Statistical decisions and risks

M1 MAS

Pierre Pudlo
Aix-Marseille Université / Faculté des Sciences

Statistical Models and Parameters

The Random Dataset

Let D be the **random dataset** - a random object taking values in the data space:

- **Single number** (e.g., number of satisfied students): $D \in \mathbb{R}$
- **Vector** (e.g., gene expression in n individuals): $D \in \mathbb{R}^n$
- **Matrix** (e.g., p variables on n individuals): $D \in \mathbb{R}^{n \times p}$

In all cases: D is a random variable with distribution P_θ

Statistical Model Components

We set a **statistical model** on D:

- **Model parameter:** θ (unknown)
- **Parameter of interest:** $\beta = \eta(\theta)$ (deterministic function of θ)
- **Goal:** Make decisions about β using estimator $S = s(D)$

> **Important**
>
> Key insight: β and θ can be different!

Examples of Statistical Models

1. Binomial Model

- Data: $D \in \{0, 1, \ldots, n\}, D \sim \mathcal{B}(n, \theta)$
- Parameter: $\theta \in [0, 1]$ (probability of success)
- Interest: $\beta = \theta$

2. Gaussian Sample

- Data: $D \in \mathbb{R}^n, D \sim \mathcal{N}(\mu, \sigma^2)^{\otimes n}$
- Parameters: $\theta = (\mu, \sigma^2)$ where $\mu \in \mathbb{R}, \sigma^2 > 0$

3. Mixture Model

- Data: $D \sim \left(\alpha\mathcal{N}(\mu_1, \sigma_1^2) + (1-\alpha)\mathcal{N}(\mu_2, \sigma_2^2)\right)^{\otimes n}$
- Parameters: $\theta = (\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$

Understanding the Examples

Examples where $\theta \neq \beta$

Gaussian case: $\theta = (\mu, \sigma^2)$ but $\beta = \mu$

$$\eta((\mu, \sigma^2)) = \mu$$

Signal-to-noise ratio: $\beta = \mu/\sigma$

$$\eta((\mu, \sigma^2)) = \frac{\mu}{\sigma}$$

Mixture component: $\beta = \alpha$ (mixing probability)

$$\eta((\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)) = \alpha$$

Special Cases: Important Interpretations

When $\beta \in \{0, 1\}$: We're doing **hypothesis testing**!

- $\beta = 0$: Null hypothesis H_0
- $\beta = 1$: Alternative hypothesis H_1

Mixture model interpretation:

- Population 1: probability α, mean μ_1, variance σ_1^2
- Population 2: probability $1 - \alpha$, mean μ_2, variance σ_2^2
- Each data point comes from one of the two populations

The Decision Problem

Statistical Decision Framework:

1. **Observe:** Dataset realization d
2. **Estimate:** Compute $s_{\text{obs}} = s(d)$
3. **Decide:** Make conclusion about β

Central Question

What makes an estimator $S = s(D)$ “good”?

Part 1: Frequentist Approach

Frequentist Philosophy

Core principle: A good estimator is **close** to the parameter of interest.

Problem: We can't evaluate $L(s_{\text{obs}}, \beta)$ since β is unknown.

Solution: Study $L(S, \beta)$ as a random variable.

Risk Function

$$R_\theta(S) = \mathbb{E}_\theta[L(S, \beta)]$$

Expected loss under model P_θ

Loss Functions and Risks

Loss function: $L(\beta, s)$ measures cost of estimating β by s

- Non-negative: $L(\beta, s) \geq 0$
- Minimized when correct: $L(\beta, \beta) = 0$

Common choices:

- **Squared error:** $L(s, \beta) = (s - \beta)^2$ → Risk = **MSE**
- **Absolute error:** $L(s, \beta) = |s - \beta|$ → Risk = **MAE**
- **0-1 loss:** $L(s, \beta) = \mathbf{1}\{s \neq \beta\}$ → Risk = **Error probability**

MSE Derivation: Binomial Case

Setup: $D \sim \mathcal{B}(n, \theta)$, $\beta = \theta$, estimator $S_{0,0} = D/n$

Step-by-step calculation

Given: $\mathbb{E}_\theta[D] = n\theta$ and $\mathrm{Var}_\theta[D] = n\theta(1-\theta)$

$$\begin{aligned}
\mathrm{MSE} &= \mathbb{E}_\theta\left[\left(\frac{D}{n} - \theta\right)^2\right] \\
&= \mathrm{Var}_\theta\left[\frac{D}{n}\right] + \left(\mathbb{E}_\theta\left[\frac{D}{n}\right] - \theta\right)^2 \\
&= \frac{\mathrm{Var}_\theta[D]}{n^2} + (0)^2 \quad \text{(unbiased)} \\
&= \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}
\end{aligned}$$

Generalized Binomial Estimator

Estimator: $S_{\alpha,\beta} = \frac{D+\alpha}{n+\alpha+\beta}$ where $\alpha, \beta \geq 0$

Complete MSE derivation

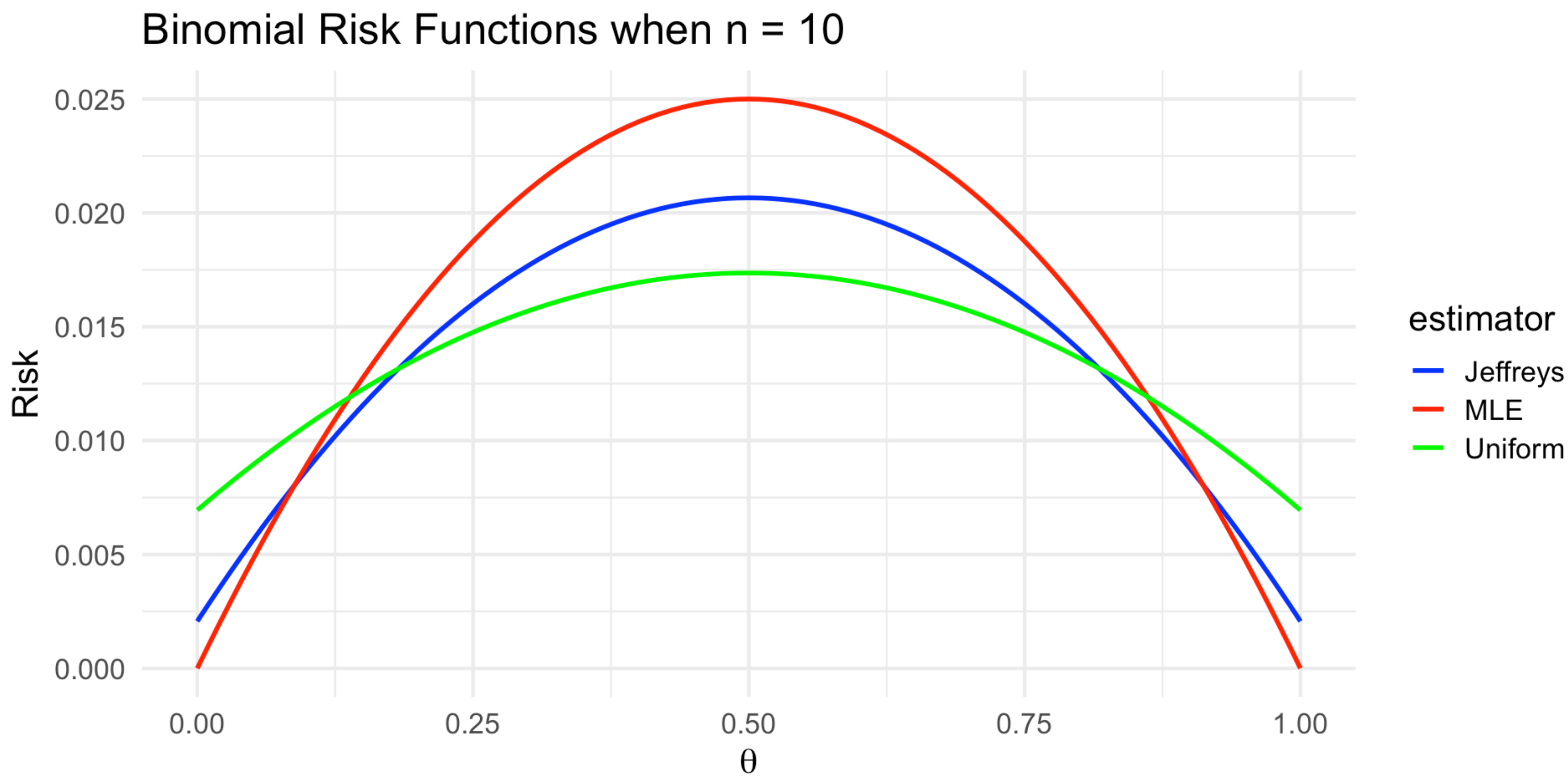
Bias: $\mathbb{E}_\theta[S_{\alpha,\beta}] - \theta = \dfrac{n\theta + \alpha}{n + \alpha + \beta} - \theta = \dfrac{\alpha(1-\theta) - \theta\beta}{n + \alpha + \beta}$

Variance: $\mathrm{Var}_\theta[S_{\alpha,\beta}] = \dfrac{n\theta(1-\theta)}{(n+\alpha+\beta)^2}$

MSE = Bias^2 + Variance:

$$\mathrm{MSE} = \frac{(\alpha(1-\theta) - \theta\beta)^2 + n\theta(1-\theta)}{(n+\alpha+\beta)^2}$$

Risk Function Visualization



Observation: Different estimators dominate in different regions! Which one would you choose?

Gaussian Sample MSE

Setup: $D \sim \mathcal{N}(\mu, \sigma^2)^{\otimes n}$, $\beta = \mu$, estimator $S_{\kappa,m} = \dfrac{\kappa m + n\bar{D}}{\kappa + n}$

Complete derivation

Sample mean: $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \sim \mathcal{N}(\mu, \sigma^2/n)$

Expected value: $\mathbb{E}_\mu[S_{\kappa,m}] = \dfrac{\kappa m + n\mu}{\kappa + n}$

Bias: $\dfrac{\kappa m + n\mu}{\kappa + n} - \mu = \dfrac{\kappa(m - \mu)}{\kappa + n}$

Variance: $\mathrm{Var}_\mu[S_{\kappa,m}] = \dfrac{n^2\sigma^2/n}{(\kappa + n)^2} = \dfrac{n\sigma^2}{(\kappa + n)^2}$

Gaussian MSE: Final Result

MSE formula

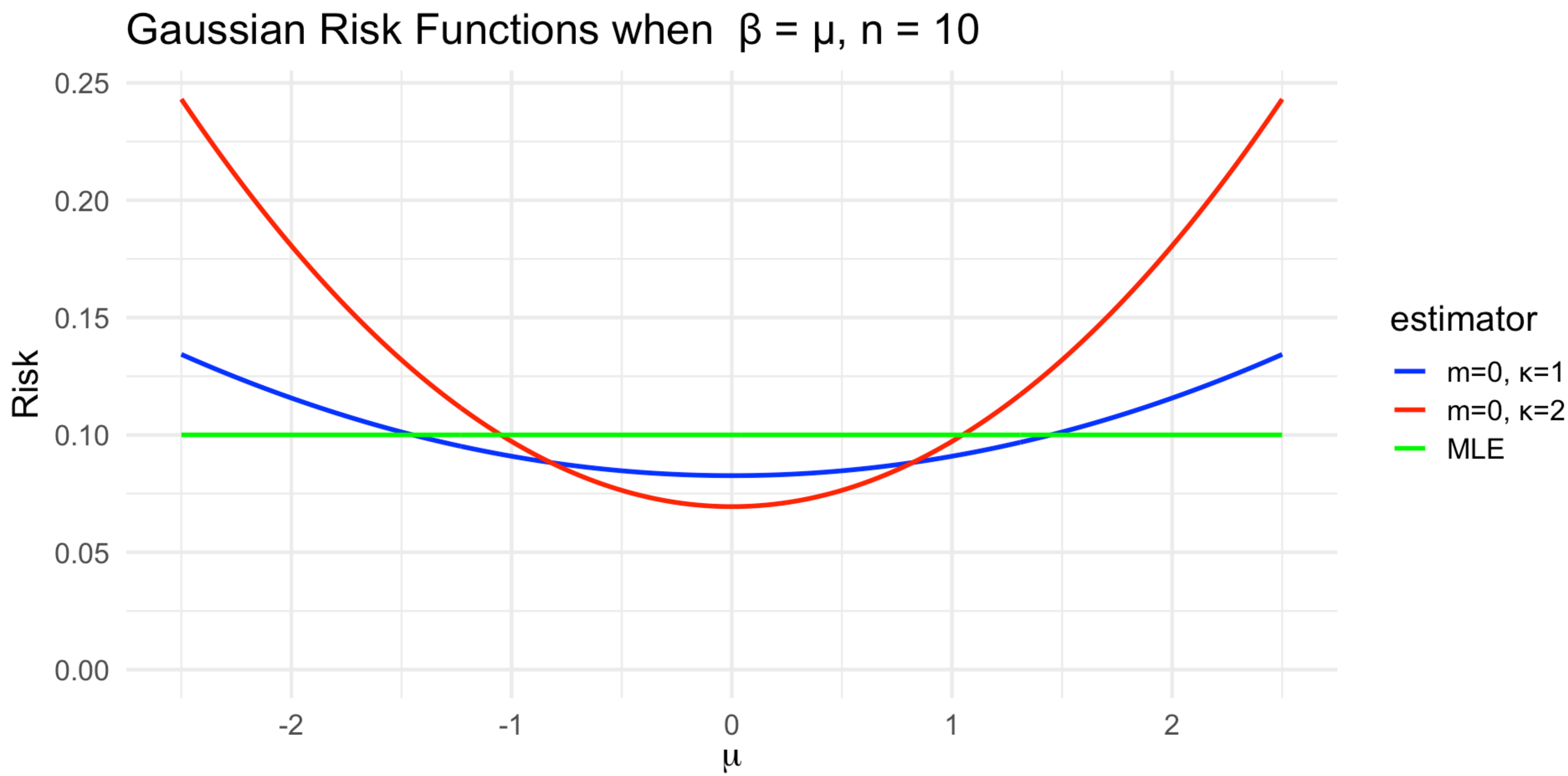
$$\text{MSE} = \frac{\kappa^2(m-\mu)^2 + n\sigma^2}{(\kappa+n)^2}$$

Key insights:

- As $\kappa \to 0$: $S_{\kappa,m} \to \bar{D}$ (sample mean)
- As $\kappa \to \infty$: $S_{\kappa,m} \to m$ (prior mean)
- **Bias-variance trade-off:** κ controls the balance

Question: What happens when $\sigma^2 = 1, 4$, or 10?
Answer: Larger σ^2 makes sample mean less reliable, so optimal κ increases.

Risk Function Visualization (2)



Observation: Different estimators dominate in different regions!
Which one would you choose?

Admissibility Theory

Definition: Admissible Estimator

$S^\star$ is **admissible** if there is no other estimator S such that:

- $R_\theta(S) \leq R_\theta(S^\star)$ for all θ
- $R_\theta(S) < R_\theta(S^\star)$ for at least one θ

Graphical interpretation: Risk curve is not **uniformly better**

- The above definition depends on the loss function L (through the risk $R_\theta(S)$)
- If $S^\star$ is admissible, and has better risk than S at $\theta = \theta_0$, then S must have worse risk than $S^\star$ at some other θ_1
- It is difficult to prove admissibility in practice, given a formula for $S^\star$ and its risk function

Admissibility Example

Conclusion: S_1 is **not admissible** because
S_2 is uniformly better! (with a difference around 0)

Minimax Theory

Definition: Minimax Estimator

$S^\star$ is **minimax** if, for any other estimator S, $\sup_\theta R_\theta(S^\star) \leq \sup_\theta R_\theta(S)$

- The **worst case** occurs at different θ-values for different estimators.
- We care only about the worse-case scenario of an estimator: this is a **pessimistic** viewpoint.
- Minimax = **Best worst-case performance**
- **Game theory connection:** Statistician (choose S) vs. Nature (choose θ)

Binomial Maximum Risk

Question: Compute $\sup_\theta R_\theta(S)$ for our estimators in the binomial case

Binomial case: For $S_{\alpha,\beta}$ with squared loss: $\sup_\theta \dfrac{(\alpha(1-\theta)-\theta\beta)^2+n\theta(1-\theta)}{(n+\alpha+\beta)^2}$

Solution approach

Let $f(\theta) = (\alpha(1-\theta)-\theta\beta)^2 + n\theta(1-\theta)$

Taking derivative: $f'(\theta) = -2(\alpha(1-\theta)-\theta\beta)(\alpha+\beta) + n(1-2\theta)$

Setting to zero and solving gives critical point at:

$$\theta^* = \frac{n+2\alpha(\alpha+\beta)}{2n+2(\alpha+\beta)^2}$$

The maximum risk occurs at this critical point or boundary.

Gaussian Maximum Risk

Gaussian case: For $S_{\kappa,m}$, if $\kappa > 0$,

$$\sup_{\mu} \frac{\kappa^2(m-\mu)^2 + n\sigma^2}{(\kappa+n)^2} = \infty$$

Key insight

- If $\kappa = 0$ (estimator=sample mean): Risk = σ^2/n (constant!)
- If $\kappa > 0$: Risk → ∞ as $|\mu - m| \to \infty$

Conclusion: Sample mean is minimax

Minimax Results

Key Minimax Estimators

Gaussian: When the parameter of interest is μ, $\bar{D}$ is minimax

General principle: Constant risk often indicates minimax property, see later.

Part 2: Bayesian Approach

Bayesian Philosophy

Fundamental difference: Uncertainty on θ is modeled as random variable θ_{rv}.
Let $\pi(\theta)$ denote the density of θ_{rv}.

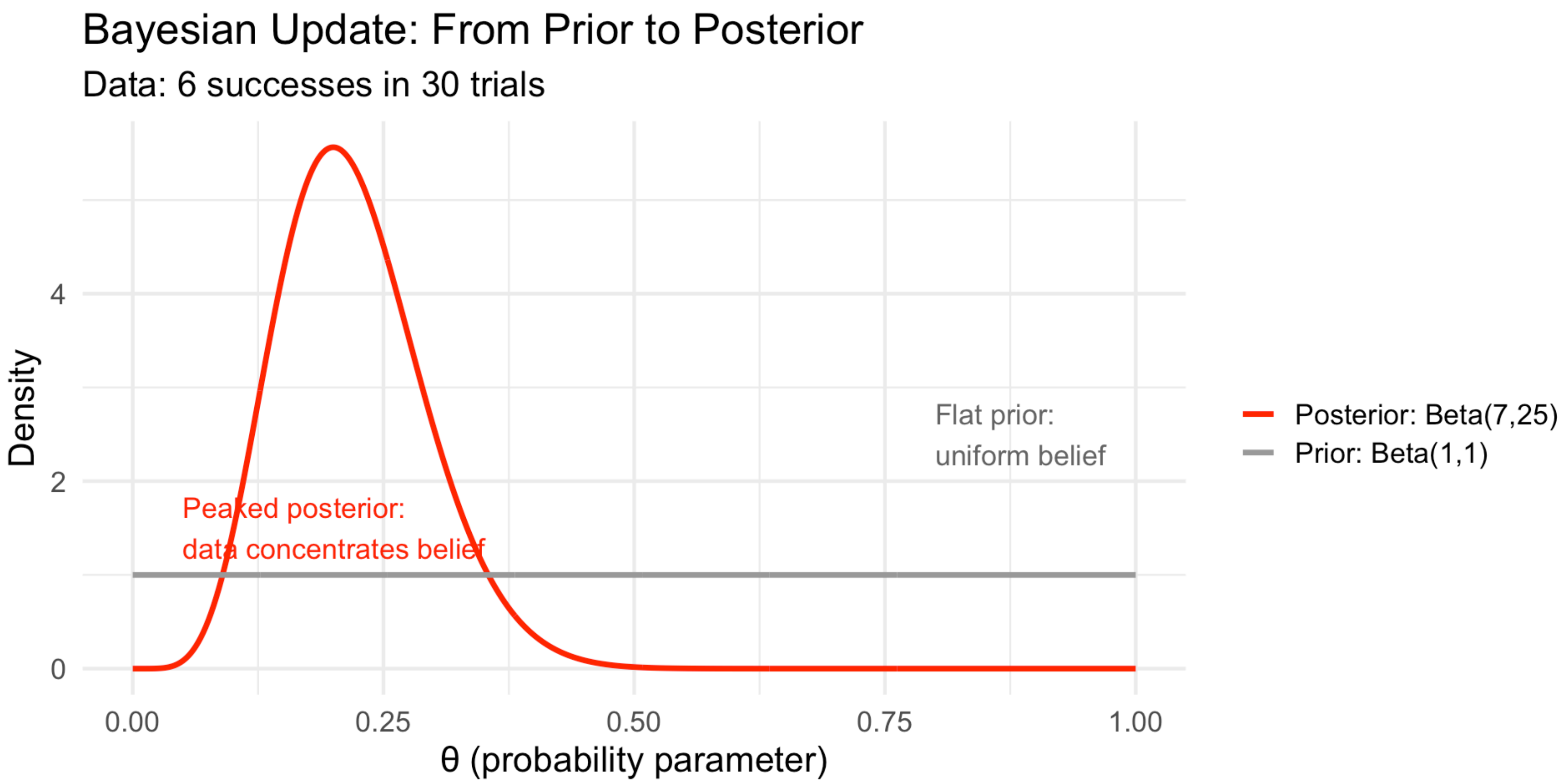
Interpretation: $\mathbb{P}_\theta(\cdot) = \mathbb{P}(\cdot|\theta_{rv} = \theta)$
I.e., $\mathbb{P}_\theta(\cdot)$ is the distribution of the data given $\theta_{rv} = \theta$.

- **Prior knowledge:** θ_{rv} has density $\pi(\theta)$ before seeing data
  $\pi(\theta)$ represents the weight of our **beliefs** about the value θ **before** seeing data
- **Posterior knowledge:** θ_{rv} has density $\pi(\theta|d)$ after seeing data $D = d$
  $\pi(\theta|d)$ represents the weight of our **beliefs** about the value θ **after** seeing data $D = d$
- The observed data are used to **update** our **beliefs** regarding the possible values of θ.

Remark

In this framework, the “true” value of the parameter is θ_{rv}

Prior vs Posterior Example



Key insight:

Observing $D = d$ transforms our **uniform ignorance** into **concentrated knowledge**!

Understanding Bayes' Theorem

Bayes' theorem:

$$\pi(\theta|d) = \frac{p(d|\theta)\pi(\theta)}{p(d)}$$

where $p(d) = \int p(d|\theta)\pi(\theta)d\theta$ is the **evidence**

- $\pi(\theta)$: **prior** (what we believed before)
- $p(d|\theta)$: **likelihood** (how well θ explains data d)
- $\pi(\theta|d)$: **posterior** (updated beliefs after seeing data d)
- $p(d)$: **evidence** (normalizing constant)

Simulation Algorithm

Question: How to understand the posterior distribution?

Look at the algorithm for binomial case with uniform prior

0. Set $N = 10000$
1. Draw many θ_i's ($i = 1, \ldots, N$) uniformly in $[0, 1]$ *(prior)*
2. For each θ_i, draw one $d_i \sim \mathrm{Binomial}(n, \theta_i)$ *(likelihood)*
3. **Return:** the set of θ_i's for which $d_i = d$ *(observed data)*

- **Question:** What is the distribution of returned θ_i's?
- **Answer:** They follow the posterior $\pi(\theta|d) \propto p(d|\theta)\pi(\theta)$

Example: Binomial, $d = 6, n = 30$

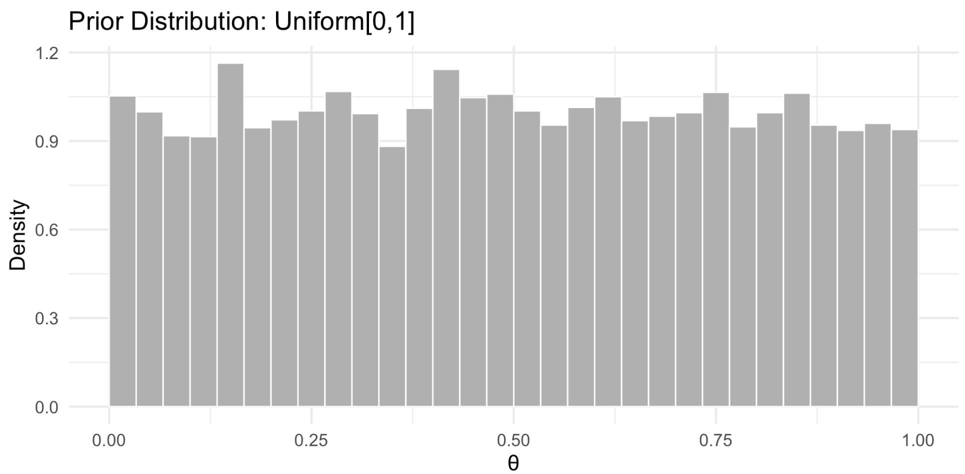
theta	d_sim
0.288	7
0.788	25
0.409	15
0.883	27
0.940	30
0.046	1
0.528	12
0.892	25
0.551	14
0.457	8
0.957	28

theta	d_sim
0.453	16
0.678	19
0.573	20
0.103	3
0.900	27
0.246	6
0.042	1
0.328	9
0.955	27
0.890	28
0.693	22

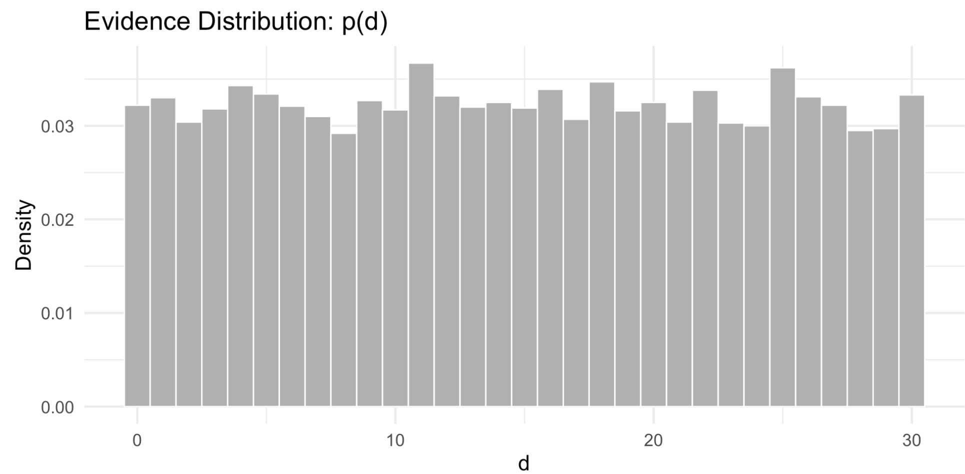
theta	d_sim
0.641	16
0.994	30
0.656	18
0.709	22
0.544	17
0.594	18
0.289	6
0.147	3
0.963	29
0.902	26
0.691	23

Example (continued)

The distribution of the `theta`-column in $\pi(\theta)$, the uniform prior

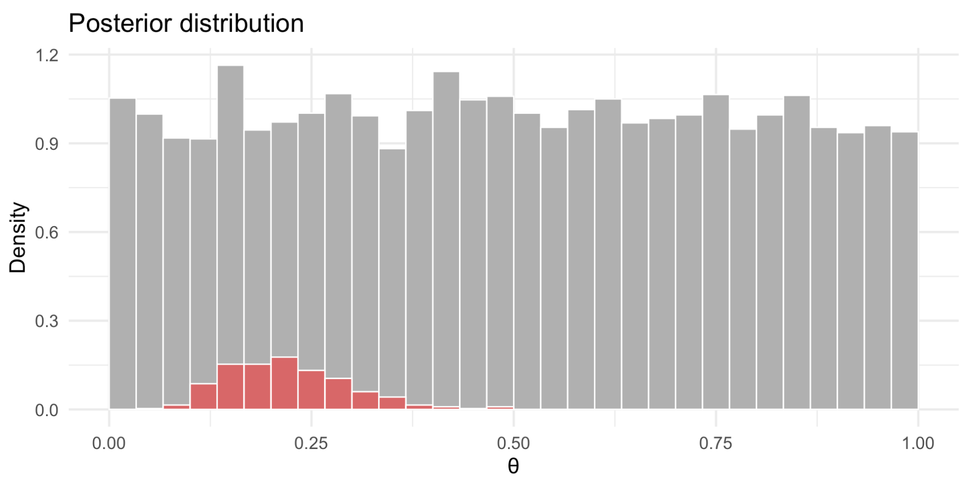


The distribution of the `d_sim`-column in $p(d)$, the evidence

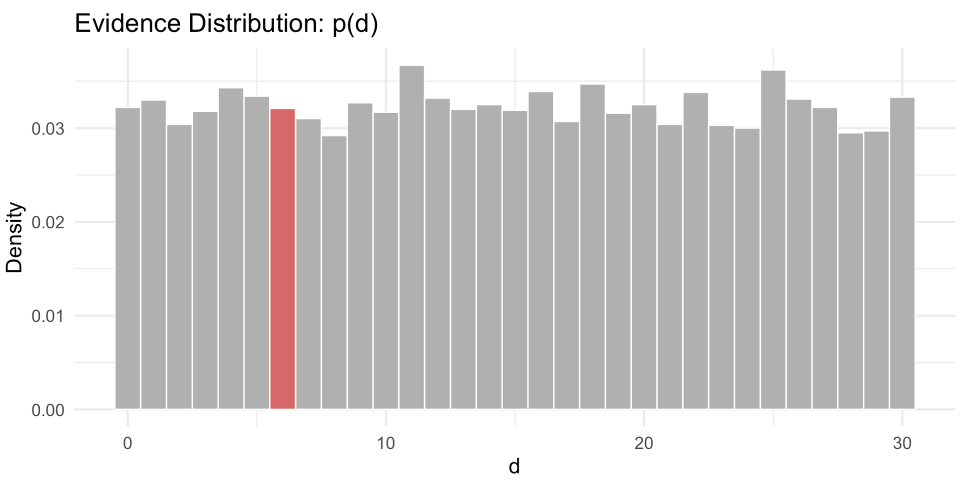


Example (continued)

Kept values of θ in red…



… when $d = 6$



Gaussian Posterior Derivation

Setup: $D \sim \mathcal{N}(\mu, \sigma^2)^{\otimes n}$ with σ^2 known, prior $\mu \sim \mathcal{N}(m, \kappa^{-2})$

Step-by-step derivation

Likelihood: $p(d|\mu) \propto \exp\left(-\frac{n(\bar{d}-\mu)^2}{2\sigma^2}\right)$

Prior: $\pi(\mu) \propto \exp\left(-\frac{\kappa(\mu-m)^2}{2}\right)$

Posterior: $\pi(\mu|d) \propto \exp\left(-\frac{1}{2}\left[\frac{n(\bar{d}-\mu)^2}{\sigma^2} + \kappa^2(\mu-m)^2\right]\right)$

Completing the square:

$$\mu|d \sim \mathcal{N}\left(\frac{\kappa^2 m + n\bar{d}/\sigma^2}{\kappa^2 + n/\sigma^2}, \frac{1}{\kappa^2 + n/\sigma^2}\right)$$

Precision Parameterization

Key insight: Work with **precisions** (inverse variances)

- **Prior precision:** κ^2
- **Data precision:** n/σ^2
- **Posterior precision:** $\kappa^2 + n/\sigma^2$

Posterior mean: Precision-weighted average of prior and data means

$$\widehat{\mu}_{\text{post}} = \frac{\kappa^2 \cdot m + (n/\sigma^2) \cdot \bar{d}}{\kappa^2 + n/\sigma^2}$$

Posterior variance: harmonic mean of prior and sample-mean variances

$$\widehat{\sigma}^2_{\text{post}} = \frac{1}{\kappa^2 + n/\sigma^2}$$

Beta-Binomial Conjugacy

Setup: $D \sim \mathcal{B}(n, \theta)$, prior $\theta \sim \mathrm{Beta}(\alpha, \beta)$

Prior density: $\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

Conjugate family calculation

Likelihood: $p(d|\theta) = \binom{n}{d}\theta^{d}(1-\theta)^{n-d}$

Posterior: $\pi(\theta|d) \propto \theta^{d}(1-\theta)^{n-d} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\alpha+d-1}(1-\theta)^{\beta+n-d-1}$

Result: $\theta|d \sim \mathrm{Beta}(\alpha+d, \beta+n-d)$

Posterior mean: $\dfrac{\alpha+d}{\alpha+\beta+n}$

Understanding Beta Priors

Common Beta priors

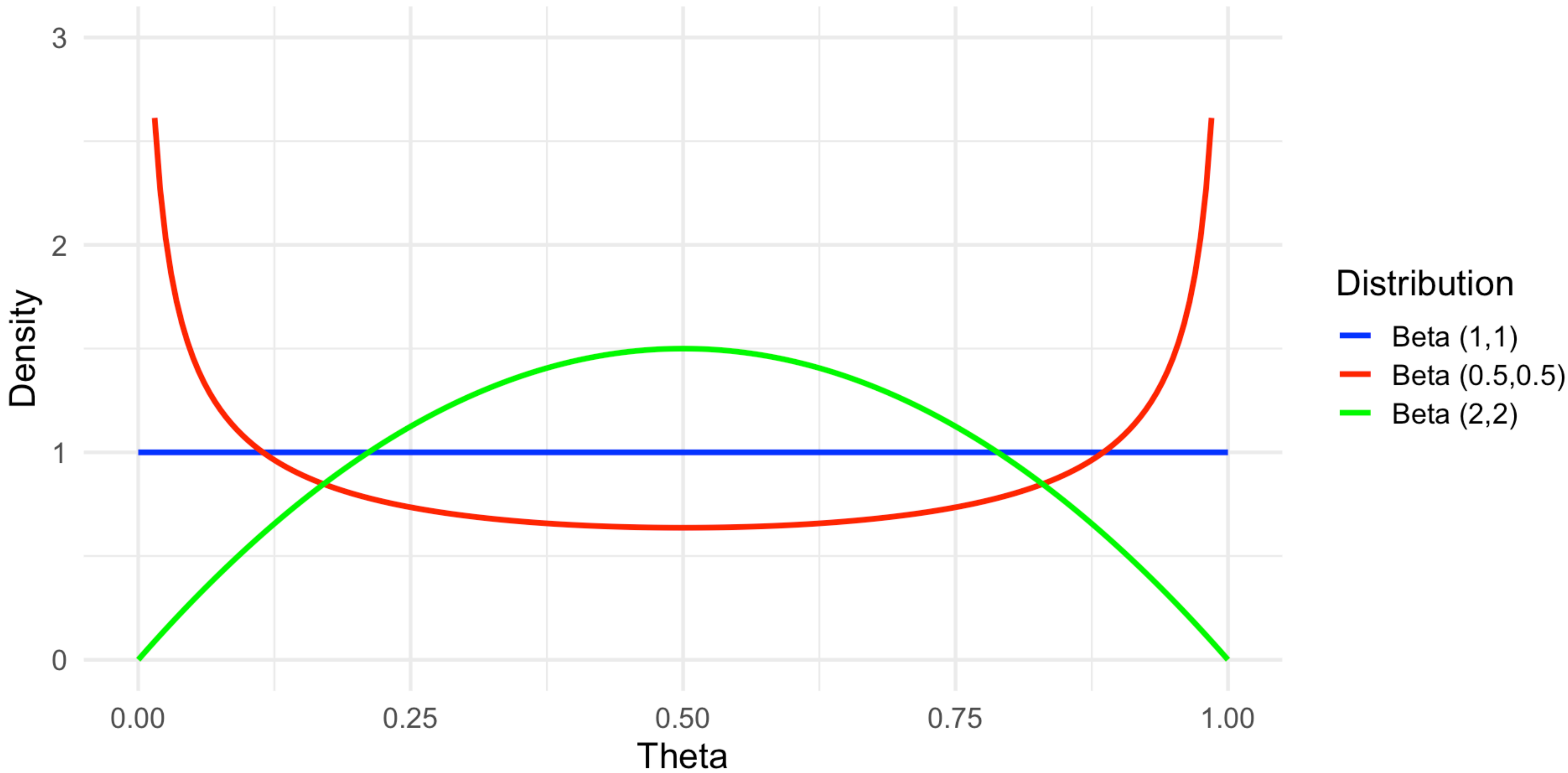
Uniform prior: $\alpha = \beta = 1$ → $\mathrm{Beta}(1, 1) = \mathrm{Uniform}[0, 1]$

Jeffreys prior: $\alpha = \beta = 1/2$ → U-shaped, more weight at extremes

Symmetric informative: $\alpha = \beta = 2$ → Bell-shaped, peaked at 0.5

Interpretation: $\alpha - 1$ = “prior successes”, $\beta - 1$ = “prior failures”

Beta densities



Bayesian Risk

Definition: The **Bayesian risk** of estimator $S = s(D)$ is:

$$R_\pi(S) = \int R_\theta(S)\pi(\theta)d\theta = \mathbb{E}\left[L(s(D), \eta(\theta_{rv}))\right]$$

Key advantage: Single number to compare estimators (vs. entire risk functions)

- **Bayes risk:** $\inf_s R_\pi(S)$ (minimum possible Bayesian risk)
- An estimator that reaches the Bayes risk is named a **Bayes estimator**

Key Theorem: Bayes Estimator

Fundamental Result

The Bayesian risk can be written as:

$$R_\pi(S) = \int \mathbb{E}\Big[L(s(d), \eta(\theta_{rv}))\Big| D = d\Big]\, p(d)\, \mathrm{d}d$$

Therefore, we can minimize $R_\pi(S)$ and get a **Bayes estimator** by solving:

$$s(d) = \arg\min_s \mathbb{E}[L(s, \eta(\theta_{rv}))|D = d]$$

for each observed dataset d

Practical importance: We only need to solve this optimization once, for our observed data!

Optimal Estimators by Loss

Loss-Specific Solutions

Squared loss: $L(s, \beta) = (s - \beta)^2$

$$\text{Bayes estimator} = \mathbb{E}[\beta|\text{data}] = \text{posterior mean}$$

Absolute loss: $L(s, \beta) = |s - \beta|$

$$\text{Bayes estimator} = \text{median}(\beta|\text{data}) = \text{posterior median}$$

Limit case: consider $L_a(s, \beta) = |s - \beta|^a$, and let $a \to 0$

$$\lim_{a \to 0} \text{Bayes estimator}_a = \text{mode}(\beta|\text{data}) = \text{posterior mode}$$

Computing Bayes Estimators

Binomial + Beta(α, β)-prior + squared loss:

$$\text{Bayes estimator} = \mathbb{E}[\theta|d] = \frac{\alpha + d}{\alpha + \beta + n}$$

Gaussian + Normal(m, κ^{-2})-prior + squared loss:

$$\text{Bayes estimator} = \mathbb{E}[\mu|d] = \frac{\kappa^2 m + n\bar{d}/\sigma^2}{\kappa^2 + n/\sigma^2}$$

Amazing fact: These match our frequentist estimators $S_{\alpha,\beta}$ and $S_{\kappa,m}$!

Bayes-Frequentist Connections

Admissibility Theorem

Any **Bayes estimator** S is **admissible**.

Intuition: If a Bayes estimator were not admissible, we could find another estimator S' with lower risk function, thus lower Bayesian risk, contradicting the definition of Bayes estimator.

Minimax Connections

Key Result

If estimator S is admissible with **constant risk function** $R_\theta(S) = c$, then S is the **unique minimax estimator**.

Applications:

- **Gaussian:** When σ^2 known, $S = \overline{D}$ is minimax (constant risk!)
- **Binomial:** Which of our $S_{\alpha,\beta}$ is minimax?

Minimax estimator in Binomial case

- The risk function of $S_{\alpha,\beta}$

$$R_\theta(S_{\alpha,\beta}) = \frac{(\alpha(1-\theta) - \theta\beta)^2 + n\theta(1-\theta)}{(n+\alpha+\beta)^2}$$

  is a quadratic function of θ.

- It is a constant function of θ iff $\alpha = \beta = \sqrt{n} \Big/ 2$.

- Hence the minimax estimator is

$$S_{\sqrt{n}/2,\sqrt{n}/2} = \frac{D + \sqrt{n}/2}{n + \sqrt{n}}$$

Advanced Theory Summary

Connections Web

Bayes estimators with proper priors are **admissible**

Constant risk frequently indicates **minimax property**

Jeffreys priors often yield **minimax estimators**

Part 3: Hypothesis Testing as Decision Theory

From Estimation to Testing

Key observation: When β takes only a **finite** number of values, we're doing **hypothesis testing**!

Classic setup: $\beta \in \{0, 1\}$

- $\beta = 0$: **Null hypothesis** H_0 is true
- $\beta = 1$: **Alternative hypothesis** H_1 is true

Decision: Based on data d, choose $s(d) \in \{0, 1\}$

0-1 loss: $L(s, \beta) = \mathbf{1}\{s \neq \beta\}$

- Cost = 0 if correct decision
- Cost = 1 if wrong decision

Bayesian Hypothesis Testing

Bayesian setup: $\beta = \eta(\theta_{rv})$ where $\theta_{rv} \sim \pi(\theta)$

Posterior probabilities:

$$p_0(d) = \mathbb{P}(\beta = 0 | D = d), \quad p_1(d) = \mathbb{P}(\beta = 1 | D = d)$$

Bayes Decision Rule

Under 0-1 loss, the **Bayes estimator** is:

$$s^*(d) = \arg\max_i p_i(d)$$

For K=2: Choose $s = 1$ if $p_1(d) > p_0(d)$, otherwise $s = 0$

Equivalently: Choose $s = 1$ if $p_1(d) > 1/2$

Example: Testing $\mu > 0$ vs $\mu \leq 0$

Setup:

- Statistical model: $D \sim \mathcal{N}(\mu, \sigma^2)^{\otimes n}$ with σ^2 known
- Hypotheses: $H_0 : \mu \leq 0$ vs $H_1 : \mu > 0$
- Parameter of interest: $\beta = \mathbf{1}\{\mu > 0\}$

Prior: $\mu \sim \mathcal{N}(0, \tau^2)$ (symmetric around 0)

Question: What is $p_1(d) = \mathbb{P}(\mu > 0 | D = d)$?

Posterior Calculation

Posterior: $\mu|d \sim \mathcal{N}\left(\frac{\tau^2 n\bar{d}}{\tau^2 n + \sigma^2}, \frac{\tau^2\sigma^2}{\tau^2 n + \sigma^2}\right)$

Posterior probability:

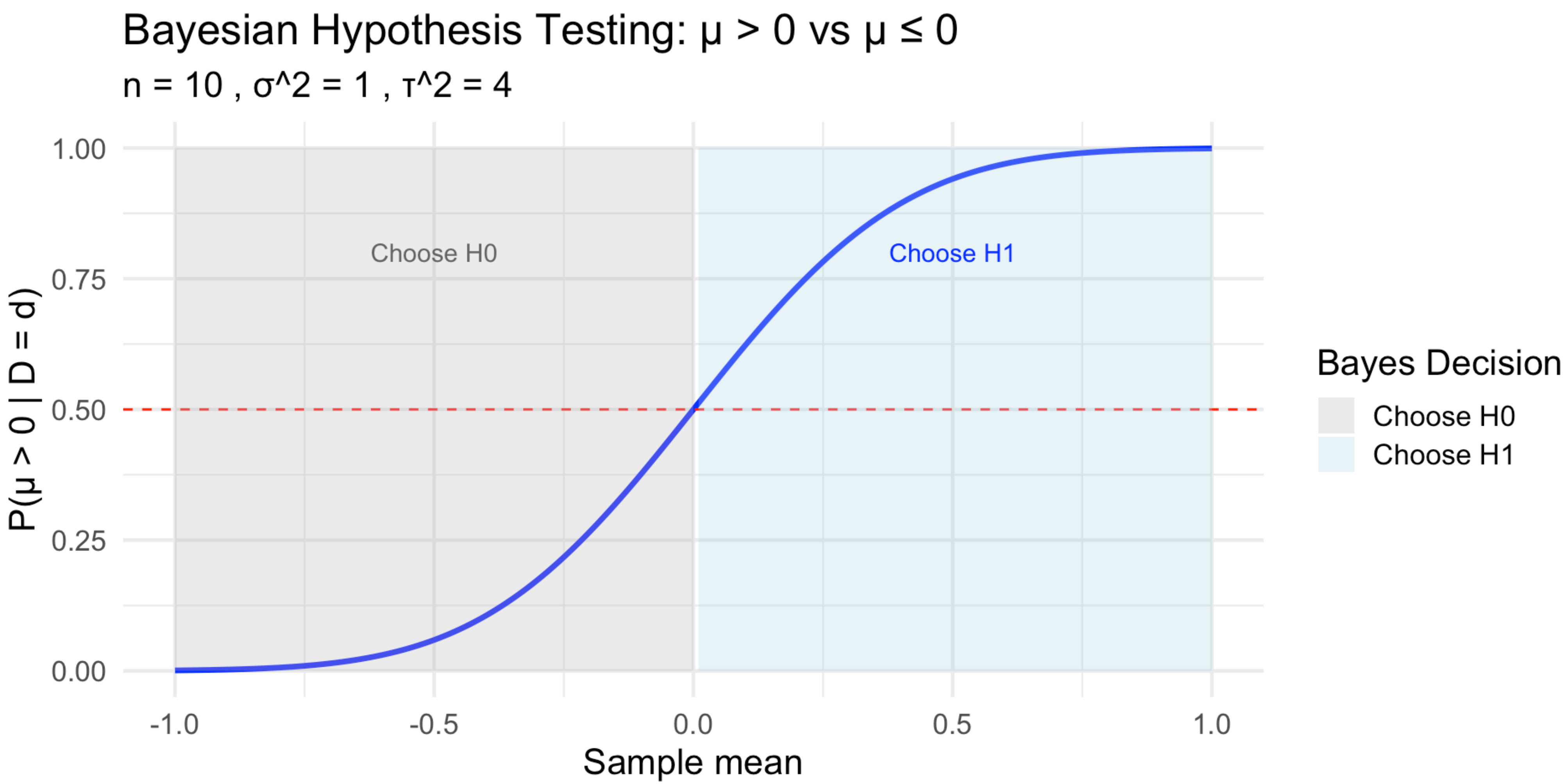
$$p_1(d) = \mathbb{P}(\mu > 0|D = d) = \Phi\left(\frac{\tau^2 n\bar{d}/\sigma^2}{\sqrt{\tau^2 n + \sigma^2}/\sigma}\right)$$

where Φ is the standard normal CDF.

Bayes Decision Rule

Choose $H_1 : \mu > 0$ if: $\Phi\left(\frac{\tau^2 n\bar{d}/\sigma^2}{\sqrt{\tau^2 n + \sigma^2}/\sigma}\right) > 0.5 \quad \Longleftrightarrow \quad \bar{d} > 0$

Visualization: Posterior Probabilities



Key insight: Note the symmetry of the decision around 0.

Frequentist Risk Interpretation

For estimator $S \in \{0,1\}$ and $\beta \in \{0,1\}$:

$$R_\theta(S) = \mathbb{E}_\theta[\mathbf{1}\{S \neq \beta\}] = \mathbb{P}_\theta(S \neq \beta)$$

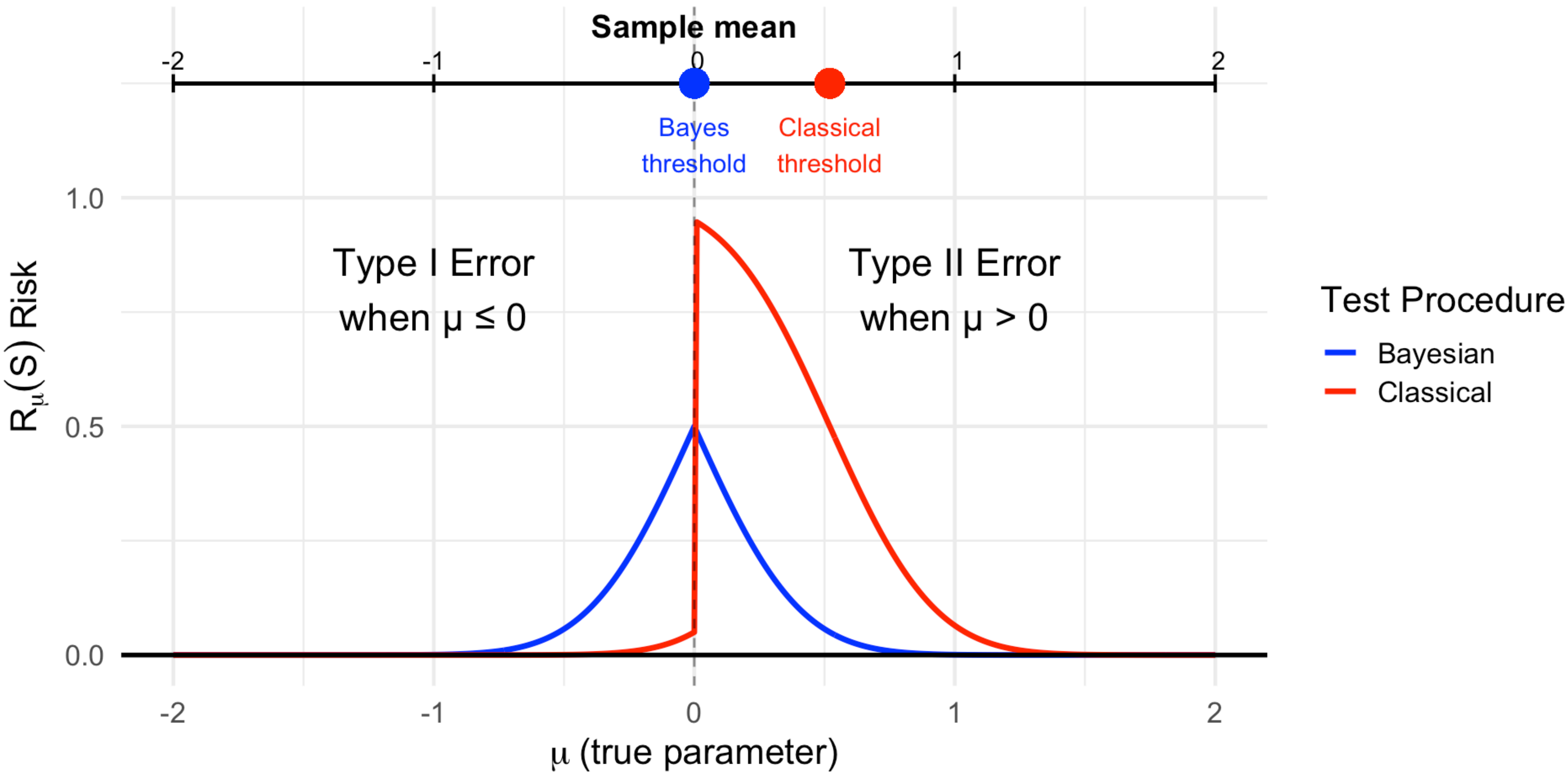
Connection to classical testing:

- When θ s.t. $\beta = \eta(\theta) = 0$: $R_\theta(S) = \mathbb{P}_\theta(S = 1)$ = **Type I error** (size)
- When θ s.t. $\beta = \eta(\theta) = 1$: $R_\theta(S) = \mathbb{P}_\theta(S = 0) = 1 - \mathbb{P}_\theta(S = 1)$ = **1 - Power**

- **Classical testing trade-off:** Lower Type I error ↔ Higher Type II error.
- **Decision theory:** Find estimator minimizing overall risk

Risk Functions for Testing

Gaussian data, $n = 10, \sigma^2 = 1$



Conclusions on risk functions

- Note that Bayesian and Classical tests are both **admissible** here
- **Classical** tests
  - control the **risk** function (less than 5%) under H_0, i.e., when $\beta = 0$, by design
  - **dyssymmetrize** the roles of H_0 and H_1: more **conservative** for H_0
- **Bayesian** tests
  - have **no dyssymmetry** if prior probabilities of H_0 and H_1 are equal
  - return our **belief that H_1 is true** given the observed data d, i.e. $P(\mu > 0 | D = d)$
  - thus typically have **overall better risk** performance!

General K-Class Testing

Extension: $\beta \in \{0, 1, \ldots, K-1\}$ (K hypotheses)

Bayes rule: Choose $s^* = \arg\max_i p_i(d)$

Example applications:

- **Model selection:** Which of K models fits best?
- **Classification:** Which of K classes does observation belong to?
- **Multiple comparisons:** Which treatments differ from control?

Key advantage: Decision theory provides **unified framework** for all these problems!

Connection to Classical Testing

Relationship Summary

Classical hypothesis testing: Fix Type I error, minimize Type II error

Decision theory approach:

- Choose loss function reflecting costs of different errors
- Find estimator minimizing expected loss
- Naturally balances all types of errors

Bayesian testing: Prior probabilities weight different hypotheses

Frequentist testing: Focus on worst-case or minimax performance

Binary Hypothesis Testing

When $K = 2$: Choose $s = 1$ if $p_1(d) > 1/2$, otherwise $s = 0$

Frequentist risk interpretation:

$$R_\theta(S) = \mathbb{P}_\theta(S \neq \beta)$$

- When $\beta = 0$: Risk = $\mathbb{P}_\theta(S = 1)$ = **Type I error**
- When $\beta = 1$: Risk = $\mathbb{P}_\theta(S = 0)$ = **1 - Power**

Connecting All Approaches

Summary

Frequentist: Compare risk functions, seek admissible/minimax estimators

Bayesian: Integrate over prior, minimize expected loss

Hypothesis testing: Special case with 0-1 loss, connects to power/size

Key insight: Choice of loss function determines optimal estimator!

Applications and Extensions

Real-world considerations:

- Asymmetric losses
- Multiple decision problems
- Robustness to model misspecification
- Computational aspects

Take-away: Statistical decision theory provides unified framework for estimation and testing problems.

Sum up

Key Takeaways

Main Messages

1. **Loss function choice is crucial** - different losses lead to different optimal estimators
2. **Frequentist vs Bayesian**: Different philosophies, but decision theory unifies both
3. **Bias-variance trade-off**: Prior information can reduce variance at cost of bias
4. **Hypothesis testing**: Special case of general decision framework

Thank you!