Statistical Inference

Longhai Li

2026-02-06

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.

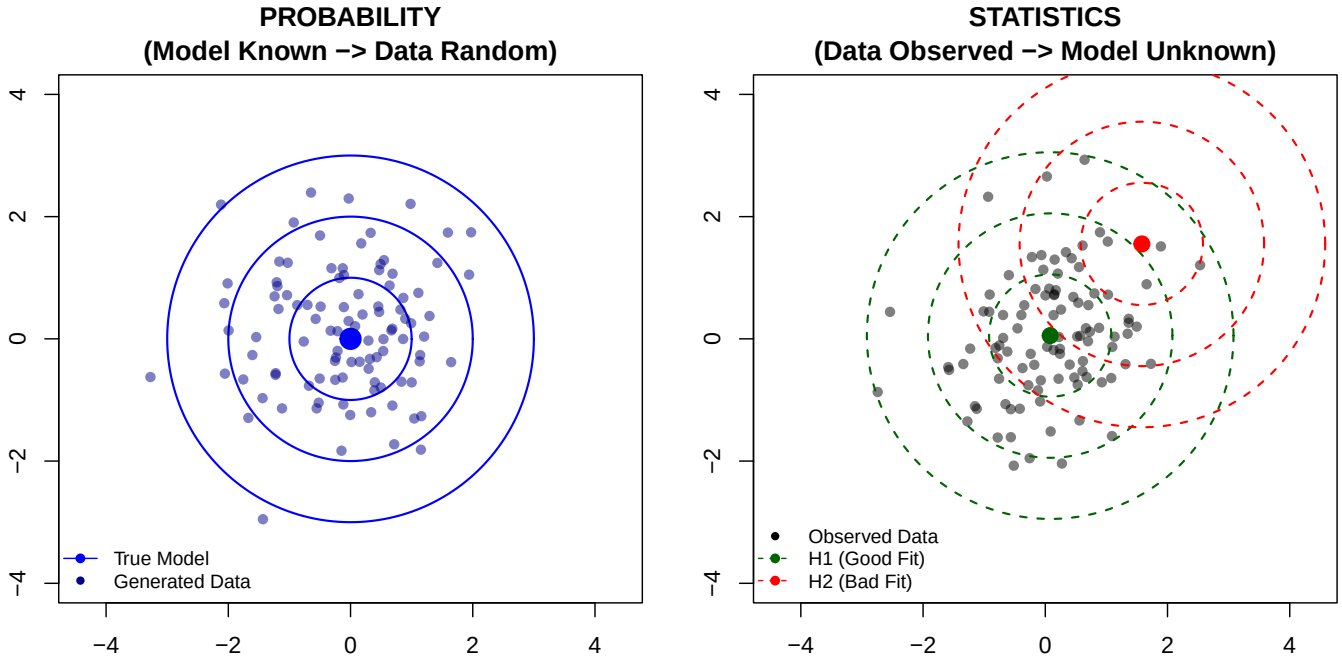


Figure 1.1: Probability vs Statistics. Left: Probability—The model is fixed (Blue center/contours), generating random data. Right: Statistics—Data is fixed (Black points); we test two hypothesized models: H1 (Green) centered at the sample mean (Good Fit) and H2 (Red) shifted by (1.5, 1.5) (Bad Fit).

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: $\hat{\theta} = \theta(X_1, \dots, X_n)$.

- *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 $f(x_1, \dots, x_n; \theta)$ 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 $x_1, \dots, x_n$, we call it the **likelihood function**, denoted $L(\theta)$.

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

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

1.5.2 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.

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)

θ	Calculation $\binom{40}{28}\theta^{28}(1-\theta)^{12}$	$L(\theta)$
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

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.

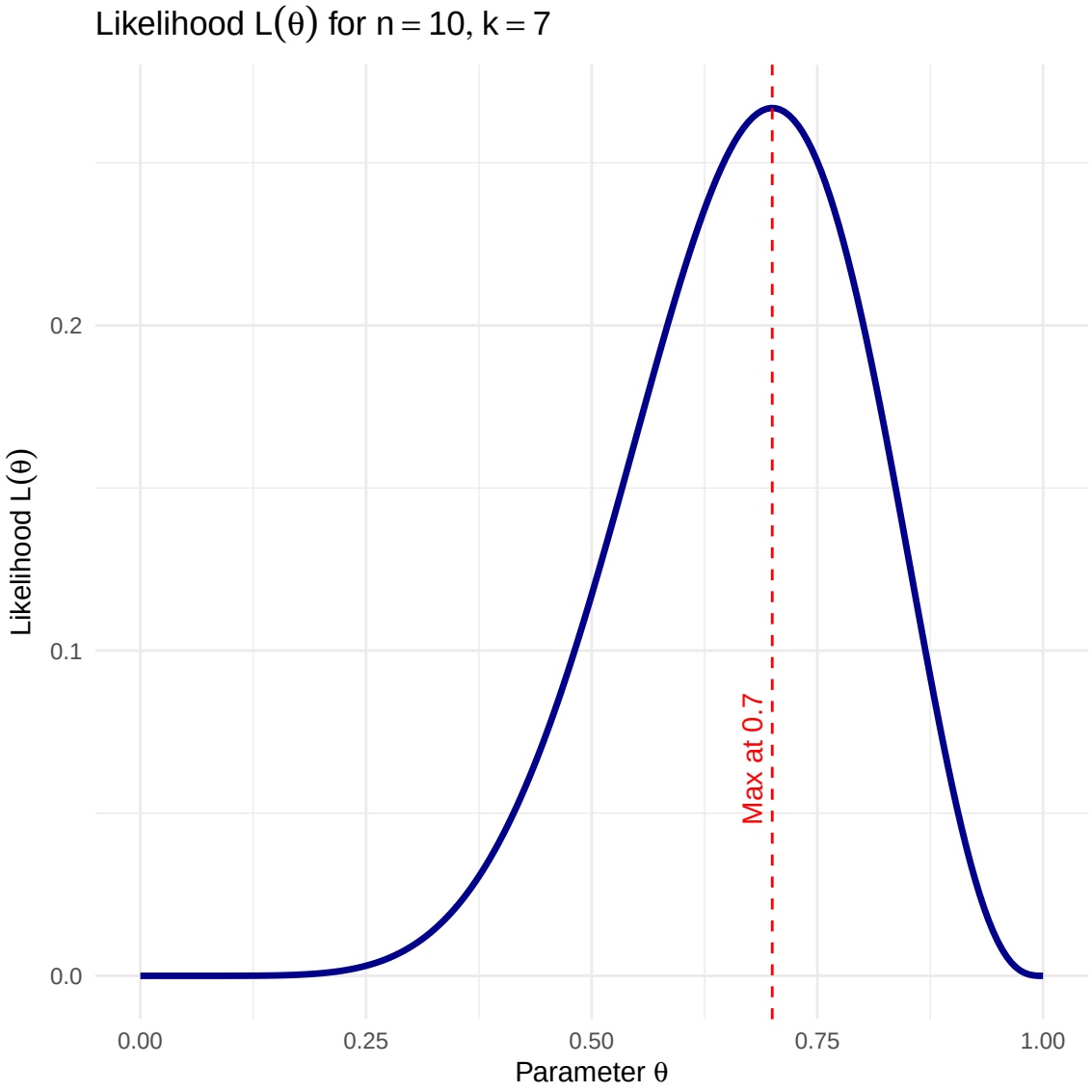


Figure 1.2: Likelihood Function (n= 10)

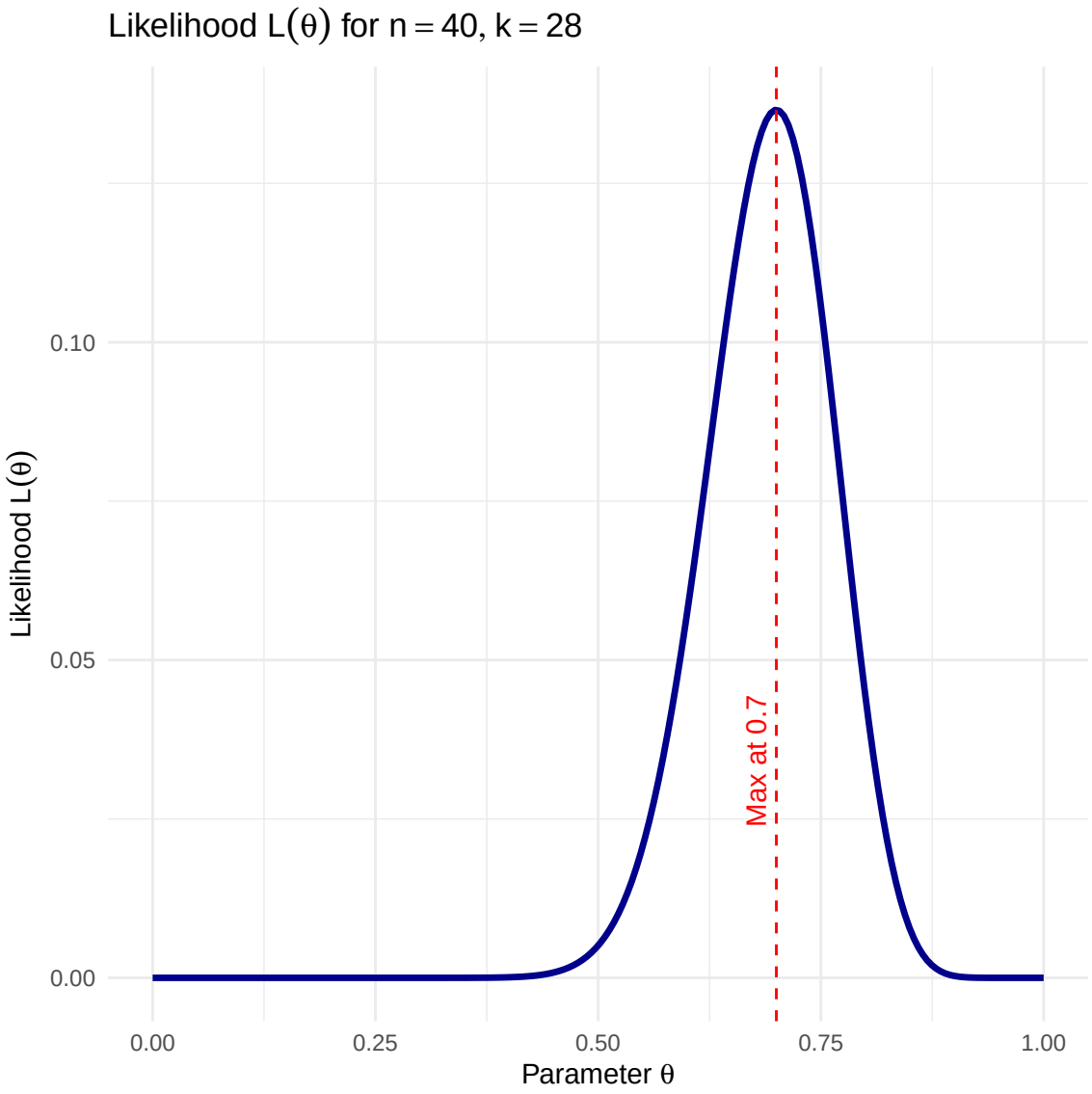


Figure 1.3: Likelihood Function (n= 40)

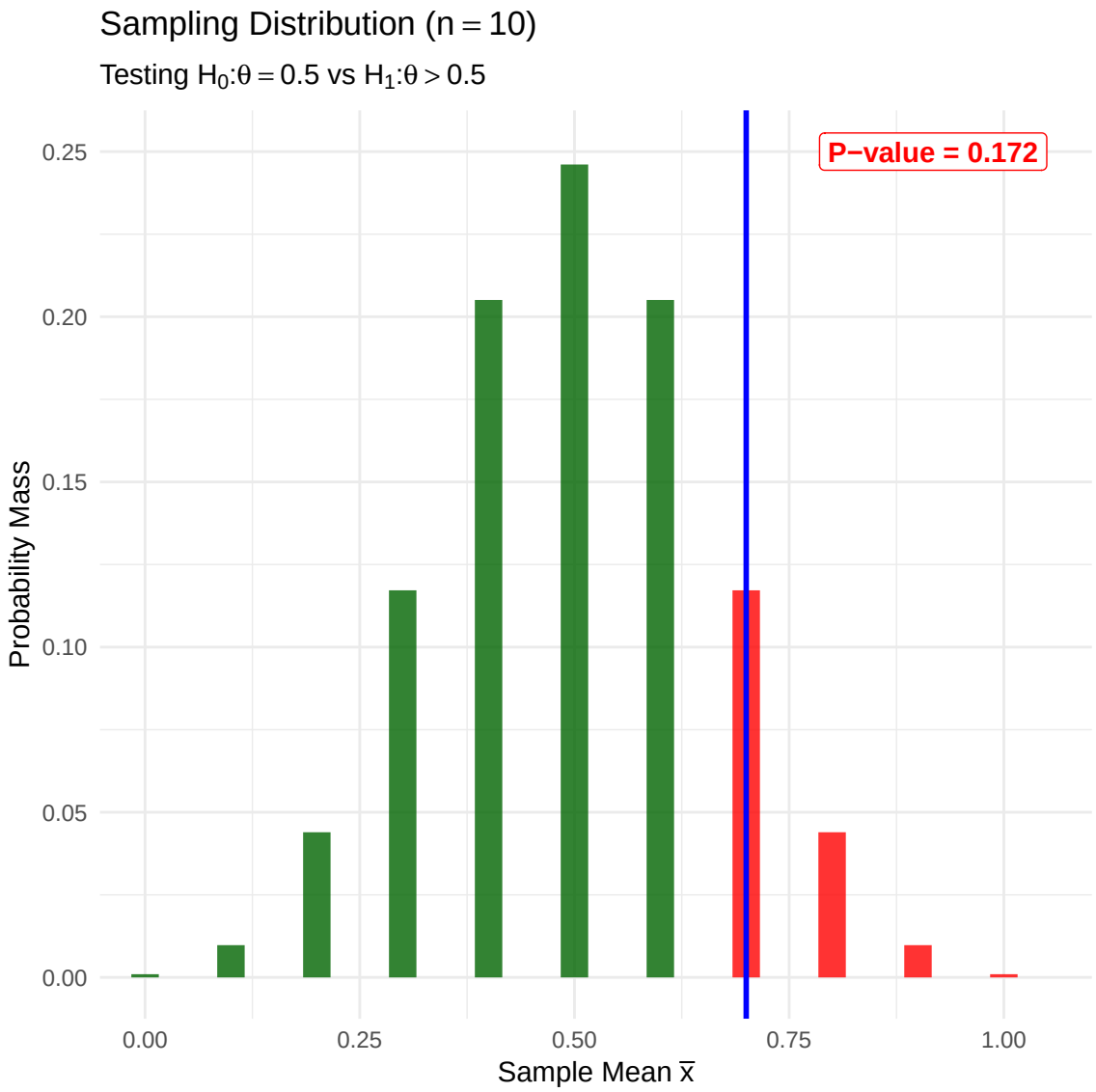


Figure 1.4: Sampling Distribution (n= 10)

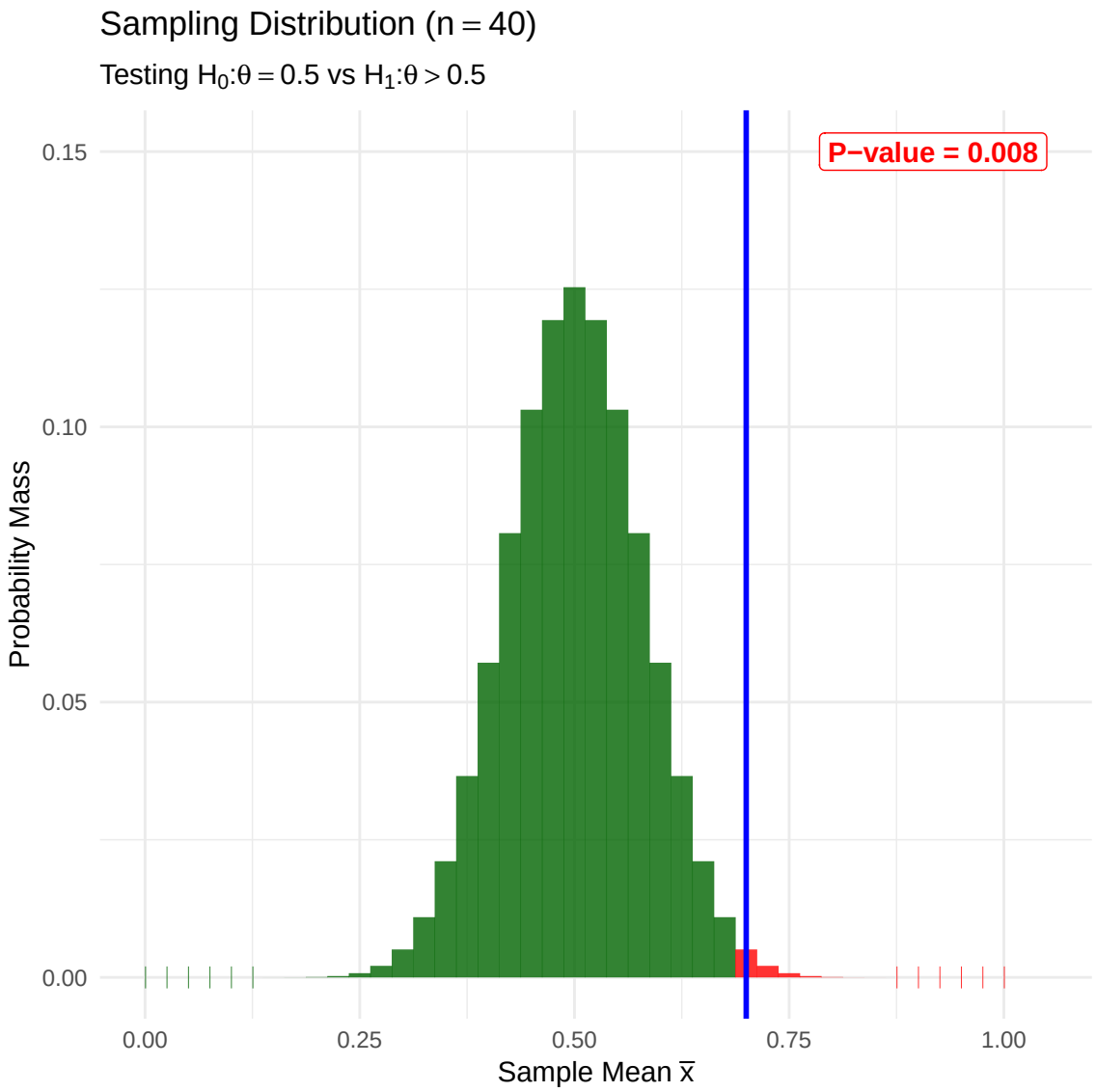


Figure 1.5: Sampling Distribution (n= 40)

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?

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?

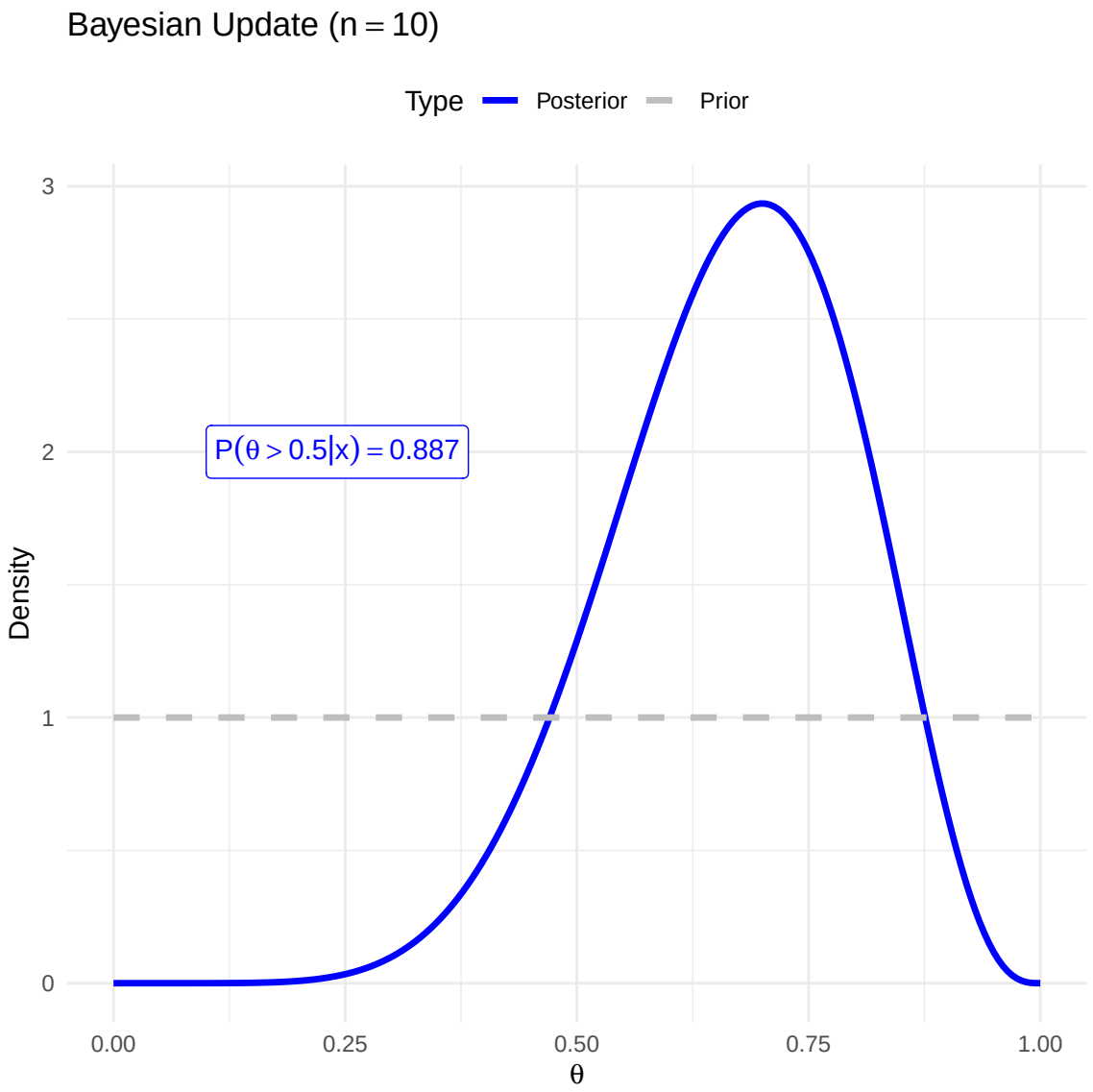


Figure 1.6: Bayesian Update (n= 10)

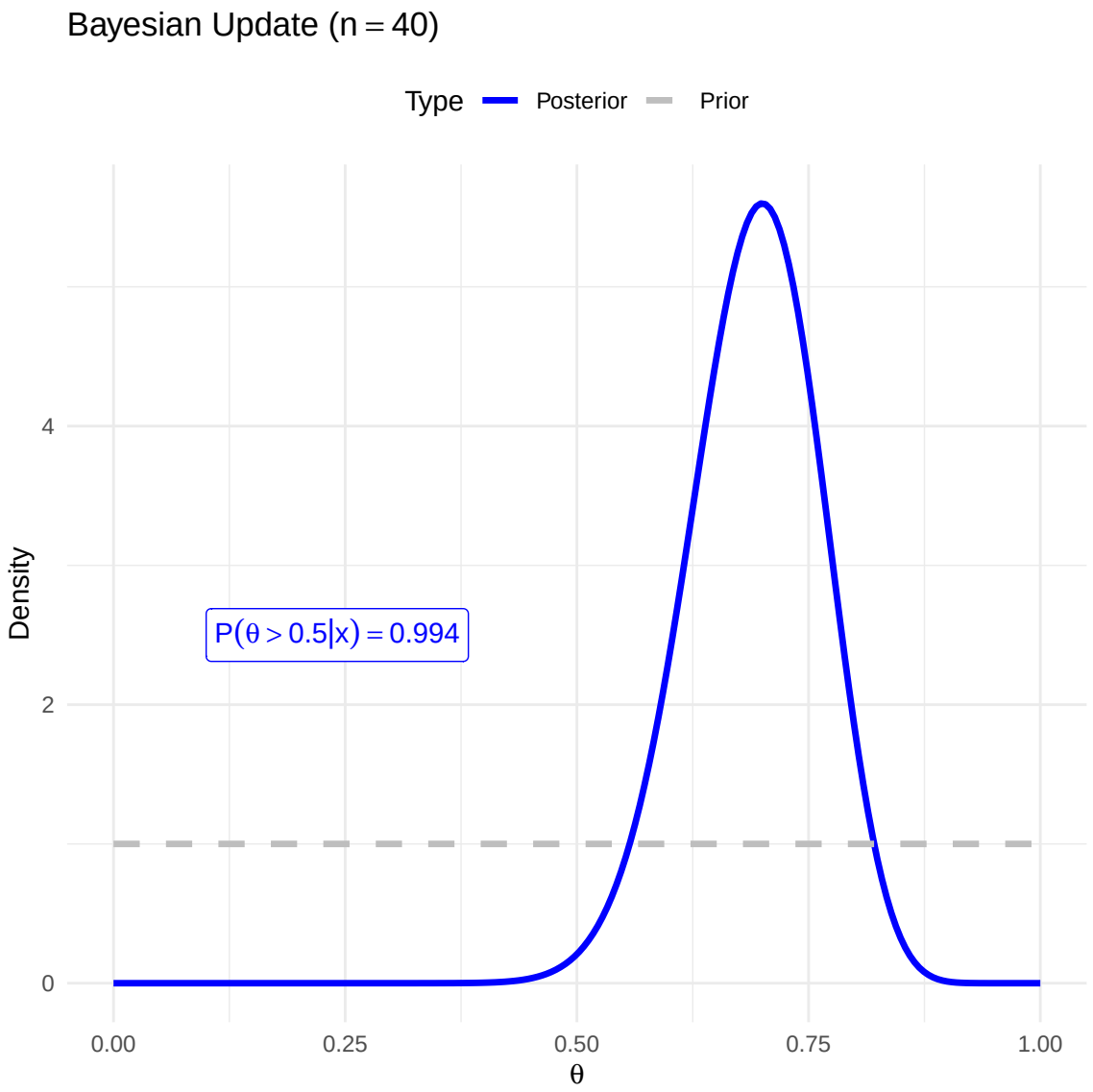


Figure 1.7: Bayesian Update (n= 40)

- **Model Comparison and Assessment:** How to assess a Bayesian model?

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

2.4 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 ($50/50$).

2.4.1 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.

2.4.2 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	$R_2 = 0.5$

2.5 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.

2.5.1 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).

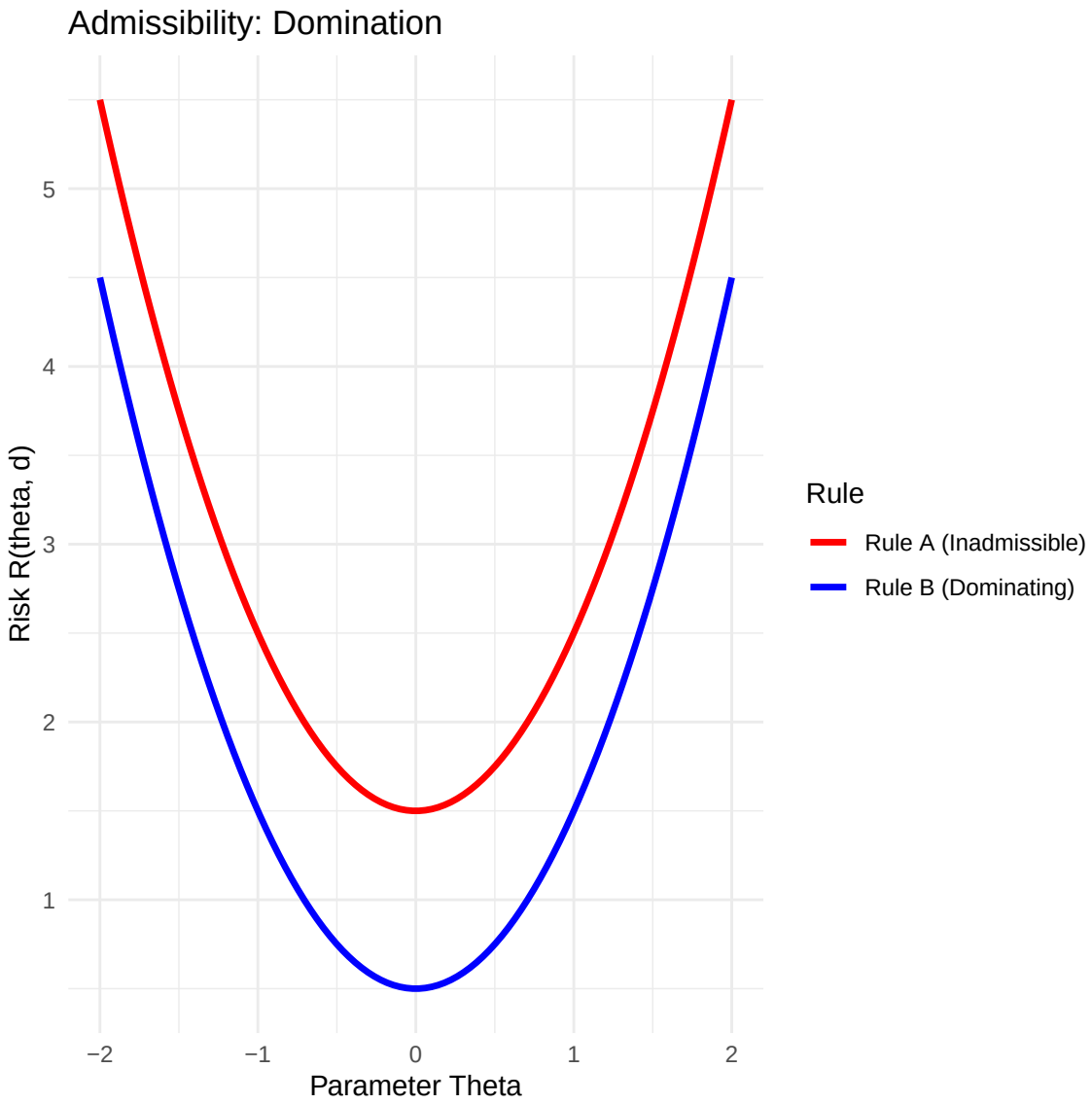


Figure 2.1: Illustration of Domination: Rule A (Red) is inadmissible because Rule B (Blue) has lower risk for all values of theta.

2.5.2 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.

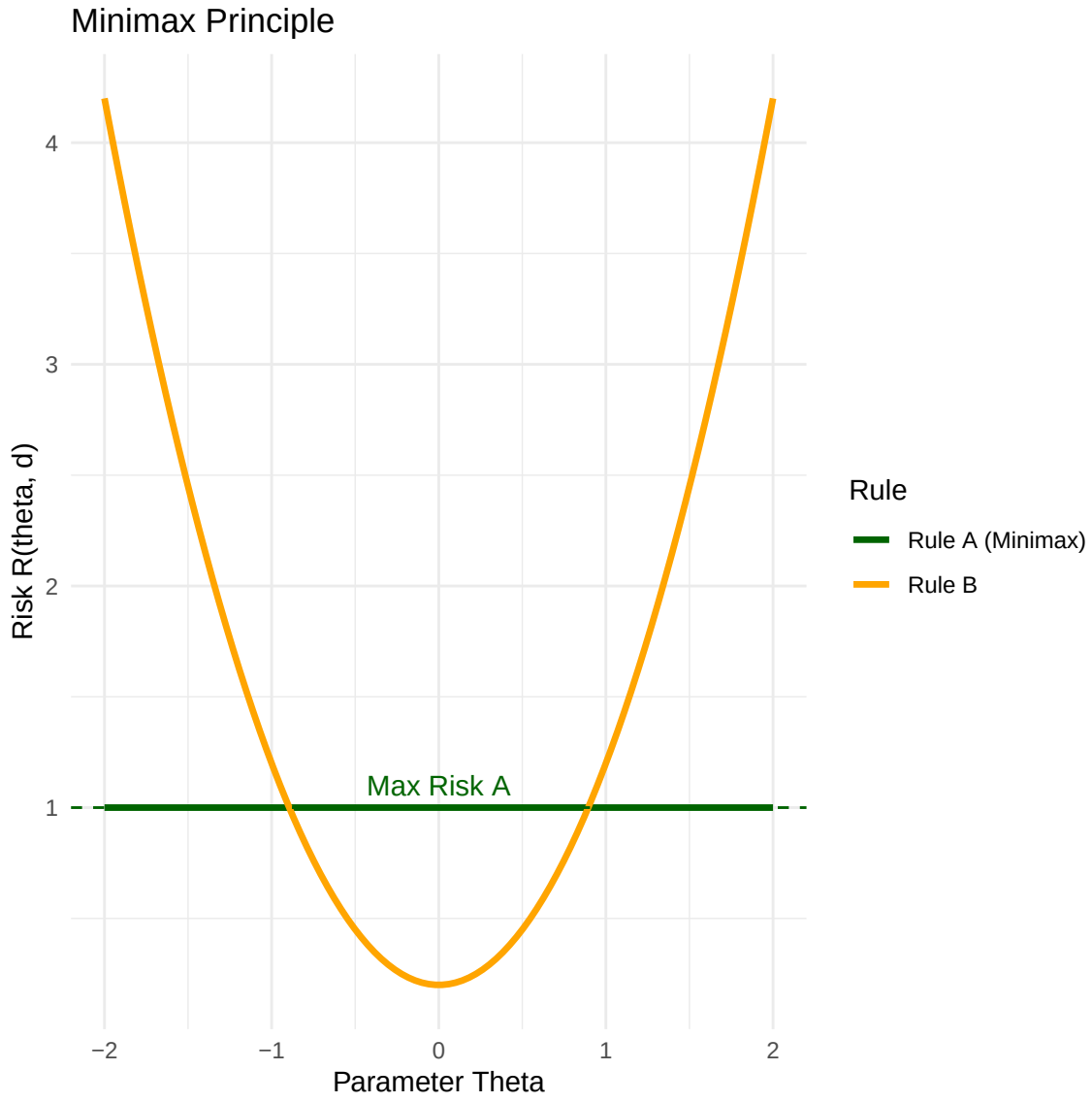


Figure 2.2: Illustration of Minimax: Rule A has a lower peak risk than Rule B, making Rule A the Minimax choice.

2.5.3 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.

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

2.6 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 $R_2 = R(\theta_2)$.

2.6.1 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.

2.6.2 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).

2.6.3 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.

2.6.4 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 $-\pi_1/\pi_2$.

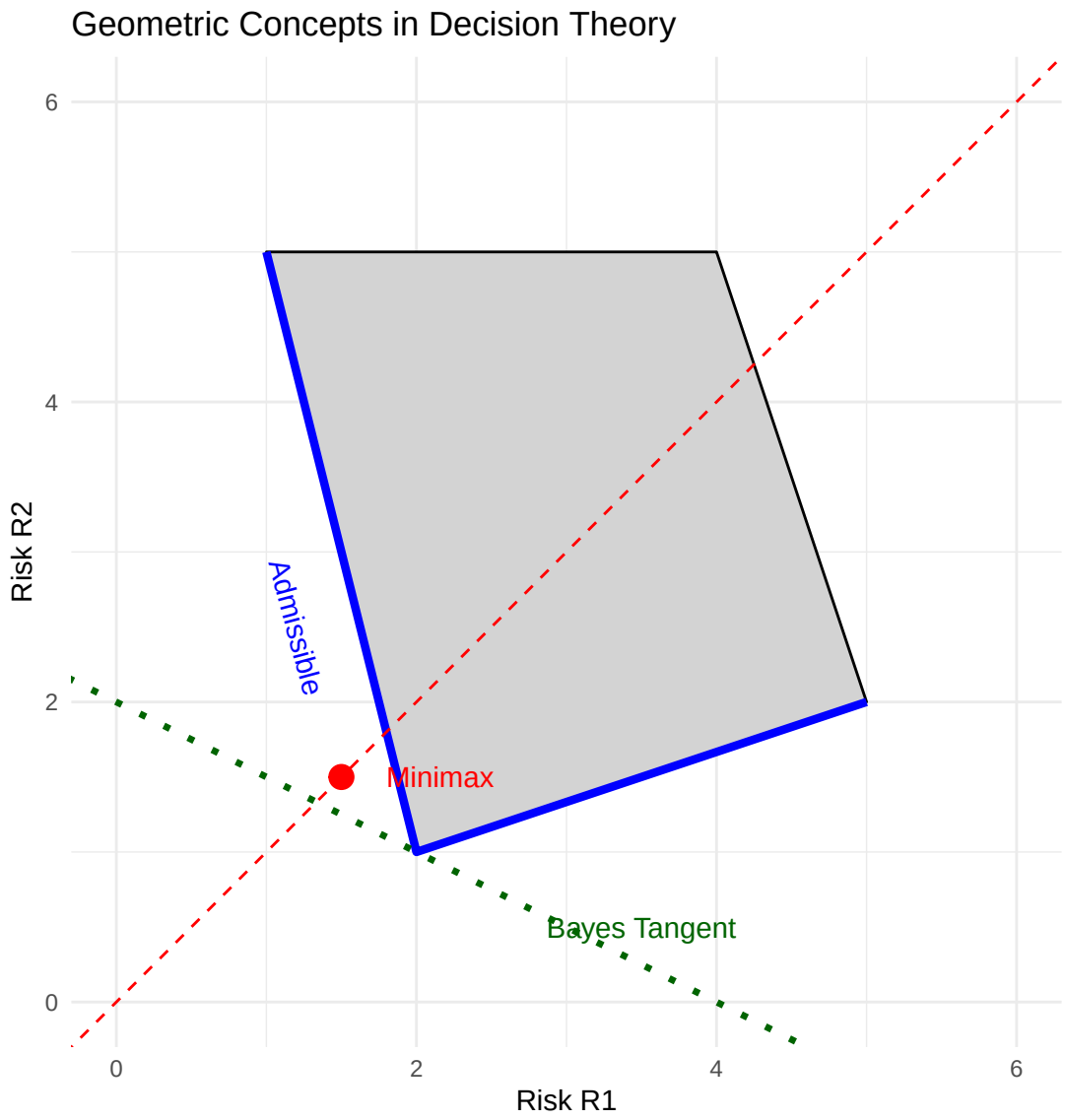


Figure 2.3: Geometric Interpretation: The gray polygon is the Risk Set S. The blue boundary represents admissible rules. The red point is the Minimax rule. The green line represents a Bayes rule for a specific prior.

2.7 Revisiting the Necklace Example: Geometric Solution

We now apply the geometric interpretation to the Necklace problem using the risks calculated in Section 2.4.

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

2.7.1 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 $1/3$.

2.8 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.

2.8.1 Constant Risk Bayes Rule Is Minimax (Proof by Contradiction)

Theorem 2.1 (Constant Risk Bayes Rule Is Minimax)**.** *Let δ^π be a Bayes estimator with respect to a prior π. If the risk function of δ^π is constant on the parameter space Θ, such that $R(\theta, \delta^\pi) = c$ for all $\theta \in \Theta$, then δ^π is a minimax estimator.*

Proof. Assume, for the sake of contradiction, that δ^π is **not** a minimax estimator.
By definition, if δ^π is not minimax, there must exist some other estimator δ' that has a strictly smaller maximum risk. That is:

$$\sup_{\theta \in \Theta} R(\theta, \delta') < \sup_{\theta \in \Theta} R(\theta, \delta^\pi) \tag{2.4}$$

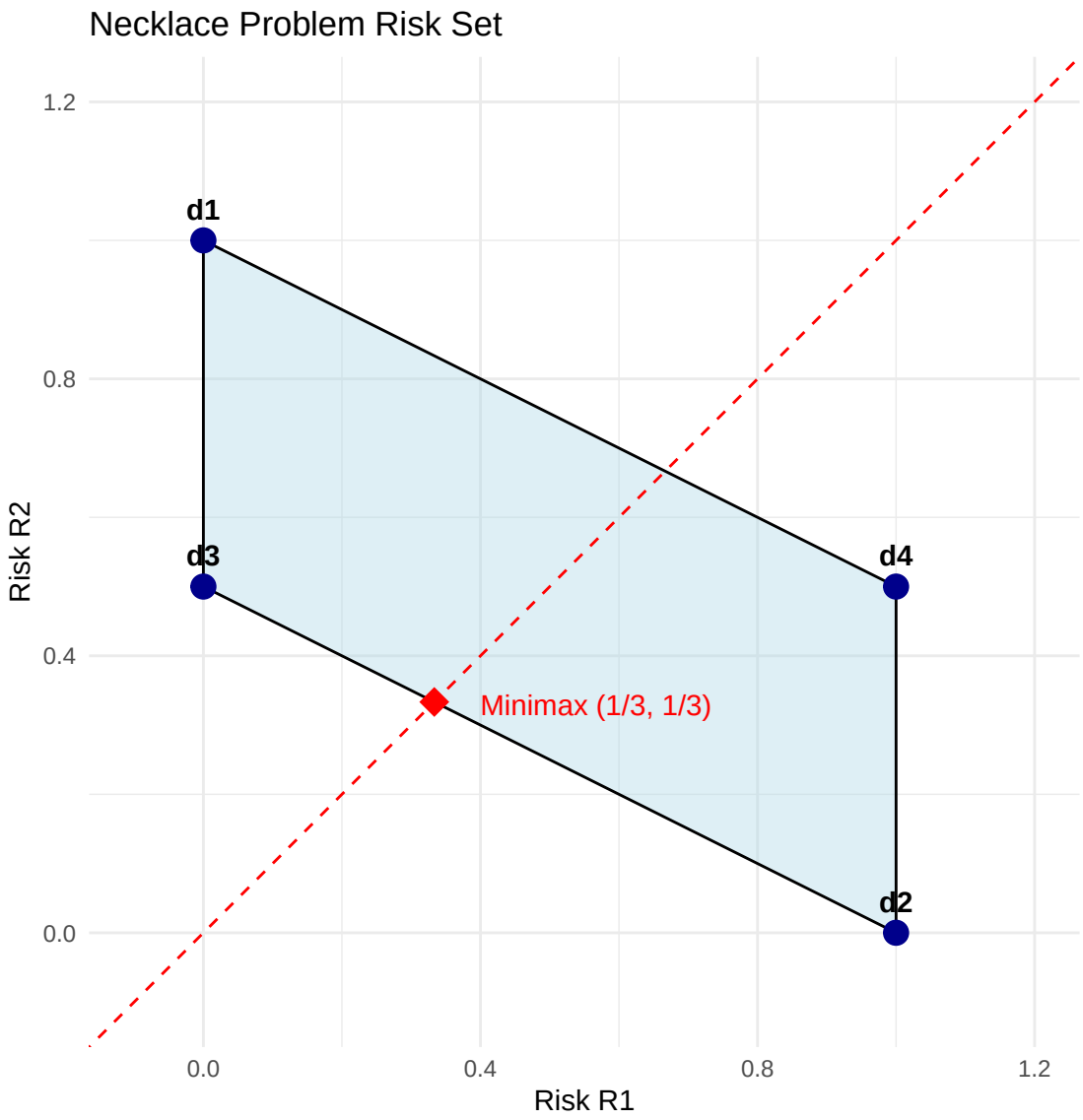


Figure 2.4: Necklace Problem Solution. The Minimax rule (red diamond) is the specific randomized combination of d3 and d2 that equalizes the risk.

Since we are given that $R(\theta, \delta^\pi) = c$ for all $\theta \in \Theta$, its supremum is simply c. Therefore, our assumption implies:

$$\sup_{\theta \in \Theta} R(\theta, \delta') < c \tag{2.5}$$

Now, consider the Bayes risk of δ' with respect to the prior π. The Bayes risk is the weighted average of the risk function:

$$r(\pi, \delta') = \int_\Theta R(\theta, \delta')\pi(\theta)d\theta \tag{2.6}$$

Since $R(\theta, \delta') \leq \sup_\theta R(\theta, \delta')$ for all θ, and we assumed this supremum is strictly less than c, it follows that:

$$r(\pi, \delta') \leq \sup_{\theta \in \Theta} R(\theta, \delta') < c \tag{2.7}$$

However, we know that c is the Bayes risk of δ^π:

$$r(\pi, \delta^\pi) = \int_\Theta c\,\pi(\theta)d\theta = c \tag{2.8}$$

Substituting this into our inequality, we get:

$$r(\pi, \delta') < r(\pi, \delta^\pi) \tag{2.9}$$

This result contradicts the fact that δ^π is a **Bayes estimator**. By definition, a Bayes estimator must minimize the Bayes risk, meaning $r(\pi, \delta^\pi) \leq r(\pi, \delta)$ for any estimator δ.
Because our assumption that δ^π is not minimax leads to a contradiction of the Bayes optimality of δ^π, the assumption must be false. Thus, δ^π must be minimax. □

The plot below visualizes this logic. If an estimator δ' (Blue) were to be "better" in a minimax sense than δ^π (Red), its entire curve would have to stay below the maximum value c. However, if it stays below c everywhere, its average (Bayes risk) would necessarily be lower than c, which is impossible if δ^π is the Bayes estimator.

```r
# Define Parameter Space Theta
theta <- seq(0, 1, length.out = 200)

# 1. Constant Risk Bayes Estimator (risk = C)
c_val <- 0.6
risk_bayes <- rep(c_val, length(theta))

# 2. An estimator that would contradict Bayes optimality
# (Always below the constant risk line)
risk_contradiction <- 0.5 + 0.05 * cos(2 * pi * theta)

# Plotting
plot(theta, risk_bayes, type = 'l', lwd = 3, col = "red",
     ylim = c(0, 1), ylab = "Risk R(theta, d)", xlab = expression(theta),
     main = "Proof by Contradiction Geometry")
```

```r
# Add the "Better" Estimator (which is impossible)
lines(theta, risk_contradiction, col = "blue", lwd = 2, lty = 2)

# Shaded area showing the "Impossible" Bayes Risk improvement
polygon(c(theta, rev(theta)), c(risk_contradiction, rev(risk_bayes)),
        col = rgb(0, 0, 1, 0.1), border = NA)

# Add Legend
legend("topright",
       legend = c("Constant Risk Bayes (c)", "Hypothetical 'Better' Est."),
       col = c("red", "blue"), lwd = 2, lty = c(1, 2))
```

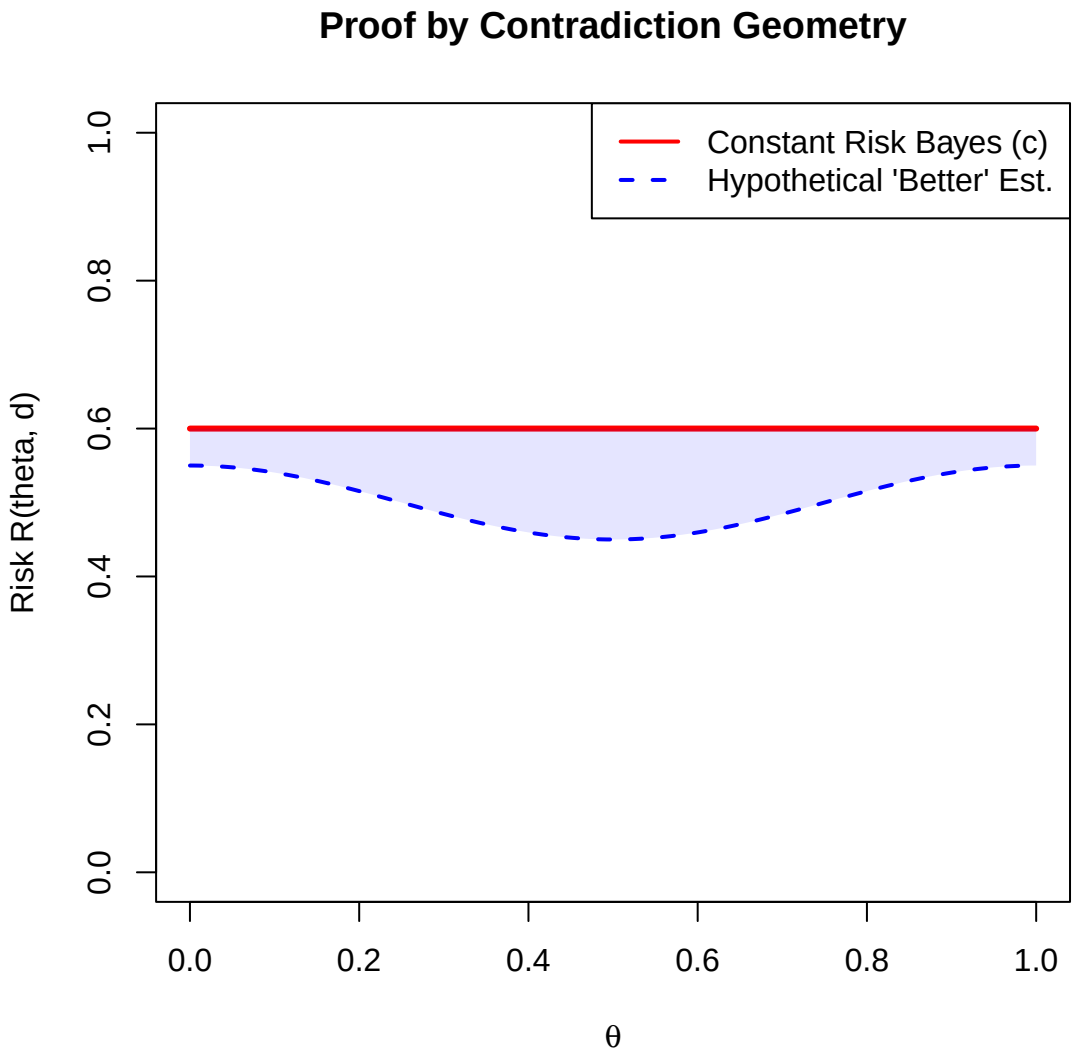


Figure 2.5: Visualizing the Contradiction: If the blue curve's maximum were below the red line, its average risk would be lower than the Bayes risk of the red estimator.

2.8.2 Minimaxity via Limiting Bayes Risks

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 (Minimaxity of Limit-Attaining 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:*

$$\sup_\theta R(\theta, \delta_0) \leq \lim_{n\to\infty} r(\pi_n, \delta_n) \tag{2.10}$$

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:

   $$\sup_\theta R(\theta, \delta^*) < \sup_\theta R(\theta, \delta_0) \leq V \tag{2.11}$$

   Let $\sup_\theta R(\theta, \delta^*) = V - \epsilon$ for some $\epsilon > 0$.
3. **Bounded Risk of δ^*:** The Bayes risk of δ^* is bounded by its maximum risk:

   $$r(\pi_n, \delta^*) = \int R(\theta, \delta^*)\pi_n(\theta)d\theta \leq V - \epsilon \tag{2.12}$$

   Therefore, $\lim_{n\to\infty} r(\pi_n, \delta^*) \leq V - \epsilon$.
4. **Optimality of δ_n:** Since δ_n is the Bayes rule for π_n, it minimizes Bayes risk. This creates the inequality pair shown in the figure (Orange $\leq$ Blue):

   $$r(\pi_n, \delta_n) \leq r(\pi_n, \delta^*) \tag{2.13}$$
5. **The Contradiction:** Combining the inequalities, we get:

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

   This implies $V \leq V - \epsilon$, which is impossible. Thus δ_0 must be Minimax. ■

□

2.8.3 Procedure: Verifying Minimaxity

The theorem above provides a practical recipe for identifying Minimax rules, particularly in unbounded parameter spaces (where a standard Least Favorable Prior often does not exist). The procedure is often used "backwards"—we

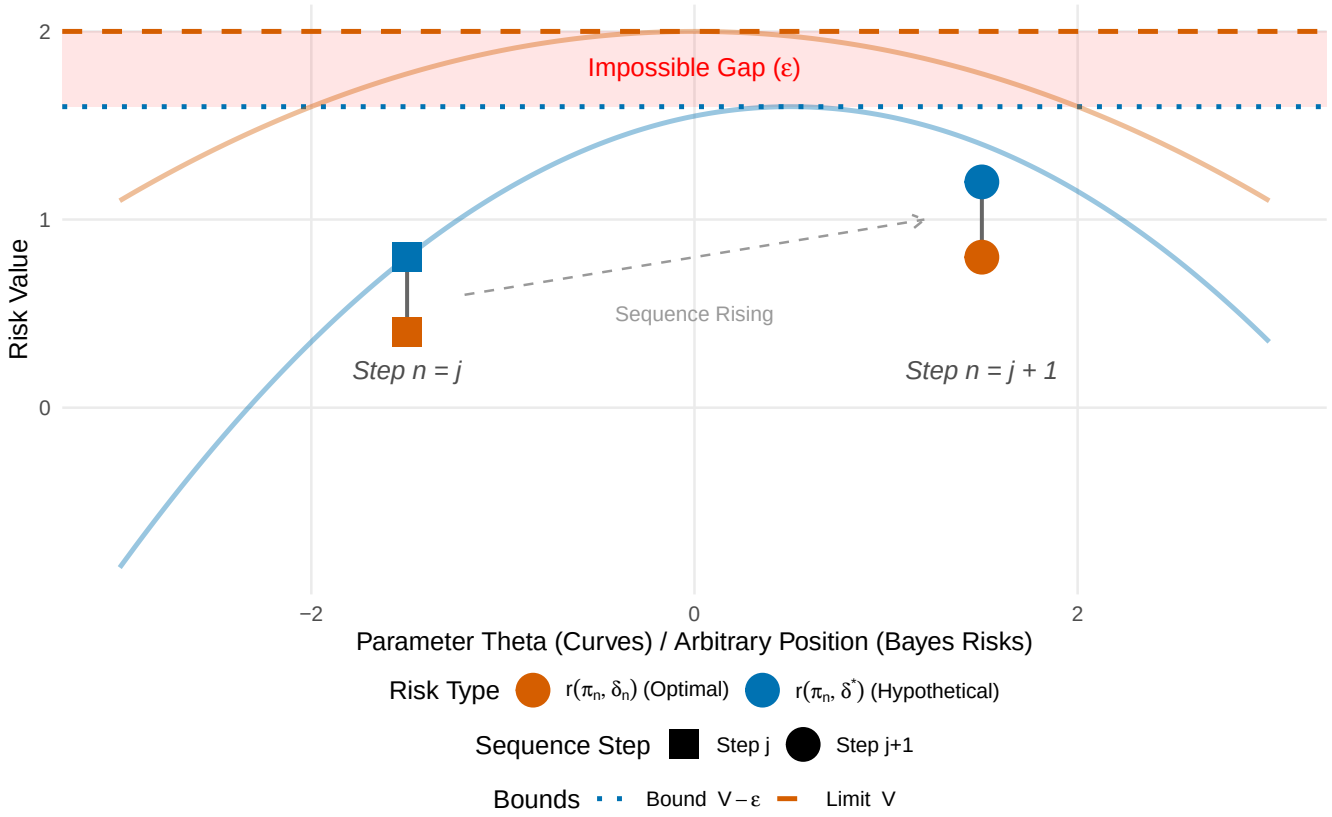


Figure 2.6: Visual Proof: We examine the Bayes risks at two steps, $n = j$ (squares) and $n = j + 1$ (circles). In both steps, the optimal risk $r(\pi_n, \delta_n)$ (orange) must be lower than the hypothetical risk $r(\pi_n, \delta^*)$ (blue). Even as the sequence rises (j+1 is higher than j), the blue points are capped by the bound $V - \epsilon$. This 'traps' the orange points, making it impossible for them to ever reach the Limit V.

guess a rule and then construct a sequence to prove it is Minimax.

1. **Propose a Candidate Rule (δ_0):** Identify a rule that intuitively seems robust. Typically, we look for an **Equalizer Rule**, which is a rule with constant risk ($R(\theta, \delta_0) = C$ for all θ). If the risk is constant, then $\sup_\theta R(\theta, \delta_0) = C$.
2. **Construct a Sequence of Priors (π_n):** Choose a sequence of priors that becomes increasingly "diffuse" or "flat" as $n \to \infty$ (e.g., Uniform on $[-n, n]$ or Normal with variance n). These approximate the "improper" prior corresponding to the candidate rule.
3. **Compute Bayes Risks (r_n):** Calculate the Bayes risk $r(\pi_n, \delta_n)$ for each prior in the sequence. Note that you do not necessarily need the formula for the Bayes rule δ_n itself, only its associated risk.
4. **Verify the Condition:** Check if the limit of the Bayes risks approaches the maximum risk of your candidate:
$$\lim_{n\to\infty} r(\pi_n, \delta_n) = \sup_\theta R(\theta, \delta_0) \tag{2.15}$$
If this holds, δ_0 is Minimax.

Example 2.1 (The Normal Mean)**.** Consider a single observation $X \sim N(\theta, 1)$ with squared error loss $L(\theta, \delta) = (\theta - \delta)^2$. We suspect the sample mean (in this case, just X itself) is the Minimax estimator.
Step 1: Candidate Rule
Let $\delta_0(X) = X$. The risk is the variance of the estimator:

$$R(\theta, \delta_0) = E[(\theta - X)^2] = \text{Var}(X) = 1 \tag{2.16}$$

Since the risk is constant (1) for all θ, $\sup_\theta R(\theta, \delta_0) = 1$.
Step 2: Sequence of Priors
We choose a sequence of Normal priors $\pi_n \sim N(0, n)$. As n increases, the variance increases, making the prior flatter over the real line.
Step 3: Bayes Risks
For a Normal prior $\theta \sim N(0, \tau^2)$ and data $X \sim N(\theta, \sigma^2)$, the Bayes risk is known to be:

$$r(\pi, \delta_\pi) = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} \tag{2.17}$$

Substituting our values ($\sigma^2 = 1, \tau^2 = n$):

$$r(\pi_n, \delta_n) = \frac{1 \cdot n}{1 + n} = \frac{n}{n+1} \tag{2.18}$$

Step 4: Verification
We take the limit of the sequence of Bayes risks:

$$\lim_{n\to\infty} r(\pi_n, \delta_n) = \lim_{n\to\infty} \frac{n}{n+1} = 1 \tag{2.19}$$

Comparing this to our candidate:

$$\sup_\theta R(\theta, \delta_0) = 1 \leq 1 \tag{2.20}$$

The condition holds. Therefore, $\delta_0(X) = X$ is the Minimax estimator for θ.

2.8.4 Bayes Rule as a Working Horse to Find a Minimax Rule

2.8.4.1 The Minimax Theorem (Saddle Point)

This theorem connects the search for a Minimax rule to the search for a Least Favorable Prior. It justifies the strategy of "finding the worst prior and solving it."

Theorem 2.3 (The Minimax Theorem)**.** *Let $\mathcal{D}$ be the set of all decision rules and Π be the set of all prior distributions. Let $r(\pi, \delta)$ denote the Bayes risk.*
The Minimax value equals the Maximin Bayes value:

$$\inf_{\delta \in \mathcal{D}} \sup_{\pi \in \Pi} r(\pi, \delta) = \sup_{\pi \in \Pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) \tag{2.21}$$

*Furthermore, a pair (δ_0, π_0) is a **Saddle Point** if for all $\delta \in \mathcal{D}$ and $\pi \in \Pi$:*

$$r(\pi_0, \delta) \geq r(\pi_0, \delta_0) \geq r(\pi, \delta_0) \tag{2.22}$$

If such a saddle point exists, then:

1. *δ_0 is a **Minimax rule**.*
2. *π_0 is a **Least Favorable Prior**.*

Proof. **Goal:** We wish to show that if (δ_0, π_0) is a saddle point, then $\sup_\theta R(\theta, \delta_0) \leq \sup_\theta R(\theta, \delta)$ for any other rule δ.
1. Interpret the Saddle Point Inequalities: The condition is given as two simultaneous inequalities:

$$\begin{aligned} &(A) \quad r(\pi_0, \delta_0) \leq r(\pi_0, \delta) \quad \text{for all } \delta \\ &(B) \quad r(\pi, \delta_0) \leq r(\pi_0, \delta_0) \quad \text{for all } \pi \end{aligned} \tag{2.23}$$

2. Analyze Inequality (A): Since $r(\pi_0, \delta_0) \leq r(\pi_0, \delta)$ for all δ, δ_0 minimizes the Bayes risk with respect to π_0.

- Therefore, δ_0 **is the Bayes rule** for π_0.

3. Analyze Inequality (B): Since $r(\pi, \delta_0) \leq r(\pi_0, \delta_0)$ for all π, the prior π_0 maximizes the average risk of δ_0.

- Since the supremum over all priors includes point-mass priors (which yield the risk at a single θ), maximizing over π is equivalent to maximizing over θ:

$$\sup_\pi r(\pi, \delta_0) = \sup_\theta R(\theta, \delta_0) \tag{2.24}$$

- Therefore, Inequality (B) implies:

$$\sup_\theta R(\theta, \delta_0) = r(\pi_0, \delta_0) \tag{2.25}$$

4. Combine to Prove Minimaxity: Let δ^* be any arbitrary decision rule. We compute its worst-case risk:

$$\begin{aligned} \sup_\theta R(\theta, \delta^*) &= \sup_\pi r(\pi, \delta^*) && \text{(Max risk = Max average risk)} \\ &\geq r(\pi_0, \delta^*) && \text{(Supremum } \geq \text{ specific value)} \\ &\geq r(\pi_0, \delta_0) && \text{(From Inequality A: } \delta_0 \text{ is Bayes for } \pi_0) \\ &= \sup_\theta R(\theta, \delta_0) && \text{(From Step 3)} \end{aligned} \tag{2.26}$$

5. Conclusion: We have shown that for any δ^*:

$$\sup_\theta R(\theta, \delta^*) \geq \sup_\theta R(\theta, \delta_0) \tag{2.27}$$

Thus, δ_0 minimizes the maximum risk. δ_0 is Minimax. ■ □

2.8.4.2 Alternating Optimization on the Risk Surface

The Minimax solution can be found computationally by iteratively optimizing one variable while holding the other fixed.

1. **Fix Prior π, Minimize Risk:** We search the valley bottom for the current π.
2. **Fix Rule δ, Maximize Risk:** We search the hill top for the current δ.

This creates a "zigzag" path on the surface that converges to the saddle point.

2.9 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 2.4 (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:

$$r(\pi, \delta_\pi) - r(\pi, \delta') = \sum_{\theta \in \Theta} \pi(\theta)[R(\theta, \delta_\pi) - R(\theta, \delta')] \tag{2.28}$$

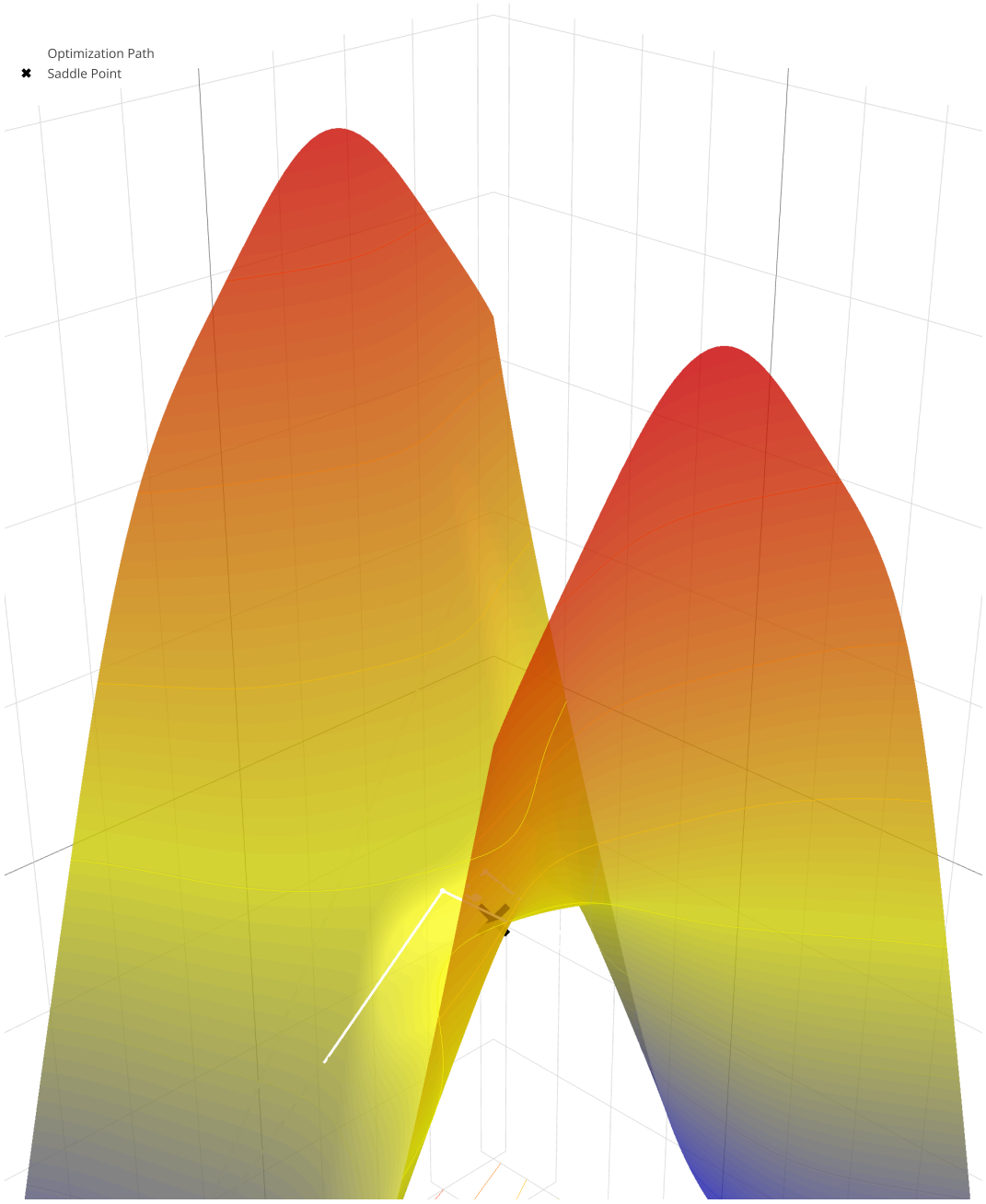


Figure 2.7: The 'Wiggle Mountain' of Risk. The surface represents Bayes Risk $r(\pi, \delta)$. The white zigzag line shows the iterative algorithm: starting from an arbitrary prior, we alternate between finding the best δ (moving along the valley) and the worst π (climbing the hill). This path spirals inward, converging to the red Saddle Point (Minimax solution) in the center.

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.
$$r(\pi, \delta_\pi) - r(\pi, \delta') > 0 \implies r(\pi, \delta') < r(\pi, \delta_\pi) \tag{2.29}$$
5. **Conclusion:** This contradicts the definition that δ_π is a Bayes rule (which must minimize Bayes risk). Therefore, δ_π is admissible. ■

□

2.9.1 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 2.5 (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:
$$r(\pi, \delta') = \int R(\theta, \delta')\pi(\theta)d\theta \leq \int R(\theta, \delta_\pi)\pi(\theta)d\theta = r(\pi, \delta_\pi) \tag{2.30}$$
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 Posterior Distributions

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:

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

3.1.1 Discrete Posterior Calculation

Example 3.1 (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 $\{0, 1, 2, ...\}$.
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.

3.1.2 Binomial-beta Conjugacy

Example 3.2 (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:

$$E^{\theta|X}[\theta] = \frac{a+X}{a+b+n} \tag{3.9}$$

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

- The posterior variance is:

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

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.

3.1.3 Normal-normal Conjugacy (known Variance)

Example 3.3 (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:

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

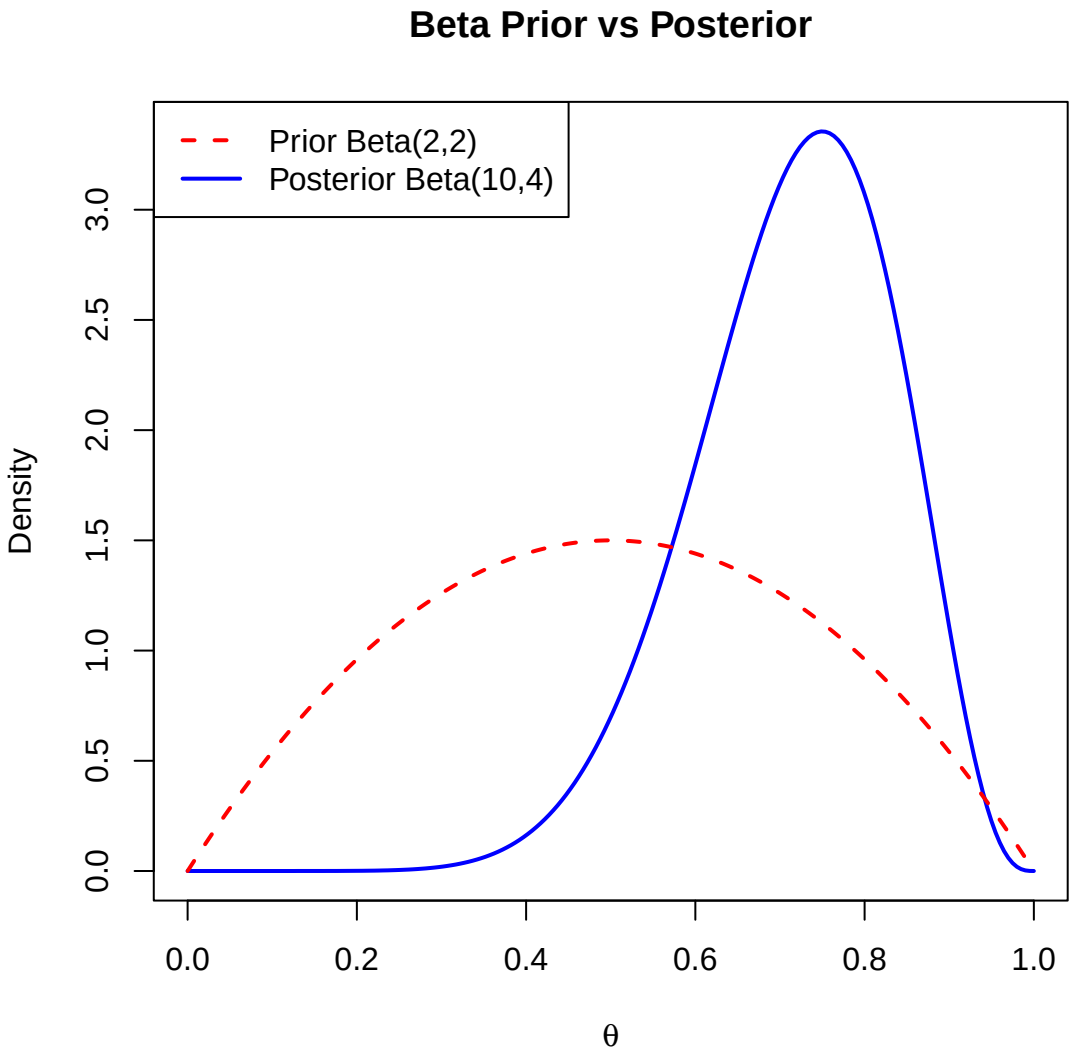


Figure 3.1: Prior vs Posterior for Beta-Binomial Example

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 $\mu_1 = \frac{175(0.04)+180(0.1)}{0.14} \approx 178.6$.

Figure 3.2 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)
```

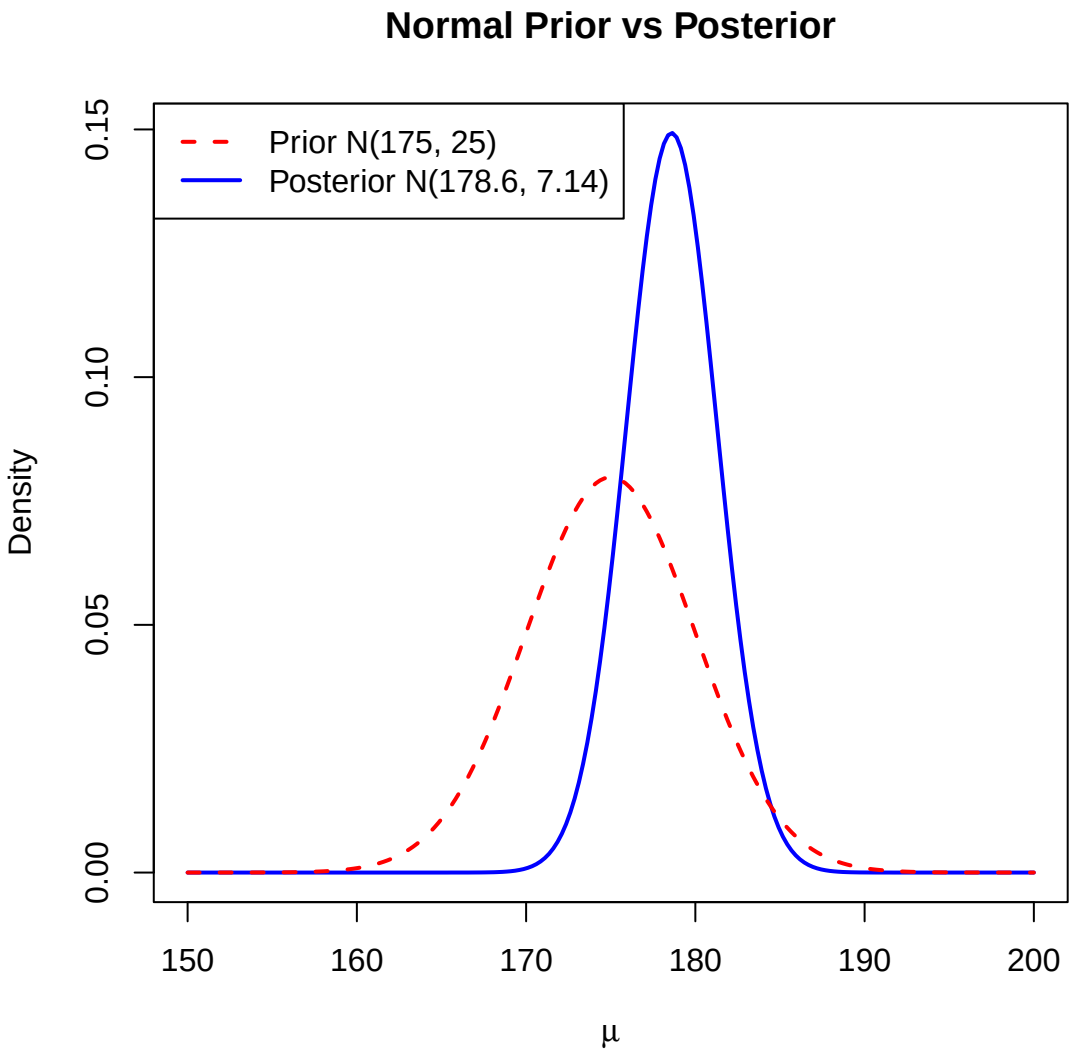


Figure 3.2: Prior vs Posterior for Normal-Normal Example

3.1.4 Normal with Unknown Mean and Variance

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 with parameters $\mu_0, \tau_0, \alpha_0, w_0$:

- $\tau \sim \text{Gamma}(\alpha_0/2, \alpha_0 w_0/2)$

$$\pi(\tau) \propto \tau^{\alpha_0/2-1} \exp\left\{-\frac{\alpha_0 w_0}{2}\tau\right\} \tag{3.17}$$

- $\mu|\tau \sim N(\mu_0, 1/(\tau_0\tau))$

$$\pi(\mu|\tau) \propto \tau^{1/2} \exp\left\{-\frac{\tau_0\tau}{2}(\mu-\mu_0)^2\right\} \tag{3.18}$$

The joint prior is:

$$\pi(\mu,\tau) \propto \tau^{(\alpha_0+1)/2-1} \exp\left\{-\frac{\tau}{2}\left(\alpha_0 w_0 + \tau_0(\mu-\mu_0)^2\right)\right\} \tag{3.19}$$

The Likelihood:
To derive the Maximum Likelihood Estimators (MLEs), we work with the log-likelihood function $l(\mu,\tau) = \log L(\mu,\tau)$:

$$\begin{aligned} l(\mu,\tau) &= \log\left(\tau^{n/2}\exp\left\{-\frac{\tau}{2}[S_{xx}+n(\bar{x}-\mu)^2]\right\}\right) \\ &= \frac{n}{2}\log\tau - \frac{\tau}{2}\left[S_{xx}+n(\bar{x}-\mu)^2\right] + \text{const} \end{aligned} \tag{3.20}$$

MLE for μ
Differentiating $l(\mu,\tau)$ with respect to μ and setting to zero:

$$\frac{\partial l}{\partial \mu} = n\tau(\bar{x}-\mu) = 0 \implies \hat{\mu}_{\text{MLE}} = \bar{x} \tag{3.21}$$

MLE for σ^2
Differentiating $l(\mu,\tau)$ with respect to τ, setting to zero, and substituting $\mu = \bar{x}$:

$$\frac{\partial l}{\partial \tau} = \frac{n}{2\tau} - \frac{S_{xx}}{2} = 0 \implies \hat{\tau}_{\text{MLE}} = \frac{n}{S_{xx}} \tag{3.22}$$

Using the invariance property ($\sigma^2 = 1/\tau$):

$$\hat{\sigma}^2_{\text{MLE}} = \frac{S_{xx}}{n} = \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n} \tag{3.23}$$

Derivation of the Posterior:
Multiplying the prior by the likelihood gives the joint posterior density. We organize the terms to separate the marginal distribution of τ from the conditional distribution of μ:

$$\begin{aligned} \pi(\mu,\tau|x) \propto & \underbrace{\tau^{(\alpha_0+n)/2-1}\exp\left\{-\frac{\tau}{2}\left[\alpha_0 w_0 + S_{xx} + \frac{n\tau_0}{n+\tau_0}(\bar{x}-\mu_0)^2\right]\right\}}_{\text{Marginal of } \tau} \\ & \times \underbrace{\tau^{1/2}\exp\left\{-\frac{(n+\tau_0)\tau}{2}\left(\mu - \frac{\tau_0\mu_0 + n\bar{x}}{n+\tau_0}\right)^2\right\}}_{\text{Conditional of } \mu|\tau} \end{aligned} \tag{3.24}$$

Results:

- **Conditional Posterior of $\mu|\tau,x$:**

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

$$E^{\mu|\tau,X}[\mu] = \frac{\tau_0\mu_0 + n\bar{x}}{\tau_0+n} \tag{3.26}$$

where

$$\tau' = \tau_0 + n \tag{3.27}$$

$$\mu' = \frac{\tau_0\mu_0 + n\bar{x}}{\tau_0+n} \tag{3.28}$$

- **Marginal Posterior of $\tau|x$**: The marginal posterior is $\tau|x \sim \text{Gamma}(\alpha', \beta')$ with:

$$\alpha' = \frac{\alpha_0 + n}{2}, \quad \beta' = \frac{\alpha_0 w_0 + n\hat{\sigma}^2_{\text{MLE}} + \frac{n\tau_0}{n+\tau_0}(\bar{x} - \mu_0)^2}{2} \tag{3.29}$$

  Using the approximation $E^{\sigma^2|X}[\sigma^2] \approx 1/E^{\tau|X}[\tau] = \beta'/\alpha'$, the posterior expectation of the variance is a weighted average of the prior variance, the data variance, and the discrepancy between the prior and data means:

$$E^{\sigma^2|X}[\sigma^2] \approx \frac{\alpha_0 w_0 + n\hat{\sigma}^2_{\text{MLE}} + \frac{1}{1/n+1/\tau_0}(\bar{x} - \mu_0)^2}{\alpha_0 + n} \tag{3.30}$$

- **Conditional Posterior of $\tau|\mu, x$**: If μ is considered known, the posterior for τ combines the prior α_0, w_0 with the deviations from μ. Note that the prior term $\pi(\mu|\tau)$ contributes an extra factor of $\tau^{1/2}$ to the shape.

$$\tau|\mu, x \sim \text{Gamma}(\alpha'', \beta'') \tag{3.31}$$

  Where:

$$\alpha'' = \frac{\alpha_0 + n + 1}{2}, \quad \beta'' = \frac{\alpha_0 w_0 + \sum_{i=1}^n (x_i - \mu)^2 + \tau_0(\mu - \mu_0)^2}{2} \tag{3.32}$$

  The approximate expectation of the variance is:

$$E^{\sigma^2|\mu,X}[\sigma^2] \approx \frac{\alpha_0 w_0 + \sum_{i=1}^n (x_i - \mu)^2 + \tau_0(\mu - \mu_0)^2}{\alpha_0 + n + 1} \tag{3.33}$$

3.2 Finding Bayes Rules via Minimizing Posterior Expected Loss

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

Definition 3.2 (Risk Function and Bayes Risk)**. Setup and Notation:**

- $\theta \in \Theta$: parameter of interest (unknown state of nature)
- $x \in X$: observed data
- $\pi(\theta)$: prior probability distribution over the parameter space
- $f(x; \theta)$: likelihood or sampling distribution of the data given the parameter
- $d : X \to A$: decision rule mapping observed data to an action/decision
- $\mathcal{L}(\theta, a)$: loss function measuring the loss incurred when the true parameter is θ and action a is taken

Definition:

- **Risk Function:** For a given decision rule d and parameter value θ,

$$R(\theta, d) = \int_X \mathcal{L}(\theta, d(x)) f(x; \theta) dx = E^{X|\theta}[\mathcal{L}(\theta, d(X))] \tag{3.34}$$

  is the expected loss with respect to the sampling distribution when the true parameter is θ.

- **Bayes Risk:** For a decision rule d and prior distribution π,

$$r(\pi, d) = \int_\Theta R(\theta, d)\pi(\theta) d\theta = E^{\theta}[R(\theta, d)] \tag{3.35}$$

  is the expected risk averaging over both the parameter uncertainty (prior) and the data variability (likelihood).

- **Posterior Bayes Loss:** The minimum possible expected loss given observed data x is denoted as $\rho^{\text{Bayes}}(\pi, x)$. It represents the expected posterior loss of the Bayes rule:

$$\rho^{\text{Bayes}}(\pi, x) = \inf_d E^{\theta|x}[\mathcal{L}(\theta, d)] \tag{3.36}$$

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 $d(x)$ is defined as*

$$d^{Bayes}(x) = \arg\min_a E^{\theta|X}[\mathcal{L}(\theta, a)] \tag{3.37}$$

The value of the minimum expected posterior loss is $\rho^{Bayes}(\pi, x)$.

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 $R(\theta, d)$:

$$\begin{aligned} r(\pi, d) &= \int_\Theta R(\theta, d)\pi(\theta) d\theta \\ &= \int_\Theta \left[\int_X \mathcal{L}(\theta, d(x)) f(x|\theta) dx \right] \pi(\theta) d\theta \end{aligned} \tag{3.38}$$

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

$$r(\pi, d) = \int_X \left[\int_\Theta \mathcal{L}(\theta, d(x)) f(x|\theta)\pi(\theta) d\theta \right] dx \tag{3.39}$$

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:

$$
\begin{aligned}
r(\pi, d) &= \int_X \left[\int_\Theta \mathcal{L}(\theta, d(x)) \pi(\theta|x) m(x) d\theta \right] dx \\
&= \int_X m(x) \left[\int_\Theta \mathcal{L}(\theta, d(x)) \pi(\theta|x) d\theta \right] dx
\end{aligned}
\tag{3.40}
$$

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**:

$$
\int_\Theta \mathcal{L}(\theta, d(x)) \pi(\theta|x) d\theta = E^{\theta|X}[\mathcal{L}(\theta, d(X))] \tag{3.41}
$$

□

! Important

Therefore, to minimize the Bayes risk, one effectively minimizes the posterior expected loss for each x. This relationship relies on the key identity for the total expectation of the loss:

$$
\boxed{r(\pi, d) = E^X \left[E^{\theta|X}(\mathcal{L}(\theta, d(X))) \right] = E^\theta \left[E^{X|\theta}(\mathcal{L}(\theta, d(X))) \right]} \tag{3.42}
$$

In the first expression, the outer expectation E^X is taken with respect to the **marginal density of the data**, $m(x)$, defined as:

$$
m(x) = \int_\Theta f(x|\theta) \pi(\theta) d\theta \tag{3.43}
$$

In the second expression, the outer expectation E^θ is taken with respect to the **prior density** $\pi(\theta)$.

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

3.3 Special Bayes Rules

3.3.1 Squared Error Loss (point Estimate)

$$
\mathcal{L}(\theta, a) = (\theta - a)^2 \tag{3.44}
$$

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:

$$
-2E^{\theta|X}(\theta - d) = 0 \implies d(X) = E^{\theta|X}[\theta] \tag{3.45}
$$

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

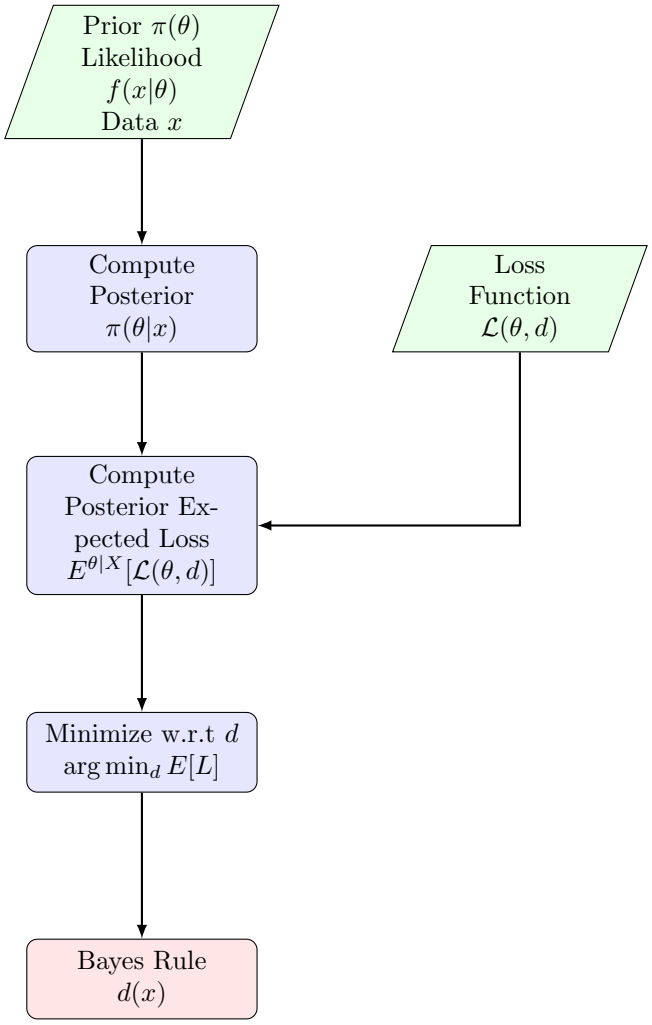


Figure 3.3: Workflow for Finding the Bayes Rule

3.3.2 Scale-Invariant Squared Error Loss

Consider the loss function that penalizes relative errors rather than absolute errors. This is particularly useful when the magnitude of the parameter θ varies significantly, and an error of 1.0 is "worse" when $\theta = 1$ than when $\theta = 1000$.

$$\mathcal{L}(\theta, d) = \left(\frac{d-\theta}{\theta}\right)^2 = \left(\frac{d}{\theta} - 1\right)^2 \tag{3.46}$$

To find the Bayes rule, we minimize the posterior expected loss $E^{\theta|X}[\mathcal{L}(\theta, d)]$:

$$Q(d) = E^{\theta|X}\left[\frac{d^2}{\theta^2} - \frac{2d}{\theta} + 1\right] = d^2 E^{\theta|X}[\theta^{-2}] - 2dE^{\theta|X}[\theta^{-1}] + 1 \tag{3.47}$$

Differentiating with respect to d and setting to zero:

$$\frac{\partial Q}{\partial d} = 2dE^{\theta|X}[\theta^{-2}] - 2E^{\theta|X}[\theta^{-1}] = 0 \tag{3.48}$$

Solving for d:

$$d(X) = \frac{E^{\theta|X}[\theta^{-1}]}{E^{\theta|X}[\theta^{-2}]} \tag{3.49}$$

Result: The Bayes rule under scale-invariant squared error loss is the ratio of the posterior mean of θ^{-1} to the posterior mean of θ^{-2}.

3.3.3 Absolute Error Loss

$$\mathcal{L}(\theta, d) = |\theta - d| \tag{3.50}$$

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

$$\psi(d) = E^{\theta|X}[|\theta - d|] = \int_{-\infty}^{\infty} |\theta - d| \, dF(\theta|x) \tag{3.51}$$

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

$$\psi(d) = \int_{-\infty}^{d} (d - \theta) \, dF(\theta|x) + \int_{d}^{\infty} (\theta - d) \, dF(\theta|x) \tag{3.52}$$

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):

$$\frac{\partial}{\partial d}\psi(d) = \int_{-\infty}^{d} 1\, dF(\theta|x) - \int_{d}^{\infty} 1\, dF(\theta|x) = P(\theta \leq d|x) - P(\theta > d|x) \tag{3.53}$$

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

$$P(\theta \leq d|x) = P(\theta > d|x) \tag{3.54}$$

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**.

3.3.4 Weighted Absolute Error Loss (min-normalization)

$$\mathcal{L}(\theta, d) = \frac{|\theta - d|}{\min(\theta, 1-\theta)} \tag{3.55}$$

This loss function penalizes errors extremely heavily when the true parameter θ is near the boundaries (0 or 1). Because the denominator approaches zero at the boundaries, the "cost" of an error becomes infinite, forcing the estimator to be very cautious (conservative) if the posterior has significant mass near 0 or 1.

To find the Bayes rule, we minimize the posterior expected loss. Let $\pi(\theta|x)$ denote the posterior density.

$$\psi(d) = E^{\theta|X}\left[\frac{|\theta - d|}{\min(\theta, 1-\theta)}\right] = \int \frac{|\theta - d|}{\min(\theta, 1-\theta)}\pi(\theta|x)\, d\theta \tag{3.56}$$

Let $w(\theta) = \frac{1}{\min(\theta, 1-\theta)}$. We can view this integral as an expectation with respect to a **weighted posterior density** $\pi^*(\theta|x)$:

$$\pi^*(\theta|x) \propto w(\theta)\pi(\theta|x) = \frac{\pi(\theta|x)}{\min(\theta, 1-\theta)} \tag{3.57}$$

Result: The Bayes rule is the **median** of the weighted posterior distribution $\pi^*(\theta|x)$.

3.3.4.0.1 Importance Sampling for Weighted Median

Goal: Estimate the median of $\pi^*(\theta|x) \propto w(\theta)\pi(\theta|x)$ using samples from $\pi(\theta|x)$.

1. **Sample:** Generate M independent draws $\theta_1, \dots, \theta_M$ from the standard posterior $\pi(\theta|x)$.
2. **Weight:** For each $i = 1, \dots, M$, compute the importance weight:

$$W_i = w(\theta_i) = \frac{1}{\min(\theta_i, 1 - \theta_i)} \tag{3.58}$$

3. **Sort:** Reorder the samples such that $\theta_{(1)} \leq \theta_{(2)} \leq \dots \leq \theta_{(M)}$. Permute the weights $W_{(1)}, \dots, W_{(M)}$ to match this ordering.
4. **Accumulate:** Compute the cumulative weights:

$$S_k = \sum_{j=1}^{k} W_{(j)} \quad \text{for } k = 1, \dots, M \tag{3.59}$$

5. **Select:** Find the smallest index k^* such that the cumulative weight exceeds half the total weight:

$$k^* = \min\{k : S_k \geq 0.5 \times S_M\} \tag{3.60}$$

6. **Output:** Return the estimator $\hat{\delta} = \theta_{(k^*)}$.

Numerical Example: Beta(2, 10)

We compare the "exact" weighted median (found by numerical integration) with the Monte Carlo estimate for a skewed distribution.

```
# 1. Setup
set.seed(2025)
M <- 10
alpha <- 2
beta <- 10

theta_samples <- rbeta(M, alpha, beta)
w <- function(theta) { 1 / pmin(theta, 1 - theta) }

# 2. Process
weights <- w(theta_samples)
ord <- order(theta_samples)
sorted_theta <- theta_samples[ord]
sorted_weights <- weights[ord]
cum_weights <- cumsum(sorted_weights)
total_weight <- sum(sorted_weights)
threshold <- 0.5 * total_weight
```

```r
# Find k
k_idx <- which(cum_weights >= threshold)[1]

# 3. Create Data Frame
selection_table <- data.frame(
  idx = 1:M,
  theta = sorted_theta,
  weight = sorted_weights,
  cum_weight = cum_weights,
  check = ifelse(cum_weights >= threshold, "$\\ge$ Threshold", "$<$ Threshold"),
  sel = ifelse(1:M == k_idx, "$\\leftarrow k$ (Median)", "")
)

# 4. Set LaTeX Column Names
# Note: We use double backslashes \\ for LaTeX commands inside R strings
colnames(selection_table) <- c(
  "$i$",
  "$\\theta_{(i)}$",        # Sorted Theta
  "$w_{(i)}$",              # Sorted Weight
  "$\\sum_{j=1}^i w_{(j)}$",# Cumulative Sum
  "Condition",
  "Selection"
)

# Print Context
cat("Total Weight ($\\sum w_i$):", total_weight, "\n")
```

```
Total Weight ($\sum w_i$): 48.91008
```

```r
cat("Threshold ($0.5 \\times \\sum w_i$):", threshold, "\n\n")
```

```
Threshold ($0.5 \times \sum w_i$): 24.45504
```

```r
# 5. Render Table with escape = FALSE
# escape = FALSE is crucial; otherwise, it prints the dollar signs literally
knitr::kable(selection_table,
             digits = 4,
             align = "c",
             escape = FALSE)
```

i	$\theta_{(i)}$	$w_{(i)}$	$\sum_{j=1}^{i} w_{(j)}$	Condition	Selection
1	0.1256	7.9601	7.9601	$<$ Threshold	
2	0.1462	6.8412	14.8013	$<$ Threshold	

i	$\theta_{(i)}$	$w_{(i)}$	$\sum_{j=1}^{i} w_{(j)}$	Condition	Selection
3	0.1563	6.3965	21.1978	$<$ Threshold	
4	0.1714	5.8329	27.0307	$\geq$ Threshold	$\leftarrow k$ (Median)
5	0.2221	4.5024	31.5330	$\geq$ Threshold	
6	0.2265	4.4144	35.9475	$\geq$ Threshold	
7	0.2676	3.7376	39.6850	$\geq$ Threshold	
8	0.2840	3.5214	43.2064	$\geq$ Threshold	
9	0.2990	3.3445	46.5509	$\geq$ Threshold	
10	0.4239	2.3592	48.9101	$\geq$ Threshold	

Actual Weighted Median with 1000 draws of θ

```
# 1. Setup Parameters
set.seed(123)
M <- 1000
alpha <- 2
beta <- 10

# 2. Generate Samples and Weights
theta_samples <- rbeta(M, alpha, beta)
# Weight function: w(theta) = 1 / min(theta, 1-theta)
w <- function(theta) { 1 / pmin(theta, 1 - theta) }
weights <- w(theta_samples)

# 3. Sort and Calculate Cumulative Weights
ord <- order(theta_samples)
sorted_theta <- theta_samples[ord]
sorted_weights <- weights[ord]

cum_weights <- cumsum(sorted_weights)
total_weight <- sum(sorted_weights)
threshold <- 0.5 * total_weight

# 4. Find the Weighted Median Index k
k_idx <- which(cum_weights >= threshold)[1]
mc_weighted_median <- sorted_theta[k_idx]

# 5. Compare with Theoretical Value (calculated previously)
# Re-calculating theoretical for completeness of this chunk
weighted_dens_unnorm <- function(theta) { w(theta) * dbeta(theta, alpha, beta) }
C <- integrate(weighted_dens_unnorm, 0, 1)$value
weighted_cdf <- function(q) { integrate(weighted_dens_unnorm, 0, q)$value / C }
theo_median <- uniroot(function(x) weighted_cdf(x) - 0.5, c(0.001, 0.999))$root
```

```r
# 6. Display Results
results <- data.frame(
  "Method" = c("Theoretical (Integration)", "Monte Carlo (M=1000)", "Standard Median (Unweighted)
  "Value" = c(theo_median, mc_weighted_median, qbeta(0.5, alpha, beta))
)

knitr::kable(results, digits = 4, align = "l", caption = "Weighted Median Estimation")
```

Table 3.3: Weighted Median Estimation

Method	Value
Theoretical (Integration)	0.0670
Monte Carlo (M=1000)	0.0692
Standard Median (Unweighted)	0.1480

3.3.5 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:

Table 3.4: Standard 0-1 Loss Function

State of Nature (θ)	Action $a = 0$ (Accept H_0)	Action $a = 1$ (Reject H_0)
$\theta \in \Theta_0$ **(H_0 True)**	0 (Correct)	1 (Type I Error)
$\theta \in \Theta_1$ **(H_1 True)**	1 (Type II Error)	0 (Correct)

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

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

$$E^{\theta|X}[\mathcal{L}(\theta, 0)] = 0 \cdot P(\theta \in \Theta_0|x) + 1 \cdot P(\theta \in \Theta_1|x) = P(\theta \in \Theta_1|x) \tag{3.61}$$

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

$$E^{\theta|X}[\mathcal{L}(\theta, 1)] = 1 \cdot P(\theta \in \Theta_0|x) + 0 \cdot P(\theta \in \Theta_1|x) = P(\theta \in \Theta_0|x) \tag{3.62}$$

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

$$P(\theta \in \Theta_0|x) \leq P(\theta \in \Theta_1|x) \tag{3.63}$$

This confirms that under 0-1 loss, the Bayes rule simply selects the hypothesis with the higher posterior probability. The optimal Bayes decision rule $d(x)$ is given by:

$$d(x) = \begin{cases} 1 & \text{if } P(\Theta_0|x) \leq \frac{1}{2} \quad \text{(Reject } H_0\text{)} \\ 0 & \text{if } P(\Theta_0|x) > \frac{1}{2} \quad \text{(Accept } H_0\text{)} \end{cases} \tag{3.64}$$

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.

Table 3.5: Loss Function with Type I Error Cost c

State of Nature (θ)	Action $a = 0$ (Accept H_0)	Action $a = 1$ (Reject H_0)
$\theta \in \Theta_0$ (H_0 **True**)	0	c (Type I Error)
$\theta \in \Theta_1$ (H_1 **True**)	1 (Type II Error)	0

We again calculate the posterior expected loss:

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

$$E^{\theta|X}[\mathcal{L}(\theta, 0)] = 0 \cdot P(\Theta_0|x) + 1 \cdot P(\Theta_1|x) = P(\Theta_1|x) \tag{3.65}$$

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

$$E^{\theta|X}[\mathcal{L}(\theta, 1)] = c \cdot P(\Theta_0|x) + 0 \cdot P(\Theta_1|x) = cP(\Theta_0|x) \tag{3.66}$$

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

$$cP(\Theta_0|x) \leq P(\Theta_1|x) \tag{3.67}$$

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

$$cP(\Theta_0|x) \leq 1 - P(\Theta_0|x) \implies (1 + c)P(\Theta_0|x) \leq 1 \tag{3.68}$$

$$P(\Theta_0|x) \leq \frac{1}{1 + c} \tag{3.69}$$

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. The optimal Bayes decision rule $d(x)$ is given by:

$$d(x) = \begin{cases} 1 & \text{if } P(\Theta_0|x) \leq \frac{1}{1+c} \quad \text{(Reject } H_0\text{)} \\ 0 & \text{if } P(\Theta_0|x) > \frac{1}{1+c} \quad \text{(Accept } H_0\text{)} \end{cases} \tag{3.70}$$

3.3.6 Classification Prediction

In classification problems, the parameter of interest is a discrete class label y taking values in a set of categories $\{1, 2, ..., K\}$. 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:

$$\mathcal{L}(y, \hat{y}) = \begin{cases} 0 & \text{if } \hat{y} = y \text{ (Correct Classification)} \\ 1 & \text{if } \hat{y} \neq y \text{ (Misclassification)} \end{cases} \tag{3.71}$$

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

$$E^{Y|X}[\mathcal{L}(y, \hat{y})] = \sum_{y} \mathcal{L}(y, \hat{y}) P(y|x) \tag{3.72}$$

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

$$E^{Y|X}[\mathcal{L}(y, \hat{y})] = \sum_{y \neq \hat{y}} 1 \cdot P(y|x) = P(y \neq \hat{y}|x) = 1 - P(y = \hat{y}|x) \tag{3.73}$$

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

Result:

The Bayes rule for classification is to predict the class with the highest posterior **predictive** probability. In the context of machine learning and pattern recognition, this decision rule is known as the **Bayes Optimal Classifier**.

$$\hat{y}_{\text{Bayes}}(x) = \underset{k \in \{1,...,K\}}{\arg\max} \; P(y = k|x) \tag{3.74}$$

3.3.7 Interval Estimation as a Decision Problem

We can motivate the choice of a Credible Interval by defining a specific loss function for interval estimation. We define the **action space** $\mathcal{A}$ as the set of all intervals of fixed radius $\delta > 0$ centered at d, i.e., $\mathcal{A} = \{[d - \delta, d + \delta] \mid d \in \mathbb{R}\}$.

The loss function is defined as:

$$\mathcal{L}(\theta, d) = \begin{cases} 0 & \text{if } |\theta - d| \leq \delta \quad (\theta \in [d - \delta, d + \delta]) \\ 1 & \text{if } |\theta - d| > \delta \quad (\theta \notin [d - \delta, d + \delta]) \end{cases} \tag{3.75}$$

Derivation of the Bayes Rule

We minimize the **Expected Posterior Loss**, which is simply the probability that θ falls outside the interval:

$$E^{\theta|X}[\mathcal{L}(\theta, d)] = 1 \cdot P(|\theta - d| > \delta|x) = 1 - P(d - \delta \leq \theta \leq d + \delta|x) \tag{3.76}$$

Minimizing this loss is equivalent to maximizing the posterior probability mass contained within the interval. Thus, the Bayes estimator d is:

$$d_{\text{Bayes}} = \arg\max_{d} \int_{d-\delta}^{d+\delta} \pi(\theta|x)\, d\theta \tag{3.77}$$

To find the optimal d, we differentiate the integral with respect to d and set it to zero:

$$\frac{\partial}{\partial d}\left(\int_{d-\delta}^{d+\delta} \pi(\theta|x)\, d\theta\right) = \pi(d + \delta|x) - \pi(d - \delta|x) = 0 \tag{3.78}$$

This yields the condition $\pi(d + \delta|x) = \pi(d - \delta|x)$.

The optimal d centers the interval such that the posterior density heights at the two endpoints are equal. This is the defining characteristic of a **Highest Posterior Density (HPD)** interval.

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)**.

3.4 Finding Minimax Rules with Bayes Rules

Theorem 2.1 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.

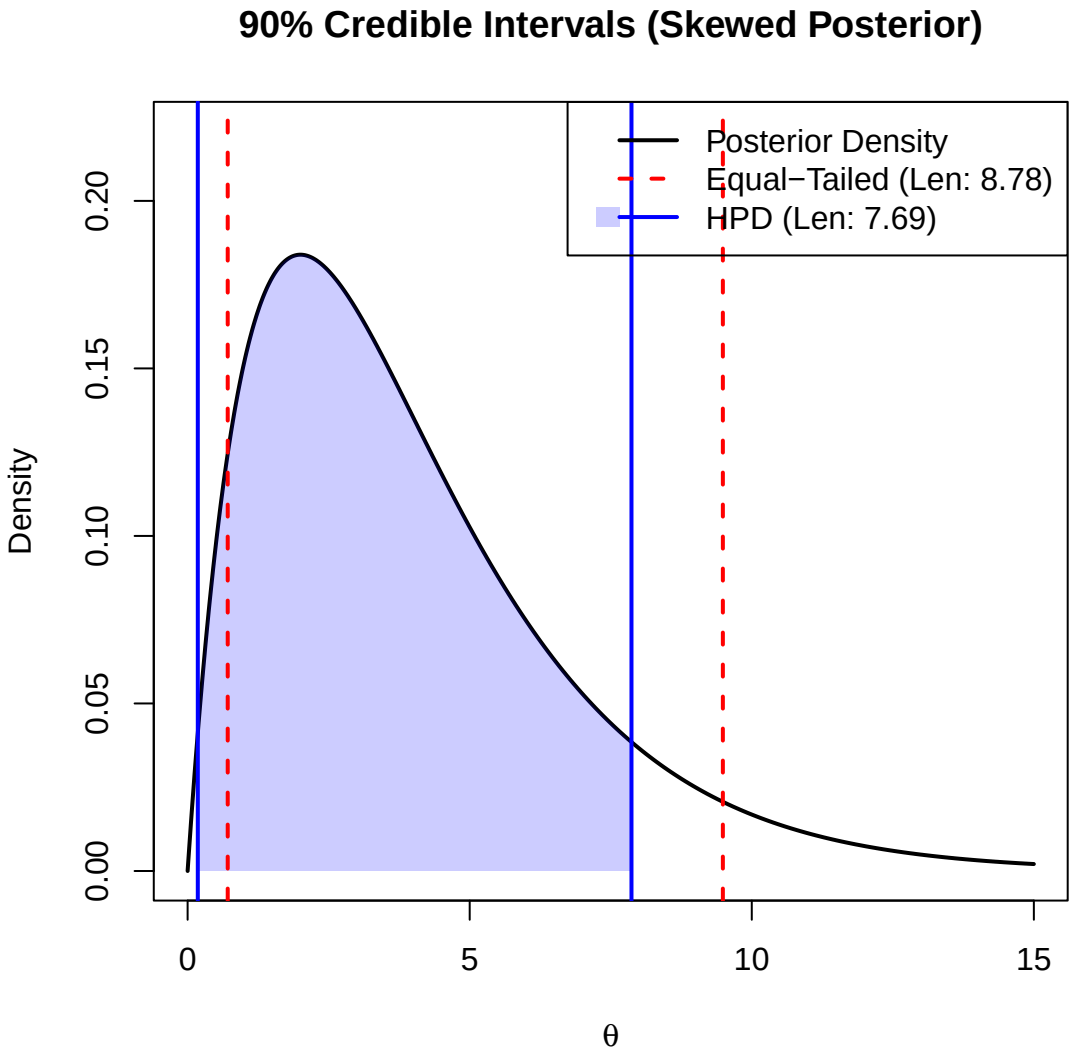


Figure 3.4: Comparison of HPD and Equal-Tailed Intervals for a Skewed Distribution

3.4.1 Binomial Minimax Estimator

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

$$d(X) = \frac{a + X}{a + b + n} \tag{3.79}$$

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

$$R(\theta, d) = E^{X|\theta}\left[\left(\theta - \frac{a + X}{a + b + n}\right)^2\right] \tag{3.80}$$

Let $c = a + b + n$.

$$R(\theta, d) = \frac{1}{c^2} E^{X|\theta}\left[(c\theta - a - X)^2\right] \tag{3.81}$$

Using the bias-variance decomposition and knowing $E^{X|\theta}[X] = n\theta$ and $E^{X|\theta}[X^2] = (n\theta)^2 + n\theta(1 - \theta)$, 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:

$$a = b = \frac{\sqrt{n}}{2} \tag{3.82}$$

Thus, the minimax estimator is:

$$d(X) = \frac{X + \sqrt{n}/2}{n + \sqrt{n}} \tag{3.83}$$

This differs from the standard MLE $\hat{p} = X/n$ and the uniform prior Bayes estimator ($a = b = 1$).
According to Theorem 2.2, let $\{\delta_n\}$ be a sequence of Bayes rules with respect to priors $\{\pi_n\}$, and let $r(\pi_n, \delta_n)$ be the associated Bayes risks. If there exists a rule δ_0 such that

$$\sup_\theta R(\theta, \delta_0) \leq \lim_{n \to \infty} r(\pi_n, \delta_n) \tag{3.84}$$

then δ_0 is a minimax estimator.
We can rewrite the Minimax estimator $d(X)$ as a linear combination of the sample proportion (MLE) $\hat{p} = X/n$ and the prior mean $p_0 = 1/2$:

$$d(X) = \underbrace{\left(\frac{n}{n + \sqrt{n}}\right)}_{w} \underbrace{\left(\frac{X}{n}\right)}_{\hat{p}} + \underbrace{\left(\frac{\sqrt{n}}{n + \sqrt{n}}\right)}_{1-w} \underbrace{\left(\frac{1}{2}\right)}_{p_0} \tag{3.85}$$

$$d(X) = w\hat{p} + (1 - w)p_0 \tag{3.86}$$

Interpretation:

- $p_0 = 0.5$: The estimator shrinks the data toward a neutral prior mean of 0.5 (representing maximum uncertainty).
- $w = \frac{n}{n+\sqrt{n}}$: The weight assigned to the data. As the sample size n increases, $w \to 1$, and the minimax estimator converges to the MLE.

Example 3.6 (Exponential Minimax Estimation)**.** Let $X_1, \dots, X_n$ be a sample from an Exp(θ) distribution with mean θ. We consider the **Scale-Invariant Loss Function**:

$$\mathcal{L}(\theta, d) = \left(\frac{d}{\theta} - 1\right)^2 \tag{3.87}$$

Likelihood and MLE
The probability density function for a single observation is $f(x_i|\theta) = \frac{1}{\theta}e^{-x_i/\theta}$. The likelihood function for the sample is:

$$L(\theta|x) = \theta^{-n}e^{-\frac{1}{\theta}\sum_{i=1}^{n} x_i} \tag{3.88}$$

The Maximum Likelihood Estimator is standard: $\hat{\theta}_{\text{MLE}} = \bar{X}$.
Minimax Estimation Setup
We propose the estimator $d_0(X) = \frac{\sum X_i}{n+1}$. To show this is a minimax estimator, we consider a sequence of priors π_k and examine the limit of their Bayes risks.
Prior Density
We assume the prior $\pi_k(\theta)$ follows an **Inverse-Gamma** distribution with shape α_k and scale β_k. The density is given by:

$$\pi_k(\theta) = \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)}\theta^{-\alpha_k-1}e^{-\beta_k/\theta}, \quad \theta > 0 \tag{3.89}$$

Posterior Analysis
Let $T = \sum X_i$. The posterior density is proportional to:

$$\pi(\theta|x) \propto \left(\theta^{-n}e^{-T/\theta}\right) \cdot \left(\theta^{-\alpha_k-1}e^{-\beta_k/\theta}\right) \propto \theta^{-(n+\alpha_k)-1}e^{-(T+\beta_k)/\theta} \tag{3.90}$$

This is an Inverse-Gamma distribution with parameters $\alpha^* = n + \alpha_k$ and $\beta^* = T + \beta_k$.
Calculation of the Bayes Estimator
Using the result derived in the **Scale-Invariant Squared Error Loss** section, the Bayes estimator is:

$$d_{\pi_k}(X) = \frac{E^{\theta|X}[\theta^{-1}]}{E^{\theta|X}[\theta^{-2}]} \tag{3.91}$$

For an Inverse-Gamma(α^*, β^*) variable, the required moments are:

- $E^{\theta|X}[\theta^{-1}] = \frac{\alpha^*}{\beta^*}$
- $E^{\theta|X}[\theta^{-2}] = \frac{\alpha^*(\alpha^*+1)}{(\beta^*)^2}$

Substituting these into the estimator formula:

$$d_{\pi_k}(X) = \frac{\frac{\alpha^*}{\beta^*}}{\frac{\alpha^*(\alpha^*+1)}{(\beta^*)^2}} = \frac{\beta^*}{\alpha^*+1} = \frac{T+\beta_k}{n+\alpha_k+1} \tag{3.92}$$

Bayes Risk Limit
The Bayes risk $r(\pi_k, d_{\pi_k})$ is the expected value of the minimum posterior loss. Substituting d_{π_k} back into the loss equation:

$$r(\pi_k, d_{\pi_k}) = 1 - \frac{(E^{\theta|X}[\theta^{-1}])^2}{E^{\theta|X}[\theta^{-2}]} = 1 - \frac{n+\alpha_k}{n+\alpha_k+1} = \frac{1}{n+\alpha_k+1} \tag{3.93}$$

Taking the limit as the prior parameters approach zero ($\alpha_k \to 0$):

$$\lim_{k\to\infty} r(\pi_k, d_{\pi_k}) = \frac{1}{n+1} \tag{3.94}$$

Minimax Verification
We compute the frequentist risk of our candidate estimator $d_0(X) = \frac{T}{n+1}$. Let $Y = T/\theta \sim \text{Gamma}(n, 1)$. Note that $E^{X|\theta}[Y] = n$ and $\text{Var}^{X|\theta}(Y) = n$.

$$\begin{aligned} R(\theta, d_0) &= E^{X|\theta}\left[\left(\frac{d_0}{\theta} - 1\right)^2\right] = E^{X|\theta}\left[\left(\frac{Y}{n+1} - 1\right)^2\right] \\ &= \text{Var}^{X|\theta}\left(\frac{Y}{n+1}\right) + \left(E^{X|\theta}\left[\frac{Y}{n+1}\right] - 1\right)^2 \\ &= \frac{n}{(n+1)^2} + \left(\frac{n}{n+1} - 1\right)^2 \\ &= \frac{1}{n+1} \end{aligned} \tag{3.95}$$

Since $R(\theta, d_0) = \lim_{k\to\infty} r(\pi_k, d_{\pi_k}) = \frac{1}{n+1}$ for all θ, d_0 is a **minimax estimator**.

3.5 Stein's Paradox and the James-stein Estimator

3.5.1 The Problem of Estimating Normal Mean

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: $\theta \sim N_p(0, \sigma^2 I)$
- Jame-Stein Estimator:

$$d^{JS}(X) = \left(1 - \frac{p-2}{||X||^2}\right) X \tag{3.96}$$

The James-Stein estimator improves upon the MLE by shrinking the individual observations toward a common mean (usually zero). The magnitude of this shrinkage depends on the total sum of squares of the observations.

- When the variance of θ is large, $||X||^2$ tends to be large, resulting in less shrinkage.
- When the variance of θ is small, $||X||^2$ is smaller, leading to a larger shrinkage factor.

The following R code simulates these two cases and displays them side-by-side with a shared y-axis for direct comparison.

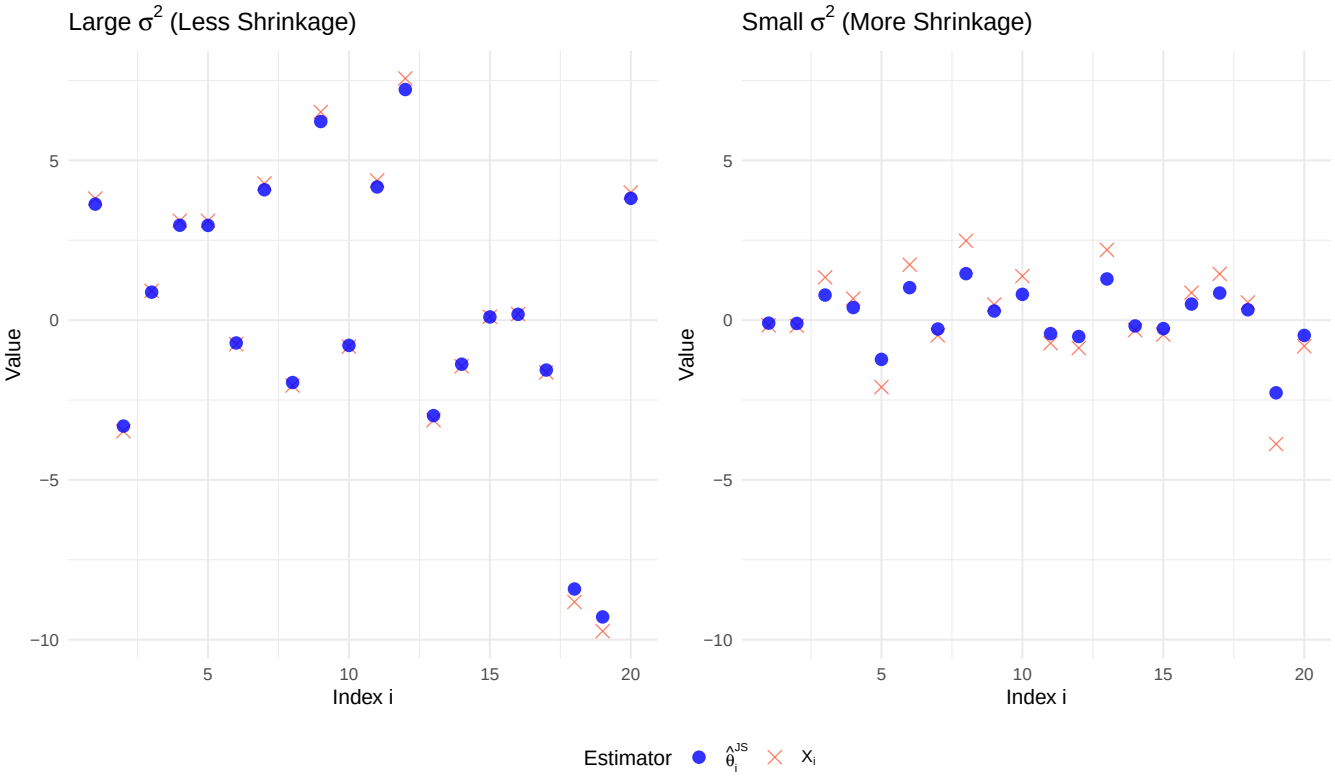


Figure 3.5: Visualization of JS Estimator

3.5.2 The Maximum Likelihood Estimator

Since the observations have the covariance matrix I (the identity matrix), the individual components $X_1, \ldots, X_p$ are independent, with $X_i \sim N(\theta_i, 1)$.

The joint likelihood function is the product of the individual probability density functions:

$$L(\theta; x) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_i - \theta_i)^2}{2}\right) \tag{3.97}$$

To find the estimator, we maximize the log-likelihood function $\ell(\theta)$:

$$\begin{aligned} \ell(\theta) &= \ln\left(\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_i - \theta_i)^2}{2}\right)\right) \\ &= \sum_{i=1}^{p} \left[\ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{(X_i - \theta_i)^2}{2}\right] \end{aligned} \tag{3.98}$$

Maximizing this sum is equivalent to minimizing the sum of squared errors $\sum(X_i - \theta_i)^2$. We can solve for each component θ_i separately. Differentiating with respect to θ_i:

$$\frac{\partial \ell}{\partial \theta_i} = (X_i - \theta_i) \tag{3.99}$$

Setting the derivative to zero gives the critical point:

$$X_i - \hat{\theta}_i = 0 \implies \hat{\theta}_{i,\text{MLE}} = X_i \tag{3.100}$$

Since this holds for every component $i = 1, \dots, p$, the Maximum Likelihood Estimator for the entire vector is simply the observation vector itself:

$$d^{\text{MLE}}(X) = X \tag{3.101}$$

3.5.3 A Bayes Rule

We first derive the Bayes rule with respect to a specific conjugate prior. Instead of using matrix notation, we can look at the problem component-wise, as the observations are independent.

Consider the model where we observe p independent components:

$$X_i|\theta_i \sim N(\theta_i, 1), \quad \text{for } i = 1, \dots, p \tag{3.102}$$

We place independent centered normal priors on each unknown parameter θ_i:

$$\theta_i \sim N(0, \sigma^2), \quad \text{for } i = 1, \dots, p \tag{3.103}$$

Since the components are independent, we can derive the Bayes rule for a single scalar component X_i estimating θ_i. The total risk will simply be the sum of the component risks.

For a single component, the posterior distribution of θ_i given X_i is Normal, with parameters determined by the standard conjugate formulas:

- **Posterior Precision (inverse variance):** The posterior precision is the sum of the prior precision and the data precision.

$$\frac{1}{v_{\text{post}}} = \frac{1}{\sigma^2} + \frac{1}{1} = \frac{1 + \sigma^2}{\sigma^2} \tag{3.104}$$

  Therefore, the posterior variance is:

$$v_{\text{post}} = \frac{\sigma^2}{1 + \sigma^2} \tag{3.105}$$

- **Posterior Mean (Bayes Estimator):** The posterior mean is the precision-weighted average of the prior mean (0) and the data mean (X_i).

$$\begin{aligned} E^{\theta_i|X_i}[\theta_i] &= v_{\text{post}} \left(\frac{0}{\sigma^2} + \frac{X_i}{1} \right) \\ &= \frac{\sigma^2}{1+\sigma^2} X_i \\ &= \left(1 - \frac{1}{1+\sigma^2} \right) X_i \end{aligned} \tag{3.106}$$

Since this holds for all i, the Bayes rule for the vector θ is applying this shrinkage factor to each component:

$$d^{\text{Bayes}}(X) = \left(1 - \frac{1}{1+\sigma^2} \right) X \tag{3.107}$$

3.5.3.1 Bayes Risk of the Bayes Rule

To compute the Bayes risk, we sum the risks of the individual components. For squared error loss, the posterior expected loss for one component is simply the posterior variance derived above:

$$E^{\theta_i|X_i}[(\theta_i - d^{\text{Bayes}}(X_i))^2] = v_{\text{post}} = \frac{\sigma^2}{1+\sigma^2} \tag{3.108}$$

The total Bayes risk is the sum of these variances over p components:

$$r(\pi, d^{\text{Bayes}}) = \sum_{i=1}^{p} \frac{\sigma^2}{1+\sigma^2} = \frac{p\sigma^2}{1+\sigma^2} \tag{3.109}$$

3.5.3.2 Minimaxity of the MLE

The James-Stein result is particularly striking when compared to the performance of the standard estimator.

Theorem 3.2 (Minimaxity of the Maximum Likelihood Estimator)**.** *Let $X \sim N_p(\theta, I)$. Under the squared error loss function $\mathcal{L}(\theta, d) = ||\theta - d||^2$, the standard Maximum Likelihood Estimator $d^0(X) = X$ is a **minimax rule**. That is, it minimizes the maximum possible risk over the parameter space:*

$$\sup_{\theta \in \mathbb{R}^p} R(\theta, d^0) = \inf_d \sup_{\theta \in \mathbb{R}^p} R(\theta, d) = p \tag{3.110}$$

Proof.
Click to view proof by least favorable prior
The risk of the MLE is $R(\theta, d^0) = p$ for all θ. Since it is a constant risk estimator, its maximum risk is simply p.
To prove it is minimax, we show that p is the limit of the Bayes risks for a sequence of conjugate priors $\theta_i \sim$

$N(0, \sigma^2)$. As derived above, the Bayes risk for the optimal Bayes estimator d^{Bayes} is:

$$r(\pi_{\sigma^2}, d^{\text{Bayes}}) = \frac{p\sigma^2}{1+\sigma^2} \tag{3.111}$$

As $\sigma^2 \to \infty$ (the prior becomes "flat"), the Bayes risk approaches p:

$$\lim_{\sigma^2 \to \infty} \frac{p\sigma^2}{1+\sigma^2} = p \tag{3.112}$$

By the property that the maximum risk of an estimator is always at least the Bayes risk of any prior, and specifically greater than or equal to the limit of Bayes risks for a sequence of priors, we establish that no estimator can have a maximum risk lower than p. Since d^0 achieves this maximum risk, it is minimax.

□

3.5.4 Stein's Lemma

i Notation: The Divergence Operator

The symbol $\nabla \cdot g(X)$ (read as "divergence of g") is simply a shorthand notation for the sum of the partial derivatives:

$$\nabla \cdot g(X) \equiv \sum_{i=1}^{p} \frac{\partial g_i(X)}{\partial X_i} \tag{3.113}$$

It represents the total "outward flow" of the vector field g from a local point.

Lemma 3.1 (Stein's Lemma)**.** *Let $X \sim N_p(\theta, I)$ be a multivariate normal random vector, and let $g : \mathbb{R}^p \to \mathbb{R}^p$ be a continuously differentiable function such that $E^{X|\theta}[|\partial g_i / \partial X_i|] < \infty$. Then:*

$$E^{X|\theta}\left[(X-\theta)^T g(X)\right] = E^{X|\theta}\left[\nabla \cdot g(X)\right] = E^{X|\theta}\left[\sum_{i=1}^{p} \frac{\partial g_i(X)}{\partial X_i}\right] \tag{3.114}$$

*The term $\nabla \cdot g(X)$ represents the **divergence** of the vector field g, which intuitively measures the local rate of expansion or outward flux of the function g at the point X; in this statistical context, it quantifies the aggregate sensitivity of the function components to changes in the data.*

Proof.
It suffices to show the result for a single component in 1 dimension, as the multivariate case follows by summation due to independence. Let $X_i \sim N(\theta_i, 1)$ and let $\phi(t)$ be the standard normal density function. The joint density is $f(x) = \prod \phi(x_j - \theta_j)$.
Consider the expectation of the i-th term:

$$E^{X|\theta}[(X_i - \theta_i) g_i(X)] = \int_{\mathbb{R}^p} (x_i - \theta_i) g_i(x) \left(\prod_{j=1}^{p} \phi(x_j - \theta_j)\right) dx \tag{3.115}$$

Focusing on the integral with respect to x_i:

$$\int_{-\infty}^{\infty}(x_i-\theta_i)\phi(x_i-\theta_i)g_i(x)dx_i \tag{3.116}$$

Recall that $\phi'(z)=-z\phi(z)$. Therefore, $(x_i-\theta_i)\phi(x_i-\theta_i)=-\frac{\partial}{\partial x_i}\phi(x_i-\theta_i)$. We use integration by parts with:

$$u=g_i(x) \quad \text{and} \quad dv=-\frac{\partial}{\partial x_i}\phi(x_i-\theta_i)dx_i \tag{3.117}$$

Thus:

$$\int_{-\infty}^{\infty}g_i(x)(x_i-\theta_i)\phi(x_i-\theta_i)dx_i=[-g_i(x)\phi(x_i-\theta_i)]_{-\infty}^{\infty}+\int_{-\infty}^{\infty}\frac{\partial g_i(x)}{\partial x_i}\phi(x_i-\theta_i)dx_i \tag{3.118}$$

Assuming $g(x)$ does not grow exponentially fast, the boundary term vanishes. The remaining integral is the expectation of the partial derivative. Summing over all $i=1\dots p$ gives the divergence $\nabla\cdot g(X)$.

□

In high-dimensional statistics, Stein's Lemma is often expressed using the inner product of the random vector and the function vector field, which highlights the alignment between the data and the transformation.

Corollary 3.1 (Stein's Lemma (Vector Form)). *Let $X\sim N_p(\theta,I)$ and $g:\mathbb{R}^p\to\mathbb{R}^p$ be a weakly differentiable function. Then:*

$$E^{X|\theta}[X^Tg(X)]=\theta^TE^{X|\theta}[g(X)]+E^{X|\theta}[\nabla\cdot g(X)] \tag{3.119}$$

Remark: Connection to Non-Central χ^2 Moments

This identity provides an elegant way to derive the mean of a non-central chi-square distribution without performing complex integration.
Consider the case where $g(X)=X$. Here, $\nabla\cdot X=p$. Plugging this into the vector form:

$$E^{X|\theta}[X^TX]=\theta^TE^{X|\theta}[X]+E^{X|\theta}[p] \tag{3.120}$$

Since $E^{X|\theta}[X]=\theta$, we immediately obtain:

$$E^{X|\theta}[\|X\|^2]=\|\theta\|^2+p \tag{3.121}$$

This is precisely the mean of a $\chi^2_p(\lambda)$ distribution with non-centrality parameter $\lambda=\|\theta\|^2$. Essentially, Stein's Lemma decomposes the second moment into the **signal component** ($\|\theta\|^2$) and the **geometric noise component** (p).

Lemma 3.2 (Stein's Lemma for Radial Fields). *Let $X\sim N_p(\theta,I)$ and consider a radial vector field of the form $g(X)=c(\|X\|^2)X$, where $c:\mathbb{R}\to\mathbb{R}$ is a differentiable scalar function. Then:*

$$E^{X|\theta}\left[(X-\theta)^Tg(X)\right]=E^{X|\theta}\left[p\cdot c(\|X\|^2)+2\|X\|^2\cdot c'(\|X\|^2)\right] \tag{3.122}$$

where $c'(z) = \frac{d}{dz}c(z)$.

Proof.
We apply the general Stein's Lemma by calculating the divergence of the radial field $g(X) = c(\|X\|^2)X$. Using the product rule for divergence:

$$\nabla \cdot (c(\|X\|^2)X) = c(\|X\|^2)(\nabla \cdot X) + X^T(\nabla c(\|X\|^2)) \tag{3.123}$$

Step 1: The geometric spread. The divergence of the identity map X in p dimensions is simply the sum of the partial derivatives of each component with respect to itself:

$$\nabla \cdot X = \sum_{i=1}^{p} \frac{\partial X_i}{\partial X_i} = p \tag{3.124}$$

Step 2: The radial stretch. To find $\nabla c(\|X\|^2)$, we use the chain rule. Let $h(X) = \|X\|^2 = \sum X_i^2$. Then $\nabla h(X) = 2X$.

$$\nabla c(\|X\|^2) = c'(\|X\|^2)\nabla(\|X\|^2) = 2c'(\|X\|^2)X \tag{3.125}$$

Substituting this back into the divergence formula:

$$\begin{aligned} \nabla \cdot g(X) &= p \cdot c(\|X\|^2) + X^T(2c'(\|X\|^2)X) \\ &= p \cdot c(\|X\|^2) + 2c'(\|X\|^2)\|X\|^2 \end{aligned} \tag{3.126}$$

Taking the expectation of both sides completes the proof.
Connection to the χ^2 Distribution
This version of the lemma is particularly useful because when $\theta = 0$, the quantity $\|X\|^2$ follows a central χ_p^2 distribution.

- **Verifying the Mean:** If we set $c(\|X\|^2) = 1$, then $g(X) = X$. The lemma gives $E^{X|\theta=0}[\|X\|^2] = p + 2\|X\|^2(0) = p$, which is the expected value of a χ_p^2 variable.
- **The James-Stein Weight:** If we set $c(\|X\|^2) = \frac{1}{\|X\|^2}$, then $c'(z) = -\frac{1}{z^2}$.

$$\nabla \cdot g(X) = \frac{p}{\|X\|^2} + 2\|X\|^2\left(-\frac{1}{\|X\|^4}\right) = \frac{p-2}{\|X\|^2} \tag{3.127}$$

This explains why the $p - 2$ constant appears in the James-Stein estimator—it is the net result of p dimensions of "spreading" minus 2 dimensions of "radial thinning."

□

Example 3.7 (An Example for Verifying Stein's Lemma)**.** Let $X \sim N(\theta, 1)$ be a univariate normal random variable with unit variance. Let $g(x) = x^2$. Stein's Lemma states that:

$$E^{X|\theta}\left[(X-\theta)g(X)\right] = E^{X|\theta}\left[g'(X)\right] \tag{3.128}$$

Step 1: Calculate the Right-Hand Side (RHS) First, we find the derivative of $g(x)$:

$$g'(x) = \frac{d}{dx}(x^2) = 2x \tag{3.129}$$

Now, compute the expectation of the derivative:

$$\text{RHS} = E^{X|\theta}[g'(X)] = E^{X|\theta}[2X] = 2E^{X|\theta}[X] = 2\theta \tag{3.130}$$

Step 2: Calculate the Left-Hand Side (LHS) We evaluate the expectation of the cross-product term. Substitute $X = Z + \theta$, where $Z \sim N(0,1)$ is a standard normal variable. Then $X - \theta = Z$.

$$\begin{aligned} \text{LHS} &= E^{X|\theta}\left[(X-\theta)X^2\right] \\ &= E^{X|\theta}\left[Z(Z+\theta)^2\right] \quad \text{where } Z \sim N(0,1) \\ &= E^{X|\theta}\left[Z(Z^2 + 2\theta Z + \theta^2)\right] \\ &= E^{X|\theta}[Z^3] + 2\theta E^{X|\theta}[Z^2] + \theta^2 E^{X|\theta}[Z] \end{aligned} \tag{3.131}$$

We use the known moments of the standard normal distribution Z:

- $E[Z] = 0$ (mean)
- $E[Z^2] = 1$ (variance)
- $E[Z^3] = 0$ (skewness of symmetric distribution)

Substituting these values back:

$$\text{LHS} = 0 + 2\theta(1) + \theta^2(0) = 2\theta \tag{3.132}$$

Conclusion We observe that:

$$\text{LHS} = 2\theta \quad \text{and} \quad \text{RHS} = 2\theta \tag{3.133}$$

Thus, Stein's Lemma holds for this specific case.

Example 3.8 (A Radial Field Example Verifying Stein's Lemma)**.** Let $X \sim N_p(\theta, I)$ and $g(X) = \|X\|^2 X$. We verify the Radial Field Lemma:

$$E^{X|\theta}\left[(X-\theta)^T g(X)\right] = E^{X|\theta}\left[p \cdot c(\|X\|^2) + 2\|X\|^2 \cdot c'(\|X\|^2)\right] \tag{3.134}$$

Here, $c(z) = z$, which implies $c'(z) = 1$.
RHS (Divergence): Using the radial formula:

$$\begin{aligned} \nabla \cdot g(X) &= p(\|X\|^2) + 2\|X\|^2(1) \\ &= (p+2)\|X\|^2 \end{aligned} \tag{3.135}$$

The expectation is $(p+2)E^{X|\theta}[\|X\|^2]$. Since $\|X\|^2$ is a non-central χ^2_p with non-centrality parameter $\|\theta\|^2$, we know $E^{X|\theta}[\|X\|^2] = p + \|\theta\|^2$. Thus, $\text{RHS} = (p+2)(p+\|\theta\|^2)$.
LHS (Alignment):

$$\begin{aligned} E^{X|\theta}\left[(X-\theta)^T(\|X\|^2 X)\right] &= E^{X|\theta}\left[\|X\|^2(X^T X - \theta^T X)\right] \\ &= E^{X|\theta}\left[\|X\|^4 - \theta^T X \|X\|^2\right] \end{aligned} \tag{3.136}$$

To simplify, let $X = \theta + Z$ where $Z \sim N_p(0, I)$. Recall the moments of the non-central chi-square distribution or expand the terms: $E[\|X\|^4] = p(p+2) + 2(p+2)\|\theta\|^2 + \|\theta\|^4$. For the cross term $E[\theta^T X \|X\|^2]$, we find it equals $\|\theta\|^4 + (p+2)\|\theta\|^2$.
Subtracting these:

$$\begin{aligned} \text{LHS} &= [p(p+2) + 2(p+2)\|\theta\|^2 + \|\theta\|^4] - [\|\theta\|^4 + (p+2)\|\theta\|^2] \\ &= p(p+2) + (p+2)\|\theta\|^2 \\ &= (p+2)(p + \|\theta\|^2) \end{aligned} \tag{3.137}$$

Conclusion
The results match exactly. The alignment of the cubic radial field with the noise is perfectly predicted by the sum of its geometric expansion ($p\|X\|^2$) and its radial stretch ($2\|X\|^2$).

3.5.5 Inadmissibility of the MLE in High Dimensions (Stein's Phenomenon)

Theorem 3.3. *Let $X \sim N_p(\theta, I)$ be a p-dimensional random vector with $p \geq 3$. Under the squared error loss function $\mathcal{L}(\theta, d) = ||\theta - d||^2$, the standard Maximum Likelihood Estimator $d^0(X) = X$ is **inadmissible**.*

Proof of Inadmissibility. To show that $d^0(X) = X$ is inadmissible, we compare its risk to that of the James-Stein estimator $d^{JS}(X)$.
Let $g(X) = c(\|X\|^2)X$ where $c(\|X\|^2) = \frac{p-2}{\|X\|^2}$. We can write the James-Stein estimator as $d^{JS}(X) = X - g(X)$.
The risk is the expected squared error loss:

$$\begin{aligned} R(\theta, d^{JS}) &= E^{X|\theta}\left[||(X-\theta) - g(X)||^2\right] \\ &= E^{X|\theta}\left[||X-\theta||^2\right] - 2E^{X|\theta}\left[(X-\theta)^T g(X)\right] + E^{X|\theta}\left[||g(X)||^2\right] \end{aligned} \tag{3.138}$$

The first term is the risk of the MLE, which is p.
For the second term, we apply **Stein's Lemma for Radial Fields** (Lemma 3.2). We first compute the scalar function and its derivative:

$$c(z) = \frac{p-2}{z} \implies c'(z) = -\frac{p-2}{z^2} \tag{3.139}$$

Substituting these into the radial divergence formula from Lemma 3.2:

$$\begin{aligned} \nabla \cdot g(X) &= p \cdot c(\|X\|^2) + 2\|X\|^2 \cdot c'(\|X\|^2) \\ &= p\left(\frac{p-2}{\|X\|^2}\right) + 2\|X\|^2\left(-\frac{p-2}{\|X\|^4}\right) \\ &= \frac{p(p-2)}{\|X\|^2} - \frac{2(p-2)}{\|X\|^2} \\ &= \frac{(p-2)^2}{\|X\|^2} \end{aligned} \tag{3.140}$$

Applying the lemma to the cross-term:

$$2E^{X|\theta}\left[(X-\theta)^T g(X)\right] = 2E^{X|\theta}[\nabla \cdot g(X)] = 2(p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] \tag{3.141}$$

The third term in the risk expansion is the squared magnitude of the shrinkage:

$$\|g(X)\|^2 = \left\|\frac{p-2}{\|X\|^2}X\right\|^2 = \frac{(p-2)^2}{\|X\|^4}\|X\|^2 = \frac{(p-2)^2}{\|X\|^2} \tag{3.142}$$

Substituting these results back into the risk equation:

$$\begin{aligned} R(\theta, d^{JS}) &= p - 2(p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] + (p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] \\ &= p - (p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] \end{aligned} \tag{3.143}$$

Since $p \geq 3$, the constant $(p-2)^2$ is strictly positive. Because $1/\|X\|^2 > 0$ with probability 1, the risk of the James-Stein estimator is strictly less than p for all $\theta \in \mathbb{R}^p$.
Thus, d^{JS} dominates the MLE, proving the MLE is inadmissible. □

3.5.6 How much JS Estimator Improves over MLE

The exact risk function of the James-Stein estimator $d^{JS}(X)$ under squared error loss for $X \sim N_p(\theta, I)$ is given by:

$$R(\theta, d^{JS}) = p - (p-2)^2 E\left[\frac{1}{\chi^2_p(\|\theta\|^2/2)}\right] \tag{3.144}$$

where $\chi^2_p(\|\theta\|^2/2)$ is a non-central chi-square random variable with p degrees of freedom and non-centrality parameter $\lambda = \|\theta\|^2/2$.

Using the approximation $E[1/\chi^2_p(\lambda)] \approx 1/E[\chi^2_p(\lambda)] = 1/(p+\|\theta\|^2)$, we can see the approximate behavior of the risk:

$$R(\theta, d^{JS}) \approx p - \frac{(p-2)^2}{p+\|\theta\|^2} \tag{3.145}$$

- **Aggressive Shrinkage near the Origin**: When $\|\theta\|^2$ is small, the denominator $p+\|\theta\|^2$ is small, making the subtracted term large. This results in a risk substantially lower than the MLE risk of p.
- **Diminishing Improvement with Large Signal**: As $\|\theta\|^2$ becomes large, the term $\frac{(p-2)^2}{p+\|\theta\|^2}$ approaches zero. Consequently, the risk of the James-Stein estimator approaches p, and the improvement over the MLE becomes negligible.

Risk Ratio near the Origin: As the true parameter vector shrinks to zero ($\|\theta\| \to 0$), the ratio of the risks converges to a constant fraction. Using the exact expectation $E^{X|\theta=0}[1/\|X\|^2] = 1/(p-2)$:

$$\lim_{\|\theta\|\to 0} \frac{R(\theta, d^{JS})}{R(\theta, d^{\text{MLE}})} = \frac{p - (p-2)^2 \left(\frac{1}{p-2}\right)}{p} = \frac{p-(p-2)}{p} = \frac{2}{p} \tag{3.146}$$

For a dimension like $p = 10$, the James-Stein estimator incurs only 20% of the risk of the MLE near the origin.

> **i** Is d^{JS} Minimax?
>
> Yes. Since the MLE is minimax with constant risk p, the minimax risk value for this problem is p. Because $R(\theta, d^{JS}) < p$ for all θ and $\lim_{\|\theta\|\to\infty} R(\theta, d^{JS}) = p$, the maximum risk of the James-Stein estimator is exactly p. Therefore, d^{JS} achieves the minimax risk level and is a minimax estimator.

3.5.7 Using Normalized Loss (Optional)

We consider the **Normalized Squared Error Loss** function, which penalizes errors relative to the magnitude of the true parameter vector:

$$\mathcal{L}(\theta, d) = \frac{\|d-\theta\|^2}{\|\theta\|^2}, \quad \theta \neq 0 \tag{3.147}$$

1. Risk of the MLE ($d^{\text{MLE}} = X$)

The risk of the Maximum Likelihood Estimator is straightforward because the standard Mean Squared Error (MSE) of X is constant (p):

$$R(\theta, d^{\text{MLE}}) = E^{X|\theta}\left[\frac{\|X-\theta\|^2}{\|\theta\|^2}\right] = \frac{1}{\|\theta\|^2} E^{X|\theta}\left[\|X-\theta\|^2\right] \tag{3.148}$$

Since $X \sim N_p(\theta, I)$, we have $E^{X|\theta}[\|X-\theta\|^2] = p$.

$$R(\theta, d^{\text{MLE}}) = \frac{p}{\|\theta\|^2} \tag{3.149}$$

2. Risk of the James-Stein Estimator (d^{JS})

For the James-Stein estimator $d^{JS} = \left(1 - \frac{p-2}{\|X\|^2}\right) X$, we utilize the known result for its standard MSE risk:

$$E^{X|\theta}[\|d^{JS} - \theta\|^2] = p - (p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] \tag{3.150}$$

The risk under the normalized loss is simply this term scaled by $1/\|\theta\|^2$:

$$R(\theta, d^{JS}) = \frac{1}{\|\theta\|^2}\left(p - (p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right]\right) \tag{3.151}$$

3. Comparison and Dominance

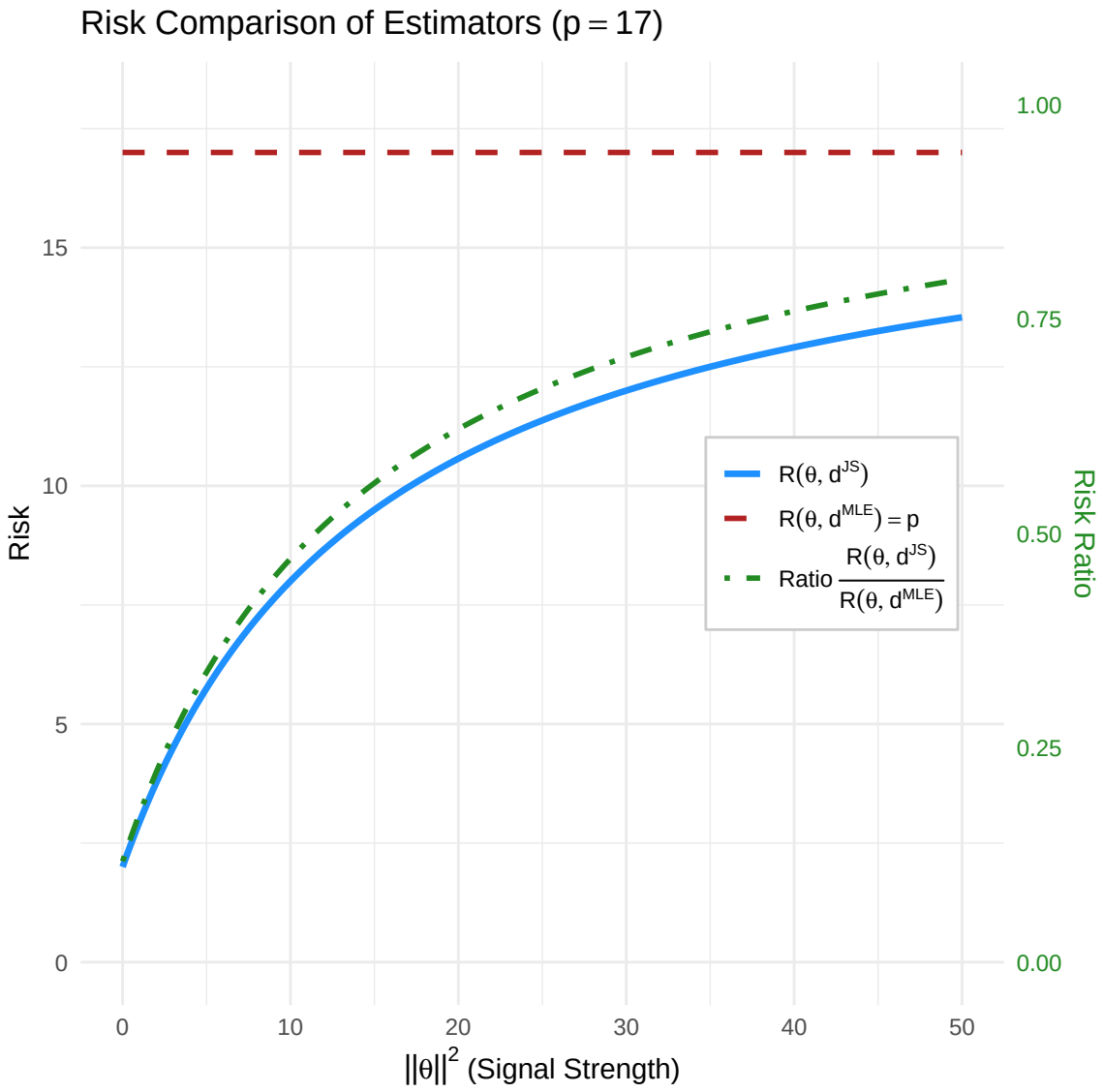


Figure 3.6: Risk comparison of James-Stein vs MLE (p=10). Note: Ratio uses the right axis.

We compare the risks by taking the difference:

$$R(\theta, d^{\text{MLE}}) - R(\theta, d^{JS}) = \frac{1}{\|\theta\|^2}(p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] \tag{3.152}$$

Since $(p-2)^2 > 0$ (for $p \geq 3$) and the expectation of a positive random variable is positive, this difference is strictly positive for all $\theta \neq 0$.

$$R(\theta, d^{JS}) < R(\theta, d^{\text{MLE}}) \tag{3.153}$$

- **Global Dominance:** The James-Stein estimator dominates the MLE under this loss function as well, achieving lower risk everywhere in the parameter space.
- **Behavior near $\theta \approx 0$:** As $\|\theta\| \to 0$, both risks diverge to infinity. We analyze their relative performance by examining the ratio of the risks:

$$\frac{R(\theta, d^{JS})}{R(\theta, d^{\text{MLE}})} = \frac{p - (p-2)^2 E^{X|\theta}[1/\|X\|^2]}{p} = 1 - \frac{(p-2)^2}{p} E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] \tag{3.154}$$

At $\theta = 0$, $\|X\|^2 \sim \chi^2_p$, and the expectation is $E^{X|\theta=0}[1/\|X\|^2] = 1/(p-2)$. Substituting this into the ratio:

$$\lim_{\|\theta\|\to 0} \frac{R(\theta, d^{JS})}{R(\theta, d^{\text{MLE}})} = 1 - \frac{(p-2)^2}{p(p-2)} = 1 - \frac{p-2}{p} = \frac{2}{p} \tag{3.155}$$

Thus, near the origin, the James-Stein estimator reduces the risk by a factor of $p/2$. For large dimensions (e.g., $p = 10$), the JS estimator has only 20% of the risk of the MLE.

3.5.8 Bayes Risk of James-stein Estimator (Optional)

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 3.4 (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 $\theta \sim N(0, \sigma^2 I)$ is:*

$$r(\pi, d^{JS}) = \frac{p\sigma^2 + 2}{\sigma^2 + 1} \tag{3.156}$$

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:

$$R(\theta, d^{JS}) = p - (p-2)^2 E^{X|\theta}\left[\frac{1}{||X||^2}\right] \tag{3.157}$$

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

$$r(\pi, d^{JS}) = \int R(\theta, d^{JS})\pi(\theta)d\theta = p - (p-2)^2 E^{\pi}\left[E^{X|\theta}\left(\frac{1}{||X||^2}\right)\right] \tag{3.158}$$

By the law of iterated expectations, $E^{\pi}[E^{X|\theta}(\cdot)]$ 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 $X \sim N(0, (1+\sigma^2)I)$.

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

$$E^X\left[\frac{1}{||X||^2}\right] = \frac{1}{1+\sigma^2}E\left[\frac{1}{\chi^2_p}\right] = \frac{1}{1+\sigma^2}\cdot\frac{1}{p-2} \tag{3.159}$$

Substituting this back into the risk equation:

$$\begin{aligned} r(\pi, d^{JS}) &= p - (p-2)^2 \cdot \frac{1}{(p-2)(1+\sigma^2)} \\ &= p - \frac{p-2}{1+\sigma^2} \\ &= \frac{p(1+\sigma^2) - (p-2)}{1+\sigma^2} \\ &= \frac{p\sigma^2 + p - p + 2}{1+\sigma^2} = \frac{p\sigma^2+2}{\sigma^2+1} \end{aligned} \tag{3.160}$$

□

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:

$$r(\pi, d) = E^X[E^{\theta|X}[\mathcal{L}(\theta, d(X))]] \tag{3.161}$$

Step 1: Posterior Expected Loss

The posterior distribution of θ given x is:

$$\theta|x \sim N\left(\frac{\sigma^2}{1+\sigma^2}x, \frac{\sigma^2}{1+\sigma^2}I\right) \tag{3.162}$$

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

$$E^{\theta|X}[||\theta - d^{JS}(X)||^2] = \text{tr}(\text{Var}^{\theta|X}(\theta)) + ||E^{\theta|X}[\theta] - d^{JS}(X)||^2 \tag{3.163}$$

- **Trace term:**

$$\text{tr}\left(\frac{\sigma^2}{1+\sigma^2}I_p\right) = \frac{p\sigma^2}{1+\sigma^2} \tag{3.164}$$

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

$$E^{\theta|X}[\theta] - d^{JS}(X) = \left((1-B) - \left(1 - \frac{p-2}{||X||^2}\right)\right) X = \left(\frac{p-2}{||X||^2} - B\right) X \tag{3.165}$$

Squaring the norm gives:

$$\left(\frac{p-2}{||X||^2} - B\right)^2 ||X||^2 = \frac{(p-2)^2}{||X||^2} - 2B(p-2) + B^2||X||^2 \tag{3.166}$$

Step 2: Expectation with respect to Marginal X
We now take the expectation $E^X[\cdot]$ of the posterior loss. Recall $X \sim N(0, (1+\sigma^2)I)$, so $E^X[||X||^2] = p(1+\sigma^2)$ and $E^X[1/||X||^2] = \frac{1}{(p-2)(1+\sigma^2)}$.

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

$$\begin{aligned} E^X\left[\frac{(p-2)^2}{||X||^2} - \frac{2(p-2)}{1+\sigma^2} + \frac{||X||^2}{(1+\sigma^2)^2}\right] &= (p-2)^2 \frac{1}{(p-2)(1+\sigma^2)} - \frac{2(p-2)}{1+\sigma^2} + \frac{p(1+\sigma^2)}{(1+\sigma^2)^2} \\ &= \frac{p-2}{1+\sigma^2} - \frac{2p-4}{1+\sigma^2} + \frac{p}{1+\sigma^2} \\ &= \frac{p-2-2p+4+p}{1+\sigma^2} \\ &= \frac{2}{1+\sigma^2} \end{aligned} \tag{3.167}$$

Step 3: Combine Terms

$$r(\pi, d^{JS}) = \underbrace{\frac{p\sigma^2}{1+\sigma^2}}_{\text{Variance Part}} + \underbrace{\frac{2}{1+\sigma^2}}_{\text{Bias Part}} = \frac{p\sigma^2+2}{\sigma^2+1} \tag{3.168}$$

Both methods yield the same result.

□

3.5.9 Practical Application: One-way ANOVA and "Borrowing Strength"

Example 3.9. 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:

$$\hat{\theta}_i^{JS} = \bar{X} + \left(1 - \frac{(p-3)\sigma^2}{\sum(X_i - \bar{X})^2}\right)(X_i - \bar{X}) \tag{3.169}$$

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.

- **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.
- **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.

3.5.10 Why Is This Paradoxical?

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

- **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.
- **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.
- **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.

3.5.11 What We Learned

- **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.
- **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.

- **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.

3.6 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.

3.6.1 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:
$$m(x|\eta) = \int f(x|\theta)\pi(\theta|\eta)d\theta \tag{3.170}$$
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:
$$\hat{\eta} = \arg\max_{\eta} m(x|\eta) \tag{3.171}$$
4. **Posterior Inference:** We proceed with standard Bayesian inference, but we substitute the estimated estimate $\hat{\eta}$ into the posterior:
$$\pi(\theta|x, \hat{\eta}) \propto f(x|\theta)\pi(\theta|\hat{\eta}) \tag{3.172}$$

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.

3.6.2 Deriving James-Stein as Empirical Bayes

The James-Stein estimator can be viewed as an **Empirical Bayes** procedure, where the hyperparameters of the prior are estimated directly from the data rather than being specified *a priori*.

Model:

- **Likelihood:** $X_i|\mu_i \sim N(\mu_i, 1)$ for $i = 1, \dots, p$.
- **Prior:** $\mu_i \sim N(0, \sigma^2)$, where σ^2 is an unknown hyperparameter.

Step 1: The Ideal Bayes Estimator

If σ^2 were known, the posterior distribution of μ_i would be Normal. The optimal estimator under squared error loss is the posterior mean:

$$E^{\mu_i|X_i,\sigma^2}[\mu_i] = \frac{\sigma^2}{1+\sigma^2} X_i = \left(1 - \frac{1}{1+\sigma^2}\right) X_i \tag{3.173}$$

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

Step 2: Marginal Estimation

The marginal distribution of the data (integrating out μ_i) is:

$$X_i \sim N(0, 1+\sigma^2) \tag{3.174}$$

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

$$S \sim (1+\sigma^2)\chi_p^2 \tag{3.175}$$

Step 3: Estimating the Shrinkage Factor

We need an estimator for $B = \frac{1}{1+\sigma^2}$. From the properties of the inverse Chi-square distribution, we know $E^X[1/\chi_p^2] = \frac{1}{p-2}$ for $p > 2$. Therefore:

$$E^X\left[\frac{p-2}{S}\right] = \frac{p-2}{1+\sigma^2} E^X\left[\frac{1}{\chi_p^2}\right] = \frac{p-2}{1+\sigma^2} \cdot \frac{1}{p-2} = \frac{1}{1+\sigma^2} = B \tag{3.176}$$

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

Step 4: The Empirical Bayes Rule

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

$$\delta^{EB}(X) = \left(1 - \hat{B}\right) X = \left(1 - \frac{p-2}{||X||^2}\right) X \tag{3.177}$$

Remarks:

(1) **Adaptive Shrinkage:** The James-Stein estimator automatically adjusts the amount of shrinkage based on the observed total magnitude $\|X\|^2$. If the data suggests the true means are spread far from zero, $\|X\|^2$ will be large, $\hat{B}$ will be small, and we shrink less.

(2) **Unbiasedness of B:** Interestingly, while $\hat{B}$ is an unbiased estimator of the shrinkage factor, the resulting James-Stein estimator itself is biased toward the origin. This is a classic example of sacrificing unbiasedness to minimize total risk.

3.7 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.

3.7.1 Hierarchical Model Structure

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

$$\begin{aligned} \text{Level 1 (Data Likelihood):} \quad & X_i|\mu_i, \sigma^2 \sim f(x_i|\mu_i, \sigma^2) \\ \text{Level 2 (Parameters):} \quad & \mu_i|\theta, \tau^2 \sim \pi(\mu_i|\theta, \tau^2) \\ \text{Level 3 (Hyperparameters):} \quad & \theta, \tau^2 \sim \pi(\theta, \tau^2) \end{aligned} \tag{3.178}$$

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

$$p(\mu, \theta, \tau^2|X) \propto \left[\prod_{i=1}^{n} f(x_i|\mu_i, \sigma^2)\pi(\mu_i|\theta, \tau^2)\right] \pi(\theta, \tau^2) \tag{3.179}$$

3.7.2 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.

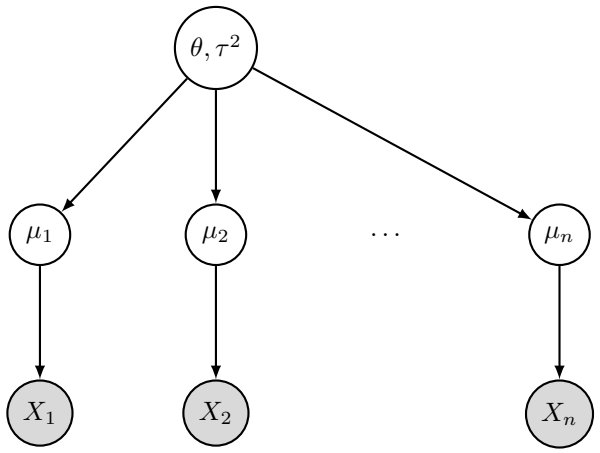


Figure 3.7: Hierarchical Tree Structure

3.7.3 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.

3.7.3.1 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:

$$\begin{aligned}
\theta_1^{(t+1)} &\sim p(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_k^{(t)}) \\
\theta_2^{(t+1)} &\sim p(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_k^{(t)}) \\
&\vdots \\
\theta_k^{(t+1)} &\sim p(\theta_k|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \dots, \theta_{k-1}^{(t+1)})
\end{aligned} \tag{3.180}$$

Example 3.10 (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:

$$\mu_i^{(t+1)} \sim N\left(\frac{\tau^2 X_i + \sigma^2 \theta^{(t)}}{\sigma^2 + \tau^2}, \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}\right) \tag{3.181}$$

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

$$\theta^{(t+1)} \sim N\left(\frac{1}{n}\sum_{i=1}^{n} \mu_i^{(t+1)}, \frac{\tau^2}{n}\right) \tag{3.182}$$

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

3.7.3.2 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:

$$\alpha = \min\left(1, \frac{p(\theta^*|X)q(\theta^{(t)}|\theta^*)}{p(\theta^{(t)}|X)q(\theta^*|\theta^{(t)})}\right) \tag{3.183}$$

- 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.

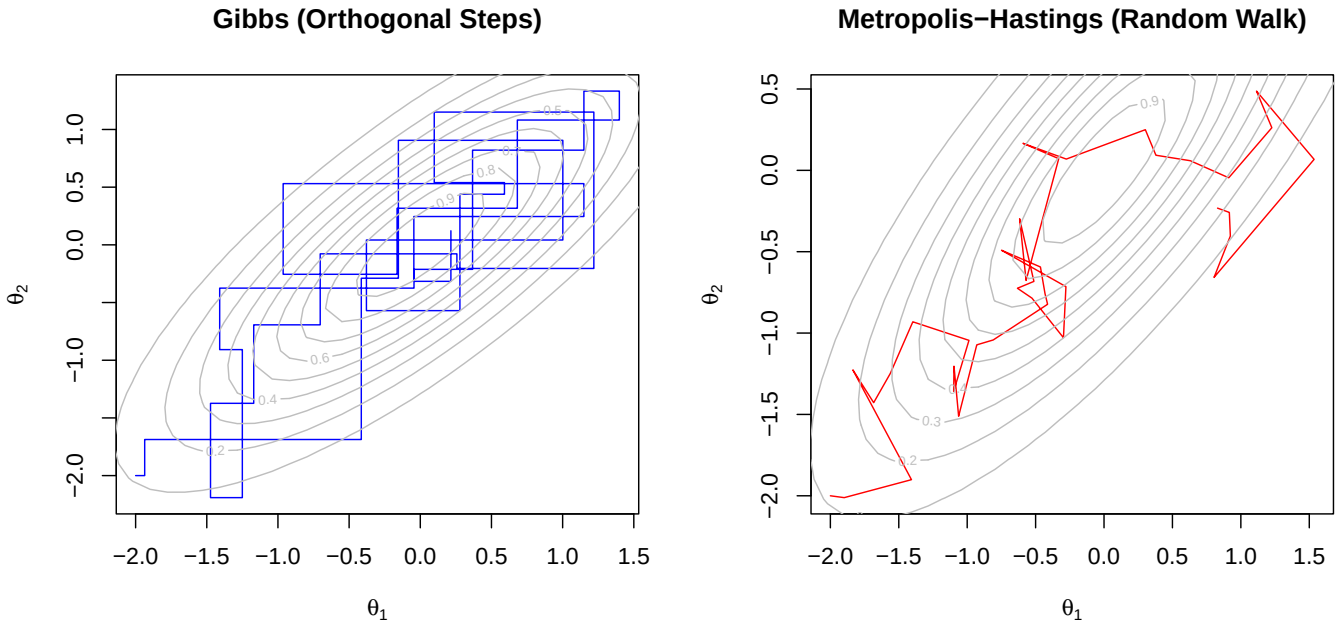


Figure 3.8: Comparison of Sampling Paths

3.8 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.

3.8.1 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 $\sigma^2 = 1$.

$$x_i = \sqrt{n_i} \arcsin\left(2\frac{y_i}{n_i} - 1\right) \tag{3.184}$$

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

3.8.2 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:

$$\mu_i^{\text{season}} = \sqrt{n_i} \arcsin\left(2\frac{Y_i}{N_i} - 1\right) \tag{3.185}$$

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.

Table 3.6: 1998 MLB Statistics: Raw Counts, Probabilities, and Transformed Data

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

3.8.3 Methods for Estimating μ_i (transformed Scale)

3.8.3.1 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.

$$\hat{\mu}_i^{MLE} = X_i \tag{3.186}$$

```
# 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)
```

3.8.3.2 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.

$$\hat{\mu}_i^{JS} = \bar{X} + \left(1 - \frac{k-3}{\sum(X_i - \bar{X})^2}\right)(X_i - \bar{X}) \tag{3.187}$$

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

```
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)
```

3.8.3.3 Method 3: Fully Bayesian MCMC (brms)

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

$$\begin{aligned} X_i &\sim N(\mu_i, 1) \\ \mu_i &\sim N(\theta, \tau^2) \\ \theta &\sim N(0, 10) \\ \tau &\sim \text{Cauchy}(0, 2) \end{aligned} \tag{3.188}$$

```
baseball_data$sei <- rep(1, length(baseball_data$x))
# 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)
```

3.8.4 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.

Table 3.7: 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

Table 3.7: Comparison of Estimates (Sorted by Pre-season x_i)

Player	x_i (MLE)	$\widehat{\mu}_{JS}$	$\widehat{\mu}_{Bayes}$	μ_{true}
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 Squared 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")

# 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))
```

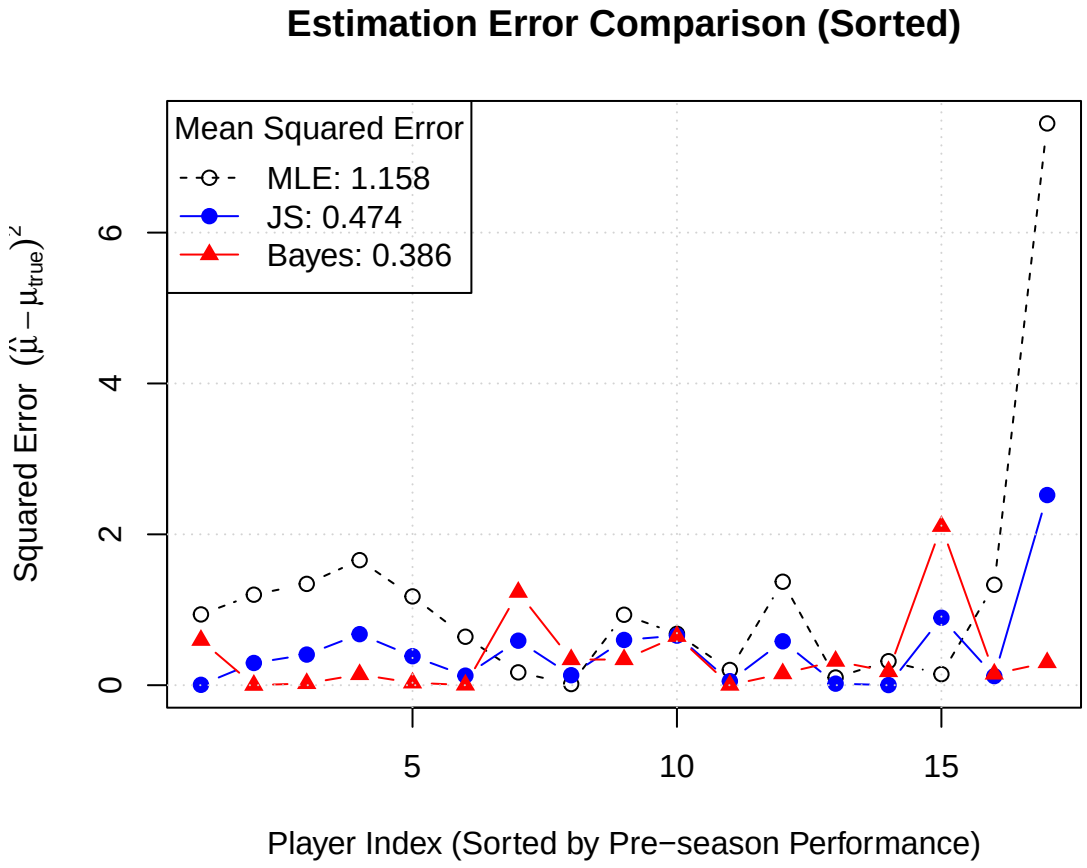


Figure 3.9: Squared Error by Sorted Player Index (Transformed Scale)

3.8.5 Methods for Estimating p_i Directly

3.8.5.1 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:

$$\hat{p}_i = \frac{1}{2}\left(\sin\left(\frac{\hat{\mu}_i}{\sqrt{n_i}}\right) + 1\right) \tag{3.189}$$

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.

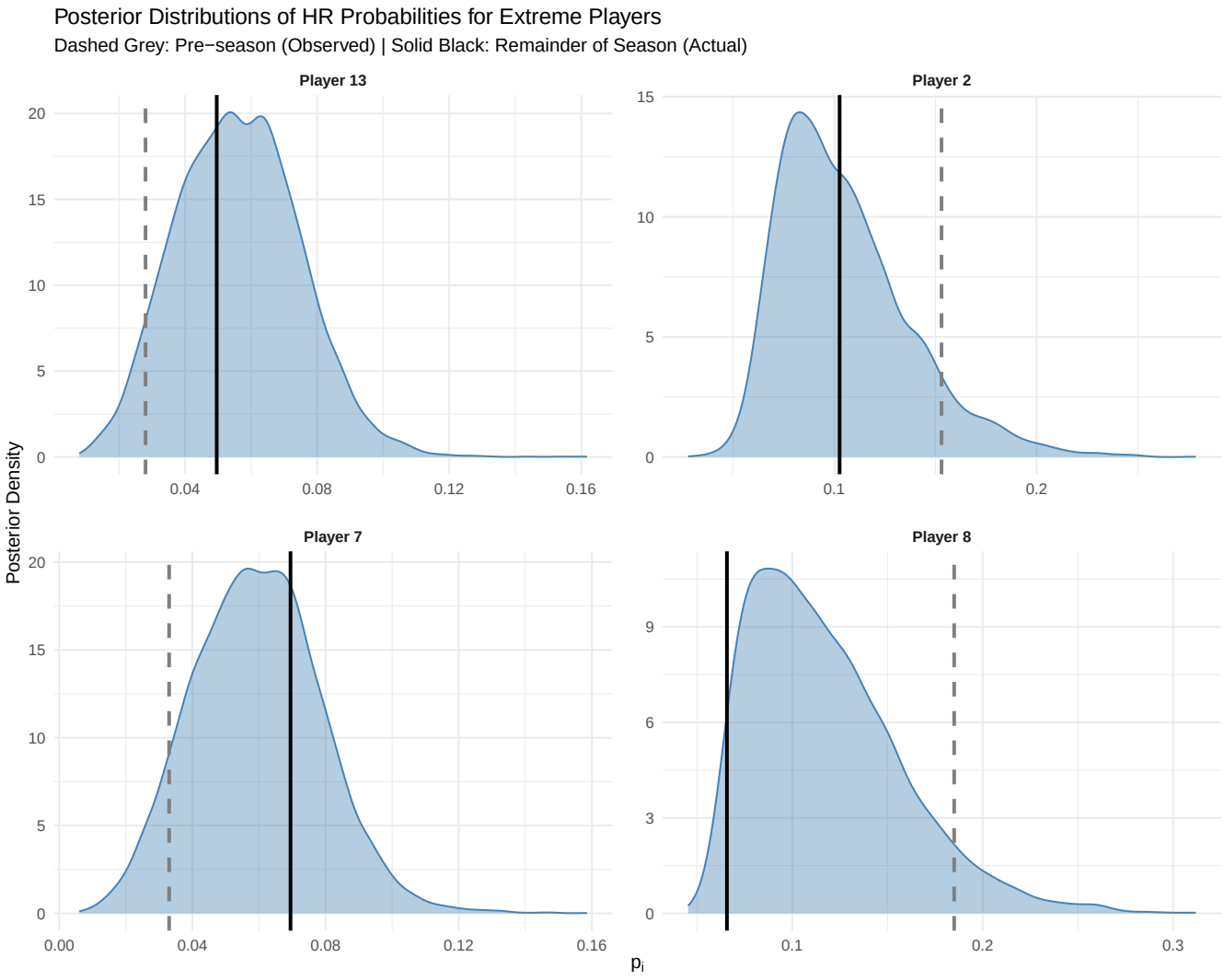
3.8.5.2 Method 4: Hierarchical Logistic Regression (logit-normal)

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.

$$\begin{aligned} y_i | p_i &\sim \text{Binomial}(n_i, p_i) \\ \text{logit}(p_i) &\sim N(\mu_0, \tau_0^2) \\ \mu_0 &\sim N(0, 10) \\ \tau_0 &\sim \text{Cauchy}(0, 2) \end{aligned} \tag{3.190}$$

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.



3.8.5.3 Method 5: Optimal Bayes Estimator w.r.t. Relative Absolute Error

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:

$$L(p, \hat{p}) = \frac{|p - \hat{p}|}{\min(p, 1-p)} \tag{3.191}$$

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.

```r
# 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. Extract Posterior Means (Method 4)
# This provides the missing p_hat_logit variable
p_hat_logit <- colMeans(p_samples)

# 3. 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])
}

# 4. Apply to All Players (Method 5)
p_hat_optimal <- apply(p_samples, 2, get_weighted_median)
```

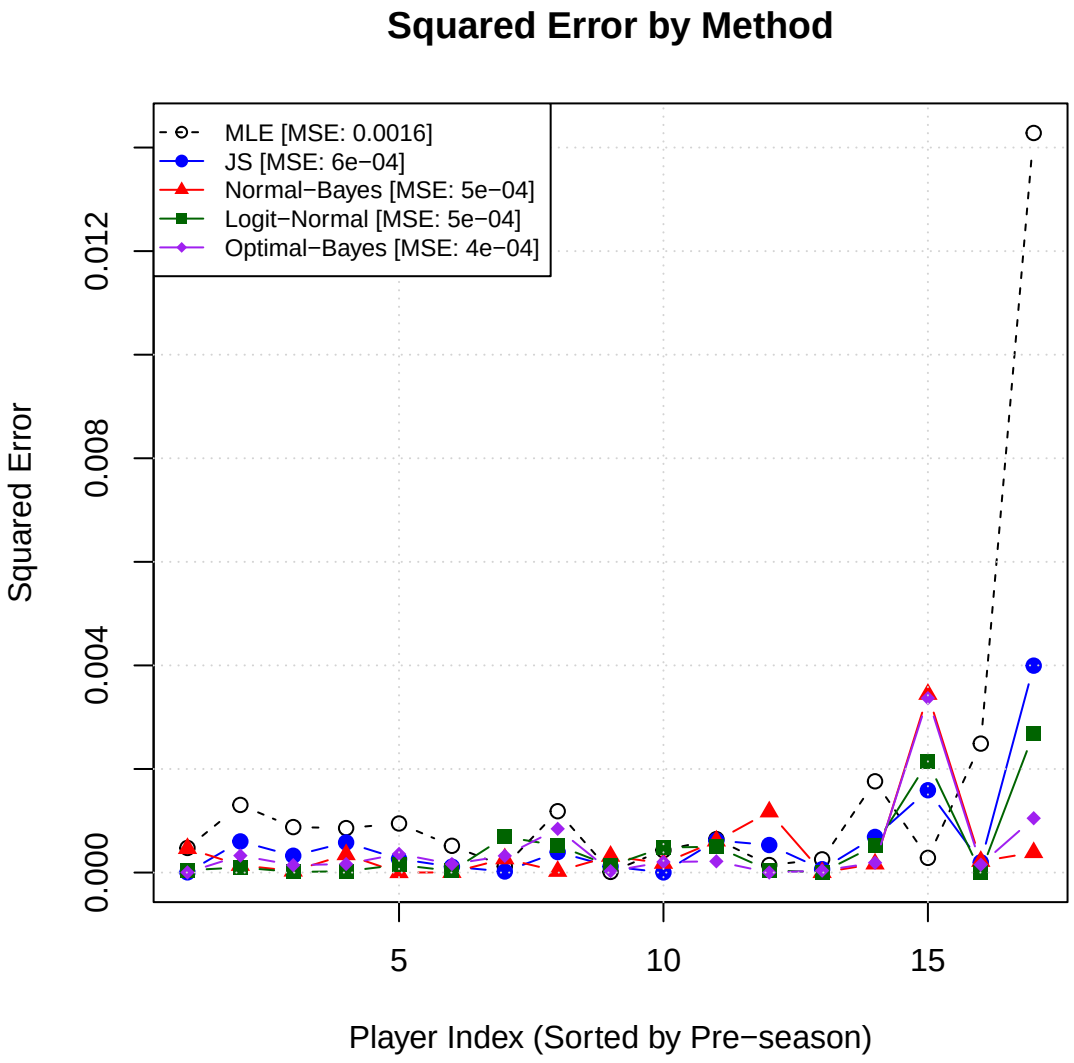
3.8.5.4 Comparison of All Five Estimates of p_i

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

Table 3.8: Comparison of Estimated Probabilities (p_i) across Five Methods

Player	Season Avg (p_i)	MLE	James-Stein	Normal-Bayes	Logit-Normal	Optimal-Bayes
13	0.050	0.028	0.048	0.071	0.056	0.048
7	0.069	0.033	0.045	0.058	0.060	0.051
10	0.063	0.033	0.045	0.058	0.060	0.051
9	0.067	0.038	0.043	0.048	0.062	0.054
5	0.074	0.043	0.058	0.074	0.062	0.055
12	0.068	0.045	0.058	0.070	0.063	0.055
16	0.042	0.053	0.037	0.025	0.068	0.060
3	0.088	0.054	0.069	0.083	0.066	0.059
11	0.057	0.061	0.068	0.075	0.068	0.062
15	0.051	0.071	0.052	0.037	0.073	0.065
14	0.053	0.078	0.078	0.077	0.075	0.067
4	0.071	0.083	0.094	0.106	0.078	0.071
6	0.079	0.095	0.087	0.081	0.082	0.073
17	0.061	0.103	0.088	0.074	0.084	0.075
1	0.138	0.121	0.098	0.079	0.091	0.079
2	0.103	0.153	0.117	0.088	0.105	0.090
8	0.066	0.185	0.129	0.085	0.118	0.098

1. MSE Comparison



2. Comparison of Relative Absolute Error

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

$$\text{Metric}_i = \frac{|p_i^{\text{true}} - \hat{p}_i|}{\min(p_i^{\text{true}}, 1 - p_i^{\text{true}})} \tag{3.192}$$

3.9 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

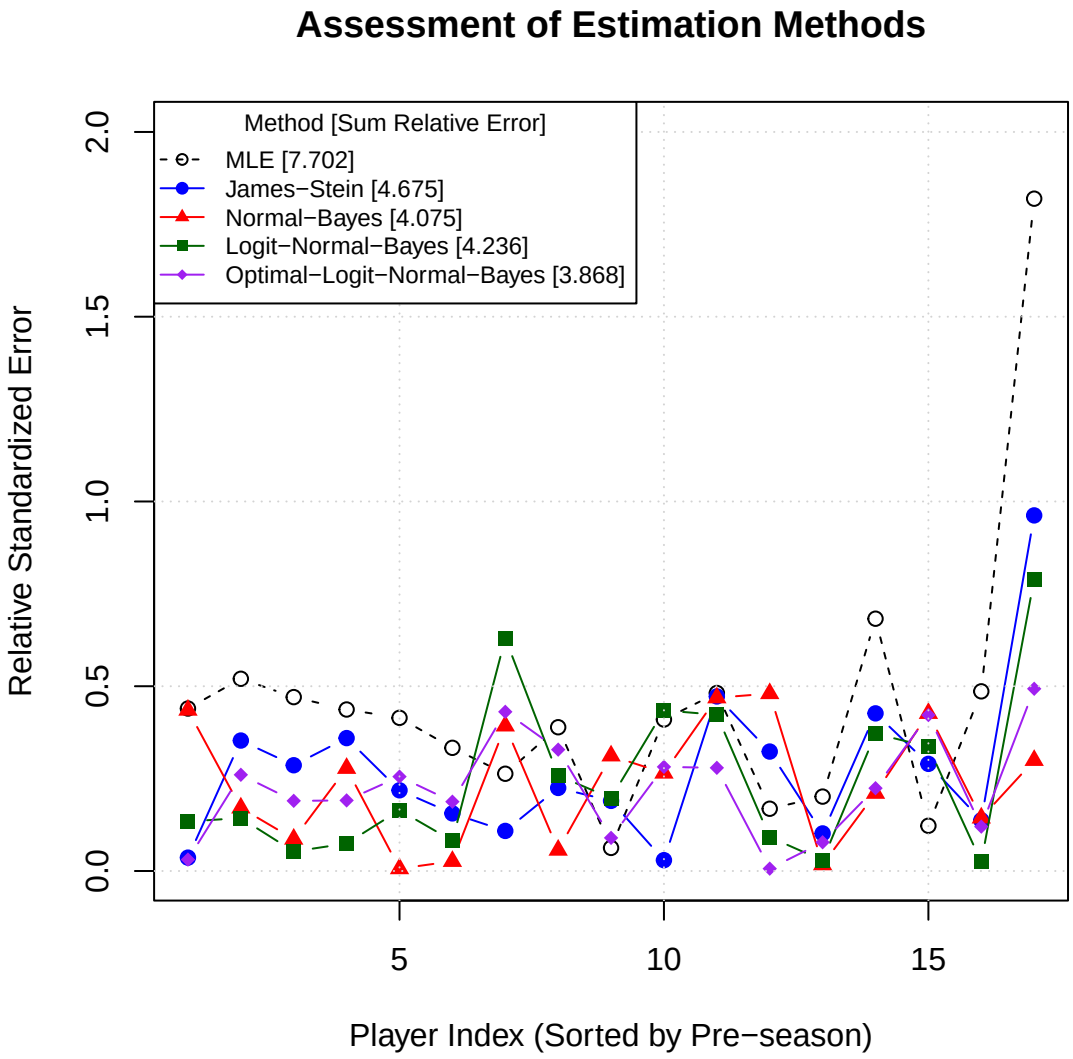


Figure 3.10: Relative Error Assessment: Five Methods

predictive density is obtained by marginalizing over the parameter θ:

$$f(y^*|y) = \int_\Theta f(y^*|\theta)\pi(\theta|y)\, d\theta \tag{3.193}$$

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 3.11 (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:

$$Y_i|\mu \sim N(\mu, \sigma^2) \tag{3.194}$$

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:

$$\begin{aligned} f(y^*|y) &= \int_{-\infty}^{\infty} f(y^*|\mu)\pi(\mu|y)\, d\mu \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^*-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\mu-\mu_n)^2}{2\sigma_n^2}}\, d\mu \end{aligned} \tag{3.195}$$

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

$$y^*|y \sim N(\mu_n, \sigma^2 + \sigma_n^2) \tag{3.196}$$

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

4 Sufficient Statistic

4.1 Sufficient Statistics

Definition 4.1 (Sufficient Statistic)**.** A statistic $T = T(\mathbf{X})$ is sufficient for θ if one of the following three equivalent conditions holds:

1. **Parallel Log-Likelihood**

   For any pair of data sets $\mathbf{x}$ and $\mathbf{y}$ such that $T(\mathbf{x}) = T(\mathbf{y})$, the difference in their log-likelihoods ($\ell(\theta; \mathbf{x}) = \ln(f(\mathbf{x}; \theta))$) is constant with respect to θ:

   $$\ell(\theta; \mathbf{x}) - \ell(\theta; \mathbf{y}) = c(\mathbf{x}, \mathbf{y}) \quad \text{for all } \theta \tag{4.1}$$

   where $c(\mathbf{x}, \mathbf{y})$ depends only on $\mathbf{x}$ and $\mathbf{y}$, not on θ.

2. **Factorization of Likelihood**

   The likelihood function of θ given $\mathbf{x}$ can be expressed as:

   $$L(\theta; \mathbf{x}) = h(\mathbf{x})g(T(\mathbf{x}); \theta) \tag{4.2}$$

   where $h(\mathbf{x})$ is irrelevant to θ.

3. **Non-informative Conditional Distribution of $\mathbf{X}|T(\mathbf{X})$**

   The conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = t$, denoted as $f(\mathbf{x}|t, \theta)$, is independent of θ.

   $$f(\mathbf{x}|T(\mathbf{x}) = t, \theta) = f(\mathbf{x}|t) \tag{4.3}$$

Theorem 4.1 (Factorization Theorem)**.** *The three conditions in the definitions of Definition 4.1 are equivalent.*

Click to view Complete Proof of Equivalence

Proof. **Proof of Equivalence**
We show the equivalence by proving the implications in a cycle or pairs: $(2 \Rightarrow 1)$, $(1 \Rightarrow 2)$, $(2 \Rightarrow 3)$, and $(3 \Rightarrow 2)$.

1. **Factorization $\Rightarrow$ Log-Likelihood Difference** $(2 \Rightarrow 1)$

   Assume the **Factorization Theorem** holds: $L(\theta; \mathbf{x}) = h(\mathbf{x})g(T(\mathbf{x}); \theta)$. Consider any pair $\mathbf{x}, \mathbf{y}$ such that $T(\mathbf{x}) = T(\mathbf{y})$.

   $$\ell(\theta; \mathbf{x}) - \ell(\theta; \mathbf{y}) = [\ln h(\mathbf{x}) + \ln g(T(\mathbf{x}); \theta)] - [\ln h(\mathbf{y}) + \ln g(T(\mathbf{y}); \theta)] \tag{4.4}$$

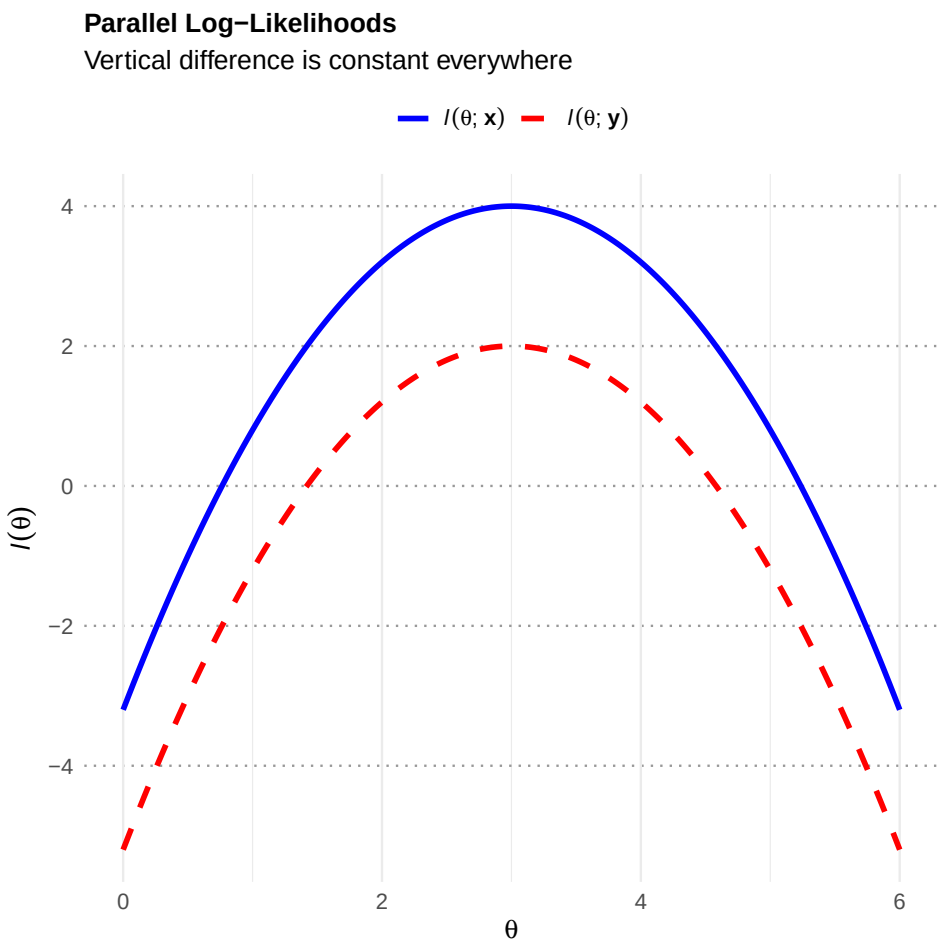


Figure 4.1: Visualizing Parallel Log-Likelihoods

Since $T(\mathbf{x}) = T(\mathbf{y})$, the terms $\ln g(T(\mathbf{x}); \theta)$ and $\ln g(T(\mathbf{y}); \theta)$ are identical and cancel out.

$$\ell(\theta; \mathbf{x}) - \ell(\theta; \mathbf{y}) = \ln h(\mathbf{x}) - \ln h(\mathbf{y}) \tag{4.5}$$

This difference depends only on $\mathbf{x}$ and $\mathbf{y}$ (via h), and is independent of θ. Thus, condition 1 holds.

2. **Log-Likelihood Difference $\Rightarrow$ Factorization** $(1 \Rightarrow 2)$

   Assume Condition 1 holds. For any $\mathbf{x}$ and $\mathbf{y}$ with $T(\mathbf{x}) = T(\mathbf{y})$, $\ell(\theta; \mathbf{x}) - \ell(\theta; \mathbf{y}) = c(\mathbf{x}, \mathbf{y})$. Exponentiating, we get $L(\theta; \mathbf{x}) = k(\mathbf{x}, \mathbf{y}) L(\theta; \mathbf{y})$, where k is independent of θ.

   For each value t in the range of T, select a fixed representative data point $\mathbf{x}_t$ such that $T(\mathbf{x}_t) = t$. For any data point $\mathbf{x}$, let $t = T(\mathbf{x})$. Using the relation above:

$$L(\theta; \mathbf{x}) = k(\mathbf{x}, \mathbf{x}_t) L(\theta; \mathbf{x}_t) \tag{4.6}$$

   Define $h(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}_{T(\mathbf{x})})$ and $g(t; \theta) = L(\theta; \mathbf{x}_t)$. Then:

$$L(\theta; \mathbf{x}) = h(\mathbf{x}) g(T(\mathbf{x}); \theta) \tag{4.7}$$

   This is exactly the Factorization form.

3. **Factorization $\Rightarrow$ Conditional Distribution** $(2 \Rightarrow 3)$

   Assume $f(\mathbf{x}; \theta) = h(\mathbf{x}) g(T(\mathbf{x}); \theta)$. We derive the conditional distribution $P(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t)$. If $T(\mathbf{x}) \neq t$, the probability is 0 (independent of θ). If $T(\mathbf{x}) = t$:

$$P(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t) = \frac{P(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t)}{P(T(\mathbf{X}) = t)} = \frac{f(\mathbf{x}; \theta)}{\sum_{\{\mathbf{y}: T(\mathbf{y}) = t\}} f(\mathbf{y}; \theta)} \tag{4.8}$$

   Substitute the factorization:

$$= \frac{h(\mathbf{x}) g(t; \theta)}{\sum_{\{\mathbf{y}: T(\mathbf{y}) = t\}} h(\mathbf{y}) g(t; \theta)} = \frac{h(\mathbf{x}) g(t; \theta)}{g(t; \theta) \sum_{\{\mathbf{y}: T(\mathbf{y}) = t\}} h(\mathbf{y})} \tag{4.9}$$

   The term $g(t; \theta)$ cancels out:

$$= \frac{h(\mathbf{x})}{\sum_{\{\mathbf{y}: T(\mathbf{y}) = t\}} h(\mathbf{y})} \tag{4.10}$$

   This expression depends only on $\mathbf{x}$ and $h(\cdot)$, and is entirely free of θ. Thus, Condition 3 holds.

4. **Conditional Distribution $\Rightarrow$ Factorization** $(3 \Rightarrow 2)$

   Assume $f(\mathbf{x} | T(\mathbf{x}); \theta) = k(\mathbf{x})$, where $k(\mathbf{x})$ is independent of θ. We can write the joint distribution as:

$$f(\mathbf{x}; \theta) = f(\mathbf{x} | T(\mathbf{x}) = t; \theta) \cdot P(T(\mathbf{X}) = t; \theta) \tag{4.11}$$

   Substitute the assumption:

$$f(\mathbf{x}; \theta) = k(\mathbf{x}) \cdot P(T(\mathbf{X}) = T(\mathbf{x}); \theta) \tag{4.12}$$

   Let $h(\mathbf{x}) = k(\mathbf{x})$ and $g(t; \theta) = P(T(\mathbf{X}) = t; \theta)$. Then:

$$f(\mathbf{x}; \theta) = h(\mathbf{x}) g(T(\mathbf{x}); \theta) \tag{4.13}$$

   This recovers the Factorization form.

□

Example 4.1 (Uniform Distribution $U(\theta-1,\theta+1)$)**.** Consider a random sample $\mathbf{X}=(X_1,\ldots,X_n)$ from a Uniform distribution with range $(\theta-1,\theta+1)$.
The density for a single observation is:

$$f(x_i|\theta)=\frac{1}{(\theta+1)-(\theta-1)}I(\theta-1<x_i<\theta+1)=\frac{1}{2}I(\theta-1<x_i<\theta+1) \tag{4.14}$$

The joint PDF (likelihood) for the vector **x** is:

$$L(\theta;\mathbf{x})=\prod_{i=1}^{n}\frac{1}{2}I(\theta-1<x_i<\theta+1) \tag{4.15}$$

$$L(\theta;\mathbf{x})=2^{-n}\cdot I(\min(x_i)>\theta-1)\cdot I(\max(x_i)<\theta+1) \tag{4.16}$$

Using order statistics notation where $X_{(1)}=\min(X_i)$ and $X_{(n)}=\max(X_i)$:

$$L(\theta;\mathbf{x})=2^{-n}\cdot I(\theta<X_{(1)}+1)\cdot I(\theta>X_{(n)}-1) \tag{4.17}$$

$$L(\theta;\mathbf{x})=2^{-n}\cdot I(X_{(n)}-1<\theta<X_{(1)}+1) \tag{4.18}$$

By the **Factorization Theorem**, we can define:

- $h(\mathbf{x})=2^{-n}$ (or simply 1, grouping constants into g)
- $g(T(\mathbf{x}),\theta)=I(X_{(n)}-1<\theta<X_{(1)}+1)$

Thus, the sufficient statistic is the pair of order statistics:

$$T(\mathbf{X})=(X_{(1)},X_{(n)}) \tag{4.19}$$

Example 4.2 (Gamma Distribution)**.** Let $\mathbf{X}=(X_1,\ldots,X_n)$ be i.i.d. $\Gamma(\alpha,\beta)$. The pdf is:

$$f(x_i|\alpha,\beta)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x_i^{\alpha-1}e^{-x_i/\beta},\quad x_i>0 \tag{4.20}$$

The joint likelihood is:

$$L(\alpha,\beta;\mathbf{x})=\left(\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\right)^{n}\left(\prod_{i=1}^{n}x_i\right)^{\alpha-1}\exp\left(-\frac{1}{\beta}\sum_{i=1}^{n}x_i\right) \tag{4.21}$$

By the Factorization Theorem, we can identify the parts that depend on the data and the parameters:

$$g(T(\mathbf{x}),\square)=\left(\prod_{i=1}^{n}x_i\right)^{\alpha}\exp\left(-\frac{1}{\beta}\sum_{i=1}^{n}x_i\right) \tag{4.22}$$

Thus, the sufficient statistics are:

$$T(\mathbf{X})=\left(\prod_{i=1}^{n}X_i,\sum_{i=1}^{n}X_i\right) \tag{4.23}$$

Example 4.3 (Sufficient Statistic of Exponential Family)**.** Many common distributions (Normal, Poisson, Gamma, Binomial) belong to the **Exponential Family**, which has a density in the form:

$$f(\mathbf{x}|\theta) = h(\mathbf{x})c(\theta)\exp\left(\sum_{j=1}^{k}\pi_j(\theta)t_j(\mathbf{x})\right) \tag{4.24}$$

Then, by the Factorization Theorem, the statistic:

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n}t_1(x_i), \dots, \sum_{i=1}^{n}t_k(x_i)\right) \tag{4.25}$$

is a sufficient statistic for θ.

Example 4.4 (Bernoulli as Exponential Family)**.** Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} \text{Bernoulli}(p)$. To find the sufficient statistic, we write the **Joint PDF** of the sample in the canonical Exponential Family form:

$$f(\mathbf{x}|\theta) = h(\mathbf{x})c(\theta)\exp\left(\sum_{j=1}^{k}\pi_j(\theta)T_j(\mathbf{x})\right) \tag{4.26}$$

1. **Write the Joint PDF**

$$f(\mathbf{x}|p) = \prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i} \tag{4.27}$$

2. **Convert to Exponential Form**

$$\begin{aligned} f(\mathbf{x}|p) &= \exp\left(\sum_{i=1}^{n}\left[x_i\ln p + (1-x_i)\ln(1-p)\right]\right) \\ &= \exp\left(\sum_{i=1}^{n}\left[x_i\ln p + \ln(1-p) - x_i\ln(1-p)\right]\right) \\ &= \exp\left(\sum_{i=1}^{n}\ln(1-p) + \sum_{i=1}^{n}x_i\left[\ln p - \ln(1-p)\right]\right) \end{aligned} \tag{4.28}$$

3. **Factor into Components** We separate the terms to match the definition:

$$f(\mathbf{x}|p) = \underbrace{1}_{h(\mathbf{x})}\cdot\underbrace{(1-p)^n}_{c(p)}\cdot\exp\left(\underbrace{\ln\left(\frac{p}{1-p}\right)}_{\pi_1(p)}\underbrace{\sum_{i=1}^{n}x_i}_{T_1(\mathbf{x})}\right) \tag{4.29}$$

Conclusion: By inspection of the exponent, the statistic coupled with the parameter $\pi_1(p)$ is the sufficient statistic:

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i \tag{4.30}$$

Remark 4.1 (Sufficient Statistic is the sufficient "Parameter" of Likelihood). There is a dual relationship between the sufficient statistic and the parameter θ. Conventionally, we view $f(x|\theta)$ as a function of x parameterized by θ.

However, in Bayesian inference or likelihood theory, we often view the likelihood $L(\theta; x)$ as a function of θ determined by the observed data x. The Factorization Theorem implies:

$$L(\theta; \mathbf{x}) \propto g(T(\mathbf{x})|\theta) \tag{4.31}$$

This suggests that $T(\mathbf{x})$ completely determines the shape of the likelihood function. In this specific sense, the sufficient statistic $T(\mathbf{x})$ acts as the **"parameter"** of the likelihood function itself.

For the exponential family that we will discuss below, this duality is explicit:

$$\log L(\theta; \mathbf{x}) = \text{const} + \sum_{i=1}^{k} \eta_i(\theta) T_i(\mathbf{x}) - nA(\theta) \tag{4.32}$$

Here, $T_i(\mathbf{x})$ serves as the coefficient (or parameter) for the function $\eta_i(\theta)$.

4.2 Minimal Sufficient Statistics

Definition 4.2 (Minimal Sufficient Statistic (MSS)). A statistic $T(X)$ is a **Minimal Sufficient Statistic** if:

1. **Sufficiency:** $T(X)$ is a sufficient statistic for θ.
2. **Minimality:** For any other sufficient statistic $S(X)$, $T(X)$ is a function of $S(X)$.

$$T(X) = g(S(X)) \tag{4.33}$$

(This implies that $T(X)$ provides the greatest possible data reduction without losing information about θ. If $S(x) = S(y)$, then it must be that $T(x) = T(y)$).

Theorem 4.2 (MSS Condition Theorem). *Let $T(X)$ be a **sufficient statistic**. $T(X)$ is a **Minimal Sufficient Statistic (MSS)** if and only if for any pair of data sets x and y:*

$$\ell(\theta; x) = \ell(\theta; y) + c(x, y) \text{ for all } \theta \implies T(x) = T(y) \tag{4.34}$$

where $c(x, y)$ is a constant independent of θ.

Click to view the Proof

Proof. **Direction 1: Sufficiency (Implication holds $\implies$ T is MSS)**
Assume that for any x, y, $[\ell(\theta; x) = \ell(\theta; y) + c(x, y)] \implies T(x) = T(y)$. We must show that T is a function of *any* sufficient statistic U.

1. Let $U(X)$ be any sufficient statistic. Assume $U(x) = U(y)$.
2. By the **Factorization Theorem**, the likelihoods are:
$$L(\theta; x) = h(x)g(U(x), \theta) \tag{4.35}$$
$$L(\theta; y) = h(y)g(U(y), \theta) \tag{4.36}$$
3. Since $U(x) = U(y)$, the factor $g(U(x), \theta)$ is identical to $g(U(y), \theta)$. Taking the log-ratio:
$$\ell(\theta; x) - \ell(\theta; y) = \ln h(x) - \ln h(y) \tag{4.37}$$
The term $\ln h(x) - \ln h(y)$ depends only on x and y, not on θ. Let this be $c(x, y)$.
$$\ell(\theta; x) = \ell(\theta; y) + c(x, y) \tag{4.38}$$
4. By our main assumption, this condition implies $T(x) = T(y)$.
5. Thus, we have shown that $U(x) = U(y) \implies T(x) = T(y)$. This means T is a function of U. Since U is arbitrary, T is Minimal Sufficient.

Direction 2: Necessity (T is MSS $\implies$ Implication holds)
Assume T is Minimal Sufficient. We must prove that if $\ell(\theta; x) = \ell(\theta; y) + c(x, y)$ for all θ, then $T(x) = T(y)$.

1. **Define the Statistic $S(x)$:** Let $S(x)$ be the set of all possible datasets z which give the same log-likelihood shape as x:
$$S(x) = \{z \mid \ell(\theta; z) = \ell(\theta; x) + c_z \text{ for all } \theta\} \tag{4.39}$$
This statistic $S(x)$ represents the equivalence class of x under the parallel log-likelihood relationship. If the condition $\ell(\theta; x) = \ell(\theta; y) + c(x, y)$ holds, then by definition x and y generate the same equivalence class, so $S(x) = S(y)$.
2. **Show $S(x)$ is Sufficient (Directly via Likelihood Ratio):** To prove S is sufficient, we check the **Likelihood Ratio Condition** (Condition 2 from Section 1.1). Suppose $S(x) = S(y)$. By the definition of S, this implies:
$$\ell(\theta; x) - \ell(\theta; y) = c(x, y) \tag{4.40}$$
By the definition of sufficiency, $S(X)$ is a sufficient statistic.
3. **Use Minimality of T:** Since T is a **Minimal** Sufficient Statistic, it is a function of *any* sufficient statistic. Therefore, T must be a function of S. That is, $T(x) = f(S(x))$.
4. **Conclusion:** Assume $\ell(\theta; x) = \ell(\theta; y) + c(x, y)$. Then $S(x) = S(y)$. Consequently, $T(x) = f(S(x)) = f(S(y)) = T(y)$.

□

Example 4.5 (Checking Minimality via Log-Likelihood Condition)**.** Let $X_1, X_2, X_3 \overset{i.i.d.}{\sim}$ Bernoulli(p). We determine the MSS by checking the implication from the **MSS Condition Theorem**:

$$\text{Parallel Log-Likelihoods} \implies T(x) = T(y) \tag{4.41}$$

Step 1: Establishing the MSS
First, we find the condition under which two log-likelihoods are parallel.

$$\ell(p; x) = (\sum x_i) \ln p + (n - \sum x_i) \ln(1 - p) \tag{4.42}$$

The difference $\ell(p; x) - \ell(p; y)$ depends on p only through the term $(\sum x_i - \sum y_i) \ln \frac{p}{1-p}$. For this difference to be constant (independent of p), the coefficient must be zero:

$$\text{Parallel Log-Likelihoods} \iff \sum x_i = \sum y_i \tag{4.43}$$

The statistic that corresponds exactly to this condition is $T(X) = \sum X_i$. Since $\sum x_i = \sum y_i$ trivially implies $T(x) = T(y)$, $T(X)$ is the **Minimal Sufficient Statistic**.
Step 2: Why $S(X) = (X_1, \sum_{i=2}^{3} X_i)$ is NOT Minimal
Now consider the "richer" statistic $S(X)$. If S were minimal, the parallel condition must imply $S(x) = S(y)$. We check:

$$\sum x_i = \sum y_i \overset{?}{\implies} (x_1, \sum_{i=2}^{3} x_i) = (y_1, \sum_{i=2}^{3} y_i) \tag{4.44}$$

Counter-Example:
Let $x = (1, 0, 1)$ and $y = (0, 1, 1)$.

1. **Check Parallel Condition:**
   $\sum x_i = 2$ and $\sum y_i = 2$. The sums are equal, so the log-likelihoods are parallel.
2. **Check Statistic Equality:**

$$S(x) = (1, 1) \tag{4.45}$$

$$S(y) = (0, 2) \tag{4.46}$$

$$S(x) \neq S(y) \tag{4.47}$$

Conclusion: The parallel condition holds, but $S(x) \neq S(y)$. The implication fails. This proves that $S(X)$ is **not** minimal—it retains "extra" information (the position of the first success) that is not relevant to the likelihood shape.

5 Likelihood Theory

5.1 Definitions and Notations

5.1.1 Regular Family

Definition 5.1 (Regular Families)**.** A family of probability density functions is said to be a **Regular Family** if the support $\{\mathbf{x} : f(\mathbf{x}|\theta) > 0\}$ does not depend on the parameter vector θ.
This condition allows for the interchange of differentiation and integration:

$$\nabla_\theta \int \exp\{\ell(\theta; \mathbf{x})\} d\mathbf{x} = \int \nabla_\theta \exp\{\ell(\theta; \mathbf{x})\} d\mathbf{x} \tag{5.1}$$

5.1.2 Score and Fisher Information

Before stating the theorem, we define the following notations for the score and information in the context of a parameter vector $\theta = (\theta_1, \dots, \theta_p)^T \in \mathbb{R}^p$:

Definition 5.2.

1. **Score Vector (U):** The gradient of the log-likelihood. It is a random column vector of dimension $p \times 1$.

$$\mathbf{U}(\theta; \mathbf{X}) = \nabla \ell(\theta; \mathbf{X}) = \frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta} = \begin{bmatrix} \frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta_1} \\ \frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta_p} \end{bmatrix} \tag{5.2}$$

2. **Observed Information Matrix (J):** The negative Hessian of the log-likelihood. It is a symmetric random matrix of dimension $p \times p$, measuring the curvature of the log-likelihood surface.

$$\mathbf{J}(\theta; \mathbf{X}) = -\nabla^2 \ell(\theta; \mathbf{X}) = -\frac{\partial^2 \ell(\theta; \mathbf{X})}{\partial \theta \partial \theta^T} = -\begin{bmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_p^2} \end{bmatrix} \tag{5.3}$$

3. **(Expected) Fisher Information Matrix (I):** The covariance matrix of the score vector. It is a deterministic $p \times p$ matrix (for a fixed θ).

$$\mathbf{I}(\theta) = E_\theta \left[\mathbf{J}(\theta; \mathbf{X}) \right] \tag{5.4}$$

5.2 Mean and Covariance of Score Vector

Theorem 5.1 (Bartlett's Identities: Mean and Covariance of Score Vector)**.** *Let $\{f(\mathbf{x}|\theta) : \theta \in \Theta\}$ be a regular family of probability density functions. The following identities hold relating the moments of the score vector $\mathbf{U}(\theta; \mathbf{X})$ and the observed information matrix $\mathbf{J}(\theta; \mathbf{X})$:*

1. ***First Moment Identity:*** *The expected score is zero vector.*

$$E_\theta[\mathbf{U}(\theta; \mathbf{X})] = \mathbf{0} \tag{5.5}$$

2. ***Second Moment Identity:*** *The expected observed information equals the covariance of the score vector (Fisher Information).*

$$Cov_\theta \left(\mathbf{U}(\theta; \mathbf{X}) \right) = E_\theta[\mathbf{J}(\theta; \mathbf{X})] = \mathbf{I}(\theta) \tag{5.6}$$

Remark 5.1. The only assumption in the theorem above is that the families are regular. Therefore, we do not need to assume the log-likelihood $\ell(\theta)$ is "well-behaved" (e.g., approximately quadratic or independence within $\mathbf{X}$) for these two identities to hold.

Proof.

1. Proof of the First Moment Identity

We start with the fundamental property that a density function integrates to 1 over the sample space of $\mathbf{X}$:

$$\int f(\mathbf{x}|\theta) \, d\mathbf{x} = 1 \tag{5.7}$$

Differentiating both sides with respect to the parameter vector θ:

$$\nabla_\theta \int f(\mathbf{x}|\theta) \, d\mathbf{x} = \mathbf{0} \tag{5.8}$$

Assuming regularity allows us to interchange differentiation and integration:

$$\int \nabla_\theta f(\mathbf{x}|\theta) \, d\mathbf{x} = \mathbf{0} \tag{5.9}$$

Using the identity $\nabla_\theta f(\mathbf{x}|\theta) = f(\mathbf{x}|\theta) \nabla_\theta \log f(\mathbf{x}|\theta) = f(\mathbf{x}|\theta) \mathbf{U}(\theta; \mathbf{x})$:

$$\int \mathbf{U}(\theta; \mathbf{x}) f(\mathbf{x}|\theta) \, d\mathbf{x} = \mathbf{0} \tag{5.10}$$

This is precisely the definition of the expectation:

$$E_\theta[\mathbf{U}(\theta; \mathbf{X})] = \mathbf{0} \tag{5.11}$$

2. Proof of the Second Moment Identity

We differentiate the result of the First Moment Identity ($E[\mathbf{U}(\theta;\mathbf{X})] = \mathbf{0}$) with respect to θ^T.

$$\nabla_{\theta^T} \int \mathbf{U}(\theta;\mathbf{x}) f(\mathbf{x}|\theta)\, d\mathbf{x} = \mathbf{0} \tag{5.12}$$

Applying the product rule inside the integral (remembering $\mathbf{U}$ is a vector):

$$\int \left[\left(\nabla_{\theta^T} \mathbf{U}(\theta;\mathbf{x})\right) f(\mathbf{x}|\theta) + \mathbf{U}(\theta;\mathbf{x}) \left(\nabla_{\theta^T} f(\mathbf{x}|\theta)\right) \right] d\mathbf{x} = \mathbf{0} \tag{5.13}$$

We analyze the two terms in the bracket:

- **Term 1:** $\nabla_{\theta^T} \mathbf{U}(\theta;\mathbf{x})$ is the Jacobian of the score, which is the Hessian of the log-likelihood, $\nabla^2 \ell(\theta;\mathbf{x})$. By definition, this is $-\mathbf{J}(\theta;\mathbf{x})$.
- **Term 2:** We use the identity $\nabla_{\theta^T} f(\mathbf{x}|\theta) = f(\mathbf{x}|\theta)(\nabla_\theta \log f(\mathbf{x}|\theta))^T = f(\mathbf{x}|\theta)\mathbf{U}(\theta;\mathbf{x})^T$.

Substituting these back into the integral:

$$\int \left[-\mathbf{J}(\theta;\mathbf{x}) f(\mathbf{x}|\theta) + \mathbf{U}(\theta;\mathbf{x})\mathbf{U}(\theta;\mathbf{x})^T f(\mathbf{x}|\theta) \right] d\mathbf{x} = \mathbf{0} \tag{5.14}$$

This simplifies to expectations:

$$-E_\theta[\mathbf{J}(\theta;\mathbf{X})] + E_\theta[\mathbf{U}(\theta;\mathbf{X})\mathbf{U}(\theta;\mathbf{X})^T] = \mathbf{0} \tag{5.15}$$

Rearranging gives:

$$E_\theta[\mathbf{J}(\theta;\mathbf{X})] = E_\theta[\mathbf{U}(\theta;\mathbf{X})\mathbf{U}(\theta;\mathbf{X})^T] \tag{5.16}$$

Finally, recall the definition of the covariance matrix for a random vector with zero mean. Since $E_\theta[\mathbf{U}(\theta;\mathbf{X})] = \mathbf{0}$, we have:

$$\text{Cov}_\theta(\mathbf{U}(\theta;\mathbf{X})) = E_\theta[\mathbf{U}(\theta;\mathbf{X})\mathbf{U}(\theta;\mathbf{X})^T] - E_\theta[\mathbf{U}(\theta;\mathbf{X})]E_\theta[\mathbf{U}(\theta;\mathbf{X})]^T = E_\theta[\mathbf{U}(\theta;\mathbf{X})\mathbf{U}(\theta;\mathbf{X})^T] \tag{5.17}$$

Therefore, we conclude:

$$\text{Cov}_\theta(\mathbf{U}(\theta;\mathbf{X})) = E_\theta[\mathbf{J}(\theta;\mathbf{X})] = \mathbf{I}(\theta) \tag{5.18}$$

□

5.3 Cramer-Rao Lower Bound

In estimation theory, we often wish to know the limit of how well a parameter can be estimated. The following theorem provides a lower bound on the variance of any estimator.

Theorem 5.2 (Cramer-Rao Lower Bound for Scalar Estimator)**.** *Let X be a random variable with probability density function (or probability mass function) $f(x|\theta)$, where $\theta \in \Theta$ is a scalar unknown parameter. Let $T(X)$*

be any estimator with finite variance, and let $m(\theta) = E_\theta[T(X)]$ denote its expectation.
*Assume the following **regularity conditions** hold:*

1. *The support of X, denoted $\mathcal{X} = \{x : f(x|\theta) > 0\}$, does not depend on θ.*
2. *The differentiation with respect to θ and integration (or summation) with respect to x can be interchanged.*

Then, the variance of $T(X)$ satisfies:

$$Var_\theta(T(X)) \geq \frac{[m'(\theta)]^2}{I(\theta)} \tag{5.19}$$

where $I(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]$ is the scalar Fisher Information.
***Particular Case:** If $T(X)$ is an **unbiased** estimator of θ (i.e., $m(\theta) = \theta$ and $m'(\theta) = 1$), then:*

$$Var_\theta(T(X)) \geq \frac{1}{I(\theta)} \tag{5.20}$$

Proof. Let $U = \frac{\partial}{\partial\theta}\log f(X|\theta)$ be the scalar Score function. From the properties of the Score function under the stated regularity conditions, we know that the score has mean zero and variance equal to the Fisher Information:

$$E_\theta[U] = 0 \quad \text{and} \quad \text{Var}_\theta(U) = I(\theta) \tag{5.21}$$

Consider the covariance between the estimator $T(X)$ and the Score U. By the Cauchy-Schwarz inequality (applied to covariance), we have:

$$[\text{Cov}_\theta(T, U)]^2 \leq \text{Var}_\theta(T)\text{Var}_\theta(U) \tag{5.22}$$

We now evaluate the covariance term explicitly. By definition:

$$\begin{aligned}
\text{Cov}_\theta(T, U) &= E_\theta[T(X)U] - E_\theta[T]E_\theta[U] \\
&= E_\theta\left[T(X)\frac{\partial}{\partial\theta}\log f(X|\theta)\right] - m(\theta)\cdot 0 \\
&= \int T(x)\left(\frac{\partial \log f(x|\theta)}{\partial\theta}\right) f(x|\theta)\,dx \\
&= \int T(x)\left(\frac{1}{f(x|\theta)}\frac{\partial f(x|\theta)}{\partial\theta}\right) f(x|\theta)\,dx \\
&= \int T(x)\frac{\partial f(x|\theta)}{\partial\theta}\,dx
\end{aligned} \tag{5.23}$$

Invoking the regularity condition that allows the interchange of derivative and integral, we move the derivative outside the integral:

$$\text{Cov}_\theta(T, U) = \frac{\partial}{\partial\theta}\int T(x)f(x|\theta)\,dx = \frac{\partial}{\partial\theta}E_\theta[T(X)] = m'(\theta) \tag{5.24}$$

Substituting this result and $\text{Var}_\theta(U) = I(\theta)$ back into the covariance inequality:

$$[m'(\theta)]^2 \leq \text{Var}_\theta(T) \cdot I(\theta) \tag{5.25}$$

Rearranging the terms yields the desired lower bound:

$$\text{Var}_\theta(T(X)) \geq \frac{[m'(\theta)]^2}{I(\theta)} \tag{5.26}$$

□

5.4 Multivariate Cramer-Rao Lower Bound

Theorem 5.3 (Multivariate Cramer-Rao Lower Bound)**.** *Let* $\mathbf{X}$ *be a random vector with density* $f(\mathbf{x}|\theta)$*, where* $\theta \in \mathbb{R}^p$ *is a vector of unknown parameters. Let* $\mathbf{T}(\mathbf{X}) \in \mathbb{R}^k$ *be any estimator with finite covariance matrix, and let* $\mathbf{m}(\theta) = E_\theta[\mathbf{T}(\mathbf{X})]$ *denote its expectation vector.*
Let $\mathbf{I}(\theta)$ *be the* $p \times p$ *Fisher Information Matrix:*

$$\mathbf{I}(\theta) = E_\theta\left[\mathbf{U}(\theta;\mathbf{X})\mathbf{U}(\theta;\mathbf{X})^\top\right] \tag{5.27}$$

Let $\mathbf{D}(\theta) = \frac{\partial \mathbf{m}(\theta)}{\partial \theta}$ *be the* $k \times p$ *Jacobian matrix of the expectation, where* $D_{ij} = \frac{\partial m_i}{\partial \theta_j}$*.*
Under standard regularity conditions, the covariance matrix of $\mathbf{T}$ *satisfies the inequality:*

$$Var_\theta(\mathbf{T}) \succeq \mathbf{D}(\theta)[\mathbf{I}(\theta)]^{-1}\mathbf{D}(\theta)^\top \tag{5.28}$$

Here, $\mathbf{A} \succeq \mathbf{B}$ *means that the matrix* $\mathbf{A} - \mathbf{B}$ *is positive semi-definite (i.e., for any vector* $\mathbf{v}$*,* $\mathbf{v}^\top(\mathbf{A} - \mathbf{B})\mathbf{v} \geq 0$*).*

Proof. Let $\mathbf{U} = \nabla_\theta \log f(\mathbf{X}|\theta)$ be the $p \times 1$ Score vector. We know that $E[\mathbf{U}] = \mathbf{0}$ and $\text{Var}(\mathbf{U}) = \mathbf{I}(\theta)$. Consider the covariance between the estimator $\mathbf{T}$ and the Score $\mathbf{U}$. By an argument similar to the scalar case (interchanging derivative and integral), we find:

$$\text{Cov}(\mathbf{T}, \mathbf{U}) = E[\mathbf{T}\mathbf{U}^\top] = \mathbf{D}(\theta) \tag{5.29}$$

Now, define the block vector $\mathbf{Z} = \begin{pmatrix} \mathbf{T} \\ \mathbf{U} \end{pmatrix}$. The covariance matrix of $\mathbf{Z}$ is necessarily positive semi-definite:

$$\text{Var}(\mathbf{Z}) = \begin{pmatrix} \text{Var}(\mathbf{T}) & \text{Cov}(\mathbf{T}, \mathbf{U}) \\ \text{Cov}(\mathbf{U}, \mathbf{T}) & \text{Var}(\mathbf{U}) \end{pmatrix} = \begin{pmatrix} \Sigma_\mathbf{T} & \mathbf{D} \\ \mathbf{D}^\top & \mathbf{I} \end{pmatrix} \succeq 0 \tag{5.30}$$

For this block matrix to be positive semi-definite, the Schur complement of the block $\mathbf{I}$ must be positive semi-definite (assuming $\mathbf{I}$ is positive definite/invertible):

$$\Sigma_\mathbf{T} - \mathbf{D}\mathbf{I}^{-1}\mathbf{D}^\top \succeq 0 \tag{5.31}$$

Thus, $\text{Var}(\mathbf{T}) \succeq \mathbf{D}\mathbf{I}^{-1}\mathbf{D}^\top$. □

Corollary 5.1 (Corollary: Scalar Estimator)**.** *Consider the case where* $T(\mathbf{X})$ *is a scalar estimator (*$k = 1$*) for a scalar parameter* θ *(*$p = 1$*).*

1. ***Matrices to Scalars:*** *The covariance matrix* $Var(\mathbf{T})$ *becomes the scalar variance* $Var(T)$*. The Fisher Information matrix* $\mathbf{I}(\theta)$ *becomes the scalar* $I(\theta)$*.*

2. ***Jacobian to Derivative:*** *The Jacobian matrix* $\mathbf{D}(\theta)$ *reduces to the scalar derivative* $m'(\theta)$*.*

Substituting these into the multivariate bound:

$$Var(T) - m'(\theta)[I(\theta)]^{-1}m'(\theta) \geq 0 \tag{5.32}$$

$$Var(T) \geq \frac{[m'(\theta)]^2}{I(\theta)} \tag{5.33}$$

Remark 5.2 (Generality of the Lower Bound). The power of the Cramer-Rao Lower Bound lies in its independence from the specific method of estimation. It relies solely on the properties of the underlying probability model (specifically, the curvature of the log-likelihood function) and the bias of the estimator. Consequently, it provides a universal benchmark for precision:

1. **Fundamental Limit** It represents the limit of "extractable information" about θ contained in the data $\mathbf{X}$. No matter how clever the estimation algorithm is (e.g., Method of Moments, Bayes estimators, etc.), the variance cannot be reduced beyond this intrinsic bound determined by the Fisher Information.

2. **Efficiency Standard** It allows us to define the concept of an *efficient estimator*. Any unbiased estimator that attains this lower bound is the Uniformly Minimum Variance Unbiased Estimator (UMVUE).

3. **Asymptotic Justification** While finite-sample estimators may not always achieve this bound, the Maximum Likelihood Estimator (MLE) is asymptotically efficient. This means that as the sample size $n \to \infty$, the variance of the MLE approaches the CRLB, justifying the popularity of likelihood-based inference.

5.4.1 Example with Exponential Likelihood

Example 5.1. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\theta)$, where the density is $f(x|\theta) = \frac{1}{\theta}e^{-x/\theta}$. We illustrate the likelihood identities and the efficiency of the sample mean.

1. The Score Function (U) The log-likelihood function is:

$$\ell(\theta; \mathbf{x}) = \sum_{i=1}^{n} \left(-\log\theta - \frac{x_i}{\theta} \right) = -n\log\theta - \frac{1}{\theta}\sum_{i=1}^{n} x_i \tag{5.34}$$

   The Score function is the first derivative with respect to θ:

$$U(\theta; \mathbf{x}) = \frac{\partial \ell}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum x_i}{\theta^2} \tag{5.35}$$

   Check First Moment: $E[U] = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum E[X_i] = -\frac{n}{\theta} + \frac{n\theta}{\theta^2} = 0$. (Verified)

2. Fisher Information ($I(\theta)$) We calculate the information using two different definitions to verify Bartlett's identity.

- **Method A: Negative Expected Hessian**

$$U'(\theta) = \frac{\partial U}{\partial \theta} = \frac{n}{\theta^2} - \frac{2\sum x_i}{\theta^3} \tag{5.36}$$

$$I(\theta) = -E[U'(\theta)] = -\left(\frac{n}{\theta^2} - \frac{2n\theta}{\theta^3}\right) = -\left(\frac{n}{\theta^2} - \frac{2n}{\theta^2}\right) = \frac{n}{\theta^2} \tag{5.37}$$

- **Method B: Variance of the Score**

$$\text{Var}(U) = \text{Var}\left(-\frac{n}{\theta} + \frac{\sum X_i}{\theta^2}\right) = \frac{1}{\theta^4}\text{Var}\left(\sum X_i\right) \tag{5.38}$$

Since X_i are independent with $\text{Var}(X_i) = \theta^2$:

$$\text{Var}(U) = \frac{1}{\theta^4}(n\theta^2) = \frac{n}{\theta^2} \tag{5.39}$$

Result: $\text{Var}(U) = -E[U'] = I(\theta)$. (Identity Verified)

3. Cramer-Rao Lower Bound (CRLB) Consider the estimator $T(\mathbf{X}) = \bar{X}$.

- **Expectation:** $m(\theta) = E[\bar{X}] = \theta$. Thus, T is unbiased and $m'(\theta) = 1$.
- **Actual Variance:**

$$\text{Var}(T(\mathbf{X})) = \text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{\theta^2}{n} \tag{5.40}$$

- **Theoretical Lower Bound:**

$$\text{CRLB} = \frac{[m'(\theta)]^2}{I(\theta)} = \frac{1^2}{n/\theta^2} = \frac{\theta^2}{n} \tag{5.41}$$

Conclusion:

$$\text{Var}(T(\mathbf{X})) = \frac{\theta^2}{n} \geq \frac{\theta^2}{n} \tag{5.42}$$

The variance of $T(\mathbf{X})$ achieves the lower bound exactly. Therefore, $\bar{X}$ is an **efficient estimator** for θ.

5.5 Exponential Families

Definition 5.3 (Exponential Family)**.** A family of probability density functions (or probability mass functions) is said to be an **Exponential Family** if the log-likelihood function, denoted by $\ell(\theta; \mathbf{x}) = \log f(\mathbf{x}|\theta)$, can be expressed as the sum of three distinct terms:

$$\ell(\theta; \mathbf{x}) = \sum_{i=1}^{k} \eta_i(\theta) T_i(\mathbf{x}) - A(\theta) + \log h(\mathbf{x}) \tag{5.43}$$

Exponentiating this yields the density form:

$$f(\mathbf{x}|\theta) = h(\mathbf{x}) \exp\left\{ \sum_{i=1}^{k} \eta_i(\theta) T_i(\mathbf{x}) - A(\theta) \right\} \tag{5.44}$$

where:

- $\theta = (\theta_1, \dots, \theta_d)$ is the vector of model parameters.
- $\eta_i(\theta)$ are the **natural parameter functions**.
- $\mathbf{T}(\mathbf{x}) = (T_1(\mathbf{x}), \dots, T_k(\mathbf{x}))$ constitutes the vector of **sufficient statistics** for θ.
- $A(\theta)$ is the **log-partition function** (or cumulant function), which ensures the density integrates to 1.
- $h(\mathbf{x})$ is the base measure.

5.5.1 Examples

Exponential Distribution

Example 5.2 (Exponential Distribution)**.** Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Exp}(\theta)$, where θ is the scale parameter.

$$f(\mathbf{x}|\theta) = \theta^{-n} \exp\left\{ -\frac{1}{\theta} \sum_{i=1}^{n} x_i \right\} \tag{5.45}$$

The log-likelihood is:

$$\ell(\theta; \mathbf{x}) = -\frac{1}{\theta} \sum_{i=1}^{n} x_i - n \log \theta \tag{5.46}$$

Identifying the components:

- $\eta_1(\theta) = -\frac{1}{\theta}$
- $T_1(\mathbf{x}) = \sum_{i=1}^{n} x_i$
- $A(\theta) = n \log \theta$
- $\log h(\mathbf{x}) = 0$

Gamma Distribution

Example 5.3 (Gamma Distribution)**.** Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$. The density is:

$$f(\mathbf{x}|\theta) = [\Gamma(\alpha)\beta^{\alpha}]^{-n} \left(\prod_{i=1}^{n} x_i \right)^{\alpha - 1} \exp\left\{ -\frac{1}{\beta} \sum_{i=1}^{n} x_i \right\} \tag{5.47}$$

The log-likelihood is:

$$\ell(\theta; \mathbf{x}) = (\alpha - 1) \sum_{i=1}^{n} \log x_i - \frac{1}{\beta} \sum_{i=1}^{n} x_i - [n \log \Gamma(\alpha) + n\alpha \log \beta] \tag{5.48}$$

Identifying the components:

- $\eta_1(\theta) = \alpha - 1, \quad T_1(\mathbf{x}) = \sum \log x_i$
- $\eta_2(\theta) = -\frac{1}{\beta}, \quad T_2(\mathbf{x}) = \sum x_i$
- $A(\theta) = n \log \Gamma(\alpha) + n\alpha \log \beta$

Beta Distribution

Example 5.4 (Beta Distribution)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Beta}(a, b)$ with $\theta = (a, b)$.

$$\ell(\theta; \mathbf{x}) = (a - 1) \sum_{i=1}^{n} \log x_i + (b - 1) \sum_{i=1}^{n} \log(1 - x_i) - n \log B(a, b) \tag{5.49}$$

This is an exponential family with $k = 2$.

- $\eta_1 = a - 1, T_1 = \sum \log x_i$
- $\eta_2 = b - 1, T_2 = \sum \log(1 - x_i)$
- $A(\theta) = n \log B(a, b)$

Normal Distribution

Example 5.5 (Normal Distribution)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The log-likelihood is:

$$\ell(\theta; \mathbf{x}) = \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 - \left[\frac{n\mu^2}{2\sigma^2} + \frac{n}{2} \log(2\pi\sigma^2) \right] \tag{5.50}$$

Identifying the components:

- $\eta_1 = \frac{\mu}{\sigma^2}, T_1 = \sum x_i$
- $\eta_2 = -\frac{1}{2\sigma^2}, T_2 = \sum x_i^2$
- $A(\theta) = \frac{n\mu^2}{2\sigma^2} + n \log \sigma + \frac{n}{2} \log(2\pi)$

5.5.2 Examples of Non-exponential Families

A model is **not** in the exponential family if the support depends on the parameter.

Uniform Distribution

Example 5.6 (Uniform Distribution)**.** Let $X \sim U(0, \theta)$.

$$\ell(\theta; x) = -\log\theta + \log I(0 < x < \theta) \tag{5.51}$$

The term $\log I(0 < x < \theta)$ couples x and θ in a way that cannot be separated into a sum $\sum \eta_i(\theta)T_i(x)$.

5.5.3 Moments of Sufficient Statistics of Exponential Families

5.5.3.1 Means of Sufficient Statistics (General Case)

Theorem 5.4 (Means via the Score Function)**.** *For a regular exponential family with log-likelihood* $\ell(\theta; \mathbf{x}) = \sum \eta_i(\theta)T_i(\mathbf{x}) - A(\theta) + \log h(\mathbf{x})$*, the expectation of the sufficient statistics can be found by setting the expected score to zero:*

$$E_\theta\left[\frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta_j}\right] = 0 \tag{5.52}$$

Substituting the specific form of $\ell(\theta; \mathbf{X})$*:*

$$\sum_{i=1}^{k} \frac{\partial \eta_i(\theta)}{\partial \theta_j} E[T_i(\mathbf{X})] = \frac{\partial A(\theta)}{\partial \theta_j} \quad \textit{for } j = 1, \dots, d \tag{5.53}$$

Proof. The log-likelihood is:

$$\ell(\theta; \mathbf{x}) = \sum_{i=1}^{k} \eta_i(\theta)T_i(\mathbf{x}) - A(\theta) + \log h(\mathbf{x}) \tag{5.54}$$

Differentiating with respect to θ_j:

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{k} \frac{\partial \eta_i(\theta)}{\partial \theta_j} T_i(\mathbf{x}) - \frac{\partial A(\theta)}{\partial \theta_j} \tag{5.55}$$

Taking the expectation and using the regularity condition $E[\frac{\partial \ell}{\partial \theta_j}] = 0$:

$$E\left[\sum_{i=1}^{k} \frac{\partial \eta_i(\theta)}{\partial \theta_j} T_i(\mathbf{X}) - \frac{\partial A(\theta)}{\partial \theta_j}\right] = 0 \tag{5.56}$$

$$\sum_{i=1}^{k} \frac{\partial \eta_i(\theta)}{\partial \theta_j} E[T_i(\mathbf{X})] = \frac{\partial A(\theta)}{\partial \theta_j} \tag{5.57}$$

□

5.5.3.2 Natural Parameterization

Definition 5.4 (Natural Parameterization (Canonical Form))**.** If the parameterization is chosen such that the natural parameters are the components of the parameter vector itself (i.e., $\eta(\theta) = \theta$), the exponential family is said to be in **Canonical Form** or **Natural Parameterization**.
The log-likelihood for the natural parameter vector $\eta = (\eta_1, \dots, \eta_k)^T$ simplifies to:

$$\ell(\eta; \mathbf{x}) = \sum_{i=1}^{k} \eta_i T_i(\mathbf{x}) - A(\eta) + \log h(\mathbf{x}) \tag{5.58}$$

or in vector notation:

$$\ell(\eta; \mathbf{x}) = \eta^T \mathbf{T}(\mathbf{x}) - A(\eta) + \log h(\mathbf{x}) \tag{5.59}$$

where $A(\eta)$ is the log-partition function.

Definition 5.5 (Full vs. Curved Exponential Families)**.**

- **Full Exponential Family:** When the natural parameters η can vary independently in an open set of $\mathbb{R}^k$ (i.e., $d = k$ and the mapping is a bijection).
- **Curved Exponential Family:** When the dimension of the parameter vector θ is smaller than the number of sufficient statistics ($d < k$), forcing the natural parameters $\eta(\theta)$ to lie on a non-linear curve or surface within the natural parameter space.

Example 5.7 (Curved Exponential Family Example)**.** Consider the $N(\theta, \theta^2)$ distribution ($d = 1$). The log-likelihood is:

$$\ell(\theta; \mathbf{x}) = -\frac{1}{2\theta^2} \sum x_i^2 + \frac{1}{\theta} \sum x_i - n \log \theta - \text{const} \tag{5.60}$$

Here:

- $\eta_1(\theta) = -\frac{1}{2\theta^2}, T_1 = \sum x_i^2$
- $\eta_2(\theta) = \frac{1}{\theta}, T_2 = \sum x_i$

Since $d = 1$ but $k = 2$, and $\eta_1 = -\frac{1}{2}\eta_2^2$, the parameters are constrained to a parabola. This is a **Curved Exponential Family**.

5.5.3.3 Mean and Variance of Sufficient Statistics

Theorem 5.5 (Mean and Variance of Sufficient Statistics)**.** *For an exponential family in canonical form, the log-partition function $A(\eta)$ acts as the **Cumulant Generating Function** for the sufficient statistic vector $\mathbf{T}(\mathbf{X})$. The derivatives of $A(\eta)$ yield the moments of $\mathbf{T}(\mathbf{X})$ as follows:*

1. ***Mean (First Derivative):***

$$E[\mathbf{T}(\mathbf{X})] = \nabla A(\eta) \tag{5.61}$$

2. ***Covariance (Second Derivative):***

$$Var(\mathbf{T}(\mathbf{X})) = \nabla^2 A(\eta) \tag{5.62}$$

Link to Fisher Information: *In the canonical parameterization, the observed information matrix is constant (non-stochastic) and equals the Hessian of $A(\eta)$. Therefore, the covariance of the sufficient statistics is exactly the Fisher Information Matrix:*

$$Var(\mathbf{T}(\mathbf{X})) = \mathbf{I}(\eta) \tag{5.63}$$

This implies that $\mathbf{T}(\mathbf{X})$ is an efficient estimator for the mean parameter $\mathbf{m}(\eta) = E[\mathbf{T}(\mathbf{X})]$, as it achieves the Cramer-Rao Lower Bound with equality (identity link).

Proof. **Derivation**
These results follow directly from Bartlett's Identities (Theorem Theorem 5.1) applied to the canonical log-likelihood:

$$\ell(\eta; \mathbf{x}) = \eta^T \mathbf{T}(\mathbf{x}) - A(\eta) + \log h(\mathbf{x}) \tag{5.64}$$

For the Mean: The score function (gradient of ℓ) is:

$$\mathbf{U}(\eta) = \nabla_\eta \ell(\eta; \mathbf{x}) = \mathbf{T}(\mathbf{x}) - \nabla A(\eta) \tag{5.65}$$

By the First Moment Identity, $E[\mathbf{U}(\eta)] = \mathbf{0}$:

$$E[\mathbf{T}(\mathbf{X}) - \nabla A(\eta)] = \mathbf{0} \implies E[\mathbf{T}(\mathbf{X})] = \nabla A(\eta) \tag{5.66}$$

For the Covariance: The observed information (negative Hessian of ℓ) is:

$$\mathbf{J}(\eta) = -\nabla^2_\eta \ell(\eta; \mathbf{x}) = -\nabla_\eta (\mathbf{T}(\mathbf{x}) - \nabla A(\eta)) = \nabla^2 A(\eta) \tag{5.67}$$

Note that $\mathbf{T}(\mathbf{x})$ is constant with respect to η, so its derivative vanishes. By the Second Moment Identity, $\mathbf{I}(\eta) = E[\mathbf{J}(\eta)] = \text{Cov}(\mathbf{U}(\eta))$. Since $\mathbf{U}(\eta) = \mathbf{T}(\mathbf{X}) - \text{constant}$, $\text{Cov}(\mathbf{U}(\eta)) = \text{Cov}(\mathbf{T}(\mathbf{X}))$. Therefore:

$$\text{Cov}(\mathbf{T}(\mathbf{X})) = E[\nabla^2 A(\eta)] = \nabla^2 A(\eta) \tag{5.68}$$

□

5.5.3.4 Examples

Moments of the Binomial Distribution

Example 5.8 (Moments of the Binomial Distribution)**.** Consider n independent coin flips $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. We find the mean and variance of $T = \sum X_i$.

1. **Log-Likelihood Form**

The standard log-likelihood is:

$$\ell(p;\mathbf{x}) = \log\left(\frac{p}{1-p}\right)\sum x_i + n\log(1-p) \tag{5.69}$$

- Natural Parameter: $\eta = \log\left(\frac{p}{1-p}\right) \implies p = \frac{e^\eta}{1+e^\eta}$.
- Log-Partition Function: $A(\eta) = -n\log(1-p) = n\log(1+e^\eta)$.

Canonical Log-Likelihood $\ell(\eta)$:

$$\ell(\eta;\mathbf{x}) = \eta\left(\sum x_i\right) - n\log(1+e^\eta) \tag{5.70}$$

2. **Calculating Moments**

$$E[T] = \frac{\partial A}{\partial \eta} = n\frac{e^\eta}{1+e^\eta} = np \tag{5.71}$$

$$\text{Var}(T) = \frac{\partial^2 A}{\partial \eta^2} = n\frac{e^\eta(1+e^\eta) - e^\eta(e^\eta)}{(1+e^\eta)^2} = n\frac{e^\eta}{(1+e^\eta)^2} = np(1-p) \tag{5.72}$$

Moments of the Gamma Sufficient Statistic

Example 5.9 (Moments of the Gamma Sufficient Statistic)**.** Consider $X_i \sim \text{Exp}(\lambda)$. We find the moments of $T = \sum X_i$.

1. **Log-Likelihood Form**

   The standard log-likelihood is:

$$\ell(\lambda;\mathbf{x}) = -\lambda\sum x_i + n\log\lambda \tag{5.73}$$

   - Natural Parameter: $\eta = -\lambda$.
   - Log-Partition Function: $A(\eta) = -n\log\lambda = -n\log(-\eta)$.

   Canonical Log-Likelihood $\ell(\eta)$:

$$\ell(\eta;\mathbf{x}) = \eta\left(\sum x_i\right) - [-n\log(-\eta)] = \eta\sum x_i + n\log(-\eta) \tag{5.74}$$

2. **Calculating Moments**

$$E[T] = \frac{\partial A}{\partial \eta} = -n\frac{1}{-\eta}(-1) = -\frac{n}{\eta} = \frac{n}{\lambda} \tag{5.75}$$

$$\text{Var}(T) = \frac{\partial^2 A}{\partial \eta^2} = \frac{\partial}{\partial \eta}\left(-\frac{n}{\eta}\right) = \frac{n}{\eta^2} = \frac{n}{\lambda^2} \tag{5.76}$$

Moments of Normal Sufficient Statistics

Example 5.10 (Moments of Normal Sufficient Statistics)**.** Consider $X_i \sim N(\mu, \sigma^2)$.

1. **Log-Likelihood Form**

   The standard log-likelihood is:

$$\ell(\theta; \mathbf{x}) = \frac{\mu}{\sigma^2} \sum x_i - \frac{1}{2\sigma^2} \sum x_i^2 - \left[\frac{n\mu^2}{2\sigma^2} + \frac{n}{2} \log(2\pi\sigma^2) \right] \tag{5.77}$$

   - Natural Parameters: $\eta_1 = \frac{\mu}{\sigma^2}, \quad \eta_2 = -\frac{1}{2\sigma^2}$.
   - Log-Partition Function (in terms of η): Using $\sigma^2 = -\frac{1}{2\eta_2}$ and $\mu = -\frac{\eta_1}{2\eta_2}$:

$$A(\eta) = -\frac{n\eta_1^2}{4\eta_2} - \frac{n}{2} \log(-2\eta_2) + \frac{n}{2} \log(2\pi) \tag{5.78}$$

   Canonical Log-Likelihood $\ell(\eta)$:

$$\ell(\eta; \mathbf{x}) = \eta_1 \left(\sum x_i \right) + \eta_2 \left(\sum x_i^2 \right) - \left[-\frac{n\eta_1^2}{4\eta_2} - \frac{n}{2} \log(-2\eta_2) \right] \tag{5.79}$$

2. **First Moments (Means)**

$$E[T_1] = E\left[\sum X_i \right] = \frac{\partial A}{\partial \eta_1} = -\frac{2n\eta_1}{4\eta_2} = -\frac{n\eta_1}{2\eta_2} = n\mu \tag{5.80}$$

$$E[T_2] = E\left[\sum X_i^2 \right] = \frac{\partial A}{\partial \eta_2} = \frac{n\eta_1^2}{4\eta_2^2} - \frac{n}{2(-2\eta_2)}(-2) = \frac{n\eta_1^2}{4\eta_2^2} - \frac{n}{2\eta_2} \tag{5.81}$$

   Subbing back μ, σ:

$$= n\mu^2 + n\sigma^2 = n(\mu^2 + \sigma^2) \tag{5.82}$$

3. **Second Moment (Covariance)**

$$\text{Cov}(T_1, T_2) = \frac{\partial^2 A}{\partial \eta_1 \partial \eta_2} = \frac{\partial}{\partial \eta_2} \left(-\frac{n\eta_1}{2\eta_2} \right) = \frac{n\eta_1}{2\eta_2^2} = 2n\mu\sigma^2 \tag{5.83}$$

4. **Independence of $\bar{X}$ and S^2** We verify that $\text{Cov}(\bar{X}, S^2) = 0$.

   Express $\bar{X}$ and S^2 in terms of T_1 and T_2:

$$\bar{X} = \frac{1}{n} T_1 \tag{5.84}$$

$$S^2 = \frac{1}{n-1} \left(\sum X_i^2 - n\bar{X}^2 \right) = \frac{1}{n-1} \left(T_2 - \frac{1}{n} T_1^2 \right) \tag{5.85}$$

   Now compute the covariance (ignoring constants $\frac{1}{n(n-1)}$ for now):

$$\text{Cov}\left(T_1, T_2 - \frac{1}{n} T_1^2 \right) = \text{Cov}(T_1, T_2) - \frac{1}{n} \text{Cov}(T_1, T_1^2) \tag{5.86}$$

We need $\text{Cov}(T_1, T_1^2)$. Since $T_1 = \sum X_i \sim N(n\mu, n\sigma^2)$, we use the property of the normal distribution that for $Y \sim N(\theta, \tau^2)$, $\text{Cov}(Y, Y^2) = 2\theta\tau^2$. Here $\theta = n\mu$ and $\tau^2 = n\sigma^2$:

$$\text{Cov}(T_1, T_1^2) = 2(n\mu)(n\sigma^2) = 2n^2\mu\sigma^2 \tag{5.87}$$

Substituting this back into the expression:

$$\text{Cov}\left(T_1, T_2 - \frac{1}{n}T_1^2\right) = \underbrace{2n\mu\sigma^2}_{\text{From Part 3}} - \frac{1}{n}(2n^2\mu\sigma^2) = 2n\mu\sigma^2 - 2n\mu\sigma^2 = 0 \tag{5.88}$$

Since $\bar{X}$ and S^2 are uncorrelated and derived from normally distributed data, they are **independent**.

6 Maximum Likelihood Estimation

6.1 Definitions

1. **Likelihood Function** Let $f(x|\theta)$ be the probability density function (or mass function). The likelihood function is:
$$L(\theta; x) = f(x|\theta) \tag{6.1}$$

2. **Log-likelihood**
$$l(\theta; x) = \log L(\theta; x) = \log f(x|\theta) \tag{6.2}$$

3. **Score Function** The score function is the derivative of the log-likelihood with respect to the parameter θ:
$$S(\theta; x) = \frac{\partial}{\partial\theta} l(\theta; x) = \frac{\partial}{\partial\theta} \log L(\theta; x) \tag{6.3}$$

4. **Maximum Likelihood Estimator (MLE)** The MLE is the value that maximizes the likelihood function:
$$\hat{\theta}_{\text{MLE}}(x) = \text{argmax}_\theta\, L(\theta; x) = \text{argmax}_\theta\, l(\theta; x) \tag{6.4}$$
An approach to finding $\hat{\theta}$ is to solve the score equation:
$$\forall_\theta, \quad S(\theta; x) = 0 \tag{6.5}$$

Example 6.1 (Uniform Distribution MLE)**.** Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$.
The likelihood function is:
$$L(\theta; x) = \prod_{i=1}^{n} f(x_i|\theta) = \frac{1}{\theta^n} I(X_{(n)} < \theta) \tag{6.6}$$
where $X_{(n)} = \max\{X_1, \dots, X_n\}$.
To maximize this function, we observe that $L(\theta)$ decreases as θ increases, but θ must be at least $X_{(n)}$. Therefore:
$$\hat{\theta}_{\text{MLE}}(x) = X_{(n)} \tag{6.7}$$

Properties of this estimator: The CDF of $X_{(n)}$ is:
$$P(X_{(n)} \le x) = [P(X_1 \le x)]^n = \left(\frac{x}{\theta}\right)^n \quad \text{for } 0 < x < \theta \tag{6.8}$$

The PDF is $f_{X_{(n)}}(x) = n\left(\frac{x}{\theta}\right)^{n-1} \frac{1}{\theta}$.

The expected value is:

$$E(X_{(n)}) = \int_0^\theta x \cdot \frac{nx^{n-1}}{\theta^n} dx = \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{\theta^n} \left[\frac{x^{n+1}}{n+1} \right]_0^\theta = \frac{n}{n+1}\theta < \theta \tag{6.9}$$

Thus, it is a biased estimator.

Example 6.2 (Normal Distribution MLE)**.** Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Let $\theta = (\mu, \sigma^2)$.
The likelihood is:

$$L(\theta; x) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum (x_i - \mu)^2}{2\sigma^2} \right) \tag{6.10}$$

The log-likelihood is:

$$l(\theta; x) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{\sum (x_i - \mu)^2}{2\sigma^2} \tag{6.11}$$

Score Functions:

1. With respect to μ:

$$\frac{\partial l}{\partial \mu} = \frac{2\sum (x_i - \mu)}{2\sigma^2} = \frac{\sum (x_i - \mu)}{\sigma^2} = 0 \implies \hat{\mu}_{\text{MLE}} = \bar{x} \tag{6.12}$$

2. With respect to σ^2:

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum (x_i - \mu)^2}{2(\sigma^2)^2} = 0 \tag{6.13}$$

   Solving for σ^2:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{\sum (x_i - \hat{\mu})^2}{n} = \frac{\sum (x_i - \bar{x})^2}{n} \tag{6.14}$$

   Bias:

   - $E(S^2) = \sigma^2$ (Unbiased)
   - $E(\hat{\sigma}^2_{\text{MLE}}) = E\left(\frac{n-1}{n} S^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$ (Biased)

6.2 Properties of Score and Fisher Information

Definition 6.1 (Definition of Fisher Information)**.** Some properties about the Score function $S(\theta; x)$:

1. **Mean:** $E[S(\theta; x)|\theta] = 0$.
2. **Variance/Covariance:**

$$\text{Cov}(S_i(\theta; x), S_j(\theta; x)) = -E\left[\frac{\partial^2 l(\theta; x)}{\partial \theta_i \partial \theta_j} \right] \tag{6.15}$$

The Fisher Information matrix $I(\theta)$ is defined as:

$$I(\theta) = \text{Cov}(S(\theta; x)) = E[S(\theta; x)S(\theta; x)^T] = -E\left[\frac{\partial^2 l(\theta; x)}{\partial \theta^2}\right] \tag{6.16}$$

Note: $J(\theta, x) = -\frac{\partial^2}{\partial \theta_i \partial \theta_k} l(\theta; x)$ is the Observed Fisher Information. $I(\theta) = E[J(\theta, x)]$.

Theorem 6.1 (Properties of Score Function)**.** *Given the support of X is free of θ:*

1. $E_X[S(\theta; x)] = 0$
2. $Cov(S(\theta; x)) = I(\theta)$

Proof. **Proof of Mean 0:**

$$\begin{aligned} E[S(\theta; X)] &= \int \frac{\partial \log f(x|\theta)}{\partial \theta} f(x|\theta) dx \\ &= \int \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta} f(x|\theta) dx \\ &= \int \frac{\partial f(x|\theta)}{\partial \theta} dx \\ &= \frac{\partial}{\partial \theta} \int f(x|\theta) dx \quad \text{(assuming regularity conditions allow interchange)} \\ &= \frac{\partial}{\partial \theta}(1) = 0 \end{aligned} \tag{6.17}$$

Proof of Variance: Differentiating $\int f(x|\theta)dx = 1$ twice with respect to θ leads to the identity:

$$\text{Var}(S(\theta)) = E\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 l}{\partial \theta^2}\right] \tag{6.18}$$

□

Remark 6.1. If $X_1, \ldots, X_n$ are i.i.d with $f(x|\theta)$, then:

$$l(\theta; x) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{6.19}$$

The score function is the sum of individual score functions. Since variance of a sum of independent variables is the sum of variances:

$$I_n(\theta) = nI_1(\theta) \tag{6.20}$$

where $I_1(\theta) = E\left[-\frac{\partial^2}{\partial \theta^2} \log f(X_1|\theta)\right]$.

6.3 Asymptotic Properties of MLE

There are three main asymptotic properties:

1. **Consistency:** $\hat{\theta}_n \xrightarrow{p} \theta_0$
2. **Asymptotic Normality:** $\hat{\theta}_n \sim N(\theta_0, 1/I(\theta_0))$ roughly.
3. **Efficiency:** Variance achieves CRLB asymptotically.

6.4 Review of Convergence

1. **LLN (Law of Large Numbers):** $\bar{X} \xrightarrow{p} E(X)$.
2. **CLT (Central Limit Theorem):** $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2)$.
3. **Slutsky's Theorem:** If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} a$ (constant), then: * $X_n + Y_n \xrightarrow{d} X + a$ * $X_n Y_n \xrightarrow{d} aX$ * $X_n/Y_n \xrightarrow{d} X/a$

Example 6.3. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. $\hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2}{n}$. We can show:

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4) \tag{6.21}$$

Using CLT on $Y_i = (X_i - \mu)^2$ and Slutsky's theorem.

6.5 Consistency

Theorem 6.2 (Consistency)**.** *Under regularity conditions, let θ_0 be the true parameter. Then $\hat{\theta}_n \xrightarrow{p} \theta_0$.*

Proof Idea: $\hat{\theta}_n$ maximizes $\frac{1}{n} l(\theta; x)$. By LLN, $\frac{1}{n} l(\theta; x) \xrightarrow{p} E[\log f(X|\theta)]$. We compare the expected log-likelihood at θ vs θ_0:

$$E\left[\log \frac{f(X|\theta)}{f(X|\theta_0)}\right] \leq \log E\left[\frac{f(X|\theta)}{f(X|\theta_0)}\right] \quad \text{(Jensen's Inequality)} \tag{6.22}$$

$$= \log \int f(x|\theta_0) \frac{f(x|\theta)}{f(x|\theta_0)} dx = \log \int f(x|\theta) dx = \log(1) = 0 \tag{6.23}$$

Thus $E[\log f(X|\theta)]$ is maximized at $\theta = \theta_0$.

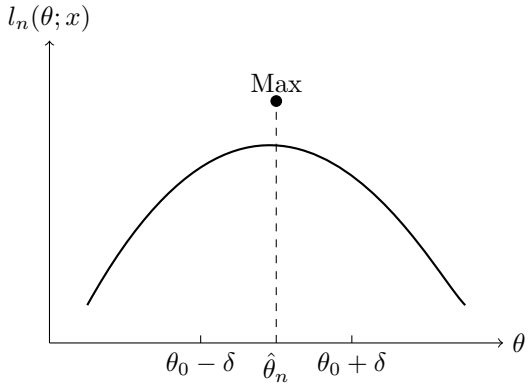


Figure 6.1: Log-likelihood function maximized at MLE

6.6 Asymptotic Normality

Theorem 6.3 (Asymptotic Normality of MLE)**.** *Under regularity conditions:*

1. *Support of $f(x|\theta)$ does not depend on θ.*
2. *Likelihood is twice continuously differentiable.*
3. *Fisher Information exists and is positive.*

Then:

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I_1(\theta_0)}\right) \tag{6.24}$$

Proof. **Taylor Expansion Method:** Expand the score function $l'(\theta)$ around the true parameter θ_0:

$$l'(\hat{\theta}_n) = l'(\theta_0) + (\hat{\theta}_n - \theta_0)l''(\theta_n^*) \tag{6.25}$$

where θ_n^* lies between $\hat{\theta}_n$ and θ_0. Since $\hat{\theta}_n$ is the MLE, $l'(\hat{\theta}_n) = 0$.

$$0 = l'(\theta_0) + (\hat{\theta}_n - \theta_0)l''(\theta_n^*) \tag{6.26}$$

Rearranging:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\frac{1}{\sqrt{n}}l'(\theta_0)}{\frac{1}{n}l''(\theta_n^*)} \tag{6.27}$$

We analyze the numerator and denominator:

1. **Numerator:** $E[l'(\theta_0)] = 0$, $\text{Var}(l'(\theta_0)) = nI_1(\theta_0)$. By CLT: $\frac{1}{\sqrt{n}}l'(\theta_0) \xrightarrow{d} N(0, I_1(\theta_0))$.
2. **Denominator:** By LLN and Consistency, $\frac{1}{n}l''(\theta_n^*) \xrightarrow{p} E[l''(\theta_0)] = -I_1(\theta_0)$.

Combining via Slutsky's Theorem:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \frac{N(0, I_1(\theta_0))}{I_1(\theta_0)} \sim N\left(0, \frac{1}{I_1(\theta_0)}\right) \tag{6.28}$$

□

Remark 6.2. For a vector parameter $\theta \in \mathbb{R}^p$:

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} N_p(0, I(\theta_0)^{-1}) \tag{6.29}$$

where $I(\theta_0)$ is the Fisher Information Matrix.

6.7 Hypothesis Testing: Likelihood Ratio Test

Consider testing:

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta \ \ \Theta_0 \tag{6.30}$$

where $\dim(\Theta) = p$ and $\dim(\Theta_0) = p - m$.

The Likelihood Ratio Statistic is:

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \tag{6.31}$$

Theorem 6.4 (Wilks' Theorem)**.** *Under regularity conditions, under H_0:*

$$-2 \log \Lambda \xrightarrow{d} \chi^2_m \tag{6.32}$$

where m is the difference in dimensions (number of restrictions).

Example 6.4. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$. Here $p = 2$ (μ, σ^2) and under H_0, free parameters = 1 (σ^2). So $m = 2 - 1 = 1$.
The statistic Λ:

$$\Lambda = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \right)^{n/2} = \left(\frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \mu_0)^2} \right)^{n/2} \tag{6.33}$$

It can be shown that:

$$\Lambda = \left(1 + \frac{(\bar{x} - \mu_0)^2}{\hat{\sigma}^2} \right)^{-n/2} \tag{6.34}$$

The rejection region $-2 \log \Lambda > \chi^2_{1,\alpha}$ is equivalent to the t-test:

$$\left| \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \right| > t_{n-1,\alpha/2} \tag{6.35}$$

6.8 Illustration of Wilks' Theorem

We can approximate the statistic using Taylor expansion. Consider the scalar case.

$$-2 \log \Lambda = 2[l(\hat{\theta}) - l(\theta_0)] \tag{6.36}$$

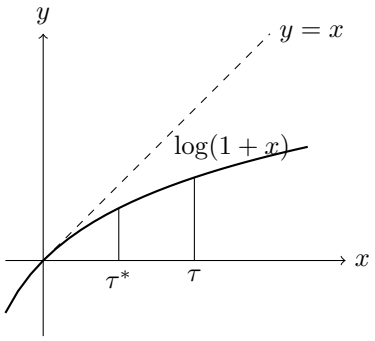


Figure 6.2: Approximation of Likelihood Ratio

Using Taylor expansion around the MLE $\hat{\theta}$:

$$l(\theta_0) \approx l(\hat{\theta}) + (\theta_0 - \hat{\theta})l'(\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})^2 l''(\hat{\theta}) \tag{6.37}$$

Since $l'(\hat{\theta}) = 0$:

$$2[l(\hat{\theta}) - l(\theta_0)] \approx -(\theta_0 - \hat{\theta})^2 l''(\hat{\theta}) = (\hat{\theta} - \theta_0)^2[-\frac{1}{n}l''(\hat{\theta})] \cdot n \tag{6.38}$$

As $n \to \infty$, $-\frac{1}{n}l'' \to I(\theta_0)$ and $\sqrt{n}(\hat{\theta} - \theta_0) \to N(0, 1/I)$. Thus, the expression behaves like $Z^2 \sim \chi_1^2$.

6.9 An Example: Poisson Regression

In this example, we explore the Generalized Linear Model (GLM) for count data using the Poisson distribution. Let $Y_1, \ldots, Y_n$ be independent count variables where $Y_i \sim \text{Poisson}(\lambda_i)$. The expected count λ_i is related to a vector of covariates $\mathbf{x}_i$ and parameters β via the **canonical log link function**:

$$\log(\lambda_i) = \eta_i = \mathbf{x}_i^\top \beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \tag{6.39}$$

6.9.1 Canonical Representation

We begin by expressing the log-likelihood in the canonical exponential family form. The probability mass function for the Poisson distribution is $P(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$. The log-likelihood is:

$$\ell(\beta) = \sum_{i=1}^{n} (y_i \log(\lambda_i) - \lambda_i - \log(y_i!)) \tag{6.40}$$

Substituting the log link $\log(\lambda_i) = \mathbf{x}_i^\top \beta$ and $\lambda_i = e^{\mathbf{x}_i^\top \beta}$:

$$\ell(\beta) = \sum_{i=1}^{n} \left(y_i(\mathbf{x}_i^\top \beta) - e^{\mathbf{x}_i^\top \beta} \right) + \text{const} \tag{6.41}$$

Rearranging terms to isolate the parameters β_j:

$$\ell(\beta; \mathbf{y}) = \sum_{j=0}^{k} \beta_j \underbrace{\left(\sum_{i=1}^{n} y_i x_{ij} \right)}_{T_j(\mathbf{y})} - \underbrace{\sum_{i=1}^{n} e^{\mathbf{x}_i^\top \beta}}_{A(\beta)} + \text{const} \tag{6.42}$$

From this form, we identify:

- **Sufficient Statistics:** $\mathbf{T}(\mathbf{y}) = \mathbf{X}^\top \mathbf{y}$.
- **Log-Partition Function:** $A(\beta) = \sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} e^{\mathbf{x}_i^\top \beta}$.

6.9.2 The Score and Information

Next, we derive the gradient and Hessian of the log-likelihood.

- **Score Vector (U):** The gradient of $\ell(\beta)$ is the difference between the observed sufficient statistics and their expectations (derived from ∇A).

$$\mathbf{U}(\beta) = \nabla_\beta \ell = \mathbf{T}(\mathbf{y}) - \nabla A(\beta) = \sum_{i=1}^{n} y_i \mathbf{x}_i - \sum_{i=1}^{n} \lambda_i \mathbf{x}_i = \mathbf{X}^\top (\mathbf{y} - \lambda) \tag{6.43}$$

- **Fisher Information Matrix ($\mathcal{J}$):** This is the negative Hessian of the log-likelihood, or equivalently the Hessian of the log-partition function $A(\beta)$.

$$\mathcal{J}(\beta) = \nabla^2 A(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \frac{\partial \lambda_i}{\partial \beta^\top} = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^\top \tag{6.44}$$

  In matrix notation, $\mathcal{J}(\beta) = \mathbf{X}^\top \mathbf{W} \mathbf{X}$, where $\mathbf{W} = \text{diag}(\lambda_i)$. Note that for the Poisson model, the variance equals the mean λ_i.

6.9.3 Asymptotic Distributions (Theory)

a. **Normality of the Score:** The Score vector $\mathbf{U}(\beta) = \sum_{i=1}^{n} (Y_i - \lambda_i)\mathbf{x}_i$ is a sum of independent mean-zero random vectors.

   - Mean: $E[\mathbf{U}] = \mathbf{0}$.
   - Variance: $\text{Var}(\mathbf{U}) = \sum \text{Var}(Y_i)\mathbf{x}_i \mathbf{x}_i^\top = \mathcal{J}(\beta)$. By the **Multivariate Central Limit Theorem**, as $n \to \infty$:

$$\frac{1}{\sqrt{n}} \mathbf{U}(\beta) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}(\beta)) \tag{6.45}$$

     where $\mathcal{I}$ is the limit of $\frac{1}{n}\mathcal{J}$.

b. **Normality of the Estimator:** The MLE $\hat{\beta}$ satisfies $\mathbf{U}(\hat{\beta}) = \mathbf{0}$. Taking a first-order Taylor expansion around the true parameter β:

$$\mathbf{0} = \mathbf{U}(\hat{\beta}) \approx \mathbf{U}(\beta) - \mathcal{J}(\beta)(\hat{\beta} - \beta) \tag{6.46}$$

Rearranging gives $(\hat{\beta} - \beta) \approx \mathcal{J}^{-1}(\beta)\mathbf{U}(\beta)$. Since $\mathbf{U}$ is asymptotically normal, the linear transformation implies:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{I}^{-1}(\beta)\right) \tag{6.47}$$

6.9.4 Numerical Application in R

We now implement the Newton-Raphson algorithm in R to estimate parameters for a Poisson regression model with a **continuous covariate**. Let $\log(\lambda_i) = \beta_0 + \beta_1 x_i$.

a. Data Generation We simulate $n = 100$ observations using a continuous predictor x drawn from a Uniform distribution.

```r
set.seed(123)
n <- 100
x <- runif(n, 0, 1)

# True parameters: Intercept=0.5, Slope=2.0
beta_true <- c(0.5, 2.0)

# Generate response Y
lambda_true <- exp(beta_true[1] + beta_true[2] * x)
y <- rpois(n, lambda_true)

# Design Matrix (intercept column + covariate)
X <- cbind(1, x)
```

b. Newton-Raphson Implementation The update rule is $\beta^{(t+1)} = \beta^{(t)} + \mathcal{J}^{-1}\mathbf{U}$. We implement this iteratively.

```r
newton_raphson_poisson <- function(X, y, tol = 1e-6, max_iter = 100) {
  beta <- rep(0, ncol(X)) # Start at 0

  for (i in 1:max_iter) {
    # 1. Compute means and weights
    eta <- X %*% beta
    lambda <- as.vector(exp(eta))

    # 2. Compute Score U and Info J
    U <- crossprod(X, y - lambda)
    J <- crossprod(X * lambda, X) # Efficient t(X) %*% W %*% X

    # 3. Update beta
```

```
    delta <- solve(J, U)
    beta_new <- beta + as.vector(delta)

    # 4. Check convergence
    diff <- sum((beta_new - beta)^2)
    cat(sprintf("Iter %d: beta0=%.4f, beta1=%.4f, diff=%.6f\n",
                i, beta_new[1], beta_new[2], diff))

    if (diff < tol) return(beta_new)
    beta <- beta_new
  }
}

beta_mle <- newton_raphson_poisson(X, y)
```

Output:

```
Iter 1: beta0=1.1718, beta1=0.6234, diff=1.761592
Iter 2: beta0=0.6970, beta1=1.5794, diff=1.139433
Iter 3: beta0=0.4996, beta1=2.0101, diff=0.224446
Iter 4: beta0=0.4725, beta1=2.0838, diff=0.006173
Iter 5: beta0=0.4722, beta1=2.0845, diff=0.000001
```

c. Discussion & Verification We compare our manual estimates with R's built-in `glm` function.

```
cat("Manual MLE: ", beta_mle, "\n")
cat("GLM Output: ", coef(glm(y ~ x, family = "poisson")))
```

Observation: The manual implementation converges to the exact same values as the built-in function ($\hat{\beta}_0 \approx 0.47, \hat{\beta}_1 \approx 2.08$). Notice that we started at $\beta = (0, 0)$. In the first iteration, the algorithm took a large step because the log-likelihood surface is steep. The quadratic convergence of Newton-Raphson is evident in the rapid decrease of the difference term (from 1.13 to 0.22 to 0.006).

7 Minimum Variance Estimators

7.1 Completeness

Definition 7.1 (Complete Statistic)**.** A statistic T is said to be **complete** if for any real-valued function g,

$$E[g(T)|\theta] = 0 \quad \text{for all } \theta \tag{7.1}$$

implies

$$P(g(T) = 0|\theta) = 1 \quad \text{for all } \theta \tag{7.2}$$

Significance: If T is complete, then there exists at most one unbiased estimator for θ that is a function of T.

Example 7.1 (Uniform Distribution (Not Complete))**.** Let $X_1, \dots, X_n \sim \text{Unif}(\theta - 1, \theta + 1)$. The density is:

$$f(x) = \prod I(\theta - 1 < x_i < \theta + 1) = I(\theta \in (x_{(n)} - 1, x_{(1)} + 1)) \tag{7.3}$$

The statistic $T(X) = (X_{(1)}, X_{(n)})$ is a Minimal Sufficient Statistic. However, it is **not complete**.
Consider the range $R = X_{(n)} - X_{(1)}$. The distribution of R does not depend on θ (it is an ancillary statistic). Let $g(T) = X_{(n)} - X_{(1)} - c$, where $c = E[X_{(n)} - X_{(1)}]$. Then $E[g(T)] = 0$ for all θ, but $g(T)$ is not identically zero.

Lemma 7.1 (Exponential Family Completeness)**.** *If $T = (T_1, \dots, T_k)$ is the natural statistic of an exponential family that contains an open rectangle in the parameter space, then T is complete.*

7.2 UMVUE

Definition 7.2 (Uniformly Minimum Variance Unbiased Estimator (UMVUE))**.** A statistic $T(x)$ is a UMVUE for θ if:

1. $E(T(x)|\theta) = \theta$ for all θ (Unbiased).
2. $Var(T(x)|\theta) \leq Var(d(x)|\theta)$ for all θ and for all other unbiased estimators $d(x)$.

The relationship between statistics types is visualized below:

Theorem 7.1 (Lehmann-Scheffe Theorem)**.** *If T is a complete and sufficient statistic, and there is an unbiased estimator $d(X)$ such that $E[d(X)] = \theta$, then $\phi(T) = E[d(X)|T]$ is the unique UMVUE for θ.*

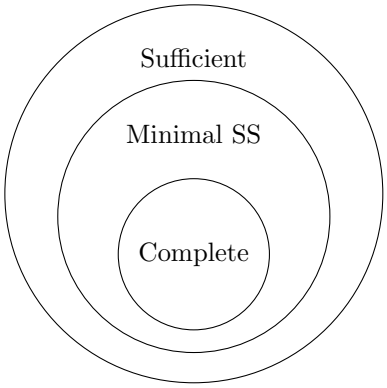


Figure 7.1: Hierarchy of Statistics

Theorem 7.2 (Rao-Blackwell Theorem)**.** *Given that T is a sufficient statistic and $d_1(x)$ is an unbiased estimator ($E[d_1(x)] = \theta$). Define $g(T) = E[d_1(x)|T]$. Then:*

1. *$g(T)$ is a statistic (free of θ because T is sufficient).*
2. *$E[g(T)] = \theta$ (Unbiased).*
3. *$Var(g(T)) \leq Var(d_1(x))$.*

Proof. **Proof of Rao-Blackwell:**

1. Since T is sufficient, the conditional distribution $X|T$ is independent of θ, so $g(T)$ is a valid statistic.
2. By the Law of Iterated Expectations:

$$E[g(T)] = E_T[E_X(d_1(X)|T)] = E_X[d_1(X)] = \theta \tag{7.4}$$

3. By the variance decomposition formula:

$$Var(d_1(X)) = Var(E[d_1(X)|T]) + E[Var(d_1(X)|T)] \tag{7.5}$$

$$Var(d_1(X)) = Var(g(T)) + E[(d_1(X) - g(T))^2|T] \tag{7.6}$$

Since $(d_1(X) - g(T))^2 \geq 0$, we have $Var(g(T)) \leq Var(d_1(X))$.

□

7.3 Methods for Finding UMVUE

To find the UMVUE for a parameter θ:

1. **Find a Complete Sufficient Statistic:** Identify T which is complete and sufficient for θ (often using the Exponential Family properties).

2. **Find an Unbiased Estimator:** Find any simple statistic $d(X)$ such that $E[d(X)] = \theta$.
3. **Rao-Blackwellize:** Compute $g(T) = E[d(X)|T]$. The result $g(T)$ is the UMVUE.

Example 7.2 (Poisson UMVUE)**.** Let $X_1, \ldots, X_n \sim$ Poisson(λ). Find the UMVUE for λ and λ^2.

1. For λ: $T = \sum X_i$ is a complete sufficient statistic (Poisson is exponential family). Let $d_1(X) = X_1$. $E[X_1] = \lambda$. We compute $g(T) = E[X_1|T]$. Since the conditional distribution of X_1 given $T = t$ is Binomial$(t, 1/n)$:

$$E[X_1|T] = t \cdot \frac{1}{n} = \frac{T}{n} = \bar{X} \tag{7.7}$$

Thus, $\bar{X}$ is the UMVUE for λ.

2. For λ^2: We know $Var(X_1) = \lambda = E(X_1^2) - (E(X_1))^2$. So $E(X_1^2) - \lambda = \lambda^2$, which implies $E(X_1^2 - X_1) = \lambda^2$. Let $d_2(X) = X_1^2 - X_1$. This is an unbiased estimator for λ^2.

We calculate $g(T) = E[X_1^2 - X_1|T]$.

$$g(T) = E[X_1^2|T] - E[X_1|T] \tag{7.8}$$

Using the second moment of the Binomial distribution $Bin(T, 1/n)$: $E[X_1^2|T] = Var(X_1|T) + (E[X_1|T])^2 = T\frac{1}{n}(1 - \frac{1}{n}) + (\frac{T}{n})^2$.

$$g(T) = \left[\frac{T}{n} - \frac{T}{n^2} + \frac{T^2}{n^2}\right] - \frac{T}{n} = \frac{T^2 - T}{n^2} = \frac{T(T-1)}{n^2} \tag{7.9}$$

Thus, $\frac{T(T-1)}{n^2}$ is the UMVUE for λ^2.

8 Hypothesis Testing

8.1 General Terminologies

8.1.1 Hypothesis Testing

We formulate the problem of hypothesis testing as deciding between two competing claims about a parameter θ:

$$H_0 : \theta \in \Theta_0 \quad \text{(Null Hypothesis)} \tag{8.1}$$

$$H_1 : \theta \in \Theta_1 \quad \text{(Alternative Hypothesis)} \tag{8.2}$$

Definition 8.1 (Simple and Composite Hypotheses)**.** A hypothesis is called **simple** if it specifies a single value for the parameter (e.g., Θ_0 contains only one point). It is called **composite** if it specifies more than one value.

Example 8.1 (Normal Mean Test)**.** Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

- If σ^2 is known, $H_0 : \mu = \mu_0$ is a simple hypothesis.
- If σ^2 is unknown, $H_0 : \mu = \mu_0$ is a composite hypothesis (since σ^2 can vary).

8.1.2 Test Functions and Size

A test is defined by a **critical region** C_α such that we reject H_0 if the data $x \in C_\alpha$. Equivalently, we can define a **test function** $\phi(x)$ representing the probability of rejecting H_0 given data x.

- A non-randomized test is given as follows:

$$\phi(x) = I(x \in C_\alpha) = \begin{cases} 1 & \text{if } x \in C_\alpha \text{ (Reject } H_0\text{)} \\ 0 & \text{otherwise} \end{cases} \tag{8.3}$$

- A randomized test, $\phi(x)$ can take values in $[0, 1]$, which can be expressed typically as follows:

$$\phi(x) = \begin{cases} 1 & \text{if } x \in C_1 \\ \gamma & \text{if } x \in C_* \\ 0 & \text{otherwise} \end{cases} \tag{8.4}$$

where:

  - C_1 is the region where we strictly reject H_0.
  - C_* is the boundary region (often where $T(x) = k$) where we reject H_0 with probability γ.

- More generally, $\phi(x)$ is just a function of x with values in $[0, 1]$, which represents the probability that we will reject H_0.

Example 8.2 (Randomized Test for Binomial)**.** Let $X \sim \text{Bin}(n = 10, \theta)$. Consider testing $H_0 : \theta = 1/2$ vs $H_1 : \theta > 1/2$ with target size $\alpha = 0.05$.
Suppose we choose a critical region $X \geq k$.

- If $k = 9$, $P(X \geq 9|\theta = 0.5) \approx 0.0107$.
- If $k = 8$, $P(X \geq 8|\theta = 0.5) \approx 0.0547$.

Since we cannot achieve exactly 0.05 with a non-randomized test (the survival function jumps over 0.05), we must use a randomized test function.
The randomized test is defined as:

$$\phi(x) = \begin{cases} 1 & \text{if } x \in C_1 \text{ (i.e., } x \geq 9) \\ \gamma & \text{if } x \in C_* \text{ (i.e., } x = 8) \\ 0 & \text{otherwise} \end{cases} \tag{8.5}$$

From the figure, we see that $\alpha = 0.05$ lies between $P(X \geq 9)$ and $P(X \geq 8)$. We always reject the "tail" where probabilities are strictly less than α (here $x \geq 9$). At the boundary $x = 8$, we cannot reject with probability 1 (which would give total size 0.0547), nor with probability 0 (which would give total size 0.0107). We choose γ to bridge this gap:

$$\begin{aligned} \alpha &= P(X \geq 9) + \gamma \cdot P(X = 8) \\ 0.05 &= 0.01074 + \gamma \cdot (P(X \geq 8) - P(X \geq 9)) \\ 0.05 &= 0.01074 + \gamma \cdot (0.05469 - 0.01074) \end{aligned} \tag{8.6}$$

Solving for γ:

$$\gamma = \frac{0.05 - 0.01074}{0.04395} \approx \frac{39}{44} \approx 0.89 \tag{8.7}$$

8.2 Power and Size Function

Definition 8.2 (Size of a Test)**.** The **size** of a test $\phi(x)$ is the maximum probability of rejecting the null hypothesis when it is true:

$$\text{Size}(\phi) = \sup_{\theta \in \Theta_0} W_\phi(\theta) = \sup_{\theta \in \Theta_0} E_\theta[\phi(X)] \tag{8.8}$$

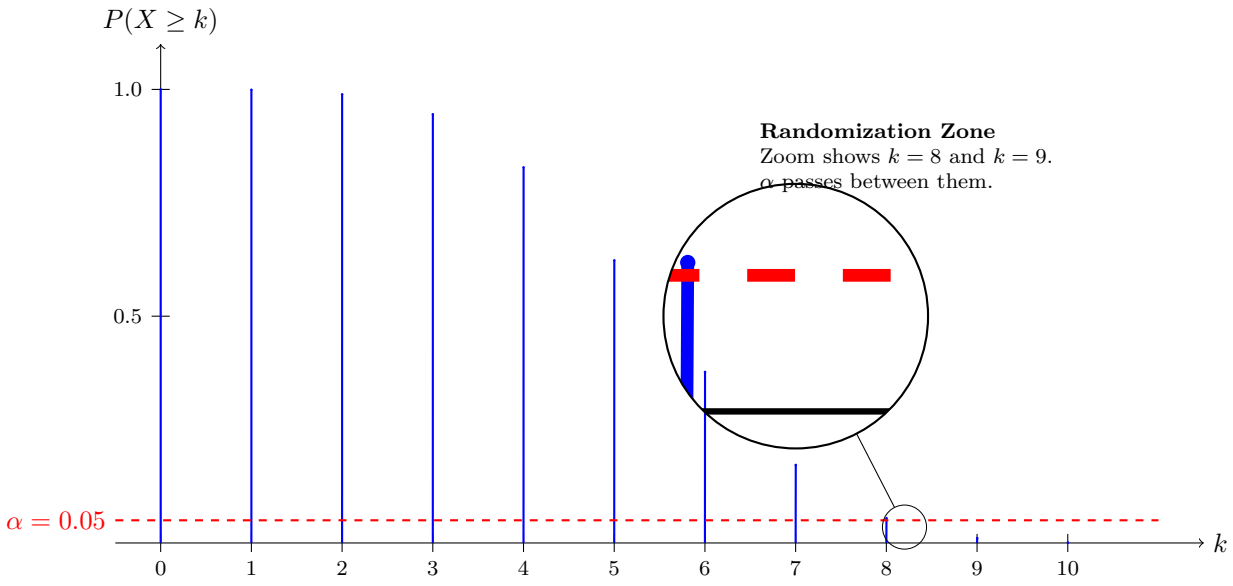


Figure 8.1: Survival Function P(X >= k) with randomization detail

8.2.1 Power Definitions

We distinguish between the power function varying over parameters and the power metric of a specific test.

1. Power Function ($W_\phi(\theta)$) The probability of rejecting H_0 as a function of the parameter θ:

$$W_\phi(\theta) = E_\theta[\phi(X)] \tag{8.9}$$

2. Power of the Test (Power(ϕ)) In the context of a specific alternative hypothesis (e.g., $H_1 : \theta = \theta_1$), we define the power as a scalar functional of ϕ:

$$\text{Power}(\phi) = E_{\theta_1}[\phi(X)] \tag{8.10}$$

Ideally, we want:

- $W_\phi(\theta) \leq \text{Size}(\phi)$ for all $\theta \in \Theta_0$ (Control Type I error).
- Power(ϕ) to be as large as possible (Maximize sensitivity to H_1).

8.3 The Neyman-Pearson Lemma

Consider testing a simple null against a simple alternative: $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$.

We define the **Likelihood Ratio** $\Lambda(x)$ as:

$$\Lambda(x) = \frac{f_1(x)}{f_0(x)} = \frac{f(x;\theta_1)}{f(x;\theta_0)} \tag{8.11}$$

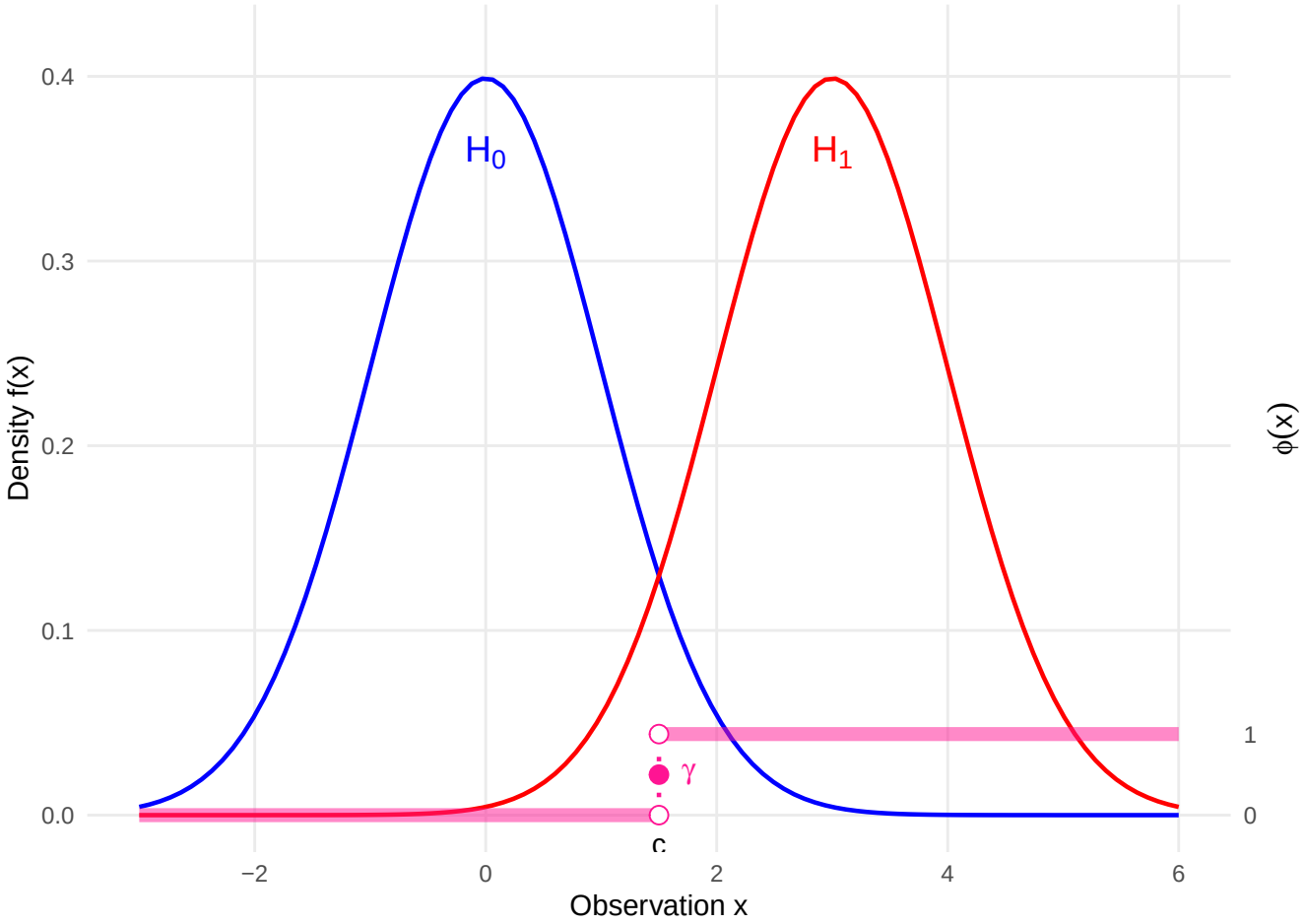


Figure 8.2: Illustration of a Test Function $\phi(x)$ (Pink) relative to Size (H_0, Blue) and Power (H_1, Red).

Definition 8.3 (Likelihood Ratio Test (LRT))**.** A test $\phi(x)$ is a Likelihood Ratio Test if it has the form:

$$\phi(x) = \begin{cases} 1 & \text{if } \Lambda(x) > k \\ \gamma(x) & \text{if } \Lambda(x) = k \\ 0 & \text{if } \Lambda(x) < k \end{cases} \tag{8.12}$$

where $k \geq 0$ is a constant and $0 \leq \gamma(x) \leq 1$.

8.3.1 Neyman-Pearson Lemma

Theorem 8.1 (Neyman-Pearson Lemma)**.**

a) ***Optimality:*** *For any k and $\gamma(x)$, the LRT $\phi_0(x)$ defined above has maximum power among all tests whose size is less than or equal to the size of $\phi_0(x)$.*

b) ***Existence:*** *Given $\alpha \in (0, 1)$, there exist constants k and γ_0 such that the LRT defined by this k and $\gamma(x) = \gamma_0$ has size exactly α.*

c) ***Uniqueness:*** *If a test ϕ has size α and is of maximum power among all tests of size α, then ϕ is necessarily an LRT, except possibly on a set of measure zero under H_0 and H_1.*

8.3.2 A Derivation with The Lagrange Multiplier Approach

To make the optimality of the Likelihood Ratio Test (LRT) intuitive, we can frame the search for the best test function $\phi(x)$ as a constrained optimization problem.

We want to maximize the power of the test:

$$\text{Power}(\phi) = \int \phi(x) f_1(x) dx \tag{8.13}$$

subject to the constraint on the size of the test α:

$$\text{Size}(\phi) = \int \phi(x) f_0(x) dx = \alpha \tag{8.14}$$

Using the method of Lagrange multipliers, we define the objective function L with a multiplier k:

$$L(\phi, k) = \int \phi(x) f_1(x) dx - k \left(\int \phi(x) f_0(x) dx - \alpha \right) \tag{8.15}$$

Rearranging the terms inside the integral, we get:

$$L(\phi, k) = \int \phi(x)[f_1(x) - kf_0(x)]dx + k\alpha \tag{8.16}$$

To maximize L with respect to $\phi(x)$, we look at the integrand. Since $0 \leq \phi(x) \leq 1$, we should choose $\phi(x)$ to be as large as possible whenever its coefficient is positive, and as small as possible whenever its coefficient is negative:

- If $f_1(x) - kf_0(x) > 0$, set $\phi(x) = 1$.
- If $f_1(x) - kf_0(x) < 0$, set $\phi(x) = 0$.
- If $f_1(x) - kf_0(x) = 0$, the value of $\phi(x)$ does not affect the integral (this is where γ comes in).

This decision rule is equivalent to:

$$\phi(x) = \begin{cases} 1 & \text{if } \frac{f_1(x)}{f_0(x)} > k \\ 0 & \text{if } \frac{f_1(x)}{f_0(x)} < k \end{cases} \tag{8.17}$$

This is precisely the form of the Likelihood Ratio Test. The "shadow price" or Lagrange multiplier k represents the critical threshold that balances the gain in power against the cost of increasing the Type I error.

8.3.3 Proof of NP Lemma

Proof. **Proof of (a) Optimality:** Let ϕ_{LRT} be the LRT with size α, and ϕ be any other test with size $\leq \alpha$. Define the function $U(x)$ as the difference in test functions weighted by the linear combination of densities:

$$U(x) = (\phi_{LRT}(x) - \phi(x))(f_1(x) - kf_0(x)) \tag{8.18}$$

We analyze the sign of $U(x)$ by looking at the sign of its two factors in three cases:

- If $f_1(x) - kf_0(x) > 0$ (implies $\Lambda(x) > k$). Since $\phi_{LRT}(x) = 1$ and $\phi(x) \leq 1$, we have:

$$\begin{aligned} \phi_{LRT}(x) - \phi(x) &\geq 0 \\ U(x) = (\phi_{LRT}(x) - \phi(x))(f_1(x) - kf_0(x)) &\geq 0 \end{aligned} \tag{8.19}$$

- If $f_1(x) - kf_0(x) < 0$ (implies $\Lambda(x) < k$). Since $\phi_{LRT}(x) = 0$ and $\phi(x) \geq 0$, we have:

$$\begin{aligned} \phi_{LRT}(x) - \phi(x) &\leq 0 \\ U(x) = (\phi_{LRT}(x) - \phi(x))(f_1(x) - kf_0(x)) &\geq 0 \end{aligned} \tag{8.20}$$

- If $f_1(x) - kf_0(x) = 0$. The product is zero regardless of the test functions.

$$U(x) = 0 \tag{8.21}$$

Combining these cases, we conclude that the product is non-negative for all x:

$$U(x) = (\phi_{LRT}(x) - \phi(x))(f_1(x) - kf_0(x)) \geq 0 \tag{8.22}$$

Therefore, integrating $U(x)$ over the entire domain:

$$\int U(x)dx = \int (\phi_{LRT} - \phi)(f_1 - kf_0)dx \geq 0 \tag{8.23}$$

Expanding the integral:

$$\int \phi_{LRT} f_1 - \int \phi f_1 - k\left(\int \phi_{LRT} f_0 - \int \phi f_0\right) \geq 0 \tag{8.24}$$

Converting to expectations:

$$E_{\theta_1}[\phi_{LRT}] - E_{\theta_1}[\phi] - k(E_{\theta_0}[\phi_{LRT}] - E_{\theta_0}[\phi]) \geq 0 \tag{8.25}$$

Since $E_{\theta_0}[\phi_{LRT}] = \alpha$ and $E_{\theta_0}[\phi] \leq \alpha$, the difference $(E_{\theta_0}[\phi_{LRT}] - E_{\theta_0}[\phi]) \geq 0$. Given that $k \geq 0$:

$$E_{\theta_1}[\phi_{LRT}] - E_{\theta_1}[\phi] \geq 0 \implies \text{Power}(\phi_{LRT}) \geq \text{Power}(\phi) \tag{8.26}$$

Proof of (b) Existence: Let $G(k) = P_{\theta_0}(\Lambda(X) \leq k)$. $G(k)$ is the cumulative distribution function of the random variable $\Lambda(X)$, so it is non-decreasing. We seek k_0 such that $1 - G(k_0) \approx \alpha$. Because of discrete jumps, we might not hit α exactly. We choose k_0 such that:

$$P_{\theta_0}(\Lambda(X) > k_0) \leq \alpha \leq P_{\theta_0}(\Lambda(X) \geq k_0) \tag{8.27}$$

Set $\gamma_0 = \frac{\alpha - P_{\theta_0}(\Lambda(X) > k_0)}{P_{\theta_0}(\Lambda(X) = k_0)}$. □

8.4 Uniformly Most Powerful (UMP) Tests

When the alternative hypothesis is composite ($H_1 : \theta \in \Theta_1$), we seek a test that is "best" for *all* $\theta \in \Theta_1$.

Definition 8.4 (Uniformly Most Powerful Test)**.** A test $\phi_0(x)$ of size α is **Uniformly Most Powerful (UMP)** if:

1. $E_\theta[\phi_0(X)] \leq \alpha$ for all $\theta \in \Theta_0$.
2. For any other test $\phi(x)$ satisfying (1), $E_\theta[\phi_0(X)] \geq E_\theta[\phi(X)]$ for all $\theta \in \Theta_1$.

8.5 Monotone Likelihood Ratio (MLR)

Definition 8.5 (Monotone Likelihood Ratio)**.** A family of densities $\{f(x;\theta)\}$ has a **Monotone Likelihood Ratio (MLR)** with respect to a statistic $T(x)$ if for any $\theta_1 > \theta_0$, the ratio:

$$\frac{f(x;\theta_1)}{f(x;\theta_0)} \tag{8.28}$$

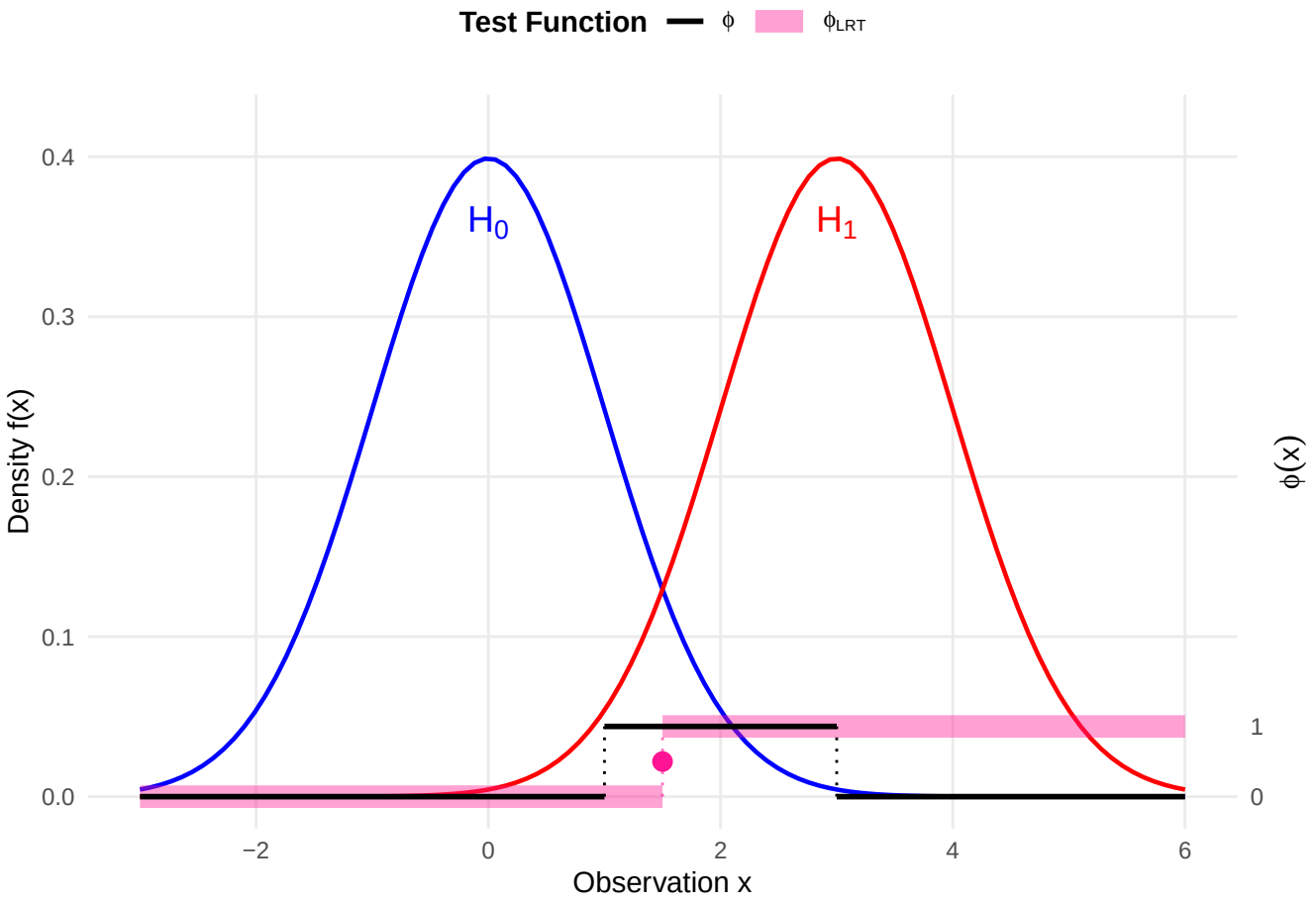


Figure 8.3: Visualizing the NP Lemma: The thick, transparent pink line is ϕ_{LRT}. The thin solid black line is ϕ. Overlap is visible as a black line inside pink.

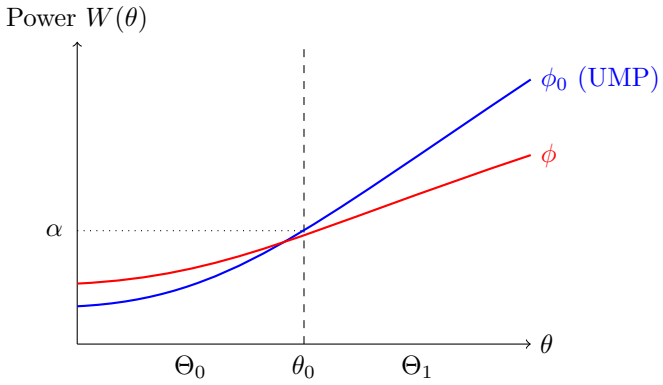


Figure 8.4: Power functions of UMP test vs. another test

is a non-decreasing function of $T(x)$.

Common examples include the one-parameter Exponential Family: $f(x;\theta) = h(x)c(\theta)\exp\{w(\theta)T(x)\}$. If $w(\theta)$ is increasing, the family has MLR w.r.t $T(x)$.

Karlin-Rubin Theorem

Theorem 8.2 (Karlin-Rubin Theorem)**.** *Suppose X has a distribution from a family with MLR with respect to $T(X)$, and the distribution of $T(X)$ is continuous. Consider testing $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$.*
The test:

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > t_0 \\ 0 & \text{if } T(x) \leq t_0 \end{cases} \tag{8.29}$$

where t_0 is determined by $P_{\theta_0}(T(X) > t_0) = \alpha$, is the UMP size α test.

Proof of Theorem 4.2 (UMP for MLR Families). **The Test:** Define the test $\phi(x)$ as:

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > t_0 \\ 0 & \text{if } T(x) \leq t_0 \end{cases} \tag{8.30}$$

where t_0 is determined such that the power at the boundary is α, i.e., $W_{LR}(\theta_0) = \alpha$.
(1) Application of Neyman-Pearson Lemma Because the family has a Monotone Likelihood Ratio (MLR) in $T(x)$, for any specific alternative $\theta_1 > \theta_0$, the likelihood ratio $\Lambda(x)$ is an increasing function of $T(x)$. Therefore, the rejection region $T(x) > t_0$ corresponds to $\Lambda(x) > k$. By the Neyman-Pearson Lemma, $\phi(x)$ is the Most Powerful (MP) test for testing $H_0' : \theta = \theta_0$ vs $H_1' : \theta = \theta_1$.
(2) Monotonicity of the Power Function We claim that $W_{LR}(\theta)$ is a non-decreasing function of θ.
Proof: Let $\theta_2 < \theta_1$. Consider testing $\theta = \theta_2$ vs $\theta = \theta_1$. Let $\beta = W_{LR}(\theta_2)$. Define a constant dummy test $\phi^*(x) = \beta$ for all x. The power of this test is constant: $W_{\phi^*}(\theta) = \beta$. Since $\phi(x)$ corresponds to the likelihood ratio test form (reject for large T) for θ_2 vs θ_1, it is the MP test of size β. By the optimality of the NP Lemma, the power of ϕ at θ_1 must be at least the power of the competitor ϕ^*:

$$W_{LR}(\theta_1) \geq W_{\phi^*}(\theta_1) = \beta = W_{LR}(\theta_2) \tag{8.31}$$

Thus, $W_{LR}(\theta)$ is non-decreasing.
(3) Size Control Since $W_{LR}(\theta)$ is non-decreasing and we set $W_{LR}(\theta_0) = \alpha$:

$$W_{LR}(\theta) \leq W_{LR}(\theta_0) = \alpha \quad \text{for all } \theta \leq \theta_0 \tag{8.32}$$

This proves the test satisfies the size constraint for the composite null $H_0 : \theta \leq \theta_0$.
(4) UMP Property Let $\phi'(x)$ be any other test with size $\leq \alpha$ for $H_0 : \theta \leq \theta_0$. This implies $W_{\phi'}(\theta_0) \leq \alpha$. For any specific $\theta_1 > \theta_0$, we treat the problem as testing θ_0 vs θ_1. Since $\phi(x)$ is the MP test for θ_0 vs θ_1 (from Step 1), and ϕ' is a valid competitor (size $\leq \alpha$ at θ_0), we have:

$$W_{LR}(\theta_1) \geq W_{\phi'}(\theta_1) \tag{8.33}$$

Since this holds for all $\theta_1 > \theta_0$, $\phi(x)$ is the Uniformly Most Powerful (UMP) test. □

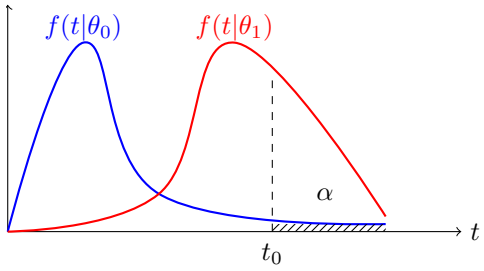


Figure 8.5: Distribution of statistic T under H0 and H1 with MLR

Example 8.3 (UMP Test for Exponential/Gamma)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\theta)$ with pdf $f(x) = \frac{1}{\theta}e^{-x/\theta}$. Test $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$. The sum $T = \sum X_i$ is a sufficient statistic, and $T \sim \text{Gamma}(n, \theta)$.
The Likelihood Ratio for $\theta_1 > \theta_0$ is:

$$\frac{L(\theta_1)}{L(\theta_0)} = \frac{\theta_1^{-n} e^{-\sum x_i/\theta_1}}{\theta_0^{-n} e^{-\sum x_i/\theta_0}} = \left(\frac{\theta_0}{\theta_1}\right)^n \exp\left\{\left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right)\sum x_i\right\} \tag{8.34}$$

Since $\theta_1 > \theta_0$, the term $(\frac{1}{\theta_0} - \frac{1}{\theta_1})$ is positive. Thus, $\Lambda(x)$ is an increasing function of $\sum x_i$. Rejecting for large $\Lambda(x)$ is equivalent to rejecting for $\sum x_i > C$.
This test form does not depend on the specific θ_1, so it is UMP for all $\theta > \theta_0$.

8.5.1 Note on Two-Sided Hypotheses

For testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ (e.g., in a Normal distribution), a UMP test generally **does not exist**. This is because the "best" rejection region for $\theta > \theta_0$ (right tail) is completely different from the "best" region for $\theta < \theta_0$ (left tail).