Introduction to Bayesian Inference

Paul Bürkner

Types of Probabilities

- Probability as the limit of relative frequencies

"In 52% of the presidental elections between Clinton and Trump, Clinton will win."

- Probability as the representation of our beliefs about the world

"Clinton will win against Trump with probability of 52%."

Types of Probabilities

- Probability as the limit of relative frequencies

"The true parameter value lies within 95% of the confidence intervals"

- Probability as the representation of our beliefs about the world

"With 95% probability, the parameter value lies within the credible interval"

Prior and Posterior Beliefs

- Before the data collection, we hold certain **prior** beliefs about the effects under study.
- After collecting the **data**, we update our beliefs, which then become our **posterior** beliefs.
- Bayesian inference gets us from prior to posterior beliefs.

Bayes Rule

- The posterior probability of the parameters given data is

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- Likelihood: $p(D|\theta)$
- Prior: $p(\theta)$
- Evidence / Marginal likelihood: $p(D)$

A Simple Example

- Assume a multiple-choice task with 4 alternatives of which only 1 is correct
- We want to model the rate θ (our parameter) that participants correctly solve the task

We need:

- The likelihood / generative model (probability of the data given the parameters)
- The prior (probability of the parameters before seing the data)

The Binomial Likelihood

- We call y_i the response of participant i
- The response may be correct ($y_i = 1$) or incorrect ($y_i = 0$)
- A correct response is obtained with probability θ
- An incorrect response is obtained with probability $1 - \theta$
- A total of N participants complete the task

The Binomial likelihood for the number of correct responses y:

$$p(y|\theta, N) = \binom{N}{y} \theta^y (1 - \theta)^{N-y}$$

An Artifical Prior

- Suppose that θ can only take on the values .10, .30, .50, .70.
- We expect the following probabilites for these values:

$$p(\theta = .10) = .10$$

$$p(\theta = .30) = .20$$

$$p(\theta = .50) = .60$$

$$p(\theta = .70) = .10$$

Obtaining the Posterior Distribution (1)

- Suppose we have observed $y = 4$ of $N = 10$ correct responses
- We compute

$$p(y = 4|\theta = .10, N = 10) \times p(\theta = .10) = .011 \times .10$$

$$p(y = 4|\theta = .30, N = 10) \times p(\theta = .30) = .200 \times .20$$

$$p(y = 4|\theta = .50, N = 10) \times p(\theta = .50) = .205 \times .60$$

$$p(y = 4|\theta = .70, N = 10) \times p(\theta = .70) = .037 \times .10$$

$$p(y = 4) = \sum_{j=1}^{4} p(y = 4|\theta_j, N = 10) \times p(\theta_j)$$

Obtaining the Posterior Distribution (2)

- We apply Bayes Theorem:

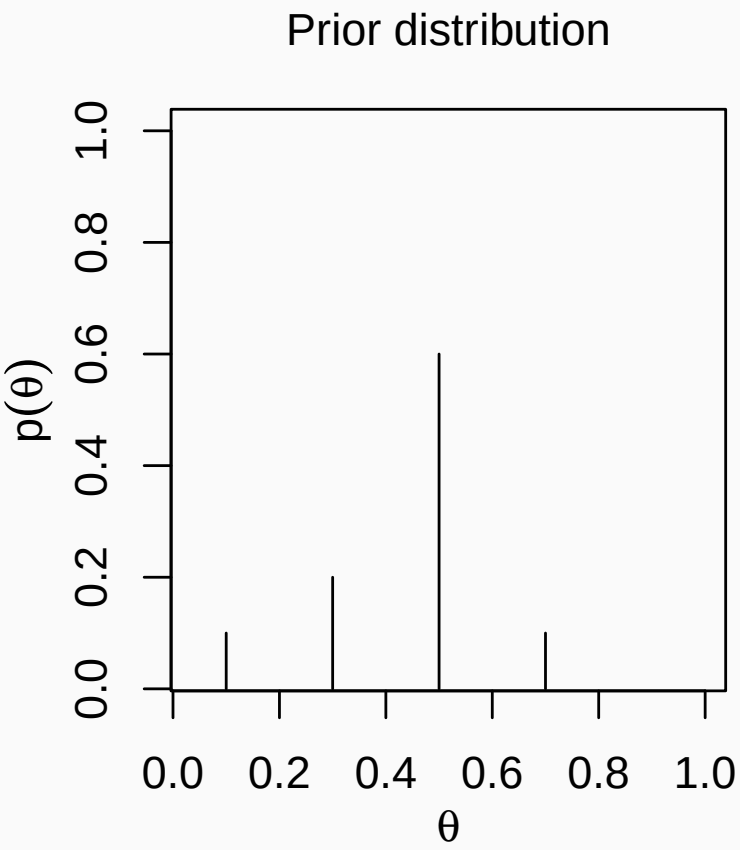
$$p(\theta = 0.10|y = 4) = .007$$

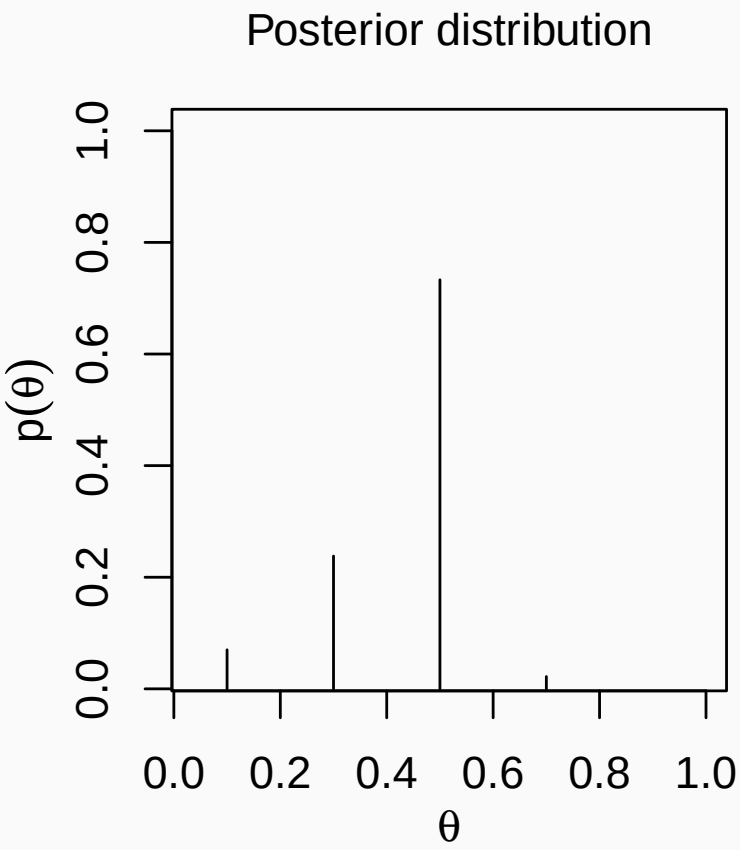
$$p(\theta = 0.30|y = 4) = .238$$

$$p(\theta = 0.50|y = 4) = .733$$

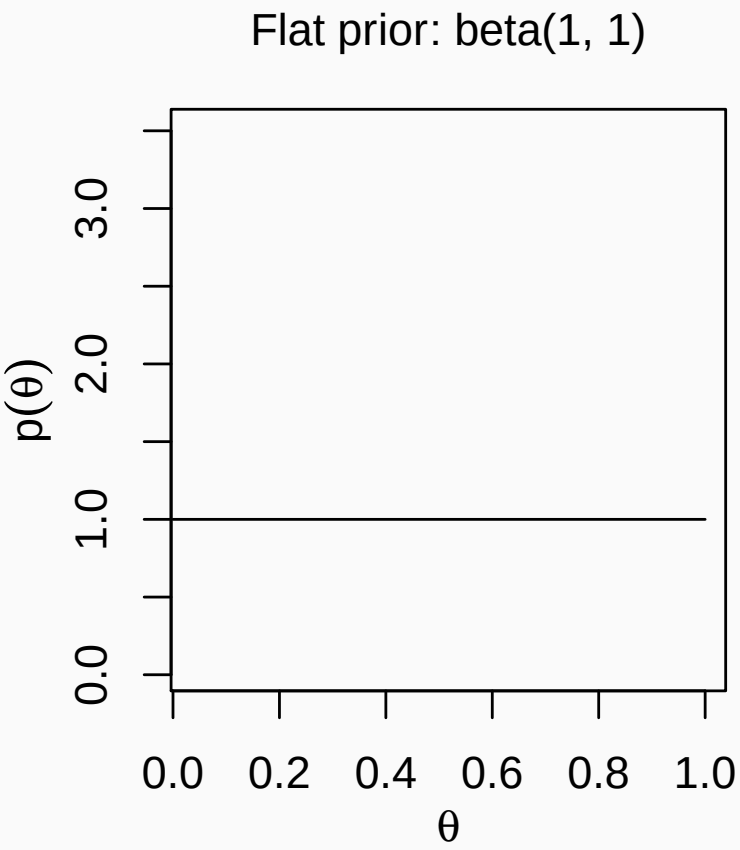
$$p(\theta = 0.70|y = 4) = .022$$

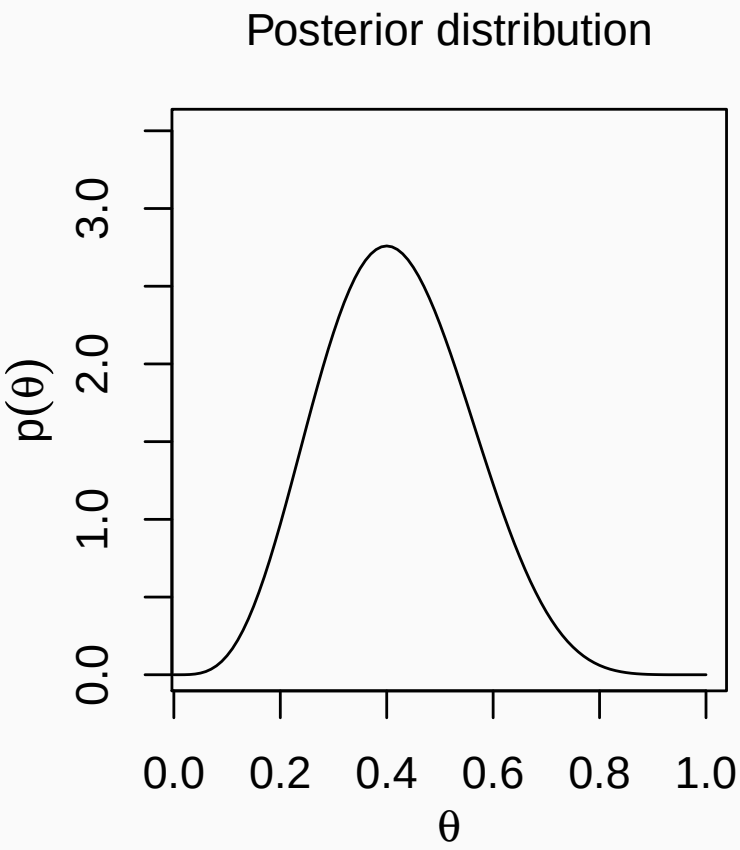
Prior vs. Posterior distribution



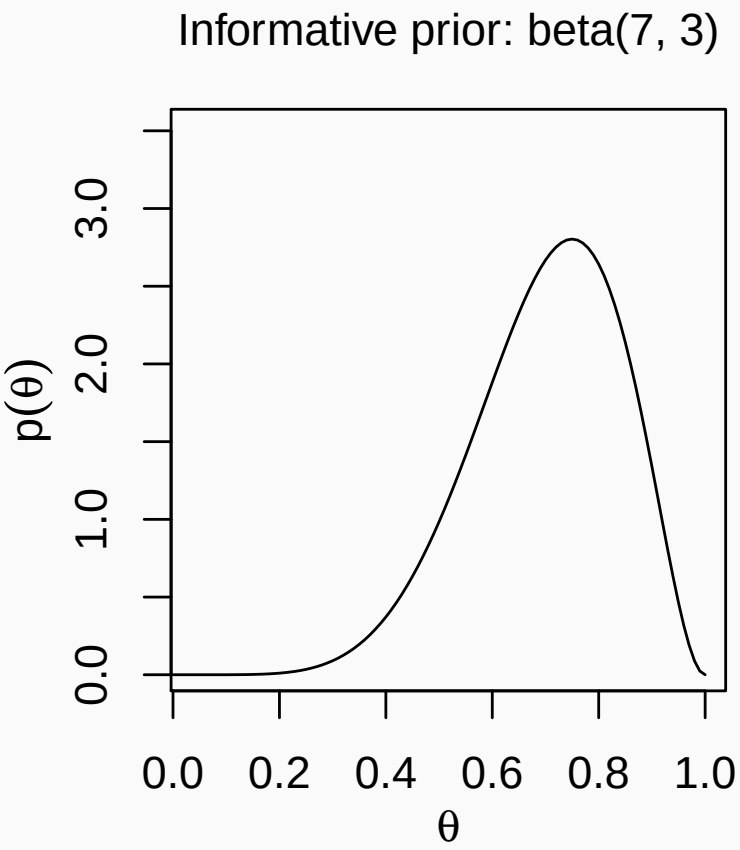


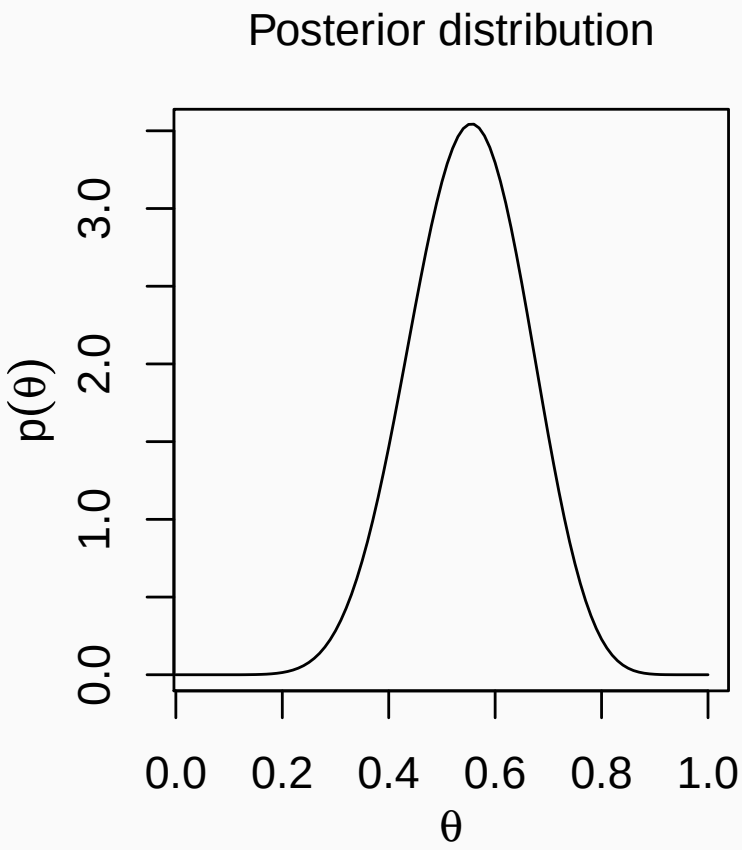
Results for a Continuous Prior





Results for a Continuous Prior





Sampling from the Posterior distribution

Why Do We need Sampling?

- For simple models, we can compute the evidence $p(D) = \int p(D|\theta)p(\theta)d\theta$ analytically.

Binomial likelihood with flat prior:

$$p(y) = \int_0^1 \binom{N}{y} \theta^y (1-\theta)^{N-y} \times 1 \, d\,\theta = \frac{1}{N+1}$$

- For a bit more complex models, integration may be done numerically.
- For more than 3 or 4 parameters, numerical computation of the evidence becomes infeasible
- We need to sample (somehow) from the posterior

Rejection Sampling

- Sample parameter values from the prior
- Sample data from the likelihood based on the sampled parameters
- Only keep those parameter values, which produced data consistent with our observed data
- Repeat this process many times
- The kept parameter values are samples from the posterior

Rejection Sampling: Examples

Hypothetical Sample 1:

- Sampling from the prior yields $\theta_s = 0.7$
- Sampling from the binomial likelihood $p(y, \theta = 0.7, N = 10)$ yields $y_s = 6$
- The sample response $y_s = 6$ is different from the observed response $y = 4$ so that $\theta_s = 0.7$ is thrown away

Hypothetical Sample 2:

- Sampling from the prior yields $\theta_s = 0.42$
- Sampling from the binomial likelihood $p(y, \theta = 0.42, N = 10)$ yields $y_s = 4$
- The sample response $y_s = 4$ is equal to the observed response $y = 4$ so that $\theta_s = 0.42$ is kept

Time for exercise

'exercise_bayesian_inference_1.R'

Why MCMC Sampling?

- Rejection Sampling is super inefficient
- There is no *efficient* way to draw independent samples from the posterior
- Thus, we have to draw dependent samples (in a certain way)
- A Markov Chain is a sequence of values where the value at position t is based only on the former value at position $t-1$
- If done correctly, the distribution of the values will converge to the posterior distribution

The Metropolis-Algorithm

- Choose an initial value θ_1. Set $t = 1$.
- Sample a possible new value θ^* based on a *proposal distribution* $g(\theta^*|\theta_t)$ – usually use $N(\theta_t, \sigma_p)$ as the proposal distribution
- (σ_p serves as a tunining parameter controlling the *step-size*)
- Compute the ratio $\alpha = p(\theta^*|D)/p(\theta_t|D)$
- If $\alpha \geq 1$, set $\theta_{t+1} = \theta^*$.
- If $\alpha < 1$, set $\theta_{t+1} = \theta^*$ with probability α
- Else, go back to step 2 and sample new value θ^*

Advantages of Bayesian Statistics

- Natural approach to probability
- Ability to incorporate prior information
- Full posterior distribution of parameters
- Modeling flexibility
- Bayes factors (?)

Disadvantages of Bayesian Statistics

More complicated to fit and apply

- Recently developed software packages make things easier

Speed (for large data sets)

- Possibly the only remaining drawback

Model Comparison / Hypothesis Testing

Three ways for hypothesis testing:

- Estimation with uncertainty
- Bayes factors
- Information criteria

Estimation with Uncertainty

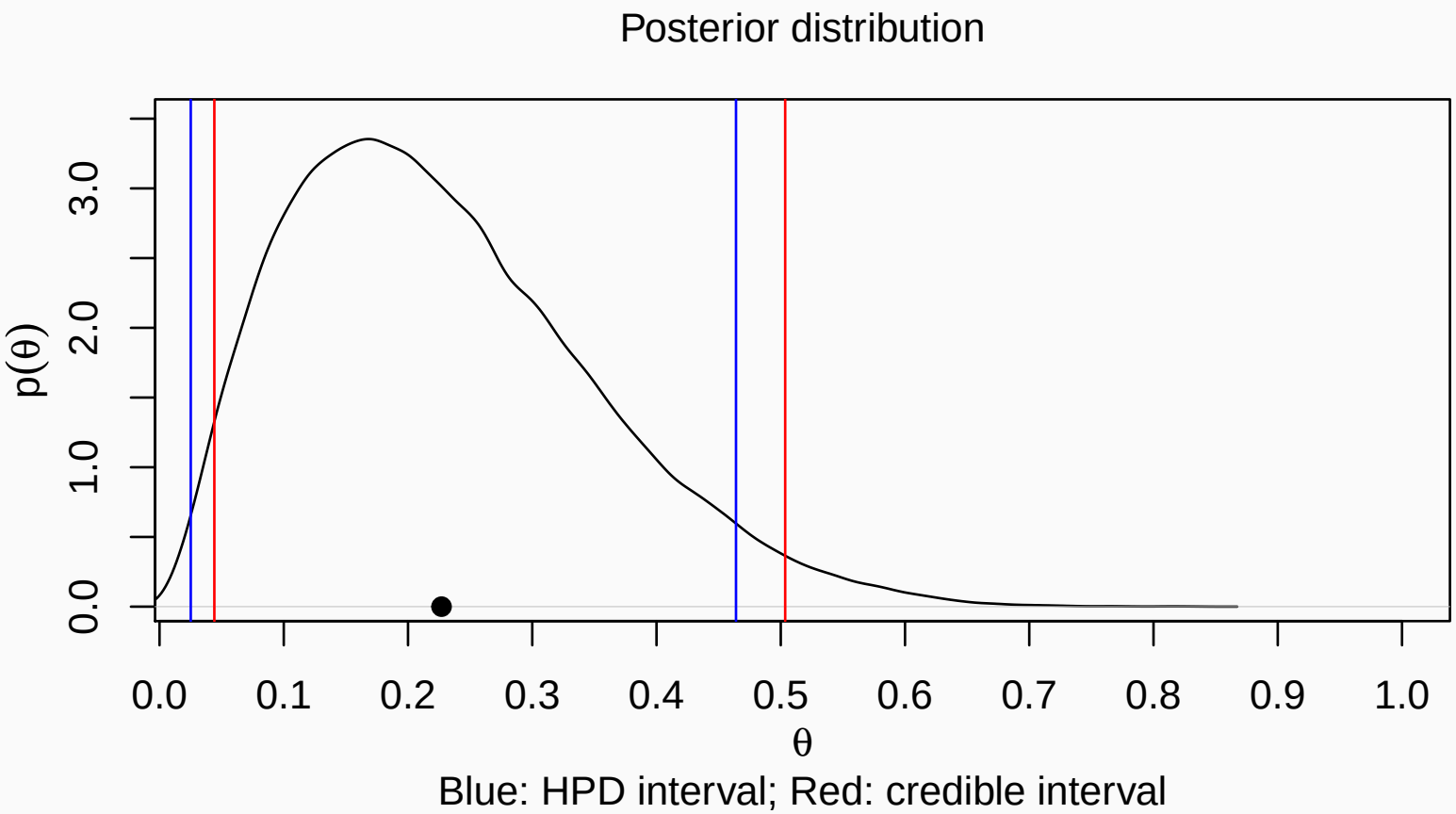
For inference use:

- Point estimates
- Uncertainty intervals (UIs)

Bayesian Uncertainty intervals:

- Credible intervals based on quantiles
- Highest posterior density (HPD) intervals (also called HDIs)

Visualization of uncertainty intervals



Blue: HPD interval; Red: credible interval

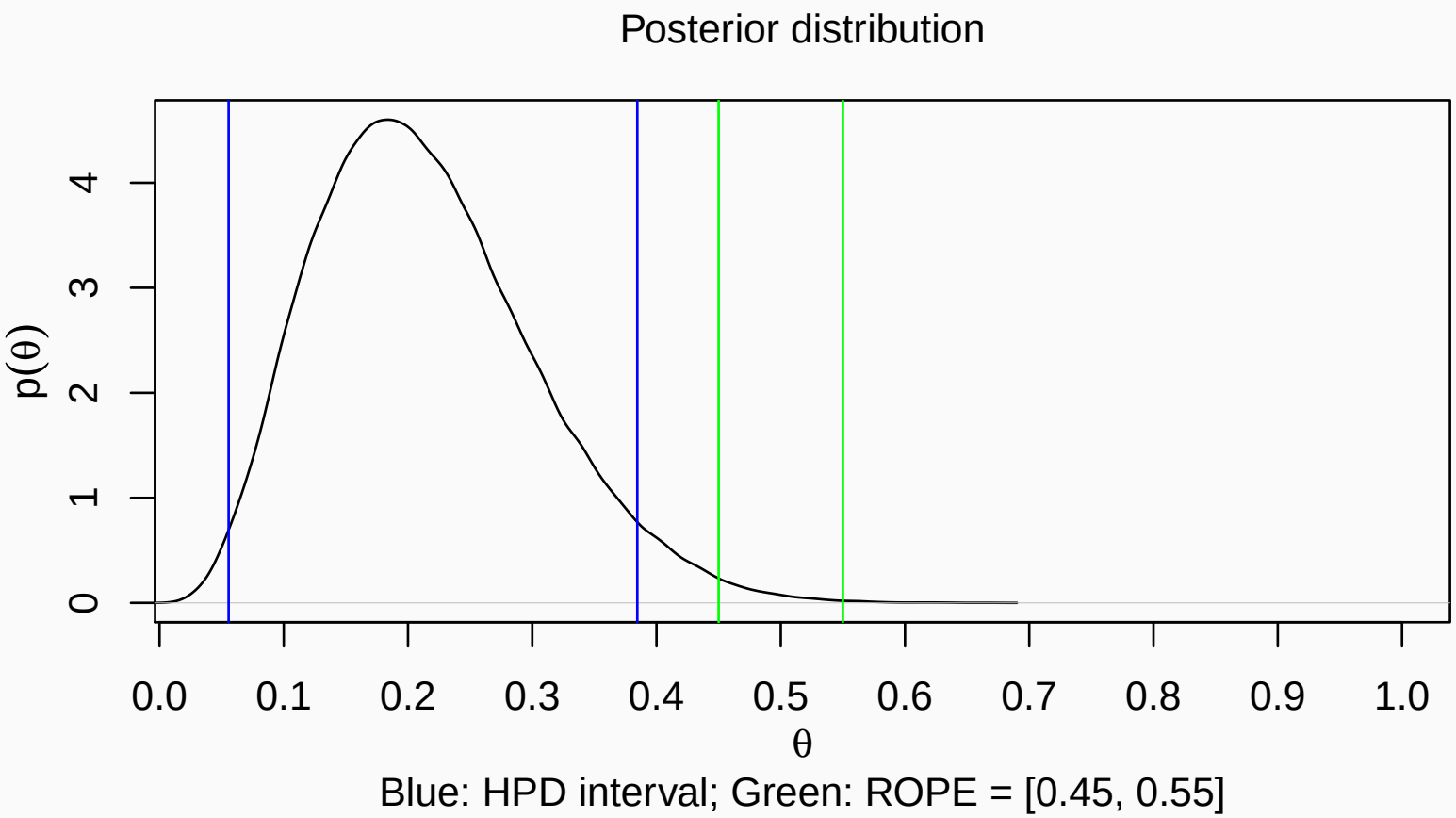
Region of Practical Equivalence (ROPE)

- Define a region that is thought to be practically equivalent to the value being tested.
- For instance $ROPE = [d = -0.1, d = 0.1]$ in intervention studies

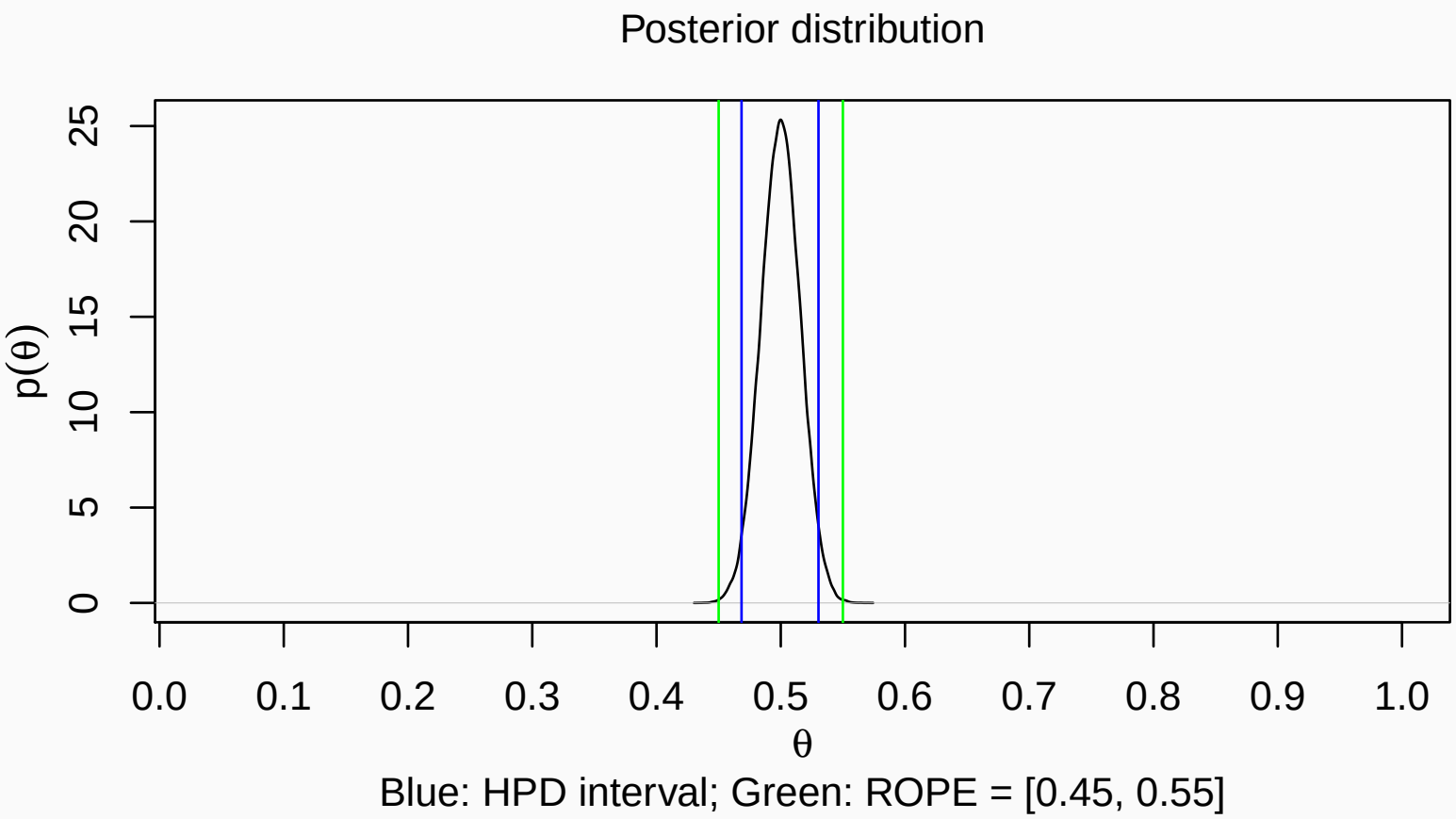
Three possible outcomes of the hypothesis:

- ROPE and UI do not intersect: Reject the null hypothesis
- UI is completely within ROPE: Accept the null hypothesis
- Else: Evidence is inconclusive

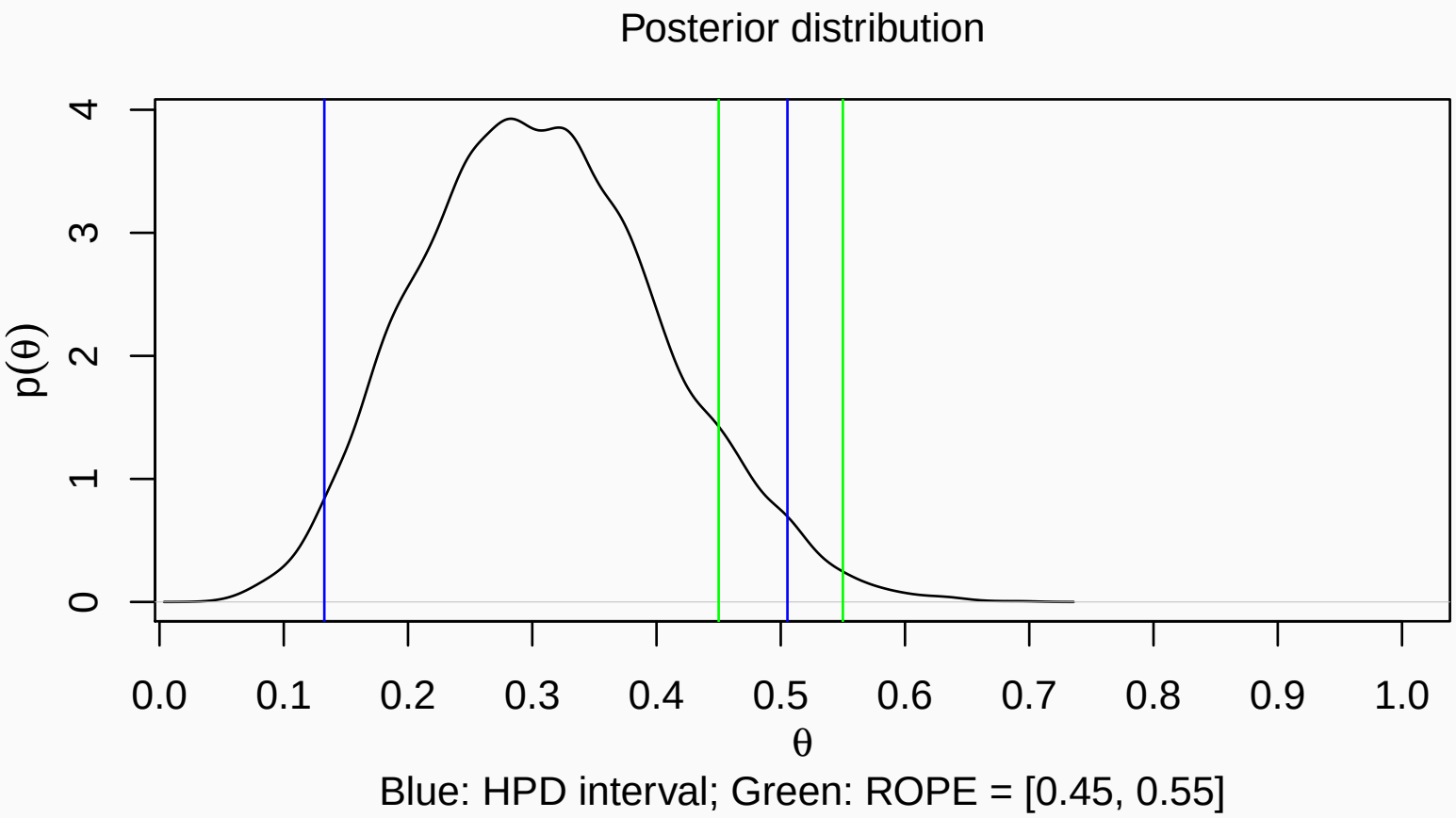
Visualization of ROPEs: Reject the Null Hypothesis



Visualization of ROPEs: Accept the Null Hypothesis



Visualization of ROPEs: Inconclusive



Evidence of Models

Evidence of model M:

$$p(D|M) = \int p(D|\theta, M)p(\theta|M)d\theta$$

- This is the probability of the data given the model
- Can be considered as a measure of model fit
- Depends heavily on the prior $p(\theta|M)$

Bayes Factors

- Used to compare two models M_1 and M_2:

$$BF_{12} = \frac{p(D|M_1)}{p(D|M_2)}$$

- Closely related to the posterior Odds:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(M_1)}{p(M_2)} BF_{12}$$

- $p(M_1)$ and $p(M_2)$ are the prior probabilites of the models M_1 and M_2
- Usually $p(M_1) = p(M_2) = 1/2$

The Savage-Dickey Ratio

- Computation of the evidence is complicated and so is the computation of the BF
- Assume we are testing $M_1 : \theta = \theta_0$ against $M_2 : \theta \neq \theta_0$
- (We could use the word 'hypothesis' instead of 'models')
- Then the Bayes factor can be computed as

$$BF_{12} = \frac{p(\theta_0|D, M_2)}{p(\theta_0|M_2)}$$

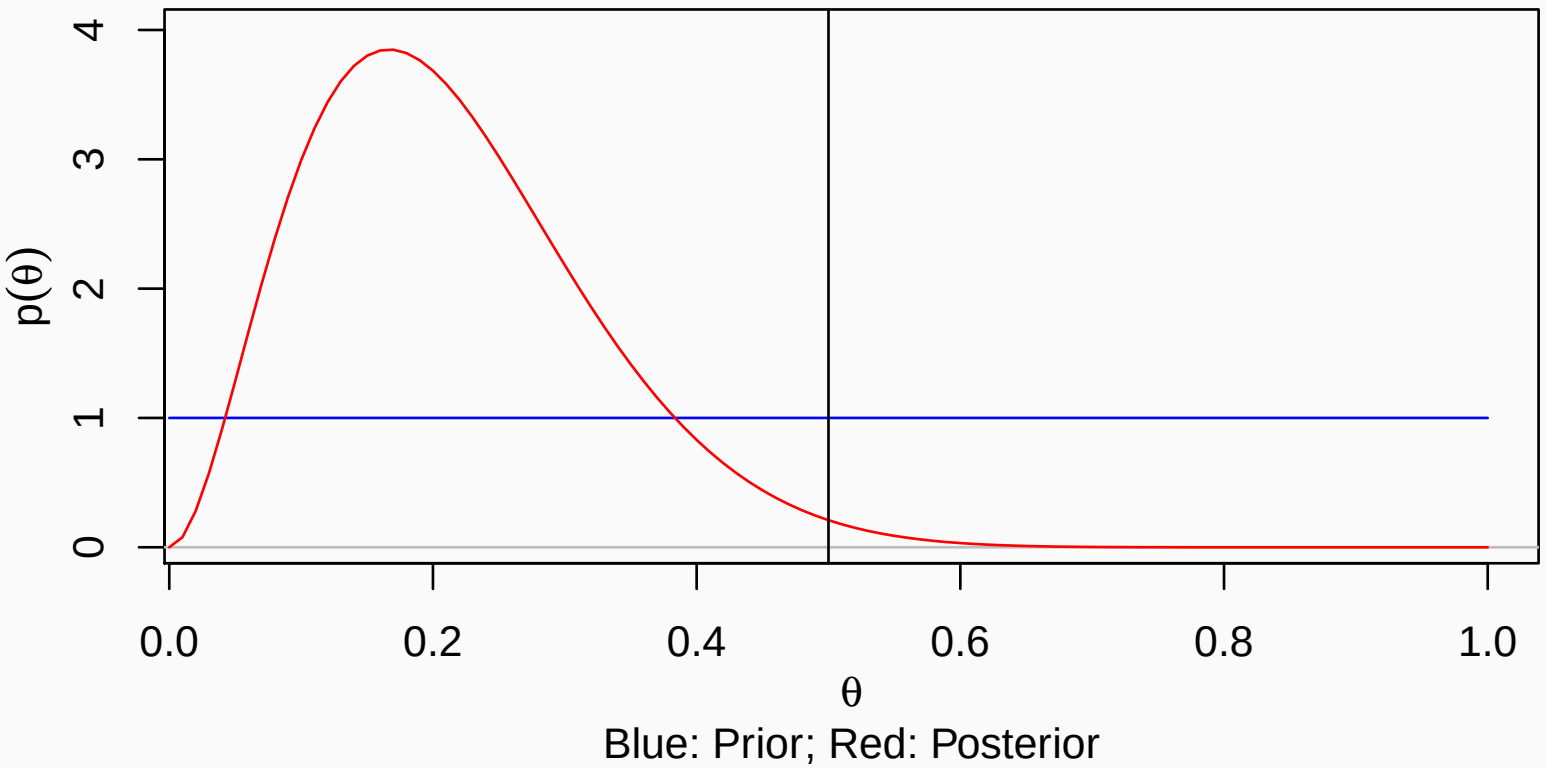
Bayes Factors: Example

- We assume a flat prior $beta(1, 1)$
- We observed $y = 2$ for $N = 12$.
- The resulting posterior (computed analytically) is $beta(3, 11)$
- We are interested in the BF at $\theta_0 = 0.5$

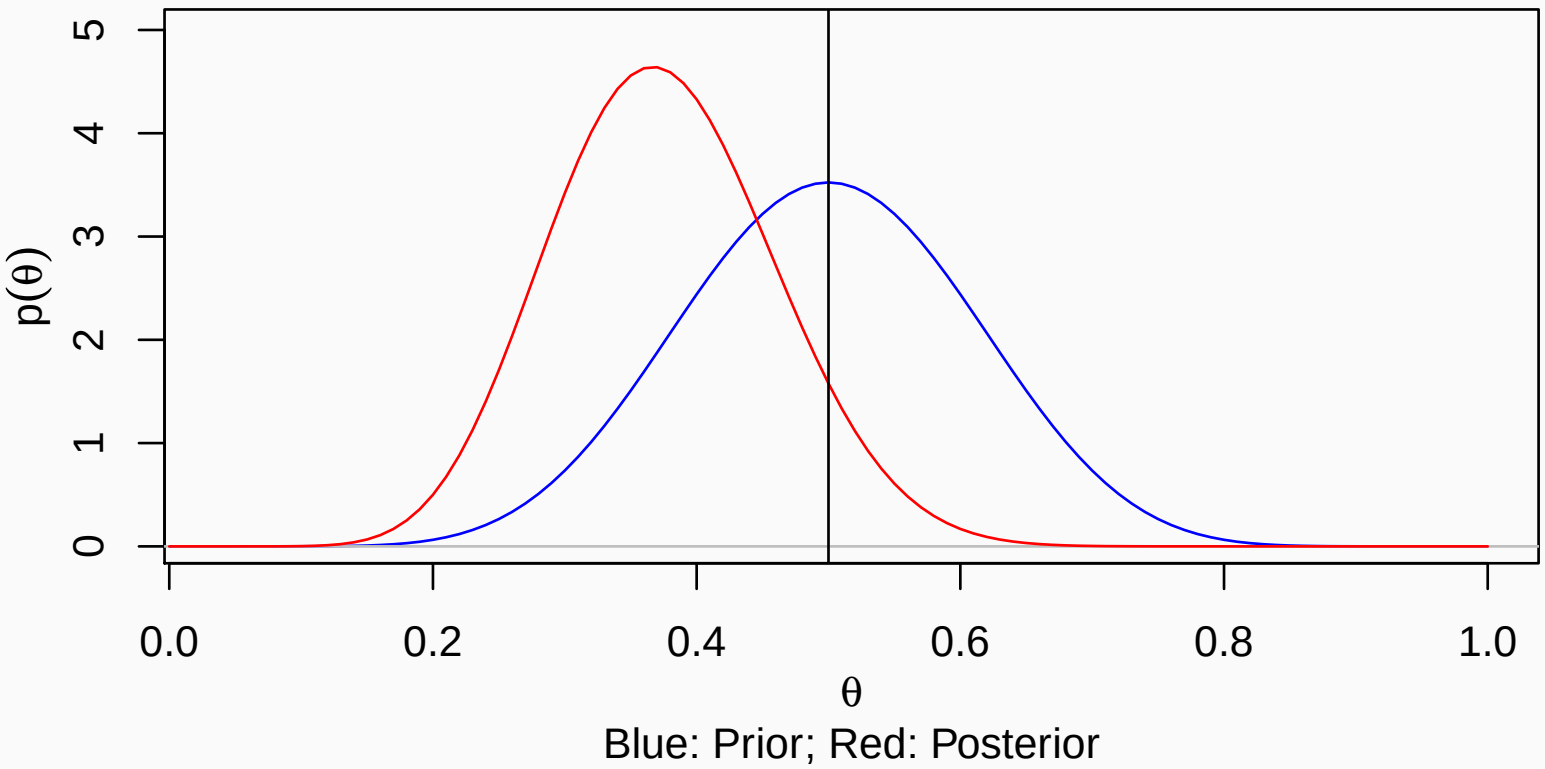
```
dbeta(0.5, 3, 11) / dbeta(0.5, 1, 1)
```

```
## [1] 0.2094727
```

Bayes Factors: Visualization



Bayes factor: Influence of Priors



Time for exercise

'exercise_bayesian_inference_2.R'

Bayes Factors: Comparison of Two Groups (1)

Simulate some data

```
set.seed(0)
dat <- data.frame(y = c(rnorm(30), rnorm(30, mean = 0.6)),
                  group = rep(0:1, each = 30))
head(dat)
```

y	group
1.2629543	0
-0.3262334	0
1.3297993	0
1.2724293	0
0.4146414	0
-1.5399500	0

Bayes Factors: Comparison of Two Groups (2)

Fit a simple model to compare both groups

```
library(brms)
fit1 <- brm(y ~ group, data = dat,
            prior = prior(normal(0, 0.5)),
            sample_prior = TRUE,
            iter = 5000, warmup = 1000)
```

Bayes Factors: Comparison of Two Groups (3)

Compute the Bayes factor

```
(hyp <- hypothesis(fit1, "group = 0"))
```

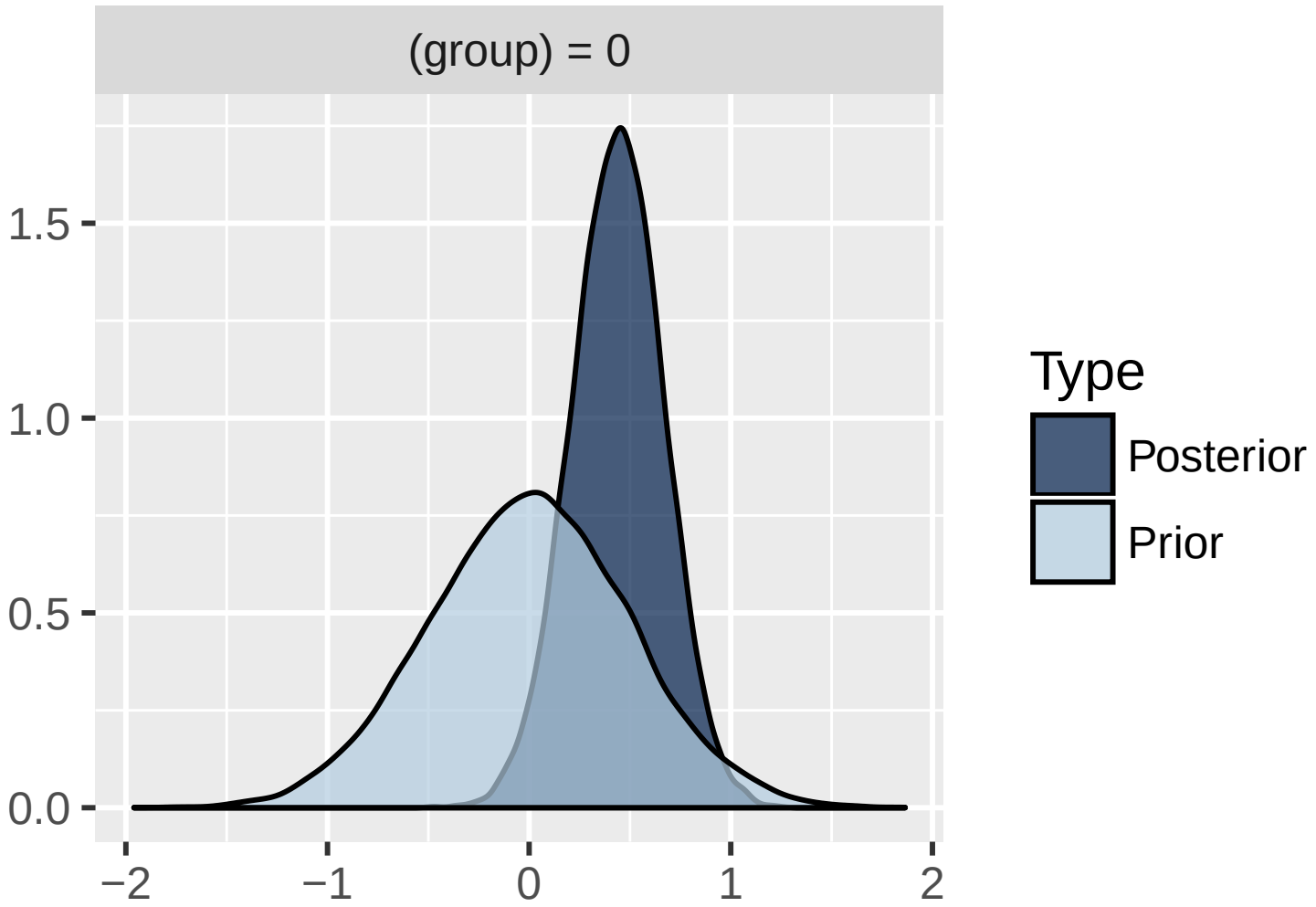
```
## Hypothesis Tests for class b:
##             Estimate Est.Error l-95% CI u-95% CI Evid.Ratio Star
## (group) = 0     0.44      0.23    -0.01     0.88       0.34
## ---
## '*': The expected value under the hypothesis lies outside the 95%
```

```
# Evidence in favour of M2
1 / hyp$hypothesis$Evid.Ratio
```

```
## [1] 2.918184
```

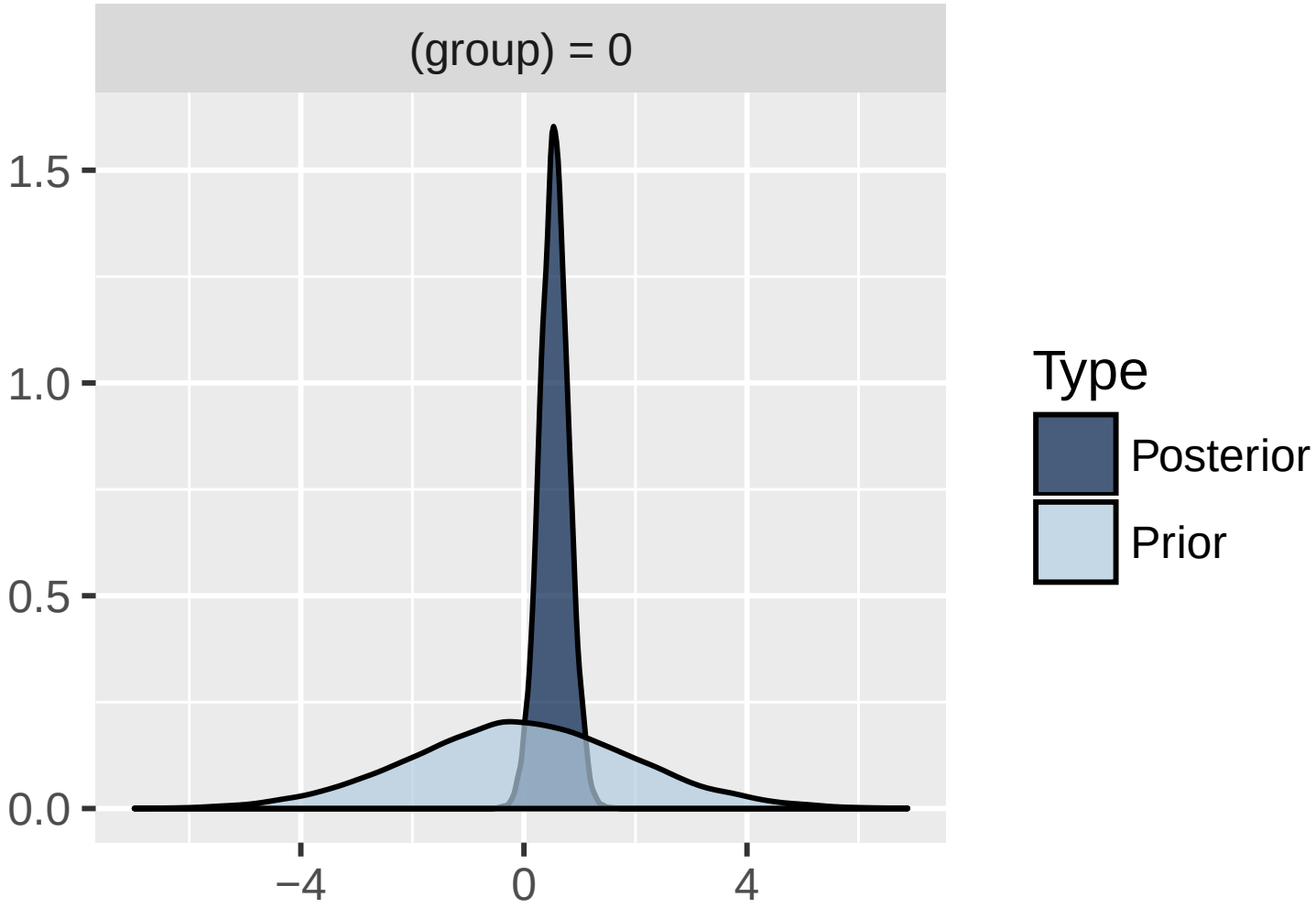
Bayes Factor: Visualization

```
plot(hyp)
```



Bayes factor: Influence of Priors

```
plot(hypothesis(fit2, "group = 0"))
```



Bayesian Information Criteria: LOO and WAIC

- Provide an estimate of the out-of-sample deviance (predictive error) of a model
- Can be interpreted as other information criteria such as the AIC (smaller values indicate better fit)

Leave-one-out cross-validation (LOO) is the current state of the art Bayesian information criterion

- Standard errors of LOO can be computed to ease comparisions
- Can be approximated throught the widely applicable information criterion (WAIC)
- Far less dependent on the prior than Bayes factors

LOO: Example

Fit a model without accounting for groups

```
fit3 <- brm(y ~ 1, data = dat)
```

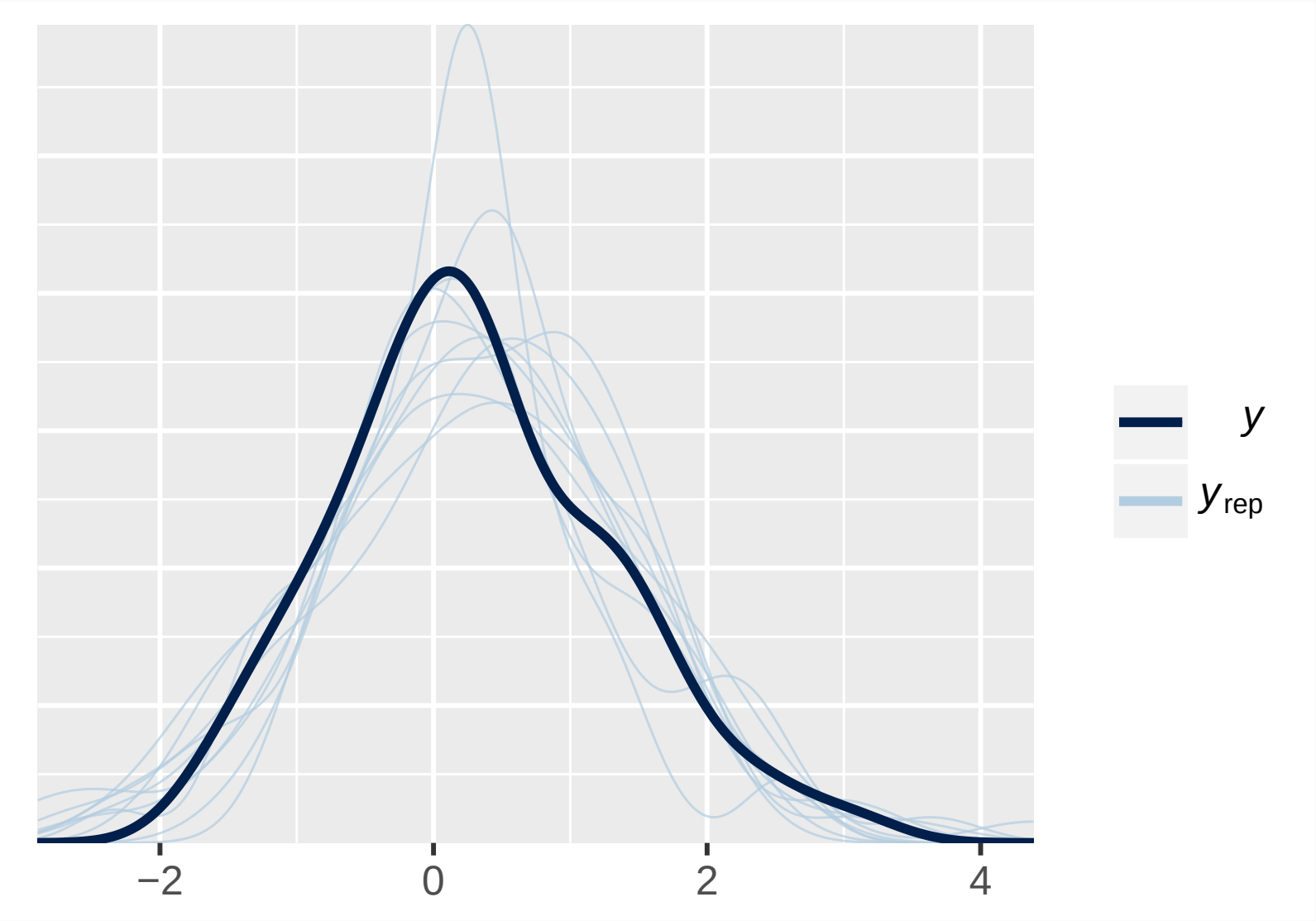
Compare models using LOO

```
LOO(fit1, fit3)
```

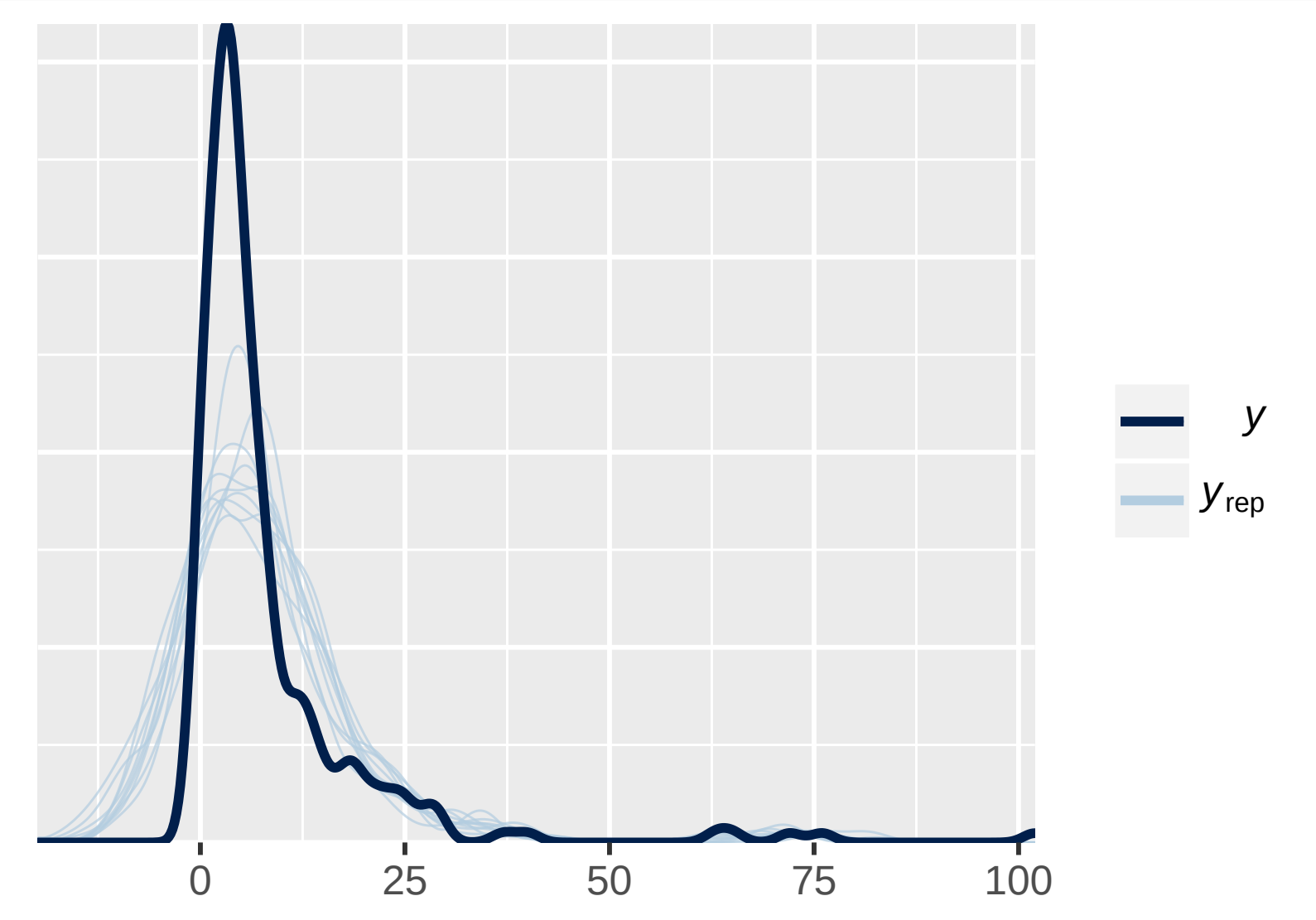
```
##                LOOIC    SE
## fit1          170.93 11.08
## fit3          173.83 10.99
## fit1 - fit3   -2.90  3.46
```

Posterior Predictive Checks

```
pp_check(fit1)
```



Posterior Predictive Checks: Number of Epileptic Seizures



Software

General software to fit Bayesian models:

- **Stan**, JAGS, and BUGS

High level R packages using Stan on the backend:

- **brms**, rstanarm, rethinking

Software to compute Bayesian versions of simple models (e.g., t-test, ANOVA):

- JASP

Further Reading

- McElreath, R. (2016). Statistical rethinking: A Bayesian course with examples in R and Stan (Vol. 122). CRC Press.
- Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press.
- Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychonomic Bulletin & Review, 1-29.
- Lee, M. D., & Wagenmakers, E. J. (2014). Bayesian cognitive modeling: A practical course. Cambridge University Press.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). Bayesian data analysis (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.