Statistical Inference

Longhai Li

2026-01-15

Preface

This is a concise course about statistical inference.

Key Features

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

Audience

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

1 Introduction to Statistical Inference

1.1 Population Model (Data Model)

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

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

Assumptions and Scope

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

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

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

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

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

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

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

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

1.2 Probabilistic Model vs. Statistical Inference

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

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

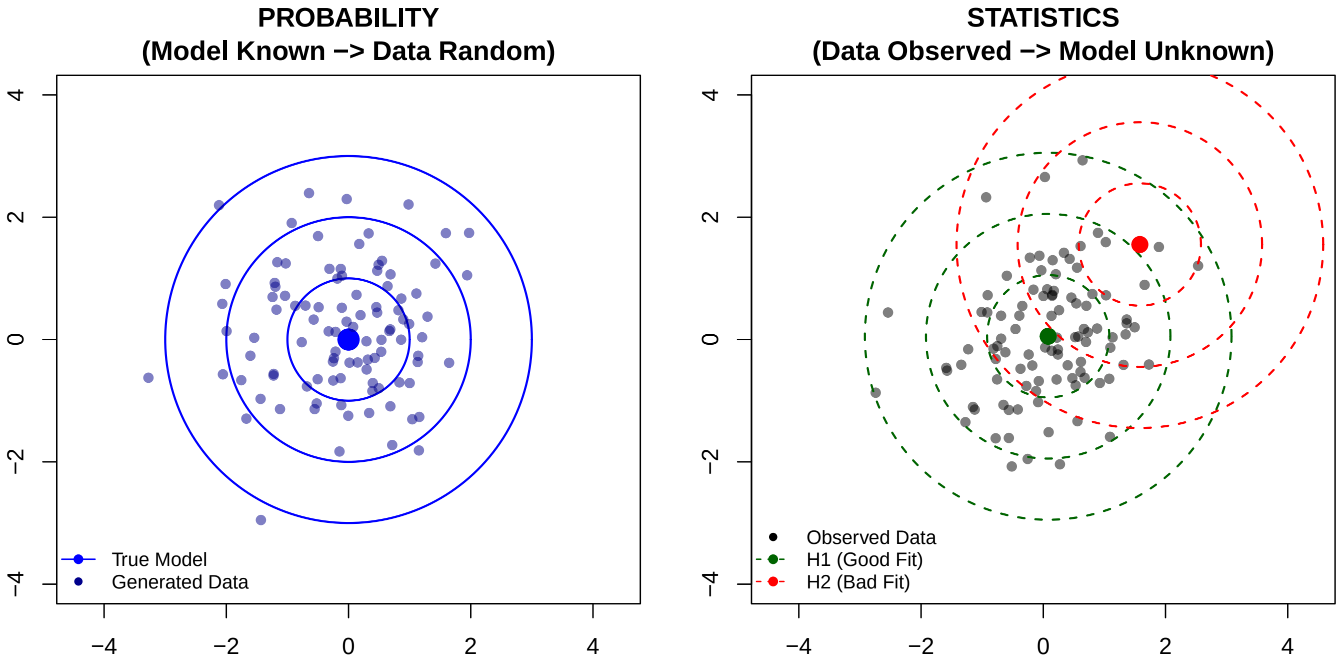

Figure 1.1: Probability vs Statistics. Left: Probability—The model is fixed (Blue center/contours), generating random data. Right: Statistics—Data is fixed (Black points); we test two hypothesized models: H1 (Green) centered at the sample mean (Good Fit) and H2 (Red) shifted by (1.5, 1.5) (Bad Fit).

1.3 A Motivating Example: The Lady Tasting Tea

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

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

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

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

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

1.3.1 Small Sample (n=10)

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

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

1.3.2 Large Sample (n=40)

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

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

1.4 Questions to Answer in Statistical Inference

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

Point Estimation

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

- *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

1.5.1.1 n=40 (k=28)

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

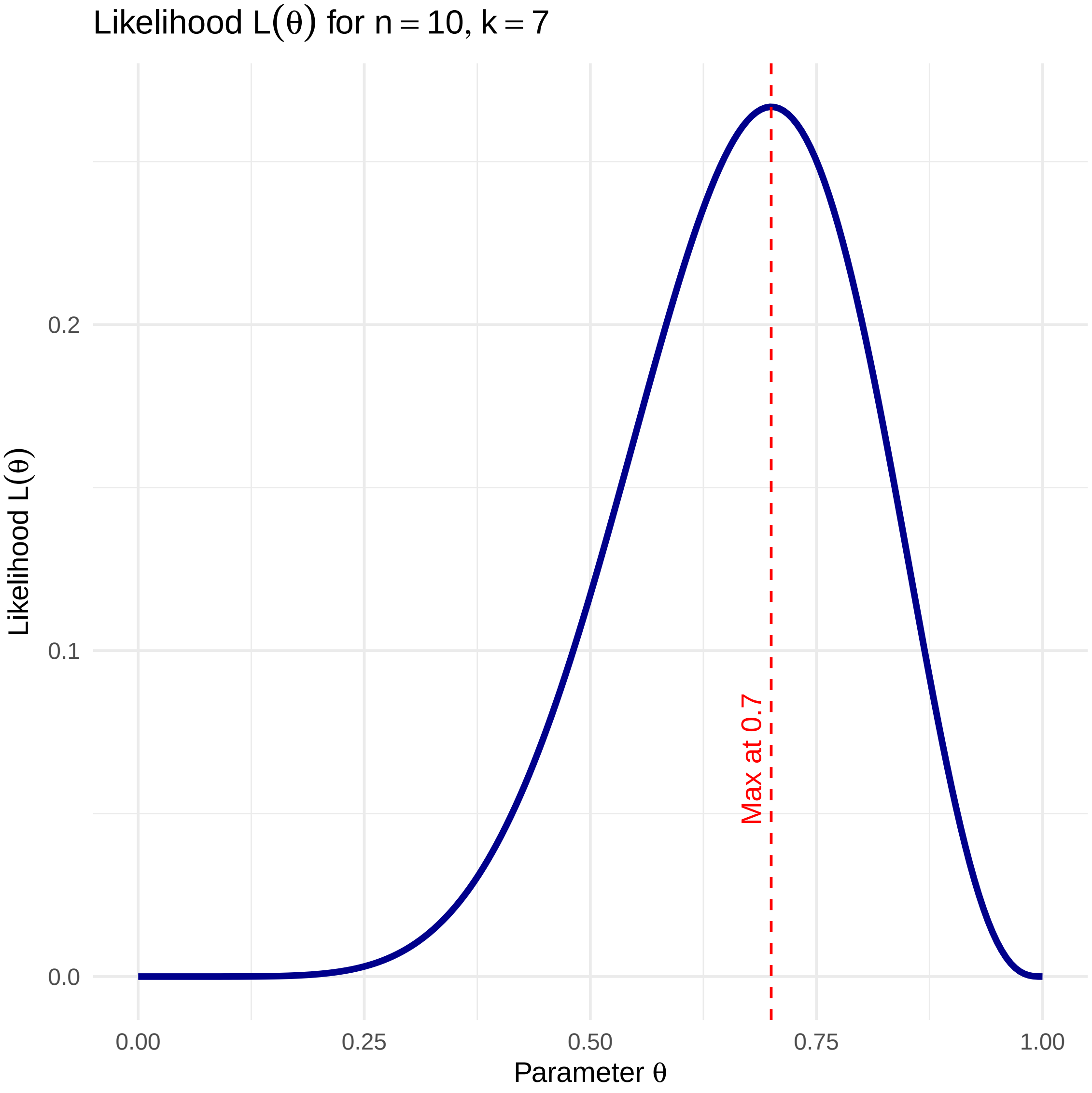

Figure 1.2: Likelihood Function (n= 10)

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

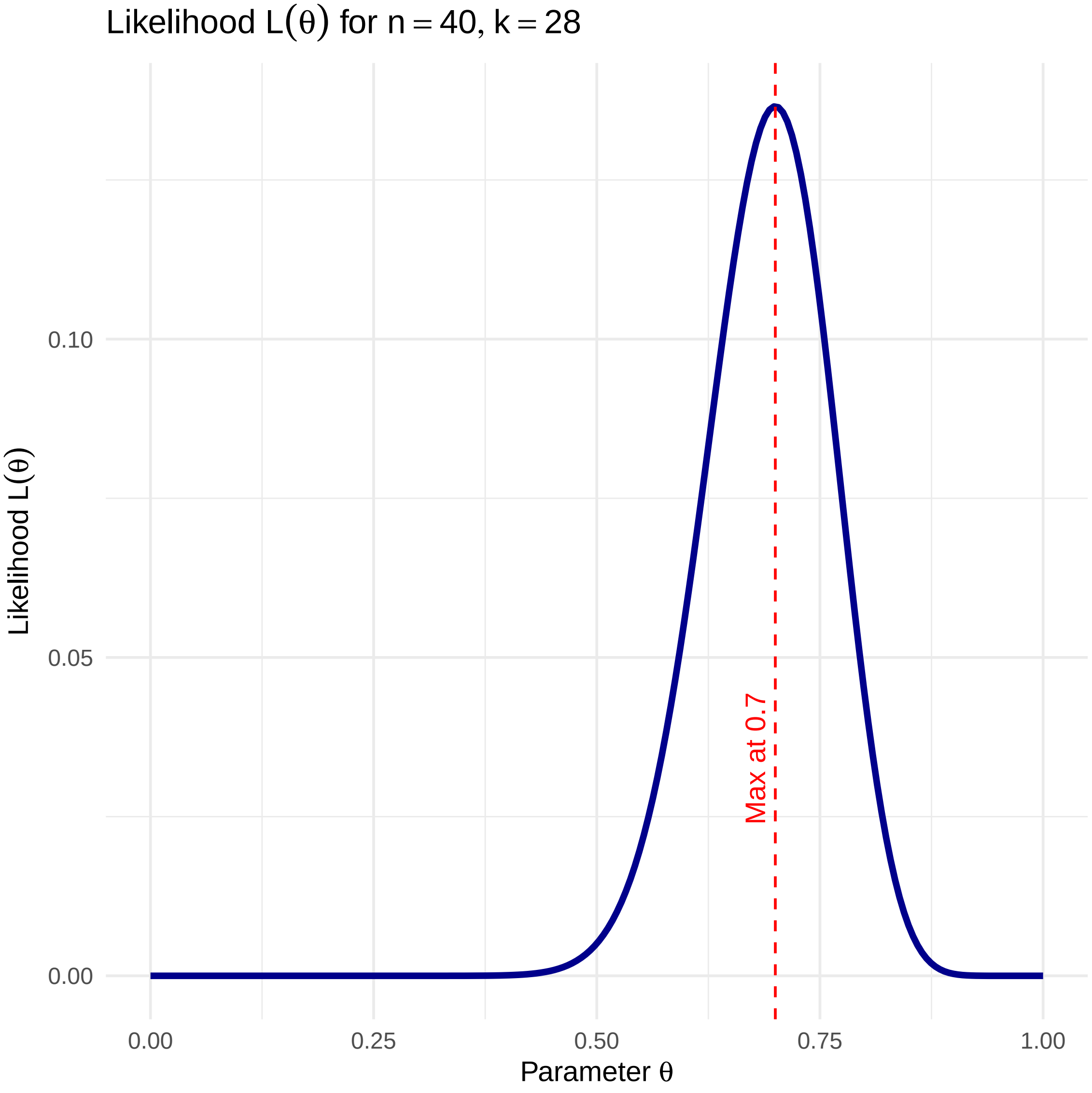

Figure 1.3: Likelihood Function (n= 40)

Questions

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

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

1.6 Frequentist Inference

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

Example: Frequentist Test of Lady Tasting Tea

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

1.6.1 n=10 (k=7)

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

1.6.2 n=40 (k=28)

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

1.6.3 Questions to Answer

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

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

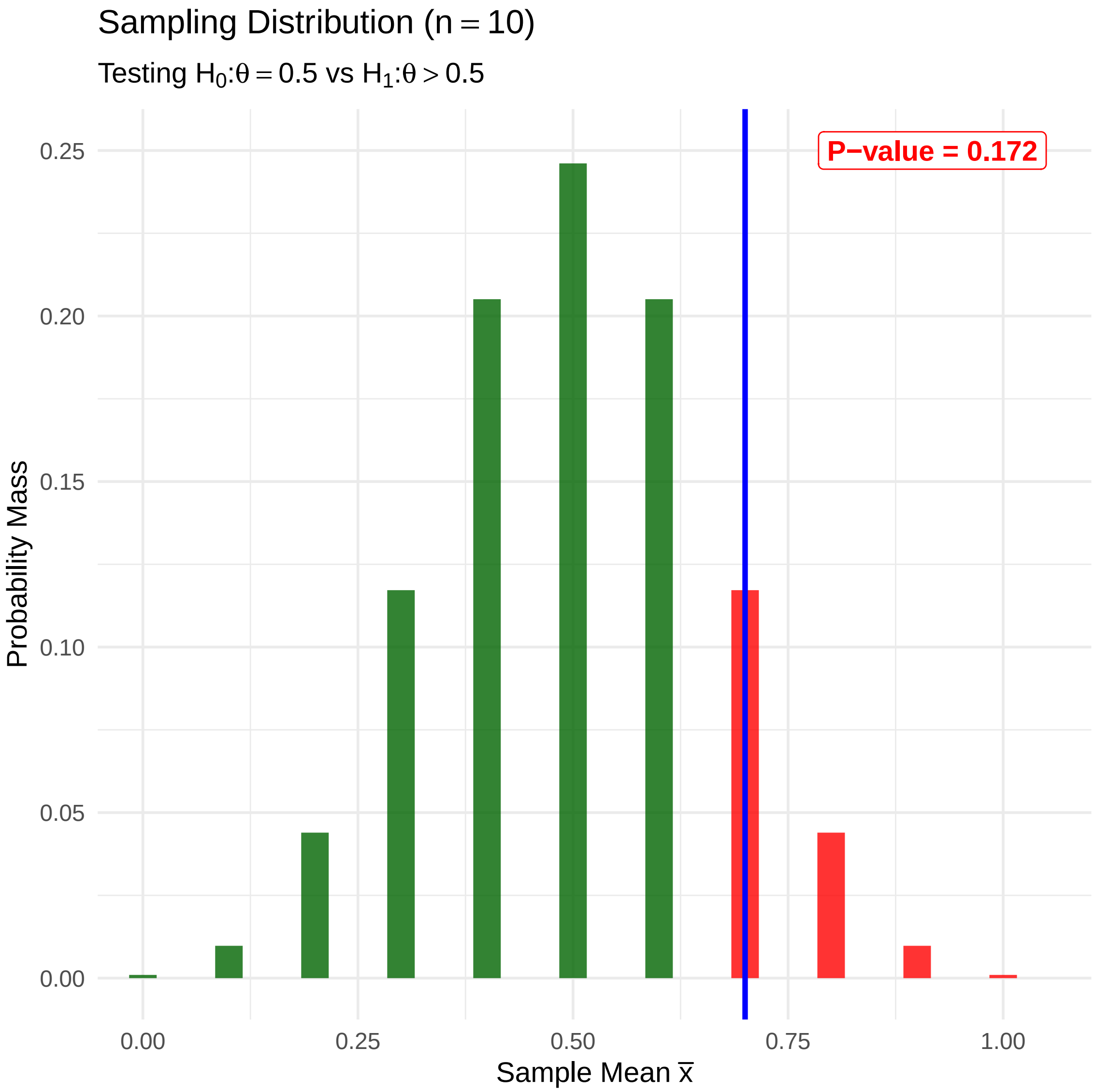

Figure 1.4: Sampling Distribution (n= 10)

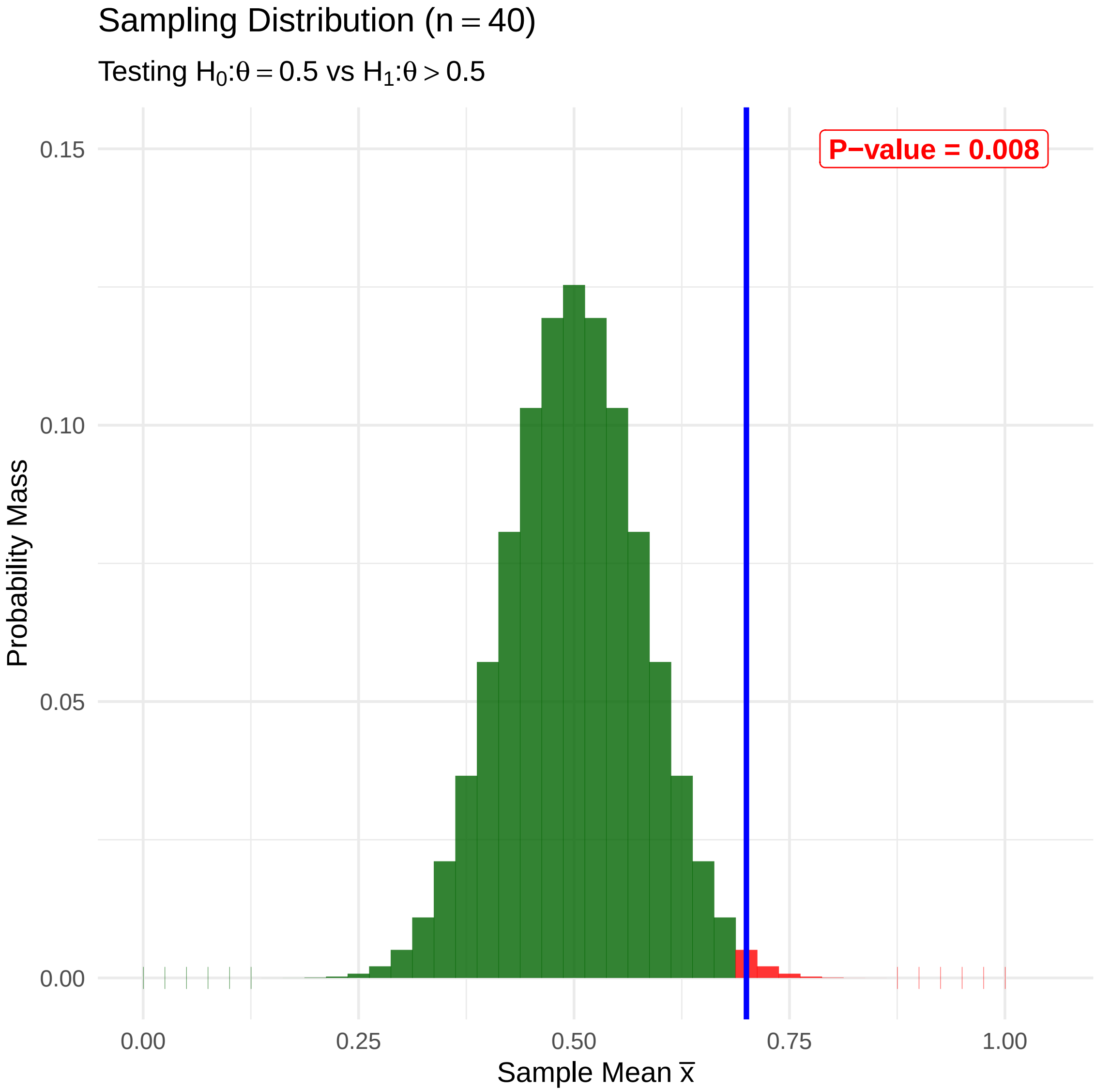

Figure 1.5: Sampling Distribution (n= 40)

1.7 Bayesian Inference

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

Example: Bayesian Analysis of the Lady Tasting Tea

Prior: $\text{Beta}(1, 1)$ (Uniform).

1.7.1 n=10 (k=7)

Posterior: $\text{Beta}(1 + 7, 1 + 3) = \text{Beta}(8, 4)$

1.7.2 n=40 (k=28)

Posterior: $\text{Beta}(1 + 28, 1 + 12) = \text{Beta}(29, 13)$.

1.7.3 Questions to Answer

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

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

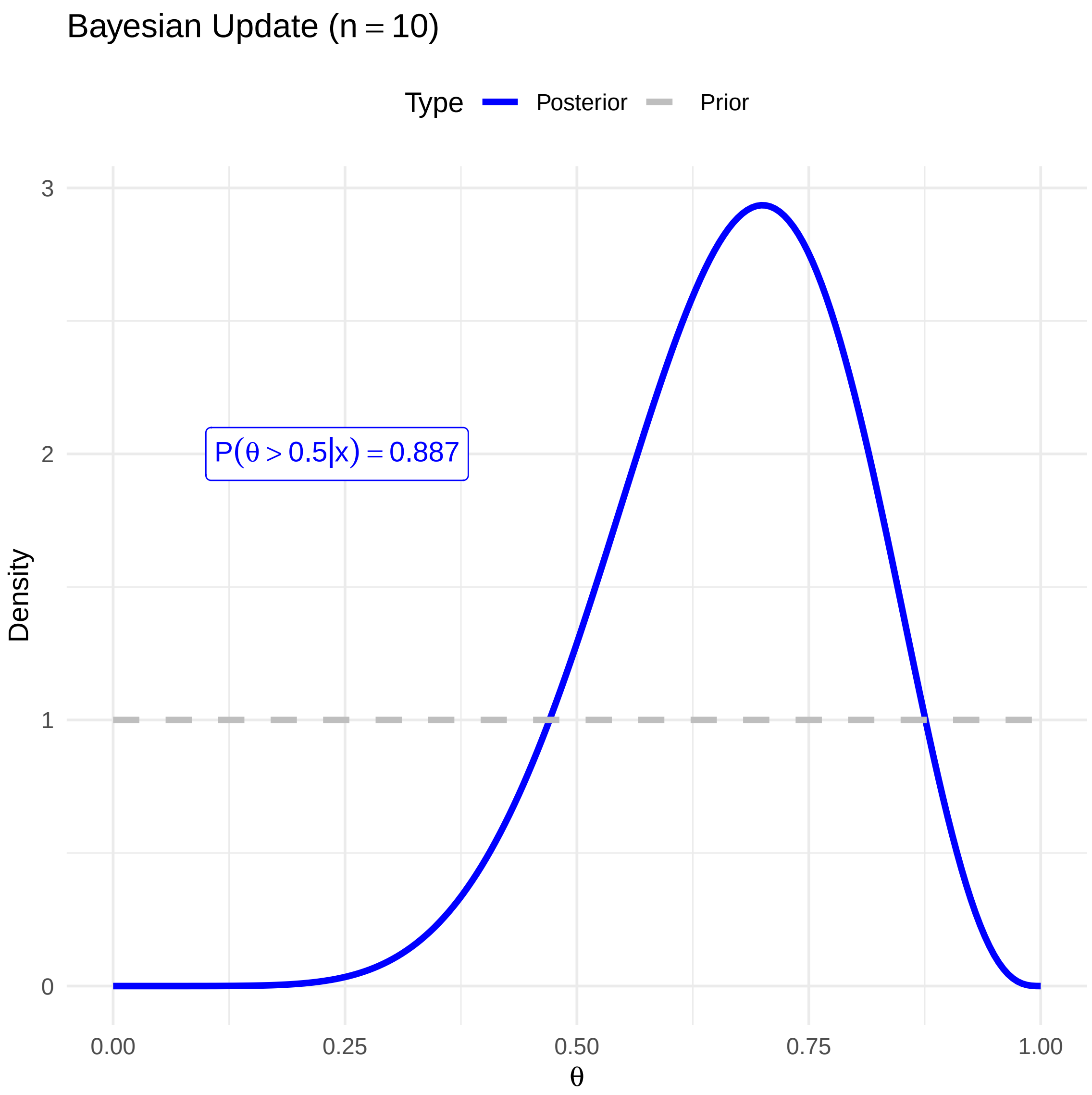

Figure 1.6: Bayesian Update (n= 10)

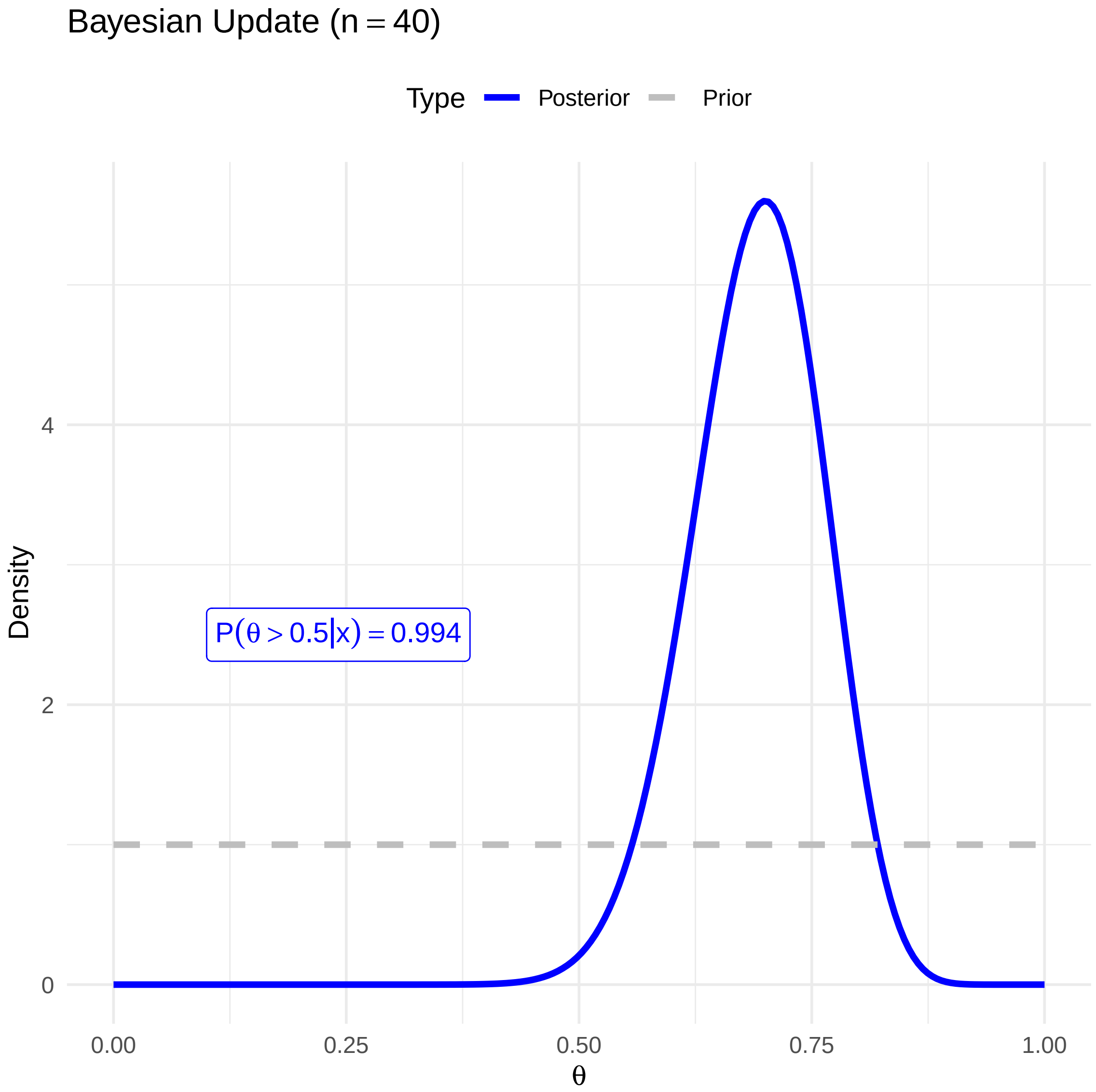

Figure 1.7: Bayesian Update (n= 40)

2 Decision Theory

2.1 Formulation of Decision Theory

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

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

2.2 Decision Rules and Risk Functions

2.2.1 Decision Rule

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

2.2.2 Risk Function

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

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

2.3 Examples of Decision Problems

2.3.1 Example 1: Hypothesis Testing

We want to test H_0 vs H_1.

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

2.3.2 Example 2: Point Estimation

We want to estimate a parameter θ.

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

2.3.3 Example 3: Interval Estimation

We want to estimate a range for the parameter.

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

2.3.4 Example 4: The Duchess and the Emerald Necklace

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

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

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

2.3.4.1 Formulation

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

2.3.4.2 Risk Calculation for Deterministic Rules

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

Rule d_1 (Always Left)

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

Rule d_2 (Always Right)

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

Rule d_3 (Follow Aunt)

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

Rule d_4 (Do Opposite)

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

2.4 Principles for Choosing a Decision Rule

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