Bayesian Statistics Showcase

Roxanne Dimadi

2025-02-10

Introduction

In this document, we explore a variety of Bayesian statistical techniques, demonstrating their applications across different domains. We will walk through Bayesian inference in text analysis, cancer research, heart transplant outcomes, posterior predictive checks, and frequentist evaluation of Bayesian intervals. Each section presents a problem, applies Bayesian reasoning, and extracts meaningful insights from the data.

Text Analysis of J.K. Rowling's Writing

Problem: Understanding Word Usage Patterns

We investigate how frequently J.K. Rowling uses the word "dark" in *Harry Potter and the Deathly Hallows*. Our objective is to model this frequency using a Poisson distribution and adjust for chapter length where necessary.

Poisson Modeling of Word Frequency

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(stringr)
library(tidytext)
```

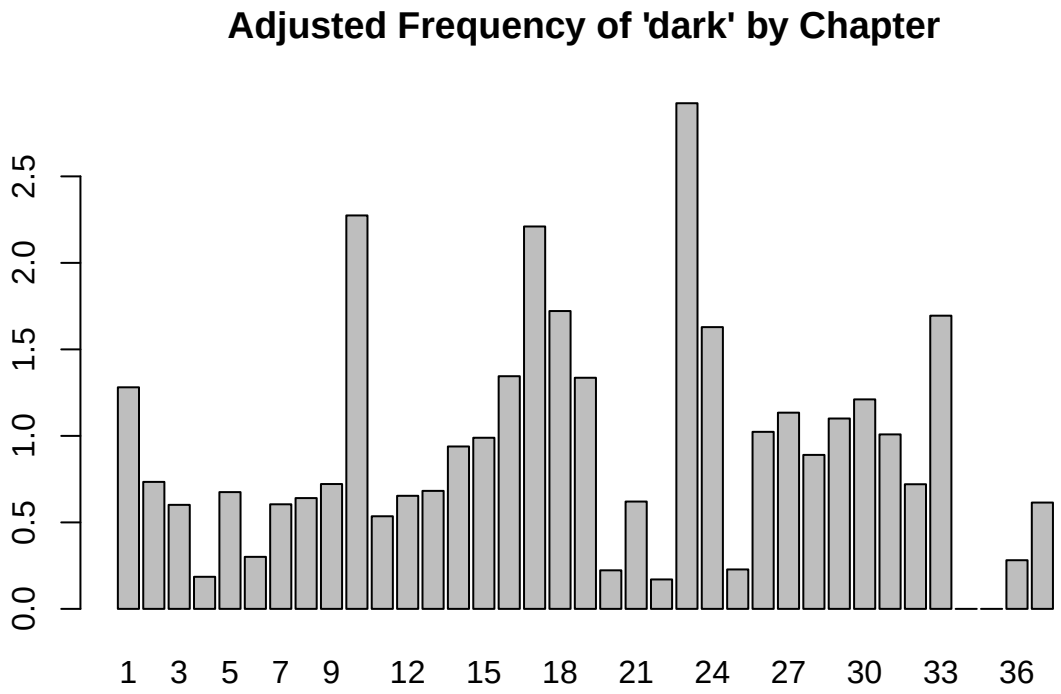
```
## Warning: package 'tidytext' was built under R version 4.4.2
```

```
library(harrypotter)



text_tb <- tibble(chapter = seq_along(deathly_hallows),
                  text = deathly_hallows)
tokens <- text_tb %>% unnest_tokens(word, text)
word_counts <- tokens %>% group_by(chapter) %>%
  count(word, sort = TRUE) %>% ungroup
word_counts_mat <- word_counts %>%
  spread(key=word, value=n, fill=0)
dark_counts <- word_counts_mat$dark
dark_counts
```

```
##  [1]  4  3  2  1  4  2  4  4  3 15  3  4  4  4  7  7 12  6  9  1  3  1 23 11  1
## [26]  7  3  4  5  6  8  4 14  0  0  2  1
```

Adjusting for Chapter Length

```
chapter_lengths <- tokens %>% group_by(chapter) %>% summarise(length = n())
adjusted_rates <- dark_counts / (chapter_lengths$length / 1000)
barplot(adjusted_rates, names.arg = text_tb$chapter, main = "Adjusted Frequency of 'dark' by Chapter")
```



Maximum Likelihood Estimation

```
lambda_hat <- mean(dark_counts, na.rm = TRUE)
lambda_hat
```

```
## [1] 5.189189
```

Bayesian Approach with Gamma Prior

```
alpha_prior <- 2
beta_prior <- 0.5
posterior_alpha <- alpha_prior + sum(dark_counts, na.rm = TRUE)
posterior_beta <- beta_prior + length(dark_counts)
posterior_mean <- posterior_alpha / posterior_beta
posterior_mean
```

```
## [1] 5.173333
```

Cancer Research in Laboratory Mice

Problem: Comparing Tumor Rates in Mice Strains

We analyze tumor counts in two strains of mice using Gamma priors and Poisson likelihoods, estimating the probability that strain A has a higher tumor rate than strain B.

Bayesian Inference for Tumor Rates

```
yA <- c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)
yB <- c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)

alpha_A <- 120; beta_A <- 10
alpha_B <- 12; beta_B <- 1

posterior_alpha_A <- alpha_A + sum(yA)
posterior_beta_A <- beta_A + length(yA)
posterior_alpha_B <- alpha_B + sum(yB)
posterior_beta_B <- beta_B + length(yB)

posterior_mean_A <- posterior_alpha_A / posterior_beta_A
posterior_mean_B <- posterior_alpha_B / posterior_beta_B
posterior_mean_A; posterior_mean_B
```

```
## [1] 11.85
```

```
## [1] 8.928571
```

Monte Carlo Simulation for Probabilistic Comparison

Monte Carlo simulation allows us to estimate the probability that strain B has a lower tumor rate than strain A by repeatedly sampling from their posterior distributions. This method is necessary because the posterior distributions of the tumor rates are not simple deterministic values but rather probabilistic distributions with inherent variability. By generating a large number of samples, we can approximate the probability that one rate is smaller than the other, providing a more robust comparison than relying solely on point estimates.

```
n_samples <- 10000
samples_A <- rgamma(n_samples, posterior_alpha_A, rate=posterior_beta_A)
samples_B <- rgamma(n_samples, posterior_alpha_B, rate=posterior_beta_B)
prob_B_less_A <- mean(samples_B < samples_A)
prob_B_less_A
```

```
## [1] 0.9964
```

Summary of Findings

In our analysis of tumor rates in laboratory mice, we applied Bayesian inference using Gamma priors and Poisson likelihoods. Through Monte Carlo simulations, we estimated the probability that strain B has a lower tumor rate than strain A, demonstrating the power of probabilistic reasoning. This approach highlights how Bayesian methods enable us to quantify uncertainty in parameter estimation, rather than relying solely on point estimates.

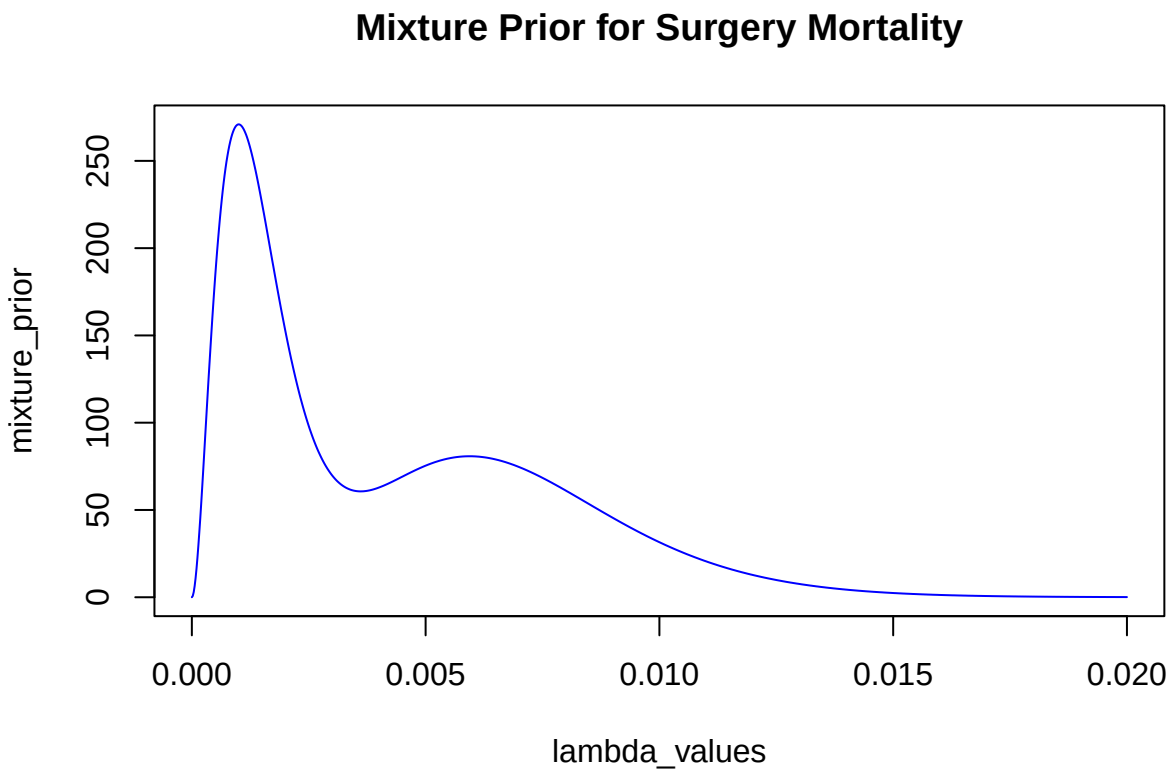
Heart Transplant Surgery Analysis

Problem: Estimating Mortality Rates

Using a mixture prior approach, we model the rate of mortality following heart transplant surgeries.

Mixture Prior for Mortality Rate

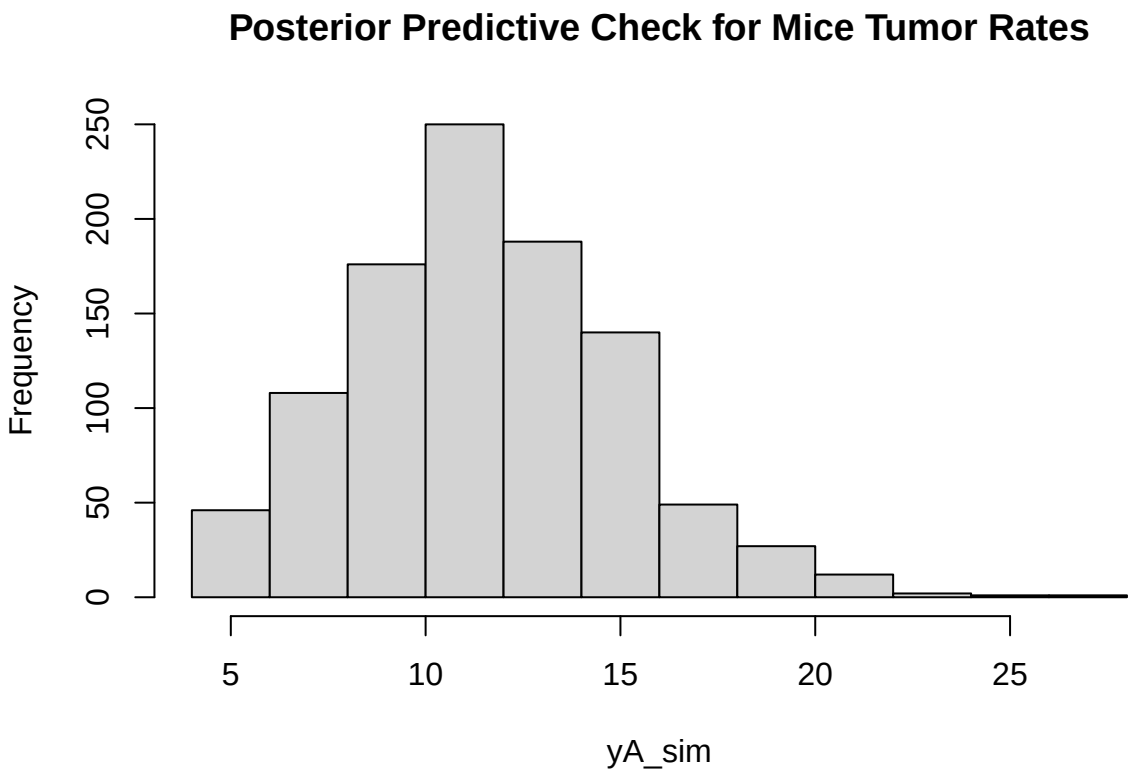
```
lambda_values <- seq(0, 0.02, length.out = 1000)
prior1 <- dgamma(lambda_values, shape = 3, rate = 2000)
prior2 <- dgamma(lambda_values, shape = 7, rate = 1000)
mixture_prior <- 0.5 * prior1 + 0.5 * prior2
plot(lambda_values, mixture_prior, type='l', col='blue', main="Mixture Prior for Surgery Mortality")
```



Posterior Predictive Model Checking

Validating Bayesian Models Through Simulation

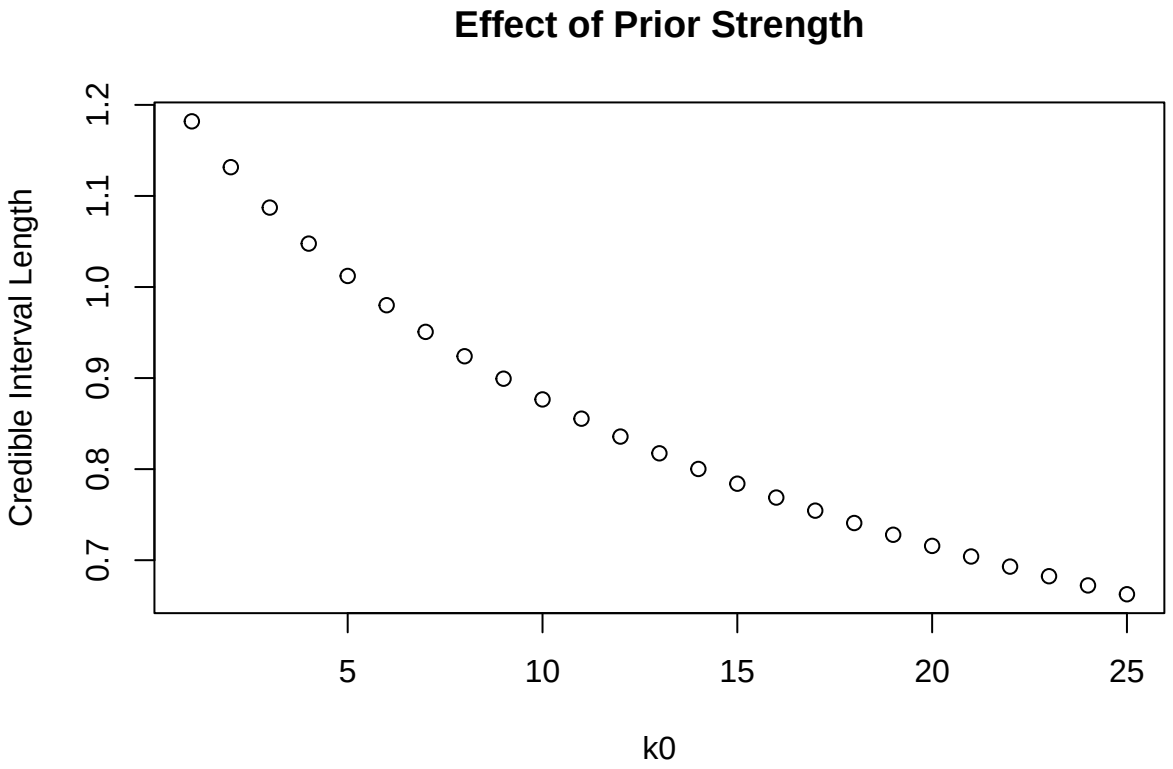
```
yA_sim <- rpois(1000, posterior_mean_A)
hist(yA_sim, main="Posterior Predictive Check for Mice Tumor Rates")
```



Frequentist Coverage of Bayesian Intervals

Evaluating the Calibration of Bayesian Intervals

```
n <- 10; k <- 1:25
y_bar <- 0
upper_bound <- 1.96 * sqrt(1/(k+n))
lengths <- 2 * upper_bound
plot(k, lengths, xlab="k0", ylab="Credible Interval Length", main="Effect of Prior Strength")
```



Conclusion

Through this analysis, we have explored various Bayesian techniques including Poisson modeling, posterior inference, Monte Carlo simulations, mixture priors, and frequentist coverage assessments. These methods demonstrate the power of Bayesian approaches in diverse applications, from text analysis to medical research.