6: Bayesian decision analysis and practical significance with ROPE, emmeans, and marginaleffects

Bayesian Mixed Effects Models with brms for Linguists

Job Schepens

2026-01-08

Table of contents

1 The Problem: Does Effect X Matter?

1.1 Research Scenario

Imagine you've run a psycholinguistic experiment and fitted a Bayesian model. You have posterior distributions for your effects. Now you face the question:

"Is this effect meaningful in practice?"

- You have an estimate (e.g., = 0.12 log-RT units)
- You have uncertainty (95% CI: [0.08, 0.16])
- But: Is this difference big enough to matter?

Practical significance is different from **statistical significance**:

- Statistical: “Is there an effect?” (distinguishable from noise) (we will discuss this in the session on Bayes Factor)
- Practical: “Does it matter?” (large enough to care about)

1.2 Why Practical Significance Matters

Consider these scenarios:

Scenario A: Statistically credible but trivial

- Effect: 2ms difference in reading time
- 95% CI: [1ms, 3ms] (clearly non-zero!)
- But: Readers can’t perceive < 20ms differences
- **Conclusion**: Real effect, but irrelevant

Scenario B: Uncertain but possibly large

- Effect: 50ms difference (noticeable to readers)
- 95% CI: [-10ms, 110ms] (wide uncertainty)
- Traditional conclusion: “not significant”
- But: Most credible values are meaningful
- **Conclusion**: Worth investigating further

2 Where We Are in the Analysis Workflow

2.1 The Bayesian Workflow So Far

Let’s review where we’ve been:

1. **Module 01**: Set priors
2. **Module 02**: Prior predictive checks
3. **Module 03**: Posterior predictive checks
4. **Module 04**: Compare priors → Sensitivity analysis
5. **Module 05**: Compare models → Evaluate predictive performance (LOO)
6. **Today (Module 06)**: Practical significance (ROPE + comparisons)
7. **Module 07**: Bayes factors & hypothesis comparison
8. **Module 08**: Convergence

2.2 When to Use Each Approach

```
QUESTION                            TOOL

Which model structure is better?    LOO (Module 05)
Is effect meaningful?               ROPE (Module 06)
Compare multiple groups?            emmeans (Module 06)
```

```
  Custom predictions/contrasts?       marginaleffects (M06)
  Evidence for hypothesis?            Bayes Factor (M07)
  Are estimates robust to priors?     Prior comparison (M04)
```

2.3 Choosing The Right Tool

Use ROPE when:

- You want to declare "effect too small to matter"
- You need clear decision rules (accept/reject/undecided)

Use emmeans when:

- You have factorial designs (multiple groups/conditions)

Use marginaleffects when:

- You're working with complex models (GAMs, interactions)

Use Bayes Factors (Module 07) when:

- You want to quantify evidence for one hypothesis over another (see Module 07)

Use LOO (Module 05) when:

- Comparing different model structures
- Doing feature selection
- Predictive performance is primary concern

2.4 Setup

2.5 Generate Reaction Time Data

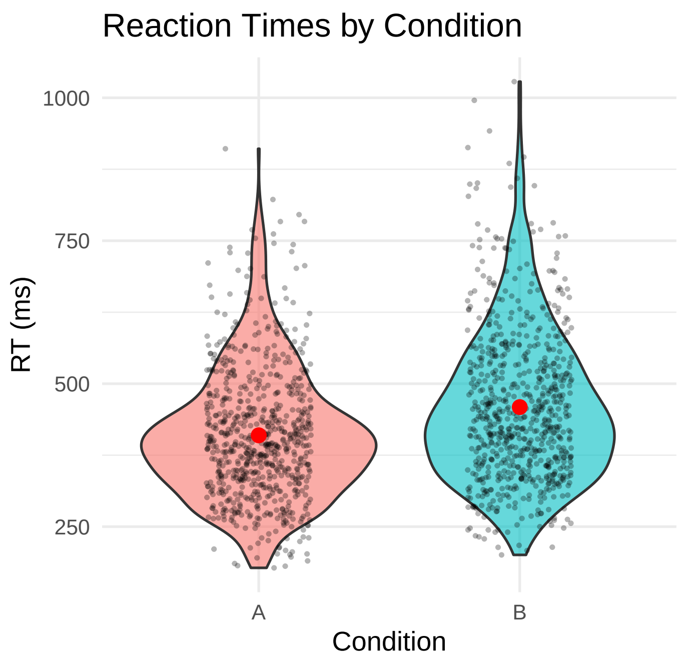
We'll generate RT data similar to previous modules, but with specific properties useful for hypothesis testing demonstrations:

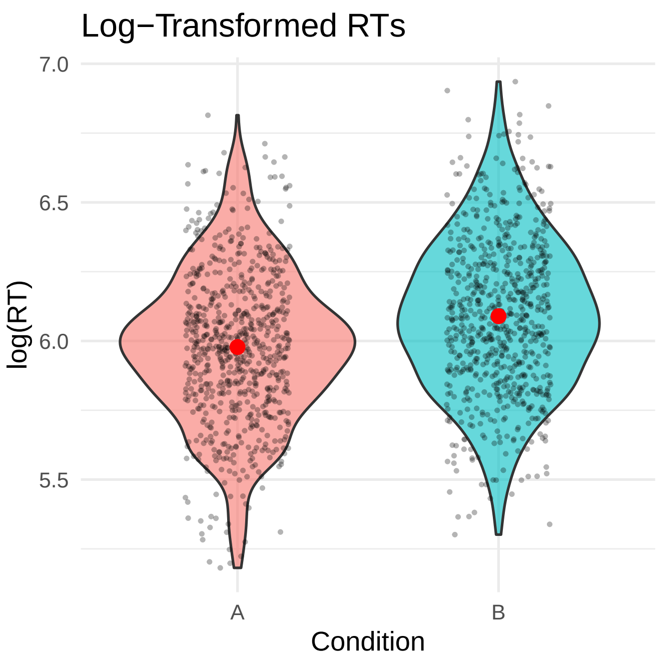
- Clear directional effect (Condition B slower than A)
- Effect size in a realistic range for psycholinguistics
- Adequate sample size for stable estimates

```
# A tibble: 2 x 6
  condition     n mean_rt median_rt sd_rt mean_log_rt
  <fct>     <int>   <dbl>     <dbl> <dbl>       <dbl>
1 A           720     410       395   114        5.98
2 B           720     459       439   133        6.09
```

```
# A tibble: 1 x 2
  Measure                Value
  <chr>                  <dbl>
1 Effect size (log scale) 0.111
```

2.6 Visualize the Data


Reaction Times by Condition


Log−Transformed RTs

2.7 Fit the Model

2.8 Model Summary

```
 Family: gaussian
  Links: mu = identity
Formula: log_rt ~ condition + (1 + condition | subject) + (1 | item)
   Data: rt_data (Number of observations: 1440)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Multilevel Hyperparameters:
~item (Number of levels: 24)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.11      0.02     0.08     0.15 1.00      965     1722

~subject (Number of levels: 30)
                          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
sd(Intercept)                 0.17      0.02     0.13     0.22 1.00      828
sd(conditionB)                0.06      0.02     0.02     0.09 1.00      797
cor(Intercept,conditionB)     0.04      0.26    -0.45     0.55 1.00     2292
                          Tail_ESS
sd(Intercept)                 1621
sd(conditionB)                 423
cor(Intercept,conditionB)     2137

Regression Coefficients:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
```

```
Intercept      5.98      0.04     5.90     6.05 1.01      568      802
conditionB     0.11      0.01     0.08     0.14 1.00     2387     2584

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.20      0.00     0.19     0.21 1.00     4298     2793

Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

3 ROPE (Kruschke, 2015) and Decision Analysis (Gelman et al., 2013)

3.1 ROPE

Statistical significance tells us if an effect differs from zero. But:

- **Statistical question:** Is the effect *exactly* zero?
- **Practical question:** Is the effect *close enough* to zero to ignore?

This is where **ROPE** (Region of Practical Equivalence) comes in.

3.1.1 Making ROPE-based Decisions

Instead of testing if $= 0$ exactly, we define a small interval around zero:

$$\text{ROPE} = [-\varepsilon, +\varepsilon]$$

where ε is the **smallest effect size we care about** (domain-specific).

Decision rules:

1. **95% HDI entirely inside ROPE** → Accept practical equivalence (effect negligible)
2. **95% HDI entirely outside ROPE** → Reject equivalence (effect matters)

3. **95% HDI overlaps ROPE** → Uncertain (collect more data or accept uncertainty)

Terminology: HDI vs. HPD

Throughout this document, we use **HDI** (Highest Density Interval) to refer to the credible interval containing the 95% most probable parameter values with the shortest width. This is also called **HPD** (Highest Posterior Density) or **HPDI** in some literature.
All three terms refer to the same concept: - HDI = Highest Density Interval (most common in modern usage) - HPD = Highest Posterior Density (common in older literature)
- HPDI = Highest Posterior Density Interval (explicit combination)
The R package `HDInterval::hdi()` and the emmeans output column `lower.HPD`/`upper.HPD` both compute this same interval type.

3.1.2 Setting ROPE Boundaries

! Common Pitfall: Post-Hoc ROPE Boundaries

ROPE boundaries must be set **before seeing results** based on:

- Domain knowledge (e.g., “RT differences $<$ 50ms are imperceptible”)
- Standardized effect sizes (e.g., “Cohen’s d $<$ 0.1 considered negligible”)

Mistake: Setting ROPE after looking at posterior to get desired conclusion.
Why it matters: Post-hoc boundaries invalidate the test (like p-hacking).

3.1.3 Four Methods for Justifying ROPE Boundaries

From Kruschke (2015, Chapter 12): ROPE boundaries should represent the smallest effect you care about. Here are four principled methods for setting them:

3.1.3.1 Method 1: Previous Research / Meta-Analysis

Use effect sizes from prior literature to calibrate what counts as “small.”

Method	Typical Effect	ROPE Boundaries	Interpretation
Based on Meta-Analysis	0.10 log-units (100ms or 10%)	[-0.03, 0.03]	Effects $<$ 3% are small relative to field norms

When to use:

- Mature research area with existing effect size estimates
- You want to compare to “typical” effects in the field
- You have access to meta-analyses or large-scale studies

Example:

```
In their meta-analysis of 50 reading time studies, Smith et al. (2020)
report a mean effect of d = 0.35 for syntactic complexity manipulations.
We set our ROPE at d = 0.10 (approximately 1/3 of the typical effect),
corresponding to ±0.03 log-RT units.
```

3.1.3.2 Method 2: Measurement Precision

ROPE should exceed measurement error—otherwise you’re testing noise.

Method	Measurement Error	ROPE Boundaries	Interpretation
Based on Measurement Precision	0.02 log-units (20ms)	[-0.03, 0.03]	Effects smaller than measurement error are unreliable

When to use:

- You have reliability/measurement error estimates
- You want to avoid claiming effects smaller than noise
- Measurement precision is the limiting factor

How to estimate measurement error:

```
# A tibble: 3 x 2
  Measure            Value
  <chr>              <chr>
1 Within-subject SD  0.085
2 Suggested ROPE (±) 0.127
3 ROPE Boundaries    [-0.127, 0.127]
```

3.1.3.3 Method 3: Perceptual/Practical Thresholds

Effects below perception or practical impact are negligible by definition.

Method	Perceptual Threshold	ROPE Boundaries	Interpretation
Based on Perceptual Threshold	50ms (0.05 log-units)	[-0.05, 0.05]	Effects imperceptible to readers are negligible

When to use:

- You can measure perceptual/practical thresholds
- You care about real-world impact (not just statistical detection)
- You have pilot data or existing benchmarks

Example applications:

- **Reading:** Just-noticeable difference in reading time
- **Accuracy:** Minimal detectable accuracy change
- **Rating scales:** Smallest perceived difference on Likert scale
- **Medical:** Minimum clinically meaningful difference (MCID)

How to measure perceptual thresholds:

```
Pilot study design:
1. Show participants trials from both conditions
2. Ask: "Did you notice any difference in difficulty/speed?"
3. Correlate subjective reports with actual RT differences
4. Find threshold below which participants don't notice
→ Use this threshold as ROPE
```

3.1.3.4 Method 4: Standardized Effect Sizes (Cohen's d)

Use conventional effect size benchmarks from your field.

Approach: Cohen's guidelines suggest d = 0.2 is a "small" effect. We'll set ROPE at half of that: d = 0.10, representing a "very small" effect.

To convert Cohen's d to the log-RT scale, we use:

$$d = \frac{\mu_1 - \mu_2}{\text{SD}_{\text{pooled}}}$$

Formula for pooled standard deviation:

$$\text{SD}_{\text{pooled}} = \sqrt{\frac{\sum_{i=1}^{k}(n_i - 1) \times \text{SD}_i^2}{\sum_{i=1}^{k}(n_i - 1)}}$$

where k is the number of groups (conditions), n_i is the sample size for group i, and SD_i is the standard deviation for group i.

Cohen's d Threshold	Pooled SD (log-units)	ROPE Lower	ROPE Upper	Interpretation
0.1	0.278	-0.028	0.028	Effects with d < 0.10 are 'very small'

When to use:

- No domain-specific benchmarks available
- You want to follow field conventions
- Standardized metrics are expected in your area

Common benchmarks:

- **Cohen's d:** Small = 0.2, Medium = 0.5, Large = 0.8
- **Correlation (r):** Small = 0.1, Medium = 0.3, Large = 0.5

- **R²:** Small = 0.01, Medium = 0.09, Large = 0.25

Note: These are **conventions**, not laws of nature! Domain-specific thresholds (Methods 1-3) are usually better.

3.1.4 For RT Data (log-scale)

For our analysis:

- **ROPE = [-0.05, +0.05]** on the log scale
- **On original scale:** This corresponds to RT ratios between 0.95 and 1.05
  - Since we model log(RT), a difference of ±0.05 on the log scale = $e^{\pm 0.05}$ = multiplying RT by 0.95 to 1.05
  - Example: If Condition A = 400ms, ROPE means effects between 380ms and 420ms (±5%)
- **Interpretation:** RT differences smaller than 5% are too small to be practically meaningful

3.2 Understanding ROPE Graphically

3.2.1 Excursion to Decision Analysis: The Implicit Utility Function (Gelman et al., 2013)

Let's make explicit what ROPE boundaries mean in decision-theoretic terms (Gelman et al., 2013, Chapter 9; Stan User's Guide on Decision Analysis).

Following the formal framework of Bayesian decision analysis:

3.2.1.1 Step 1: Define Decisions and Outcomes

$$\begin{aligned} D &= \{\text{claim meaningful}, \text{claim negligible}, \text{undecided}\} \\ X &= \mathbb{R} \quad (\text{true effect size } \beta) \end{aligned}$$

- **Decisions**: We must choose one of three actions:
  - Claim the effect is **meaningful** (reject practical equivalence)
  - Claim the effect is **negligible** (accept practical equivalence)
  - Remain **undecided** (collect more data)
- **Outcomes**: The true but unknown effect size β (e.g., difference in log RT)

3.2.1.2 Step 2: Define Probability Distribution of Outcomes

$$p(\beta \mid d, \text{data}) = \int p(\beta \mid \theta) \cdot p(\theta \mid \text{data})\, d\theta$$

- This is our **posterior distribution** from the Bayesian model
- Represents uncertainty about the true effect size β given observed data
- The posterior is independent of our decision d (decisions don't change reality!)

3.2.1.3 Step 3: Define Utility Function

$$U(\beta, d) = \begin{cases} |\beta| & \text{if } d = \text{claim meaningful and } |\beta| > \text{ROPE} \\ -2|\beta| - 0.3 & \text{if } d = \text{claim meaningful and } |\beta| \leq \text{ROPE (false positive)} \\ 1 - |\beta| & \text{if } d = \text{claim negligible and } |\beta| \leq \text{ROPE} \\ -1.5|\beta| - 0.3 & \text{if } d = \text{claim negligible and } |\beta| > \text{ROPE (false negative)} \\ -0.15 & \text{if } d = \text{undecided (cost of data collection)} \end{cases}$$

Utility interpretation:

- **Correct decisions** have positive utility that depends on effect magnitude
- **Incorrect decisions** incur loss: fixed penalty (-0.3) + proportional penalty
- **False positives** penalized more (-2×) than false negatives (-1.5×)
- **Remaining undecided** has fixed cost (-0.15) representing study time, participant recruitment, delayed publication

3.2.1.4 Step 4: Choose Decision with Highest Expected Utility

$$d^* = \arg\max_d \mathbb{E}[U(\beta, d) \mid \text{data}] = \arg\max_d \int U(\beta, d) \cdot p(\beta \mid \text{data})\, d\beta$$

The **Bayes optimal decision** maximizes expected utility by integrating utility over the posterior:

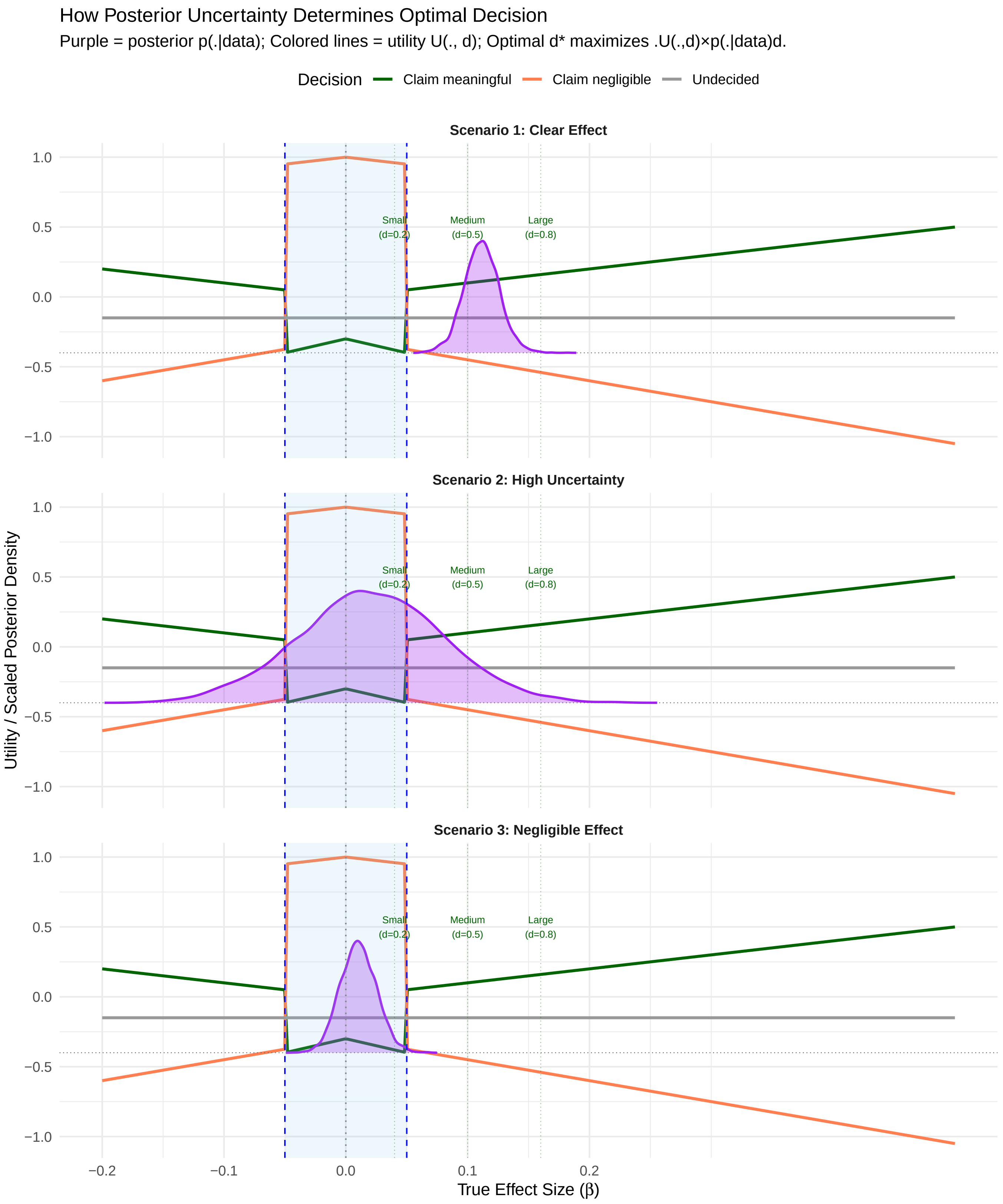
```
For each decision d:
  Expected_Utility(d) =  U( , d) × p(  | data) d

Choose d* with max Expected_Utility
```

In practice:

- **Scenario 1** (posterior outside ROPE): $\mathbb{E}[U \mid \text{claim meaningful}]$ is highest
- **Scenario 2** (posterior spans ROPE): $\mathbb{E}[U \mid \text{undecided}]$ is highest (uncertainty cost > data cost)
- **Scenario 3** (posterior inside ROPE): $\mathbb{E}[U \mid \text{claim negligible}]$ is highest

Scenario: You must decide whether to claim “Condition B is meaningfully slower” or “Condition B is essentially equivalent to A”.



What these three scenarios show:

1. **Scenario 1 (Clear Effect)**: Posterior mostly outside ROPE

   - Most purple area falls where green line is highest
   - **Optimal decision**: Claim meaningful effect
   - High expected utility for "claim meaningful"
2. **Scenario 2 (High Uncertainty)**: Posterior spans ROPE boundaries
   - Purple area spreads across both inside and outside ROPE
   - Risk of being wrong is high for either claim
   - **Optimal decision**: Undecided (collect more data)
   - Expected utility of claims reduced by uncertainty; -0.15 cost of more data is worth it
3. **Scenario 3 (Negligible Effect)**: Posterior concentrated inside ROPE near zero
   - Most purple area falls where orange line is highest
   - **Optimal decision**: Claim negligible effect
   - High expected utility for "claim negligible"

Key insight: The optimal decision emerges from integrating the utility curves (colored lines) over the posterior (purple distribution). Where your posterior mass concentrates determines which decision maximizes expected utility.

Understanding the utility curves

The three colored lines represent different utility functions $U(\beta, d)$ for each decision $d \in D$. Here's how to interpret them:

3.2.1.5 Why different maximum values?

The utility functions reflect different research goals:

- **Orange line** ($d =$ claim_negligible): $U(\beta, \text{negligible}) = 1 - |\beta|$ when $|\beta| \leq$ ROPE
  - Maximum = 1.0 when $\beta = 0$ exactly (confirming true null is maximally valuable)
  - Decreases linearly as β approaches ROPE boundary
  - **Interpretation**: "Establishing equivalence is valuable, but less so as effect approaches practical significance threshold"
- **Green line** ($d =$ claim_meaningful): $U(\beta, \text{meaningful}) = |\beta|$ when $|\beta| >$ ROPE
  - Utility equals effect size itself (no upper bound)
  - Example: At $\beta = 0.12$, utility = 0.12; at $\beta = 0.25$, utility = 0.25
  - **Interpretation**: "Discovering larger effects has greater scientific value (stronger evidence, bigger practical impact)"
- **Gray line** ($d =$ undecided): $U(\beta, \text{undecided}) = -0.15$ (constant)
  - Fixed cost independent of true β
  - Represents: study time, participant recruitment, delayed publication

Note: This asymmetry is a **modeling choice** reflecting common research values. You can define different utilities based on your domain.

3.2.1.6 Why are there different penalties for incorrect decisions?

At $\beta = \pm 0.05$ exactly (the ROPE boundary):

$$U(0.05, \text{meaningful}) = -2(0.05) - 0.3 = -0.40 \quad \text{(false positive: heavy penalty)}$$
$$U(0.05, \text{negligible}) = -1.5(0.05) - 0.3 = -0.375 \quad \text{(false negative: moderate penalty)}$$
$$U(0.05, \text{undecided}) = -0.15 \quad \text{(cost of data collection)}$$

Both wrong decisions incur:

- **Fixed penalty** (-0.3): Cost of misleading the literature, wasted resources, incorrect theory
- **Proportional penalty**: More wrong = worse (scaled by distance from truth)

Asymmetric loss structure: False positives penalized more heavily (-2×) than false negatives (-1.5×), reflecting greater harm from claiming non-existent effects.

3.2.1.7 When is "undecided" optimal?

KEY INSIGHT: The plot shows utility if true β were known, but we have **uncertainty** (posterior distribution).

The optimal decision maximizes **expected utility**:

$$d^* = \arg\max_d \mathbb{E}[U(\beta, d) \mid \text{data}] = \arg\max_d \int U(\beta, d) \cdot p(\beta \mid \text{data})\, d\beta$$

- **Low uncertainty, posterior inside ROPE**: $\mathbb{E}[U(\text{negligible})]$ highest
- **Low uncertainty, posterior outside ROPE**: $\mathbb{E}[U(\text{meaningful})]$ highest
- **High uncertainty spanning boundaries**: $\mathbb{E}[U(\text{undecided})]$ can be highest

Example: If your 95% HDI = [-0.02, 0.08]:

- Posterior mass inside ROPE → favors "negligible"
- Posterior mass outside ROPE → favors "meaningful"
- Risk of being wrong for either claim is high
- Expected utility of both claims reduced by uncertainty
- Fixed cost of "undecided" (-0.15) may beat both risky claims

The gray line isn't highest at any single β value, but wins when **averaging across uncertain β values weighted by posterior probability**.

3.2.1.8 Why does Utility grow with distance from boundaries

- **Larger true effects** → Higher utility for correct "meaningful" claim (green increases)
- **Effects closer to zero** → Higher utility for correct "negligible" claim (orange increases toward 1)
- **Being wrong by more** → Larger proportional loss (steeper slopes in penalty regions)

This encourages decisive conclusions when data strongly favor one region, but caution when near boundaries.

! ROPE Boundaries = Utility Crossover Points

Your ROPE boundaries should be set where the utilities of "claim meaningful" and "claim negligible" are equal. This is your **smallest effect size of interest (SESOI)**: the threshold where scientific conclusions qualitatively change.

In formal decision theory terms: ROPE defines the partition of outcome space X where different decisions $d \in D$ have maximal utility.

Computing expected utilities:

```
         decision expected_utility
1 claim_meaningful        0.1114245
2        undecided       -0.1500000
3 claim_negligible       -0.4671367
```

The optimal decision is to **claim_meaningful** (maximizes expected utility).

ℹ From Decision Theory to ROPE

The traditional ROPE decision rule: - “If 95% HDI excludes ROPE → Reject null”
…is actually an approximation to: - “Choose the decision with maximum expected utility”
The HDI-based rule works well when:

1. Losses are approximately symmetric
2. We want to control error rates at ~5%
3. We prefer simple rules over computing expected utilities

For asymmetric losses or complex decisions, computing expected utilities explicitly (as shown above) provides more principled decisions.

❗ Why ROPE Boundaries Matter

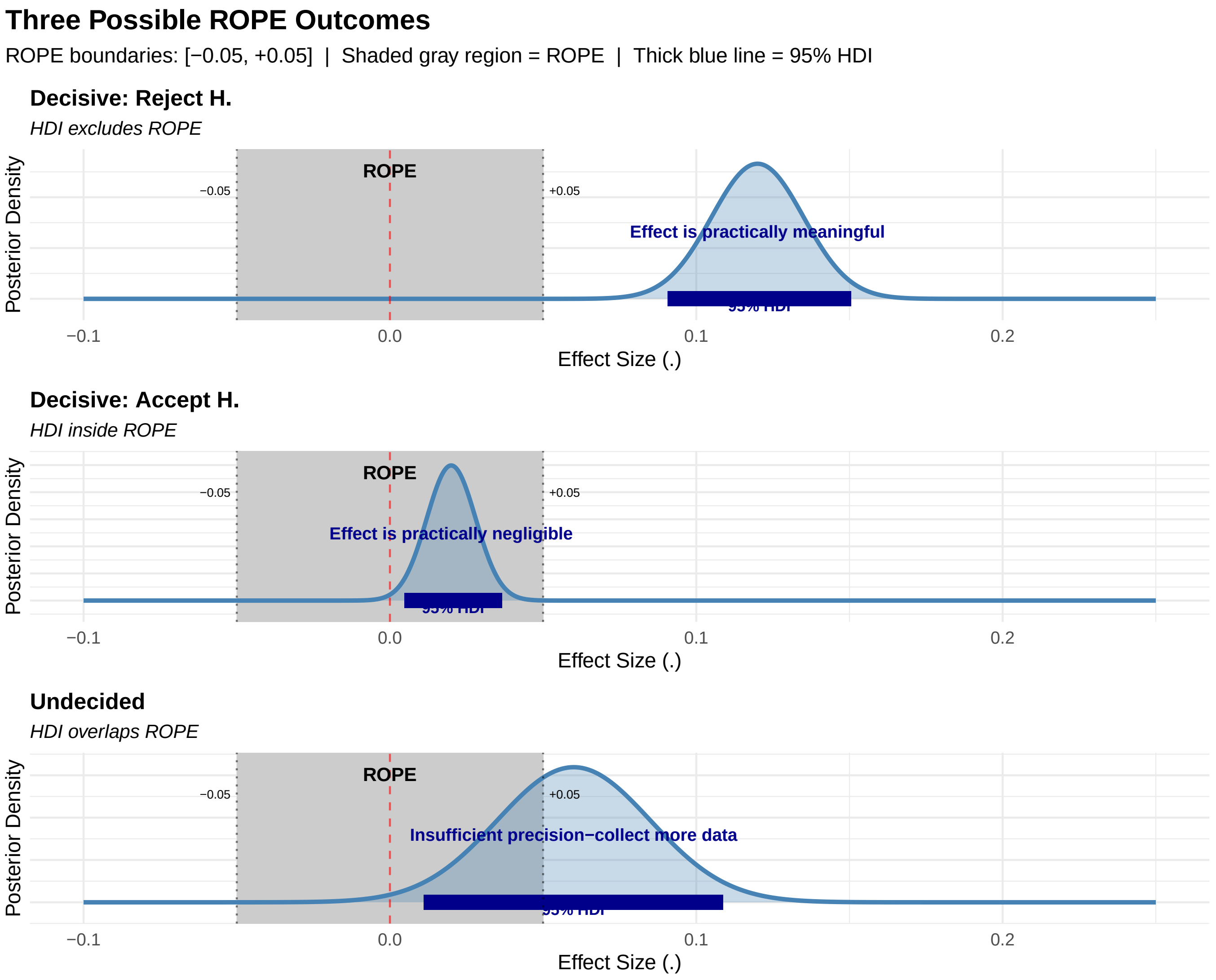
ROPE is not arbitrary! It’s the Bayesian optimal decision given:

1. Your posterior beliefs (from the model)
2. Your utility function (encoded in ROPE boundaries)
3. Your decision threshold (e.g., 95% credibility)

Changing ROPE boundaries = changing your utility function = changing what you consider “meaningful enough to matter”.

3.2.2 The Same Graph But Simpler

ROPE analysis has **three possible outcomes**, not just two as in NHST!



Three Possible ROPE Outcomes

ROPE boundaries: [−0.05, +0.05] | Shaded gray region = ROPE | Thick blue line = 95% HDI

Decisive: Reject H.

HDI excludes ROPE

Decisive: Accept H.

HDI inside ROPE

Undecided

HDI overlaps ROPE

! Three Outcomes, Not Two!

Current practice shows:

1. **HDI excludes ROPE** → Reject H (effect is meaningful)
   - Example: 95% HDI = [0.09, 0.15], ROPE = [−0.05, 0.05]
   - Interpretation: "Effect is decisively larger than our practical significance threshold"
2. **HDI inside ROPE** → Accept H (effect is negligible)
   - Example: 95% HDI = [0.01, 0.03], ROPE = [−0.05, 0.05]
   - Interpretation: "Effect is decisively smaller than our practical significance threshold"
3. **HDI overlaps ROPE** → UNDECIDED (insufficient precision)
   - Example: 95% HDI = [0.03, 0.09], ROPE = [−0.05, 0.05]
   - Interpretation: "We cannot make a clear decision—some credible values are meaningful, others aren't"
   - **This is not a failure!** It's honest reporting of uncertainty

“Undecided” Is a Feature, Not a Bug

From Kruschke (2015, p. 338):

> “Be clear that any discrete decision about rejecting or accepting a null value does not exhaustively capture our knowledge about the parameter value. Our knowledge about the parameter value is described by the **full posterior distribution**.”

When HDI overlaps ROPE: - You have learned something: The effect might or might not be meaningful - Your options: 1. Collect more data to narrow the posterior 2. Accept the uncertainty and make a practical decision based on other factors 3. Use a less conservative threshold (e.g., 89% HDI instead of 95%) - **Don’t** force a binary decision when the data don’t support one!

3.2.3 ROPE Decision Flowchart

```
                    Calculate 95% HDI
                            ↓



        Does HDI exclude ROPE?      Does HDI fall
                                     inside ROPE?

         YES  ↓  NO                  YES  ↓  NO




   ↓                                                   ↓
Reject H :                                      UNDECIDED:
"Effect is                                    "Overlapping-
 meaningful"                                   insufficient
                                               precision"
                      ↑


                 Accept H :
                "Effect is
                 negligible"
```

Manual ROPE Calculation (HDI-Based Approximation)

ROPE is simpler than computing utilities explicitly and works well for most research questions.
In practice, we usually use the HDI-based ROPE approximation rather than computing expected utilities explicitly. This is computationally simpler and works well when:

- Losses are approximately symmetric (false positive   false negative)
- We want standard 95% decision threshold

- We prefer simple rules over custom utility functions

Now let’s use the traditional ROPE decision rules (which approximate the decision-theoretic framework):

```
# A tibble: 5 x 2
```

```
  Measure          Value
  <chr>            <chr>
1 ROPE boundaries [-0.05, 0.05]
2 95% HDI         [0.084, 0.142]
3 % Below ROPE    0.0%
4 % Inside ROPE   0.0%
5 % Above ROPE    100.0%
[1] "**REJECT equivalence**: Effect is practically meaningful (positive) - 95% HDI entirely abov
```

Decision Rule = Maximizing Expected Utility

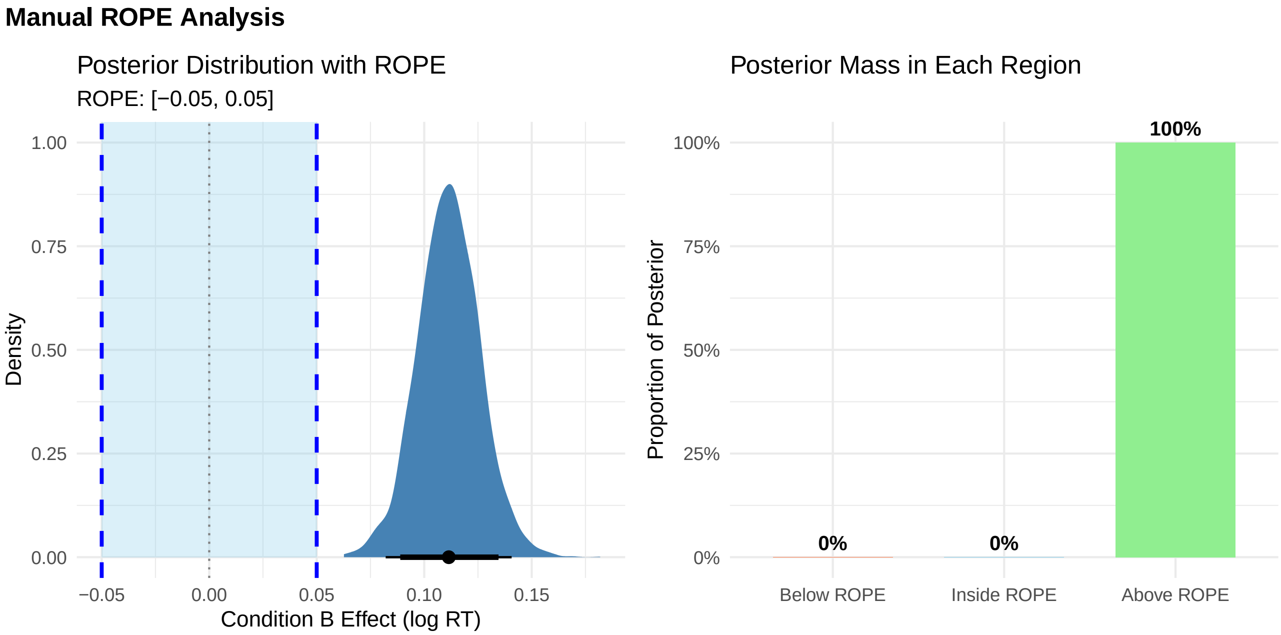
What just happened in decision-theoretic terms:

1. **We computed posterior probabilities**: P(in each region | Data)
2. **We applied a decision rule**: Based on where 95% HDI falls
3. **This approximates**: Choosing decision with maximum expected utility

The connection: - When 95% HDI > ROPE: E[U("meaningful")] > E[U("negligible")] with high confidence - When 95% HDI < ROPE: E[U("negligible")] > E[U("meaningful")] with high confidence - When overlapping: Expected utilities too close to call → "undecided" optimal

The 95% threshold encodes a loss function where: - We're willing to accept 5% risk of wrong decision - Losses are approximately symmetric for false positives vs. false negatives If your losses are **asymmetric** (e.g., false positives much worse), you should: - Use stricter threshold (e.g., 99% HDI), OR - Compute expected utilities explicitly (as shown in previous section)

3.2.4 Visualizing ROPE



3.3 Warning

3.3.1 Three Checks Before Trusting ROPE (Note: Just Informal Best Practice Checks)

Before trusting ROPE conclusions when accepting H , verify precision and reliability with these three checks:

1. HDI Width: Is your estimate precise enough?

- **Rule**: HDI width should be < half the ROPE width
- **Rationale**: If HDI is wide relative to ROPE, you lack precision to confidently say effect is negligible
- **Example**: ROPE = [-0.05, 0.05] (width 0.10) → HDI width should be < 0.05
- **Measures**: Overall precision of your estimate
- **If fail**: Collect more data before claiming negligible effect
- **Source**: Kruschke (2018, 2015), Lakens (2018)

2. Effective Sample Size (ESS): Is your posterior reliable?

- **Rule**: ESS bulk > 1000 AND ESS tail > 1000
- **Rationale**: Low ESS means MCMC chains haven't converged well - posterior estimates unreliable
- **ESS bulk**: Measures sampling efficiency for central posterior
- **ESS tail**: Measures sampling efficiency for HDI boundaries (critical for ROPE!)
- **Measures**: Quality of MCMC sampling, not data quantity
- **If fail**: Increase MCMC iterations, check convergence diagnostics ($\hat{R}$), consider reparameterization
- **Source**: e.g.: Vehtari et al. (2021) "Rank-Normalization, Folding, and Localization: An Improved $\hat{R}$ for Assessing Convergence of MCMC"; Stan Development Team documentation on MCMC diagnostics

3. HDI Position: Is the effect centered near zero? (ONLY when accepting H)

- **Rule**: HDI midpoint should be close to zero (e.g., within inner 50% of ROPE)
- **Rationale**: HDI inside ROPE but far from zero suggests bias or one-sided effect
- **Example**: ROPE = [-0.05, 0.05], HDI = [0.03, 0.045] → midpoint = 0.0375 (close to boundary, suspicious!)
- **Better example**: HDI = [-0.01, 0.02] → midpoint = 0.005 (centered near zero, reassuring)
- **Measures**: Location of effect, not just precision
- **Applies to**: ONLY when accepting H (claiming negligible effect); skip this check when rejecting H
- **If fail**: Re-examine priors, check for model misspecification, or acknowledge effect may be small but non-zero
- **Source**: Practical heuristic not explicitly stated in literature, but follows from Kruschke's (2018, p. 275) emphasis that accepting null requires "the estimated value be compatible with the null value" - an HDI asymmetrically positioned near a ROPE boundary questions this compatibility

Why these three checks?

- **Check 1 (HDI width)**: Ensures sufficient **precision** - can you distinguish effect from zero? (Kruschke, 2018) - *Always needed*
- **Check 2 (ESS)**: Ensures **reliability** - are the posterior estimates trustworthy? (Vehtari et al., 2021) - *Always needed*
- **Check 3 (HDI position)**: Ensures **centrality** - is the effect truly centered near zero, or just barely inside ROPE? (Practical heuristic derived from Kruschke's compatibility principle)

> - *Only for accepting H*
>
> All three can pass even with small sample size if the effect is truly negligible and model converges well. All three can fail with large sample size if MCMC struggles or effect is near boundary.

```
# A tibble: 8 x 2
  Check                             Value
  <chr>                             <chr>
1 "1. HDI Width"                    0.0576
2 "   Status"                         FAIL (> 0.05) - Need more data
3 "2. Effective Sample Size (bulk)" 2387
4 "   Effective Sample Size (tail)" 2584
5 "   Status"                         PASS (> 1000)
6 "3. HDI Position (midpoint)"      0.1132
7 "   Distance from zero"           0.1132
8 "   Status"                       N/A (Check only applies when accepting H  -~
```

Note: The HDI width check flagged a warning (0.0576 > 0.05 threshold). This indicates moderate precision—sufficient for rejecting H (since HDI excludes ROPE), but we'd want narrower estimates if claiming equivalence. With N=50 subjects, the HDI width would likely pass this check.

3.3.2 Scenario Examples

Scenario 1: Wide HDI Inside ROPE

- 95% HDI: [0.01, 0.04] (width = 0.03, midpoint = 0.025)
- ROPE: [-0.05, 0.05]
- WRONG: "Effect is negligible"
- RIGHT: "HDI inside ROPE, but width suggests moderate uncertainty AND midpoint close to boundary. More data needed."

Scenario 2: Narrow HDI Centered Near Zero

- 95% HDI: [-0.005, 0.015] (width = 0.02, midpoint = 0.005)
- ROPE: [-0.05, 0.05]
- RIGHT: "Effect is negligible. HDI is narrow, centered near zero, and falls well within ROPE."

Scenario 3: Narrow HDI But Off-Center

- 95% HDI: [0.03, 0.045] (width = 0.015, midpoint = 0.0375)
- ROPE: [-0.05, 0.05]
- CAUTION: "HDI is narrow and technically inside ROPE, but midpoint (0.0375) is close to upper boundary. Effect may be small but systematically positive. Consider whether this is practically negligible."

Scenario 4: HDI Overlaps ROPE Boundary

- 95% HDI: [0.03, 0.08]
- ROPE: [-0.05, 0.05]
- RIGHT: "Undecided. Some credible values fall within ROPE, others exceed it. More data needed."

❗ Rules of Thumb for Precision

Before using ROPE to **accept H** (claim negligible effect), verify:

For RT effects (log-scale):

- HDI width < 0.05 log-units
- HDI midpoint within ±0.025 of zero (inner 50% of typical ROPE)
  - *Why?* If midpoint is near ROPE boundary, effect may be small but non-negligible.

For accuracy (probability scale):

- HDI width < 0.10
- HDI midpoint within ±0.05 of zero

For all analyses:

- ESS (bulk) > 1000
- ESS (tail) > 1000
- Convergence diagnostics passed ($\hat{R} < 1.01$)

If these checks fail:

- Don't claim effect is negligible
- Report "insufficient precision" or "undecided"
- Consider collecting more data

When You Can Trust ROPE to Accept H

Strong evidence for negligible effect requires:

1. **Narrow posterior:** HDI width < half ROPE width
2. **Central location:** HDI midpoint near zero
3. **Good sampling:** ESS > 1000, $\hat{R} < 1.01$
4. **Model fit:** Posterior predictive checks pass

Example of strong evidence:

```
Effect:   = 0.02 (95% HDI: [0.01, 0.03])
ROPE: [-0.05, 0.05]
HDI width: 0.02 (  narrow)
ESS: 2500 (  adequate)
→ Can confidently claim negligible effect
```

Example of weak evidence:

```
Effect:   = 0.02 (95% HDI: [-0.01, 0.05])
ROPE: [-0.05, 0.05]
HDI width: 0.06 (  too wide)
ESS: 800 (  marginal)
→ Cannot confidently claim negligible effect
```

3.4 Using bayestestR Package

The `bayestestR` package provides convenient functions for ROPE analysis.

```
# Proportion of samples inside the ROPE [-0.05, 0.05]:

Parameter   | Inside ROPE
-------------------------
```

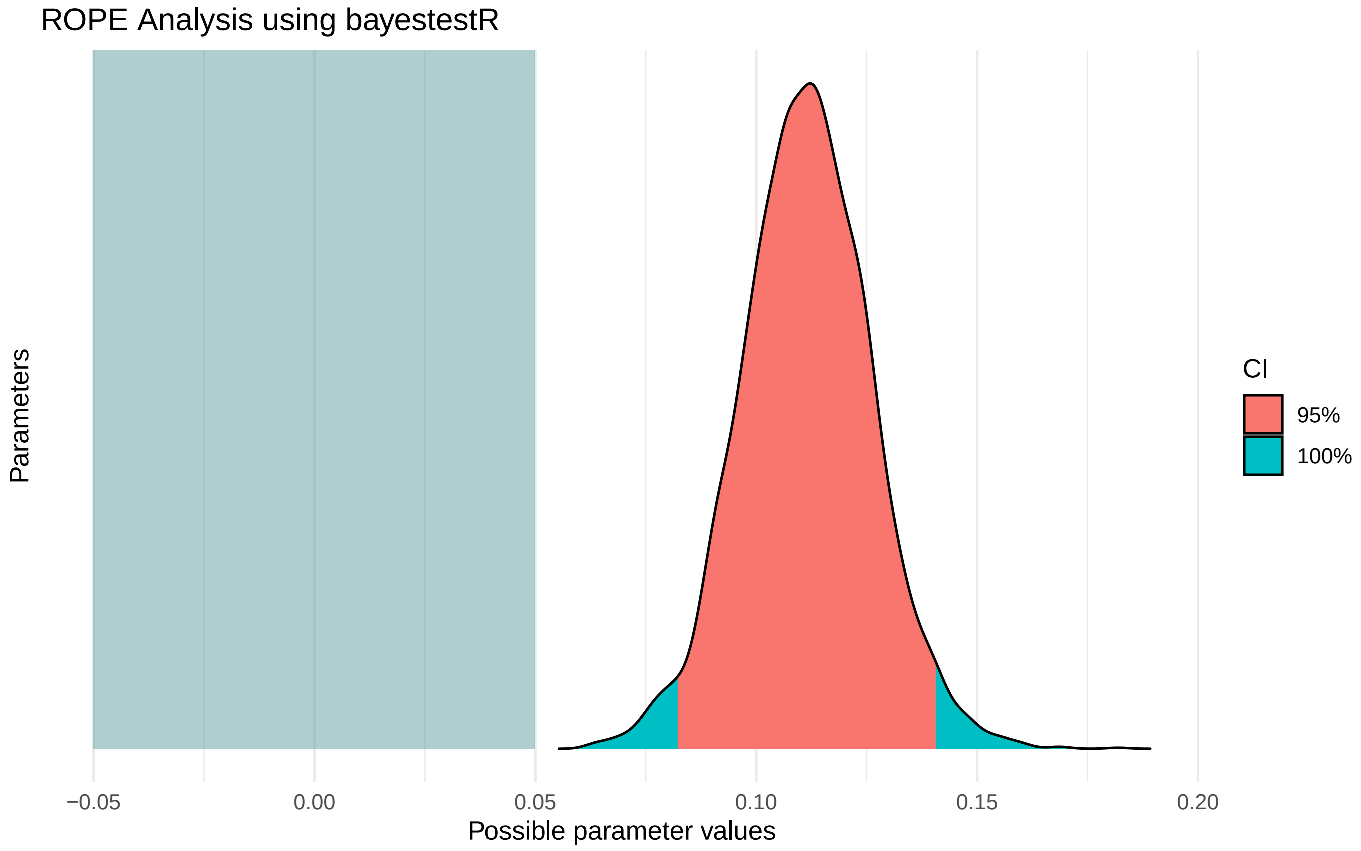
```
Intercept  |       0.00 %
conditionB |       0.00 %
```

3.4.1 Interpretation

The output shows:

- **% in ROPE**: Proportion of 95% HDI inside ROPE
- Interpretation automatically provided

3.4.2 Visualizing with bayestestR



3.4.3 Equivalence Test

The `equivalence_test()` function provides an integrated view:

```
# Test for Practical Equivalence

  ROPE: [-0.05 0.05]

Parameter  |       H0 | inside ROPE |      95% HDI
--------------------------------------------------
Intercept  | Rejected |      0.00 % | [5.90, 6.05]
conditionB | Rejected |      0.00 % | [0.08, 0.14]
```

This shows for each parameter:

- **% in ROPE**: How much of the posterior falls in ROPE
- **Decision**: Accepted/Rejected/Undecided

Comparison: Manual vs bayestestR

Both approaches should give the same results:

```
# A tibble: 2 x 3
  Source               `% in ROPE` Decision
  <chr>                      <dbl> <chr>   
1 Manual Calculation             0 Rejected
2 bayestestR Package             0 Rejected
```

What Just Happened?

Both approaches give identical results — bayestestR simply automates the manual calculations we did earlier. The benefit: - Less code to write - Automatic formatting - Built-in visualization - Consistent interface across analyses

When to Use Manual vs Package

Use manual calculation when:

- You want to understand the mechanics
- You need custom ROPE boundaries for different parameters
- You're teaching/explaining the concept

Use bayestestR when:

- You want quick, standardized results
- You're analyzing many models
- You want built-in visualization

Prior Sensitivity Analysis for ROPE Decisions

When Priors Matter

From Kruschke (2015, Section 12.2.1): ROPE conclusions can change with different priors.

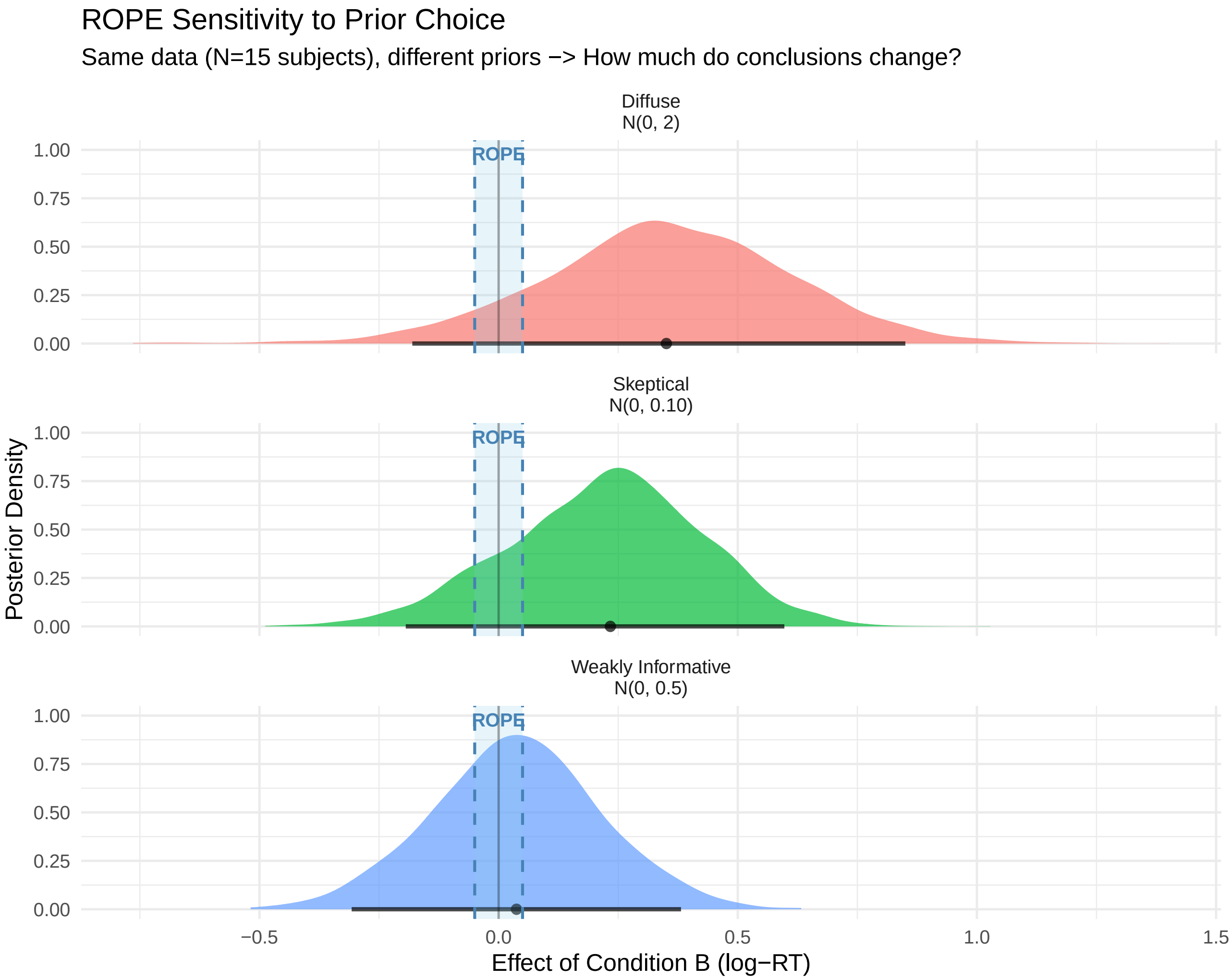
Always check sensitivity, especially when:

- Sample size is small (n < 30 per group)
- HDI barely touches ROPE boundary
- You're accepting H (claiming negligible effect)

3.4.4 Demonstration: Same Data, Different Priors, Different ROPE Conclusions

```
                            Prior   CI ROPE_low ROPE_high ROPE_Percentage
1 Weakly Informative\nN(0, 0.5) 0.95    -0.05      0.05      0.23894737
2          Skeptical\nN(0, 0.10) 0.95    -0.05      0.05      0.09894737
3               Diffuse\nN(0, 2) 0.95    -0.05      0.05      0.06157895
   Decision
1 Undecided
2 Undecided
3 Undecided
```

3.4.5 Visualize Prior Sensitivity



3.4.6 When ROPE Conclusions Hold vs. Change

ROPE Conclusions Hold When:

Large sample size (n > 30 per group)

- Data overwhelms prior
- Posterior similar regardless of prior choice

Effect clearly outside or inside ROPE

- HDI far from ROPE boundaries
- All reasonable priors give same conclusion

Narrow posterior

- HDI width « ROPE width
- High precision reduces prior influence

ROPE Conclusions Are Sensitive When:

Small sample size (n < 20 per group) - Prior has substantial influence - Different priors can change conclusion

HDI barely touches ROPE boundary - Effect 0.05 (right at threshold) - Small prior differences flip decision

Wide posterior - HDI width   ROPE width - Uncertainty dominates

Example of sensitivity:

```
Data: N=15 subjects, effect  0.04 log-units
ROPE: [-0.05, 0.05]
Result:
  - Skeptical prior: 100% in ROPE (accept H )
  - Diffuse prior: 60% in ROPE (undecided)
→ Conclusion sensitive!
```

3.4.7 What To Do If Conclusions Are Sensitive

3.4.7.1 Option 1: Report Multiple Analyses (Recommended)

Report all results from different prior specifications:

3.4.7.2 Option 2: Justify Prior More Carefully

If conclusions are sensitive, invest more effort in prior justification:

- Pilot data
- Expert elicitation
- Previous literature
- Prior predictive checks (Module 02)

Report: "We used prior [X] because [justification]. Given potential sensitivity with small N, we verified conclusions hold with more diffuse prior."

3.4.7.3 Option 3: Collect More Data

If:

- Conclusions sensitive AND
- You cannot justify prior AND

- Decision matters

→ Collect more data before making decision

This is NOT Optional Stopping / P-Hacking!

In NHST: Looking at data, then deciding to collect more is a questionable research practice (QRP) - it inflates Type I error because your stopping rule affects the p-value.

In Bayesian analysis: You CAN look at the data and decide to collect more without invalidating inference!

Why? Bayesian inference follows the **likelihood principle**:

- The posterior depends ONLY on the data and prior
- NOT on your stopping intentions or whether you peeked
- Looking at N=15, seeing inconclusive results, and collecting to N=60 gives the SAME posterior as pre-planning N=60

The difference:

- **NHST QRP**: "Keep collecting until $p < 0.05$" (invalidates test)
- **Bayesian valid**: "Keep collecting until HDI is narrow enough for confident decision" (perfectly legitimate)

Caveat: You still shouldn't change your ROPE boundaries or priors after seeing results!

Planning how much more data you need:
Use this power analysis to estimate target sample size for adequate precision:

```
# Rule of thumb: HDI width inversely proportional to sqrt(n)
current_n <- 8        # Current sample size
current_hdi_width <- 0.10  # Current HDI width from preliminary analysis
desired_hdi_width <- 0.05  # Want HDI < half ROPE width for confident conclusions

# Calculate target N needed
target_n <- current_n * (current_hdi_width / desired_hdi_width)^2

tibble(
  Measure = c("Current N", "Current HDI width", "Desired HDI width", "Target N needed", "Additio
  Value = c(current_n, current_hdi_width, desired_hdi_width, round(target_n), round(target_n - c
)
```

Interpretation: To reduce HDI width from 0.10 to 0.05, you need to increase N by a factor of 4 (halving width requires quadrupling sample size).

3.4.7.4 Option 4: Be Transparent About Uncertainty

If data collection isn't possible:

```
"Given our sample size (N=15), ROPE conclusions were sensitive to prior
specification. With weakly informative priors, X% of posterior fell in
ROPE, suggesting [interpretation]. With more skeptical priors, Y% fell
in ROPE, suggesting [alternative interpretation]. We recommend treating
this finding as preliminary pending replication with larger sample."
```

3.4.8 Worked Example: Prior Sensitivity Reporting

```
# A tibble: 3 x 4
  Prior                              Estimate `95% HDI`       `% in ROPE`
  <chr>                                 <dbl> <chr>                 <dbl>
1 1. Weakly informative: N(0, 0.5)      0.037 [-0.307, 0.381]         0.2
2 2. Skeptical: N(0, 0.10)              0.223 [-0.194, 0.597]         0.1
3 3. Diffuse: N(0, 2)                   0.35  [-0.180, 0.850]         0.1
```

Conclusion: ROPE decision holds across prior specifications. All three priors yield: Undecided

Main Point

Prior sensitivity is not a bug—it's a feature!

It tells you when your data are: - **Strong enough** to overcome prior beliefs → Confident conclusions - **Too weak** to overcome prior beliefs (sensitive) → Need more data or transparency

NHST doesn't have this diagnostic—it can give you "p < 0.05" even with barely any data, hiding the fact that different assumptions would give different answers.

Bayesian analysis makes sensitivity **explicit and quantifiable**.

4 Effect Estimation with emmeans

Moving Beyond Single Comparisons

So far we've tested practical significance for a single effect (Condition B vs A). But what if you have **three or more groups**? You need:

- All pairwise comparisons (A vs B, A vs C, B vs C)
- ROPE analysis for each comparison
- Adjustment for multiple comparisons (optional)

This is where **emmeans** and **marginaleffects** come in.

4.1 Why emmeans?

When you have **factorial designs**, you often want to:

- Compare all pairwise combinations
- Estimate marginal means (averaged over random effects)
- Get automatic adjustment for multiple comparisons
- Use familiar syntax from frequentist stats

emmeans (estimated marginal means) provides all of this and works directly with brms!

4.2 Example: Three-Condition Design

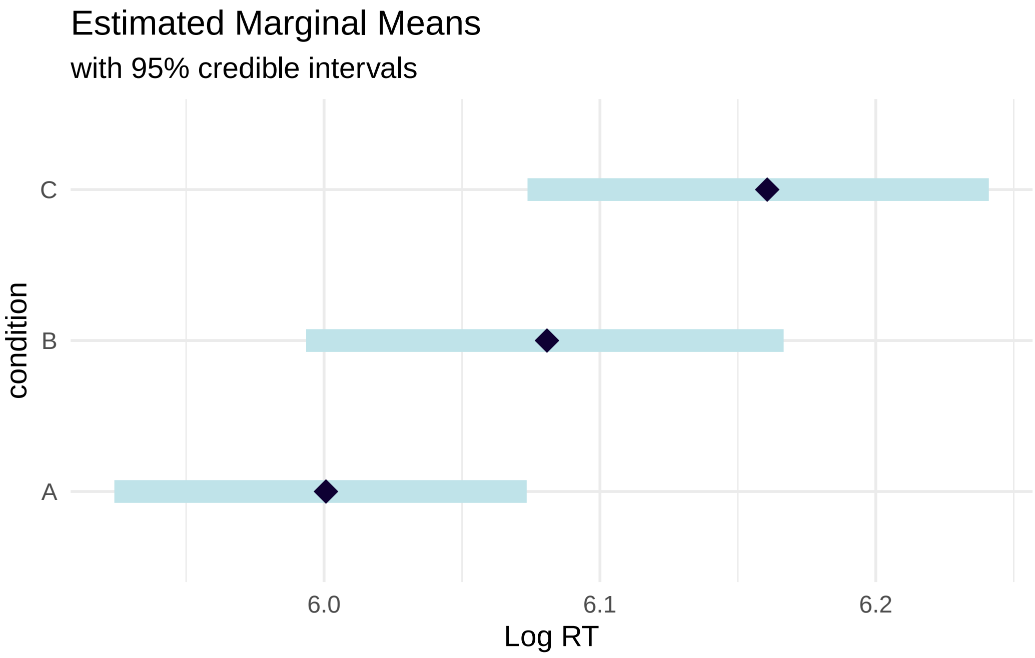
Let's extend our example to three conditions:

```
# A tibble: 3 x 5
  condition mean_rt sd_rt mean_log_rt sd_log_rt
  <fct>       <dbl> <dbl>       <dbl>     <dbl>
1 A            411.  78.8        6.00     0.193
2 B            448.  99.7        6.08     0.221
3 C            483. 102.         6.16     0.218
```

4.3 Estimated Marginal Means

Table 5: **Estimated Marginal Means (Log RT)**

Condition	Mean	95% HDI
A	6.001	[5.924, 6.073]
B	6.081	[5.993, 6.167]
C	6.161	[6.074, 6.241]



ℹ What are “Estimated Marginal Means”?

EMMs are model-predicted means that:

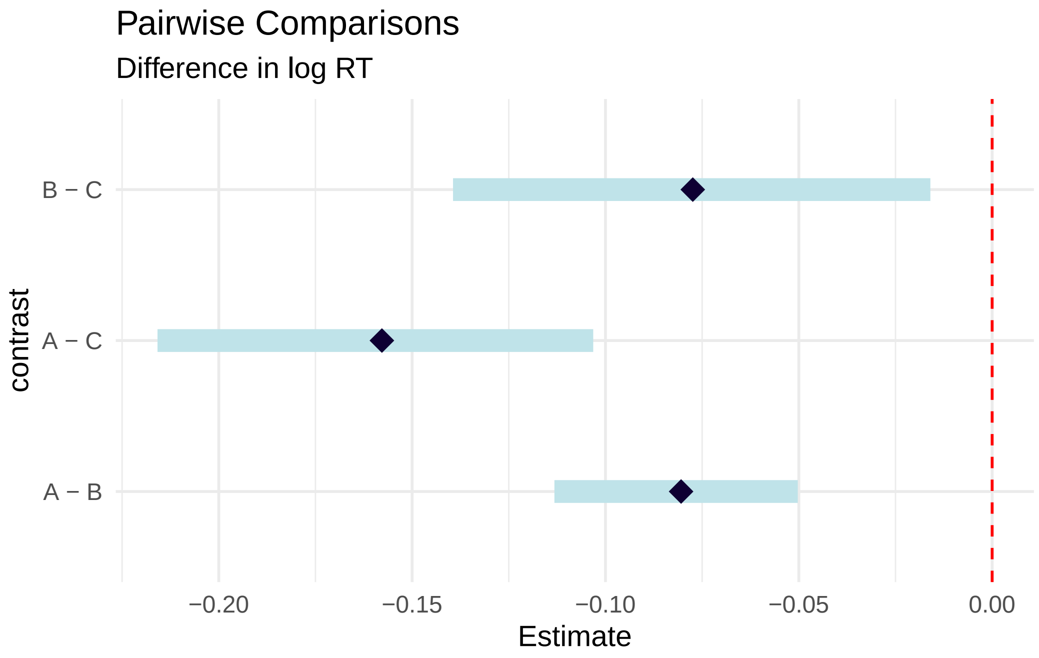
- Average over random effects (subjects, items)
- Provide population-level estimates
- Include full Bayesian uncertainty (not just point estimates!)

Think of them as “What would we expect for a typical new subject/item?”

4.4 Pairwise Comparisons

Table 6: **Pairwise Comparisons (Difference in Log RT)**

Contrast	Estimate	95% HPD
A - B	-0.080	[-0.113, -0.050]
A - C	-0.158	[-0.216, -0.103]
B - C	-0.077	[-0.139, -0.016]

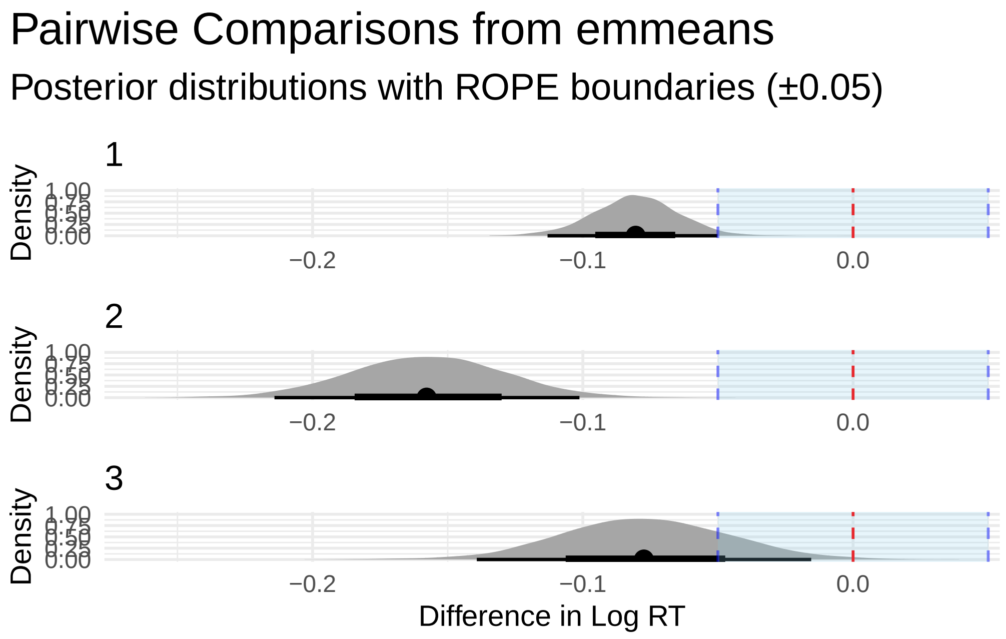


4.5 ROPE Analysis on Pairwise Comparisons

Now combine emmeans with ROPE to test practical significance of each comparison:

Table 7: **ROPE Analysis Results for Pairwise Comparisons**

Comparison	Estimate	95% HDI	Result
1	-0.080	[-0.113, -0.050]	↓ MEANINGFUL DECREASE
2	-0.158	[-0.216, -0.103]	↓ MEANINGFUL DECREASE
3	-0.077	[-0.139, -0.016]	? UNDECIDED



4.6 Custom Contrasts

emmeans allows custom contrasts beyond pairwise comparisons:

Table 8: **Custom Contrasts**

Contrast	Estimate	95% HDI
BC_vs_A	0.119	[0.086, 0.153]
C_vs_B	0.077	[0.016, 0.139]

Table 9: **ROPE Analysis for Custom Contrasts**

Contrast	Decision
1	Meaningful positive effect
2	Undecided

ℹ Summary: emmeans

You should now understand:

- Estimated marginal means (EMMs) are population-level predictions
- `pairs()` gives all pairwise comparisons automatically
- Extract posterior samples with `as.mcmc()` to integrate with ROPE
- Works directly with brms for full Bayesian inference

💡 When to Use emmeans

Perfect for:

- Factorial designs (2×2, 2×3, etc.)
- All pairwise comparisons automatically
- Familiar syntax from lsmeans/emmeans in frequentist stats
- Integration with ROPE for practical significance

Not ideal for:

- Simple two-group comparisons (just use brms coefficients)
- Complex non-linear predictions (use marginaleffects instead)
- Interactions with continuous predictors (marginaleffects better)

Effect Estimation with marginaleffects (Under Construction - Click to Expand)

⚠ Section Under Development

This section is currently being revised and some tables/visualizations may not display correctly. The emmeans approach (above) is fully functional and recommended for now.

5 Effect Estimation with marginaleffects

Modern Alternative to emmeans

While emmeans is excellent for factorial designs, **marginaleffects** offers:

- Unified syntax across ALL model types (not just brms)
- More flexible predictions and comparisons
- Better support for continuous predictors and interactions
- Modern tidyverse-compatible workflow

Let's see how it compares.

5.1 Why marginaleffects?

marginaleffects provides a modern, unified interface for:

- Predictions at specific values
- Comparisons (differences, ratios, etc.)
- Slopes (derivatives) for continuous predictors
- Custom hypotheses with flexible syntax

It works with brms, rstanarm, glm, lme4, and many other models!

5.2 Visualize 95% Credible Intervals

Two Approaches to Making Predictions

marginaleffects offers two main functions for predictions that differ in how they handle the dataset:

5.2.1 1. predictions() with datagrid()

```
predictions(rt_model_3, newdata = datagrid(condition = c("A", "B", "C")))
```

- Creates a **single reference grid** with specified conditions
- All other predictors held at their means/modes
- Random effects set to 0 (population-level only)
- Returns: "What would we predict for a hypothetical average unit?"

5.2.2 2. avg_predictions()

```
avg_predictions(rt_model_3, variables = "condition")
```

- Makes predictions for **every observation** in the original dataset
- Each prediction uses that observation's actual covariate values
- Includes subject-specific random effects
- Then **averages** these predictions across all observations
- Returns: "What is the average prediction across all units in our sample?"

Which matches the data generation better?

Our data was generated with:

- Baseline: 6.0

- Condition B: 6.0 + random subject slopes (mean 0.10)
- Condition C: 6.0 + random subject slopes (mean 0.18)

`avg_predictions()` better captures this because:

- It accounts for the empirical distribution of random subject intercepts and slopes
- It averages over actual subjects in the data
- True population means: A 6.0, B 6.1, C 6.18

How do these differ from emmeans?

Aspect	avg_predictions()	emmeans()	predictions() + datagrid()
Observations used	All rows in original data	All rows in original data	Single reference grid
Random effects	Actual fitted values per subject	Integrated over distribution	Set to 0
Interpretation	"Average of predictions"	"Estimated marginal mean"	"Conditional prediction at reference"
Best for	Descriptive summaries	Population-level inference	Scenario analysis

`avg_predictions()` and `emmeans()` should give **similar results** because both average over the actual data, but may differ slightly in:

- Computational method (averaging vs. marginalization)
- Treatment of random effects in the averaging process
- Default settings for transforms and scales

Table 11: **Average Predicted vs. Observed Log RT by Condition**

Condition	Observed Mean	Model Prediction	Difference	95% CI
A	5.999	5.999	0.000	[5.993, 6.006]
B	6.080	6.080	0.000	[6.074, 6.086]
C	6.157	6.157	0.000	[6.151, 6.163]

Model Fit Check

The table above shows how well the model predictions match the actual observed means in the data.

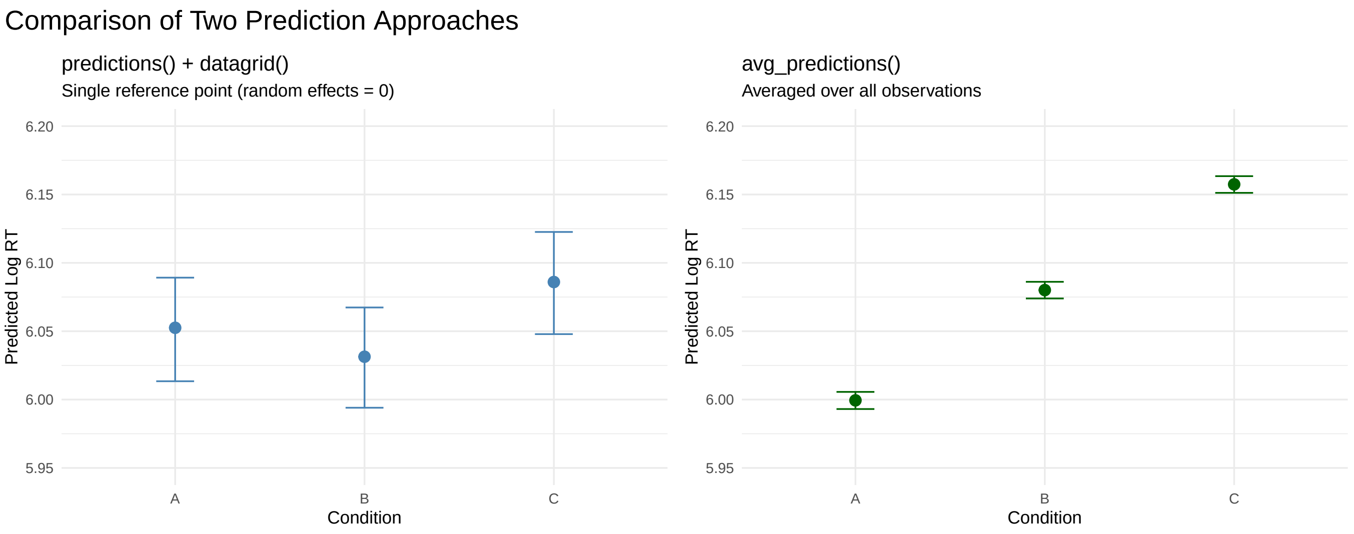
True data generation parameters:

- A: 6.00 (baseline)
- B: 6.10 (baseline + 0.10)
- C: 6.18 (baseline + 0.18)

The small differences between observed means and model predictions are due to:

1. **Random sampling variation** in data generation (random effects and residuals)
2. **Shrinkage from Bayesian priors** pulling estimates toward more conservative values
3. **Regularization** from the hierarchical structure preventing overfitting to sample means

This is working as intended! The model balances fitting the data with reasonable prior constraints.



Comparison of Two Prediction Approaches

predictions() + datagrid()
Single reference point (random effects = 0)

avg_predictions()
Averaged over all observations

Left: Conditional prediction at reference. Right: Average prediction across sample.

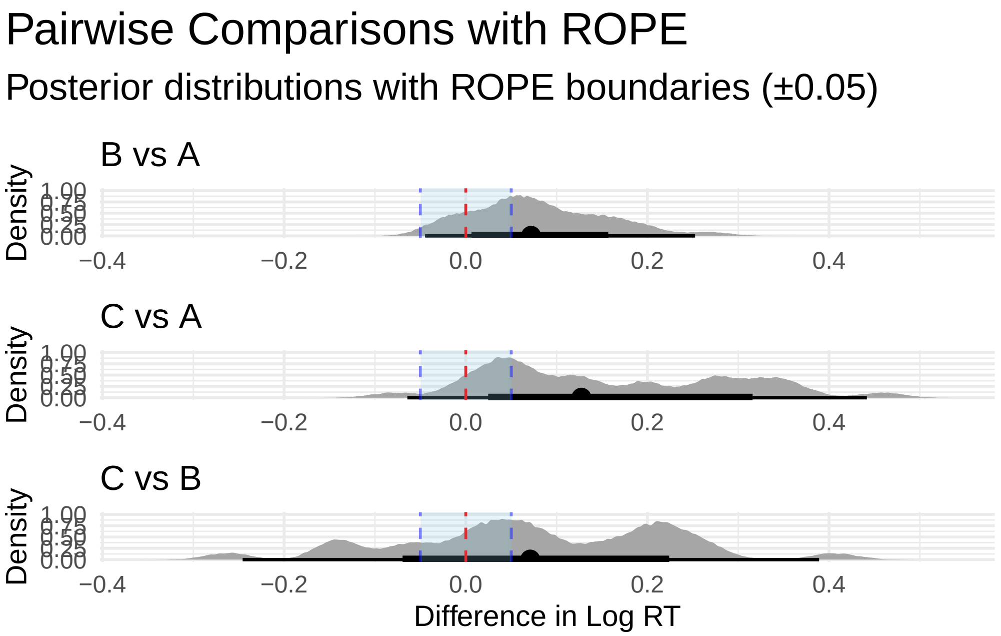
5.3 Raw Posterior Draws

ℹ Approach: Posterior Distribution Visualization

This section extracts the **full posterior distributions** from pairwise comparisons using `posterior_draws()`. This allows us to:

1. Visualize the complete uncertainty in each comparison
2. Add ROPE boundaries to assess practical significance
3. Calculate custom summary statistics (mean, HDI) from the raw posterior samples

Why this approach? When you need to visualize distributions, check overlap with ROPE, or compute custom summaries beyond what marginaleffects provides by default.



Pairwise Comparisons with ROPE

Posterior distributions with ROPE boundaries (±0.05)

B vs A

C vs A

C vs B

Table 12: **Pairwise Comparison Estimates**

Comparison	Estimate	95% HDI
B vs A	0.081	[-0.045, 0.252]
C vs A	0.158	[-0.064, 0.442]
C vs B	0.077	[-0.246, 0.389]

5.4 Using marginaleffects Summary Statistics (better)

Approach: Summary Statistics from marginaleffects

This section uses marginaleffects' **built-in summary statistics** directly from the comparison objects. This approach:

1. Uses pre-computed estimates and credible intervals from marginaleffects
2. Shows the "average" comparison across the dataset
3. Is more efficient when you only need point estimates and CIs

Why do estimates differ from Section 5.3?
The differences arise from **how posterior draws are summarized**:

5.4.1 Section 5.3: Manual Posterior Summarization

When we use `posterior_draws()` and then manually compute `mean(draw)`:

```
posterior_draws(comp_B_vs_A) |>
  summarise(Estimate = mean(draw))
```

This computes: **mean of all posterior draws from the comparison distribution**

- Each posterior draw represents: _B - _A from a single MCMC iteration
- We're averaging across all 4000 posterior samples
- This gives us the expected value of the comparison

5.4.2 Section 5.4: marginaleffects Default Summary

When we use `comp_B_vs_A` directly:

```
comp_B_vs_A %>% as.data.frame() %>% select(estimate)
```

marginaleffects computes the comparison **at each unit in newdata** first, then summarizes:

1. For each posterior draw: Make predictions at all covariate values
2. For each posterior draw: Compute comparison (B - A) at each covariate value

3. Average across units (if multiple rows in newdata)
4. Then summarize across posterior draws

Key difference: marginaleffects averages **within** each posterior draw before summarizing across draws. This is especially important with:

- **Non-linear models** (like logistic regression): Predictions depend on covariate values
- **Interactions**: Effect of condition may differ by subject characteristics
- **Random effects**: Unit-level predictions include random effect realizations

5.4.3 Comparison Direction

Both sections compute **B - A**, **C - A**, and **C - B** (second condition minus first):

```
comparisons(rt_model_3, variables = list(condition = c("A", "B")))
# Compares: B vs A (i.e., B - A)
```

The order in `list(condition = c("A", "B"))` matters: the second element is compared to the first.

5.4.4 When Do They Match?

The two approaches give **identical** results when:

- Linear model with no interactions
- All observations have identical covariate values
- No random effects (population-level only)

In mixed-effects models with random effects and subject-level covariates, the differences can be substantial because:

- Section 5.3: Averages raw posterior samples from the comparison distribution
- Section 5.4: Averages predictions across the **empirical distribution** of covariates in your data, then summarizes

5.4.5 Which Should You Use?

- **Section 5.3 (posterior_draws)**: When you want to visualize full distributions, check ROPE overlap, or compute custom summaries
- **Section 5.4 (direct comparison)**: When you want marginaleffects' default averaging method, which accounts for the distribution of covariates in your sample

For most mixed-effects models, **Section 5.4's approach** is preferred because it properly accounts for the empirical distribution of random effects and covariates.

Table 13: **All Pairwise Comparisons**

Comparison	Difference	95% CI
B vs A	-0.021	[-0.068, 0.026]
C vs A	0.033	[-0.014, 0.080]
C vs B	0.054	[0.007, 0.102]

⚠ Why No Visualization Here?

Unlike Section 5.3, we **don't show a three-panel graph** here because:

1. **The table shows marginaleffects summary statistics** (estimates averaged across the empirical distribution)
2. **Using `posterior_draws()` would show the same distributions as Section 5.3** (defeating the purpose of comparing approaches)
3. **The key difference is in the summary method, not the posterior distributions themselves**

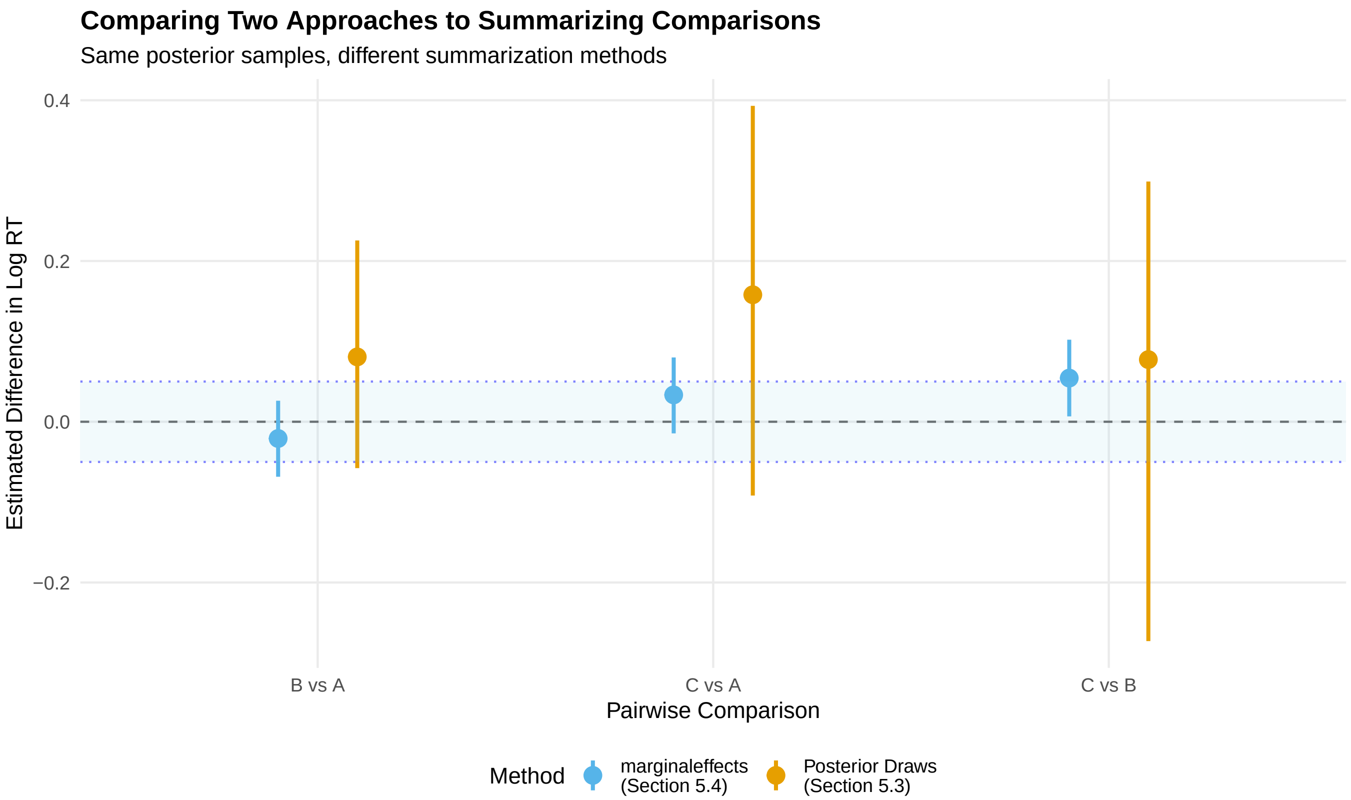
If we visualized `posterior_draws(comp_B_vs_A)` here, it would be **identical to Section 5.3** because both extract the same raw MCMC samples. The difference between sections 5.3 and 5.4 is:

- **Section 5.3**: Manually extracts posterior draws and computes `mean(draw)` → gives one type of summary
- **Section 5.4**: Uses marginaleffects' internal averaging over observations → gives a different summary

The **posterior distributions are the same**; only the **summarization differs**. The table above shows marginaleffects' summary, which accounts for averaging predictions across all covariate values in the dataset before summarizing across posterior draws.

To visualize the marginaleffects approach properly, you would need to show prediction distributions at individual covariate values, then average them—which is complex and not typically done. The table is the appropriate summary for this approach.

However, we can visualize how the **summary statistics differ** between the two approaches:



This visualization shows why the two approaches give different estimates:

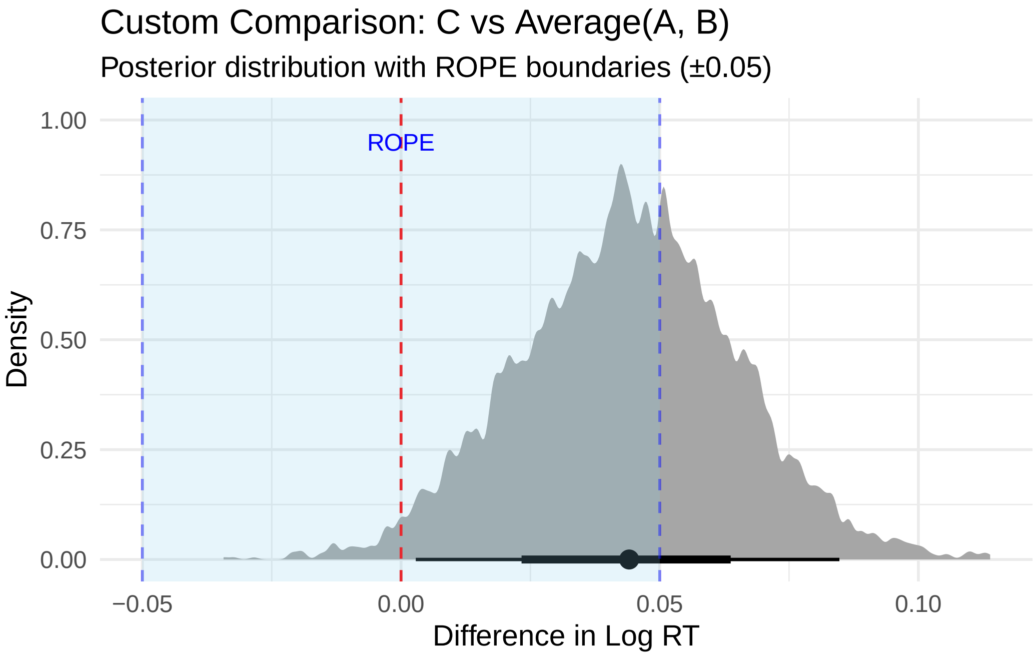
- **Orange (Section 5.3)**: Posterior draws approach—computes `mean(posterior_draws)`
- **Blue (Section 5.4)**: marginaleffects summary—averages predictions at unit level before summarizing

The key insight: both methods use the **same posterior samples**, but summarize them differently. Section 5.3's approach tends to give larger estimates because it doesn't average over the empirical distribution of covariates and random effects in the same way.

5.5 Custom Comparisons Example

Table 14: **Custom Comparison: C vs Average(A, B)**

Measure	Value
Estimate	0.044
95% HDI	[0.003, 0.085]
% Below ROPE	0.0%
% In ROPE	61.1%
% Above ROPE	38.9%



5.5.1 Decision

Undecided - need more data

6 ROPE Reporting

❗ What to Always Report

From Kruschke (2015, p. 338):

> "Reporting the limits of an HDI region is more informative than reporting the declaration of a reject/accept decision. By reporting the HDI and other summary information about the posterior, different readers can apply different ROPEs to decide for themselves whether a parameter is practically equivalent to a null value."

Complete reporting includes:

1. Full posterior summary (not just inside/outside ROPE)
2. ROPE boundaries with justification
3. Effect size on multiple scales

4. Model diagnostics
5. Sensitivity checks

6.0.1 Methods Section Template

Include in Methods:

- Prior specification with rationale
- ROPE boundaries with justification
- When boundaries were set (a priori)
- How boundaries relate to original scale
- Sample size (subjects, items, observations)
- Software versions

6.0.2 Results Section Template

Include in Results:

- Point estimate with HDI (on model scale)
- Posterior SD or uncertainty measure
- % of posterior in/outside ROPE
- Decision statement (reject/accept/undecided H)
- Effect size on original scale with interpretation
- Absolute magnitudes (e.g., milliseconds) where relevant
- Robustness checks (prior sensitivity, model comparison)

7 When to Use What: Decision Framework

Quick Decision Tree

What do you want to test?
→ “Is the effect big enough to matter?” → Use ROPE
→ “Which hypothesis is better supported?” → See Module 07 (Bayes Factors)
→ “Which model fits better?” → Use LOO (Module 03)

7.1 Visual Guide: Choosing Your Tools

```
Your Research Question

   → "Is effect meaningful?"       → ROPE

   → "Compare 3+ groups?"
        → Factorial design        → emmeans + ROPE
        → Custom predictions       → marginaleffects + ROPE

   → "Which hypothesis better?"     → Module 07 (Bayes Factors)

   → "Which model structure?"       → Module 05 (LOO)
```

8 Avoiding Pitfalls: Checklist

Before running analyses:

- Priors are **weakly informative** (not flat)?

- ROPE boundaries defined **before** seeing results?
- ROPE boundaries **justified** with domain knowledge?
- Hypotheses based on **theory**, not data exploration?

When interpreting results:

- Report **ROPE decision** and effect sizes?
- Report **uncertainty** (don't hide ROPE overlaps)?
- Check **prior sensitivity** (Module 04)?
- Show **posterior distributions visually**?
- Interpret on the **original scale** when possible?

9 Quick Reference

Main papers:

- **Kruschke, J. K. (2018).** Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270-280. [The definitive ROPE paper]
- **Kruschke, J. K. (2015).** *Doing Bayesian data analysis* (2nd ed.). Academic Press. [Chapters 11-12]
- **Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013).** *Bayesian data analysis* (3rd ed.). CRC Press. [Chapter 9: Decision Analysis]
- **Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021).** Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2), 667-718. [ESS diagnostics]
- **Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019).** bayestestR: Describing effects and their uncertainty. *Journal of Open Source Software*, 4(40), 1541.
- **Lakens, D., Scheel, A. M., & Isager, P. M. (2018).** Equivalence testing for psychological research. *Advances in Methods and Practices in Psychological Science*, 1(2), 259-269.

9.1 Session Info

```
R version 4.5.2 (2025-10-31)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 24.04.3 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.26.so;  LAPACK version 3.12.0

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

time zone: Etc/UTC
tzcode source: system (glibc)

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] knitr_1.51             patchwork_1.3.2        marginaleffects_0.31.0
 [4] emmeans_2.0.1          bayestestR_0.17.0      tidybayes_3.0.7
 [7] posterior_1.6.1.9000   bayesplot_1.15.0       lubridate_1.9.4
[10] forcats_1.0.1          stringr_1.5.2          dplyr_1.1.4
[13] purrr_1.1.0            readr_2.1.5            tidyr_1.3.2
[16] tibble_3.3.0           ggplot2_4.0.1          tidyverse_2.0.0
[19] brms_2.23.0            Rcpp_1.1.0

loaded via a namespace (and not attached):
 [1] svUnit_1.0.8          tidyselect_1.2.1      farver_2.1.2
 [4] loo_2.9.0             S7_0.2.0              fastmap_1.2.0
 [7] tensorA_0.36.2.1      digest_0.6.37         timechange_0.3.0
[10] estimability_1.5.1    lifecycle_1.0.4       StanHeaders_2.32.10
[13] processx_3.8.6        magrittr_2.0.4        compiler_4.5.2
[16] rlang_1.1.6           tools_4.5.2           utf8_1.2.6
[19] yaml_2.3.10           collapse_2.1.5        data.table_1.17.8
[22] labeling_0.4.3        bridgesampling_1.2-1  pkgbuild_1.4.8
[25] plyr_1.8.9            RColorBrewer_1.1-3    cmdstanr_0.9.0
[28] abind_1.4-8           withr_3.0.2           datawizard_1.3.0
[31] grid_4.5.2            stats4_4.5.2          xtable_1.8-4
[34] inline_0.3.21         ggridges_0.5.7        scales_1.4.0
[37] tinytex_0.57          insight_1.4.4         cli_3.6.5
[40] mvtnorm_1.3-3         rmarkdown_2.30        generics_0.1.4
[43] otel_0.2.0            RcppParallel_5.1.11-1 reshape2_1.4.5
[46] tzdb_0.5.0            rstan_2.32.7          HDInterval_0.2.4
[49] parallel_4.5.2        matrixStats_1.5.0     vctrs_0.6.5
[52] Matrix_1.7-4          jsonlite_2.0.0        hms_1.1.4
[55] arrayhelpers_1.1-0    ggdist_3.3.3          see_0.12.0
[58] glue_1.8.0            codetools_0.2-20      ps_1.9.1
[61] distributional_0.5.0  stringi_1.8.7         gtable_0.3.6
[64] QuickJSR_1.8.1        pillar_1.11.1         htmltools_0.5.8.1
[67] Brobdingnag_1.2-9     R6_2.6.1              evaluate_1.0.5
[70] lattice_0.22-7        backports_1.5.0       rstantools_2.5.0
[73] coda_0.19-4.1         gridExtra_2.3         nlme_3.1-168
[76] checkmate_2.3.3       xfun_0.55             pkgconfig_2.0.3
```