A gentle introduction to Bayesian estimation
Day 5:
More priors, simulation-based calibration & Bayes factors

Suzanne Hoogeveen

Practicalities

- Housing keys picked up between 9-10

Today

- Informative prior specification (original program)
- Simulation-based calibration (new)
- Hypothesis testing with Bayes factors (new)

- Afternoon: showcase your skills!

Prior advice

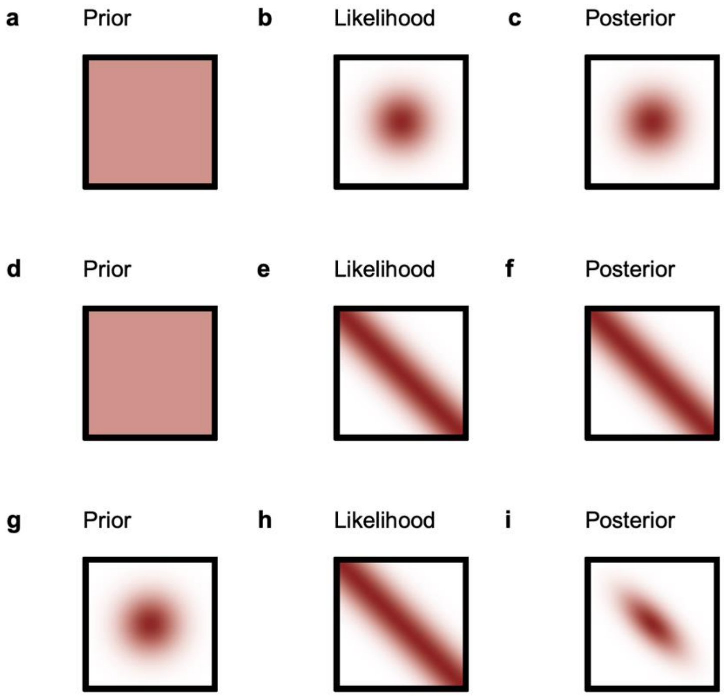
Strong advice:

1. Think about your priors! Whatever settings you choose, justify them.

Prior advice

Strong advice:

1. Think about your priors! Whatever settings you choose, justify them.



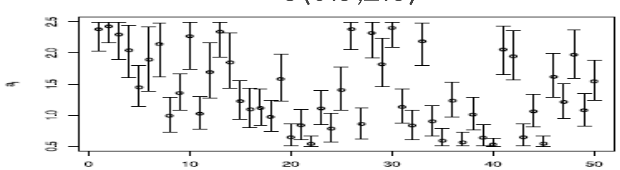
When the data provide good information via the likelihood (b), the posterior is sufficiently concentrated (c), even with a flat prior (a). However, when the data is not informative enough (e,h), a weakly-informative prior (g) is needed to help constrain the posterior (i)

Prior advice

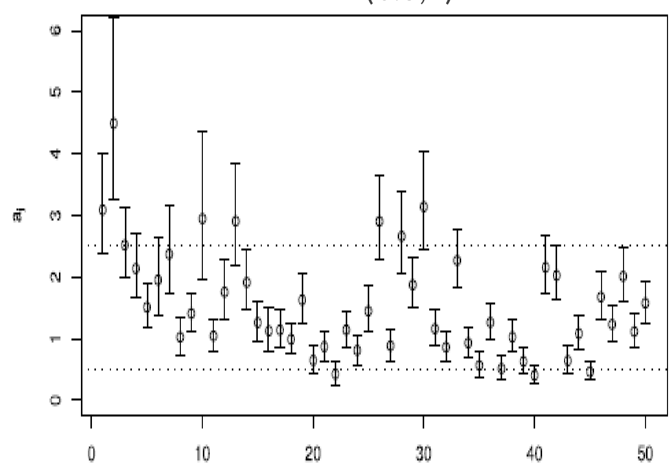
Strong advice:

1. Think about your priors! Whatever settings you choose, justify them.
2. Don't use uniform priors: they seem uninformative, but because they have fixed bounds, they can be very influential if they are (accidentally) too narrow. Normal priors with a large variance and/or bounds are often better choices.

U(0.5,2.5)



LN(0.5,1)



Veen, D., & Klugkist, I. (2019). 10.1016/j.jkss.2019.07.004

Prior advice

Strong advice:

1. Think about your priors! Whatever settings you choose, justify them.
2. Don't use uniform priors: they seem uninformative, but because they have fixed bounds, they can be very influential if they are (accidentally) too narrow. Normal priors with a large variance and/or bounds are often better choices.
3. Conduct prior predictive checks to make sure that the combination of priors leads to reasonable predictions.

Prior advice

Strong advice:

1. Think about your priors! Whatever settings you choose, justify them.
2. Don't use uniform priors: they seem uninformative, but because they have fixed bounds, they can be very influential if they are (accidentally) too narrow. Normal priors with a large variance and/or bounds are often better choices.
3. Conduct prior predictive checks to make sure that the combination of priors leads to reasonable predictions.
4. Conduct sensitivity analyses with alternative priors to assess the robustness of your results.

Prior advice

Weaker advice:

1. Don't use inverse-gamma priors on the variance; this is an historical choice due to conjugacy, but not necessary in modern implementations in Stan/brms (and unintuitive). The Stan team recommends using half-normal or half-Cauchy priors instead.
   – in brms, you don't have to worry about negative variance priors, because it automatically restricts variance parameters to have a lower bound of zero.

Prior advice

Weaker advice:

1. Don't use inverse-gamma priors on the variance; this is an historical choice due to conjugacy, but not necessary in modern implementations in Stan/brms (and unintuitive). The Stan team recommends using half-normal or half-Cauchy priors instead.
   - in brms, you don't have to worry about negative variance priors, because it automatically restricts variance parameters to have a lower bound of zero.
2. Don't use vague priors, such as N(0, 1000). These can lead to numerical problems in the estimation. Instead, use weakly informative priors that are centered around zero and have a reasonable range of values.

Prior advice

3. Consider standardizing the data if the scale of the parameters is either very large (e.g., 2000 milliseconds → 2 seconds) or very small. Values around 0 with a scale of 1 are often more stable in the algorithms.

Prior advice

3. Consider standardizing the data if the scale of the parameters is either very large (e.g., 2000 milliseconds → 2 seconds) or very small. Values around 0 with a scale of 1 are often more stable in the algorithms.
4. If you want to elicit priors from experts, use an established protocol, such as the MATCH protocol or the 5-step procedure (Veen et al., 2017). You can also use the shiny-app: https://utrecht-university.shinyapps.io/elicitation/

Veen et al. (2017). 10.3389/fpsyg.2017.02110

Simulation-based calibration

A full workflow for robust Bayesian inference focuses on the following steps:

1. assessing adequacy of priors (**prior predictive checks**): do the priors lead to reasonable predictions?

Simulation-based calibration

A full workflow for robust Bayesian inference focuses on the following steps:

1. assessing adequacy of priors (**prior predictive checks**): do the priors lead to reasonable predictions?
2. assessing computational faithfulness (through **simulation-based calibration**): can the model recover the parameters that were used to generate the data?

Simulation-based calibration

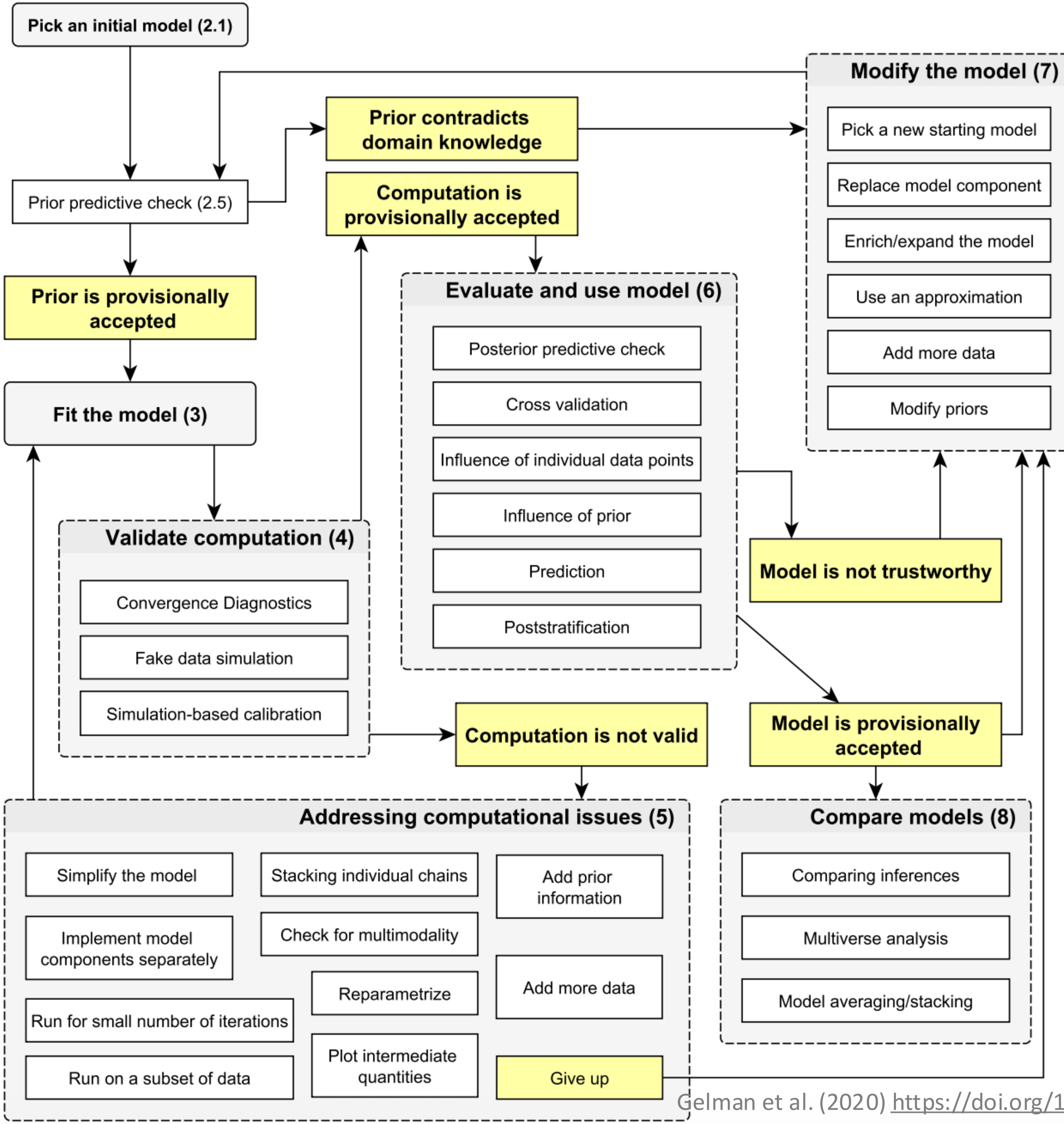
A full workflow for robust Bayesian inference focuses on the following steps:

1. assessing adequacy of priors (**prior predictive checks**): do the priors lead to reasonable predictions?
2. assessing computational faithfulness (through **simulation-based calibration**): can the model recover the parameters that were used to generate the data?
3. assessing model sensitivity: can the model return **unbiased estimates and effectively update prior beliefs** (i.e., can the model learn from data)?

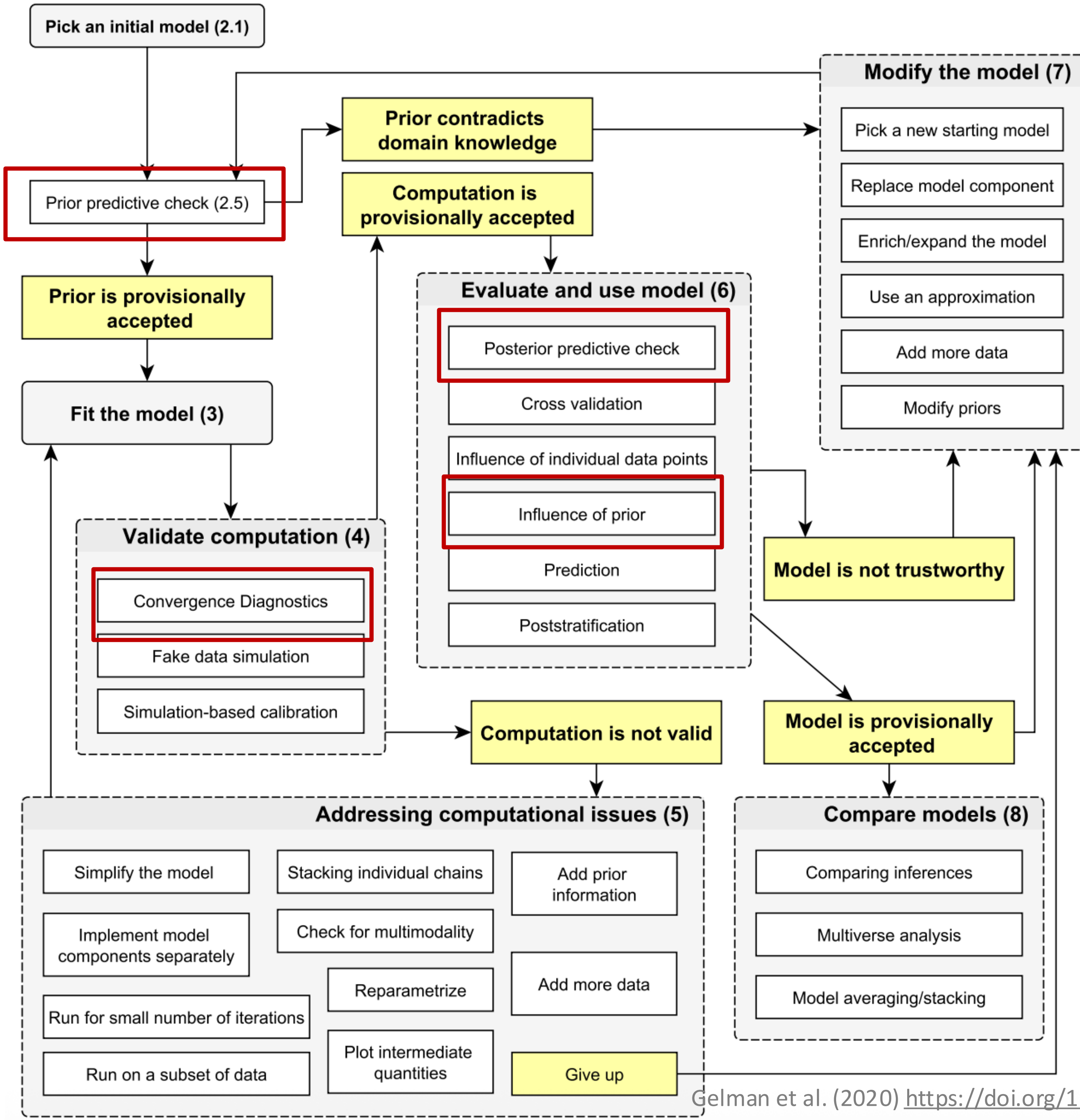
Simulation-based calibration

A full workflow for robust Bayesian inference focuses on the following steps:

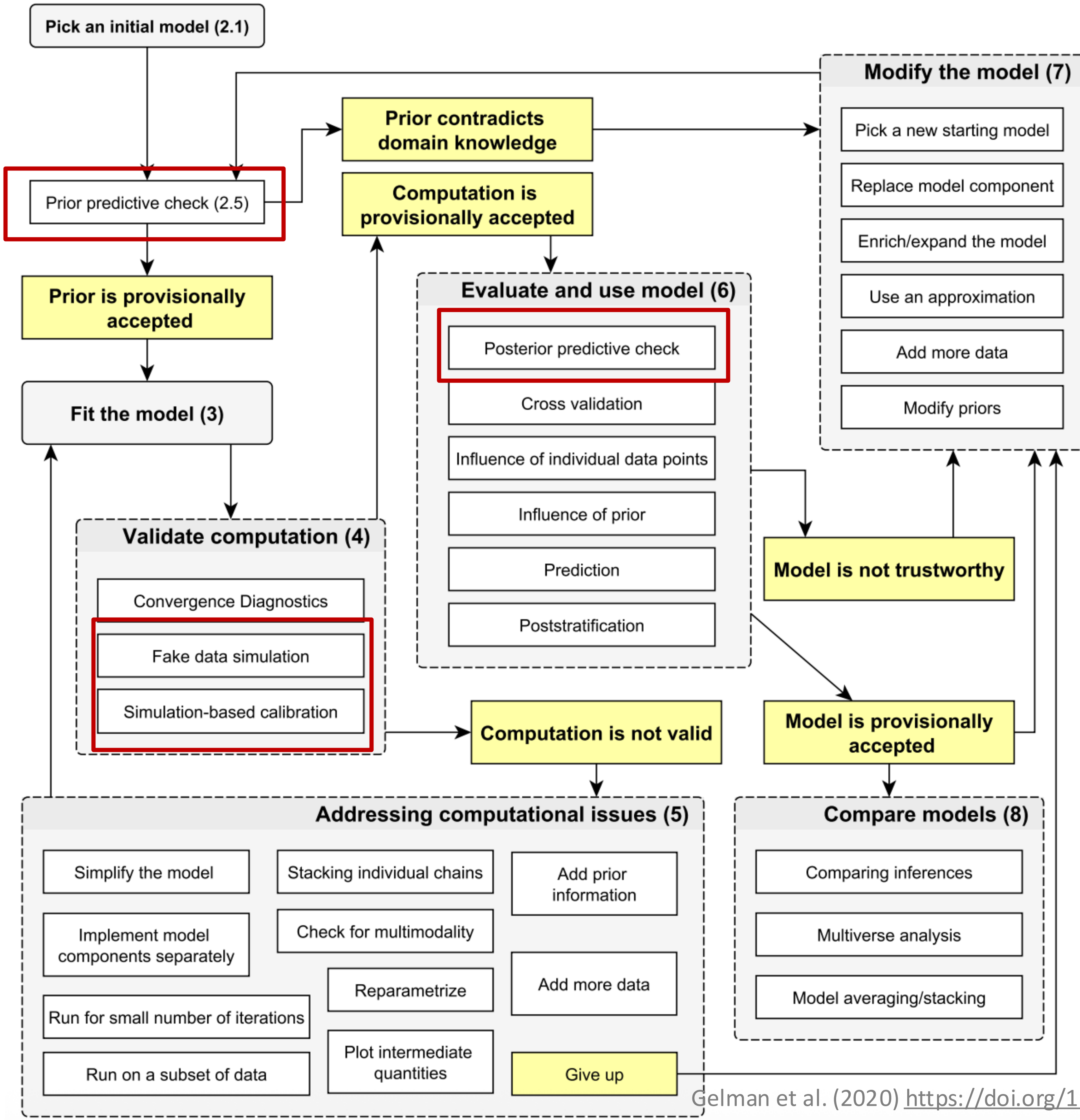
1. assessing adequacy of priors (**prior predictive checks**): do the priors lead to reasonable predictions?
2. assessing computational faithfulness (through **simulation-based calibration**): can the model recover the parameters that were used to generate the data?
3. assessing model sensitivity: can the model return **unbiased estimates and effectively update prior beliefs** (i.e., can the model learn from data)?
4. assessing adequacy of posteriors: **posterior predictive checks**: do the posterior estimates reflect reasonable predictions?



Gelman et al. (2020) https://doi.org/10.48550/arXiv.2011.(



Gelman et al. (2020) https://doi.org/10.48550/arXiv.2011.



Gelman et al. (2020) https://doi.org/10.48550/arXiv.2011.(

Model validation

We've talked about ways to check:

- Sensibility of the priors (*prior predictive checks*)
- Reliability of the sampling procedure and posterior estimates (*convergence diagnostics*)
- Sensibility of the model's predictions and model fit (*posterior predictive checks*)

Another aspect we may want to know is:

- How reliable and sensitive is the model + computational method?
  - → Can we trust the posterior inference; is it a good model?

Model validation

A good Bayesian model should:

1. be able to return parameters that the data was simulated from:
    - if we know the ground truth, because we generated data from known settings, we can validate if the model is able to converge to these ‘true’ values.

Model validation

A good Bayesian model should:

1. be able to return parameters that the data was simulated from:
   - if we know the ground truth, because we generated data from known settings, we can validate if the model is able to converge to these 'true' values.
2. give unbiased estimates:
   - not systematically over- or underestimate parameters (given the priors)

Model validation

A good Bayesian model should:

1. be able to return parameters that the data was simulated from:
   - if we know the ground truth, because we generated data from known settings, we can validate if the model is able to converge to these ‘true’ values.
2. give unbiased estimates:
   - not systematically over- or underestimate parameters (given the priors)
3. effectively learn from data:
   - posteriors should be more certain (i.e., more peaked) than priors

→ Check with **simulation-based calibration (SBC)**

Simulation-based calibration

- Idea: if a model is *computationally faithful*, it will be able to return unbiased estimates with appropriate uncertainty.

- We can assess this through simulation, because then we know the ground truth (e.g., $\text{®}_1 = 0.5$)

- When the model is computational faithful, it should be able to recover the prior distribution accurately.

Simulation-based calibration

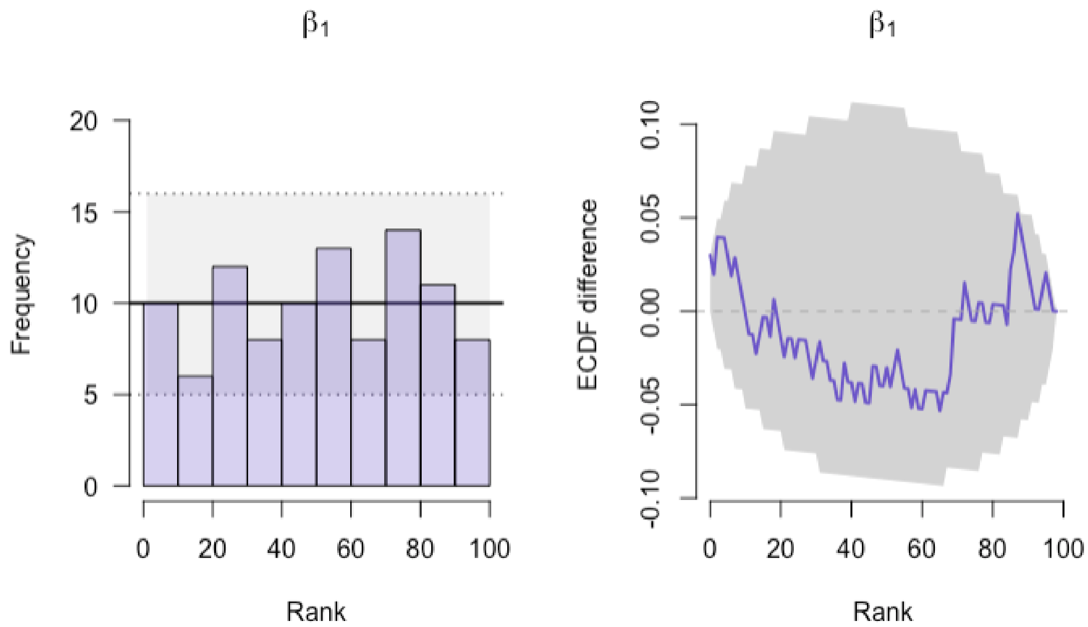
Steps:

1. Take the prior $\pi(\theta)$ and randomly draw a parameter set $\tilde{\theta}$ from it: $\tilde{\theta} \sim \pi(\theta)$
2. Use this parameter set $\tilde{\theta}$ to simulate hypothetical data $\tilde{y}$ from the model: $\tilde{y} \sim \pi(y|\tilde{\theta})$
3. Fit the model to this hypothetical data and draw samples from the posterior distribution: $\tilde{\theta}' \sim \pi(\theta|\tilde{y})$
4. Find the **rank** of the true parameter θ within the posterior samples $\tilde{\theta}'$ (that is, the count of posterior samples smaller than the generating parameter value).

Repeat steps 1-4, say, 100 times

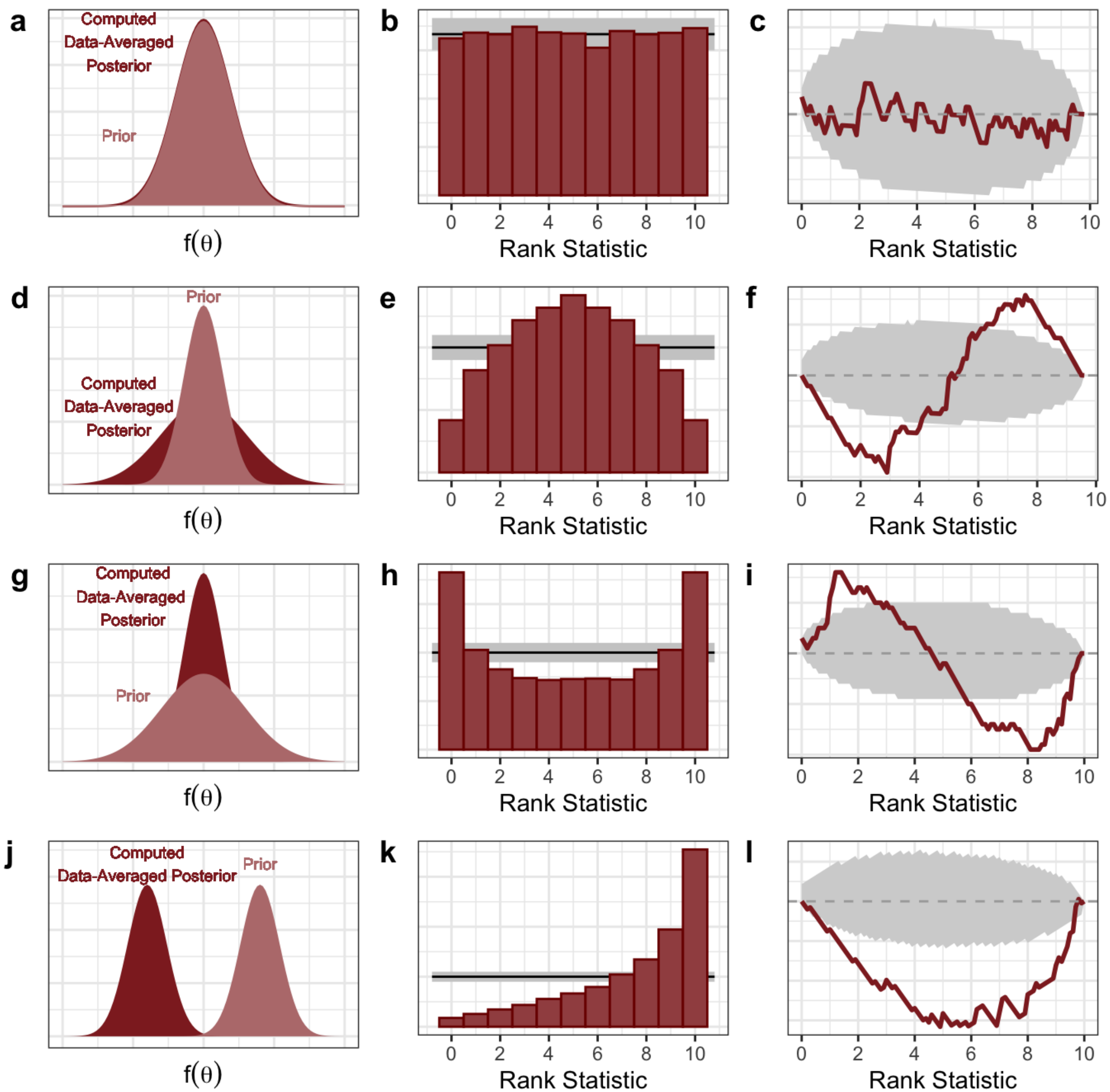
If the model is computationally faithful, every rank should occur equally often → we expect a uniform distribution of the ranks

Simulation-based calibration



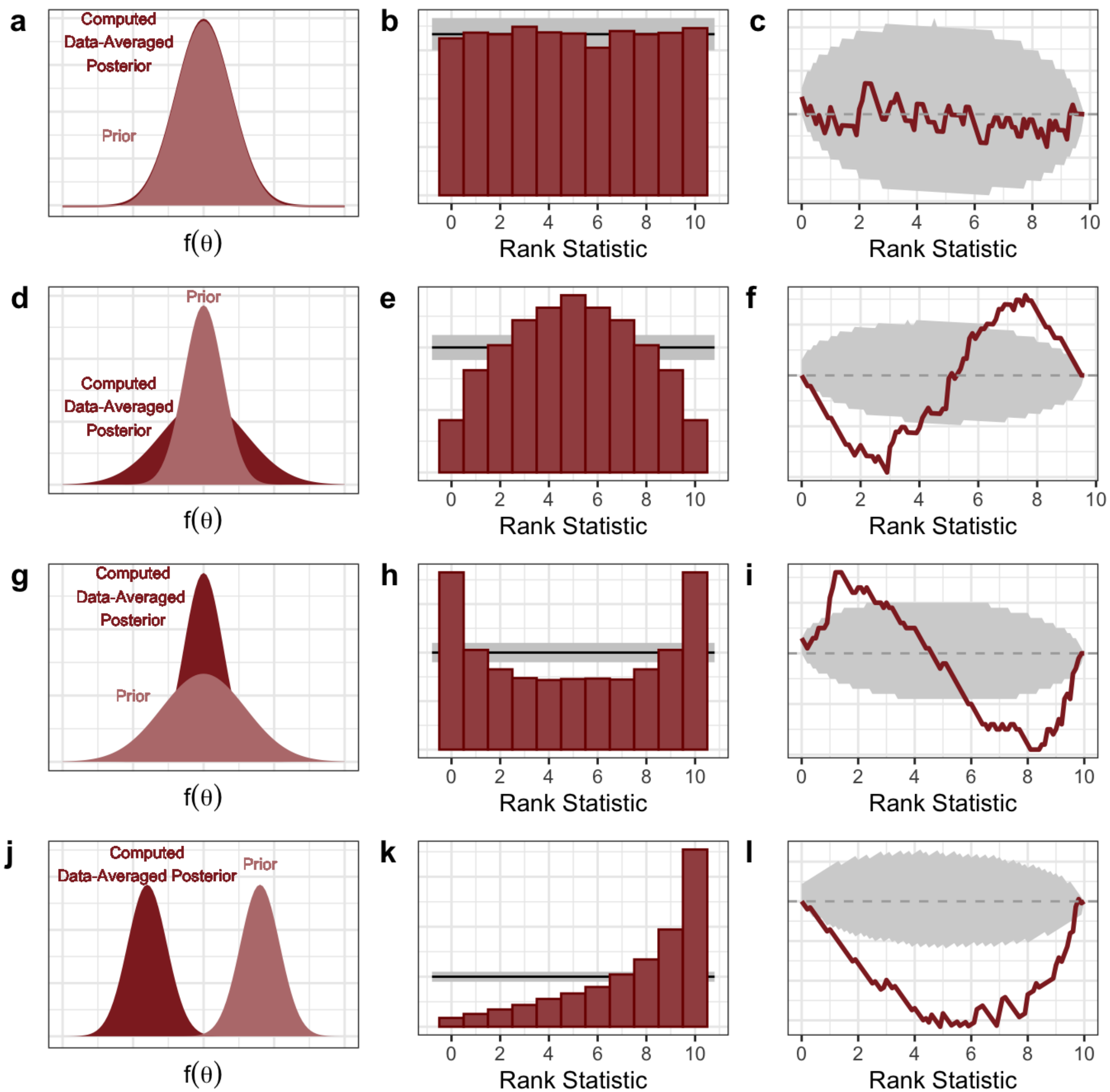
- Downside of the histogram: depends on number of bins
- Alternative: **empirical cumulative distribution function** (ECDF) of the ranks, and more specifically, the difference between the perfectly uniform CDF and the empirical CDF of the ranks, including the 95% interval of expected deviations.

Simulation-based calibration



Plots can not only show *if* something is wrong, but also give an indication of *how* it is wrong

Simulation-based calibration



Plots can not only show *if* something is wrong, but also give an indication of *how* it is wrong

a,b,c) Model well-calibrated
d,e,f) Model too uncertain
g,h,i) Model too certain
j,k,l) Model underestimates

Model sensitivity

For model sensitivity, we assess:

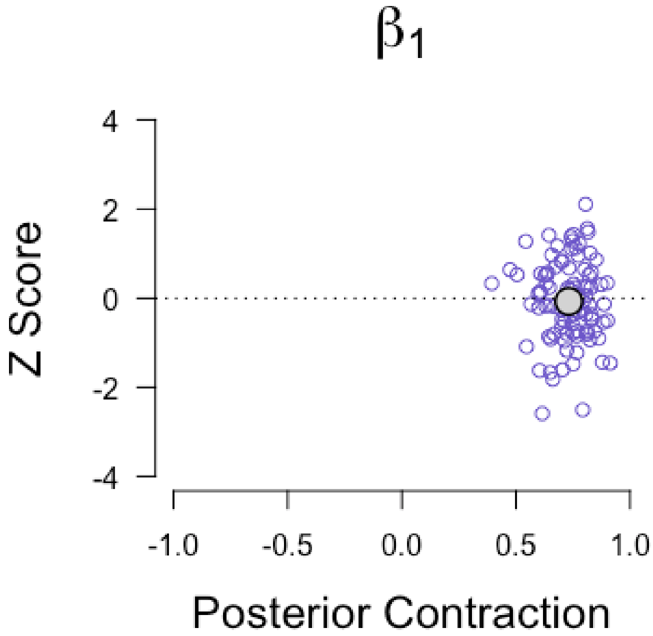
1. Are the (mean) posterior estimates unbiased?
   - How different is the mean posterior from the mean prior value (in each simulation)?
   - We don’t want a prior-likelihood mismatch (→ bias)

Model sensitivity

For model sensitivity, we assess:

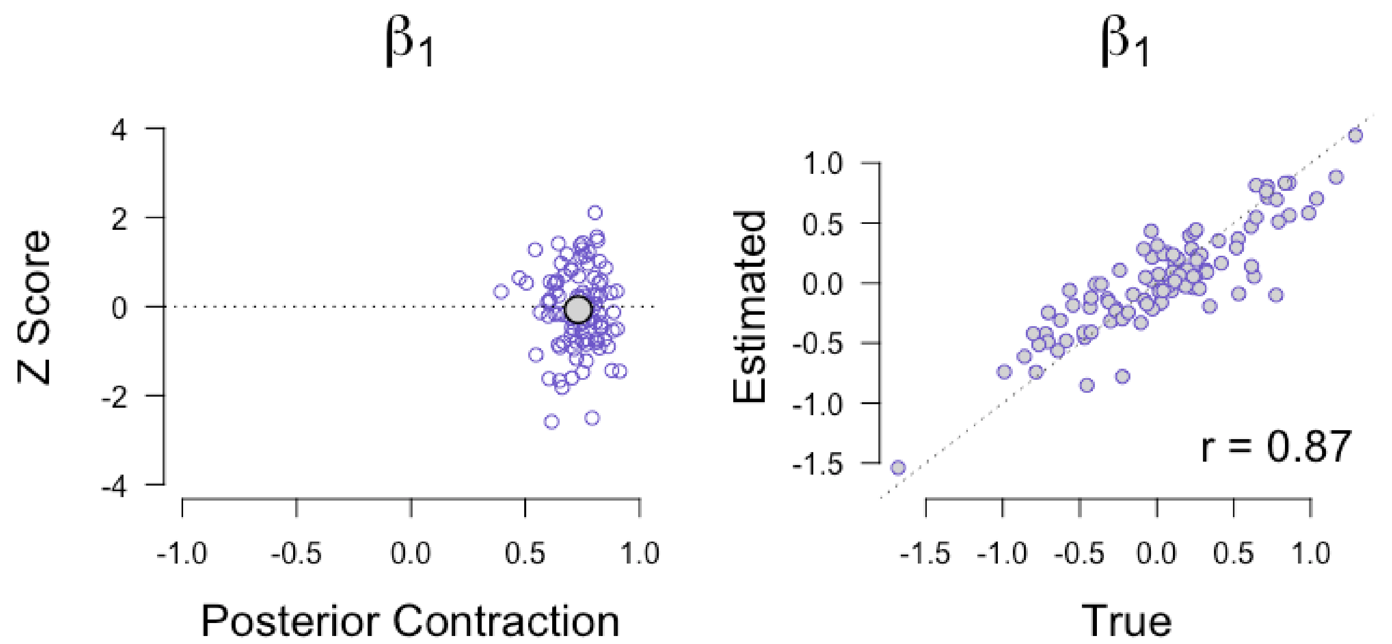
1. Are the (mean) posterior estimates unbiased?
   - How different is the mean posterior from the mean prior value (in each simulation)?
   - We don't want a prior-likelihood mismatch (→ bias)
2. Does the model learn from data? (i.e., is there posterior contraction?)
   - Is the posterior uncertainty substantially lower than the prior uncertainty?
   - In context of number of observations and model complexity
   - Posterior contraction: 1 – (var(posterior) / var(prior))
   - 0: no updating, 0.5: 50% more certain, 0.99: 99% more certain

Model sensitivity



- Z-scores (y-axis) clustering around zero: model returns unbiased estimates. Posterior contraction (x-axis) around 0.7: satisfactory updating of model parameters.

Model sensitivity



- Z-scores (y-axis) clustering around zero: model returns unbiased estimates. Posterior contraction (x-axis) around 0.7: satisfactory updating of model parameters.
- Correlation true (x-axis) and estimated (y-axis) parameter values is 0.87: good recovery of the model parameters.

Simulation-based calibration

- In conclusion: by simulating data from the prior distribution + likelihood, we can evaluate how well-calibrated our model is.

- We want:
  - Uniform ranks / ECDFs → global posterior distribution similar to prior distribution
  - Z-scores of difference between mean posterior estimates and mean prior estimates close to zero → no bias in estimates
  - Posterior contraction close to 1 → posterior uncertainty (per simulation) much smaller than prior uncertainty; model can learn from data

Simulation-based calibration

- If things go wrong, we know it has to do either with the specification of the model, the sampling algorithm or the connection between them.

- Potential problems:
  - Mismatch between data-generating model and (statistical) model
  - Problem in the algorithm (e.g., convergence, suboptimal non-MCMC methods)
  - Incorrect implementation (e.g., error in Stancode; unlikely with brms)
- Hard to debug, but at least you know there is a problem!

Hypothesis testing

- Hypothesis testing with Bayes:
  - Does the credible interval of the posterior include zero?
  - Savage-Dickey density ratio test
  - Bayes factor model comparison with *bridgesampling*

Hypothesis testing

- Hypothesis testing with Bayes:
  - Does the credible interval of the posterior include zero?
  - Savage-Dickey density ratio test
  - Bayes factor model comparison with *bridgesampling*

- The latter two involve the **Bayes factor (BF)** as the measure of evidence in the data for one hypothesis/model versus another.

- BF_{12} = probability of the data given hypothesis 1 versus the probability of the data given hypothesis 2

Bayes factor

- Remember Bayes' rule:

$$p(\theta \mid y) = \frac{p(\theta) \times p(y \mid \theta)}{p(y)}.$$

- This can be rewritten as:

$$\underbrace{p(\theta \mid y)}_{\substack{\text{Posterior for } \theta\text{:} \\ \text{new beliefs}}} = \underbrace{p(\theta)}_{\substack{\text{Prior for } \theta\text{:} \\ \text{old beliefs}}} \times \underbrace{\frac{p(y \mid \theta)}{p(y)}}_{\substack{\text{Relative predictive} \\ \text{adequacy for } \theta}}.$$

- Meaning: the posterior for theta given the data = prior for theta x the likelihood (probability of the data given theta) / prior probability of the data

Bayes factor

$$\underbrace{p(\theta \mid y)}_{\substack{\text{Posterior for } \theta: \\ \text{new beliefs}}} = \underbrace{p(\theta)}_{\substack{\text{Prior for } \theta: \\ \text{old beliefs}}} \times \underbrace{\frac{p(y \mid \theta)}{p(y)}}_{\substack{\text{Relative predictive} \\ \text{adequacy for } \theta}} .$$

- We can also use this formula to compare two hypotheses/models

$$\underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\substack{\text{Posterior uncertainty} \\ \text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\substack{\text{Prior uncertainty} \\ \text{about hypotheses}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\substack{\text{Predictive} \\ \text{updating factor}}} .$$

- The predictive updating factor

  = the ratio of marginal likelihoods

  = probability of the data under H_1 vs H_0

  = the level of evidence in the data for H_1 vs H_0

  = the Bayes factor

Extrasensory perception

- Example: Bem's (in)famous experiment (based on Heck et al. (2023))
  - n = 40 persons guess which of two cards hides an erotic picture (or the number 7)
  - Bem's ESP hypothesis: "precognitive detection of erotic stimuli."
  - Data: x = 26, that is, 26 out of 40 people selected the erotic card

Heck et al. (2023). https://doi.org/10.1037/met0000454

Extrasensory perception

- Example: Bem's (in)famous experiment (based on Heck et al. (2023))
  - n = 40 persons guess which of two cards hides an erotic picture (or the number 7)
  - Bem's ESP hypothesis: "precognitive detection of erotic stimuli."
  - Data: x = 26, that is, 26 out of 40 people selected the erotic card

  - Different competing models:
    - M_1 : x ~ Binomial(n = 40, θ = .50) → ESP does not exist, random guessing
    - M_2: x ~ Binomial(n = 40, θ ≠ .50) → ESP does exist

    - Frequentist: $\hat{\theta}$ = 26/40 = .65 with a confidence interval of [.48, .79], p = .081

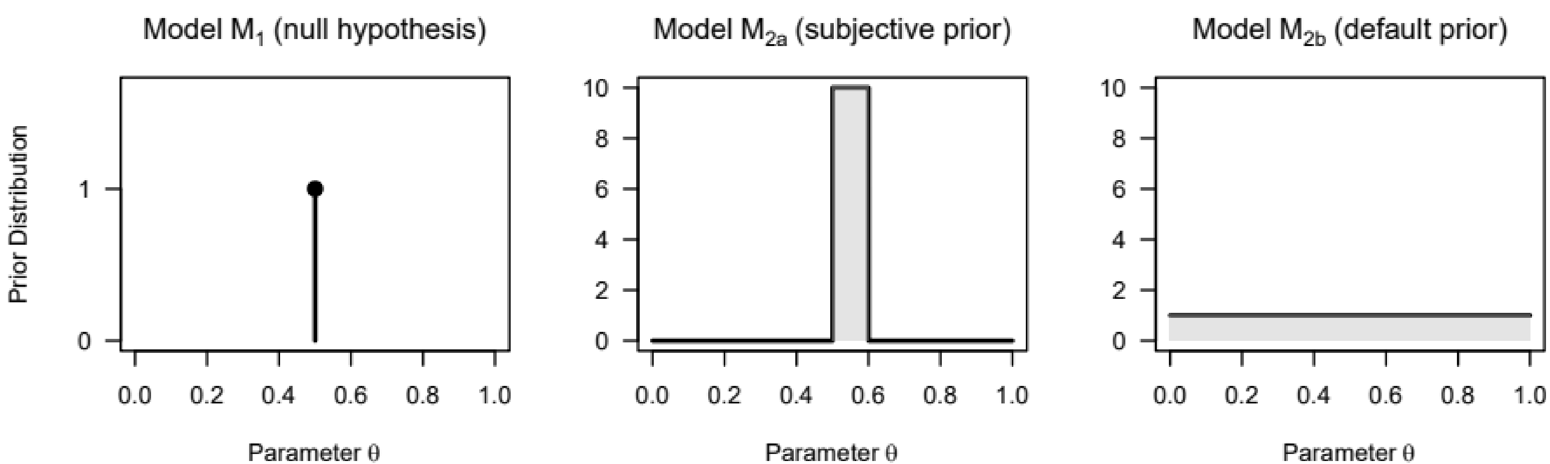
Heck et al. (2023). https://doi.org/10.1037/met0000454

Extrasensory perception

- In the Bayesian setting: we need priors for θ
  - M_1: belief: $\theta = .50$ → prior: spike at $\theta = .50$
  - M_2: belief $\theta \neq .50$ → prior?
    - M_{2a}: subjective → $\theta \sim$ Uniform(0.5, 0.6) *(ESP is weak but real)*
    - M_{2b}: default → $\theta \sim$ Uniform(0, 1) *(ignorant; let the data speak)*

Extrasensory perception

- In the Bayesian setting: we need priors for θ
  - M_1: belief: θ = .50 → prior: spike at θ = .50
  - M_2: belief θ ≠ .50 → prior?
    - M_{2a}: subjective → θ ~ Uniform(0.5, 0.6) *(ESP is weak but real)*
    - M_{2b}: default → θ ~ Uniform(0, 1) *(ignorant; let the data speak)*

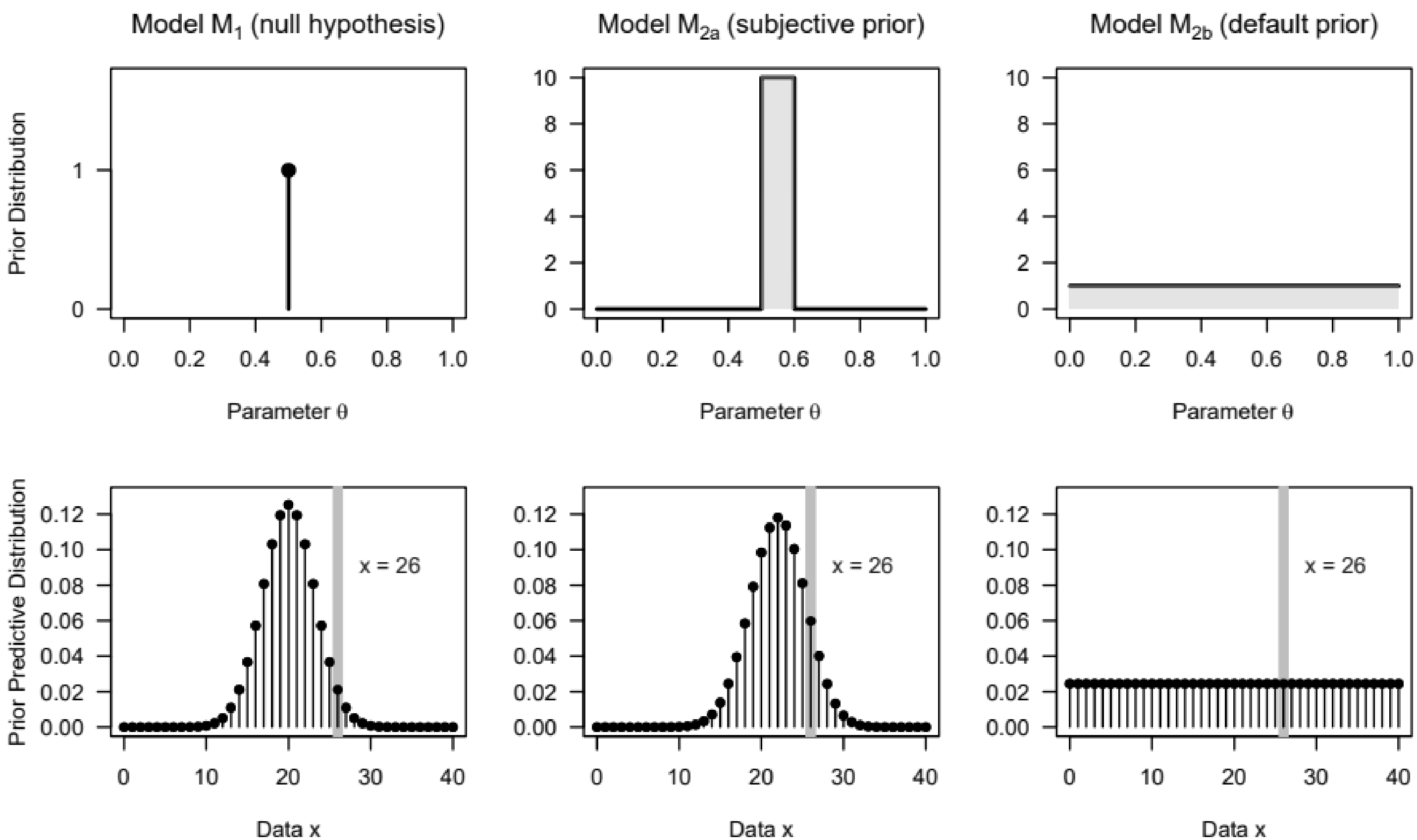


Extrasensory perception

- Now we can assess predictions from each model (before having seen the data)
- This *prior predictive distribution* provides the probability of observing a specific number of successes (x = 0, x = 1, . . . , x = 40) conditional on a model and prior.

Extrasensory perception

- Now we can assess predictions from each model (before having seen the data)
- This *prior predictive distribution* provides the probability of observing a specific number of successes (x = 0, x = 1, . . . , x = 40) conditional on a model and prior.



ESP Bayes factors

- From the prior predictive distribution, we can directly obtain the marginal likelihood of the observed data given each model.
- The marginal likelihood $P(x = 26 \mid M)$: probability of observing x = 26 "correct" guesses out of n = 40 trials given a specific model M with some prior distribution.

ESP Bayes factors

- From the prior predictive distribution, we can directly obtain the marginal likelihood of the observed data given each model.
- The marginal likelihood *P*(x = 26 | M): probability of observing x = 26 "correct" guesses out of n = 40 trials given a specific model M with some prior distribution.
- The Bayes factor compares how well two models predict the observed data; it is the ratio of the marginal likelihoods of the data for two models:

$$\mathrm{BF}_{1,2a} = \frac{P(x = 26 \mid \mathcal{M}_1)}{P(x = 26 \mid \mathcal{M}_{2a})},$$

Note: $BF_{2a,1} = 1/BF_{1,2a}$

- **Interpretation**:
  - BF > 1: More support for M_1
  - BF < 1: More support for M_{2a}

ESP Bayes factors

Here we get:

- $BF_{2a,1}$ = 2.83 → data of 26/40 “correct” is about 3 times more likely under the ESP exists but is weak model (M_{2a}) than under the ESP does not exist model (M_1)
- $BF_{2b,1}$ = 1.16 → about equal support in the data for no ESP (M_1) and no expectation (M_{2b})

ESP Bayes factors

Here we get:

- $BF_{2a,1}$ = 2.83 → data of 26/40 "correct" is about 3 times more likely under the ESP exists but is weak model (M_{2a}) than under the ESP does not exist model (M_1)
- $BF_{2b,1}$ = 1.16 → about equal support in the data for no ESP (M_1) and no expectation (M_{2b})

  - Notice the effect of 'vague' prior: vague predictions may hurt the chances of finding evidence for an effect.
  - In general, the Bayes factor penalizes complex models (e.g., models with many parameters or vague priors) if the increase in complexity does not pay off in terms of a better fit → optimal trade-off between model fit and complexity (cf. Occam's razor)

ESP Bayes factors

But we're forgetting one part of the equation:

$$\underbrace{\frac{P(\mathcal{M}_1 \mid x = 26, n = 40)}{P(\mathcal{M}_{2a} \mid x = 26, n = 40)}}_{\text{Posterior model odds}} = \underbrace{\text{BF}_{1,2a}}_{\text{Bayes factor}} \times \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_{2a})}}_{\text{Prior model odds}} .$$

- Bayes factor quantifies how to update our beliefs in light of the data, but is *independent* from the prior beliefs.
- Depending our prior beliefs about the two models, the *posterior* model probabilities may be different!

ESP Bayes factors

But we're forgetting one part of the equation:

$$\underbrace{\frac{P(\mathcal{M}_1 \mid x = 26, n = 40)}{P(\mathcal{M}_{2a} \mid x = 26, n = 40)}}_{\text{Posterior model odds}} = \underbrace{\mathrm{BF}_{1,2a}}_{\text{Bayes factor}} \times \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_{2a})}}_{\text{Prior model odds}} .$$

- Bayes factor quantifies how to update our beliefs in light of the data, but is *independent* from the prior beliefs.
- Depending our prior beliefs about the two models, the *posterior* model probabilities may be different!
- **Basically: our initial beliefs should not influence the evidence in the data, but they can influence our posterior beliefs.**

ESP Bayes factors

- Often, the default of equal prior model probabilities is used:
  - $P(M_1) = P(M_{2a}) = ½$
  - These translate into:
    - $P(M_1 \mid x = 26, n = 40) = .26$
    - $P(M_{2a} \mid x = 26, n = 40) = 1 - .26 = .74$

ESP Bayes factors

- Often, the default of equal prior model probabilities is used:
  - $P(M_1) = P(M_{2a}) = ½$
  - These translate into:
    - $P(M_1 \mid x = 26, n = 40) = .26$
    - $P(M_{2a} \mid x = 26, n = 40) = 1 - .26 = .74$
- But as with priors, we can also use *subjective prior model probabilities*, such as:
  - $P(M_1) = .90$
  - $P(M_{2a}) = .10$ → reflecting a priori scepticism for the existence of extrasensory perception (of erotic stimuli)
- This means that we need a lot of evidence in the data to shift our belief to the conviction that ESP exists.

Posterior ESP beliefs

- With $P(M_1) = .90$ and $P(M_{2a}) = .10$, we get:
  - $P(M_1 \mid x = 26, n = 40) = .74$
  - $P(M_{2a} \mid x = 26, n = 40) = .26$

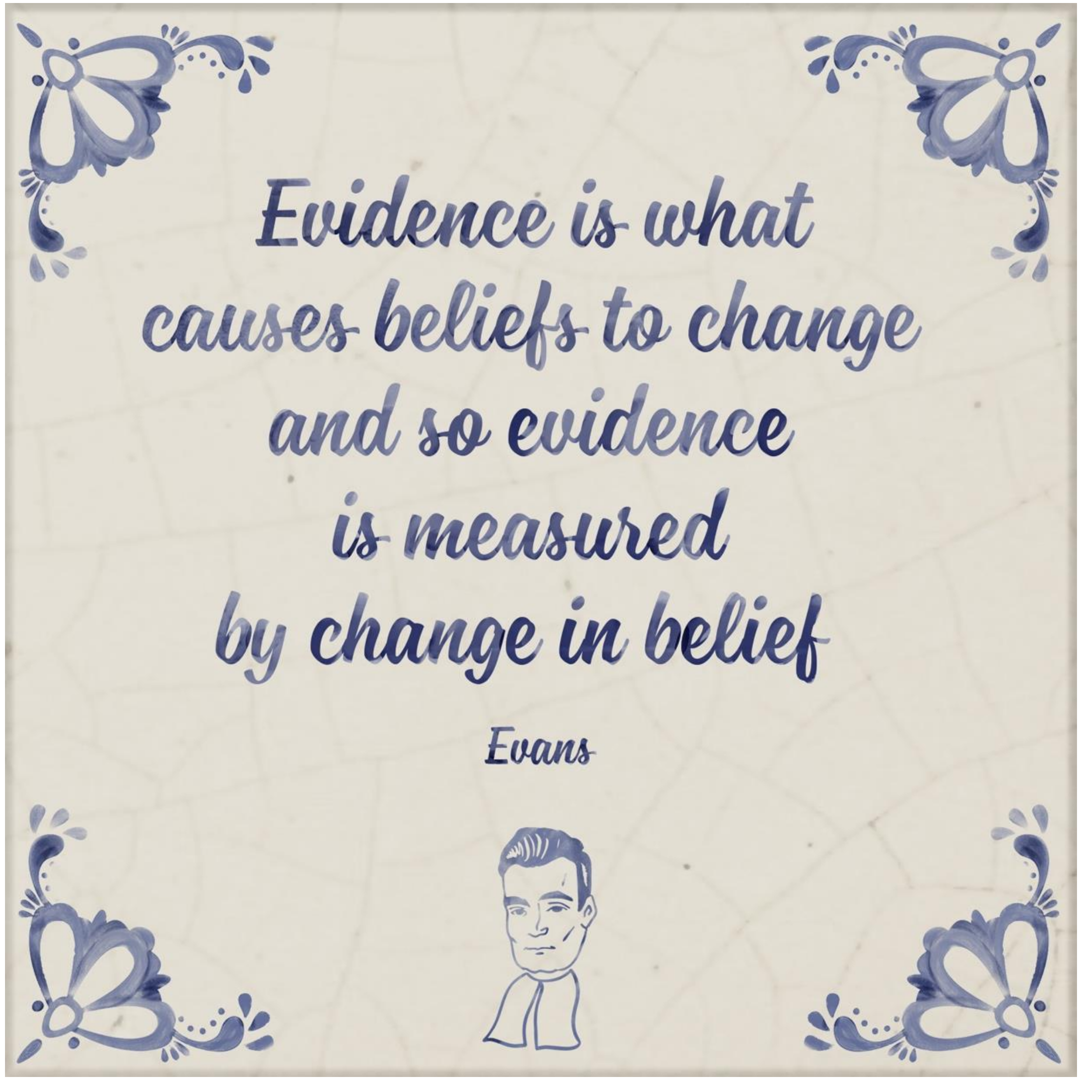
- So: the data are about 3 times more likely under the weak-but-existent-ESP model versus the no-ESP model
- However, the Bayesian framework allows us to include beliefs about the model's a priori plausibility

Posterior ESP beliefs

- With $P(M_1) = .90$ and $P(M_{2a}) = .10$, we get:
  - $P(M_1 \mid x = 26, n = 40) = .74$
  - $P(M_{2a} \mid x = 26, n = 40) = .26$

- So: the data are about 3 times more likely under the weak-but-existent-ESP model versus the no-ESP model
- However, the Bayesian framework allows us to include beliefs about the model's a priori plausibility

- This means that given (a) our initial scepticism and (b) the not-overwhelming evidence, we may update our beliefs in ESP from 1:9 odds to 1:3 odds, but still remain (rationally) unconvinced.


Evidence is what
causes beliefs to change
and so evidence
is measured
by change in belief
Evans

Bayes factors in complex models

- So Bayes factors are great, but how do we get them in more complex models?

Bayes factors in complex models

- So Bayes factors are great, but how do we get them in more complex models?
- Two options:
  - Savage-Dickey density ratio: posterior density at the point of interest divided by the prior density at that point.
    - Benefit: easy to compute, requires no additional computation
    - Downside: only for single parameters

Bayes factors in complex models

- So Bayes factors are great, but how do we get them in more complex models?
- Two options:
  - Savage-Dickey density ratio: posterior density at the point of interest divided by the prior density at that point.
    - Benefit: easy to compute, requires no additional computation
    - Downside: only for single parameters
  - Model comparison: ratio of marginal likelihoods of two models, using *bridgesampling*. Favors well-fitting models, but penalizes complexity (cf. Occam's razor)
    - Benefit: very flexible, also for multiple parameters (e.g., random effects)
    - Downside: requires many iterations (more than estimation)

Bayes factors in complex models

- Important practical considerations:
  - No improper / flat priors
  - Save all parameters when fitting the model (in brms: `save_pars = save_pars(all = TRUE)`) to keep the log-marginal-likelihood needed for bridgesampling
  - Use many iterations (~10 times more than for estimation)

Example: afterlife beliefs model

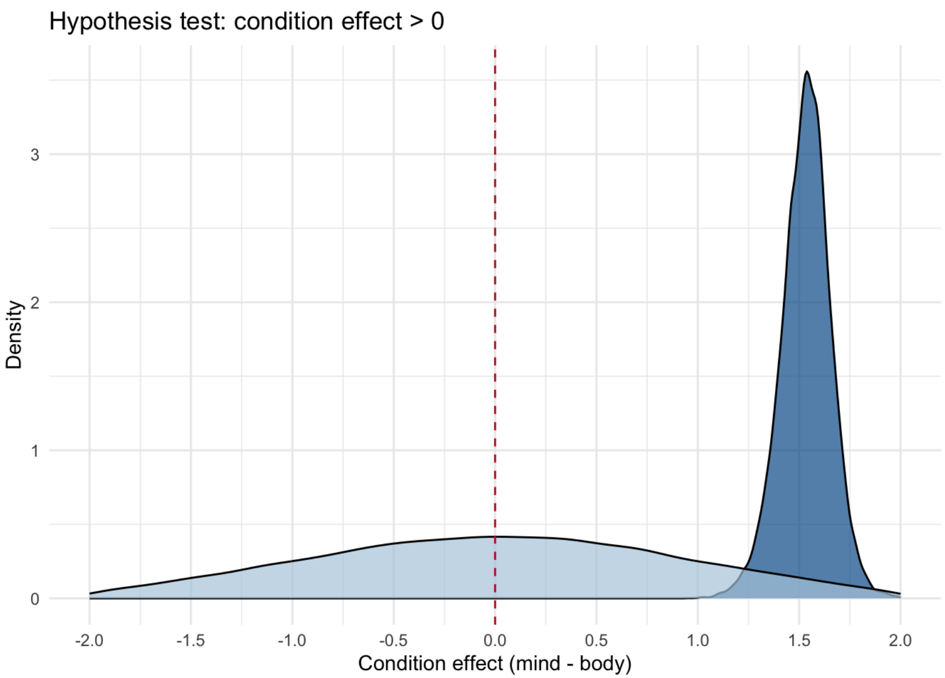
Consider: H_1: continuity judgments after biological death are more likely for mental states than bodily states

- Typically, if you want to test against a null-hypothesis, you would use a weakly informative prior centered around zero. Here we use a N(0,1) prior for the condition effect.

Savage-Dickey density ratio

Consider: H_1: continuity judgments after biological death are more likely for mental states than bodily states

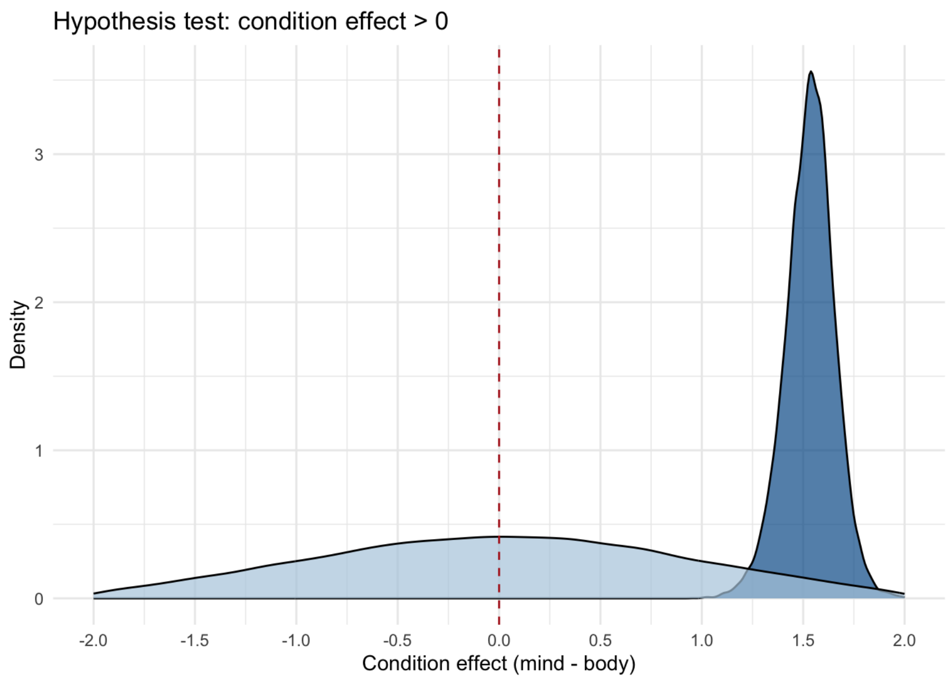
```
Hypothesis Tests for class b:
  Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio Post.Prob Star
1  (cat) > 0     1.53      0.12     1.33     1.73        Inf         1    *
---
```



Savage-Dickey density ratio

Consider: H_1: continuity judgments after biological death are more likely for mental states than bodily states

```
Hypothesis Tests for class b:
  Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio Post.Prob Star
1  (cat) > 0     1.53      0.12     1.33     1.73        Inf         1    *
---
```

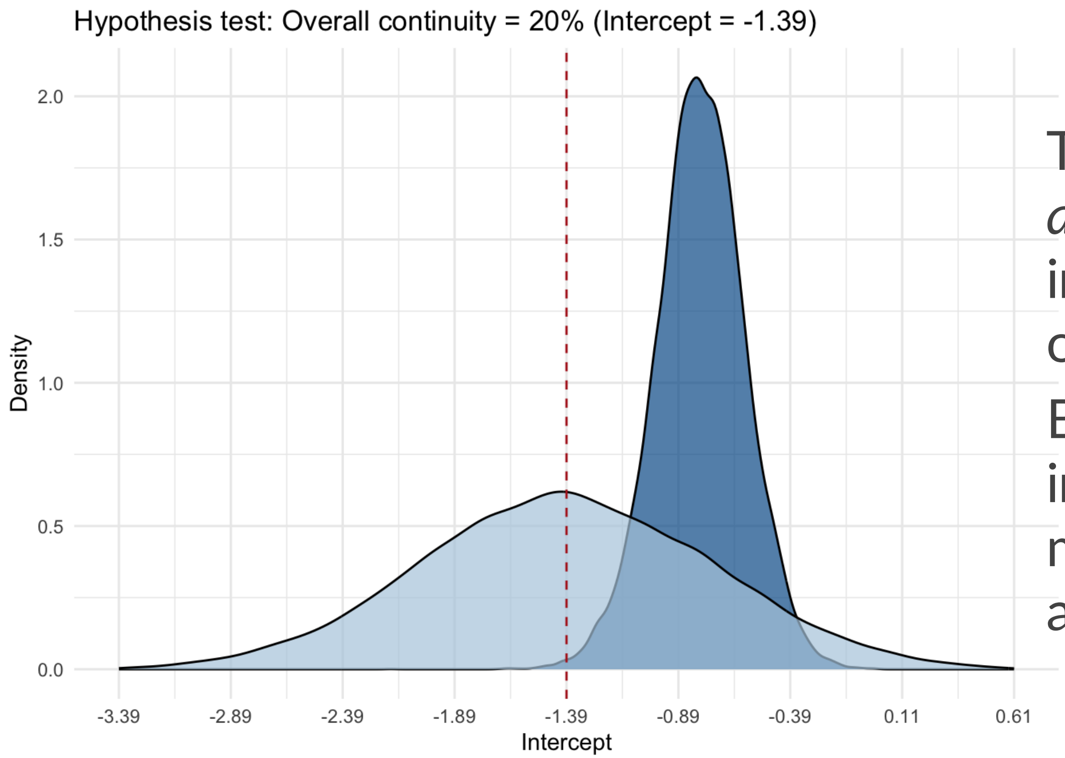


We get a Bayes factor (*Evid.Ratio*) of infinity → all posterior draws are larger than zero, indicating that the data provide strong evidence in favor of H_1.

Rather than infinity, we should read this as BF_{10}>20000, as we have 20000 posterior samples, all of which are larger than zero.

Savage-Dickey density ratio

Consider: H_2: overall continuity is around 20% on average



The data show evidence *against* the hypothesis that the intercept is at 20% (i.e., -1.39 on the logit scale):

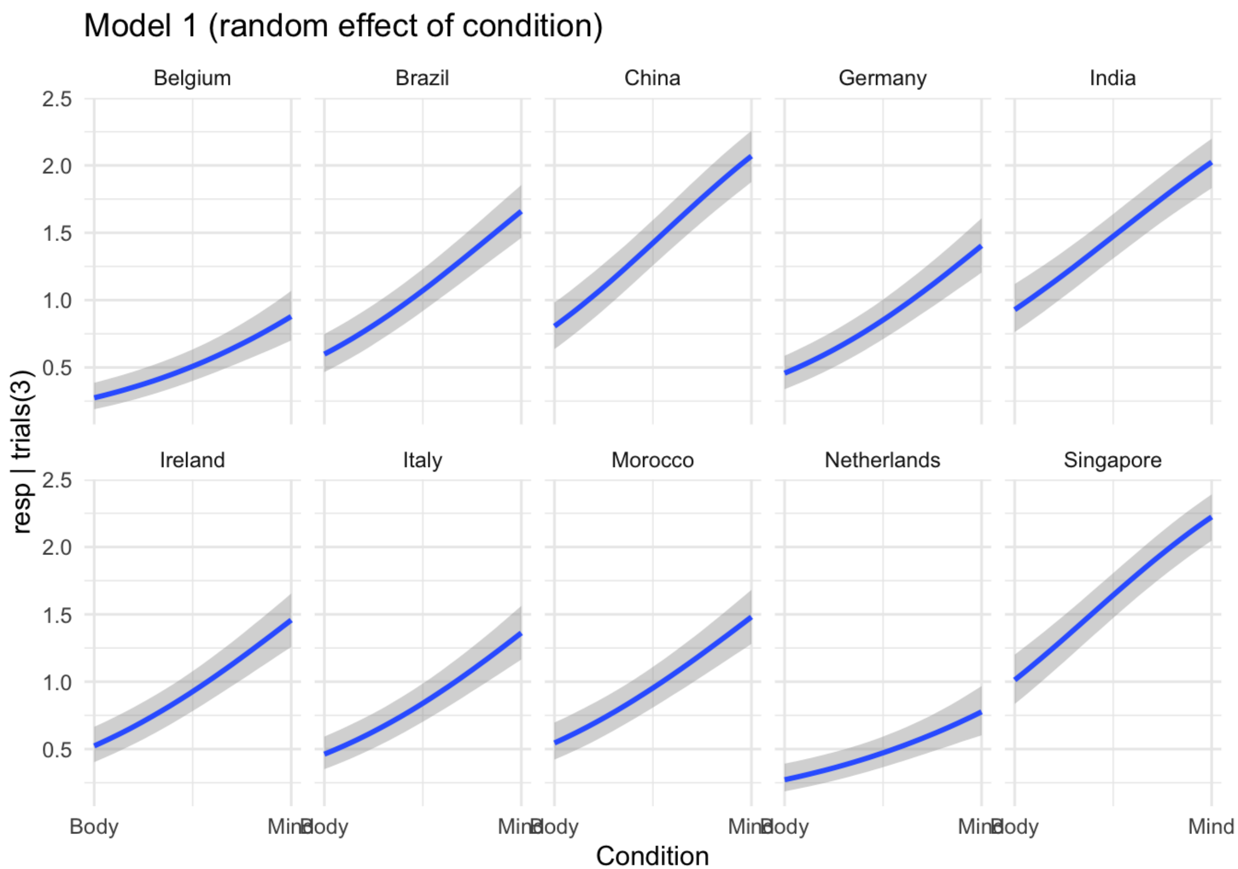
BF_{01}= 0.054; BF_{10}= 18.456, indicating that the data provide moderate to strong evidence against this hypothesis.

Model comparison

- Now we want to evaluate the evidence in the data for the inclusion of a random effect of condition (H_3); that is, is the difference between the body and mind condition different across countries?
- To do this, we can compare the model with a random effect of condition (M_1) to a model without a random effect of condition (M_2). We can then compute the Bayes factor to quantify the evidence in the data for M_1 compared to M_2.

Model comparison

- Now we want to evaluate the evidence in the data for the inclusion of a random effect of condition (H_3); that is, is the difference between the body and mind condition different across countries?



Model comparison

- Now we want to evaluate the evidence in the data for the inclusion of a random effect of condition (H_3); that is, is the difference between the body and mind condition different across countries?
- Here we get BF_{12}=0.249, or BF_{21}=4.02, which indicates that the data provide moderate evidence in favor of M_2 (no random effect) compared to M_1 (random effect).

Model comparison

- Now we want to evaluate the evidence in the data for the inclusion of a random effect of condition (H_3); that is, is the difference between the body and mind condition different across countries?
- Here we get BF_{12}=0.249, or BF_{21}=4.02, which indicates that the data provide moderate evidence in favor of M_2 (no random effect) compared to M_1 (random effect).
- We can also calculate the corresponding posterior model probabilities, that is, the probability of M_1 given the data or $P(M_1|\text{data})$, and the probability of M_2 given the data, or $P(M_2|\text{data})$.
- Assuming equal prior model probabilities, the posterior probability of M_1 is 0.199, while the posterior probability of M_2 is 0.801, which aligns with the moderate evidence for M_2 from the Bayes factor.

Useful references

Simulation based calibration:

- Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). *Bayesian Workflow* (No. arXiv:2011.01808). arXiv. https://doi.org/10.48550/arXiv.2011.01808
- Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). *Validating Bayesian Inference Algorithms with Simulation-Based Calibration* (No. arXiv:1804.06788). arXiv. https://doi.org/10.48550/arXiv.1804.06788
- Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103–126. https://doi.org/10.1037/met0000275

Useful references

Hypothesis testing with Bayes factors:

- Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H. A. L., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., … Hoijtink, H. (2023). A review of applications of the Bayes factor in psychological research. *Psychological Methods*, *28*(3), 558–579. https://doi.org/10.1037/met0000454
- Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*(5), 539–556. https://doi.org/10.1037/met0000201

Evaluation