Theory of Statistical Inference

Longhai Li

2026-04-14

Table of contents

Preface

This is a concise course about statistical inference, which was developed for the course STAT 442/851 at University of Saskatchewan.

Key Features

- **Traditional Mathematical Rigor:** Emphasis is placed on the rigorous mathematical derivation and proof of the fundamental theorems underpinning statistical inference. For example, students will engage deeply with the analytic proofs of the Neyman-Pearson Lemma, the asymptotic normality of Maximum Likelihood Estimators (MLE), and the formulation of the Cramér-Rao Lower Bound.
- **Integration of Computational Tools:** Utilization of computational simulations and graphical representations to elucidate and validate complex statistical concepts. For instance, the course employs simulation studies to empirically demonstrate the advantages of shrinkage estimators and to visualize the asymptotic distributions governed by Maximum Likelihood Estimation (MLE) theory and Wilks' Theorem. Furthermore, discussions will highlight the operational relevance of classical mathematical theorems within modern computational frameworks, such as the implications of Bartlett Identities in deep learning algorithms.
- **Priority of Topics in Light of Modern Practice in Statistics and Machine Learning:** The curriculum is strategically curated to emphasize methodologies that offer broad generalizability and utility in contemporary applications. Prominence is given to versatile frameworks such as Bayesian inference, regularization techniques, Maximum Likelihood Estimation (MLE), likelihood-based hypothesis testing, information criteria, and rigorous model assessment. Concurrently, essential classical foundations, including the Neyman-Pearson Lemma and the theory of Uniformly Minimum Variance Unbiased Estimators (UMVUE), are deliberately retained to ensure a comprehensive theoretical grounding.

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.

About the author

Longhai Li is a professor at the University of Saskatchewan in Canada. He received his Ph.D. degree in statistics from the University of Toronto. His research activities focus on developing and applying statistical machine-learning

methods for bioinformatics and epidemiology applications, with particular interests on statistical learning, cross-validation, hierarchical modelling, survival modelling, model checking, residual diagnostics, model comparison, zero-inflated models, high-throughput data, microbiome data. His research has been funded by NSERC, CANSSI, CFI, CFREF, and MITACS. His research papers have appeared in highly reputed journals, such as Journal of American Statistical Association, Bayesian Analysis, Statistics in Medicine, Statistics and Computing, American Statistician, Journal of Applied Statistics, Scientific Reports, and BMC Bioinformatics.

Notation and Symbols

Throughout this document, we use **boldface lowercase** letters (e.g., $\mathbf{x}$) to denote vectors and **boldface uppercase** letters (e.g., $\mathbf{X}$, $\mathbf{I}$) to denote matrices. Scalar variables and parameters are written in standard italics (e.g., x, θ).

Symbol	Description
Variables & Parameters	
x, X	Scalar random variable or observation
$\mathbf{x}$, $\mathbf{X}$	Vector random variable or observation column vector
θ	Scalar unknown parameter
$\boldsymbol{\theta}$	Vector of unknown parameters, $\boldsymbol{\theta} \in \mathbb{R}^p$
Θ	Parameter space
$\mathbf{0}$	Zero vector or zero matrix (dimension implied by context)
$\mathbf{I}_p$	Identity matrix of size $p \times p$
Functions & Operators	
$f(x\|\theta)$	Probability density function (pdf) or probability mass function (pmf)
$\ell(\boldsymbol{\theta})$	Log-likelihood function, $\ell(\boldsymbol{\theta}) = \log f(\mathbf{x}\|\boldsymbol{\theta})$
$E^{\mathbf{X}\|\theta}[\cdot]$ or $E_\theta[\cdot]$	Expectation w.r.t. $\mathbf{X}$ conditional on θ
$\text{Var}^{\mathbf{X}\|\theta}(\cdot)$ or $\text{Var}_\theta(\cdot)$	Variance (or Covariance Matrix) w.r.t. $\mathbf{X}$ conditional on θ
$\text{Cov}_\theta(\cdot,\cdot)$	Covariance between two random variables or vectors
$\mathbf{A}^\top$ or $\mathbf{A}^T$	Transpose of vector or matrix $\mathbf{A}$
$\text{tr}(\mathbf{A})$	Trace of matrix $\mathbf{A}$ (sum of diagonal elements)
$\mathbf{A} \succeq \mathbf{B}$	Matrix inequality; $\mathbf{A} - \mathbf{B}$ is positive semi-definite
Calculus & Gradients	
$\nabla_\theta f$	Gradient vector with respect to θ
$\nabla^2_\theta f$	Hessian matrix (second derivatives) with respect to θ
$\frac{\partial}{\partial \theta}$	Partial derivative operator
$\mathbf{J}_f(\mathbf{x})$	Jacobian matrix of a vector-valued function f
$\nabla \cdot \mathbf{g}$	Divergence of a vector field $\mathbf{g}$, $\sum \frac{\partial g_i}{\partial x_i}$
Statistical Quantities	
$\mathbf{U}(\boldsymbol{\theta})$	Score vector, $\nabla_\theta \ell(\boldsymbol{\theta})$

Symbol	Description
$\mathbf{J}(\theta)$	Observed Information matrix, $-\nabla^2_\theta \ell(\theta)$
$\mathcal{J}(\theta)$	Fisher Information matrix, $E[\mathbf{U}\mathbf{U}^\top] = E[\mathbf{J}]$
$T(\mathbf{X})$	Estimator or statistic
$m(\theta)$	Expectation of an estimator, $E[T(\mathbf{X})]$
$\mathbf{D}(\theta)$	Jacobian of the expectation vector $m(\theta)$

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.

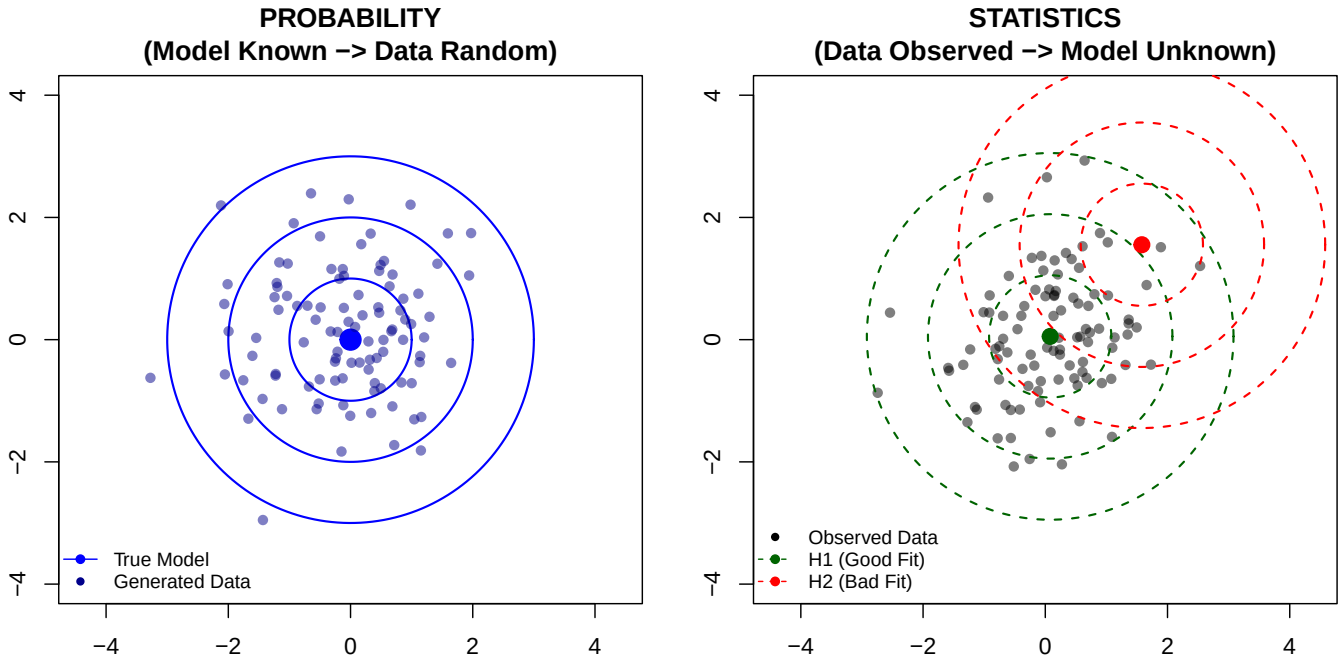


Figure 1.1: Probability vs Statistics. Left: Probability—The model is fixed (Blue center/contours), generating random data. Right: Statistics—Data is fixed (Black points); we test two hypothesized models: H1 (Green) centered at the sample mean (Good Fit) and H2 (Red) shifted by (1.5, 1.5) (Bad Fit).

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

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

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.

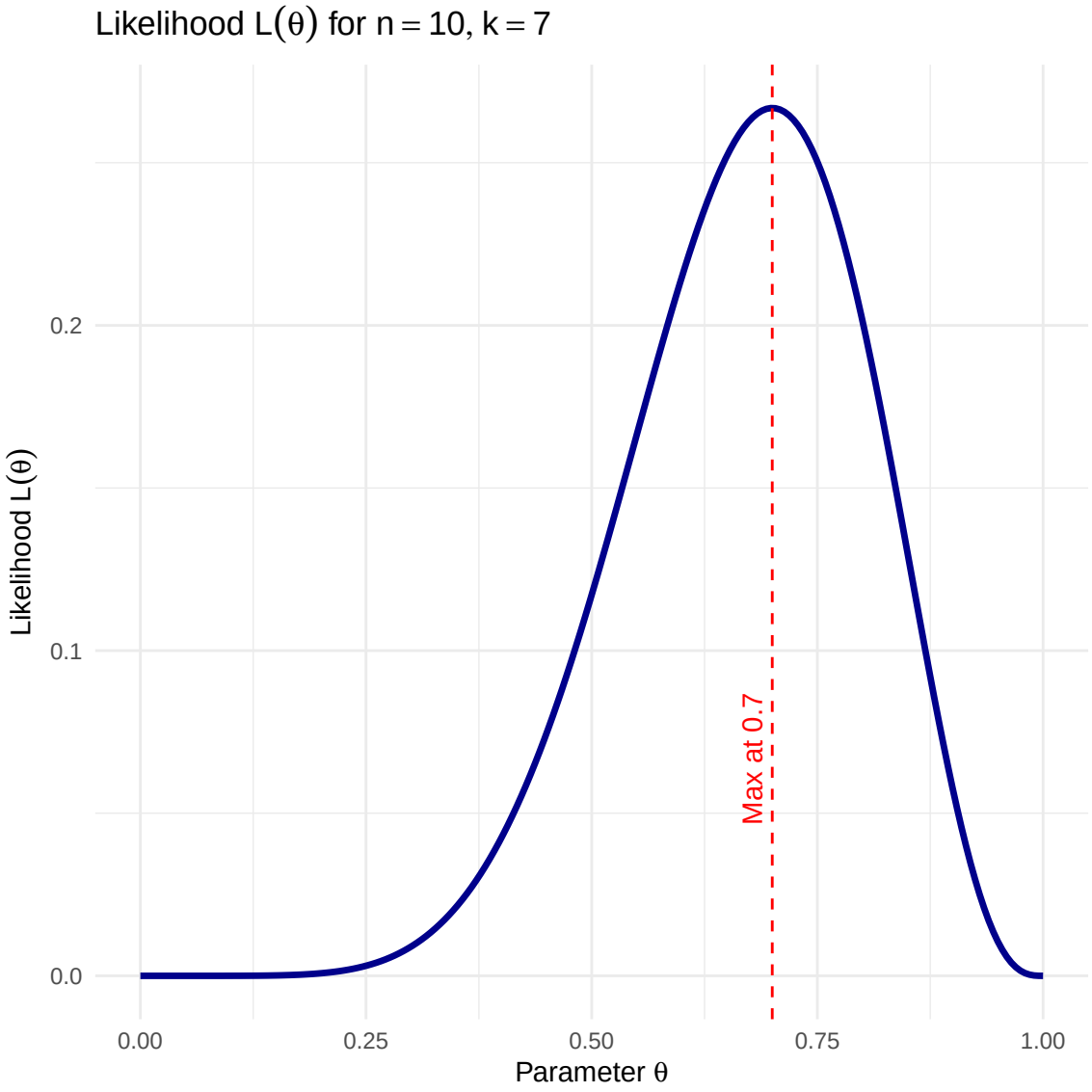


Figure 1.2: Likelihood Function (n= 10)

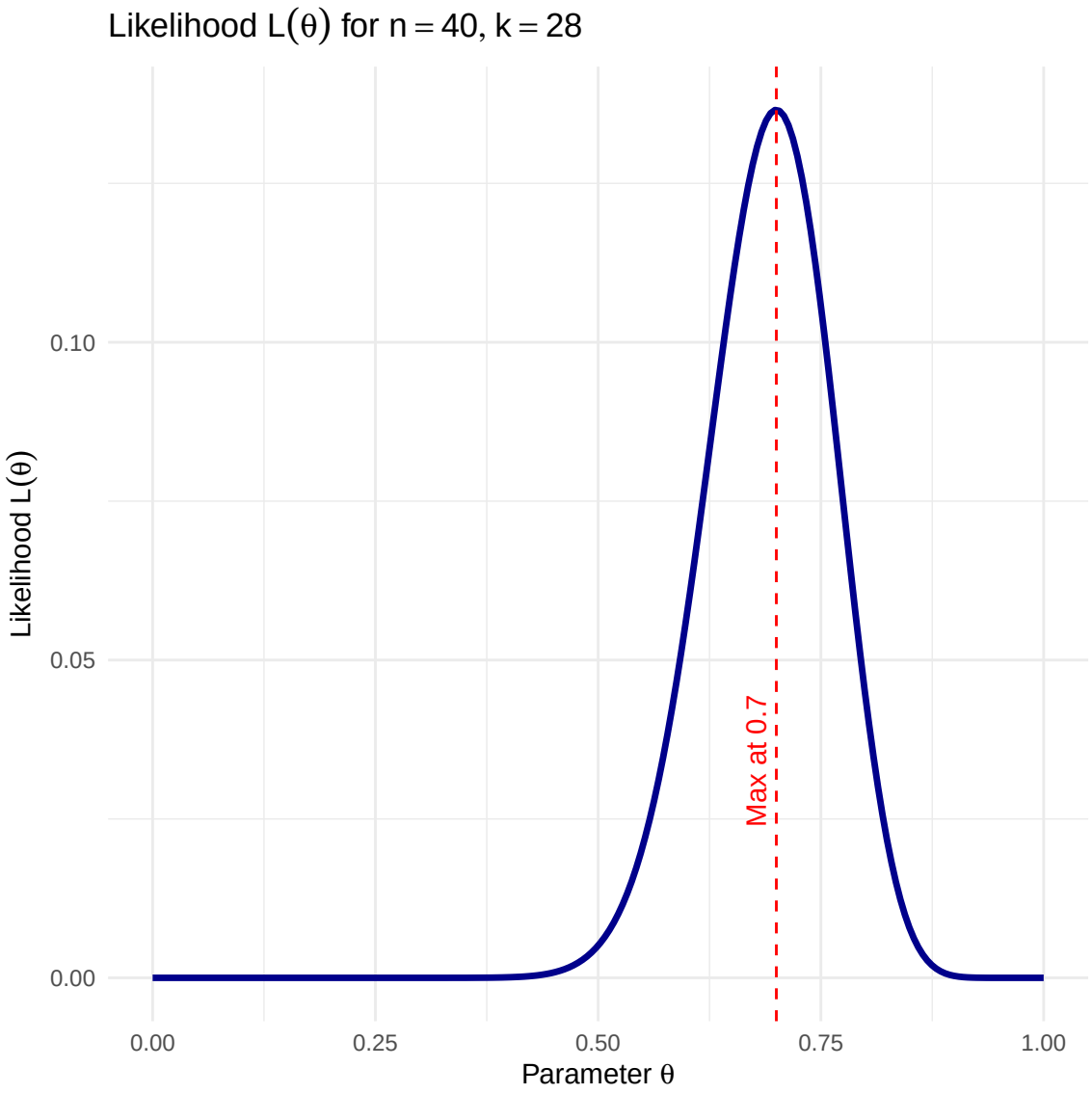


Figure 1.3: Likelihood Function (n= 40)

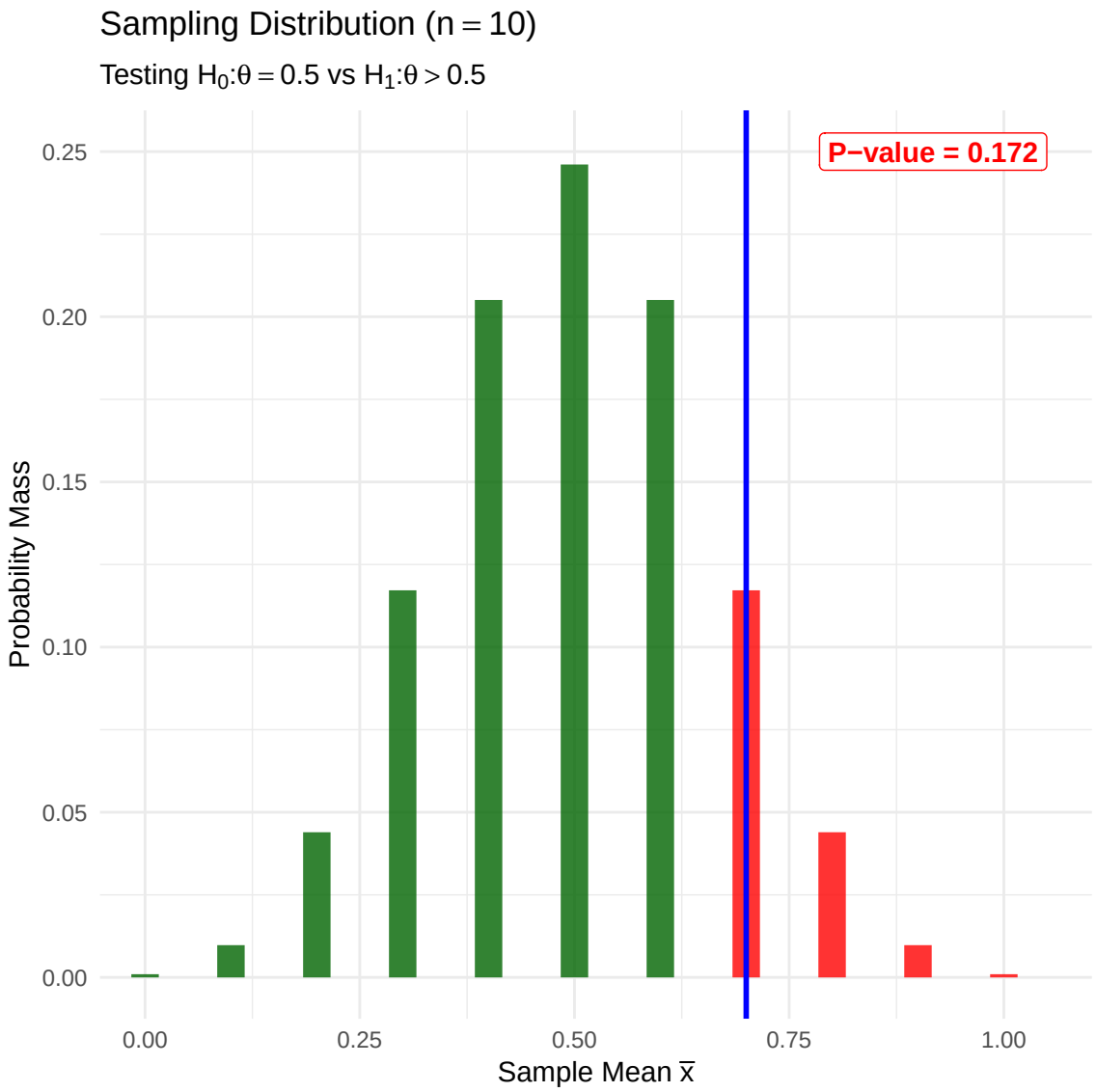


Figure 1.4: Sampling Distribution (n= 10)

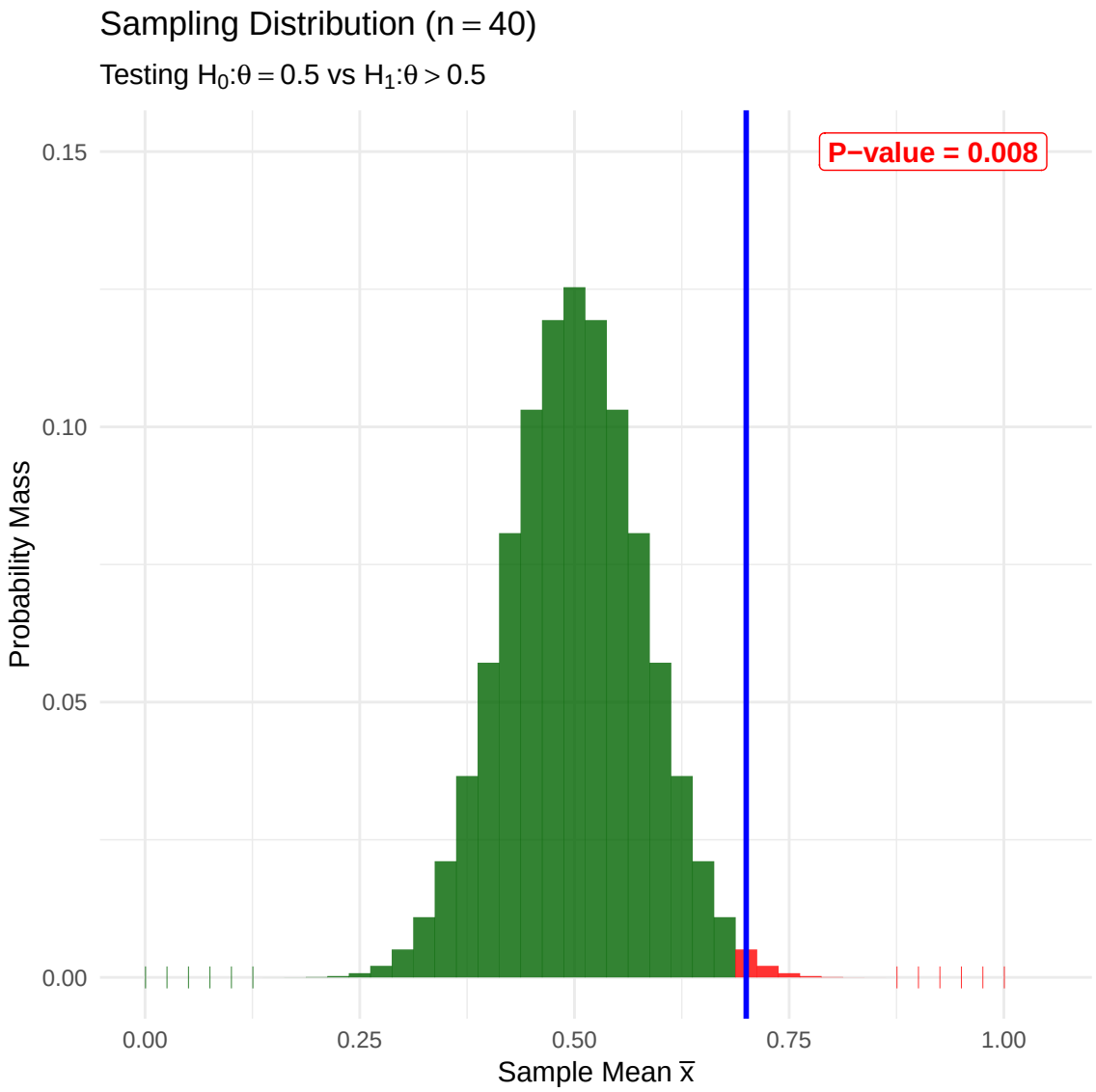


Figure 1.5: Sampling Distribution (n= 40)

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?

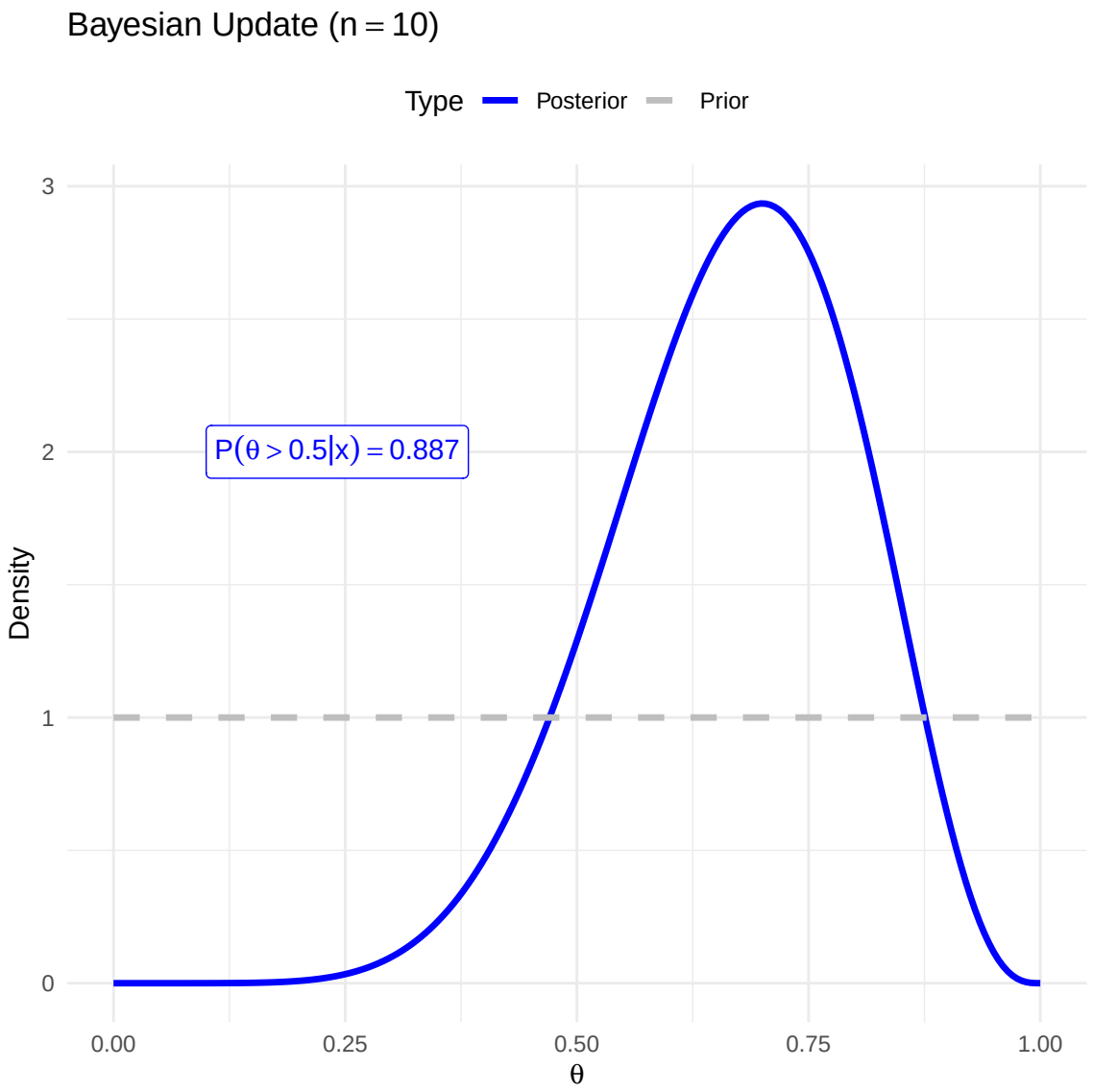


Figure 1.6: Bayesian Update (n= 10)

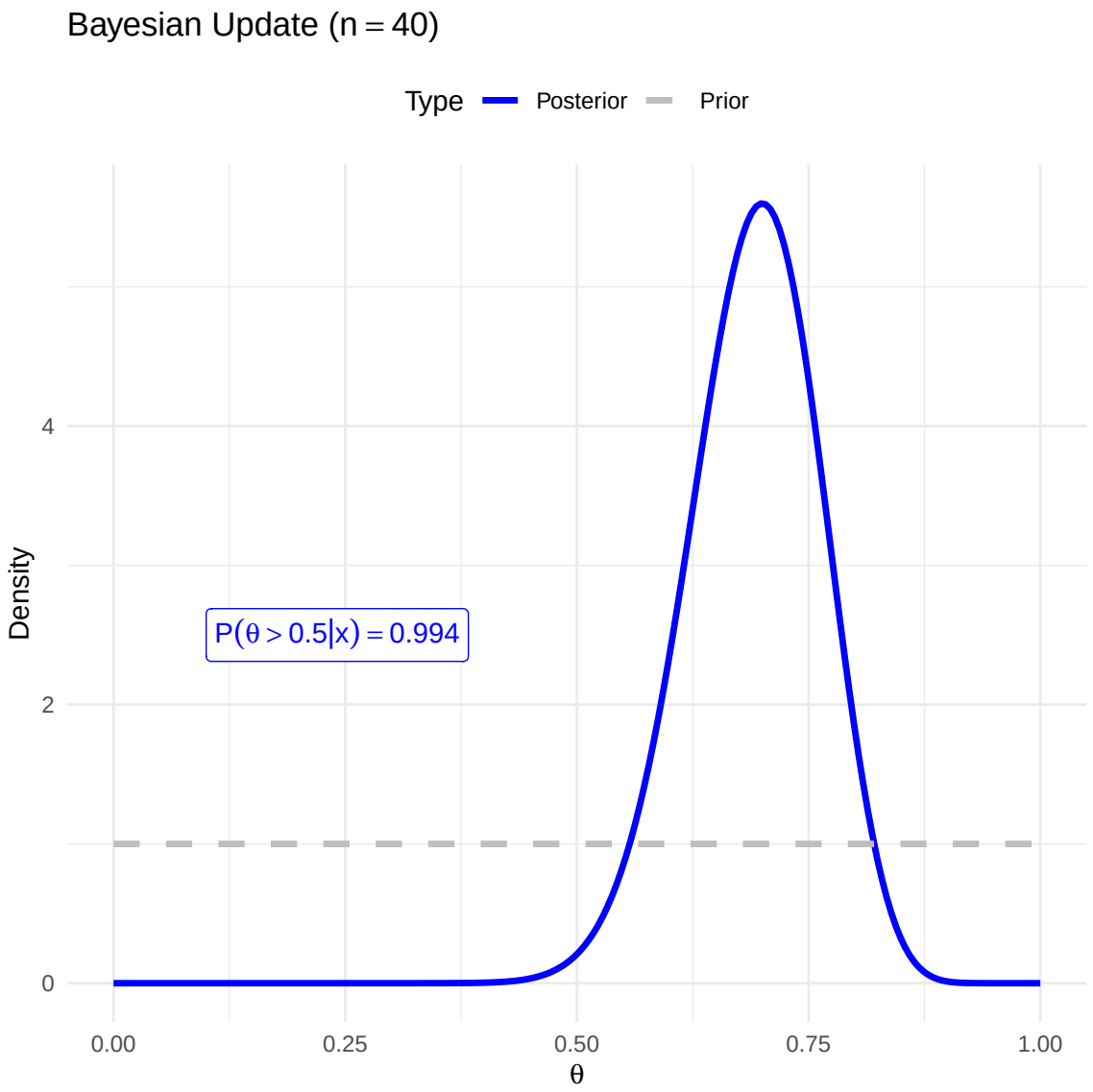


Figure 1.7: Bayesian Update (n= 40)

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

2 Sufficient Statistic

2.1 Sufficient Statistics

Definition 2.1 (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:

   $$\ell(\theta; \mathbf{x}) - \ell(\theta; \mathbf{y}) = c(\mathbf{x}, \mathbf{y}) \quad \text{for all } \theta \tag{2.1}$$

   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:

   $$L(\theta; \mathbf{x}) = h(\mathbf{x})g(T(\mathbf{x}); \theta) \tag{2.2}$$

   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.

   $$f(\mathbf{x}|T(\mathbf{x}) = t, \theta) = f(\mathbf{x}|t) \tag{2.3}$$

Theorem 2.1 (Factorization Theorem)**.** *The three conditions in the definitions of Definition 2.1 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 $T(\mathbf{x}) = T(\mathbf{y})$.

   $$\ell(\theta; \mathbf{x}) - \ell(\theta; \mathbf{y}) = [\ln h(\mathbf{x}) + \ln g(T(\mathbf{x}); \theta)] - [\ln h(\mathbf{y}) + \ln g(T(\mathbf{y}); \theta)] \tag{2.4}$$

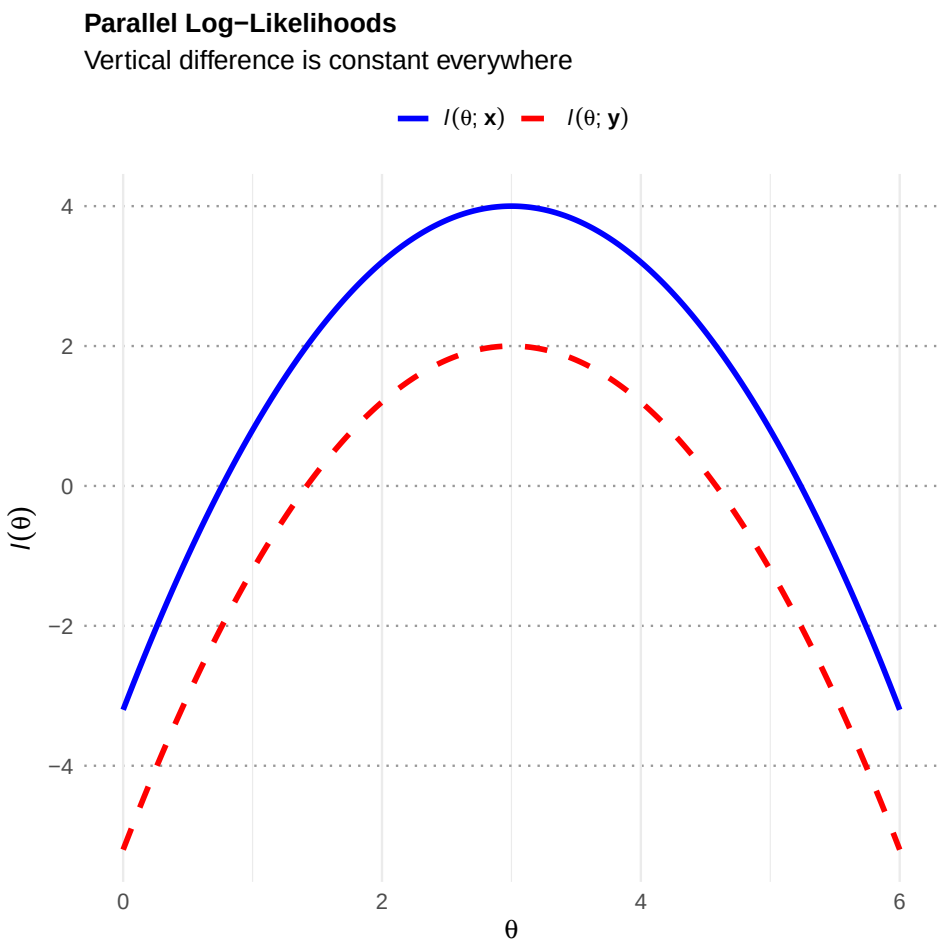


Figure 2.1: Visualizing Parallel Log-Likelihoods

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.

$$\ell(\theta; \mathbf{x}) - \ell(\theta; \mathbf{y}) = \ln h(\mathbf{x}) - \ln h(\mathbf{y}) \tag{2.5}$$

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:

   $$L(\theta; \mathbf{x}) = k(\mathbf{x}, \mathbf{x}_t) L(\theta; \mathbf{x}_t) \tag{2.6}$$

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

   $$L(\theta; \mathbf{x}) = h(\mathbf{x}) g(T(\mathbf{x}); \theta) \tag{2.7}$$

   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 $T(\mathbf{x}) = t$:

   $$P(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t) = \frac{P(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t)}{P(T(\mathbf{X}) = t)} = \frac{f(\mathbf{x}; \theta)}{\sum_{\{\mathbf{y}: T(\mathbf{y}) = t\}} f(\mathbf{y}; \theta)} \tag{2.8}$$

   Substitute the factorization:

   $$= \frac{h(\mathbf{x}) g(t; \theta)}{\sum_{\{\mathbf{y}: T(\mathbf{y}) = t\}} h(\mathbf{y}) g(t; \theta)} = \frac{h(\mathbf{x}) g(t; \theta)}{g(t; \theta) \sum_{\{\mathbf{y}: T(\mathbf{y}) = t\}} h(\mathbf{y})} \tag{2.9}$$

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

   $$= \frac{h(\mathbf{x})}{\sum_{\{\mathbf{y}: T(\mathbf{y}) = t\}} h(\mathbf{y})} \tag{2.10}$$

   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:

   $$f(\mathbf{x}; \theta) = f(\mathbf{x} | T(\mathbf{x}) = t; \theta) \cdot P(T(\mathbf{X}) = t; \theta) \tag{2.11}$$

   Substitute the assumption:

   $$f(\mathbf{x}; \theta) = k(\mathbf{x}) \cdot P(T(\mathbf{X}) = T(\mathbf{x}); \theta) \tag{2.12}$$

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

   $$f(\mathbf{x}; \theta) = h(\mathbf{x}) g(T(\mathbf{x}); \theta) \tag{2.13}$$

   This recovers the Factorization form.

□

Example 2.1 (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:

$$f(x_i|\theta)=\frac{1}{(\theta+1)-(\theta-1)}I(\theta-1<x_i<\theta+1)=\frac{1}{2}I(\theta-1<x_i<\theta+1) \tag{2.14}$$

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

$$L(\theta;\mathbf{x})=\prod_{i=1}^{n}\frac{1}{2}I(\theta-1<x_i<\theta+1) \tag{2.15}$$

$$L(\theta;\mathbf{x})=2^{-n}\cdot I(\min(x_i)>\theta-1)\cdot I(\max(x_i)<\theta+1) \tag{2.16}$$

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

$$L(\theta;\mathbf{x})=2^{-n}\cdot I(\theta<X_{(1)}+1)\cdot I(\theta>X_{(n)}-1) \tag{2.17}$$

$$L(\theta;\mathbf{x})=2^{-n}\cdot I(X_{(n)}-1<\theta<X_{(1)}+1) \tag{2.18}$$

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:

$$T(\mathbf{X})=(X_{(1)},X_{(n)}) \tag{2.19}$$

Example 2.2 (Gamma Distribution)**.** Let $\mathbf{X}=(X_1,\ldots,X_n)$ be i.i.d. $\Gamma(\alpha,\beta)$. The pdf is:

$$f(x_i|\alpha,\beta)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x_i^{\alpha-1}e^{-x_i/\beta},\quad x_i>0 \tag{2.20}$$

The joint likelihood is:

$$L(\alpha,\beta;\mathbf{x})=\left(\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\right)^{n}\left(\prod_{i=1}^{n}x_i\right)^{\alpha-1}\exp\left(-\frac{1}{\beta}\sum_{i=1}^{n}x_i\right) \tag{2.21}$$

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

$$g(T(\mathbf{x}),)=\left(\prod_{i=1}^{n}x_i\right)^{\alpha}\exp\left(-\frac{1}{\beta}\sum_{i=1}^{n}x_i\right) \tag{2.22}$$

Thus, the sufficient statistics are:

$$T(\mathbf{X})=\left(\prod_{i=1}^{n}X_i,\sum_{i=1}^{n}X_i\right) \tag{2.23}$$

Example 2.3 (Sufficient Statistic of Exponential Family)**.** Many common distributions (Normal, Poisson, Gamma, Binomial) belong to the **Exponential Family**, which has a density in the form:

$$f(\mathbf{x}|\theta) = h(\mathbf{x})c(\theta)\exp\left(\sum_{j=1}^{k}\pi_j(\theta)t_j(\mathbf{x})\right) \tag{2.24}$$

Then, by the Factorization Theorem, the statistic:

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n}t_1(x_i), \dots, \sum_{i=1}^{n}t_k(x_i)\right) \tag{2.25}$$

is a sufficient statistic for θ.

Example 2.4 (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:

$$f(\mathbf{x}|\theta) = h(\mathbf{x})c(\theta)\exp\left(\sum_{j=1}^{k}\pi_j(\theta)T_j(\mathbf{x})\right) \tag{2.26}$$

1. **Write the Joint PDF**

$$f(\mathbf{x}|p) = \prod_{i=1}^{n}p^{x_i}(1-p)^{1-x_i} \tag{2.27}$$

2. **Convert to Exponential Form**

$$\begin{aligned} f(\mathbf{x}|p) &= \exp\left(\sum_{i=1}^{n}\left[x_i\ln p + (1-x_i)\ln(1-p)\right]\right) \\ &= \exp\left(\sum_{i=1}^{n}\left[x_i\ln p + \ln(1-p) - x_i\ln(1-p)\right]\right) \\ &= \exp\left(\sum_{i=1}^{n}\ln(1-p) + \sum_{i=1}^{n}x_i\left[\ln p - \ln(1-p)\right]\right) \end{aligned} \tag{2.28}$$

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

$$f(\mathbf{x}|p) = \underbrace{1}_{h(\mathbf{x})}\cdot\underbrace{(1-p)^n}_{c(p)}\cdot\exp\left(\underbrace{\ln\left(\frac{p}{1-p}\right)}_{\pi_1(p)}\underbrace{\sum_{i=1}^{n}x_i}_{T_1(\mathbf{x})}\right) \tag{2.29}$$

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

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i \tag{2.30}$$

Remark 2.1 (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:

$$L(\theta; \mathbf{x}) \propto g(T(\mathbf{x})|\theta) \tag{2.31}$$

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:

$$\log L(\theta; \mathbf{x}) = \text{const} + \sum_{i=1}^{k} \eta_i(\theta) T_i(\mathbf{x}) - nA(\theta) \tag{2.32}$$

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

2.2 Minimal Sufficient Statistics

Definition 2.2 (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 $S(X)$.

$$T(X) = g(S(X)) \tag{2.33}$$

(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 2.2 (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:*

$$\ell(\theta; x) = \ell(\theta; y) + c(x, y) \text{ for all } \theta \implies T(x) = T(y) \tag{2.34}$$

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:
$$L(\theta; x) = h(x)g(U(x), \theta) \tag{2.35}$$
$$L(\theta; y) = h(y)g(U(y), \theta) \tag{2.36}$$
3. Since $U(x) = U(y)$, the factor $g(U(x), \theta)$ is identical to $g(U(y), \theta)$. Taking the log-ratio:
$$\ell(\theta; x) - \ell(\theta; y) = \ln h(x) - \ln h(y) \tag{2.37}$$
The term $\ln h(x) - \ln h(y)$ depends only on x and y, not on θ. Let this be $c(x, y)$.
$$\ell(\theta; x) = \ell(\theta; y) + c(x, y) \tag{2.38}$$
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:
$$S(x) = \{z \mid \ell(\theta; z) = \ell(\theta; x) + c_z \text{ for all } \theta\} \tag{2.39}$$
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:
$$\ell(\theta; x) - \ell(\theta; y) = c(x, y) \tag{2.40}$$
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 2.5 (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**:

$$\text{Parallel Log-Likelihoods} \implies T(x) = T(y) \tag{2.41}$$

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

$$\ell(p; x) = (\sum x_i) \ln p + (n - \sum x_i) \ln(1 - p) \tag{2.42}$$

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:

$$\text{Parallel Log-Likelihoods} \iff \sum x_i = \sum y_i \tag{2.43}$$

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:

$$\sum x_i = \sum y_i \overset{?}{\implies} (x_1, \sum_{i=2}^{3} x_i) = (y_1, \sum_{i=2}^{3} y_i) \tag{2.44}$$

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:**

$$S(x) = (1, 1) \tag{2.45}$$

$$S(y) = (0, 2) \tag{2.46}$$

$$S(x) \neq S(y) \tag{2.47}$$

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.

3 Likelihood Theory

3.1 Definitions and Notations

3.1.1 Regular Families

Definition 3.1 (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:

$$\nabla_\theta \int \exp\{\ell(\theta; \mathbf{x})\} d\mathbf{x} = \int \nabla_\theta \exp\{\ell(\theta; \mathbf{x})\} d\mathbf{x} \tag{3.1}$$

3.1.2 Score Vector 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 3.2.

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

$$\mathbf{U}(\theta; \mathbf{X}) = \nabla \ell(\theta; \mathbf{X}) = \frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta} = \begin{bmatrix} \frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta_1} \\ \frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta_p} \end{bmatrix} \tag{3.2}$$

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.

$$\mathbf{J}(\theta; \mathbf{X}) = -\nabla^2 \ell(\theta; \mathbf{X}) = -\frac{\partial^2 \ell(\theta; \mathbf{X})}{\partial \theta \partial \theta^T} = -\begin{bmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_p^2} \end{bmatrix} \tag{3.3}$$

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

$$\mathcal{J}(\theta) = E_\theta \left[\mathbf{J}(\theta; \mathbf{X}) \right] \tag{3.4}$$

3.2 Examples

3.2.1 Exponential Likelihood

Example 3.1. Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Exp}(\theta)$, where the density is $f(x|\theta) = \frac{1}{\theta} e^{-x/\theta}$. We use this setting to explore the theoretical connections provided by **Bartlett's Identities** and the **Cramér-Rao Lower Bound (CRLB)**.

1. **The Score Function (U)**

   The Score function is the gradient of the log-likelihood. Bartlett's first identity suggests that the expected score at the true parameter value is zero, providing the intuition for finding the maximum likelihood estimator (MLE) by setting $U(\theta) = 0$.

   The log-likelihood function is:

$$\ell(\theta; \mathbf{x}) = \sum_{i=1}^{n} \left(-\log \theta - \frac{x_i}{\theta} \right) = -n \log \theta - \frac{1}{\theta} \sum_{i=1}^{n} x_i \tag{3.5}$$

   The Score function is:

$$U(\theta; \mathbf{x}) = \frac{\partial \ell}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum x_i}{\theta^2} \tag{3.6}$$

   Bartlett's First Identity Check:

$$E[U] = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum E[X_i] = -\frac{n}{\theta} + \frac{n\theta}{\theta^2} = 0 \tag{3.7}$$

   (Verified)

2. **Fisher Information ($\mathcal{J}(\theta)$)**

   Bartlett's second identity connects the variance of the Score to the curvature (expected Hessian) of the log-likelihood. This curvature is the Fisher Information.

   - **Method A: Negative Expected Hessian**

$$U'(\theta) = \frac{\partial U}{\partial \theta} = \frac{n}{\theta^2} - \frac{2 \sum x_i}{\theta^3} \tag{3.8}$$

$$\mathcal{J}(\theta) = -E[U'(\theta)] = -\left(\frac{n}{\theta^2} - \frac{2n\theta}{\theta^3} \right) = \frac{n}{\theta^2} \tag{3.9}$$

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

$$\text{Var}(U) = \text{Var}\left(-\frac{n}{\theta} + \frac{\sum X_i}{\theta^2} \right) = \frac{1}{\theta^4} \text{Var}\left(\sum X_i \right) \tag{3.10}$$

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

$$\text{Var}(U) = \frac{1}{\theta^4}(n\theta^2) = \frac{n}{\theta^2} \tag{3.11}$$

Result: $\text{Var}(U) = \mathcal{I}(\theta)$. (Identity Verified)

3. **Cramer-Rao Lower Bound (CRLB)**

   The CRLB states that the variance of any unbiased estimator is bounded below by the inverse of the Fisher Information. We test the efficiency of the sample mean $T(\mathbf{X}) = \bar{X}$.

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

   $$\text{Var}(T(\mathbf{X})) = \text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{\theta^2}{n} \tag{3.12}$$

   - **Theoretical Lower Bound:**

   $$\text{CRLB} = \frac{[m'(\theta)]^2}{\mathcal{I}(\theta)} = \frac{1^2}{n/\theta^2} = \frac{\theta^2}{n} \tag{3.13}$$

   Conclusion: Since $\text{Var}(T(\mathbf{X})) = \text{CRLB}$, the estimator $\bar{X}$ exhausts the information in the sample and is an **efficient estimator** for θ.

3.2.2 Cauchy Likelihood (R Illustration)

To illustrate the concepts visually, we use the **Cauchy distribution**, known for its heavy tails. We assume a known scale of 1 and estimate the location parameter θ.

The probability density function is $f(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}$.

The **Score Function** $U(\theta)$ is the derivative of the log-likelihood with respect to θ:

$$U(\theta; \mathbf{x}) = \sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2} \tag{3.14}$$

We visualize the Log-Likelihoods (dashed lines) and Score functions (solid lines) for two different sample sizes ($n = 10$ and $n = 20$) using a true parameter $\theta^* = 0$.

- **Global Scaling:** The y-axis ranges are fixed across both plots. Note how the "hills" of the likelihood become sharper and the "slopes" of the score become steeper as n increases.
- **MLE ($\hat{\theta}$):** The triangles on the x-axis mark the Maximum Likelihood Estimate. Unlike the exponential case, the Cauchy MLE has no closed form and must be found numerically (where the Score crosses zero).
- **Score at True Parameter ($U(\theta^*)$):** The solid circles on the vertical dashed line mark the value of the Score function at the true parameter.

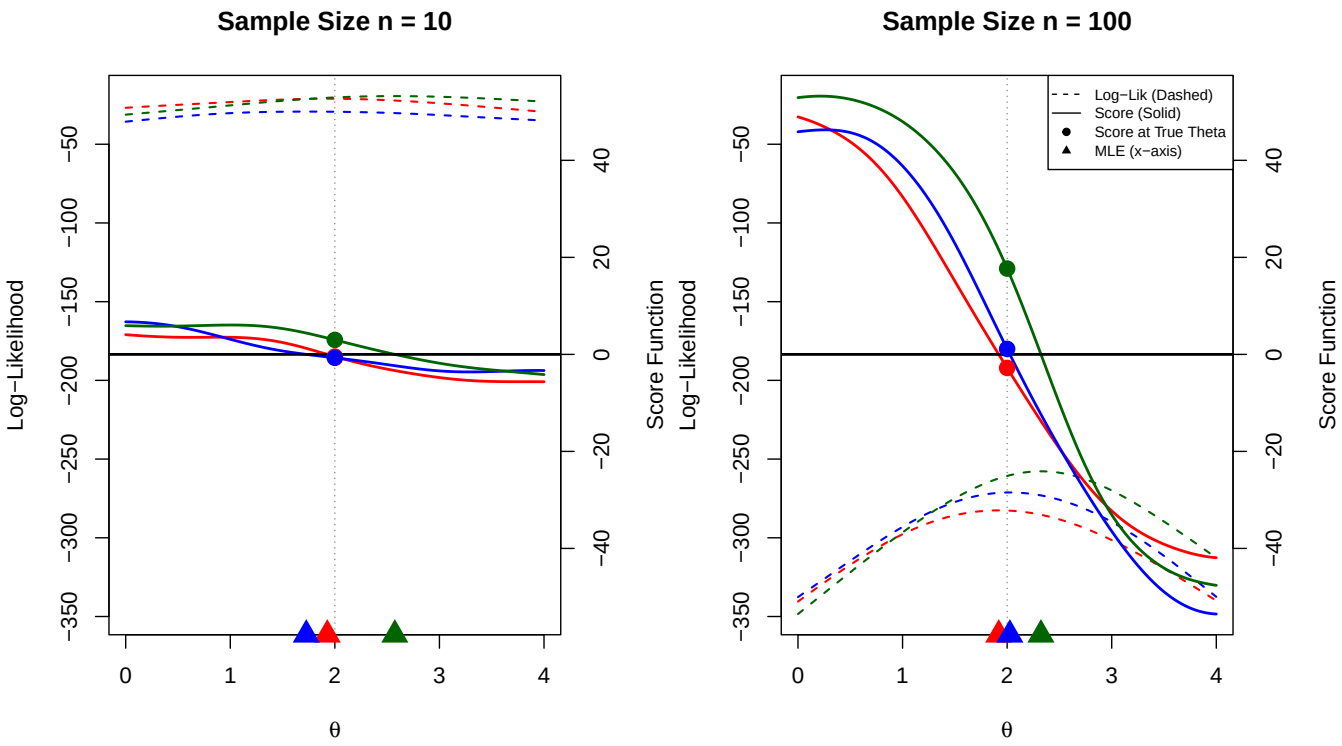


Figure 3.1: Visualization of the Log-Likelihood, Score Function, MLE of Cauchy Models.

3.3 Bartlett's Identities: Mean and Covariance of Score Vector

Theorem 3.1 (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.*

$$E_\theta[\mathbf{U}(\theta; \mathbf{X})] = \mathbf{0} \tag{3.15}$$

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

$$Cov_\theta\left(\mathbf{U}(\theta; \mathbf{X})\right) = E_\theta[\mathbf{J}(\theta; \mathbf{X})] = \mathcal{I}(\theta) \tag{3.16}$$

Remark 3.1. 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 $\mathbf{X}$:

$$\int f(\mathbf{x}|\theta)\, d\mathbf{x} = 1 \tag{3.17}$$

Differentiating both sides with respect to the parameter vector θ:

$$\nabla_\theta \int f(\mathbf{x}|\theta)\, d\mathbf{x} = \mathbf{0} \tag{3.18}$$

Assuming regularity allows us to interchange differentiation and integration:

$$\int \nabla_\theta f(\mathbf{x}|\theta)\, d\mathbf{x} = \mathbf{0} \tag{3.19}$$

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

$$\int \mathbf{U}(\theta;\mathbf{x}) f(\mathbf{x}|\theta)\, d\mathbf{x} = \mathbf{0} \tag{3.20}$$

This is precisely the definition of the expectation:

$$E_\theta[\mathbf{U}(\theta;\mathbf{X})] = \mathbf{0} \tag{3.21}$$

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.

$$\nabla_{\theta^T} \int \mathbf{U}(\theta;\mathbf{x}) f(\mathbf{x}|\theta)\, d\mathbf{x} = \mathbf{0} \tag{3.22}$$

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

$$\int \left[\left(\nabla_{\theta^T}\mathbf{U}(\theta;\mathbf{x})\right) f(\mathbf{x}|\theta) + \mathbf{U}(\theta;\mathbf{x}) \left(\nabla_{\theta^T} f(\mathbf{x}|\theta)\right) \right] d\mathbf{x} = \mathbf{0} \tag{3.23}$$

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:

$$\int \left[-\mathbf{J}(\theta;\mathbf{x}) f(\mathbf{x}|\theta) + \mathbf{U}(\theta;\mathbf{x})\mathbf{U}(\theta;\mathbf{x})^T f(\mathbf{x}|\theta) \right] d\mathbf{x} = \mathbf{0} \tag{3.24}$$

This simplifies to expectations:

$$-E_\theta[\mathbf{J}(\theta;\mathbf{X})] + E_\theta[\mathbf{U}(\theta;\mathbf{X})\mathbf{U}(\theta;\mathbf{X})^T] = \mathbf{0} \tag{3.25}$$

Rearranging gives:

$$E_\theta[\mathbf{J}(\theta;\mathbf{X})] = E_\theta[\mathbf{U}(\theta;\mathbf{X})\mathbf{U}(\theta;\mathbf{X})^T] \tag{3.26}$$

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:

$$\text{Cov}_\theta(\mathbf{U}(\theta;\mathbf{X})) = E_\theta[\mathbf{U}(\theta;\mathbf{X})\mathbf{U}(\theta;\mathbf{X})^T] - E_\theta[\mathbf{U}(\theta;\mathbf{X})]E_\theta[\mathbf{U}(\theta;\mathbf{X})]^T = E_\theta[\mathbf{U}(\theta;\mathbf{X})\mathbf{U}(\theta;\mathbf{X})^T] \tag{3.27}$$

Therefore, we conclude:

$$\text{Cov}_\theta(\mathbf{U}(\theta;\mathbf{X})) = E_\theta[\mathbf{J}(\theta;\mathbf{X})] = \mathcal{J}(\theta) \tag{3.28}$$

□

3.4 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.

3.4.1 The Score Covariance Identity

The following theorem establishes a fundamental relationship between the sensitivity of an estimator's expectation to the parameter θ and its covariance with the Score function. This identity is the engine behind the Cramer-Rao Lower Bound.

Theorem 3.2 (Covariance of Estimator and Score)**.** *Let $T(\mathbf{X})$ be any estimator with finite variance, and let $\mathbf{U}(\theta) = \nabla_\theta \log f(\mathbf{X}|\theta)$ be the Score function. Under standard regularity conditions, the covariance between the estimator and the score is equal to the derivative of the estimator's expectation:*

$$Cov_\theta(T(\mathbf{X}), \mathbf{U}(\theta)) = \nabla_\theta E_\theta[T(\mathbf{X})] \tag{3.29}$$

Proof. We evaluate the covariance term explicitly. By definition:

$$\text{Cov}_\theta(T, \mathbf{U}) = E_\theta[T(\mathbf{X})\mathbf{U}] - E_\theta[T]E_\theta[\mathbf{U}] \tag{3.30}$$

Recall that the expected score is zero ($E_\theta[\mathbf{U}] = \mathbf{0}$). Thus, the second term vanishes:

$$\begin{aligned} \text{Cov}_\theta(T, \mathbf{U}) &= E_\theta\left[T(\mathbf{X})\nabla_\theta \log f(\mathbf{X}|\theta)\right] - m(\theta) \cdot \mathbf{0} \\ &= \int T(\mathbf{x}) \left(\nabla_\theta \log f(\mathbf{x}|\theta)\right) f(\mathbf{x}|\theta)\, d\mathbf{x} \end{aligned} \tag{3.31}$$

Using the logarithmic derivative identity $(\nabla_\theta \log f)f = \frac{1}{f}(\nabla_\theta f)f = \nabla_\theta f$:

$$\text{Cov}_\theta(T, \mathbf{U}) = \int T(\mathbf{x}) \nabla_\theta f(\mathbf{x}|\theta)\, d\mathbf{x} \tag{3.32}$$

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

$$\text{Cov}_\theta(T, \mathbf{U}) = \nabla_\theta \int T(\mathbf{x}) f(\mathbf{x}|\theta)\, d\mathbf{x} = \nabla_\theta E_\theta[T(\mathbf{X})] \tag{3.33}$$

which yields the result:

$$\boxed{\text{Cov}_\theta(T, \mathbf{U}) = \nabla m(\theta)} \tag{3.34}$$

□

Theorem 3.3 (Cramer-Rao Lower Bound for Scalar Estimator)**.** *Let $\mathbf{X}$ be a random variable with probability density function (or probability mass function) $f(\mathbf{x}|\theta)$, where $\theta \in \Theta$ is a vector unknown parameter. Let $T(\mathbf{X})$ be any estimator with finite variance, and let $m(\theta) = E_\theta[T(\mathbf{X})]$ denote its expectation.*
*Assume the following **regularity conditions** hold:*

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

Then, the variance of $T(\mathbf{X})$ *satisfies:*

$$Var_\theta(T(\mathbf{X})) \geq \frac{[\nabla m(\theta)]^2}{\mathcal{I}(\theta)} \tag{3.35}$$

where $\mathcal{I}(\theta) = E_\theta\left[(\nabla_\theta \log f(\mathbf{X}|\theta))^2\right]$ *is the Fisher Information.*
Particular Case: *If* $T(\mathbf{X})$ *is an* ***unbiased*** *estimator of* θ *(i.e.,* $m(\theta) = \theta$ *and* $\nabla m(\theta) = \mathbf{I}$*), then:*

$$Var_\theta(T(\mathbf{X})) \geq \frac{1}{\mathcal{I}(\theta)} \tag{3.36}$$

Proof. Let $\mathbf{U} = \nabla_\theta \log f(\mathbf{X}|\theta)$ be the 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:

$$E_\theta[\mathbf{U}] = \mathbf{0} \quad \text{and} \quad \text{Var}_\theta(\mathbf{U}) = \mathcal{I}(\theta) \tag{3.37}$$

Consider the covariance between the estimator $T(\mathbf{X})$ and the Score $\mathbf{U}$. By the Cauchy-Schwarz inequality:

$$[\text{Cov}_\theta(T, \mathbf{U})]^2 \leq \text{Var}_\theta(T)\text{Var}_\theta(\mathbf{U}) \tag{3.38}$$

Using the **Score Covariance Identity** (Theorem 3.2), we substitute the explicit form of the covariance:

$$\text{Cov}_\theta(T, \mathbf{U}) = \nabla m(\theta) \tag{3.39}$$

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

$$[\nabla m(\theta)]^2 \leq \text{Var}_\theta(T) \cdot \mathcal{I}(\theta) \tag{3.40}$$

Rearranging the terms yields the desired lower bound:

$$\text{Var}_\theta(T(\mathbf{X})) \geq \frac{[\nabla m(\theta)]^2}{\mathcal{I}(\theta)} \tag{3.41}$$

□

Figure 3.2 illustrates the relationship between the curvature of the log-likelihood (Fisher Information) and the variance of the estimator. A sharper peak implies higher information and lower variance.

Remark 3.2 (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.

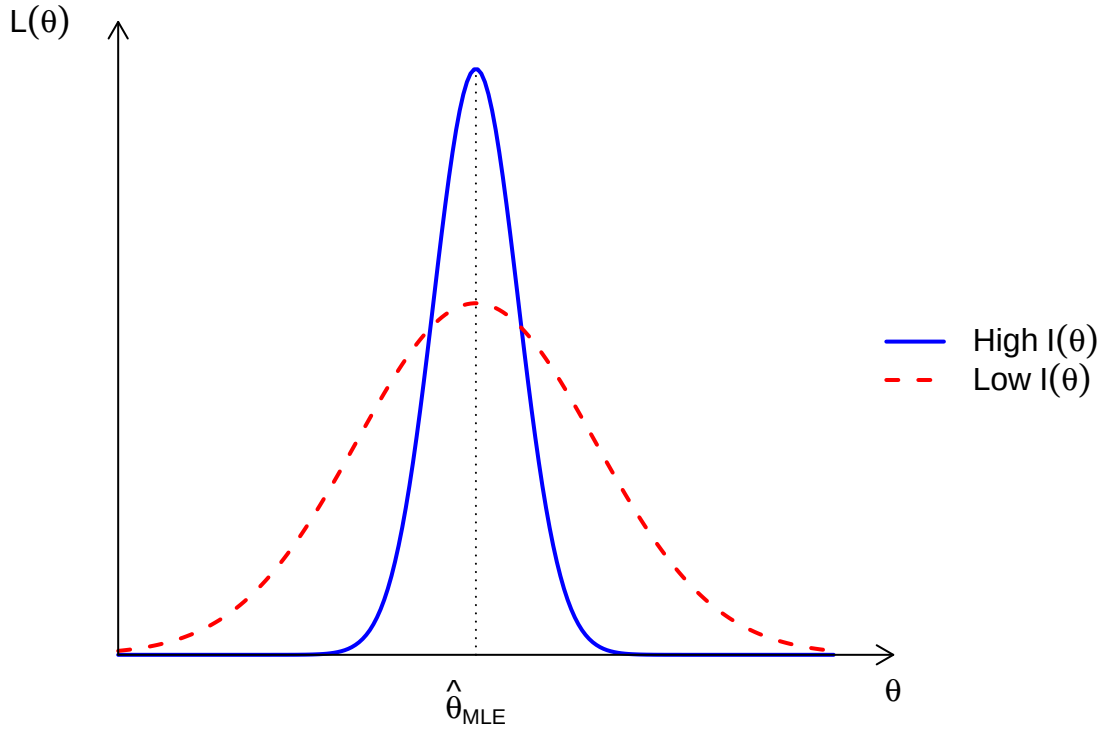


Figure 3.2: Fisher Information: Curvature vs. Variance

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.

3.4.2 Multivariate Cramer-Rao Lower Bound

Theorem 3.4 (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 $\mathcal{J}(\theta)$ *be the* $p \times p$ *Fisher Information Matrix:*

$$\mathcal{J}(\theta) = E_\theta \left[\mathbf{U}(\theta; \mathbf{X})\mathbf{U}(\theta; \mathbf{X})^\top\right] \tag{3.42}$$

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:*

$$Var_\theta(\mathbf{T}) \succeq \mathbf{D}(\theta)[\mathcal{J}(\theta)]^{-1}\mathbf{D}(\theta)^\top \tag{3.43}$$

Here, $\mathbf{A} \succeq \mathbf{B}$ *means that the matrix* $\mathbf{A} - \mathbf{B}$ *is positive semi-definite.*

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 $\text{Var}(\mathbf{U}) = \mathcal{J}(\theta)$. We apply the multivariate extension of the **Score Covariance Identity** (Theorem 3.2). This identity states that the covariance between an estimator and the score vector is the Jacobian of the estimator's expectation:

$$\text{Cov}(\mathbf{T}, \mathbf{U}) = E[\mathbf{T}\mathbf{U}^\top] = \mathbf{D}(\theta) \tag{3.44}$$

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:

$$\text{Var}(\mathbf{Z}) = \begin{pmatrix} \text{Var}(\mathbf{T}) & \text{Cov}(\mathbf{T}, \mathbf{U}) \\ \text{Cov}(\mathbf{U}, \mathbf{T}) & \text{Var}(\mathbf{U}) \end{pmatrix} = \begin{pmatrix} \Sigma_\mathbf{T} & \mathbf{D} \\ \mathbf{D}^\top & \mathcal{J} \end{pmatrix} \succeq 0 \tag{3.45}$$

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

$$\Sigma_\mathbf{T} - \mathbf{D}\mathcal{J}^{-1}\mathbf{D}^\top \succeq 0 \tag{3.46}$$

Thus, we obtain the bound:

$$\text{Var}(\mathbf{T}) \succeq \mathbf{D}\mathcal{J}^{-1}\mathbf{D}^\top \tag{3.47}$$

□

3.4.3 Connection to Stein's Lemma

The **Score Covariance Identity** ($\text{Cov}(\mathbf{T}, \mathbf{U}) = \mathbf{D}(\theta)$) derived in the proof of the CRLB is a fundamental result that links the sensitivity of an estimator to its correlation with the score. When applied to the **Normal distribution**, this identity specializes to the famous **Stein's Lemma**, which relates the covariance of a function and the random vector to the expected gradient of the function.

Theorem 3.5 (Stein's Lemma for Scalar $g(\mathbf{X})$)**.** *Let $\mathbf{X} \sim \mathcal{N}(\theta, \sigma^2\mathbf{I})$, where $\theta \in \mathbb{R}^p$ is the mean vector and $\sigma^2 > 0$ is a known scalar variance. Let $g : \mathbb{R}^p \to \mathbb{R}$ be a differentiable function such that $E[\|\nabla g(\mathbf{X})\|] < \infty$. Then, the following identity holds:*

$$Cov(g(\mathbf{X}), \mathbf{X}) = \sigma^2 E[\nabla g(\mathbf{X})] \tag{3.48}$$

where $\nabla g(\mathbf{X})$ is the gradient vector of g with respect to $\mathbf{X}$.

Proof. **Proof via Score Covariance Identity**

1. The Score Function: For $\mathbf{X} \sim \mathcal{N}(\theta, \sigma^2\mathbf{I})$, the log-likelihood is quadratic, and the score vector is linear in $\mathbf{X}$:
$$\mathbf{U}(\theta) = \nabla_\theta \log f(\mathbf{x}|\theta) = \frac{1}{\sigma^2}(\mathbf{X} - \theta) \tag{3.49}$$
2. Applying the Score Covariance Identity: From the proof of the CRLB, we established the identity $\text{Cov}(T, \mathbf{U}) = \nabla_\theta E[T]$ for any statistic T. Letting $T = g(\mathbf{X})$, we substitute the Normal score $\mathbf{U}$:
$$\text{Cov}\left(g(\mathbf{X}), \frac{\mathbf{X} - \theta}{\sigma^2}\right) = \nabla_\theta E_\theta[g(\mathbf{X})] \tag{3.50}$$

Using the linearity of covariance, the Left-Hand Side (LHS) simplifies to:

$$\text{LHS} = \frac{1}{\sigma^2}\text{Cov}(g(\mathbf{X}), \mathbf{X}) \tag{3.51}$$

3. Evaluating the Sensitivity (RHS): We evaluate the sensitivity of the expectation $\nabla_\theta E_\theta[g(\mathbf{X})]$. Using the substitution $\mathbf{z} = \mathbf{x} - \theta$ (which removes θ from the density f and puts it into g):

$$E_\theta[g(\mathbf{X})] = \int g(\mathbf{z} + \theta) f(\mathbf{z})\, d\mathbf{z} \tag{3.52}$$

Differentiating with respect to θ and noting that $\nabla_\theta g(\mathbf{z} + \theta) = \nabla_{\mathbf{x}} g(\mathbf{x})$:

$$\nabla_\theta E_\theta[g(\mathbf{X})] = \int \nabla g(\mathbf{z} + \theta) f(\mathbf{z})\, d\mathbf{z} = E[\nabla g(\mathbf{X})] \tag{3.53}$$

4. Conclusion: Equating the LHS and RHS:

$$\frac{1}{\sigma^2}\text{Cov}(g(\mathbf{X}), \mathbf{X}) = E[\nabla g(\mathbf{X})] \tag{3.54}$$

Multiplying by σ^2 yields the result.

□

3.4.4 Stein's Lemma (Multivariate Divergence Form)

Let $\mathbf{X} \sim \mathcal{N}(\theta, \sigma^2\mathbf{I})$, where $\theta \in \mathbb{R}^p$ is the mean vector and $\sigma^2 > 0$ is a known scalar variance. Let $\mathbf{g} : \mathbb{R}^p \to \mathbb{R}^p$ be a differentiable vector-valued function, denoted $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \dots, g_p(\mathbf{x}))^\top$, such that $E[|\nabla \cdot \mathbf{g}(\mathbf{X})|] < \infty$.

Then, the following identity holds:

$$E\left[(\mathbf{X} - \theta)^\top \mathbf{g}(\mathbf{X})\right] = \sigma^2 E\left[\nabla \cdot \mathbf{g}(\mathbf{X})\right] \tag{3.55}$$

where $\nabla \cdot \mathbf{g}(\mathbf{X}) = \sum_{i=1}^p \frac{\partial g_i}{\partial x_i}$ is the divergence of $\mathbf{g}$.

Proof. **Proof via Component-wise Score Identity**

1. The Score Function: The score vector for the Normal distribution is $\mathbf{U}(\theta) = \frac{1}{\sigma^2}(\mathbf{X} - \theta)$.
2. Applying the Scalar Identity Component-wise: Let's look at the i-th component of the vector function, $g_i(\mathbf{X})$. Treating g_i as a scalar estimator and X_i as the data, we apply the scalar score identity (or simply integration by parts on the i-th coordinate):

$$\text{Cov}(g_i(\mathbf{X}), X_i) = \sigma^2 E\left[\frac{\partial g_i(\mathbf{X})}{\partial x_i}\right] \tag{3.56}$$

Note that $\text{Cov}(g_i(\mathbf{X}), X_i) = E[(X_i - \theta_i) g_i(\mathbf{X})]$.

3. Summing Components: We sum the identity over all p dimensions:

$$\sum_{i=1}^{p} E\left[(X_i - \theta_i) g_i(\mathbf{X})\right] = \sigma^2 \sum_{i=1}^{p} E\left[\frac{\partial g_i(\mathbf{X})}{\partial x_i}\right] \tag{3.57}$$

4. Vector Notation: The Left-Hand Side is the expected inner product $E[(\mathbf{X}-\theta)^\top \mathbf{g}(\mathbf{X})]$. The Right-Hand Side is the scaled expected divergence $\sigma^2 E[\nabla \cdot \mathbf{g}(\mathbf{X})]$.

$$E\left[(\mathbf{X}-\theta)^\top \mathbf{g}(\mathbf{X})\right] = \sigma^2 E\left[\nabla \cdot \mathbf{g}(\mathbf{X})\right] \tag{3.58}$$

□

i Significance: Stein's Unbiased Risk Estimate (SURE)

This divergence form is famous for enabling **SURE**. If we estimate θ using $\widehat{\theta} = \mathbf{X} + \mathbf{g}(\mathbf{X})$, we can estimate the Mean Squared Error (MSE) purely from the data, because Stein's Lemma allows us to replace the unknown cross-term involving θ with the observable divergence $\nabla \cdot \mathbf{g}(\mathbf{X})$.

3.5 Differentiated Log Likelihood

3.5.1 Mean and Covariance of Score Vector $U(\theta; \mathbf{X})$.

We can also establish Bartlett's identities by analyzing the properties of the likelihood ratio expectation.

Proof. Let $\mathbf{X}$ be a random vector with density $f(\mathbf{x}|\theta)$. We define the function $M(\mathbf{t})$ as the expected value of the likelihood ratio between parameters $\theta + \mathbf{t}$ and θ, where $\mathbf{t}$ is a perturbation vector.

$$M(\mathbf{t}) = E_\theta\left[\exp\left(\ell(\theta+\mathbf{t}; \mathbf{X}) - \ell(\theta; X)\right)\right] \tag{3.59}$$

Since $E_\theta\left[\frac{f(\mathbf{X}|\theta+\mathbf{t})}{f(\mathbf{X}|\theta)}\right] = \int f(\mathbf{x}|\theta+\mathbf{t})\, d\mathbf{x} = 1$, we have the identity:

$$M(\mathbf{t}) \equiv 1 \quad \text{for all } \mathbf{t} \tag{3.60}$$

We expand the log-likelihood difference $\Delta\ell(\mathbf{t}) = \ell(\theta+\mathbf{t}) - \ell(\theta)$ using a Taylor series around $\mathbf{t} = \mathbf{0}$:

$$\Delta\ell(\mathbf{t}) = \mathbf{U}(\theta)^T\mathbf{t} - \frac{1}{2}\mathbf{t}^T\mathbf{J}(\theta)\mathbf{t} + o(\|\mathbf{t}\|^2) \tag{3.61}$$

Substituting this into $M(\mathbf{t})$ and expanding the exponential function:

$$\begin{aligned} M(\mathbf{t}) &= E_\theta\left[\exp\left(\mathbf{U}^T\mathbf{t} - \frac{1}{2}\mathbf{t}^T\mathbf{J}\mathbf{t} + o(\|\mathbf{t}\|^2)\right)\right] \\ &= E_\theta\left[1 + \left(\mathbf{U}^T\mathbf{t} - \frac{1}{2}\mathbf{t}^T\mathbf{J}\mathbf{t}\right) + \frac{1}{2}\left(\mathbf{U}^T\mathbf{t}\right)^2 + o(\|\mathbf{t}\|^2)\right] \\ &= 1 + \mathbf{t}^T E_\theta[\mathbf{U}] + \frac{1}{2}\mathbf{t}^T\left(E_\theta[\mathbf{U}\mathbf{U}^T] - E_\theta[\mathbf{J}]\right)\mathbf{t} + o(\|\mathbf{t}\|^2) \end{aligned} \tag{3.62}$$

Since $M(\mathbf{t}) \equiv 1$, the coefficients of the linear and quadratic terms in $\mathbf{t}$ must be zero:

1. **Linear Term:** $\mathbf{t}^T E_\theta[\mathbf{U}] = 0 \implies E_\theta[\mathbf{U}] = \mathbf{0}$.
2. **Quadratic Term:** $E_\theta[\mathbf{U}\mathbf{U}^T] - E_\theta[\mathbf{J}] = \mathbf{0} \implies \text{Cov}(\mathbf{U}) = E_\theta[\mathbf{J}]$.

□

3.5.2 Alternative Proof of CRLB

We can prove the CRLB using the likelihood ratio expansion method. This approach relates the sensitivity of the estimator (the derivative of its expectation) to its correlation with the score function.

Proof.

1. Setup the Perturbed Expectation: Consider the expectation of the estimator $\mathbf{T}(\mathbf{X})$ under a slightly shifted parameter $\theta + \mathbf{t}$. We can express this as an expectation under the original parameter θ using the likelihood ratio:
$$\mathbf{m}(\theta + \mathbf{t}) = E_{\theta+\mathbf{t}}[\mathbf{T}(\mathbf{X})] = E_\theta\left[\mathbf{T}(\mathbf{X})\frac{f(\mathbf{X}|\theta + \mathbf{t})}{f(\mathbf{X}|\theta)}\right] \tag{3.63}$$
This can be written using the log-likelihood difference $\Delta\ell = \ell(\theta + \mathbf{t}) - \ell(\theta)$:
$$\mathbf{m}(\theta + \mathbf{t}) = E_\theta\left[\mathbf{T}(\mathbf{X})\exp(\Delta\ell)\right] \tag{3.64}$$

2. Taylor Expansion (Left Side): We expand the expectation vector $\mathbf{m}(\theta + \mathbf{t})$ around $\mathbf{t} = \mathbf{0}$. By definition, the linear term is the Jacobian matrix $\mathbf{D}(\theta) = \frac{\partial \mathbf{m}}{\partial \theta}$:
$$\mathbf{m}(\theta + \mathbf{t}) = \mathbf{m}(\theta) + \mathbf{D}(\theta)\mathbf{t} + o(\|\mathbf{t}\|) \tag{3.65}$$

3. Taylor Expansion (Right Side): We expand the likelihood ratio term $\exp(\Delta\ell)$. Since $\Delta\ell \approx \mathbf{U}(\theta)^T\mathbf{t}$, we have $\exp(\Delta\ell) \approx 1 + \mathbf{U}(\theta)^T\mathbf{t}$.
$$\begin{aligned} E_\theta\left[\mathbf{T}(\mathbf{X})(1 + \mathbf{U}^T\mathbf{t} + \dots)\right] &= E_\theta[\mathbf{T}] + E_\theta[\mathbf{T}\mathbf{U}^T\mathbf{t}] + \dots \\ &= \mathbf{m}(\theta) + E_\theta[\mathbf{T}\mathbf{U}^T]\mathbf{t} + o(\|\mathbf{t}\|) \end{aligned} \tag{3.66}$$

4. Matching Linear Coefficients: Since the Left Side must equal the Right Side for any perturbation $\mathbf{t}$, the coefficients of the linear term $\mathbf{t}$ must be identical:
$$\mathbf{D}(\theta) = E_\theta[\mathbf{T}\mathbf{U}^T] \tag{3.67}$$

5. Linking to Covariance: Recall that the score vector has mean zero, $E[\mathbf{U}] = \mathbf{0}$. Therefore, the term $E[\mathbf{T}\mathbf{U}^T]$ is exactly the covariance between the estimator and the score:
$$\text{Cov}(\mathbf{T}, \mathbf{U}) = E[\mathbf{T}\mathbf{U}^T] - E[\mathbf{T}]\underbrace{E[\mathbf{U}]^T}_{\mathbf{0}} = E[\mathbf{T}\mathbf{U}^T] \tag{3.68}$$
Thus, we have the crucial identity:
$$\mathbf{D}(\theta) = \text{Cov}(\mathbf{T}, \mathbf{U}) \tag{3.69}$$

6. Deriving the Inequality: To obtain the bound, we construct the joint covariance matrix of the vector $\mathbf{Z} = \begin{pmatrix} \mathbf{T} \\ \mathbf{U} \end{pmatrix}$. This matrix must be positive semi-definite (PSD).

$$\text{Var}\begin{pmatrix} \mathbf{T} \\ \mathbf{U} \end{pmatrix} = \begin{pmatrix} \text{Var}(\mathbf{T}) & \text{Cov}(\mathbf{T}, \mathbf{U}) \\ \text{Cov}(\mathbf{U}, \mathbf{T}) & \text{Var}(\mathbf{U}) \end{pmatrix} = \begin{pmatrix} \Sigma_{\mathbf{T}} & \mathbf{D} \\ \mathbf{D}^T & \mathcal{J} \end{pmatrix} \succeq 0 \tag{3.70}$$

By the property of Schur complements, for a block matrix to be PSD, the complement of the bottom-right block must be PSD:

$$\Sigma_{\mathbf{T}} - \mathbf{D}\mathcal{J}^{-1}\mathbf{D}^T \succeq 0 \tag{3.71}$$

Rearranging this gives the Multivariate Cramer-Rao Lower Bound:

$$\text{Var}(\mathbf{T}) \succeq \mathbf{D}(\theta)[\mathcal{J}(\theta)]^{-1}\mathbf{D}(\theta)^T \tag{3.72}$$

□

3.6 Exponential Families

A family of probability distributions is defined by the specific mathematical way the parameters and the data interact. For most common distributions, this interaction can be factored into a form that simplifies both theoretical analysis and computational estimation.

Definition 3.3 (Exponential Family)**.** A family of probability density functions (or mass functions) $\{f(\mathbf{x}|\theta) : \theta \in \Theta\}$ is called a k-parameter **Exponential Family** if it can be expressed in the following equivalent forms:

1. **Density Form** The probability density function is written as:

$$f(\mathbf{x}|\theta) = h(\mathbf{x}) \exp\left(\sum_{i=1}^{k} \eta_i(\theta) T_i(\mathbf{x}) - A(\theta)\right) \tag{3.73}$$

2. **Log-Likelihood Form** By taking the natural logarithm, the log-likelihood for a single observation (or a joint sample) is:

$$\ell(\theta; \mathbf{x}) = \sum_{i=1}^{k} \eta_i(\theta) T_i(\mathbf{x}) - A(\theta) + \log h(\mathbf{x}) \tag{3.74}$$

where the components are defined as:

1. $h(\mathbf{x}) \geq 0$ The base measure, which is a function depending only on the data $\mathbf{x}$ and not on the parameter θ.
2. $T_i(\mathbf{x})$ The components of the sufficient statistic vector $\mathbf{T}(\mathbf{x}) = (T_1(\mathbf{x}), \ldots, T_k(\mathbf{x}))^T$.
3. $\eta_i(\theta)$ The natural parameters (or canonical parameters) which determine how the sufficient statistics contribute to the likelihood.

4. $A(\theta)$ The log-partition function (or cumulant function). It is a normalization constant that ensures the density integrates to 1, defined by:

$$A(\theta) = \log \left(\int h(\mathbf{x}) \exp \left(\sum_{i=1}^{k} \eta_i(\theta) T_i(\mathbf{x}) \right) d\mathbf{x} \right) \tag{3.75}$$

3.6.1 Examples

Exponential Distribution

Example 3.2 (Exponential Distribution)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\theta)$, where θ is the scale parameter.

$$f(\mathbf{x}|\theta) = \theta^{-n} \exp \left\{ -\frac{1}{\theta} \sum_{i=1}^{n} x_i \right\} \tag{3.76}$$

The log-likelihood is:

$$\ell(\theta; \mathbf{x}) = -\frac{1}{\theta} \sum_{i=1}^{n} x_i - n \log \theta \tag{3.77}$$

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

$$f(\mathbf{x}|\theta) = [\Gamma(\alpha)\beta^{\alpha}]^{-n} \left(\prod_{i=1}^{n} x_i \right)^{\alpha - 1} \exp \left\{ -\frac{1}{\beta} \sum_{i=1}^{n} x_i \right\} \tag{3.78}$$

The log-likelihood is:

$$\ell(\theta; \mathbf{x}) = (\alpha - 1) \sum_{i=1}^{n} \log x_i - \frac{1}{\beta} \sum_{i=1}^{n} x_i - [n \log \Gamma(\alpha) + n\alpha \log \beta] \tag{3.79}$$

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

Beta Distribution

Example 3.4 (Beta Distribution)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Beta}(a, b)$ with $\theta = (a, b)$.

$$\ell(\theta; \mathbf{x}) = (a-1)\sum_{i=1}^{n} \log x_i + (b-1)\sum_{i=1}^{n} \log(1-x_i) - n \log B(a, b) \tag{3.80}$$

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

$$\ell(\theta; \mathbf{x}) = \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 - \left[\frac{n\mu^2}{2\sigma^2} + \frac{n}{2}\log(2\pi\sigma^2)\right] \tag{3.81}$$

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

3.6.2 Examples of Non-exponential Families

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

Uniform Distribution

Example 3.6 (Uniform Distribution)**.** Let $X \sim U(0, \theta)$.

$$\ell(\theta; x) = -\log \theta + \log I(0 < x < \theta) \tag{3.82}$$

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

3.6.3 Moments of Sufficient Statistics of Exponential Families

3.6.3.1 Means of Sufficient Statistics (General Case)

Theorem 3.6 (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:

$$E_\theta \left[\frac{\partial \ell(\theta; \mathbf{X})}{\partial \theta_j} \right] = 0 \tag{3.83}$$

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

$$\sum_{i=1}^{k} \frac{\partial \eta_i(\theta)}{\partial \theta_j} E[T_i(\mathbf{X})] = \frac{\partial A(\theta)}{\partial \theta_j} \quad \textit{for } j = 1, \dots, d \tag{3.84}$$

Proof. The log-likelihood is:

$$\ell(\theta; \mathbf{x}) = \sum_{i=1}^{k} \eta_i(\theta) T_i(\mathbf{x}) - A(\theta) + \log h(\mathbf{x}) \tag{3.85}$$

Differentiating with respect to θ_j:

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{k} \frac{\partial \eta_i(\theta)}{\partial \theta_j} T_i(\mathbf{x}) - \frac{\partial A(\theta)}{\partial \theta_j} \tag{3.86}$$

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

$$E \left[\sum_{i=1}^{k} \frac{\partial \eta_i(\theta)}{\partial \theta_j} T_i(\mathbf{X}) - \frac{\partial A(\theta)}{\partial \theta_j} \right] = 0 \tag{3.87}$$

$$\sum_{i=1}^{k} \frac{\partial \eta_i(\theta)}{\partial \theta_j} E[T_i(\mathbf{X})] = \frac{\partial A(\theta)}{\partial \theta_j} \tag{3.88}$$

□

3.6.3.2 Natural Parameterization

Definition 3.4 (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:

$$\ell(\eta; \mathbf{x}) = \sum_{i=1}^{k} \eta_i T_i(\mathbf{x}) - A(\eta) + \log h(\mathbf{x}) \tag{3.89}$$

or in vector notation:

$$\ell(\eta; \mathbf{x}) = \eta^T \mathbf{T}(\mathbf{x}) - A(\eta) + \log h(\mathbf{x}) \tag{3.90}$$

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

Definition 3.5 (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 3.7 (Curved Exponential Family Example)**.** Consider the $N(\theta, \theta^2)$ distribution ($d = 1$). The log-likelihood is:

$$\ell(\theta; \mathbf{x}) = -\frac{1}{2\theta^2} \sum x_i^2 + \frac{1}{\theta} \sum x_i - n \log \theta - \text{const} \tag{3.91}$$

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**.

3.6.3.3 Mean and Variance of Sufficient Statistics

Theorem 3.7 (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):***
$$E[\mathbf{T}(\mathbf{X})] = \nabla A(\eta) \tag{3.92}$$
2. ***Covariance (Second Derivative):***
$$Var(\mathbf{T}(\mathbf{X})) = \nabla^2 A(\eta) \tag{3.93}$$

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:*

$$Var(\mathbf{T}(\mathbf{X})) = \mathcal{I}(\eta) \tag{3.94}$$

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 3.1) applied to the canonical log-likelihood:

$$\ell(\eta; \mathbf{x}) = \eta^T \mathbf{T}(\mathbf{x}) - A(\eta) + \log h(\mathbf{x}) \tag{3.95}$$

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

$$\mathbf{U}(\eta) = \nabla_\eta \ell(\eta; \mathbf{x}) = \mathbf{T}(\mathbf{x}) - \nabla A(\eta) \tag{3.96}$$

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

$$E[\mathbf{T}(\mathbf{X}) - \nabla A(\eta)] = \mathbf{0} \implies E[\mathbf{T}(\mathbf{X})] = \nabla A(\eta) \tag{3.97}$$

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

$$\mathbf{J}(\eta) = -\nabla_\eta^2 \ell(\eta; \mathbf{x}) = -\nabla_\eta (\mathbf{T}(\mathbf{x}) - \nabla A(\eta)) = \nabla^2 A(\eta) \tag{3.98}$$

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

$$\text{Cov}(\mathbf{T}(\mathbf{X})) = E[\nabla^2 A(\eta)] = \nabla^2 A(\eta) \tag{3.99}$$

□

Remark 3.3 (Application to Curved Exponential Families). Theorem Theorem 3.6 applies to both **full** and **curved** exponential families. The derivation relies only on the **regularity** of the family, which ensures that the expected score is zero:

1. **Validity of the Score Identity** As long as the support of the distribution does not depend on θ, the identity $E_\theta[\nabla_\theta \ell(\theta; \mathbf{X})] = \mathbf{0}$ remains valid regardless of whether the natural parameters η are independent or constrained.
2. **The Resulting System of Equations** In a curved exponential family, the number of parameters d is less than the number of sufficient statistics k. This means the theorem provides a system of d equations:

$$\sum_{i=1}^{k} \frac{\partial \eta_i(\theta)}{\partial \theta_j} E[T_i(\mathbf{X})] = \frac{\partial A(\theta)}{\partial \theta_j}, \quad j = 1, \dots, d \tag{3.100}$$

   While these d equations are always true, they may not be sufficient to uniquely solve for all k individual expectations $E[T_i(\mathbf{X})]$ without additional information about the structure of the distribution.
3. **Contrast with Canonical Form** In a full exponential family in canonical form ($\eta = \theta$), the Jacobian $\partial \eta_i / \partial \theta_j$ is the identity matrix, simplifying the result to the well-known $E[T_j(\mathbf{X})] = \partial A / \partial \eta_j$. In the curved case, the expectations are "mixed" by the derivatives of the mapping $\eta(\theta)$.

3.6.3.4 Examples

Moments of the Binomial Distribution

Example 3.8 (Moments of the Binomial Distribution)**.** Consider n independent coin flips $X_1, \dots, X_n \sim \text{Bernoulli}(p)$. We find the mean and variance of $T = \sum X_i$.

1. **Log-Likelihood Form** The standard log-likelihood is:

$$\ell(p; \mathbf{x}) = \log\left(\frac{p}{1-p}\right) \sum x_i + n \log(1-p) \tag{3.101}$$

- 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 $\ell(\eta)$:

$$\ell(\eta; \mathbf{x}) = \eta\left(\sum x_i\right) - n\log(1+e^\eta) \tag{3.102}$$

2. **Calculating Moments**

$$E[T] = \frac{\partial A}{\partial \eta} = n\frac{e^\eta}{1+e^\eta} = np \tag{3.103}$$

$$\text{Var}(T) = \frac{\partial^2 A}{\partial \eta^2} = n\frac{e^\eta(1+e^\eta) - e^\eta(e^\eta)}{(1+e^\eta)^2} = n\frac{e^\eta}{(1+e^\eta)^2} = np(1-p) \tag{3.104}$$

Moments of the Gamma Sufficient Statistic

Example 3.9 (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:

$$\ell(\lambda; \mathbf{x}) = -\lambda\sum x_i + n\log\lambda \tag{3.105}$$

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

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

$$\ell(\eta; \mathbf{x}) = \eta\left(\sum x_i\right) - [-n\log(-\eta)] = \eta\sum x_i + n\log(-\eta) \tag{3.106}$$

2. **Calculating Moments**

$$E[T] = \frac{\partial A}{\partial \eta} = -n\frac{1}{-\eta}(-1) = -\frac{n}{\eta} = \frac{n}{\lambda} \tag{3.107}$$

$$\text{Var}(T) = \frac{\partial^2 A}{\partial \eta^2} = \frac{\partial}{\partial \eta}\left(-\frac{n}{\eta}\right) = \frac{n}{\eta^2} = \frac{n}{\lambda^2} \tag{3.108}$$

Moments of Normal Sufficient Statistics

Example 3.10 (Moments of Normal Sufficient Statistics)**.** Consider $X_i \sim N(\mu, \sigma^2)$.

1. **Log-Likelihood Form** The standard log-likelihood is:

$$\ell(\theta; \mathbf{x}) = \frac{\mu}{\sigma^2}\sum x_i - \frac{1}{2\sigma^2}\sum x_i^2 - \left[\frac{n\mu^2}{2\sigma^2} + \frac{n}{2}\log(2\pi\sigma^2)\right] \tag{3.109}$$

- 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 $\mu = -\frac{\eta_1}{2\eta_2}$:

$$A(\eta) = -\frac{n\eta_1^2}{4\eta_2} - \frac{n}{2}\log(-2\eta_2) + \frac{n}{2}\log(2\pi) \tag{3.110}$$

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

$$\ell(\eta; \mathbf{x}) = \eta_1 \left(\sum x_i\right) + \eta_2 \left(\sum x_i^2\right) - \left[-\frac{n\eta_1^2}{4\eta_2} - \frac{n}{2}\log(-2\eta_2)\right] \tag{3.111}$$

2. **First Moments (Means)**

$$E[T_1] = E\left[\sum X_i\right] = \frac{\partial A}{\partial \eta_1} = -\frac{2n\eta_1}{4\eta_2} = -\frac{n\eta_1}{2\eta_2} = n\mu \tag{3.112}$$

$$E[T_2] = E\left[\sum X_i^2\right] = \frac{\partial A}{\partial \eta_2} = \frac{n\eta_1^2}{4\eta_2^2} - \frac{n}{2(-2\eta_2)}(-2) = \frac{n\eta_1^2}{4\eta_2^2} - \frac{n}{2\eta_2} \tag{3.113}$$

Subbing back μ, σ:

$$= n\mu^2 + n\sigma^2 = n(\mu^2 + \sigma^2) \tag{3.114}$$

3. **Second Moment (Covariance)**

$$\text{Cov}(T_1, T_2) = \frac{\partial^2 A}{\partial \eta_1 \partial \eta_2} = \frac{\partial}{\partial \eta_2}\left(-\frac{n\eta_1}{2\eta_2}\right) = \frac{n\eta_1}{2\eta_2^2} = 2n\mu\sigma^2 \tag{3.115}$$

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:

$$\bar{X} = \frac{1}{n}T_1 \tag{3.116}$$

$$S^2 = \frac{1}{n-1}\left(\sum X_i^2 - n\bar{X}^2\right) = \frac{1}{n-1}\left(T_2 - \frac{1}{n}T_1^2\right) \tag{3.117}$$

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

$$\text{Cov}\left(T_1, T_2 - \frac{1}{n}T_1^2\right) = \text{Cov}(T_1, T_2) - \frac{1}{n}\text{Cov}(T_1, T_1^2) \tag{3.118}$$

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 $\tau^2 = n\sigma^2$:

$$\text{Cov}(T_1, T_1^2) = 2(n\mu)(n\sigma^2) = 2n^2\mu\sigma^2 \tag{3.119}$$

Substituting this back into the expression:

$$\text{Cov}\left(T_1, T_2 - \frac{1}{n}T_1^2\right) = \underbrace{2n\mu\sigma^2}_{\text{From Part 3}} - \frac{1}{n}(2n^2\mu\sigma^2) = 2n\mu\sigma^2 - 2n\mu\sigma^2 = 0 \tag{3.120}$$

Moments of the Curved Normal $N(\theta, \theta^2)$

Example 3.11 (Moments of the Curved Normal $N(\theta, \theta^2)$)**.** Consider a sample $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \theta^2)$. This is a curved exponential family with $d = 1$ parameter and $k = 2$ sufficient statistics.

1. **Log-Likelihood Components** The log-likelihood is given by:

$$\ell(\theta; \mathbf{x}) = -\frac{1}{2\theta^2} \sum x_i^2 + \frac{1}{\theta} \sum x_i - n \log \theta - \text{const} \tag{3.121}$$

   From this, we identify:

   - Natural parameters: $\eta_1(\theta) = -1/(2\theta^2)$ and $\eta_2(\theta) = 1/\theta$.
   - Sufficient statistics: $T_1 = \sum X_i^2$ and $T_2 = \sum X_i$.
   - Log-partition function: $A(\theta) = n \log \theta$.

2. **Applying the Theorem** Since $d = 1$, we have one equation from the score identity $\partial A/\partial \theta = \sum (\partial \eta_i/\partial \theta) E[T_i]$:

$$\frac{n}{\theta} = \frac{\partial \eta_1}{\partial \theta} E[T_1] + \frac{\partial \eta_2}{\partial \theta} E[T_2] \tag{3.122}$$

3. **Calculating Derivatives**

   - $\frac{\partial \eta_1}{\partial \theta} = \frac{\partial}{\partial \theta}(-\frac{1}{2}\theta^{-2}) = \theta^{-3} = \frac{1}{\theta^3}$
   - $\frac{\partial \eta_2}{\partial \theta} = \frac{\partial}{\partial \theta}(\theta^{-1}) = -\theta^{-2} = -\frac{1}{\theta^2}$
   - $\frac{\partial A}{\partial \theta} = \frac{n}{\theta}$

4. **Solving the System** Substituting these into the equation:

$$\frac{n}{\theta} = \frac{1}{\theta^3} E\left[\sum X_i^2\right] - \frac{1}{\theta^2} E\left[\sum X_i\right] \tag{3.123}$$

   Multiplying by θ^3 gives:

$$n\theta^2 = E\left[\sum X_i^2\right] - \theta E\left[\sum X_i\right] \tag{3.124}$$

   We can verify this using the known moments $E[X_i] = \theta$ and $E[X_i^2] = \text{Var}(X) + (E[X])^2 = \theta^2 + \theta^2 = 2\theta^2$:

$$n\theta^2 = (2n\theta^2) - \theta(n\theta) = 2n\theta^2 - n\theta^2 = n\theta^2 \tag{3.125}$$

   The identity holds, demonstrating that the theorem correctly relates the moments even when the sufficient statistics are "mixed" by the parameter constraints.

3.6.4 Maximum Likelihood and Moment Matching Estimation Scheme

For a multiparameter exponential family, the inner product $\eta^\top \mathbf{T}(\mathbf{y})$ can be written explicitly as a sum over the p components of the canonical parameter and sufficient statistic vectors: $\sum_{j=1}^{p} \eta_j T_j(\mathbf{y})$.

Given a sample of data $\mathbf{y}$, the log-likelihood function parameterized by the canonical parameters η takes the explicit form:

$$\ell(\eta; \mathbf{y}) = \sum_{j=1}^{p} \eta_j T_j(\mathbf{y}) - A(\eta) + h(\mathbf{y}) \tag{3.126}$$

Taking the partial derivative of the log-likelihood with respect to each canonical parameter η_j yields the components of the score function. Setting these to zero gives the Maximum Likelihood Estimator (MLE) equations:

$$\frac{\partial}{\partial \eta_j} \ell(\eta; \mathbf{y}) = T_j(\mathbf{y}) - \frac{\partial A(\eta)}{\partial \eta_j} = 0 \tag{3.127}$$

Therefore, finding the MLE in a canonical exponential family is exactly equivalent to solving the system of equations formed by equating each observed sufficient statistic directly to the corresponding partial derivative of the log-partition function:

$$T_j(\mathbf{y}) = \frac{\partial A(\eta)}{\partial \eta_j} \quad \text{for } j = 1, \dots, p \tag{3.128}$$

This establishes a powerful estimating scheme: the maximum likelihood estimates for the canonical parameters are found by constructing a system of equations where the observed sufficient statistics are matched to their theoretical expectations (the derivatives of the log-partition function).

3.6.4.1 Example 1: Poisson Distribution with a Common Mean

Consider an independent and identically distributed (i.i.d.) sample $y_1, y_2, \dots, y_n \sim \text{Poisson}(\lambda)$. The joint probability mass function can be written in exponential family form:

$$f(\mathbf{y}; \lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \exp\left(\log(\lambda) \sum_{i=1}^{n} y_i - n\lambda - \sum_{i=1}^{n} \log(y_i!) \right) \tag{3.129}$$

From this expression, we can identify the components of the canonical exponential family:

- **Canonical parameter:** There is a single parameter ($p = 1$), $\eta = \log(\lambda)$. This implies $\lambda = e^\eta$.
- **Sufficient statistic:** $T(\mathbf{y}) = \sum_{i=1}^{n} y_i$.
- **Log-partition function:** Expressed in terms of the canonical parameter, $A(\eta) = n\lambda = ne^\eta$.

Applying the estimating scheme, we first find the derivative of the log-partition function with respect to the canonical parameter:

$$\frac{\partial A(\eta)}{\partial \eta} = ne^\eta \tag{3.130}$$

We then set the observed sufficient statistic equal to this derivative:

$$T(\mathbf{y}) = \frac{\partial A(\eta)}{\partial \eta} \tag{3.131}$$

$$\sum_{i=1}^{n} y_i = n e^{\hat{\eta}} \tag{3.132}$$

Substituting back $\hat{\lambda} = e^{\hat{\eta}}$, we arrive at the familiar estimator:

$$\sum_{i=1}^{n} y_i = n\hat{\lambda} \implies \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \tag{3.133}$$

3.6.5 Example: Poisson Regression (MLE)

Consider a GLM where $Y_i \sim \text{Poisson}(\lambda_i)$. We use the canonical link $\log(\lambda_i(\beta)) = \mathbf{x}_i^\top \beta$, which implies the mean function is $\lambda_i(\beta) = \exp(\mathbf{x}_i^\top \beta)$.

1. **Exponential Family Representation**

   We rewrite the total log-likelihood by grouping terms associated with each β_j:

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^{n} \left(y_i(\mathbf{x}_i^\top \beta) - \lambda_i(\beta) - \log(y_i!) \right) \\ &= \sum_{i=1}^{n} \sum_{j=0}^{p} y_i x_{ij} \beta_j - \sum_{i=1}^{n} \exp(\mathbf{x}_i^\top \beta) - \sum_{i=1}^{n} \log(y_i!) \\ &= \sum_{j=0}^{p} \beta_j \underbrace{\left(\sum_{i=1}^{n} y_i x_{ij} \right)}_{T_j(\mathbf{y})} - \underbrace{\sum_{i=1}^{n} \lambda_i(\beta)}_{A(\beta)} + \text{const} \end{aligned} \tag{3.134}$$

   This is a multivariate exponential family in **canonical form** where:

   - **Natural Parameters:** The regression coefficients $\beta = (\beta_0, \dots, \beta_p)^\top$.
   - **Sufficient Statistics:** $\mathbf{T}(\mathbf{y}) = \mathbf{X}^\top \mathbf{y}$.
   - **Log-Partition Function:** $A(\beta) = \sum_{i=1}^{n} \lambda_i(\beta)$.

2. **Derivations via Moments of Sufficient Statistics**

   According to the properties of the canonical exponential family, the expected value of the sufficient statistic vector is exactly the gradient of the log-partition function: $E[\mathbf{T}(\mathbf{Y})] = \nabla A(\beta)$.

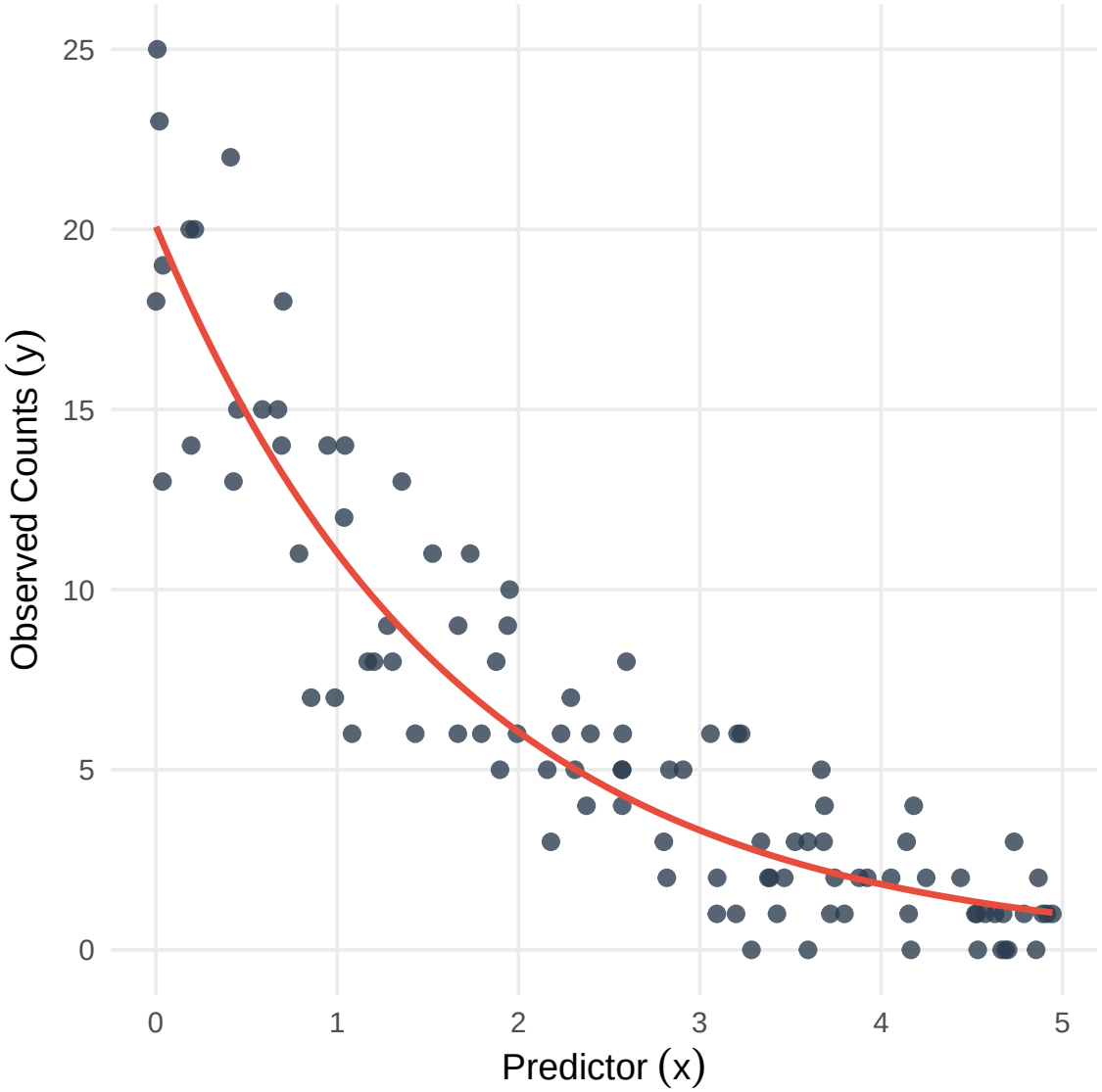


Figure 3.3: Simulated Poisson regression data ($n = 100$) with a true negative slope parameter ($\beta_1 = -0.6$). The red curve represents the true underlying expected mean function, $\lambda(x) = \exp(\beta_0 + \beta_1 x)$.

To find $\nabla A(\beta)$, we first take the partial derivative of $A(\beta)$ with respect to an individual regression coefficient β_j. Applying the chain rule to the sum of exponentials:

$$\begin{aligned}\frac{\partial A(\beta)}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \exp\left(\mathbf{x}_i^\top \beta\right) \\ &= \sum_{i=1}^{n} \exp\left(\mathbf{x}_i^\top \beta\right) \cdot \frac{\partial}{\partial \beta_j} \left(\sum_{k=0}^{p} x_{ik} \beta_k \right) \\ &= \sum_{i=1}^{n} \lambda_i(\beta) x_{ij}\end{aligned} \tag{3.135}$$

Collecting these partial derivatives for $j = 0, \dots, p$ into a single column vector allows us to express the full gradient compactly in matrix notation:

$$\nabla A(\beta) = \mathbf{X}^\top \lambda(\beta) \tag{3.136}$$

where $\lambda(\beta) = (\lambda_1(\beta), \dots, \lambda_n(\beta))^\top$ is the vector of conditional means.

The **Method of Moments** estimating scheme constructs estimators by equating the observed sufficient statistics $\mathbf{T}(\mathbf{y})$ to their theoretical expectations $\nabla A(\beta)$. Substituting our derived terms yields the parameter estimating equations:

$$\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \lambda(\beta) \tag{3.137}$$

Rearranging this system gives the explicit residual condition that must be satisfied by the estimated parameters:

$$\mathbf{X}^\top (\mathbf{y} - \lambda(\beta)) = \mathbf{0} \tag{3.138}$$

Connection to Least Squares Theory

The resulting score equation for the MLE, $\mathbf{U}(\widehat{\beta}) = \mathbf{0}$, simplifies to $\mathbf{X}^\top (\mathbf{y} - \lambda(\widehat{\beta})) = \mathbf{0}$. This reveals a profound structural and geometric connection to standard Least Squares Theory (LST) for linear models.

In Ordinary Least Squares, assuming a Gaussian error structure (which is also an exponential family) with an identity link, the estimated mean vector is given directly by the linear predictor: $\widehat{\mu} = \mathbf{X}\widehat{\beta}$. The standard normal equations used to find $\widehat{\beta}$ are typically written explicitly in terms of the residuals:

$$\mathbf{X}^\top (\mathbf{y} - \widehat{\mu}) = \mathbf{0} \tag{3.139}$$

Comparing this to the Poisson regression result, where the estimated mean is $\widehat{\lambda} = \lambda(\widehat{\beta})$, we see that the canonical link function forces the exact same algebraic condition. In both settings, the estimating equations demand that the raw residual vector—whether it is $\mathbf{y} - \widehat{\mu}$ for the Gaussian case or $\mathbf{y} - \widehat{\lambda}$ for the Poisson case—must be strictly orthogonal to every column of the design matrix $\mathbf{X}$.

Geometrically, this establishes that the maximum likelihood estimates obtained via a canonical link ensure the residuals are orthogonal to the column space of $\mathbf{X}$, denoted as $\mathcal{C}(\mathbf{X})$. The moment-matching property of the canonical exponential family thereby elegantly preserves the fundamental projection-based geometry of standard linear models, extending it into the broader framework of Generalized Linear Models.

4 Maximum Likelihood Estimation

4.1 Maximum Likelihood Estimation

4.1.1 Definitions and Notations

Definition 4.1 (Likelihood and Log-Likelihood)**.** Let $f(\mathbf{x}|\theta)$ be the joint probability density function (or mass function) of the data $\mathbf{X} = (X_1, \dots, X_n)$.

1. **Joint Likelihood Function:**

   When viewed as a function of the parameter θ given fixed data $\mathbf{x}$, it is called the likelihood function:

   $$L(\theta; \mathbf{x}) = f(\mathbf{x}|\theta) \tag{4.1}$$

2. **Joint Log-Likelihood:**

   It is usually easier to maximize the natural logarithm of the likelihood:

   $$\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) \tag{4.2}$$

3. **Independent Observations:**

   If the observations $X_1, \dots, X_n$ are independent, the joint likelihood factors into a product of individual densities. Let $\ell(\theta; x_i) = \log f(x_i|\theta)$ denote the log-likelihood for a **single observation**. The total log-likelihood is strictly the sum of the individual log-likelihoods:

   $$\ell(\theta; \mathbf{x}) = \sum_{i=1}^{n} \ell(\theta; x_i) \tag{4.3}$$

Definition 4.2 (Maximum Likelihood Estimation)**. Maximum Likelihood Estimator (MLE):** The MLE $\hat{\theta}_{\text{MLE}}$ is the value in the parameter space Θ that maximizes the likelihood (and equivalently, the log-likelihood) function:

$$\hat{\theta}_{\text{MLE}}(\mathbf{x}) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \ell(\theta; \mathbf{x}) \tag{4.4}$$

Definition 4.3 (Score Function)**.** The score function is defined as the gradient of the log-likelihood with respect to the parameter vector θ. Finding the MLE often involves solving the score equation $\mathbf{U}(\theta; \mathbf{x}) = \mathbf{0}$.

1. **Total Score Function (U):**

$$\mathbf{U}(\theta; \mathbf{x}) = \nabla_\theta \ell(\theta; \mathbf{x}) \tag{4.5}$$

2. **Single Observation Score ($\mathbf{u}_i$):**

   For a single independent observation x_i, its score is:

$$\mathbf{u}(\theta; x_i) = \nabla_\theta \ell(\theta; x_i) \tag{4.6}$$

3. **Sum for Independent Data:**

   By the linearity of the derivative operator, the total score function for independent observations is the sum of the individual score functions:

$$\mathbf{U}(\theta; \mathbf{x}) = \sum_{i=1}^{n} \mathbf{u}(\theta; x_i) \tag{4.7}$$

Definition 4.4 (Fisher Information)**.** The Fisher Information measures the curvature of the log-likelihood surface and, correspondingly, the amount of information the data carries about the unknown parameter.

1. **Observed Fisher Information (J):**

   This is the negative Hessian matrix of the log-likelihood, evaluated at the observed data.

   - **Total:** $\mathbf{J}_n(\theta; \mathbf{x}) = -\nabla^2_\theta \ell(\theta; \mathbf{x})$
   - **Single Observation:** $\mathbf{J}_1(\theta; x_i) = -\nabla^2_\theta \ell(\theta; x_i)$
   - **Sum for Independent Data:**

$$\mathbf{J}_n(\theta; \mathbf{x}) = \sum_{i=1}^{n} \mathbf{J}_1(\theta; x_i) \tag{4.8}$$

2. **Expected Fisher Information ($\mathcal{I}$):**

   This is the covariance matrix of the score vector, which simplifies to the expected value of the observed Fisher information. It is a deterministic $p \times p$ matrix.

   - **Total:** $\mathcal{I}_n(\theta) = E_\theta\left[\mathbf{J}_n(\theta; \mathbf{X})\right]$
   - **Single Observation:** $\mathcal{I}_1(\theta) = E_\theta\left[\mathbf{J}_1(\theta; X_i)\right]$
   - **Sum for Independent Data:**

$$\mathcal{I}_n(\theta) = \sum_{i=1}^{n} \mathcal{I}_1(\theta) \tag{4.9}$$

   (Note: If the independent observations are also identically distributed (i.i.d.), this simplifies further to $\mathcal{I}_n(\theta) = n\mathcal{I}_1(\theta)$).

4.1.2 Example: MLE of Normal Sample

Example 4.1 (Example: Normal Distribution (Exact Asymptotics))**.** Let $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ with σ^2 known. We explicitly derive the key quantities and map them to the four general asymptotic results under $H_0 : \theta = \theta_0$.

1. Derivation of Key Quantities

   - **Log-Likelihood:** Using the identity $\sum(x_i - \theta)^2 = \sum(x_i - \overline{x})^2 + n(\overline{x} - \theta)^2$, the log-likelihood is a perfect quadratic centered at $\overline{x}$:

$$\ell(\theta) = \underbrace{-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2}_{C} - \frac{n}{2\sigma^2}(\theta - \overline{x})^2 \tag{4.10}$$

   - **Score Function (U_n):** Differentiating with respect to θ:

$$U_n(\theta) = \ell'(\theta) = \frac{n}{\sigma^2}(\overline{X}_n - \theta) \tag{4.11}$$

   - **Maximum Likelihood Estimator ($\hat{\theta}_n$):** Setting $U_n(\hat{\theta}) = 0 \implies \hat{\theta}_n = \overline{X}_n$.
   - **Fisher Information ($\mathcal{I}_n$):** Taking the negative second derivative:

$$\mathcal{I}_n(\theta) = -E[\ell''(\theta)] = \frac{n}{\sigma^2} \tag{4.12}$$

2. Asymptotic Distributions

   We now illustrate the four main theorems using these derived forms under $H_0 : \theta = \theta_0$:

   1. **Consistency:** The MLE $\hat{\theta}_n = \overline{X}_n \xrightarrow{p} \theta_0$ by the Weak Law of Large Numbers.
   2. **Asymptotic Normality of the Score:** Evaluated at the null, the score follows a Normal distribution exactly:

$$U_n(\theta_0) \sim N(0, \mathcal{I}_n(\theta_0)) \tag{4.13}$$

   3. **Asymptotic Normality of the Estimator:** The distribution of the standardized MLE centered at the null value is:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \sim N(0, \sigma^2) = N(0, \mathcal{I}_1(\theta_0)^{-1}) \tag{4.14}$$

   4. **Wilks' Theorem (Deviance):** The Deviance is defined as $D = 2[\ell(\hat{\theta}) - \ell(\theta_0)]$.Using the quadratic form of the log-likelihood derived above, we can directly compute the values at the MLE ($\hat{\theta} = \overline{x}$) and the null ($\theta_0$):
      - $\ell(\hat{\theta}) = C - \frac{n}{2\sigma^2}(\overline{x} - \overline{x})^2 = C$
      - $\ell(\theta_0) = C - \frac{n}{2\sigma^2}(\theta_0 - \overline{x})^2$

Substituting these into the Deviance formula directly yields:

$$D = 2\left[C - \left(C - \frac{n}{2\sigma^2}(\overline{x} - \theta_0)^2\right)\right] = \frac{n}{\sigma^2}(\overline{x} - \theta_0)^2 \tag{4.15}$$

Rearranging this expression reveals the square of a standard Normal variable (Z):

$$D = \left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}}\right)^2 = Z^2 \sim \chi^2_1 \tag{4.16}$$

This confirms that for the Normal distribution, the χ^2_1 distribution of the Deviance is exact for any sample size n.

4.1.2.1 Uniform Distribution (Boundary Case)

Example 4.2. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta)$. The likelihood is:

$$L(\theta; \mathbf{x}) = \frac{1}{\theta^n} I(x_{(n)} \leq \theta) \tag{4.17}$$

This function strictly decreases for $\theta \geq x_{(n)}$ and is zero otherwise. Thus:

$$\hat{\theta}_{\text{MLE}} = x_{(n)} \tag{4.18}$$

Note that the Score equation approach fails here because the support depends on θ, making the log-likelihood discontinuous at the boundary.

4.1.3 Summary of Key Asymptotic Results for MLE

Under standard regularity conditions (e.g., i.i.d. observations, smooth density, and θ in the interior of Θ), the MLE exhibits the following asymptotic properties as $n \to \infty$:

1. **Consistency:**

   The estimator converges in probability to the true parameter θ_0.

   $$\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta_0 \tag{4.19}$$

2. **Asymptotic Normality of the Score:**

   The Score vector evaluated at the true parameter converges to a Normal distribution with variance equal to the Fisher Information.

   $$\frac{1}{\sqrt{n}}\mathbf{U}(\theta_0) \xrightarrow{d} N_p(\mathbf{0}, \mathcal{I}_1(\theta_0)) \tag{4.20}$$

   (Where $\mathcal{I}_1$ is the expected information for a single observation).

3. **Asymptotic Normality of the Estimator:**

   The MLE is asymptotically normal, unbiased, and efficient (achieving the Cramér-Rao lower bound).

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{d} N_p(\mathbf{0}, \mathcal{I}_1(\theta_0)^{-1}) \tag{4.21}$$

4. **Wilks' Theorem (Likelihood Ratio Test for a Simple Null):**

   The Likelihood Ratio Statistic (Deviance) evaluates the drop in the log-likelihood from the Maximum Likelihood Estimator ($\hat{\theta}$) to the true value (θ_0):

$$D = 2\left(\ell(\hat{\theta}) - \ell(\theta_0)\right) \tag{4.22}$$

   The statistic converges to a Chi-squared distribution:

$$D \xrightarrow{d} \chi^2_p \tag{4.23}$$

   (The degrees of freedom equal the dimension of the parameter vector p, because all p parameters are fixed under the null. For a single scalar parameter θ, this reduces to χ^2_1).

4.2 Newton-Raphson and Fisher Scoring Algorithms for Finding MLE

When the score equation $\mathbf{U}(\theta) = \mathbf{0}$ does not have a closed-form solution, we must rely on iterative numerical methods to find the Maximum Likelihood Estimator. The Newton-Raphson algorithm is the most common approach, utilizing the curvature of the log-likelihood surface to guide the search for the maximum.

4.2.1 Newton-Raphson Iteration

Algorithm 4.1 (Newton-Raphson Iteration)**.** *Starting from an initial guess $\theta^{(0)}$, the algorithm updates the estimate of the parameter vector at each step t using the following recurrence relation:*

$$\theta^{(t+1)} = \theta^{(t)} + \left[\mathbf{J}(\theta^{(t)})\right]^{-1} \mathbf{U}(\theta^{(t)}) \tag{4.24}$$

where:

1. $\mathbf{U}(\theta^{(t)})$ *is the score vector (gradient) evaluated at the current estimate.*
2. $\mathbf{J}(\theta^{(t)})$ *is the* ***Observed Fisher Information Matrix****, defined as the negative Hessian of the log-likelihood:*

$$\mathbf{J}(\theta) = -\nabla^2 \ell(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T} \tag{4.25}$$

Derivation via Taylor Expansion. The algorithm is derived by approximating the score function $\mathbf{U}(\theta)$ with a first-order Taylor expansion around the current estimate $\theta^{(t)}$:

1. **Linear Approximation**

   We approximate the score at the next step as:

$$\mathbf{U}(\theta^{(t+1)}) \approx \mathbf{U}(\theta^{(t)}) + \nabla\mathbf{U}(\theta^{(t)})(\theta^{(t+1)} - \theta^{(t)}) \tag{4.26}$$

2. **Setting to Zero**

   Since we want to find the root where $\mathbf{U}(\hat{\theta}) = \mathbf{0}$, we set the left side to zero:

$$\mathbf{0} = \mathbf{U}(\theta^{(t)}) - \mathbf{J}(\theta^{(t)})(\theta^{(t+1)} - \theta^{(t)}) \tag{4.27}$$

   Note that $\nabla\mathbf{U}(\theta) = \nabla^2\ell(\theta) = -\mathbf{J}(\theta)$.

3. **Solving for the Update**

   Rearranging the terms gives the update rule:

$$\mathbf{J}(\theta^{(t)})(\theta^{(t+1)} - \theta^{(t)}) = \mathbf{U}(\theta^{(t)}) \tag{4.28}$$

$$\theta^{(t+1)} = \theta^{(t)} + \mathbf{J}(\theta^{(t)})^{-1}\mathbf{U}(\theta^{(t)}) \tag{4.29}$$

□

4.2.2 Fisher Scoring

Algorithm 4.2 (Fisher Scoring)**.** *The Newton-Raphson algorithm specifically uses the* ***Observed Information*** $\mathbf{J}(\theta)$*. If the expected Fisher Information* $\mathcal{I}(\theta) = E[\mathbf{J}(\theta)]$ *is used instead, the method is known as* ***Fisher Scoring****.*

1. ***Convergence Speed***

   Newton-Raphson typically achieves quadratic convergence near the maximum, making it very fast when the initial guess is good.

2. ***Stability***

   Because $\mathbf{J}(\theta)$ *depends on the specific data observed, it can occasionally be non-positive definite in regions far from the MLE, which may cause the algorithm to diverge. Fisher Scoring is often more stable in these cases because* $\mathcal{I}(\theta)$ *is always positive semi-definite for regular families.*

4.2.3 Example: Poisson Regression (MLE)

Example 4.3. Consider a GLM where $Y_i \sim \text{Poisson}(\lambda_i)$. We use the canonical link $\log(\lambda_i(\beta)) = \mathbf{x}_i^\top \beta$, which implies the mean function is $\lambda_i(\beta) = \exp(\mathbf{x}_i^\top \beta)$.

1. **Exponential Family Representation**

   We rewrite the total log-likelihood by grouping terms associated with each β_j:

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{n} \left(y_i(\mathbf{x}_i^\top \beta) - \lambda_i(\beta) - \log(y_i!)\right) \\ &= \sum_{i=1}^{n}\sum_{j=0}^{p} y_i x_{ij}\beta_j - \sum_{i=1}^{n} \exp(\mathbf{x}_i^\top \beta) - \sum_{i=1}^{n} \log(y_i!) \\ &= \sum_{j=0}^{p} \beta_j \underbrace{\left(\sum_{i=1}^{n} y_i x_{ij}\right)}_{T_j(\mathbf{y})} - \underbrace{\sum_{i=1}^{n} \lambda_i(\beta)}_{A(\beta)} + \text{const}\end{aligned} \tag{4.30}$$

   This is a multivariate exponential family in **canonical form** where:

   - **Natural Parameters:** The regression coefficients $\beta = (\beta_0, \dots, \beta_p)^\top$.
   - **Sufficient Statistics:** $\mathbf{T}(\mathbf{y}) = \mathbf{X}^\top \mathbf{y}$.
   - **Log-Partition Function:** $A(\beta) = \sum_{i=1}^{n} \lambda_i(\beta)$.

2. **Derivations via Moments of Sufficient Statistics**

   Since we are in canonical form, the Score and Information are simply the derivatives of the log-partition function $A(\beta)$.

   - **Score Vector (U):**

$$\mathbf{U}(\beta) = \mathbf{T}(\mathbf{y}) - \nabla A(\beta) = \mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \lambda(\beta) = \mathbf{X}^\top (\mathbf{y} - \lambda(\beta)) \tag{4.31}$$

     where $\lambda(\beta) = (\lambda_1(\beta), \dots, \lambda_n(\beta))^\top$.

   - **Fisher Information (J):** The information is the Hessian of the log-partition function:

$$J(\beta) = \nabla^2 A(\beta) = \mathbf{X}^\top \mathbf{W}(\beta)\mathbf{X} \tag{4.32}$$

     where $\mathbf{W}(\beta) = \text{diag}(\lambda_1(\beta), \dots, \lambda_n(\beta))$.

   The details of procedure for deriving the above formula for score and fisher information matrix using scalar derivative is given in Section 4.6.

3. **Newton-Raphson Algorithm for Poisson MLE**

   To find the MLE $\widehat{\beta}$, we iterate until convergence:

$$\beta^{(t+1)} = \beta^{(t)} + [J(\beta^{(t)})]^{-1} \mathbf{U}(\beta^{(t)}) \tag{4.33}$$

Substituting the Poisson-specific forms:

$$\beta^{(t+1)} = \beta^{(t)} + (\mathbf{X}^\top \mathbf{W}(\beta^{(t)})\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{y} - \lambda(\beta^{(t)})) \tag{4.34}$$

This specific iteration is also known as **Iteratively Reweighted Least Squares (IRLS)**.

```r
library(latex2exp)

# 1. Data Generation
set.seed(42)
n <- 50
beta_true <- c(0.5, 1.5)
x_vals <- runif(n, -1, 1)
X_mat <- cbind(1, x_vals)
y_obs <- rpois(n, lambda = exp(X_mat %*% beta_true))

# 2. Log-Likelihood Function for Contours
log_lik <- function(b0, b1) {
  eta <- b0 + b1 * x_vals
  sum(y_obs * eta - exp(eta))
}

# 3. Newton-Raphson Implementation
b_curr <- c(2, 3) # Initial guess
path <- matrix(NA, nrow = 10, ncol = 2)
path[1, ] <- b_curr

for (t in 2:10) {
  lambda <- as.vector(exp(X_mat %*% b_curr))
  W <- diag(lambda)
  U <- t(X_mat) %*% (y_obs - lambda)
  J <- t(X_mat) %*% W %*% X_mat

  b_new <- b_curr + solve(J) %*% U
  path[t, ] <- b_new
  b_curr <- b_new
}
path <- na.omit(path)
beta_mle <- path[nrow(path), ]

# --- Plotting Side-by-Side ---
par(mfrow = c(1, 2))


# Plot B: Data and Fitted Curves
```

```r
plot(x_vals, y_obs, pch = 21, bg = "gray80", col = "gray50",
     xlab = "x", ylab = "y", main = "Poisson Regression Fit")

# Generate smooth lines for lambda
x_smooth <- seq(-1, 1, length.out = 100)
lambda_true <- exp(beta_true[1] + beta_true[2] * x_smooth)
lambda_mle  <- exp(beta_mle[1] + beta_mle[2] * x_smooth)

lines(x_smooth, lambda_true, col = "blue", lwd = 2, lty = 2)
lines(x_smooth, lambda_mle, col = "red", lwd = 2)

legend("topleft", legend = c("True Mean", "MLE Mean", "Observed Data"),
       col = c("blue", "red", "gray50"), lty = c(2, 1, NA),
       pch = c(NA, NA, 21), pt.bg = "gray80", bty = "n", cex = 0.8)

# Plot A: NR Optimization Path
b0_grid <- seq(-0.5, 2.5, length.out = 50)
b1_grid <- seq(0.5, 3.5, length.out = 50)
z_vals <- outer(b0_grid, b1_grid, Vectorize(log_lik))

contour(b0_grid, b1_grid, z_vals, nlevels = 20,
        xlab = TeX("$\\beta_0$"), ylab = TeX("$\\beta_1$"),
        main = "Newton-Raphson Path")
lines(path[,1], path[,2], type = "b", col = "red", pch = 19, lwd = 1.5)
points(beta_true[1], beta_true[2], col = "blue", pch = 4, lwd = 2, cex = 1.2)
legend("bottomright", legend = c("NR Path", "True Beta"),
       col = c("red", "blue"), pch = c(19, 4), bty = "n", cex = 0.8)
```

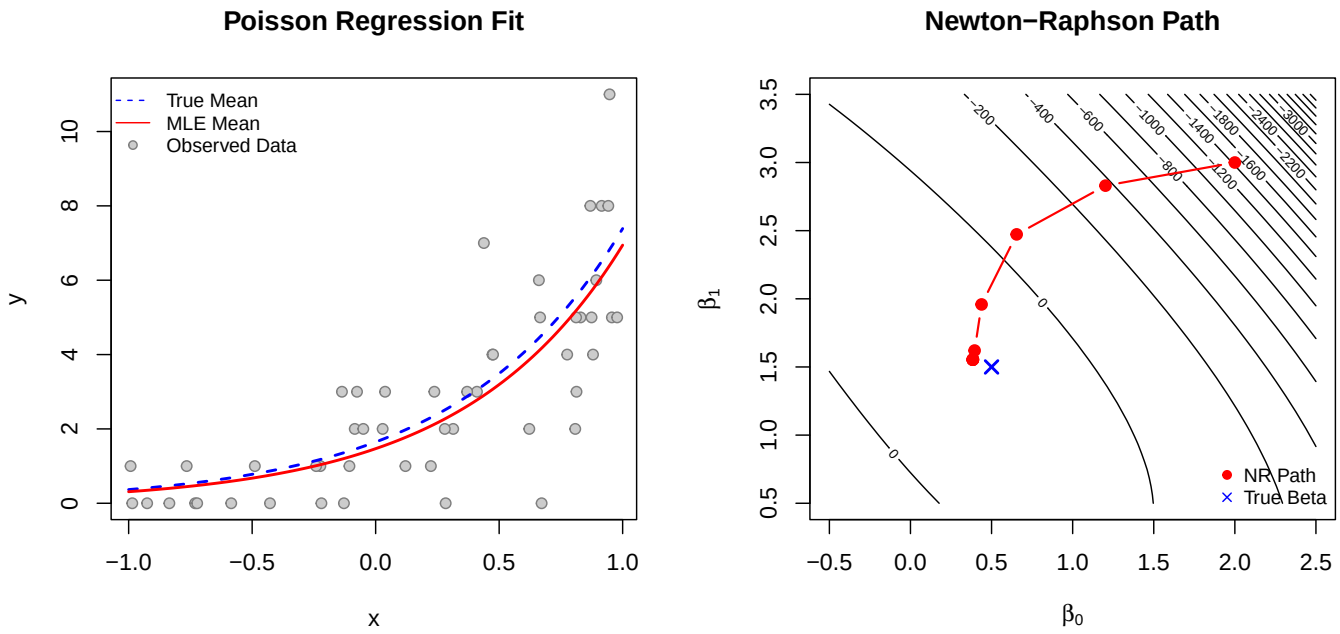


Figure 4.1: Left: Observed data with the true and estimated mean functions; Right: Newton-Raphson path on the log-likelihood surface.

4.3 Convergence Theorems in Probability Theory

4.3.1 Weak Law of Large Numbers (WLLN)

Theorem 4.1 (Weak Law of Large Numbers (WLLN)). *Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with mean $E[X_i] = \mu$ and finite variance $Var(X_i) = \sigma^2 < \infty$.*
Then, the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to μ:

$$\overline{X}_n \xrightarrow{p} \mu \quad \text{as } n \to \infty \tag{4.35}$$

Formal definition: For any $\epsilon > 0$, $\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \epsilon) = 0$.

4.3.2 Central Limit Theorem for IID Cases

Theorem 4.2 (Central Limit Theorem for IID Cases). *Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean $E[X_i] = \mu$ and finite variance $0 < Var(X_i) = \sigma^2 < \infty$.*
Then, the random variable $\sqrt{n}(\overline{X}_n - \mu)$ converges in distribution to a normal distribution with mean 0 and variance σ^2:

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \tag{4.36}$$

4.3.3 Lindeberg-Feller CLT (For Non-Identical Distributions)

Theorem 4.3 (Lindeberg-Feller CLT (For Non-Identical Distributions))**.** *This variation is crucial for regression analysis (e.g., OLS properties with fixed regressors) where variables are independent but* ***not*** *identically distributed.*
Let $X_1, \ldots, X_n$ be independent random variables with $E[X_i] = \mu_i$ and Var$(X_i) = \sigma_i^2$. Define the ***average variance*** *$\tilde{\sigma}_n^2$:*

$$\tilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2 \tag{4.37}$$

If the ***Lindeberg Condition*** *holds: For every $\epsilon > 0$,*

$$\lim_{n\to\infty}\frac{1}{n\tilde{\sigma}_n^2}\sum_{i=1}^{n}E\left[(X_i-\mu_i)^2\cdot I\left(|X_i-\mu_i| > \epsilon\sqrt{n\tilde{\sigma}_n^2}\right)\right] = 0 \tag{4.38}$$

Then the standardized sum converges to a standard normal:

$$\frac{\sum_{i=1}^{n}(X_i-\mu_i)}{\sqrt{n\tilde{\sigma}_n^2}} \xrightarrow{d} N(0,1) \tag{4.39}$$

4.3.4 Approximating Distribution for Sample Mean (Non-i.i.d.)

Corollary 4.1 (Approximating Distribution for Sample Mean (Non-i.i.d.))**.** *Under the conditions of the Lindeberg-Feller CLT, we can derive the asymptotic distribution for the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n}X_i$.*
Let $\overline{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\mu_i$ be the average mean. Note that the denominator in the CLT is simply $\sqrt{n}\tilde{\sigma}_n$.
The standardized sum converges to $N(0,1)$:

$$\frac{\sum(X_i-\mu_i)}{\sqrt{n\tilde{\sigma}_n^2}} = \frac{n(\overline{X}_n-\overline{\mu}_n)}{\sqrt{n}\tilde{\sigma}_n} = \frac{\sqrt{n}(\overline{X}_n-\overline{\mu}_n)}{\tilde{\sigma}_n} \xrightarrow{d} N(0,1) \tag{4.40}$$

This implies the following ***approximate distributions*** *for large n:*

1. ***For the Sample Mean:***

$$\overline{X}_n \dot{\sim} N\left(\overline{\mu}_n, \frac{\tilde{\sigma}_n^2}{n}\right) \tag{4.41}$$

2. ***For the Scaled Difference (Root-n consistency):***

$$\sqrt{n}(\overline{X}_n-\overline{\mu}_n) \dot{\sim} N(0,\tilde{\sigma}_n^2) \tag{4.42}$$

Note: If all X_i share the same mean μ, simply replace $\overline{\mu}_n$ with μ.

4.3.5 Slutsky's Theorem

Theorem 4.4 (Slutsky's Theorem)**.** *Let X_n and Y_n be sequences of random variables. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a constant, then:*

1. ***Sum:*** $X_n + Y_n \xrightarrow{d} X + c$
2. ***Product:*** $X_n Y_n \xrightarrow{d} cX$
3. ***Quotient:*** $X_n / Y_n \xrightarrow{d} X/c$ *(provided $c \neq 0$)*

4.3.6 Generalized Slutsky's Theorem (Continuous Mapping)

Theorem 4.5 (Generalized Slutsky's Theorem (Continuous Mapping))**.** *The arithmetic operations in Slutsky's theorem are special cases of a broader property.*
*Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a constant. Let $g : \mathbb{R}^2 \to \mathbb{R}^k$ be a function that is **continuous** at every point (x, c) where x is in the support of X.*
Then:

$$g(X_n, Y_n) \xrightarrow{d} g(X, c) \tag{4.43}$$

This implies that for any "well-behaved" algebraic combination (polynomials, exponentials, etc.) of a sequence converging in distribution and a sequence converging in probability to a constant, the limit behaves as if the constant were substituted directly.

4.3.7 Delta Methods

Theorem 4.6 (The Univariate Delta Method)**.** *Let X_n be a sequence of random variables that satisfies*

$$\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{4.44}$$

for some constant $\theta \in \mathbb{R}$ and variance $\sigma^2 > 0$. Suppose $g : \mathbb{R} \to \mathbb{R}$ is a function that is continuously differentiable at θ. Then,

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, [g'(\theta)]^2 \sigma^2) \tag{4.45}$$

Proof. By Taylor's theorem, expanding $g(X_n)$ around θ yields:

$$g(X_n) = g(\theta) + g'(\tilde{\theta})(X_n - \theta) \tag{4.46}$$

where $\tilde{\theta}$ lies strictly between X_n and θ.
Since $\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, it follows that $X_n \xrightarrow{p} \theta$. Because $\tilde{\theta}$ is squeezed between X_n and θ, we also

have $\tilde{\theta} \xrightarrow{p} \theta$.
Assuming g' is continuous at θ, the Continuous Mapping Theorem implies $g'(\tilde{\theta}) \xrightarrow{p} g'(\theta)$.
Rearranging the expansion and multiplying by $\sqrt{n}$ gives:

$$\sqrt{n}(g(X_n) - g(\theta)) = g'(\tilde{\theta})\sqrt{n}(X_n - \theta) \tag{4.47}$$

By Slutsky's theorem, the product converges in distribution to $g'(\theta)\mathcal{N}(0, \sigma^2)$, which is distributed as $\mathcal{N}(0, [g'(\theta)]^2\sigma^2)$. □

Theorem 4.7 (The Multivariate Delta Method)**.** *Let $\mathbf{X}_n$ be a sequence of random vectors in $\mathbb{R}^k$ such that:*

$$\sqrt{n}(\mathbf{X}_n - \theta) \xrightarrow{d} \mathcal{N}_k(\mathbf{0}, \Sigma) \tag{4.48}$$

Suppose $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^m$ is a mapping that is continuously differentiable at θ. Let $\nabla\mathbf{g}(\theta)$ be the $m \times k$ Jacobian matrix of partial derivatives of $\mathbf{g}$ evaluated at θ. Then,

$$\sqrt{n}(\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\theta)) \xrightarrow{d} \mathcal{N}_m(\mathbf{0}, \nabla\mathbf{g}(\theta)\Sigma\nabla\mathbf{g}(\theta)^\top) \tag{4.49}$$

4.3.8 Example: Asymptotic Normality of Sample Variance

Example 4.4 (Asymptotic Normality of Sample Variance and Its Logarithm)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. We wish to derive the asymptotic distribution of the maximum likelihood estimator for variance, $\hat{\sigma}^2$, its logarithm, and their corresponding Cumulative Distribution Functions (CDFs). The estimator is defined as:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{4.50}$$

1. **Algebraic Expansion**

   We rewrite the estimator by adding and subtracting the true mean μ:

   $$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}((X_i - \mu) - (\overline{X} - \mu))^2 \tag{4.51}$$

   Expanding the square yields:

   $$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\overline{X} - \mu)^2 \tag{4.52}$$

   Since $\sum_{i=1}^{n}(X_i - \mu) = n(\overline{X} - \mu)$:

   $$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2 \tag{4.53}$$

Dividing by n, we get:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2 \tag{4.54}$$

2. **Scaling by $\sqrt{n}$**

   We rearrange to look at the pivotal quantity $\sqrt{n}(\hat{\sigma}^2 - \sigma^2)$:

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2\right) - \sqrt{n}(\overline{X} - \mu)^2 \tag{4.55}$$

3. **Applying Convergence Theorems for $\hat{\sigma}^2$**

   Let $W_i = (X_i - \mu)^2$. Since $X_i \sim \mathcal{N}(\mu, \sigma^2)$, the scaled variable $W_i/\sigma^2 \sim \chi^2_1$. The moments of W_i are $E[W_i] = \sigma^2$ and $\text{Var}(W_i) = 2\sigma^4$. By the standard Central Limit Theorem:

$$\sqrt{n}(\overline{W} - E[W_i]) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2\right) \xrightarrow{d} \mathcal{N}(0, 2\sigma^4) \tag{4.56}$$

   Consider the term $\sqrt{n}(\overline{X} - \mu)^2$. We can rewrite this as:

$$\underbrace{\sqrt{n}(\overline{X} - \mu)}_{\xrightarrow{d}\mathcal{N}(0,\sigma^2)} \cdot \underbrace{(\overline{X} - \mu)}_{\xrightarrow{p}0 \text{ by WLLN}} \tag{4.57}$$

   By Slutsky's Theorem, the product converges in distribution to 0. Combining the terms, we establish the asymptotic distribution for the sample variance:

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, 2\sigma^4) \tag{4.58}$$

4. **Applying the Delta Method for $\log(\hat{\sigma}^2)$**

   To find the asymptotic distribution of $\log(\hat{\sigma}^2)$, we apply the Univariate Delta Method using the transformation $g(x) = \log(x)$, with derivative $g'(x) = \frac{1}{x}$. Evaluating the derivative at the true parameter value σ^2 gives $g'(\sigma^2) = \frac{1}{\sigma^2}$. By the Delta Method, the asymptotic variance is $[g'(\sigma^2)]^2(2\sigma^4)$. Calculating this gives:

$$\left(\frac{1}{\sigma^2}\right)^2 (2\sigma^4) = 2 \tag{4.59}$$

   Therefore, the asymptotic distribution for the logarithm of the sample variance is:

$$\sqrt{n}(\log(\hat{\sigma}^2) - \log(\sigma^2)) \xrightarrow{d} \mathcal{N}(0, 2) \tag{4.60}$$

5. **Exact CDF based on χ^2**

   The scaled sum of squared deviations follows a Chi-squared distribution with $n-1$ degrees of freedom, $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$. The exact CDF of $\hat{\sigma}^2$ evaluated at x is:

$$P(\hat{\sigma}^2 \leq x) = P\left(\frac{n\hat{\sigma}^2}{\sigma^2} \leq \frac{nx}{\sigma^2}\right) = F_{\chi^2_{n-1}}\left(\frac{nx}{\sigma^2}\right) \tag{4.61}$$

6. **Asymptotic Normal CDF for $\hat{\sigma}^2$**

   Using the asymptotic normal distribution from Step 3, for a finite sample size n, we approximate $\hat{\sigma}^2 \approx \mathcal{N}(\sigma^2, \frac{2\sigma^4}{n})$. The approximate CDF is:

$$P(\hat{\sigma}^2 \leq x) \approx \Phi\left(\frac{x - \sigma^2}{\sqrt{2\sigma^4/n}}\right) \tag{4.62}$$

7. **Asymptotic Normal CDF based on $\log(\hat{\sigma}^2)$**

   Using the asymptotic distribution from Step 4, for a finite sample size n, we approximate $\log(\hat{\sigma}^2) \approx \mathcal{N}(\log(\sigma^2), \frac{2}{n})$. To find the CDF for $\hat{\sigma}^2$ itself evaluated at $x > 0$:

$$P(\hat{\sigma}^2 \leq x) = P(\log(\hat{\sigma}^2) \leq \log(x)) \approx \Phi\left(\frac{\log(x) - \log(\sigma^2)}{\sqrt{2/n}}\right) \tag{4.63}$$

To provide a numerical illustration of these derivations, the following R code overlays the three CDFs. Note that this floating figure chunk is placed outside the example division to satisfy formatting constraints.

```r
# Set up plotting area for two plots side by side (1 row, 2 columns)
par(mfrow = c(1, 2))

# Create a function to avoid repeating the plotting code
plot_cdfs <- function(n, sigma2 = 1) {
  # Sequence of x values for the plot
  x <- seq(0.01, 3, length.out = 300)

  # 1. Exact Chi-squared CDF
  cdf_exact <- pchisq(n * x / sigma2, df = n - 1)

  # 2. Asymptotic Normal CDF
  sd_norm <- sqrt(2 * sigma2^2 / n)
  cdf_norm <- pnorm(x, mean = sigma2, sd = sd_norm)

  # 3. Log-Normal Approximation CDF
  sd_log <- sqrt(2 / n)
  cdf_log <- pnorm(log(x), mean = log(sigma2), sd = sd_log)
```

```r
  # Plotting the CDFs
  plot(x, cdf_exact, type = "l", col = "black", lwd = 2,
       ylab = "Cumulative Probability",
       xlab = expression(hat(sigma)^2),
       main = bquote(paste("CDFs of ", hat(sigma)^2, " (n = ", .(n), ", ", sigma^2, " = 1)")),
       ylim = c(0, 1))

  # Add the normal approximation curve
  lines(x, cdf_norm, col = "red", lwd = 2, lty = 2)

  # Add the log-normal approximation curve
  lines(x, cdf_log, col = "blue", lwd = 2, lty = 3)

  # Add a legend
  legend("bottomright",
         legend = c("Exact (Chi-squared)",
                    "Normal Approx",
                    "Log-Normal Approx"),
         col = c("black", "red", "blue"),
         lwd = 2,
         lty = c(1, 2, 3),
         bty = "n",
         cex = 0.8) # Slightly smaller text to fit well in side-by-side view
}

# Generate the plot for n = 10
plot_cdfs(n = 10)

# Generate the plot for n = 50
plot_cdfs(n = 50)

# Reset plotting parameters to default
par(mfrow = c(1, 1))
```

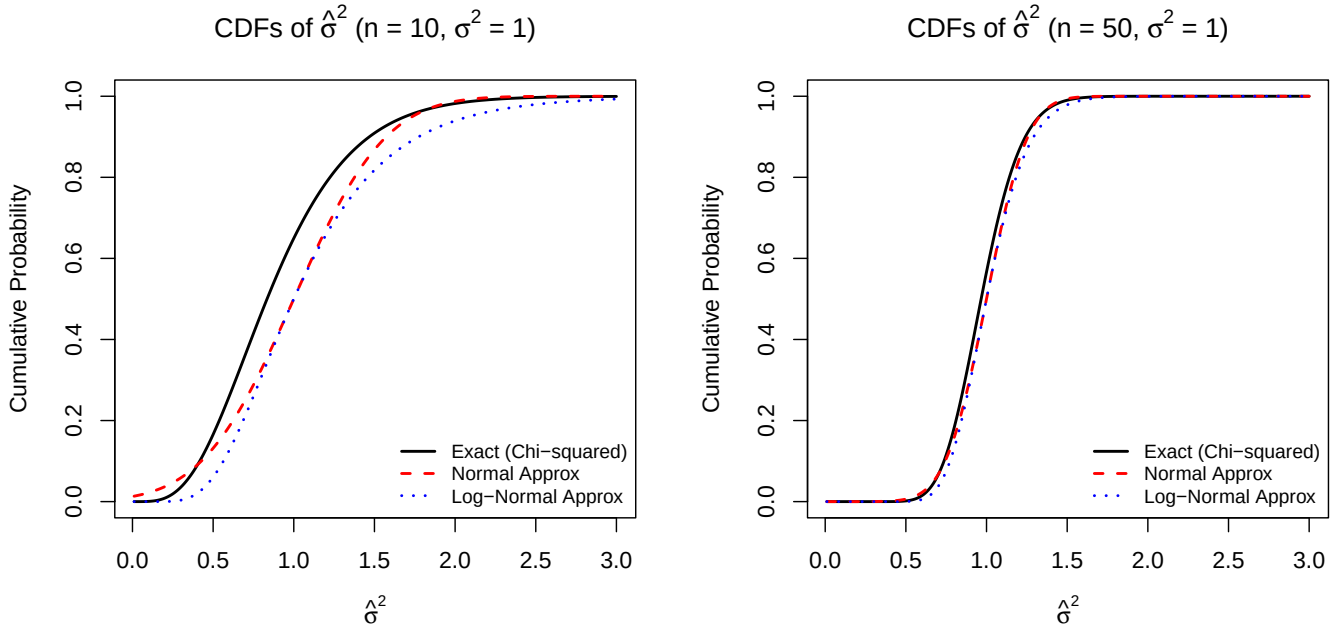


Figure 4.2: Comparison of the Exact, Asymptotic Normal, and Log-Normal Approximated CDFs of the sample variance for n = 10 and n = 50.

4.4 Asymptotic Theory of Maximized Likelihood

4.4.1 Consistency of the MLE

4.4.1.1 IID Cases

Consistency establishes that as the sample size grows, the estimator converges in probability to the true parameter value.

Lemma 4.1 (Minimization of KL Divergence at the True Density)**.** *Let $f(y|\theta_0)$ be the true density of a random variable Y, and let $f(y|\theta)$ be any other density in the family. The **Kullback-Leibler (KL) Divergence** (or relative entropy) between the true distribution and a candidate distribution is defined as:*

$$D_{KL}(f(\cdot|\theta_0)\|f(\cdot|\theta)) = E_{\theta_0}\left[\log\left(\frac{f(Y|\theta_0)}{f(Y|\theta)}\right)\right] \tag{4.64}$$

*This quantity satisfies the **Information Inequality**:*

$$D_{KL}(f(\cdot|\theta_0)\|f(\cdot|\theta)) \geq 0 \tag{4.65}$$

with equality if and only if $f(y|\theta) = f(y|\theta_0)$ almost everywhere. This implies that the expected log-likelihood $M(\theta) = E_{\theta_0}[\log f(Y|\theta)]$ is uniquely maximized at the true parameter θ_0 (or true density for Y).

Proof.

1. **Information Inequality via Jensen's Inequality**

   Consider the negative KL divergence, which is the expected log-likelihood ratio. By **Jensen's Inequality**, since the logarithm is a strictly concave function:

$$-D_{KL}(f(\cdot|\theta_0)\|f(\cdot|\theta)) = E_{\theta_0}\left[\log\left(\frac{f(Y|\theta)}{f(Y|\theta_0)}\right)\right] \leq \log E_{\theta_0}\left[\frac{f(Y|\theta)}{f(Y|\theta_0)}\right] \tag{4.66}$$

2. **Evaluating the Expectation**

   We evaluate the expectation on the right-hand side:

$$E_{\theta_0}\left[\frac{f(Y|\theta)}{f(Y|\theta_0)}\right] = \int_y \left(\frac{f(y|\theta)}{f(y|\theta_0)}\right) f(y|\theta_0)\, dy = \int_y f(y|\theta)\, dy = 1 \tag{4.67}$$

3. **Conclusion**

   Substituting the result into the inequality:

$$-D_{KL}(f(\cdot|\theta_0)\|f(\cdot|\theta)) \leq \log(1) = 0 \implies D_{KL}(f(\cdot|\theta_0)\|f(\cdot|\theta)) \geq 0 \tag{4.68}$$

   Because the log function is strictly concave, the equality $E[\log Z] = \log E[Z]$ holds if and only if Z is a constant with probability 1. Here, $Z = f(Y|\theta)/f(Y|\theta_0)$, so equality holds if and only if $f(y|\theta) = f(y|\theta_0)$ for almost all y.

□

Theorem 4.8 (Consistency of MLE)**.** *Let $Y_1, \ldots, Y_n$ be i.i.d. with density $f(y|\theta)$. Let θ_0 be the true parameter. Assume the following **Identifiability** condition holds:*

$$f(y|\theta) = f(y|\theta_0) \text{ for almost all } y \implies \theta = \theta_0 \tag{4.69}$$

Under this and other standard regularity conditions (such as compactness of the parameter space Θ), the Maximum Likelihood Estimator $\hat{\theta}_n$ satisfies:

$$\hat{\theta}_n \xrightarrow{p} \theta_0 \quad \text{as } n \to \infty \tag{4.70}$$

Remark 4.1 (The Role of Identifiability). Identifiability is a fundamental requirement for consistency because it ensures that different parameter values correspond to different probability distributions.

1. **Uniqueness of the Mapping**

   Mathematically, it means the mapping $\theta \to f(\cdot|\theta)$ is one-to-one. If the model were not identifiable, there would exist at least one $\theta_1 \neq \theta_0$ such that $f(y|\theta_1) = f(y|\theta_0)$.

2. **Uniqueness of the Maximum**

In the context of the proof, Lemma 4.1 shows that the KL divergence is zero if and only if the densities are identical. Identifiability allows us to move from the space of "densities" back to the space of "parameters," ensuring that θ_0 is the **unique** global maximum of the expected log-likelihood function $M(\theta)$.

3. **Failure of Consistency**

   Without identifiability, the data cannot distinguish between θ_0 and θ_1. The MLE would not converge to a single point but rather to the set of all parameters that produce the true density.

Proof.

1. **The Objective Function**

   The MLE maximizes the average log-likelihood:

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i|\theta) \tag{4.71}$$

   By the **Weak Law of Large Numbers (WLLN)**, for any fixed θ, $M_n(\theta)$ converges in probability to its expectation:

$$M_n(\theta) \xrightarrow{p} M(\theta) = E_{\theta_0}[\log f(Y|\theta)] \tag{4.72}$$

2. **Convergence to the Unique Maximum**

   According to Lemma 4.1, the limit function $M(\theta)$ is maximized when the KL divergence is zero. This happens uniquely at θ_0 due to identifiability.

   Since $\hat{\theta}_n$ is the maximizer of $M_n(\theta)$, and $M_n(\theta)$ converges uniformly to $M(\theta)$, the sequence of maximizers must converge to the maximizer of the limit function:

$$\hat{\theta}_n \xrightarrow{p} \theta_0 \tag{4.73}$$

□

4.4.1.2 Non-IID Cases

In regression settings, observations are independent but not identically distributed (i.n.i.d.). We generalize the consistency result by requiring that the "average" information accumulates sufficiently.

Theorem 4.9 (Consistency of MLE (General Case))**.** *Let $Y_1, \ldots, Y_n$ be independent observations with densities $f_i(y|\theta)$ (e.g., depending on covariates x_i). Let θ_0 be the true parameter.*
Under the following conditions:

1. ***Parameter Space:*** *Compact parameter space Θ.*
2. ***Identification:*** *For any $\theta \neq \theta_0$, the average Kullback-Leibler divergence is strictly positive in the limit:*

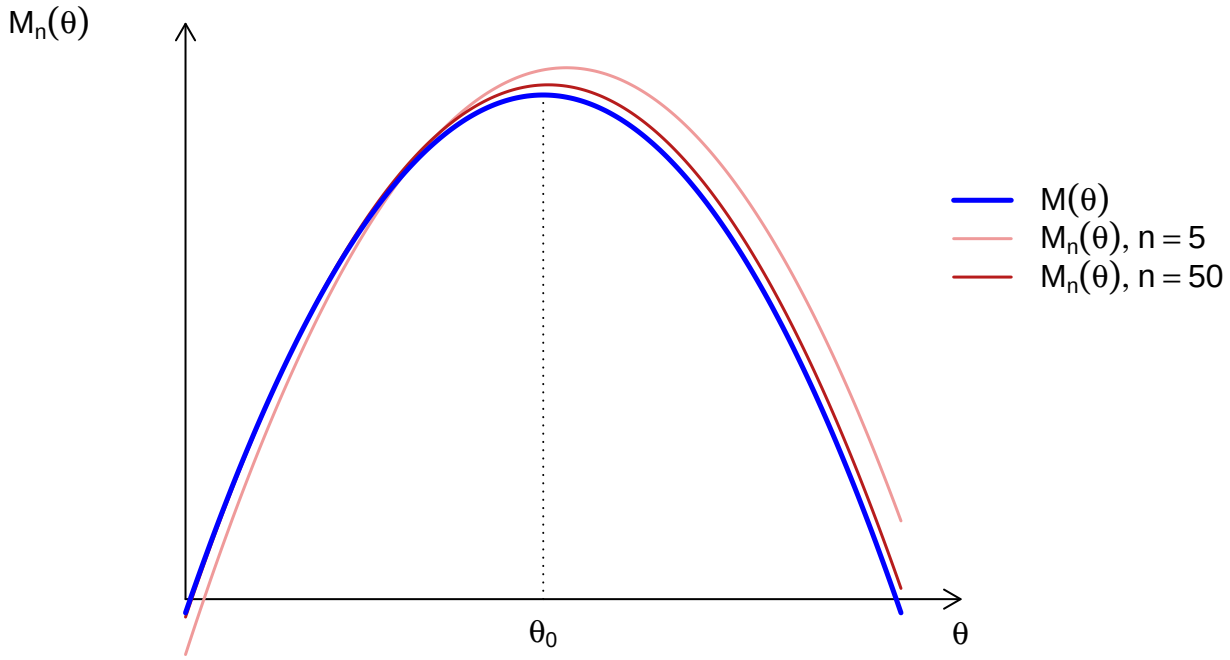


Figure 4.3: Consistency: Concentration of the Average Log-Likelihood for $N(\theta, 1)$

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} KL(f_i(\cdot|\theta_0)||f_i(\cdot|\theta)) > 0 \tag{4.74}$$

3. ***Uniform Convergence:*** *The log-likelihood satisfies a Uniform Law of Large Numbers (ULLN).*

Then $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Proof.

1. **The Objective Function**

   The MLE maximizes the average log-likelihood:

   $$M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f_i(Y_i|\theta) \tag{4.75}$$

   Unlike the i.i.d. case, the terms in the sum have different expectations. We rely on a **WLLN for Independent Variables** (e.g., Chebyshev's WLLN). Provided the variances are bounded, $M_n(\theta)$ converges in probability to the **average expectation**:

   $$M_n(\theta) \xrightarrow{p} \overline{M}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} E_{\theta_0}[\log f_i(Y_i|\theta)] \tag{4.76}$$

   Note: For consistency of the maximizer, we strictly require ***Uniform Convergence*** *over* Θ*, not just pointwise convergence.*

2. **Identifying the Maximum (Average KL Divergence)**

   We compare the limit function at θ versus θ_0. Consider the average difference:

$$\overline{M}_n(\theta) - \overline{M}_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} E_{\theta_0}\left[\log\left(\frac{f_i(Y_i|\theta)}{f_i(Y_i|\theta_0)}\right)\right] \tag{4.77}$$

   By Jensen's Inequality applied to *each* term in the sum:

$$E_{\theta_0}\left[\log\frac{f_i(Y_i|\theta)}{f_i(Y_i|\theta_0)}\right] \leq \log E_{\theta_0}\left[\frac{f_i(Y_i|\theta)}{f_i(Y_i|\theta_0)}\right] = \log(1) = 0 \tag{4.78}$$

   Summing these inequalities implies that $\overline{M}_n(\theta)$ is uniquely maximized at θ_0 (provided the identification condition holds). Since the sample function $M_n(\theta)$ converges uniformly to this maximized limit function, the argmax $\widehat{\theta}_n$ converges to θ_0.

□

4.4.2 Asymptotic Normality of the Score Vector

The Score function acts as the "engine" for the normality of the MLE. We treat the I.I.D. and non-I.I.D. cases separately.

4.4.2.1 IID Cases

Theorem 4.10 (Normality of Score (I.I.D.))**.** *Let $Y_1, \ldots, Y_n$ be i.i.d. with density $f(y|\theta)$. Define the **Fisher Information matrix** for a single observation as the expected outer product of the score:*

$$\mathcal{J}_1(\theta) = E\left[\left(\nabla_\theta \log f(Y_1|\theta)\right)\left(\nabla_\theta \log f(Y_1|\theta)\right)^T\right] \tag{4.79}$$

Then, the scaled total score vector converges to a Normal distribution:

$$\frac{1}{\sqrt{n}}\mathbf{U}_n(\theta_0;\mathbf{Y}) \xrightarrow{d} N(\mathbf{0}, \mathcal{J}_1(\theta_0)) \tag{4.80}$$

Proof. The total score is the sum of independent score contributions:

$$\mathbf{U}_n(\theta_0;\mathbf{Y}) = \sum_{i=1}^{n} \nabla \log f(Y_i|\theta_0) = \sum_{i=1}^{n} \mathbf{u}_i \tag{4.81}$$

From **Bartlett's Identities**, for each observation i:

1. **Mean**

   $E[\mathbf{u}_i] = \mathbf{0}$.

2. **Variance**

$$\text{Var}(\mathbf{u}_i) = E[\mathbf{u}_i \mathbf{u}_i^T] = \mathcal{I}_1(\theta_0).$$

Since the terms $\mathbf{u}_i$ are i.i.d. with finite variance, we apply the **Multivariate Central Limit Theorem (Lindeberg-Lévy)**:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{u}_i \xrightarrow{d} N(\mathbf{0}, \text{Var}(\mathbf{u}_i)) = N(\mathbf{0}, \mathcal{I}_1(\theta_0)) \tag{4.82}$$

□

4.4.2.2 Case B: The Non-I.I.D. Case (Regression Setting)

Theorem 4.11 (Normality of Score (Independent, Non-Identical))**.** *Let $Y_1, \dots, Y_n$ be independent but not necessarily identically distributed (e.g., due to different covariates). Let $\mathcal{I}_n(\theta_0) = \sum_{i=1}^{n} Var(\mathbf{u}_i)$ be the total information.*
*Define the **Average Fisher Information Matrix**:*

$$\overline{\mathcal{I}}_n(\theta_0) = \frac{1}{n}\mathcal{I}_n(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} E[\mathbf{u}_i \mathbf{u}_i^T] \tag{4.83}$$

*If the **Lindeberg Condition** is satisfied for the sequence of score vectors, then the standardized score converges to a standard Normal:*

$$(\mathcal{I}_n(\theta_0))^{-1/2} \mathbf{U}_n(\theta_0; \mathbf{Y}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p) \tag{4.84}$$

(Here $\mathbf{I}_p$ denotes the Identity matrix).
***Approximating Distribution:** If the average information converges to a positive definite limit $\overline{\mathcal{I}}_n \to \mathcal{I}_\infty$, we have the following approximation for the scaled average score $\sqrt{n}\overline{\mathbf{U}}_n(\theta_0; \mathbf{Y})$:*

$$\sqrt{n}\overline{\mathbf{U}}_n(\theta_0; \mathbf{Y}) = \frac{1}{\sqrt{n}}\mathbf{U}_n(\theta_0; \mathbf{Y}) \dot{\sim} N(\mathbf{0}, \overline{\mathcal{I}}_n(\theta_0)) \tag{4.85}$$

Proof. This is a direct application of the **Lindeberg-Feller Central Limit Theorem**. The total score $\mathbf{U}_n(\theta_0; \mathbf{Y}) = \sum \mathbf{u}_i$ is a sum of independent random vectors with mean $\mathbf{0}$ and variances $\text{Var}(\mathbf{u}_i)$. Provided that no single observation's score contribution dominates the sum (the Lindeberg condition), the standardized sum converges to a standard Normal distribution. □

4.4.3 Asymptotic Normality of the MLE

We now transfer the normality from the Score function to the estimator $\hat{\theta}$ using a Taylor expansion (the Delta Method logic).

Definition 4.5 (Regularity Conditions for Asymptotic Normality)**.** The following conditions are required to ensure the validity of the Taylor expansion and the convergence of the remainder term:

1. **Interiority:** The true parameter θ_0 lies in the interior of the parameter space Θ.

2. **Smoothness:** The log-likelihood function $\log f(y|\theta)$ is three times continuously differentiable with respect to θ.
3. **Boundedness:** The third derivatives are bounded by an integrable function $M(y)$ (to control the Taylor remainder).
4. **Positive Information:** The Fisher Information Matrix $\mathcal{I}_1(\theta_0)$ exists and is non-singular (positive definite).
5. **Interchangeability:** Differentiation and integration can be interchanged for the density $f(y|\theta)$.

Theorem 4.12 (Asymptotic Normality of MLE)**.** *Under the regularity conditions above, the MLE $\hat{\theta}_n$ satisfies:*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(\mathbf{0}, [\mathcal{I}_1(\theta_0)]^{-1}) \tag{4.86}$$

Proof.

1. **Taylor Expansion of Score Equation**

   We expand the Score function $\mathbf{U}_n(\theta)$ around the true parameter θ_0. Since $\hat{\theta}_n$ is the MLE, $\mathbf{U}_n(\hat{\theta}_n) = \mathbf{0}$. Define the **Observed Information Matrix** as the negative Hessian:

   $$\mathbf{J}_n(\theta) = -\nabla^2 \ell_n(\theta) \tag{4.87}$$

   The fundamental approximation linking the random Score vector to the estimation error is:

   $$\underbrace{\mathbf{U}_n(\hat{\theta}_n)}_{=\mathbf{0}} \approx \mathbf{U}_n(\theta_0) - \mathbf{J}_n(\theta_0)(\hat{\theta}_n - \theta_0) \tag{4.88}$$

   $$\implies \quad \mathbf{U}_n(\theta_0) \approx \mathbf{J}_n(\theta_0)(\hat{\theta}_n - \theta_0) \tag{4.89}$$

2. **Scaling and Inversion**

   Multiply the link equation by $\frac{1}{\sqrt{n}}$ and introduce n to the Observed Information term to stabilize the limits:

   $$\frac{1}{\sqrt{n}}\mathbf{U}_n(\theta_0) \approx \left(\frac{1}{n}\mathbf{J}_n(\theta_0)\right) \sqrt{n}(\hat{\theta}_n - \theta_0) \tag{4.90}$$

   Rearranging to solve for the estimator's distribution:

   $$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx \left[\frac{1}{n}\mathbf{J}_n(\theta_0)\right]^{-1} \left[\frac{1}{\sqrt{n}}\mathbf{U}_n(\theta_0)\right] \tag{4.91}$$

3. **Convergence of Components**

   - **Numerator (Score):** From the I.I.D. Score Normality theorem:

   $$\frac{1}{\sqrt{n}}\mathbf{U}_n(\theta_0) \xrightarrow{d} Z \sim N(\mathbf{0}, \mathcal{I}_1(\theta_0)) \tag{4.92}$$

- **Denominator (Observed Info):** By the **WLLN**, the average observed information converges to the expected information:

$$\frac{1}{n}\mathbf{J}_n(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n} \nabla^2 \log f(Y_i|\theta_0) \xrightarrow{p} -E[\nabla^2 \log f] = \mathcal{I}_1(\theta_0) \tag{4.93}$$

4. **Slutsky's Theorem**

   Combining these via Slutsky's Theorem (matrix inverse is a continuous mapping):

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} [\mathcal{I}_1(\theta_0)]^{-1} Z \tag{4.94}$$

   The variance of this limiting distribution is:

$$\text{Var}([\mathcal{I}_1]^{-1} Z) = \mathcal{I}_1^{-1}\text{Var}(Z)(\mathcal{I}_1^{-1})^T = \mathcal{I}_1^{-1}\mathcal{I}_1\mathcal{I}_1^{-1} = \mathcal{I}_1^{-1} \tag{4.95}$$

   Thus:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(\mathbf{0}, [\mathcal{I}_1(\theta_0)]^{-1}) \tag{4.96}$$

□

4.4.4 Wilks' Theorem

4.4.4.1 Wilks' Theorem (Simple Null Hypothesis)

Theorem 4.13 (Wilks' Theorem (Simple Null Hypothesis))**.** *Consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Define the **Likelihood Ratio** Λ_n as:*

$$\Lambda_n = \frac{L(\theta_0)}{L(\hat{\theta}_{MLE})} \tag{4.97}$$

*Define the **Deviance Statistic** D_n as:*

$$D_n = -2\log \Lambda_n = 2[\ell_n(\hat{\theta}) - \ell_n(\theta_0)] \tag{4.98}$$

Under the null hypothesis, the Deviance converges to a Chi-squared distribution:

$$D_n \xrightarrow{d} \chi^2_p \tag{4.99}$$

where p is the dimension of θ.

Proof. We express the likelihood difference $\Delta\ell$ in terms of the Score vector $\mathbf{U}_n(\theta_0)$ and apply the Central Limit Theorem directly to the Score.

1. **Quadratic Approximation of Deviance**

   Expand $\ell_n(\theta_0)$ around the MLE $\hat{\theta}$. Since the gradient at the MLE is zero ($\mathbf{U}_n(\hat{\theta}) = \mathbf{0}$), the first-order term

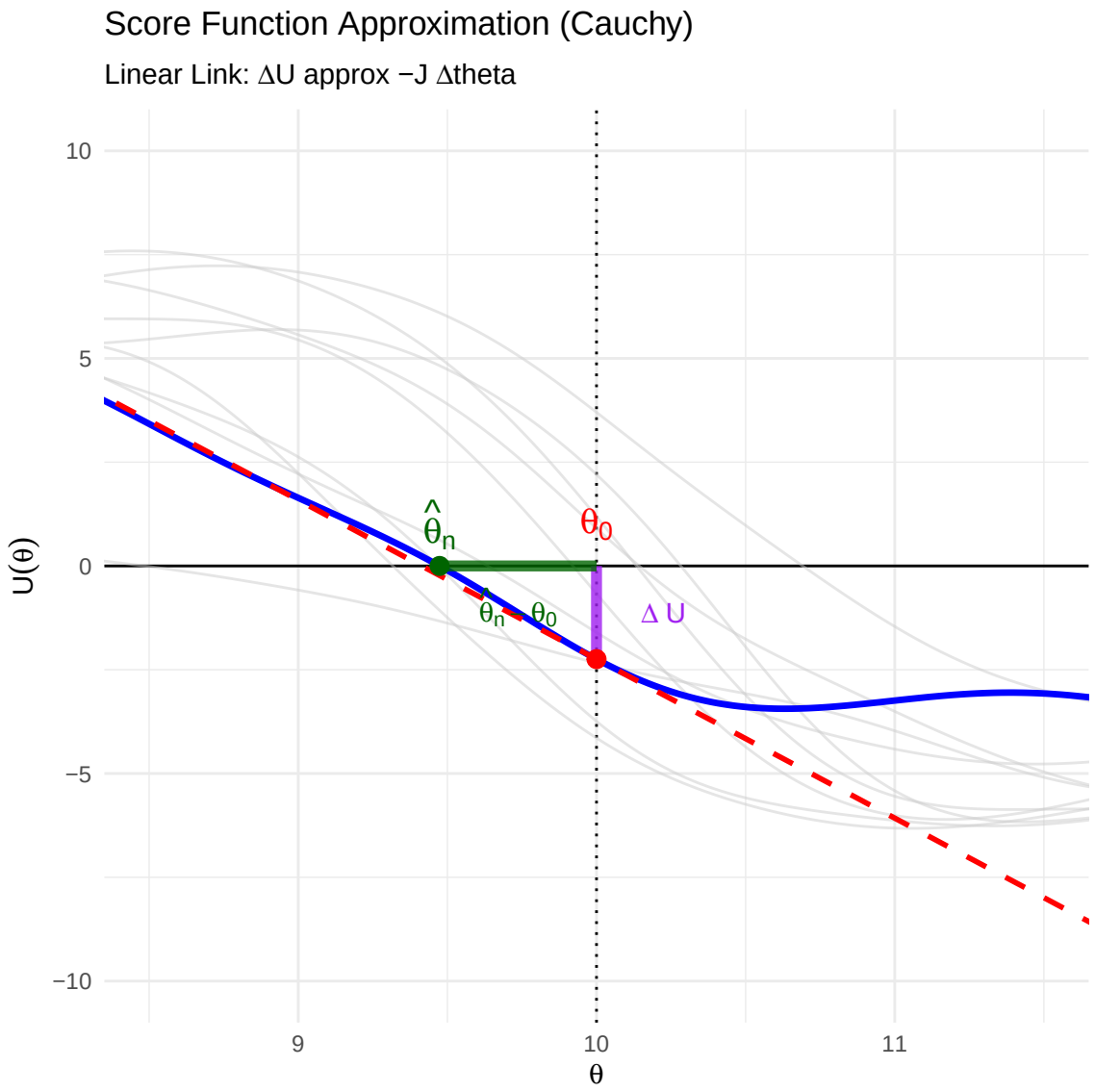


Figure 4.4: Linear Approximation of the Score Function: The vertical purple segment represents the Score at the true parameter (ΔU), while the horizontal green segment shows the estimation error ($\hat{\theta}_n - \theta_0$). The red dashed line indicates the linear approximation ($-J(\theta_0)$).

vanishes:

$$\ell_n(\theta_0) \approx \ell_n(\hat{\theta}) - \frac{1}{2}(\hat{\theta} - \theta_0)^T \mathbf{J}_n(\hat{\theta})(\hat{\theta} - \theta_0) \tag{4.100}$$

Rearranging for the Deviance (D_n):

$$D_n = 2[\ell_n(\hat{\theta}) - \ell_n(\theta_0)] \approx (\hat{\theta} - \theta_0)^T \mathbf{J}_n(\hat{\theta})(\hat{\theta} - \theta_0) \tag{4.101}$$

2. **Substituting the Score**

   Recall the linear link we established in the Normality proof: $\mathbf{U}_n(\theta_0) \approx \mathbf{J}_n(\theta_0)(\hat{\theta} - \theta_0)$. Inverting this relationship gives the parameter error in terms of the Score:

$$(\hat{\theta} - \theta_0) \approx \mathbf{J}_n^{-1} \mathbf{U}_n(\theta_0) \tag{4.102}$$

   Substitute this back into the Deviance equation:

$$D_n \approx (\mathbf{J}_n^{-1} \mathbf{U}_n)^T \mathbf{J}_n (\mathbf{J}_n^{-1} \mathbf{U}_n) \tag{4.103}$$

$$\boxed{D_n \approx \mathbf{U}_n(\theta_0)^T \mathbf{J}_n^{-1} \mathbf{U}_n(\theta_0)} \tag{4.104}$$

   This form (the Score Statistic) shows that the deviance is essentially the squared length of the Score vector, standardized by the Information.

3. **Asymptotic Convergence**

   We rewrite the expression using normalized quantities to apply the limit theorems.

   - **Score:** $\frac{1}{\sqrt{n}} \mathbf{U}_n(\theta_0) \xrightarrow{d} \mathbf{Z} \sim N(\mathbf{0}, \mathcal{I}_1)$.
   - **Information:** $\frac{1}{n} \mathbf{J}_n \xrightarrow{p} \mathcal{I}_1$.

$$D_n \approx \left(\frac{1}{\sqrt{n}} \mathbf{U}_n \right)^T \left(\frac{1}{n} \mathbf{J}_n \right)^{-1} \left(\frac{1}{\sqrt{n}} \mathbf{U}_n \right) \tag{4.105}$$

   Taking the limit:

$$D_n \xrightarrow{d} \mathbf{Z}^T \mathcal{I}_1^{-1} \mathbf{Z} \tag{4.106}$$

   Since $\mathbf{Z} \sim N(\mathbf{0}, \mathcal{I}_1)$, the quadratic form follows a Chi-squared distribution:

$$\mathbf{Z}^T \mathcal{I}_1^{-1} \mathbf{Z} \sim \chi_p^2 \tag{4.107}$$

□

4.4.5 Summary of Asymptotic Approximations

In the limit as $n \to \infty$ (or exactly for the Normal distribution), the log-likelihood behaves like a quadratic function. This leads to the following relationships between the Deviance (D), Score ($\mathbf{U}$), Error ($\Delta\theta$), and Information ($\mathcal{J}$).

Relationship	Formula	Intuition (Geometry of Parabola)
Deviance vs. Error	$D \approx \Delta\theta^T \mathcal{J} \Delta\theta$	The vertical drop (D) is proportional to the squared horizontal distance ($\Delta\theta^2$), scaled by curvature ($\mathcal{J}$). This is the **Wald Statistic**.
Score vs. Error	$\mathbf{U} \approx \mathcal{J}\Delta\theta$	The slope ($\mathbf{U}$) increases linearly with horizontal distance ($\Delta\theta$). The rate of increase is the curvature ($\mathcal{J}$).
Deviance vs. Score	$D \approx \mathbf{U}^T \mathcal{J}^{-1} \mathbf{U}$	The vertical drop (D) is proportional to the squared slope ($\mathbf{U}^2$), *divided* by curvature ($\mathcal{J}$). This is the **Score Statistic**.
Information Definition	$\mathcal{J} \approx -\nabla \mathbf{U}$	The curvature ($\mathcal{J}$) is the rate of change of the slope ($\mathbf{U}$) (negative Hessian).
Information Approximation	$\mathcal{J} \approx \sum_{i=1}^{n} \mathbf{U}_i \mathbf{U}_i^T$	The curvature ($\mathcal{J}$) is estimated by the "Sum of Squared Gradients" (Outer Products). This relies on Bartlett's 2nd Identity: $E[\mathcal{J}] = \text{Var}(\mathbf{U})$.

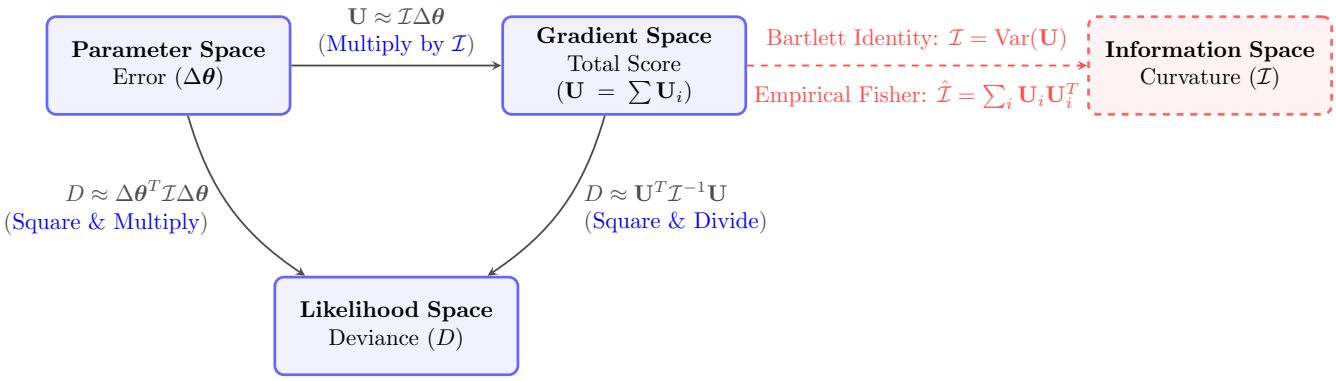


Figure 4.5: Asymptotic Relationships: The Information matrix $\mathcal{J}$ bridges the spaces. Note the use of the empirical sum of outer products $\sum \mathbf{U}_i \mathbf{U}_i^T$ to estimate curvature.

4.4.6 Asympotic Distributions of MLE of Normal Sample

Example 4.5 (Example: Normal Distribution with Unknown Variance ($N(\mu, \sigma^2)$)). Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The parameter vector is $\theta = (\mu, \sigma^2)^T$. We assume both parameters are unknown.

1. Derivation of Key Quantities

- **Log-Likelihood:**

$$\ell(\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{4.108}$$

- **Score Vector $\mathbf{U}(\theta)$:** Gradient w.r.t μ and σ^2:

$$\mathbf{U}(\theta) = \begin{pmatrix} \frac{\partial \ell}{\partial \mu} \\ \frac{\partial \ell}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2}\sum(x_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(x_i - \mu)^2 \end{pmatrix} \tag{4.109}$$

- **Maximum Likelihood Estimator ($\hat{\theta}$):** Solving $\mathbf{U}(\hat{\theta}) = \mathbf{0}$:

$$\hat{\mu} = \overline{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{4.110}$$

- **Fisher Information Matrix $\mathcal{J}_n(\theta)$:** The matrix of negative expected second derivatives (note that the cross-terms are zero, implying orthogonality of μ and σ^2):

$$\mathcal{J}_n(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \tag{4.111}$$

2. Asymptotic Properties

Based on the general theory, we state the asymptotic distributions for the MLE, Score, and Deviance (testing the full vector $H_0 : \theta = \theta_0$).

1. **Estimator Normality:**

$$\hat{\theta} \overset{a}{\sim} N_2\left(\theta, \mathcal{J}_n^{-1}\right) = N_2\left(\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}\right) \tag{4.112}$$

2. **Score Normality:**

$$\mathbf{U}(\theta) \overset{a}{\sim} N_2\left(\mathbf{0}, \mathcal{J}_n\right) = N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}\right) \tag{4.113}$$

3. **Deviance (Wilks' Theorem):**
   For testing $H_0 : (\mu, \sigma^2) = (\mu_0, \sigma_0^2)$ vs H_1 : unrestricted:

$$D = 2[\ell(\hat{\theta}) - \ell(\theta_0)] \xrightarrow{d} \chi_2^2 \tag{4.114}$$

4. Analytical Verification

We check if these asymptotic approximations match the exact finite-sample properties.

- **For the Mean (μ):**

- **Score:** $U_\mu = \frac{n}{\sigma^2}(\overline{X} - \mu)$. Since $\overline{X}$ is Normal, U_μ is **exactly Normal** for any n.
- **MLE:** $\hat{\mu} = \overline{X}$ is **exactly Normal** for any n.

- **For the Variance (σ^2):**
  - **Score:** $U_{\sigma^2} = \frac{1}{2\sigma^4}\left(\sum(X_i - \mu)^2 - n\sigma^2\right)$. The term $\sum(X_i - \mu)^2$ follows a scaled Chi-squared distribution ($\sigma^2\chi^2_n$).
  - **Verification:** The Score is **not exactly Normal** (it is a shifted Gamma/Chi-squared). However, by the Central Limit Theorem, as $n \to \infty$, a Chi-squared random variable converges to Normality. Thus, the property holds **asymptotically**.

- **For the Deviance (D):**
  - The Deviance involves sums of squared errors and logarithms of sums of squares. It is **not exactly Chi-squared** for finite n (unlike the known-variance case). It converges to χ^2_2 as n increases.

4. Simulation Verification

We simulate 5000 datasets with $n = 100$. We calculate the standardized MLEs, standardized Scores, and the Deviance to check convergence.

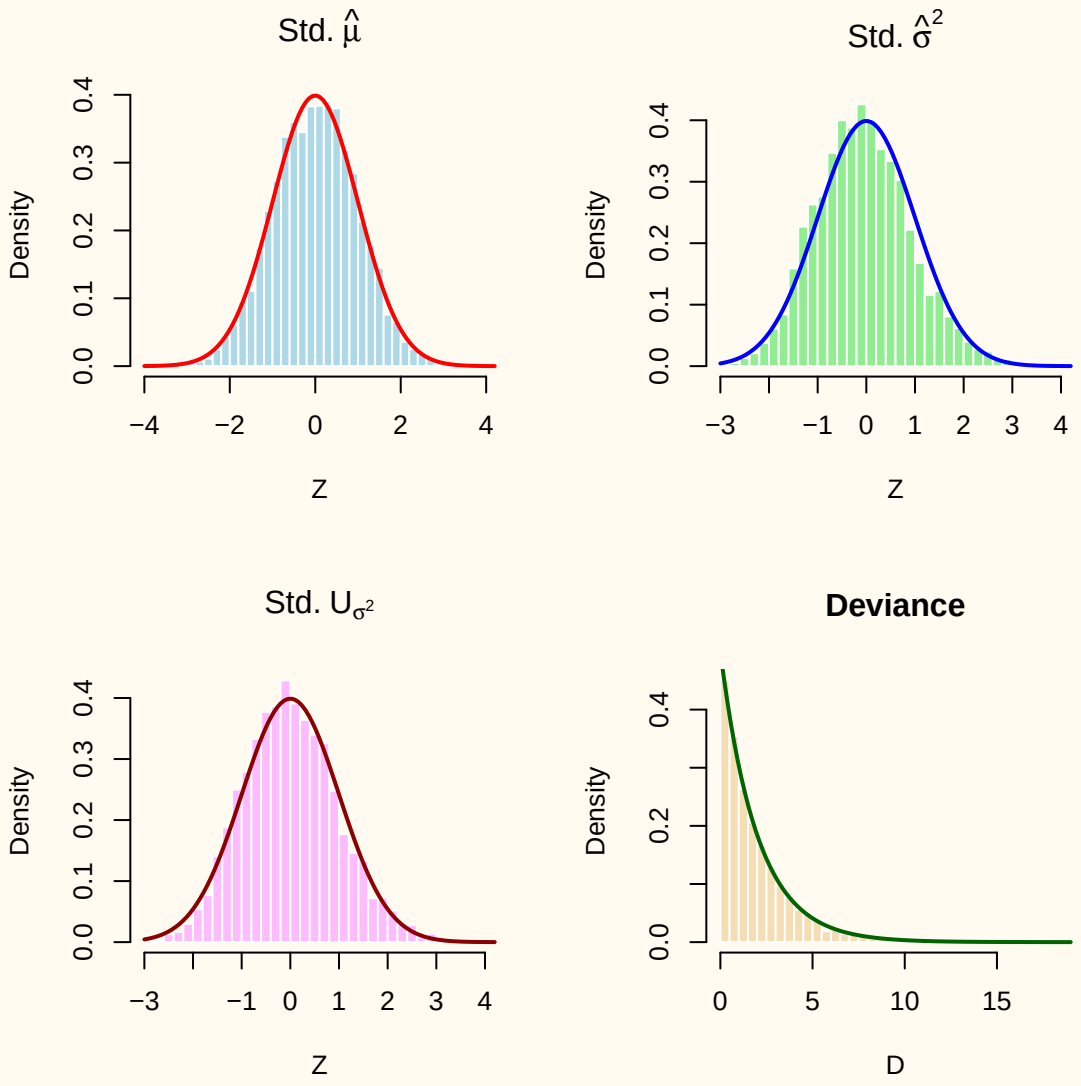


Figure 4.6: Simulation of Normal (Unknown Variance) Asymptotics (n=100). Left/Center-Left: Standardized MLEs. Center-Right: Standardized Score for Variance. Right: Deviance (2 df).

4.4.7 Asymptotic Distributions of Poisson Regression

Example 4.6 (Simulation Study: Poisson Regression Asymptotics)**.** In this example, we verify the asymptotic normality of the Maximum Likelihood Estimator (MLE) and the Chi-squared distribution of the Deviance statistic using a simulation study.

1. Simulation Setup

We consider a Generalized Linear Model (GLM) with a Poisson response and a canonical log-link function:

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \log(\lambda_i) = \beta_0 + \beta_1 x_i \tag{4.115}$$

- **True Parameters:** We set the true coefficient vector to $\beta = (0.5, 1.5)^\top$.
- **Covariates:** We generate $n = 200$ covariate values x_i from a Uniform(0, 1) distribution. These remain fixed across all simulations.
- **Data Generation:** For each simulation iteration, we generate response vector $\mathbf{y}$ using the true mean $\lambda_i = \exp(0.5 + 1.5x_i)$.

2. Model Fitting and Statistics

For each of the `n_sim = 2000` generated datasets, we perform the following:

1. **Fit the Model:** We estimate $\hat{\beta}$ using the `glm()` function in R.
2. **Calculate Standardized MLEs (Wald Statistic):**

$$Z_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_{j,\text{true}}}{\text{SE}(\hat{\beta}_j)} \tag{4.116}$$

   The standard errors are derived from the inverse of the Fisher Information matrix, $\mathcal{I}^{-1} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}$.

3. **Calculate Standardized Score:**

$$Z_{U_j} = \frac{U_j(\beta_{\text{true}})}{\sqrt{\text{Var}(U_j)}} \tag{4.117}$$

   The score is evaluated at the *true* parameter values.

4. **Calculate Deviance:**

$$D = 2\left(\ell(\hat{\beta}) - \ell(\beta_{\text{true}})\right) \tag{4.118}$$

   This statistic compares the fitted model to the true model. According to Wilks' theorem, this should follow a χ^2_p distribution (where $p = 2$ is the number of parameters).

5. Simulation Code and Results

The histograms below overlay the theoretical asymptotic densities (Red/Blue lines) on top of the empirical histograms. The close alignment confirms that even for a moderate sample size of $n = 200$, the asymptotic approximations are highly accurate.

```r
library(latex2exp)

# 1. Setup
set.seed(42)
n_sim <- 2000
n <- 200
beta_true <- c(0.5, 1.5)

# Generate fixed covariates
x_cov <- runif(n, 0, 1)
X_mat <- cbind(1, x_cov) # Design matrix

# True mean vector and Fisher Information at Truth
lambda_true <- exp(X_mat %*% beta_true)
W <- diag(as.vector(lambda_true))
I_fisher <- t(X_mat) %*% W %*% X_mat
inv_I <- solve(I_fisher)

# Storage vectors
z_mle_b0 <- numeric(n_sim)
z_mle_b1 <- numeric(n_sim)
z_score_b1 <- numeric(n_sim)
deviance_vals <- numeric(n_sim)

# 2. Simulation Loop
for(i in 1:n_sim) {
  # Generate response from true model
  y_sim <- rpois(n, lambda = lambda_true)

  # Fit GLM
  fit <- glm(y_sim ~ x_cov, family = poisson)
  beta_hat <- coef(fit)

  # A. Standardized MLEs (Wald)
  # Uses the TRUE information matrix for standardization to check theory strictly
  z_mle_b0[i] <- (beta_hat[1] - beta_true[1]) / sqrt(inv_I[1,1])
  z_mle_b1[i] <- (beta_hat[2] - beta_true[2]) / sqrt(inv_I[2,2])

  # B. Standardized Score at Truth
  # U = X^T (Y - lambda)
  U <- t(X_mat) %*% (y_sim - lambda_true)
  z_score_b1[i] <- U[2] / sqrt(I_fisher[2,2])

  # C. Deviance (Likelihood Ratio)
  # D = 2 * (ll_hat - ll_true)
  ll_hat <- as.numeric(logLik(fit))
  ll_true <- sum(dpois(y_sim, lambda_true, log = TRUE))
  deviance_vals[i] <- 2 * (ll_hat - ll_true)
}

# 3. Visualization (2x2 Grid)
par(mfrow = c(2, 2))
```

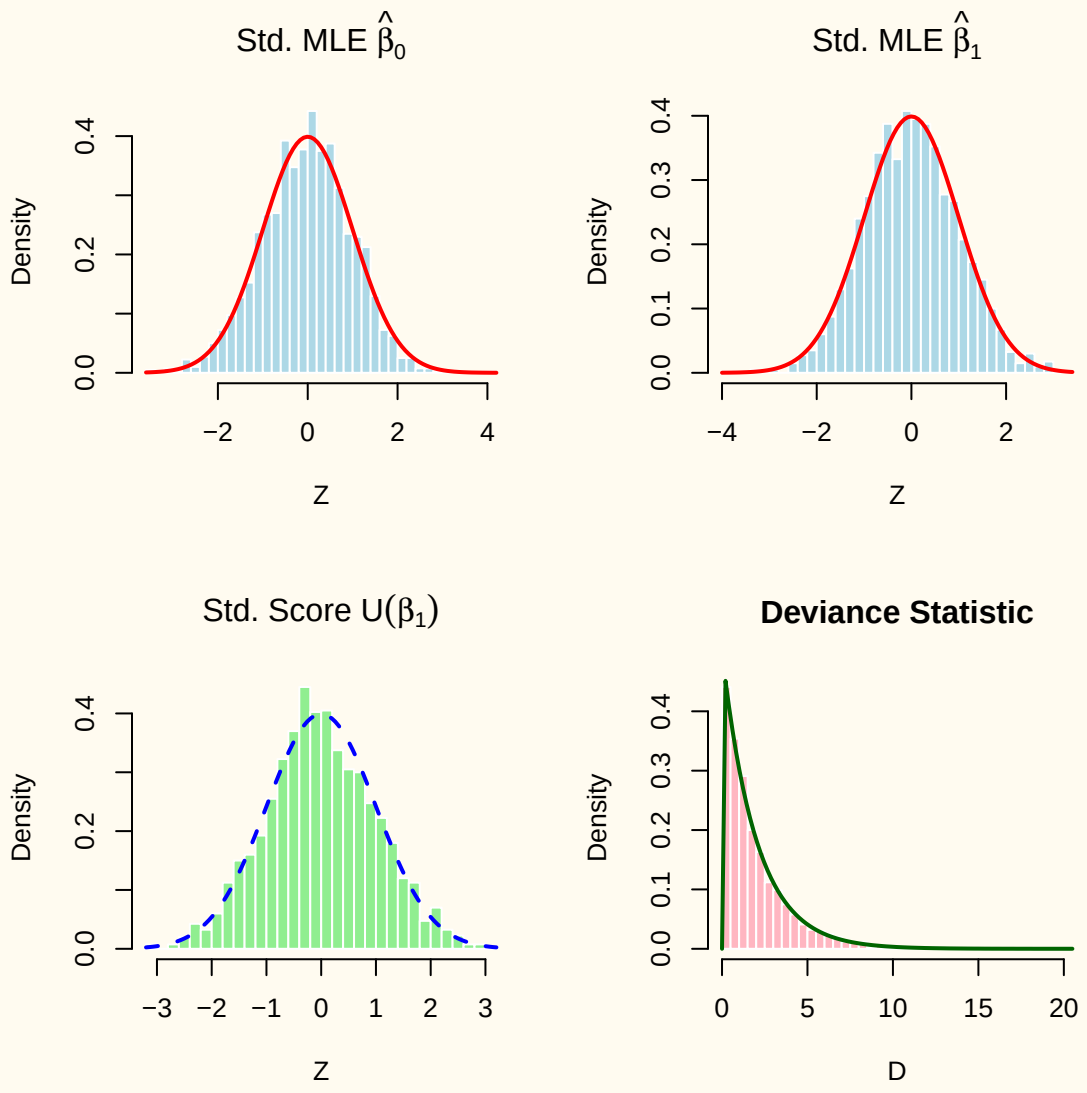


Figure 4.7: Asymptotic distributions for Poisson Regression (n=200). Top Row: Standardized MLEs for Intercept and Slope (Normal). Bottom Left: Standardized Score for Slope (Normal). Bottom Right: Deviance Statistic (Chi-squared with 2 df).

4.4.8 Alkeike Information Criterion for Estimating Out-of-sample Deviance

```
Converting page 1 to fig_expected_loglik.png... done!
```

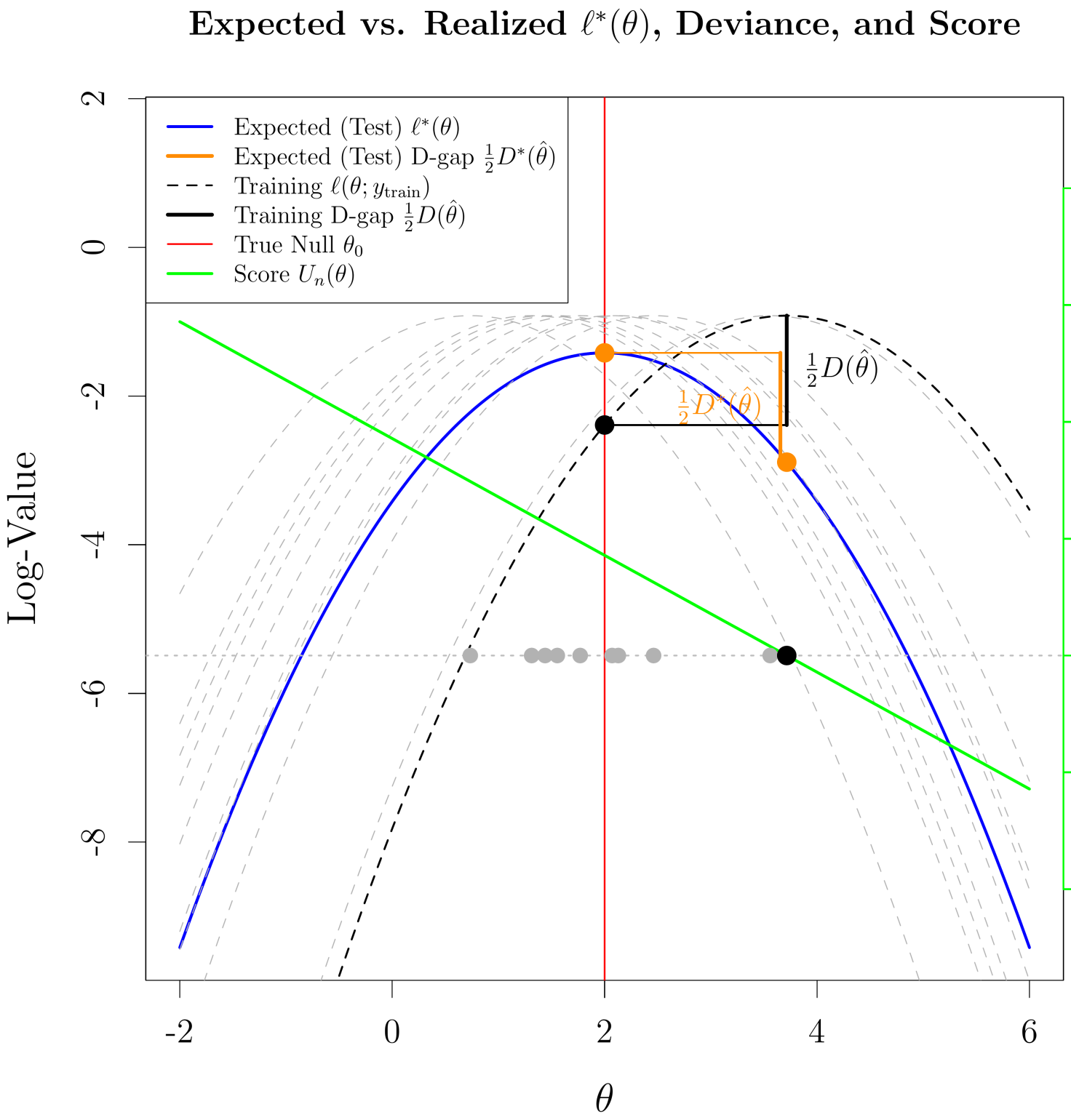


Figure 4.8: Expected Log-Likelihood $\ell^*(\theta)$ and 10 realized $\ell(\theta)$ curves. The largest sample is highlighted in black, showing the 1/2 Deviance gap. The expected out-of-sample drop is marked in orange.

The fundamental challenge in model selection is estimating how well our fitted model will perform on unseen data. When we optimize a model using our sample data Y_{train}, we climb the realized (training) log-likelihood curve to arrive at the maximum likelihood estimate, $\hat{\theta}$. Because $\hat{\theta}$ inherently accommodates the idiosyncratic noise of that specific sample, its training performance is artificially inflated compared to the true parameter θ_0. However, if we evaluate this exact same estimate $\hat{\theta}$ out-of-sample against new data Y_{test}, it performs worse than θ_0.

The Akaike Information Criterion (AIC) provides a geometric correction for this dual illusion by quantifying the total expected distance between the optimistic peak we observe and the true out-of-sample performance we actually care about.

1. The Realized Deviance Expansions

Let the deviance be defined as $D(\theta; Y) = -2\ell(\theta; Y)$, and let the geometric distance between the estimate and the truth be represented by the Wald statistic:

$$W = (\hat{\theta} - \theta_0)^T \mathcal{J}_n(\theta_0)(\hat{\theta} - \theta_0). \tag{4.119}$$

For the training data, the MLE $\hat{\theta}$ is the global minimum. By Taylor expanding the deviance of the true parameter $D(\theta_0; Y_{\text{train}})$ around the estimate $\hat{\theta}$, the first derivative vanishes, leaving:

$$D(\theta_0; Y_{\text{train}}) \approx D(\hat{\theta}; Y_{\text{train}}) + W \tag{4.120}$$

Rearranging this isolates the training deviance we observe:

$$\boxed{D(\hat{\theta}; Y_{\text{train}}) \approx D(\theta_0; Y_{\text{train}}) - W}. \tag{4.121}$$

For the test data, $\hat{\theta}$ is a fixed constant imported from the training phase. By Taylor expanding the test deviance $D(\hat{\theta}; Y_{\text{test}})$ around the true parameter θ_0, we get:

$$D(\hat{\theta}; Y_{\text{test}}) \approx D(\theta_0; Y_{\text{test}}) - 2(\hat{\theta} - \theta_0)^T U(Y_{\text{test}}, \theta_0) + W \tag{4.122}$$

where $U(Y_{\text{test}}, \theta_0) = \nabla \ell(\theta_0; Y_{\text{test}})$ is the score function evaluated on the test data.

To intuitively understand the behavior of this expansion, we can briefly average Y_{test} away. If we take the expectation strictly over the test data (holding our training estimate $\hat{\theta}$ fixed), the expected score evaluates to zero ($\mathbb{E}[U(Y_{\text{test}}, \theta_0)] = 0$). This reveals that the expected out-of-sample performance for our specific model rises above the baseline by exactly the Wald statistic:

$$\boxed{\mathbb{E}_{Y_{\text{test}}}[D(\hat{\theta}; Y_{\text{test}})] \approx \mathbb{E}_{Y_{\text{test}}}[D(\theta_0; Y_{\text{test}})] + W}. \tag{4.123}$$

2. The Raw Optimism Gap

Returning to the fully realized values, by subtracting the training deviance from the test deviance, we can express the total observed optimism gap as a function of the deviance evaluated at the true parameter θ_0:

$$D(\hat{\theta}; Y_{\text{test}}) - D(\hat{\theta}; Y_{\text{train}}) \approx [D(\theta_0; Y_{\text{test}}) - D(\theta_0; Y_{\text{train}})] - 2(\hat{\theta} - \theta_0)^T U(Y_{\text{test}}, \theta_0) + 2W \tag{4.124}$$

This single equation contains the entire story of overfitting. For any given pair of test and training samples, the optimism gap depends on the baseline difference in the data, a cross-term of the estimate and the test score, and twice the Wald statistic.

3. The Expected AIC Penalty

To find the theoretical penalty, we take the expectation of this gap with respect to both Y_{train} and Y_{test}.

Because θ_0 is a fixed, deterministic constant representing the true data-generating parameters, it carries no memory of the sampling process. As long as Y_{train} and Y_{test} are drawn from the same distribution, their expected deviance at θ_0 is identical and perfectly cancels out:

$$\mathbb{E}[D(\theta_0; Y_{\text{test}})] - \mathbb{E}[D(\theta_0; Y_{\text{train}})] = D^*(\theta_0) - D^*(\theta_0) = 0 \tag{4.125}$$

Furthermore, because the test data Y_{test} is completely independent of the estimate $\hat{\theta}$ (which was derived solely from Y_{train}), the expectation of their cross-term evaluates to zero.

This leaves only the expectation of the Wald statistic. Because $\hat{\theta}$ is asymptotically normal, W follows a χ^2_p distribution, meaning its expected value is exactly the dimension of the parameter space: $\mathbb{E}[W] = p$.

Applying the full expectation collapses the raw gap down to the pure penalty constant:

$$\mathbb{E}[D(\hat{\theta}; Y_{\text{test}})] - \mathbb{E}[D(\hat{\theta}; Y_{\text{train}})] = \mathbb{E}[2W] = 2p \tag{4.126}$$

4. The Akaike Information Criterion (AIC)

In practice, we only have access to our specific sample, Y_{train}. We cannot observe the future test data Y_{test}, which means we cannot directly compute the true out-of-sample deviance $D(\hat{\theta}; Y_{\text{test}})$.

However, using the expectation derived above, we know that our observable training deviance is, on average, structurally optimistic by exactly $2p$. To correct for this and construct an asymptotically unbiased estimator of the expected test deviance, we simply add this $2p$ penalty back to the realized training deviance.

This defines the Akaike Information Criterion:

$$\text{AIC} = D(\hat{\theta}; Y_{\text{train}}) + 2p = -2\ell(\hat{\theta}; Y_{\text{train}}) + 2p \tag{4.127}$$

By minimizing the AIC across a set of candidate models, we are no longer just finding the model that best fits the current data; we are effectively minimizing the estimated out-of-sample deviance. The $+2p$ term perfectly counterbalances the $-W$ drop in the training deviance, preventing the model from over-allocating parameters to chase the idiosyncratic noise of the training sample.

4.5 Optimization in Deep Learning

In deep learning, the objective function is typically the negative log-likelihood of the dataset, often supplemented by a penalty term (regularization) to prevent overfitting. While the theoretical foundations of Maximum Likelihood Estimation remain the same, the practical application in high-dimensional neural networks faces significant computational challenges that necessitate specialized optimization strategies.

4.5.1 A Brief Introduction to Deep Learning

See this page.

4.5.2 The Optimization Challenge

In traditional statistical models, $\ell(\theta)$ is often globally concave, allowing algorithms like Newton-Raphson to converge reliably. Deep learning models, however, present a different landscape:

1. **Non-convexity**

   The log-likelihood surface $\ell(\theta)$ is highly non-convex with numerous local minima, saddle points, and plateaus.

2. **High Dimensionality**

   The parameter vector θ can contain millions or billions of weights ($p \gg 10^6$). This makes the storage and inversion of the Observed Information Matrix $\mathbf{J}(\theta)$ (the $p \times p$ Hessian) computationally intractable, as its storage requires $O(p^2)$ memory.

4.5.3 Penalized Likelihood (Regularization)

To improve generalization, we often optimize a **penalized log-likelihood**:

$$\ell_{pen}(\theta) = \ell(\beta) - \Omega(\theta) \tag{4.128}$$

where $\Omega(\theta)$ is a penalty function. Common examples include:

1. L_2 **Regularization (Weight Decay):** $\Omega(\theta) = \frac{\lambda}{2}\|\theta\|^2$, which corresponds to a Gaussian prior in a Bayesian framework.
2. L_1 **Regularization (Lasso):** $\Omega(\theta) = \lambda\|\theta\|_1$, which encourages sparsity in the network weights.

4.5.4 Scalable First-Order and Adaptive Optimization Methods

Because the exact Hessian is computationally prohibitive in high-dimensional settings, scalable optimization strategies rely on the Score vector (the gradient of the log-likelihood) to either take direct steps or approximate local curvature. By avoiding full $\mathcal{O}(p^3)$ matrix inversions, these methods maintain a feasible memory and computational complexity of roughly $\mathcal{O}(p)$, making them the standard for large-scale problems.

4.5.4.1 Stochastic Gradient and Momentum Methods

- **Stochastic Gradient Descent (SGD):** Instead of computing the full score over the entire dataset, SGD uses a stochastic "mini-batch" estimate $\mathbf{U}_{batch}$. The standard update rule is:

$$\theta^{(t+1)} = \theta^{(t)} + \alpha \mathbf{U}_{batch}(\theta^{(t)}) \tag{4.129}$$

  While computationally trivial, SGD treats the parameter space as Euclidean (flat). It treats all dimensions equally, which leads to erratic zig-zagging if the likelihood surface is highly skewed.

- **Momentum-Based Methods:** To mitigate the noise of stochastic mini-batches and navigate steep ravines, momentum maintains an exponentially decaying moving average of past scores. This acts as a low-pass filter, dampening orthogonal oscillations and accelerating the optimizer along consistent directions.

4.5.4.2 Natural SGD with Empirical Fisher Information

To correct the flat geometry of SGD, modern optimizers leverage the **Second Moment Identity** (Bartlett's Identity) to capture curvature information without the $\mathcal{O}(p^2)$ cost of computing the actual Hessian.

According to this identity, the Fisher Information Matrix $\mathcal{I}(\theta)$ is exactly the covariance matrix of the score. Since $E[\mathbf{U}] = \mathbf{0}$, this simplifies to the expected second moment:

$$\mathcal{I}(\theta) = \text{Cov}_{\theta_0}(\mathbf{U}) = E_{\theta_0}[\mathbf{U}\mathbf{U}^\top] \tag{4.130}$$

This provides a computational lifeline: we can estimate second-order curvature using only first-order gradients.

- **The Empirical Fisher Estimate:** In practical optimization, we cannot compute the expectation over all possible data. Instead, we compute the sample average of the cross-products of the score vectors using the observed mini-batch data:

$$\hat{\mathcal{I}}_{\text{empirical}} = \frac{1}{n_{batch}} \sum_{i=1}^{n_{batch}} \mathbf{U}_i(\mathbf{y}_i)\mathbf{U}_i(\mathbf{y}_i)^\top \tag{4.131}$$

- **Full-Matrix NSGD:** Using the full $p \times p$ Empirical Fisher perfectly captures parameter correlations. The update step is $\theta^{(t+1)} = \theta^{(t)} + \alpha \hat{\mathcal{I}}^{-1}\mathbf{U}_{batch}$. However, inverting this matrix requires $\mathcal{O}(p^3)$ operations, rendering it impossible for massive models.

- **Diagonal NSGD:** To handle high dimensionality, practitioners drop the off-diagonal elements, assuming parameters are roughly independent. The diagonal elements are computed via the element-wise (Hadamard) product:

$$\text{diag}(\hat{\mathcal{I}}) = \frac{1}{n_{batch}} \sum_{i=1}^{n_{batch}} [\mathbf{U}_i \odot \mathbf{U}_i] \tag{4.132}$$

  This reduces the inversion complexity to $\mathcal{O}(p)$ scalar divisions, scaling the gradient coordinate-wise.

Remark 4.2. **A Note on the Empirical Approximation** While $\text{Cov}(\mathbf{U}) = E[\mathbf{J}]$ holds at the true parameter θ_0, it may not hold perfectly when θ is far from the truth. In these regions, the empirical Fisher is technically an approximation of the *variability* of the gradients drawn from the training data, rather than the true geometric curvature of the model.

4.5.4.3 Adam, AdamW, and Adaptive Equivalents

Algorithms like RMSProp and Adam build directly upon the concept of the Diagonal Empirical Fisher, introducing moving averages and specific geometric modifications to handle the chaotic, non-convex landscapes of deep neural networks.

- **RMSProp:** Maintains a running average of the squared scores for each parameter j:

$$v_j^{(t)} = \gamma v_j^{(t-1)} + (1-\gamma)(U_j^{(t)})^2 \tag{4.133}$$

- **Adam (Adaptive Moment Estimation):** Combines momentum (first moment) with the RMSProp variance estimate (second moment). It maintains bias-corrected running averages:
  - **First Moment (m_j):** $m_j^{(t)} = \beta_1 m_j^{(t-1)} + (1-\beta_1)U_j^{(t)}$
  - **Second Moment (v_j):** $v_j^{(t)} = \beta_2 v_j^{(t-1)} + (1-\beta_2)(U_j^{(t)})^2$

  The Adam update rule is:

$$\theta_j^{(t+1)} = \theta_j^{(t)} + \alpha \frac{\hat{m}_j}{\sqrt{\hat{v}_j} + \epsilon} \tag{4.134}$$

  The Square Root Departure: Notice that v_j serves as an estimate of the diagonal Fisher ($\mathcal{I}_{jj}$). True Natural SGD divides by the variance to maintain Riemannian invariance. Adam divides by the *square root* of the variance (the standard deviation). This makes Adam roughly bound its step sizes by $\pm\alpha$, acting more like a smoothed "sign descent" algorithm than a mathematically strict Natural Gradient method.

Remark 4.3. **AdamW: Decoupling Weight Decay from the Gradient**

The transition from Adam to **AdamW** resolves a subtle but critical mathematical conflation between L_2 regularization and weight decay that occurs in adaptive gradient methods.

1. **The Breakdown in Standard Adam**

   In standard SGD, adding an L_2 penalty to the loss ($\ell_{reg} = \ell + \frac{\lambda}{2}\|\theta\|^2$) is mathematically equivalent to decaying the weights. However, if we pass the L_2-regularized gradient ($\mathbf{U}_{reg} = \mathbf{U} + \lambda\theta$) into Adam, the regularization term $\lambda\theta$ gets divided by the adaptive denominator $\sqrt{\hat{v}_j}$:

$$\theta_j^{(t+1)} = \theta_j^{(t)} - \alpha \frac{\hat{m}_j(\mathbf{U}_{reg})}{\sqrt{\hat{v}_j(\mathbf{U}_{reg})} + \epsilon} \tag{4.135}$$

   This causes a severe structural problem: parameters with high historical gradients (large $\hat{v}_j$) receive *less* regularization penalty. The geometry of the likelihood surface inappropriately scales the geometric prior, often leading to poor generalization.

2. **The AdamW Solution**

   AdamW fixes this by explicitly separating the two mechanisms. It computes the first and second moments strictly using the unregularized gradient $\mathbf{U}$ to capture the true geometric curvature of the data. It then applies the weight decay directly to the parameter update, bypassing the adaptive scaling entirely:

$$\theta_j^{(t+1)} = \theta_j^{(t)}(1 - \eta\lambda) - \alpha \frac{\hat{m}_j(\mathbf{U})}{\sqrt{\hat{v}_j(\mathbf{U})} + \epsilon} \tag{4.136}$$

   This decoupling ensures that the regularizer decays all weights uniformly, while the gradients are still adaptively scaled. This exact modification established AdamW as the foundational optimizer for training modern Large Language Models.

4.5.4.4 Re-sampling Natural SGD

To obtain a mathematically rigorous Natural Gradient, the expectation $E[\mathbf{U}\mathbf{U}^\top]$ must be taken over the **model's predicted distribution**, not the observed training data.

Re-sampling Natural SGD achieves this by bypassing the empirical training targets:

1. At each step, it simulates "fake" target data $\mathbf{y}_{sim}$ from the current postulated model, $p(\cdot|\mathbf{X}, \theta^{(t)})$.
2. It calculates a simulated score vector $\mathbf{U}_{sim}$ based on this generated data.
3. It estimates the true Fisher Information using the cross-products of the simulated scores: $\hat{\mathcal{J}}_{\text{true}} \approx \mathbf{U}_{sim}\mathbf{U}_{sim}^\top$.

This Monte Carlo approach ensures the optimizer is navigating the actual probability manifold defined by the model, though it often requires a smaller learning rate to manage the variance introduced by the simulation step.

4.5.5 Example: The Poisson Regression Problem and Experimental Setup

In Poisson regression, we model count data $y_i \in \{0, 1, 2, ...\}$ given a vector of covariates $\mathbf{x}_i \in \mathbb{R}^p$. Assuming $y_i \sim \text{Poisson}(\lambda_i)$, the canonical log-link function connects the expected mean to the linear predictor:

$$\lambda_i = \exp(\mathbf{x}_i^\top \beta) \tag{4.137}$$

The goal of optimization is to find the parameter vector β that minimizes the negative log-likelihood. The true gradient of this objective is $\mathbf{g} = \mathbf{X}^\top(\lambda - \mathbf{y})$, and the true Fisher Information Matrix (FIM), which acts as the Hessian, is $\mathcal{J} = \mathbf{X}^\top \text{diag}(\lambda)\mathbf{X}$.

4.5.5.1 Data Generation Strategy

To test these algorithms in a realistic, challenging environment, we simulate a high-dimensional dataset with significant collinearity:

- **Dimensions:** $n = 5000$ observations and $p = 100$ features (plus an intercept).

- **Collinearity:** The covariates $\mathbf{X}$ are drawn from a Multivariate Normal distribution where every pair of variables has a high correlation of $\rho = 0.8$. This ill-conditioned design matrix makes the curvature of the optimization landscape highly skewed, punishing methods that ignore parameter correlations.
- **True Parameters:** The underlying ground-truth vector β^* is sparse. Only the intercept and the first three predictors are active, with alternating signs and magnitudes: $\beta^* = (0.5, -1.0, -0.5, 1.0, 0, ..., 0)^\top$. The remaining 97 parameters are identically zero.

4.5.5.2 Algorithm Comparison and Hyperparameters

All stochastic methods attempt to approximate the optimal update step $\beta^{(t+1)} = \beta^{(t)} - \eta\mathcal{H}^{-1}\mathbf{g}$ using a mini-batch of data. They differ fundamentally in how they compute the curvature matrix $\mathcal{H}$, whether they leverage the model's statistical properties, and if they rely on simulated data.

To stabilize the variance across iterations, the Natural SGD methods utilize an Exponential Moving Average (EMA) with a decay rate of $\rho = 0.95$ to smooth their Fisher Information estimates over time.

Method	Gradient ($\mathbf{g}$) Computation	Fisher/Curvature ($\mathcal{H}$) Computation	Re-sampling Used?	Tuning Settings
1. Standard SGD	Analytical on real data: $\mathbf{X}_b^\top(\lambda_b - \mathbf{y}_b)$	None (Identity Matrix $\mathbf{I}$)	No	$\eta = 0.0001$, Batch = 50, Iter = 10k
2. Diagonal NSGD	Analytical on real data: $\mathbf{X}_b^\top(\lambda_b - \mathbf{y}_b)$	Analytical Diagonal: $\mathcal{H}_{jj} = \sum_{i \in b} x_{ij}^2 \lambda_i$	No	$\eta = 0.01$, Batch = 50, Iter = 3.5k
3. Full-Matrix NSGD	Analytical on real data: $\mathbf{X}_b^\top(\lambda_b - \mathbf{y}_b)$	Analytical Full FIM: $\mathbf{X}_b^\top \text{diag}(\lambda_b)\mathbf{X}_b$	No	$\eta = 0.05$, Batch = 50, Iter = 1.5k
4. Re-sampling NSGD	Analytical on real data: $\mathbf{X}_b^\top(\lambda_b - \mathbf{y}_b)$	Monte Carlo Diagonal: $\mathcal{H}_{jj} \approx g_{sim,j}^2$	Yes ($y_{sim} \sim$ Poisson(λ))	$\eta = 0.005$, Batch = 50, Iter = 8k

4.5.5.3 Optimization Trajectories by Wall-Clock Time

By plotting the parameter estimates against actual execution time (in seconds) rather than iteration count, we visualize the fundamental trade-off in high-dimensional optimization: the cost of accurate curvature calculation versus the speed of frequent updates.

4.5.5.4 Interpretation of Convergence Behaviors

The visual results clearly underscore the weaknesses of uncorrected gradients in highly collinear spaces. **Standard SGD** requires a severely restricted learning rate to prevent divergence, causing the parameters to drift aimlessly; the inactive coefficients (in grey) completely fail to shrink back to zero within the allotted time limit.

In contrast, incorporating the Fisher Information corrects for the warped geometry. **Diagonal NSGD** rapidly pushes the grey noise parameters toward zero, though it chatters slightly around the true active values due to ignoring the off-diagonal correlations. The **Full-Matrix NSGD** provides the most direct, stable path to convergence with minimal variance, but takes noticeably longer per step to invert the full 100×100 matrix. Finally, the **Re-sampling**

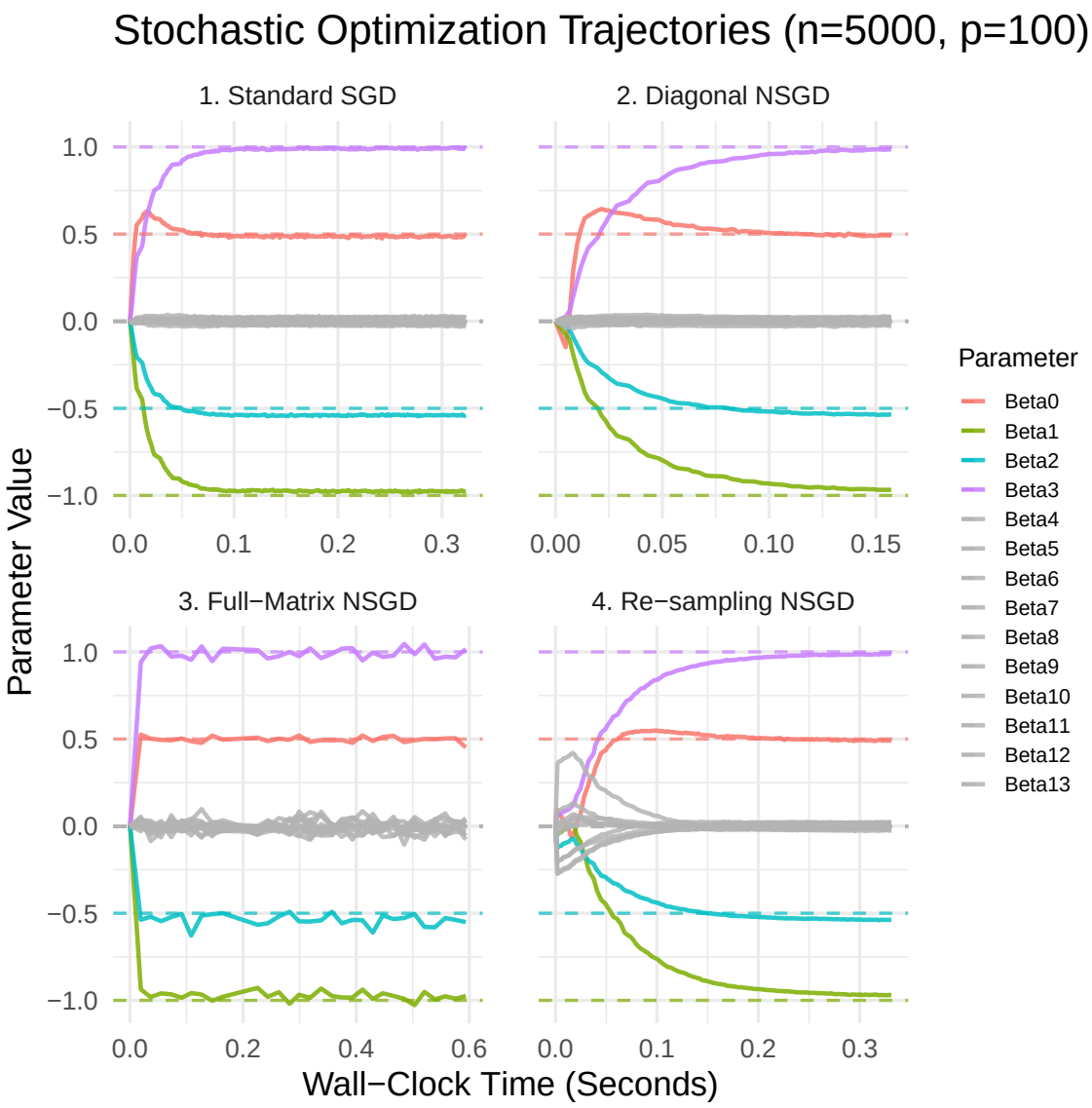


Figure 4.9: Informal demonstration of parameter traces over wall-clock execution time for four stochastic optimization strategies ($n = 5000, p = 100$). Dashed lines represent the true parameter values. Inactive parameters (Beta4 to Beta13) are plotted in grey to demonstrate shrinkage and noise around zero.

NSGD demonstrates that we can successfully approximate the diagonal Fisher geometrically by injecting simulated model noise, achieving shrinkage comparable to the analytical diagonal method despite requiring a slightly more conservative learning rate to manage the Monte Carlo variance.

4.6 Appendix: Derivation of Score and Information in Poisson Regression

In a Poisson Generalized Linear Model with a canonical link, the mean is modeled as $\lambda_i(\beta) = \exp(\mathbf{x}_i^\top \beta)$.

Throughout these derivations, we will utilize the **Hadamard product** (element-wise multiplication), denoted by the operator $\odot$, to provide concise vectorized representations. For two vectors **a** and **b** of the same dimension, their Hadamard product $\mathbf{a} \odot \mathbf{b}$ yields a vector with elements $a_i b_i$. When applied between a column vector and a matrix, the operation broadcasts the element-wise multiplication down the columns of the matrix.

The derivation of the Score vector and the Fisher Information matrix relies on the chain rule and the properties of the exponential function.

4.6.1 Derivation of the Score Vector $\mathbf{U}(\beta)$

The Score vector is defined as the gradient of the log-likelihood with respect to the parameter vector β.

1. **Differentiate the Log-Likelihood** Starting from the log-likelihood function:

$$\ell(\beta) = \sum_{i=1}^{n} \left(y_i(\mathbf{x}_i^\top \beta) - \exp(\mathbf{x}_i^\top \beta) - \log(y_i!) \right) \tag{4.138}$$

   We take the partial derivative with respect to a single component β_j:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \left(y_i \frac{\partial (\mathbf{x}_i^\top \beta)}{\partial \beta_j} - \frac{\partial \exp(\mathbf{x}_i^\top \beta)}{\partial \beta_j} \right) \tag{4.139}$$

2. **Apply the Chain Rule** Note that $\frac{\partial (\mathbf{x}_i^\top \beta)}{\partial \beta_j} = x_{ij}$. Using the chain rule on the exponential term:

$$\frac{\partial \exp(\mathbf{x}_i^\top \beta)}{\partial \beta_j} = \exp(\mathbf{x}_i^\top \beta) \cdot x_{ij} = \lambda_i(\beta) x_{ij} \tag{4.140}$$

3. **Form the Vector** Substituting these back, we get:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \lambda_i(\beta)) x_{ij} \tag{4.141}$$

   Let $\mathbf{x}^{(j)}$ denote the $n \times 1$ vector of the j-th covariate across all observations. Using the Hadamard product $\odot$, the j-th component of the score vector can be written equivalently as the sum of the elements in the resulting vector:

$$U_j(\beta) = \sum \left((\mathbf{y} - \lambda(\beta)) \odot \mathbf{x}^{(j)} \right) \tag{4.142}$$

In full matrix notation, this corresponds to:

$$\mathbf{U}(\beta) = \mathbf{X}^\top(\mathbf{y} - \lambda(\beta)) \tag{4.143}$$

4.6.2 Derivation of the Observed Fisher Information $J(\beta)$

The Observed Information is the negative Hessian matrix of the log-likelihood.

1. **Differentiate the Score** We take the derivative of the j-th component of the score, $U_j = \sum_{i=1}^{n}(y_i - \lambda_i(\beta))x_{ij}$, with respect to β_k:

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = \frac{\partial}{\partial \beta_k} \sum_{i=1}^{n}(y_i - \lambda_i(\beta))x_{ij} = \sum_{i=1}^{n}\left(-\frac{\partial \lambda_i(\beta)}{\partial \beta_k}\right) x_{ij} \tag{4.144}$$

2. **Differentiate the Mean Function** Recall $\lambda_i(\beta) = \exp(\mathbf{x}_i^\top \beta)$. Its derivative with respect to β_k is:

$$\frac{\partial \lambda_i(\beta)}{\partial \beta_k} = \lambda_i(\beta)x_{ik} \tag{4.145}$$

3. **Construct the Information Matrix** The (j, k) element of the Hessian is:

$$H_{jk} = -\sum_{i=1}^{n} \lambda_i(\beta)x_{ij}x_{ik} \tag{4.146}$$

The Observed Information $J(\beta) = -H$ is therefore:

$$J(\beta)_{jk} = \sum_{i=1}^{n} \lambda_i(\beta)x_{ij}x_{ik} \tag{4.147}$$

Using the Hadamard product $\odot$, the (j, k)-th element is elegantly expressed as the sum of the element-wise product of three vectors (the predicted means and the two covariate columns):

$$J(\beta)_{jk} = \sum \left(\lambda(\beta) \odot \mathbf{x}^{(j)} \odot \mathbf{x}^{(k)}\right) \tag{4.148}$$

In matrix form, let $\mathbf{W}(\beta)$ be a diagonal matrix where $W_{ii} = \lambda_i(\beta)$. Then:

$$J(\beta) = \mathbf{X}^\top \mathbf{W}(\beta)\mathbf{X} \tag{4.149}$$

4.6.3 Derivation of the Expected Fisher Information $\mathcal{I}(\beta)$ via Var(U)

By the fundamental properties of likelihoods, the Expected Fisher Information matrix $\mathcal{I}(\beta)$ is equivalent to the variance-covariance matrix of the Score vector: $\mathcal{I}(\beta) = \text{Var}(\mathbf{U}(\beta))$.

1. **Set Up the Variance Expression** Let $u_{ij}(\beta)$ denote the gradient contribution of the i-th observation with respect to the parameter β_j:

$$u_{ij}(\beta) = (y_i - \lambda_i(\beta))x_{ij} \tag{4.150}$$

   The j-th component of the total Score vector is the sum of these individual observation gradients:

$$U_j(\beta) = \sum_{i=1}^{n} u_{ij}(\beta) \tag{4.151}$$

   We find the (j, k)-th element of the Expected Information by taking the covariance between the j-th and k-th score components. Since the individual observations i are independent, the cross-covariances between different observations ($i \neq i'$) are zero:

$$\mathcal{I}(\beta)_{jk} = \text{Cov}(U_j(\beta), U_k(\beta)) = \sum_{i=1}^{n} \text{Cov}(u_{ij}(\beta), u_{ik}(\beta)) \tag{4.152}$$

2. **Apply Variance Properties** Expanding the covariance term for the i-th observation using our definition of u_{ij} and u_{ik}:

$$\text{Cov}(u_{ij}(\beta), u_{ik}(\beta)) = \text{Cov}((y_i - \lambda_i(\beta))x_{ij}, (y_i - \lambda_i(\beta))x_{ik}) \tag{4.153}$$

   Because the predicted mean $\lambda_i(\beta)$ and the covariates x_{ij}, x_{ik} are treated as fixed constants, the covariance simplifies to the variance of the random variable y_i scaled by the covariates. Therefore, summing over all observations gives:

$$\mathcal{I}(\beta)_{jk} = \sum_{i=1}^{n} x_{ij}x_{ik}\text{Var}(y_i) \tag{4.154}$$

3. **Substitute the Poisson Variance** For a Poisson distribution, the variance is equal to the mean: $\text{Var}(y_i) = \lambda_i(\beta)$. Substituting this known variance into our equation gives:

$$\mathcal{I}(\beta)_{jk} = \sum_{i=1}^{n} \lambda_i(\beta)x_{ij}x_{ik} \tag{4.155}$$

   Just as with the Observed Information, this (j, k)-th element takes the exact same form as the sum of the Hadamard product of the three vectors:

$$\mathcal{I}(\beta)_{jk} = \sum \left(\lambda(\beta) \odot \mathbf{x}^{(j)} \odot \mathbf{x}^{(k)}\right) \tag{4.156}$$

4. **Construct the Matrix Form** The full matrix can therefore be written using the exact same weight matrix $\mathbf{W}(\beta)$:

$$\mathcal{I}(\beta) = \mathbf{X}^\top \mathbf{W}(\beta)\mathbf{X} \tag{4.157}$$

Remark 4.4 (Canonical Link Advantage). Notice that because we used the canonical link ($\log \lambda = \eta$), the second derivative of the log-likelihood (the Hessian) does not depend on the observed data y_i. It only depends on the predicted means $\lambda_i(\beta)$. Consequently, the **Observed Information** and the **Expected Information** are identical:

$$J(\beta) = \mathcal{J}(\beta) \tag{4.158}$$

This is a unique property of canonical links in Generalized Linear Models.

5 Most Powerful Tests

5.1 General Termininologies of Hypothesis Testing

5.1.1 Hypothesis

We formulate the problem of hypothesis testing as deciding between two competing claims about a parameter θ:

$$H_0 : \theta \in \Theta_0 \quad \text{(Null Hypothesis)} \tag{5.1}$$

$$H_1 : \theta \in \Theta_1 \quad \text{(Alternative Hypothesis)} \tag{5.2}$$

Definition 5.1 (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 5.1 (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).

5.1.2 Test Functions

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:

$$\phi(x) = I(x \in C_\alpha) = \begin{cases} 1 & \text{if } x \in C_\alpha \text{ (Reject } H_0) \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

- A **randomized** test, $\phi(x)$ can take values in $[0, 1]$, which can be expressed typically as follows:

$$\phi(x) = \begin{cases} 1 & \text{if } x \in C_1 \\ \gamma & \text{if } x \in C_* \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

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 5.2 (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:

$$\phi(x) = \begin{cases} 1 & \text{if } x \in C_1 \text{ (i.e., } x \geq 9) \\ \gamma & \text{if } x \in C_* \text{ (i.e., } x = 8) \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

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:

$$\begin{aligned} \alpha &= P(X \geq 9) + \gamma \cdot P(X = 8) \\ 0.05 &= 0.01074 + \gamma \cdot (P(X \geq 8) - P(X \geq 9)) \\ 0.05 &= 0.01074 + \gamma \cdot (0.05469 - 0.01074) \end{aligned} \tag{5.6}$$

Solving for γ:

$$\gamma = \frac{0.05 - 0.01074}{0.04395} \approx \frac{39}{44} \approx 0.89 \tag{5.7}$$

5.1.3 Size

Definition 5.2 (Size of a Test)**.** The **size** of a test $\phi(x)$ is the maximum probability of rejecting the null hypothesis when it is true:

$$\text{Size}(\phi) = \sup_{\theta \in \Theta_0} W_\phi(\theta) = \sup_{\theta \in \Theta_0} E_\theta[\phi(X)] \tag{5.8}$$

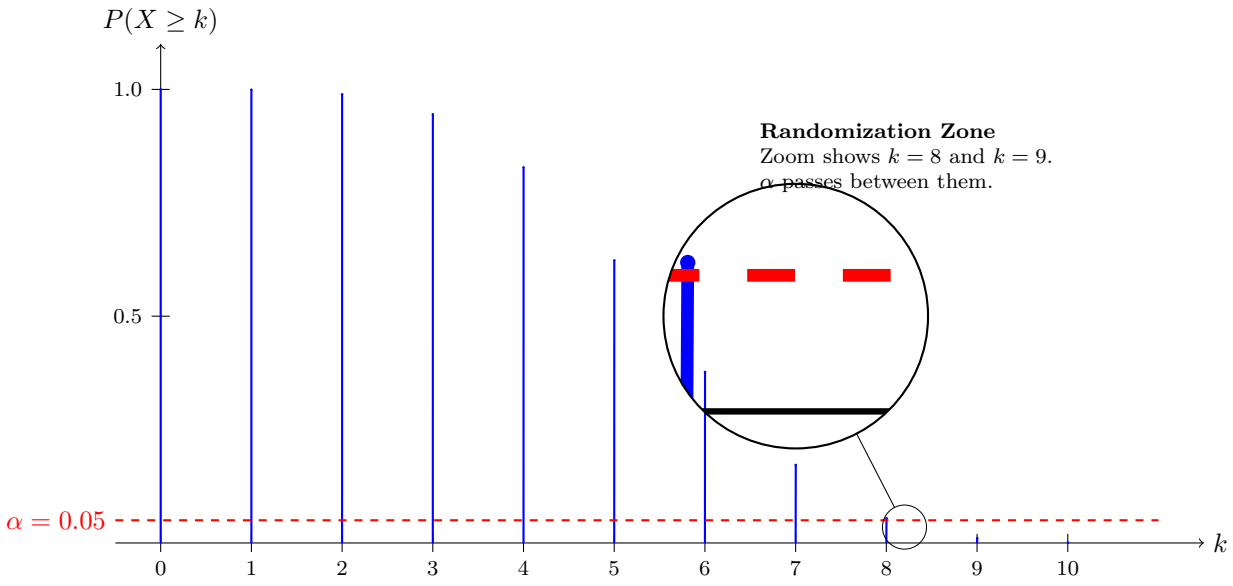


Figure 5.1: Survival Function P(X >= k) with randomization detail

5.1.4 Power

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:

$$W_\phi(\theta) = E_\theta[\phi(X)] \tag{5.9}$$

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:

$$\text{Power}(\phi) = E_{\theta_1}[\phi(X)] \tag{5.10}$$

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).

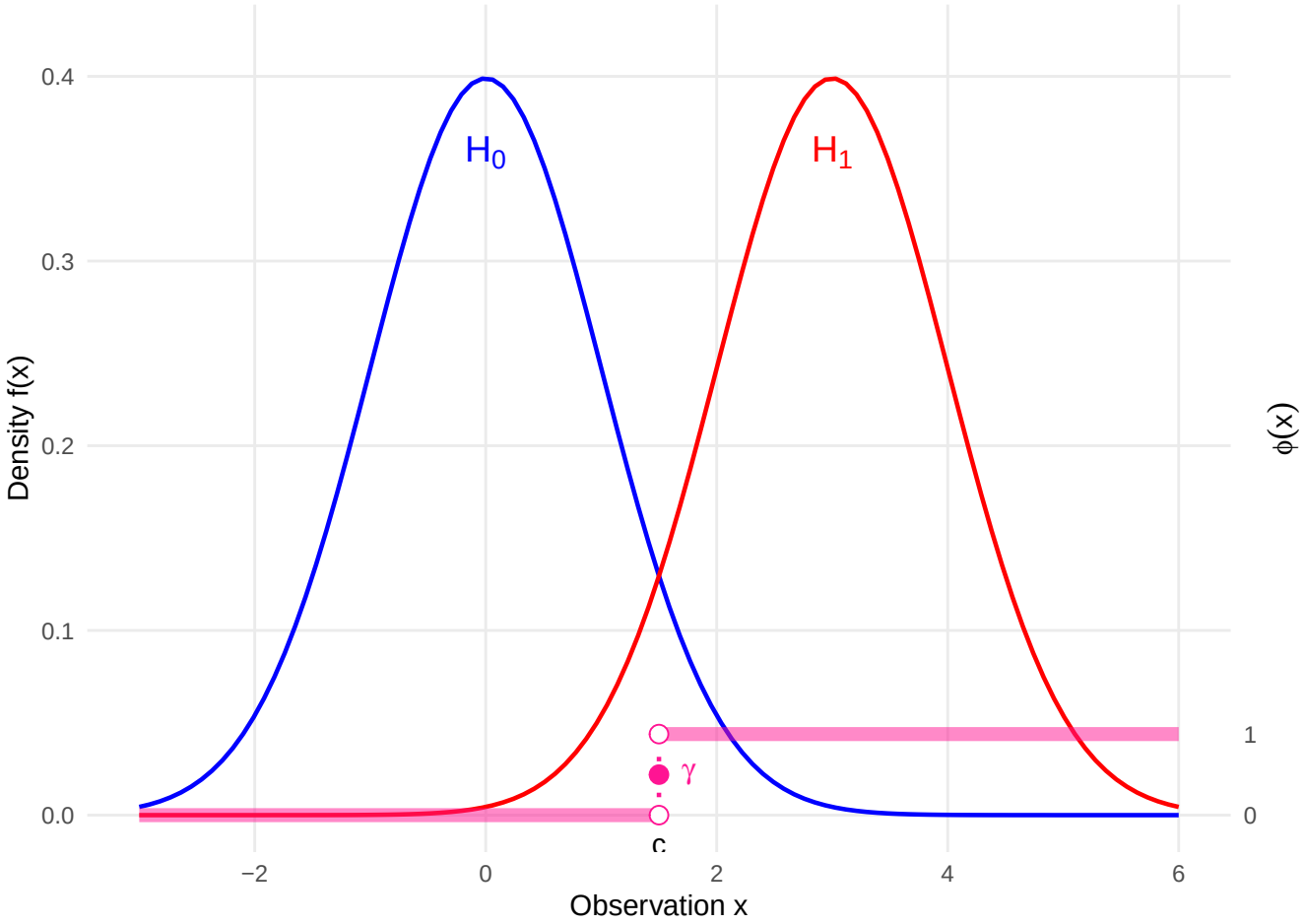


Figure 5.2: Illustration of a Test Function $\phi(x)$ (Pink) relative to Size (H_0, Blue) and Power (H_1, Red).

5.2 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:

$$\Lambda(x) = \frac{f_1(x)}{f_0(x)} = \frac{f(x;\theta_1)}{f(x;\theta_0)} \tag{5.11}$$

Definition 5.3 (Likelihood Ratio Test (LRT))**.** A test $\phi(x)$ is a Likelihood Ratio Test if it has the form:

$$\phi_{\text{LRT}}(x) = \begin{cases} 1 & \text{if } \Lambda(x) > k \\ \gamma(x) & \text{if } \Lambda(x) = k \\ 0 & \text{if } \Lambda(x) < k \end{cases} \tag{5.12}$$

where $k \geq 0$ is a constant and $0 \leq \gamma(x) \leq 1$.

5.2.1 Neyman-Pearson Lemma

Theorem 5.1 (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.*

5.2.2 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:

$$\text{Power}(\phi) = \int \phi(x) f_1(x) dx \tag{5.13}$$

subject to the constraint on the size of the test α:

$$\text{Size}(\phi) = \int \phi(x) f_0(x) dx = \alpha \tag{5.14}$$

Using the method of Lagrange multipliers, we define the objective function L with a multiplier k:

$$L(\phi, k) = \int \phi(x) f_1(x) dx - k \left(\int \phi(x) f_0(x) dx - \alpha \right) \tag{5.15}$$

Rearranging the terms inside the integral, we get:

$$L(\phi, k) = \int \phi(x)[f_1(x) - k f_0(x)] dx + k\alpha \tag{5.16}$$

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 $f_1(x) - k f_0(x) > 0$, set $\phi(x) = 1$.
- If $f_1(x) - k f_0(x) < 0$, set $\phi(x) = 0$.
- If $f_1(x) - k f_0(x) = 0$, the value of $\phi(x)$ does not affect the integral (this is where γ comes in).

This decision rule is equivalent to:

$$\phi(x) = \begin{cases} 1 & \text{if } \frac{f_1(x)}{f_0(x)} > k \\ 0 & \text{if } \frac{f_1(x)}{f_0(x)} < k \end{cases} \tag{5.17}$$

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.

5.2.3 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:

$$U(x) = (\phi_{\text{LRT}}(x) - \phi(x))(f_1(x) - k f_0(x)) \tag{5.18}$$

We analyze the sign of $U(x)$ by looking at the sign of its two factors in three cases:

- If $f_1(x) - k f_0(x) > 0$ (implies $\Lambda(x) > k$). Since $\phi_{\text{LRT}}(x) = 1$ and $\phi(x) \leq 1$, we have:

$$\begin{aligned} \phi_{\text{LRT}}(x) - \phi(x) &\geq 0 \\ U(x) = (\phi_{\text{LRT}}(x) - \phi(x))(f_1(x) - k f_0(x)) &\geq 0 \end{aligned} \tag{5.19}$$

- If $f_1(x) - k f_0(x) < 0$ (implies $\Lambda(x) < k$). Since $\phi_{\text{LRT}}(x) = 0$ and $\phi(x) \geq 0$, we have:

$$\begin{aligned} \phi_{\text{LRT}}(x) - \phi(x) &\leq 0 \\ U(x) = (\phi_{\text{LRT}}(x) - \phi(x))(f_1(x) - k f_0(x)) &\geq 0 \end{aligned} \tag{5.20}$$

- If $f_1(x) - k f_0(x) = 0$. The product is zero regardless of the test functions.

$$U(x) = 0 \tag{5.21}$$

Combining these cases, we conclude that the product is non-negative for all x:

$$U(x) = (\phi_{\text{LRT}}(x) - \phi(x))(f_1(x) - kf_0(x)) \geq 0 \tag{5.22}$$

Therefore, integrating $U(x)$ over the entire domain:

$$\int U(x)dx = \int (\phi_{\text{LRT}}(x) - \phi(x))(f_1(x) - kf_0(x))dx \geq 0 \tag{5.23}$$

Expanding the integral:

$$\int \phi_{\text{LRT}}(x) f_1(x)\, dx - \int \phi(x) f_1(x)\, dx - k\left(\int \phi_{\text{LRT}}(x) f_0(x)\, dx - \int \phi(x) f_0(x)\, dx\right) \geq 0 \tag{5.24}$$

Converting to expectations:

$$E_{\theta_1}[\phi_{\text{LRT}}] - E_{\theta_1}[\phi] \geq k(E_{\theta_0}[\phi_{\text{LRT}}] - E_{\theta_0}[\phi]) \tag{5.25}$$

Since $E_{\theta_0}[\phi_{\text{LRT}}] = \text{Size}(\phi_{\text{LRT}}) = \alpha$ and we require that $E_{\theta_0}[\phi] = \text{Size}(\phi) \leq \alpha$,

$$E_{\theta_0}[\phi_{\text{LRT}}] - E_{\theta_0}[\phi] \geq 0 \tag{5.26}$$

Thereore, given that $k \geq 0$:

$$\text{Power}(\phi_{\text{LRT}}) \geq \text{Power}(\phi) \tag{5.27}$$

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:

$$P_{\theta_0}(\Lambda(X) > k_0) \leq \alpha \leq P_{\theta_0}(\Lambda(X) \geq k_0) \tag{5.28}$$

Set $\gamma_0 = \frac{\alpha - P_{\theta_0}(\Lambda(X) > k_0)}{P_{\theta_0}(\Lambda(X) = k_0)}$.
Proof of (c) Uniqueness
Let ϕ_{LRT} be the LRT of size α. Suppose there exists another test ϕ that is also Most Powerful (MP) with size $\leq \alpha$. We wish to show that $\phi(x) = \phi_{\text{LRT}}(x)$ for almost all x where $f_1(x) \neq kf_0(x)$.
As established in the optimality proof, the function:

$$U(x) = (\phi_{\text{LRT}}(x) - \phi(x))(f_1(x) - kf_0(x)) \tag{5.29}$$

is non-negative for all x. Since both tests are MP, they have the same power: $E_{\theta_1}[\phi_{\text{LRT}}] = E_{\theta_1}[\phi]$.
From the integral of $U(x)$, we have:

$$0 \leq \int U(x)dx = (E_{\theta_1}[\phi_{\text{LRT}}] - E_{\theta_1}[\phi]) - k(E_{\theta_0}[\phi_{\text{LRT}}] - E_{\theta_0}[\phi]) \tag{5.30}$$

Substituting the equality of power:

$$0 \leq -k(\alpha - E_{\theta_0}[\phi]) \tag{5.31}$$

Since $k > 0$ and $E_{\theta_0}[\phi] \leq \alpha$, the term $-k(\alpha - E_{\theta_0}[\phi])$ is ≤ 0. The only way for the integral of a non-negative function $U(x)$ to be ≤ 0 is if the integral is exactly zero:

$$\int (\phi_{\text{LRT}}(x) - \phi(x))(f_1(x) - kf_0(x))\, dx = 0 \tag{5.32}$$

For the integral of a non-negative function to be zero, the integrand must be zero almost everywhere:

$$(\phi_{\text{LRT}}(x) - \phi(x))(f_1(x) - kf_0(x)) = 0 \quad \text{a.e.} \tag{5.33}$$

This implies that for any x where $f_1(x) - kf_0(x) \neq 0$, we must have:

$$\phi_{\text{LRT}}(x) - \phi(x) = 0 \implies \phi(x) = \phi_{\text{LRT}}(x) \tag{5.34}$$

Thus, the test is unique except possibly on the boundary set $\{x : f_1(x) = kf_0(x)\}$. If $P_{\theta_0}(\Lambda(X) = k) = 0$ (as in continuous distributions like the Normal), the MP test is unique almost everywhere. □

5.3 Uniformly Most Powerful (UMP) Tests via MLR

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 5.4 (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 $\theta \in \Theta_1$.

5.3.1 Monotone Likelihood Ratio (MLR)

Definition 5.5 (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:

$$\frac{f(x;\theta_1)}{f(x;\theta_0)} \tag{5.35}$$

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 $T(x)$.

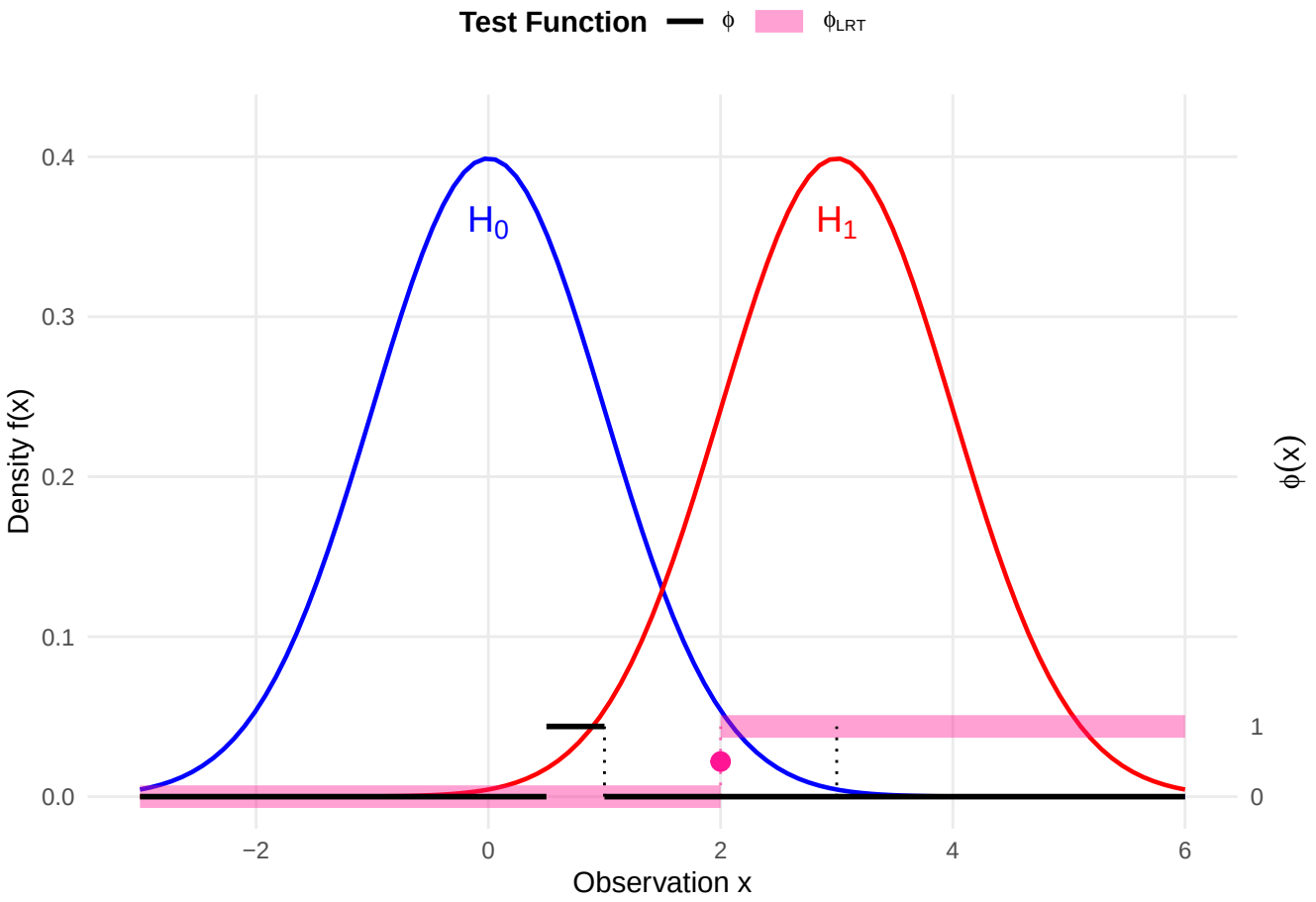


Figure 5.3: Visualizing the NP Lemma. The thick, transparent pink line is ϕ_{extLRT}. The thin solid black line is ϕ. Overlap is visible as a black line inside pink.

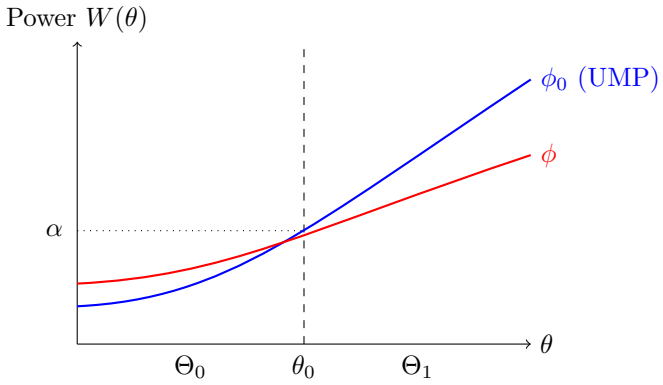


Figure 5.4: Diagram of UMP Tests

Example 5.3. Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Exp}(\theta)$ with pdf $f(x) = \frac{1}{\theta}e^{-x/\theta}$. Test

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta > \theta_0. \tag{5.36}$$

The Likelihood Ratio for $\theta_1 > \theta_0$ is:

$$\frac{L(\theta_1)}{L(\theta_0)} = \frac{\theta_1^{-n}e^{-\sum x_i/\theta_1}}{\theta_0^{-n}e^{-\sum x_i/\theta_0}} = \left(\frac{\theta_0}{\theta_1}\right)^n \exp\left\{\left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right)\sum x_i\right\} \tag{5.37}$$

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 the sum $T(x) = \sum x_i$. Rejecting for large $\Lambda(x)$ is equivalent to rejecting for $T(x) = \sum x_i > C$.
Under H_0, $X_i \sim \text{Exp}(\theta_0)$, which is equivalent to Gamma$(1, \theta_0)$. By the reproductive property of the Gamma distribution:

$$T = \sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \theta_0) \tag{5.38}$$

Alternatively, using the relationship with the Chi-square distribution:

$$\frac{2T}{\theta_0} \sim \chi^2_{2n} \tag{5.39}$$

To find C for a significance level α, we set $P(T > C|\theta_0) = \alpha$. Using the χ^2 transformation:

$$P\left(\frac{2T}{\theta_0} > \frac{2C}{\theta_0}\right) = \alpha \implies \frac{2C}{\theta_0} = \chi^2_{2n,\alpha} \tag{5.40}$$

Thus, the critical value is:

$$C = \frac{\theta_0}{2}\chi^2_{2n,\alpha} \tag{5.41}$$

where $\chi^2_{2n,\alpha}$ is the upper-α quantile of a Chi-square distribution with $2n$ degrees of freedom.

> **!** Important
>
> We note that the value C does not depend on θ_1.

5.3.2 Karlin-Rubin Theorem

Theorem 5.2 (Chebyshev's Association Inequality)**.**

Theorem 5.3. *Let X be a random variable, and let $f(x)$ and $g(x)$ be two functions that are both non-decreasing (or both non-increasing). Then:*

$$E[f(X)g(X)] \geq E[f(X)] \cdot E[g(X)] \tag{5.42}$$

Equivalently, the covariance is non-negative: $Cov(f(X), g(X)) \geq 0$.

Proof. Let Y be an independent copy of X (i.e., X and Y are i.i.d.). Consider the quantity:

$$\Delta = (f(X) - f(Y))(g(X) - g(Y)) \tag{5.43}$$

Since f and g are both non-decreasing (or both non-increasing), the terms $(f(X) - f(Y))$ and $(g(X) - g(Y))$ always share the same sign. Thus, their product is always non-negative:

$$\Delta \geq 0 \tag{5.44}$$

Taking the expectation:

$$E[(f(X) - f(Y))(g(X) - g(Y))] \geq 0 \tag{5.45}$$

Expanding the product and using linearity of expectation:

$$E[f(X)g(X)] - E[f(X)g(Y)] - E[f(Y)g(X)] + E[f(Y)g(Y)] \geq 0 \tag{5.46}$$

Since X and Y are i.i.d.:

1. $E[f(Y)g(Y)] = E[f(X)g(X)]$
2. $E[f(X)g(Y)] = E[f(X)]E[g(Y)] = E[f(X)]E[g(X)]$ (Independence)
3. $E[f(Y)g(X)] = E[f(Y)]E[g(X)] = E[f(X)]E[g(X)]$ (Independence)

Substituting these back yields:

$$2E[f(X)g(X)] - 2E[f(X)]E[g(X)] \geq 0 \tag{5.47}$$

Dividing by 2 proves the inequality:

$$E[f(X)g(X)] \geq E[f(X)]E[g(X)] \tag{5.48}$$

□

Theorem 5.4 (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 $H_0 : \theta \leq \theta^*$ vs $H_1 : \theta > \theta^*$.*
The test:

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > t^* \\ 0 & \text{if } T(x) \leq t^* \end{cases} \tag{5.49}$$

where t^* is determined by $P_{\theta^*}(T(X) > t^*) = \alpha$, is the UMP size α test.

Proof. **The Test:** Define the test $\phi^*(x)$ as:

$$\phi^*(x) = \begin{cases} 1 & \text{if } T(x) > t^* \\ 0 & \text{if } T(x) \leq t^* \end{cases} \tag{5.50}$$

where t^* is determined such that the power at the boundary is α, i.e., $W_{\phi^*}(\theta^*) = \alpha$.

1. **Monotonicity of the Power Function**

   We first establish that $W_{\phi^*}(\theta)$ is non-decreasing over the entire parameter space. Let θ_0 and θ_1 be any two arbitrary parameter values such that $\theta_1 > \theta_0$. Define the test indicator function $h(t) = \mathbb{I}(t > t^*)$ and the likelihood ratio $\Lambda(t) = \frac{f_{\theta_1}(t)}{f_{\theta_0}(t)}$.

   Because of the Monotone Likelihood Ratio (MLR) property, $\Lambda(t)$ is a non-decreasing function of t. The indicator function $h(t)$ is clearly non-decreasing.

   The power at θ_1 can be written as an integral involving the density under θ_0:

   $$W_{\phi^*}(\theta_1) = \int_{-\infty}^{\infty} h(t) f_{\theta_1}(t)\, dt = \int_{-\infty}^{\infty} h(t) \frac{f_{\theta_1}(t)}{f_{\theta_0}(t)} f_{\theta_0}(t)\, dt = \int_{-\infty}^{\infty} h(t)\Lambda(t) f_{\theta_0}(t)\, dt \tag{5.51}$$

   This integral is the expectation $E_{\theta_0}[h(T)\Lambda(T)]$. By Chebyshev's Association Inequality (Covariance Inequality), since both $h(t)$ and $\Lambda(t)$ are non-decreasing, the expectation of their product is at least the product of their expectations:

   $$\int_{-\infty}^{\infty} h(t)\Lambda(t) f_{\theta_0}(t)\, dt \geq \left(\int_{-\infty}^{\infty} h(t) f_{\theta_0}(t)\, dt \right) \left(\int_{-\infty}^{\infty} \Lambda(t) f_{\theta_0}(t)\, dt \right) \tag{5.52}$$

   We evaluate the two integrals on the right-hand side:

   1. The first term is the power at θ_0: $\int h(t) f_{\theta_0}(t)\, dt = W_{\phi^*}(\theta_0)$.
   2. The second term integrates the likelihood ratio: $\int \frac{f_{\theta_1}(t)}{f_{\theta_0}(t)} f_{\theta_0}(t)\, dt = \int f_{\theta_1}(t)\, dt = 1$.

   Substituting these back, we get:

   $$W_{\phi^*}(\theta_1) \geq W_{\phi^*}(\theta_0) \cdot 1 \tag{5.53}$$

   Thus, $W_{\phi^*}(\theta)$ is non-decreasing for any $\theta_1 > \theta_0$.

2. **Size Control**

   For the composite null $H_0 : \theta \leq \theta^*$, we require the size to be at most α. Since $W_{\phi^*}(\theta)$ is non-decreasing (established in Step 1) and we explicitly set $W_{\phi^*}(\theta^*) = \alpha$:

   $$W_{\phi^*}(\theta) \leq W_{\phi^*}(\theta^*) = \alpha \quad \text{for all } \theta \leq \theta^* \tag{5.54}$$

   This confirms ϕ^* is a valid level-α test.

3. **Uniformly Most Powerful (UMP) via Neyman-Pearson Lemma**

   Let $\phi'(x)$ be any other valid test of size α for H_0. This implies its power at the boundary satisfies $W_{\phi'}(\theta^*) \leq \alpha$.

   Consider any specific alternative $\theta_1 > \theta^*$. Because the family has MLR, the likelihood ratio $\Lambda(x) = \frac{f_{\theta_1}(x)}{f_{\theta^*}(x)}$ is increasing in $T(x)$. Therefore, the test ϕ^* (which rejects for large T) is identified by the Neyman-Pearson Lemma as the Most Powerful (MP) test for the simple hypotheses θ^* vs θ_1.

   Comparing the power of ϕ^* and ϕ' at this specific alternative θ_1:

$$W_{\phi^*}(\theta_1) \geq W_{\phi'}(\theta_1) \tag{5.55}$$

   Since θ_1 was an arbitrary value strictly greater than θ^*, this inequality holds for all $\theta > \theta^*$. Thus, ϕ^* is the UMP test.

□

Example 5.4 (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:

$$\frac{L(\theta_1)}{L(\theta_0)} = \frac{\theta_1^{-n} e^{-\sum x_i/\theta_1}}{\theta_0^{-n} e^{-\sum x_i/\theta_0}} = \left(\frac{\theta_0}{\theta_1}\right)^n \exp\left\{\left(\frac{1}{\theta_0} - \frac{1}{\theta_1}\right)\sum x_i\right\} \tag{5.56}$$

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 $\sum x_i > C$.
$T \sim \text{Gamma}(n, \theta)$
This test form does not depend on the specific θ_1, so it is UMP for all $\theta > \theta_0$.

5.4 Non-Existence of UMP for Two-Sided Hypotheses

For testing a point null hypothesis $H_0 : \theta = \theta^*$ against a two-sided alternative $H_1 : \theta \neq \theta^*$ in a family with a monotone likelihood ratio (e.g., Normal, Exponential), a Uniformly Most Powerful (UMP) test generally **does not exist**. The non-existence proof relies on the uniqueness of the Most Powerful (MP) test derived from the Neyman-Pearson Lemma:

1. **Conflict of Optimal Regions:** Consider a specific alternative $\theta_1 > \theta^*$. By the Neyman-Pearson Lemma, the MP test ϕ_1 rejects H_0 for large values of the sufficient statistic $T(\mathbf{X}) > k_1$. Conversely, consider an alternative $\theta_2 < \theta^*$. The MP test ϕ_2 rejects H_0 for small values of the statistic $T(\mathbf{X}) < k_2$.
2. **Failure of Uniformity:** A UMP test ϕ^* would need to be the MP test for *every* $\theta \in H_1$.
   - For ϕ^* to be most powerful against θ_1, it must be equivalent to ϕ_1 (rejecting in the right tail).
   - For ϕ^* to be most powerful against θ_2, it must be equivalent to ϕ_2 (rejecting in the left tail).

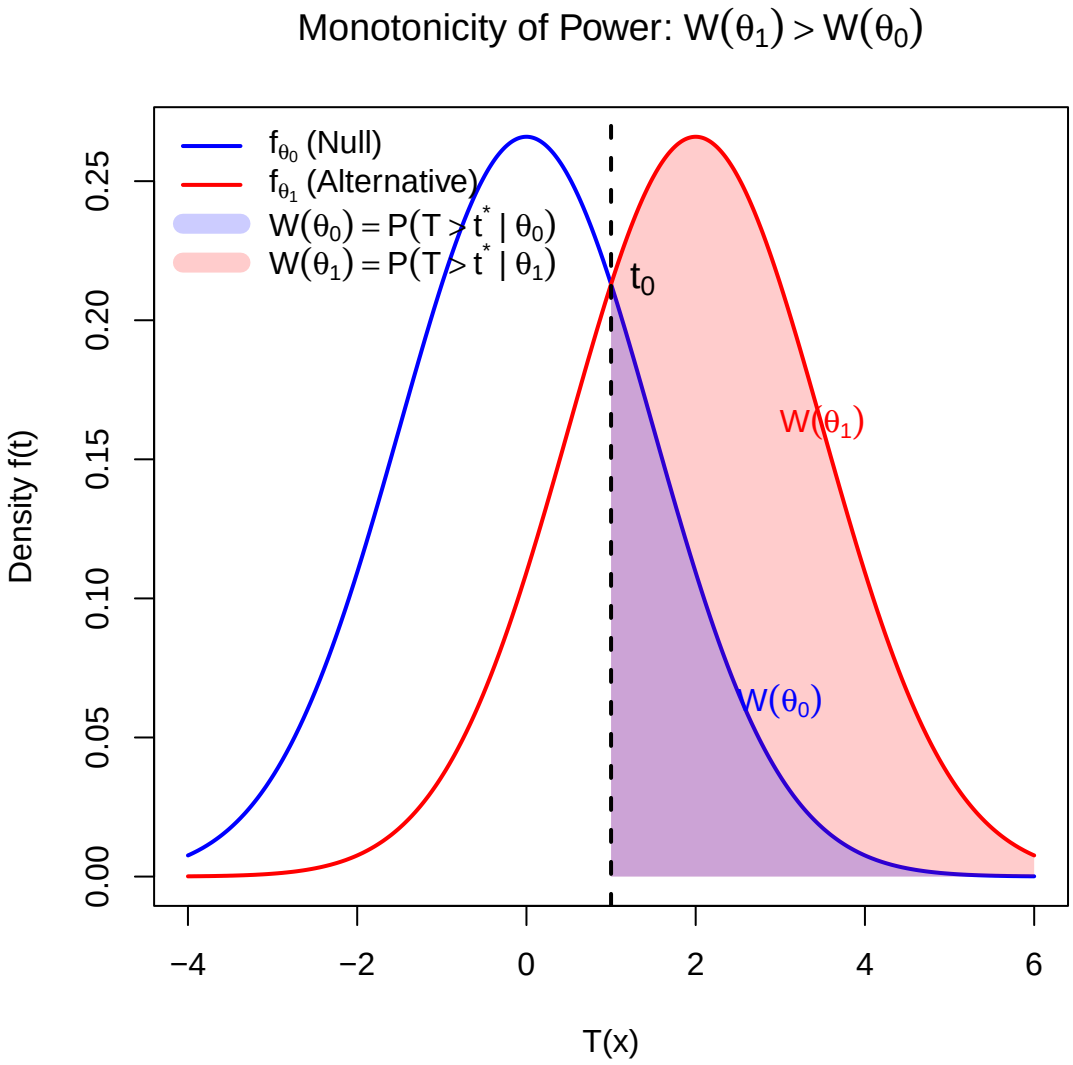


Figure 5.5: Visualizing Monotonicity of Power Function $W(\theta)$ for MLR Distributions

3. **Biased Power Function:** The MP test ϕ_1 (Right-Sided) has a power function that drops below the size α for values $\theta < \theta^*$. Therefore, it cannot be the most powerful test for θ_2, as there exists a valid test (e.g., ϕ_2) with power strictly greater than α at θ_2.

Since no single critical region can simultaneously maximize power for both $\theta > \theta^*$ and $\theta < \theta^*$, no UMP test exists. We typically restrict our search to **Unbiased** tests (UMPU) to resolve this.

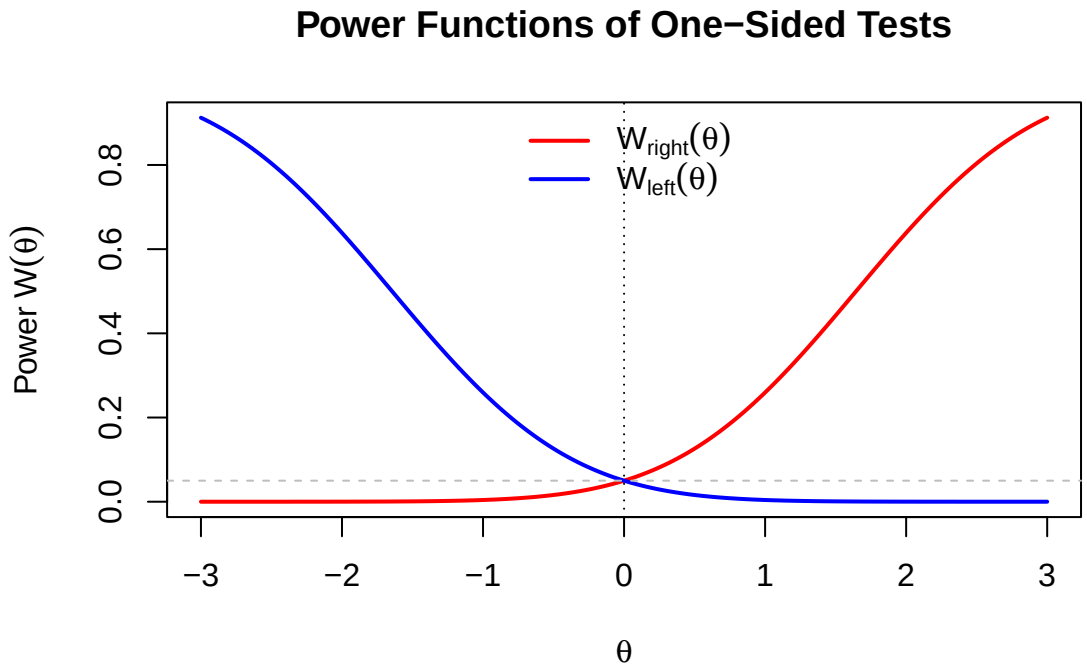


Figure 5.6: Illustration of the Conflict of Optimal Regions: Power functions for the Right-Tailed (red) and Left-Tailed (blue) UMP tests.

Example 5.5 (Non-Existence of UMP for the Normal Mean)**.** Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with known variance σ^2. We wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at significance level α.
Because the Normal distribution belongs to the exponential family, the sufficient statistic $\overline{X}$ possesses the Monotone Likelihood Ratio (MLR) property with respect to μ. By the Karlin-Rubin Theorem, this guarantees the existence of UMP tests for one-sided alternatives:

- **Right-Tailed UMP (ϕ_R):** For $H_1 : \mu > \mu_0$, the UMP test of size α rejects H_0 when $\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma} > z_{1-\alpha}$.
- **Left-Tailed UMP (ϕ_L):** For $H_1 : \mu < \mu_0$, the UMP test of size α rejects H_0 when $\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma} < -z_{1-\alpha}$.

Now consider the standard **Two-Sided Test (ϕ_{Two})** of size α, which splits the rejection region to cover both directions. It rejects when $\left|\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma}\right| > z_{1-\alpha/2}$.
The Power Deficit: Because $z_{1-\alpha/2} > z_{1-\alpha}$ (e.g., for $\alpha = 0.05$, $1.96 > 1.645$), the two-sided test requires more extreme evidence to reject in either specific direction.

- For any true mean $\mu > \mu_0$, the two-sided test has **strictly lower power** than the right-tailed UMP test.
- For any true mean $\mu < \mu_0$, the two-sided test has **strictly lower power** than the left-tailed UMP test.

Since ϕ_{Two} is beaten by ϕ_R for positive shifts and beaten by ϕ_L for negative shifts, no single test is "uniformly" most powerful across the entire alternative space $H_1 : \mu \neq \mu_0$.

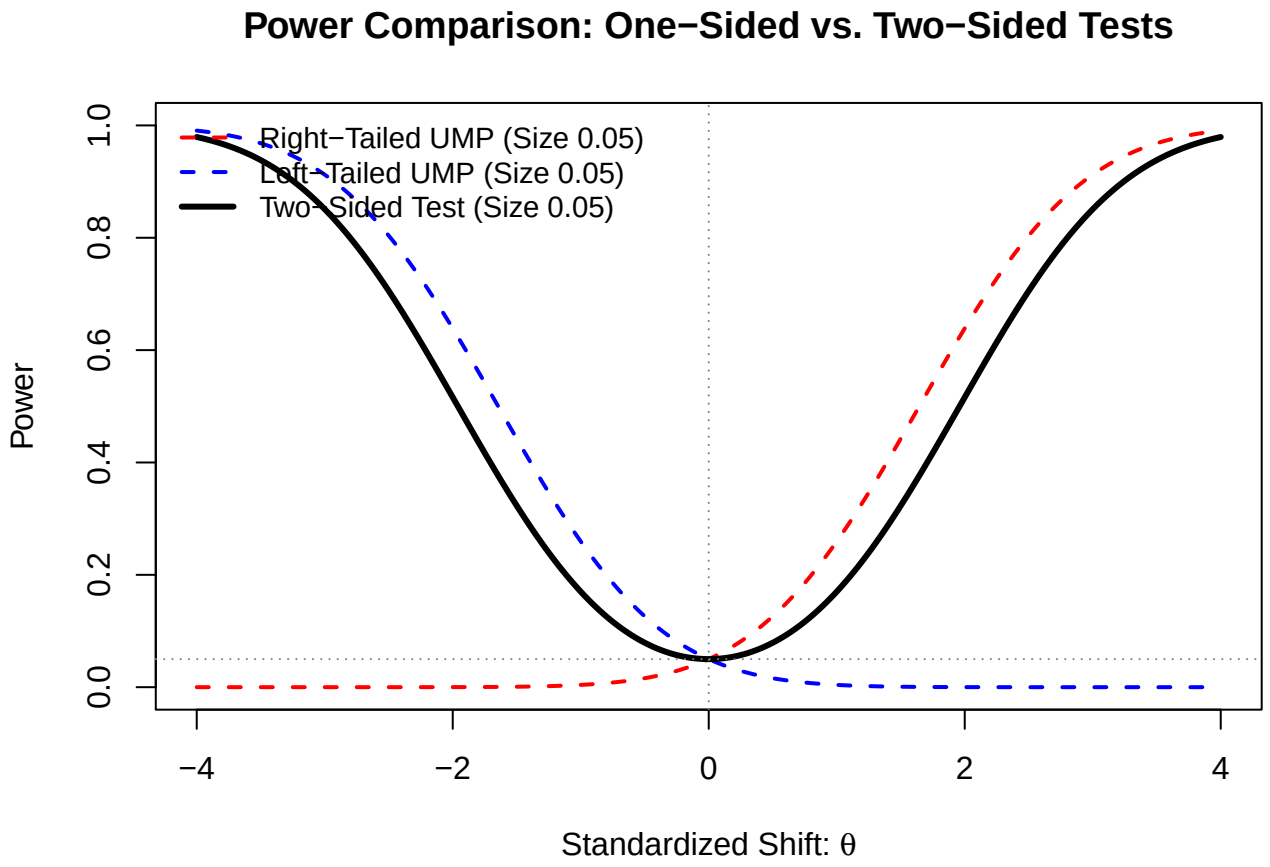


Figure 5.7: Power functions of size alpha=0.05 tests for a Normal mean. The two-sided test (black) is everywhere less powerful than the respective optimal one-sided test (red or blue) for any true difference.

6 Likelihood-based Tests

6.1 The Geometry of Mahalanobis Distance

6.1.1 Pythagorean Theorem for Mahalanobis Distance

Definitions and Setup

Let x be a random vector partitioned as $x = (x_0, x_1)^T$, and let μ be its corresponding center vector $\mu = (\mu_0, \mu_1)^T$. Let Σ be a symmetric, positive-definite covariance matrix governing the joint space, partitioned into corresponding blocks:

$$\Sigma = \begin{pmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{pmatrix} \tag{6.1}$$

Define the generalized squared Mahalanobis distance of any vector v to a center c with respect to a positive-definite matrix S as:

$$D^2(v, c, S) = (v - c)^T S^{-1} (v - c) \tag{6.2}$$

For a symmetric, positive-definite matrix S, the Mahalanobis inner product between two vectors u and v is defined as $\langle u, v \rangle_{S^{-1}} = u^T S^{-1} v$. Two vectors are orthogonal in this space if their inner product is zero.

Define the conditional expectation (the linear regression surface) of x_1 given x_0 as $\mu_{1|0}(x_0)$:

$$\mu_{1|0}(x_0) = \mu_1 + \Sigma_{10}\Sigma_{00}^{-1}(x_0 - \mu_0) \tag{6.3}$$

Define the conditional covariance matrix (the Schur complement of Σ_{00} in Σ) as $\Sigma_{1|0}$:

$$\Sigma_{1|0} = \Sigma_{11} - \Sigma_{10}\Sigma_{00}^{-1}\Sigma_{01} \tag{6.4}$$

Finally, define the **Projection Point ($\tilde{x}$)** as the point where x projects exactly onto the regression surface: $\tilde{x} = (x_0, \tilde{x}_1)^T$, where $\tilde{x}_1 = \mu_{1|0}(x_0)$.

Lemma 6.1 (Fundamental Orthogonal Decomposition Lemma)**.** *Any deviation from the center, $(x - \mu)$, can be expressed as the sum of a conditional residual vector $(x - \tilde{x})$ and a marginal projection vector $(\tilde{x} - \mu)$.*

These two vectors are strictly orthogonal with respect to the precision matrix Σ^{-1}:

$$(x - \tilde{x})^T \Sigma^{-1} (\tilde{x} - \mu) = 0 \tag{6.5}$$

Furthermore, the joint Mahalanobis distances of these individual vectors perfectly isolate the conditional and marginal spaces:

1. $(x - \tilde{x})^T \Sigma^{-1} (x - \tilde{x}) = D^2(x_1, \tilde{x}_1, \Sigma_{1|0})$
2. $(\tilde{x} - \mu)^T \Sigma^{-1} (\tilde{x} - \mu) = D^2(x_0, \mu_0, \Sigma_{00})$

Proof. **Proof via Block Factorization**
We can express the partitioned covariance matrix Σ using a block LDL^T factorization. The precision matrix is then $\Sigma^{-1} = (L^{-1})^T D^{-1} L^{-1}$, where:

$$L^{-1} = \begin{pmatrix} I & 0 \\ -\Sigma_{10}\Sigma_{00}^{-1} & I \end{pmatrix}, \quad D^{-1} = \begin{pmatrix} \Sigma_{00}^{-1} & 0 \\ 0 & \Sigma_{1|0}^{-1} \end{pmatrix} \tag{6.6}$$

Evaluating Mahalanobis inner products with Σ^{-1} is equivalent to applying the shearing transformation L^{-1} to the vectors, and then taking their inner product with respect to the block-diagonal metric D^{-1}. Let us evaluate the transformed vectors for the residual $u = (x - \tilde{x})$ and the projection $v = (\tilde{x} - \mu)$.

1. Transforming the Residual Vector ($L^{-1}u$):

$$L^{-1}(x - \tilde{x}) = \begin{pmatrix} I & 0 \\ -\Sigma_{10}\Sigma_{00}^{-1} & I \end{pmatrix} \begin{pmatrix} 0 \\ x_1 - \tilde{x}_1 \end{pmatrix} = \begin{pmatrix} 0 \\ x_1 - \tilde{x}_1 \end{pmatrix} \tag{6.7}$$

The transformation leaves the residual vector unchanged; it lies entirely in the conditional subspace.

2. Transforming the Projection Vector ($L^{-1}v$):

$$L^{-1}(\tilde{x} - \mu) = \begin{pmatrix} I & 0 \\ -\Sigma_{10}\Sigma_{00}^{-1} & I \end{pmatrix} \begin{pmatrix} x_0 - \mu_0 \\ \Sigma_{10}\Sigma_{00}^{-1}(x_0 - \mu_0) \end{pmatrix} = \begin{pmatrix} x_0 - \mu_0 \\ 0 \end{pmatrix} \tag{6.8}$$

The transformation cleanly removes the induced correlation, flattening the projection vector so it lies entirely in the marginal subspace.
Evaluating the Inner Products: Because D^{-1} is block-diagonal, the inner product of these transformed vectors is simply the sum of the inner products of their corresponding blocks:

$$(x - \tilde{x})^T \Sigma^{-1} (\tilde{x} - \mu) = \begin{pmatrix} 0 \\ x_1 - \tilde{x}_1 \end{pmatrix}^T \begin{pmatrix} \Sigma_{00}^{-1} & 0 \\ 0 & \Sigma_{1|0}^{-1} \end{pmatrix} \begin{pmatrix} x_0 - \mu_0 \\ 0 \end{pmatrix} = 0 + 0 = 0 \tag{6.9}$$

This proves strict orthogonality.
Applying the same logic to the self-inner products evaluates the quadratic forms:

$$(x - \tilde{x})^T \Sigma^{-1} (x - \tilde{x}) = \begin{pmatrix} 0 \\ x_1 - \tilde{x}_1 \end{pmatrix}^T \begin{pmatrix} \Sigma_{00}^{-1} & 0 \\ 0 & \Sigma_{1|0}^{-1} \end{pmatrix} \begin{pmatrix} 0 \\ x_1 - \tilde{x}_1 \end{pmatrix} = (x_1 - \tilde{x}_1)^T \Sigma_{1|0}^{-1} (x_1 - \tilde{x}_1) \tag{6.10}$$

$$(\tilde{x}-\mu)^T\Sigma^{-1}(\tilde{x}-\mu) = \begin{pmatrix} x_0-\mu_0 \\ 0 \end{pmatrix}^T \begin{pmatrix} \Sigma_{00}^{-1} & 0 \\ 0 & \Sigma_{1|0}^{-1} \end{pmatrix} \begin{pmatrix} x_0-\mu_0 \\ 0 \end{pmatrix} = (x_0-\mu_0)^T\Sigma_{00}^{-1}(x_0-\mu_0) \quad (6.11)$$

which yields exactly $D^2(x_1, \tilde{x}_1, \Sigma_{1|0})$ and $D^2(x_0, \mu_0, \Sigma_{00})$, respectively. □

Theorem 6.1 (The Generalized Pythagorean Decomposition for Mahalanobis Distance)**.** *For any point x, the total joint squared Mahalanobis distance from x to the center μ decomposes precisely into the sum of the squared distances between x and its projection $\tilde{x}$ (the Conditional Residual), and between $\tilde{x}$ and the center μ (the Marginal Projection):*

$$D^2(x, \mu, \Sigma) = D^2(x, \tilde{x}, \Sigma) + D^2(\tilde{x}, \mu, \Sigma) \quad (6.12)$$

which, mapping to the marginal and conditional subspaces, is equivalent to:

$$D^2(x, \mu, \Sigma) = D^2(x_1, \tilde{x}_1, \Sigma_{1|0}) + D^2(x_0, \mu_0, \Sigma_{00}) \quad (6.13)$$

Proof. **Proof**
We construct the vector from the origin to the target as the sum of the residual and the projection:

$$x - \mu = (x - \tilde{x}) + (\tilde{x} - \mu) \quad (6.14)$$

We evaluate the total joint distance by expanding the quadratic form:

$$D^2(x, \mu, \Sigma) = \Big((x - \tilde{x}) + (\tilde{x} - \mu)\Big)^T \Sigma^{-1} \Big((x - \tilde{x}) + (\tilde{x} - \mu)\Big) \quad (6.15)$$

$$= (x - \tilde{x})^T\Sigma^{-1}(x - \tilde{x}) + (\tilde{x} - \mu)^T\Sigma^{-1}(\tilde{x} - \mu) + 2(x - \tilde{x})^T\Sigma^{-1}(\tilde{x} - \mu) \quad (6.16)$$

By the **Fundamental Orthogonal Decomposition Lemma**, the cross-term evaluates to zero due to orthogonality, leaving only the isolated quadratic forms:

$$D^2(x, \mu, \Sigma) = (x - \tilde{x})^T\Sigma^{-1}(x - \tilde{x}) + (\tilde{x} - \mu)^T\Sigma^{-1}(\tilde{x} - \mu) \quad (6.17)$$

This is exactly $D^2(x, \tilde{x}, \Sigma) + D^2(\tilde{x}, \mu, \Sigma)$. Furthermore, as established by the Lemma, substituting the subspace identities for these two terms yields $D^2(x_1, \tilde{x}_1, \Sigma_{1|0}) + D^2(x_0, \mu_0, \Sigma_{00})$. □

6.1.2 Interactive Illustration

```
#| '!! shinylive warning !!': |
#|   shinylive does not work in self-contained HTML documents.
#|   Please set `embed-resources: false` in your metadata.
#| standalone: true
```

```
#| viewerHeight: 750
#| echo: false
#| column: screen
#| label: fig-MD-decomp-unified

library(shiny)
library(ggplot2)

ui <- fluidPage(
  titlePanel("Pythagorean Decomposition of Mahalanobis Distance"),

  # A relative container holding the plot and absolute panels
  div(style = "position: relative;",

      plotOutput("distPlot", height = "700px", width = "100%"),

      # Floating Control Panel (Top Left)
      absolutePanel(
        top = 10, left = 10, width = 320, draggable = TRUE,
        style = "background-color: rgba(255, 255, 255, 0.85); padding: 10px 15px; border-radius:

        h4("Parameters", style = "margin-top: 0; margin-bottom: 10px; font-size: 16px;"),

        fluidRow(
          column(6, sliderInput("mu_0", "\\(\\mu_0\\)", min = 0, max = 5, value = 2.2, step = 0.1
          column(6, sliderInput("mu_1", "\\(\\mu_1\\)", min = 0, max = 5, value = 2.2, step = 0.1
        ),
        fluidRow(
          column(6, sliderInput("x_0", "\\(x_0\\)", min = 0, max = 5, value = 0.9, step = 0.1, ti
          column(6, sliderInput("x_1", "\\(x_1\\)", min = 0, max = 5, value = 1.9, step = 0.1, ti
        ),
        hr(style = "margin: 5px 0; border-top: 1px solid #ddd;"),
        fluidRow(
          column(6, sliderInput("sigma_0", "\\(\\sigma_0\\)", min = 0.5, max = 3, value = 1.25, s
          column(6, sliderInput("sigma_1", "\\(\\sigma_1\\)", min = 0.5, max = 3, value = 1.25, s
        ),
        sliderInput("rho", "Correlation \\(\\rho\\)", min = -0.95, max = 0.95, value = -0.6, step
      ),

      # Floating Legend Panel (Bottom Right)
      absolutePanel(
        bottom = 20, right = 20, width = 340, draggable = TRUE,
        style = "background-color: rgba(255, 255, 255, 0.9); padding: 10px; border-radius: 8px; b
        uiOutput("legendUI")
      )
```

```
  )
)

server <- function(input, output, session) {

  # Reactive calculations
  calc_data <- reactive({
    mu <- c(input$mu_0, input$mu_1)
    x <- c(input$x_0, input$x_1)

    # Construct Covariance Matrix
    Sigma <- matrix(c(
      input$sigma_0^2,
      input$rho * input$sigma_0 * input$sigma_1,
      input$rho * input$sigma_0 * input$sigma_1,
      input$sigma_1^2
    ), nrow = 2)

    Sigma_inv <- solve(Sigma)

    # Marginal and Conditional variances
    Sigma_00 <- Sigma[1, 1]
    Sigma_1given0 <- Sigma[2, 2] - (Sigma[1, 2]^2) / Sigma[1, 1]

    # Conditional Mean
    mu_1given0 <- mu[2] + (Sigma[1, 2] / Sigma[1, 1]) * (x[1] - mu[1])

    # Projected Point p = \tilde{x} = (x_0, \mu_{1|0}(x_0))
    x_tilde <- c(x[1], mu_1given0)

    # Calculate Distances
    diff_total <- x - mu
    d2_total <- as.numeric(t(diff_total) %*% Sigma_inv %*% diff_total)

    # By theorem, marginal maps to D^2(\tilde{x}, \mu, \Sigma)
    d2_marg <- (x[1] - mu[1])^2 / Sigma_00

    # By theorem, conditional maps to D^2(x, \tilde{x}, \Sigma)
    d2_cond <- (x[2] - mu_1given0)^2 / Sigma_1given0

    list(mu = mu, x = x, x_tilde = x_tilde, Sigma = Sigma, Sigma_inv = Sigma_inv,
         d2_total = d2_total, d2_marg = d2_marg, d2_cond = d2_cond)
  })

  output$legendUI <- renderUI({
```

```
  res <- calc_data()
  withMathJax(
    HTML(paste0(
      "<div style='font-size: 14px;'>",
      "<b>Distance Legend</b><br><br>",
      "<span style='color: black;'><b>Total:</b> \\( D^2(x, \\mu, \\Sigma) = \\) ", round(res$d
      "<span style='color: blue;'><b>Marginal:</b> \\( D^2(\\tilde{x}, \\mu, \\Sigma) = \\) ",
      "<span style='color: red;'><b>Conditional:</b> \\( D^2(x, \\tilde{x}, \\Sigma) = \\) ", r
      "</div>"
    ))
  )
})

output$distPlot <- renderPlot({
  res <- calc_data()
  mu <- res$mu; x <- res$x; x_tilde <- res$x_tilde; Sigma <- res$Sigma; Sigma_inv <- res$Sigma_

  # Generate grid for quadratic contours (extended for fixed bounds)
  grid_vals <- seq(-2, 7, length.out = 100)
  grid <- expand.grid(x0 = grid_vals, x1 = grid_vals)
  grid$z <- apply(grid, 1, function(row) {
    diff <- row - mu
    as.numeric(t(diff) %*% Sigma_inv %*% diff)
  })

  # Regression surface line parameters
  slope <- Sigma[1, 2] / Sigma[1, 1]
  intercept <- mu[2] - slope * mu[1]

  # Midpoints for labels
  mid_marg <- (mu + x_tilde) / 2
  mid_cond <- (x_tilde + x) / 2
  mid_total <- (mu + x) / 2

  ggplot(grid, aes(x = x0, y = x1)) +
    # Contours
    geom_contour(aes(z = z), bins = 25, color = "gray85") +

    # The Regression Surface (Light Blue Line extending across plot)
    geom_abline(intercept = intercept, slope = slope, color = "lightblue", linewidth = 2, alpha

    # The Distance Segments
    geom_segment(aes(x = mu[1], y = mu[2], xend = x_tilde[1], yend = x_tilde[2]), color = "blue
    geom_segment(aes(x = x_tilde[1], y = x_tilde[2], xend = x[1], yend = x[2]), color = "red",
    geom_segment(aes(x = mu[1], y = mu[2], xend = x[1], yend = x[2]), color = "black", linetype
```

```
      # Labels on Segments using unified notation
      annotate("text", x = mid_marg[1] + 0.25, y = mid_marg[2] - 0.25, label = "D^2*(list(tilde(x
      annotate("text", x = mid_cond[1] + 0.35, y = mid_cond[2], label = "D^2*(list(x, tilde(x), S
      annotate("text", x = mid_total[1] - 0.25, y = mid_total[2] + 0.25, label = "D^2*(list(x, mu

      # Points and their labels
      annotate("point", x = mu[1], y = mu[2], color = "blue", size = 4) +
      annotate("text", x = mu[1] - 0.2, y = mu[2] + 0.2, label = "mu", parse = TRUE, color = "blu

      annotate("point", x = x_tilde[1], y = x_tilde[2], color = "purple", shape = 3, size = 4, st
      annotate("text", x = x_tilde[1] + 0.3, y = x_tilde[2] - 0.2, label = "tilde(x)", parse = TR

      annotate("point", x = x[1], y = x[2], color = "black", shape = 4, size = 4, stroke = 2) +
      annotate("text", x = x[1] - 0.15, y = x[2] + 0.2, label = "x", parse = TRUE, size = 6) +

      # Strict Fixed Coordinates to prevent re-shaping
      coord_fixed(xlim = c(-1, 6), ylim = c(-1, 6), expand = FALSE) +
      labs(x = expression(x[0]), y = expression(x[1])) +
      theme_minimal(base_size = 16) +
      theme(panel.grid.minor = element_blank())
  })
}

shinyApp(ui, server)
```

6.2 Likelihood Ratio Test for General Nested Models

Theorem 6.2 (Wilks' Theorem for General Nested Models)**.** *Let* $\mathbf{Y} = (Y_1, \dots, Y_n)$ *be an i.i.d. sample from a distribution* $f(\mathbf{y}; \theta)$ *with* $\theta \in \Theta \subseteq \mathbb{R}^p$*. Let* Θ_0 *be a lower-dimensional subspace of* Θ *defined by* q *independent constraints, such that* $\dim(\Theta_0) = p - q$*.*
Consider testing the nested hypotheses:

$$H_0 : \theta_0 = \theta_0^* \quad vs \quad H_1 : \theta_0 \neq \theta_0^* \tag{6.18}$$

Let $\theta^* = (\theta_0^*, \theta_1^*)^T$ *be the true parameter vector generating the data under* H_0*. Let* $\hat{\theta}$ *be the unrestricted Maximum Likelihood Estimator (MLE) over* Θ*, and let* $\tilde{\theta} = (\theta_0^*, \tilde{\theta}_1)^T$ *be the restricted MLE over* Θ_0*.*
The likelihood ratio test statistic D *is defined as the difference between the unrestricted and restricted maximum log-likelihoods:*

$$D = 2\left[\ell(\hat{\theta}) - \ell(\tilde{\theta})\right] \tag{6.19}$$

Under H_0*, as* $n \to \infty$*, the statistic* D *converges in distribution to a chi-square distribution with degrees of freedom equal to the number of constraints:*

$$D \xrightarrow{d} \chi^2_q \tag{6.20}$$

Proof.

1. Mapping Likelihood to Mahalanobis Geometry

By a second-order Taylor expansion around the unrestricted peak $\hat{\theta}$, log-likelihood deviances are asymptotically equivalent to squared Mahalanobis distances. The center of this space is the unrestricted MLE, and the metric is the inverse Fisher Information matrix $\Sigma = \mathcal{J}^{-1}$.
We map the likelihood parameters directly to our geometric definitions $(x, \mu, \tilde{x})$:

- **The Origin ($\mu \to \hat{\theta}$):** The global MLE sits at the center of the quadratic contours.
- **The Target ($x \to \theta^*$):** The true parameter value being evaluated.
- **The Projection ($\tilde{x} \to \tilde{\theta}$):** The restricted MLE maximizes the likelihood given θ_0^*. Geometrically, it minimizes the distance to the center $\hat{\theta}$ subject to the null constraint, which strictly defines it as the projection of x onto the regression surface: $\tilde{\theta}_1 = \mu_{1|0}(\theta_0^*)$.

2. Evaluating the Deviances

Using this mapping, we can express the total deviance (D_1), the restricted deviance (D_0), and the Wilks statistic (D) strictly as geometric distances:

- **Total Deviance:** $D_1 = 2[\ell(\hat{\theta}) - \ell(\theta^*)] \approx D^2(\theta^*, \hat{\theta}, \Sigma) \iff D^2(x, \mu, \Sigma)$
- **Restricted Deviance:** $D_0 = 2[\ell(\tilde{\theta}) - \ell(\theta^*)] \approx D^2(\theta^*, \tilde{\theta}, \Sigma) \iff D^2(x, \tilde{x}, \Sigma)$
- **Wilks Statistic:** $D = 2[\ell(\hat{\theta}) - \ell(\tilde{\theta})] \approx D^2(\tilde{\theta}, \hat{\theta}, \Sigma) \iff D^2(\tilde{x}, \mu, \Sigma)$

3. The Pythagorean Identity and Marginal Decomposition

By the **Generalized Pythagorean Decomposition**, we know the distances strictly relate via:

$$D^2(x, \mu, \Sigma) = D^2(x, \tilde{x}, \Sigma) + D^2(\tilde{x}, \mu, \Sigma) \tag{6.21}$$

which confirms the algebraic identity of the likelihoods: $D_1 \approx D_0 + D$.
Furthermore, Wilks' statistic D is precisely the **Marginal Projection** component: $D^2(\tilde{x}, \mu, \Sigma)$. According to the **Fundamental Orthogonal Decomposition Lemma**, this projection perfectly isolates the marginal distance of the constrained block:

$$D \approx D^2(\tilde{\theta}, \hat{\theta}, \Sigma) = D^2(\theta_0^*, \hat{\theta}_0, \Sigma_{00}) = (\hat{\theta}_0 - \theta_0^*)^T \Sigma_{00}^{-1} (\hat{\theta}_0 - \theta_0^*) \tag{6.22}$$

4. Asymptotic Distribution

Under H_0, the asymptotic normality of the unrestricted MLE implies $\sqrt{n}(\hat{\theta}_0 - \theta_0^*) \xrightarrow{d} N(\mathbf{0}, \Sigma_{00})$. Therefore, the isolated quadratic form D follows a chi-square distribution with q degrees of freedom:

$$D \xrightarrow{d} \chi^2_q \tag{6.23}$$

□

Remark (Geometric Intuition). By centering the Mahalanobis space at the unrestricted MLE $\hat{\theta}$, the proof of Wilks' Theorem collapses into a trivial geometric observation. The restricted MLE $\tilde{\theta}$ is simply the projection of the true parameter θ^* onto the regression surface. The Wilks statistic D measures the squared length of the projection leg of the resulting right triangle. Because it is a projection onto the marginal subspace, it depends *only* on the parameters of interest, entirely stripping away the nuisance parameters.

Diagram to Illustrate Wilks' Theorem

The following R code visualizes the likelihood ratio test as a projection. We illustrate the unrestricted MLE $\hat{\theta}$ (the center of the contours), the hypothesized true value θ^*, the geometric center $\hat{\theta}_{1,\Sigma}$, and a jittered point representing the actual restricted MLE $\tilde{\theta}_1$.

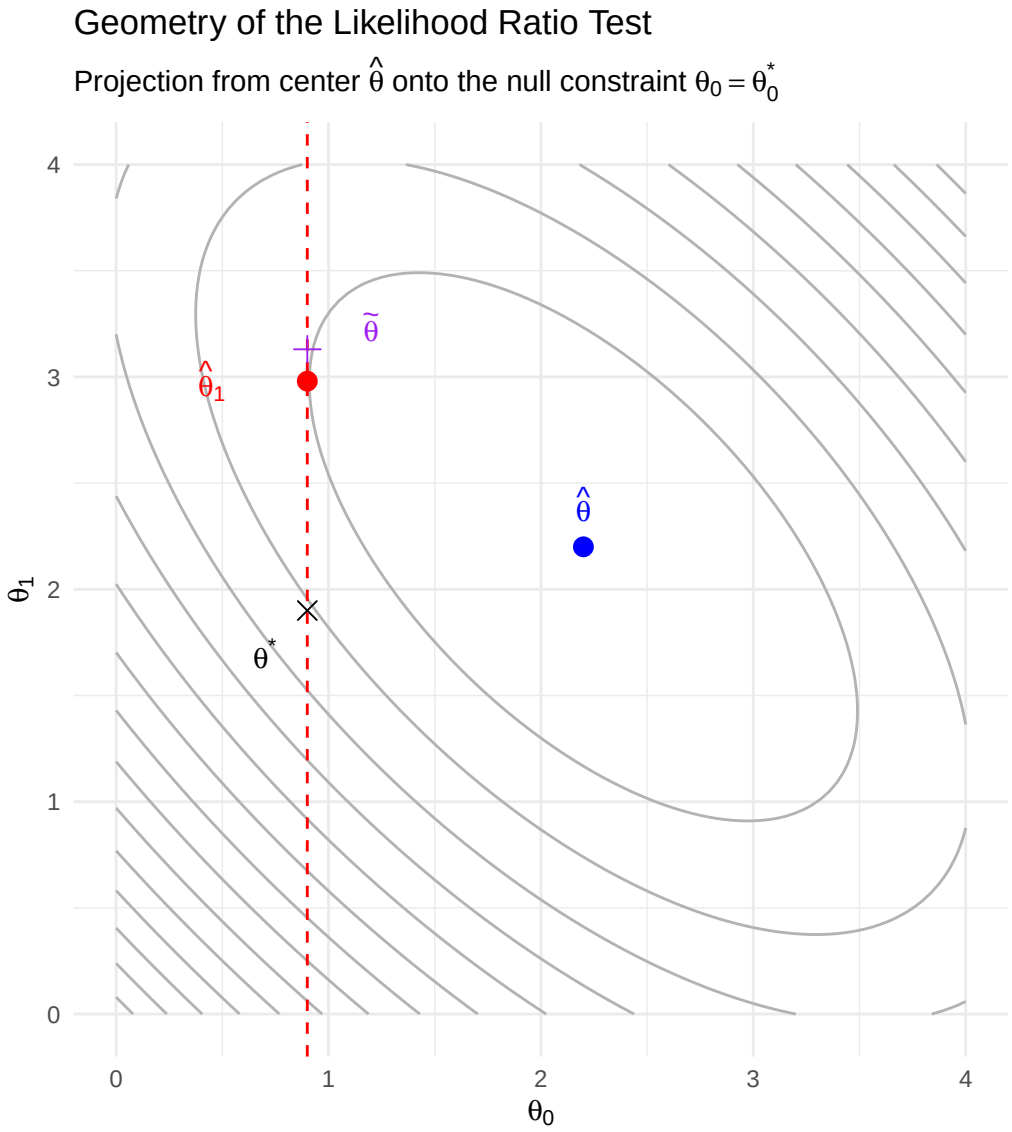


Figure 6.1: The geometry of restricted optimization. Blue dot is the unrestricted MLE (center). The red dot represents the projected nuisance parameter center on the null line. The purple cross is the restricted MLE.

Example 6.1. LRT for Normal Mean with Unknown Variance
Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. The parameter vector is $\theta = (\mu, \sigma^2)$. We wish to test:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0 \tag{6.24}$$

1. **Find the Maximum Likelihood Estimators:** The log-likelihood function is $\ell(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$.

   - **Unrestricted MLEs ($\hat{\theta}$):** $\hat{\mu} = \overline{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. The maximized log-likelihood is:

   $$\ell(\hat{\mu}, \hat{\sigma}^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{n}{2} \tag{6.25}$$

   - **Restricted MLEs ($\hat{\theta}_0$):** Under $H_0, \hat{\mu}_0 = \mu_0, \quad \hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2$. The restricted maximized log-likelihood is:

   $$\ell(\hat{\mu}_0, \hat{\sigma}_0^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_0^2) - \frac{n}{2} \tag{6.26}$$

2. **Define the Test Statistic D:** The test statistic D is defined as twice the difference between the unrestricted and restricted log-likelihoods:

$$D = 2\left[\ell(\hat{\mu}, \hat{\sigma}^2) - \ell(\hat{\mu}_0, \hat{\sigma}_0^2)\right] \tag{6.27}$$

Substituting the expressions above, the constant terms cancel out:

$$D = 2\left[-\frac{n}{2}\ln(\hat{\sigma}^2) - \left(-\frac{n}{2}\ln(\hat{\sigma}_0^2)\right)\right] = n\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right) \tag{6.28}$$

3. **Relate to the t-statistic:** Using the sum of squares decomposition $\hat{\sigma}_0^2 = \hat{\sigma}^2 + (\overline{X} - \mu_0)^2$, we have:

$$\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} = 1 + \frac{(\overline{X} - \mu_0)^2}{\hat{\sigma}^2} \tag{6.29}$$

Substituting the t-statistic $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$, where $S^2 = \frac{n}{n-1}\hat{\sigma}^2$, we find:

$$\frac{(\overline{X} - \mu_0)^2}{\hat{\sigma}^2} = \frac{T^2}{n-1} \tag{6.30}$$

Thus, D is a monotone increasing function of T^2:

$$D = n\ln\left(1 + \frac{T^2}{n-1}\right) \tag{6.31}$$

Rejecting H_0 for large D is equivalent to rejecting for large $|T|$.

4. **Asymptotic Distribution (Wilks' Theorem):** By Wilks' Theorem, $D \xrightarrow{d} \chi_1^2$. For large n, using the approximation $\ln(1 + x) \approx x$:

$$D = n\ln\left(1 + \frac{T^2}{n-1}\right) \approx n\left(\frac{T^2}{n-1}\right) \approx T^2 \tag{6.32}$$

As $n \to \infty$, $T^2 \xrightarrow{d} \chi_1^2$, confirming the theorem.

6.3 Wald's Theorem for Testing Parameter Restrictions

The Wald Test evaluates the distance between the unrestricted Maximum Likelihood Estimator (MLE) and the null hypothesis value, standardized by the curvature of the log-likelihood at the unrestricted estimate. Unlike the Score test, the Wald test requires estimating the full, unrestricted model.

Theorem 6.3 (Asymptotic Distribution of the Wald Statistic)**.** *Let $\mathbf{Y} = (Y_1, \dots, Y_n)$ be an i.i.d. sample from a regular family of distributions with parameters $\theta \in \Theta \subseteq \mathbb{R}^p$. Partition the parameter vector as $\theta = (\theta_0^T, \theta_1^T)^T$, where θ_0 is the q-dimensional parameter of interest and θ_1 is the $(p-q)$-dimensional nuisance parameter. Consider the test:*

$$H_0 : \theta_0 = \theta_0^* \quad vs \quad H_1 : \theta_0 \neq \theta_0^* \tag{6.33}$$

Let $\hat{\theta} = (\hat{\theta}_0^T, \hat{\theta}_1^T)^T$ be the unrestricted MLE under H_1. Let $\Sigma(\theta) = \mathcal{I}^{-1}(\theta)$ be the asymptotic covariance matrix of the unrestricted MLE, where $\mathcal{I}(\theta)$ is the expected Fisher Information matrix, and let Σ_{00} denote its top-left $q \times q$ marginal block.
The Wald statistic is defined as the quadratic form:

$$W = (\hat{\theta}_0 - \theta_0^*)^T \Sigma_{00}^{-1}(\hat{\theta})(\hat{\theta}_0 - \theta_0^*) \tag{6.34}$$

Under standard regularity conditions, if H_0 is true, as $n \to \infty$:

$$W \xrightarrow{d} \chi_q^2 \tag{6.35}$$

where q is the dimension of the parameter of interest θ_0.

Proof. **Algebraic Proof via Mahalanobis Decomposition**
We can derive the Wald statistic directly by borrowing the geometric decomposition used in Wilks' Theorem, isolating the marginal behavior of the unrestricted MLE.

1. The Total Mahalanobis Distance

By standard large-sample theory, the unrestricted MLE $\hat{\theta}$ is asymptotically normal. Its asymptotic covariance matrix is $\Sigma = \mathcal{I}^{-1}(\theta)$. The total squared Mahalanobis distance between the unrestricted MLE $\hat{\theta}$ and the full hypothesized parameter vector $\theta^* = ({\theta_0^*}^T, {\theta_1^*}^T)^T$ under the precision metric Σ^{-1} is:

$$D_{total} = (\hat{\theta} - \theta^*)^T \Sigma^{-1} (\hat{\theta} - \theta^*) \tag{6.36}$$

2. The Pythagorean Decomposition (From Wilks' Theorem)

As established in the geometric proof of Wilks' Theorem, any Mahalanobis distance in a partitioned parameter space orthogonally decomposes into a marginal distance for the parameter of interest and a conditional distance for the nuisance parameter:

$$D_{total} = (\hat{\theta}_0 - \theta_0^*)^T \Sigma_{00}^{-1} (\hat{\theta}_0 - \theta_0^*) + (\hat{\theta}_{1,\Sigma} - \theta_1^*)^T \mathcal{I}_{11} (\hat{\theta}_{1,\Sigma} - \theta_1^*) \tag{6.37}$$

where Σ_{00} is the top-left block of Σ, and $\hat{\theta}_{1,\Sigma}$ is the geometric center of the nuisance parameter constraint.

3. Isolating the Marginal Distance (The Wald Statistic)

The Wald test zeroes in exclusively on the first term of this decomposition. Rather than evaluating the full likelihood surface (as Wilks does) or the gradient at the boundary (as the Score test does), the Wald test directly measures the isolated **marginal Mahalanobis distance** of the parameter of interest from its null value:

$$W_{true} = (\hat{\theta}_0 - \theta_0^*)^T \Sigma_{00}^{-1} (\hat{\theta}_0 - \theta_0^*) \tag{6.38}$$

Because the marginal asymptotic distribution of the estimator under the null hypothesis is $\hat{\theta}_0 \overset{a}{\sim} N(\theta_0^*, \Sigma_{00})$, standardizing this deviation by its precision matrix Σ_{00}^{-1} produces a quadratic form that follows a chi-square distribution:

$$W_{true} \xrightarrow{d} \chi_q^2 \tag{6.39}$$

4. Slutsky's Substitution for the Unknown Covariance

In practice, the true marginal covariance Σ_{00} is unknown because it depends on the true parameter θ. However, because the Fisher Information matrix is a continuous function of the parameters, and $\hat{\theta}$ is a consistent estimator of θ under H_1, we can invoke Slutsky's Theorem. Substituting the estimated covariance matrix $\Sigma_{00}(\hat{\theta})$ yields the computable Wald statistic without altering its asymptotic distribution:

$$W = (\hat{\theta}_0 - \theta_0^*)^T \Sigma_{00}^{-1}(\hat{\theta})(\hat{\theta}_0 - \theta_0^*) \xrightarrow{d} \chi_q^2 \tag{6.40}$$

□

Example 6.2. Wald Test for Normal Mean with Unknown Variance
Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. The parameter vector is $\theta = (\mu, \sigma^2)^T$. We wish to test:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0 \tag{6.41}$$

1. **Find the Unrestricted MLEs:** Maximizing the full log-likelihood with respect to both parameters yields the unrestricted estimators:

$$\hat{\mu} = \overline{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{6.42}$$

Our unrestricted parameter estimate is $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)^T$.

2. **Determine the Marginal Covariance at the MLE:** The expected Fisher Information matrix for a Normal distribution is diagonal:

$$\mathcal{J}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \tag{6.43}$$

Because the matrix is diagonal, the asymptotic covariance matrix $\Sigma(\theta) = \mathcal{I}^{-1}(\theta)$ is simply the matrix of reciprocals. The marginal variance for μ (the top-left block Σ_{00}) evaluated at the unrestricted MLE $\hat{\theta}$ is:

$$\Sigma_{00}(\hat{\theta}) = \frac{\hat{\sigma}^2}{n} \tag{6.44}$$

3. **Compute the Wald Statistic W:** We construct the quadratic form using the parameter of interest (μ), its null value (μ_0), and the inverse of its estimated marginal variance (Σ_{00}^{-1}):

$$W = (\hat{\mu} - \mu_0)^T \Sigma_{00}^{-1}(\hat{\theta})(\hat{\mu} - \mu_0) \tag{6.45}$$

Substituting our values:

$$W = (\overline{X} - \mu_0) \left(\frac{n}{\hat{\sigma}^2} \right) (\overline{X} - \mu_0) = \frac{n(\overline{X} - \mu_0)^2}{\hat{\sigma}^2} \tag{6.46}$$

4. **Contrast with the Score Statistic:**

   By Wald's Theorem, this statistic follows a χ_1^2 distribution asymptotically. Note the critical difference from the Score test derived previously. The Wald statistic uses the unrestricted variance estimator $\hat{\sigma}^2$, whereas the Score statistic uses the restricted variance estimator $\hat{\sigma}_0^2 = \frac{1}{n}\sum(X_i - \mu_0)^2$. Because $\hat{\sigma}_0^2 \geq \hat{\sigma}^2$, it guarantees that mathematically $W \geq S$ for this model.

6.4 Rao's Score (Lagrange Multiplier) Theorem for Restricted Models

The asymptotic distribution of the Score Test statistic relies on the behavior of the gradient of the log-likelihood (the Score vector) evaluated at the restricted Maximum Likelihood Estimator. Because it explicitly evaluates the "force" pushing against the null hypothesis constraint, it is equivalently known as the **Lagrange Multiplier (LM) Test**. This theorem provides a powerful method for hypothesis testing that only requires fitting the model under the null hypothesis.

Theorem 6.4 (Asymptotic Distribution of the Score Statistic)**.** *Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be an i.i.d. sample from a regular family of distributions with parameters $\theta \in \Theta \subseteq \mathbb{R}^p$. Partition the parameter vector as $\theta = (\theta_0^T, \theta_1^T)^T$, where θ_0 is the q-dimensional parameter of interest and θ_1 is the $(p-q)$-dimensional nuisance parameter. Consider the test:*

$$H_0 : \theta_0 = \theta_0^* \quad vs \quad H_1 : \theta_0 \neq \theta_0^* \tag{6.47}$$

Let $\hat{\theta}_0 = (\theta_0^{*T}, \hat{\theta}_1^T)^T$ be the restricted MLE under H_0. Let $\mathbf{U}(\theta)$ be the total Score vector partitioned as $\mathbf{U} = [\mathbf{U}_0^T, \mathbf{U}_1^T]^T$, and let $\Sigma = \mathcal{I}^{-1}(\theta)$ be the asymptotic covariance matrix of the unrestricted MLE, with Σ_{00} representing its $q \times q$ top-left marginal block.
The Score statistic is $S = \mathbf{U}_0(\hat{\theta}_0)^T \Sigma_{00}(\hat{\theta}_0) \mathbf{U}_0(\hat{\theta}_0)$. Under H_0, as $n \to \infty$:

$$S \xrightarrow{d} \chi_q^2 \tag{6.48}$$

where q is the dimension of the parameter of interest θ_0.

Proof. **Algebraic Proof via Joint Distribution and Conditioning**
This proof establishes the behavior of the Score vector using constrained optimization, and then derives its sampling distribution by conditioning the joint asymptotic distribution of the true scores.

1. Constrained Optimization (The Lagrange Multipliers)

   To find the restricted MLE $\hat{\theta}_0$, we maximize the log-likelihood subject to the constraint $\theta_0 = \theta_0^*$. We form the Lagrangian with a q-dimensional multiplier vector λ:

   $$\mathcal{L}(\theta, \lambda) = \ell(\theta_0, \theta_1) - \lambda^T(\theta_0 - \theta_0^*) \tag{6.49}$$

   Taking the partial derivatives with respect to the parameters and evaluating them at the restricted optimum $\hat{\theta}_0$ yields the first-order conditions:

   - $\nabla_{\theta_1}\mathcal{L} = \mathbf{U}_1(\hat{\theta}_0) = \mathbf{0}$
   - $\nabla_{\theta_0}\mathcal{L} = \mathbf{U}_0(\hat{\theta}_0) - \hat{\lambda} = \mathbf{0} \implies \mathbf{U}_0(\hat{\theta}_0) = \hat{\lambda}$

   This proves that at the restricted peak, the observed nuisance score $\mathbf{U}_1$ is exactly zero. Furthermore, the evaluated score of interest $\mathbf{U}_0$ is exactly the Lagrange multiplier $\hat{\lambda}$, representing the gradient force pulling toward the unrestricted peak.

2. Joint Asymptotic Distribution of the True Scores

   Let $\mathbf{U}_0$ and $\mathbf{U}_1$ denote the unobservable true score vectors evaluated at the true parameters under the null hypothesis, $\theta^* = ({\theta_0^*}^T, \theta_1^T)^T$. By the Central Limit Theorem, the joint score vector is asymptotically normal, centered at zero, with a covariance matrix given by the expected Fisher Information:

   $$\begin{pmatrix} \mathbf{U}_0 \\ \mathbf{U}_1 \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathcal{I}_{00} & \mathcal{I}_{01} \\ \mathcal{I}_{10} & \mathcal{I}_{11} \end{pmatrix} \right) \tag{6.50}$$

3. The Conditional Distribution

   Using the standard properties of the multivariate normal distribution, the conditional distribution of the true score of interest $\mathbf{U}_0$ given the true nuisance score $\mathbf{U}_1$ is:

   $$\mathbf{U}_0 \mid \mathbf{U}_1 \sim N(\mathcal{I}_{01}\mathcal{I}_{11}^{-1}\mathbf{U}_1, \mathcal{I}_{00} - \mathcal{I}_{01}\mathcal{I}_{11}^{-1}\mathcal{I}_{10}) \tag{6.51}$$

4. Evaluating at the Restricted MLE (Conditioning on Zero)

   As established in Step 1, evaluating the model at the restricted MLE $\hat{\theta}_0$ structurally forces the nuisance score to zero ($\mathbf{U}_1 = \mathbf{0}$). Therefore, the asymptotic distribution of the evaluated score $\mathbf{U}_0(\hat{\theta}_0)$ is equivalent to the conditional distribution of $\mathbf{U}_0$ given $\mathbf{U}_1 = \mathbf{0}$.

   Substituting $\mathbf{U}_1 = \mathbf{0}$ into the conditional distribution, the mean vanishes entirely:

   $$\mathbf{U}_0(\hat{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}_{00} - \mathcal{I}_{01}\mathcal{I}_{11}^{-1}\mathcal{I}_{10}) \tag{6.52}$$

5. Connection to Marginal Covariance and the Final Statistic

The conditional variance matrix $\mathcal{J}_{00} - \mathcal{J}_{01}\mathcal{J}_{11}^{-1}\mathcal{J}_{10}$ is the Schur complement of the Information matrix. By the formula for the inverse of a partitioned matrix, this is exactly the inverse of the top-left block of the asymptotic covariance matrix $\Sigma = \mathcal{J}^{-1}$:

$$\text{Var}(\mathbf{U}_0(\hat{\theta}_0)) = \Sigma_{00}^{-1} \tag{6.53}$$

Because the evaluated score $\mathbf{U}_0(\hat{\theta}_0)$ is asymptotically normal with mean zero and variance Σ_{00}^{-1}, standardizing it by its inverse-variance creates a quadratic form. Since the inverse of the precision is the covariance itself ($(\Sigma_{00}^{-1})^{-1} = \Sigma_{00}$), we obtain the test statistic:

$$S = \mathbf{U}_0(\hat{\theta}_0)^T \Sigma_{00} \mathbf{U}_0(\hat{\theta}_0) \xrightarrow{d} \chi_q^2 \tag{6.54}$$

□

Example 6.3. Score Test for Normal Mean with Unknown Variance
Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. The parameter vector is $\theta = (\mu, \sigma^2)^T$. We wish to test:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0 \tag{6.55}$$

1. **Find the Restricted MLEs:** Under H_0, we fix $\mu = \mu_0$. Maximizing the restricted log-likelihood with respect to the nuisance parameter σ^2 yields:

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2 \tag{6.56}$$

The restricted parameter vector is $\hat{\theta}_0 = (\mu_0, \hat{\sigma}_0^2)^T$.

2. **Evaluate the Full Score Vector:** The full log-likelihood function for the normal sample is:

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{6.57}$$

The total Score vector $\mathbf{U}(\theta)$ is the gradient with respect to both parameters:

$$\mathbf{U}(\mu, \sigma^2) = \begin{pmatrix} U_\mu(\mu, \sigma^2) \\ U_{\sigma^2}(\mu, \sigma^2) \end{pmatrix} = \begin{pmatrix} \frac{n(\overline{X} - \mu)}{\sigma^2} \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2 \end{pmatrix} \tag{6.58}$$

Now we evaluate this full vector at the restricted MLE $\hat{\theta}_0 = (\mu_0, \hat{\sigma}_0^2)^T$:

$$\mathbf{U}(\hat{\theta}_0) = \begin{pmatrix} \frac{n(\overline{X} - \mu_0)}{\hat{\sigma}_0^2} \\ -\frac{n}{2\hat{\sigma}_0^2} + \frac{1}{2\hat{\sigma}_0^4}\sum_{i=1}^{n}(X_i - \mu_0)^2 \end{pmatrix} \tag{6.59}$$

Because $\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2$, the second component perfectly cancels out. By definition, the score for the nuisance parameter evaluated at the restricted MLE is zero:

$$\mathbf{U}(\hat{\theta}_0) = \begin{pmatrix} \frac{n(\overline{X} - \mu_0)}{\hat{\sigma}_0^2} \\ 0 \end{pmatrix} \tag{6.60}$$

3. **Determine the Marginal Covariance:** The expected Fisher Information matrix for a Normal distribution is diagonal because the mean and variance are orthogonal parameters:

$$\mathcal{J}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \tag{6.61}$$

Because the matrix is diagonal, the asymptotic covariance matrix $\Sigma(\theta) = \mathcal{J}^{-1}(\theta)$ is simply the matrix of reciprocals. The marginal variance for μ evaluated at the restricted MLE $\hat{\theta}_0$ is:

$$\Sigma_{00}(\hat{\theta}_0) = \frac{\hat{\sigma}_0^2}{n} \tag{6.62}$$

4. **Compute the Score Statistic S:** Constructing the quadratic form for the parameter of interest using its marginal variance Σ_{00}:

$$S = U_\mu(\hat{\theta}_0)^T \Sigma_{00}(\hat{\theta}_0) U_\mu(\hat{\theta}_0) \tag{6.63}$$

$$S = \left(\frac{n(\overline{X} - \mu_0)}{\hat{\sigma}_0^2}\right)\left(\frac{\hat{\sigma}_0^2}{n}\right)\left(\frac{n(\overline{X} - \mu_0)}{\hat{\sigma}_0^2}\right) = \frac{n(\overline{X} - \mu_0)^2}{\hat{\sigma}_0^2} \tag{6.64}$$

5. **Asymptotic Distribution:** By the Score Theorem, since we are testing $q = 1$ parameter restriction, the statistic follows a χ_1^2 distribution:

$$S = \frac{n(\overline{X} - \mu_0)^2}{\hat{\sigma}_0^2} \xrightarrow{d} \chi_1^2 \tag{6.65}$$

This is highly analogous to the squared t-statistic, but specifically uses the variance estimated under the null hypothesis constraint.

6.5 A Comparison of Finite-Sample Distributions of the Three Asymptotic Tests in OLS

6.5.0.1 Exact Finite-Sample Distributions via the F-Statistic

When testing linear restrictions in Ordinary Least Squares (OLS) under the assumption of normal errors, we do not need to rely on asymptotic approximations. The true, exact finite-sample test statistic follows an F-distribution.

Because the Wald (W), Likelihood Ratio (LR), and Score (S) statistics are strictly monotonic algebraic transformations of the exact finite-sample F-statistic, we can use the transformation of variables to derive their exact finite-sample Cumulative Distribution Functions (CDFs).

Let us consider testing a linear hypothesis involving q restrictions in an OLS model with k total parameters and n observations. Let F denote the exact finite-sample F-statistic, which follows an $F_{q,n-k}$ distribution. Let $G_{q,n-k}(x)$ denote the CDF of this F-distribution.

By isolating F in the algebraic definitions of W, LR, and S, we derive their exact finite-sample CDFs:

Expressing the Test Statistics via Residual Sums of Squares

Let RSS_1 denote the residual sum of squares from the **full** (unrestricted) model (estimating all k parameters under H_1), and let RSS_0 denote the residual sum of squares from the **restricted** model (imposing the q constraints of H_0).

Because imposing constraints can only degrade the fit of the model to the training data, we know that $RSS_0 \geq RSS_1$. The difference $(RSS_0 - RSS_1)$ isolates the exact loss of fit caused by enforcing the null hypothesis.

The exact finite-sample F-statistic is the ratio of this loss of fit to the unrestricted noise, scaled by their respective degrees of freedom:

$$F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-k)} \tag{6.66}$$

By substituting this definition of F into the algebraic transformations derived previously, the "Holy Trinity" of asymptotic tests can be expressed beautifully as functions of the residual sums of squares:

1. Wald Statistic (W)

The Wald test standardizes the loss of fit using the **unrestricted** variance estimate ($\hat{\sigma}^2 = RSS_1/n$). Substituting the F formula into $W = \frac{nq}{n-k}F$:

$$W = n\left(\frac{RSS_0 - RSS_1}{RSS_1}\right) \tag{6.67}$$

2. Likelihood Ratio Statistic (LR)

The Likelihood Ratio test evaluates the natural logarithm of the ratio of the two likelihoods, which simplifies directly to the log-ratio of the residual sums of squares. Substituting F into $LR = n\ln\left(1 + \frac{q}{n-k}F\right)$:

$$LR = n\ln\left(\frac{RSS_0}{RSS_1}\right) \tag{6.68}$$

3. Score Statistic (S)

The Score test standardizes the loss of fit using the **restricted** variance estimate ($\hat{\sigma}_0^2 = RSS_0/n$). Substituting F into $S = \frac{nqF}{n-k+qF}$:

$$S = n\left(\frac{RSS_0 - RSS_1}{RSS_0}\right) \tag{6.69}$$

6.5.0.2 Asymptotic Equivalence in OLS

Having established the exact algebraic relationship between the three test statistics and the finite-sample F-statistic, we can use standard OLS large-sample theory to rigorously prove that all three converge to the exact same χ_q^2 asymptotic distribution.

Proof.

1. The Asymptotic Limit of the Scaled F-Statistic

Under the null hypothesis H_0 (and assuming standard regularity conditions for OLS), the numerator of the F-statistic represents the loss of fit, which asymptotically follows a Chi-square distribution scaled by the true error variance σ^2:

$$(RSS_0 - RSS_1) \xrightarrow{d} \sigma^2 \chi_q^2 \tag{6.70}$$

The denominator of the F-statistic is the unbiased estimator of the error variance, which converges in probability to the true variance by the Law of Large Numbers:

$$\frac{RSS_1}{n-k} \xrightarrow{p} \sigma^2 \tag{6.71}$$

By Slutsky's Theorem, dividing the convergent numerator by the convergent denominator cancels out the unknown σ^2, yielding the foundational limit:

$$qF = \frac{RSS_0 - RSS_1}{RSS_1/(n-k)} \xrightarrow{d} \chi_q^2 \tag{6.72}$$

2. Convergence of the Wald Statistic (W)

We previously defined the Wald statistic algebraically as:

$$W = \frac{n}{n-k}(qF) \tag{6.73}$$

As $n \to \infty$, the ratio of the sample size to the degrees of freedom $(\frac{n}{n-k})$ converges to 1. Applying Slutsky's Theorem again, the Wald statistic converges directly to the scaled F-statistic:

$$W \xrightarrow{d} 1 \cdot \chi_q^2 = \chi_q^2 \tag{6.74}$$

3. Convergence of the Likelihood Ratio Statistic (LR)

The LR statistic is defined via the natural logarithm:

$$LR = n \ln\left(1 + \frac{qF}{n-k}\right) \tag{6.75}$$

Because qF converges to a χ_q^2 distribution, it is bounded in probability, meaning $\frac{qF}{n-k} = O_p(n^{-1})$. As $n \to \infty$, this term approaches zero. We can apply the first-order Taylor expansion $\ln(1+x) \approx x$ for small x:

$$LR \approx n\left(\frac{qF}{n-k}\right) = \frac{n}{n-k}(qF) \tag{6.76}$$

This first-order approximation is exactly the Wald statistic. Therefore, the difference between LR and W vanishes asymptotically, giving:

$$LR \xrightarrow{d} \chi_q^2 \tag{6.77}$$

4. Convergence of the Score Statistic (S)

The Score statistic is algebraically defined as:

$$S = \frac{nqF}{n - k + qF} \tag{6.78}$$

Dividing the numerator and denominator by $n - k$ yields:

$$S = \frac{\frac{n}{n-k}(qF)}{1 + \frac{qF}{n-k}} = \frac{W}{1 + \frac{qF}{n-k}} \tag{6.79}$$

As established, $\frac{qF}{n-k} \xrightarrow{p} 0$. The denominator converges in probability to 1, meaning the Score statistic perfectly tracks the Wald statistic in the limit:

$$S \xrightarrow{d} \frac{W}{1} \xrightarrow{d} \chi^2_q \tag{6.80}$$

Conclusion: While the algebraic inequality $W \geq LR \geq S$ holds strictly in any finite sample, the differences between the statistics are of order $O_p(n^{-1})$. As $n \to \infty$, the "gaps" between them shrink to zero, and all three provide asymptotically equivalent tests against the χ^2_q distribution. □

6.5.0.3 Visualizing Finite-Sample Calibration (Size Distortion)

To visualize how badly the asymptotic χ^2_q approximation distorts the Type I error rates in finite samples, we can plot the true Cumulative Distribution Function (CDF) of the asymptotic p-values.

If a test is perfectly calibrated, the probability that its p-value is less than a nominal threshold α should be exactly α, forming a perfect $y = x$ diagonal line (the Uniform(0,1) CDF).

The event that the asymptotic p-value is less than α is mathematically equivalent to the test statistic being greater than the χ^2_q critical value (c_α). Therefore, the true rejection probability $P(p < \alpha)$ is exactly $1 - F(c_\alpha)$, where F is the exact finite-sample CDF derived above.

6.5.0.4 Exact Finite-Sample Distribution Calibration Plot

The following code allows you to adjust the sample size (n) and the number of parameters (k) to observe how the finite-sample rejection rates of the three asymptotic tests (Wald, Likelihood Ratio, and Score) converge toward the ideal Uniform(0,1) calibration.

Summary

In OLS with finite samples, the χ^2 distribution is an optimistic approximation that leads to **over-rejection** (liberal bias) across all three tests. However:

- The **Wald test** is the most liberal because it does nothing to counteract the too-low χ^2 threshold.
- The **Score test** is the most conservative of the three because its mathematical structure provides a natural downward "correction" that happens to land it closest to the ideal uniform distribution at common significance levels like $\alpha = 0.05$.

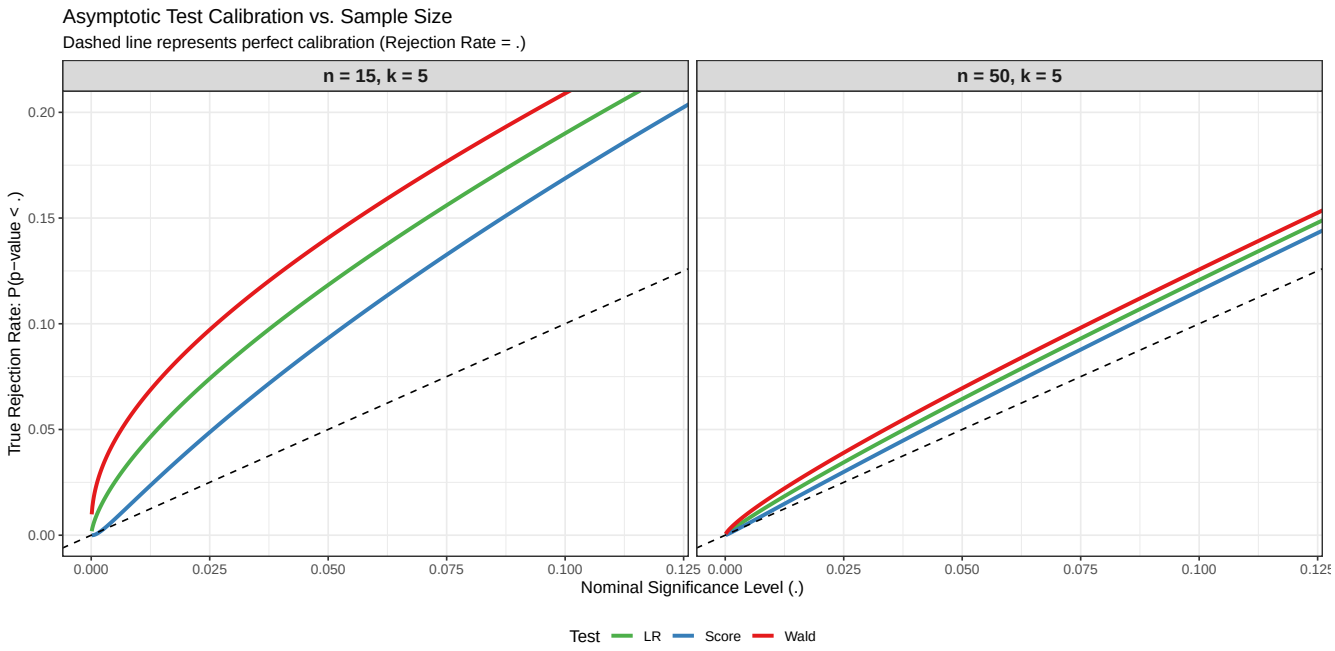


Figure 6.2: Exact Analytical CDFs of Asymptotic P-values

7 Minimum Variance Estimators

7.1 Completeness

7.1.1 Complete Statistic

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

$$P(g(T) = 0|\theta) = 1 \quad \text{for all } \theta \tag{7.2}$$

Corollary 7.1 (Uniqueness of Unbiased Estimator)**.** *Let T be a complete statistic for a family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$. If there exists an unbiased estimator for θ (or a parametric function $\tau(\theta)$) that is strictly a function of T, then this estimator is unique almost everywhere.*

Proof. Suppose there are two functions of T, say $g_1(T)$ and $g_2(T)$, that are both unbiased estimators for θ. By the definition of unbiasedness, we have:

$$E_\theta[g_1(T)] = \theta \tag{7.3}$$

and

$$E_\theta[g_2(T)] = \theta \tag{7.4}$$

for all $\theta \in \Omega$.
Subtracting the second equation from the first yields:

$$E_\theta[g_1(T) - g_2(T)] = 0 \tag{7.5}$$

for all $\theta \in \Omega$.
Let $w(T) = g_1(T) - g_2(T)$. The equation above establishes that:

$$E_\theta[w(T)] = 0 \tag{7.6}$$

for all θ.
Because T is a complete statistic, the definition of completeness dictates that the only measurable function of T with an expected value of zero for all θ is the function that is zero almost everywhere. Therefore, we conclude:

$$P_\theta(w(T) = 0) = 1 \tag{7.7}$$

which directly implies:

$$P_\theta(g_1(T) = g_2(T)) = 1 \tag{7.8}$$

for all $\theta \in \Omega$. Thus, $g_1(T)$ and $g_2(T)$ are identical with probability 1, proving that the unbiased estimator which is a function of T is unique. □

Example 7.1 (Uniform Distribution (Not Complete))**.** Let $X_1, \ldots, X_n \sim \text{Unif}(\theta - 1, \theta + 1)$. The statistic $T(X) = (X_{(1)}, X_{(n)})$ is a Minimal Sufficient Statistic. However, it is **not complete**.

Properties of Order Statistics via Transformation

Instead of deriving the joint distribution of the order statistics using calculus, we can simplify the problem by standardizing the variables. Let $U_i = X_i - \theta$. The transformed variables $U_1, \ldots, U_n$ are independent and identically distributed as $\text{Unif}(-1, 1)$.

The order statistics preserve this additive transformation:

$$X_{(k)} = U_{(k)} + \theta \tag{7.9}$$

For a uniform distribution on $(-1, 1)$, the expected values of the minimum and maximum order statistics are standard results:

$$E[U_{(1)}] = -\frac{n-1}{n+1} \tag{7.10}$$

and

$$E[U_{(n)}] = \frac{n-1}{n+1} \tag{7.11}$$

Consequently, the expected values for the extreme order statistics of our original sample X are simply shifted by θ:

$$E[X_{(1)}] = \theta - \frac{n-1}{n+1} \tag{7.12}$$

and

$$E[X_{(n)}] = \theta + \frac{n-1}{n+1} \tag{7.13}$$

The range R of the sample is invariant to the location shift θ:

$$R = X_{(n)} - X_{(1)} = U_{(n)} - U_{(1)} \tag{7.14}$$

Because the distribution of the range depends only on the U_i variables, it has no dependence on θ, making R an ancillary statistic. Its expected value is directly obtained from the expectations above without needing to integrate the joint density:

$$E[R] = E[X_{(n)}] - E[X_{(1)}] = \frac{n-1}{n+1} - \left(-\frac{n-1}{n+1}\right) = \frac{2(n-1)}{n+1} \tag{7.15}$$

Two Distinct Unbiased Estimators
Because the minimal sufficient statistic $T(X) = (X_{(1)}, X_{(n)})$ is not complete, we can construct multiple unbiased estimators for θ that are purely functions of T:

1. **The Midrange Estimator (W_1)** Consider the midrange, $W_1 = \frac{X_{(1)}+X_{(n)}}{2}$. Using the expectations derived above, we take the expected value:

$$E[W_1] = \frac{E[X_{(1)}] + E[X_{(n)}]}{2} = \frac{\left(\theta - \frac{n-1}{n+1}\right) + \left(\theta + \frac{n-1}{n+1}\right)}{2} = \theta \tag{7.16}$$

   Thus, W_1 is an unbiased estimator of θ and is strictly a function of T.

2. **The Adjusted Estimator (W_2)** To elegantly illustrate the consequence of T being incomplete, we can identify a non-trivial function of the statistic that has an expected value of zero. Let $V(T)$ be a function based on the range R. Since we know $E[R] = \frac{2(n-1)}{n+1}$, the following statistic has a mean of zero for all θ:

$$V(T) = R - \frac{2(n-1)}{n+1} = X_{(n)} - X_{(1)} - \frac{2(n-1)}{n+1} \tag{7.17}$$

   We can now form a new unbiased estimator W_2 by adding this zero-mean statistic $V(T)$ to our first estimator W_1:

$$W_2 = W_1 + V(T) = \frac{X_{(1)} + X_{(n)}}{2} + X_{(n)} - X_{(1)} - \frac{2(n-1)}{n+1} \tag{7.18}$$

   Taking the expectation yields $E[W_2] = E[W_1] + E[V(T)] = \theta + 0 = \theta$.

Both W_1 and W_2 are functions of the minimal sufficient statistic T, and both are unbiased for θ. However, since $V(T)$ is not identically zero, $W_1 \neq W_2$ almost everywhere. According to the Lehmann-Scheffé theorem, an unbiased estimator based on a complete sufficient statistic must be unique. By demonstrating that we can construct two distinct unbiased estimators based on T, we have proven that T is not complete.

7.1.2 Exponential Family Completeness

Lemma 7.1 (Exponential Family Completeness)**.** *If $T = (T_1, \ldots, T_k)$ is the natural statistic of an exponential family that contains an open rectangle in the parameter space, then T is complete.*

Proof. **Proof of Lemma Lemma 7.1:**
Consider a k-parameter exponential family in canonical form with density:

$$f(y|\eta) = h(y) \exp\left(\sum_{j=1}^{k} \eta_j T_j(y) - A(\eta)\right) \tag{7.19}$$

where η belongs to a natural parameter space Ξ that contains a k-dimensional open rectangle.

Let $g(T)$ be a function such that $E_\eta[g(T)] = 0$ for all $\eta \in \Xi$. The expectation is defined as:

$$\int g(t)h(t) \exp\left(\sum_{j=1}^{k} \eta_j t_j - A(\eta)\right) dt = 0 \tag{7.20}$$

Since $\exp(-A(\eta))$ is a strictly positive constant with respect to t, we can divide it out:

$$\int g(t)h(t) \exp\left(\sum_{j=1}^{k} \eta_j t_j\right) dt = 0 \tag{7.21}$$

We can decompose $g(t)$ into its positive and negative parts, $g(t) = g^+(t) - g^-(t)$. Substituting this in gives:

$$\int g^+(t)h(t)e^{\sum \eta_j t_j} dt = \int g^-(t)h(t)e^{\sum \eta_j t_j} dt \tag{7.22}$$

These integrals represent the **multivariate Laplace transforms** of the measures $m^+(t) = g^+(t)h(t)$ and $m^-(t) = g^-(t)h(t)$. A fundamental property of Laplace transforms is that if two transforms are equal over an open set (in this case, the open rectangle in Ξ), then the underlying measures must be identical almost everywhere. Therefore, $g^+(t)h(t) = g^-(t)h(t)$, which implies $g(t) = 0$ almost everywhere with respect to the distribution of T. Thus, T is complete. □

7.1.3 Relationship Between Completeness and Minimality

While sufficiency ensures that a statistic captures all information about θ, and minimality ensures it does so without redundancy, completeness is a stronger property that guarantees the statistic is "uniquely informative" for unbiased estimation.

Theorem 7.1 (Completeness to Minimal Sufficiency)**.** *If T is a **complete sufficient statistic** for a family of distributions $\{f(x|\theta) : \theta \in \Theta\}$, then T is also a **minimal sufficient statistic**.*

Proof. To prove T is minimal sufficient, we must show that for any other sufficient statistic S, T is a function of S.

1. **Define a Conditional Expectation:**

   Let S be any sufficient statistic. Define the statistic $\phi(S) = E[T|S]$. Because S is sufficient, $\phi(S)$ is a valid statistic (it does not depend on θ).

2. **Compare Expectations:**

   By the Law of Iterated Expectations:

   $$E[\phi(S)] = E[E[T|S]] = E[T] \tag{7.23}$$

   This implies that $E[\phi(S) - T] = 0$ for all $\theta \in \Theta$.

3. **Construct a Function of T:**

Define $h(T) = E[\phi(S)|T]$. Since T is sufficient, $h(T)$ is a statistic. By the properties of expectation:

$$E[h(T)] = E[E[\phi(S)|T]] = E[\phi(S)] = E[T] \tag{7.24}$$

Therefore, $E[h(T) - T] = 0$ for all $\theta \in \Theta$.

4. **Apply Completeness:**

   Since $g(T) = h(T) - T$ is a function of T whose expectation is zero for all θ, and T is **complete**, it follows that:

$$g(T) = 0 \implies h(T) = T \quad \text{with probability 1.} \tag{7.25}$$

5. **Conclusion:**

   Since T is a function of $\phi(S)$ (which is a function of S), T must be a function of S. Since this holds for any sufficient statistic S, T is minimal sufficient.

□

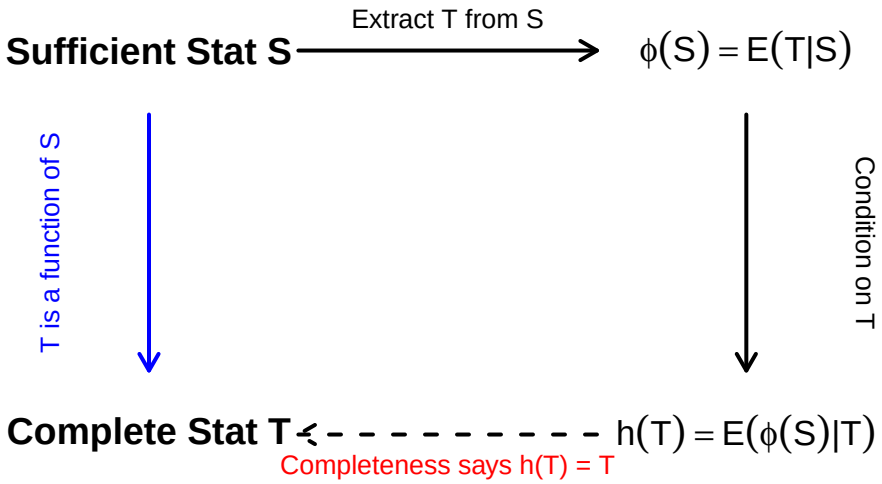


Figure 7.1: Logical flow of the proof: Showing T is a function of any sufficient statistic S.

The relationship between statistics types is visualized by Figure 7.2:

Example 7.2 (Minimal Sufficient Statistic is Not Necessarily Complete)**.** It is important to note that the converse is not generally true: a **minimal sufficient statistic is not necessarily complete**.
A standard counterexample is a random sample $X_1, \ldots, X_n \sim \text{Unif}(\theta - 1, \theta + 1)$. For this distribution, the statistic $T(X) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic for θ. However, it is not complete because the range $R = X_{(n)} - X_{(1)}$ is an ancillary statistic. Its distribution does not depend on θ, meaning we can construct a non-zero function of T that has an expected value of zero for all θ.

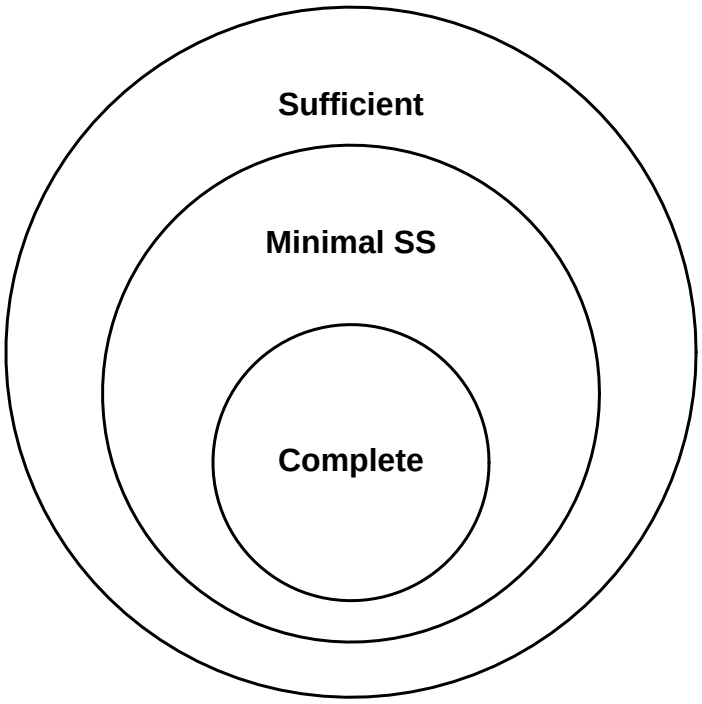


Figure 7.2: Hierarchy of Statistics: Completeness implies Minimal Sufficiency under regularity.

7.2 UMVUE

7.2.1 Definition

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. $\text{Var}(T(x)|\theta) \leq \text{Var}(d(x)|\theta)$ for all θ and for all other unbiased estimators $d(x)$.

7.2.2 Rao-Blackwell Theorem

The Rao-Blackwell theorem provides a mechanism for improving an estimator by utilizing a sufficient statistic.

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 (independent of θ).*
2. *$E[g(T)] = \theta$ (Unbiased).*
3. *Var$(g(T)) \leq$ Var$(d_1(x))$.*

Proof.

1. **Statistic Property:** By the definition of sufficiency, the conditional distribution of X given T does not depend on θ. Thus, $g(T) = E[d_1(x)|T]$ does not involve θ and is a valid statistic.
2. **Unbiasedness:** By the Law of Iterated Expectations:

$$E[g(T)] = E_T[E_X(d_1(X)|T)] = E_X[d_1(X)] = \theta \tag{7.26}$$

3. **Variance Reduction:** Using the Law of Total Variance:

$$\text{Var}(d_1(X)) = \text{Var}(E[d_1(X)|T]) + E[\text{Var}(d_1(X)|T)] = \text{Var}(g(T)) + E[\text{Var}(d_1(X)|T)] \tag{7.27}$$

Since $\text{Var}(d_1(X)|T) \geq 0$, it follows that $E[\text{Var}(d_1(X)|T)] \geq 0$. Therefore, $\text{Var}(g(T)) \leq \text{Var}(d_1(X))$. Equality holds only if $d_1(X) = g(T)$ almost surely. □

7.2.3 Lehmann-Scheffé Theorem

While Rao-Blackwell improves an estimator, Lehmann-Scheffé ensures that we have found the *best* one.

Theorem 7.3 (Lehmann-Scheffé 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.*

Proof.

1. **Existence:** From Rao-Blackwell, $\phi(T)$ is an unbiased estimator with variance at most that of any $d(X)$.
2. **Uniqueness:** Suppose there exists another unbiased estimator $\psi(T)$ that is also a function of the same complete sufficient statistic T. Then:

$$E[\phi(T) - \psi(T)|\theta] = E[\phi(T)|\theta] - E[\psi(T)|\theta] = \theta - \theta = 0 \tag{7.28}$$

By the property of **completeness** for T, $E[g(T)] = 0$ implies $P(g(T) = 0) = 1$. Thus, $\phi(T) = \psi(T)$ almost surely. This proves that $\phi(T)$ is the unique unbiased estimator based on T, and therefore the unique UMVUE. □

7.3 A Procedure to Find UMVUE

Example 7.3 (Poisson UMVUE for λ^2)**.** Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. $T = \sum X_i$ is complete sufficient. $E[X_1^2 - X_1] = E[X_1^2] - E[X_1] = (\lambda^2 + \lambda) - \lambda = \lambda^2$. So $d(X) = X_1^2 - X_1$ is unbiased. The UMVUE is $g(T) = E[X_1(X_1 - 1)|T]$. Using the moments of $X_1|T \sim \text{Bin}(T, 1/n)$:

$$g(T) = \frac{T(T-1)}{n^2} \tag{7.29}$$

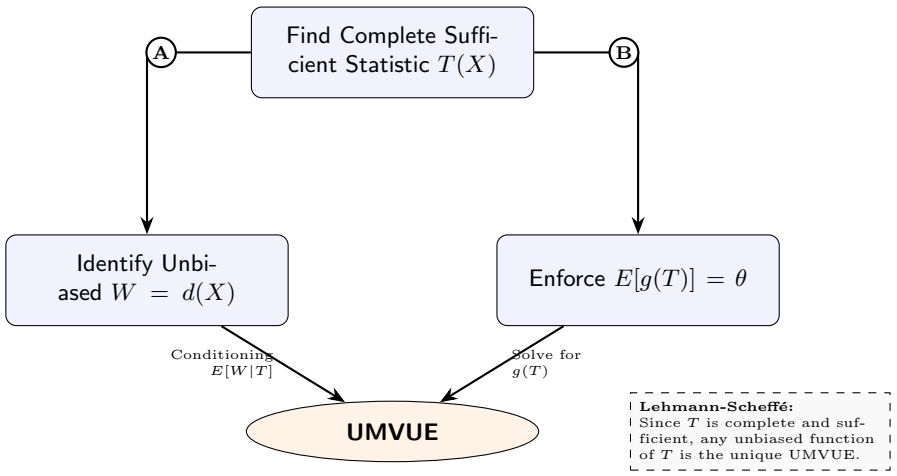


Figure 7.3: The two paths to finding the UMVUE: (A) Rao-Blackwellization via conditioning an initial unbiased estimator W on T; (B) The Direct Method by solving for an unbiased function $g(T)$ directly.

Sol A: Rao-Blackwellization (Conditioning)

1. **Find a simple unbiased estimator**: Consider $W = X_1(X_1 - 1)$.

$$E[W] = E[X_1^2] - E[X_1] = (\lambda^2 + \lambda) - \lambda = \lambda^2 \tag{7.30}$$

Thus, W is unbiased for λ^2.

2. **Condition on the complete sufficient statistic**:

   The UMVUE is $\phi(T) = E[X_1(X_1 - 1)|T]$. Recall that the conditional distribution of X_1 given $T = t$ is Binomial$(t, 1/n)$.

3. **Compute the conditional expectation**:

   Using the second factorial moment of a Binomial(t, p) distribution, $E[Y(Y-1)] = t(t-1)p^2$:

$$\phi(T) = E[X_1(X_1 - 1)|T] = T(T-1)\left(\frac{1}{n}\right)^2 = \frac{T(T-1)}{n^2} \tag{7.31}$$

Sol B: The Direct Method (Enforcing Unbiasedness)

1. **Identify the distribution of T**:

   Since $X_i \sim$ Poisson(λ), the sum $T = \sum X_i$ follows:

$$T \sim \text{Poisson}(n\lambda) \tag{7.32}$$

2. **Propose a functional form for $g(T)$**:

   For any $Y \sim$ Poisson(μ), $E[Y(Y-1)] = \mu^2$. Here, $\mu = n\lambda$, so:

$$E[T(T-1)] = (n\lambda)^2 = n^2\lambda^2 \tag{7.33}$$

3. **Adjust for unbiasedness**:

   To get an expected value of exactly λ^2, we scale by $1/n^2$:

$$E\left[\frac{T(T-1)}{n^2}\right] = \frac{1}{n^2}(n^2\lambda^2) = \lambda^2 \tag{7.34}$$

4. **Conclusion**:

   By the Lehmann-Scheffé Theorem, $g(T) = \frac{T(T-1)}{n^2}$ is the unique UMVUE.

7.3.1 Example: Joint UMVUE in the Normal Family

The power of the Exponential Family lemma is most evident when dealing with distributions involving multiple parameters, such as the Normal distribution.

Example 7.4 (Joint UMVUE for μ and σ^2)**.** Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ where both parameters are unknown. The joint density can be written in the canonical exponential form:

$$f(\mathbf{x}|\mu,\sigma^2) = \underbrace{(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{n\mu^2}{2\sigma^2}\right)}_{C(\eta)}\exp\left(\frac{\mu}{\sigma^2}\sum x_i - \frac{1}{2\sigma^2}\sum x_i^2\right) \tag{7.35}$$

1. **Identify Natural Statistics:**

   The natural sufficient statistics are $T_1 = \sum X_i$ and $T_2 = \sum X_i^2$. Because the natural parameter space contains an open rectangle in $\mathbb{R} \times \mathbb{R}^-$, the vector (T_1, T_2) is **complete and sufficient**.

2. **Transformation to Common Estimators:**

   Any one-to-one mapping of a complete sufficient statistic is also complete and sufficient. We can define:

$$\overline{X} = \frac{T_1}{n} \quad \text{and} \quad S^2 = \frac{1}{n-1}\left(T_2 - \frac{T_1^2}{n}\right) \tag{7.36}$$

   Since $(\overline{X}, S^2)$ is a one-to-one function of (T_1, T_2), it is also a **complete sufficient statistic**.

3. **Applying Lehmann-Scheffé:**

   - Since $E[\overline{X}] = \mu$ and $\overline{X}$ is a function of the complete sufficient statistic, $\overline{X}$ is the **UMVUE for** μ.
   - Since $E[S^2] = \sigma^2$ and S^2 is a function of the complete sufficient statistic, S^2 is the **UMVUE for** σ^2.

7.3.2 Example: UMVUE for $\log(\sigma^2)$ in the Normal Family

Example 7.5 (UMVUE for $\log(\sigma^2)$ in the Normal Family)**.** While we saw that no UMVUE exists for $\log(\lambda)$ in the Poisson case, an unbiased estimator for the log-variance *does* exist in the Normal distribution $N(\mu, \sigma^2)$.

1. **The Target:** We want to estimate $g(\sigma^2) = \log(\sigma^2)$.
2. **Distributional Fact:** Recall that for a sample of size n, the statistic $Y = \frac{(n-1)S^2}{\sigma^2}$ follows a χ^2_{n-1} distribution (which is a Gamma$(\alpha = \frac{n-1}{2}, \beta = 2)$ distribution).
3. **Log-Moments of Gamma:** For a variable $W \sim \text{Gamma}(\alpha, \beta)$, the expected value of its logarithm is:

$$E[\log(W)] = \psi(\alpha) + \log(\beta) \tag{7.37}$$

   where $\psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the **digamma function**.
4. **Deriving the Estimator:**

   Expanding the expectation of Y and rearranging to isolate $\log(\sigma^2)$:

$$E\left[\log(S^2) + \left(\log\left(\frac{n-1}{2}\right) - \psi\left(\frac{n-1}{2}\right)\right)\right] = \log(\sigma^2) \tag{7.38}$$

5. **Conclusion:**

   Since S^2 is a complete sufficient statistic, the **Lehmann-Scheffé Theorem** implies:

$$\widehat{\log(\sigma^2)}_{\text{UMVUE}} = \log(S^2) + C(n) \tag{7.39}$$

   where $C(n) = \log\left(\frac{n-1}{2}\right) - \psi\left(\frac{n-1}{2}\right)$ is the **positive** correction term.

Because the logarithm is a concave function, Jensen's Inequality tells us that $E[\log(S^2)] < \log(E[S^2]) = \log(\sigma^2)$. Therefore, $\log(S^2)$ is "too small," and we must add a positive value $C(n)$ to reach unbiasedness.

Numerical Values of the Correction

As seen in Table 7.1, the bias is quite severe for small samples but vanishes as $n \to \infty$.

Table 7.1: Positive correction factors $C(n)$ to be added to $\log(S^2)$

n	$\alpha = \frac{n-1}{2}$	Positive Correction $C(n)$
2	0.5	1.2704
5	2.0	0.2704
10	4.5	0.1152
30	14.5	0.0349
100	49.5	0.0101

7.4 Asymptotic Optimality: UMVUE, CRLB, and the MLE

7.4.1 The Cramér-Rao Lower Bound as an Optimality Check

While the Lehmann-Scheffé theorem provides a constructive path to finding the UMVUE via conditioning on a complete sufficient statistic, the **Cramér-Rao Lower Bound (CRLB)** provides a theoretical variance floor to verify optimality directly.

Recall that for any unbiased estimator $T(X)$ of θ, its variance is bounded below by the inverse of the Fisher Information:

$$\text{Var}(T(X)) \geq \mathcal{I}_n(\theta)^{-1} \tag{7.40}$$

If you identify an unbiased estimator whose variance exactly equals $\mathcal{I}_n(\theta)^{-1}$, you have mathematically proven it is the UMVUE. It has hit the absolute floor of information extraction; no other unbiased estimator can possibly possess a smaller variance.

7.4.2 The Asymptotic Triumph of the MLE

In finite samples, finding an estimator that achieves the exact Cramér-Rao Lower Bound is rare (typically only occurring for specific parameterizations within the Exponential Family). Furthermore, the Maximum Likelihood Estimator (MLE) is often biased in finite samples (for example, the MLE for normal variance divides by n instead of $n-1$), meaning it is frequently *not* the UMVUE for a fixed sample size n.

However, as $n \to \infty$, the MLE becomes the undisputed champion of estimation. Recall the established asymptotic sampling distribution of the MLE:

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta^*) \xrightarrow{d} \mathcal{N}\Big(0, \mathcal{I}_1(\theta^*)^{-1}\Big) \tag{7.41}$$

This result reveals three profound asymptotic properties of the MLE:

1. **Consistency:** The bias strictly vanishes ($\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta^*$).
2. **Asymptotic Normality:** The sampling distribution converges to a Gaussian centered exactly on the truth.
3. **Asymptotic Efficiency:** The variance of this limiting distribution shrinks to exactly match the Cramér-Rao Lower Bound.

Because its asymptotic covariance perfectly matches the inverse Fisher Information matrix, **the MLE is asymptotically UMVUE**.

To understand this intuitively, we can think of the Fisher Information, $\mathcal{I}(\theta^*)$, as a measure of how sharply the log-likelihood function peaks around the true parameter. A sharper peak means the data provides clearer, more precise information. The Cramér-Rao Lower Bound (CRLB) translates this "sharpness" into a strict mathematical floor: it represents the absolute minimum variance any unbiased estimator can possibly achieve.The fundamental takeaway of Maximum Likelihood estimation is this: while the MLE might not be the perfect UMVUE for a small sample, as the sample size grows ($n \to \infty$), its variance shrinks until it exactly hits the CRLB floor. Because it achieves this theoretical limit of precision, the MLE is called asymptotically efficient. For large datasets, no other unbiased estimator can extract more information or provide a tighter variance than the Maximum Likelihood Estimator.

8 Decision Theory

8.1 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, $L(\theta, a) \geq 0$.

8.2 Decision Rules and Risk Functions

8.2.1 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.

8.2.2 Risk Function

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

$$R(\theta, d) = E_\theta[L(\theta, d(X))] \tag{8.1}$$

8.3 Examples of Decision Problems

8.3.1 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 $\theta \in H_1$: $R(\theta, d) = P(\text{Type II Error})$.

8.3.2 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):** $R(\theta, d) = \text{Var}(\bar{x}) + \text{Bias}^2$.

8.3.3 Example 3: Interval Estimation

We want to estimate a range for the parameter.

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

8.4 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 ($50/50$).

8.4.1 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.

8.4.2 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	$R_2 = 0.5$

8.5 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.

8.5.1 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).

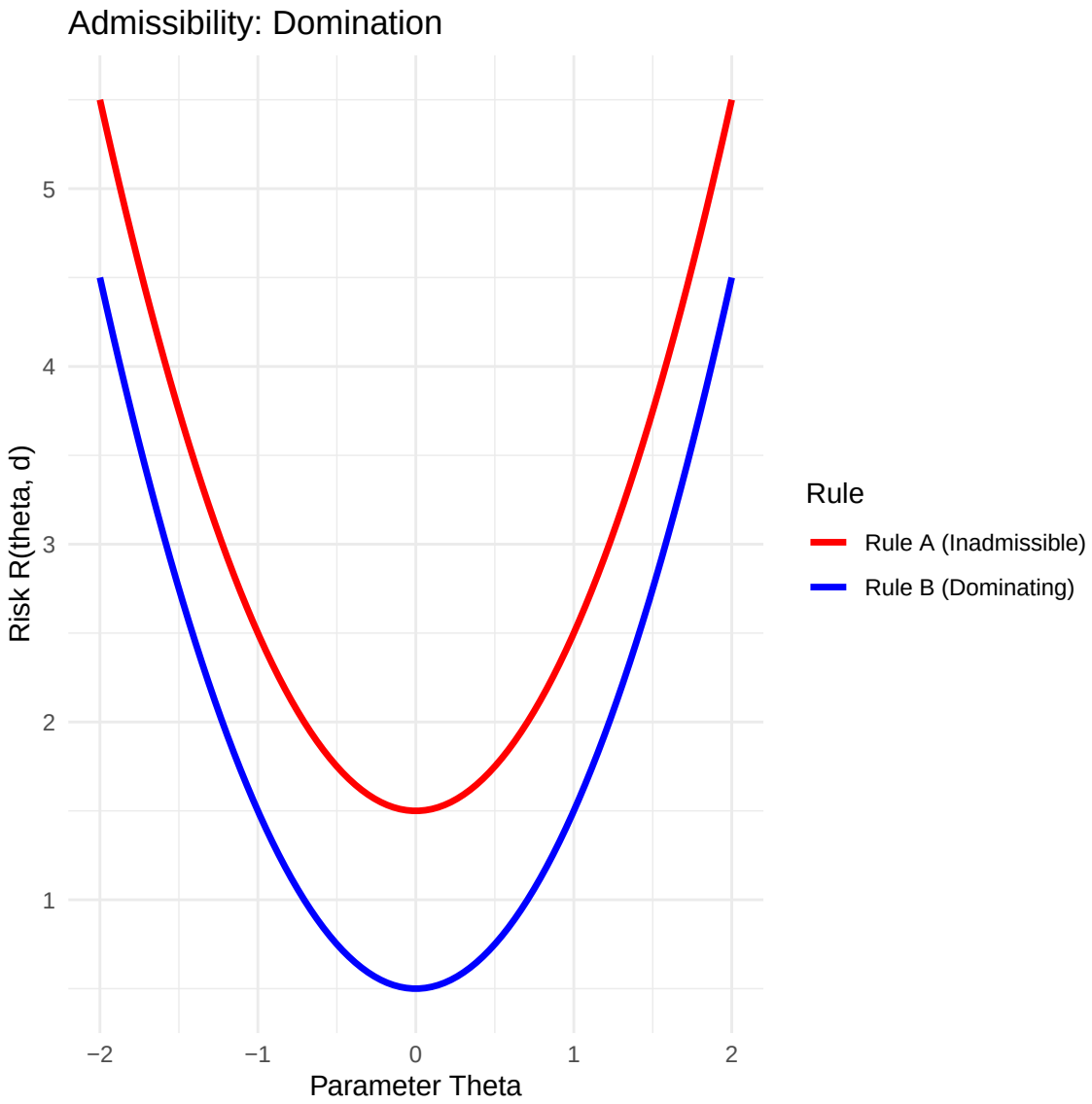


Figure 8.1: Illustration of Domination: Rule A (Red) is inadmissible because Rule B (Blue) has lower risk for all values of theta.

8.5.2 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.

$$\min_{d} \left[\sup_{\theta} R(\theta, d) \right] \tag{8.2}$$

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.

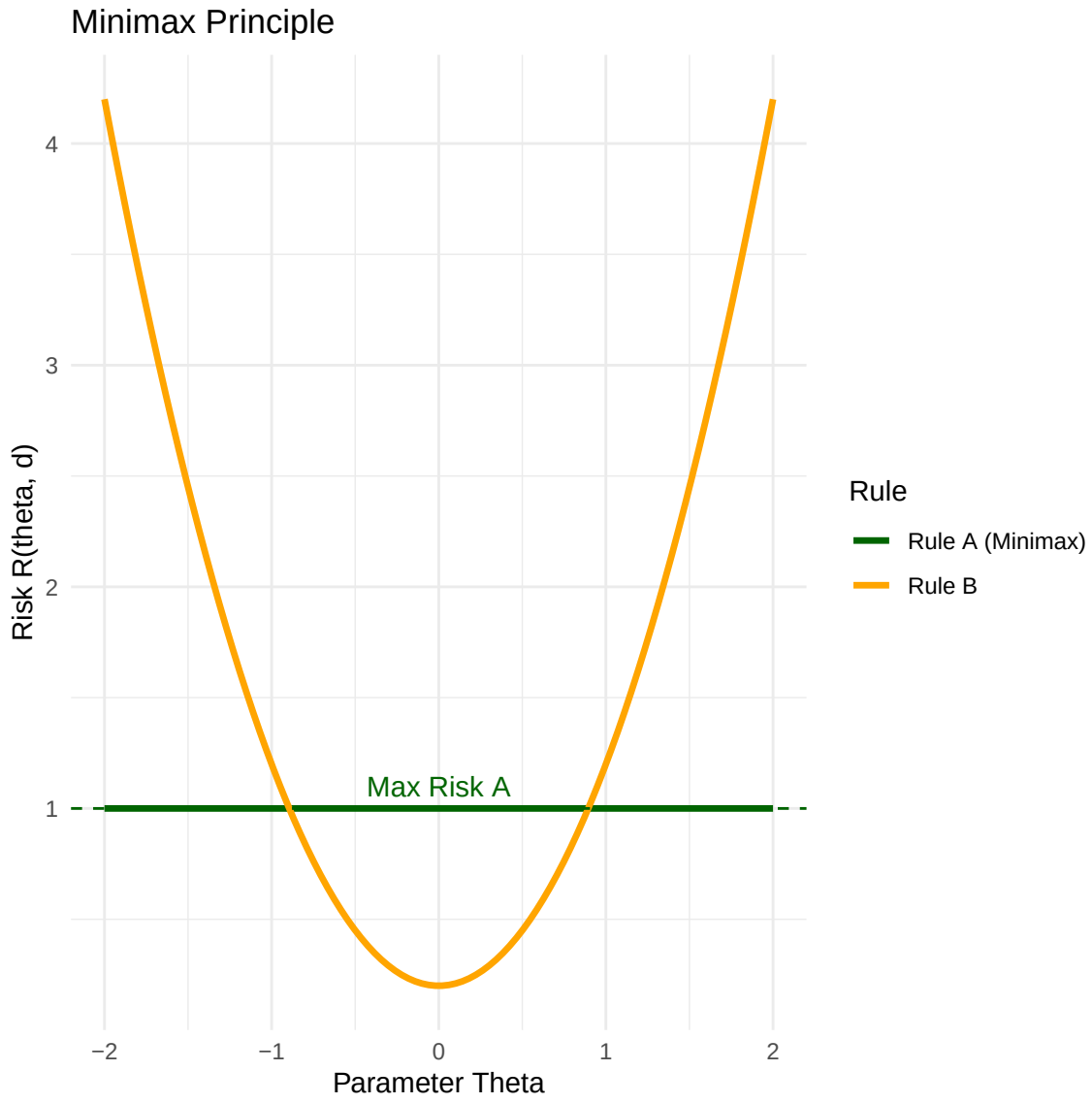


Figure 8.2: Illustration of Minimax: Rule A has a lower peak risk than Rule B, making Rule A the Minimax choice.

8.5.3 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.

$$r(\pi, d) = E_{\pi}[R(\theta, d)] = \int_{\Theta} R(\theta, d)\pi(\theta)d\theta \tag{8.3}$$

8.6 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 $R_2 = R(\theta_2)$.

8.6.1 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.

8.6.2 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).

8.6.3 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.

8.6.4 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 $-\pi_1/\pi_2$.

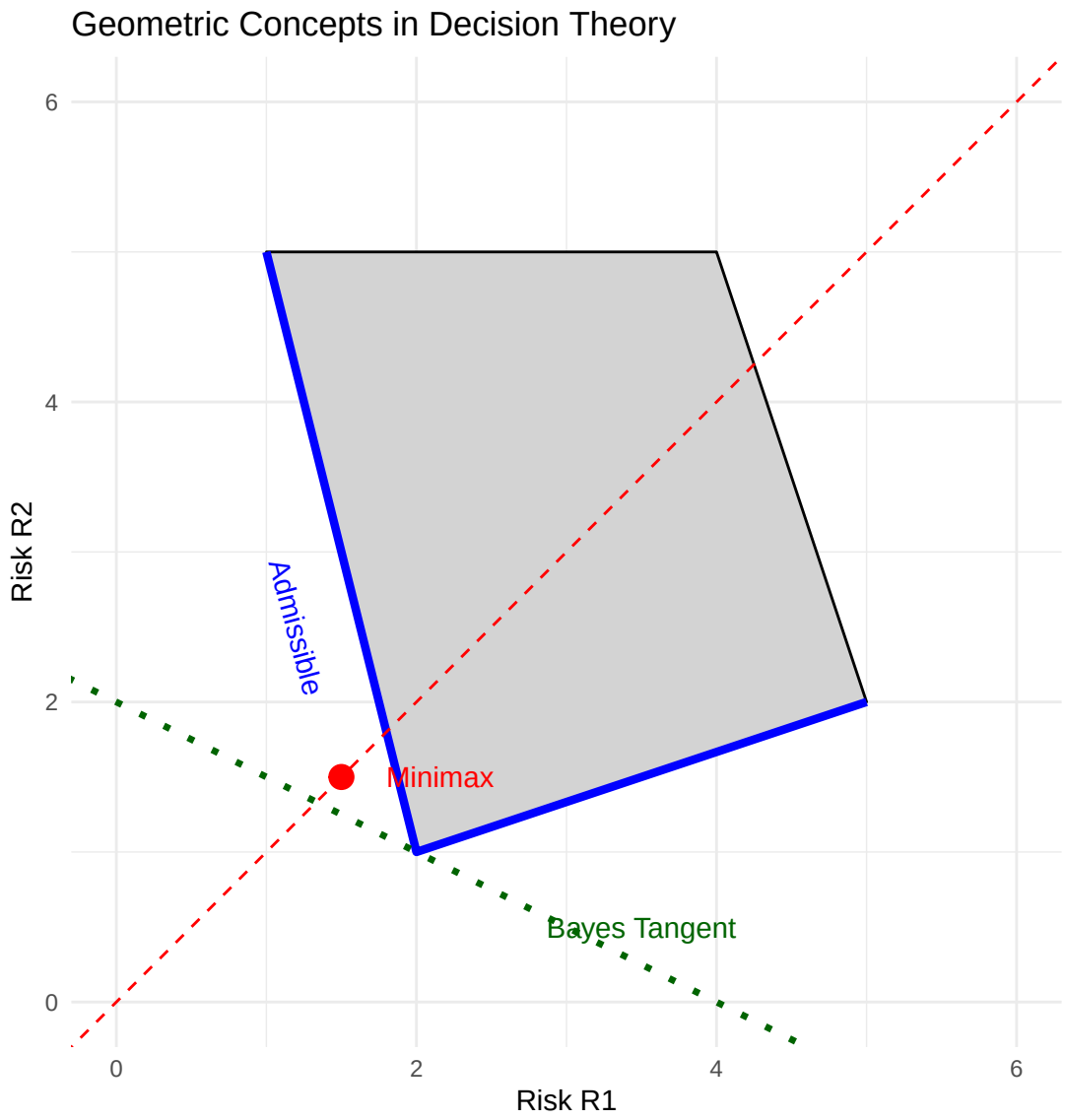


Figure 8.3: Geometric Interpretation: The gray polygon is the Risk Set S. The blue boundary represents admissible rules. The red point is the Minimax rule. The green line represents a Bayes rule for a specific prior.

8.7 Revisiting the Necklace Example: Geometric Solution

We now apply the geometric interpretation to the Necklace problem using the risks calculated in Section 8.4.

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

8.7.1 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 $1/3$.

8.8 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.

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

Theorem 8.1 (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:

$$\sup_{\theta \in \Theta} R(\theta, \delta') < \sup_{\theta \in \Theta} R(\theta, \delta^\pi) \tag{8.4}$$

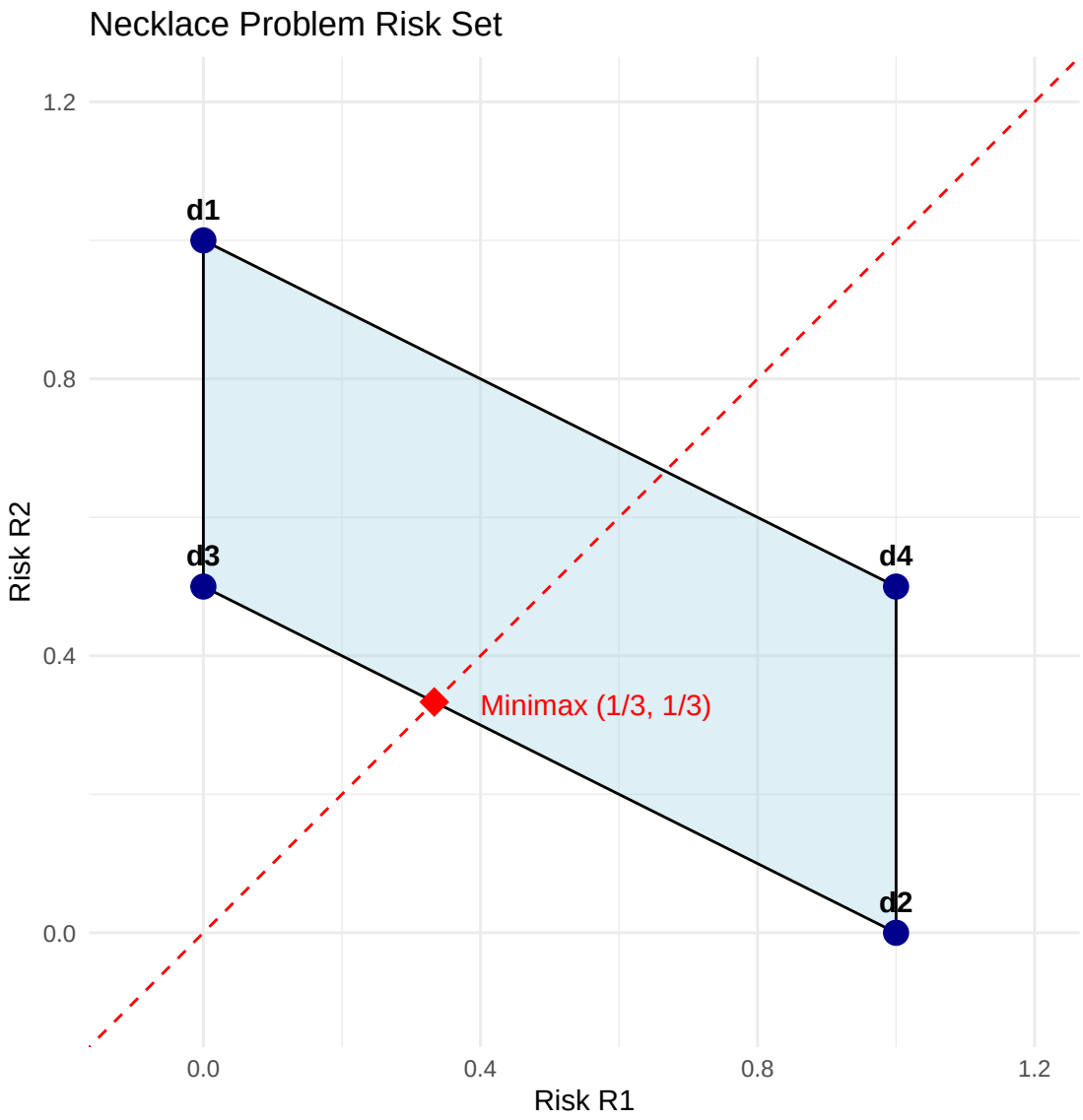


Figure 8.4: Necklace Problem Solution. The Minimax rule (red diamond) is the specific randomized combination of d3 and d2 that equalizes the risk.

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

$$\sup_{\theta \in \Theta} R(\theta, \delta') < c \tag{8.5}$$

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

$$r(\pi, \delta') = \int_\Theta R(\theta, \delta')\pi(\theta)d\theta \tag{8.6}$$

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:

$$r(\pi, \delta') \leq \sup_{\theta \in \Theta} R(\theta, \delta') < c \tag{8.7}$$

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

$$r(\pi, \delta^\pi) = \int_\Theta c\,\pi(\theta)d\theta = c \tag{8.8}$$

Substituting this into our inequality, we get:

$$r(\pi, \delta') < r(\pi, \delta^\pi) \tag{8.9}$$

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))
```

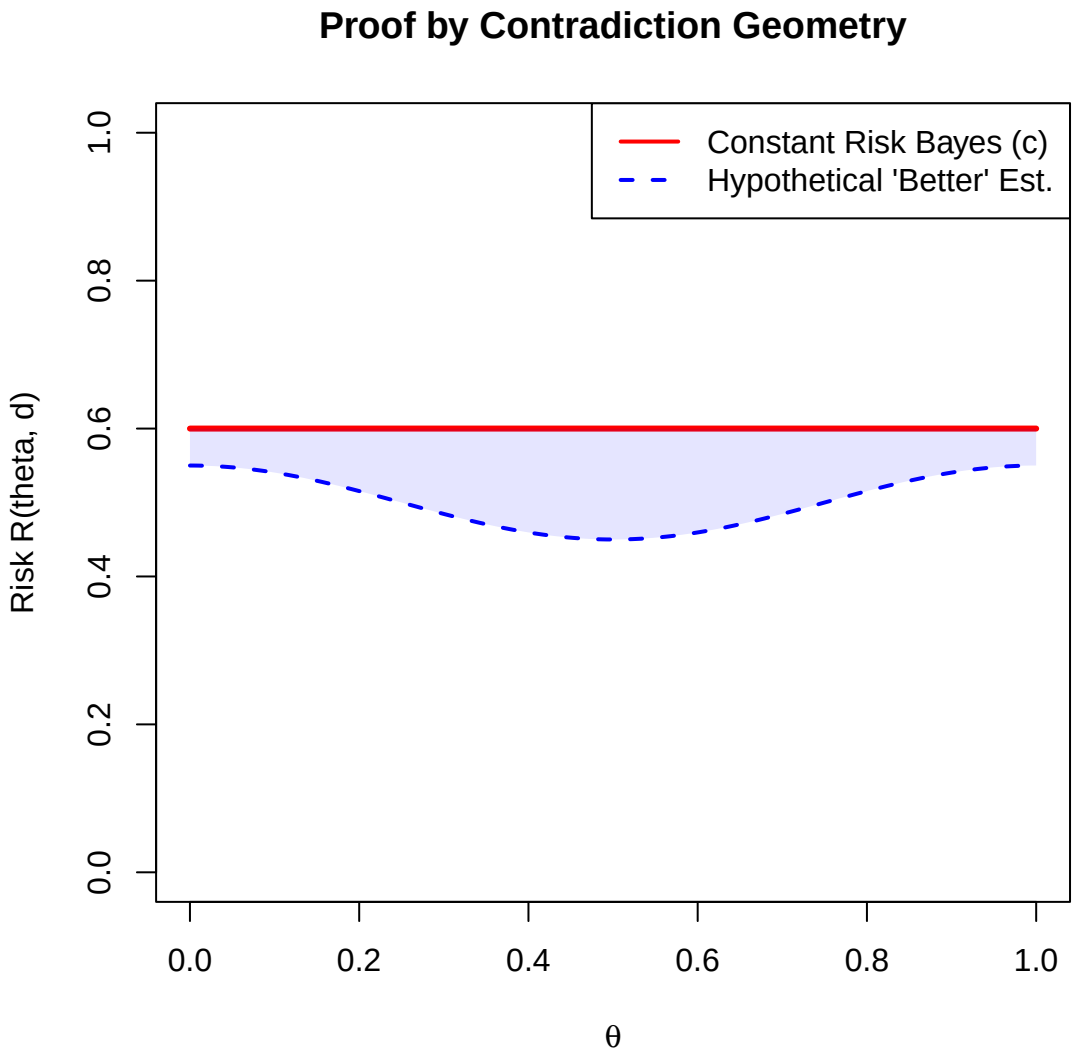


Figure 8.5: Visualizing the Contradiction: If the blue curve's maximum were below the red line, its average risk would be lower than the Bayes risk of the red estimator.

8.8.2 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 8.2 (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:*

$$\sup_\theta R(\theta, \delta_0) \leq \lim_{n\to\infty} r(\pi_n, \delta_n) \tag{8.10}$$

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:

   $$\sup_\theta R(\theta, \delta^*) < \sup_\theta R(\theta, \delta_0) \leq V \tag{8.11}$$

   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:

   $$r(\pi_n, \delta^*) = \int R(\theta, \delta^*)\pi_n(\theta)d\theta \leq V - \epsilon \tag{8.12}$$

   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):

   $$r(\pi_n, \delta_n) \leq r(\pi_n, \delta^*) \tag{8.13}$$
5. **The Contradiction:** Combining the inequalities, we get:

   $$\lim_{n\to\infty} r(\pi_n, \delta_n) \leq \lim_{n\to\infty} r(\pi_n, \delta^*) \leq V - \epsilon \tag{8.14}$$

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

□

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

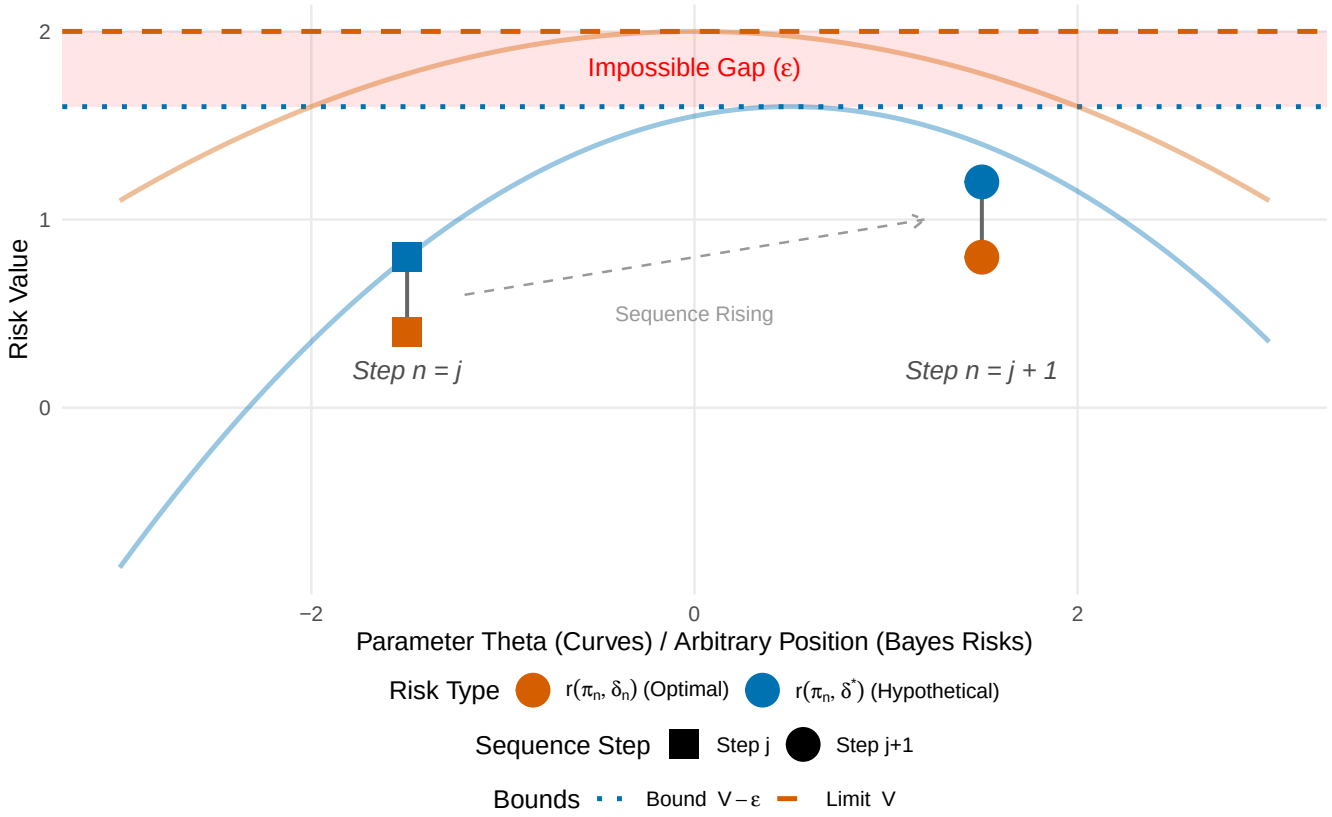


Figure 8.6: Visual Proof: We examine the Bayes risks at two steps, $n = j$ (squares) and $n = j + 1$ (circles). In both steps, the optimal risk $r(\pi_n, \delta_n)$ (orange) must be lower than the hypothetical risk $r(\pi_n, \delta^*)$ (blue). Even as the sequence rises (j+1 is higher than j), the blue points are capped by the bound $V - \epsilon$. This 'traps' the orange points, making it impossible for them to ever reach the Limit V.

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:

$$\lim_{n\to\infty} r(\pi_n, \delta_n) = \sup_\theta R(\theta, \delta_0) \tag{8.15}$$

   If this holds, δ_0 is Minimax.

Example 8.1 (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:

$$R(\theta, \delta_0) = E[(\theta - X)^2] = \text{Var}(X) = 1 \tag{8.16}$$

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:

$$r(\pi, \delta_\pi) = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} \tag{8.17}$$

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

$$r(\pi_n, \delta_n) = \frac{1 \cdot n}{1 + n} = \frac{n}{n+1} \tag{8.18}$$

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

$$\lim_{n\to\infty} r(\pi_n, \delta_n) = \lim_{n\to\infty} \frac{n}{n+1} = 1 \tag{8.19}$$

Comparing this to our candidate:

$$\sup_\theta R(\theta, \delta_0) = 1 \leq 1 \tag{8.20}$$

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

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

8.8.4.1 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 8.3 (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:

$$\inf_{\delta \in \mathcal{D}} \sup_{\pi \in \Pi} r(\pi, \delta) = \sup_{\pi \in \Pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) \tag{8.21}$$

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

$$r(\pi_0, \delta) \geq r(\pi_0, \delta_0) \geq r(\pi, \delta_0) \tag{8.22}$$

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:

$$\begin{aligned} &(A) \quad r(\pi_0, \delta_0) \leq r(\pi_0, \delta) \quad \text{for all } \delta \\ &(B) \quad r(\pi, \delta_0) \leq r(\pi_0, \delta_0) \quad \text{for all } \pi \end{aligned} \tag{8.23}$$

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:

$$\sup_\pi r(\pi, \delta_0) = \sup_\theta R(\theta, \delta_0) \tag{8.24}$$

- Therefore, Inequality (B) implies:

$$\sup_\theta R(\theta, \delta_0) = r(\pi_0, \delta_0) \tag{8.25}$$

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

$$\begin{aligned} \sup_\theta R(\theta, \delta^*) &= \sup_\pi r(\pi, \delta^*) && \text{(Max risk = Max average risk)} \\ &\geq r(\pi_0, \delta^*) && \text{(Supremum } \geq \text{ specific value)} \\ &\geq r(\pi_0, \delta_0) && \text{(From Inequality A: } \delta_0 \text{ is Bayes for } \pi_0) \\ &= \sup_\theta R(\theta, \delta_0) && \text{(From Step 3)} \end{aligned} \tag{8.26}$$

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

$$\sup_\theta R(\theta, \delta^*) \geq \sup_\theta R(\theta, \delta_0) \tag{8.27}$$

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

8.8.4.2 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.

8.9 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 8.4 (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:

$$r(\pi, \delta_\pi) - r(\pi, \delta') = \sum_{\theta \in \Theta} \pi(\theta)[R(\theta, \delta_\pi) - R(\theta, \delta')] \tag{8.28}$$

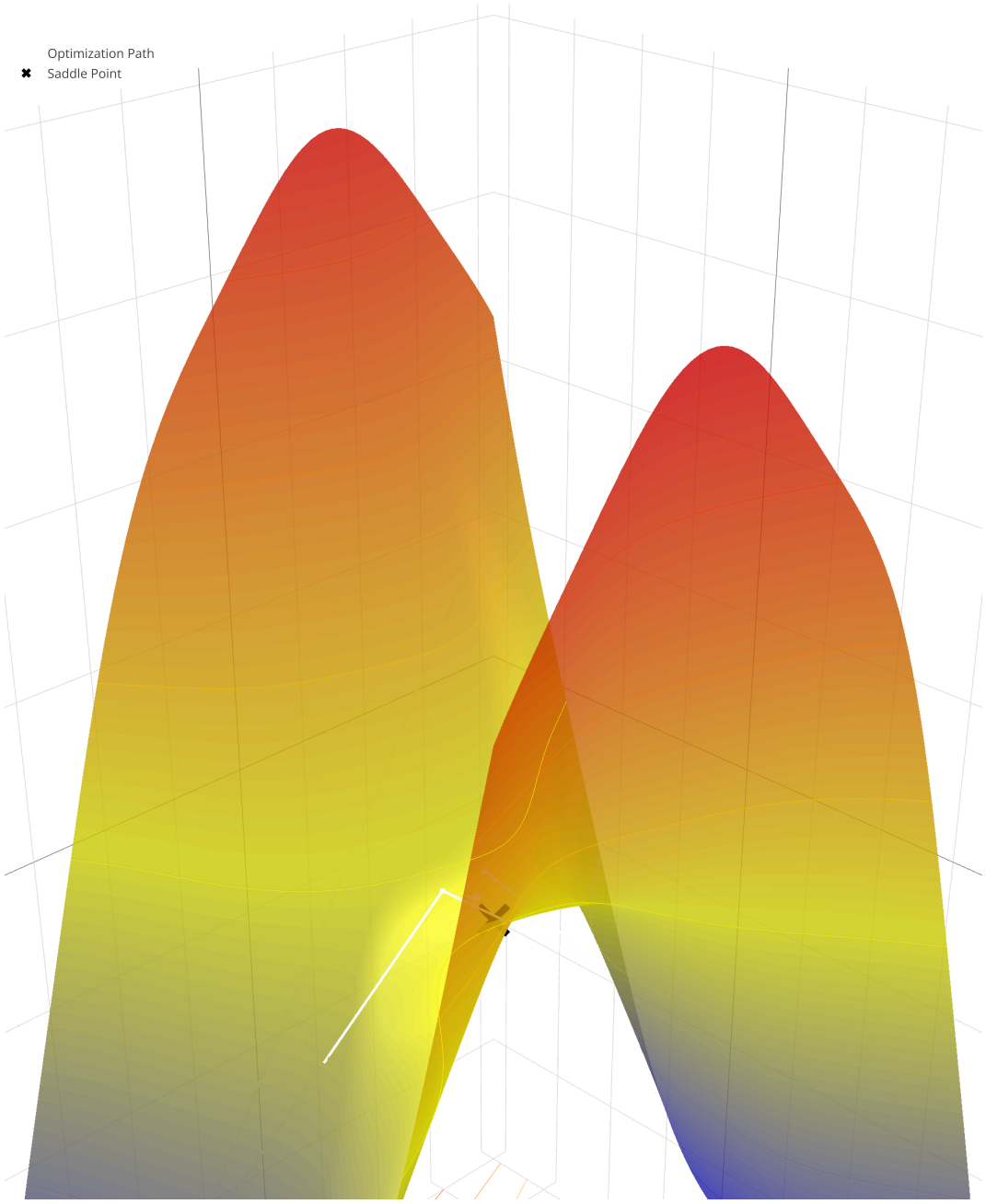


Figure 8.7: The 'Wiggle Mountain' of Risk. The surface represents Bayes Risk $r(\pi, \delta)$. The white zigzag line shows the iterative algorithm: starting from an arbitrary prior, we alternate between finding the best δ (moving along the valley) and the worst π (climbing the hill). This path spirals 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.
$$r(\pi, \delta_\pi) - r(\pi, \delta') > 0 \implies r(\pi, \delta') < r(\pi, \delta_\pi) \tag{8.29}$$
5. **Conclusion:** This contradicts the definition that δ_π is a Bayes rule (which must minimize Bayes risk). Therefore, δ_π is admissible. ■

□

8.9.1 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 8.5 (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:
$$r(\pi, \delta') = \int R(\theta, \delta')\pi(\theta)d\theta \leq \int R(\theta, \delta_\pi)\pi(\theta)d\theta = r(\pi, \delta_\pi) \tag{8.30}$$
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. ■

□

9 Bayesian Inference

9.1 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 9.1 (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:

$$\pi(\theta|x) = \frac{\pi(\theta)f(x;\theta)}{m(x)} \tag{9.1}$$

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

$$m(x) = \int_{\Theta} \pi(\theta)f(x;\theta)d\theta \tag{9.2}$$

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:

$$\pi(\theta|x) \propto \pi(\theta)f(x;\theta) \tag{9.3}$$

9.1.1 Discrete Posterior Calculation

Example 9.1 (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.

9.1.2 Binomial-beta Conjugacy

Example 9.2 (Binomial-beta Conjugacy)**.** Consider an experiment where $x|\theta \sim \text{Bin}(n, \theta)$. The likelihood function is:

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \tag{9.4}$$

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

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \tag{9.5}$$

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:

$$\pi(\theta|x) \propto \theta^{a-1}(1-\theta)^{b-1} \cdot \theta^x (1-\theta)^{n-x} \tag{9.6}$$

Combining terms with the same base:

$$\pi(\theta|x) \propto \theta^{a+x-1}(1-\theta)^{b+n-x-1} \tag{9.7}$$

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

$$\theta|x \sim \text{Beta}(a + x, b + n - x) \tag{9.8}$$

Properties of the Posterior:

- The posterior mean is:

$$E^{\theta|X}[\theta] = \frac{a + X}{a + b + n} \tag{9.9}$$

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

- The posterior variance is:

$$\text{Var}^{\theta|X}(\theta) = \frac{(a+X)(n+b-X)}{(a+b+n)^2(a+b+n+1)} \tag{9.10}$$

  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.

9.1.3 Normal-normal Conjugacy (known Variance)

Example 9.3 (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:

$$\pi(\mu|x) \propto \pi(\mu) \cdot f(x|\mu) \tag{9.11}$$

$$\propto \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \cdot \exp\left\{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right\} \tag{9.12}$$

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:

$$\tau_1 = \tau_0 + n\tau \tag{9.13}$$

$$\text{Posterior Precision} = \text{Prior Precision} + \text{Precision of Data} \tag{9.14}$$

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

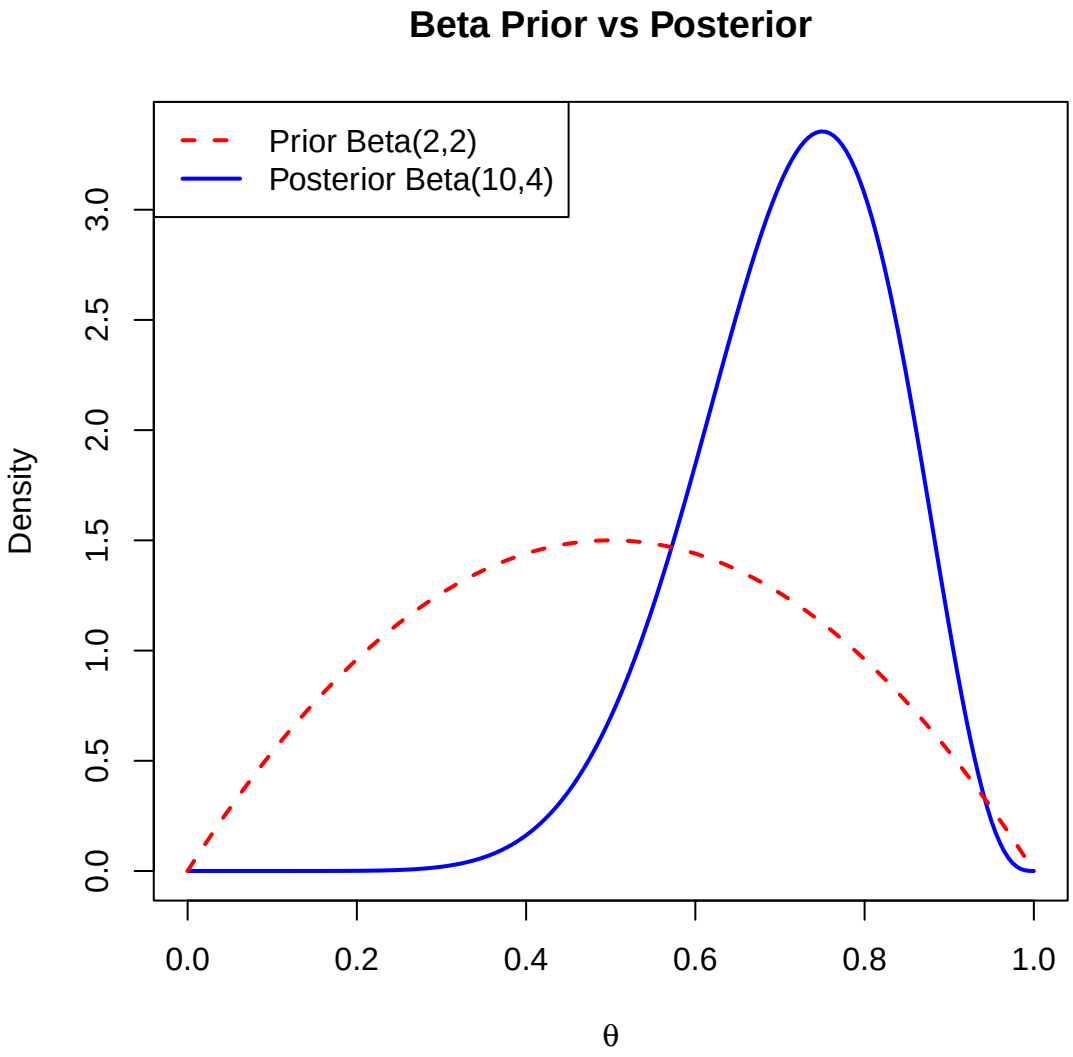


Figure 9.1: Prior vs Posterior for Beta-Binomial Example

$$\mu_1 = \frac{\mu_0\tau_0 + n\bar{x}\tau}{\tau_0 + n\tau} \tag{9.15}$$

So, the posterior distribution is:

$$\mu|x_1, \ldots, x_n \sim N\left(\frac{\mu_0\tau_0 + n\bar{x}\tau}{\tau_0 + n\tau}, \frac{1}{\tau_0 + n\tau}\right) \tag{9.16}$$

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 9.2 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)
```

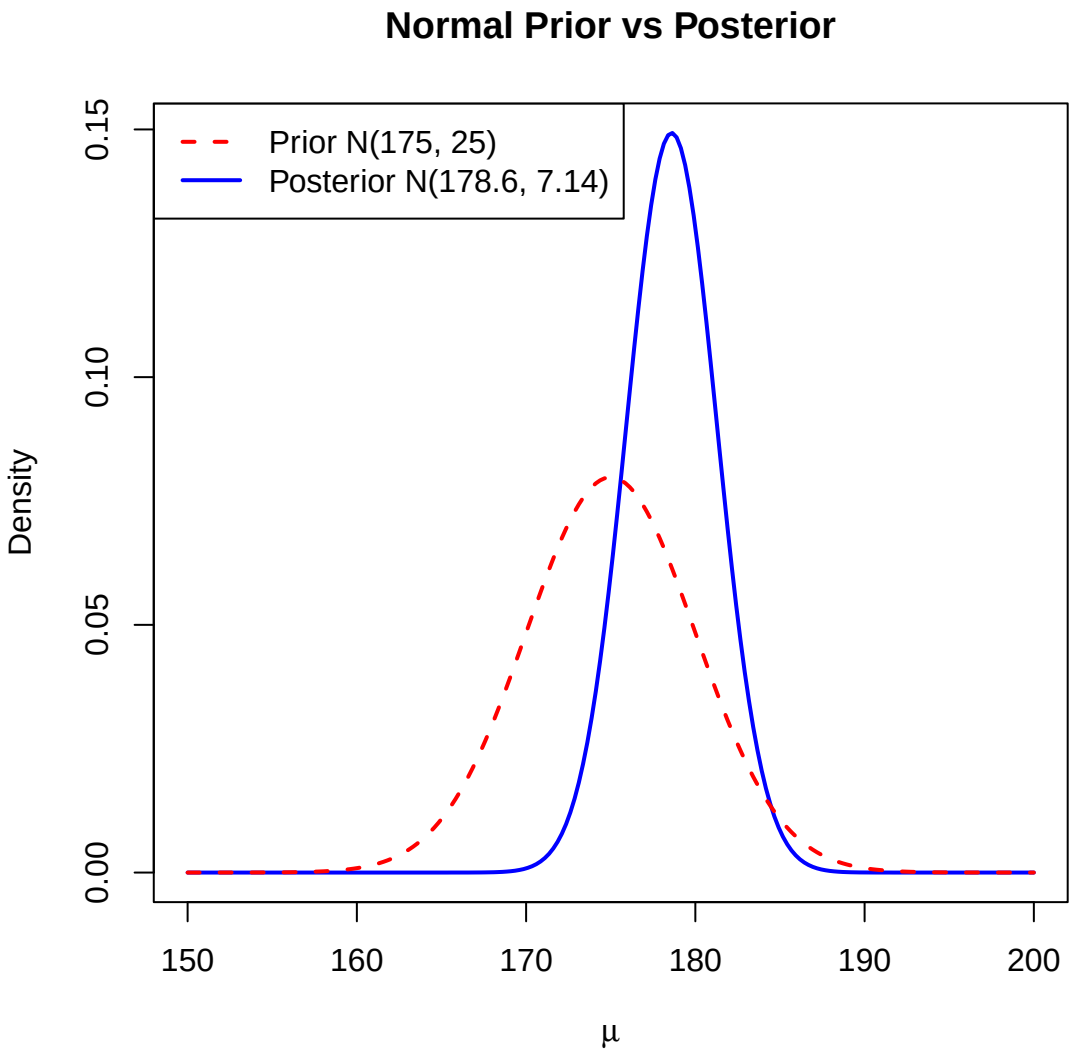


Figure 9.2: Prior vs Posterior for Normal-Normal Example

9.1.4 Normal with Unknown Mean and Variance

Example 9.4 (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 $\mu_0, \tau_0, \alpha_0, w_0$:

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

$$\pi(\tau) \propto \tau^{\alpha_0/2-1} \exp\left\{-\frac{\alpha_0 w_0}{2}\tau\right\} \tag{9.17}$$

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

$$\pi(\mu|\tau) \propto \tau^{1/2} \exp\left\{-\frac{\tau_0\tau}{2}(\mu-\mu_0)^2\right\} \tag{9.18}$$

The joint prior is:

$$\pi(\mu,\tau) \propto \tau^{(\alpha_0+1)/2-1} \exp\left\{-\frac{\tau}{2}\left(\alpha_0 w_0 + \tau_0(\mu-\mu_0)^2\right)\right\} \tag{9.19}$$

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

$$\begin{aligned} l(\mu,\tau) &= \log\left(\tau^{n/2}\exp\left\{-\frac{\tau}{2}[S_{xx}+n(\bar{x}-\mu)^2]\right\}\right) \\ &= \frac{n}{2}\log\tau - \frac{\tau}{2}\left[S_{xx}+n(\bar{x}-\mu)^2\right] + \text{const} \end{aligned} \tag{9.20}$$

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

$$\frac{\partial l}{\partial \mu} = n\tau(\bar{x}-\mu) = 0 \implies \hat{\mu}_{\text{MLE}} = \bar{x} \tag{9.21}$$

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

$$\frac{\partial l}{\partial \tau} = \frac{n}{2\tau} - \frac{S_{xx}}{2} = 0 \implies \hat{\tau}_{\text{MLE}} = \frac{n}{S_{xx}} \tag{9.22}$$

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

$$\hat{\sigma}^2_{\text{MLE}} = \frac{S_{xx}}{n} = \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n} \tag{9.23}$$

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:

$$\begin{aligned} \pi(\mu,\tau|x) \propto & \underbrace{\tau^{(\alpha_0+n)/2-1}\exp\left\{-\frac{\tau}{2}\left[\alpha_0 w_0 + S_{xx} + \frac{n\tau_0}{n+\tau_0}(\bar{x}-\mu_0)^2\right]\right\}}_{\text{Marginal of } \tau} \\ & \times \underbrace{\tau^{1/2}\exp\left\{-\frac{(n+\tau_0)\tau}{2}\left(\mu - \frac{\tau_0\mu_0 + n\bar{x}}{n+\tau_0}\right)^2\right\}}_{\text{Conditional of } \mu|\tau} \end{aligned} \tag{9.24}$$

Results:

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

$$\mu|\tau,x \sim N(\mu', 1/(\tau'\tau)) \tag{9.25}$$

$$E^{\mu|\tau,X}[\mu] = \frac{\tau_0\mu_0 + n\bar{x}}{\tau_0+n} \tag{9.26}$$

where

$$\tau' = \tau_0 + n \tag{9.27}$$

$$\mu' = \frac{\tau_0\mu_0 + n\bar{x}}{\tau_0+n} \tag{9.28}$$

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

$$\alpha' = \frac{\alpha_0 + n}{2}, \quad \beta' = \frac{\alpha_0 w_0 + n\hat{\sigma}^2_{\text{MLE}} + \frac{n\tau_0}{n+\tau_0}(\bar{x} - \mu_0)^2}{2} \tag{9.29}$$

  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:

$$E^{\sigma^2|X}[\sigma^2] \approx \frac{\alpha_0 w_0 + n\hat{\sigma}^2_{\text{MLE}} + \frac{1}{1/n+1/\tau_0}(\bar{x} - \mu_0)^2}{\alpha_0 + n} \tag{9.30}$$

- **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.

$$\tau|\mu, x \sim \text{Gamma}(\alpha'', \beta'') \tag{9.31}$$

  Where:

$$\alpha'' = \frac{\alpha_0 + n + 1}{2}, \quad \beta'' = \frac{\alpha_0 w_0 + \sum_{i=1}^n (x_i - \mu)^2 + \tau_0(\mu - \mu_0)^2}{2} \tag{9.32}$$

  The approximate expectation of the variance is:

$$E^{\sigma^2|\mu,X}[\sigma^2] \approx \frac{\alpha_0 w_0 + \sum_{i=1}^n (x_i - \mu)^2 + \tau_0(\mu - \mu_0)^2}{\alpha_0 + n + 1} \tag{9.33}$$

9.2 Finding Bayes Rules via Minimizing Posterior Expected Loss

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

Definition 9.2 (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,

$$R(\theta, d) = \int_X \mathcal{L}(\theta, d(x)) f(x; \theta) dx = E^{X|\theta}[\mathcal{L}(\theta, d(X))] \tag{9.34}$$

  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,

$$r(\pi, d) = \int_\Theta R(\theta, d)\pi(\theta) d\theta = E^{\theta}[R(\theta, d)] \tag{9.35}$$

  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:

$$\rho^{\text{Bayes}}(\pi, x) = \inf_d E^{\theta|x}[\mathcal{L}(\theta, d)] \tag{9.36}$$

Theorem 9.1 (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*

$$d^{Bayes}(x) = \arg\min_a E^{\theta|X}[\mathcal{L}(\theta, a)] \tag{9.37}$$

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 $R(\theta, d)$:

$$\begin{aligned} r(\pi, d) &= \int_\Theta R(\theta, d)\pi(\theta) d\theta \\ &= \int_\Theta \left[\int_X \mathcal{L}(\theta, d(x)) f(x|\theta) dx \right] \pi(\theta) d\theta \end{aligned} \tag{9.38}$$

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

$$r(\pi, d) = \int_X \left[\int_\Theta \mathcal{L}(\theta, d(x)) f(x|\theta)\pi(\theta) d\theta \right] dx \tag{9.39}$$

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:

$$
\begin{aligned}
r(\pi, d) &= \int_X \left[\int_\Theta \mathcal{L}(\theta, d(x)) \pi(\theta|x) m(x) d\theta \right] dx \\
&= \int_X m(x) \left[\int_\Theta \mathcal{L}(\theta, d(x)) \pi(\theta|x) d\theta \right] dx
\end{aligned}
\tag{9.40}
$$

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**:

$$
\int_\Theta \mathcal{L}(\theta, d(x)) \pi(\theta|x) d\theta = E^{\theta|X}[\mathcal{L}(\theta, d(X))] \tag{9.41}
$$

□

! 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:

$$
\boxed{r(\pi, d) = E^X \left[E^{\theta|X}(\mathcal{L}(\theta, d(X))) \right] = E^\theta \left[E^{X|\theta}(\mathcal{L}(\theta, d(X))) \right]} \tag{9.42}
$$

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

$$
m(x) = \int_\Theta f(x|\theta) \pi(\theta) d\theta \tag{9.43}
$$

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:

9.3 Special Bayes Rules

9.3.1 Squared Error Loss (point Estimate)

$$
\mathcal{L}(\theta, a) = (\theta - a)^2 \tag{9.44}
$$

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:

$$
-2E^{\theta|X}(\theta - d) = 0 \implies d(X) = E^{\theta|X}[\theta] \tag{9.45}
$$

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

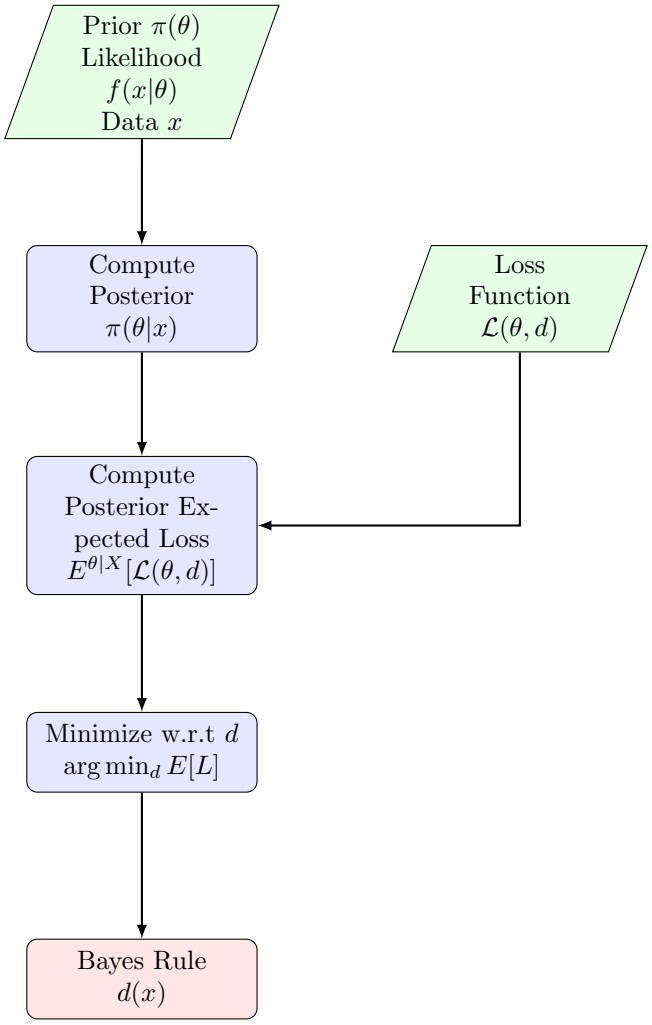


Figure 9.3: Workflow for Finding the Bayes Rule

9.3.2 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 $\theta = 1000$.

$$\mathcal{L}(\theta, d) = \left(\frac{d-\theta}{\theta}\right)^2 = \left(\frac{d}{\theta} - 1\right)^2 \tag{9.46}$$

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

$$Q(d) = E^{\theta|X}\left[\frac{d^2}{\theta^2} - \frac{2d}{\theta} + 1\right] = d^2 E^{\theta|X}[\theta^{-2}] - 2dE^{\theta|X}[\theta^{-1}] + 1 \tag{9.47}$$

Differentiating with respect to d and setting to zero:

$$\frac{\partial Q}{\partial d} = 2dE^{\theta|X}[\theta^{-2}] - 2E^{\theta|X}[\theta^{-1}] = 0 \tag{9.48}$$

Solving for d:

$$d(X) = \frac{E^{\theta|X}[\theta^{-1}]}{E^{\theta|X}[\theta^{-2}]} \tag{9.49}$$

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}.

9.3.3 Absolute Error Loss

$$\mathcal{L}(\theta, d) = |\theta - d| \tag{9.50}$$

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

$$\psi(d) = E^{\theta|X}[|\theta - d|] = \int_{-\infty}^{\infty} |\theta - d| \, dF(\theta|x) \tag{9.51}$$

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

$$\psi(d) = \int_{-\infty}^{d} (d - \theta) \, dF(\theta|x) + \int_{d}^{\infty} (\theta - d) \, dF(\theta|x) \tag{9.52}$$

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):

$$\frac{\partial}{\partial d}\psi(d) = \int_{-\infty}^{d} 1\, dF(\theta|x) - \int_{d}^{\infty} 1\, dF(\theta|x) = P(\theta \leq d|x) - P(\theta > d|x) \tag{9.53}$$

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

$$P(\theta \leq d|x) = P(\theta > d|x) \tag{9.54}$$

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**.

9.3.4 Weighted Absolute Error Loss (min-normalization)

$$\mathcal{L}(\theta, d) = \frac{|\theta - d|}{\min(\theta, 1-\theta)} \tag{9.55}$$

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.

$$\psi(d) = E^{\theta|X}\left[\frac{|\theta - d|}{\min(\theta, 1-\theta)}\right] = \int \frac{|\theta - d|}{\min(\theta, 1-\theta)}\pi(\theta|x)\, d\theta \tag{9.56}$$

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

$$\pi^*(\theta|x) \propto w(\theta)\pi(\theta|x) = \frac{\pi(\theta|x)}{\min(\theta, 1-\theta)} \tag{9.57}$$

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

Algorithm 9.1 (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:*

$$W_i = w(\theta_i) = \frac{1}{\min(\theta_i, 1 - \theta_i)} \tag{9.58}$$

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:*

$$S_k = \sum_{j=1}^{k} W_{(j)} \quad \text{for } k = 1, \dots, M \tag{9.59}$$

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

$$k^* = \min\{k : S_k \geq 0.5 \times S_M\} \tag{9.60}$$

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	
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) }
```

```r
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

# 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 9.1: Weighted Median Estimation

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

9.3.5 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 9.2: 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):**

$$E^{\theta|X}[\mathcal{L}(\theta, 0)] = 0 \cdot P(\theta \in \Theta_0|x) + 1 \cdot P(\theta \in \Theta_1|x) = P(\theta \in \Theta_1|x) \tag{9.61}$$

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

$$E^{\theta|X}[\mathcal{L}(\theta, 1)] = 1 \cdot P(\theta \in \Theta_0|x) + 0 \cdot P(\theta \in \Theta_1|x) = P(\theta \in \Theta_0|x) \tag{9.62}$$

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

$$P(\theta \in \Theta_0|x) \leq P(\theta \in \Theta_1|x) \tag{9.63}$$

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:

$$d(x) = \begin{cases} 1 & \text{if } P(\Theta_0|x) \leq \frac{1}{2} \quad \text{(Reject } H_0\text{)} \\ 0 & \text{if } P(\Theta_0|x) > \frac{1}{2} \quad \text{(Accept } H_0\text{)} \end{cases} \tag{9.64}$$

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 9.3: 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** $a = 0$**:**

$$E^{\theta|X}[\mathcal{L}(\theta, 0)] = 0 \cdot P(\Theta_0|x) + 1 \cdot P(\Theta_1|x) = P(\Theta_1|x) \tag{9.65}$$

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

$$E^{\theta|X}[\mathcal{L}(\theta, 1)] = c \cdot P(\Theta_0|x) + 0 \cdot P(\Theta_1|x) = cP(\Theta_0|x) \tag{9.66}$$

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

$$cP(\Theta_0|x) \leq P(\Theta_1|x) \tag{9.67}$$

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

$$cP(\Theta_0|x) \leq 1 - P(\Theta_0|x) \implies (1 + c)P(\Theta_0|x) \leq 1 \tag{9.68}$$

$$P(\Theta_0|x) \leq \frac{1}{1 + c} \tag{9.69}$$

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:

$$d(x) = \begin{cases} 1 & \text{if } P(\Theta_0|x) \leq \frac{1}{1+c} \quad \text{(Reject } H_0\text{)} \\ 0 & \text{if } P(\Theta_0|x) > \frac{1}{1+c} \quad \text{(Accept } H_0\text{)} \end{cases} \tag{9.70}$$

9.3.6 Classification Prediction

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

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

$$\mathcal{L}(y, \hat{y}) = \begin{cases} 0 & \text{if } \hat{y} = y \text{ (Correct Classification)} \\ 1 & \text{if } \hat{y} \neq y \text{ (Misclassification)} \end{cases} \tag{9.71}$$

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

$$E^{Y|X}[\mathcal{L}(y, \hat{y})] = \sum_{y} \mathcal{L}(y, \hat{y}) P(y|x) \tag{9.72}$$

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

$$E^{Y|X}[\mathcal{L}(y, \hat{y})] = \sum_{y \neq \hat{y}} 1 \cdot P(y|x) = P(y \neq \hat{y}|x) = 1 - P(y = \hat{y}|x) \tag{9.73}$$

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**.

$$\hat{y}_{\text{Bayes}}(x) = \underset{k \in \{1, \dots, K\}}{\arg\max} \; P(y = k|x) \tag{9.74}$$

9.3.7 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:

$$\mathcal{L}(\theta, d) = \begin{cases} 0 & \text{if } |\theta - d| \leq \delta \quad (\theta \in [d - \delta, d + \delta]) \\ 1 & \text{if } |\theta - d| > \delta \quad (\theta \notin [d - \delta, d + \delta]) \end{cases} \tag{9.75}$$

Derivation of the Bayes Rule

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

$$E^{\theta|X}[\mathcal{L}(\theta, d)] = 1 \cdot P(|\theta - d| > \delta|x) = 1 - P(d - \delta \leq \theta \leq d + \delta|x) \tag{9.76}$$

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

$$d_{\text{Bayes}} = \arg\max_{d} \int_{d-\delta}^{d+\delta} \pi(\theta|x)\, d\theta \tag{9.77}$$

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

$$\frac{\partial}{\partial d}\left(\int_{d-\delta}^{d+\delta} \pi(\theta|x)\, d\theta\right) = \pi(d + \delta|x) - \pi(d - \delta|x) = 0 \tag{9.78}$$

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)**.

9.4 Finding Minimax Rules with Bayes Rules

Theorem 8.1 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.

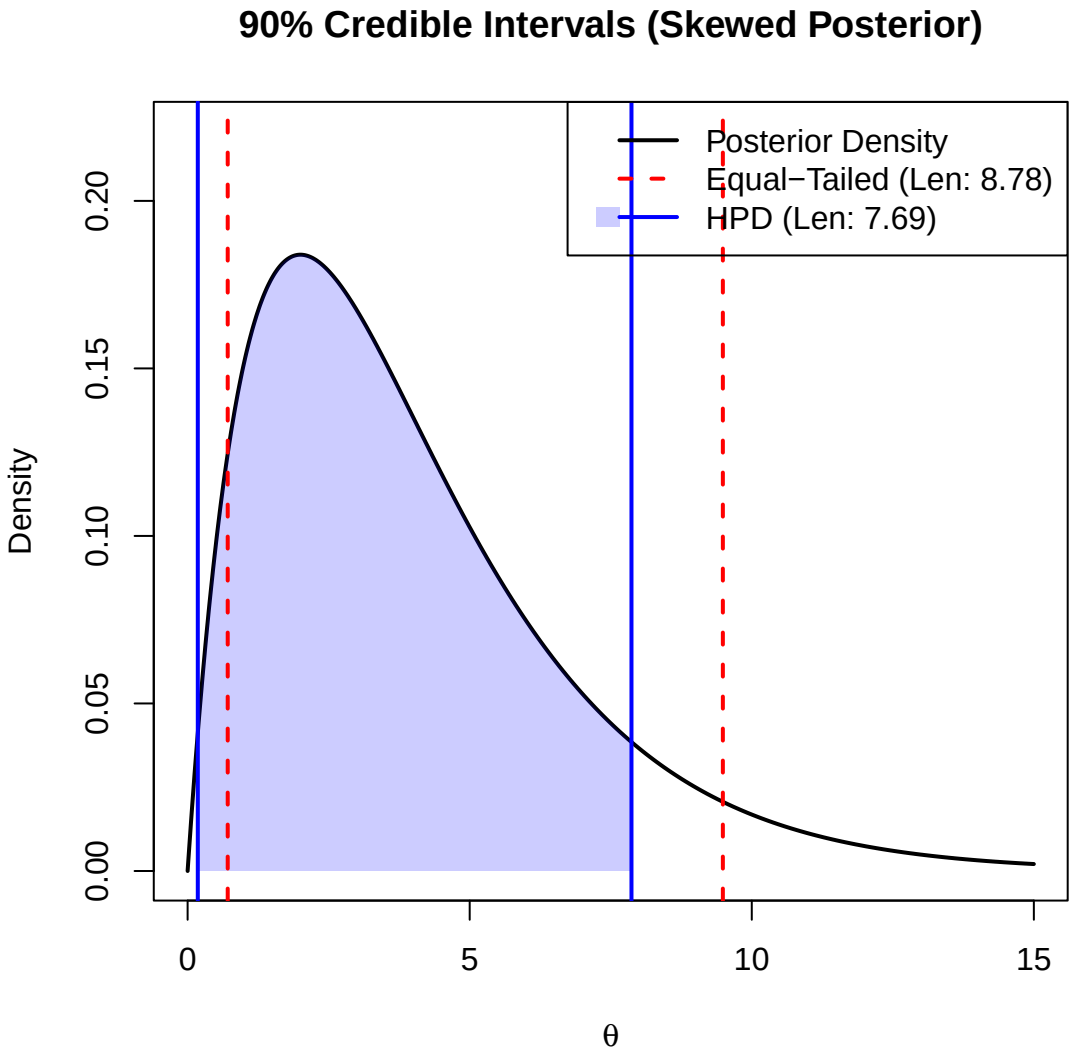


Figure 9.4: Comparison of HPD and Equal-Tailed Intervals for a Skewed Distribution

9.4.1 Binomial Minimax Estimator

Example 9.5. 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:

$$d(X) = \frac{a + X}{a + b + n} \tag{9.79}$$

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

$$R(\theta, d) = E^{X|\theta}\left[\left(\theta - \frac{a + X}{a + b + n}\right)^2\right] \tag{9.80}$$

Let $c = a + b + n$.

$$R(\theta, d) = \frac{1}{c^2} E^{X|\theta}\left[(c\theta - a - X)^2\right] \tag{9.81}$$

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:

$$a = b = \frac{\sqrt{n}}{2} \tag{9.82}$$

Thus, the minimax estimator is:

$$d(X) = \frac{X + \sqrt{n}/2}{n + \sqrt{n}} \tag{9.83}$$

This differs from the standard MLE $\hat{p} = X/n$ and the uniform prior Bayes estimator ($a = b = 1$).
According to Theorem 8.2, 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

$$\sup_\theta R(\theta, \delta_0) \leq \lim_{n \to \infty} r(\pi_n, \delta_n) \tag{9.84}$$

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 $p_0 = 1/2$:

$$d(X) = \underbrace{\left(\frac{n}{n + \sqrt{n}}\right)}_{w} \underbrace{\left(\frac{X}{n}\right)}_{\hat{p}} + \underbrace{\left(\frac{\sqrt{n}}{n + \sqrt{n}}\right)}_{1-w} \underbrace{\left(\frac{1}{2}\right)}_{p_0} \tag{9.85}$$

$$d(X) = w\hat{p} + (1 - w)p_0 \tag{9.86}$$

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.

9.4.2 Exponential Minimax Estimation

Let's recall Theorem 8.2. If there exists a rule δ_0 such that:

$$\sup_{\theta} R(\theta, \delta_0) \leq \lim_{n\to\infty} r(\pi_n, \delta_n), \tag{9.87}$$

where $\{\delta_n\}$ is a sequence of Bayes rules with respect to priors $\{\pi_n\}$ and $r(\pi_n, \delta_n)$ is the associated Bayes risks. Then δ_0 is Minimax. This theorem is frequently used when a minimax estimator corresponds to an "improper" prior (a prior that does not integrate to 1, such as a uniform distribution on an infinite interval). Since Bayes rules cannot be directly defined for improper priors in the standard risk framework, we approximate the improper prior with a sequence of proper priors $\{\pi_k\}$. If the risk of our proposed estimator δ_0 acts as a ceiling that the Bayes risks approach from below, δ_0 effectively guards against the "least favorable" conditions, satisfying the minimax criterion.

Example 9.6 (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**:

$$\mathcal{L}(\theta, d) = \left(\frac{d}{\theta} - 1\right)^2 \tag{9.88}$$

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:

$$L(\theta|x) = \theta^{-n} e^{-\frac{1}{\theta}\sum_{i=1}^{n} x_i} \tag{9.89}$$

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:

$$\pi_k(\theta) = \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \theta^{-\alpha_k - 1} e^{-\beta_k/\theta}, \quad \theta > 0 \tag{9.90}$$

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

$$\pi(\theta|x) \propto \left(\theta^{-n} e^{-T/\theta}\right) \cdot \left(\theta^{-\alpha_k - 1} e^{-\beta_k/\theta}\right) \propto \theta^{-(n+\alpha_k)-1} e^{-(T+\beta_k)/\theta} \tag{9.91}$$

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:

$$d_{\pi_k}(X) = \frac{E^{\theta|X}[\theta^{-1}]}{E^{\theta|X}[\theta^{-2}]} \tag{9.92}$$

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:

$$d_{\pi_k}(X) = \frac{\frac{\alpha^*}{\beta^*}}{\frac{\alpha^*(\alpha^*+1)}{(\beta^*)^2}} = \frac{\beta^*}{\alpha^*+1} = \frac{T+\beta_k}{n+\alpha_k+1} \tag{9.93}$$

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:

$$r(\pi_k, d_{\pi_k}) = 1 - \frac{(E^{\theta|X}[\theta^{-1}])^2}{E^{\theta|X}[\theta^{-2}]} = 1 - \frac{n+\alpha_k}{n+\alpha_k+1} = \frac{1}{n+\alpha_k+1} \tag{9.94}$$

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

$$\lim_{k\to\infty} r(\pi_k, d_{\pi_k}) = \frac{1}{n+1} \tag{9.95}$$

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 $\text{Var}^{X|\theta}(Y) = n$.

$$\begin{aligned} R(\theta, d_0) &= E^{X|\theta}\left[\left(\frac{d_0}{\theta} - 1\right)^2\right] = E^{X|\theta}\left[\left(\frac{Y}{n+1} - 1\right)^2\right] \\ &= \text{Var}^{X|\theta}\left(\frac{Y}{n+1}\right) + \left(E^{X|\theta}\left[\frac{Y}{n+1}\right] - 1\right)^2 \\ &= \frac{n}{(n+1)^2} + \left(\frac{n}{n+1} - 1\right)^2 \\ &= \frac{1}{n+1} \end{aligned} \tag{9.96}$$

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**.

9.5 Stein's Paradox and the James-stein Estimator

9.5.1 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:

$$d^{JS}(X) = \left(1 - \frac{p-2}{||X||^2}\right) X \tag{9.97}$$

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.

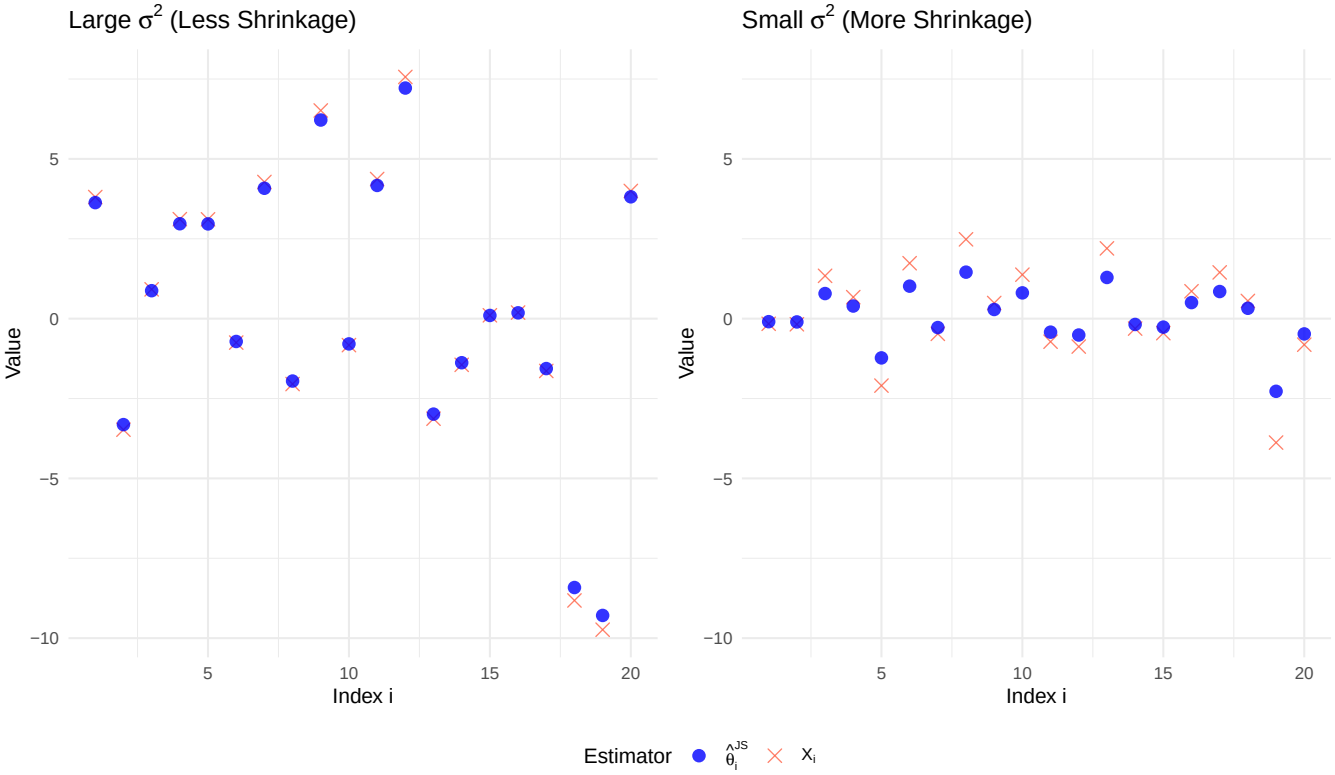


Figure 9.5: Visualization of JS Estimator

9.5.2 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:

$$L(\theta; x) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_i - \theta_i)^2}{2}\right) \tag{9.98}$$

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

$$\begin{aligned} \ell(\theta) &= \ln\left(\prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_i - \theta_i)^2}{2}\right)\right) \\ &= \sum_{i=1}^{p} \left[\ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{(X_i - \theta_i)^2}{2}\right] \end{aligned} \tag{9.99}$$

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:

$$\frac{\partial \ell}{\partial \theta_i} = (X_i - \theta_i) \tag{9.100}$$

Setting the derivative to zero gives the critical point:

$$X_i - \hat{\theta}_i = 0 \implies \hat{\theta}_{i,\text{MLE}} = X_i \tag{9.101}$$

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

$$d^{\text{MLE}}(X) = X \tag{9.102}$$

9.5.3 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:

$$X_i | \theta_i \sim N(\theta_i, 1), \quad \text{for } i = 1, \dots, p \tag{9.103}$$

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

$$\theta_i \sim N(0, \sigma^2), \quad \text{for } i = 1, \dots, p \tag{9.104}$$

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.

$$\frac{1}{v_{\text{post}}} = \frac{1}{\sigma^2} + \frac{1}{1} = \frac{1+\sigma^2}{\sigma^2} \tag{9.105}$$

  Therefore, the posterior variance is:

$$v_{\text{post}} = \frac{\sigma^2}{1+\sigma^2} \tag{9.106}$$

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

$$\begin{aligned} E^{\theta_i|X_i}[\theta_i] &= v_{\text{post}} \left(\frac{0}{\sigma^2} + \frac{X_i}{1} \right) \\ &= \frac{\sigma^2}{1+\sigma^2} X_i \\ &= \left(1 - \frac{1}{1+\sigma^2} \right) X_i \end{aligned} \tag{9.107}$$

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

$$d^{\text{Bayes}}(X) = \left(1 - \frac{1}{1+\sigma^2} \right) X \tag{9.108}$$

9.5.3.1 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:

$$E^{\theta_i|X_i}[(\theta_i - d^{\text{Bayes}}(X_i))^2] = v_{\text{post}} = \frac{\sigma^2}{1+\sigma^2} \tag{9.109}$$

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

$$r(\pi, d^{\text{Bayes}}) = \sum_{i=1}^{p} \frac{\sigma^2}{1+\sigma^2} = \frac{p\sigma^2}{1+\sigma^2} \tag{9.110}$$

9.5.3.2 Minimaxity of the MLE

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

Theorem 9.2 (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:*

$$\sup_{\theta \in \mathbb{R}^p} R(\theta, d^0) = \inf_d \sup_{\theta \in \mathbb{R}^p} R(\theta, d) = p \tag{9.111}$$

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:

$$r(\pi_{\sigma^2}, d^{\text{Bayes}}) = \frac{p\sigma^2}{1+\sigma^2} \tag{9.112}$$

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

$$\lim_{\sigma^2 \to \infty} \frac{p\sigma^2}{1+\sigma^2} = p \tag{9.113}$$

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.

□

9.5.4 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:

$$\nabla \cdot g(X) \equiv \sum_{i=1}^{p} \frac{\partial g_i(X)}{\partial X_i} \tag{9.114}$$

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

Lemma 9.1 (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:*

$$E^{X|\theta}\left[(X-\theta)^T g(X)\right] = E^{X|\theta}\left[\nabla \cdot g(X)\right] = E^{X|\theta}\left[\sum_{i=1}^{p} \frac{\partial g_i(X)}{\partial X_i}\right] \tag{9.115}$$

*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:

$$E^{X|\theta}[(X_i - \theta_i)g_i(X)] = \int_{\mathbb{R}^p} (x_i - \theta_i)g_i(x) \left(\prod_{j=1}^{p} \phi(x_j - \theta_j) \right) dx \tag{9.116}$$

Focusing on the integral with respect to x_i:

$$\int_{-\infty}^{\infty} (x_i - \theta_i)\phi(x_i - \theta_i)g_i(x)dx_i \tag{9.117}$$

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:

$$u = g_i(x) \quad \text{and} \quad dv = -\frac{\partial}{\partial x_i}\phi(x_i - \theta_i)dx_i \tag{9.118}$$

Thus:

$$\int_{-\infty}^{\infty} g_i(x)(x_i - \theta_i)\phi(x_i - \theta_i)dx_i = [-g_i(x)\phi(x_i - \theta_i)]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{\partial g_i(x)}{\partial x_i}\phi(x_i - \theta_i)dx_i \tag{9.119}$$

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 9.1 (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:*

$$E^{X|\theta}[X^T g(X)] = \theta^T E^{X|\theta}[g(X)] + E^{X|\theta}[\nabla \cdot g(X)] \tag{9.120}$$

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:

$$E^{X|\theta}[X^T X] = \theta^T E^{X|\theta}[X] + E^{X|\theta}[p] \tag{9.121}$$

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

$$E^{X|\theta}[\|X\|^2] = \|\theta\|^2 + p \tag{9.122}$$

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 9.2 (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:*

$$E^{X|\theta}\left[(X-\theta)^T g(X)\right] = E^{X|\theta}\left[p \cdot c(\|X\|^2) + 2\|X\|^2 \cdot c'(\|X\|^2)\right] \tag{9.123}$$

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:

$$\nabla \cdot (c(\|X\|^2)X) = c(\|X\|^2)(\nabla \cdot X) + X^T(\nabla c(\|X\|^2)) \tag{9.124}$$

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:

$$\nabla \cdot X = \sum_{i=1}^{p} \frac{\partial X_i}{\partial X_i} = p \tag{9.125}$$

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 $\nabla h(X) = 2X$.

$$\nabla c(\|X\|^2) = c'(\|X\|^2)\nabla(\|X\|^2) = 2c'(\|X\|^2)X \tag{9.126}$$

Substituting this back into the divergence formula:

$$\begin{aligned} \nabla \cdot g(X) &= p \cdot c(\|X\|^2) + X^T(2c'(\|X\|^2)X) \\ &= p \cdot c(\|X\|^2) + 2c'(\|X\|^2)\|X\|^2 \end{aligned} \tag{9.127}$$

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^2_p 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^2_p variable.
  - **The James-Stein Weight:** If we set $c(\|X\|^2) = \frac{1}{\|X\|^2}$, then $c'(z) = -\frac{1}{z^2}$.

$$\nabla \cdot g(X) = \frac{p}{\|X\|^2} + 2\|X\|^2\left(-\frac{1}{\|X\|^4}\right) = \frac{p-2}{\|X\|^2} \tag{9.128}$$

    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 9.7 (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:

$$E^{X|\theta}\left[(X-\theta)g(X)\right] = E^{X|\theta}\left[g'(X)\right] \tag{9.129}$$

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

$$g'(x) = \frac{d}{dx}(x^2) = 2x \tag{9.130}$$

Now, compute the expectation of the derivative:

$$\text{RHS} = E^{X|\theta}[g'(X)] = E^{X|\theta}[2X] = 2E^{X|\theta}[X] = 2\theta \tag{9.131}$$

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 $X - \theta = Z$.

$$\begin{aligned} \text{LHS} &= E^{X|\theta}\left[(X-\theta)X^2\right] \\ &= E^{X|\theta}\left[Z(Z+\theta)^2\right] \quad \text{where } Z \sim N(0,1) \\ &= E^{X|\theta}\left[Z(Z^2 + 2\theta Z + \theta^2)\right] \\ &= E^{X|\theta}[Z^3] + 2\theta E^{X|\theta}[Z^2] + \theta^2 E^{X|\theta}[Z] \end{aligned} \tag{9.132}$$

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:

$$\text{LHS} = 0 + 2\theta(1) + \theta^2(0) = 2\theta \tag{9.133}$$

Conclusion We observe that:

$$\text{LHS} = 2\theta \quad \text{and} \quad \text{RHS} = 2\theta \tag{9.134}$$

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

Example 9.8 (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:

$$E^{X|\theta}\left[(X-\theta)^T g(X)\right] = E^{X|\theta}\left[p \cdot c(\|X\|^2) + 2\|X\|^2 \cdot c'(\|X\|^2)\right] \tag{9.135}$$

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

$$\begin{aligned} \nabla \cdot g(X) &= p(\|X\|^2) + 2\|X\|^2(1) \\ &= (p+2)\|X\|^2 \end{aligned} \tag{9.136}$$

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, RHS $= (p+2)(p+\|\theta\|^2)$.
LHS (Alignment):

$$\begin{aligned} E^{X|\theta}\left[(X-\theta)^T(\|X\|^2 X)\right] &= E^{X|\theta}\left[\|X\|^2(X^TX - \theta^TX)\right] \\ &= E^{X|\theta}\left[\|X\|^4 - \theta^TX\|X\|^2\right] \end{aligned} \tag{9.137}$$

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 $E[\theta^TX\|X\|^2]$, we find it equals $\|\theta\|^4 + (p+2)\|\theta\|^2$.
Subtracting these:

$$\begin{aligned} \text{LHS} &= [p(p+2) + 2(p+2)\|\theta\|^2 + \|\theta\|^4] - [\|\theta\|^4 + (p+2)\|\theta\|^2] \\ &= p(p+2) + (p+2)\|\theta\|^2 \\ &= (p+2)(p+\|\theta\|^2) \end{aligned} \tag{9.138}$$

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 ($2\|X\|^2$).

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

Theorem 9.3. *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:

$$\begin{aligned} R(\theta, d^{JS}) &= E^{X|\theta}\left[||(X-\theta) - g(X)||^2\right] \\ &= E^{X|\theta}\left[||X-\theta||^2\right] - 2E^{X|\theta}\left[(X-\theta)^Tg(X)\right] + E^{X|\theta}\left[||g(X)||^2\right] \end{aligned} \tag{9.139}$$

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 9.2). We first compute the scalar function and its derivative:

$$c(z) = \frac{p-2}{z} \implies c'(z) = -\frac{p-2}{z^2} \tag{9.140}$$

Substituting these into the radial divergence formula from Lemma 9.2:

$$\begin{aligned}\nabla \cdot g(X) &= p \cdot c(\|X\|^2) + 2\|X\|^2 \cdot c'(\|X\|^2) \\ &= p\left(\frac{p-2}{\|X\|^2}\right) + 2\|X\|^2\left(-\frac{p-2}{\|X\|^4}\right) \\ &= \frac{p(p-2)}{\|X\|^2} - \frac{2(p-2)}{\|X\|^2} \\ &= \frac{(p-2)^2}{\|X\|^2}\end{aligned} \tag{9.141}$$

Applying the lemma to the cross-term:

$$2E^{X|\theta}\left[(X-\theta)^T g(X)\right] = 2E^{X|\theta}[\nabla \cdot g(X)] = 2(p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] \tag{9.142}$$

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

$$\|g(X)\|^2 = \left\|\frac{p-2}{\|X\|^2}X\right\|^2 = \frac{(p-2)^2}{\|X\|^4}\|X\|^2 = \frac{(p-2)^2}{\|X\|^2} \tag{9.143}$$

Substituting these results back into the risk equation:

$$\begin{aligned}R(\theta, d^{JS}) &= p - 2(p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] + (p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] \\ &= p - (p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right]\end{aligned} \tag{9.144}$$

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. □

9.5.6 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:

$$R(\theta, d^{JS}) = p - (p-2)^2 E\left[\frac{1}{\chi^2_p(\|\theta\|^2/2)}\right] \tag{9.145}$$

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:

$$R(\theta, d^{JS}) \approx p - \frac{(p-2)^2}{p + \|\theta\|^2} \tag{9.146}$$

- **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 $E^{X|\theta=0}[1/\|X\|^2] = 1/(p-2)$:

$$\lim_{\|\theta\|\to 0} \frac{R(\theta, d^{JS})}{R(\theta, d^{\text{MLE}})} = \frac{p - (p-2)^2\left(\frac{1}{p-2}\right)}{p} = \frac{p-(p-2)}{p} = \frac{2}{p} \tag{9.147}$$

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.

9.5.7 Using Normalized Loss (Optional)

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

$$\mathcal{L}(\theta, d) = \frac{\|d-\theta\|^2}{\|\theta\|^2}, \quad \theta \neq 0 \tag{9.148}$$

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):

$$R(\theta, d^{\text{MLE}}) = E^{X|\theta}\left[\frac{\|X-\theta\|^2}{\|\theta\|^2}\right] = \frac{1}{\|\theta\|^2} E^{X|\theta}\left[\|X-\theta\|^2\right] \tag{9.149}$$

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

$$R(\theta, d^{\text{MLE}}) = \frac{p}{\|\theta\|^2} \tag{9.150}$$

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

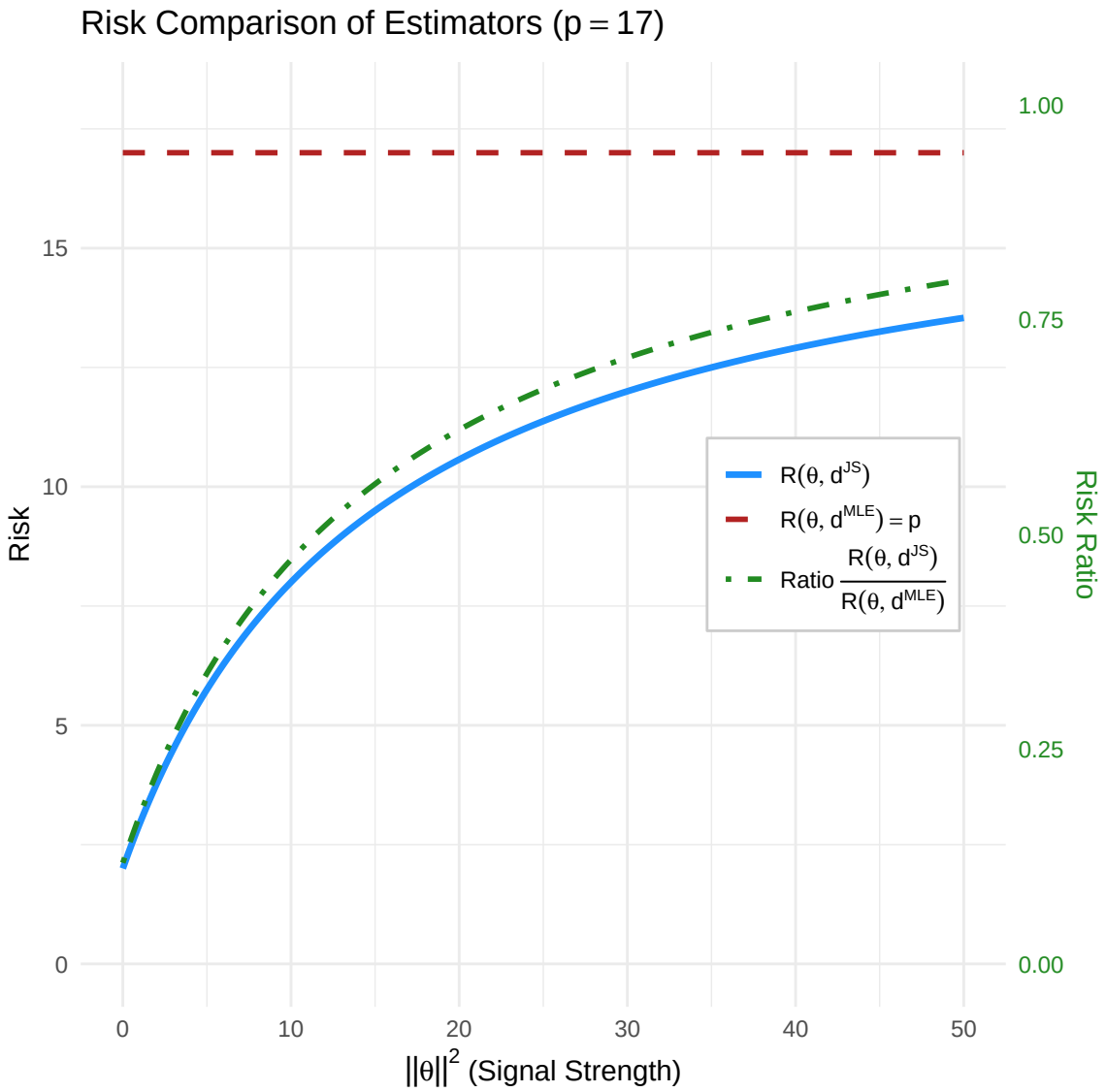


Figure 9.6: Risk comparison of James-Stein vs MLE (p=10). Note: Ratio uses the right axis.

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:

$$E^{X|\theta}[\|d^{JS} - \theta\|^2] = p - (p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] \tag{9.151}$$

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

$$R(\theta, d^{JS}) = \frac{1}{\|\theta\|^2}\left(p - (p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right]\right) \tag{9.152}$$

3. Comparison and Dominance

We compare the risks by taking the difference:

$$R(\theta, d^{\text{MLE}}) - R(\theta, d^{JS}) = \frac{1}{\|\theta\|^2}(p-2)^2 E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] \tag{9.153}$$

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 $\theta \neq 0$.

$$R(\theta, d^{JS}) < R(\theta, d^{\text{MLE}}) \tag{9.154}$$

- **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:

$$\frac{R(\theta, d^{JS})}{R(\theta, d^{\text{MLE}})} = \frac{p - (p-2)^2 E^{X|\theta}[1/\|X\|^2]}{p} = 1 - \frac{(p-2)^2}{p} E^{X|\theta}\left[\frac{1}{\|X\|^2}\right] \tag{9.155}$$

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:

$$\lim_{\|\theta\| \to 0} \frac{R(\theta, d^{JS})}{R(\theta, d^{\text{MLE}})} = 1 - \frac{(p-2)^2}{p(p-2)} = 1 - \frac{p-2}{p} = \frac{2}{p} \tag{9.156}$$

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.

9.5.8 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 9.4 (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:*

$$r(\pi, d^{JS}) = \frac{p\sigma^2 + 2}{\sigma^2 + 1} \tag{9.157}$$

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:

$$R(\theta, d^{JS}) = p - (p-2)^2 E^{X|\theta}\left[\frac{1}{||X||^2}\right] \tag{9.158}$$

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

$$r(\pi, d^{JS}) = \int R(\theta, d^{JS})\pi(\theta)d\theta = p - (p-2)^2 E^{\pi}\left[E^{X|\theta}\left(\frac{1}{||X||^2}\right)\right] \tag{9.159}$$

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:

$$E^X\left[\frac{1}{||X||^2}\right] = \frac{1}{1+\sigma^2}E\left[\frac{1}{\chi^2_p}\right] = \frac{1}{1+\sigma^2} \cdot \frac{1}{p-2} \tag{9.160}$$

Substituting this back into the risk equation:

$$\begin{aligned} r(\pi, d^{JS}) &= p - (p-2)^2 \cdot \frac{1}{(p-2)(1+\sigma^2)} \\ &= p - \frac{p-2}{1+\sigma^2} \\ &= \frac{p(1+\sigma^2) - (p-2)}{1+\sigma^2} \\ &= \frac{p\sigma^2 + p - p + 2}{1+\sigma^2} = \frac{p\sigma^2 + 2}{\sigma^2 + 1} \end{aligned} \tag{9.161}$$

□

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:

$$r(\pi, d) = E^X[E^{\theta|X}[\mathcal{L}(\theta, d(X))]] \tag{9.162}$$

Step 1: Posterior Expected Loss
The posterior distribution of θ given x is:

$$\theta|x \sim N\left(\frac{\sigma^2}{1+\sigma^2}x, \frac{\sigma^2}{1+\sigma^2}I\right) \tag{9.163}$$

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

$$E^{\theta|X}[||\theta - d^{JS}(X)||^2] = \text{tr}(\text{Var}^{\theta|X}(\theta)) + ||E^{\theta|X}[\theta] - d^{JS}(X)||^2 \tag{9.164}$$

- **Trace term:**

$$\text{tr}\left(\frac{\sigma^2}{1+\sigma^2}I_p\right) = \frac{p\sigma^2}{1+\sigma^2} \tag{9.165}$$

- **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:

$$E^{\theta|X}[\theta] - d^{JS}(X) = \left((1-B) - \left(1 - \frac{p-2}{||X||^2}\right)\right)X = \left(\frac{p-2}{||X||^2} - B\right)X \tag{9.166}$$

  Squaring the norm gives:

$$\left(\frac{p-2}{||X||^2} - B\right)^2 ||X||^2 = \frac{(p-2)^2}{||X||^2} - 2B(p-2) + B^2||X||^2 \tag{9.167}$$

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:**

$$\begin{aligned} E^X\left[\frac{(p-2)^2}{||X||^2} - \frac{2(p-2)}{1+\sigma^2} + \frac{||X||^2}{(1+\sigma^2)^2}\right] &= (p-2)^2\frac{1}{(p-2)(1+\sigma^2)} - \frac{2(p-2)}{1+\sigma^2} + \frac{p(1+\sigma^2)}{(1+\sigma^2)^2} \\ &= \frac{p-2}{1+\sigma^2} - \frac{2p-4}{1+\sigma^2} + \frac{p}{1+\sigma^2} \\ &= \frac{p-2-2p+4+p}{1+\sigma^2} \\ &= \frac{2}{1+\sigma^2} \end{aligned} \tag{9.168}$$

Step 3: Combine Terms

$$r(\pi, d^{JS}) = \underbrace{\frac{p\sigma^2}{1+\sigma^2}}_{\text{Variance Part}} + \underbrace{\frac{2}{1+\sigma^2}}_{\text{Bias Part}} = \frac{p\sigma^2+2}{\sigma^2+1} \tag{9.169}$$

Both methods yield the same result.

□

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

Example 9.9. 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:

$$\hat{\theta}_i^{JS} = \bar{X} + \left(1 - \frac{(p-3)\sigma^2}{\sum(X_i - \bar{X})^2}\right)(X_i - \bar{X}) \tag{9.170}$$

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.

9.5.10 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.

9.5.11 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.

9.5.12 Bias-Variance Decomposition for James-Stein Estimator

Based on the derivations in your handwritten notes, here are the corresponding LaTeX equations formatted for your Quarto document.

$$X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{9.171}$$

The Mean Squared Error (MSE) of the James-Stein estimator $d^{JS}(X)$ with respect to the parameter θ is decomposed by adding and subtracting the expected value $u_{d^{JS}} = E^{X|\theta}[d^{JS}(X)]$:

$$\begin{aligned} E^{X|\theta}\left[\|d^{JS}(X) - \theta\|^2\right] &= E^{X|\theta}\left[\|d^{JS} - u_{d^{JS}} + u_{d^{JS}} - \theta\|^2\right] \\ &= V(d^{JS}(X)) + \left[Bias(d^{JS}(X))\right]^2 \end{aligned} \tag{9.172}$$

where the components are defined as:

1. **Variance Component**
$$V(d^{JS}(X)) = E^{X|\theta}\left[\|d^{JS}(X) - u_{d^{JS}}\|^2\right] \tag{9.173}$$
2. **Bias Component**
$$Bias(d^{JS}(X)) = u_{d^{JS}} - \theta \tag{9.174}$$

Remark. Note that while the ordinary MLE X is unbiased ($E^{X|\theta}(X) - \theta = 0$), the James-Stein estimator introduces a deliberate bias to significantly reduce the variance, resulting in a lower total MSE when the dimension is three or greater.

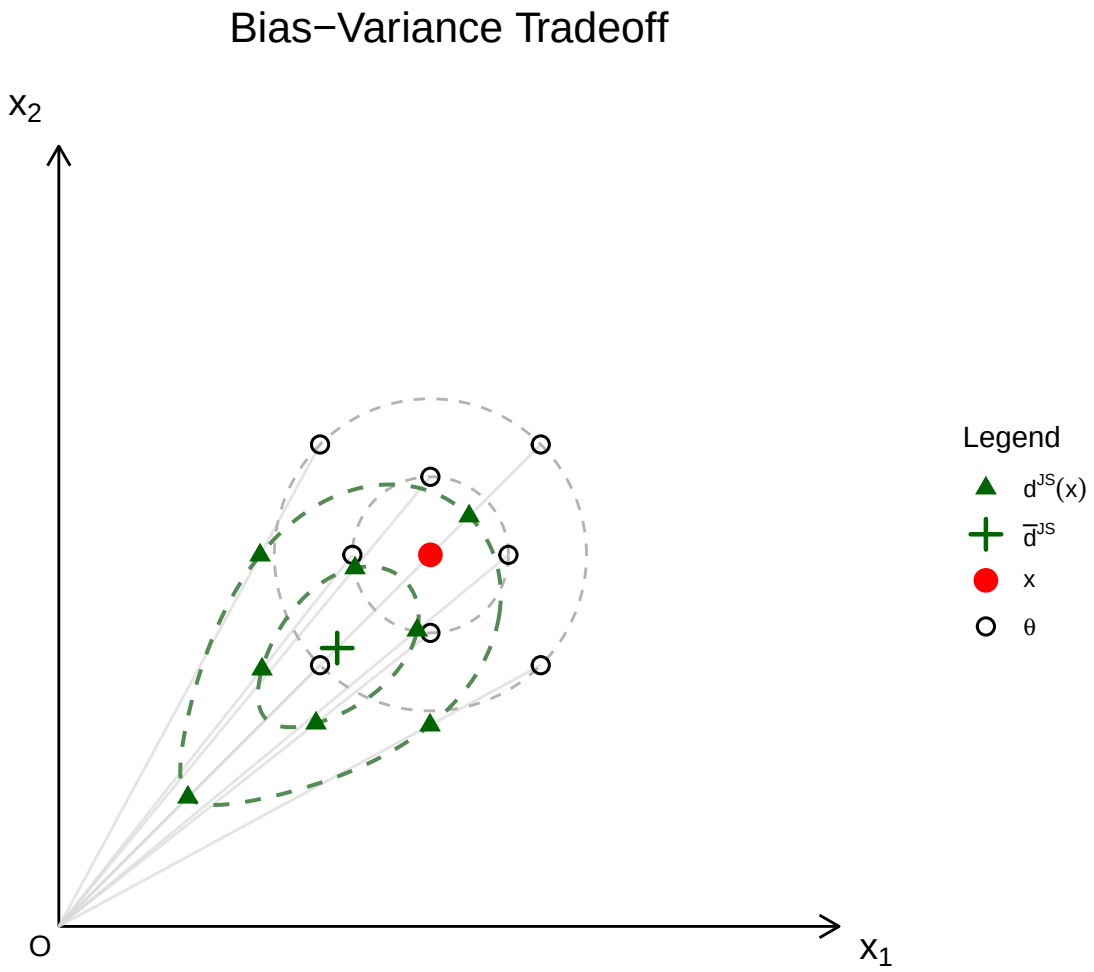


Figure 9.7: Variance-Bias Tradeoff: 8 Alternated Points with Origin Vectors

9.6 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.

9.6.1 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:
$$m(x|\eta) = \int f(x|\theta)\pi(\theta|\eta)d\theta \tag{9.175}$$
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:
$$\hat{\eta} = \arg\max_{\eta} m(x|\eta) \tag{9.176}$$
4. **Posterior Inference:** We proceed with standard Bayesian inference, but we substitute the estimated estimate $\hat{\eta}$ into the posterior:
$$\pi(\theta|x, \hat{\eta}) \propto f(x|\theta)\pi(\theta|\hat{\eta}) \tag{9.177}$$

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.

9.6.2 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:

$$E^{\mu_i|X_i,\sigma^2}[\mu_i] = \frac{\sigma^2}{1+\sigma^2}X_i = \left(1 - \frac{1}{1+\sigma^2}\right)X_i \tag{9.178}$$

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:

$$X_i \sim N(0, 1+\sigma^2) \tag{9.179}$$

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

$$S \sim (1+\sigma^2)\chi_p^2 \tag{9.180}$$

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:

$$E^X\left[\frac{p-2}{S}\right] = \frac{p-2}{1+\sigma^2}E^X\left[\frac{1}{\chi_p^2}\right] = \frac{p-2}{1+\sigma^2}\cdot\frac{1}{p-2} = \frac{1}{1+\sigma^2} = B \tag{9.181}$$

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:

$$\delta^{EB}(X) = \left(1 - \hat{B}\right)X = \left(1 - \frac{p-2}{||X||^2}\right)X \tag{9.182}$$

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.

9.7 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.

9.7.1 Hierarchical Model Structure

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

$$\begin{aligned} \text{Level 1 (Data Likelihood):} \quad & X_i|\mu_i, \sigma^2 \sim f(x_i|\mu_i, \sigma^2) \\ \text{Level 2 (Parameters):} \quad & \mu_i|\theta, \tau^2 \sim \pi(\mu_i|\theta, \tau^2) \\ \text{Level 3 (Hyperparameters):} \quad & \theta, \tau^2 \sim \pi(\theta, \tau^2) \end{aligned} \tag{9.183}$$

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

$$p(\mu, \theta, \tau^2|X) \propto \left[\prod_{i=1}^{n} f(x_i|\mu_i, \sigma^2)\pi(\mu_i|\theta, \tau^2)\right] \pi(\theta, \tau^2) \tag{9.184}$$

9.7.2 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.

9.7.3 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.

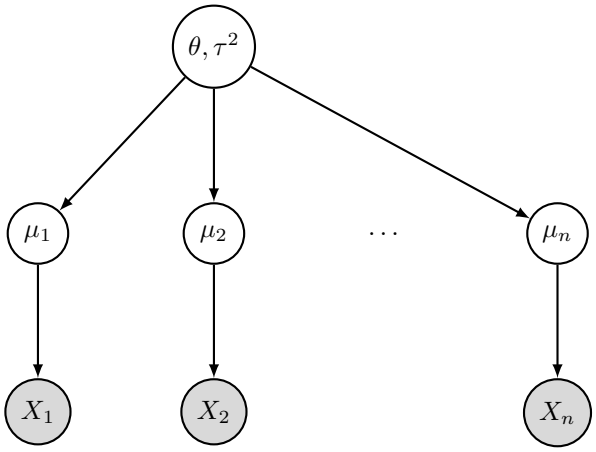


Figure 9.8: Hierarchical Tree Structure

9.7.3.1 Gibbs Sampling Algorithm

Algorithm 9.2. *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:*

$$\begin{aligned} \theta_1^{(t+1)} &\sim p(\theta_1 | \theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_k^{(t)}) \\ \theta_2^{(t+1)} &\sim p(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_k^{(t)}) \\ &\vdots \\ \theta_k^{(t+1)} &\sim p(\theta_k | \theta_1^{(t+1)}, \theta_2^{(t+1)}, \dots, \theta_{k-1}^{(t+1)}) \end{aligned} \tag{9.185}$$

Example 9.10 (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:

$$\mu_i^{(t+1)} \sim N\left(\frac{\tau^2 X_i + \sigma^2 \theta^{(t)}}{\sigma^2 + \tau^2}, \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}\right) \tag{9.186}$$

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

$$\theta^{(t+1)} \sim N\left(\frac{1}{n}\sum_{i=1}^{n} \mu_i^{(t+1)}, \frac{\tau^2}{n}\right) \tag{9.187}$$

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

9.7.3.2 Metropolis-hastings (MH) Sampling

Algorithm 9.3. *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:*

$$\alpha = \min\left(1, \frac{p(\theta^*|X)q(\theta^{(t)}|\theta^*)}{p(\theta^{(t)}|X)q(\theta^*|\theta^{(t)})}\right) \tag{9.188}$$

- *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.*

9.8 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.

9.8.1 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 $\sigma^2 = 1$.

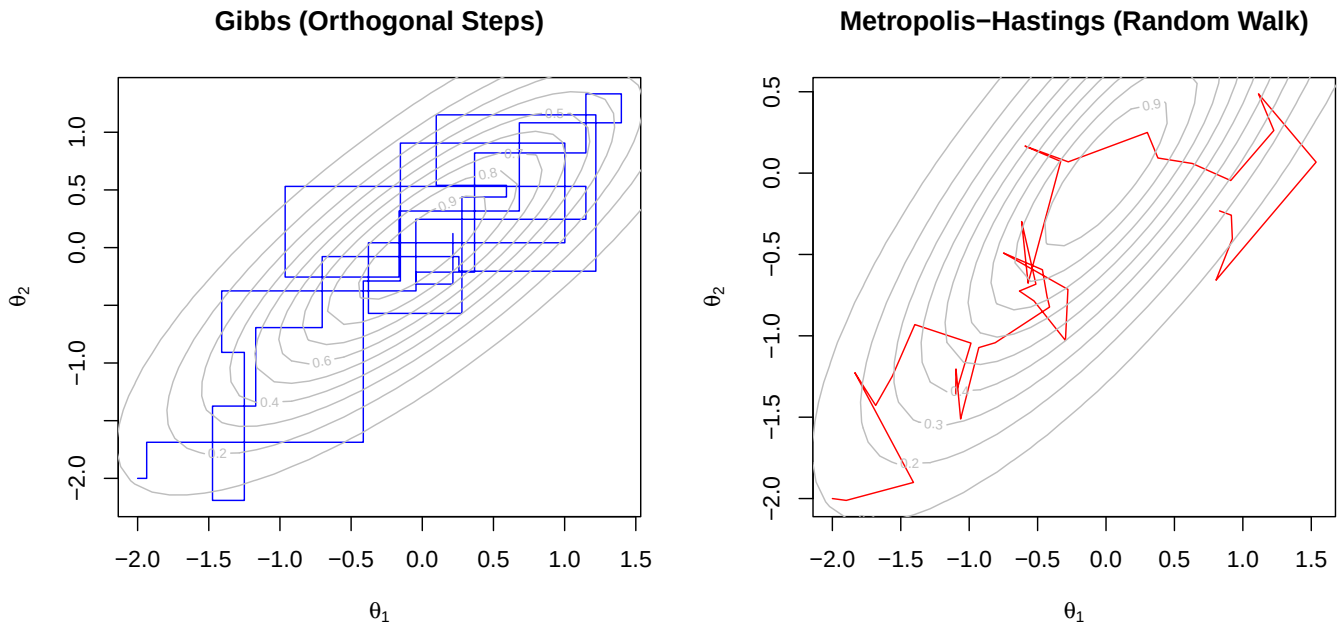


Figure 9.9: Comparison of Sampling Paths

$$x_i = \sqrt{n_i} \arcsin\left(2\frac{y_i}{n_i} - 1\right) \tag{9.189}$$

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

9.8.2 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:

$$\mu_i^{\text{season}} = \sqrt{n_i} \arcsin\left(2\frac{Y_i}{N_i} - 1\right) \tag{9.190}$$

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 9.4: 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

Player	y_i	n_i	p_i^{pre}	x_i	Y_i	N_i	p_i^{seas}	μ_i
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.

9.8.3 Methods for Estimating μ_i (transformed Scale)

9.8.3.1 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.

$$\hat{\mu}_i^{MLE} = X_i \tag{9.191}$$

```
# 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)
```

9.8.3.2 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.

$$\hat{\mu}_i^{JS} = \bar{X} + \left(1 - \frac{k-3}{\sum (X_i - \bar{X})^2}\right)(X_i - \bar{X}) \tag{9.192}$$

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)
```

9.8.3.3 Method 3: Fully Bayesian MCMC (brms)

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

$$\begin{aligned} X_i &\sim N(\mu_i, 1) \\ \mu_i &\sim N(\theta, \tau^2) \\ \theta &\sim N(0, 10) \\ \tau &\sim \text{Cauchy}(0, 2) \end{aligned} \tag{9.193}$$

```
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)
```

9.8.4 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 9.5: Comparison of Estimates (Sorted by Pre-season x_i)

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

Plots of 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))
```

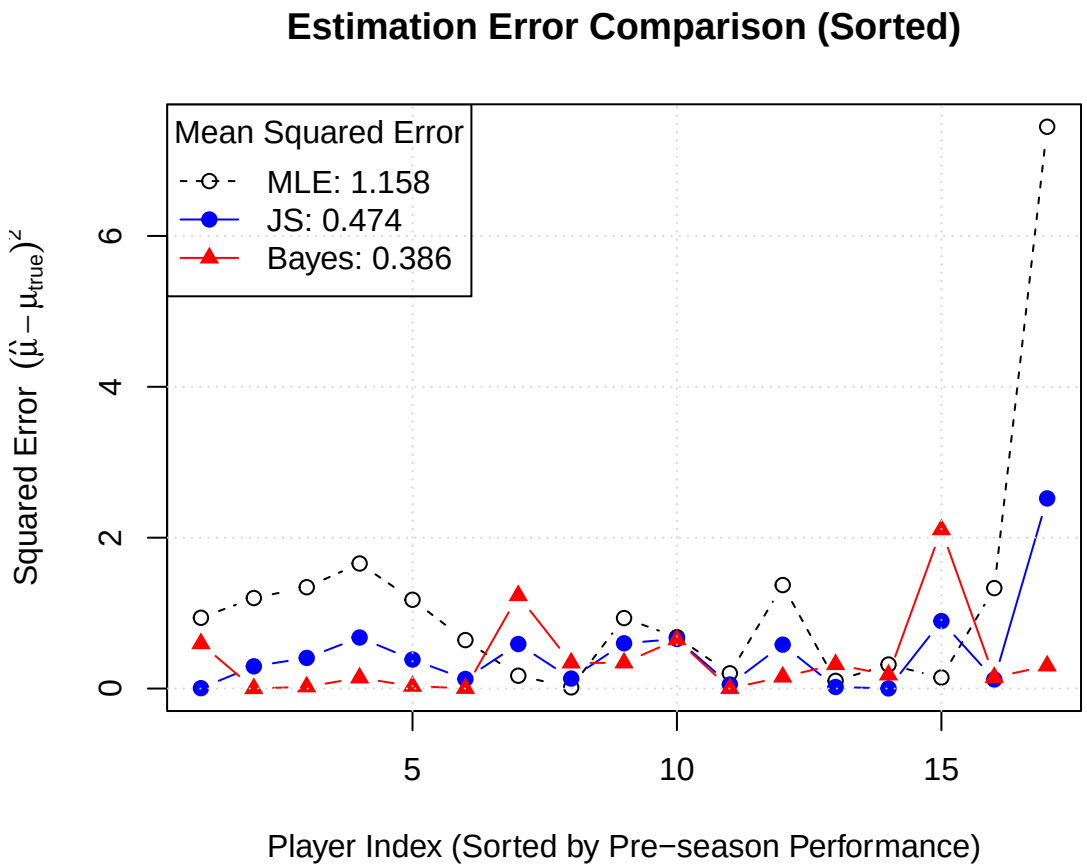


Figure 9.10: Squared Error by Sorted Player Index (Transformed Scale)

9.8.5 Methods for Estimating p_i Directly

9.8.5.1 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:

$$\hat{p}_i = \frac{1}{2}\left(\sin\left(\frac{\hat{\mu}_i}{\sqrt{n_i}}\right) + 1\right) \tag{9.194}$$

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.

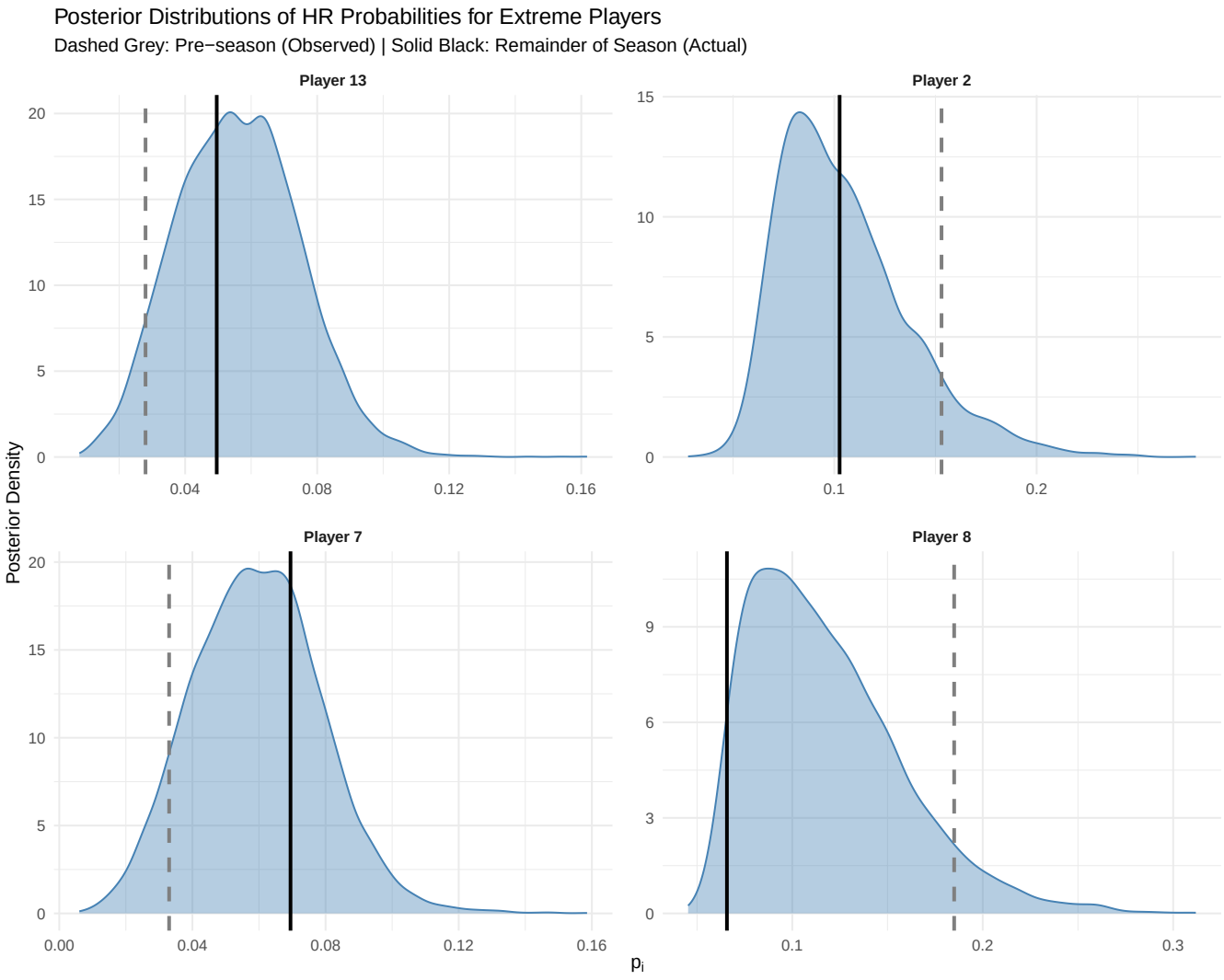
9.8.5.2 Method 4: Hierarchical Logistic Regression (logit-normal)

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.

$$\begin{aligned} y_i|p_i &\sim \text{Binomial}(n_i, p_i) \\ \text{logit}(p_i) &\sim N(\mu_0, \tau_0^2) \\ \mu_0 &\sim N(0, 10) \\ \tau_0 &\sim \text{Cauchy}(0, 2) \end{aligned} \tag{9.195}$$

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.



9.8.5.3 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:

$$L(p, \hat{p}) = \frac{|p - \hat{p}|}{\min(p, 1-p)} \tag{9.196}$$

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)
```

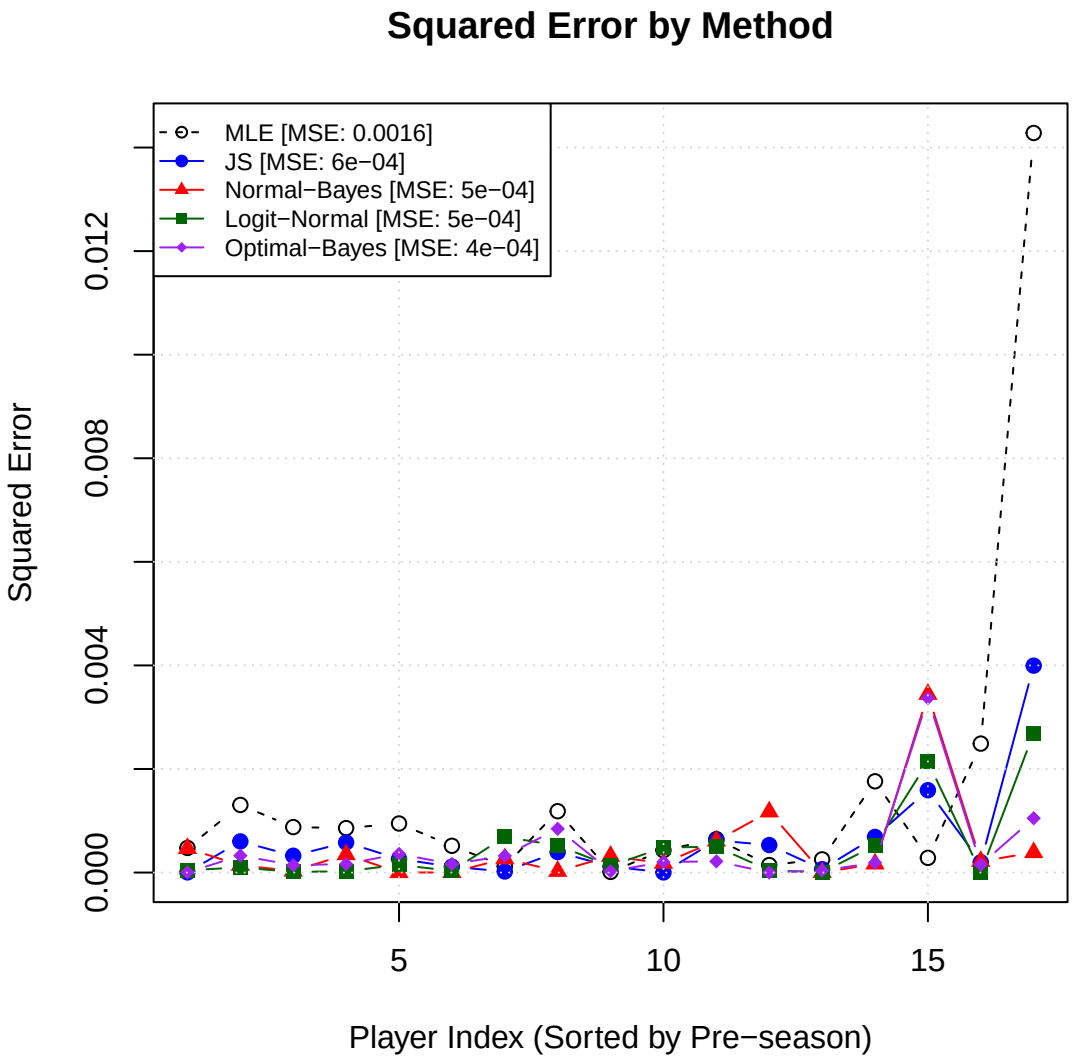
9.8.5.4 Comparison of All Five Estimates of p_i

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 9.6: 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:

$$\text{Metric}_i = \frac{|p_i^{\text{true}} - \hat{p}_i|}{\min(p_i^{\text{true}}, 1 - p_i^{\text{true}})} \tag{9.197}$$

9.9 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 9.3 (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

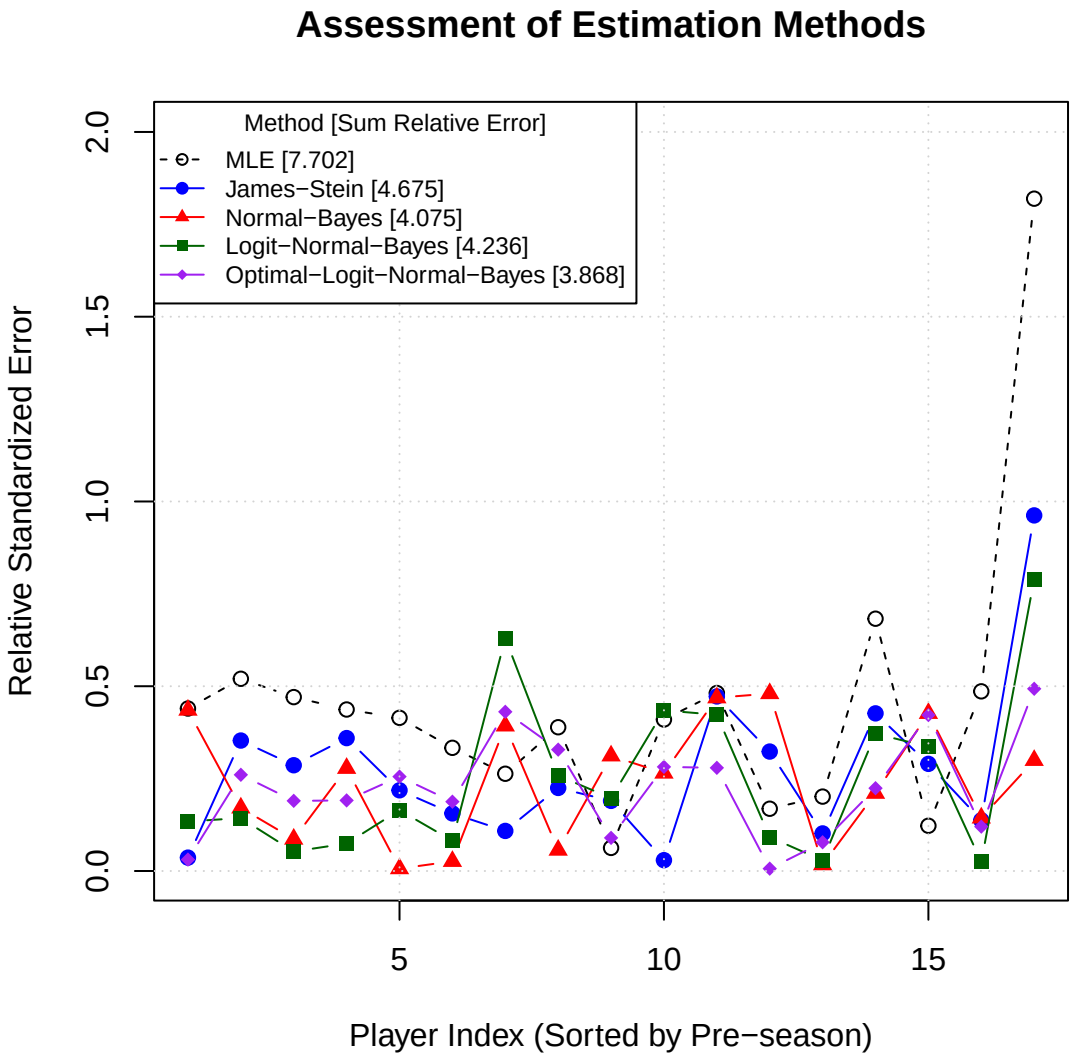


Figure 9.11: Relative Error Assessment: Five Methods

predictive density is obtained by marginalizing over the parameter θ:

$$f(y^*|y) = \int_\Theta f(y^*|\theta)\pi(\theta|y)\, d\theta \tag{9.198}$$

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 9.11 (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:

$$Y_i|\mu \sim N(\mu, \sigma^2) \tag{9.199}$$

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:

$$\begin{aligned} f(y^*|y) &= \int_{-\infty}^{\infty} f(y^*|\mu)\pi(\mu|y)\, d\mu \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^*-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\mu-\mu_n)^2}{2\sigma_n^2}}\, d\mu \end{aligned} \tag{9.200}$$

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

$$y^*|y \sim N(\mu_n, \sigma^2 + \sigma_n^2) \tag{9.201}$$

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

10 Appendices

10.1 A Short List of Contributors to Statistical Inference Based On Likelihood

1. **R.A. Fisher (1922, 1925)**
    - *Affiliation:* Rothamsted / University College London (UCL)
    - *Connection:* The "father" of the field; his work at UCL influenced the entire next generation, including Rao and Pearson.
    - **Contribution:** Fisher formally defined **Likelihood** as a function distinct from probability ($L(\theta; x)$ vs $f(x; \theta)$). He introduced the **Maximum Likelihood Estimator (MLE)**, derived the **Fisher Information** measure, and established the asymptotic efficiency of MLEs.
2. **J. Neyman & E.S. Pearson (1928, 1933)**
    - *Affiliations:* UC Berkeley (Neyman) / UCL (Pearson)
    - *Connection:* Egon Pearson was Karl Pearson's son (Fisher's rival). Neyman founded the Berkeley Statistics department, which later hosted Lehmann, Scheffé, and Le Cam.
    - **Contribution:** They introduced the **Likelihood Ratio** criterion (λ) as a general method for hypothesis testing. Their 1933 Lemma proved that for simple hypotheses, the Likelihood Ratio Test is the **Uniformly Most Powerful (UMP)** test, establishing the optimality of likelihood-based methods.
3. **S.S. Wilks (1938)**
    - *Affiliation:* Princeton University
    - *Connection:* A student of Henry Rietz, Wilks spent time at UCL with Pearson and Wishart before establishing the Princeton program.
    - **Contribution:** Wilks derived the asymptotic distribution of the Likelihood Ratio statistic for composite hypotheses. He proved that $-2 \ln \Lambda$ converges to a **Chi-square distribution** with degrees of freedom equal to the difference in the number of free parameters.
4. **M.S. Bartlett (1937)**
    - *Affiliation:* UCL / University of Manchester
    - *Connection:* A student of Fisher and colleague of Pearson at UCL; famously debated with Fisher on conditional inference.
    - **Contribution:** Bartlett improved the accuracy of the Likelihood Ratio Test for small samples. He introduced the **Bartlett Correction**, a scaling factor for the test statistic that aligns its expected value with that of the limiting Chi-square distribution.
5. **A. Wald (1943)**
    - *Affiliation:* Columbia University

- *Connection:* A student of Karl Menger in Vienna; at Columbia, he worked with Wolfowitz and influenced the decision-theoretic approach adopted by Stein and Karlin.
- **Contribution:** Wald developed the **Wald Test**, which tests hypotheses based on the distance between the unrestricted MLE and the hypothesized value. He demonstrated the asymptotic equivalence of the Likelihood Ratio, Wald, and Score tests for standard models.

6. **H. Cramér (1946)**

   - *Affiliation:* Stockholm University
   - *Connection:* Hosted Rao in Stockholm; his textbook synthesized the British (Fisher/Pearson) and American (Wilks/Wald) schools into one rigorous framework.
   - **Contribution:** In his influential text *Mathematical Methods of Statistics*, Cramér provided rigorous proofs for the consistency and asymptotic normality of the MLE. Independently of Rao, he established the **Cramér-Rao Lower Bound**.

7. **C.R. Rao (1948)**

   - *Affiliation:* Indian Statistical Institute (ISI)
   - *Connection:* A PhD student of Fisher at Cambridge; he unified the testing theories of Fisher, Neyman, and Wald.
   - **Contribution:** Rao introduced the **Score Test** (or Lagrange Multiplier Test), which allows for hypothesis testing using only the restricted MLE (under the null hypothesis). He also independently derived the lower bound for estimator variance.

8. **E.L. Lehmann & H. Scheffé (1950, 1955)**

   - *Affiliation:* UC Berkeley
   - *Connection:* Lehmann was Neyman's student at Berkeley; Scheffé was a colleague. They solidified the "Berkeley School" of optimality.
   - **Contribution:** Building on the concept of **Sufficiency**, they established the **Lehmann-Scheffé Theorem**. This result connects complete sufficient statistics to **Uniformly Minimum Variance Unbiased Estimators (UMVUE)**.

9. **S. Karlin & H. Rubin (1956)**

   - *Affiliation:* Stanford University
   - *Connection:* Part of the "Stanford School" (along with Stein) that focused on decision theory, heavily influenced by Wald's earlier work.
   - **Contribution:** They formalized the property of **Monotone Likelihood Ratio (MLR)** for families of distributions. The **Karlin-Rubin Theorem** extended the Neyman-Pearson Lemma to one-sided composite hypotheses.

10. **W. James & C. Stein (1961)**

    - *Affiliation:* Stanford University
    - *Connection:* Stein was a student of Wald at Columbia before moving to Stanford.
    - **Contribution:** They discovered **Stein's Paradox**, showing that the MLE (sample mean) is **inadmissible** for estimating the mean of a multivariate Normal distribution ($p \geq 3$). They introduced the **James-Stein Estimator**, a shrinkage estimator that dominates the MLE.

11. **L. Le Cam (1960s)**

- *Affiliation:* UC Berkeley
- *Connection:* A student of Neyman at Berkeley; he took the asymptotic torch from Wald and Wilks and placed it on a rigorous topological foundation.
- **Contribution:** Le Cam modernized asymptotic theory by introducing **Local Asymptotic Normality (LAN)**. He showed that under general conditions, the likelihood ratio of a complex experiment behaves asymptotically like that of a Normal shift experiment.