mathcal{H}(\theta_0)\big]}_{\mathcal{H}} \;=\; \mathcal{I}(\theta_0).$$

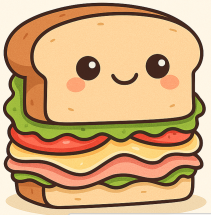
$$\sqrt{n}\,(\hat{\theta}-\theta_0) \;\xrightarrow[n\to\infty]{d}\; \mathcal{N}\big(0,\, \mathcal{I}(\theta_0)^{-1}\big).$$

Robust (Misspezifikation) Sandwich Estimator

$$\sqrt{n}\,(\hat{\theta}-\theta_0) \;\xrightarrow[n\to\infty]{d}\; \mathcal{N}\big(0,\, \mathcal{H}^{-1}\,\mathcal{J}\,\mathcal{H}^{-1}\big),$$

$$\mathcal{H} \;:=\; -\,\mathbb{E}_{\theta_0}\Big[\tfrac{1}{n}\mathcal{H}_n(\theta_0)\Big], \qquad \mathcal{J} \;:=\; \mathbb{V}\mathrm{ar}_{\theta_0}\Big(\tfrac{1}{\sqrt{n}}\mathcal{S}_n(\theta_0)\Big).$$

Klassisch gilt die „Information identity“ $\mathcal{J}=\mathcal{H}=\mathcal{I}(\theta_0)$; robust i. A. $\mathcal{J}\neq\mathcal{H}$ (Sandwich-Varianz). Empirisch: ersetze θ_0 durch $\hat{\theta}$.



Figures created with ChatGPT 5

Einige Regularitätsannahmen

Für die Herleitung benutzt

- **Glattheit & Austausch:** $f(x \mid \theta)$ messbar in x, in einer Umgebung von θ_0 zweifach stetig differenzierbar (also Ableitungen exisitiern und sind gutartig); Erwartung und Ableitung dürfen vertauscht werden
  $\Rightarrow\ \mathbb{E}_{\theta_0}[\mathcal{S}(\theta_0)] = 0,\ \mathbb{E}_{\theta_0}\big[-\frac{1}{n}\mathcal{H}_n(\theta_0)\big] = \mathcal{I}(\theta_0).$
- **GGZ/ZGWS für Score & Hesse:** (X_i) i.i.d., $\mathbb{E}_{\theta_0}\|\mathcal{S}(\theta_0)\|^2 < \infty$. Dann $\frac{1}{\sqrt{n}}\mathcal{S}_n(\theta_0) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \mathcal{I}(\theta_0))$ und $-\frac{1}{n}\mathcal{H}_n(\theta_0) \xrightarrow[n\to\infty]{\mathbb{P}} \mathcal{I}(\theta_0)$.
- **Identifikation & Krümmung:** θ_0 ist Innenpunkt von Θ; die Fisher-Information $\mathcal{I}(\theta_0) = \mathbb{E}_{\theta_0}\big[-\frac{1}{n}\mathcal{H}_n(\theta_0)\big]$ ist **positiv definit** (also invertierbar).

Sichert: Score-Bedingung & Taylor gültig, $\sqrt{n}(\hat{\theta} - \theta_0)$ linear in $\frac{1}{\sqrt{n}}\mathcal{S}_n(\theta_0)$, Varianz $\mathcal{I}(\theta_0)^{-1}$ existiert.

Figures created with ChatGPT 5

Wozu brauchen wir die asymptotische Normalität?

Idee

Für $n \to \infty$ gilt

$$\sqrt{n}\,(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0,\, \mathcal{V}),$$

wobei $\mathcal{V} = \mathcal{I}(\theta_0)^{-1}$ **(klassisch)** bzw. $\mathcal{V} = \mathcal{H}^{-1}\mathcal{J}\mathcal{H}^{-1}$ **(robust)**.

Was bringt das konkret?

- **Standardfehler:** $\mathbb{V}\mathrm{ar}(\hat{\theta}) \approx \mathcal{V}/n \;\Rightarrow\; \mathrm{SE}_j = \sqrt{(\widehat{\mathcal{V}}/n)_{jj}}$.
- **Konfidenzintervalle:** $\hat{\theta}_j \pm z_{1-\alpha/2} \cdot \mathrm{SE}_j$.
- **Tests:** Wald-/Score-/LR-Tests nutzen die Normal- bzw. χ^2-Grenzverteilungen.

θ_0 ist unbekannt: per **Plug-in** ersetzen wir $\mathcal{I}(\theta_0)$ durch $\widehat{\mathcal{I}}(\hat{\theta})$ (oder robust: $\widehat{\mathcal{H}}^{-1}\widehat{\mathcal{J}}\,\widehat{\mathcal{H}}^{-1}$).
Begründung: $\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0$ (Konsistenz) $\Rightarrow$ Plug-in ist asymptotisch korrekt.

Figures from pxhere.com under CC0

R-Übung: Logistische Regression verstehen

Modell

- Dichotome Antwortvariable $Y_i \in \{0, 1\}$, ein Prädiktor X_i.
- **Logit-Modell:**

$$p_\beta(X_i) \;:=\; \mathbb{P}(Y_i = 1 \mid X_i) \;=\; \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}.$$

- **Log-Likelihood-Beiträge:**

$$\log \mathcal{L}_i(\beta) \;=\; Y_i \log\big[p_\beta(X_i)\big] + (1 - Y_i) \log\big[1 - p_\beta(X_i)\big].$$

Intuition

- $p_\beta(X)$ bildet lineare Prädiktoren $\beta_0 + \beta_1 X$ auf Wahrscheinlichkeiten in $(0, 1)$ ab.
- Die Steigung ist in der Mitte der S-Kurve am größten (nichtlinearer Effekt).

R-Übung: Eigenen Maximum-Likelihood-Schätzer bauen

Schritte zur Log-Likelihood

1. Log-Likelihood pro Beobachtung: $\log \mathcal{L}_i(\beta)$ (Vektor).
2. Gesamte Log-Likelihood:

$$\log \mathcal{L}(\beta) = \sum_{i=1}^{n} \log \mathcal{L}_i(\beta).$$

3. Maximierung $\Rightarrow \quad \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$.

Inference aus der Asymptotik

4. Hesse-Matrix von $\log \mathcal{L}(\beta)$ im Optimum.
5. Standardfehler aus $(-\mathcal{H}(\hat{\beta}))^{-1}$.
6. Empirische Varianz der Scores $\mathcal{S}_i(\hat{\beta})$.
7. Robuste SEs via Sandwich:

$$\widehat{\mathbb{V}\mathrm{ar}}(\hat{\beta}) \;=\; \mathcal{H}(\hat{\beta})^{-1}\, \widehat{\mathbb{V}\mathrm{ar}}\big(\mathcal{S}(\hat{\beta})\big)\, \mathcal{H}(\hat{\beta})^{-1}.$$

8. Vergleiche mit `glm` und `sandwich::vcovHC`.

$\rightarrow$ Idee: θ_0 unbekannt, daher nutzen wir $\hat{\theta}$, das wegen Konsistenz „nah dran“ ist.

Zeitplan

Zeit	Titel
9:00–9:15	Begrüßung
9:15–10:30	Grundlagen: Wie kommt der Zufall in die Mathematik?
10:30–10:45	Kaffeepause
10:45–12:15	Der Mittelwert, Asymptotiken und zentrale mathematische Sätze
12:15–13:15	Mittagspause
13:15–14:45	Schätzer am Beispiel von Maximum Likelihood
14:45–15:00	Kaffeepause
15:00–16:45	Von robusten Maximum Likelihood zu Machine Learning
16:45–17:00	Abschlussrunde

Figures created with ChatGPT 5

Zeitplan

Zeit	Titel
9:00–9:15	Begrüßung
9:15–10:30	Grundlagen: Wie kommt der Zufall in die Mathematik?
10:30–10:45	Kaffeepause
10:45–12:15	Der Mittelwert, Asymptotiken und zentrale mathematische Sätze
12:15–13:15	Mittagspause
13:15–14:45	Schätzer am Beispiel von Maximum Likelihood
14:45–15:00	Kaffeepause
15:00–16:45	Von robusten Maximum Likelihood zu Machine Learning
16:45–17:00	Abschlussrunde

Figures from pxhere.com under CC0

R-Übung

- Erweitert die logistische Regression auf beliebig viele Prädiktoren.
- `4_Aufgaben_ML_vektorisiert.R`

Teaser: Machine Learning für bessere Vorhersagen

$$\text{MSE} = \text{Bias} + \text{Variance} + \text{Error}$$

Zentrale Fragen

- Was hat das Bias-Varianz-Dilemma mit Machine Learning zu tun?
- Ist das Ziel der Psychologie Vorhersagen zu machen?
- Warum funktionieren Neuronale Netze so gut und wo sind ihre aktuellen Schwachstellen?

Bias-Varianz-Dilemma

Der wahre Wert y ist eine Funktion von $f(x)$ mit Fehler ϵ: $y = f(x) + \epsilon$ mit $\mathbb{E}[y|x] = f(x), \mathbb{E}[\varepsilon|x] = 0$.
Dann ist die mittlere quadratische Abweichung (MSE):

$$\begin{aligned}
\text{MSE} &= \mathbb{E}\left[(y - \hat{f}(x))^2\right] \\
&= \text{Bias}(\hat{f}(x))^2 + \mathbb{V}\text{ar}[\hat{f}(x)] + \mathbb{E}[\epsilon^2] \qquad \text{mit} \\
\textit{Bias}(\hat{f}(x)) &= f(x) - \mathbb{E}[\hat{f}(x)] \\
\mathbb{V}\text{ar}[\hat{f}(x)] &= \left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2
\end{aligned}$$

Implikationen

Der MSE lässt sich aufteilen in Bias, Varianz und irreduziblen Fehler (siehe folgende Herleitung).

Bias-Varianz-Dilemma

$$
\begin{aligned}
\mathrm{MSE} =& \mathbb{E}\left[(y - \hat{f}(x))^2\right] \\
=& \mathbb{E}\left[\left(f(x) + \epsilon - \hat{f}(x)\right)^2\right] \qquad y = f(x) + \epsilon \\
=& \mathbb{E}\left[\left(f(x) - \hat{f}(x)\right)^2\right] + 2\mathbb{E}\left[f(x) - \hat{f}(x)\right]\mathbb{E}[\epsilon] + \mathbb{E}[\epsilon^2] \qquad \mathbb{E}[\epsilon] = 0 \\
=& \mathbb{E}\left[\left(f(x) - \mathbb{E}[\hat{f}(x)] + \mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right] + \mathbb{E}[\epsilon^2] \\
=& \mathbb{E}\left[\left(f(x) - \mathbb{E}[\hat{f}(x)]\right)^2\right] + 2\mathbb{E}\left[\left(f(x) - \mathbb{E}[\hat{f}(x)]\right)\left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)\right] + \\
& \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right] + \mathbb{E}[\epsilon^2] \\
=& \mathrm{Bias}\left(\hat{f}(x)\right)^2 + 0 + \mathbb{V}\mathrm{ar}\left[\hat{f}(x)\right] + \mathbb{E}[\epsilon^2]
\end{aligned}
$$

Bias-Varianz-Dilemma

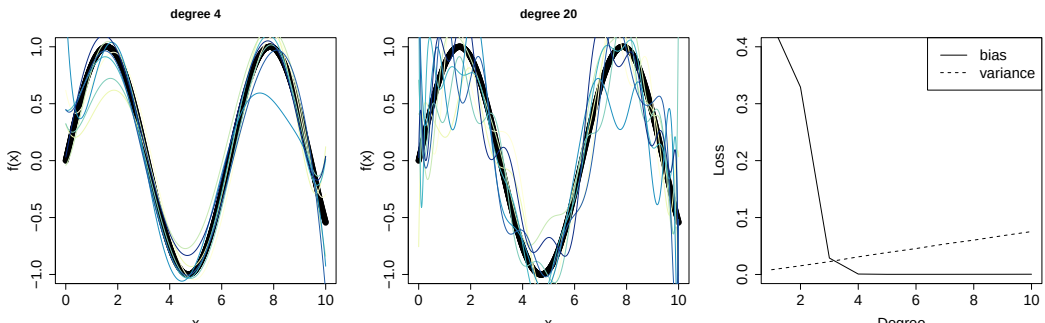
$$
\begin{aligned}
\text{MSE} =& \mathbb{E}\left[(y - \hat{f}(x))^2\right] \\
=& \left(f(x) - \mathbb{E}[\hat{f}(x)]\right)^2 + \left(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)\right)^2 + \mathbb{E}[\epsilon^2] \\
=& \text{Bias}\left(\hat{f}(x)\right)^2 + \mathbb{V}\text{ar}[\hat{f}(x)] + \mathbb{E}\left[\epsilon^2\right]
\end{aligned}
$$

Implikationen

Der MSE lässt sich aufteilen in Bias, Varianz und irreduziblen Fehler.
Klassische Statistik fokussiert auf geringen Bias, Machine Learning auf geringe Varianz (Vorhersagefehler auf neuen Daten).

Bias-Varianz-Dilemma - Illustration

100 Datensätze generiert aus $y = sin(x) + N(0, 2)$ mit $x \in \{0, 0.01, \ldots, 10\}$.
Polynomfunktionen x bis x^{20} gefittet.



To explain or to predict? (Shmueli, 2010)

- Yarkoni und Westfall (2017): Psychologie sollte sich (auch) darauf konzentrieren, zukünftiges Verhalten vorherzusagen. Machine Learning kann dabei hilfreich sein.
- Ist das eine Frage der Anwendung- versus Grundlagenforschung?
- Philosophie/Erkenntnistheorie: Beantwortet eine Theorie die Frage, wie etwas *ist* oder wie etwas *funktioniert/sich verändern lässt*?
- The worst of both worlds (Hullman et al., 2022) – Ist ML Engineering für die beste Vorhersage und damit ähnlich zu p-Hacking?

Diskussion: Wie seht ihr das?

Regularisierung

Standard-Regression $\hat{\beta} = \arg\min_{\beta} -\frac{1}{n} \sum_i \log \mathcal{L}_i(\beta)$

Lasso Regularisierung $\hat{\beta} = \arg\min_{\beta} -\frac{1}{n} \sum_i \log \mathcal{L}_i(\beta) + \lambda \sum |\beta|$

Ridge Regularisierung $\hat{\beta} = \arg\min_{\beta} -\frac{1}{n} \sum_i \log \mathcal{L}_i(\beta) + \lambda \sum \beta^2$

Elastic Net $\hat{\beta} = \arg\min_{\beta} -\frac{1}{n} \sum_i \log \mathcal{L}_i(\beta) + \lambda \left(\alpha \sum |\beta| + \frac{1-\alpha}{2} \sum \beta^2 \right)$

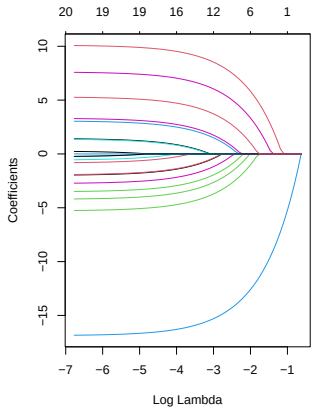
Übersetzung/Beispiel

Wir lassen Bias in den Koeffizienten β zu, um die Generalisierbarkeit zu verbessern (= Vorhersagefehler verbessern, Varianz verringen).

Parametrisierung orientiert an `glmnet`.

Beispiel: Regularisierung

Lasso



Ridge

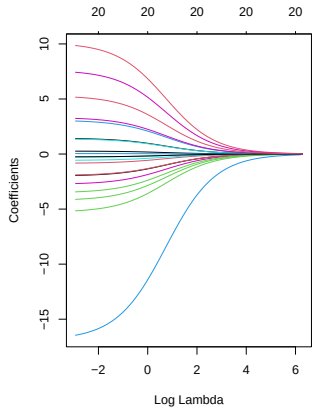


Abbildung: Polynom 20. Grades mit Regularisierung auf sin(x)

Figures from pxhere.com under CC0

R-Aufgabe

Multiple logistische Regression mit Regularisierung:

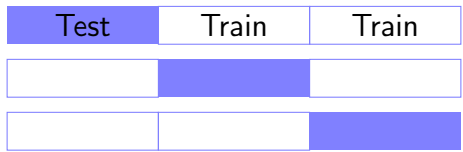
- Datengenerierendes Modell: Logistische Regression mit Regressionskoeffizienten $\beta = (-0.5, 1.2, 0, 0, 0)$.
  (Intercept = 0.5, 4 Prädiktoren, nur der Erste hat einen Effekt)
- Baut eine Ridge-Penalty in die negative Log-Likelihood ein.
- Vergleicht die Ergebnisse mit dem Output von `glmnet::glmnet(..., standardize = FALSE)`.
- Wie verhalten sich die geschätzten Regressionskoeffizienten $\hat{\beta}$ für verschiedene Werte des Regularisierungsparameters λ?

Implementation in `glmnet`

- Lasso Regularisierung ist ein beschränktes Optimierungsproblem!
- λ-Sequenz finden
- Regressionskoeffizient von vorherigem λ als Startwert
- zunächst Koeffizienten schätzen, die bei vorherigem λ nicht null waren
- Trick: Karush-Kuhn-Tucker Bedingungen beschreiben diese Situation
- Bedingungen prüfen, ggf. Menge der nicht-null Koeffizienten erweitern

Out-of-Sample Error und Kreuzvalidierung

- Wie können wir so schätzen, dass die Varianz des Schätzers $\hat{\theta} = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$ minimiert wird?
  - In-Sample Error: Vorhersagefehler auf den Daten, auf denen das Modell geschätzt ("trainiert") wurde
  - Out-Of-Sample Error: Vorhersagefehler auf neuen Daten
- Mit nur einem Datensatz: Aufteilung in Test- und Trainingsdaten
  - mehrfache, systematische Aufteilung $\rightarrow$ Kreuzvalidierung



Ein Machine Learning Rezept

- Wie finde ich die richtigen Hyperparameter?
  - Teile den Datensatz in k Teile
  - Definiere/Wähle mögliche Hyperparameter(kombinationen) = Grid
  - Für jeden Teil, für jeden Wert der Hyperparameter: Fitte das Modell auf $\mathrm{D} \setminus k$, Bestimme Fehler auf k
  - Wähle Hyperparameter mit dem kleinsten Out-of-Sample Error.
  - Alternative: Optimierungsalgorithmus für Hyperparameter statt Grid
- Wende dieses Modell auf neue Daten an.

Figures created with ChatGPT 5

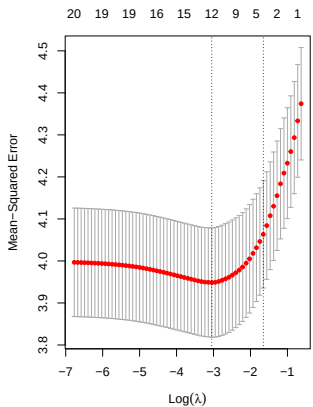
Ein Machine Learning Rezept

- Hyperparameter finden (Tuning).
- Modell damit schätzen (trainieren).
- Wende dieses Modell auf neue Daten an.
- Und wie interpretiere ich die geschätzten Parameter?
  - Gar nicht (Koeffizienten haben Bias, damit sind auch die SEs falsch).
  - Alternativen: Post-selection inference oder gleich bayesianisch schätzen (siehe Quasi-Äquivalenz von Penalties und Priors).

Figures created with ChatGPT 5

Beispiel: Kreuzvalidierung von λ

Lasso



Ridge

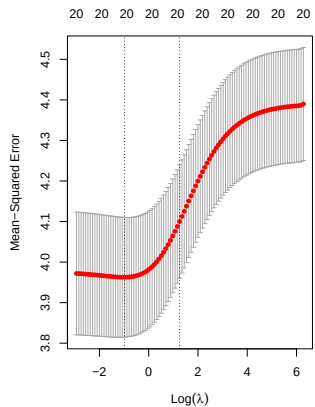


Abbildung: Polynom 20. Grades mit Regularisierung auf sin(x)

Vorhersagefehler bei Hyperparameter-Tuning - Nested Crossvalidation

- Der Vorhersagefehler aus einer einfachen Kreuzvalidierung ist überoptimistisch (Bates et al., 2024, Stichwort: Data/Information leakage)
- Nested Crossvalidation: Kreuzvalidiere die Hyperparameter innerhalb jedes Splits.
- Dann wird der Vorhersagefehler jeweils auf ungesehenen Daten bestimmt. Aber wir haben für jeden äußeren Split ein anderes Modell.

Figures from pxhere.com under CC0

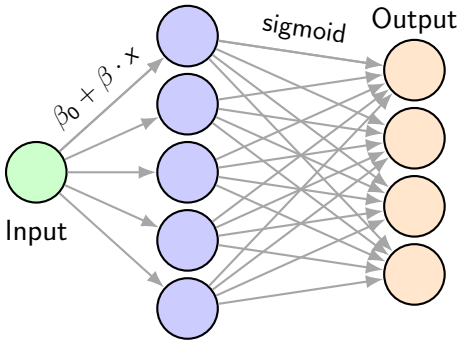
R-Aufgabe

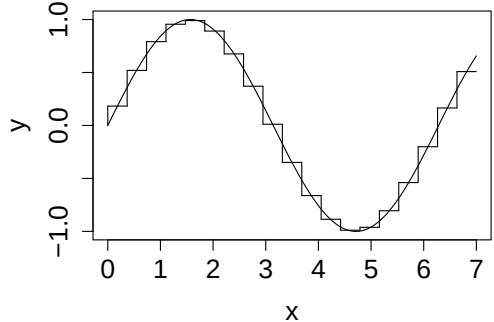
Erweitert die Regression um eine Kreuzvalidierung des Lambda-Parameters:

- Teilt den Datensatz zufällig in 10 Teile (`sample`). Je 1 Teil dient als Testdaten, 9 als Trainingsdaten.
- Testet 10 verschiedene Werte für λ.
- Plotted den Out-of-Sample MSE und vergleicht mit den Ergebnissen von `glmnet::cv.glmnet`

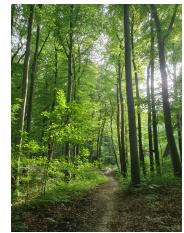
Neuronale Netze und Universal Approximation Theorem

- ein Feed-Forward Neuronales Netz kann jede funktionale Form approximieren (mit nur einem Layer, beliebig vielen Einheiten; Cybenko, 1989)
- Intuition: Verkettung von Treppenfunktionen
- Warum die heutigen neuronalen Netzwerke so gut funktionieren, ist mathematisch erst ansatzweise erklärt (Zhang et al., 2023).





Weitere Entwicklungen



Uncertainty Quantification

- Ensemble-Methoden: Varianz über die einzelnen Modelle
- Bayesianisch: Variational Inference
- ...

Trustworthy AI (Thiebes et al., 2021)

- Bias in den Trainingsdaten, unterrepräsentierte Gruppen berücksichtigen
- erklärbar, interpretierbar, transparent
- ...

Zeitplan

Zeit	Titel
9:00–9:15	Begrüßung
9:15–10:30	Grundlagen: Wie kommt der Zufall in die Mathematik?
10:30–10:45	Kaffeepause
10:45–12:15	Der Mittelwert, Asymptotiken und zentrale mathematische Sätze
12:15–13:15	Mittagspause
13:15–14:45	Schätzer am Beispiel von Maximum Likelihood
14:45–15:00	Kaffeepause
15:00–16:45	Von robusten Maximum Likelihood zu Machine Learning
16:45–17:00	Abschlussrunde

Abschlussrunde

Figures created with ChatGPT 5

Zeitplan

Zeit	Titel
9:00–9:15	Begrüßung
9:15–10:30	Grundlagen: Wie kommt der Zufall in die Mathematik?
10:30–10:45	Kaffeepause
10:45–12:15	Der Mittelwert, Asymptotiken und zentrale mathematische Sätze
12:15–13:15	Mittagspause
13:15–14:45	Schätzer am Beispiel von Maximum Likelihood
14:45–15:00	Kaffeepause
15:00–16:45	Von robusten Maximum Likelihood zu Machine Learning
16:45–17:00	Abschlussrunde

Zum Nachlesen

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2024). *Statistik: Der Weg zur Datenanalyse* (9. Aufl. 2023). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-67526-7

Klenke, A. (2020). *Wahrscheinlichkeitstheorie*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-62089-2

Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden: Lehrbuch. Mit Online-Material* (Originalausgabe, 5., korrigierte Aufl). Beltz

Zum Vertiefen I

- Bates, S., Hastie, T., & Tibshirani, R. (2024).Cross-Validation: What Does It Estimate and How Well Does It Do It? *Journal of the American Statistical Association*, *119*(546), 1434–1445. https://doi.org/10.1080/01621459.2023.2197686
- Cybenko, G. (1989).Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274
- Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und Forschungsmethoden: Lehrbuch. Mit Online-Material* (Originalausgabe, 5., korrigierte Aufl). Beltz.
- Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2024). *Statistik: Der Weg zur Datenanalyse* (9. Aufl. 2023). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-67526-7

Zum Vertiefen II

- Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022).The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 335–348.
- Klenke, A. (2020). *Wahrscheinlichkeitstheorie*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-62089-2
- Shmueli, G. (2010).To Explain or to Predict? *Statistical Science*, *25*(3). https://doi.org/10.1214/10-STS330
- Thiebes, S., Lins, S., & Sunyaev, A. (2021).Trustworthy Artificial Intelligence. *Electronic Markets*, *31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4
- Yarkoni, T., & Westfall, J. (2017).Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Zum Vertiefen III


Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into Deep Learning* [https://D2L.ai]. Cambridge University Press.