Statistical Inference

Longhai Li

2026-01-15

Preface

This is a concise course about statistical inference.

Key Features

- Use simulation and graphs to illustrate the concepts in probability theory and statistical inference
- Rigourous derivation of the key theorems in statistical inference

Audience

This course requires a strong command of multivariate calculus, alongside a rigorous foundation in intermediate probability theory including asymptotic theorey for probability. Students should also possess prior exposure to applied statistical methods and familiar with basic statistical concepts such as p-value and confidence internal.

1 Introduction to Statistical Inference

1.1 Population Model (Data Model)

We begin with observations (units) $X_1, X_2, \dots, X_n$. These may be vectors. We regard these observations as a realization of random variables.

Definition 1.1 (Population Distribution)**.** We assume that $X_1, X_2, \dots, X_n \sim f(x)$. The function $f(x)$ is called the **population distribution**.

Assumptions and Scope

For simplicity, we often assume the data are Independent and Identically Distributed (i.i.d.). The assumption of identical distribution can be relaxed to regression settings in which the distributions of x_i's are independent but dependent on covariate x_i.

In **Parametric Statistics**, we assume $f(x)$ is of a known analytic form but involves unknown parameters.

Example 1.1 (Parametric Model: Normal)**.** Consider the Normal distribution:

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.1}$$

Here, the parameter space is $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma \in [0, +\infty)\}$. The goal is to learn aspects of the unknown θ from observations $X_1, \dots, X_n$.

Example 1.2 (Parametric Model: Bernoulli)**.** Consider a sequence of binary outcomes (e.g., Success/Failure) where each $X_i \in \{0, 1\}$. We assume $X_i \sim$ Bernoulli(θ). The probability mass function is:

$$f(x;\theta) = \theta^x (1-\theta)^{1-x} \tag{1.2}$$

Here, the parameter space is $\Theta = [0, 1]$, where θ represents the probability of success.

1.2 Probabilistic Model vs. Statistical Inference

There is a fundamental distinction between probability and statistics regarding the parameter θ. We can visualize this using a "shooting target" analogy:

- θ **(The Center):** The true, unknown bullseye location.
- x **(The Shots):** The observed holes on the target board.
- **Probability (Deductive):** The center θ is **known**. We predict where the shots x will land.
- **Statistics (Inductive):** The shots x are **observed** on the board. The center θ is unknown. We hypothesize different potential centers to see which one best explains the shots.

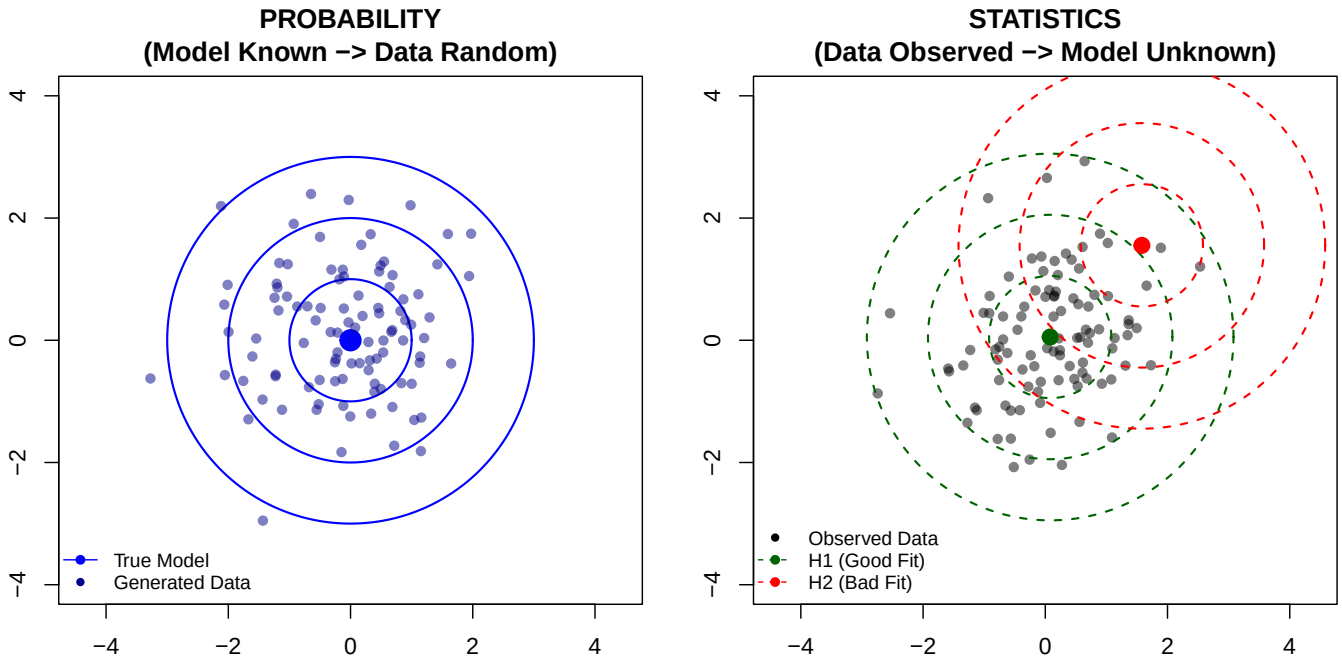


Figure 1.1: Probability vs Statistics. Left: Probability—The model is fixed (Blue center/contours), generating random data. Right: Statistics—Data is fixed (Black points); we test two hypothesized models: H1 (Green) centered at the sample mean (Good Fit) and H2 (Red) shifted by (1.5, 1.5) (Bad Fit).

1.3 A Motivating Example: The Lady Tasting Tea

To illustrate the concepts of statistical inference, we consider the famous experiment described by R.A. Fisher.

A lady claims she can distinguish whether milk was poured into the cup before or after the tea. To test this claim, we prepare n cups of tea.

- **Random Variable:** Let $X_i = 1$ if she identifies the cup correctly, and 0 otherwise.
- **Parameter:** Let θ be the probability that she correctly identifies a cup.

- **The Data:** Suppose we observe that she identifies **70%** of cups correctly ($\bar{x} = 0.7$), which is a summary of the observed vector of x_i, for example,

$$x = (0, 1, 1, 0, 1, 1, 0, 1, 1, 1) \tag{1.3}$$

1.3.1 Small Sample (n=10)

We observe **7 out of 10** correct ($k = 7$).

$$\bar{x} = 0.7 \tag{1.4}$$

1.3.2 Large Sample (n=40)

We observe **28 out of 40** correct ($k = 28$).

$$\bar{x} = 0.7 \tag{1.5}$$

1.4 Questions to Answer in Statistical Inference

Using this example, we identify the four main types of statistical inference.

Point Estimation

We want to use a single number to capture the parameter: $\hat{\theta} = \theta(X_1, \ldots, X_n)$.

- *Tea Example:* Our best guess for her success rate is $\hat{\theta} = 0.7$.

Hypothesis Testing

We want to test a theory about the parameter: H_0 vs H_1.

- *Tea Example:* Is she just guessing? We test $H_0 : \theta = 0.5$ vs $H_1 : \theta > 0.5$.

Model Assessment

We want to test a theory about the parameter: H_0 vs H_1.

- *Example:* Can we use a reduced model? What level of complexity of $f(x; \theta)$ is necessary?

Interval Estimation

We want to construct an interval likely to contain the parameter: $\theta \in (L, U)$.

- *Tea Example:* We might say her true skill θ is likely between 0.45 and 0.95.

Prediction

We want to predict a new observation Y_{n+1} given previous data.

- *Tea Example:* If we give her an $(n+1)$-th cup, what is the probability she identifies it correctly?

1.5 The Likelihood Function

The bridge between probability and statistics is the Likelihood Function.

Definition 1.2 (Likelihood Function)**.** Let $f(x_1, \ldots, x_n; \theta)$ be the joint probability density (or mass) function of the data given the parameter θ. When we view this function as a function of θ for fixed observed data $x_1, \ldots, x_n$, we call it the **likelihood function**, denoted $L(\theta)$.

$$L(\theta) = f(x_1, \ldots, x_n; \theta) \tag{1.6}$$

Example: Lady Tasting Tea

For our Tea Tasting data, the likelihood is proportional to the Binomial probability:

$$L(\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{1.7}$$

1.5.1 n=10 (k=7)

Here, $L(\theta) = \binom{10}{7}\theta^7(1-\theta)^3$.

θ	Calculation $\binom{10}{7}\theta^7(1-\theta)^3$	$L(\theta)$
0.0	$120 \times 0^7 \times 1^3$	0.0000
0.2	$120 \times 0.2^7 \times 0.8^3$	0.0008
0.4	$120 \times 0.4^7 \times 0.6^3$	0.0425
0.6	$120 \times 0.6^7 \times 0.4^3$	0.2150
0.7	$120 \times 0.7^7 \times 0.3^3$	**0.2668** (Max)
0.8	$120 \times 0.8^7 \times 0.2^3$	0.2013
1.0	$120 \times 1^7 \times 0^3$	0.0000

1.5.1.1 n=40 (k=28)

Here, $L(\theta) = \binom{40}{28}\theta^{28}(1-\theta)^{12}$. Notice how the likelihood becomes **narrower** (more peaked) with more data, even though the peak remains at 0.7.

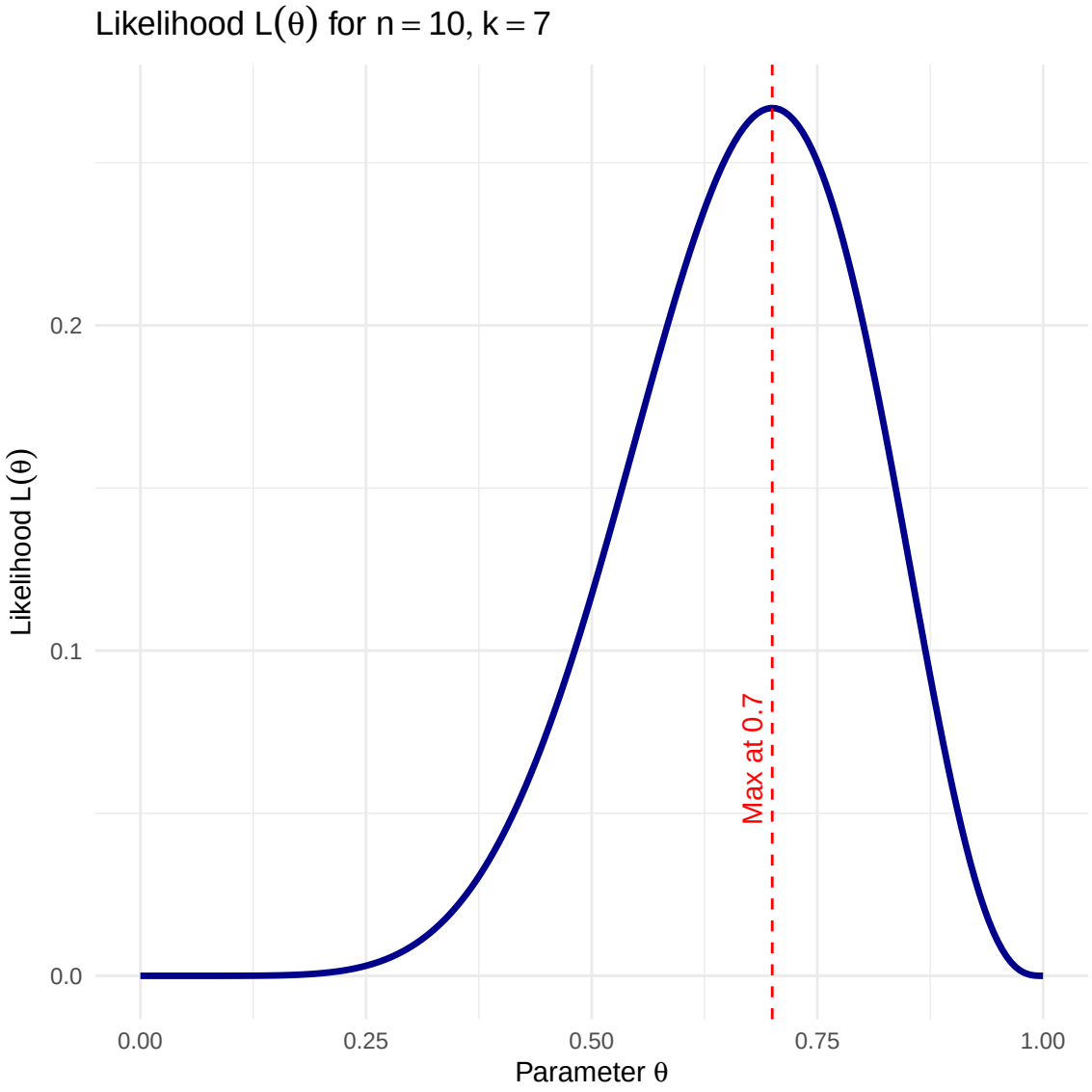


Figure 1.2: Likelihood Function (n= 10)

θ	Calculation $\binom{40}{28}\theta^{28}(1-\theta)^{12}$	$L(\theta)$
0.0	$5.5868535 \times 10^9 \times 0^{28} \times 1^{12}$	0.0000
0.2	$5.5868535 \times 10^9 \times 0.2^{28} \times 0.8^{12}$	0.0000
0.4	$5.5868535 \times 10^9 \times 0.4^{28} \times 0.6^{12}$	0.0001
0.6	$5.5868535 \times 10^9 \times 0.6^{28} \times 0.4^{12}$	0.0576
0.7	$5.5868535 \times 10^9 \times 0.7^{28} \times 0.3^{12}$	**0.1366** (Max)
0.8	$5.5868535 \times 10^9 \times 0.8^{28} \times 0.2^{12}$	0.0443
1.0	$5.5868535 \times 10^9 \times 1^{28} \times 0^{12}$	0.0000

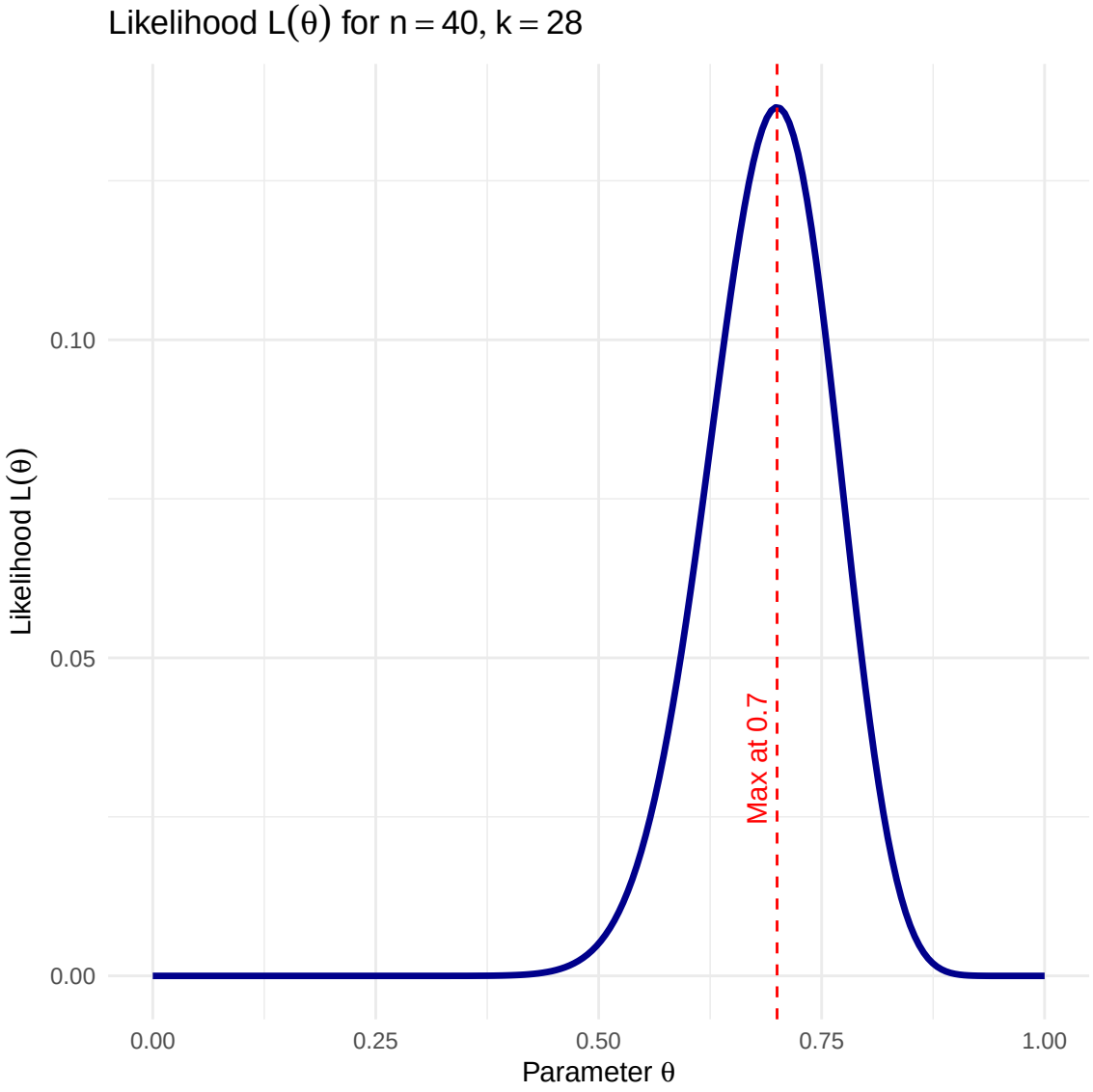


Figure 1.3: Likelihood Function (n= 40)

Questions

- Is an estimator like $\bar{x}$, which is called Maximum Likelihood Estimator (MLE), a good estimator in general?
- What do you discover from actually observing the two likelihood unctions of different sample size n?
- Is the likelihood function central to all inference problems?
- What are the essential 'parameters' of the likelihood function?

There are two primary frameworks for "How" to perform these inferences.

1.6 Frequentist Inference

- **Concept:** θ is unknown but fixed; Data X is random.
- **Sampling Distribution:** We analyze how $\hat{\theta}$ behaves under hypothetical repeated sampling.

Example: Frequentist Test of Lady Tasting Tea

We test $H_0 : \theta = 0.5$ (Guessing) vs $H_1 : \theta > 0.5$ (Skill). We analyze the behavior of $\bar{X}$ assuming H_0 is true. The rejection region (one-sided) is shaded red.

1.6.1 n=10 (k=7)

We calculate the P-value: Probability of observing ≥ 7 correct out of 10, assuming $\theta = 0.5$.

1.6.2 n=40 (k=28)

We calculate the P-value: Probability of observing ≥ 28 correct out of 40. With a larger sample size, the same proportion (0.7) provides **stronger evidence** against the null.

1.6.3 Questions to Answer

In this course, we will answer several challenging questions related to general parametric models in the Frequentist framework.

- **MLE:** Can we use the Maximum Likelihood Estimator (MLE) $\hat{\theta}$ for general models even no closed-form solution exists? Is MLE a good method?
- **Sampling Distributions:** What is the distribution of $\hat{\theta}_{\text{MLE}}$? What's its mean and standard deviation?
- **Confidence Intervals:** How to construct CI with $\hat{\theta}$?
- **Hypothesis Testing:** How do we derive powerful tests from the likelihood function? How to assess goodness-of-fit of parametric models with their likelhiood information?

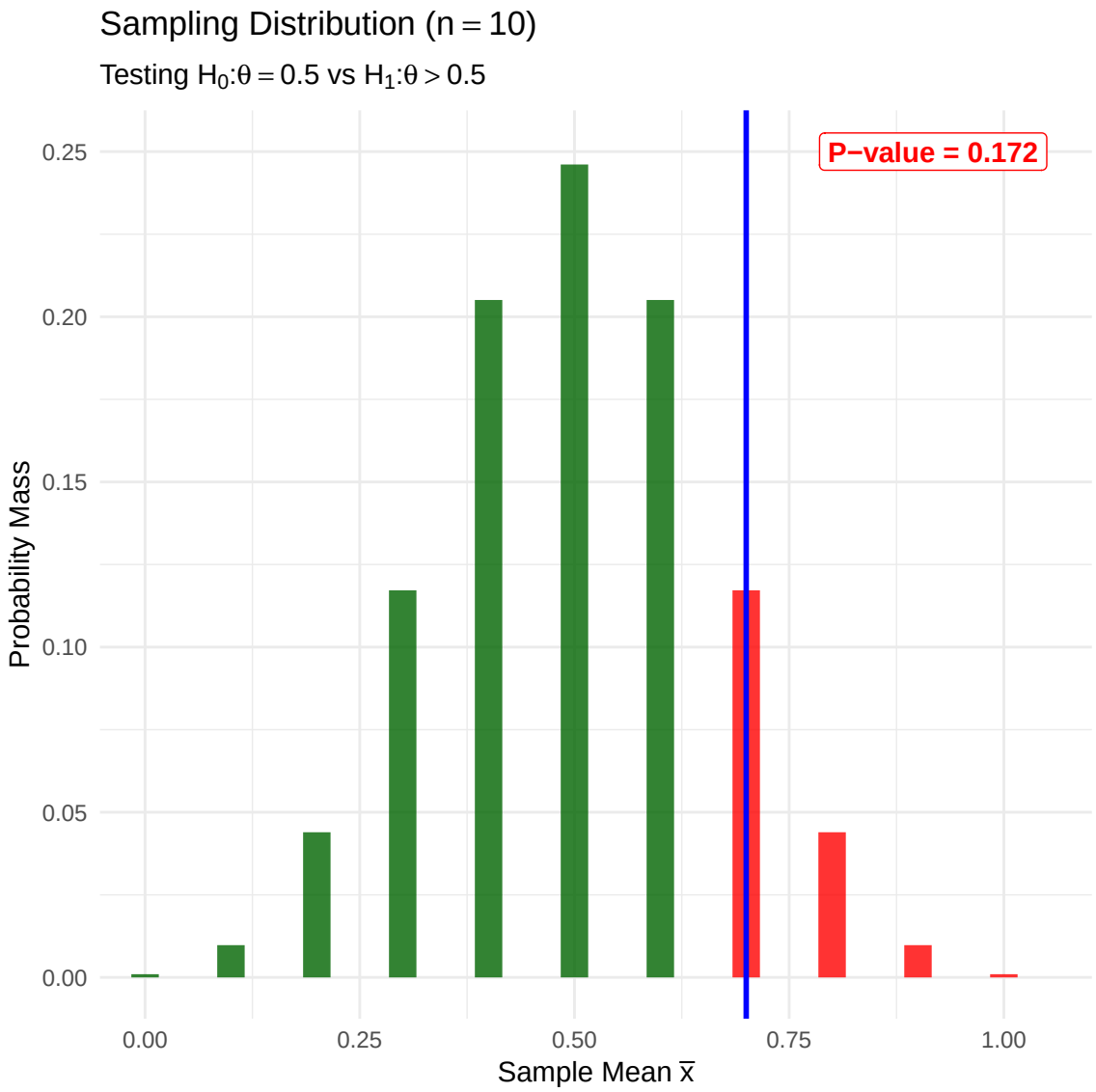


Figure 1.4: Sampling Distribution (n= 10)

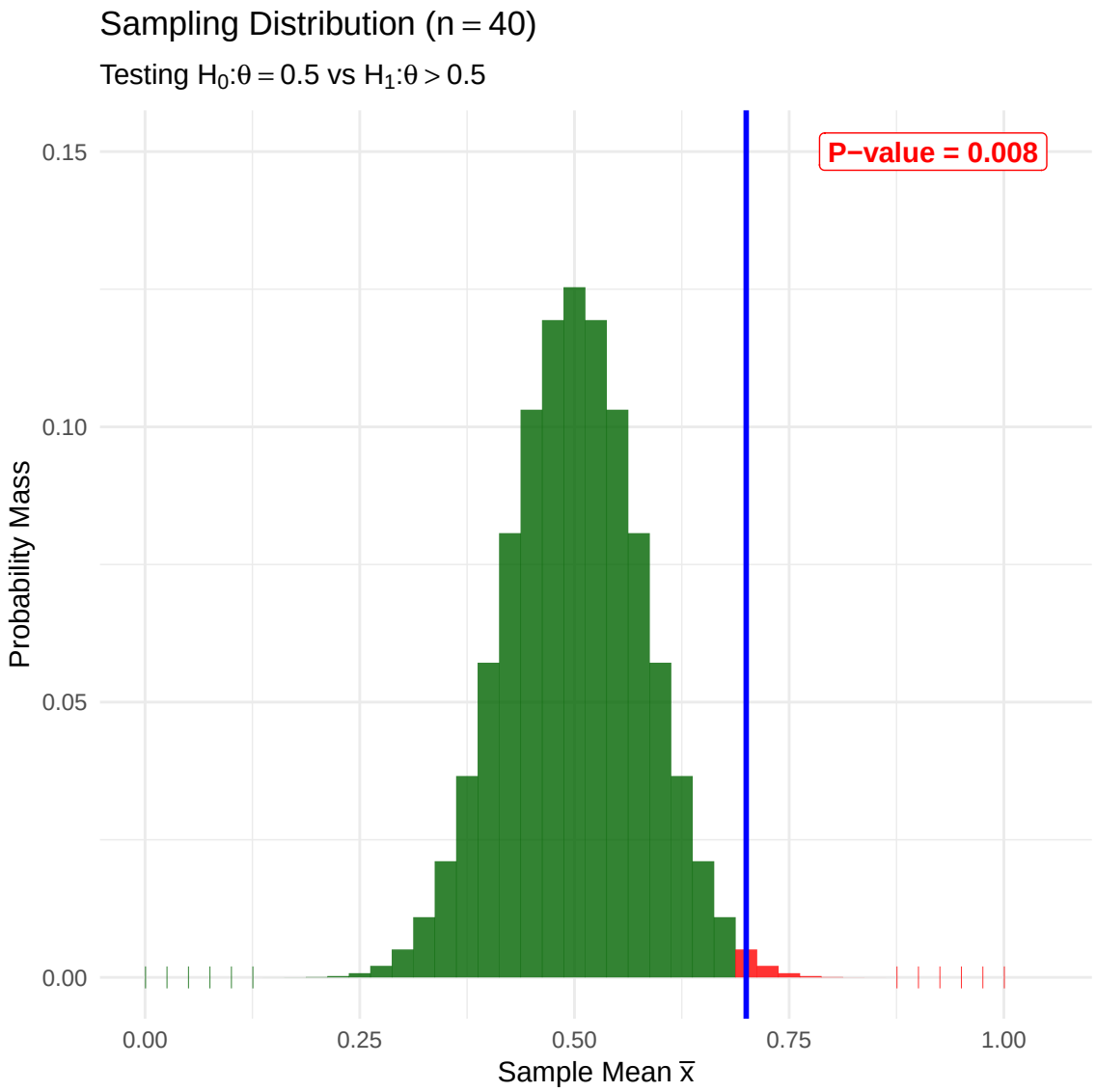


Figure 1.5: Sampling Distribution (n= 40)

1.7 Bayesian Inference

- **Concept:** θ is regarded as a random variable.
- **Posterior:** Posterior $\propto$ Likelihood $\times$ Prior.

Example: Bayesian Analysis of the Lady Tasting Tea

Prior: Beta$(1, 1)$ (Uniform).

1.7.1 n=10 (k=7)

Posterior: Beta$(1 + 7, 1 + 3) =$ Beta$(8, 4)$

1.7.2 n=40 (k=28)

Posterior: Beta$(1 + 28, 1 + 12) =$ Beta$(29, 13)$.

1.7.3 Questions to Answer

We will also tackle the specific technical challenges involved in Bayesian analysis.

- **Posterior Derivation:** How do we derive the posterior distribution $f(\theta|x)$ for various likelihoods and priors?
- **Comparing with Other methods:** Are Bayesain methods good or not or general inference?
- **Computation:** When the posterior cannot be derived analytically, how do we use computational techniques like Markov Chain Monte Carlo (MCMC) to sample from it?
- **Summarization:** How do we construct Credible Intervals (e.g., Highest Posterior Density regions) from posterior samples?
- **Prediction:** How do we solve the integral required to compute the posterior predictive distribution for future data?
- **Prior:** How to choose our prior? What's its effect on our inference?
- **Model Comparison and Assessment:** How to assess a Bayesian model?

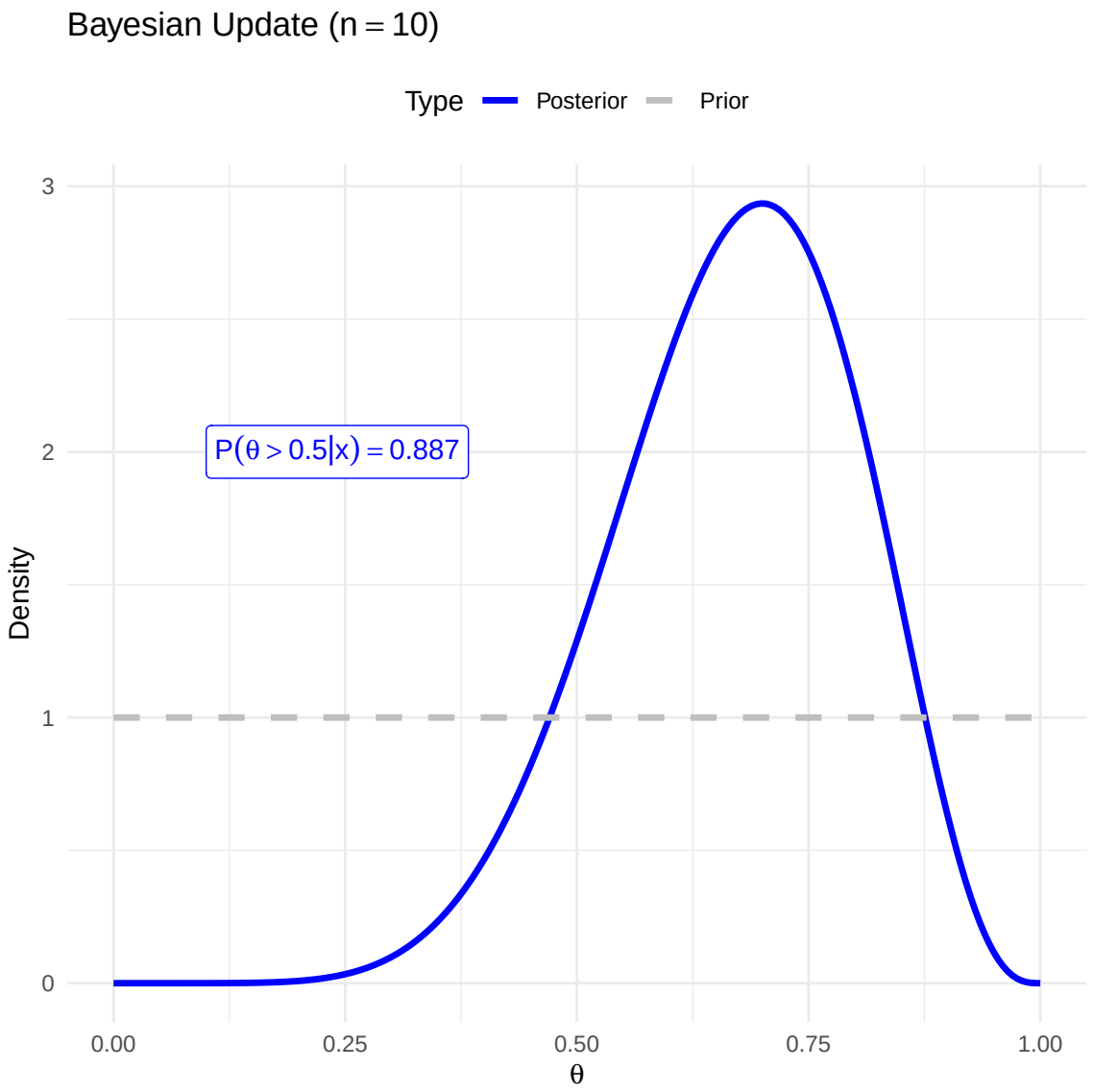


Figure 1.6: Bayesian Update (n= 10)

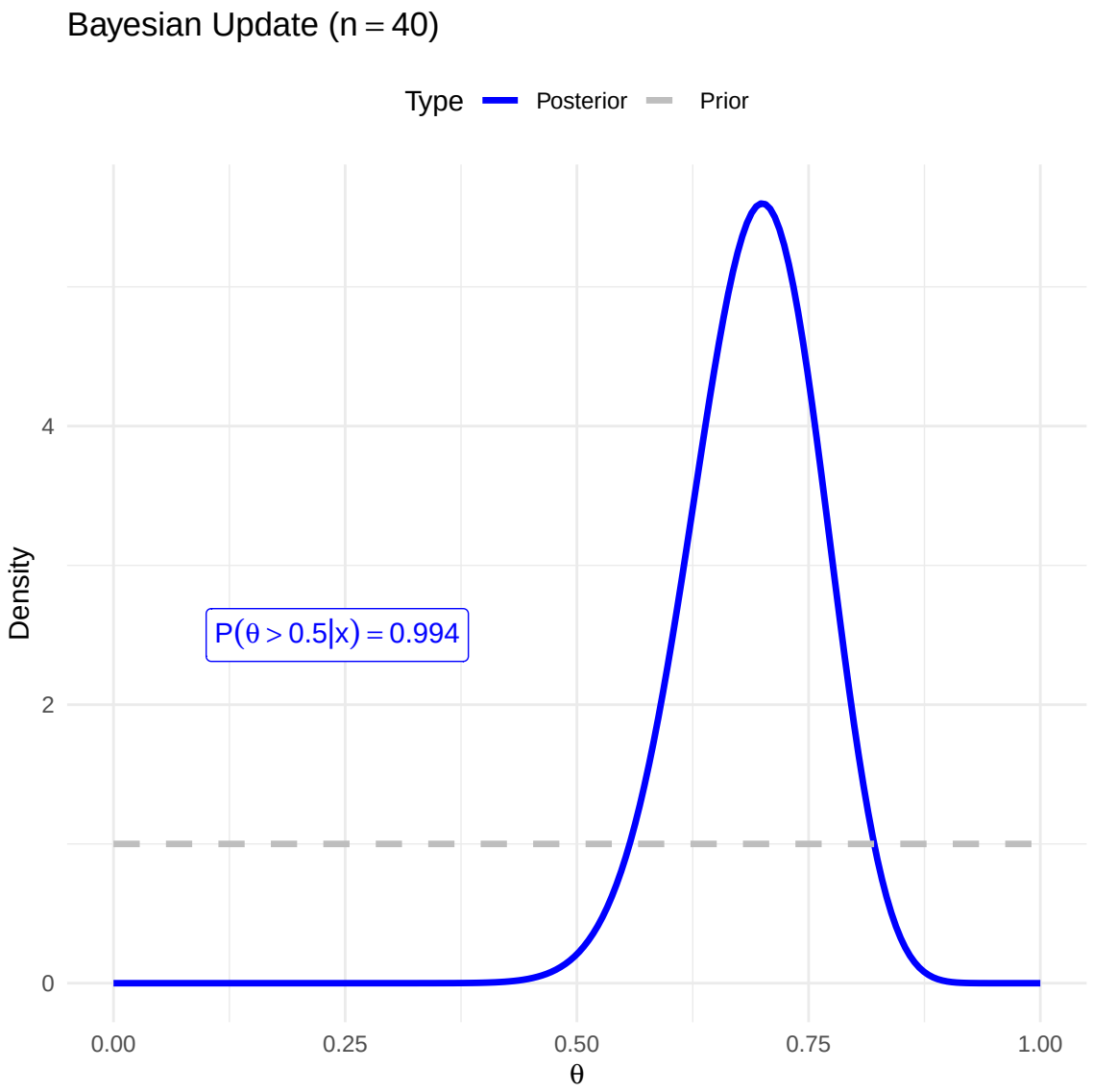


Figure 1.7: Bayesian Update (n= 40)

2 Decision Theory

2.1 Formulation of Decision Theory

In decision theory, we formalize the process of making decisions under uncertainty using the following components:

1. **Parameter Space (Θ):** The set of all possible states of nature or values that the parameter can take. $\theta \in \Theta$ (e.g., mean, variance).
2. **Sample Space ($\mathcal{X}$):** The space where the data X lies. Example: $X = (X_1, X_2, \dots, X_n)$ where $X_i \in \mathbb{R}$. So $\mathcal{X} \in \mathbb{R}^n$.
3. **Family of Probability Distributions:** $\{P_\theta(x) : \theta \in \Theta\}$. This describes how likely we are to see the data X given a specific parameter θ.
    - If X is continuous: $P_\theta(x) = f(x, \theta)$ (Probability Density Function).
    - If X is discrete: $P_\theta(x) = f(x, \theta)$ (Probability Mass Function).
4. **Action Space ($\mathcal{A}$):** The set of all actions or decisions available to the experimenter.
5. **Loss Function:** $L : \Theta \times \mathcal{A} \to \mathbb{R}$. $L(\theta, a)$ specifies the loss incurred if the true parameter is θ and we take action a. Generally, $L(\theta, a) \geq 0$.

2.2 Decision Rules and Risk Functions

2.2.1 Decision Rule

A decision rule is a function $d : \mathcal{X} \to \mathcal{A}$. It dictates the action $d(x)$ we take when we observe data x.

2.2.2 Risk Function

The risk function is the expected loss for a given decision rule d as a function of the parameter θ.

$$R(\theta, d) = E_\theta[L(\theta, d(X))] \tag{2.1}$$

2.3 Examples of Decision Problems

2.3.1 Example 1: Hypothesis Testing

We want to test H_0 vs H_1.

- **Action Space:** $\mathcal{A} = \{0, 1\}$ (0="Accept H_0", 1="Reject H_0").
- **Loss Function (0-1 Loss):** 0 if correct, 1 if wrong.
- **Risk Function:**
  - If $\theta \in H_0$: $R(\theta, d) = P(\text{Type I Error})$.
  - If $\theta \in H_1$: $R(\theta, d) = P(\text{Type II Error})$.

2.3.2 Example 2: Point Estimation

We want to estimate a parameter θ.

- **Action Space:** $\mathcal{A} = \Theta$.
- **Loss Function (Squared Error):** $L(\theta, a) = (\theta - a)^2$.
- **Risk Function (MSE):** $R(\theta, d) = \text{Var}(\bar{x}) + \text{Bias}^2$.

2.3.3 Example 3: Interval Estimation

We want to estimate a range for the parameter.

- **Action Space:** $\mathcal{A} = \{(l, u) : l \in \mathbb{R}, u \in \mathbb{R}, l \leq u\}$.

2.3.4 Example 4: The Duchess and the Emerald Necklace

Scenario: You are the Duchess of Omnium. You have two necklaces: a priceless **Real** one and a valueless **Imitation**. They are indistinguishable to you. One is in the **Left Drawer (Box 1)**, the other is in the **Right Drawer (Box 2)**.

The Data (Great Aunt): You consult your Great Aunt. She inspects the Left Drawer first, then the Right.

- If the **Real** necklace is in the **Left** ($\theta = 1$): She identifies it correctly. (Infallible).
- If the **Real** necklace is in the **Right** ($\theta = 2$): She sees the fake first, gets confused, and guesses randomly ($50/50$).

2.3.4.1 Formulation

1. **Parameter Space:** $\Theta = \{1, 2\}$ (1=Real Left, 2=Real Right).
2. **Action Space:** $\mathcal{A} = \{1, 2\}$ (1=Wear Left, 2=Wear Right).
3. **Loss Function:** 0 if correct, 1 if wrong.

2.3.4.2 Risk Calculation for Deterministic Rules

We consider four deterministic rules $d(X)$. We calculate the risk (R_1 for $\theta = 1$ and R_2 for $\theta = 2$) for each.

Rule d_1 (Always Left)

State	Component	$X = 1$	$X = 2$	Risk (Sum)
$\theta = 1$	Loss $L(1, d)$	0	0	
	Prob $P(X \mid \theta = 1)$	1	0	$R_1 = 0$
$\theta = 2$	Loss $L(2, d)$	1	1	
	Prob $P(X \mid \theta = 2)$	0.5	0.5	$R_2 = 1$

Rule d_2 (Always Right)

State	Component	$X = 1$	$X = 2$	Risk (Sum)
$\theta = 1$	Loss $L(1, d)$	1	1	
	Prob $P(X \mid \theta = 1)$	1	0	$R_1 = 1$
$\theta = 2$	Loss $L(2, d)$	0	0	
	Prob $P(X \mid \theta = 2)$	0.5	0.5	$R_2 = 0$

Rule d_3 (Follow Aunt)

State	Component	$X = 1$	$X = 2$	Risk (Sum)
$\theta = 1$	Loss $L(1, d)$	0	1	
	Prob $P(X \mid \theta = 1)$	1	0	$R_1 = 0$
$\theta = 2$	Loss $L(2, d)$	1	0	
	Prob $P(X \mid \theta = 2)$	0.5	0.5	$R_2 = 0.5$

Rule d_4 (Do Opposite)

State	Component	$X = 1$	$X = 2$	Risk (Sum)
$\theta = 1$	Loss $L(1, d)$	1	0	
	Prob $P(X \mid \theta = 1)$	1	0	$R_1 = 1$
$\theta = 2$	Loss $L(2, d)$	0	1	
	Prob $P(X \mid \theta = 2)$	0.5	0.5	$R_2 = 0.5$

2.4 Principles for Choosing a Decision Rule

Since no single rule minimizes risk for all θ, we rely on several principles to order and select decision rules.

2.4.1 Admissibility

A decision rule d is **admissible** if it is not "dominated" by any other rule.

- **Domination:** A rule d dominates d' if $R(\theta, d) \leq R(\theta, d')$ for all θ, with strict inequality for at least one θ.
- **Inadmissibility:** If a rule is dominated, it is inadmissible and can be discarded (we can do better or equal in every possible state).

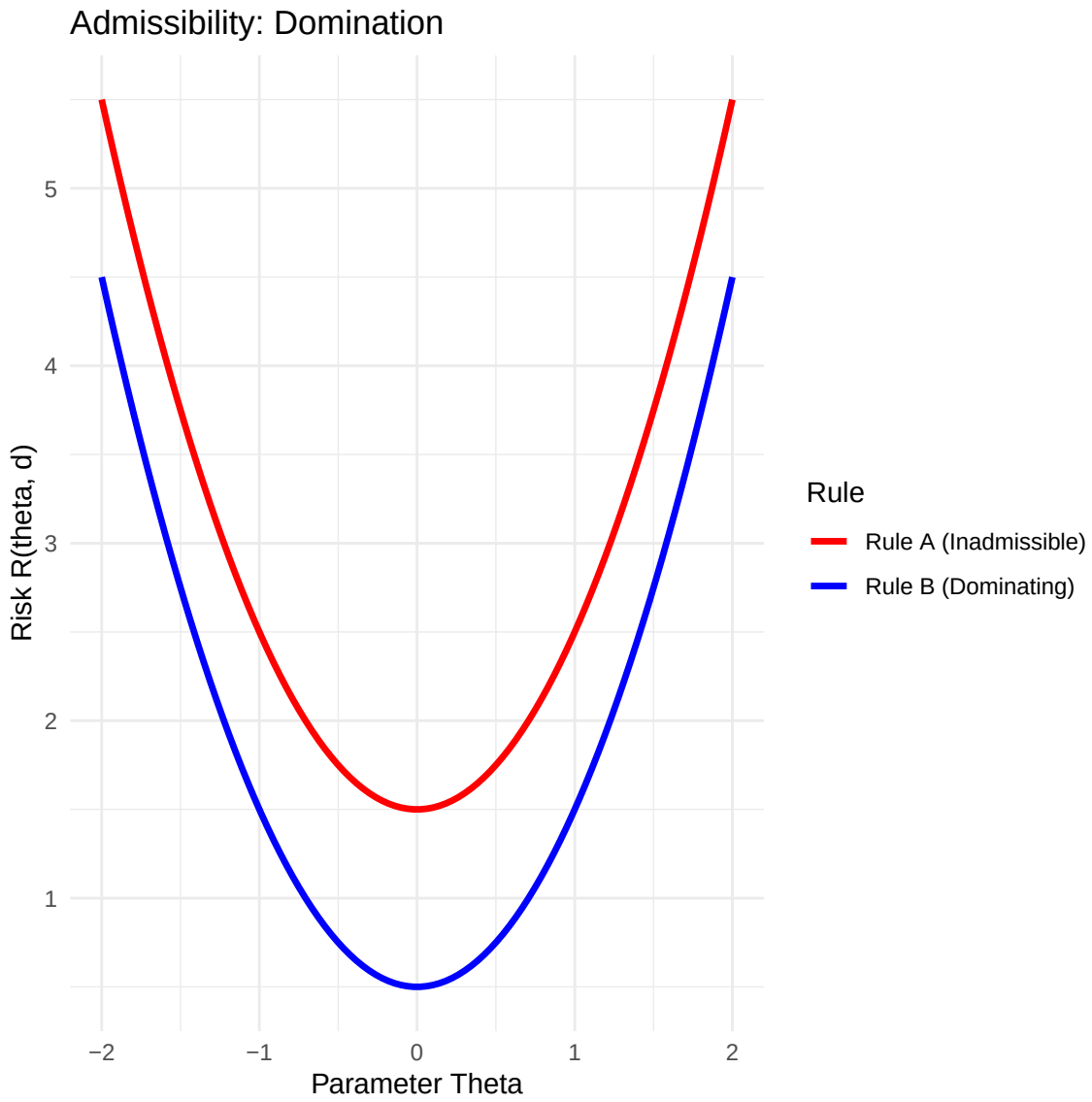


Figure 2.1: Illustration of Domination: Rule A (Red) is inadmissible because Rule B (Blue) has lower risk for all values of theta.

2.4.2 Minimax Principle

The Minimax principle is a conservative approach that guards against the worst-case scenario. It selects the rule that minimizes the maximum risk.

$$\min_{d} \left[\sup_{\theta} R(\theta, d) \right] \tag{2.2}$$

In the plot below, while Rule B has lower risk in the center, it has a very high maximum risk. Rule A is "flatter" and has a lower maximum value, making it the **Minimax** choice.

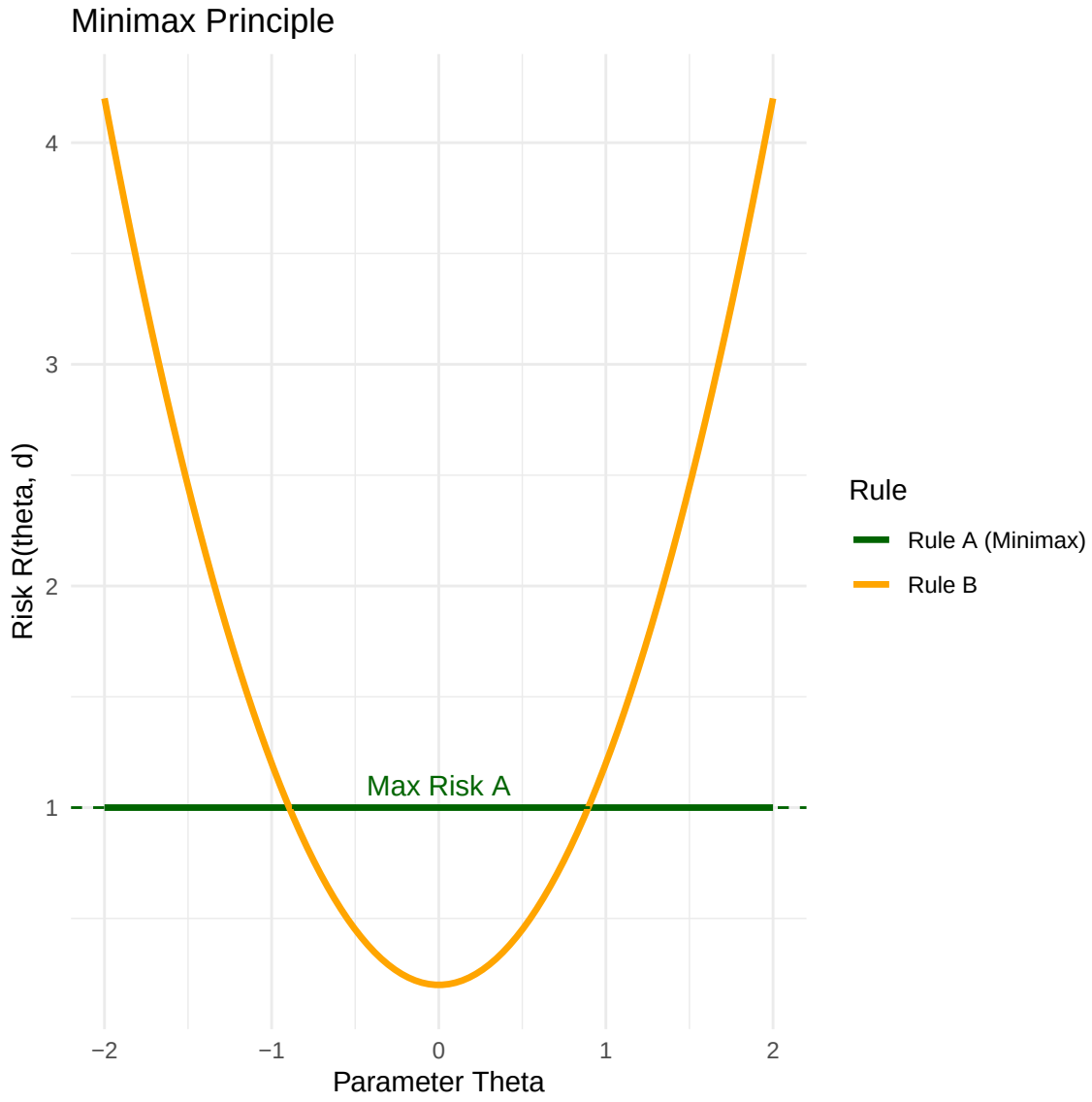


Figure 2.2: Illustration of Minimax: Rule A has a lower peak risk than Rule B, making Rule A the Minimax choice.

2.4.3 Bayes Decision Rules

The Bayes principle incorporates prior knowledge. If we assign a probability distribution (prior) $\pi(\theta)$ to the parameter, we can calculate the **Bayes Risk**, which is the weighted average of the risk function. We choose the rule that minimizes this average.

$$r(\pi, d) = E_\pi[R(\theta, d)] = \int_\Theta R(\theta, d)\pi(\theta)d\theta \tag{2.3}$$

2.5 Risk Set for Finite Parameter Space

For finite parameter spaces (e.g., $\Theta = \{1, 2\}$), we can visualize the problem in 2D space where the axes are $R_1 = R(\theta_1)$ and $R_2 = R(\theta_2)$.

2.5.1 The Risk Set (S)

The set of all possible risk vectors is called the Risk Set S.

- **Deterministic Rules:** These are the vertices of the set.
- **Randomized Rules:** By choosing rule d_i with probability p and d_j with probability $1 - p$, we can achieve any risk on the line segment connecting them.
- **Convexity:** The Risk Set is the **convex hull** of the deterministic rules.

2.5.2 Visualizing Admissibility

The admissible rules lie on the **lower-left boundary** of the set. Any point to the "north-east" of another point is dominated (inadmissible).

2.5.3 Visualizing Minimax

The Minimax rule is found by intersecting the Risk Set with the line $y = x$ ($R_1 = R_2$).

- We look for the point in S that touches the 45° line at the lowest value.
- If the set is entirely below the line, we minimize R_2. If entirely above, we minimize R_1.

2.5.4 Visualizing Bayes Rules

A Bayes rule minimizes $\pi_1 R_1 + \pi_2 R_2 = k$. This equation represents a line with slope $m = -\pi_1/\pi_2$.

- To find the Bayes rule, we find the **tangent line** to the Risk Set S with slope $-\pi_1/\pi_2$.

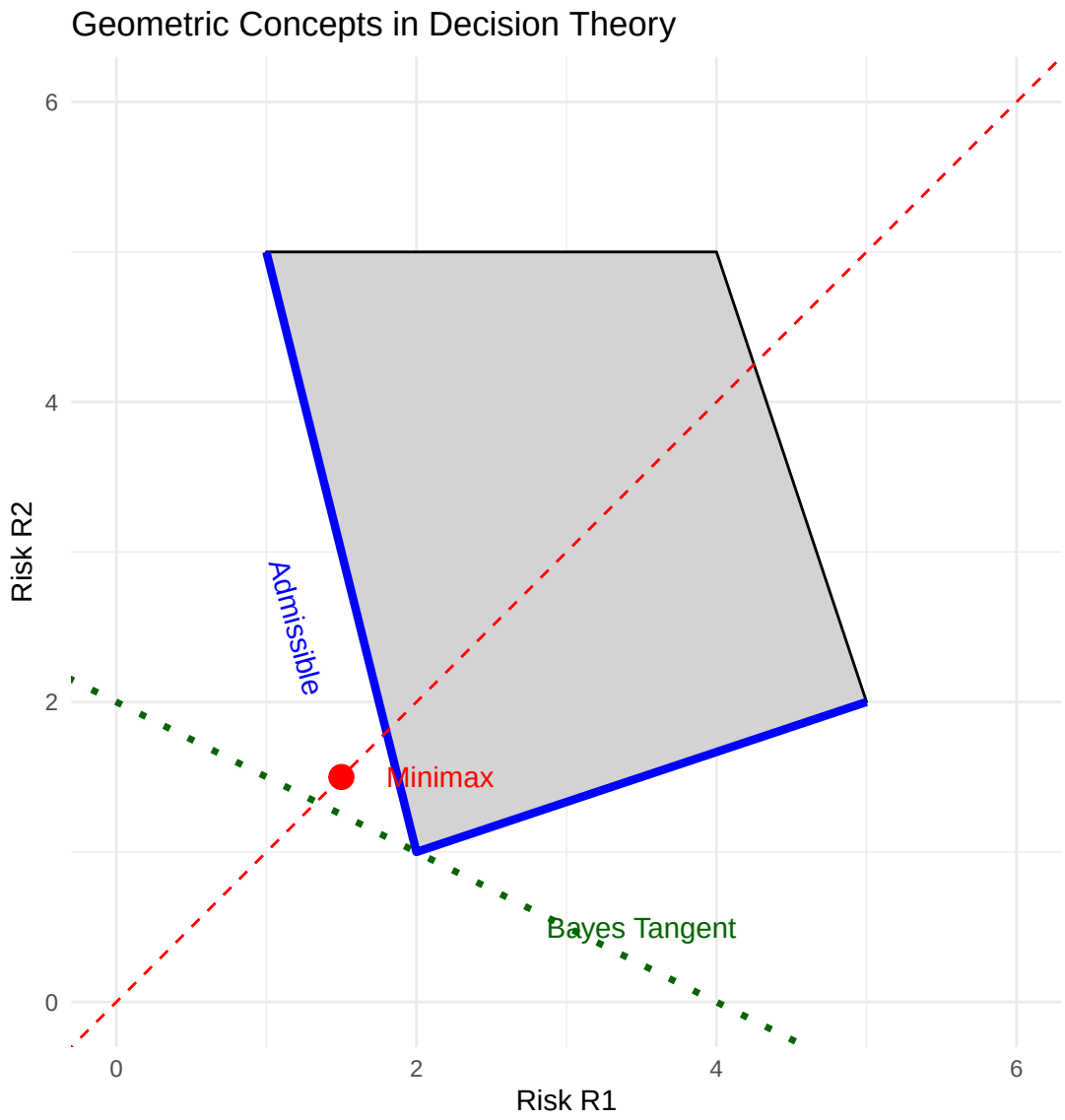


Figure 2.3: Geometric Interpretation: The gray polygon is the Risk Set S. The blue boundary represents admissible rules. The red point is the Minimax rule. The green line represents a Bayes rule for a specific prior.

2.6 Revisiting the Necklace Example: Geometric Solution

We now apply the geometric interpretation to the Necklace problem using the risks calculated in Section 2.3.4.

- d_1: $(0, 1)$
- d_2: $(1, 0)$
- d_3: $(0, 0.5)$
- d_4: $(1, 0.5)$

2.6.1 Analysis

1. **Admissibility:**
   - d_4 has risk $(1, 0.5)$. d_3 has risk $(0, 0.5)$. Since $0 < 1$, d_3 strictly dominates d_4. Thus d_4 is **inadmissible**.
   - The efficient frontier connects d_3 and d_2.
2. **Minimax Solution:** The Minimax rule lies on the segment connecting $d_3(0, 0.5)$ and $d_2(1, 0)$.
   - Let the randomized rule be $\delta^* = pd_3 + (1-p)d_2$.
   - $R(\delta^*) = p \begin{pmatrix} 0 \\ 0.5 \end{pmatrix} + (1-p) \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1-p \\ 0.5p \end{pmatrix}$.
   - Set $R_1 = R_2$: $1 - p = 0.5p \Rightarrow 1 = 1.5p \Rightarrow p = 2/3$.
   - **Result:** The Minimax rule is to choose d_3 with probability $2/3$ and d_2 with probability $1/3$.

2.7 Theorems Relating Minimax and Bayes Rules

In practice, finding a Minimax rule directly is mathematically difficult. A standard strategy is to "guess" a Least Favorable Prior π—defined as the prior distribution that maximizes the minimum Bayes risk (i.e., the prior against which it is hardest to defend)—find the corresponding Bayes rule, and then check if it satisfies specific conditions to confirm it is Minimax.

2.7.1 Equalizer Rules

Theorem 2.1 (The Equalizer Rule Strategy)**.** *If δ^* is a Bayes rule with respect to some prior π, and if δ^* is an **equalizer** rule (meaning $R(\theta, \delta^*) = C$ for some constant C for all $\theta \in \Theta$), then δ^* is Minimax.*

Proof.

1. **Bayes Risk Definition:** Since δ^* is an equalizer rule with risk C, its Bayes risk with respect to π is:

$$r(\pi, \delta^*) = \int_\Theta R(\theta, \delta^*)\pi(\theta)d\theta = \int_\Theta C\pi(\theta)d\theta = C \cdot 1 = C \tag{2.4}$$

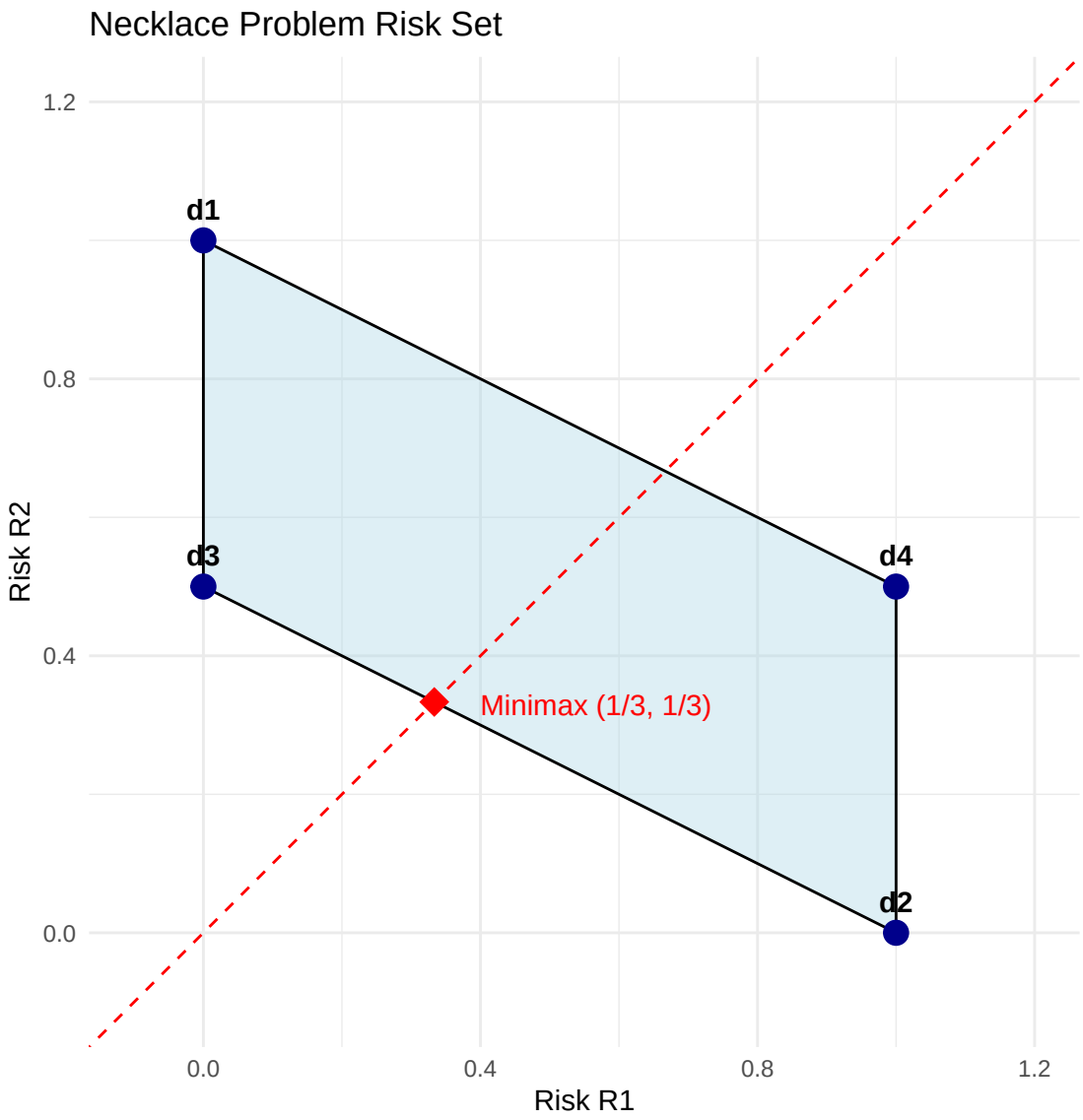


Figure 2.4: Necklace Problem Solution. The Minimax rule (red diamond) is the specific randomized combination of d3 and d2 that equalizes the risk.

2. **Minimax Contradiction:** Suppose, for the sake of contradiction, that δ^* is *not* Minimax. This implies there exists another rule δ' such that:
$$\sup_\theta R(\theta, \delta') < \sup_\theta R(\theta, \delta^*) \tag{2.5}$$
Since $R(\theta, \delta^*) = C$ for all θ, the supremum is C. Thus:
$$\sup_\theta R(\theta, \delta') < C \tag{2.6}$$

3. **Inequality:** This implies that for all θ, $R(\theta, \delta') < C$.

4. **Bayes Risk Comparison:** Now, consider the Bayes risk of this alternative rule δ':
$$r(\pi, \delta') = \int_\Theta R(\theta, \delta')\pi(\theta)d\theta \tag{2.7}$$
Since $R(\theta, \delta') < C$ for all θ, it follows that:
$$r(\pi, \delta') < \int_\Theta C\pi(\theta)d\theta = C \tag{2.8}$$

5. **Conclusion:** We have established that $r(\pi, \delta') < C$. However, we established in step 1 that $r(\pi, \delta^*) = C$. This yields $r(\pi, \delta') < r(\pi, \delta^*)$. This contradicts the assumption that δ^* is a Bayes rule (since a Bayes rule must minimize the Bayes risk). Therefore, no such δ' exists, and δ^* is Minimax. ■

□

2.7.2 Limits of Bayes Rules

Sometimes the Minimax rule corresponds to an "improper" prior (a prior that does not integrate to 1, like a uniform distribution on the real line). We approach these via a limiting sequence.

Theorem 2.2 (Limits of Bayes Rules)**.** *Let $\{\delta_n\}$ be a sequence of Bayes rules with respect to priors $\{\pi_n\}$. Let $r(\pi_n, \delta_n)$ be the associated Bayes risks. If there exists a rule δ_0 such that:*
$$\sup_\theta R(\theta, \delta_0) \leq \lim_{n\to\infty} r(\pi_n, \delta_n) \tag{2.9}$$
Then δ_0 is Minimax.

Proof.

1. **Define Limit:** Let $V = \lim_{n\to\infty} r(\pi_n, \delta_n)$. We are given that $\sup_\theta R(\theta, \delta_0) \leq V$.

2. **Contradiction Setup:** Suppose δ_0 is *not* Minimax. Then there exists a rule δ^* such that:
$$\sup_\theta R(\theta, \delta^*) < \sup_\theta R(\theta, \delta_0) \leq V \tag{2.10}$$
Let $\sup_\theta R(\theta, \delta^*) = V - \epsilon$ for some $\epsilon > 0$.

3. **Bayes Risk Bound:** For any prior π_n, the Bayes risk of δ^* cannot exceed its maximum risk:

$$r(\pi_n, \delta^*) = \int R(\theta, \delta^*)\pi_n(\theta)d\theta \leq \int (V - \epsilon)\pi_n(\theta)d\theta = V - \epsilon \tag{2.11}$$

4. **Optimality of δ_n:** Since δ_n is the Bayes rule for π_n, it minimizes Bayes risk. Thus:

$$r(\pi_n, \delta_n) \leq r(\pi_n, \delta^*) \tag{2.12}$$

5. **Combining Inequalities:** Combining steps 3 and 4:

$$r(\pi_n, \delta_n) \leq V - \epsilon \tag{2.13}$$

6. **Taking Limits:** Taking the limit as $n \to \infty$:

$$\lim_{n\to\infty} r(\pi_n, \delta_n) \leq V - \epsilon \tag{2.14}$$

$$V \leq V - \epsilon \tag{2.15}$$

This is a contradiction since $\epsilon > 0$. Therefore, δ_0 must be Minimax. ■

□

2.7.3 Admissibility of Bayes Rules

Bayes rules are generally good candidates for admissibility. If a rule is Bayes, it is likely efficient, provided the prior doesn't ignore parts of the parameter space.

Theorem 2.3 (Admissibility of Bayes Rules (Finite Support))**.** *If the parameter space Θ is finite (or countable) and the prior π assigns positive probability to every $\theta \in \Theta$ (i.e., $\pi(\theta) > 0$ for all θ), then any Bayes rule δ_π is admissible.*

Proof.

1. **Contradiction Setup:** Suppose δ_π is inadmissible. Then there exists a rule δ' that dominates it. By definition of domination:
   - $R(\theta, \delta') \leq R(\theta, \delta_\pi)$ for all θ.
   - $R(\theta_k, \delta') < R(\theta_k, \delta_\pi)$ for at least one θ_k.
2. **Bayes Risk Difference:** Consider the difference in Bayes risk:

$$r(\pi, \delta_\pi) - r(\pi, \delta') = \sum_{\theta\in\Theta} \pi(\theta)[R(\theta, \delta_\pi) - R(\theta, \delta')] \tag{2.16}$$

3. **Strict Positivity:**
   - Since δ' dominates δ_π, each term $[R(\theta, \delta_\pi) - R(\theta, \delta')]$ is non-negative (≥ 0).

- At θ_k, the term is strictly positive (> 0).
- We assumed the prior has full support, so $\pi(\theta) > 0$ for all θ.

4. **Summation:** A sum of non-negative terms where at least one term is strictly positive must be strictly positive.

$$r(\pi, \delta_\pi) - r(\pi, \delta') > 0 \implies r(\pi, \delta') < r(\pi, \delta_\pi) \tag{2.17}$$

5. **Conclusion:** This contradicts the definition that δ_π is a Bayes rule (which must minimize Bayes risk). Therefore, δ_π is admissible. ■

□

2.7.4 Admissibility of Unique Bayes Rules

If the Bayes rule is unique, we can drop the requirement that the parameter space be discrete or finite.

Theorem 2.4 (Admissibility of Unique Bayes Rules)**.** *Let δ_π be a Bayes rule with respect to π. If δ_π is the **unique** Bayes rule (up to risk equivalence), then δ_π is admissible.*

Proof.

1. **Contradiction Setup:** Suppose δ_π is inadmissible. Then there exists a rule δ' such that: $R(\theta, \delta') \leq R(\theta, \delta_\pi)$ for all θ, with strict inequality for some set of θ.
2. **Bayes Risk Inequality:** Taking the expectation with respect to π:

$$r(\pi, \delta') = \int R(\theta, \delta')\pi(\theta)d\theta \leq \int R(\theta, \delta_\pi)\pi(\theta)d\theta = r(\pi, \delta_\pi) \tag{2.18}$$

3. **Minimality:** Since δ_π is Bayes, it minimizes the risk, so $r(\pi, \delta_\pi) \leq r(\pi, \delta')$. Combining these gives $r(\pi, \delta') = r(\pi, \delta_\pi)$.
4. **Uniqueness:** This implies that δ' is also a Bayes rule. However, we assumed that δ_π is the **unique** Bayes rule. Therefore, δ' must be equal to δ_π (in terms of risk functions).
5. **Conclusion:** If δ' and δ_π have identical risk functions, then δ' cannot strictly dominate δ_π. This contradicts the assumption of inadmissibility. Thus, δ_π is admissible. ■

□

3 Bayesian Methods

3.1 Fundamental Elements of Bayesian Inference

The foundation of Bayesian inference relies on the relationship between the prior distribution, the likelihood of the data, and the posterior distribution. This relationship is governed by Bayes' Theorem (or Law).

Definition 3.1 (Posterior Distribution)**.** Suppose we have a parameter θ with a prior distribution denoted by $\pi(\theta)$. If we observe data x drawn from a distribution with probability density function (pdf) $f(x;\theta)$, then the **posterior density** of θ given the data x is defined as:

$$\pi(\theta|x) = \frac{\pi(\theta)f(x;\theta)}{m(x)} \tag{3.1}$$

where $m(x)$ is the **marginal distribution** (or marginal likelihood) of the data, calculated as:

$$m(x) = \int_{\Theta} \pi(\theta)f(x;\theta)d\theta \tag{3.2}$$

In this context, $m(x)$ acts as a normalizing constant. Since it depends only on the data x and not on the parameter θ, it ensures that the posterior density integrates to 1 but does not influence the **shape** of the posterior distribution.

Thus, we often state the proportional relationship:

$$\pi(\theta|x) \propto \pi(\theta)f(x;\theta) \tag{3.3}$$

Example 3.1 (Binomial-beta Conjugacy)**.** Consider an experiment where $x|\theta \sim \text{Bin}(n,\theta)$. The likelihood function is:

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \tag{3.4}$$

Suppose we choose a Beta distribution as the prior for θ, such that $\theta \sim \text{Beta}(a,b)$. The prior density is:

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \tag{3.5}$$

where $B(a,b)$ is the Beta function defined as $\int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta$.

To find the posterior, we multiply the prior and the likelihood:

$$\pi(\theta|x) \propto \theta^{a-1}(1-\theta)^{b-1} \cdot \theta^{x}(1-\theta)^{n-x} \tag{3.6}$$

Combining terms with the same base:

$$\pi(\theta|x) \propto \theta^{a+x-1}(1-\theta)^{b+n-x-1} \tag{3.7}$$

We can recognize this kernel as a Beta distribution. Therefore, we conclude that the posterior distribution is:

$$\theta|x \sim \text{Beta}(a+x, b+n-x) \tag{3.8}$$

Properties of the Posterior:

- The posterior mean is:

$$E(\theta|x) = \frac{a+x}{a+b+n} \tag{3.9}$$

  As $n \to \infty$, this approximates the maximum likelihood estimate $\frac{x}{n}$.

- The posterior variance is:

$$\text{Var}(\theta|x) = \frac{(a+x)(n+b-x)}{(a+b+n)^2(a+b+n+1)} \tag{3.10}$$

  For large n, this approximates $\frac{x(n-x)}{n^3} = \frac{\hat{p}(1-\hat{p})}{n}$.

Numerical Illustration:

Suppose we are estimating a probability θ.

- **Prior:** $\theta \sim \text{Beta}(2, 2)$ (Mean = 0.5).
- **Data:** 10 trials, 8 successes ($n = 10, x = 8$).
- **Posterior:** $\theta|x \sim \text{Beta}(2+8, 2+2) = \text{Beta}(10, 4)$ (Mean ≈ 0.71).

The plot below shows the prior (dashed) and posterior (solid) densities.

```r
theta <- seq(0, 1, length.out = 200)

# Prior: Beta(2, 2)
prior <- dbeta(theta, shape1 = 2, shape2 = 2)

# Posterior: Beta(10, 4)
posterior <- dbeta(theta, shape1 = 10, shape2 = 4)

plot(theta, posterior, type = 'l', lwd = 2, col = "blue",
     xlab = expression(theta), ylab = "Density",
     main = "Beta Prior vs Posterior", ylim = c(0, max(c(prior, posterior))))
lines(theta, prior, col = "red", lty = 2, lwd = 2)
```

```r
legend("topleft", legend = c("Prior Beta(2,2)", "Posterior Beta(10,4)"),
       col = c("red", "blue"), lty = c(2, 1), lwd = 2)
```

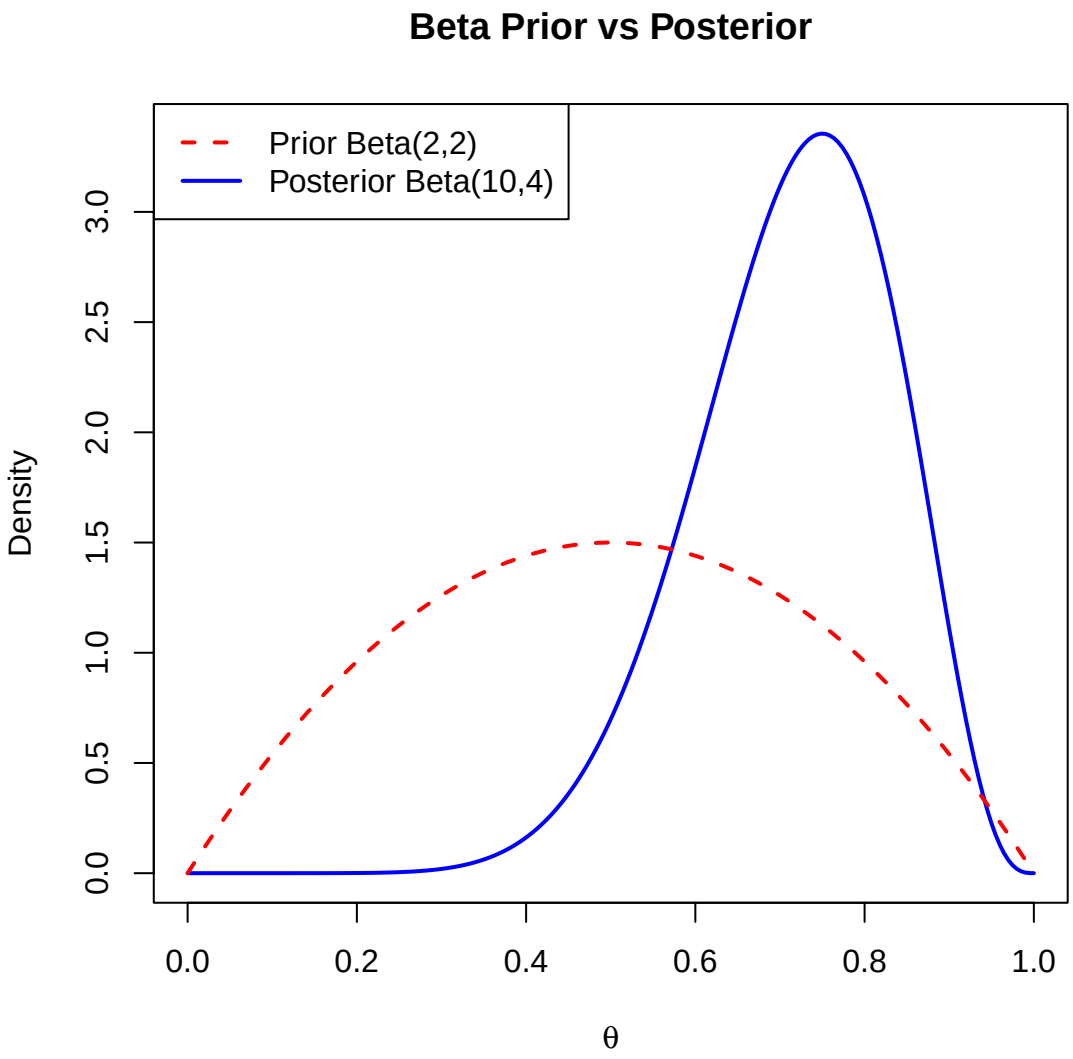


Figure 3.1: Prior vs Posterior for Beta-Binomial Example

Example 3.2 (Normal-normal Conjugacy (known Variance))**.** Let $X_1, X_2, \dots, X_n$ be independent and identically distributed (i.i.d.) variables such that $X_i \sim N(\mu, \sigma^2)$, where σ^2 is known.

We assign a Normal prior to the mean μ: $\mu \sim N(\mu_0, \sigma_0^2)$.

To find the posterior $\pi(\mu|x_1, \dots, x_n)$, let $x = (x_1, \dots, x_n)$. The posterior is proportional to:

$$\pi(\mu|x) \propto \pi(\mu) \cdot f(x|\mu) \tag{3.11}$$

$$\propto \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \cdot \exp\left\{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \tag{3.12}$$

Posterior Precision:

It is often more convenient to work with **precision** (the inverse of variance). Let:

- $\tau_0 = 1/\sigma_0^2$ (Prior precision)
- $\tau = 1/\sigma^2$ (Data precision)
- $\tau_1 = 1/\sigma_1^2$ (Posterior precision)

The relationship is additive:

$$\tau_1 = \tau_0 + n\tau \tag{3.13}$$

$$\text{Posterior Precision} = \text{Prior Precision} + \text{Precision of Data} \tag{3.14}$$

The posterior mean μ_1 is a weighted average of the prior mean and the sample mean:

$$\mu_1 = \frac{\mu_0\tau_0 + n\bar{x}\tau}{\tau_0 + n\tau} \tag{3.15}$$

So, the posterior distribution is:

$$\mu|x_1, \ldots, x_n \sim N\left(\frac{\mu_0\tau_0 + n\bar{x}\tau}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right) \tag{3.16}$$

Numerical Illustration:

Suppose we estimate a mean height μ.

- **Known Variance:** $\sigma^2 = 100$ ($\tau = 0.01$).
- **Prior:** $\mu \sim N(175, 25)$ (Precision $\tau_0 = 0.04$).
- **Data:** $n = 10, \bar{x} = 180$. (Total data precision $n\tau = 0.1$).
- **Posterior:**
  - Precision $\tau_1 = 0.04 + 0.1 = 0.14$.
  - Variance $\sigma_1^2 \approx 7.14$.
  - Mean $\mu_1 = \frac{175(0.04)+180(0.1)}{0.14} \approx 178.6$.

The plot below illustrates the prior (dashed) and posterior (solid) normal densities.

```r
mu_vals <- seq(150, 200, length.out = 200)

# Prior: N(175, 25) -> SD = 5
prior_norm <- dnorm(mu_vals, mean = 175, sd = 5)

# Posterior: N(178.6, 7.14) -> SD = Sqrt(7.14) Approx 2.67
posterior_norm <- dnorm(mu_vals, mean = 178.6, sd = sqrt(7.14))

plot(mu_vals, posterior_norm, type = 'l', lwd = 2, col = "blue",
     xlab = expression(mu), ylab = "Density",
     main = "Normal Prior vs Posterior",
     ylim = c(0, max(c(prior_norm, posterior_norm))))
lines(mu_vals, prior_norm, col = "red", lty = 2, lwd = 2)
legend("topleft", legend = c("Prior N(175, 25)", "Posterior N(178.6, 7.14)"),
       col = c("red", "blue"), lty = c(2, 1), lwd = 2)
```

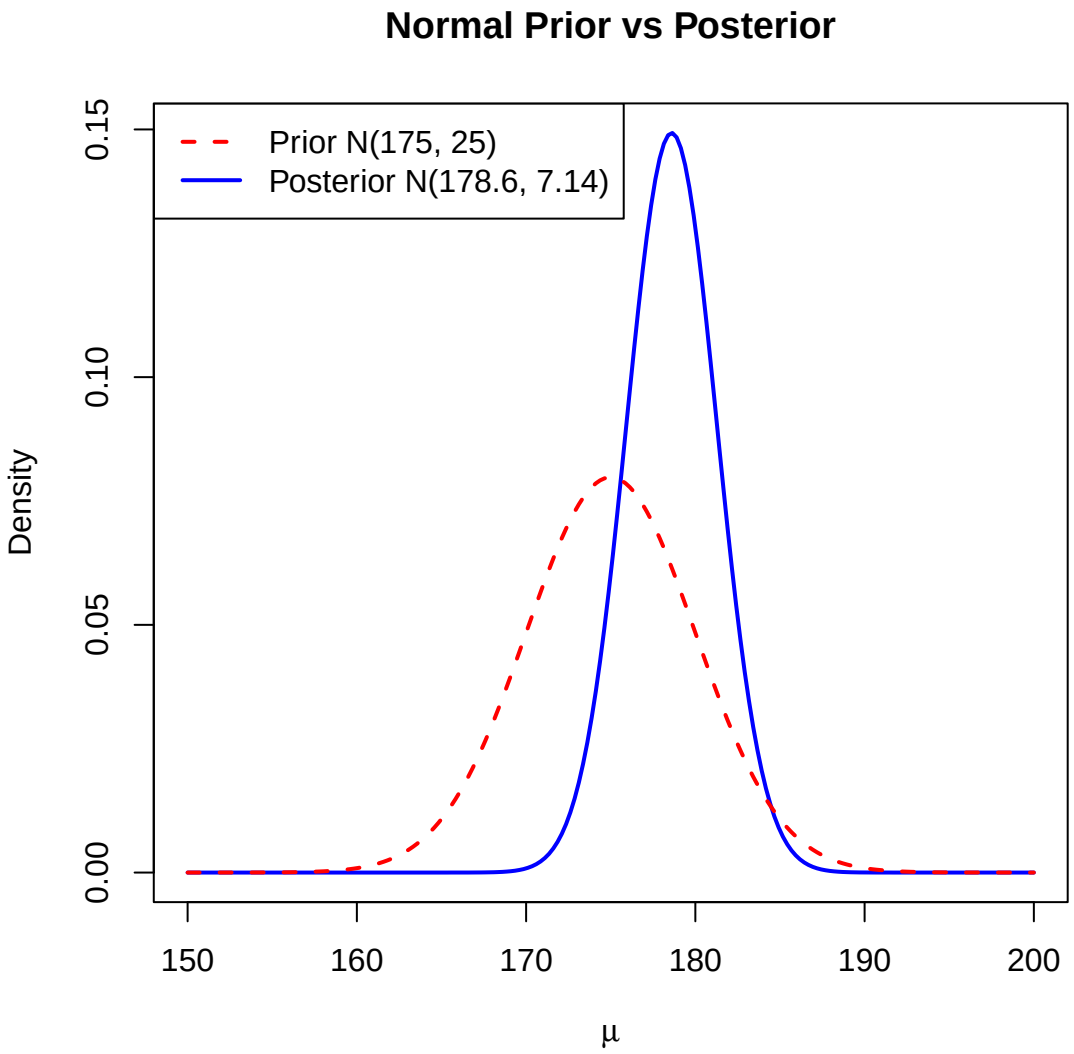


Figure 3.2: Prior vs Posterior for Normal-Normal Example

Example 3.3 (Discrete Posterior Calculation)**.** Consider the following table where we calculate the posterior probabilities for a discrete parameter space.

Let the parameter θ take values $\{1, 2, 3\}$ with prior probabilities $\pi(\theta)$. Let the data x take values $\{0, 1, 2, \dots\}$.

Given:

- Prior $\pi(\theta)$: $\pi(1) = 1/3, \pi(2) = 1/3, \pi(3) = 1/3$.
- Likelihood $\pi(x|\theta)$:
  - If $\theta = 1$, $x \sim$ Uniform on $\{0, 1\}$ (Prob = 1/2).
  - If $\theta = 2$, $x \sim$ Uniform on $\{0, 1, 2\}$ (Prob = 1/3).
  - If $\theta = 3$, $x \sim$ Uniform on $\{0, 1, 2, 3\}$ (Prob = 1/4).

Suppose we observe $x = 2$. The calculation of the posterior probabilities is summarized in the table below:

	$\theta = 1$	$\theta = 2$	$\theta = 3$	Sum
Prior $\pi(\theta)$	$1/3$	$1/3$	$1/3$	1
Likelihood $\pi(x = 2\|\theta)$	0	$1/3$	$1/4$	-
Product $\pi(\theta)\pi(x\|\theta)$	0	$1/9$	$1/12$	$7/36$
Posterior $\pi(\theta\|x)$	0	$4/7$	$3/7$	1

The marginal sum (evidence) is calculated as $0 + 1/9 + 1/12 = 4/36 + 3/36 = 7/36$. The posterior values are obtained by dividing the product row by this sum.

Example 3.4 (Normal with Unknown Mean and Variance)**.** Consider $X_1, \ldots, X_n \sim N(\mu, 1/\tau)$, where both μ and the precision τ are unknown.

We use a **Normal-Gamma** conjugate prior:

1. $\tau \sim \text{Gamma}(\alpha, \beta)$
$$\pi(\tau) \propto \tau^{\alpha-1} e^{-\beta\tau} \tag{3.17}$$
2. $\mu|\tau \sim N(\nu, 1/(k\tau))$
$$\pi(\mu|\tau) \propto \tau^{1/2} e^{-\frac{k\tau}{2}(\mu-\nu)^2} \tag{3.18}$$

The joint prior is the product of the conditional and the marginal:

$$\pi(\mu, \tau) \propto \tau^{\alpha-1/2} \exp\left\{-\tau\left(\beta + \frac{k}{2}(\mu-\nu)^2\right)\right\} \tag{3.19}$$

Derivation of the Posterior:

First, we write the likelihood in terms of the sufficient statistics $\bar{x}$ and $S_{xx} = \sum(x_i - \bar{x})^2$:

$$L(\mu, \tau|x) \propto \tau^{n/2} \exp\left\{-\frac{\tau}{2}\left[S_{xx} + n(\bar{x}-\mu)^2\right]\right\} \tag{3.20}$$

Multiplying the prior by the likelihood gives the joint posterior:

$$\begin{aligned}\pi(\mu, \tau|x) &\propto \tau^{\alpha-1/2} e^{-\beta\tau} e^{-\frac{k\tau}{2}(\mu-\nu)^2} \cdot \tau^{n/2} e^{-\frac{\tau}{2}S_{xx}} e^{-\frac{n\tau}{2}(\mu-\bar{x})^2} \\ &\propto \tau^{\alpha+n/2-1/2} \exp\left\{-\tau\left[\beta + \frac{S_{xx}}{2} + \frac{1}{2}\left(k(\mu-\nu)^2 + n(\mu-\bar{x})^2\right)\right]\right\}\end{aligned} \tag{3.21}$$

Next, we complete the square for the terms involving μ inside the brackets. It can be shown that:

$$k(\mu-\nu)^2 + n(\mu-\bar{x})^2 = (k+n)\left(\mu - \frac{k\nu + n\bar{x}}{k+n}\right)^2 + \frac{nk}{n+k}(\bar{x}-\nu)^2 \tag{3.22}$$

Substituting this back into the joint density and grouping terms that do not depend on μ:

$$\pi(\mu,\tau|x) \propto \underbrace{\tau^{\alpha+n/2-1}\exp\left\{-\tau\left[\beta+\frac{S_{xx}}{2}+\frac{nk}{2(n+k)}(\bar{x}-\nu)^2\right]\right\}}_{\text{Marginal of }\tau}\cdot\underbrace{\tau^{1/2}\exp\left\{-\frac{(k+n)\tau}{2}\left(\mu-\frac{k\nu+n\bar{x}}{k+n}\right)^2\right\}}_{\text{Conditional of }\mu|\tau} \tag{3.23}$$

Results:

By inspecting the factored equation above, we identify the updated parameters:

- **Marginal Posterior of τ:** The first part corresponds to a Gamma kernel $\tau^{\alpha'-1}e^{-\beta'\tau}$.

$$\tau|x \sim \text{Gamma}(\alpha',\beta') \tag{3.24}$$

  where $\alpha' = \alpha + n/2$ and $\beta' = \beta + \frac{1}{2}\sum(x_i-\bar{x})^2 + \frac{nk}{2(n+k)}(\bar{x}-\nu)^2$.

- **Conditional Posterior of μ:** The second part corresponds to a Normal kernel with precision $k'\tau$.

$$\mu|\tau,x \sim N(\nu', 1/(k'\tau)) \tag{3.25}$$

  where $k' = k+n$ and $\nu' = \frac{k\nu+n\bar{x}}{k+n}$.

3.2 Bayes Rules

The general form of Bayes rule is derived by minimizing risk.

Definition 3.2 (Risk Function and Bayes Risk)**.**

- **Risk Function:** $R(\theta,d) = \int_X L(\theta,d(x))f(x;\theta)dx$
- **Bayes Risk:** The expected risk with respect to the prior.

$$r(\pi,d) = \int_\Theta R(\theta,d)\pi(\theta)d\theta \tag{3.26}$$

Theorem 3.1 (Minimization of Bayes Risk)**.** *Minimizing the Bayes risk $r(\pi,d)$ is equivalent to minimizing the posterior expected loss for each observed x. That is, the Bayes rule $d(x)$ satisfies:*

$$d(x) = \arg\min_a E_{\theta|x}[L(\theta,a)] \tag{3.27}$$

Proof. We start by writing the Bayes risk essentially as a double integral over the parameters and the data. Substituting the definition of the risk function $R(\theta,d)$:

$$\begin{aligned} r(\pi,d) &= \int_\Theta R(\theta,d)\pi(\theta)d\theta \\ &= \int_\Theta \left[\int_X L(\theta,d(x))f(x|\theta)dx\right]\pi(\theta)d\theta \end{aligned} \tag{3.28}$$

Assuming the conditions for Fubini's Theorem are met, we switch the order of integration:

$$r(\pi, d) = \int_X \left[\int_\Theta L(\theta, d(x)) f(x|\theta) \pi(\theta) d\theta \right] dx \tag{3.29}$$

Recall that the joint density can be factored as $f(x, \theta) = f(x|\theta)\pi(\theta) = \pi(\theta|x)m(x)$, where $m(x)$ is the marginal density of the data. Substituting this into the inner integral:

$$\begin{aligned} r(\pi, d) &= \int_X \left[\int_\Theta L(\theta, d(x)) \pi(\theta|x) m(x) d\theta \right] dx \\ &= \int_X m(x) \left[\int_\Theta L(\theta, d(x)) \pi(\theta|x) d\theta \right] dx \end{aligned} \tag{3.30}$$

Since the marginal density $m(x)$ is non-negative, minimizing the total integral $r(\pi, d)$ with respect to the decision rule $d(\cdot)$ is equivalent to minimizing the term inside the brackets for every x (specifically where $m(x) > 0$).

The term inside the brackets is the **Posterior Expected Loss**:

$$\int_\Theta L(\theta, d(x)) \pi(\theta|x) d\theta = E_{\theta|x}[L(\theta, d(x))] \tag{3.31}$$

□

> **!** Important
>
> Therefore, to minimize the Bayes risk, one just need to choose $d(x)$ to minimize the posterior expected loss for each x.

The following diagram summarizes the general workflow for deriving a Bayes estimator:

3.2.1 Common Loss Functions and Bayes Estimators

3.2.1.1 Squared Error Loss (point Estimate)

$$L(\theta, a) = (\theta - a)^2 \tag{3.32}$$

To find the optimal estimator $d(x)$, we minimize the posterior expected loss $E_{\theta|x}[(\theta - d(x))^2]$. Taking the derivative with respect to d and setting it to 0:

$$-2E_{\theta|x}(\theta - d) = 0 \implies d(x) = E(\theta|x) \tag{3.33}$$

Result: The Bayes rule under squared error loss is the **posterior mean**.

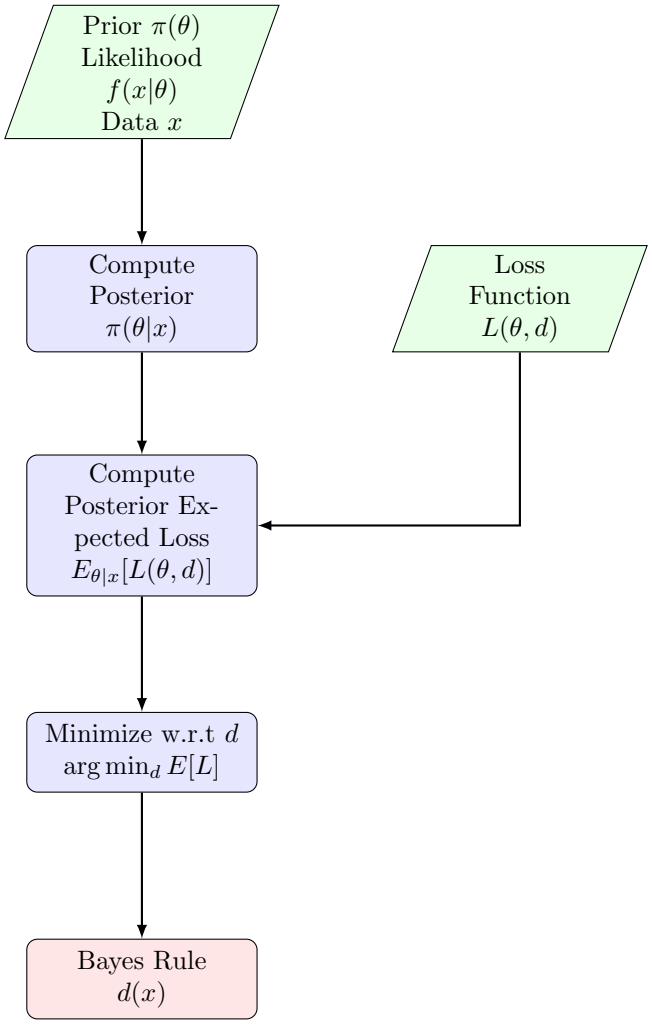


Figure 3.3: Workflow for Finding the Bayes Rule

3.2.1.2 Absolute Error Loss

$$L(\theta, d) = |\theta - d| \tag{3.34}$$

To find the Bayes rule, we minimize the posterior expected loss:

$$\psi(d) = E_{\theta|x}[|\theta - d|] = \int_{-\infty}^{\infty} |\theta - d| \, dF(\theta|x) \tag{3.35}$$

where $F(\theta|x)$ is the cumulative distribution function (CDF) of the posterior. Splitting the integral at the decision point d:

$$\psi(d) = \int_{-\infty}^{d} (d - \theta) \, dF(\theta|x) + \int_{d}^{\infty} (\theta - d) \, dF(\theta|x) \tag{3.36}$$

We find the minimum by analyzing the rate of change of $\psi(d)$ with respect to d. Differentiating (or taking the subgradient for non-differentiable points):

$$\frac{d}{dd}\psi(d) = \int_{-\infty}^{d} 1 \, dF(\theta|x) - \int_{d}^{\infty} 1 \, dF(\theta|x) = P(\theta \leq d|x) - P(\theta > d|x) \tag{3.37}$$

Setting this derivative to zero implies we seek a point where the probability mass to the left equals the probability mass to the right:

$$P(\theta \leq d|x) = P(\theta > d|x) \tag{3.38}$$

Since the total probability is 1, this condition simplifies to finding d such that the cumulative probability is $1/2$.

General Case (Discrete or Mixed Distributions)

In cases where the posterior distribution is discrete or has jump discontinuities (e.g., the CDF jumps from 0.4 to 0.6 at a specific value), an exact solution to $F(d) = 0.5$ may not exist. To generalize, the Bayes rule is defined as any **median** m of the posterior distribution.

A median is formally defined as any value m that satisfies the following two conditions simultaneously:

- $P(\theta \leq m|x) \geq \frac{1}{2}$
- $P(\theta \geq m|x) \geq \frac{1}{2}$

Result: The Bayes rule under absolute error loss is the **posterior median**.

3.2.1.3 Hypothesis Testing (0-1 Loss)

Consider the hypothesis test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. We define the decision space as $\mathcal{A} = \{0, 1\}$, where $a = 0$ means accepting H_0 and $a = 1$ means rejecting H_0 (accepting H_1).

Case 1: 0-1 Loss

The standard 0-1 loss function assigns a penalty of 1 for an incorrect decision and 0 for a correct one:

$$L(\theta, a) = \begin{cases} 0 & \text{if } \theta \in \Theta_0, a = 0 \text{ (Correct } H_0) \\ 1 & \text{if } \theta \in \Theta_0, a = 1 \text{ (Type I Error)} \\ 1 & \text{if } \theta \in \Theta_1, a = 0 \text{ (Type II Error)} \\ 0 & \text{if } \theta \in \Theta_1, a = 1 \text{ (Correct } H_1) \end{cases} \tag{3.39}$$

To find the Bayes rule, we minimize the **posterior expected loss** for a given x, denoted as $E_{\theta|x}[L(\theta, a)]$.

- **Expected Loss for choosing $a = 0$ (Accept H_0):**

$$E_{\theta|x}[L(\theta, 0)] = 0 \cdot P(\theta \in \Theta_0|x) + 1 \cdot P(\theta \in \Theta_1|x) = P(\theta \in \Theta_1|x) \tag{3.40}$$

- **Expected Loss for choosing $a = 1$ (Reject H_0):**

$$E_{\theta|x}[L(\theta, 1)] = 1 \cdot P(\theta \in \Theta_0|x) + 0 \cdot P(\theta \in \Theta_1|x) = P(\theta \in \Theta_0|x) \tag{3.41}$$

The Bayes rule selects the action with the smaller expected loss. Thus, we choose $a = 1$ if:

$$P(\theta \in \Theta_0|x) \leq P(\theta \in \Theta_1|x) \tag{3.42}$$

This confirms that under 0-1 loss, the Bayes rule simply selects the hypothesis with the higher posterior probability.

Case 2: General Loss (Asymmetric Costs)

In many practical applications, the cost of errors is not symmetric. For example, a Type I error (false rejection) might be more costly than a Type II error. Let c_1 be the cost of a Type I error and c_2 be the cost of a Type II error. Usually, we normalize one cost to 1.

Suppose the loss function is:

$$L(\theta, a) = \begin{cases} 0 & \text{if } \theta \in \Theta_0, a = 0 \\ c & \text{if } \theta \in \Theta_0, a = 1 \text{ (Cost of Type I Error)} \\ 1 & \text{if } \theta \in \Theta_1, a = 0 \text{ (Cost of Type II Error)} \\ 0 & \text{if } \theta \in \Theta_1, a = 1 \end{cases} \tag{3.43}$$

We again calculate the posterior expected loss:

- **Expected Loss for $a = 0$:**

$$E[L(\theta, 0)|x] = 0 \cdot P(\Theta_0|x) + 1 \cdot P(\Theta_1|x) = P(\Theta_1|x) \tag{3.44}$$

- **Expected Loss for $a = 1$:**

$$E[L(\theta, 1)|x] = c \cdot P(\Theta_0|x) + 0 \cdot P(\Theta_1|x) = cP(\Theta_0|x) \tag{3.45}$$

We reject H_0 ($a = 1$) if the expected loss of doing so is lower:

$$cP(\Theta_0|x) \leq P(\Theta_1|x) \tag{3.46}$$

Since $P(\Theta_1|x) = 1 - P(\Theta_0|x)$, we can rewrite this condition as:

$$cP(\Theta_0|x) \leq 1 - P(\Theta_0|x) \implies (1 + c)P(\Theta_0|x) \leq 1 \tag{3.47}$$

$$P(\Theta_0|x) \leq \frac{1}{1 + c} \tag{3.48}$$

Result: With asymmetric costs, we accept H_1 only if the posterior probability of the null hypothesis is sufficiently small (below the threshold $\frac{1}{1+c}$). If the cost of false rejection c is high, we require stronger evidence against H_0.

3.2.1.4 Classification Prediction (categorical Parameter)

In classification problems, the parameter of interest is a discrete class label θ (often denoted as y) taking values in a set of categories $\{1, 2, \dots, K\}$. The goal is to predict the true class label based on observed features x.

We typically employ the **0-1 loss function**, which assigns a penalty of 1 for a misclassification and 0 for a correct prediction:

$$L(\theta, \hat{\theta}) = \begin{cases} 0 & \text{if } \hat{\theta} = \theta \text{ (Correct Classification)} \\ 1 & \text{if } \hat{\theta} \neq \theta \text{ (Misclassification)} \end{cases} \tag{3.49}$$

To find the optimal classification rule (the Bayes Classifier), we minimize the posterior expected loss, which is equivalent to minimizing the probability of misclassification.

$$E_{\theta|x}[L(\theta, \hat{\theta})] = \sum_{\theta} L(\theta, \hat{\theta})\pi(\theta|x) \tag{3.50}$$

Since the loss is 1 only when the predicted class $\hat{\theta}$ differs from the true class θ, this sum simplifies to:

$$E_{\theta|x}[L(\theta, \hat{\theta})] = \sum_{\theta \neq \hat{\theta}} 1 \cdot \pi(\theta|x) = P(\theta \neq \hat{\theta}|x) = 1 - P(\theta = \hat{\theta}|x) \tag{3.51}$$

Minimizing the misclassification rate $1 - P(\theta = \hat{\theta}|x)$ is mathematically equivalent to maximizing the probability of being correct, $P(\theta = \hat{\theta}|x)$.

Result: The Bayes rule for classification is to predict the class with the highest posterior probability. While this is technically the **Maximum A Posteriori (MAP)** estimator, in the context of machine learning and pattern recognition, this decision rule is known as the **Bayes Optimal Classifier**.

$$\hat{\theta}_{\text{Bayes}}(x) = \underset{k \in \{1,\ldots,K\}}{\arg\max}\ P(\theta = k|x) \tag{3.52}$$

3.2.1.5 Interval Estimation and Highest Posterior Density (HPD)

We can motivate the choice of a Credible Interval by defining a specific loss function for interval estimation. Suppose we seek an estimate d and specify a tolerance $\delta > 0$. We define the loss function as:

$$L(\theta, d) = \begin{cases} 0 & \text{if } |\theta - d| \leq \delta \\ 1 & \text{if } |\theta - d| > \delta \end{cases} \tag{3.53}$$

The **Expected Posterior Loss** is the posterior probability that θ lies outside the interval $(d - \delta, d + \delta)$. Therefore, minimizing the loss is equivalent to finding the interval of fixed length 2δ that **maximizes the posterior probability**:

$$P(d - \delta \leq \theta \leq d + \delta \mid x) \tag{3.54}$$

In practice, we typically reverse this formulation: instead of fixing the length 2δ, we fix the coverage probability $1 - \alpha$ (e.g., 0.95) and seek the interval with the **shortest possible length**. This results in the **Highest Posterior Density (HPD)** interval, defined as the region where the posterior density exceeds a certain threshold c:

$$C_{HPD} = \{\theta : \pi(\theta \mid x) \geq c\} \tag{3.55}$$

This HPD interval is optimal because it includes the "most likely" values of θ and, for a unimodal distribution, provides the narrowest interval for a given confidence level.

Comparison with Equal-Tailed Intervals:

- **Equal-Tailed Interval:** We simply cut off $\alpha/2$ probability from each tail of the distribution. This is easy to compute but may not be the shortest interval if the distribution is skewed.
- **HPD Interval:** This is the shortest possible interval for the given coverage. For unimodal distributions, the probability density at the two endpoints of the HPD interval is identical.

The plot below illustrates a skewed posterior distribution (Gamma). Notice how the **HPD Interval (Blue)** is shifted toward the mode (the peak) to capture the highest density values, resulting in a shorter interval length compared to the **Equal-Tailed Interval (Red)**.

```
## Define a Skewed Distribution: Gamma(shape=2, Rate=0.5)
x_vals <- seq(0, 15, length.out = 1000)
y_vals <- dgamma(x_vals, shape = 2, rate = 0.5)

## Target Coverage
alpha <- 0.10
target_prob <- 1 - alpha
```

```r
## 1. Equal-tailed Interval (quantiles)
eq_lower <- qgamma(alpha/2, shape = 2, rate = 0.5)
eq_upper <- qgamma(1 - alpha/2, shape = 2, rate = 0.5)

## 2. HPD Interval (density Threshold Optimization)
## We Look for a Density Threshold K Such That the Area Above K Is 0.90
find_hpd <- function(dist_vals, density_vals, probability) {
  ## Sort density values
  ord <- order(density_vals, decreasing = TRUE)
  sorted_dens <- density_vals[ord]
  sorted_dist <- dist_vals[ord]

  ## Accumulate probability (approximation)
  dx <- diff(dist_vals)[1]
  cum_prob <- cumsum(sorted_dens * dx)

  ## Find cutoff index
  cutoff_idx <- which(cum_prob >= probability)[1]

  ## Get the subset of x values
  hpd_set <- sorted_dist[1:cutoff_idx]
  return(c(min(hpd_set), max(hpd_set)))
}

hpd_bounds <- find_hpd(x_vals, y_vals, target_prob)
hpd_lower <- hpd_bounds[1]
hpd_upper <- hpd_bounds[2]

## Plotting
plot(x_vals, y_vals, type = 'l', lwd = 2, col = "black",
     main = "90% Credible Intervals (Skewed Posterior)",
     xlab = expression(theta), ylab = "Density",
     ylim = c(0, max(y_vals) * 1.2))

## Shade HPD
polygon(c(x_vals[x_vals >= hpd_lower & x_vals <= hpd_upper], hpd_upper, hpd_lower),
        c(y_vals[x_vals >= hpd_lower & x_vals <= hpd_upper], 0, 0),
        col = rgb(0, 0, 1, 0.2), border = NA)

## Draw Equal-tailed Lines (red)
abline(v = c(eq_lower, eq_upper), col = "red", lwd = 2, lty = 2)
## Draw HPD Lines (blue)
abline(v = c(hpd_lower, hpd_upper), col = "blue", lwd = 2, lty = 1)

legend("topright",
```

```
       legend = c("Posterior Density",
                  paste0("Equal-Tailed (Len: ", round(eq_upper - eq_lower, 2), ")"),
                  paste0("HPD (Len: ", round(hpd_upper - hpd_lower, 2), ")")),
       col = c("black", "red", "blue"),
       lty = c(1, 2, 1), lwd = 2,
       fill = c(NA, NA, rgb(0, 0, 1, 0.2)), border = NA)
```

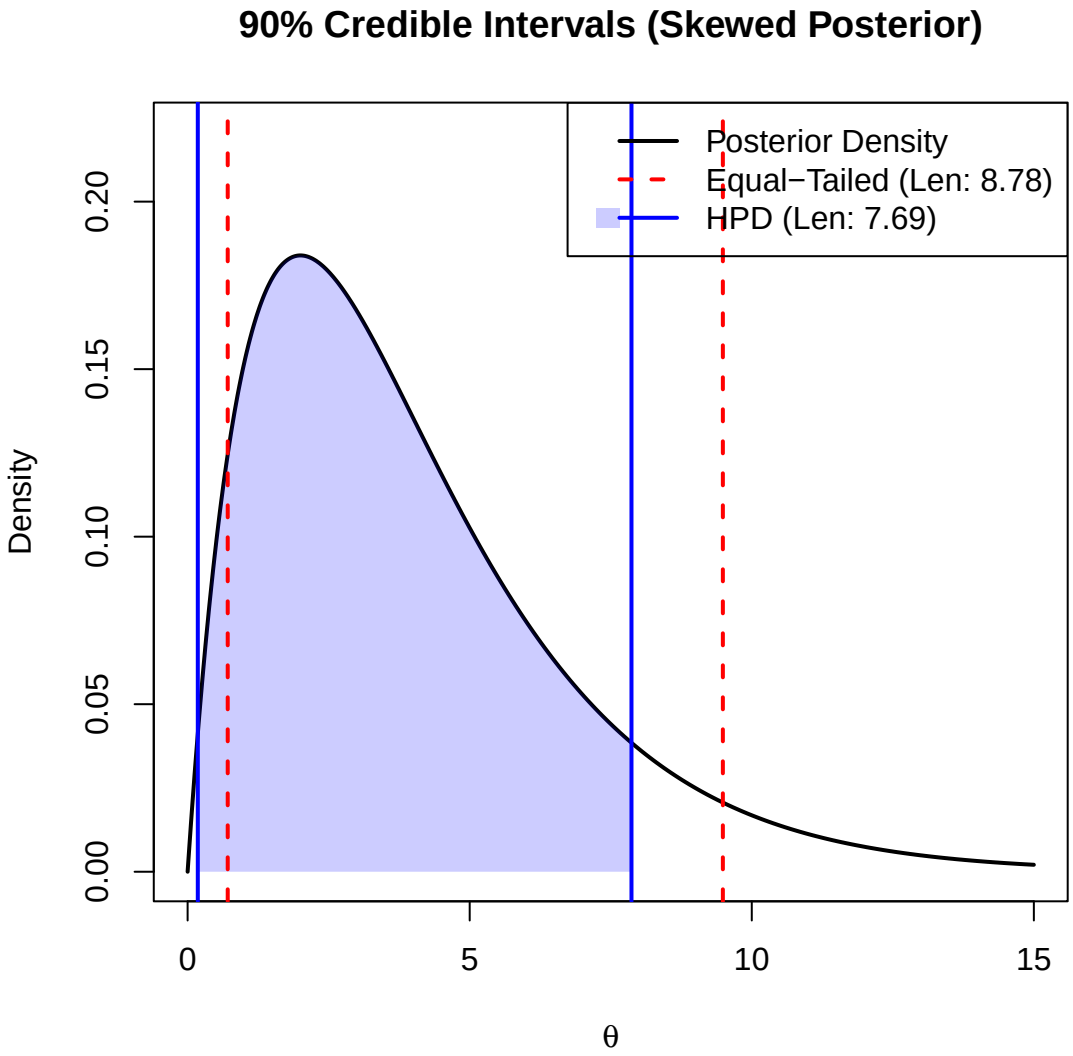


Figure 3.4: Comparison of HPD and Equal-Tailed Intervals for a Skewed Distribution

3.2.2 Finding Minimax Rules with Bayes Rules

?@thm-minimax-constant states that if a Bayes estimator δ^π (derived from a prior π) yields a constant risk $R(\theta, \delta^\pi) = c$ across the entire parameter space Θ, then that estimator is necessarily minimax.

This result is a cornerstone of decision theory because it provides a sufficient condition for minimaxity. While the minimax criterion focuses on the "worst-case scenario" by minimizing the maximum possible risk, the Bayes criterion focuses on the "average-case scenario" relative to a prior. When the risk is constant, these two perspectives align: the average risk equals the maximum risk, and no other estimator can achieve a lower maximum without also having a lower Bayes risk, which would contradict the optimality of the Bayes rule.

In practice, this means that to find a minimax estimator, one can often search for a "least favorable prior" that results in a Bayes estimator with a flat risk profile.

Example 3.5 (Binomial Minimax Estimator)**.** Let $X \sim \text{Bin}(n, \theta)$ and $\theta \sim \text{Beta}(a, b)$. The squared error loss is $L(\theta, d) = (\theta - d)^2$. The Bayes estimator is the posterior mean:

$$d(x) = \frac{a + x}{a + b + n} \tag{3.56}$$

We calculate the risk $R(\theta, d)$:

$$R(\theta, d) = E_x \left[\left(\theta - \frac{a + x}{a + b + n} \right)^2 \right] \tag{3.57}$$

Let $c = a + b + n$.

$$R(\theta, d) = \frac{1}{c^2} E \left[(c\theta - a - x)^2 \right] \tag{3.58}$$

Using the bias-variance decomposition and knowing $E(x) = n\theta$ and $E(x^2) = (n\theta)^2 + n\theta(1-\theta)$, we expand the risk function. To make the risk constant (independent of θ), we set the coefficients of θ and θ^2 to zero.

Solving the resulting system of equations yields:

$$a = b = \frac{\sqrt{n}}{2} \tag{3.59}$$

Thus, the minimax estimator is:

$$d(x) = \frac{x + \sqrt{n}/2}{n + \sqrt{n}} \tag{3.60}$$

This differs from the standard MLE $\hat{p} = x/n$ and the uniform prior Bayes estimator ($a = b = 1$).

3.3 Stein's Paradox and the James-stein Estimator

In high-dimensional estimation ($p \geq 3$), the Maximum Likelihood Estimator (MLE) is inadmissible under squared error loss. The **James-Stein Estimator** dominates the MLE, meaning it achieves lower risk for all values of θ.

Consider the setting:

- Data: $X \sim N_p(\theta, I)$
- Prior: $\theta \sim N_p(0, \tau^2 I)$

- Estimator: $d^{JS}(x) = \left(1 - \frac{p-2}{||x||^2}\right) x$

We can derive the Bayes Risk $r(\pi, d^{JS})$ of this estimator using two equivalent methods: minimizing the expected frequentist risk, or minimizing the expected posterior loss.

Theorem 3.2 (Bayes Risk of James-stein Estimator)**.** *For $p \geq 3$, the Bayes risk of the James-Stein estimator d^{JS} with respect to the prior $\theta \sim N(0, \tau^2 I)$ is:*

$$r(\pi, d^{JS}) = \frac{p\tau^2 + 2}{\tau^2 + 1} \tag{3.61}$$

Proof. **Method 1: Integration over the Prior (Frequentist Risk approach)**

The Bayes risk is defined as $r(\pi, d) = E_\pi[R(\theta, d)]$.

First, recall the frequentist risk of the James-Stein estimator for a fixed θ. Using Stein's Lemma, the risk is given by:

$$R(\theta, d^{JS}) = p - (p-2)^2 E_\theta \left[\frac{1}{||X||^2}\right] \tag{3.62}$$

To find the Bayes risk, we take the expectation of this risk with respect to the prior $\pi(\theta)$:

$$r(\pi, d^{JS}) = \int R(\theta, d^{JS})\pi(\theta)d\theta = p - (p-2)^2 E_\pi \left[E_\theta \left(\frac{1}{||X||^2}\right)\right] \tag{3.63}$$

By the law of iterated expectations, $E_\pi[E_\theta(\cdot)]$ is equivalent to the expectation with respect to the marginal distribution of X, denoted as $m(x)$. Under the conjugate prior, the marginal distribution is $X \sim N(0, (1+\tau^2)I)$.

Consequently, the quantity $\frac{||X||^2}{1+\tau^2}$ follows a Chi-squared distribution with p degrees of freedom (χ^2_p). The expectation of the inverse chi-square is:

$$E\left[\frac{1}{||X||^2}\right] = \frac{1}{1+\tau^2} E\left[\frac{1}{\chi^2_p}\right] = \frac{1}{1+\tau^2} \cdot \frac{1}{p-2} \tag{3.64}$$

Substituting this back into the risk equation:

$$\begin{aligned} r(\pi, d^{JS}) &= p - (p-2)^2 \cdot \frac{1}{(p-2)(1+\tau^2)} \\ &= p - \frac{p-2}{1+\tau^2} \\ &= \frac{p(1+\tau^2) - (p-2)}{1+\tau^2} \\ &= \frac{p\tau^2 + p - p + 2}{1+\tau^2} = \frac{p\tau^2 + 2}{\tau^2 + 1} \end{aligned} \tag{3.65}$$

□

Proof. **Method 2: Integration over the Marginal (Posterior Loss approach)**

Alternatively, we can compute the Bayes risk by first finding the posterior expected loss for a given x, and then averaging over the marginal distribution of x:

$$r(\pi, d) = E_m[E_{\theta|x}[L(\theta, d(x))]] \tag{3.66}$$

Step 1: Posterior Expected Loss

The posterior distribution of θ given x is:

$$\theta|x \sim N\left(\frac{\tau^2}{1+\tau^2}x, \frac{\tau^2}{1+\tau^2}I\right) \tag{3.67}$$

The expected squared error loss can be decomposed into the variance (trace) and the squared bias:

$$E_{\theta|x}[||\theta - d^{JS}(x)||^2] = \text{tr}(\text{Var}(\theta|x)) + ||E[\theta|x] - d^{JS}(x)||^2 \tag{3.68}$$

- **Trace term:**

$$\text{tr}\left(\frac{\tau^2}{1+\tau^2}I_p\right) = \frac{p\tau^2}{1+\tau^2} \tag{3.69}$$

- **Squared Bias term:** Let $B = \frac{1}{1+\tau^2}$. Then $E[\theta|x] = (1-B)x$. The estimator is $d^{JS}(x) = (1 - \frac{p-2}{||x||^2})x$. The difference is:

$$E[\theta|x] - d^{JS}(x) = \left((1-B) - \left(1 - \frac{p-2}{||x||^2}\right)\right)x = \left(\frac{p-2}{||x||^2} - B\right)x \tag{3.70}$$

  Squaring the norm gives:

$$\left(\frac{p-2}{||x||^2} - B\right)^2 ||x||^2 = \frac{(p-2)^2}{||x||^2} - 2B(p-2) + B^2||x||^2 \tag{3.71}$$

Step 2: Expectation with respect to Marginal X

We now take the expectation $E_m[\cdot]$ of the posterior loss. Recall $X \sim N(0, (1+\tau^2)I)$, so $E[||X||^2] = p(1+\tau^2)$ and $E[1/||X||^2] = \frac{1}{(p-2)(1+\tau^2)}$.

- **Expectation of Trace term:** Constant, remains $\frac{p\tau^2}{1+\tau^2}$.
- **Expectation of Bias term:**

$$\begin{aligned} E_m\left[\frac{(p-2)^2}{||X||^2} - \frac{2(p-2)}{1+\tau^2} + \frac{||X||^2}{(1+\tau^2)^2}\right] &= (p-2)^2\frac{1}{(p-2)(1+\tau^2)} - \frac{2(p-2)}{1+\tau^2} + \frac{p(1+\tau^2)}{(1+\tau^2)^2} \\ &= \frac{p-2}{1+\tau^2} - \frac{2p-4}{1+\tau^2} + \frac{p}{1+\tau^2} \\ &= \frac{p-2-2p+4+p}{1+\tau^2} \\ &= \frac{2}{1+\tau^2} \end{aligned} \tag{3.72}$$

Step 3: Combine Terms

$$r(\pi, d^{JS}) = \underbrace{\frac{p\tau^2}{1+\tau^2}}_{\text{Variance Part}} + \underbrace{\frac{2}{1+\tau^2}}_{\text{Bias Part}} = \frac{p\tau^2+2}{\tau^2+1} \tag{3.73}$$

Both methods yield the same result. □

Theorem 3.3 (Inadmissibility of the MLE in High Dimensions (stein's Phenomenon)). *Let $X \sim N_p(\theta, I)$ be a p-dimensional random vector with $p \geq 3$. Under the squared error loss function $L(\theta, d) = ||\theta - d||^2$, the standard Maximum Likelihood Estimator $d^0(X) = X$ is **inadmissible**.*

Proof. To show that $d^0(X) = X$ is inadmissible, we must find another estimator that dominates it (i.e., has equal or lower risk for all θ, and strictly lower risk for at least one θ).

First, consider the risk of the standard estimator d^0. Since $X_i \sim N(\theta_i, 1)$ are independent:

$$R(\theta, d^0) = E_\theta[||X - \theta||^2] = \sum_{i=1}^{p} E[(X_i - \theta_i)^2] = \sum_{i=1}^{p} \text{Var}(X_i) = p \tag{3.74}$$

Now consider the James-Stein estimator $d^{JS}(X) = \left(1 - \frac{p-2}{||X||^2}\right) X$. As established in the derivation of the Bayes Risk in Theorem 3.2, the frequentist risk function of d^{JS} is:

$$R(\theta, d^{JS}) = p - (p-2)^2 E_\theta \left[\frac{1}{||X||^2}\right] \tag{3.75}$$

Since the random variable $||X||^2$ is non-negative and not identically infinity, the expectation $E_\theta[1/||X||^2]$ is strictly positive for all θ. Therefore:

$$R(\theta, d^{JS}) < p = R(\theta, d^0) \quad \text{for all } \theta \in \mathbb{R}^p \tag{3.76}$$

Because d^{JS} achieves a strictly lower risk than d^0 everywhere in the parameter space, d^0 is dominated by d^{JS} and is thus inadmissible. □

3.3.1 Practical Application: One-way ANOVA and "borrowing Strength"

Example 3.6. Consider a One-Way ANOVA setting where we wish to estimate the means of p different independent groups (e.g., the true batting averages of $p = 10$ baseball players, or the efficacy of $p = 5$ different hospital treatments).

- **Model:** Let $X_i \sim N(\theta_i, \sigma^2)$ be the observed sample mean for group i, for $i = 1, \dots, p$.
- **Goal:** Estimate the vector of true means $\theta = (\theta_1, \dots, \theta_p)$ simultaneously. The loss is the sum of squared errors: $L(\theta, \hat{\theta}) = \sum(\theta_i - \hat{\theta}_i)^2$.

The MLE Approach (Total Separation): The standard estimator is $\hat{\theta}_i^{\text{MLE}} = X_i$. This estimates each group entirely independently, using only data from that specific group. If a specific player has a lucky streak, their estimate is very high; if they are unlucky, it is very low.

The James-Stein Approach (Shrinkage / Pooling): In this context, the James-Stein estimator (specifically the variation shrinking toward the grand mean $\bar{X}$) is:

$$\hat{\theta}_i^{JS} = \bar{X} + \left(1 - \frac{(p-3)\sigma^2}{\sum(X_i - \bar{X})^2}\right)(X_i - \bar{X}) \tag{3.77}$$

Why is this better? Even though the groups might be physically independent (e.g., distinct hospitals), the James-Stein estimator **"borrows strength"** from the ensemble.

1. **Noise Reduction:** Extreme observations X_i are likely to contain more positive noise than signal. Shrinking them toward the global average $\bar{X}$ reduces this variance.
2. **Stein's Paradox:** While $\hat{\theta}_i^{JS}$ introduces bias (estimates are pulled toward the center), the reduction in variance is so significant that the **Total Risk** (sum of squared errors over all groups) is strictly lower than that of the MLE, provided $p \geq 3$.

Thus, estimating the groups *together* yields a more accurate global picture than estimating them *separately*, even if the groups are independent.

3.3.2 Why is this Paradoxical?

The result that d^{JS} dominates d^0 is called **Stein's Paradox** because it defies intuition in several ways:

1. **Independence Irrelevance:** The result holds even if the components X_i are completely unrelated (e.g., X_1 is the price of tea in China, X_2 is the temperature in Saskatoon, and X_3 is the weight of a local cat). It seems absurd that combining unrelated data improves the estimate of each, but the combined risk is indeed lower.
2. **No "Free Lunch":** The James-Stein estimator does not improve every individual component θ_i simultaneously for every realization. Instead, it minimizes the **total** risk $\sum E(\hat{\theta}_i - \theta_i)^2$. It sacrifices accuracy on outliers (by biasing them) to gain significant stability on the bulk of the data.
3. **Destruction of Symmetry:** The MLE is invariant under translation and rotation. The James-Stein estimator breaks this symmetry by shrinking toward an arbitrary point (usually the origin or the grand mean), yet it yields a better objective performance.

3.3.3 What We Learned

1. **Bias-Variance Tradeoff:** This is the most famous example where introducing **bias** (shrinkage) leads to a massive reduction in **variance**, thereby reducing the overall Mean Squared Error (MSE). Unbiasedness is not always a virtue in estimation.
2. **Inadmissibility in High Dimensions:** Intuitions formed in 1D or 2D (where MLE is admissible) fail in higher dimensions ($p \geq 3$). The volume of space grows so fast that "standard" diffuse priors or MLEs become inefficient.

3. **Hierarchical Modeling:** Stein's result provides the theoretical foundation for **Hierarchical Bayesian Models**. When we assume parameters come from a common distribution (e.g., $\theta_i \sim N(\mu, \tau^2)$), we naturally derive shrinkage estimators that "borrow strength" across groups, formalized as Empirical Bayes or fully Bayesian methods.

3.4 Empirical Bayes Rules

The James-Stein estimator provides a natural entry point into the concept of **Empirical Bayes (EB)**. While the Stein estimator was originally derived using frequentist risk arguments, it can be intuitively understood as a Bayesian estimator where the parameters of the prior distribution are estimated from the data itself.

3.4.1 The General Empirical Bayes Framework

In a standard Bayesian analysis, the hyperparameters of the prior are fixed based on subjective belief or external information. In contrast, Empirical Bayes uses the observed data to "learn" the prior.

The workflow typically follows these steps:

1. **Hierarchical Model:** We assume the data X comes from a distribution $f(x|\theta)$, and the parameter θ comes from a prior $\pi(\theta|\eta)$ controlled by hyperparameters η.
2. **Marginal Likelihood (Evidence):** We integrate out the parameter θ to obtain the marginal distribution of the data given the hyperparameters:

$$m(x|\eta) = \int f(x|\theta)\pi(\theta|\eta)d\theta \tag{3.78}$$

3. **Estimation of Hyperparameters:** Instead of fixing η, we estimate it by maximizing the marginal likelihood (Type-II Maximum Likelihood) or using method-of-moments:

$$\widehat{\eta} = \arg\max_{\eta} m(x|\eta) \tag{3.79}$$

4. **Posterior Inference:** We proceed with standard Bayesian inference, but we substitute the estimated estimate $\widehat{\eta}$ into the posterior:

$$\pi(\theta|x, \widehat{\eta}) \propto f(x|\theta)\pi(\theta|\widehat{\eta}) \tag{3.80}$$

Discussion:

- **"Borrowing Strength":** EB allows us to pool information across independent groups to estimate the common structure (the prior) governing them.
- **The Critique:** A purist Bayesian might object that using the data twice (once to estimate the prior, once to estimate θ) underestimates the uncertainty. A fully Bayesian Hierarchical model would instead place a "hyperprior" on η and integrate it out.

3.4.2 Deriving James-Stein as Empirical Bayes

We can derive the James-Stein rule explicitly using this framework.

Model:

1. Likelihood: $X_i|\mu_i \sim N(\mu_i, 1)$ for $i = 1, \dots, p$.
2. Prior: $\mu_i \sim N(0, \tau^2)$. Here, τ^2 is the unknown hyperparameter.

Step 1: The Ideal Bayes Estimator

If we knew τ^2, the posterior distribution of μ_i would be Normal with mean:

$$E(\mu_i|x_i, \tau^2) = \frac{\tau^2}{1+\tau^2}x_i = \left(1 - \frac{1}{1+\tau^2}\right)x_i \tag{3.81}$$

We define the shrinkage factor $B = \frac{1}{1+\tau^2}$.

Step 2: Marginal Estimation

Since μ_i and $X_i - \mu_i$ are independent normals, the marginal distribution of the data is:

$$X_i \sim N(0, 1+\tau^2) \tag{3.82}$$

Consequently, the sum of squares $S = ||X||^2 = \sum X_i^2$ follows a scaled Chi-squared distribution:

$$S \sim (1+\tau^2)\chi_p^2 \tag{3.83}$$

Step 3: Estimating the Shrinkage Factor

We need to estimate $B = \frac{1}{1+\tau^2}$. Note that the expected value of an inverse Chi-square variable is $E[1/\chi_p^2] = \frac{1}{p-2}$. Therefore:

$$E\left[\frac{p-2}{S}\right] = \frac{p-2}{1+\tau^2}E\left[\frac{1}{\chi_p^2}\right] = \frac{p-2}{1+\tau^2} \cdot \frac{1}{p-2} = \frac{1}{1+\tau^2} = B \tag{3.84}$$

Thus, $\hat{B} = \frac{p-2}{||X||^2}$ is an unbiased estimator of the optimal shrinkage factor.

Step 4: The Empirical Bayes Rule

Plugging $\hat{B}$ into the ideal Bayes estimator recovers the James-Stein rule:

$$\delta^{EB}(X) = \left(1 - \hat{B}\right)X = \left(1 - \frac{p-2}{||X||^2}\right)X \tag{3.85}$$

3.5 Hierarchical Modeling via MCMC

In complex Bayesian settings where the posterior distribution cannot be derived analytically, we utilize hierarchical structures to represent levels of uncertainty and Markov Chain Monte Carlo (MCMC) to approximate the resulting distributions.

3.5.1 Hierarchical Model Structure

A hierarchical model decomposes a complex joint distribution into a series of conditional levels. The general mathematical form is:

$$\begin{aligned} \text{Level 1 (Data Likelihood):} \quad & X_i|\mu_i, \sigma^2 \sim f(x_i|\mu_i, \sigma^2) \\ \text{Level 2 (Parameters):} \quad & \mu_i|\theta, \tau^2 \sim \pi(\mu_i|\theta, \tau^2) \\ \text{Level 3 (Hyperparameters):} \quad & \theta, \tau^2 \sim \pi(\theta, \tau^2) \end{aligned} \tag{3.86}$$

The goal is to compute the joint posterior distribution of all unobserved parameters given the data $X = \{X_1, ..., X_n\}$:

$$p(\mu, \theta, \tau^2|X) \propto \left[\prod_{i=1}^{n} f(x_i|\mu_i, \sigma^2)\pi(\mu_i|\theta, \tau^2)\right] \pi(\theta, \tau^2) \tag{3.87}$$

3.5.2 Graphical Model Representation (Tree Structure)

The following tree diagram illustrates the conditional dependencies. Note that the parameters μ_i are conditionally independent given the hyperparameter θ, which facilitates "borrowing strength" across groups.

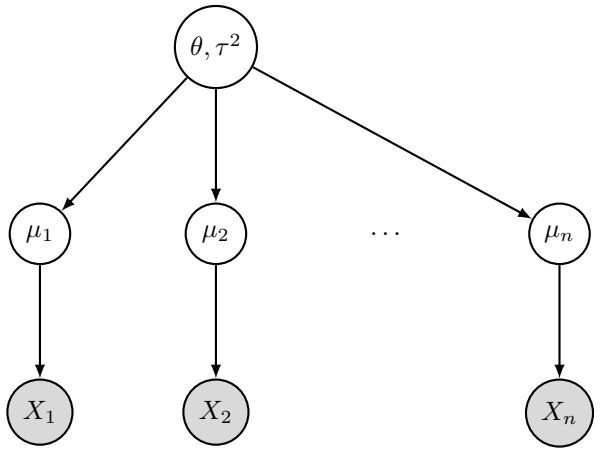


Figure 3.5: Hierarchical Tree Structure

3.5.3 MCMC Estimation

In hierarchical models, the joint posterior distribution $p(\mu, \theta|X)$ often lacks a closed-form analytical solution due to the integration required for the normalizing constant. We use **Markov Chain Monte Carlo (MCMC)** to draw sequence of samples $\{\mu^{(t)}, \theta^{(t)}\}$ that converge to the target posterior distribution.

3.5.3.1 Gibbs Sampling Algorithm

Gibbs sampling is an algorithm for sampling from a multivariate distribution by sequentially sampling from the **full conditional distributions**. To sample from a target distribution $p(\theta_1, \theta_2, \dots, \theta_k)$, the algorithm iterates through each variable, updating it conditioned on the current values of all other variables:

$$\begin{aligned}
\theta_1^{(t+1)} &\sim p(\theta_1 | \theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_k^{(t)}) \\
\theta_2^{(t+1)} &\sim p(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_k^{(t)}) \\
&\vdots \\
\theta_k^{(t+1)} &\sim p(\theta_k | \theta_1^{(t+1)}, \theta_2^{(t+1)}, \dots, \theta_{k-1}^{(t+1)})
\end{aligned} \tag{3.88}$$

Example 3.7 (Gibbs Sampling for Groups of Normal Data)**. The Model**

To apply the general Gibbs sampling framework $\theta_1, \theta_2, \dots, \theta_k$ to our specific hierarchical model, we identify the variables as follows:

- **Data Observations (X_i):** These are the known, measured values at the lowest level of the hierarchy (e.g., test scores of students in school i). In the Gibbs sampler, these remain fixed and condition the updates of the parameters.
- **Group-Level Parameters ($\theta_1 = \mu_i$):** These represent the latent means for each specific group or cluster. In the update step, μ_i acts as the first block of variables. It is updated by "compromising" between the local data X_i and the global characteristic θ.
- **Global Hyperparameter ($\theta_2 = \theta$):** This represents the common mean across all groups. It acts as the second block in the sampler. Its update depends on the current state of all μ_i values, effectively "pooling" information from all groups to estimate the overall population center.

Gibbs Update in Hierarchical Models

In the hierarchical tree structure provided earlier, let our parameter vector be (μ_i, θ). The "orthogonality" of the updates becomes clear when we derive the full conditionals for a Gaussian case:

- **Case $\theta_1 = \mu_i$:** Sample $\mu_i^{(t+1)}$ from $p(\mu_i | X_i, \theta^{(t)})$. This is a normal distribution with:

$$\mu_i^{(t+1)} \sim N\left(\frac{\tau^2 X_i + \sigma^2 \theta^{(t)}}{\sigma^2 + \tau^2}, \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} \right) \tag{3.89}$$

- **Case $\theta_2 = \theta$:** Sample $\theta^{(t+1)}$ from $p(\theta | \mu^{(t+1)})$. Assuming a flat prior $\pi(\theta) \propto 1$:

$$\theta^{(t+1)} \sim N\left(\frac{1}{n} \sum_{i=1}^{n} \mu_i^{(t+1)}, \frac{\tau^2}{n} \right) \tag{3.90}$$

Visual Characteristic: Gibbs sampling moves along the coordinate axes because it updates one parameter at a time while holding others constant.

3.5.3.2 Metropolis-Hastings (MH) Sampling

When the full conditional distributions are not easy to sample from, we use the Metropolis-Hastings algorithm. At each step t:

- **Propose:** Draw a candidate state θ^* from a proposal distribution $q(\theta^*|\theta^{(t)})$.
- **Accept/Reject:** Calculate the acceptance probability:

$$\alpha = \min\left(1, \frac{p(\theta^*|X)q(\theta^{(t)}|\theta^*)}{p(\theta^{(t)}|X)q(\theta^*|\theta^{(t)})}\right) \tag{3.91}$$

- Set $\theta^{(t+1)} = \theta^*$ with probability α; otherwise, set $\theta^{(t+1)} = \theta^{(t)}$.

Visual Characteristic: MH sampling moves in arbitrary directions and can "stay put" if a proposal is rejected, exploring the space via a random walk.

```
set.seed(123)
rho <- 0.8
log_target <- function(x, y) { -0.5 * (x^2 - 2*rho*x*y + y^2) / (1 - rho^2) }

# Gibbs Path (Step-wise update)
gx <- -2; gy <- -2
gx_path <- gx; gy_path <- gy
for(i in 1:25) {
  gx <- rnorm(1, rho * gy, sqrt(1 - rho^2))
  gx_path <- c(gx_path, gx, gx); gy_path <- c(gy_path, gy, gy) # Horizontal move
  gy <- rnorm(1, rho * gx, sqrt(1 - rho^2))
  gx_path <- c(gx_path, gx); gy_path <- c(gy_path, gy) # Vertical move
}

# MH Path (Random Walk)
mx <- numeric(50); my <- numeric(50)
mx[1] <- -2; my[1] <- -2
for(i in 2:50) {
  px <- mx[i-1] + rnorm(1, 0, 0.4); py <- my[i-1] + rnorm(1, 0, 0.4)
  acc <- exp(log_target(px, py) - log_target(mx[i-1], my[i-1]))
  if(runif(1) < acc) { mx[i] <- px; my[i] <- py } else { mx[i] <- mx[i-1]; my[i] <- my[i-1] }
}

par(mfrow = c(1, 2))
t_seq <- seq(-3, 3, length=50); z <- outer(t_seq, t_seq, function(x,y) exp(log_target(x,y)))
plot(gx_path, gy_path, type="l", col="blue", main="Gibbs (Orthogonal Steps)", xlab=expression(the
contour(t_seq, t_seq, z, add=TRUE, col="gray")
plot(mx, my, type="l", col="red", main="Metropolis-Hastings (Random Walk)", xlab=expression(theta
contour(t_seq, t_seq, z, add=TRUE, col="gray")
```

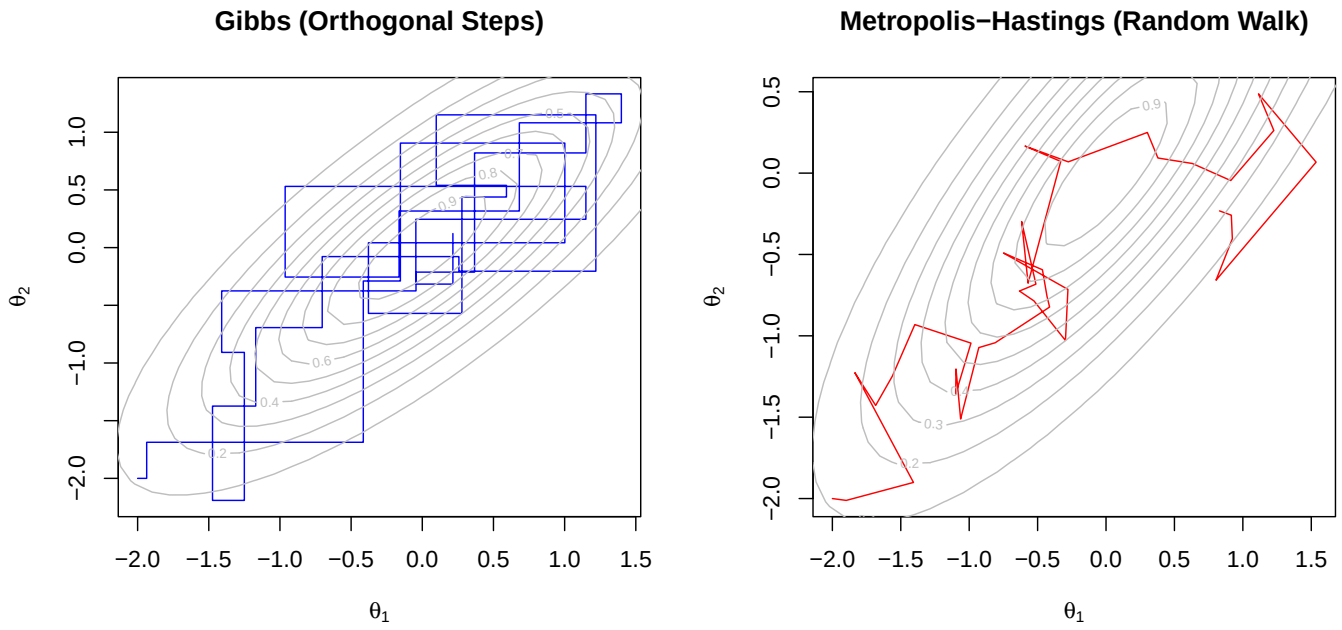


Figure 3.6: Comparison of Sampling Paths

3.6 Case Study: 1998 Major League Baseball Home Run Race

In 1998, the baseball world was captivated by Mark McGwire and Sammy Sosa as they chased Roger Maris' 1961 record of 61 home runs in a single season. While McGwire and Sosa finished with 70 and 66 home runs respectively, we consider whether such performance could have been predicted using pre-season exhibition data.

For a set of $i = 1, \ldots, 17$ players (including McGwire and Sosa), we observe their batting records in pre-season exhibition matches. Our goal is to estimate each player's home run "strike rate" for the competitive season.

3.6.1 Transforming Data

We utilize the pre-season home runs (y_i) and at-bats (n_i) for 17 players. The data is transformed using a variance-stabilizing transformation to approximate a normal distribution with known variance $\sigma^2 = 1$.

$$x_i = \sqrt{n_i} \arcsin\left(2\frac{y_i}{n_i} - 1\right) \tag{3.92}$$

The goal is to estimate the latent parameter μ_i for each player and compare it to the "true" regular season performance.

3.6.2 True Season Parameter (μ_i or p_i^{Season})

To validate our estimates, we define the "true" parameter value μ_i using the player's performance over the full competitive season. Let Y_i be the total home runs and N_i be the total at-bats in the regular season. The true transformed rate is calculated as:

$$\mu_i^{\text{season}} = \sqrt{n_i} \arcsin\left(2\frac{Y_i}{N_i} - 1\right) \tag{3.93}$$

Note that while we use the season-long probability (Y_i/N_i), we scale it by the pre-season sample size ($\sqrt{n_i}$). This ensures that μ_i^{season} is on the same scale as our observations x_i, allowing for direct comparison of the estimation error.

```
library(ggplot2)
library(brms)
library(dplyr)
library(tidyr)

# 1. Input Raw Data
ni <- c(58, 59, 74, 84, 69, 63, 60, 54, 53, 60, 66, 66, 72, 64, 42, 38, 58)
yi <- c(7, 9, 4, 7, 3, 6, 2, 10, 2, 2, 4, 3, 2, 5, 3, 2, 6)
Ni <- c(509, 643, 633, 645, 606, 555, 619, 609, 552, 540, 561, 440, 585, 531, 454, 504, 244)
Yi <- c(70, 66, 56, 46, 45, 44, 43, 40, 37, 34, 32, 30, 29, 28, 23, 21, 15)

# 2. Calculate Derived Values
p_pre   <- yi / ni                          # Pre-season Probability
x       <- sqrt(ni) * asin(2 * p_pre - 1) # Transformed Pre-season (x_i)

p_season <- Yi / Ni                         # Season Probability
true_mu  <- sqrt(ni) * asin(2 * p_season - 1) # Transformed Season (mu_i)

# 3. Create Main Data Frame
baseball_data <- data.frame(
  Player = 1:17,
  Pre_HR = yi,
  Pre_AtBats = ni,
  p_pre = round(p_pre, 3),
  x = x,
  sei = 1, # Known standard error for transformed data
  Season_HR = Yi,
  Season_AtBats = Ni,
  p_season = p_season,
  true_mu = true_mu
)

# 4. Display the Data
```

```
knitr::kable(baseball_data,
             col.names = c("Player", "$y_i$", "$n_i$", "$p_i^{\\text{pre}}$", "$x_i$", "SE",
                           "$Y_i$", "$N_i$", "$p_i^{\\text{seas}}$", "$\\mu_i$"),
             align = "c",
             digits = 3,
             caption = "1998 MLB Statistics: Raw Counts, Probabilities, and Transformed Data")
```

Table 3.2: 1998 MLB Statistics: Raw Counts, Probabilities, and Transformed Data

Player	y_i	n_i	p_i^{pre}	x_i	SE	Y_i	N_i	p_i^{seas}	μ_i
1	7	58	0.121	-6.559	1	70	509	0.138	-6.176
2	9	59	0.153	-5.901	1	66	643	0.103	-7.055
3	4	74	0.054	-9.476	1	56	633	0.088	-8.317
4	7	84	0.083	-9.029	1	46	645	0.071	-9.441
5	3	69	0.043	-9.558	1	45	606	0.074	-8.463
6	6	63	0.095	-7.488	1	44	555	0.079	-7.937
7	2	60	0.033	-9.323	1	43	619	0.069	-8.035
8	10	54	0.185	-5.005	1	40	609	0.066	-7.734
9	2	53	0.038	-8.589	1	37	552	0.067	-7.622
10	2	60	0.033	-9.323	1	34	540	0.063	-8.238
11	4	66	0.061	-8.720	1	32	561	0.057	-8.843
12	3	66	0.045	-9.270	1	30	440	0.068	-8.469
13	2	72	0.028	-10.487	1	29	585	0.050	-9.518
14	5	64	0.078	-8.034	1	28	531	0.053	-8.859
15	3	42	0.071	-6.673	1	23	454	0.051	-7.237
16	2	38	0.053	-6.829	1	21	504	0.042	-7.149
17	6	58	0.103	-6.975	1	15	244	0.061	-8.146

In this analysis, we model the home run strike rates of 17 Major League Baseball players using pre-season exhibition data from 1998. We apply five statistical methods ranging from simple independent estimation to advanced Bayesian decision theory.

3.6.3 Methods for Estimating μ_i (Transformed Scale)

3.6.3.1 Method 1: Simple Estimation (MLE)

The Maximum Likelihood Estimator (MLE) assumes each player's performance is independent. It relies solely on the observed pre-season data.

$$\hat{\mu}_i^{MLE} = X_i \tag{3.94}$$

```r
# Simple Estimate is just the data itself
mu_mle <- baseball_data$x

# MSE Calculation (Transformed Scale)
mse_mle <- mean((mu_mle - baseball_data$true_mu)^2)
```

3.6.3.2 Method 2: Empirical Bayes (James-Stein)

The James-Stein estimator introduces a global mean $\bar{X}$ and shrinks individual estimates toward it. This assumes the players come from a common population distribution.

$$\hat{\mu}_i^{JS} = \bar{X} + \left(1 - \frac{k-3}{\sum(X_i - \bar{X})^2}\right)(X_i - \bar{X}) \tag{3.95}$$

where $k = 17$ is the number of players.

```r
theta_hat <- mean(baseball_data$x)
S <- sum((baseball_data$x - theta_hat)^2)
shrinkage_factor <- 1 - (14 / S)

mu_js <- theta_hat + shrinkage_factor * (baseball_data$x - theta_hat)

# MSE Calculation (Transformed Scale)
mse_js <- mean((mu_js - baseball_data$true_mu)^2)
```

3.6.3.3 Method 3: Fully Bayesian MCMC (brms)

We use a hierarchical Bayesian model where parameters are treated as random variables. We implement this using `brms`.

$$\begin{aligned} X_i &\sim N(\mu_i, 1) \\ \mu_i &\sim N(\theta, \tau^2) \\ \theta &\sim N(0, 10) \\ \tau &\sim \text{Cauchy}(0, 2) \end{aligned} \tag{3.96}$$

```r
# Fit Random Intercept Model: x | se(1) ~ 1 + (1|Player)
fit_brms <- brm(
  formula = x | se(sei, sigma = TRUE) ~ 1 + (1 | Player),
  data = baseball_data,
  prior = c(
    prior(normal(0, 10), class = "Intercept"),
    prior(cauchy(0, 2), class = "sd")
```

```
  ),
  chains = 2, iter = 4000, warmup = 1000, seed = 123,
  refresh = 0
)

# Extract Point Estimates (Posterior Means)
post_means <- fitted(fit_brms)[, "Estimate"]
mu_brms <- post_means

# MSE Calculation (Transformed Scale)
mse_brms <- mean((mu_brms - baseball_data$true_mu)^2)
```

3.6.4 Comparison of Estimates of μ_i

Full Comparison of Estimates (Transformed Scale)

The following table presents the transformed data (x_i) and the true season parameter (μ_i) alongside the estimates from the three methods. The rows are sorted by x_i to visualize how the shrinkage methods (James-Stein and Bayesian) pull the estimates away from the extremes and toward the population mean compared to the raw MLE.

```
# 1. Compile all estimates into a single data frame
df_estimates <- data.frame(
  Player = 1:17,
  ni = baseball_data$Pre_AtBats,
  x_i = baseball_data$x,                # MLE Estimate (Raw Data)
  mu_js = mu_js,                        # James-Stein Estimate
  mu_bayes = mu_brms,                   # Fully Bayesian Estimate
  mu_true = baseball_data$true_mu       # True Season Parameter
)

# 2. Sort by x_i (ascending)
df_sorted <- df_estimates[order(df_estimates$x_i), ]

# 3. Display the table
df_display_mu <- df_sorted
df_display_mu[, 3:6] <- round(df_display_mu[, 3:6], 3)

knitr::kable(df_display_mu[, c("Player", "x_i", "mu_js", "mu_bayes", "mu_true")],
             row.names = FALSE,
             col.names = c("Player", "$x_i$ (MLE)", "$\\hat{\\mu}_{JS}$",
                           "$\\hat{\\mu}_{Bayes}$", "$\\mu_{true}$"),
             align = "c",
             caption = "Comparison of Estimates (Sorted by Pre-season $x_i$)")
```

Table 3.3: Comparison of Estimates (Sorted by Pre-season x_i)

Table 3.3: Comparison of Estimates (Sorted by Pre-season x_i)

Player	x_i (MLE)	$\hat{\mu}_{JS}$	$\hat{\mu}_{Bayes}$	μ_{true}
13	-10.487	-9.589	-8.746	-9.518
5	-9.558	-9.006	-8.478	-8.463
3	-9.476	-8.954	-8.470	-8.317
7	-9.323	-8.858	-8.412	-8.035
10	-9.323	-8.858	-8.415	-8.238
12	-9.270	-8.825	-8.412	-8.469
4	-9.029	-8.673	-8.331	-9.441
11	-8.720	-8.479	-8.260	-8.843
9	-8.589	-8.397	-8.206	-7.622
14	-8.034	-8.048	-8.054	-8.859
6	-7.488	-7.705	-7.897	-7.937
17	-6.975	-7.384	-7.754	-8.146
16	-6.829	-7.292	-7.714	-7.149
15	-6.673	-7.194	-7.663	-7.237
1	-6.559	-7.122	-7.628	-6.176
2	-5.901	-6.709	-7.441	-7.055
8	-5.005	-6.146	-7.186	-7.734

Plots of Errors (Sorted by x_i)

This plot displays the Squared Error for each player. The x-axis represents the players sorted from lowest pre-season performance to highest.

```
# Calculate Squared Errors using the SORTED dataframe
err_mle  <- (df_sorted$x_i - df_sorted$mu_true)^2
err_js   <- (df_sorted$mu_js - df_sorted$mu_true)^2
err_brms <- (df_sorted$mu_bayes - df_sorted$mu_true)^2

# Determine Y-axis range
y_max <- max(c(err_mle, err_js, err_brms))

# Plot MLE Errors (Baseline)
plot(1:17, err_mle, type = "b", pch = 1, col = "black", lty = 2,
     xlab = "Player Index (Sorted by Pre-season Performance)",
     ylab = expression(Squared~Error~~(hat(mu) - mu[true])^2),
     main = "Estimation Error Comparison (Sorted)",
     ylim = c(0, y_max))

# Add James-Stein Errors
lines(1:17, err_js, type = "b", pch = 19, col = "blue")
```

```r
# Add Bayesian (brms) Errors
lines(1:17, err_brms, type = "b", pch = 17, col = "red")

# Add Grid and Legend
grid()
legend("topleft",
       title = "Mean Squared Error",
       legend = c(paste0("MLE: ", round(mse_mle, 3)),
                  paste0("JS: ", round(mse_js, 3)),
                  paste0("Bayes: ", round(mse_brms, 3))),
       col = c("black", "blue", "red"),
       pch = c(1, 19, 17),
       lty = c(2, 1, 1))
```

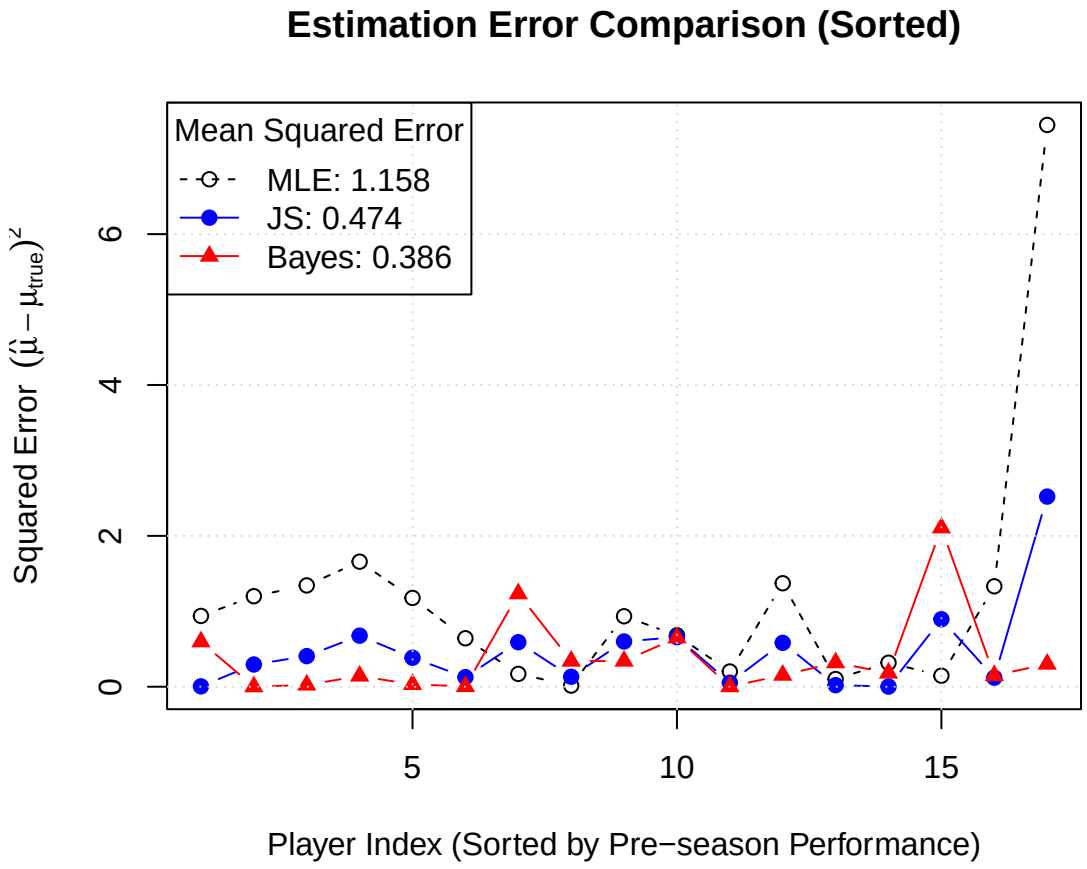


Figure 3.7: Squared Error by Sorted Player Index (Transformed Scale)

3.6.5 Methods for Estimating p_i directly

3.6.5.1 Method 1-3: Converting $\hat{\mu}_i$ back to p_i

The first three methods (MLE, James-Stein, and Normal-Normal Bayes) estimated the parameter μ_i on the transformed scale. To obtain the probability estimates $\hat{p}_i$, we apply the inverse of the variance-stabilizing transformation:

$$\hat{p}_i = \frac{1}{2}\left(\sin\left(\frac{\hat{\mu}_i}{\sqrt{n_i}}\right) + 1\right) \tag{3.97}$$

where $\hat{\mu}_i$ corresponds to the estimate derived from Method 1, 2, or 3, and n_i is the number of pre-season at-bats for player i.

3.6.5.2 Method 4: Hierarchical Logistic Regression (Logit-Normal)

In this fourth method, we model the probability p_i directly using a hierarchical structure on the log-odds scale, rather than transforming the data.

We assume the count y_i follows a Binomial distribution. The log-odds (logit) of the success rate p_i are drawn from a common Normal distribution with unknown mean μ_0 and standard deviation τ_0.

$$\begin{aligned} y_i|p_i &\sim \text{Binomial}(n_i, p_i) \\ \text{logit}(p_i) &\sim N(\mu_0, \tau_0^2) \\ \mu_0 &\sim N(0, 10) \\ \tau_0 &\sim \text{Cauchy}(0, 2) \end{aligned} \tag{3.98}$$

We implement this in `brms` using the `binomial` family with a logit link. The individual point estimate $\hat{p}_i$ is the **posterior mean** of p_i. Note that because the inverse-logit function is non-linear, the posterior mean of p_i is not simply the inverse-logit of the posterior mean of the random effect; `brms` handles this integration automatically via the `fitted()` function.

```
# 1. Fit Hierarchical Logistic Regression
# Formula: y | trials(n) ~ 1 + (1 | Player)
# This estimates a global intercept (mu_0) and random intercepts for each player (logit(p_i))
fit_logit <- brm(
  formula = Pre_HR | trials(Pre_AtBats) ~ 1 + (1 | Player),
  data = baseball_data,
  family = binomial(link = "logit"),
  prior = c(
    prior(normal(0, 5), class = "Intercept"),
    prior(cauchy(0, 2), class = "sd")
  ),
  chains = 2, iter = 4000, warmup = 1000, seed = 123,
```

```
  refresh = 0
)

# 2. Extract Posterior Means of p_i
# fitted() returns the posterior expectations of the response (Expected Count).
fitted_counts <- fitted(fit_logit)
p_hat_logit <- fitted_counts[, "Estimate"] / baseball_data$Pre_AtBats
```

3.6.5.3 Method 5: Optimal Bayes Estimator (Weighted Median)

While the posterior mean (Method 4) minimizes the Mean Squared Error (MSE), it is not necessarily optimal for the **Relative Standardized Error** metric we defined earlier:

$$L(p, \hat{p}) = \frac{|p - \hat{p}|}{\min(p, 1-p)} \tag{3.99}$$

This is a form of weighted absolute error loss, where the weight is $w(p) = \frac{1}{\min(p,1-p)}$. Theoretical derivation shows that the estimator minimizing the expected posterior loss for this function is the **Weighted Posterior Median**.

We compute this by extracting the full posterior samples from the Logit-Normal model (Method 4) and calculating the weighted median for each player.

```
# 1. Extract Posterior Samples (N_samples x 17 players)
# posterior_epred gives samples of the expected count (N * p)
post_counts <- posterior_epred(fit_logit)

# Convert to probability scale by dividing by trials
p_samples <- sweep(post_counts, 2, baseball_data$Pre_AtBats, "/")

# 2. Define Function for Weighted Median
# Finds the value 'q' such that sum(weights where x <= q) >= 0.5 * total_weight
get_weighted_median <- function(samples) {
  # Calculate weights based on the loss function denominator
  # Avoid division by exact zero (unlikely but safer)
  denom <- pmin(samples, 1 - samples)
  denom[denom < 1e-6] <- 1e-6
  weights <- 1 / denom

  # Normalize weights
  weights_norm <- weights / sum(weights)

  # Sort samples and weights
  ord <- order(samples)
  samp_sorted <- samples[ord]
  w_sorted <- weights_norm[ord]
```

```
  # Find cutoff
  cum_w <- cumsum(w_sorted)
  idx <- which(cum_w >= 0.5)[1]

  return(samp_sorted[idx])
}

# 3. Apply to all players
p_hat_optimal <- apply(p_samples, 2, get_weighted_median)
```

3.6.5.4 Comparison of All Five Estimates (Probability Scale)

We now compare all five methods: MLE, James-Stein (transformed), Bayes Normal-Normal (transformed), Hierarchical Logit-Normal (Posterior Mean), and Optimal Bayes (Weighted Median).

1. MSE Comparison

```
# 1. Prepare Estimates from Previous Steps
inv_trans <- function(mu, n) { 0.5 * (sin(mu / sqrt(n)) + 1) }

# Convert transformed estimates back to probability scale
p_mle    <- inv_trans(baseball_data$x, baseball_data$Pre_AtBats)
p_js     <- inv_trans(mu_js, baseball_data$Pre_AtBats)
p_normal <- inv_trans(mu_brms, baseball_data$Pre_AtBats)
# p_hat_logit (Method 4) and p_hat_optimal (Method 5) are already calculated

# 2. Combine into DataFrame
df_compare <- data.frame(
  Player = baseball_data$Player,
  x_i    = baseball_data$x, # For sorting
  p_true = baseball_data$p_season,
  p_mle  = p_mle,
  p_js   = p_js,
  p_norm = p_normal,
  p_logit = p_hat_logit,
  p_opt   = p_hat_optimal
)

# Sort by initial performance
df_compare_sorted <- df_compare[order(df_compare$x_i), ]

# 3. Calculate MSE
mse_p_mle   <- mean((df_compare_sorted$p_mle - df_compare_sorted$p_true)^2)
mse_p_js    <- mean((df_compare_sorted$p_js - df_compare_sorted$p_true)^2)
```

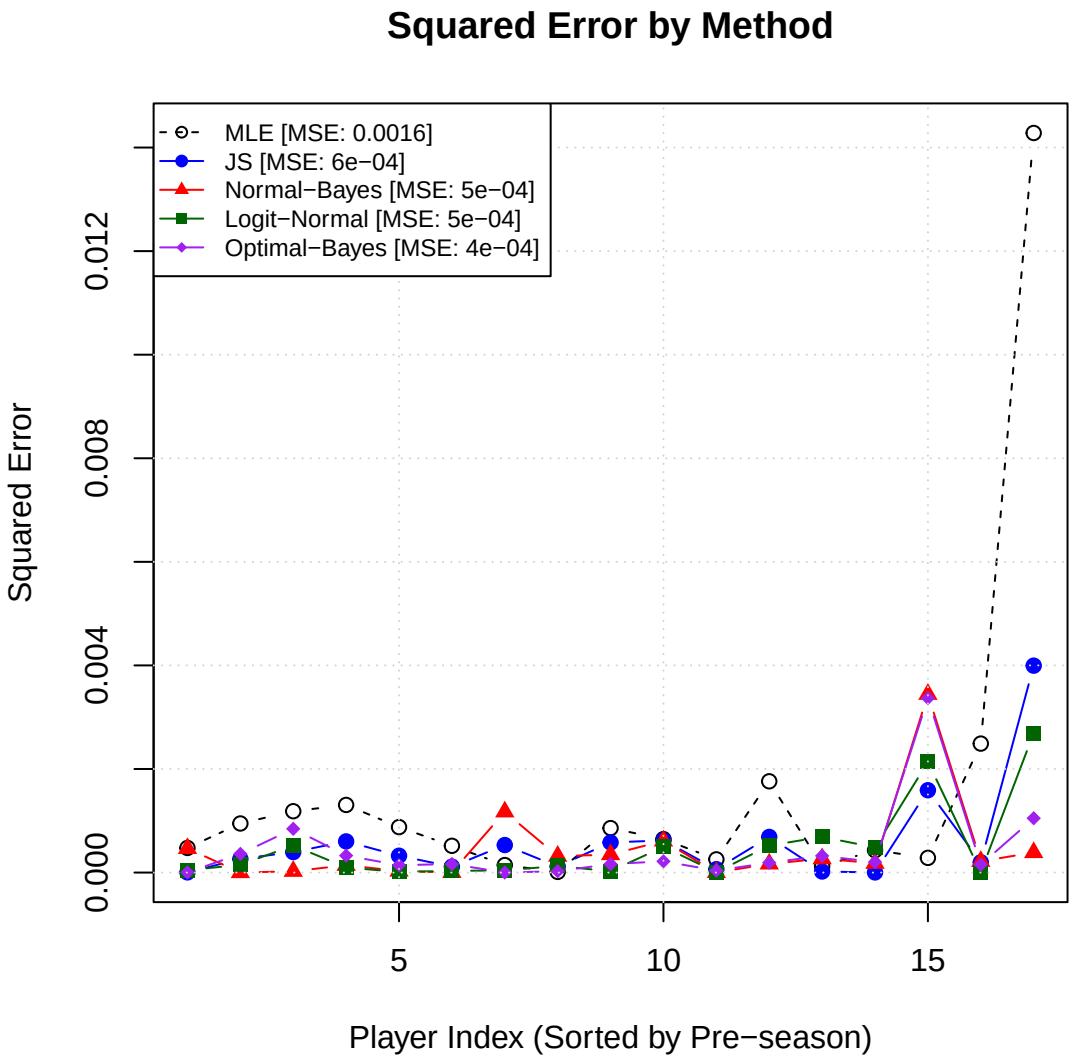
```r
mse_p_norm  <- mean((df_compare_sorted$p_norm - df_compare_sorted$p_true)^2)
mse_p_logit <- mean((df_compare_sorted$p_logit - df_compare_sorted$p_true)^2)
mse_p_opt   <- mean((df_compare_sorted$p_opt - df_compare_sorted$p_true)^2)

# 4. Plot MSE
y_max <- max((df_compare_sorted$p_mle - df_compare_sorted$p_true)^2)

plot(1:17, (df_compare_sorted$p_mle - df_compare_sorted$p_true)^2,
     type = "b", pch = 1, col = "black", lty = 2,
     xlab = "Player Index (Sorted by Pre-season)",
     ylab = "Squared Error",
     main = "Squared Error by Method",
     ylim = c(0, y_max))

lines(1:17, (df_compare_sorted$p_js - df_compare_sorted$p_true)^2, type = "b", pch = 19, col = "b
lines(1:17, (df_compare_sorted$p_norm - df_compare_sorted$p_true)^2, type = "b", pch = 17, col =
lines(1:17, (df_compare_sorted$p_logit - df_compare_sorted$p_true)^2, type = "b", pch = 15, col =
lines(1:17, (df_compare_sorted$p_opt - df_compare_sorted$p_true)^2, type = "b", pch = 18, col = "

grid()
legend("topleft",
       legend = c(paste0("MLE [MSE: ", round(mse_p_mle, 4), "]"),
                  paste0("JS [MSE: ", round(mse_p_js, 4), "]"),
                  paste0("Normal-Bayes [MSE: ", round(mse_p_norm, 4), "]"),
                  paste0("Logit-Normal [MSE: ", round(mse_p_logit, 4), "]"),
                  paste0("Optimal-Bayes [MSE: ", round(mse_p_opt, 4), "]")),
       col = c("black", "blue", "red", "darkgreen", "purple"),
       pch = c(1, 19, 17, 15, 18),
       lty = c(2, 1, 1, 1, 1),
       cex = 0.75,
       bg = "white")
```



2. Relative Standardized Error

We also evaluate the methods using the relative error metric that penalizes deviations based on the rarity of the event:

$$\text{Metric}_i = \frac{|p_i^{\text{true}} - \hat{p}_i|}{\min(p_i^{\text{true}}, 1 - p_i^{\text{true}})} \tag{3.100}$$

```
# 1. Define Metric
calc_metric <- function(p_hat, p_true) {
  denom <- pmin(p_true, 1 - p_true)
  abs(p_hat - p_true) / denom
}

# 2. Calculate Metric
rel_mle   <- calc_metric(df_compare_sorted$p_mle, df_compare_sorted$p_true)
rel_js    <- calc_metric(df_compare_sorted$p_js, df_compare_sorted$p_true)
rel_norm  <- calc_metric(df_compare_sorted$p_norm, df_compare_sorted$p_true)
```

```r
rel_logit <- calc_metric(df_compare_sorted$p_logit, df_compare_sorted$p_true)
rel_opt   <- calc_metric(df_compare_sorted$p_opt, df_compare_sorted$p_true)

# 3. Sum of Errors
sum_rel_mle   <- sum(rel_mle)
sum_rel_js    <- sum(rel_js)
sum_rel_norm  <- sum(rel_norm)
sum_rel_logit <- sum(rel_logit)
sum_rel_opt   <- sum(rel_opt)

# 4. Plot
y_max_rel <- max(c(rel_mle, rel_js, rel_norm, rel_logit, rel_opt)) * 1.1

plot(1:17, rel_mle, type = "b", pch = 1, col = "black", lty = 2,
     xlab = "Player Index (Sorted by Pre-season)",
     ylab = "Relative Standardized Error",
     main = "Assessment of Estimation Methods",
     ylim = c(0, y_max_rel))

lines(1:17, rel_js, type = "b", pch = 19, col = "blue")
lines(1:17, rel_norm, type = "b", pch = 17, col = "red")
lines(1:17, rel_logit, type = "b", pch = 15, col = "darkgreen")
lines(1:17, rel_opt, type = "b", pch = 18, col = "purple")

grid()

legend("topleft",
       title = "Method [Sum Relative Error]",
       legend = c(paste0("MLE [", round(sum_rel_mle, 3), "]"),
                  paste0("James-Stein [", round(sum_rel_js, 3), "]"),
                  paste0("Normal-Bayes [", round(sum_rel_norm, 3), "]"),
                  paste0("Logit-Normal [", round(sum_rel_logit, 3), "]"),
                  paste0("Optimal-Bayes [", round(sum_rel_opt, 3), "]")),
       col = c("black", "blue", "red", "darkgreen", "purple"),
       pch = c(1, 19, 17, 15, 18),
       lty = c(2, 1, 1, 1, 1),
       cex = 0.75,
       bg = "white")
```

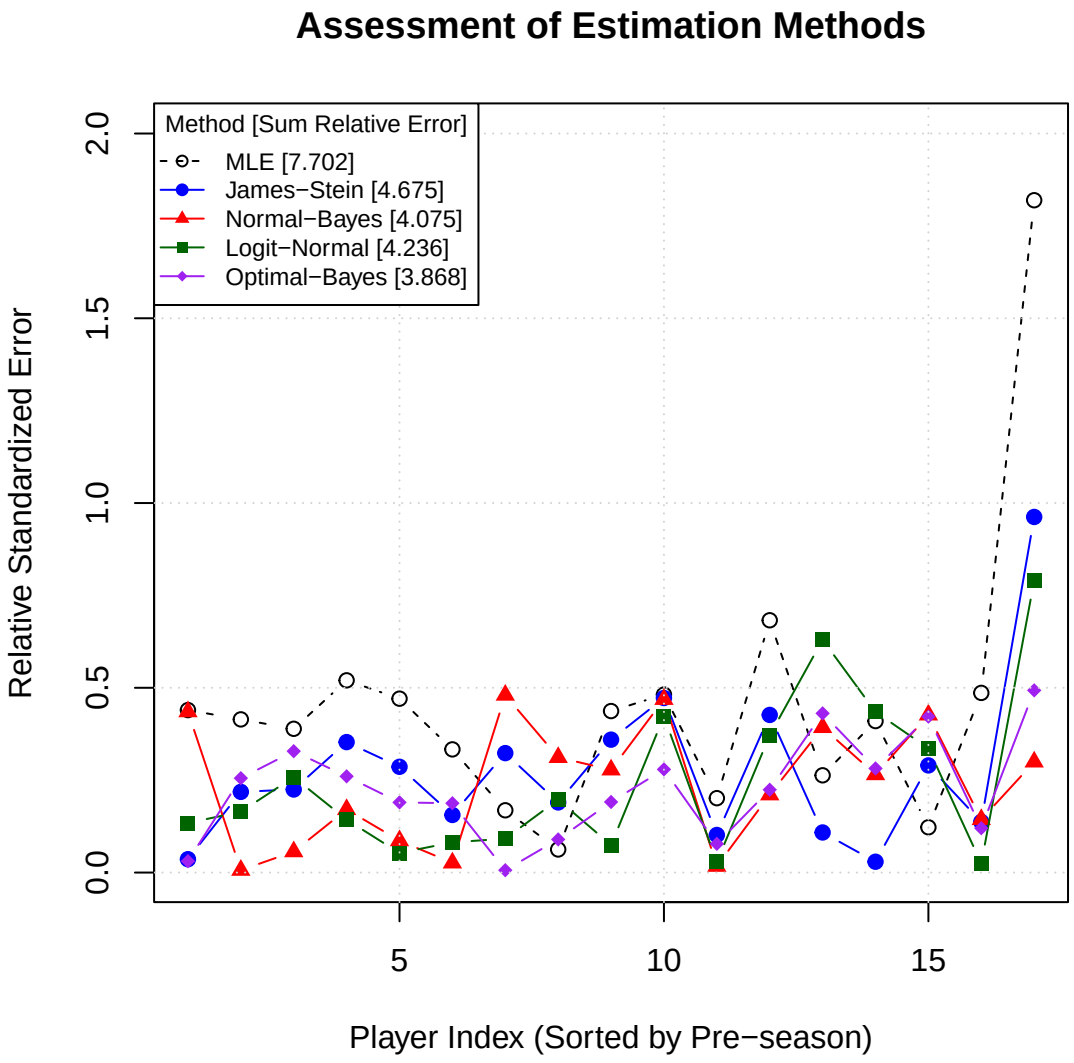


Figure 3.8: Relative Error Assessment: Five Methods

3.7 Bayesian Predictive Distributions

A key feature of Bayesian analysis is the ability to make inference about future observations, rather than just the model parameters. The **posterior predictive distribution** describes the probability of observing a new data point y^* given the observed data y.

Definition 3.3 (Posterior Predictive Distribution)**.** Let $f(y^*|\theta)$ be the sampling distribution of a future observation y^* given parameter θ, and let $\pi(\theta|y)$ be the posterior distribution of θ given observed data y. The posterior predictive density is obtained by marginalizing over the parameter θ:

$$f(y^*|y) = \int_{\Theta} f(y^*|\theta)\pi(\theta|y)\, d\theta \tag{3.101}$$

This distribution incorporates two distinct sources of uncertainty:

- **Sampling Uncertainty (Aleatoric):** The inherent variability of the data generation process, represented by the variance in $f(y^*|\theta)$.
- **Parameter Uncertainty (Epistemic):** The uncertainty regarding the true value of θ, represented by the variance in the posterior $\pi(\theta|y)$.

As sample size $n \to \infty$, the parameter uncertainty vanishes (the posterior approaches a point mass), and the predictive distribution converges to the true data-generating distribution.

Example 3.8 (Normal-Normal Predictive Distribution)**.** Consider a case where the data $y_1, \dots, y_n$ are independent and normally distributed with unknown mean μ and known variance σ^2:

$$Y_i|\mu \sim N(\mu, \sigma^2) \tag{3.102}$$

Assume a conjugate prior for the mean: $\mu \sim N(\mu_0, \sigma_0^2)$. The posterior distribution is $\mu|y \sim N(\mu_n, \sigma_n^2)$, where μ_n and σ_n^2 are the updated posterior hyperparameters.

The predictive distribution for a new observation y^* is derived as:

$$\begin{aligned} f(y^*|y) &= \int_{-\infty}^{\infty} f(y^*|\mu)\pi(\mu|y)\, d\mu \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^*-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\mu-\mu_n)^2}{2\sigma_n^2}}\, d\mu \end{aligned} \tag{3.103}$$

This convolution of two Gaussians results in a new Gaussian distribution:

$$y^*|y \sim N(\mu_n, \sigma^2 + \sigma_n^2) \tag{3.104}$$

Here, the total predictive variance is the sum of the data variance (σ^2) and the posterior uncertainty about the mean (σ_n^2).