Applied Statistical Modeling (SoSe 2021) – Homework, Week 3

Marc Blauert

2021-04-29

Tasks: McElreath Chapter 3.5, 3H1-3 + Replication of 3H1-3 with brms-package

Load the data for the tasks

Description of the data: The two vectors indicate the gender (male=1, female=0) of officially reported first and second born children in 100 two-child families. So for example, the first family in the data reported a boy (1) and then a girl (0). The second family reported a girl (0) and then a boy (1). The third family reported two girls.

```
data(homeworkch3)
```

```
str(birth1)
```

```
##  num [1:100] 1 0 0 0 1 1 0 1 0 1 ...
```

```
str(birth2)
```

```
##  num [1:100] 0 1 0 1 0 1 1 1 0 0 ...
```

Part 1: Do tasks 3H1-3H3 based on the grid approximation / Rethinking-package

Task 3H1 – Compute the posterior distribution for the probability of a birth being a boy under the assumption of a uniform prior distribution

```
# Count the number of boys and the total number of births
number_boys <- sum(birth1) + sum(birth2); number_boys
```

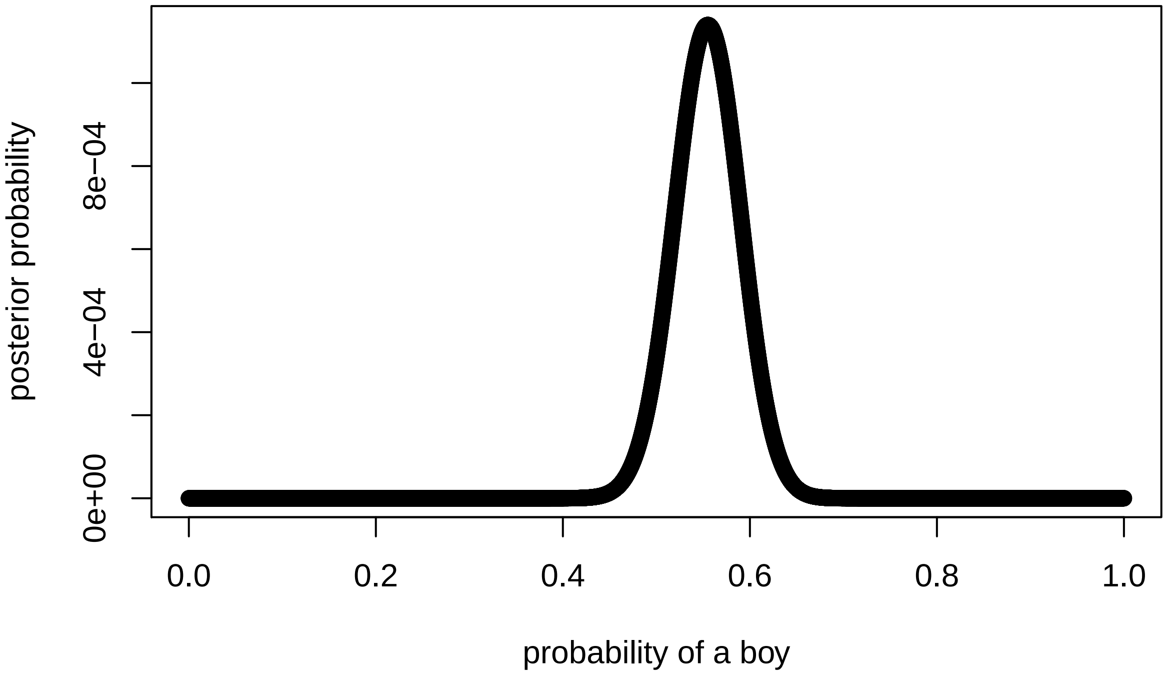
```
## [1] 111
```

```
number_births <- length(birth1) + length(birth2); number_births
```

```
## [1] 200
```

```
# Define grid resolution
res <- 10000

# Step 1: Define grid
p_grid <- seq(from = 0, to = 1, length.out = res)

# Step 2: Define prior (Assumption: Uniform distribution)
prior <- rep(1, res)

# Step 3: Compute likelihood at each value in grid
likelihood <- dbinom(number_boys, size = number_births, prob = p_grid)

# Step 4: Multiply prior with likelihood to get unstandardized posterior
```

```
unstd.posterior <- likelihood * prior

# Step 5: Standardize the posterior
posterior <- unstd.posterior / sum(unstd.posterior)

# Plot posterior distribution
plot(p_grid, posterior, type="b" ,
      xlab="probability of a boy", ylab="posterior probability")
```



```
# Find the maximum
max <- p_grid %>% as.data.frame() %>% filter(posterior == max(posterior)); max
```

```
##           .
## 1 0.5549555
```

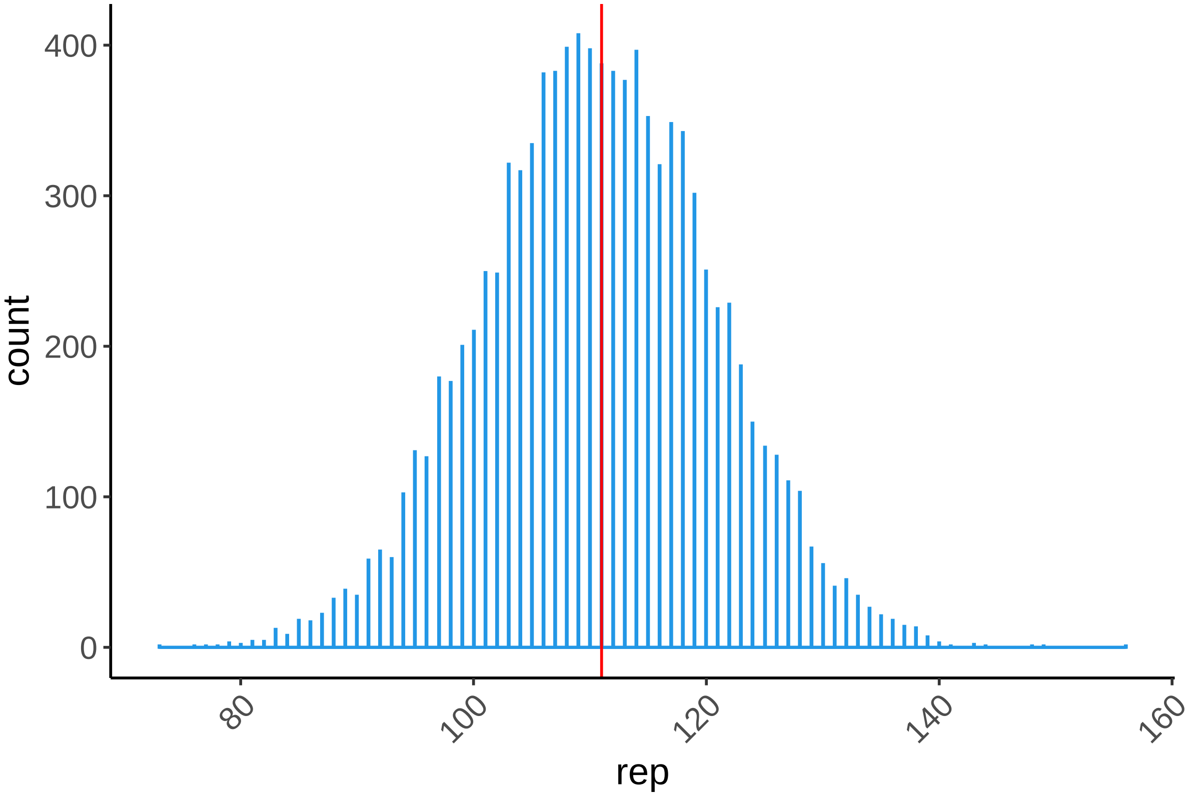
Task 3H2 – Draw 10,000 random parameter values from the posterior distribution in 3H1; What are the 50%, 89%, and 97% highest posterior density intervals?

```
samples <- sample(p_grid, size=res, replace=TRUE, prob=posterior)

HPDI(samples , prob=c(.50, .89, .97))
```

```
##     |0.97     |0.89      |0.5      0.5|     0.89|     0.97|
## 0.4804480 0.5020502 0.5317532 0.5782578 0.6121612 0.6307631
```

Task 3H3 – Simulate 10,000 replicates of 200 births based on the samples in 3H2

```
rep <- rbinom(res, size=number_births, prob=samples)

rep_df <- as.data.frame(rep)

plot <- ggplot(rep_df, aes(x=rep)) +
  geom_histogram(bins = 1000, color = 4) +
  geom_vline(xintercept = 111, color = "red") +
  theme_plots(); plot
```



Comment: The resulting histogram/density plot nicely illustrates the uncertainty inherent to the relatively small sample of 200 births together with the unspecific assumption of the uniform prior. However, from the evidence we have moved from a uniform distribution to the posterior predictive distribution shown above.

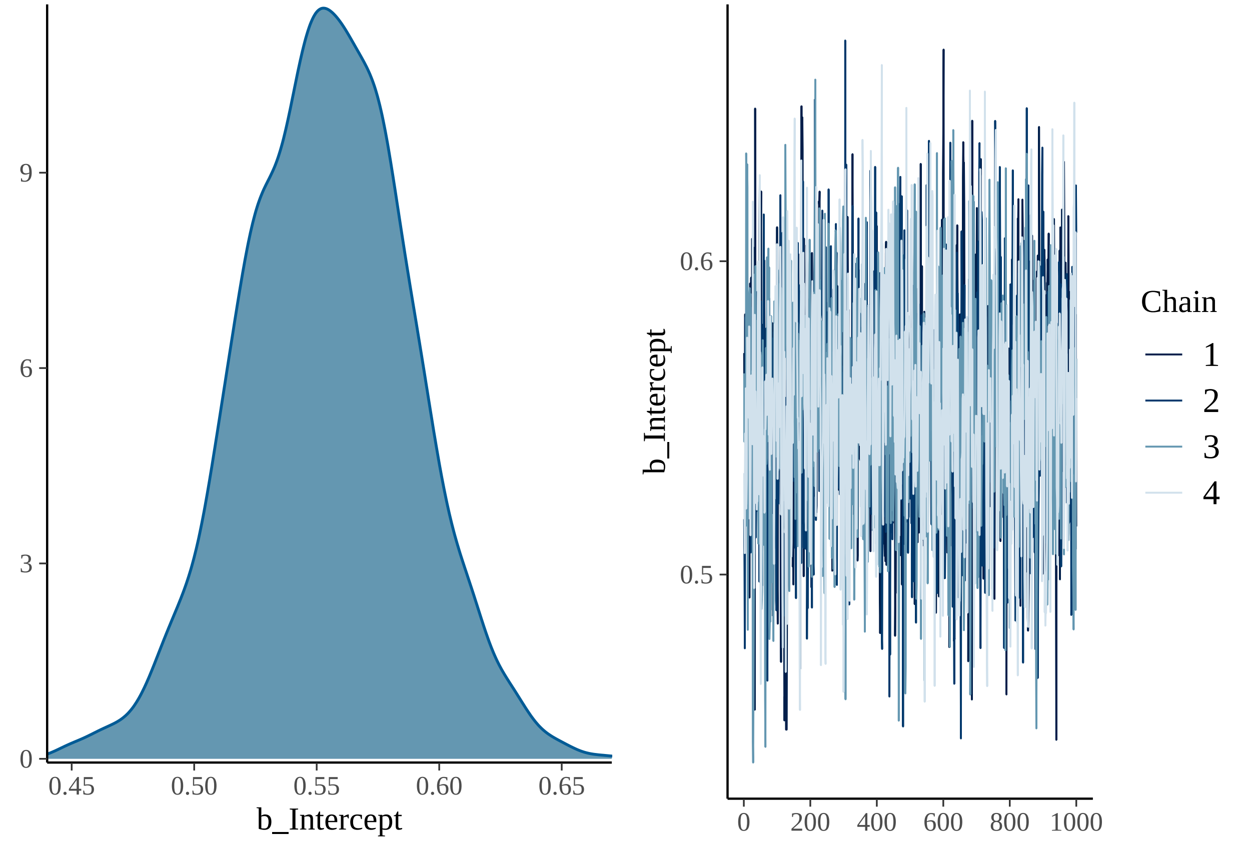
Part 2: Do tasks 3H1-3H3 again but now by using brms

Task 3H1 – Compute the posterior distribution for the probability of a birth being a boy under the assumption of a uniform prior distribution (brms version)

```
summary(model_3H1)
```

```
##  Family: binomial
##   Links: mu = identity
## Formula: boys | trials(births) ~ 0 + Intercept
##    Data: list(boys = number_boys, births = number_births) (Number of observations: 1)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.55      0.03     0.49     0.62 1.00     1502     1925
##
## Samples were drawn using sampling(NUTS). 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).
```

```
plot(model_3H1)
```

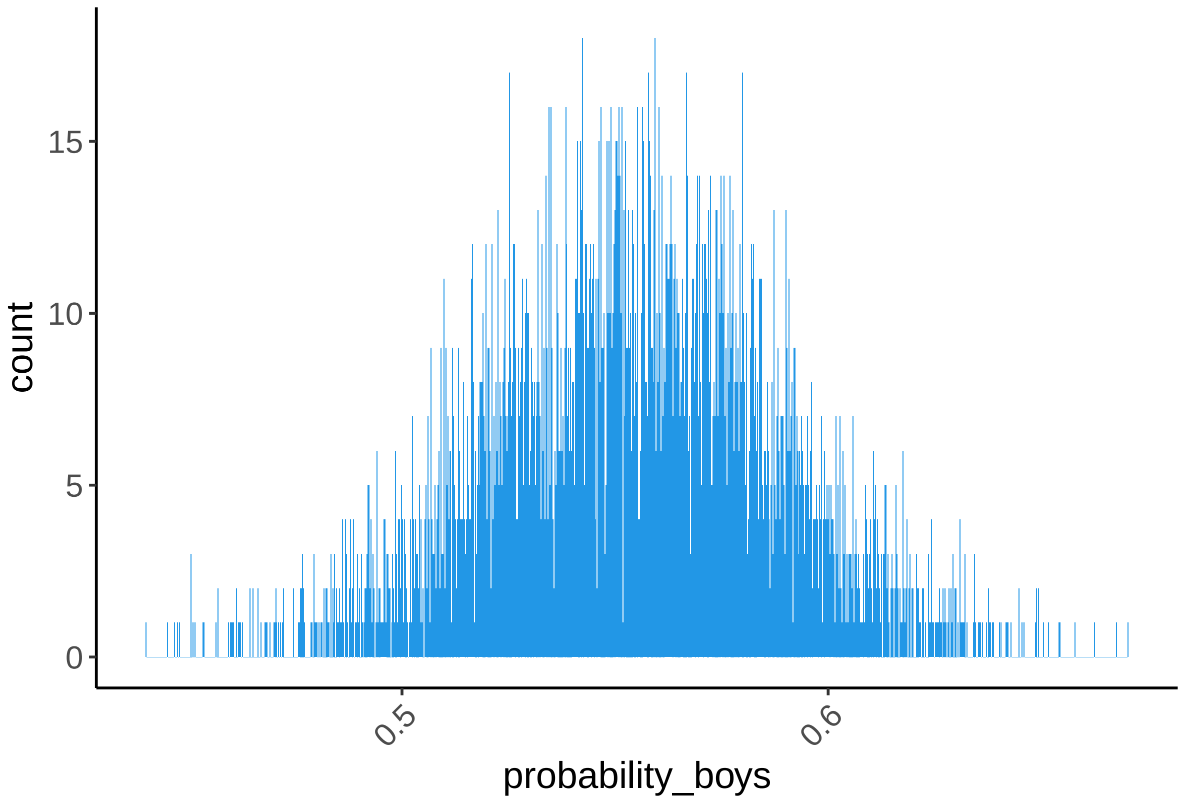


Task 3H2 – Draw 10,000 random parameter values from the posterior distribution in 3H1; What are the 50%, 89%, and 97% highest posterior density intervals? (brms version)

```
samples <- fitted(model_3H1, # fitted returns the line (mean) (=uncertainty from the mean and not from
                  summary = FALSE,
                  scale = "linear") %>%
  as_tibble() %>%
  set_names("probability_boys")
```

```
## Warning: The ‘x‘ argument of ‘as_tibble.matrix()‘ must have unique column names if ‘.name_repair‘ is
## Using compatibility ‘.name_repair‘.
```

```
plot_samples <- ggplot(samples, aes(x=probability_boys)) +
  geom_histogram(bins = 1000, fill = 4, alpha = 1) +
  theme_plots(); plot_samples
```

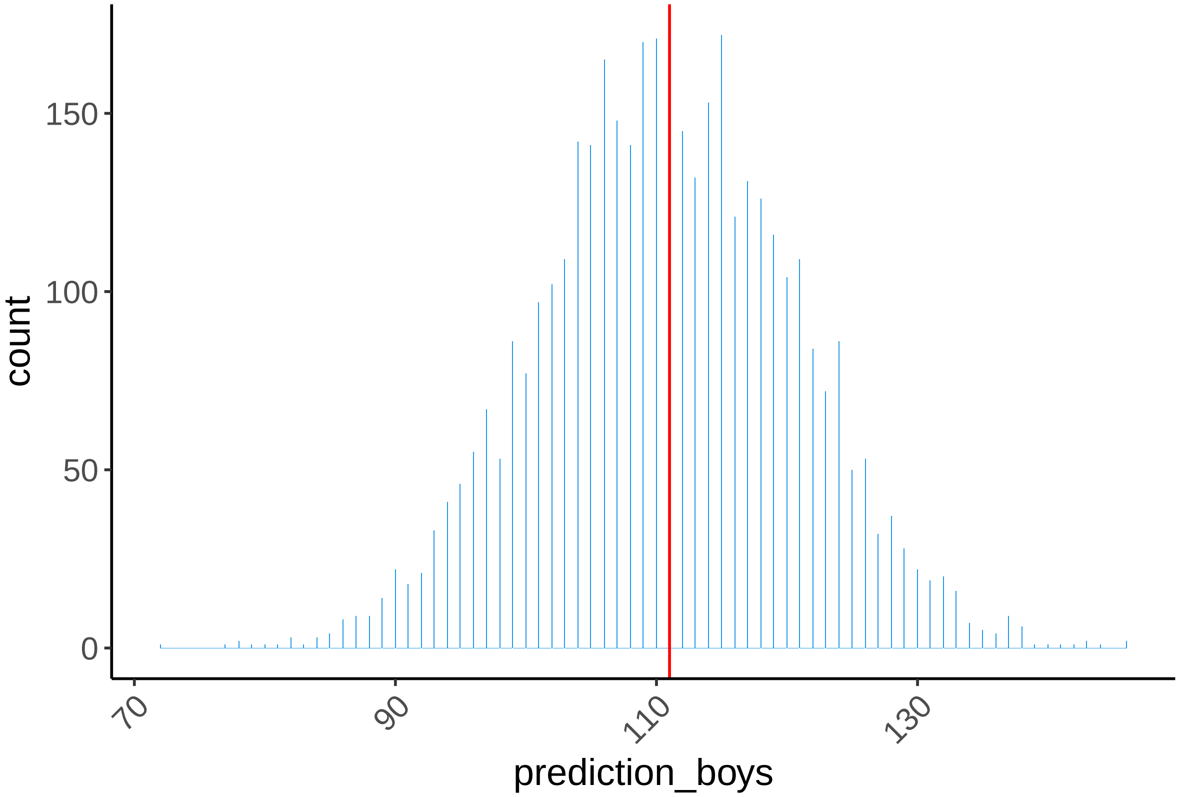


```
HPDI(samples , prob=c(.50, .89, .97))
```

```
##     |0.97     |0.89      |0.5      0.5|     0.89|     0.97|
## 0.4817450 0.5023426 0.5382008 0.5840380 0.6112300 0.6309795
```

Task 3H3 – Simulate 10,000 replicates of 200 births based on the samples in 3H2 (brms version)

```
predict_boys <- predict(model_3H1, summary = FALSE, nsamples = 4000) %>%
  as_tibble() %>%
  set_names("prediction_boys")

plot_prediction <- ggplot(predict_boys, aes(x=prediction_boys)) +
  geom_histogram(bins = 1000, fill = 4, alpha = 1) +
  geom_vline(xintercept = 111, color = "red") +
  theme_plots(); plot_prediction
```



```
max2 <- predict_boys %>% as.data.frame() %>% filter(samples == max(samples)); max2
```

```
##   prediction_boys
## 1             112
```

(Note: The histogram appears a little off but I could not find the reason for it. Maybe because of the relatively low number of only 4,000 samples as compared to 10,000 in the Rethinking-version of the code?)