Bayesian Inference

Introduction and Ideas

Abhinavan, Anant, Prateek, Mukesh

Indian Statistical Institute

May 2026

What is Bayesian Inference?

There are two main schools of thought regarding the approach to statistical inference (also called learning). We are familiar to an extent with frequentist methods. The key difference lies in the interpretation of "probability" in the two schools of thought.

- **Frequentist Inference.** Probabilities are interpreted as long-run frequencies. The goal is to create procedures with long-run frequency guarantees.
- **Bayesian Inference.** Probabilities are interpreted as subjective degrees of belief. The goal is to state and analyze one's beliefs.

Essentially, the Bayesian approach requires us to drop the idea of probability as a long-run frequency and instead think of it in a subjective sense, conditioned on known facts.

Following from this philosophy, the parameter θ of any probability distribution is taken to be *fixed* in the frequentist approach. On the contrary, in the Bayesian approach θ is treated as a *random variable* with its own distribution — chosen subjectively based on prior knowledge — called the **prior**.

The Prior and the Posterior

The prior density is represented by $\pi(\theta)$.

Given a sample $X_1, X_2, \ldots, X_n$, we wish to find a new density for θ that incorporates the information contained in the observations. This new density, conditioned on the data, is known as the **posterior** and is denoted

$$p(\theta \mid X_1, X_2, \ldots, X_n)\,.$$

Computing the Posterior via Bayes' Theorem

We recall Bayes' theorem for probability densities:

$$f_X(x \mid y) = \frac{f_Y(y \mid x)\, f_X(x)}{f_Y(y)} = \frac{f_Y(y \mid x)\, f_X(x)}{\displaystyle\int_{-\infty}^{\infty} f_Y(y \mid x)\, f_X(x)\, dx}. \tag{1}$$

Setting $x = \theta$ and $y = \mathbf{x} = (x_1, \ldots, x_n)$, we obtain:

$$f_{\theta|\mathbf{X}}(\theta \mid \mathbf{x}) = \frac{f_{\mathbf{X}|\theta}(\mathbf{x} \mid \theta)\; f_\theta(\theta)}{f_{\mathbf{X}}(\mathbf{x})}. \tag{2}$$

Renaming each factor by its role in the Bayesian framework:

$$\underbrace{p(\theta \mid \mathbf{X})}_{\text{posterior}} = \frac{\underbrace{p(\mathbf{X} \mid \theta)}_{\text{likelihood}}\;\underbrace{\pi(\theta)}_{\text{prior}}}{\underbrace{p(\mathbf{X})}_{\text{evidence}}}. \tag{3}$$

This can equivalently be written as:

$$p(\theta \mid X_1, X_2, \ldots, X_n) = \frac{p(X_1, X_2, \ldots, X_n \mid \theta)\ \pi(\theta)}{p(X_1, X_2, \ldots, X_n)} = \frac{L_n(\theta)\ \pi(\theta)}{c_n} \ \propto\ L_n(\theta)\ \pi(\theta), \quad (4)$$

where $L_n(\theta)$ is the familiar likelihood function and the denominator $c_n = p(X_1, \ldots, X_n)$ is called the **evidence**. Since the evidence does not depend on θ, the posterior is proportional to the product of the likelihood and the prior.

Inference Using the Posterior

The posterior $p(\theta \mid X_1, \ldots, X_n)$ is the central object used for inference in the Bayesian approach.

Posterior Mean

A natural point estimate of θ is the posterior mean:

$$\bar{\theta} = \int \theta\ p(\theta \mid X_1, X_2, \ldots, X_n)\ d\theta = \frac{\int \theta\ L_n(\theta)\ \pi(\theta)\ d\theta}{\int L_n(\theta)\ \pi(\theta)\ d\theta}. \quad (5)$$

Interval Estimation: Frequentist vs. Bayesian

To illustrate the difference between the two approaches, consider interval estimation. Suppose

$$X_1, X_2, \ldots, X_n\ \sim\ N(\theta, 1),$$

and we wish to construct an interval estimate for θ.

Frequentist Confidence Interval

The classical $(1-\alpha)$ confidence interval is:

$$C_1 = \left[\bar{X}_n - \frac{\tau_{\alpha/2}}{\sqrt{n}},\ \bar{X}_n + \frac{\tau_{\alpha/2}}{\sqrt{n}}\right], \quad (6)$$

where τ_α denotes the upper α-quantile of the $N(0,1)$ distribution.

Bayesian Credible Interval

In the Bayesian approach, we seek an interval $C_2 = [a, b]$ such that:

$$\int_{C_2} p(\theta \mid X_1, X_2, \ldots, X_n)\ d\theta = 1 - \alpha, \quad (7)$$

which is equivalently specified by choosing a and b satisfying:

$$\int_{-\infty}^{a} p(\theta \mid X_1, \ldots, X_n)\ d\theta\ =\ \int_{b}^{\infty} p(\theta \mid X_1, \ldots, X_n)\ d\theta\ =\ \frac{\alpha}{2}. \quad (8)$$

This interval C_2 is known as an **equal-tailed credible interval** (or Bayesian credible set). Unlike a frequentist confidence interval, a credible interval carries the direct probabilistic interpretation: given the observed data, θ lies in C_2 with posterior probability $1-\alpha$.

Conjugate Priors: Definition and Mechanics

A family of prior distributions $\mathcal{P}$ is said to be **conjugate** to a family of sampling distributions $\mathcal{F}$ if, for every prior $\pi \in \mathcal{P}$, the resulting posterior $p(\cdot \mid x)$ also belongs to $\mathcal{P}$. Essentially, the parametric family of the distribution remains closed under sampling.

This property allows for the instant calculation of posterior distributions. Instead of performing complex integrations, one simply updates the hyperparameters of the distribution.

Conjugacy in the Exponential Family

Most standard conjugate pairs arise from the exponential family. Suppose the sampling model takes the form:

$$p(x \mid \theta) = \exp\big(\theta^T x - A(\theta)\big)\,. \tag{1}$$

The conjugate prior for this model is defined as:

$$\pi_{x_0,\,n_0}(\theta) \propto \exp\big(n_0 x_0^T \theta - n_0 A(\theta)\big)\,. \tag{2}$$

Here, n_0 can be interpreted as the number of "virtual observations" and x_0 as their average. After observing n real data points with mean $\bar{X}$, the posterior hyperparameters update as:

- $n_0' = n + n_0$
- $x_0' = \dfrac{n\bar{X} + n_0 x_0}{n + n_0}$

Example: Poisson–Gamma Hyperparameter Update

Consider a Poisson model $X \sim \text{Poisson}(\lambda)$ with a $\text{Gamma}(\alpha, \beta)$ prior. Suppose our prior belief is based on $n_0 = 5$ virtual observations with a virtual average $x_0 = 2$.

1. We observe $n = 10$ new data points with sample mean $\bar{X} = 4$.
2. The updated virtual sample size is: $n_0' = 10 + 5 = 15$.
3. The updated mean is: $x_0' = \dfrac{(10 \times 4) + (5 \times 2)}{15} = \dfrac{50}{15} \approx 3.33$.

The posterior is then a Gamma distribution defined by these updated parameters.

Example: The Poisson–Gamma Conjugate Pair

Suppose we are counting events, so each observation follows $X_i \sim \text{Poisson}(\lambda)$.

The Likelihood: For n observations, the joint likelihood is proportional to:

$$\mathcal{L}(\lambda) \propto \lambda^{\sum X_i} e^{-n\lambda}. \tag{3}$$

The Conjugate Prior: We choose the Gamma distribution for λ. With $\lambda \sim \text{Gamma}(\alpha, \beta)$:

$$\pi(\lambda) \propto \lambda^{\alpha-1} e^{-\beta\lambda}. \tag{4}$$

Observe that the likelihood and the prior share the exact same algebraic structure: $\lambda^{(\cdot)} \times e^{(\cdot)\lambda}$. Multiplying them together:

$$p(\lambda \mid X_1, \ldots, X_n) \propto \left(\lambda^{\sum X_i} e^{-n\lambda}\right) \times \left(\lambda^{\alpha-1} e^{-\beta\lambda}\right). \tag{5}$$

Combining the exponents:

$$p(\lambda \mid X_1, \ldots, X_n) \propto \lambda^{(\sum X_i + \alpha) - 1}\, e^{-(\beta+n)\lambda}. \tag{6}$$

This is exactly the kernel of a new Gamma distribution — no complex integration required:

$$\lambda \mid X_1, \ldots, X_n \;\sim\; \text{Gamma}\big(\alpha + n\bar{X},\; \beta + n\big). \tag{7}$$

Technical Foundations and Proofs

The Linearity of Posterior Means

A significant theoretical result by Diaconis (1979) states that in the continuous case, the posterior mean of the gradient of the log-normalising constant $\nabla A(\theta)$ is linear if and only if the prior is conjugate. Specifically:

$$\mathbb{E}(\nabla A(\theta) \mid X) = aX + b. \tag{8}$$

This linearity ensures that the posterior mean is always a weighted average of the prior information and the observed data.

Non-Exponential Family: Uniform/Pareto

While conjugacy is most common in the exponential family, it exists elsewhere. For the model $X \sim \text{Uniform}(0, \theta)$, the conjugate prior is the Pareto distribution. The posterior updates based on the maximum observed value $X_{(n)}$:

$$\theta \mid X_1, \ldots, X_n \;\sim\; \text{Pareto}\big(n + k,\; \max\{X_{(n)}, \nu_0\}\big). \tag{9}$$

Summary of Key Conjugate Pairs

The following table summarises the most important conjugate relationships used in statistical machine learning.

Sampling Model	**Conjugate Prior**	**Posterior Distribution**
Bernoulli(θ)	Beta(α, β)	Beta($\alpha + n\bar{X},\ \beta + n(1 - \bar{X})$)
Poisson(λ)	Gamma(α, β)	Gamma($\alpha + n\bar{X},\ \beta + n$)
Normal (known σ^2)	Normal(μ_0, σ_0^2)	Normal (weighted average)
Normal (known μ)	Inverse-Gamma	Inverse-Gamma
Multinomial(θ)	Dirichlet(α)	Dirichlet($\alpha + n\bar{X}$)
Uniform($0, \theta$)	Pareto(ν_0, k)	Pareto($\max\{\nu_0, X_{(n)}\},\ n + k$)

Conjugate Priors in Machine Learning

Conjugate priors are not merely mathematical conveniences; they underpin classical machine learning algorithms.

- **Text Classification (Spam Filters):** Naïve Bayes classifiers often use the Dirichlet–Multinomial conjugate pair. The model counts word occurrences (Multinomial); the Dirichlet prior acts as Laplace smoothing, adding virtual word counts so the model does not break on unseen vocabulary. The posterior update is fast enough to allow real-time learning.
- **Topic Modelling (LDA):** Latent Dirichlet Allocation relies on the same Dirichlet–Multinomial conjugacy to infer document-topic and topic-word distributions at scale.
- In Big Data settings, running full MCMC simulations for every new data point is computationally infeasible. Conjugate updates provide the closed-form arithmetic required for real-time inference.

The Limitations of Conjugacy

If conjugate priors are so easy to compute with, why do we not always use them?

- **Inflexibility:** The prior must be chosen for mathematical convenience rather than to faithfully reflect genuine subjective beliefs.
- **Complex Models:** In modern hierarchical models or deep learning, conjugate priors simply do not exist for the intricate relationships we wish to model.
- **The Solution:** When we step outside the bounds of conjugacy, the posterior integral becomes analytically intractable. This is precisely the motivation for computational simulation methods such as MCMC.

Conclusion

Conjugate priors serve as a mathematically convenient tool that bypasses the need for high-dimensional integration. By providing closed-form posterior updates, they power modern applications such as Topic Modelling (LDA) and Naïve Bayes classifiers, where computational efficiency is paramount. Understanding their theoretical foundations — and their limitations — is essential for knowing when to use them and when to turn to simulation-based methods instead.

Example 1: Bernoulli/Beta Conjugate Model

In the Bernoulli/Beta conjugate model, we observe n Bernoulli trials with S_n successes. Taking a Uniform prior on θ, the posterior is Beta$(S_n + 1,\ n - S_n + 1)$. The following R code constructs and plots the prior, scaled likelihood, and posterior for $n = 15$, $S_n = 7$.

```r
n <- 15
S_n <- 7
alpha <- S_n + 1  # 8
beta <- n - S_n + 1  # 9

# Create theta values (grid from 0 to 1)
theta <- seq(0, 1, length.out = 1000)

# 1. Prior (Uniform)
prior <- dunif(theta, min = 0, max = 1)

# 2. Likelihood (scaled to maximum = 1)
likelihood <- theta^S_n * (1 - theta)^(n - S_n)
likelihood_scaled <- likelihood / max(likelihood)

# 3. Posterior (Beta(8, 9))
posterior <- dbeta(theta, shape1 = alpha, shape2 = beta)

# Create the plot
plot(theta, posterior, type = "l",
     col = "red", lwd = 3,
     xlab = expression(theta),
     ylab = "Density / Scaled Value",
     main = paste0("Prior, Likelihood (Scaled), and Posterior for n = ", n,
                   ", S_n = ", S_n, "\nPosterior: Beta(", alpha, ", ", beta, ")"),
     ylim = c(0, max(posterior, prior, likelihood_scaled) * 1.1))

# Add prior (uniform)
lines(theta, prior, col = "blue", lwd = 3, lty = 2)

# Add scaled likelihood
lines(theta, likelihood_scaled, col = "green", lwd = 3, lty = 3)

# Add legend
legend("topright",
       legend = c("Prior (Uniform)",
                  "Likelihood (scaled)",
                  paste0("Posterior: Beta(", alpha, ", ", beta, ")")),
       col = c("blue", "green", "red"),
       lty = c(2, 3, 1),
       lwd = 3,
       bty = "o",
       bg = "white")
```

Example 2: Normal/Normal Conjugate Model

In the Normal/Normal conjugate model, the prior is $\theta \sim N(a, b^2)$ and the likelihood comes from $X_i \,|\, \theta \sim N(\theta, \sigma^2)$. The posterior is also Normal, with mean a weighted combination

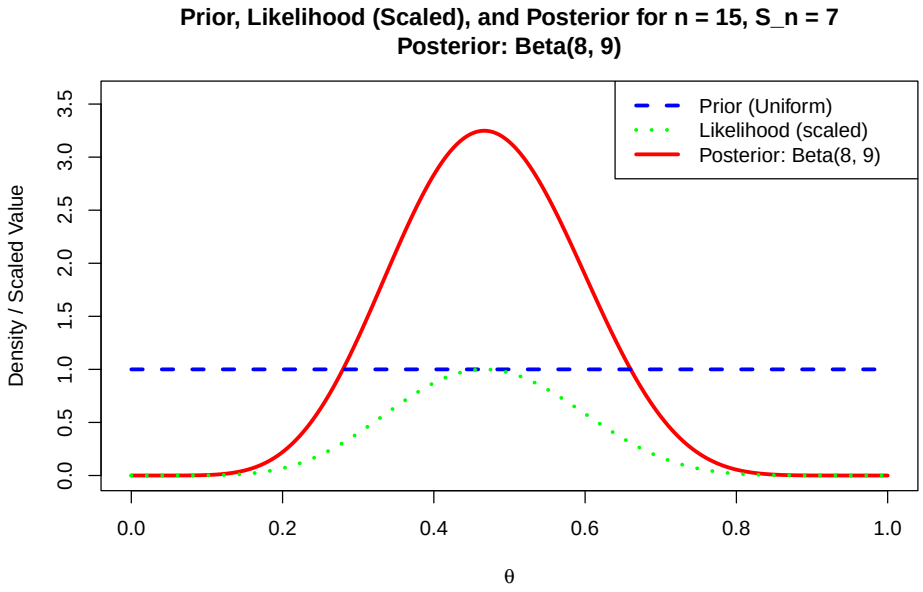

Figure 1: Bernoulli/Beta Conjugate Model: Prior (Uniform), Likelihood (scaled), and Posterior (Beta$(8, 9)$).

of the prior mean and the sample mean, and a variance that shrinks with increasing sample size. The following R code computes and plots the three densities for $n = 10$, $\bar{X} = 0.5$, $\sigma = 1$, $a = 0$, $b = 1$.

```r
n <- 10
Xbar <- 0.5
sigma <- 1
a <- 0
b <- 1
w <- (n / sigma^2) / (n / sigma^2 + 1 / b^2)
theta_bar <- w * Xbar + (1 - w) * a
tau <- sqrt(1 / (n / sigma^2 + 1 / b^2))
cat("Posterior mean (theta_bar):", theta_bar, "\n")
cat("Posterior standard deviation (tau):", tau, "\n")
cat("Weight (w):", w, "\n")

theta <- seq(-3, 3, length.out = 1000)

# 1. Prior: N(a, b^2)
prior <- dnorm(theta, mean = a, sd = b)

# 2. Likelihood: proportional to N(Xbar, sigma^2/n)
likelihood <- dnorm(theta, mean = Xbar, sd = sigma / sqrt(n))
likelihood_scaled <- likelihood / max(likelihood)

# 3. Posterior: N(theta_bar, tau^2)
posterior <- dnorm(theta, mean = theta_bar, sd = tau)

# Create the plot
plot(theta, posterior, type = "l",
     col = "red", lwd = 3,
     xlab = expression(theta),
     ylab = "Density / Scaled Value",
```

```r
     main = paste0("Prior, Likelihood (Scaled), and Posterior\n",
                   "Normal/Normal Model (n = ", n, ", Xbar = ", Xbar,
                   ", sigma = ", sigma, ")"),
     ylim = c(0, max(posterior, prior, likelihood_scaled) * 1.1))
lines(theta, prior, col = "blue", lwd = 3, lty = 2)
lines(theta, likelihood_scaled, col = "green", lwd = 3, lty = 3)
legend("topright",
       legend = c(paste0("Prior: N(", a, ", ", b^2, ")"),
                  "Likelihood (scaled)",
                  paste0("Posterior: N(", round(theta_bar, 3), ", ",
                         round(tau^2, 3), ")")),
       col = c("blue", "green", "red"),
       lty = c(2, 3, 1),
       lwd = 3,
       bty = "o",
       bg = "white")
```

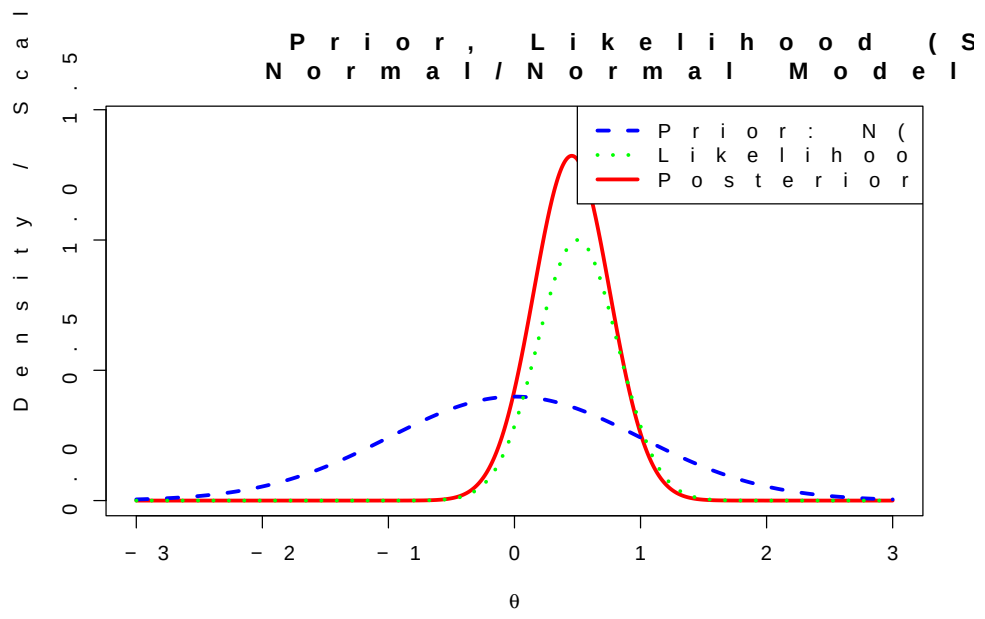

Figure 2: Normal/Normal Conjugate Model: Prior ($\theta \sim N(a, b^2)$), Likelihood (scaled), and Posterior (Normal distribution).

Inference about Functions of Parameters

Definition

A common inferential goal is to reason about a derived quantity $\tau = g(\theta)$, where g is a known (possibly nonlinear) function. The posterior distribution of τ is characterised by its CDF:

$$H(t \mid D_n) = \int_A p(\theta \mid D_n)\, d\theta, \qquad A = \{\theta : g(\theta) \le t\}.$$

When g is monotone, the posterior density of τ follows by the change-of-variables formula:

$$p(\tau \mid D_n) = p\big(g^{-1}(\tau) \mid D_n\big) \cdot \left|\frac{d\theta}{d\tau}\right|.$$

Simulation Approach

Analytical computation of $H(t \mid D_n)$ is often intractable. A universal simulation-based algorithm proceeds as follows:

1. Draw $\theta_1, \ldots, \theta_B \sim p(\theta \mid D_n)$.
2. Transform: $\tau_b = g(\theta_b)$ for $b = 1, \ldots, B$.
3. The empirical distribution of $\{\tau_b\}$ approximates $p(\tau \mid D_n)$.

This approach works for *any* function g; credible intervals are obtained directly as empirical quantiles of $\{\tau_b\}$, and it extends naturally to multiparameter settings.

Multiparameter Problems

General Framework

Let $\theta = (\theta_1, \ldots, \theta_d)^T$ with prior $\pi(\theta)$. The joint posterior satisfies

$$p(\theta \mid D_n) \propto L_n(\theta)\, \pi(\theta).$$

Interest often centres on a single component, say θ_1. Its marginal posterior requires integrating out all other parameters:

$$p(\theta_1 \mid D_n) = \int \cdots \int p(\theta_1, \ldots, \theta_d \mid D_n)\, d\theta_2 \cdots d\theta_d,$$

which is generally intractable in closed form. The simulation solution avoids this entirely:

1. Draw $\theta^{(1)}, \ldots, \theta^{(B)} \sim p(\theta \mid D_n)$.
2. Collect the first components $\theta_1^{(1)}, \ldots, \theta_1^{(B)}$.
3. These form a sample from $p(\theta_1 \mid D_n)$ — no numerical integration required.

Example: Treatment Effect $\tau = \theta_2 - \theta_1$

Consider two independent Bernoulli experiments: Group 1 (control) with n_1 trials and X_1 successes, and Group 2 (treatment) with n_2 trials and X_2 successes. Under independent Uniform priors, the posteriors are:

$$\theta_1 \mid X_1 \sim \text{Beta}(X_1 + 1,\ n_1 - X_1 + 1), \qquad \theta_2 \mid X_2 \sim \text{Beta}(X_2 + 1,\ n_2 - X_2 + 1).$$

The treatment effect $\tau = \theta_2 - \theta_1$ has no simple closed-form posterior, so we proceed by simulation:

1. Draw $\theta_1^{(b)} \sim \text{Beta}(X_1 + 1,\ n_1 - X_1 + 1)$ for $b = 1, \ldots, B$.
2. Draw $\theta_2^{(b)} \sim \text{Beta}(X_2 + 1,\ n_2 - X_2 + 1)$ independently.
3. Compute $\tau^{(b)} = \theta_2^{(b)} - \theta_1^{(b)}$.
4. $\{\tau^{(1)}, \ldots, \tau^{(B)}\}$ is a sample from $p(\tau \mid X_1, X_2)$.

R Implementation

```r
# Group 1: Control group
n1 <- 40
X1 <- 25
# Group 2: Treatment group
n2 <- 40
X2 <- 30
B <- 10000

theta1 <- rbeta(B, shape1 = X1 + 1, shape2 = n1 - X1 + 1)
theta2 <- rbeta(B, shape1 = X2 + 1, shape2 = n2 - X2 + 1)
tau <- theta2 - theta1

post_mean <- mean(tau)
cred_interval <- quantile(tau, probs = c(0.025, 0.975))
post_median <- median(tau)
post_sd <- sd(tau)

cat("========================================\n")
cat("Treatment Effect (tau = theta2 - theta1)\n")
cat("========================================\n")
cat("Group 1 (Control):", n1, "trials,", X1, "successes\n")
cat("Group 2 (Treatment):", n2, "trials,", X2, "successes\n")
cat("----------------------------------------\n")
cat("Posterior Mean     :", round(post_mean, 4), "\n")
cat("Posterior Median   :", round(post_median, 4), "\n")
cat("Posterior SD       :", round(post_sd, 4), "\n")
cat("95% Credible Interval: [", round(cred_interval[1], 4), ",",
    round(cred_interval[2], 4), "]\n")
cat("----------------------------------------\n")
if (cred_interval[1] < 0 && cred_interval[2] > 0) {
  cat("Conclusion: 0 is INSIDE the 95% credible interval\n")
  cat("No significant treatment effect (at 95% level)\n")
} else if (cred_interval[2] < 0) {
  cat("Conclusion: 0 is BELOW the 95% credible interval\n")
  cat("Treatment has significant NEGATIVE effect\n")
} else {
  cat("Conclusion: 0 is ABOVE the 95% credible interval\n")
  cat("Treatment has significant POSITIVE effect\n")
}
cat("========================================\n\n")

par(mfrow = c(2, 2))

# Plot 1: Histogram of tau with 95% CI
hist(tau,
     breaks = 50,
     col = "lightblue",
     main = "Posterior of Treatment Effect\n(tau = theta2 - theta1)",
     xlab = expression(tau),
     ylab = "Frequency",
     border = "white",
     probability = TRUE)
lines(density(tau), col = "darkblue", lwd = 2)
abline(v = 0, col = "red", lwd = 3, lty = 2)
abline(v = cred_interval[1], col = "darkgreen", lwd = 2, lty = 3)
abline(v = cred_interval[2], col = "darkgreen", lwd = 2, lty = 3)
abline(v = post_mean, col = "purple", lwd = 2, lty = 4)
```

```r
legend("topright",
       legend = c("Posterior Density", "0 (No effect)", "95% CI", "
    Posterior Mean"),
       col = c("darkblue", "red", "darkgreen", "purple"),
       lty = c(1, 2, 3, 4),
       lwd = 2,
       bty = "n",
       cex = 0.8)

# Plot 2: Density of tau with shaded 95% CI
plot(density(tau),
     main = "Posterior Density with 95% Credible Interval",
     xlab = expression(tau),
     ylab = "Density",
     col = "darkblue",
     lwd = 2)
dens <- density(tau)
x_shade <- dens$x[dens$x >= cred_interval[1] & dens$x <= cred_interval
    [2]]
y_shade <- dens$y[dens$x >= cred_interval[1] & dens$x <= cred_interval
    [2]]
polygon(c(x_shade[1], x_shade, x_shade[length(x_shade)]),
        c(0, y_shade, 0),
        col = rgb(0, 0, 1, 0.3), border = NA)
abline(v = 0, col = "red", lwd = 2, lty = 2)
abline(v = post_mean, col = "purple", lwd = 2, lty = 4)

# Plot 3: Trace plots of theta1 and theta2 (first 1000 samples)
plot(theta1[1:1000], type = "l",
     col = "blue",
     main = "Trace Plot (First 1000 samples)",
     xlab = "Iteration",
     ylab = expression(theta),
     ylim = range(c(theta1[1:1000], theta2[1:1000])))
lines(theta2[1:1000], col = "red")
legend("topright",
       legend = c(expression(theta[1]), expression(theta[2])),
       col = c("blue", "red"),
       lty = 1,
       lwd = 2,
       bty = "n")

# Plot 4: Scatter plot of theta2 vs theta1
plot(theta1, theta2,
     col = rgb(0, 0, 0, 0.1),
     pch = 16,
     main = "Joint Posterior Distribution",
     xlab = expression(theta[1] * " (Control)"),
     ylab = expression(theta[2] * " (Treatment)"))
abline(0, 1, col = "red", lwd = 2, lty = 2)
cor_val <- cor(theta1, theta2)
text(max(theta1) * 0.7, max(theta2) * 0.9,
     paste("Correlation =", round(cor_val, 3)),
     cex = 0.9)

par(mfrow = c(1, 1))
```

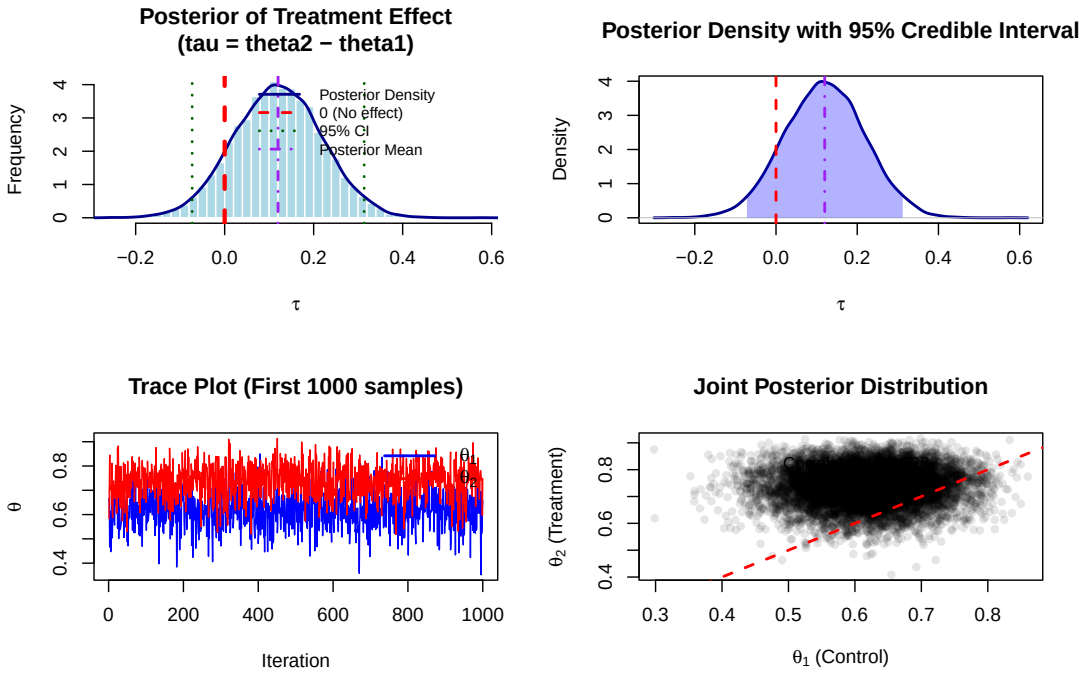

Figure 3: Multiparameter Example: Posterior analysis of the treatment effect $\tau = \theta_2 - \theta_1$, showing the posterior histogram with 95% credible interval, shaded density, trace plots, and joint posterior distribution.

Frequentist vs. Bayesian Testing

Frequentist inference and Bayesian inference are two fundamentally different processes. The core difference lies in how the data are used to perform inference.

Frequentist Testing	Bayesian Testing
Probability based on repeated sampling	Probability interpreted as degree of belief
"How unusual is the observed data under H_0?"	"How probable is the hypothesis after observing the data?"
Probability assigned to procedures (p-value, rejection region)	Probability assigned directly to hypotheses

Our Goal: Study Bayesian testing, Bayes factors, and the BIC.

Development of Bayesian Testing

We test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ with prior belief $P(H_0) = p_0$. Under H_1, θ has a prior density $\pi(\theta)$. After observing data D_n, we apply Bayes' Theorem:

$$P(H_0 \mid D_n) = \frac{p(D_n \mid H_0)\, P(H_0)}{p(D_n \mid H_0)\, P(H_0) + p(D_n \mid H_1)\, P(H_1)}.$$

Assuming equal prior beliefs $P(H_0) = P(H_1) = 1/2$, this simplifies to:

$$P(H_0 \mid D_n) = \frac{p(D_n \mid \theta_0)}{p(D_n \mid \theta_0) + \displaystyle\int p(D_n \mid \theta)\, \pi(\theta)\, d\theta}. \tag{1}$$

This yields a direct probability statement about the truth of H_0.

Interpretation and Marginal Likelihood

The numerator $p(D_n \mid \theta_0)$ measures how well the null hypothesis explains the observed data. The integral in the denominator,

$$\int p(D_n \mid \theta)\, \pi(\theta)\, d\theta,$$

represents the *average* explanatory power of all parameter values under the alternative hypothesis. This leads to the central question of Bayesian testing:

> "How well does the null explain the data compared to the entire alternative model?"

The quantity $P(D_n \mid M) = \int L(\theta)\, \pi(\theta)\, d\theta$ is called the **Marginal Likelihood**.

Balance: Fit vs. Complexity

The Bayesian Information Criterion (BIC) approximates twice the log marginal likelihood and embodies a trade-off between two competing forces:

$$\text{BIC} = \underbrace{\log L(\hat{\theta})}_{\text{Model Fit}} - \underbrace{\frac{d}{2}\log n}_{\text{Complexity Penalty}} .$$

- **Model Fit:** Measures how well the model explains the data.
- **Complexity Penalty:** Penalizes the model for each additional parameter. Larger models have larger parameter spaces and greater "uncertainty volume."

Between two models, the one with the *higher* BIC is preferred.

Summary and Limitations

- **Computational Problem:** Exact marginal likelihoods are difficult to calculate in high dimensions.
- **Asymptotics:** BIC is only accurate as $n \to \infty$. In small samples, the constant terms ignored in the BIC derivation may matter.
- **Prior Sensitivity:** Bayesian testing results can change significantly based on the choice of $\pi(\theta)$.