Tutorial on Bayesian Statistics 19 and 21 April 2022

https://vasishth.github.io/EMLAR2022BayesTutorial/

Shravan Vasishth
Cognitive Science / Linguistics
University of Potsdam, Germany
www.ling.uni-potsdam.de/~vasishth

Objectives

By the end of this 4 hour tutorial (Bayes 1 and 2), you will be able to:

1. Understand the logic of Bayesian methodology
2. Fit and interpret linear mixed models in brms/Stan
3. Understand uncertainty quantification and regularization in modeling
4. Start to apply Bayesian methods to your own research
5. Become informed on how to teach yourself more about statistical methods more generally.

This course is not for complete beginners in statistics

1. You have a basic competence in R
2. You have fit linear mixed models using lme4
3. You have experience in designing and running experiments

Beginner course (free):
https://vasishth.github.io/IntroductionStatistics/

The Bayesian approach

Imagine that you have some independent and identically distributed data: $x_1, x_2, \ldots, x_n$

$$X \sim Normal(\mu, \sigma)$$

1. Define **priors** for the parameters (here, μ, σ)
2. Derive **posterior distribution** of the parameter(s) of interest using Bayes' rule:

$$\underbrace{f(\mu \mid data)}_{\text{posterior}} \propto \underbrace{f(data \mid \mu)}_{\text{likelihood}} \times \underbrace{f(\mu)}_{\text{prior}}$$

3. Carry out inference based on the posterior

The Bayesian approach

$$P(A|B) = \frac{P(A, B)}{P(B)} \text{ where } P(B) > 0$$

$$P(A, B) = P(A|B)P(B) \text{ and } P(B, A) = P(B|A)P(A)$$

Equating the two expansions, we get:

$$P(A|B)P(B) = P(B|A)P(A)$$

Dividing both sides by $P(B)$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$p(\boldsymbol{\Theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta})}{p(\boldsymbol{y})}$$

posterior likelihood prior

$$f(\mu \,|\, data) \propto f(data \,|\, \mu) \times f(\mu)$$

Example: Modeling complications after surgery

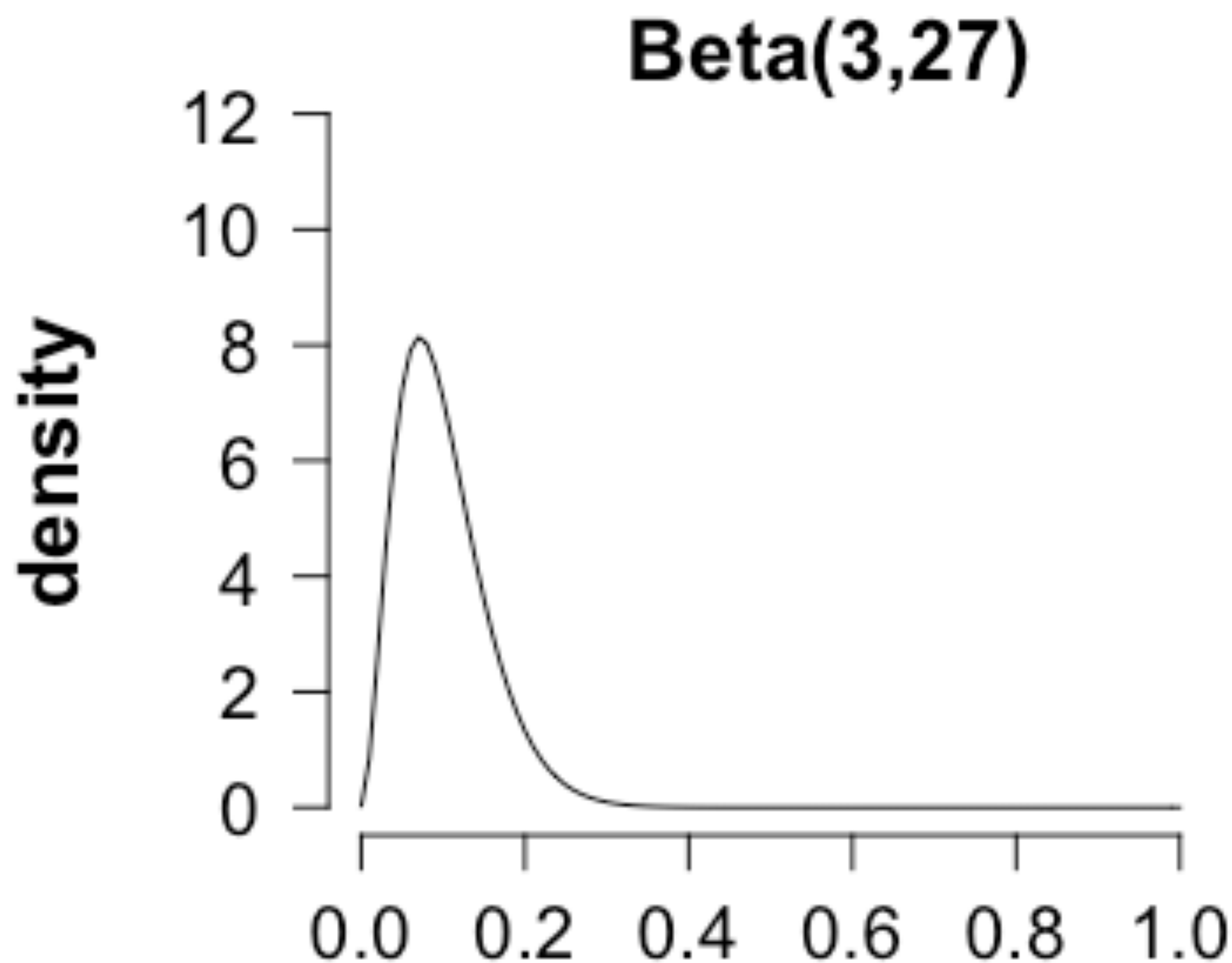
Modeling prior knowledge:

- Suppose we know that 3 out of 30 patients will experience complications after a particular operation
- This prior knowledge can be represented as a Beta(a=3,b=27) distribution:

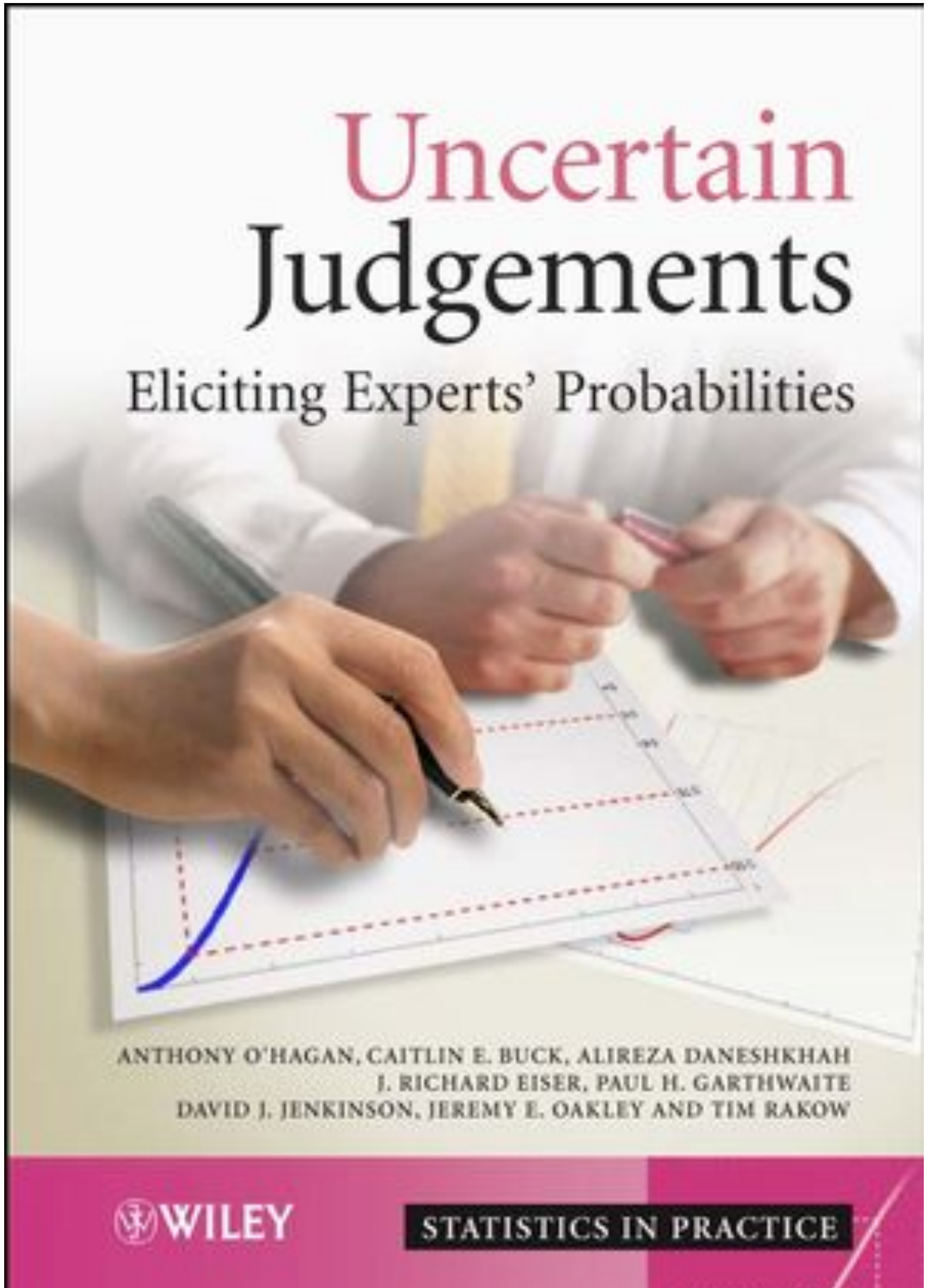
$$p(\theta|a,b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

Example: Modeling complications after surgery

Modeling prior knowledge:



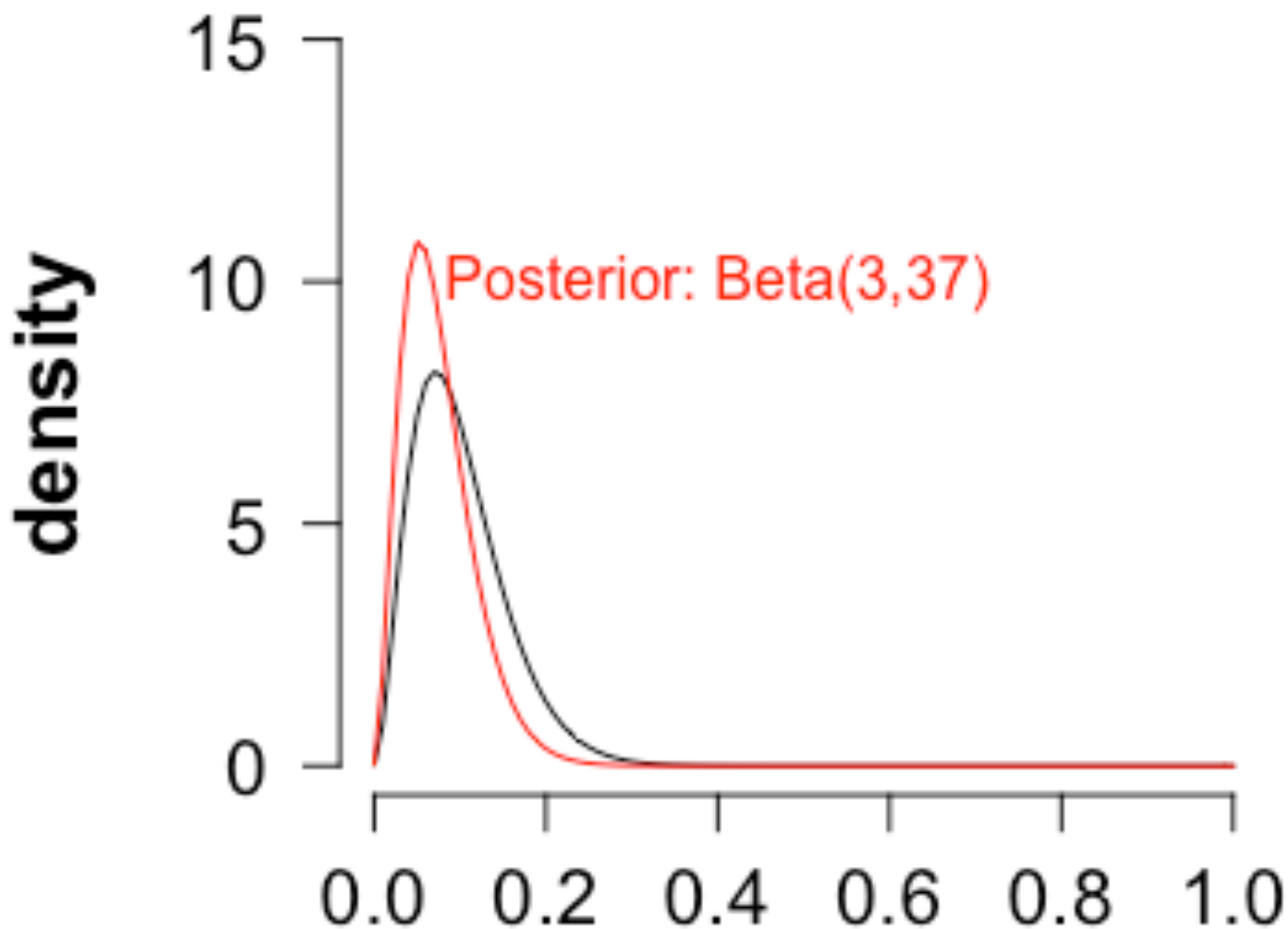
Example: Modeling complications after surgery



Example: Modeling complications after surgery

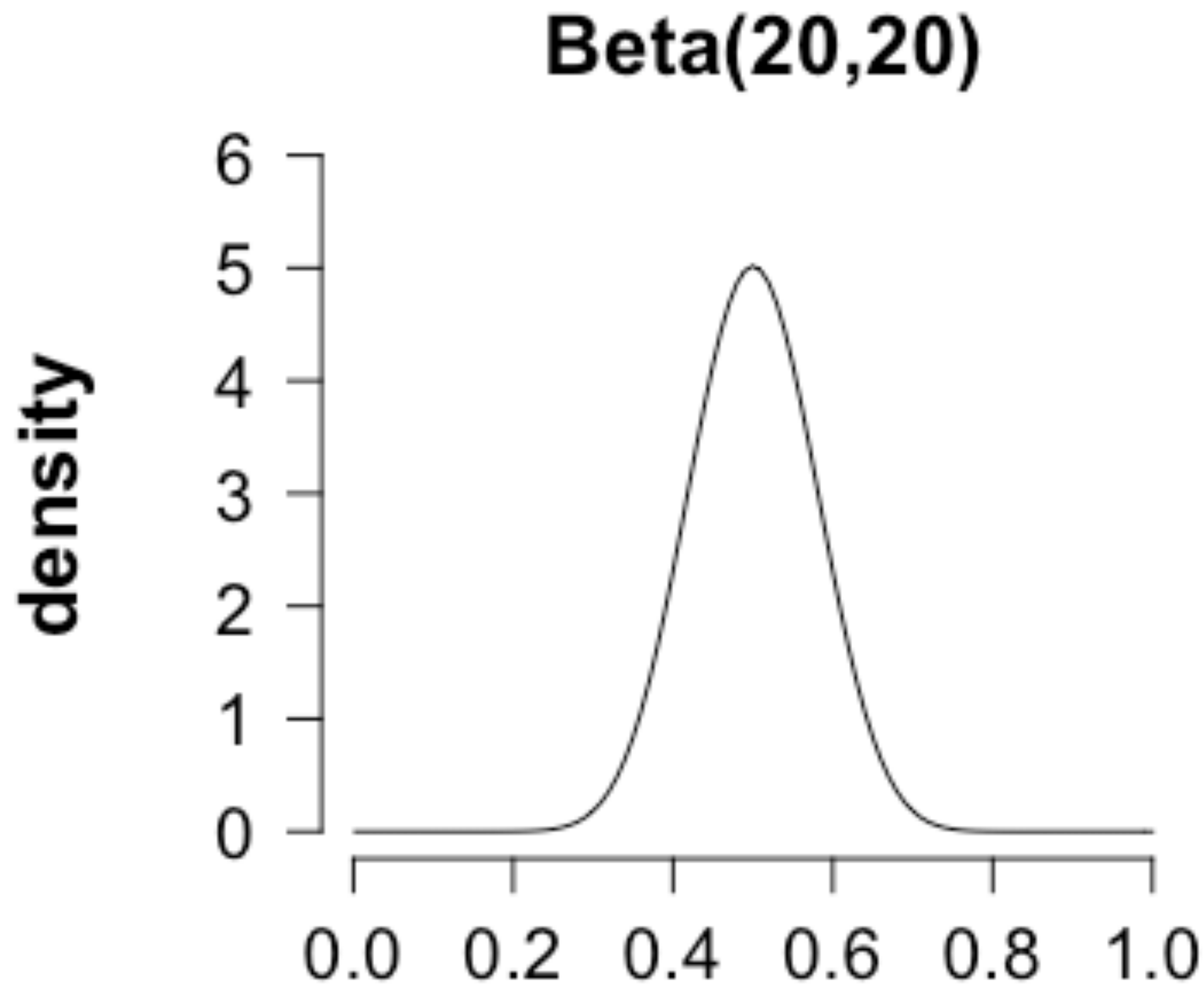
The data: 0 complications in the next 10 operations.
The **posterior distribution** of the probability of complications:

$$Posterior \propto Likelihood \times Prior$$



Example: Modeling complications after surgery

Suppose that the prior probability of complications is higher (50%):

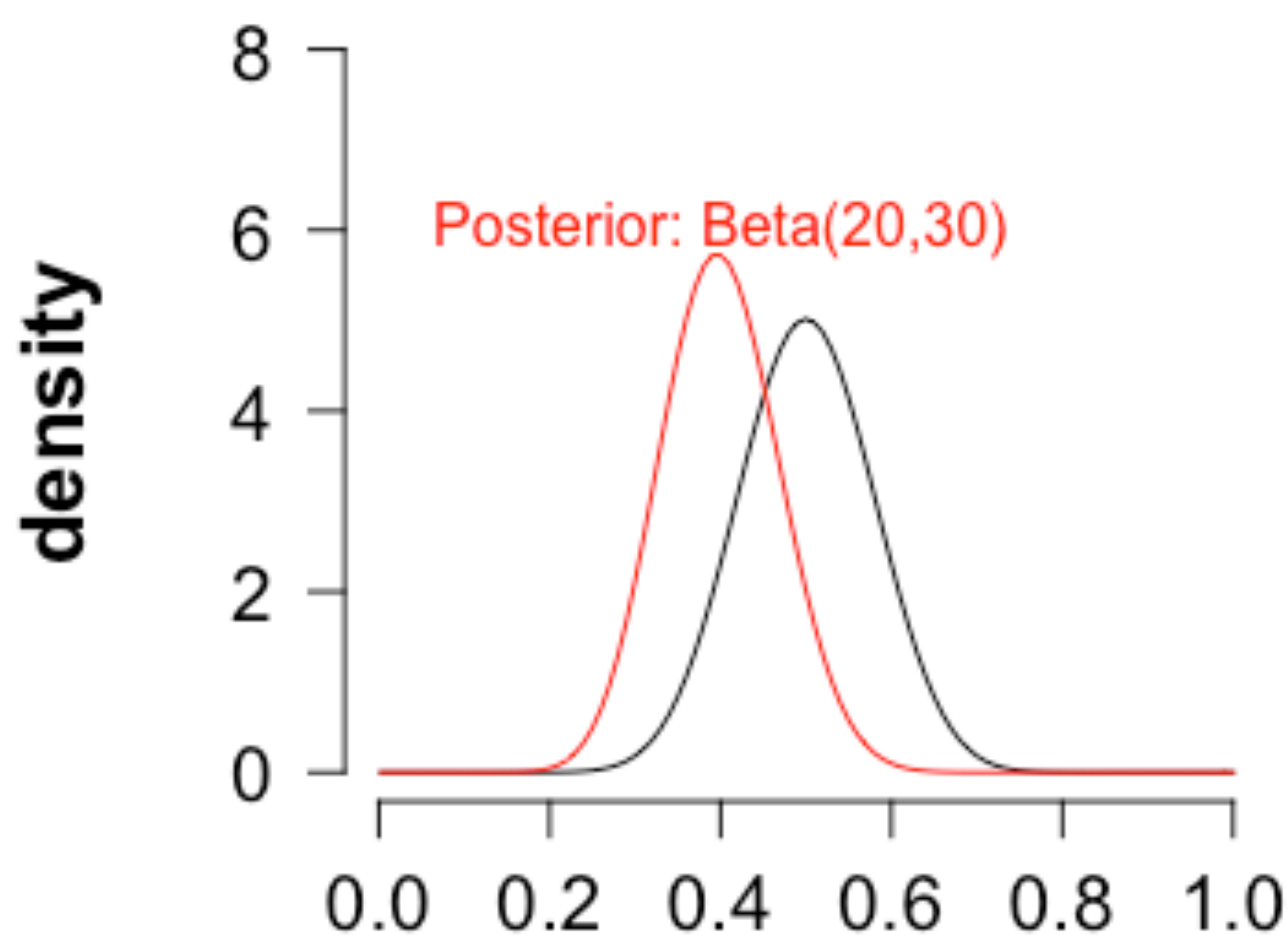


Example: Modeling complications after surgery

The data: 0 complications in the next 10 operations.

The **posterior distribution** of the probability of complications:

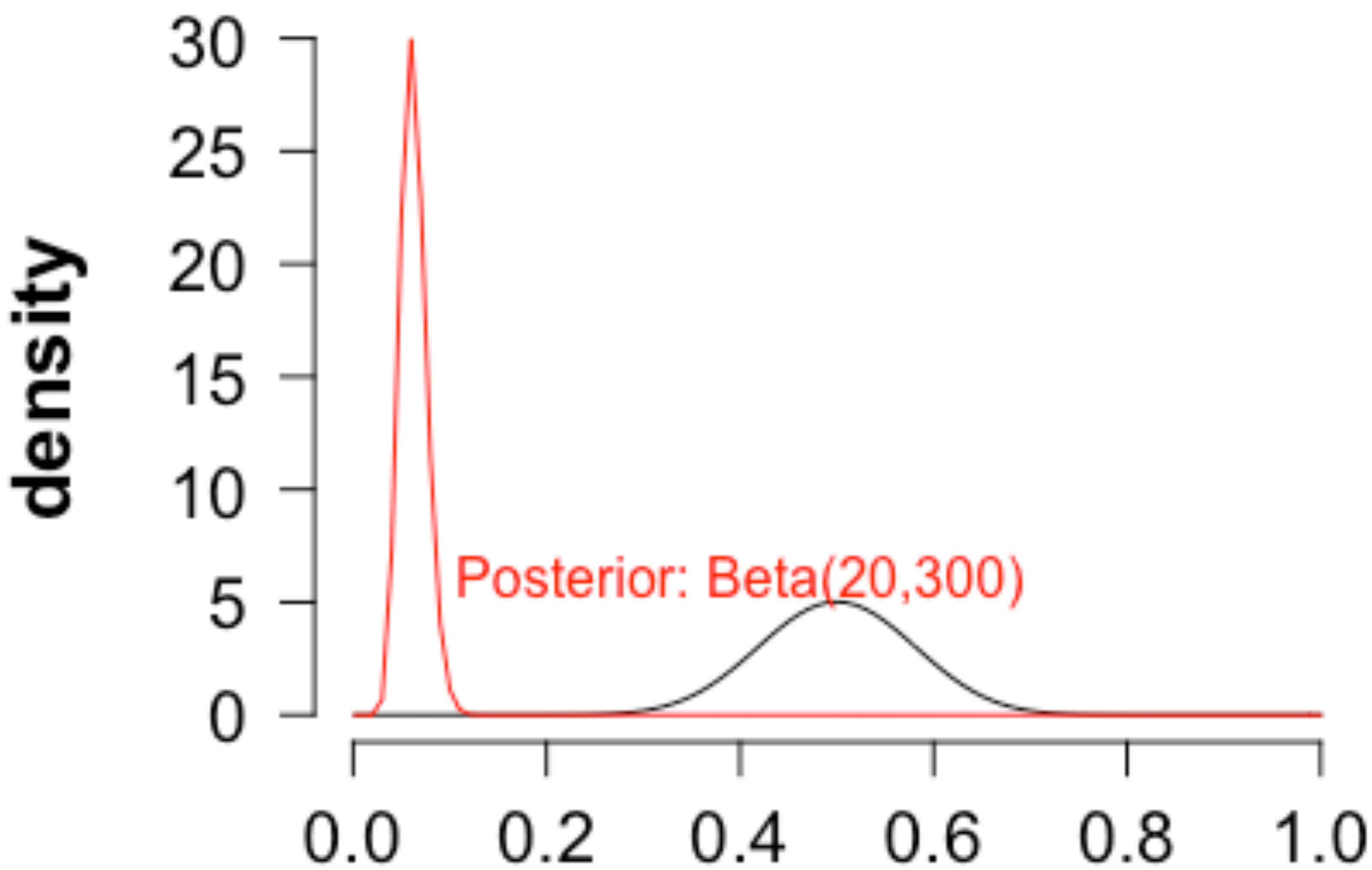
$$Posterior \propto Likelihood \times Prior$$



Example: Modeling complications after surgery

The data: 0 complications in the next **300** operations.

The **posterior distribution** of the probability of complications:



Summary

The posterior is a compromise between
the prior and the data

When data are sparse, the posterior reflects the prior

When a lot of data is available, the posterior reflects the likelihood

The prior can have a **regularizing** function

Hypothesis testing using the Bayes factor

We may want to compare two alternative models:

$$Model\ 1: \text{Probability of complications} = 0.5$$

$$Model\ 2: \text{Probability of complications} \sim Beta(1,1)$$

Bayes factor: $$BF_{12} = \frac{Prob(Data \mid Model\ 1)}{Prob(Data \mid Model\ 2)}$$

Hypothesis testing using the Bayes factor

$Model\ 1$: Probability of complications $= 0.5$

$$\binom{n}{k}\theta^0(1-\theta)^{10} = \binom{10}{0}0.5^{10} = 0.000977$$

$Model\ 2$: Probability of complications $\sim Beta(1,1)$

```
# Calculus using R:
plik1 <- function(theta) {
  dbinom(x = 0, size = 10, prob = theta) *
    dbeta(x = theta, shape1 = 1, shape2 = 1) }
# Then we integrate (compute the area under the curve):
(MargLik1 <- integrate(f = plik1, lower = 0, upper = 1)$value)
library(MASS)
fractions(MargLik1)
1/11
```

Hypothesis testing using the Bayes factor

$$Model\ 1: \text{Probability of complications} = 0.5$$

$$\binom{n}{k}\theta^0(1-\theta)^{10} = \binom{10}{0}0.5^{10} = 0.000977$$

$$Model\ 2: \text{Probability of complications} \sim Beta(1,1)$$

(see R code) $\dfrac{1}{11}$

$$BF_{12} = \frac{Prob(Data\,|\,Model\ 1)}{Prob(Data\,|\,Model\ 2)} = \frac{0.000977}{1/11} = 0.01$$

Model 2 is 100 times more likely than Model 1

Comparison of Frequentist vs Bayesian approaches

	Frequentist	Bayesian
Parameters	Fixed	Random
Data	Random	Fixed *
Prior knowledge used	No	Yes
Type I, II error	relevant	irrelevant *
Hypothesis testing	reject null	Bayes factor
Uncertainty quantification	No	Yes

* Under the classical Bayesian view

Some advantages of the Bayesian approach

1. Handles sparse data **robustly** given priors
2. Highly **customised** models can be defined
3. The focus is on **uncertainty quantification**
4. Answers the research question **directly**

Some “disadvantages” of the Bayesian approach

1. **You have to understand what you are doing**
   - Distribution theory
   - Random variable theory
   - Maximum likelihood estimation
   - Linear modeling theory
2. **Requires programming ability**
   - Statistical computing using Stan (mc-stan.org)
3. **Computational cost**
   - Cluster computing is sometimes needed
4. **Priors require thought**
   - Eliciting priors from experts
   - Adversarial/sensitivity analyses
5. **Low “power” an ever-present danger!**

Statistics is not a magic bullet, despite what you see in psychology and psycholinguistic journals.

A complete example of a Bayesian analysis

Self-paced reading data:
subject vs. object relatives in Chinese (Gibson & Wu 2013)

row	subj	item	so	rt
1	1	13	o	1561
2	1	6	s	959
3	1	5	o	582
4	1	9	o	294
5	1	14	s	438
6	1	4	s	286
⋮	⋮	⋮	⋮	
547	9	11	o	350

A complete example of a Bayesian analysis

Self-paced reading data:
subject vs. object relatives in Chinese (Gibson & Wu 2013)

Standard lmer syntax:

```
lmer(log(rt) ~ so + (1+so|subj) + (1+so|item),dat)
```

A complete example of a Bayesian analysis

Self-paced reading data:
subject vs. object relatives in Chinese (Gibson & Wu 2013)

The underlying mathematical model:

lmer(log(rt) ~ so + (1+so|subj + (1+so|item),dat)

$$rt_n \sim LogN(\alpha + u_1, subj[n] + w_1, item[n] + so_n \cdot (u_2, subj[n] + w_2, item[n]), \sigma)$$

$$\begin{pmatrix} u_{i,1} \\ u_{i,2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma_u}\right) \qquad \boldsymbol{\Sigma_u} = \begin{pmatrix} \tau_{u_1}^2 & \rho_u \tau_{u_1} \tau_{u_2} \\ \rho_u \tau_{u_1} \tau_{u_2} & \tau_{u_2}^2 \end{pmatrix}$$

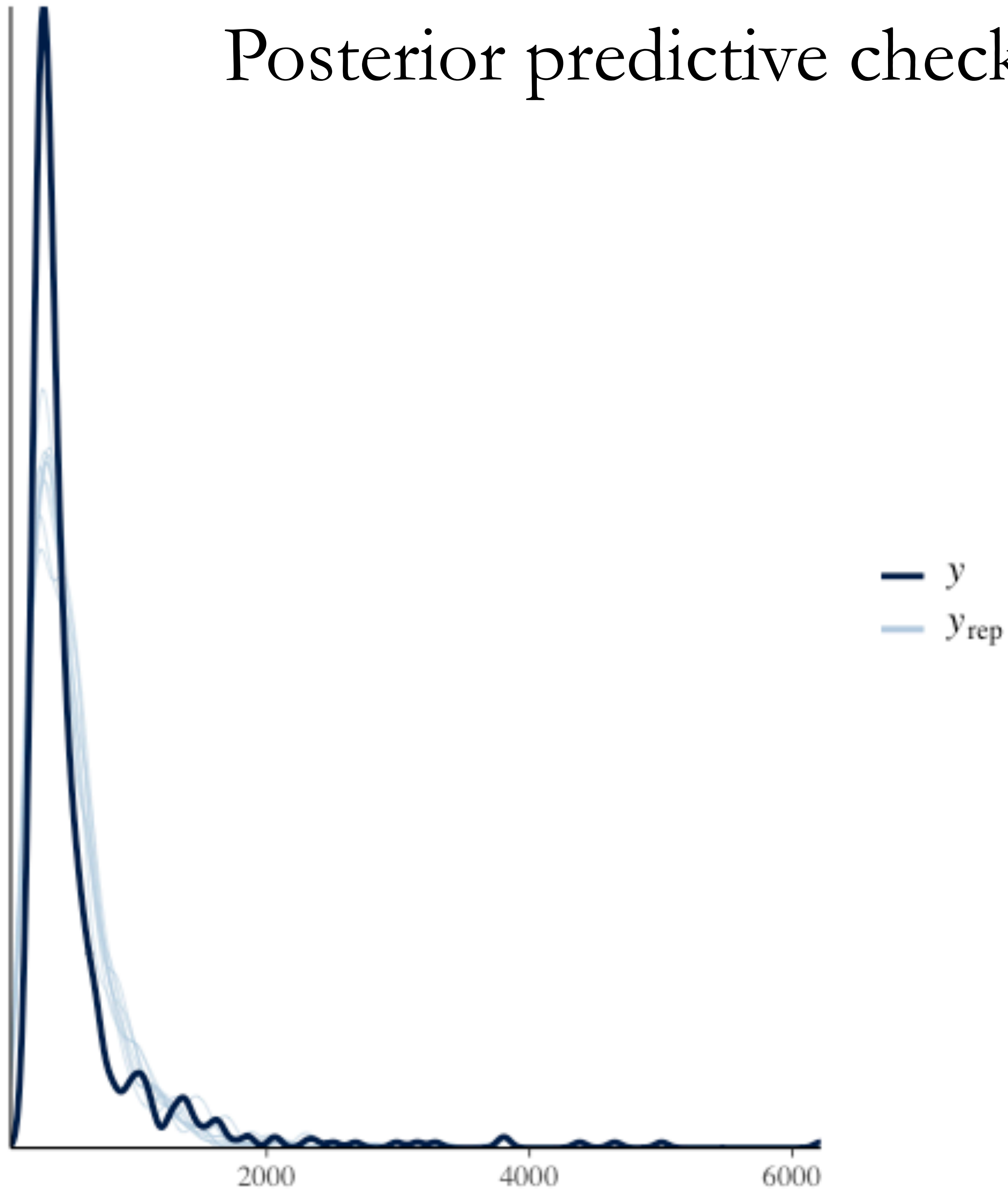
$$\begin{pmatrix} w_{j,1} \\ w_{j,2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma_w}\right) \qquad \boldsymbol{\Sigma_w} = \begin{pmatrix} \tau_{w_1}^2 & \rho_w \tau_{w_1} \tau_{w_2} \\ \rho_w \tau_{w_1} \tau_{w_2} & \tau_{w_2}^2 \end{pmatrix}$$

A complete example of a Bayesian analysis

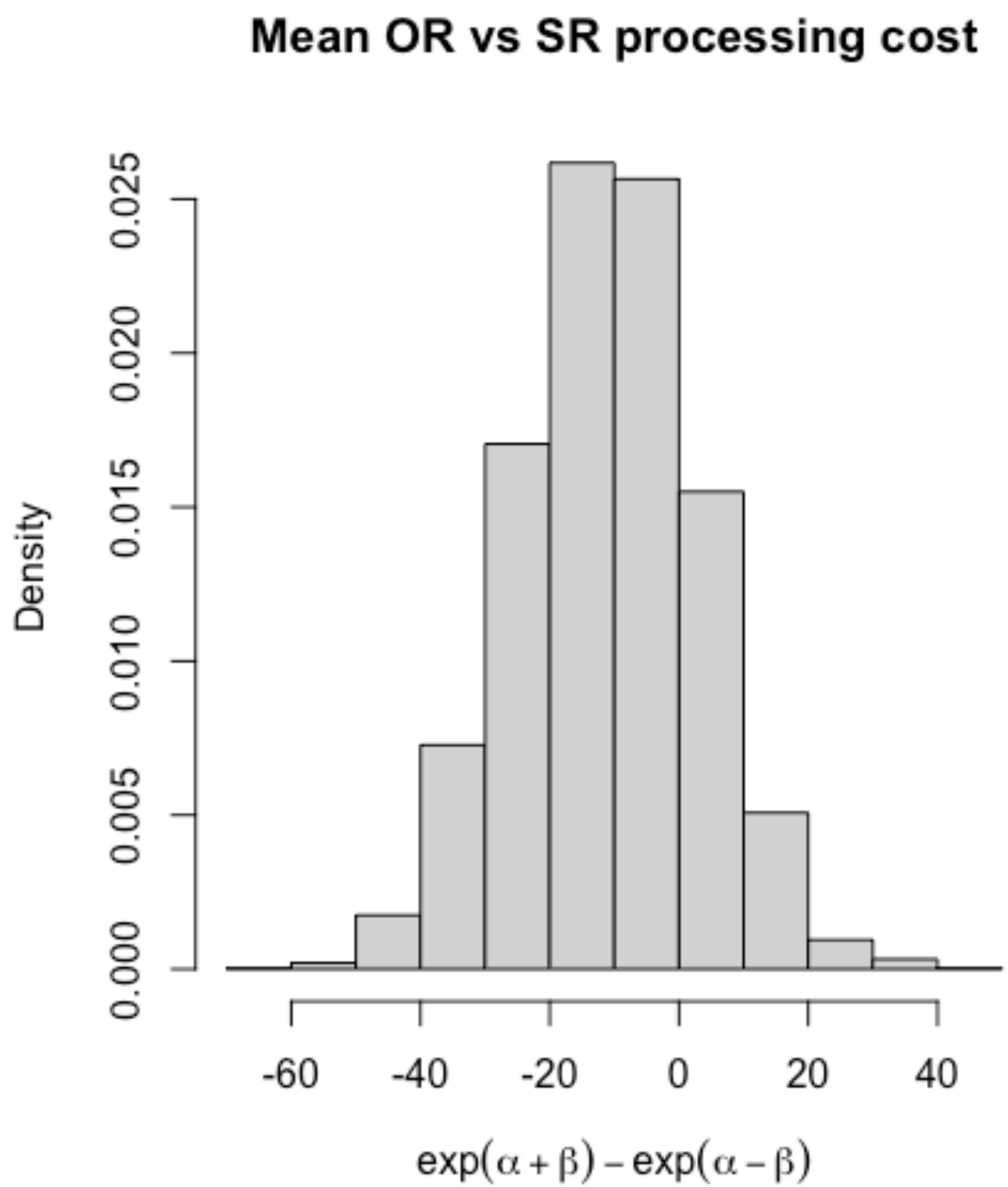
```
library(brms)
fit <- brm(rt ~ 1+so + (1+so | subj) + (1+so|item),
           family = lognormal(),
           prior =
             c(prior(normal(6, 1.5), class = Intercept),
               prior(normal(0, .01), class = b),
               prior(normal(0, 1), class = sigma),
               prior(normal(0, 1), class = sd),
               prior(lkj(2), class = cor)),
           iter = 4000,
           data = dat)
```

A complete example of a Bayesian analysis

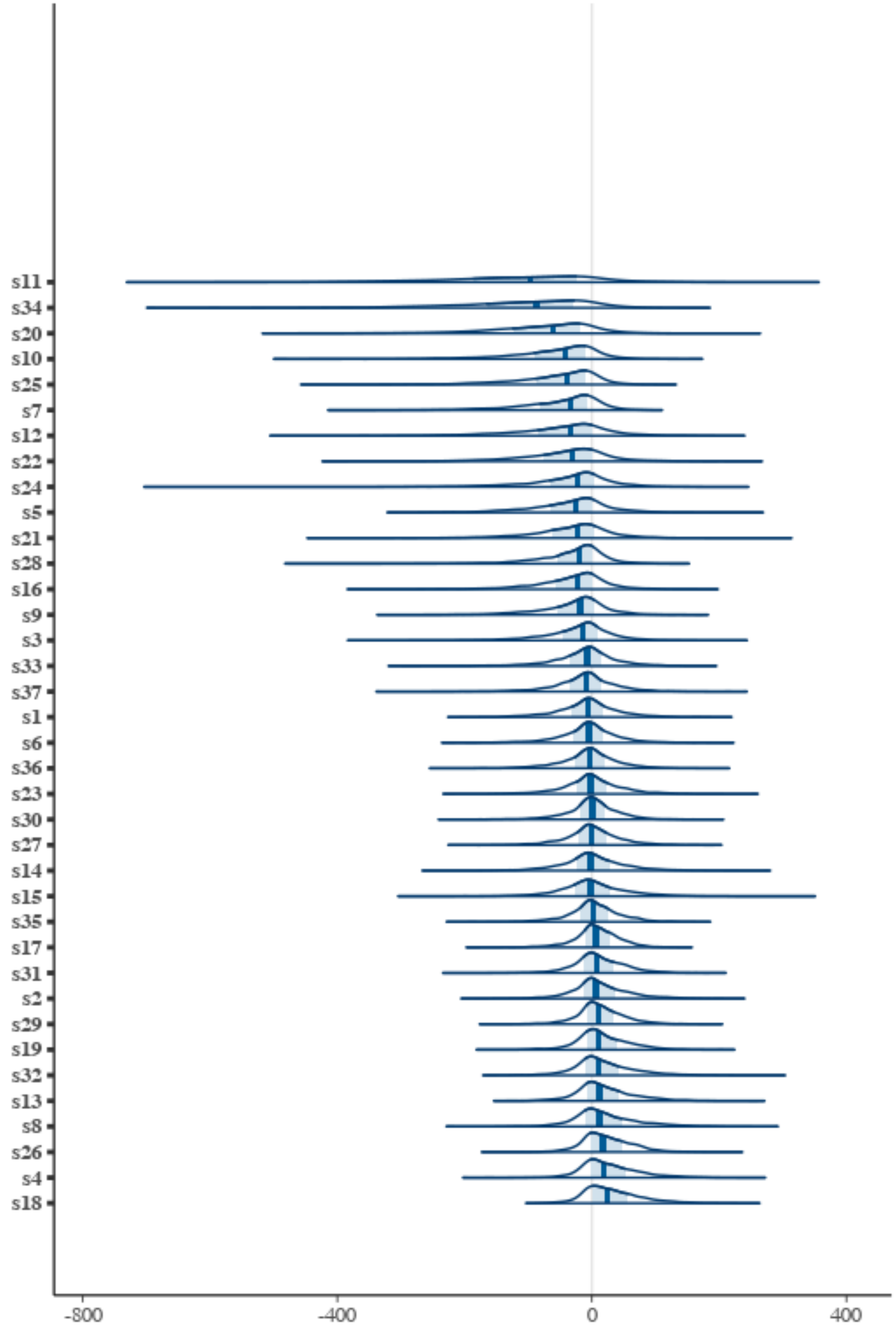
Posterior predictive check



A complete example of a Bayesian analysis



A complete example of a Bayesian analysis



A complete example of a Bayesian analysis

Bayes factor analysis:

```
m_gw<-brm(rt~1+so + (1+so|subj) +(1+so item),
          data=dat,family=lognormal(),
         prior=priors,warmup=5000,
         iter=20000,
         save_all_pars = TRUE)

m_gw0<-brm(rt~1 + (1+so|subj)+(1+so|item),
          data=dat,family=lognormal(),
          prior=priors0,warmup=5000,iter=20000,
          save_all_pars = TRUE)
```

A complete example of a Bayesian analysis

Bayes factor analysis:

bayes_factor(m_gw,m_gw0)

Estimated Bayes factor in favor of m_gw over m_gw0:
1.08125

Conclusion: No evidence for difference between RC types

Homework for Thursday

Carry out a Bayes factor analysis for the second data-set provided: gibson_wu2.

This data-set is intended to be an exact replication of the original study.

Common concerns about Bayesian methods

A post on twitter:

> Doing statistics should be like going to the bathroom. Yes, you have to do it. Yes, when you do it, you want to do it right. But don't make a big deal out of it, be careful about telling other people how to do it, and if your whole life is centered on it, there's something wrong.

I think that the main problem here is that we have been taught in school that statistics should be as easy as:

t.test(diff)

It's not.
It's hard to unlearn this lesson.

Common concerns about Bayesian methods

A post on twitter:

> Doing statistics should be like going to the bathroom. Yes, you have to do it. Yes, when you do it, you want to do it right. But don't make a big deal out of it, be careful about telling other people how to do it, and if your whole life is centered on it, there's something wrong.

We can make data analysis as easy as going to the bathroom, but we should not be surprised if what comes out is crap.

General complaints about statistical methods

Another post on twitter
(written by the acting director of ZAS, Berlin):

> Over the last 20 years where I am in research, statistical models have changed so much and have questioned many results that I sometimes think, the most reliable way is to report descriptive statistics.
>
> 11:50 PM · Apr 20, 2022 · Twitter Web App

This would be like saying:

Over the years, syntactic theory has changed so much and questioned so many results that the most reliable way is to just describe the syntax of a language.

Examples of quick analyses that went wrong

Gibson and Wu 2013:

41,285, $p < .05$; $F_2(1, 14) = 2.23$, $MS_{within} = 22{,}120$, $p = .16$]. The next word consisted of the head noun for the RC, N2. This region was read more slowly in the SRC condition [$F_1(1, 36) = 6.92$, $MS_{within} = 280{,}810$, $p = .01$; $F_2(1, 14) = 4.62$, $MS_{within} = 110{,}132$, $p < .05$]. When these two regions – the RC

Levy and Keller 2013:

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

The statistical significance filter leads to overoptimistic expectations of replicability

Shravan Vasishth[a,*], Daniela Mertzen[a], Lena A. Jäger[a], Andrew Gelman[b]

[a] Department of Linguistics, University of Potsdam, Potsdam, Germany
[b] Department of Statistics, Columbia University, New York, USA

Recommended readings

Textbooks

- Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Elsevier.
- McElreath, R. (2020). Statistical rethinking: A Bayesian course with examples in R and Stan. CRC Press.
- Lambert, B. (2018). A student's guide to Bayesian statistics. Sage.
- Our book: https://vasishth.github.io/bayescogsci/book/

Recommended readings

Textbooks that assume calculus

- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.
- Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer Science & Business Media.

Further suggested readings

https://vasishth.github.io/IntroductionBayes/

Almost anything written by:

- Wagenmakers
- Rouder
- Morey

Further suggested readings

Bayes factors tutorial

Statistics > Methodology

[Submitted on 15 Mar 2021 (v1), last revised 18 Mar 2021 (this version, v2)]

Workflow Techniques for the Robust Use of Bayes Factors

Daniel J. Schad, Bruno Nicenboim, Paul-Christian Bürkner, Michael Betancourt, Shravan Vasishth

Bayesian workflow tutorial

Statistics > Methodology

[Submitted on 29 Apr 2019 (v1), last revised 28 Feb 2020 (this version, v3)]

Toward a principled Bayesian workflow in cognitive science

Daniel J. Schad, Michael Betancourt, Shravan Vasishth

Annual summer school in Statistical Methods for Linguistics and Psychology at Potsdam

https://vasishth.github.io/smlp2021/