Start:Bayes!

Schätzen vs. Testen

Sebastian Sauer



Lernziele

Nach diesem Kapitel können Sie …

- den Unterschied zwischen dem *Schätzen* von Modellparametern und dem *Testen* von Hypothesen erläutern
- Vor- und Nachteile des Schätzens und Testens diskutieren
- das ROPE-Konzept erläutern und anwenden
- die Güte von Regressionsmodellen einschätzen und berechnen

Begleitliteratur: Kruschke (2018)

Einstieg

Einstieg: Zwei Arten von Forschungsfragen

1. “Lernen für die Klausur bringt etwas!” — eine **Behauptung**
2. “Wie viel bringt Lernen für die Klausur?” — eine **Frage**

Einmal Behauptung, einmal Frage — was macht das für einen Unterschied?

Was ist Testen? Was ist Schätzen?

Zwei Arten, Forschungsfragen zu beantworten

1. **Hypothesen testen**: “Die Daten widerlegen die Hypothese (nicht)”
2. **Parameter schätzen**: “Der Effekt von X auf Y liegt zwischen A und B”

Hypothesen testen: drei mögliche Ergebnisse

Beispiel: “Lernen erhöht den Prüfungserfolg!”

1. 🟥 Die Daten *widersprechen* der Hypothese — Ablehnung (Falsifikation)
2. 🟢 Die Daten *unterstützen* die Hypothese — Annahme
3. ❓ Die Datenlage ist *unklar* — kein Erkenntnisgewinn

Hypothesentests sind **binär**: Sie führen zu Schwarz-Weiß-Ergebnissen.

Die Nullhypothese

Wichtig

Häufigste Variante: Test der Hypothese “Es liegt kein (null) Effekt vor” — die **Nullhypothese**, H_0.

Beispiele für Nullhypothesen:

- “Lernen bringt nichts”
- “Frauen und Männer parken gleich schnell ein”
- “Es gibt keinen Zusammenhang von Babies und Störchen”
- “Bei Frauen ist der Anteil, die Statistik mögen, gleich hoch wie bei Männern”

Vorteil des Hypothesentestens: klares, einfaches Ergebnis — reduziert Komplexität für Entscheidungen.



Popper: Man kann Hypothesen nicht bestätigen

Karl Poppers Falsifikationismus

Hypothesen lassen sich nicht bestätigen (verifizieren), nur falsifizieren (Popper, 2013).

Beispiel H: “Alle Schwäne sind weiß.” Auch viele weiße Schwäne beweisen H nicht — der eine schwarze Schwan könnte noch fehlen. Aber ein einziger zuverlässig beobachteter schwarzer Schwan widerlegt H.

In der Praxis: Hypothesen werden meist **probabilistisch** untersucht — komplett sichere Belege wie bei Popper sind selten. Sowohl Bestätigung als auch Widerlegung sind daher (graduell) möglich (Kruschke, 2018; Morey & Rouder, 2011).

Parameter schätzen

Beim Schätzen fragt man **wie groß** ein Effekt ist — ein “wieviel”, kein “ja/nein”.

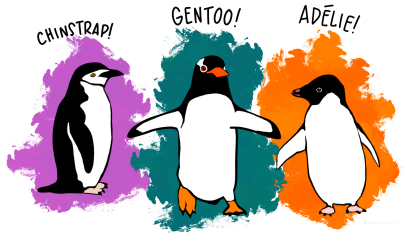
Zwei Varianten:

a. ⚫ **Punktschätzung**: ein einzelner Parameterwert (“Best Guess”)

b. 📏 **Bereichsschätzung**: ein Bereich plausibler Parameterwerte

Liegt ein bestimmter Wert (z. B. Null) *nicht* im Schätzbereich, kann man die entsprechende Hypothese verwerfen. Das Hypothesentesten steckt also implizit im Schätzen.

Beispiel: Geschlecht und Pinguingewicht



Forschungsfrage: Um welchen Wert sind männliche Pinguine im Schnitt schwerer als weibliche?

- Modell: `body_mass_g ~ sex`
- Ergebnis: Punktschätzer ca. 700 g Unterschied
- 95 %-ETI: ca. 500 g bis 800 g

Parameterschätzung als Nullhypothesentest

Forschungsfrage: Sind männliche Pinguine im Schnitt schwerer als weibliche?

$$H : \mu_M \geq \mu_F \rightarrow d = \mu_M - \mu_F \geq 0$$

- $d = 0 \Leftrightarrow \beta_1 = 0$ im Regressionsmodell
- Anteil der Post-Stichproben mit `sexmale > 0`: 100 % (4000 von 4000)
- Also: Modell räumt der Hypothese $p_H = 1$ ein (in der *kleinen Welt* des Modells)

Vorteil der Parameterschätzung: nuancierteres Ergebnis, das der Komplexität realer Systeme besser gerecht wird.



ROPE: Bereich von “praktisch Null”



Nullhypothesen sind fast immer falsch



> We do not generally use null hypothesis significance testing in our own work. […] just about every treatment one might consider will have *some* effect, and no comparison or regression coefficient of interest will be exactly zero. (Gelman et al., 2021)

ROPE: Bereich von “praktisch Null”

Alternativen zu Nullhypothesentests

Nullhypothesen wie $\rho = 0, \mu_1 = \mu_2, \beta_1 = 0$ sind streng genommen fast nie exakt wahr. Alternativen:

- Posteriori-Intervalle (ETI oder HDI) berichten
- **ROPE**-Konzept (Kruschke, 2018)
- Wahrscheinlichkeit inhaltlich bedeutsamer Hypothesen quantifizieren (z. B. $\beta_1 > 0{,}42$)
- *Probability of direction* (pd): Wahrscheinlichkeit für ein positives/negatives Vorzeichen

ROPE: Bereich von “praktisch Null”



Definition: ROPE

Definition: ROPE

ROPE (Region of Practical Equivalence) ist ein Bereich um den Nullwert, der als **praktisch äquivalent zu Null** angesehen wird.

Beispiele:

- Preisunterschied bis 100 € gilt als “praktisch gleich”
- Gewichtsunterschied bis 100 g bei Pinguinen als “vernachlässigbar”
- Umsatzunterschied bis 100k € als “praktisch irrelevant”

ROPE: Bereich von “praktisch Null”

ROPE-Entscheidungsregel

Nach Kruschke (2018):

- Großteil der Post-Verteilung liegt **innerhalb** des ROPE $\rightarrow$ *Annahme* der “praktisch kein Effekt”-Hypothese
- Großteil liegt **außerhalb** des ROPE $\rightarrow$ *Ablehnung* dieser Hypothese
- Sonst $\rightarrow$ **keine Entscheidung** — Datenlage uneindeutig

“Großteil” = per Default das 95 %-HDI der Posteriori-Verteilung.

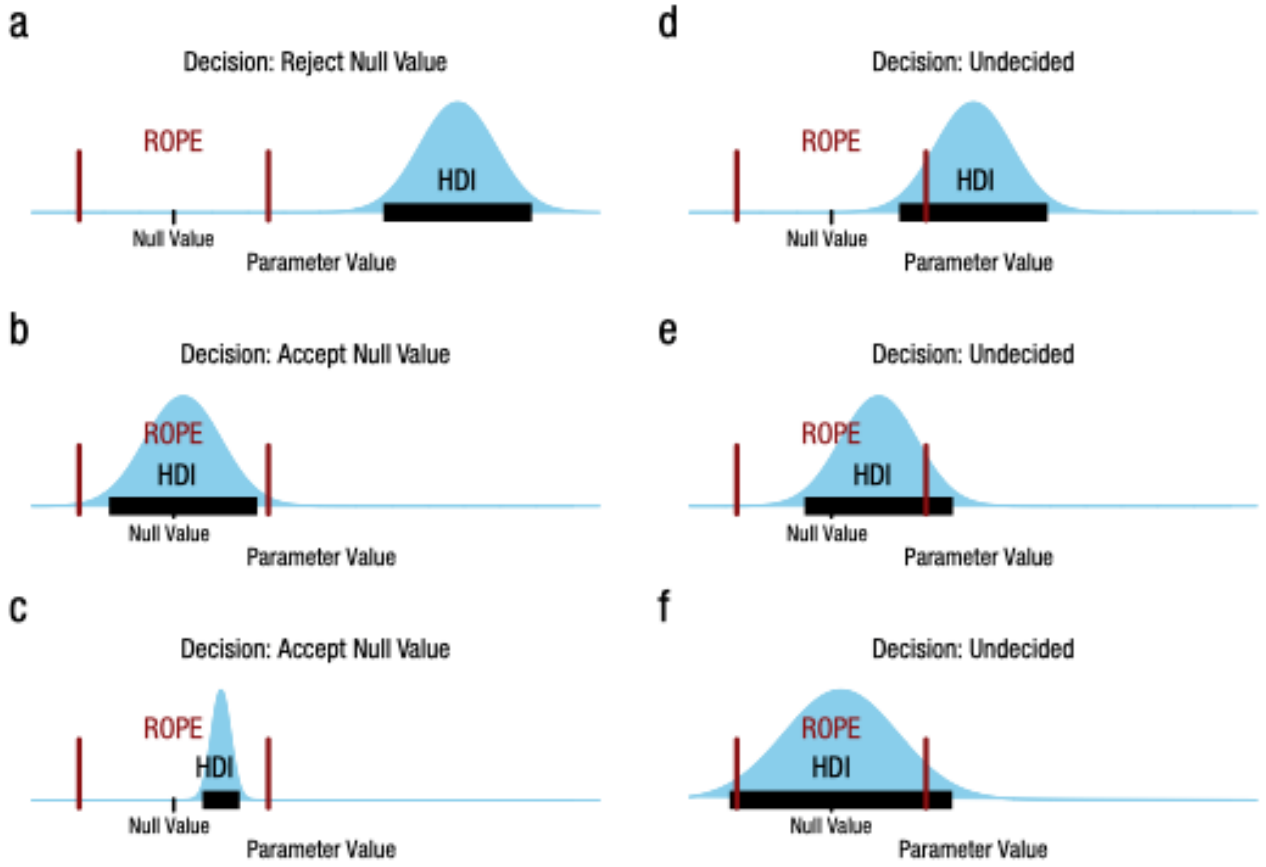
ROPE-Grenzen festlegen

- Die Grenze d ist **subjektiv** — geleitet von inhaltlichen Überlegungen
- Für Regressionskoeffizienten schlägt Kruschke (2018) vor: vernachlässigbarer Effekt = $\hat{y} = \pm 0{,}1 \; sd_y$

ROPE-Default

Voreinstellung: ROPE-Größe = ±10 % der Standardabweichung der AV.

ROPE visualisiert



ROPE: Bereich von “praktisch Null”

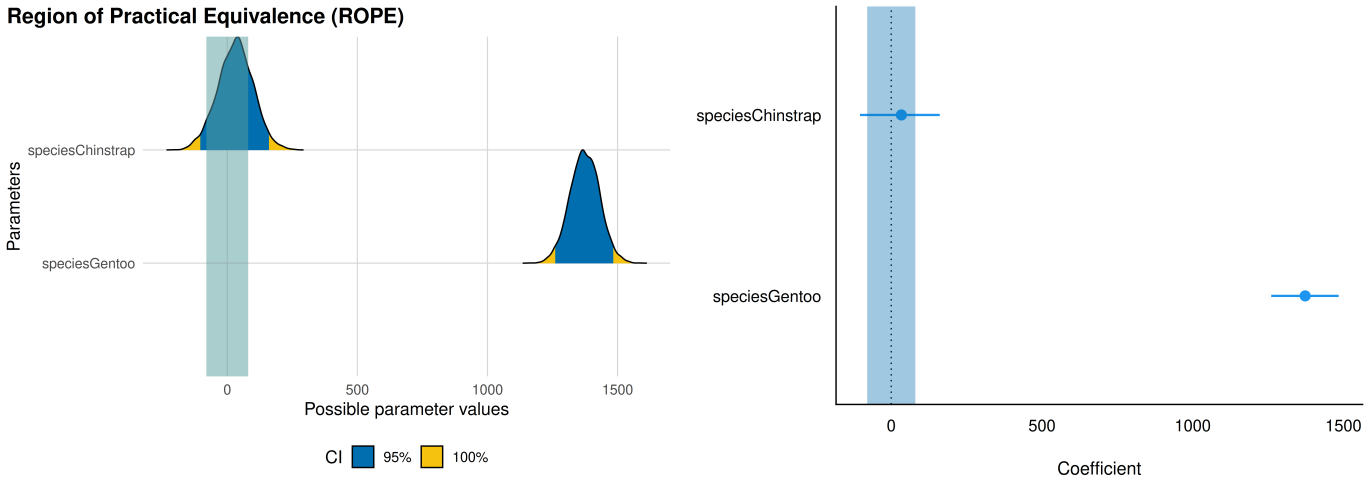
Rope berechnen: Pinguinarten

Modell: `body_mass_g ~ species` (Referenz: Adelie)

- **Chinstrap**: beträchtlicher Anteil der Post-Verteilung liegt im ROPE $\rightarrow$ ROPE-Hypothese kann *nicht verworfen* werden — **unklare Datenlage**
- **Gentoo**: 0 % im ROPE $\rightarrow$ ROPE-Hypothese wird *verworfen* — substanziell schwerer als Adelie

ROPE-Grenzwert hier: ±80 g (10 % der SD der AV)

Rope-Werte visualisiert

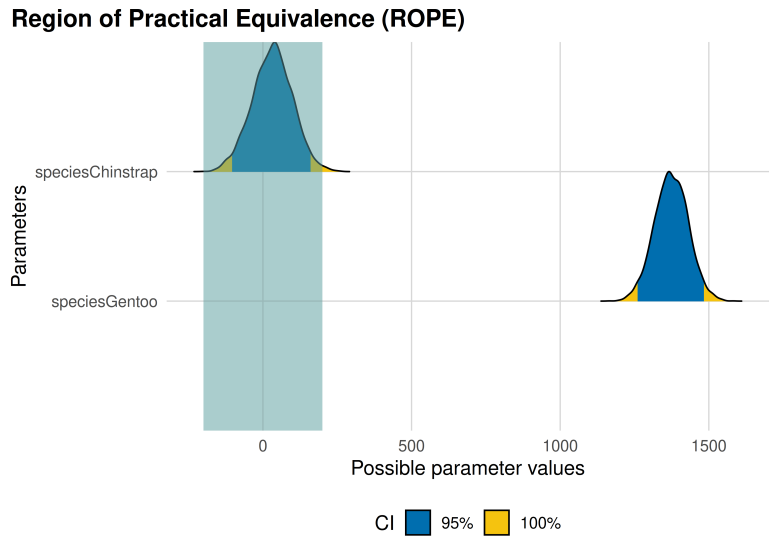


- Chinstrap: ROPE durchkreuzt die Verteilung deutlich — HDI liegt nicht komplett im ROPE $\longrightarrow$ keine klare Entscheidung
- Gentoo: liegt weit außerhalb des ROPE-Bands $\longrightarrow$ substanzieller Unterschied zu Adelie

ROPE: Bereich von “praktisch Null”

Finetuning des ROPE

Die ROPE-Grenzen lassen sich anpassen, z. B. auf ±200 g:



Auch das Kredibilitätsintervall ist wählbar (z. B. 89 % statt 95 %).

Modellgüte

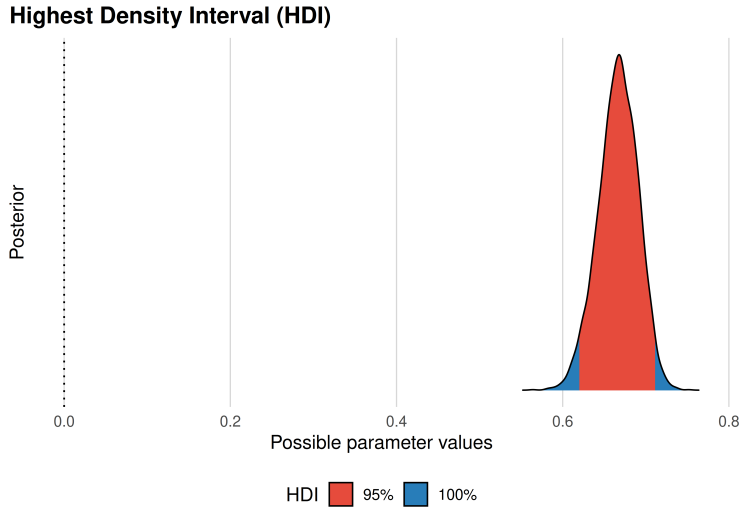


Wozu Modellgüte?

- Man möchte wissen, wie **belastbar** die Ergebnisse eines Modells sind
- Kennwerte der Modellgüte schätzen die **Präzision** der Modellaussagen
- Ein Modell kann aus mehreren unterschiedlich präzisen Parameterschätzungen bestehen $\longrightarrow$ zusammenfassender Kennwert hilfreich
- Gebräuchliche Kennzahl: R^2

Modellgüte mit R^2

- R^2 gibt den Anteil der Gesamtvarianz der AV an, den das Modell erklärt
- Höhere Werte $\rightarrow$ das Modell erklärt die Daten besser
- Klassisches R^2: nur ein Punktschätzer — Ungewissheit bleibt unberücksichtigt
- **Bayes-R^2**: liefert eine ganze Post-Verteilung von R^2



R^2 und σ

- R^2 und σ sind **negativ assoziiert**: hohes $R^2 \leftrightarrow$ geringes σ
- Beide Kennwerte dienen demselben Zweck: Abschätzung der Modellgüte
- Bayesianisch: nicht nur ein Punktschätzer, sondern ein Schätzbereich für R^2



Fazit



Schätzen oder Testen: Ein Résumé

- Hypothesentests sind aktuell verbreiteter, aber die **Parameterschätzung** hat wichtige Vorzüge (Nuanciertheit, Präzisionsangabe)
- Wer dennoch (aus Anschlussfähigkeit an bestehende Forschung) einen Nullhypothesentest möchte: das **ROPE-Verfahren** bietet eine sinnvolle Brücke zwischen beiden Ansätzen



Vielen Dank!

Fragen?

Gelman, A., Hill, J., & Vehtari, A. (2021). *Regression and Other Stories*. Cambridge University Press.

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Morey, R. D., & Rouder, J. N. (2011). Bayes Factor Approaches for Testing Interval Null Hypotheses. *Psychological Methods*, *16*(4), 406–419. https://doi.org/10.1037/a0024377

Popper, K. (2013). *Logik Der Forschung* (H. Keuth, Hrsg.). Akademie Verlag. https://doi.org/10.1524/9783050063782

