Statistical Inference

Longhai Li

2026-01-06

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.

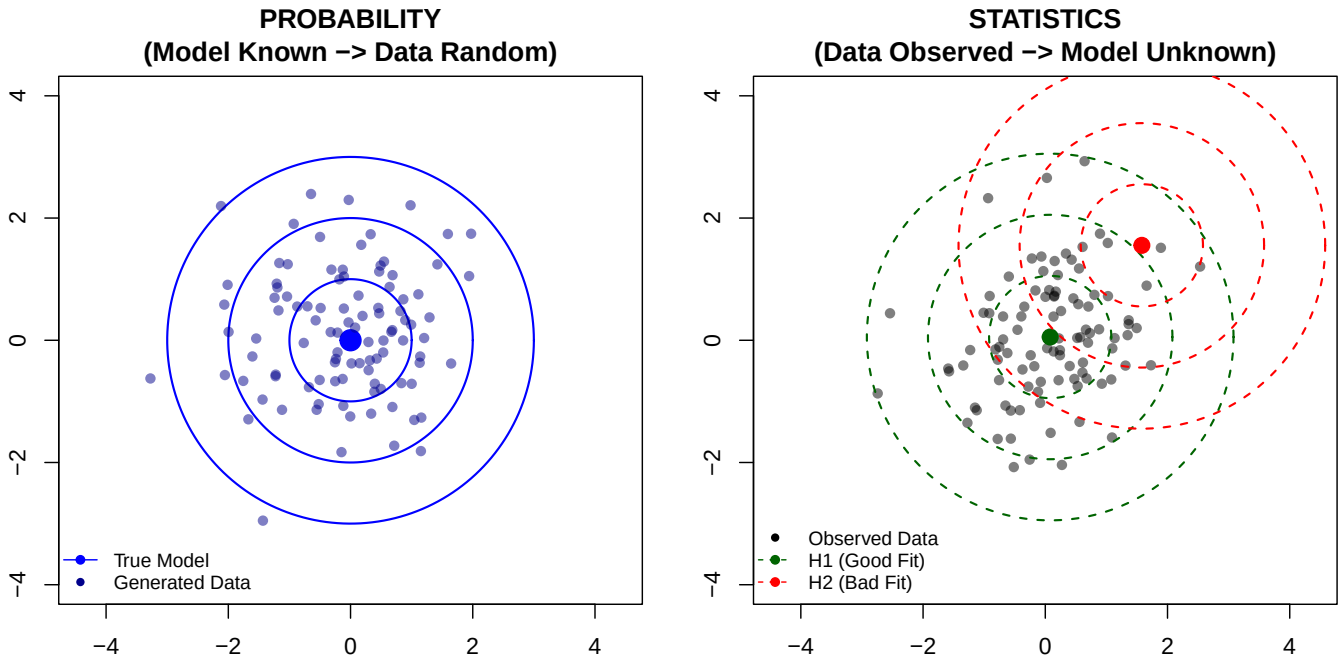

Figure 1.1: Probability vs Statistics. Left: Probability—The model is fixed (Blue center/contours), generating random data. Right: Statistics—Data is fixed (Black points); we test two hypothesized models: H1 (Green) centered at the sample mean (Good Fit) and H2 (Red) shifted by (1.5, 1.5) (Bad Fit).

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.

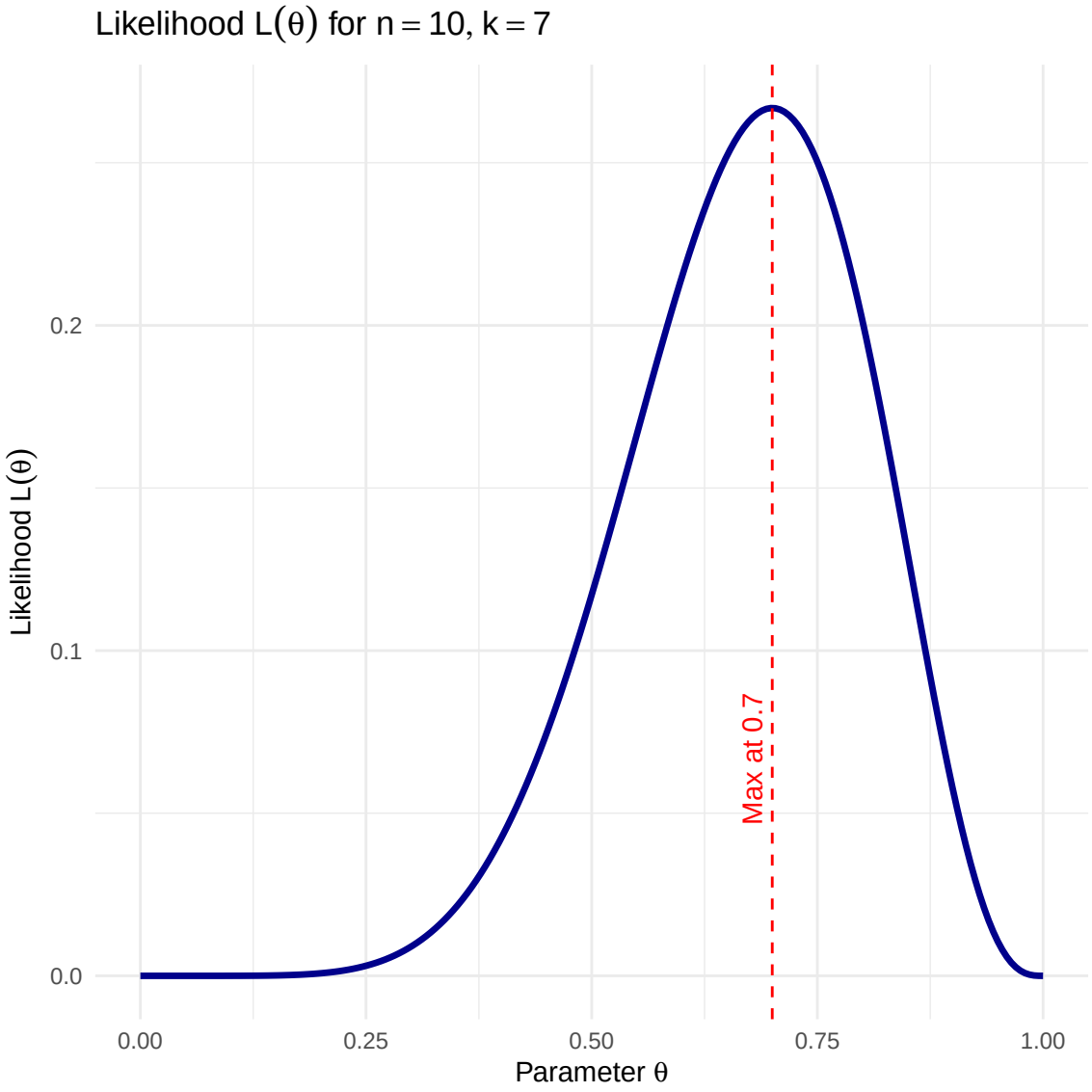

Figure 1.2: Likelihood Function (n= 10)

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

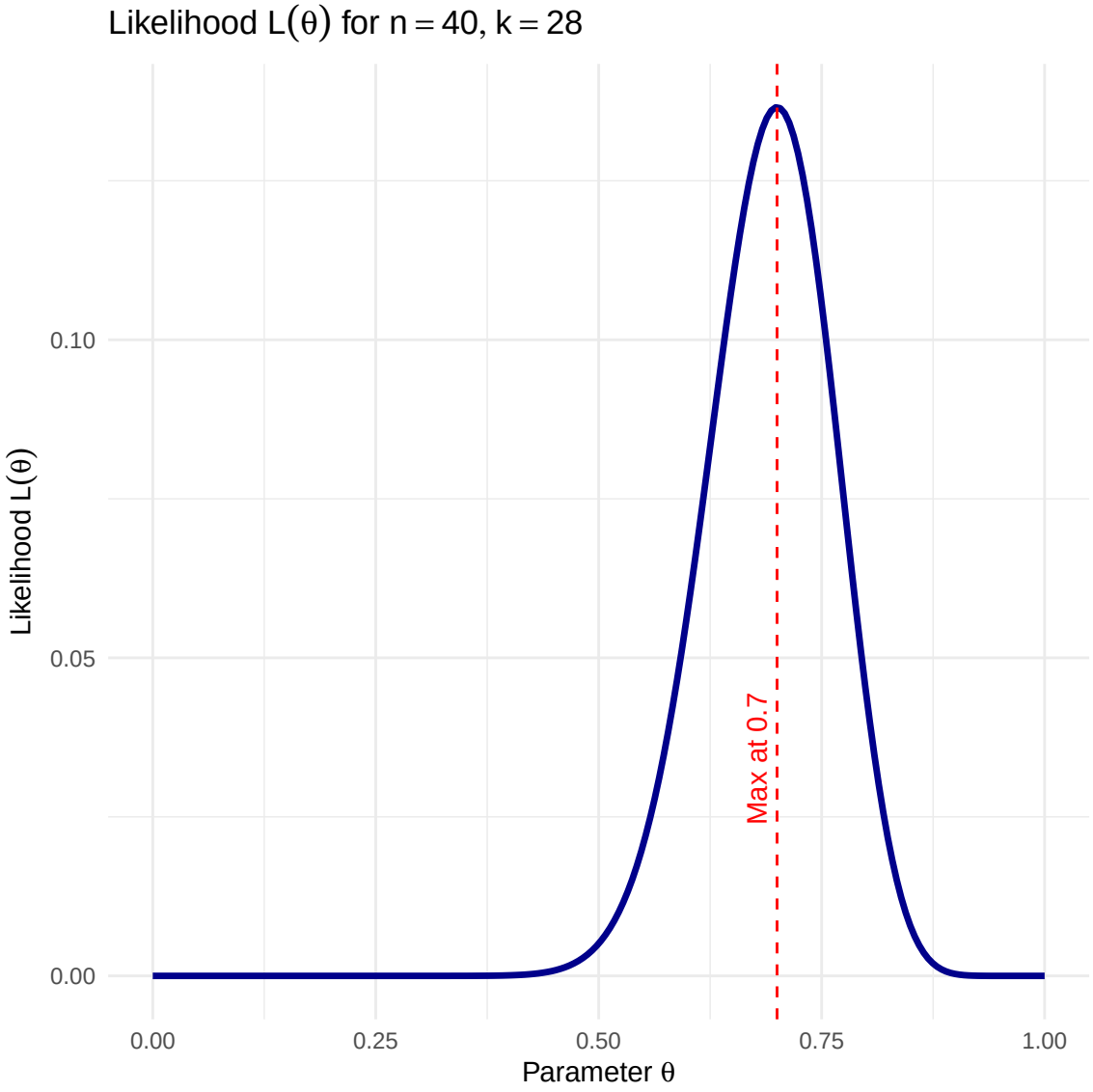

Figure 1.3: Likelihood Function (n= 40)

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?

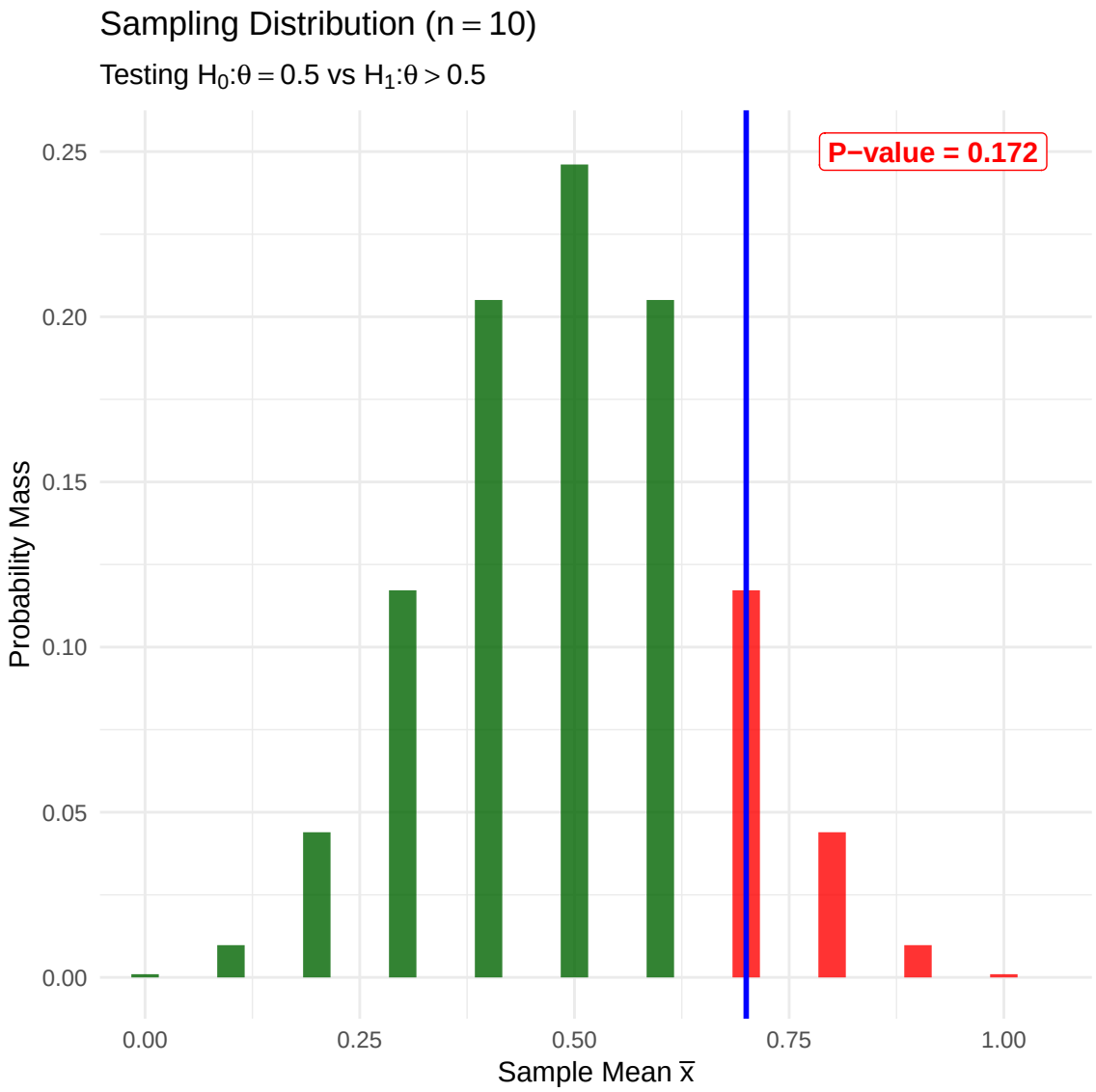

Figure 1.4: Sampling Distribution (n= 10)

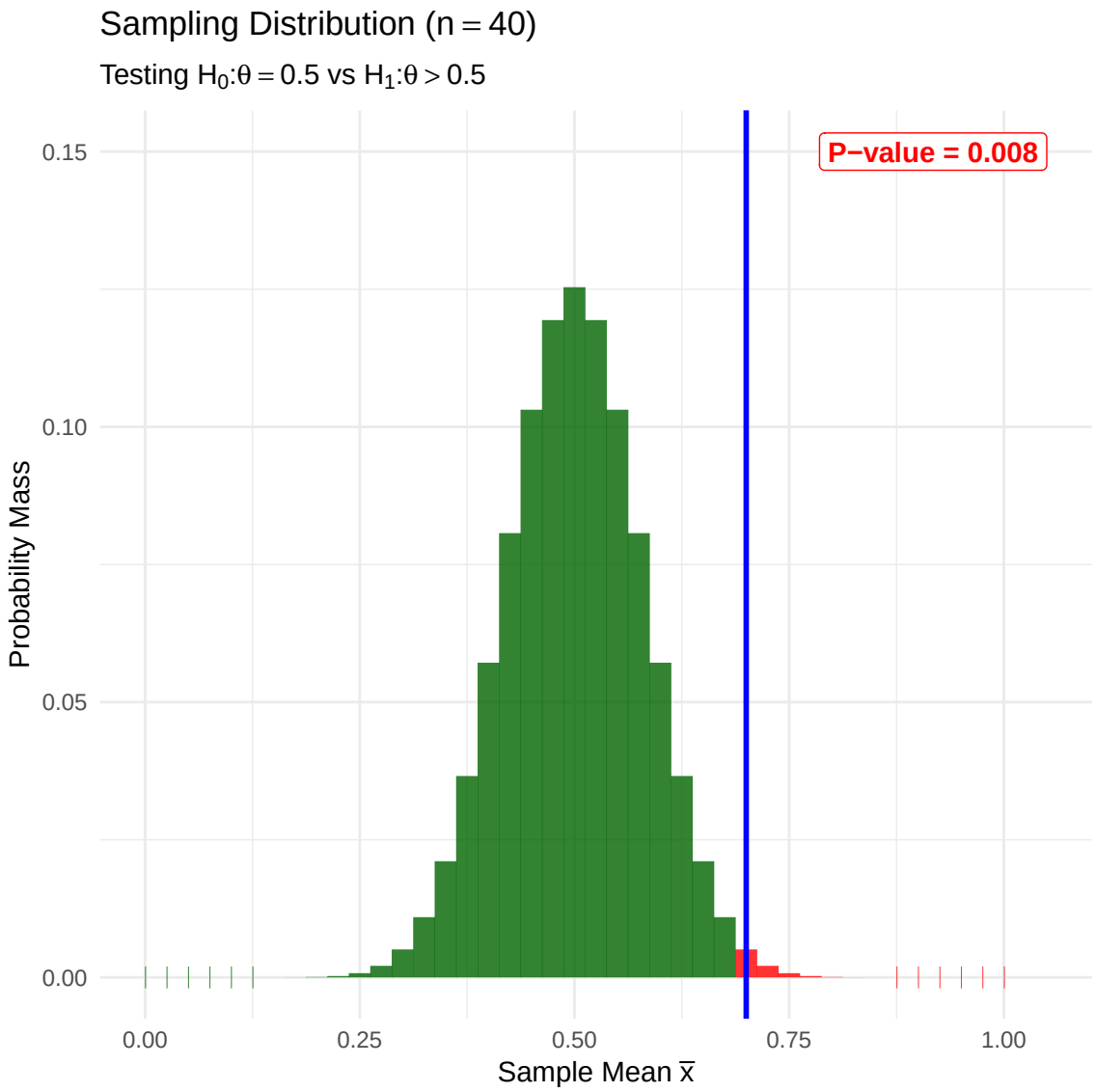

Figure 1.5: Sampling Distribution (n= 40)

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?

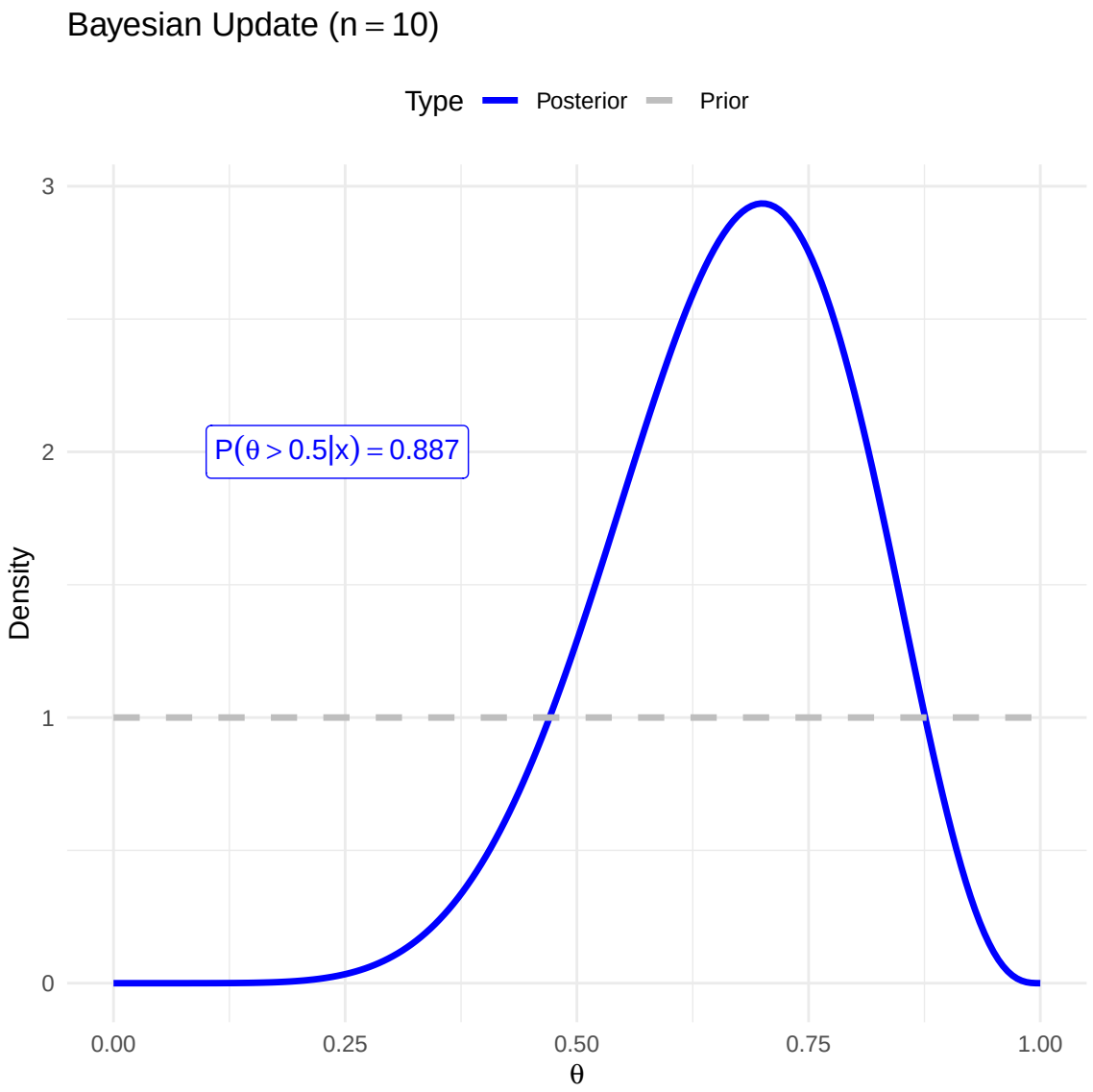

Figure 1.6: Bayesian Update (n= 10)

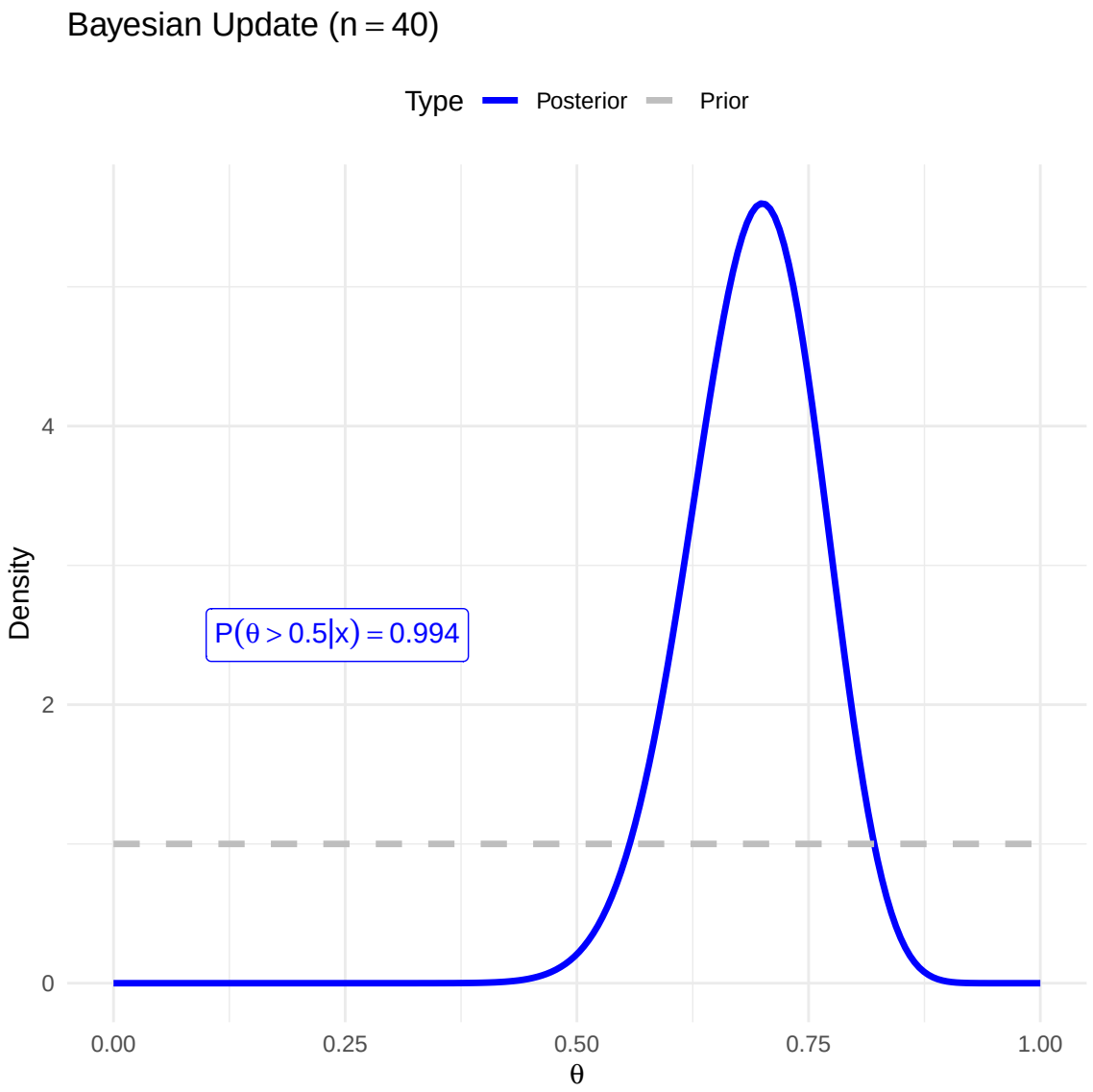

Figure 1.7: Bayesian Update (n= 40)

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).

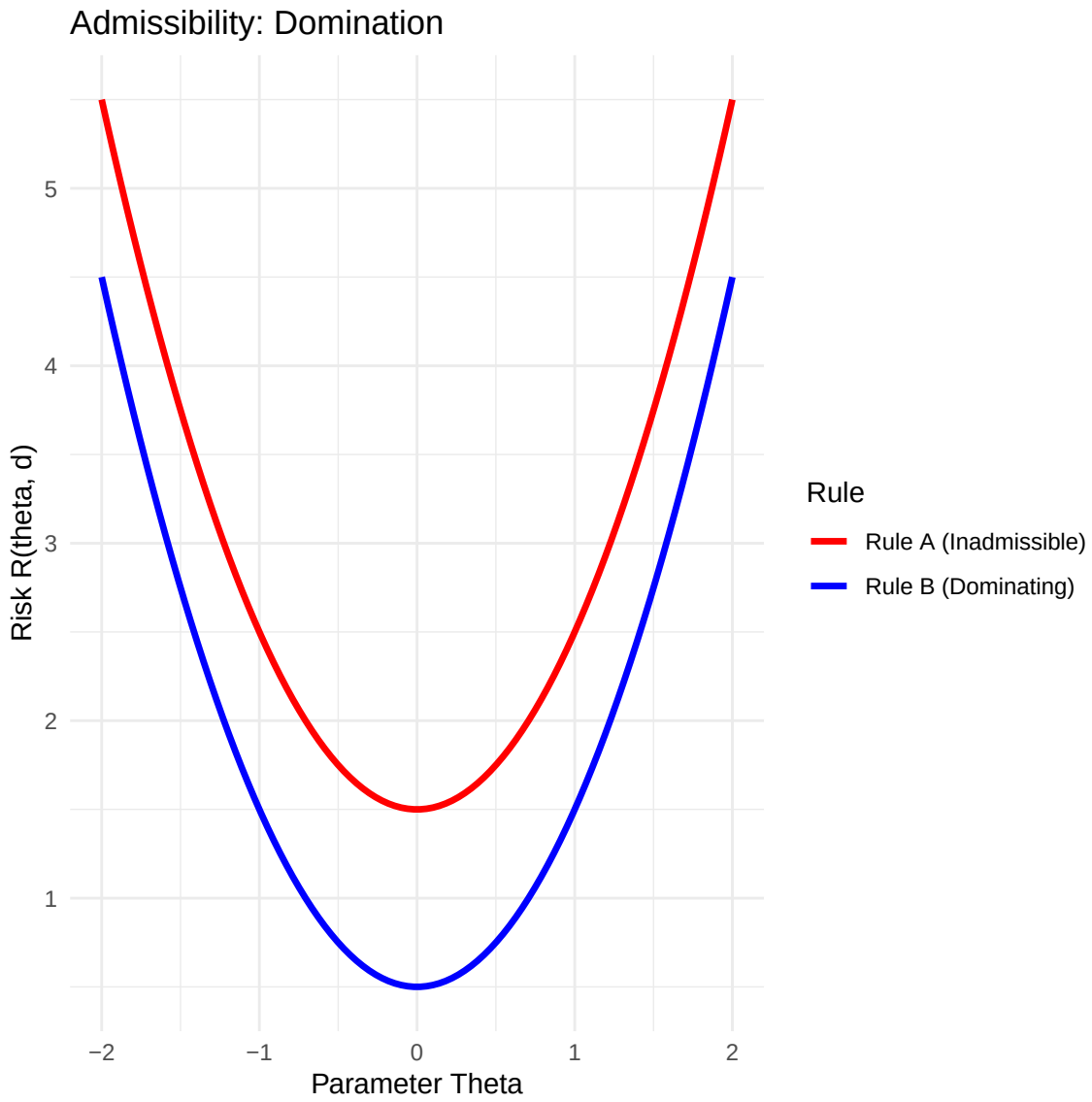

Figure 2.1: Illustration of Domination: Rule A (Red) is inadmissible because Rule B (Blue) has lower risk for all values of theta.

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.

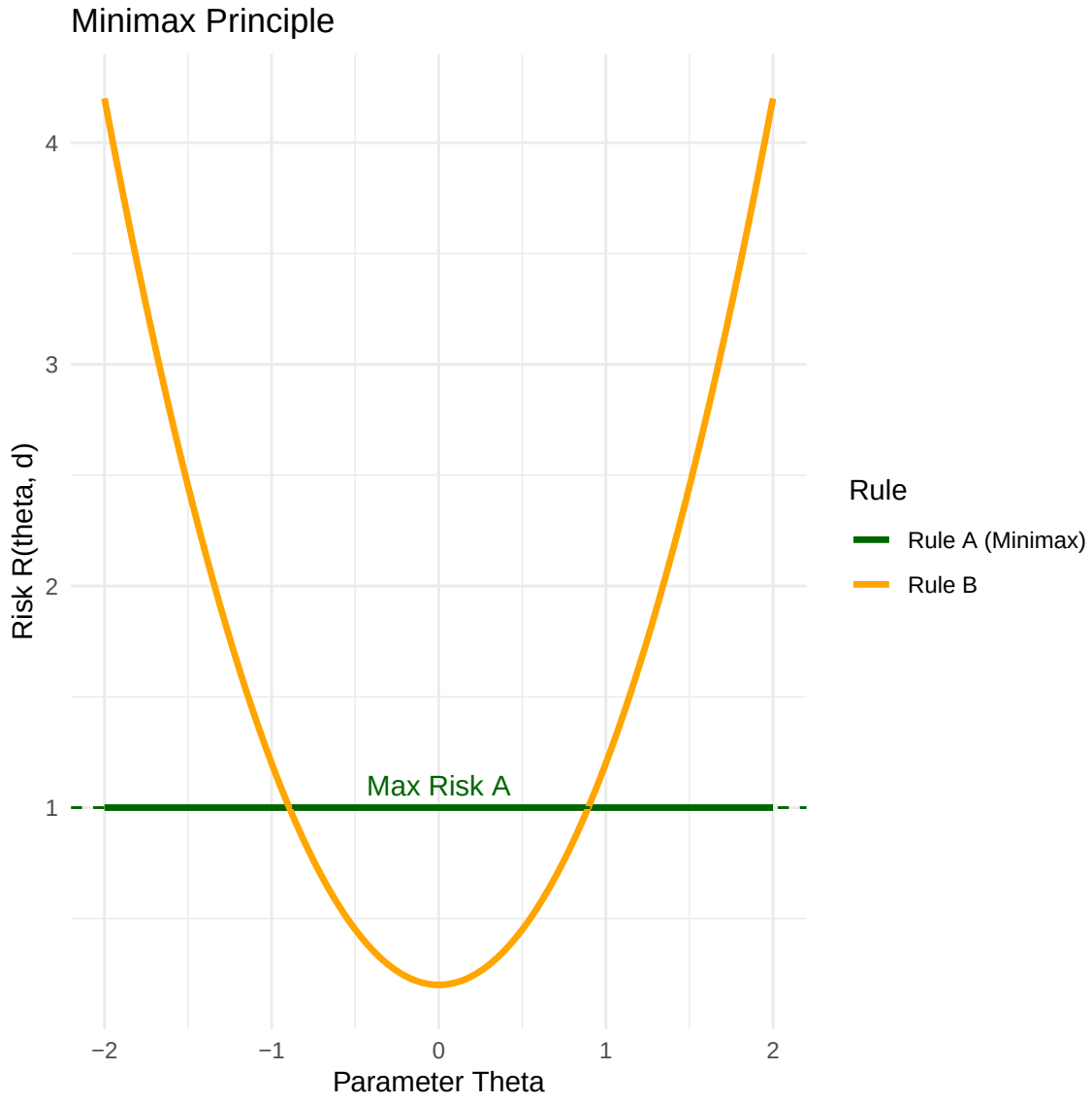

Figure 2.2: Illustration of Minimax: Rule A has a lower peak risk than Rule B, making Rule A the Minimax choice.

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 $-\pi_1/\pi_2$.

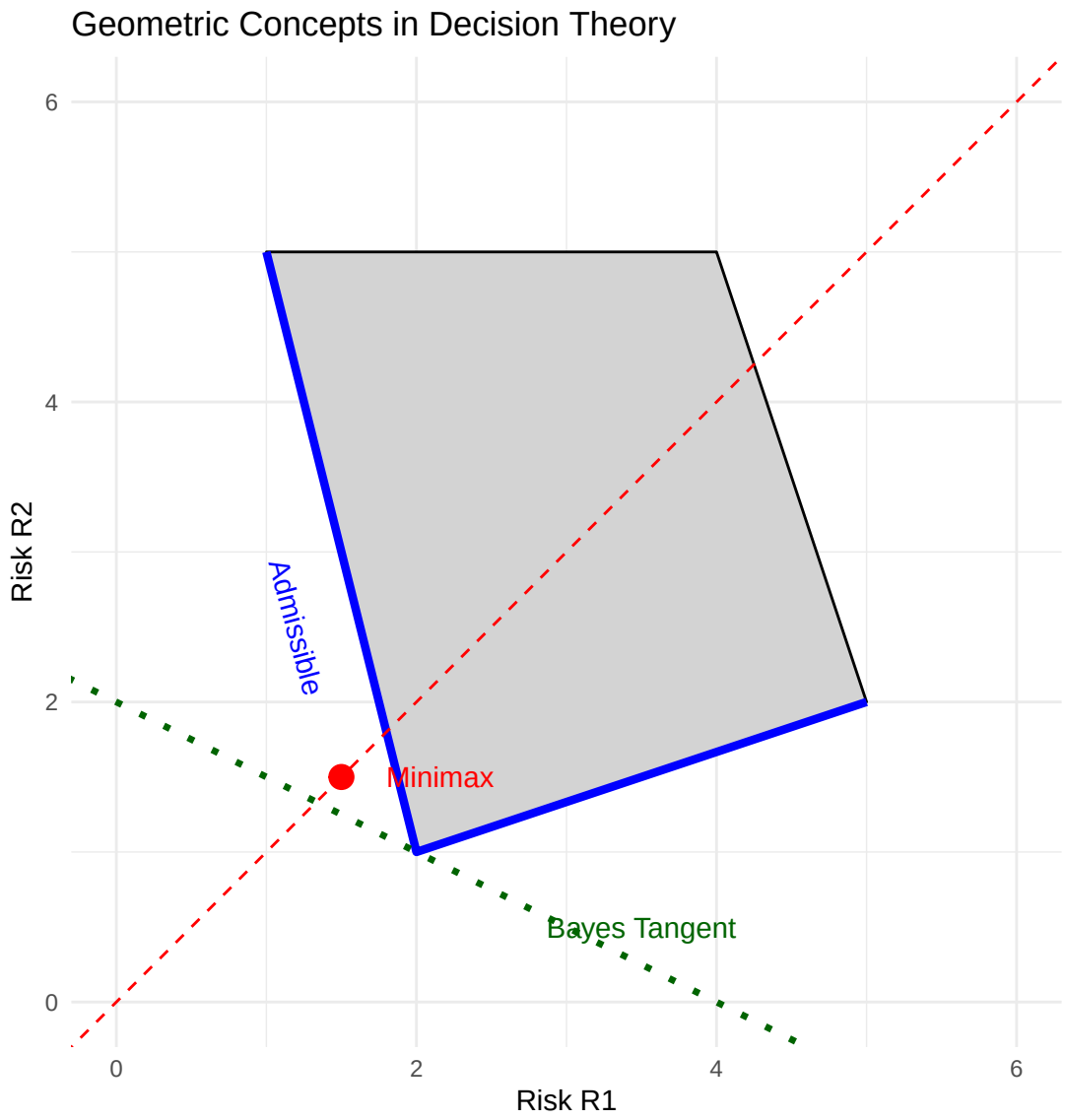

Figure 2.3: Geometric Interpretation: The gray polygon is the Risk Set S. The blue boundary represents admissible rules. The red point is the Minimax rule. The green line represents a Bayes rule for a specific prior.

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:

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

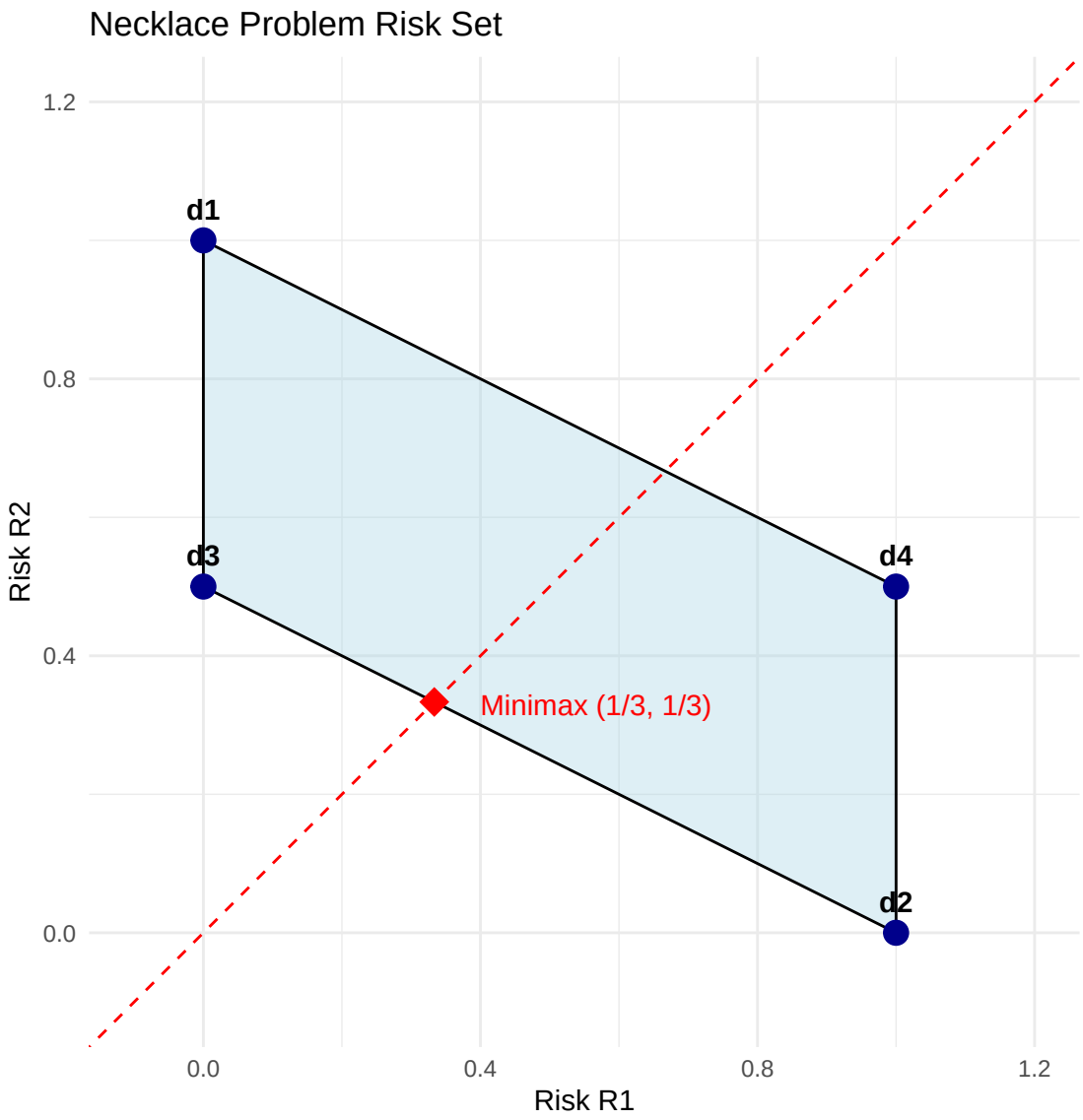

Figure 2.4: Necklace Problem Solution. The Minimax rule (red diamond) is the specific randomized combination of d3 and d2 that equalizes the risk.

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 Inference

3.1 Bayes Theorem

Suppose $\theta \sim \pi(\theta)$ and $X \sim f(x;\theta)$. The posterior density of θ given X is:

$$\pi(\theta|x) = \frac{\pi(\theta)f(x;\theta)}{\int_\Theta \pi(\theta)f(x;\theta)d\theta} \propto \pi(\theta) \cdot L(\theta;x) \tag{3.1}$$

where $L(\theta;x)$ is the likelihood.

3.2 Examples

3.2.1 1. Binomial-Beta

- $X|\theta \sim \text{Bin}(n,\theta) \Rightarrow f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$
- Prior $\theta \sim \text{Beta}(a,b) \Rightarrow \pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$

Posterior:

$$\pi(\theta|x) \propto \theta^{a-1}(1-\theta)^{b-1} \cdot \theta^x(1-\theta)^{n-x} = \theta^{a+x-1}(1-\theta)^{b+n-x-1} \tag{3.2}$$

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

Moments:

- Mean: $E(\theta|x) = \frac{a+x}{a+b+n} \approx \frac{x}{n}$ (for small n)
- Variance: $\text{Var}(\theta|x) = \frac{(a+x)(b+n-x)}{(a+b+n)^2(a+b+n+1)}$

3.2.2 2. Normal-Normal (Known Variance)

- $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where σ^2 is known.
- Prior $\mu \sim N(\mu_0, \sigma_0^2)$.

Let $\tau_0 = 1/\sigma_0^2$ (prior precision), $\tau = 1/\sigma^2$ (data precision). The posterior precision is $\tau_1 = \tau_0 + n\tau$.

Posterior:

$$\mu|x \sim N\left(\frac{\tau_0\mu_0 + n\tau\bar{x}}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right) \tag{3.3}$$

This shows the posterior mean is a weighted average of the prior mean and the sample mean.

---

4 Part 3: Bayes Estimators and Loss Functions

To find a Bayes rule $d(x)$, we minimize the posterior expected loss:

$$\min_d \int_\Theta L(\theta, d)\pi(\theta|x)d\theta \tag{4.1}$$

4.1 1. Squared Error Loss: $L(\theta, a) = (\theta - a)^2$

Minimizing $E_{\theta|x}[(\theta - d)^2]$ leads to:

$$d(x) = E(\theta|x) \quad \text{(Posterior Mean)} \tag{4.2}$$

4.2 2. Absolute Error Loss: $L(\theta, a) = |\theta - a|$

Minimizing $E_{\theta|x}[|\theta - d|]$ leads to:

$$\int_{-\infty}^{d} \pi(\theta|x)d\theta = \int_{d}^{\infty} \pi(\theta|x)d\theta = 0.5 \tag{4.3}$$

So, $d(x) =$ Median of $\pi(\theta|x)$.

4.3 3. 0-1 Loss (Hypothesis Testing)

- Loss is 1 if error, 0 if correct.
- Testing Θ_0 vs Θ_1.
- Bayes Rule: Choose class with highest posterior probability.
  - Reject H_0 if $P(\theta \in \Theta_1|x) > P(\theta \in \Theta_0|x)$.

4.4 4. Interval Estimation

We want an interval $A = (d - \delta, d + \delta)$ minimizing risk (maximizing coverage probability $1 - \alpha$).

Highest Posterior Density (HPD) Interval: The set $C = \{\theta : \pi(\theta|x) \geq k\}$ where $P(\theta \in C|x) = 1 - \alpha$. This is the shortest interval for a given confidence level if the posterior is unimodal.

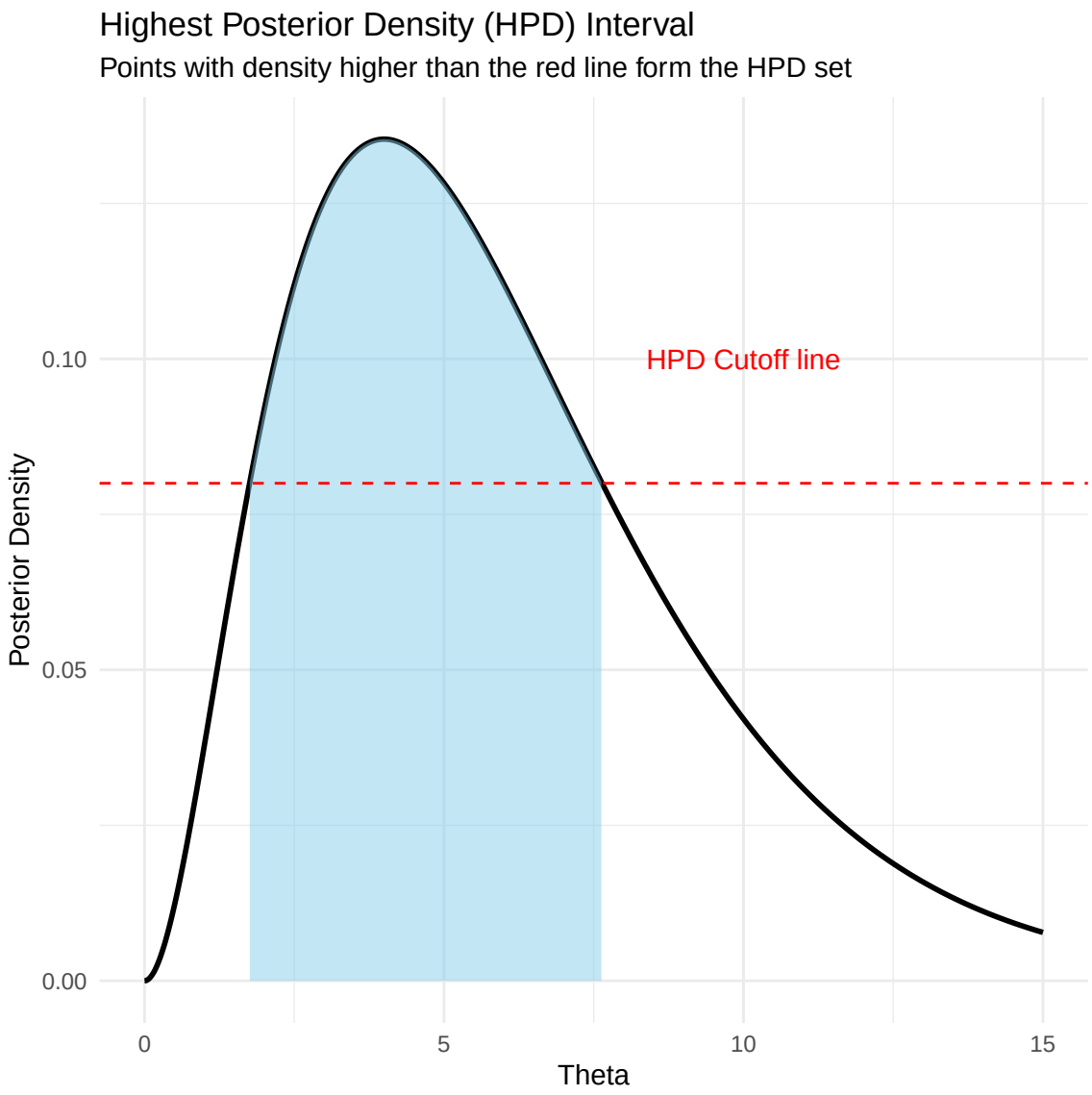

Figure 4.1: Illustration of Highest Posterior Density (HPD) Interval vs Equi-tailed Interval on a skewed posterior.

5 Part 4: Minimax Rules via Bayes

Goal: Find a minimax estimator for θ where $X \sim \text{Bin}(n, \theta)$. **Loss:** Squared Error $L(\theta, d) = (\theta - d)^2$.

Strategy: Find a prior Beta(a, b) such that the Bayes risk $R(\theta, d_{\text{Bayes}})$ is constant for all θ. By Theorem 2.2 (Equalizer Rule), if an extended Bayes rule has constant risk, it is Minimax.

The Bayes estimator is $d(x) = \frac{a+x}{a+b+n}$. The risk is:

$$R(\theta, d) = E\left[\left(\theta - \frac{a + X}{a + b + n}\right)^2\right] \tag{5.1}$$

Let $c = a + b + n$.

$$R(\theta, d) = \frac{1}{c^2}\left[(c\theta - a)^2 - 2(c\theta - a)n\theta + n\theta(1 - \theta) + n^2\theta^2\right] \tag{5.2}$$

To make this constant (independent of θ), the coefficients of θ and θ^2 must vanish or balance out. Solving the resulting system yields:

$$a = b = \frac{\sqrt{n}}{2} \tag{5.3}$$

Minimax Estimator:

$$d_{\text{minimax}}(x) = \frac{x + \sqrt{n}/2}{n + \sqrt{n}} \tag{5.4}$$

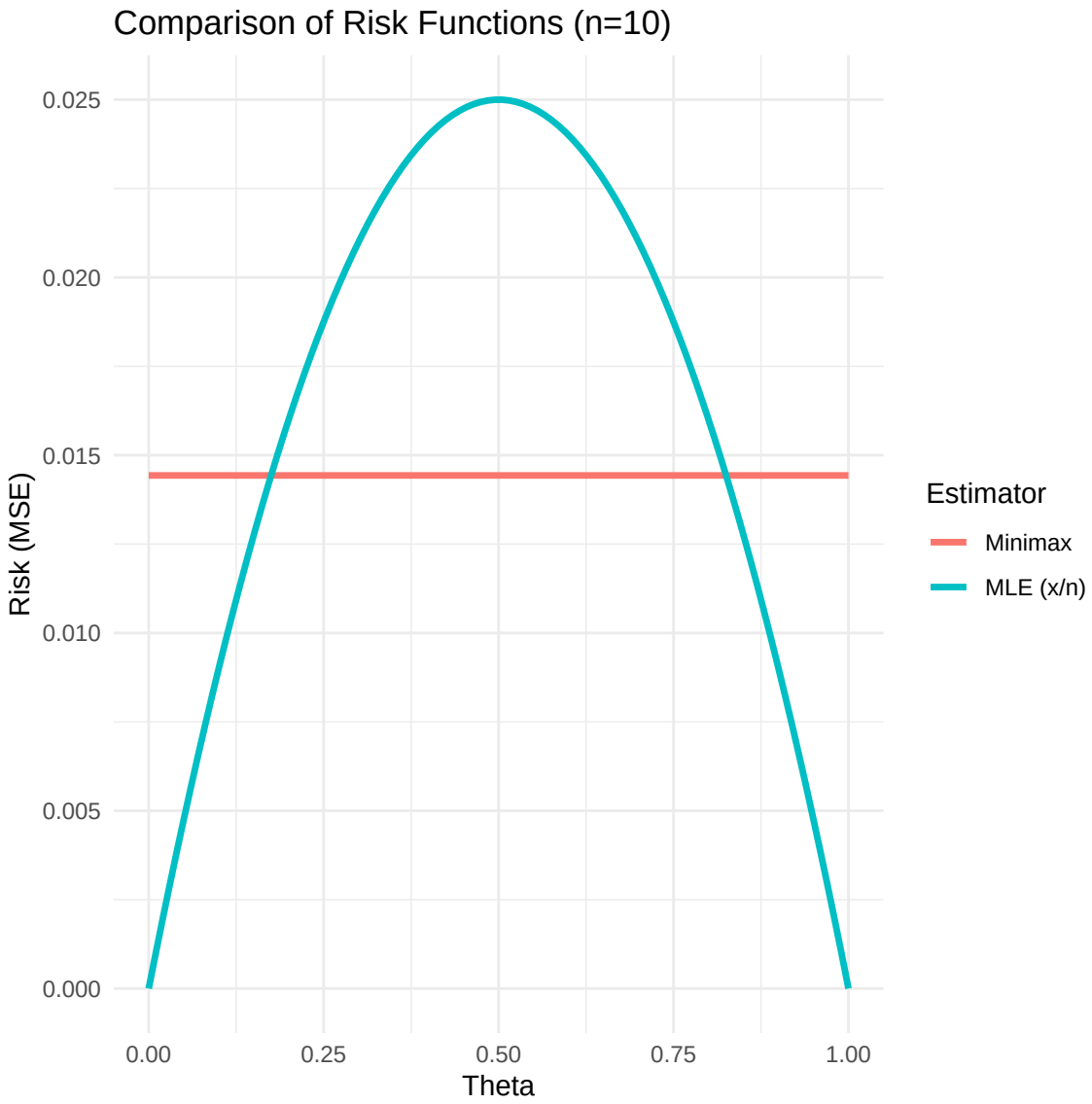

Figure 5.1: Risk Functions: MLE vs Minimax for Binomial(n=10).

6 Part 5: Stein Estimation

Context: Estimating a multivariate normal mean $\mu = (\mu_1, \dots, \mu_p)^T$ where $X \sim N_p(\mu, I)$. **Loss:** Sum of squared errors $L(\mu, d) = ||\mu - d||^2$.

Stein's Lemma: If $Y \sim N(\mu, 1)$ and $h(y)$ is differentiable:

$$E[(Y - \mu)h(Y)] = E[h'(Y)] \tag{6.1}$$

James-Stein Estimator:

$$d^{JS}(X) = \left(1 - \frac{p-2}{||X||^2}\right) X \tag{6.2}$$

This estimator shrinks the observation vector X towards the origin (or a grand mean).

Result: If $p \geq 3$, the James-Stein estimator dominates the MLE ($d^0(X) = X$).

$$R(\mu, d^{JS}) < R(\mu, d^0) = p \quad \text{for all } \mu \tag{6.3}$$

6.1 Baseball Example (Efron & Morris)

We observe batting averages for $p = 18$ players.

- MLE: Individual batting averages.
- JS: Shrinks individual averages toward the global average.

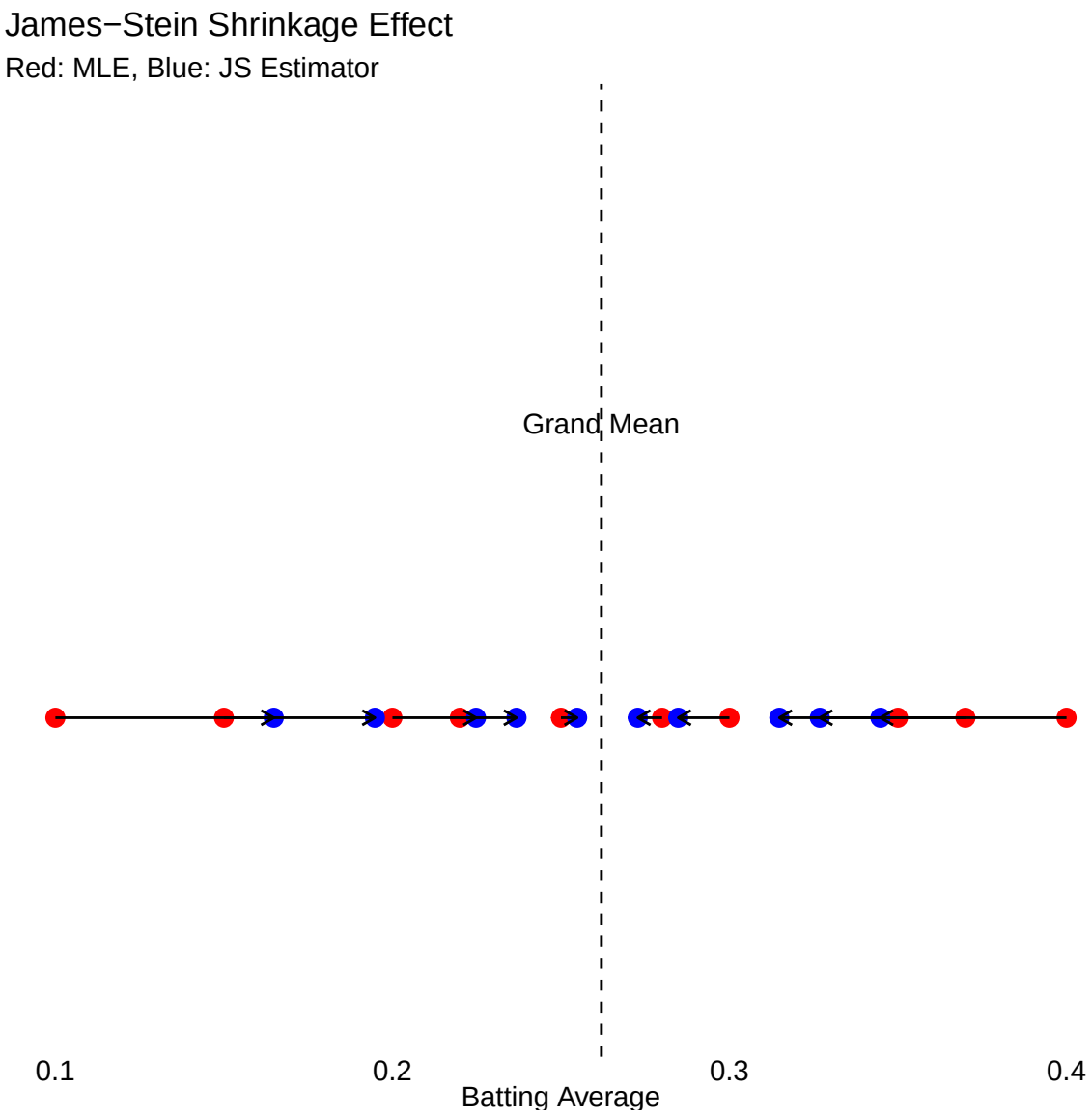

Figure 6.1: Visualizing James-Stein Shrinkage (Mock Data based on Baseball Example). The arrows show MLEs being pulled toward the Grand Mean.

7 Part 6: Empirical Bayes & Hierarchical Models

7.1 Empirical Bayes

Instead of fixing hyperparameters (μ_0, σ_0^2), we estimate them from the marginal distribution of the data.

$$m(x) = \int f(x|\theta)\pi(\theta|\eta)d\theta \tag{7.1}$$

We estimate η by maximizing $m(x)$ (Type-II MLE) or method of moments.

Example: If $X_i \sim N(\mu_i, 1)$ and $\mu_i \sim N(0, \tau^2)$, then marginally $X_i \sim N(0, 1 + \tau^2)$. We can use $S = \sum X_i^2$ to estimate τ^2.

7.2 Hierarchical Models

We assume a multistage structure:

1. Data model: $X|\theta \sim f(x|\theta)$
2. Parameter model: $\theta|\lambda \sim \pi(\theta|\lambda)$
3. Hyperparameter model: $\lambda \sim h(\lambda)$

Computation: Since analytical solutions are often impossible, we use **Markov Chain Monte Carlo (MCMC)**.

7.2.1 Gibbs Sampling

To sample from the joint posterior $f(\theta, \lambda|x)$, we sample iteratively from the full conditional distributions:

1. Draw $\theta^{(k+1)} \sim f(\theta|\lambda^{(k)}, x)$
2. Draw $\lambda^{(k+1)} \sim f(\lambda|\theta^{(k+1)}, x)$

7.2.2 Metropolis-Hastings

If a conditional distribution is hard to sample from directly:

1. Propose θ^* from a proposal density $q(\theta^*|\theta^{(t)})$.
2. Calculate acceptance ratio $\alpha = \min\left(1, \frac{f(\theta^*|x)q(\theta^{(t)}|\theta^*)}{f(\theta^{(t)}|x)q(\theta^*|\theta^{(t)})}\right)$.
3. Accept θ^* with probability α.

---

8 Part 7: Predictive Distributions

We want to predict a new observation Y^*.

$$f(y^*|y) = \int f(y^*|\theta)\pi(\theta|y)d\theta \tag{8.1}$$

Numerical Methods: If we have posterior samples $\theta^{(1)}, \ldots, \theta^{(N)}$ from MCMC:

Method 1: Density Averaging

$$\hat{f}(y^*|y) \approx \frac{1}{N}\sum_{i=1}^{N} f(y^*|\theta^{(i)}) \tag{8.2}$$

This is a Rao-Blackwellized estimator and usually has lower variance.

Method 2: Direct Sampling For each $\theta^{(i)}$, draw $Y^{*(i)} \sim f(y^*|\theta^{(i)})$. The histogram of $Y^{*(i)}$ approximates the predictive density.