

Beta–Binomial: Elicitation, Posterior and Credible Intervals

Proportion of Ph.D. students who sleep at least 6 hours per day

Hedibert Freitas Lopes

Insper

Disclaimer. These slides were prepared with AI assistance (Claude Opus 5.5, Anthropic).
All results were independently verified by the author through hand computation.

The exercise

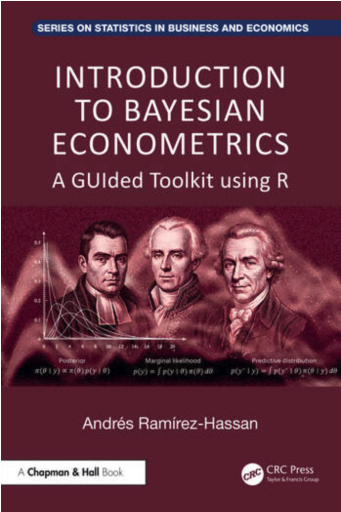
15. **Ph.D. students sleeping hours** [6, Chap. 2]

We are interested in learning about the proportion of Ph.D. students who sleep at least 6 hours per day. We have a sample of 52 students, where 15 report sleeping at least 6 hours, and the remaining 37 report not sleeping at least 6 hours. The prior distribution is a Beta distribution, with hyperparameters calibrated so that the prior probabilities of the proportion of students who sleep less than 6 hours being less than 0.4 and 0.75 are 0.6 and 0.95, respectively. Estimate the 95% posterior credible interval for the proportion of Ph.D. students who sleep at least 6 hours per day. Then, assume there is a group of experts whose beliefs about the proportion of Ph.D. students sleeping at least 6 hours are represented in the following table:

Use [Table 3.1](#) as prior information, and find the posterior distribution of the proportion of students who sleep at least 6 hours.

TABLE 3.1
Probability distribution: Ph.D students who sleep at least 6 hours per day. ↗

h	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55
$P(p = h)$	0.05	0.07	0.10	0.12	0.15	0.17	0.15	0.11	0.06	0.01	0.01



The exercise

- θ : proportion of Ph.D. students who sleep at least 6 hours per day.
- Data: $n = 52$ Bernoulli trials, $y = 15$ successes, so $\hat{\theta} = 15/52 \approx 0.288$.
- Likelihood:

$$p(y \mid \theta) \propto \theta^{15}(1 - \theta)^{37}.$$

- Two ways of encoding prior information:
 1. a **Beta** prior elicited from two quantiles;
 2. a **discrete** prior given by a table of expert beliefs.

Question 1: eliciting a Beta prior from two quantiles

Q. Find a and b for a $\text{Beta}(a, b)$ prior for θ when all we know is

$$P(\theta < 0.40) = 0.60 \quad \text{and} \quad P(\theta < 0.75) = 0.95.$$

Question 1: eliciting a Beta prior from two quantiles

Q. Find a and b for a $\text{Beta}(a, b)$ prior for θ when all we know is

$$P(\theta < 0.40) = 0.60 \quad \text{and} \quad P(\theta < 0.75) = 0.95.$$

A. Solve the two nonlinear equations

$$F_{a,b}(0.40) = 0.60, \quad F_{a,b}(0.75) = 0.95,$$

where $F_{a,b}$ is the Beta cdf (solve in $(\log a, \log b)$ to keep $a, b > 0$):

$a \approx 1.445, \quad b \approx 2.568$
--

Implied prior: mean 0.360, median 0.335, mode 0.221, sd 0.214.

Question 1: R code

```
f <- function(p) {  
  a <- exp(p[1]); b <- exp(p[2])  
  c(pbeta(0.40, a, b) - 0.60,  
    pbeta(0.75, a, b) - 0.95)  
}  
ab <- exp(nleqslv::nleqslv(c(0, 0.5), f)$x)  
ab                                     # 1.445 2.568  
pbeta(c(0.40, 0.75), ab[1], ab[2])   # 0.60 0.95
```

Question 2: posterior under the Beta prior

Q. With $n = 52$ trials and $y = 15$ successes, what is the posterior?

Question 2: posterior under the Beta prior

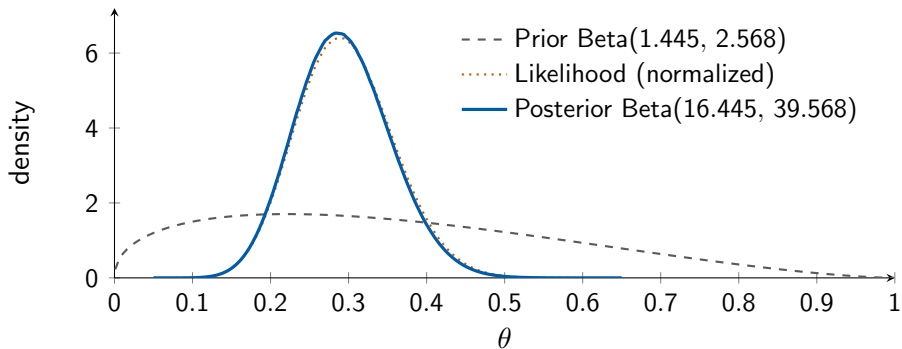
Q. With $n = 52$ trials and $y = 15$ successes, what is the posterior?

A. Conjugacy:

$$\theta \mid y \sim \text{Beta}(a+y, b+n-y) = \text{Beta}(1.445+15, 2.568+37) = \boxed{\text{Beta}(16.445, 39.568)}$$

	Mean	Median	Mode	SD
Prior	0.360	0.335	0.221	0.214
Posterior	0.294	0.291	0.286	0.060

Prior, likelihood and posterior



Question 3: 95% credible intervals

Q. Find the 95% credible interval for θ under the posterior $\text{Beta}(16.445, 39.568)$.
What about the HPD interval?

Question 3: 95% credible intervals

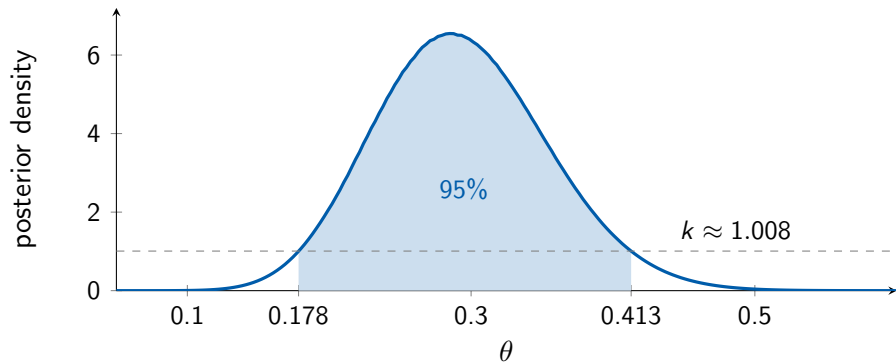
Q. Find the 95% credible interval for θ under the posterior $\text{Beta}(16.445, 39.568)$. What about the HPD interval?

A.

Interval	Lower	Upper	Length
Equal-tailed (2.5%, 97.5%)	0.1830	0.4182	0.2352
HPD	0.1784	0.4128	0.2344

- The posterior is right-skewed, so the HPD shifts left and is slightly shorter.
- HPD check: the density is equal at both endpoints ($p(0.1784 | y) \approx p(0.4128 | y) \approx 1.008$).

HPD interval: the horizontal-line picture



Questions 2–3: R code

```
a1 <- ab[1] + 15; b1 <- ab[2] + 37      # 16.445, 39.568

# Equal-tailed 95% interval
qbeta(c(0.025, 0.975), a1, b1)        # 0.1830 0.4182

# HPD: shortest interval with 95% probability
len <- function(p) qbeta(p + 0.95, a1, b1) - qbeta(p, a1, b1)
p    <- optimize(len, c(0, 0.05), tol = 1e-10)$minimum
qbeta(c(p, p + 0.95), a1, b1)        # 0.1784 0.4128

# or: HDInterval::hdi(qbeta, 0.95, shape1 = a1, shape2 = b1)
```

Question 4: a discrete prior from expert opinion

A group of experts express beliefs about θ (Table 3.1):

h	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55
$P(\theta = h)$	0.05	0.07	0.10	0.12	0.15	0.17	0.15	0.11	0.06	0.01	0.01

Q. With the same data ($n = 52$, $y = 15$), find the posterior of θ .

Question 4: a discrete prior from expert opinion

A group of experts express beliefs about θ (Table 3.1):

h	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55
$P(\theta = h)$	0.05	0.07	0.10	0.12	0.15	0.17	0.15	0.11	0.06	0.01	0.01

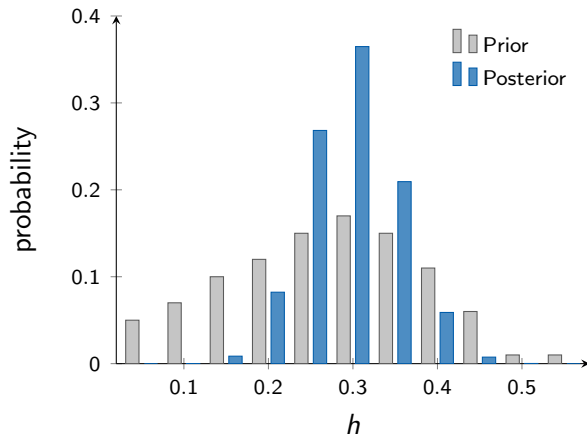
Q. With the same data ($n = 52$, $y = 15$), find the posterior of θ .

A. Bayes' theorem on the grid (binomial coefficient cancels):

$$P(\theta = h_i | y) = \frac{P(\theta = h_i) h_i^{15} (1 - h_i)^{37}}{\sum_{j=1}^{11} P(\theta = h_j) h_j^{15} (1 - h_j)^{37}}.$$

Question 4: discrete posterior

h	Prior	Posterior
0.05	0.05	0.0000
0.10	0.07	0.0001
0.15	0.10	0.0086
0.20	0.12	0.0822
0.25	0.15	0.2683
0.30	0.17	0.3648
0.35	0.15	0.2094
0.40	0.11	0.0589
0.45	0.06	0.0075
0.50	0.01	0.0002
0.55	0.01	0.0000



Question 4: posterior summaries

- Posterior mean: $E(\theta \mid y) = 0.2946$.
- Posterior sd: 0.0543.
- Posterior mode: $\theta = 0.30$ (probability 0.365).
- 95% credible set (highest posterior probability):

$$\{0.20, 0.25, 0.30, 0.35, 0.40\}, \quad P = 0.984.$$

Dropping 0.40 leaves only $0.925 < 0.95$. With a discrete grid we cannot hit 95% exactly.

Question 4: R code

```
h      <- seq(0.05, 0.55, by = 0.05)
prior  <- c(.05,.07,.10,.12,.15,.17,.15,.11,.06,.01,.01)

post   <- prior * dbinom(15, 52, h)
post   <- post / sum(post)
round(post, 4)

m <- sum(h * post)           # 0.2946
sqrt(sum(h^2 * post) - m^2) # 0.0543

o <- order(post, decreasing = TRUE) # 95% HPD set
h[o][1:which(cumsum(post[o]) >= 0.95)[1]]
```

Comparison of the two analyses

	Beta prior	Discrete prior
Prior mean	0.360	0.271
Posterior mean	0.294	0.295
Posterior sd	0.060	0.054
95% region	HPD (0.178, 0.413)	$\{0.20, \dots, 0.40\}$

- Two quite different priors lead to nearly the same posterior mean: $n = 52$ observations dominate the prior.
- The discrete prior restricts θ to a grid; the posterior can only put mass on those 11 values.