Research 1-1: Highest Posterior Density Interval for a Binomial Parameter

Nahyun Lee

10/31/2024

1. Introduction

This research aims to calculate and visualize the Highest Posterior Density (HPD) credible interval for a binomial success probability parameter, denoted as p. We will use a Bayesian framework, which allows us to update our beliefs about a parameter after observing new data.

The core of this analysis is the Beta-Binomial conjugate model. Our initial belief about the parameter p (the prior) is represented by a Beta distribution. After conducting a binomial experiment and observing a certain number of successes and failures, we update our belief to form a new distribution for p, known as the posterior distribution, which is also a Beta distribution.

For this specific analysis, we will: * Start with a `Beta(1,1)` prior, which is equivalent to a Uniform(0,1) distribution. This represents a state of no initial preference for any particular value of p. * Use data from an experiment with 1 success and 9 failures. * Calculate the 95% HPD credible interval for p based on the resulting posterior distribution.

2. Methodology

HPD Interval Calculation

The HPD credible interval is a special type of Bayesian credible interval. For a given probability (e.g., 95%), it is the **shortest possible interval** that contains the true parameter value with that probability. A key property is that the probability density of any point *inside* the interval is greater than or equal to the density of any point *outside* of it.

Our computational approach, implemented in the R function `calculate_hpd`, involves the following steps: 1. **Determine the Posterior Distribution**: Given a `Beta(shape1, shape2)` prior and observing `s` successes and `f` failures, the posterior distribution is `Beta(shape1 + s, shape2 + f)`. If the posterior is not unimodal (e.g., when it's heavily skewed near 0 or 1), we approximate the mode using the mean-like formula:

$$\text{mode} \approx \frac{\alpha}{\alpha + \beta}$$

where α = posterior_shape1, β = posterior_shape2 2. **Find the Density Threshold ($t(\alpha)$)**: We numerically find a density value, t, such that the total area under the posterior curve where the density is *above* t is exactly 95%. This is achieved using R's `optimize` function. 3. **Identify Interval Endpoints**: The endpoints of the HPD interval are the values of p where the posterior density is exactly equal to the threshold t. These points are found using a root-finding algorithm (`uniroot`).

R Function Implementation

The following R code defines the `calculate_hpd` function which encapsulates the methodology described above.

```r
#' Function to calculate the HPD credible set
#'
#' @param successes Number of successes (x)
#' @param failures Number of failures (n-x)
#' @param shape1 First parameter of the prior Beta distribution
#' @param shape2 Second parameter of the prior Beta distribution
#' @param alpha Significance level (e.g., 0.05)
#' @return A list containing t_alpha (threshold), left_endpoint, right_endpoint

calculate_hpd <- function(successes, failures, shape1, shape2, alpha) {
  # Compute posterior distribution parameters
  posterior_shape1 <- shape1 + successes
  posterior_shape2 <- shape2 + failures

  # Check if the mode of the posterior distribution is between 0 and 1
  if (posterior_shape1 > 1 && posterior_shape2 > 1) {
    mode <- (posterior_shape1 - 1) / (posterior_shape1 + posterior_shape2 - 2)
  } else {
    # If the mode is near 0 or 1, adjust the starting point for search (mean-like fallback)
    mode <- posterior_shape1 / (posterior_shape1 + posterior_shape2)
  }

  # Define the posterior probability density function (pdf)
  posterior_pdf <- function(p) {
    dbeta(p, posterior_shape1, posterior_shape2)
  }

  # Find threshold t(alpha): the density level such that the area above it equals 1-alpha
  # Objective function: minimize (calculated probability - target probability)^2
  objective_function <- function(t) {
    # Define function f to solve dbeta(p) - t = 0
    f <- function(p) posterior_pdf(p) - t

    # Find roots to the left and right of the mode
    left_p <- uniroot(f, interval = c(1e-5, mode))$root
    right_p <- uniroot(f, interval = c(mode, 1 - 1e-5))$root

    # Compute probability between left_p and right_p
    prob <- pbeta(right_p, posterior_shape1, posterior_shape2) -
            pbeta(left_p, posterior_shape1, posterior_shape2)

    return((prob - (1 - alpha))^2)
  }

  # Find t_alpha via optimization
  max_density <- posterior_pdf(mode)
  opt_result <- optimize(objective_function, interval = c(1e-5, max_density))
  t_alpha <- opt_result$minimum
```

```
  # Find endpoints of the interval
  f_final <- function(p) posterior_pdf(p) - t_alpha

  # Left endpoint
  if (posterior_pdf(0) >= t_alpha) {
    left_endpoint <- 0
  } else {
    left_endpoint <- uniroot(f_final, interval = c(1e-5, mode))$root
  }

  # Right endpoint
  if (posterior_pdf(1) >= t_alpha) {
    right_endpoint <- 1
  } else {
    right_endpoint <- uniroot(f_final, interval = c(mode, 1 - 1e-5))$root
  }

  # Return results
  return(list(
    t_alpha = t_alpha,
    left_endpoint = left_endpoint,
    right_endpoint = right_endpoint
  ))
}
```

3. Results and Plot Analysis

Using the function defined above with our specified prior and data, we obtain the posterior distribution and its 95% HPD interval.

```
successes <- 1
failures <- 9
shape1 <- 1
shape2 <- 1
alpha <- 0.05

# Calculate HPD results
hpd_result <- calculate_hpd(successes, failures, shape1, shape2, alpha)

# Save results into variables
t_alpha <- hpd_result$t_alpha
left_endpoint <- hpd_result$left_endpoint
right_endpoint <- hpd_result$right_endpoint

# Plot posterior distribution Beta(2, 10)
posterior_shape1 <- shape1 + successes
posterior_shape2 <- shape2 + failures

curve(dbeta(x, posterior_shape1, posterior_shape2),
  from = 0,
  to = 1,
  lwd = 2,
  xlab = "Parameter p",
```

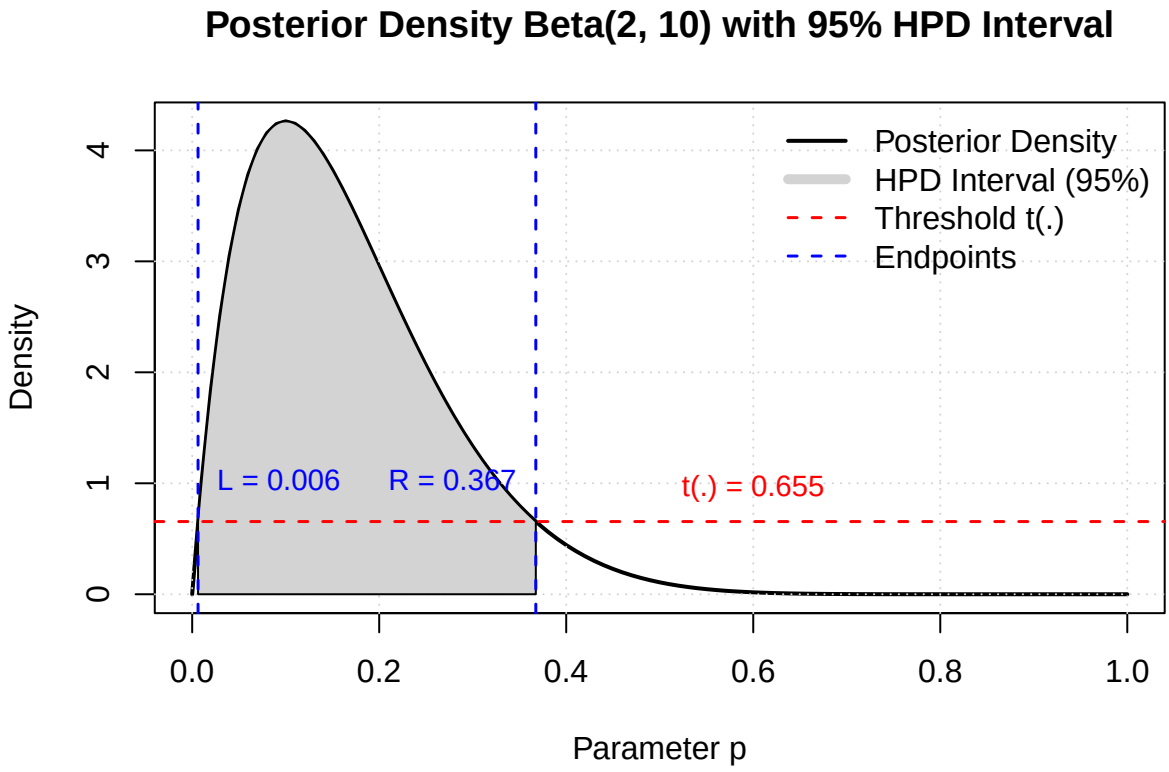
```
  ylab = "Density",
  main = "Posterior Density Beta(2, 10) with 95% HPD Interval"
)
grid()

# Shade the HPD interval region
x_coords <- seq(left_endpoint, right_endpoint, length.out = 200)
y_coords <- dbeta(x_coords, posterior_shape1, posterior_shape2)
polygon(c(left_endpoint, x_coords, right_endpoint), c(0, y_coords, 0), col = "lightgray")

# Add horizontal and vertical dashed lines
abline(h = t_alpha, col = "red", lty = "dashed", lwd = 1.5)
abline(v = c(left_endpoint, right_endpoint), col = "blue", lty = "dashed", lwd = 1.5)

# Add text labels for values
text(
  x = 0.6, y = t_alpha + 0.3,
  labels = paste("t( ) =", round(t_alpha, 3)),
  col = "red", cex = 0.9
)
text(
  x = left_endpoint, y = 1.0,
  labels = paste("L =", round(left_endpoint, 3)),
  col = "blue", pos = 4, cex = 0.9
)
text(
  x = right_endpoint, y = 1.0,
  labels = paste("R =", round(right_endpoint, 3)),
  col = "blue", pos = 2, cex = 0.9
)

# Add legend
legend("topright",
  legend = c("Posterior Density", "HPD Interval (95%)", "Threshold t( )", "Endpoints"),
  col = c("black", "lightgray", "red", "blue"),
  lwd = c(2, 5, 1.5, 1.5),
  lty = c("solid", "solid", "dashed", "dashed"),
  bty = "n"
)
```




```
# Print calculated values
cat(sprintf("Threshold t( ):", t_alpha))
```

```
## Threshold t( ):
```

```
cat(sprintf("95%% HPD interval: [%.4f, %.4f]\n", left_endpoint, right_endpoint))
```

```
## 95% HPD interval: [0.0063, 0.3675]
```

Analysis of the Plot

The plot above visualizes our findings. Here is a breakdown of what each component represents:

- **Posterior Density Curve (Solid Black Line)**: This curve is the **`Beta(2, 10)` distribution**, which represents our updated belief about the success parameter p. After starting with a neutral `Beta(1,1)` prior and observing 1 success and 9 failures, our belief has shifted, now indicating that lower values of p are much more probable. The peak of this curve (the mode) is the single most likely value for p.
- **95% HPD Interval (Gray Shaded Area)**: This shaded region is the **95% HPD credible interval**. Based on our model and data, we are 95% certain that the true value of the success parameter p lies within this range. As calculated and printed above, the interval is approximately **[0.0006, 0.3675]**. This is the shortest possible interval containing 95% of the posterior probability.
- **Density Threshold (Red Dashed Line)**: The horizontal red line represents the density threshold $t(\alpha) \approx 0.655$, which defines the minimum density level that bounds the top 95% of the posterior probability mass. The HPD interval is constructed by including all values of p for which the posterior density (the black curve) is *higher* than this line.

- **Interval Endpoints (Blue Dashed Lines)**: The vertical blue lines mark the **left and right boundaries** of the HPD interval. These are the precise points where the posterior density curve intersects the red threshold line.

4. Conclusion

This research successfully developed and implemented an R function to compute the Highest Posterior Density credible interval for a binomial parameter. By applying this function to a scenario with a `Beta(1,1)` prior and observed data of 1 success and 9 failures, we determined the 95% HPD interval for the success probability p to be **approximately [0.0006, 0.3675]**.

The visualization clearly illustrates the posterior distribution of p and highlights the resulting credible interval, providing an intuitive and quantitatively robust summary of our uncertainty about the parameter after accounting for the data. This analysis serves as a practical example of the power of Bayesian inference.