9 Schätzen vs. Testen

9.1 Lernsteuerung

9.1.1 Position im Modulverlauf

Abbildung 1.1 gibt einen Überblick zum aktuellen Standort im Modulverlauf.

9.1.2 Lernziele

Nach Absolvieren des jeweiligen Kapitels sollen folgende Lernziele erreicht sein.

Sie können…

- 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

9.1.3 Begleitliteratur

Der Stoff dieses Kapitels orientiert sich an (Kruschke 2018).

9.1.4 R-Pakete

In diesem Kapitel werden folgende R-Pakete benötigt:

```
library(rstanarm)   # Bayes-Modelle
library(tidyverse)
library(easystats)
```

9.1.5 Benötigte Daten

Wir benötigen in diesem Kapitel folgenden Datensatz: `penguins`.

Sie können den Datensatz `penguins` entweder via dem Pfad importieren:

```
penguins_url <- "https://vincentarelbundock.github.io/Rdatasets/csv/palmerpeng

penguins <- read.csv(penguins_url)
```

DOWNLOAD

Oder via dem zugehörigen R-Paket:

```
data("penguins", package = "palmerpenguins")
```

Beide Möglichkeit sind okay.

9.1.6 Einstieg

Betrachten Sie die zwei folgenden Aussagen, die jeweils ein Forschungsziel angeben:

1. "Lernen für die Klausur bringt etwas!"
2. "Wie viel bringt Lernen für die Klausur?"

Beispiel 9.1 Diskutieren Sie die epistemologische Ausrichtung sowie mögliches Für und Wider der beiden Ausrichtungen! ☐

9.2 Schätzen oder Testen?

Forschungsfragen kann man, allgemein gesprochen, auf zwei Arten beantworten:

1. *Hypothesen prüfend*: "Die Daten widerlegen die Hypothese (nicht)"
2. *Parameter schätzend*: "Der Effekt von X auf Y liegt zwischen A und B".

9.2.1 Hypothesen prüfen

Hypothesen prüfende Analysen kommen zu einer Ja-Nein-Aussage bzgl. einer Hypothese. Genauer muss man sagen: Im besten Fall kommen sie zu einer Ja-Nein-Aussage. Es kann natürlich sein, dass die Datenlage so nebelig oder das Problem so knifflig ist, dass man ehrlicherweise zugeben muss, dass man sich nicht sicher ist oder sagar komplett im Dunkeln tappt.

Beispiel 9.2 ("Lernen erhöht den Prüfungserfolg") Die Hypothese *Lernen erhöht den Prüfungserfolg* kann durch eine Studie und eine entsprechende Analyse grundsätzlich folgende drei Ergebnisse finden. 1) Die Daten widersprechen der Hypothese: Lernen bringt offenbar doch nichts für den Klausurerfolg. 2) Die Daten unterstützen die Hypothese: Lernen erhöht den Prüfungserfolg. 3) Die Daten sind uneindeutig, es ist keine Aussage zum Einfluss von Lernen auf den Prüfungserfolg möglich. ☐

Das Prüfen einer Hypothese kann zu drei Arten von Ergebnissen führen. Die ersten beiden sind informationsreich, die dritte ist informationsarm.

1. 🟢 Die Daten *widersprechen* der Hypothese: Auf Basis der Daten (und des Modells) muss man die Hypothese ablehnen (verwerfen, sagt man), also als falsch (falsifziert) betrachten oder zumindest hat die Glaubwürdigkeit der Hypothese gelitten.

2. 🟥 Die Daten *unterstützen* die Hypothese: Auf Basis der Daten (und des Modells) muss man die Hypothese annehmen (oder kann die Gegenthese zumindest nicht verwerfen). Oder zumindest hat die Hypothese an Glaubwürdigkeit gewonnen.

3. ❓ Die Datenlage ist *unklar*; zum Teil unterstützen die Daten die Hypothese zum Teil widersprechen sie ihr. Man kann keine oder kaum Schlüsse aus den Daten ziehen. In diesem Fall gibt es keinen Erkenntnisgewinn.

Hypothesen prüfen ist *binär* in dem Sinne, dass sie zu "Schwarz-Weiß-Ergebnissen" führen (sofern die Datenlage stark genug ist).

Wichtig

Eine gängige Variante des Hypothesen testen[1] ist das Testen der Hypothese “kein Effekt” (Null Effekt), man spricht vom *Nullhypothesen testen*. ☐

9.2.2 Beispiele für Nullhypothesen

- “Lernen bringt nichts”
- “Frauen und Männer parken gleich schnell ein”
- “Es gibt keinen Zusammenhang von Babies und Störchen”
- “Früher war es auch nicht besser (sondern gleich gut)”
- “Bei Frauen ist der Anteil, derer, die Statistik mögen gleich hoch wie bei Männern” (Null Unterschied zwischen den Geschlechtern) ☐

Vorteil des Hypothesen testen ist das klare, einfache Ergebnisse, was die Entscheidungsfindung unterstützen kann, da es die Komplexität reduziert.

Man kann Hypothesen nicht bestätigen

Karl Poppers These, dass man Hypothesen nicht bestätigen (verifizieren) kann, hat großen Einfluss auf die Wissenschaftstheorie (und Epistemologie allgemein) ausgeübt (Popper 2013). Schlagend ist das Beispiel zur Hypothese “Alle Schwäne sind weiß”. Auch eine große Stichprobe an weißen Schwänen kann die Wahrheit der Hypothese nicht beweisen. Schließlich ist es möglich, dass wir den schwarzen Schwan einfach noch nicht gefunden haben.[2] Umgekehrt reicht die (zuverlässige) Beobachtung eines einzelnen schwarzen Schwans, um die Hypothese zu widerlegen (falsifizieren). ☐

Wirklich nicht?

In der Wissenschaftspraxis werden die meisten Hypothesen probabilistisch untersucht. Komplett sichere Belege, wie in Poppers Beispiel mit dem schwarzen Schwan, gibt es nicht. Das bedeutet, dass Evidenz im bestätigenden wie im widerlegenden Sinne tendenziell (probabilistisch) zu betrachten ist. Auf dieser Basis und der Basis zuverlässiger, repräsentativer Daten erscheint plausibel, dass Hypothesen sowohl bestätigt als auch widerlegt werden können (Kruschke 2018; Morey und Rouder 2011). ☐

9.2.3 Parameter schätzen

Beim Parameter schätzen untersucht man, *wie groß* ein Effekt ist, etwa der Zusammenhang zwischen X und Y. Es geht also um eine Skalierung, um ein *wieviel* und nicht um ein “ja/nein”, was beim Hypothesen testen der Fall ist.

Beim Parameter schätzen gibt es zwei Varianten:

a. ⚫ Punktschätzung: Das Schätzen eines einzelnen Parameterwerts, sozusagen ein “Best Guess”

b. 📏 Bereichsschätzung: Das Schätzen eines Bereichs plausibler oder wahrscheinlicher Parameterwerte

Allerdings kann man das Parameter schätzen auch wie einen Hypothesentest nutzen: Ist ein bestimmter Wert, etwa die Null, nicht im Schätzbereich enthalten, so kann man die Hypothese verwerfen, dass der Parameter gleich diesem Wert (etwa Null) ist.

Beispiel 9.3 (Parameterschätzen als Nullhypothesentest)

> Forschungsfrage: Sind männliche Pinguine im Schnitt schwerer als weibliche Tiere?

Gleichung 9.1 formalisiert diese Forschungsfrage als statistische Hypothese H.

$$H : \mu_M \geq \mu_F \to d = \mu_M - \mu_F \geq 0 \qquad (9.1)$$

Der Unterschied zwischen den Mittelwerten, d, ist genau dann Null, wenn β_1 in unserem Regressionsmodell `m1` gleich Null ist. Entsprechend gilt $d \geq 0$ wenn $\beta_1 \geq 0$.

```
m1 <- stan_glm(
  body_mass_g ~ sex,
  data = penguins,
  refresh = 0,  # unterdrückt Ausgabe der Posteriori-Stichproben
  seed = 42  # zur Reproduzierbarkeit
)
```

Dann zählen wir einfach den Anteil der Stichproben in der Post-Verteilung für die UV `sexmale`, die einen Wert größer Null aufweisen:

```
m1_post <-
  m1 |>
  as_tibble()

m1_post |>
  count(sexmale < 0)
```

sexmale < 0	n
<lgl>	<int>
FALSE	4000

1 row

100% (4000 von 4000) Stichproben finden einen Wert größer Null für `sexmale`, dass also weibliche Tiere leichter bzw. männliche Tiere schwerer sind. Entsprechend finden 0% der Stichproben einen Wert, der für das Gegenteil spricht (das weibliche Tiere schwerer wären). Damit resümieren wir, dass unser Modell 100% Wahrscheinlichkeit für die Hypothese einräumt: $p_H = 1$. □
Vorteil der Parameterschätzung ist die Nuanciertheit des Ergebnisses, die der Komplexität echter Systeme besser Rechnung trägt.

9.3 ROPE: Bereich von "praktisch Null"

📺 Teil 2

Nullhypothesen sind fast immer falsch, s. Abbildung 9.1.

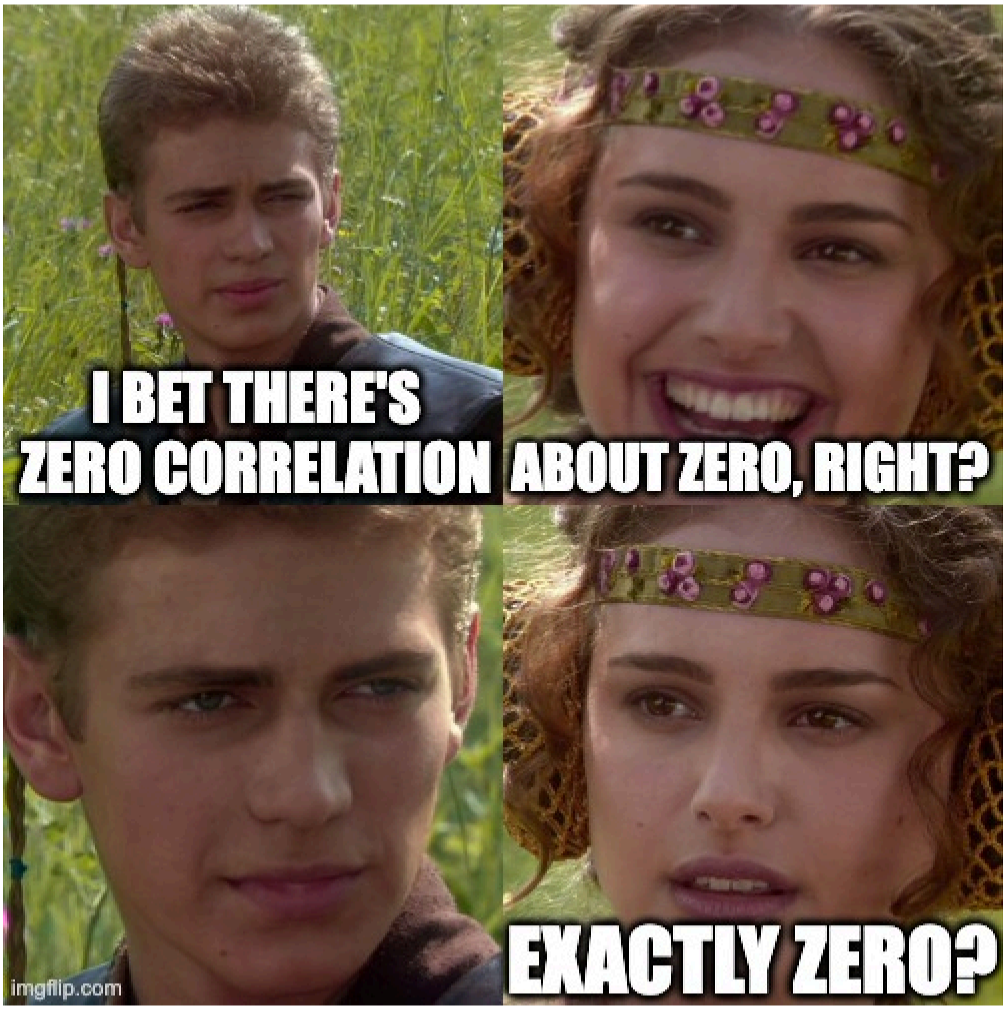


Abbildung 9.1: Du testest Nullhypothesen?

[Quelle: Imgflip Meme Generator]

> We do not generally use null hypothesis significance testing in our own work. In the fields in which we work, we do not generally think null hyptheses can be true: in social science and public health, just about every treatment one might consider will have *some* effect, and no comparison or regression coefficient of interest will be exactly zero. We do not find it particularly helpful to formulate and test null hypothess that we knowe ahead of time cannot be true. (Gelman, Hill, und Vehtari 2021)

9.3.1 Alternativen zu Nullhypothesen

Nullhypothesen, H_0, sind z. B.: $\rho = 0, \rho_1 = \rho_2, \mu_1 = \mu_2, \mu = 0, \beta_1 = 0$. Nullhypothesen zu testen, ist sehr verbreitet. Ein Grund ist, dass in der Frequentistischen Statistik keine andere Art von Hypothesentest möglich ist.[3]

Ein anderer Grund ist vermutlich, … wir haben es schon immer so gemacht. 🤷‍♀️

Alternativen zum Testen von Nullhypothesen sind:

- Posteriori-Intervalle (PI oder HDI) berichten
- *Rope*-Konzept (Kruschke 2018)
- Wahrscheinlichkeit von inhaltlich bedeutsamen Hypothesen quantifizieren.
- Wahrscheinlichkeit quantifizieren, dass der Effekt ein positives bzw. ein negatives Vorzeichen hat.

9.3.2 “Praktisch” kein Unterschied: Das Rope-Konzept

📺 ROPE-Video

Beispiel 9.4 (Beispiele für ROPE) Sagen wir, wenn sich zwei Preismittelwerte um höchstens $d = 100€$ unterscheiden, gilt dieser Unterschied für uns als "praktisch gleich", "praktisch kein Unterschied" bzw. vernachlässigbar.

Bei Pinguinarten definiert eine Biologin nach umfangreichem Studium der Literatur, dass ein Unterschied von max. 100g "vernachlässigbar wenig" ist.

Eine findige Geschäftsfrau entscheidet für ihre Firma, dass ein Umssatzunterschied von 100k Euro "praktisch irrelevant" sei. □

Nimmt man (praktisch) keinen Unterschied/Zusammenhang/Effekt an, spricht man von einer *Nullhypothese*: H_0. Die Wahl von d ist *subjektiv* in dem Sinne als sie von inhaltlichen Überlegungen geleitet sein sollte. Diesen Bereich bezeichnen wir den *Indifferenzbereich* (Äquivalenzzone, Bereich eines vernachlässigbaren Unterschieds oder *Region of practical equivalence*, Rope). Jetzt prüfen wir, ob ein "Großteil" der Posteriori-Stichproben im Rope liegt. Unter "Großteil" wird häufig das *95%-HDI* verstanden (das ist auch der Standard der R-Funktion `rope()`, die wir hier nutzen).

Entscheidungsregel nach Kruschke (2018):

- Großteil liegt *innerhalb* von Rope ➡️ *Annahme* der Nullhypothese "praktisch kein Effekt", H_0
- Großteil liegt *außerhalb* von Rope ➡️ *Ablehnung* der Nullhypothese "praktisch kein Effekt", H_0
- Ansonsten ➡️ keine Entscheidung

Mit "Großteil" meinen wir (per Default) das 95%-HDI (der Posteriori-Verteilung).

9.3.3 Vernachlässigbarer Regressionseffekt

Kruschke (2018) schlägt vor, einen Regressionskoeffizienten unter folgenden Umständen als "praktisch Null" zu bezeichnen:

Wenn eine Veränderung über "praktisch den ganzen Wertebereich" von x nur einen vernachlässigbaren Effekt auf y hat. Ein vernachlässigbarer Effekt ist dabei $\hat{y} = \pm 0.1 sd_y$. Der "praktisch ganze Wertebereich" von x sei $\bar{x} \pm 2sd_x$. Resultiert der Vergleich von $\bar{x} - 2sd$ mit $\bar{x} + 2sd$ nur eine Veränderung in $\hat{y}$ von $\bar{y} - 0.1sd_y$ auf $\bar{y} + 0.1sd_y$, so ist der Regressionskoeffizient praktisch Null, der Effekt also vernachlässigbar. Das impliziert Rope-Grenzen von $\beta_x = \pm 0.05$ für z-standardisierte Variablen.

> **ROPE-Defaults**
>
> Im der Voreinstellung umfasst die Größe des ROPE ±5% der SD der AV. □

9.3.4 HDI-Rope-Entscheidungsregel visualisiert

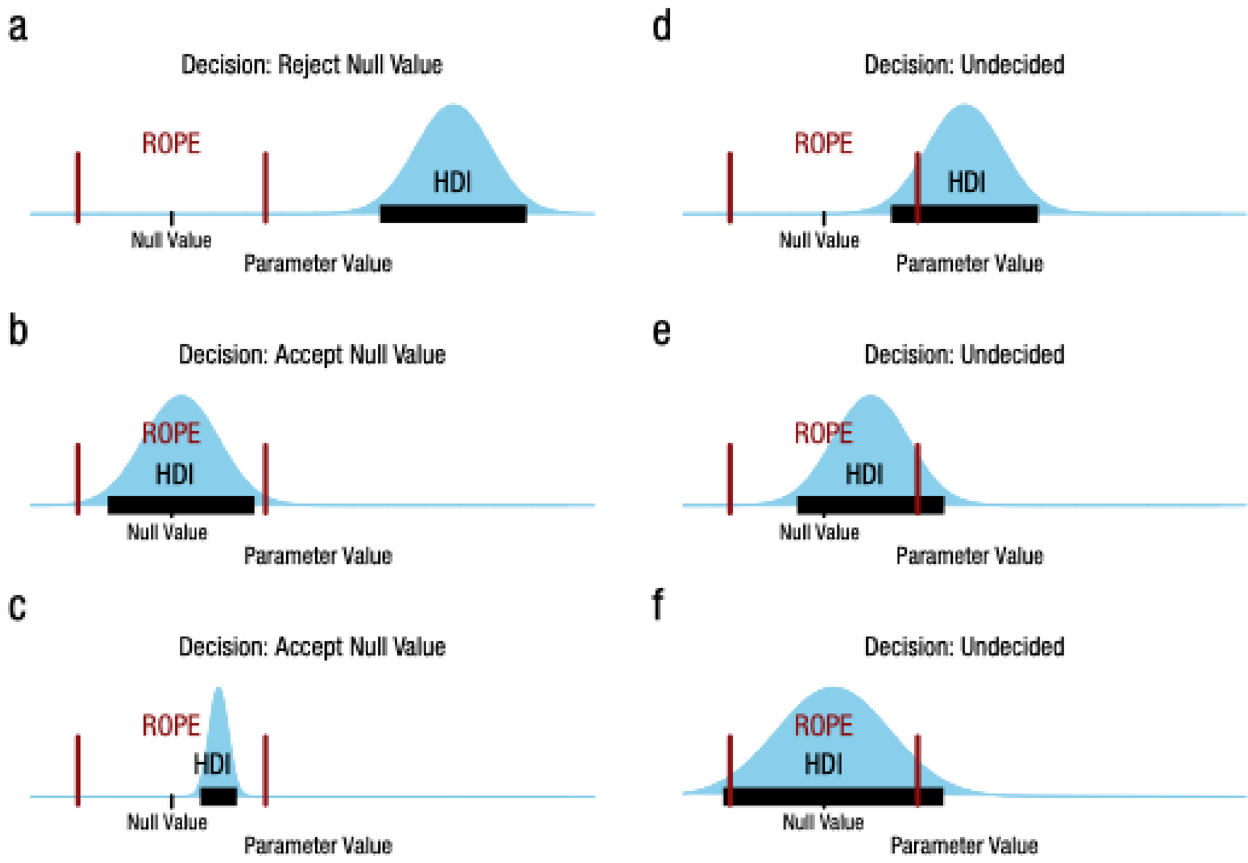


Abbildung 9.2: Die Entscheidungsregeln zum ROPE illustiert.

Abbildung 9.2 illustriert die Entscheidungsregel zum ROPE für mehrere Situatioenen (Kruschke 2018, Abb. 1, S. 272):

- Liegt das HDI komplett außerhalb des ROPE, verwirft man die Nullhypothese.
- Liegt das HDI komplett innerhalb des ROPE, akzeptiert man die Nullhypothese.
- Ansonsten ist keine Entscheidung möglich; die Datenlage ist unklar.

9.3.5 Rope berechnen

Hier ist das Modell, das Gewicht als Funktion der Pinguinart erklärt (`m10.6`).

```
m10.6 <- stan_glm(body_mass_g ~ species,
                  data = penguins,
                  refresh = 0,  # unterdrückt Ausgabe der Posteriori-Stichprob
                  seed = 42  # zur Reproduzierbarkeit
                  )
```

Den Rope berechnet man mit `rope(model)`.

```
rope(m10.6)
```

Parameter	CI	ROPE_low	ROPE_high	ROPE_Percentage	Effects
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1 (Intercept)	0.95	-80.19545	80.19545	0.0000000	fixed
2 speciesChinstrap	0.95	-80.19545	80.19545	0.7528947	fixed

Parameter	CI	ROPE_low	ROPE_high	ROPE_Percentage	Effects
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
3 speciesGentoo	0.95	-80.19545	80.19545	0.0000000	fixed

3 rows | 1-7 of 8 columns

Die Faktorstufe `Chinstrap` von `species` hat doch einen beträchtlichen Teil ihrer Wahrscheinlichkeitsmasse der Posteriori-Verteilung im ROPE. Wir können d. h.r für diese Gruppe das ROPE *nicht* verwerfen. Die Datenlage ist *unklar*. Es ist keine abschließende Entscheidung über die Hypothese möglich.

Aber: `Gentoo` liegt zu 0% im Rope. Für Gentoo können wir das Rope *verwerfen*.

Hinweis

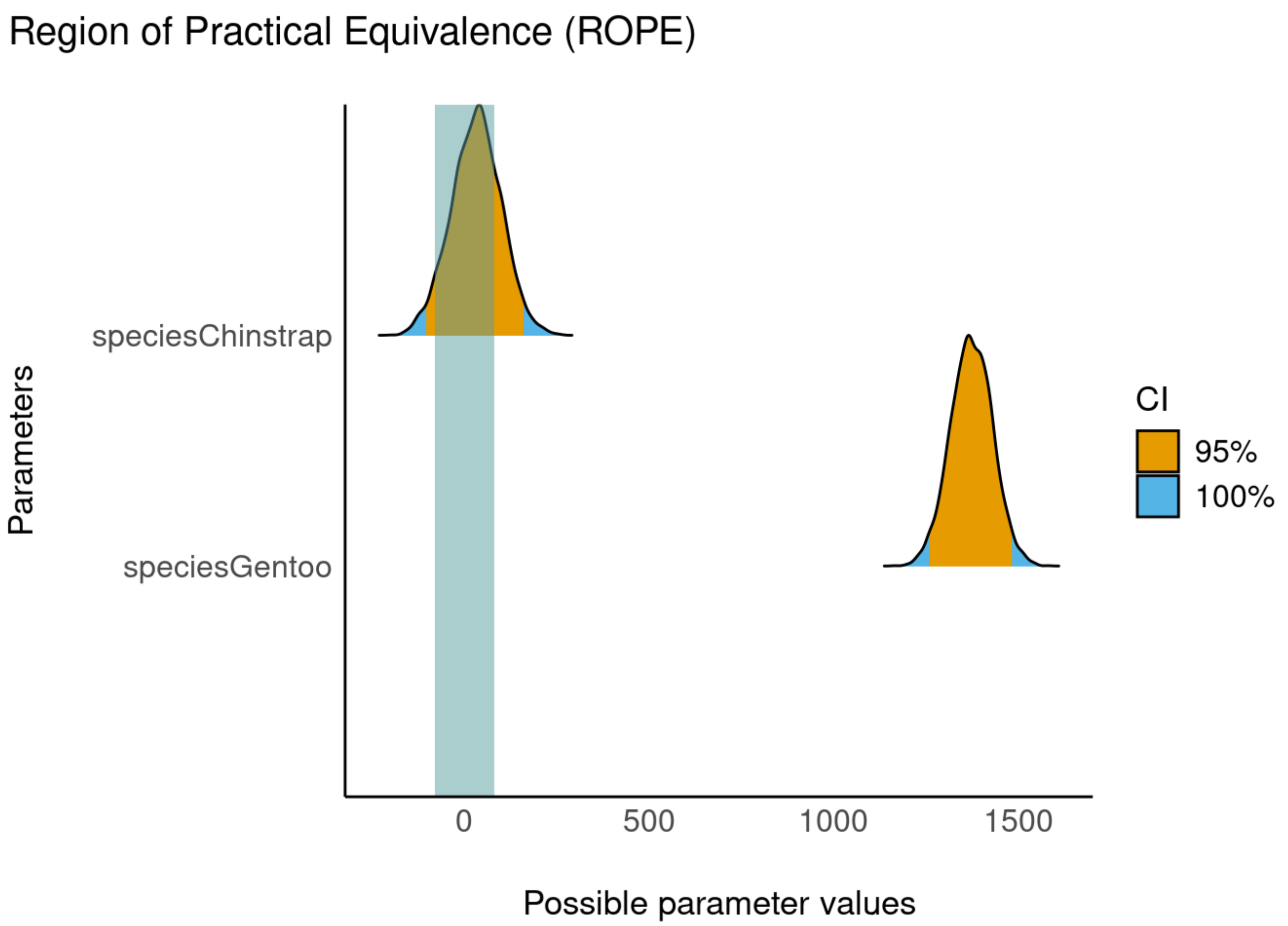
Die angegebenen Prozentwerte beziehen sich nicht auf die 100% der Post-Verteilung, sondern (in der Voreinstellung) auf das 95%-ETI, s. `help(rope)`.

Das hört sich abstrakt an? Dann lassen Sie uns das lieber visualisieren. 🎨

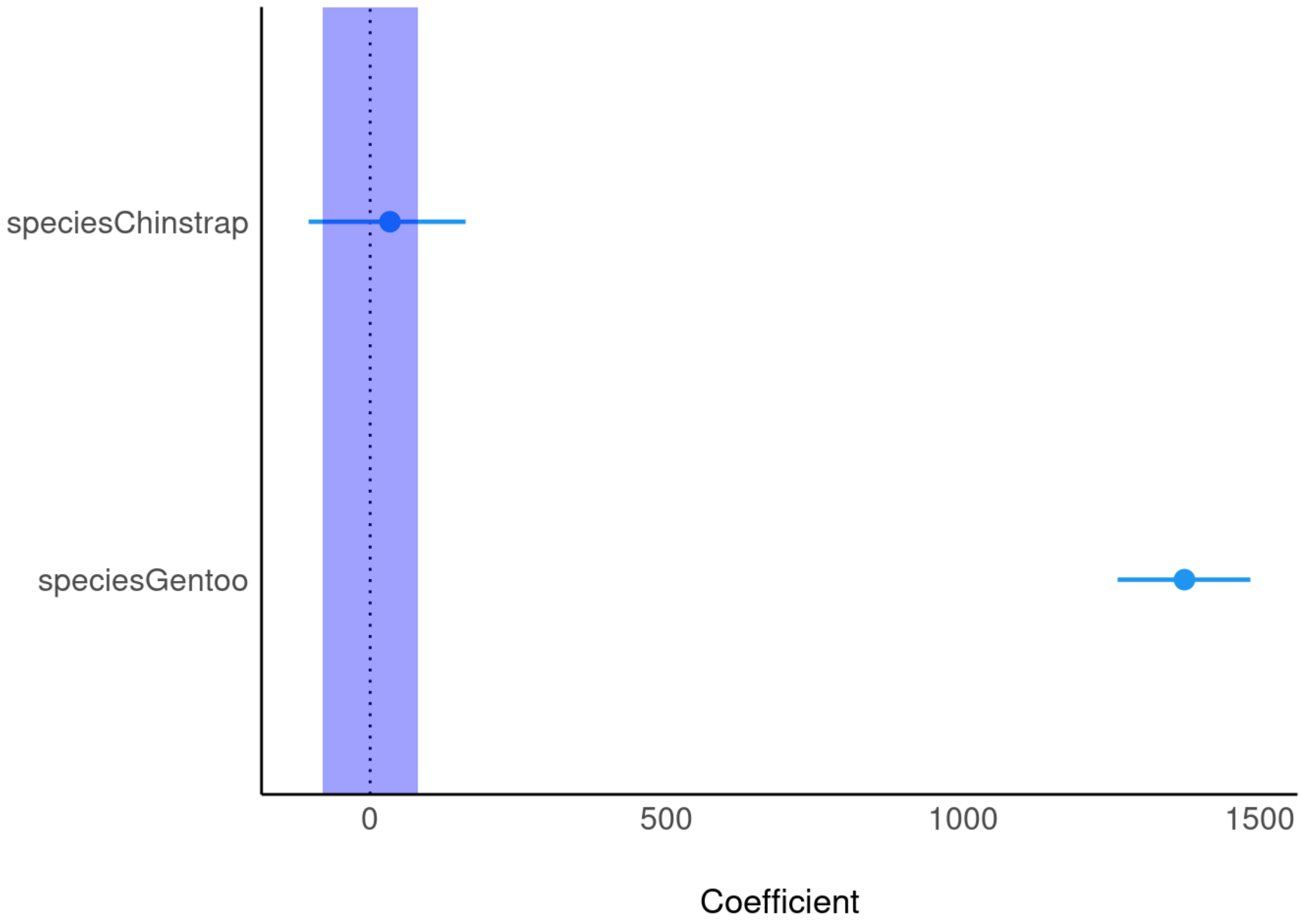
9.3.6 Visualisierung unserer Rope-Werte, m10.6

Ein Großteil der Posteriori-Masse von `m10.6` liegt *nicht* innerhalb des Rope. Aber können wir umgekehrt sagen, dass ein Großteil außerhalb liegt? Das erkennt man optisch ganz gut, s. Abbildung 9.3.

```
rope(m10.6) %>% plot()
```



(a) Diagramm mit `rope(m10.6) %>% plot()`



(b) Diagramm mit `parameters(m10.6) %>% plot()`

Abbildung 9.3: Rope und HDI überlappen bei Chinstrap, aber nicht bei Gentoo. Im ersten Fall nehmen wir die Rope-Null-Hypothese an, im zweiten Fall verwerfen wir sie.

Das ROPE druchkreuzt die "Berge" der Posteriori-Verteilung für Chinstrap deutlich. Aber: Das 95%-HDI liegt nicht komplett innerhalb des Rope. Wir können das Nullhypothese für Chinstrap *nicht verwerfen*, aber auch *nicht bestätigen*.

Gentoo hingegen wird vom vom Rope nicht durchkreuzt, es ist weit entfernt vom "blauen Fluss" des Rope: Gentoo liegt außerhalb des Rope. Es gibt einen "substanziellen" Unterschied, größer als das ROPE. Wir verwerfen die "Praktisch-Null-Hypothese" in diesem Fall.

9.3.7 Finetuning des Rope

Wir können festlegen, was wir unter "praktischer Äquivalenz" verstehen, also die Grenzen des Ropes verändern. Sagen wir, 100 Gramm sind unsere Grenze für einen vernachlässigbaren Effekt, s. Abbildung 9.4.

```
rope(m10.6, range = c(-100, 100))
plot(rope(m10.6, range = c(-100, 100))) + scale_fill_okabeito()
```

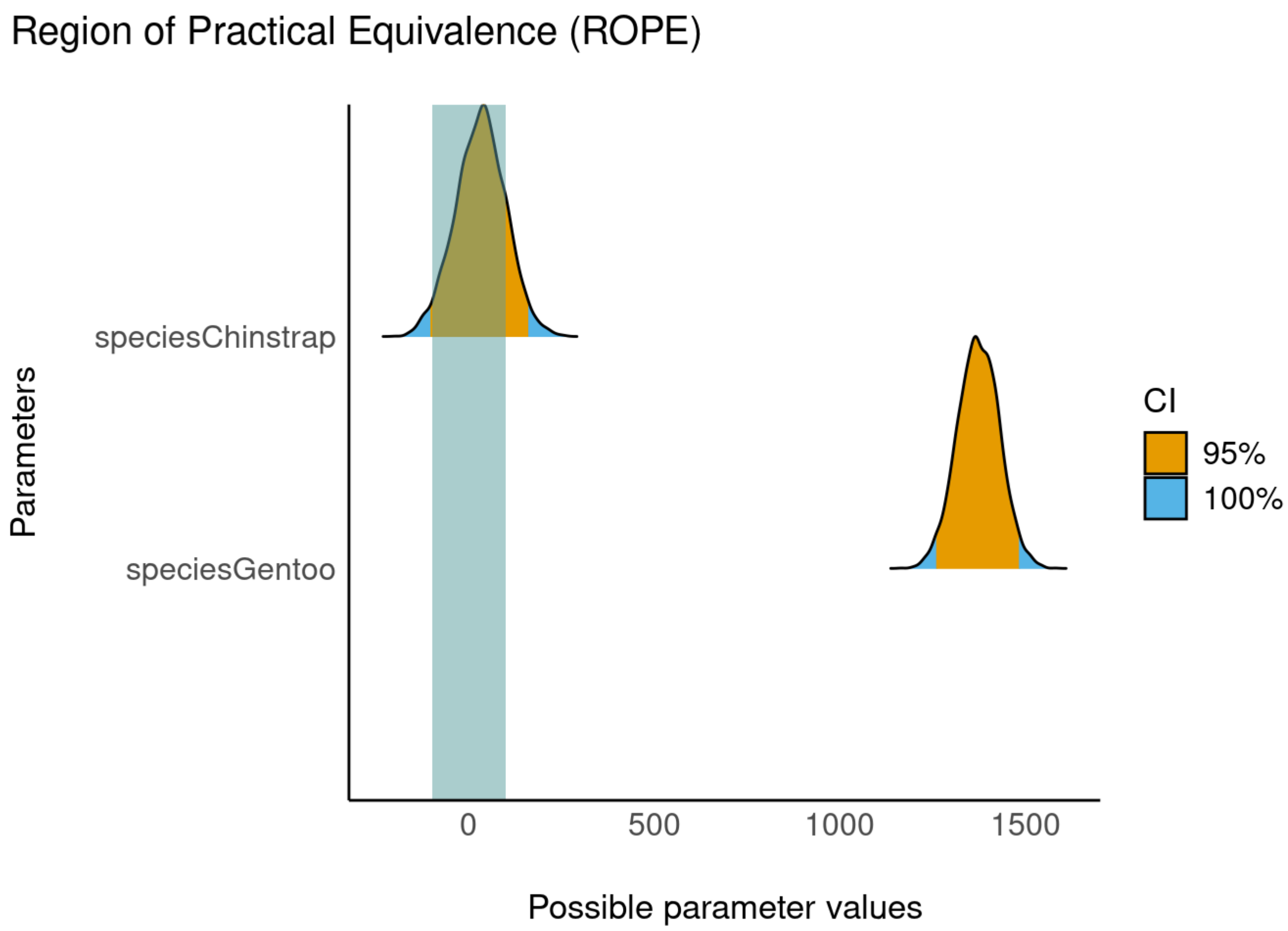


Abbildung 9.4: ROPE mit selber eingestellter Grenze von ±100 (Gramm)

Im Standard werden 95%-HDI berichtet, das kann man so ändern, wenn man möchte:

```
rope(m10.6, range = c(-100,100), ci = .89, ci_method = "ETI")
```

`ETI` (equal tails interval) steht für ein PI. Jetzt wird berichtet, welcher Teil eines 89%-CI[4] sich im Rope befindet.

9.3.8 Beantwortung der Forschungsfrage

Für die Spezeis *Gentoo* wurde ein substanzieller Gewichtsunterschied zur Referenzgruppe, *Adelie*, vom Modell entdeckt. Für *Chinstrap* hingegen ist keine klare inferenzstatistische Aussage hinsichtlich eines Indifferenzbereichs möglich: Es ist plausibel, laut dem Modell, dass es einen praktisch bedeutsamen Unterschied gibt, aber es ist auch plausibel, dass es keinen praktisch bedeutsamen Unterschied gibt.

9.4 Modellgüte

9.4.1 Wozu Modellgüte?

Hat man ein Modell aufgestellt und geprüft und Ergebnisse erhalten, möchte man wissen, wie belastbar diese Ergebnisse sind. Ein Hinweis zur Belastbarkeit des Modellergebnisse liefern Kennwerte der Modellgüte. Diese Kennwerte zielen z. B. darauf ab, wie *präzise* die Aussagen des Modells sind. Je präziser die Aussagen eines Modells, desto nützlicher ist es natürlich. Bei einer Parameterschätzung erhält man auch Informationen zur Präzision der Schätzung: Ist der Schätzbereich schmal, so ist die Schätzung präzise (und vice versa). Allerdings könnte ein Modell aus mehreren Parameterschätzungen bestehen, die unterschiedlich präzise sind. Da kann es helfen, eine zusammenfassen Beurteilung zur Präzision, oder allgemeiner zur Güte des Modells, zu erhalten.

Im Folgenden ist eine Kennzahl von mehreren gebräuchlichen und sinnvollen vorgestellt, R^2.

9.4.2 Modellgüte mit R^2 bestimmen

R^2 gibt den Anteil der Gesamtvarianz (der AV) an, den das Modell erklärt. - Höhere Wert von R^2 bedeuten, dass das Modell die Daten besser erklärt. R^2 wird normalerweise auf Basis eines Punktschätzers definiert. Solch eine Definition lässt aber viel Information - über die Ungewissheit der Schätzung - außen vor. Daher ist es wünschenswert, diese Information in R^2 einfließen zu lassen: *Bayes-R-Quadrat*.

Möchte man es ausführlicher, und im Komfort einer Bayes-Analyse schwelgen, so kann man sich die Posteriori-Verteilung von $R2$ ausgeben lassen, s. Abbildung 9.5.

```
m10.6_r2 <-
  m10.6 %>%
  r2_posterior() %>%
  as_tibble()

hdi(m10.6_r2) %>%
  plot()
```

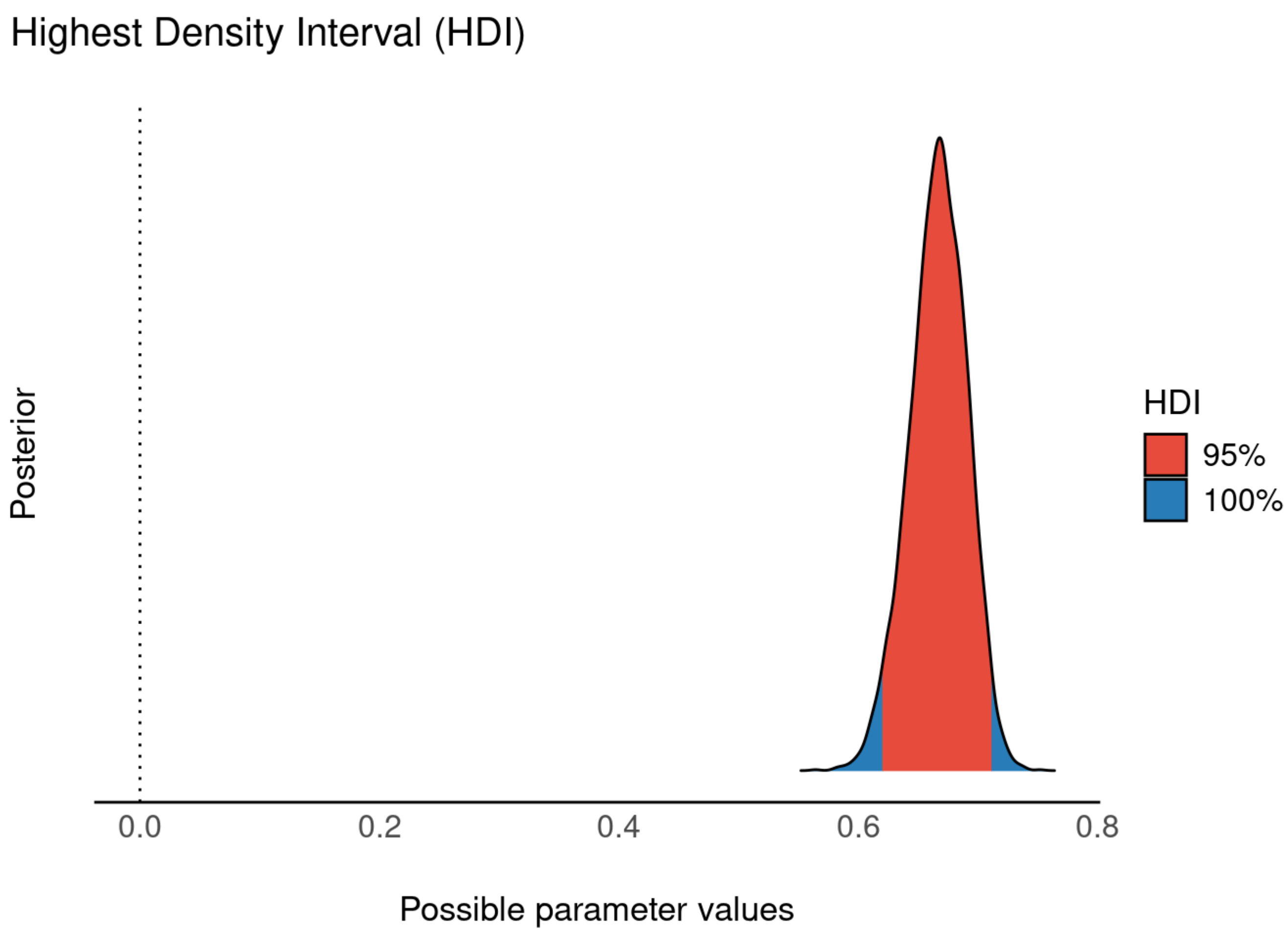


Abbildung 9.5: Die Verteilung von R-Quadrat im Modell m10.6

9.4.3 Definition vom "klassischen" R^2

Wie genau sind die Vorhersagen des Modells? σ (Vorhersagefehler) quantifiziert die Streuung der Residuen $r_i = y_i - X_i\hat{\beta}$, mit $\hat{y}_i = X_i\hat{\beta}$. Anders gesagt: $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots = X\hat{\beta}$. Anders gesagt gibt σ die "typische" Abweichung einer Beobachtung vom vorhergesagten Wert an. Es ist nützlich, σ in Bezug zu setzen zur Streuung der AV, $sd_y = s_y$: $R^2 = 1 - (\hat{\sigma}^2/s_y^2)$. $R2$ gibt damit den Anteil der vom Modell erklärten Varianz, V, an. Berechnet man das Modell mit der Methode der kleinsten Quadrate (nicht mit Bayes), dann ist der obige Ausdruck äquivalent zu: $R^2 = V_{i=1}^n \hat{y}_i / s_y^2$ Die beiden obigen Ausdrücke nehmen $\hat{y}_i$ als fix (sicher) an und vernachlässigen Ungewissheit; sie sind übergewiss aus Bayes-Sicht.

9.4.4 Bayes' R^2

Besser ist es (aus Bayes-Sicht), die Ungewissheit der Vorhersagen bei der Berechnung der Modellgüte miteinzubeziehen: $\text{Bayes } R^2 = \frac{\text{erkärte Varianz}}{\text{Erklärte Varianz + Residualvarianz}} = \frac{V_{mod}}{V_{mod}+V_{res}}$.

V_{mod} ist die Varianz in der PPV mit $s = 1, \ldots, S$ simulierten Stichproben, $V(\hat{y}_i)$ und V_{res} ist die Residualvarianz im Modell. Für jede Stichprobe s berechnet man die vorhergesagten Werte, $\hat{y}_i^s$, die Residualvarianz σ_s^2 und den Anteil der erklärten Varianz: $\text{Bayes } R_s^2 = \frac{V(\hat{y}_i^s)}{V(\hat{y}_i^s + \sigma_s^2)}$, vgl. Gelman u. a. (2019), Gelman, Hill, und Vehtari (2021), Kap. 11.7.

9.5 Fazit

Obwohl das Testen von Hypothesen im Moment verbreiteter ist, spricht einiges zugunsten der Vorzüge der Parameterschätzung. Möchte man aber, um sich bestimmter bestehender Forschung anzunähern, einen Hypothesentest, speziell den Test einer Nullhypothese verwenden, so bietet sich das ROPE-Verfahren an.

9.6 Aufgaben

1. Wskt-Schluckspecht
2. wskt-mtcars-1l
3. rope-regr
4. rope1
5. rope2
6. rope3

---

1. vor allem in der Frequentistischen Statistik ↩︎

2. Tatsächlich gibt es schwarze Schwäne, aber nicht in Europa: https://en.wikipedia.org/wiki/Black_swan ↩︎

3. Mittlerweile gibt es neue Frequentistische Ansätze für ein Verfahren ähnlich dem ROPE-Ansatz, der weiter unten vorgestellt wird. ↩︎

4. 89 ist die nächst kleinste Primzahl unter 95; und 95 wird gemeinhin als Grenzwert für Schätzbereiche verwendet. Damit ist 95 hier eine "magic number", ein Defacto-Standard ohne hinreichende Begründung. Um darauf hinzuweisen, benutzen einige Forschis mit ähem subtilen Humor lieber die 89 als die 95. 🤷‍♂️ ↩︎