Week 01 solutions

Caio Geraldes

Exercice 1

Suppose the globe tossing data (Lecture 2, Chapter 2) had turned out to be 4 water and 11 land. Construct the posterior distribution.

Answer

Using Grid Approximation:

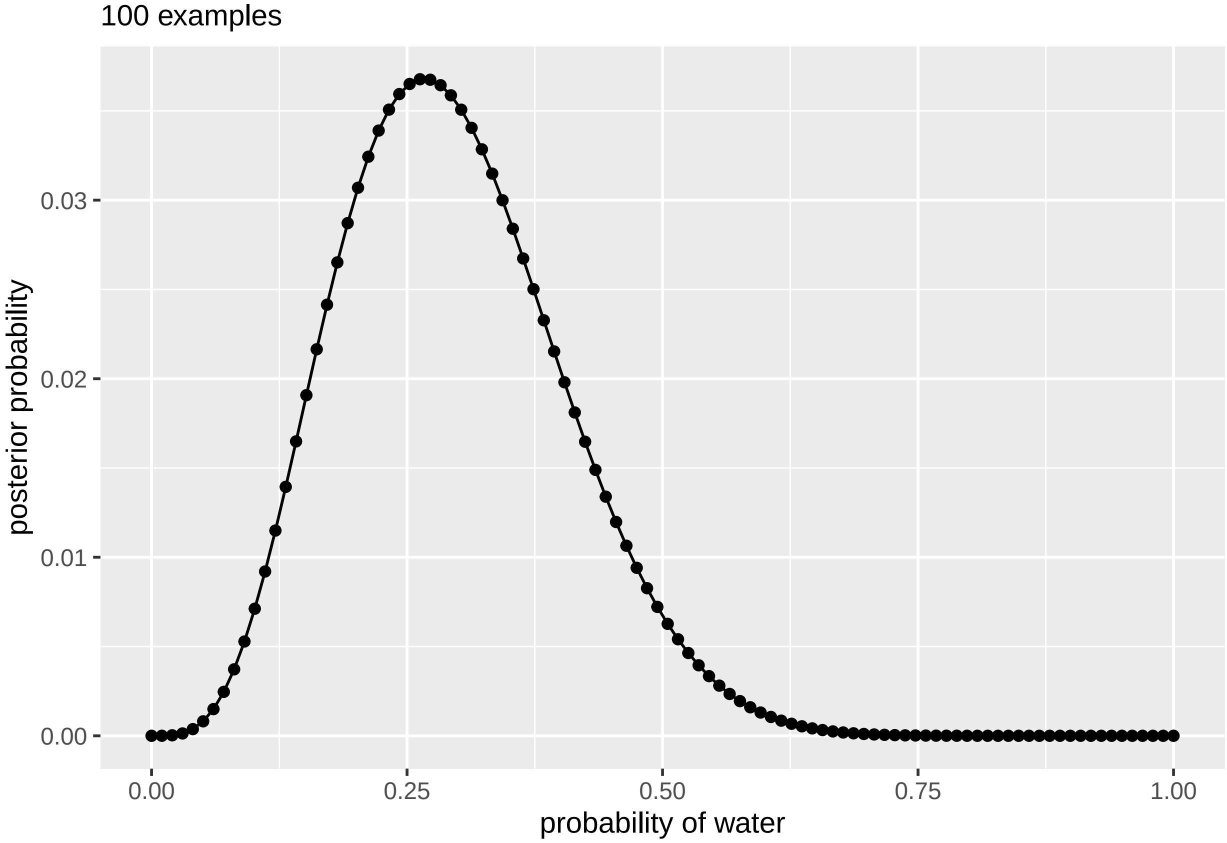
```
grid_approximation_pW <- function(W, L, size) {
    N <- W + L
    p_grid <- seq(from = 0, to = 1, length.out = size)
    prior <- rep(1, size)
    likelihood <- dbinom(W, size = N, prob = p_grid)
    unstd_post <- likelihood * prior
    posterior <- unstd_post / sum(unstd_post)
    tibble(p = p_grid, post = posterior)
}
```

```
size <- 100
W <- 4
L <- 11

model <- grid_approximation_pW(W, L, size)
model
```

```
## # A tibble: 100 x 2
##         p       post
##     <dbl>      <dbl>
##   1 0      0
##   2 0.0101 0.00000205
##   3 0.0202 0.0000294
##   4 0.0303 0.000133
##   5 0.0404 0.000373
##   6 0.0505 0.000812
##   7 0.0606 0.00150
##   8 0.0707 0.00246
##   9 0.0808 0.00372
## 10 0.0909 0.00528
## # ... with 90 more rows
```

```
model %>%
  ggplot(aes(x = p, y = post)) +
  geom_line() + geom_point() +
  xlab("probability of water") +
  ylab("posterior probability") +
```

```
  labs(subtitle = str_c(as.character(size), " examples"))
```

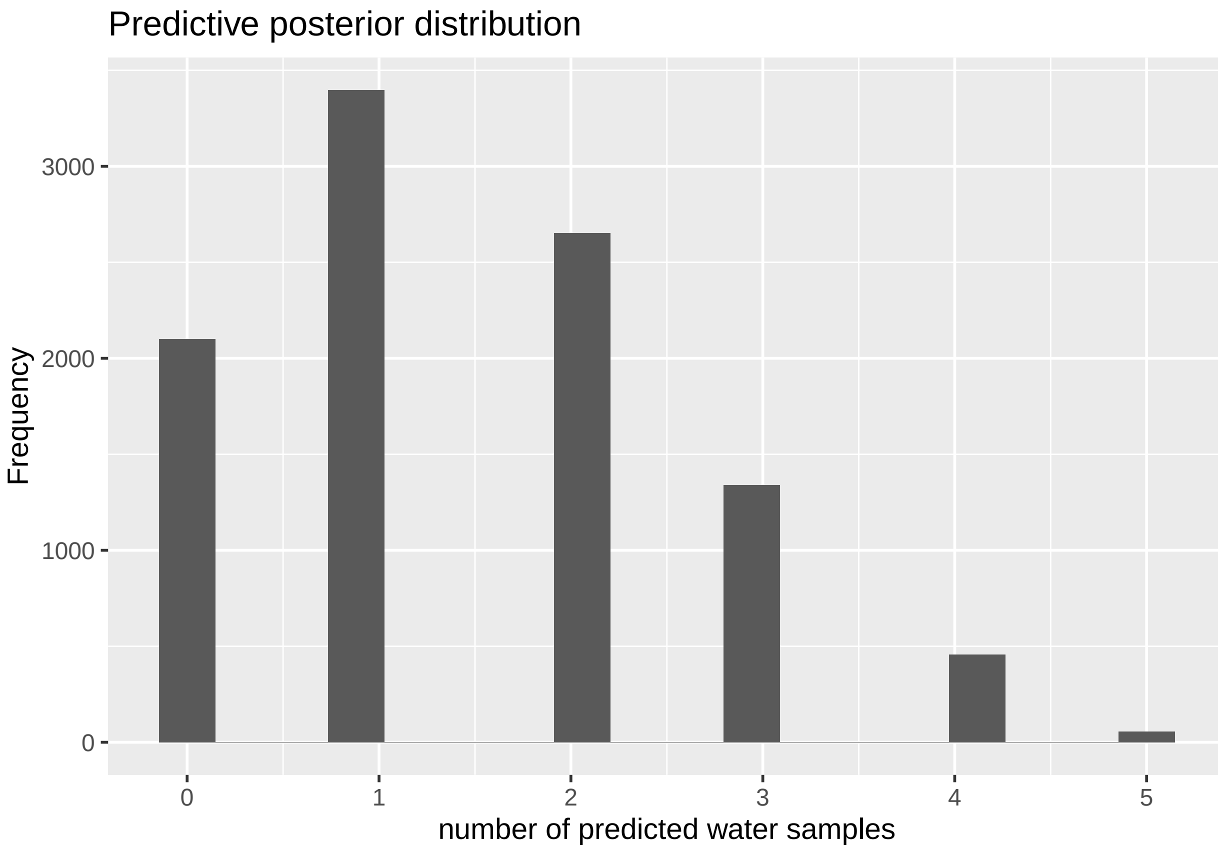


Exercice 2:

Using the posterior distribution from 1, compute the posterior predictive distribution for the next 5 tosses of the same globe. I recommend you use the sampling method.

```
n_simulations <- 1e4
sample_size <- 5
p_grid <- seq(from = 0, to = 1, length.out = size)

post_samples <- sample(
   p_grid,
   size = n_simulations,
   replace = TRUE,
   prob = model$post
)

w_pred <- rbinom(
  n_simulations,
  size = sample_size,
  prob = post_samples
)

tibble(w_pred) %>%
  ggplot(aes(x = w_pred)) +
  geom_histogram(bins = 18) +
```

```
  xlab("number of predicted water samples") +
  ylab("Frequency") +
  labs(title = "Predictive posterior distribution")
```



Exercice 3

Use the posterior predictive distribution from 2 to calculate the probability of 3 or more water samples in the next 5 tosses.

```
sum(w_pred >= 3) / n_simulations
```

```
## [1] 0.1851
```

Exercice 4

This problem is an optional challenge for people who are taking the course for a second or third time. Suppose you observe $W = 5$ water points, but you forgot to write down how many times the globe was tossed, so you don't know the number of land points L. Assume that $p = 0.7$ and compute the posterior distribution of the number of tosses N. Hint: Use the binomial distribution.

Answer

The probability of round of tosses having N tosses can be approximated with a Grid Approximation, just like the probability of water p_W was in the last exercises. Since the limit of N is $1 \ldots \infty$ and N must be an integer, in contrast to p_W's $0 \ldots 1$, the grid can not be expressed by non-integer values and must have an upper bound N_{max}.

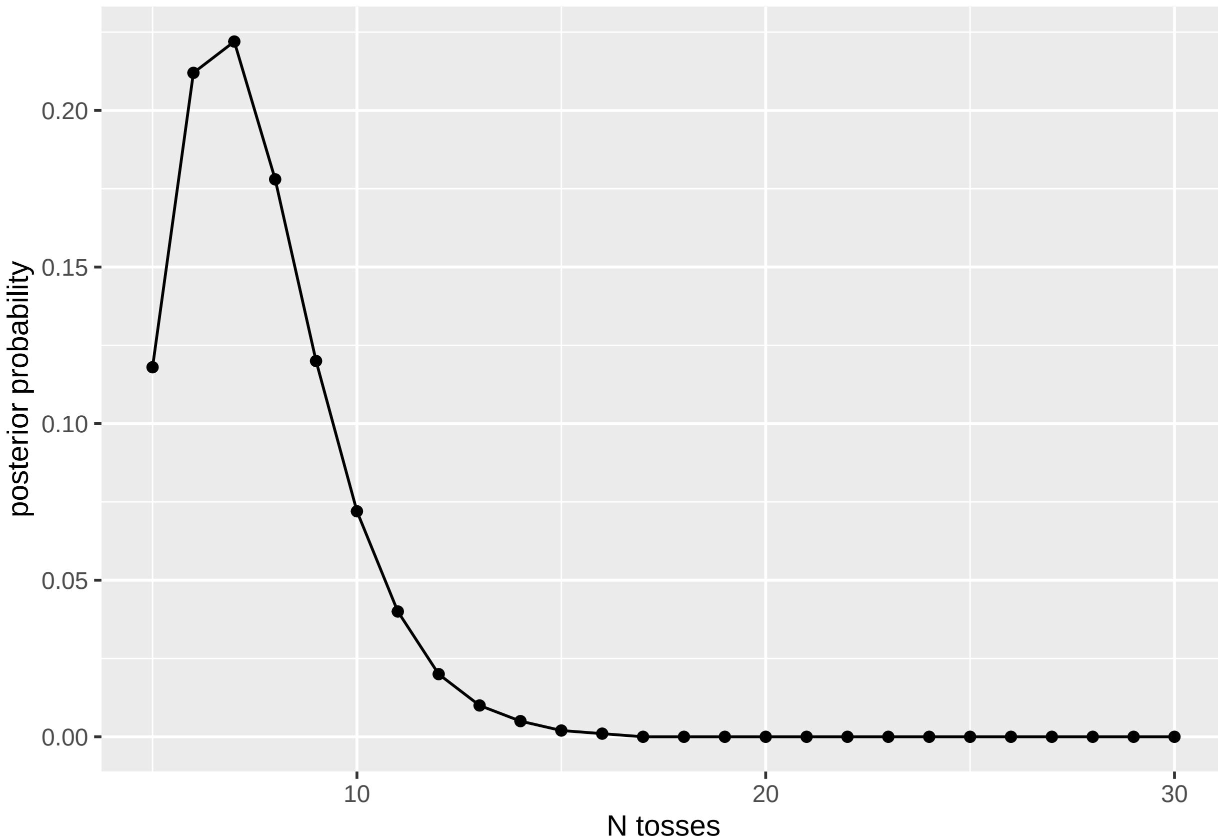
The Grid Approximation for a range of tosses is produced with help of the functions `make_n_grid` and `grid_approximation_tosses`:

```
make_n_grid <- function(W, N_max) {
  seq(from = W, to = N_max, length.out = (N_max - W + 1))
}

grid_approximation_tosses <- function(W, p_W, N_max) {
  n_grid <- make_n_grid(W, N_max)
  likelihood_w_given_n <- dbinom(W, n_grid, p_W)
  unstd_post <- likelihood_w_given_n
  posterior <- unstd_post / sum(unstd_post)
  tibble(n = n_grid, post = round(posterior,3))
}
```

In this answer we use $N_{max} = 30$, as by 20 tosses, the probability of N with $p_W = 0,7$ and $W = 5$ is presumably very low.

```
N_max <- 30
W <- 5
p_W <- 0.7

model <- grid_approximation_tosses(W, p_W, N_max)
model
```

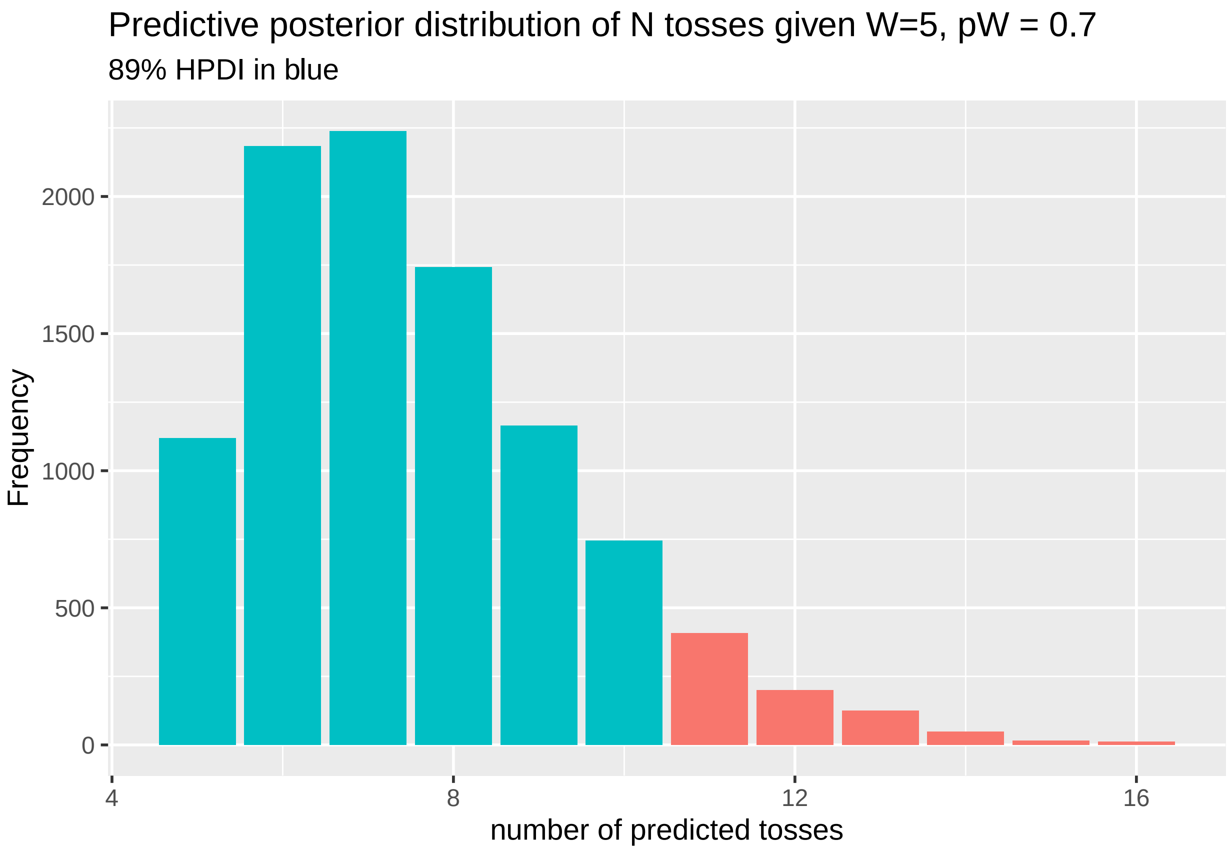
```
## # A tibble: 26 x 2
##        n  post
##    <dbl> <dbl>
##  1     5 0.118
##  2     6 0.212
##  3     7 0.222
##  4     8 0.178
##  5     9 0.12
##  6    10 0.072
##  7    11 0.04
##  8    12 0.02
##  9    13 0.01
## 10    14 0.005
## # ... with 16 more rows
```



Drawing 10.000 samples from the posterior distribution:

```
n_simulations <- 1e4
n_grid <- make_n_grid(W, N_max)
post_samples <- sample(
   n_grid,
   size = n_simulations,
   replace = TRUE,
    prob = model$post
)
hpdi <- HPDI(post_samples)
hpdi
```

```
## |0.89 0.89|
##     5    10
```

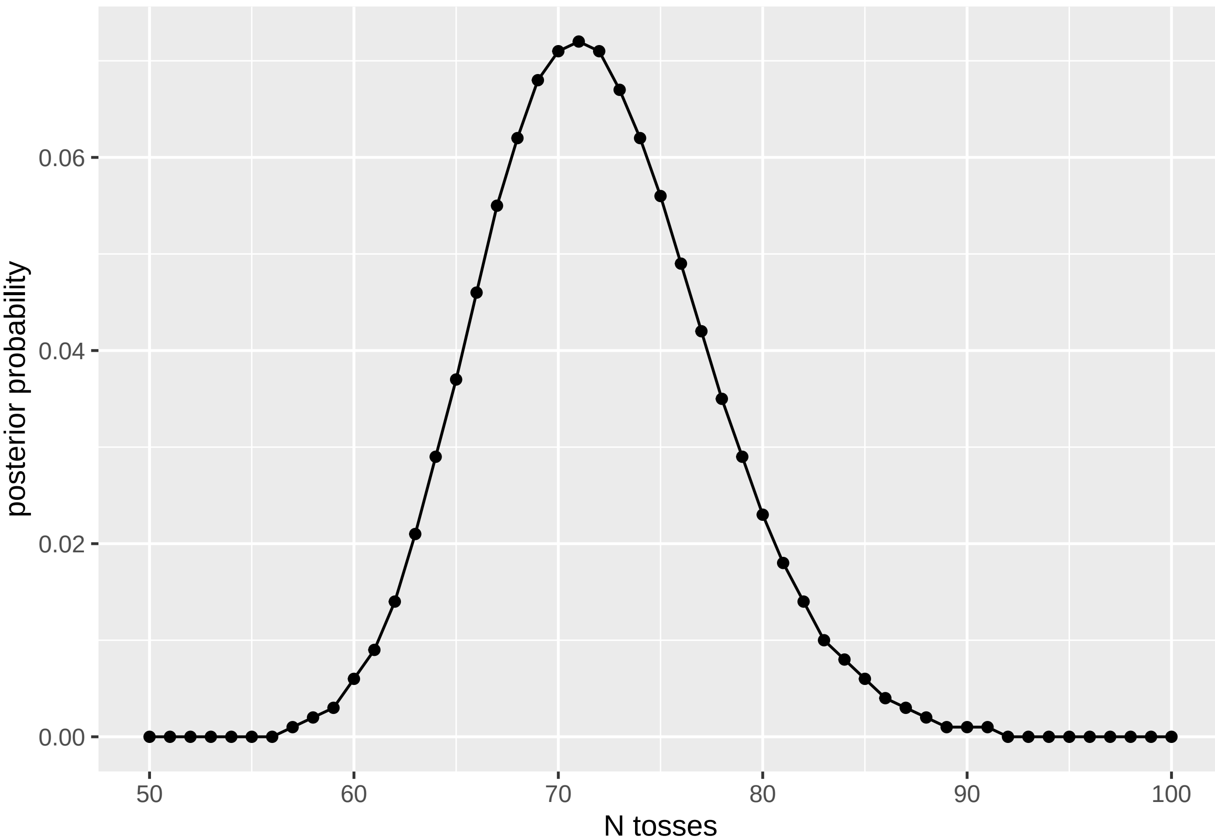


Testing

To test the model, we will assume a longer tossing run, with $W = 50$ and a $N_{max} = 100$:

```
N_max <- 100
W <- 50
p_W <- 0.7

model <- grid_approximation_tosses(W, p_W, N_max)
model
```

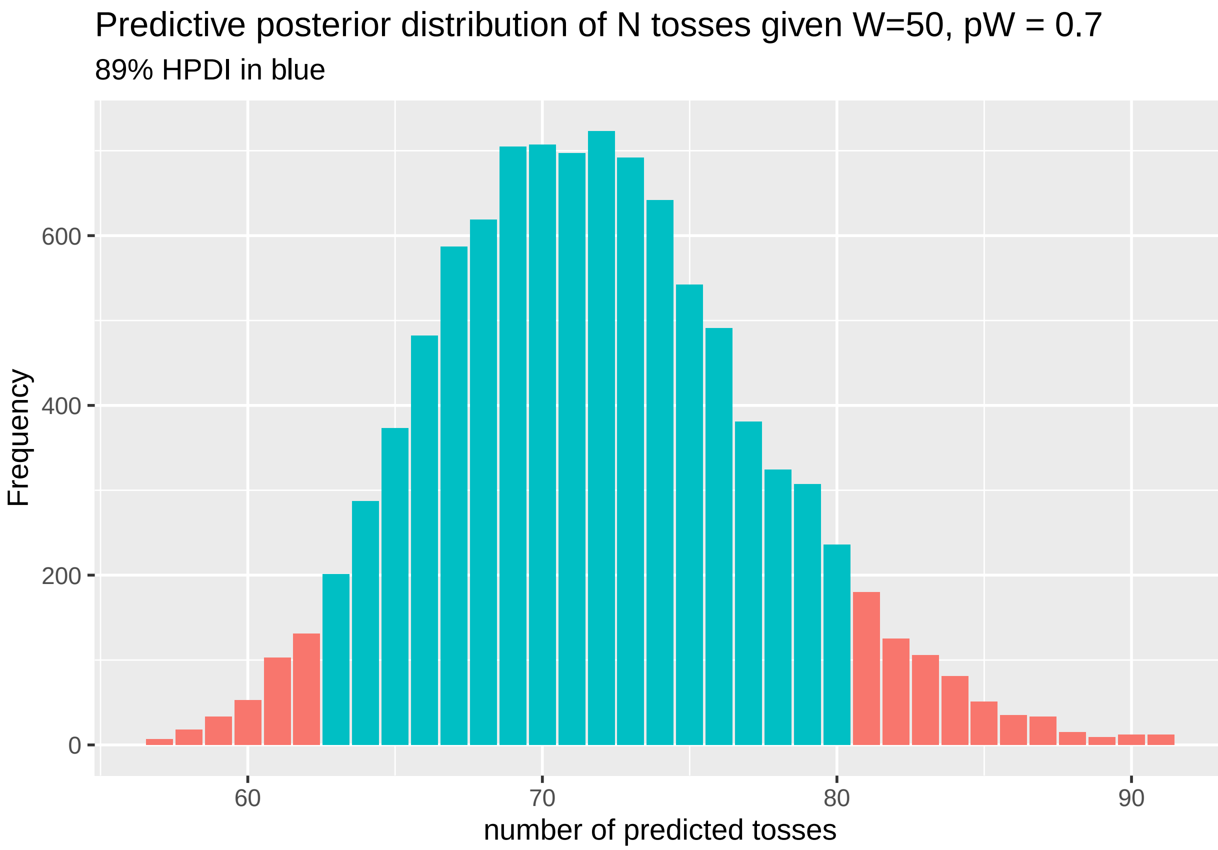
```
## # A tibble: 51 x 2
##         n  post
##     <dbl> <dbl>
##  1     50 0
##  2     51 0
##  3     52 0
##  4     53 0
##  5     54 0
##  6     55 0
##  7     56 0
##  8     57 0.001
##  9     58 0.002
## 10     59 0.003
## # ... with 41 more rows
```



Drawing 10.000 samples from the posterior distribution:

```r
n_simulations <- 1e4
n_grid <- make_n_grid(W, N_max)
post_samples <- sample(
   n_grid,
   size = n_simulations,
   replace = TRUE,
    prob = model$post
)

hpdi <- HPDI(post_samples)
hpdi
```

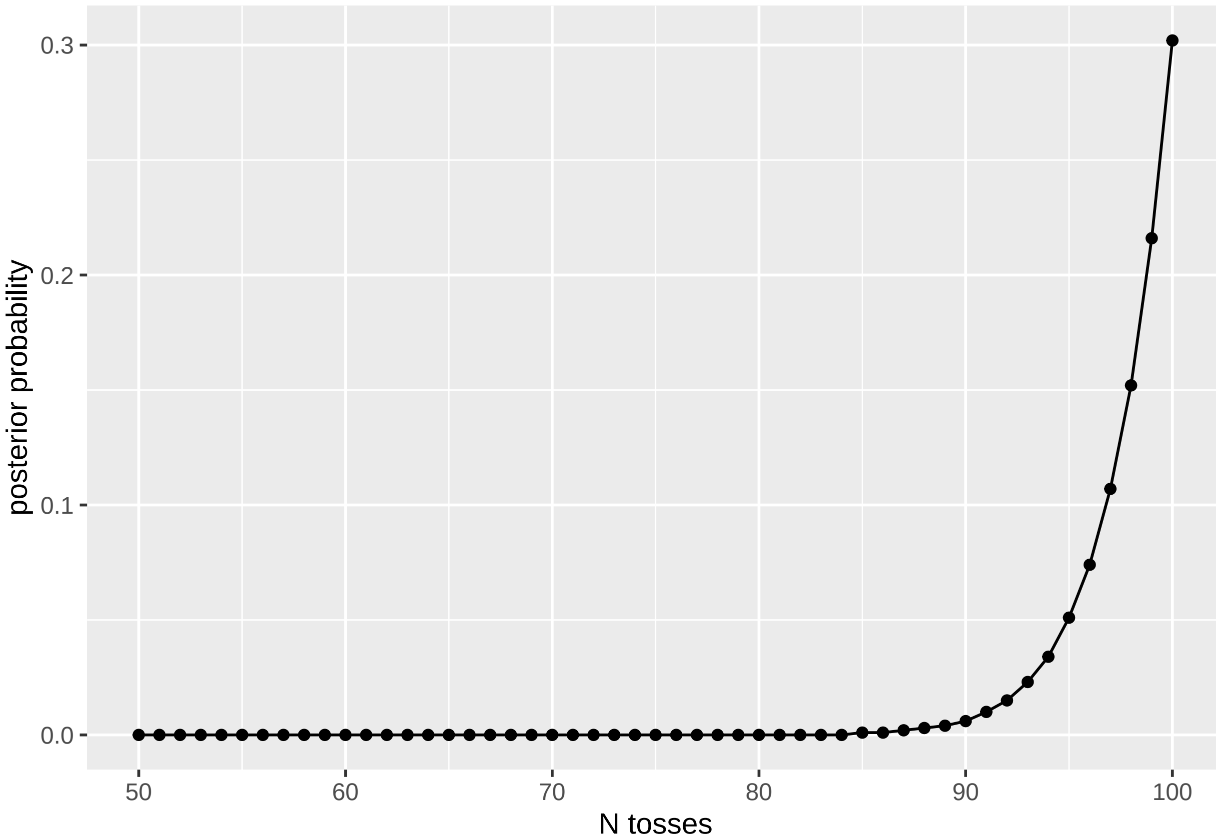
```
## |0.89 0.89|
##    63    80
```



By changing p_W for a value lower than 0.5, say $p_W = 0.3$, the likelihood of longer tossing runs increases:

```
N_max <- 100
W <- 50
p_W <- 0.3

model <- grid_approximation_tosses(W, p_W, N_max)
model
```

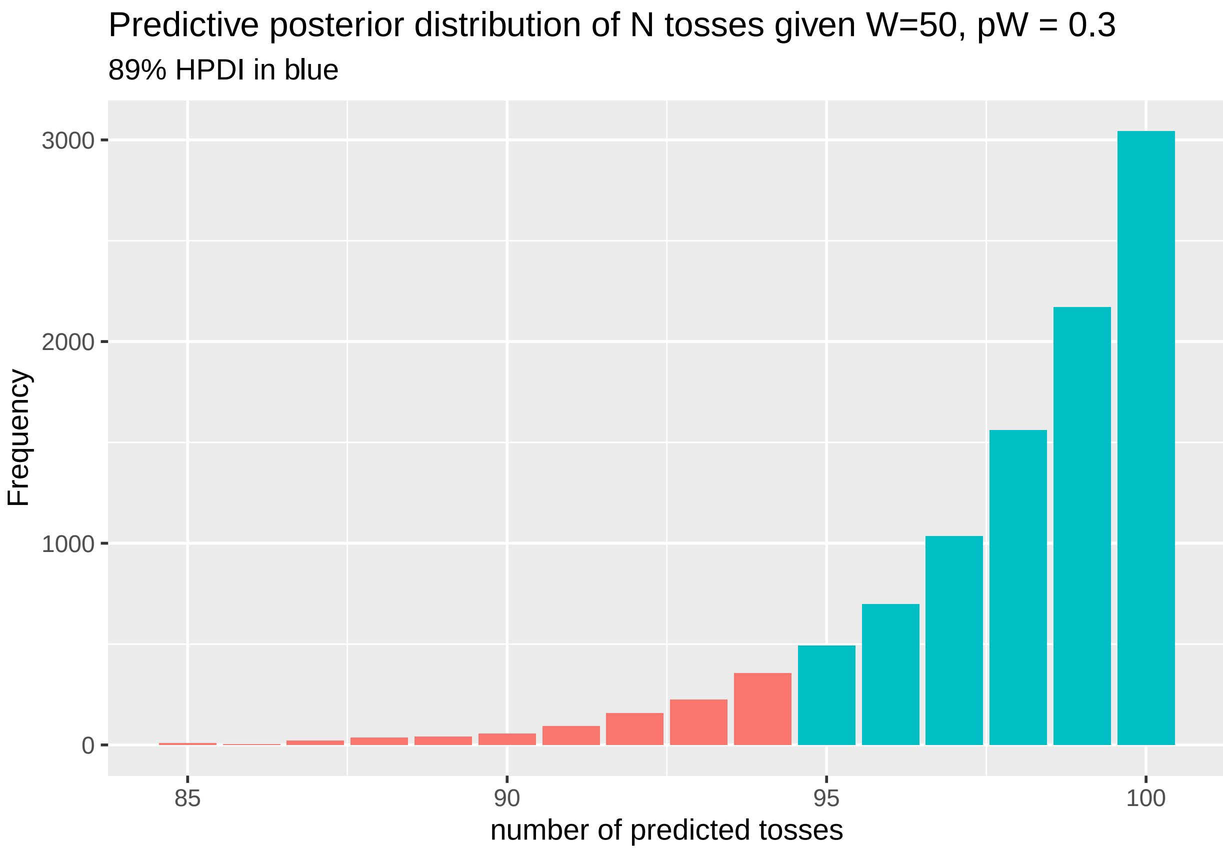
```
## # A tibble: 51 x 2
##         n  post
##     <dbl> <dbl>
##  1    50     0
##  2    51     0
##  3    52     0
##  4    53     0
##  5    54     0
##  6    55     0
##  7    56     0
##  8    57     0
##  9    58     0
## 10    59     0
## # ... with 41 more rows
```



Drawing 10.000 samples from the posterior distribution:

```
n_simulations <- 1e4
n_grid <- make_n_grid(W, N_max)
post_samples <- sample(
   n_grid,
   size = n_simulations,
   replace = TRUE,
    prob = model$post
)
hpdi <- HPDI(post_samples)
hpdi
```

```
## |0.89 0.89|
##    95   100
```


Predictive posterior distribution of N tosses given W=50, pW = 0.3
89% HPDI in blue
Frequency
3000
2000
1000
0
85
90
95
100
number of predicted tosses