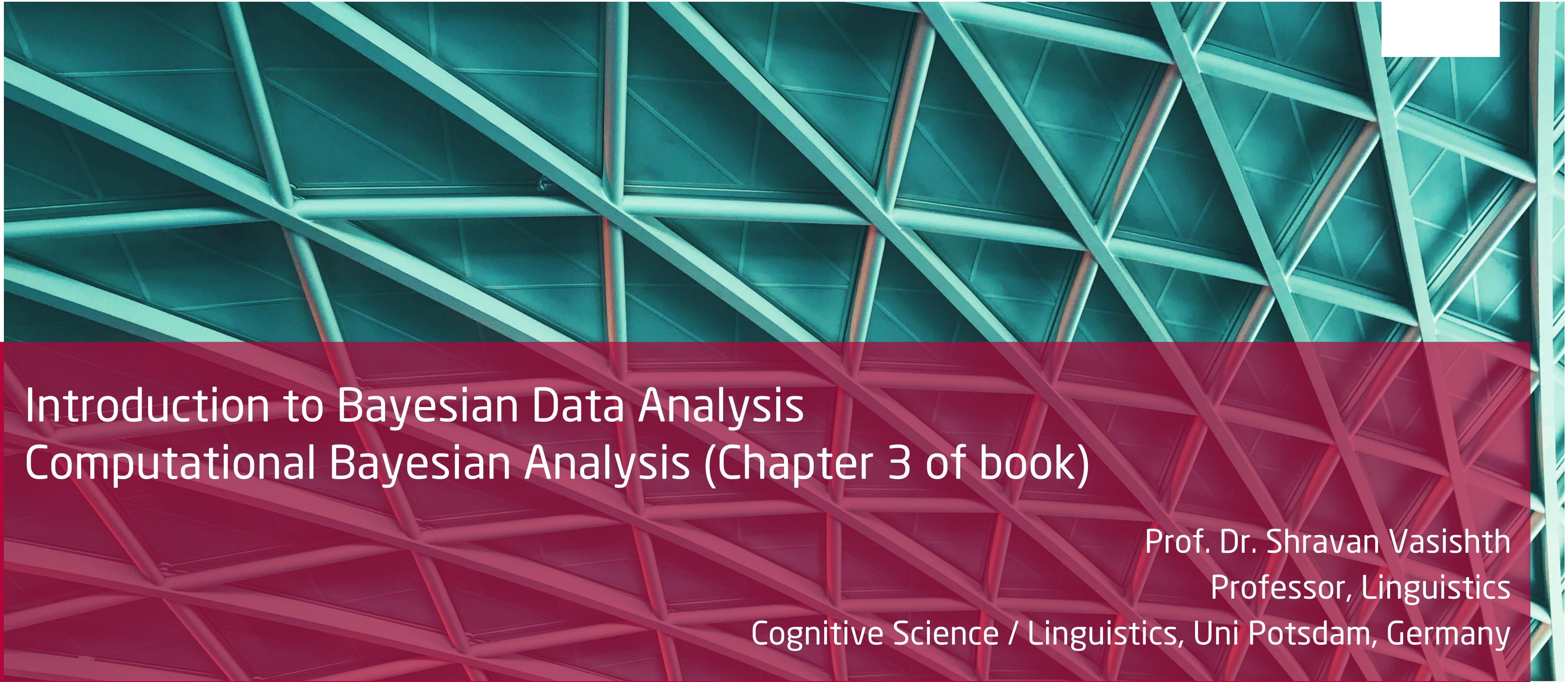


Introduction to Bayesian Data Analysis
Computational Bayesian Analysis (Chapter 3 of book)

Prof. Dr. Shravan Vasishth
Professor, Linguistics
Cognitive Science / Linguistics, Uni Potsdam, Germany

Quick review: Bayes' rule

A and B are discrete events. Bayes' rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{1}$$

Bayes' rule with distributions ($\boldsymbol{\Theta}$ can be a vector of parameters):

$$\begin{aligned} p(\boldsymbol{\Theta}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta})}{p(\boldsymbol{y})} \\ p(\boldsymbol{\Theta}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta})}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{y}|\boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}) d\boldsymbol{\Theta}} \end{aligned} \tag{2}$$

- $p(\boldsymbol{\Theta}|\boldsymbol{y})$ is the (joint) **posterior distribution** of the parameters given the data
- The posterior distribution is of central interest to us

Quick review: An example with the Poisson-Gamma conjugate case

Recall that in the Poisson-Gamma conjugate case, we had:

- The Poisson likelihood: $\frac{\exp(-n\lambda)\lambda^{\sum_i^n x_i}}{\prod_{i=1}^n x_i!}$
- A prior for the parameter: $\lambda \sim Gamma(a, b)$

$$\text{Posterior} = \underset{\substack{\uparrow \\ Likelihood}}{\left[\frac{\exp(-n\lambda)\lambda^{\sum_i^n x_i}}{\prod_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i}!}\right]} \underset{\substack{\uparrow \\ Prior}}{\left[\frac{\mathbf{b^a}\lambda^{a-1}\exp(-b\lambda)}{\mathbf{\Gamma(a)}}\right]} \tag{3}$$

$$\begin{aligned}\text{Posterior} &\propto \exp(-n\lambda)\lambda^{\sum_i^n x_i}\lambda^{a-1}\exp(-b\lambda) \\ &= \lambda^{a-1+\sum_i^n x_i}\exp(-\lambda(b+n))\end{aligned} \tag{4}$$

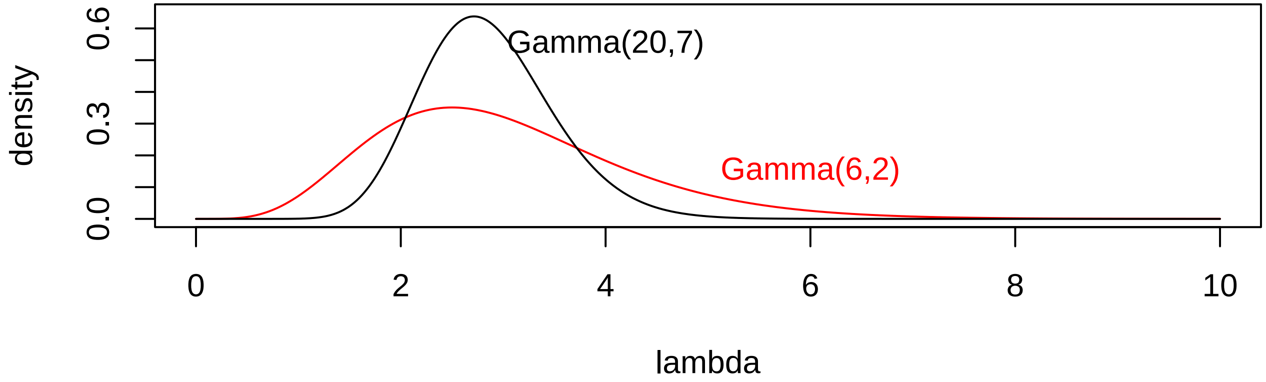
The above expression has the kernel of a Gamma distribution with new parameters $a* = a + \sum x$, and $b* = b + n$.

Posterior samples

Example:

- $\lambda \sim Gamma(a = 6, b = 2)$
- Data: 2, 4, 3, 6, 1
- Posterior: $\lambda \sim Gamma(20, 7)$

The prior and posterior seen together:



Posterior samples

We can draw inferences from the posterior analytically:

```
## 95% credible interval
qgamma(c(0.025,0.975),shape=20,rate=7)

## [1] 1.745217 4.238693
```

But if we just had a large number of samples from this distribution, we could have used those samples to draw essentially the same conclusions:

```
lambda_posterior<-rgamma(4000,shape=20,rate=7)
quantile(lambda_posterior,c(0.025,0.975))

##     2.5%    97.5%
## 1.745801 4.193481
```

That is what I mean by obtaining samples from the posterior distributions.

Our main goal: Obtaining posterior samples

Our main goal is to obtain this posterior distribution:

$$p(\boldsymbol{\Theta}|\boldsymbol{y})$$

- In the beta-binomial and Poisson-gamma cases, we could derive the posterior analytically.
- In more complex Bayesian models, we need to use some sampling method (MCMC sampling) to obtain **posterior samples** of the parameter(s).
- An example will help.

Example 1: A simple linear model

The data:

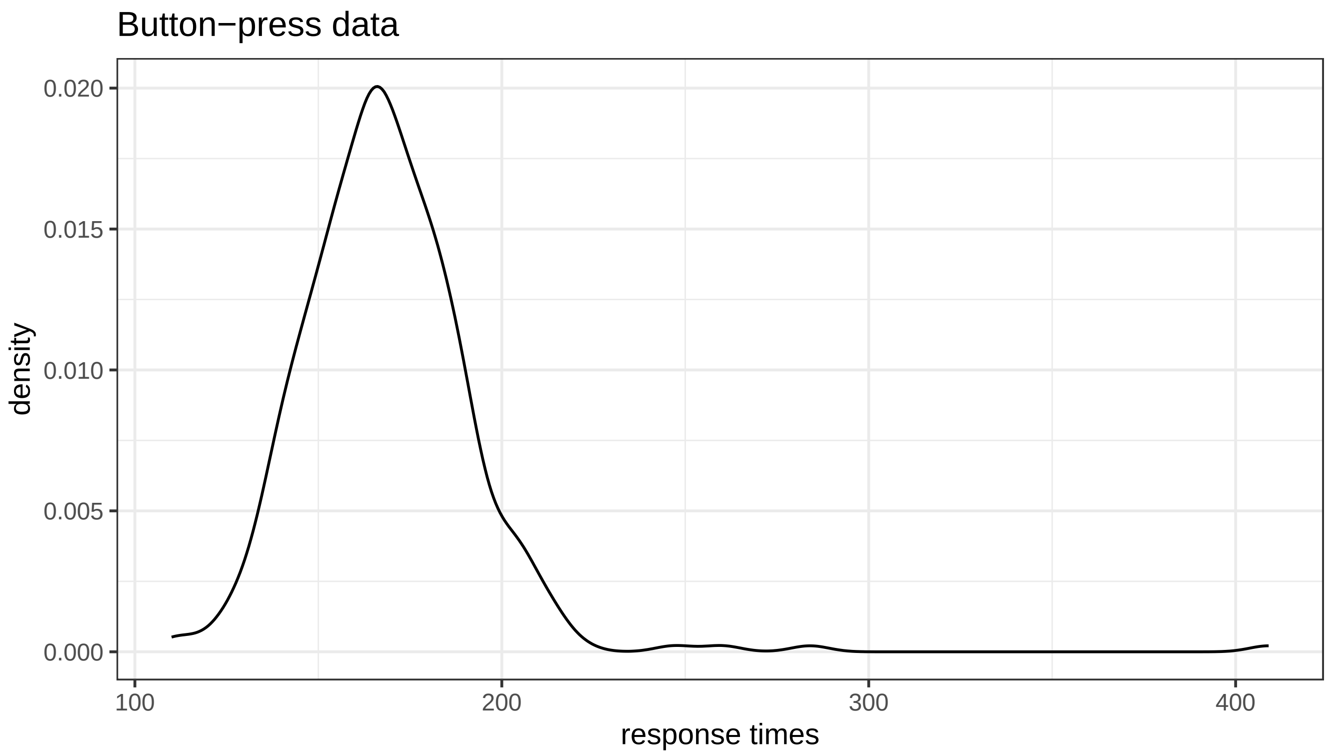
- Suppose we have data from a single subject repeatedly pressing the space bar as fast as possible, without paying attention to any stimuli.

```
library(bcogsci)
data("df_spacebar")
head(df_spacebar,n=2)
```

```
##      t trial
## 1 141     1
## 2 138     2
```

- The data are response times in milliseconds in each trial.
- We would like to know how long it takes to press a key for this subject.

Example 1: Visualize the data



Example 1: A first attempt at a statistical model

Assume this statistical model (n: the n-th row in the data frame):

$$t_n \sim Normal(\mu, \sigma) \tag{5}$$

We unpack this model next.

Example 1: A first attempt at a statistical model

Assume this statistical model (n: the n-th row in the data frame):

$$t_n \sim Normal(\mu, \sigma) \tag{6}$$

Example 1: A simple linear model

$$t_n \sim Normal(\mu, \sigma) \tag{7}$$

$$t_n = \mu + \varepsilon, \text{ where } \varepsilon_n \stackrel{iid}{\sim} Normal(0, \sigma) \tag{8}$$

Assumptions:

1. There is a true (unknown) underlying time, μ ms, that the subject needs to press the space bar.
2. There is some noise in this process.
3. The noise is normally distributed (this assumption is questionable given that response times are generally skewed; we will fix this assumption later).

Example 1: A simple linear model

A frequentist linear model:

```
m<-lm(t~1,df_spacebar)
coef(m)

## (Intercept)
##    168.6399

sigma(m)

## [1] 24.9118

mean(df_spacebar$t)

## [1] 168.6399

sd(df_spacebar$t)

## [1] 24.9118
```

Example 1: A simple linear model

There are two parameters here that we have MLEs of:

- μ (the intercept, mean button pressing time)
- σ the standard deviation of the residual noise
- μ, σ are fixed, unknown point values in frequentist models.
- In a Bayesian model, μ, σ are random variables and need prior distributions specified for them.

Example 1: Prior specification

Let's start with (unrealistic) flat priors:

$$\mu \sim Uniform(0, 60000)$$
$$\sigma \sim Uniform(0, 2000) \tag{9}$$

What beliefs are these prior distributions expressing?

Example 1: Fitting the brms model

```
fit_press <- brm(t ~ 1,
  data = df_spacebar,
  family = gaussian(),
  prior = c(
    prior(uniform(0, 60000), class = Intercept, lb = 0,
          ub = 60000),
    prior(uniform(0, 2000), class = sigma, lb = 0,
          ub = 2000)
  ),
  chains = 4,
  iter = 2000,
  warmup = 1000
)
```

Example 1: Fitting the brms model

The components of the code:

- The model specification:

```
brm(t ~ 1,data = df_spacebar)
```

- The likelihood assumed:

```
family = gaussian()
```

- The prior specification:

```
prior = c(
  prior(uniform(0, 60000), class = Intercept),
  prior(uniform(0, 2000), class = sigma)
)
```

- Sampling specifications:

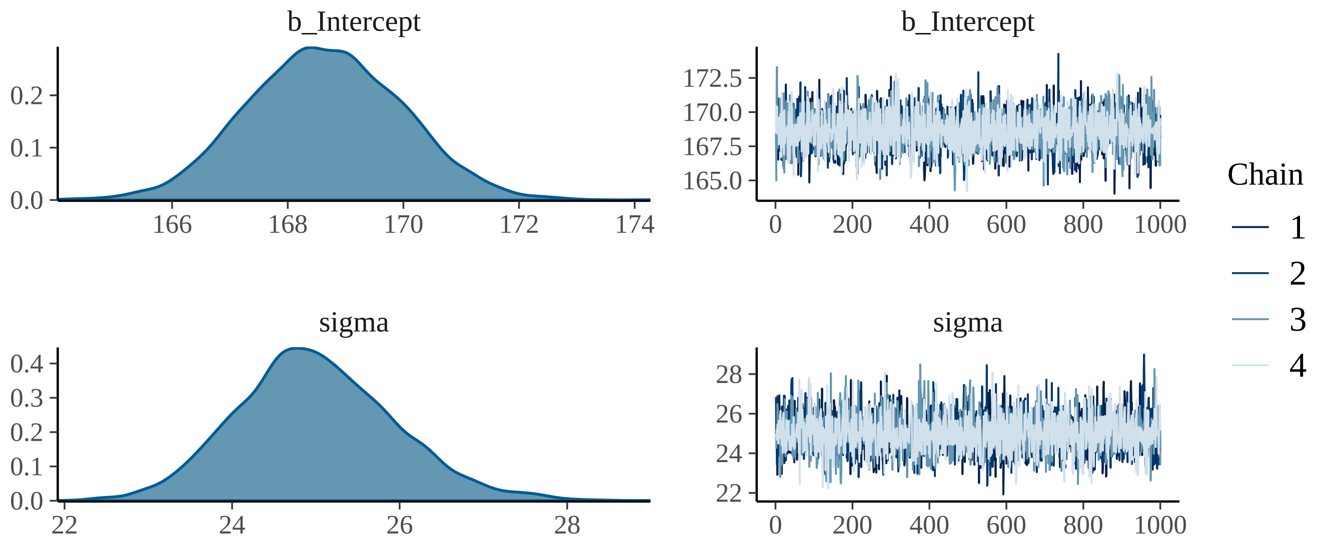
```
chains = 4,
iter = 2000,
warmup = 1000
```

Example 1: Chains, iterations, warm-up

1. The term ‘chains’ refers to the number of independent runs for sampling (by default four).
2. The term ‘iter’ refers to the number of iterations that the sampler makes to sample from the posterior distribution of each parameter (by default 2000).
3. The term ‘warmup’ refers to the number of iterations from the start of sampling that are eventually discarded (by default half of ‘iter‘). (in WinBUGS/JAGS this is called burn-in)

Example 1: Visualize the posteriors

```
plot(fit_press)
```



Example 1: Visualize the posteriors

Also, try this using shinystan:

```
library(shinystan)
launch_shinystan(fit_press)
```

Example 1: Extract the posteriors from the model and compute summary statistics

```
as_draws_df(fit_press) %>% head(3)

## # A draws_df: 3 iterations, 1 chains, and 4 variables
##   b_Intercept sigma lprior  lp__
## 1         169    25    -19 -1683
## 2         169    25    -19 -1683
## 3         167    24    -19 -1683
## # ... hidden reserved variables {'.chain', '.iteration', '.draw'}

as_draws_df(fit_press)$b_Intercept %>% mean()

## [1] 168.6456

as_draws_df(fit_press)$b_Intercept %>% quantile(c(0.025, .975))

##     2.5%    97.5%
## 166.0983 171.2367
```

Example 1: Extract the posteriors from the model and compute summary statistics

```
as_draws_df(fit_press)$sigma %>% mean()

## [1] 25.00157

as_draws_df(fit_press)$sigma %>% quantile(c(0.025, .975))

##     2.5%    97.5%
## 23.28623 26.96696
```

Next steps

The next important topics I will discuss are

- Prior predictive distributions
- Posterior predictive distributions

Example 1: Prior predictive distributions

The model specification again:

$$
\begin{aligned}
\mu &\sim Uniform(0, 60000) \\
\sigma &\sim Uniform(0, 2000)
\end{aligned} \tag{10}
$$

$$
t_n \sim Normal(\mu, \sigma) \tag{11}
$$

We can generate the **prior predictive distribution** given the above model ($\boldsymbol{\Theta} =< \mu, \sigma >$):

$$
\begin{aligned}
p(\boldsymbol{y_{pred}}) &= p(y_{pred_1}, \dots, y_{pred_n}) \\
&= \int_{\boldsymbol{\Theta}} p(y_{pred_1}|\boldsymbol{\Theta}) \cdot p(y_{pred_2}|\boldsymbol{\Theta}) \cdots p(y_{pred_N}|\boldsymbol{\Theta}) p(\boldsymbol{\Theta}) \, d\boldsymbol{\Theta}
\end{aligned} \tag{12}
$$

Example 1: Prior predictive distributions

Repeat the following many times:

1. Take one sample from each of the priors.

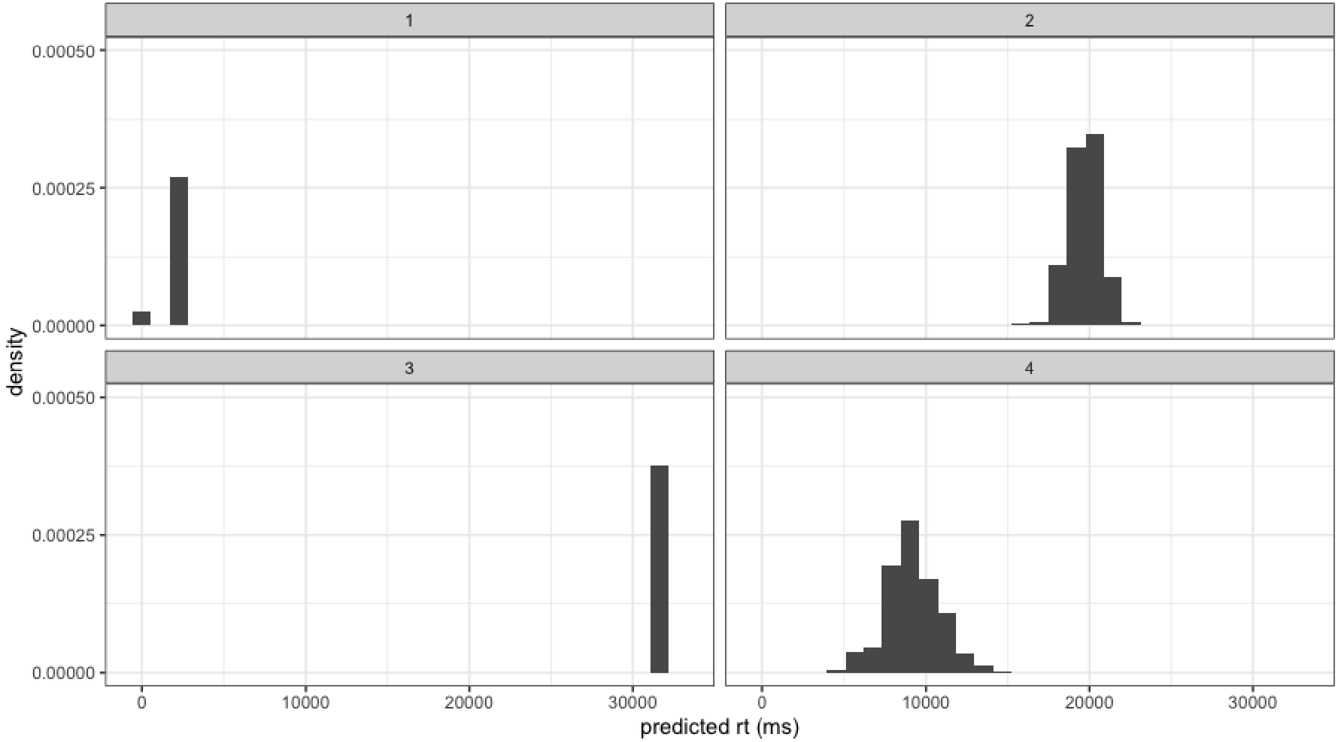
```
mu<-runif(1,min=0,max=60000)
sigma<-runif(1, 0, 2000)
```

2. Plug those samples into the probability density/mass function used as the likelihood in the model to generate a data set $y_{pred_1}, \ldots, y_{pred_n}$.

```
y_pred_1<-rnorm(n=5,mu,sigma)
y_pred_1
## [1] 60881.21 58260.20 58940.50 59362.79 57728.92
```

- Each sample is an imaginary or potential data set.
- In the textbook, you will find code for generating prior predictive data using R.

Example 1: Visualizing prior predictive distributions



Not very realistic button pressing time distributions!

Example 1: Reconsidering our options regarding the prior specifications

What are our options regarding the priors?
In our book, we classify priors as follows (these constitute a continuum):

- Flat, uninformative priors: E.g., $\mu \sim Uniform(-10^{20}, 10^{20})$
- Regularizing priors: E.g., $\mu \sim Normal_{+}(0, 1000)$
- Principled priors: E.g., $\mu \sim Normal_{+}(250, 100)$
- Informative priors: E.g., $\mu \sim Normal_{+}(200, 20)$

There is no standard terminology for types of priors!

Example 1: Reconsidering our options regarding the prior specifications

Next: Sensitivity analysis using different priors.

Example 1: Reconsidering our options regarding the prior specifications

Let's refit the model with flat, uninformative priors:

$$\begin{aligned} \mu &\sim Uniform(-10^6, 10^6) \\ \sigma &\sim Uniform(0, 10^6) \end{aligned} \tag{13}$$

Example 1: Reconsidering our options regarding the prior specifications

Earlier model's estimates:

```
as_draws_df(fit_press)$b_Intercept %>% quantile(c(0.025, .975))

##      2.5%    97.5%
## 166.0983 171.2367
```

With our new flat, uninformative priors:

```
as_draws_df(fit_press_unif)$b_Intercept %>% quantile(c(0.025, .975))

##      2.5%    97.5%
## 165.9976 171.2361
```

Example 1: Reconsidering our options regarding the prior specifications

Let's refit the model with very informative priors:

$$
\begin{aligned}
\mu &\sim Normal(400, 10) \\
\sigma &\sim Normal_{+}(100, 10)
\end{aligned}
\tag{14}
$$

Example 1: Reconsidering our options regarding the prior specifications

Compare the posteriors:

```
as_draws_df(fit_press)$b_Intercept %>% quantile(c(0.025, .975))

##      2.5%    97.5%
## 166.0983 171.2367

as_draws_df(fit_press_unif)$b_Intercept %>% quantile(c(0.025, .975))

##      2.5%    97.5%
## 165.9976 171.2361

as_draws_df(fit_press_inf)$b_Intercept %>% quantile(c(0.025, .975))

##      2.5%    97.5%
## 170.2847 175.6845
```

Example 1: Reconsidering our options regarding the prior specifications

Let's refit the model with principled priors:

$$\begin{aligned} \mu &\sim Normal(200, 100) \\ \sigma &\sim Normal_{+}(50, 50) \end{aligned} \tag{15}$$

Example 1: Reconsidering our options regarding the prior specifications

```
as_draws_df(fit_press)$b_Intercept %>% quantile(c(0.025, .975))

##     2.5%    97.5%
## 166.0983 171.2367

as_draws_df(fit_press_unif)$b_Intercept %>% quantile(c(0.025, .975))

##     2.5%    97.5%
## 165.9976 171.2361

as_draws_df(fit_press_inf)$b_Intercept %>% quantile(c(0.025, .975))

##     2.5%    97.5%
## 170.2847 175.6845

as_draws_df(fit_press_prin)$b_Intercept %>% quantile(c(0.025, .975))

##     2.5%    97.5%
## 166.0720 171.2418
```

Example 1: Reconsidering our options regarding the prior specifications

- This **sensitivity analysis** showed that the posterior is not overly affected by the choice of prior.
- Recall the earlier discussion: The posterior is a compromise between the prior and the likelihood.
- Informative prior with sparse data $\rightarrow$ the prior will dominate in determining the posterior.
- A lot of data $\rightarrow$ the likelihood will dominate in determining the posterior.

It is a good practice to carry out a sensitivity analysis (with increasing experience, you will know when this is absolutely necessary).

Example 1: Reconsidering our options regarding the prior specifications

Exercise:

- Try out some other priors.
- Produce prior predictive distributions with each prior specification to decide whether the priors make sense.
- A general strategy: use the prior predictive distribution to decide on reasonable priors.

Recommended: Read chapter 3 first!

Next topic

Next, we will look at posterior predictive distributions.

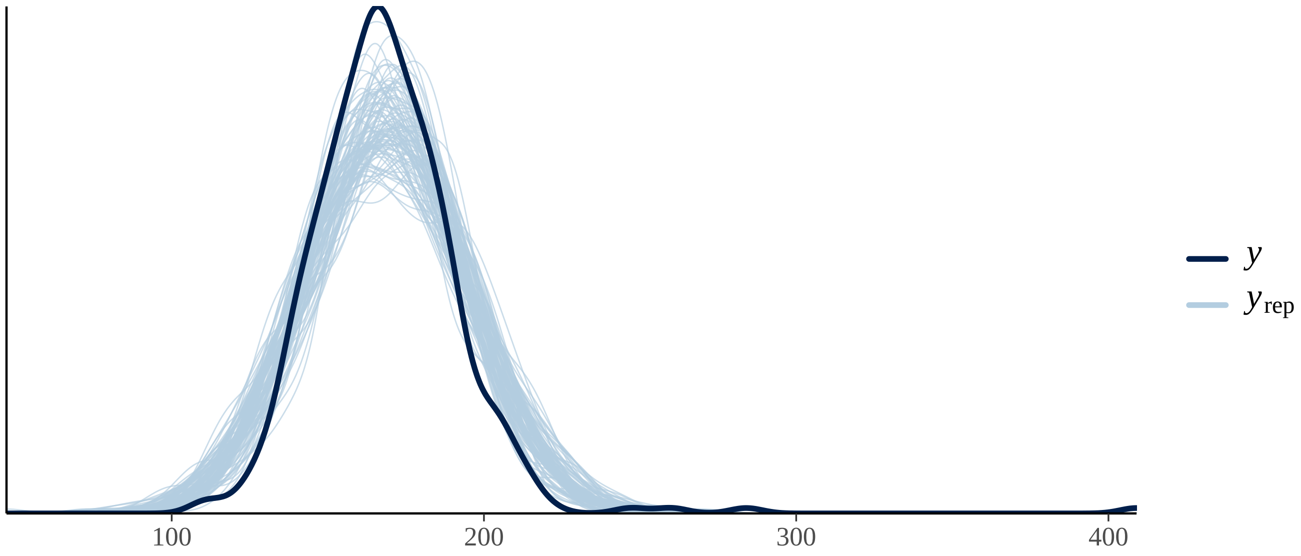
Example 1: Posterior predictive distribution

- The posterior predictive distribution is a collection of data sets generated from the model (the likelihood and the priors).
- Having obtained the posterior distributions of the parameters after taking into account the data, the posterior distributions can be used to generate future data from the model.
- In other words, given the posterior distributions of the parameters of the model, the posterior predictive distribution gives us some indication of what future data might look like.

$$p(\boldsymbol{y_{pred}} \mid \boldsymbol{y}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{y_{pred}}, \boldsymbol{\Theta} \mid \boldsymbol{y})\, d\boldsymbol{\Theta} = \int_{\boldsymbol{\Theta}} p(\boldsymbol{y_{pred}} \mid \boldsymbol{\Theta}, \boldsymbol{y}) p(\boldsymbol{\Theta} \mid \boldsymbol{y})\, d\boldsymbol{\Theta} \qquad (16)$$

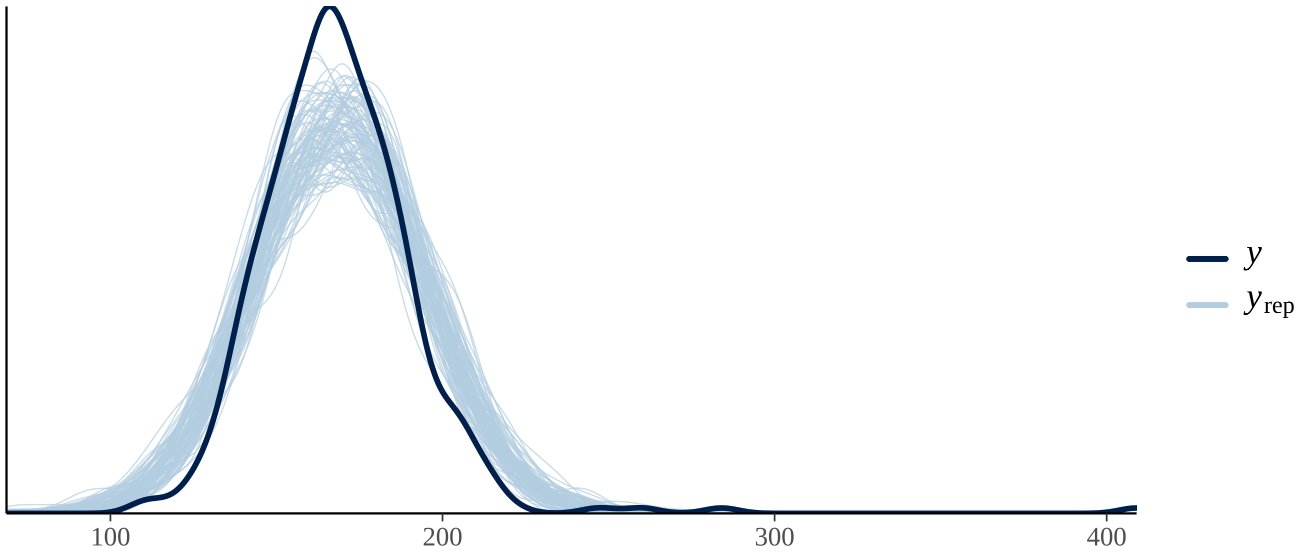
Example 1: Posterior predictive distribution

```r
pp_check(fit_press, ndraws = 100, type = "dens_overlay")
```



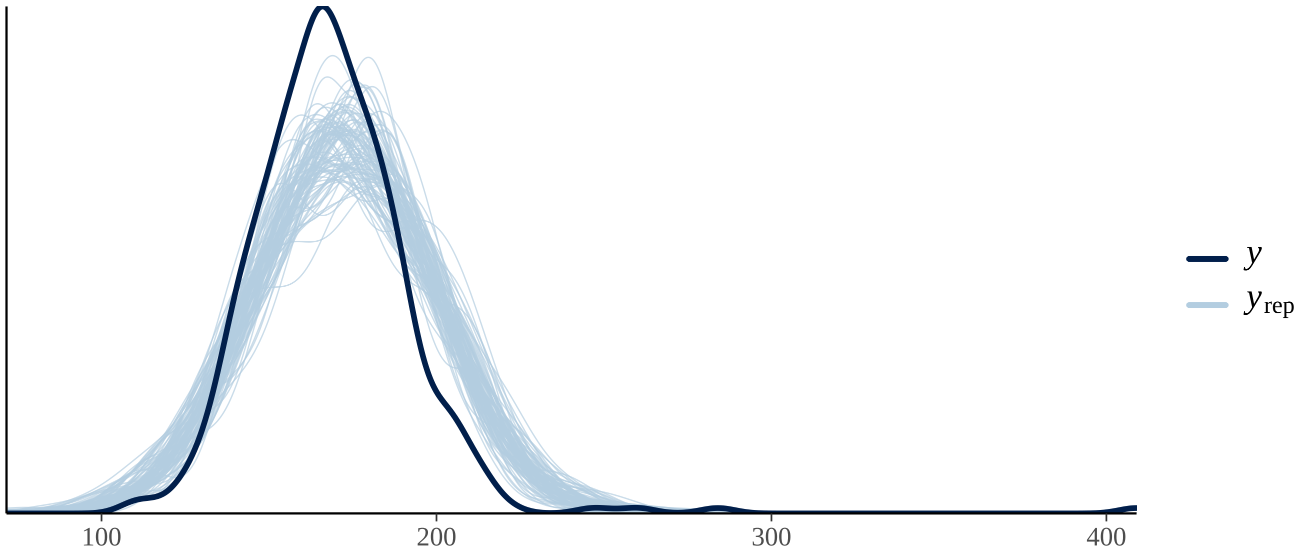
Example 1: Posterior predictive distribution

```r
pp_check(fit_press_unif, ndraws = 100, type = "dens_overlay")
```



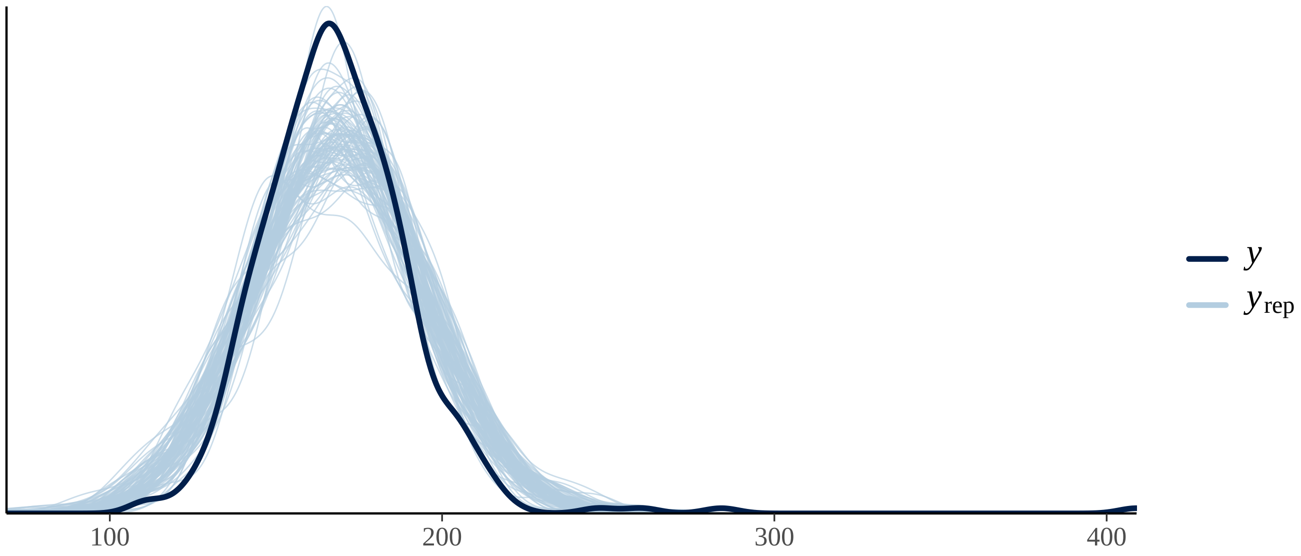
Example 1: Posterior predictive distribution

```
pp_check(fit_press_inf, ndraws = 100, type = "dens_overlay")
```



Example 1: Posterior predictive distribution

```
pp_check(fit_press_prin, ndraws = 100, type = "dens_overlay")
```



Example 1: Prior vs. posterior predictive distributions

- The prior predictive distribution shows us the predicted data given the priors and the likelihood, **before** the data were observed.
- The posterior predictive distribution shows us the predicted data given the priors and the likelihood, **after** the data were observed.
- Both are used for understanding whether the model makes sense for the research problem at hand.

Optional reading: Chapters 6 (Priors) and 7 (Workflow)

The next step: Improving the model

Next: using the log-normal instead of the normal likelihood.

Example 2: Using the log-normal likelihood

- If $\boldsymbol{y}$ is log-normally distributed, this means that $\log(\boldsymbol{y})$ is normally distributed.
- The log-normal distribution is also defined using the parameters location, μ, and scale, σ, but these are on the log ms scale.
- The relationship between the log-normal and the normal:

$$\begin{aligned} \log(\boldsymbol{y}) &\sim Normal(\mu, \sigma) \\ \boldsymbol{y} &\sim LogNormal(\mu, \sigma) \end{aligned} \tag{17}$$

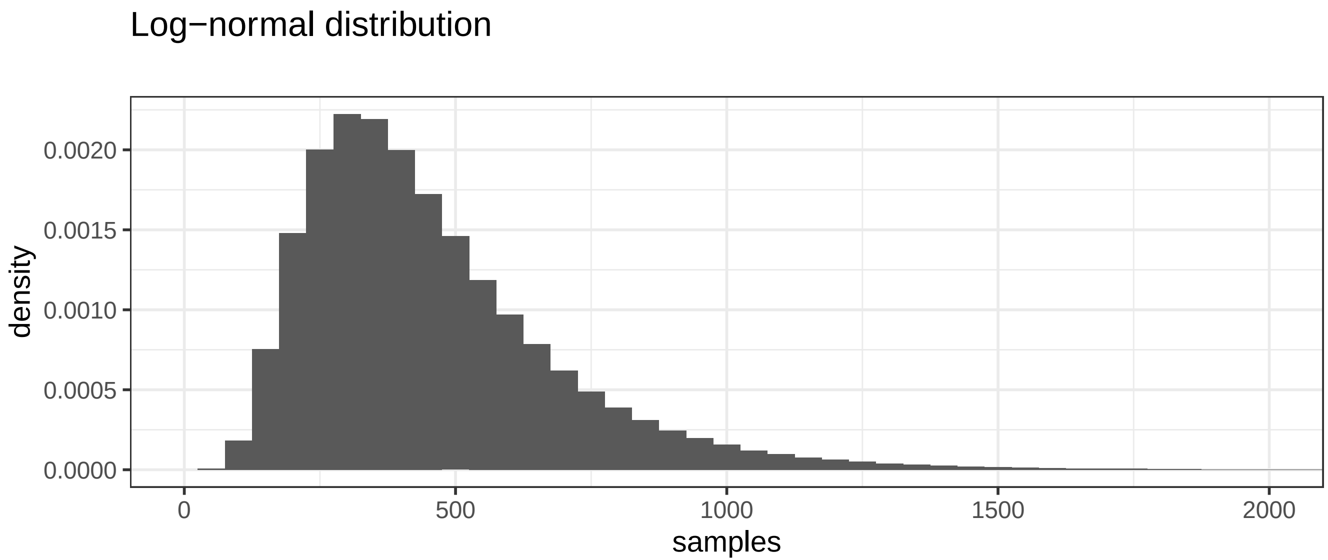
- We can obtain samples from the log-normal distribution, using the normal distribution by first setting an auxiliary variable, z, so that $z = \log(y)$. This means that $z \sim Normal(\mu, \sigma)$.
- Then we can just use $exp(z)$ as samples from the $LogNormal(\mu, \sigma)$, since $\exp(z) = \exp(\log(y)) = y$.

Example 2: Using the log-normal likelihood

Generating simulated data from a lognormal:

```
mu <- 6
sigma <- 0.5
N <- 500000
# Generate N random samples from a log-normal distribution
sl <- rlnorm(N, mu, sigma)
```

Example 2: Using the log-normal likelihood



Example 2: Using the log-normal likelihood

If we assume that response times are log-normally distributed, we'll need to change our likelihood function as follows:

$$t_n \sim LogNormal(\mu, \sigma) \tag{18}$$

- But now the scale of our priors needs to change!
- Uniform priors:

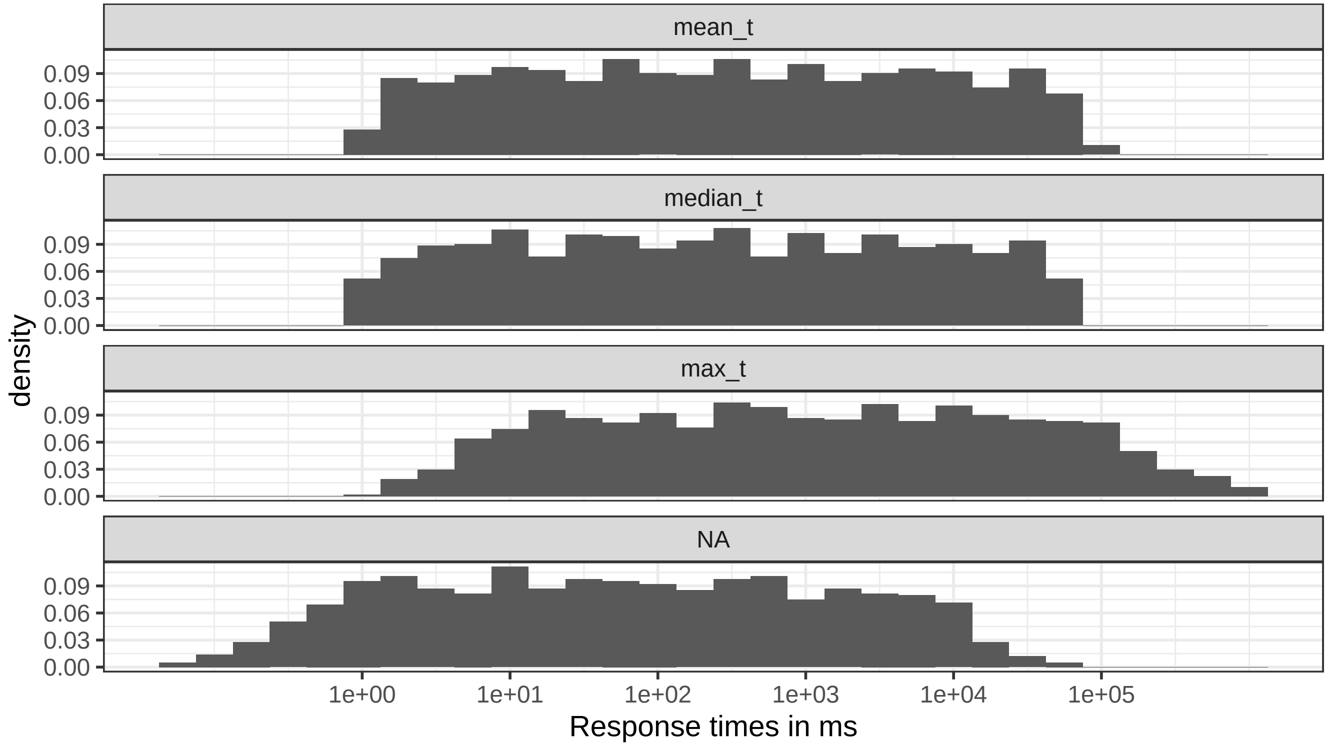
$$\begin{aligned} \mu &\sim Uniform(0, 11) \\ \sigma &\sim Uniform(0, 1) \end{aligned} \tag{19}$$

Example 2: Using the log-normal likelihood

Generate simulated data:

```
N_samples <- 1000
N_obs <- nrow(df_spacebar)
mu_samples <- runif(N_samples, 0, 11)
sigma_samples <- runif(N_samples, 0, 1)
prior_pred_ln <- normal_predictive_distribution(
  mu_samples = mu_samples,
  sigma_samples = sigma_samples,
  N_obs = N_obs
) %>%
  mutate(t_pred = exp(t_pred))
```

Example 2: Using the log-normal likelihood



Example 2: Using the log-normal likelihood

More informative priors:

$$\begin{aligned}\mu &\sim Normal(6, 1.5)\\ \sigma &\sim Normal_{+}(0, 1)\end{aligned} \tag{20}$$

Example 2: Using the log-normal likelihood

Prior predictive distribution:

```
df_spacebar_ref <- df_spacebar %>%
  mutate(t = rep(1, n()))
fit_prior_press_ln <- brm(t ~ 1,
  data = df_spacebar_ref,
  family = lognormal(),
  prior = c(
    prior(normal(6, 1.5), class = Intercept),
    prior(normal(0, 1), class = sigma)
  ),
  sample_prior = "only",
  control = list(adapt_delta = .9)
)

##
```

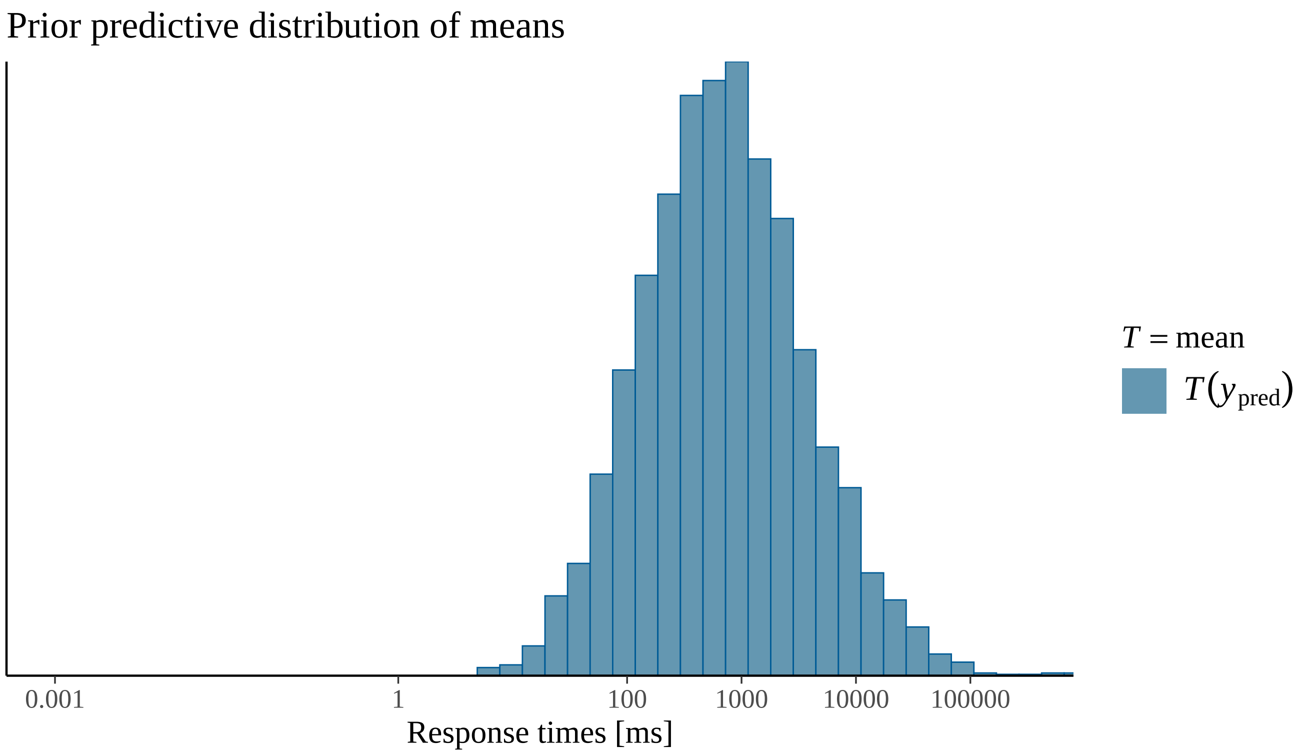
Example 2: Using the log-normal likelihood

New control parameter:

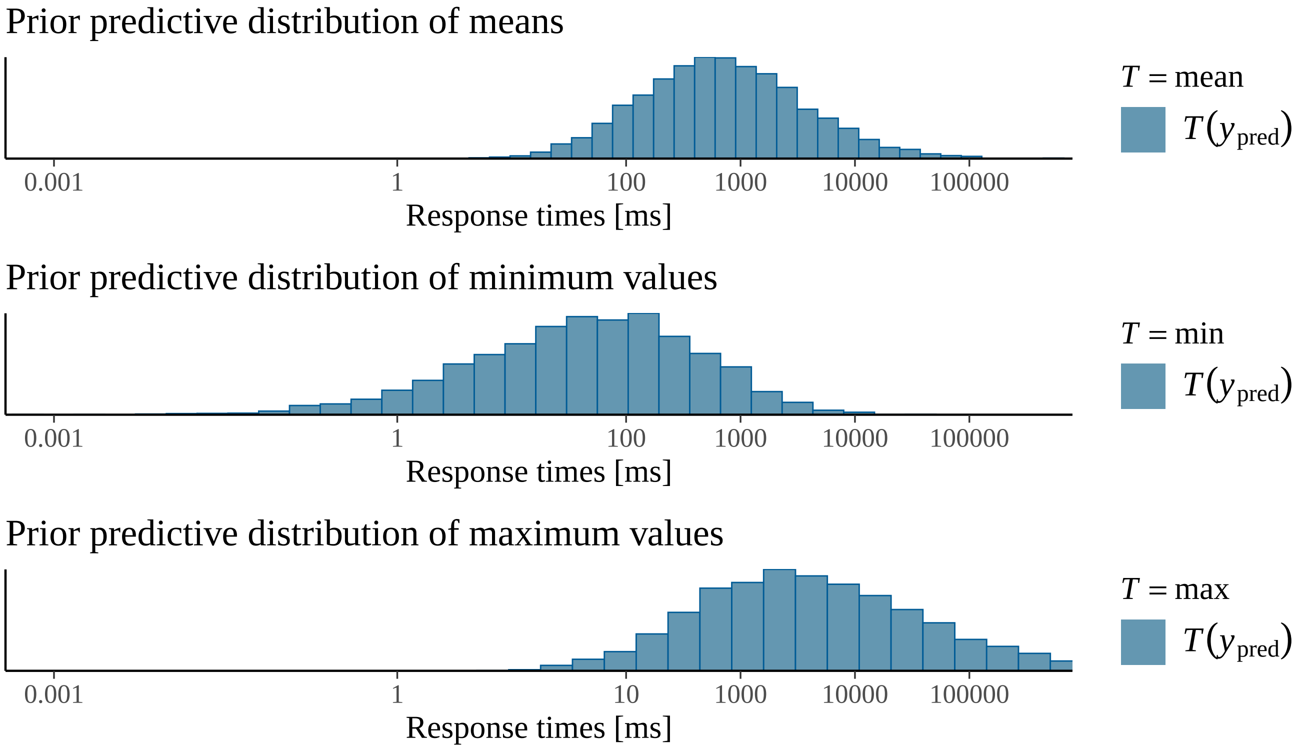
```
control = list(adapt_delta = .9)
```

Example 2: Using the log-normal likelihood

Prior predictive distribution:



Example 2: Using the log-normal likelihood



Example 2: Using the log-normal likelihood

```
fit_press_ln <- brm(t ~ 1,
  data = df_spacebar,
  family = lognormal(),
  prior = c(
    prior(normal(6, 1.5), class = Intercept),
    prior(normal(0, 1), class = sigma)
  )
)
```

Example 2: Using the log-normal likelihood

Verbose summary:

```
fit_press_ln
```

Example 2: Using the log-normal likelihood

Back-transforming to ms:

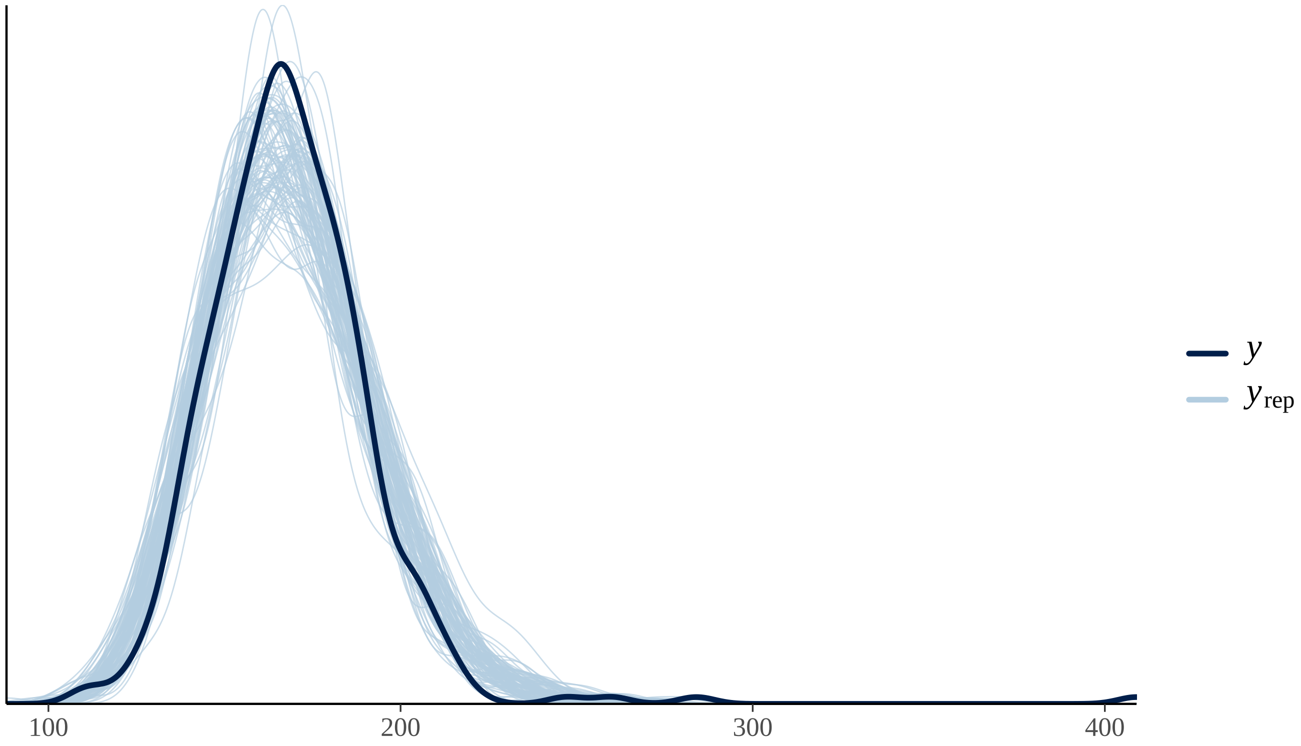
```
estimate_ms <- exp(as_draws_df(fit_press_ln)$b_Intercept)

c(mean = mean(estimate_ms), quantile(estimate_ms, probs = c(.025, .975)))

##     mean     2.5%    97.5%
## 167.0792 164.7716 169.4468
```

Example 2: Using the log-normal likelihood

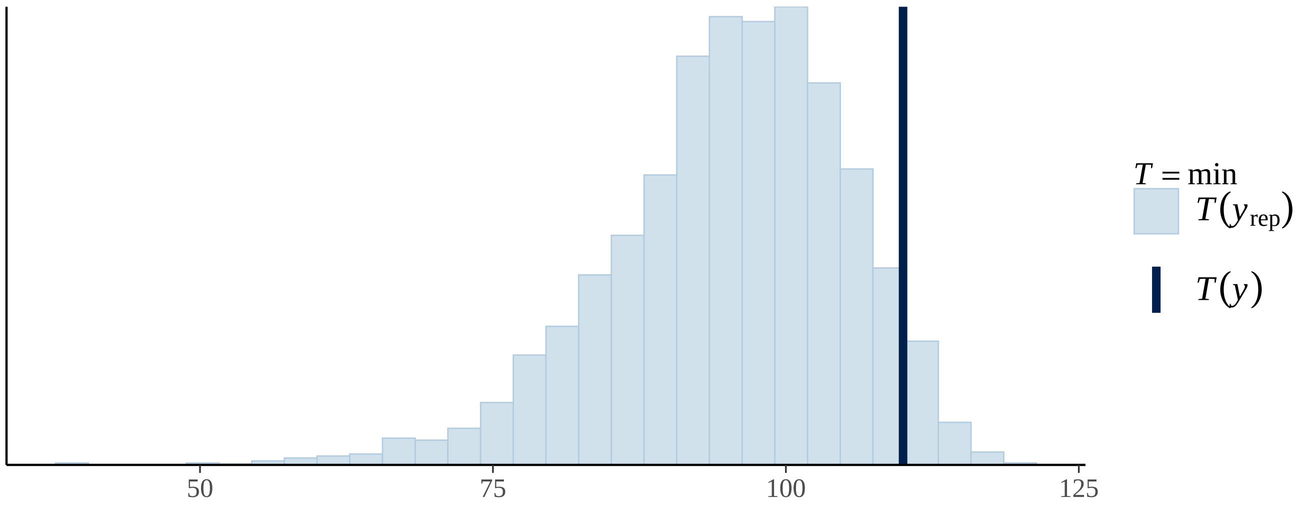
Posterior predictive check:



Example 2: Using the log-normal likelihood

Question: Are the posterior predicted data now more similar to the observed data, compared to the case where we had a Normal likelihood?

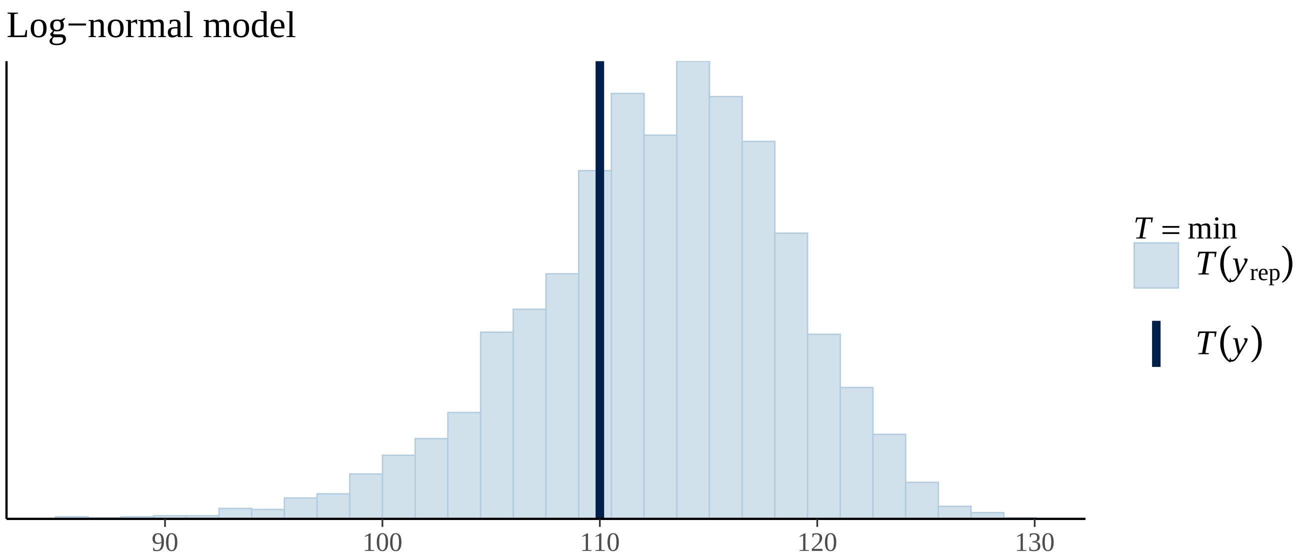
Normal model



Bayesian Data Analysis

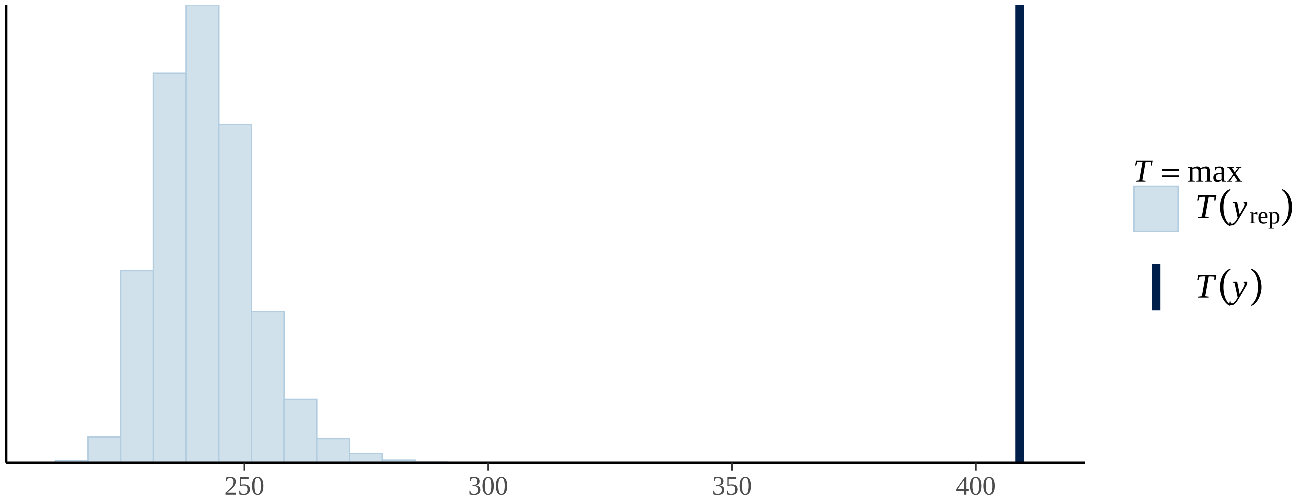
Shravan Vasishth
vasishth.github.io

Example 2: Using the log-normal likelihood

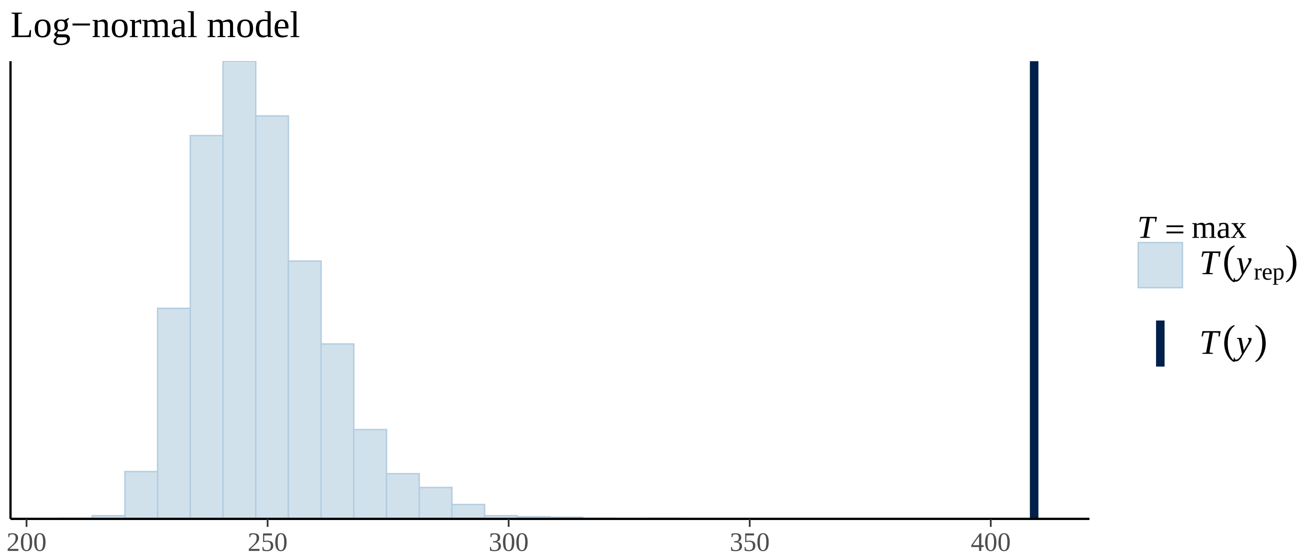


Example 2: Using the log-normal likelihood

Normal model



Example 2: Using the log-normal likelihood



Summary and next steps

- We saw two simple examples of a linear model, with two different likelihoods.
- One key skill we learned was to examine and interpet the prior predictive distribution graphically.
- Another key skill: interpreting the posterior predictive distribution.
- These two distributions tell us how well the model represents the reality, both before and after observing the particular data we have.

Next:

- Adding a predictor: $y = \alpha + \beta \times x + \varepsilon$
- First steps in modeling repeated measures (dependent) data with hierarchical models: $y = \alpha + u_0 + \beta \times x + \varepsilon$