

# Decision Theory Basics

Andrew Nobel

August 2025

# Decision Theory Overview

Decision theory provides a framework for specifying and formalizing inference problems. Components include

- ▶ Statistical model: Family of distributions governing observations
- ▶ Parameters indexing distributions in the statistical model
- ▶ Observation drawn from an unknown distribution in the model
- ▶ Decision rules: inference about unknown distribution based on observations
- ▶ Loss function: relates actions and index of unknown distribution
- ▶ Risk function: means of comparing decision rules

# Statistical Inference

**Statistical model:** Family  $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$  of distributions on fixed set  $\mathcal{X}$

- ▶ Index set  $\Theta$  called *parameter space*. Typically  $\Theta \subseteq \mathbb{R}^d$
- ▶ Elements  $\theta \in \Theta$  called *parameters*
- ▶ Set  $\mathcal{X}$  called *sample space*. Examples  $\mathcal{X} \subseteq \mathbb{R}^n$ ,  $\mathcal{X} \subseteq \mathbb{N}^n$ , or  $\mathcal{X} \subseteq \{0, 1\}^n$

**Observation:** Random variable/vector  $X \in \mathcal{X}$

- ▶ Key assumption:  $X \sim f(\cdot|\theta^*)$  where  $f(\cdot|\theta^*)$  is an unknown element of model  $\mathcal{P}$

**Inference:** Reason about true parameter  $\theta$  based on observation  $X$

- ▶ Reasoning involves actions, formalized by decision rules
- ▶ Quality of actions measured by a loss function
- ▶ Quality of decision rules measured by risk function

# Decision Theory: Basic Ingredients

1. **Sample space.** Set  $\mathcal{X}$  of all possible outcomes of the experiment
2. **Statistical model.** Family  $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$  of distributions on  $\mathcal{X}$
3. **Parameter space.** Set  $\Theta$  of parameters  $\theta$  that index elements of  $\mathcal{P}$
4. **Observation.** Random variable (or vector) with  $X \sim f(\cdot|\theta) \in \mathcal{P}$
5. **Action space.** Set  $\mathcal{A}$  of possible actions in response to observation  $X$
6. **Decision rule.** A map  $d : \mathcal{X} \rightarrow \mathcal{A}$  from observations to actions
7. **Allowable rules.** Family  $\mathcal{D}$  of decision rules  $d$  depending on inference problem
8. **Loss function.** Function  $\ell : \Theta \times \mathcal{A} \rightarrow \mathbb{R}$ . Interpretation

$$\ell(\theta, a) = \text{cost if we make decision } a \text{ when true state of nature is } \theta$$

## Example: Point Estimation

### Set-Up

- ▶ Observation  $X$  in sample space  $\mathcal{X}$  (typically  $\mathbb{R}^n$ ,  $\mathbb{N}^n$ , or  $\{0, 1\}^n$ )
- ▶ Model  $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$  with parameter space  $\Theta \subseteq \mathbb{R}^d$
- ▶ Assume  $X \sim f(\cdot|\theta^*)$  where  $\theta^* \in \Theta$  is unknown

**Inference:** Estimate  $\theta^*$  based on observed value  $x$  of  $X$

- ▶ Action space  $\mathcal{A} = \Theta$
- ▶ Decision rule  $d : \mathcal{X} \rightarrow \Theta$  is an *estimator*. Common to write  $d(x)$  as  $\hat{\theta}(x)$
- ▶ When  $d = 1$ , squared loss  $\ell(\theta, \theta') = (\theta - \theta')^2$  or absolute loss  $\ell(\theta, \theta') = |\theta - \theta'|$

## Example: Hypothesis Testing

### Set-Up

- ▶ Observation  $X$  in sample space  $\mathcal{X}$  (typically  $\mathbb{R}^n$ ,  $\mathbb{N}^n$ , or  $\{0, 1\}^n$ )
- ▶ Model  $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$  with parameter space  $\Theta$
- ▶ Partition  $\Theta = \Theta_0 \cup \Theta_1$  of parameter space
- ▶ Observation  $X \sim f(\cdot|\theta^*)$  where  $\theta^* \in \Theta$  is unknown

**Inference:** Decide if  $\theta^* \in \Theta_0$  or  $\theta^* \in \Theta_1 = \Theta_0^c$  based on value  $x$  of  $X$

- ▶ Action space  $\mathcal{A} = \{0, 1\}$
- ▶ Decision rule  $d : \mathcal{X} \rightarrow \{0, 1\}$  is an *hypothesis test*
- ▶ Zero-one loss  $\ell(\theta, a) = \mathbb{I}(\theta \notin \Theta_a)$  (1 if decision is incorrect, 0 otherwise)

# Example: Confidence Interval/Set Estimation

## Set-Up

- ▶ Observation  $X$  in sample space  $\mathcal{X}$  (typically  $\mathbb{R}^n$ ,  $\mathbb{N}^n$ , or  $\{0, 1\}^n$ )
- ▶ Model  $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$  with parameter space  $\Theta \subseteq \mathbb{R}^d$
- ▶ Assume  $X \sim f(\cdot|\theta^*)$  where  $\theta^* \in \Theta$  is unknown

**Inference:** Find *confidence set*  $C \subseteq \Theta$  likely to contain  $\theta^*$  based on value  $x$  of  $X$

- ▶ Action space  $\mathcal{A}$  = family of subsets of  $\Theta$  (e.g., intervals when  $\Theta \subseteq \mathbb{R}$ )
- ▶ Decision rule  $d : \mathcal{X} \rightarrow \mathcal{A}$  is a confidence set estimate
- ▶ Weighted 0-1 loss  $\ell(\theta, C) = \mathbb{I}(\theta \notin C) + \lambda \text{length}(C)$ , some  $\lambda > 0$

# The Risk Function

**Idea:** The *risk function* of a decision rule tells us how well that rule performs for each possible parameter  $\theta \in \Theta$ . It is the basis for comparing decision rules

**Definition:** The *risk function* of a decision rule  $d : \mathcal{X} \rightarrow \mathcal{A}$  is defined by

$$R(\theta, d) = \mathbb{E}_\theta \ell(\theta, d(X)) \quad \theta \in \Theta$$

- ▶ Notation:  $\mathbb{E}_\theta h(X)$  is the expectation of  $h(X)$  when  $X \sim f(\cdot|\theta)$
- ▶  $R(\theta, d)$  = expected loss of rule  $d$  when applied to observation  $X \sim f(\cdot|\theta)$
- ▶ Continuous case  $R(\theta, d) = \int \ell(\theta, d(x)) f(x|\theta) dx$
- ▶ Discrete case  $R(\theta, d) = \sum_x \ell(\theta, d(x)) p(x|\theta)$

## Point Estimation Under Squared Loss

Given family  $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$  with  $\Theta \subseteq \mathbb{R}$ , and an estimator  $\hat{\theta} : \mathcal{X} \rightarrow \Theta$

- ▶ The *bias* of  $\hat{\theta}$  at  $\theta$  is  $\text{bias}_\theta(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}(X)] - \theta$
- ▶ The *variance* of  $\hat{\theta}$  at  $\theta$  is  $\text{Var}_\theta(\hat{\theta}) = \mathbb{E}_\theta \left[ \hat{\theta}(X) - \mathbb{E}_\theta \hat{\theta}(X) \right]^2$
- ▶ Say  $\hat{\theta}$  is *unbiased* if  $\text{bias}_\theta(\hat{\theta}) = 0$  for all  $\theta$

**Bias-Variance Decomposition:** Under the squared loss  $\ell(\theta, a) = (\theta - a)^2$

$$R(\theta, \hat{\theta}) = (\text{bias}_\theta(\hat{\theta}))^2 + \text{Var}_\theta(\hat{\theta})$$

**Upshot:** For  $\hat{\theta}$  to perform well when  $X \sim f(\cdot|\theta)$  it should

- ▶ Be centered near the true parameter (small bias)
- ▶ Not be too spread out (small variance)

# Example: Estimation of a Normal Mean

## Setting

- ▶ Single observation  $X \in \mathbb{R}$
- ▶  $\mathcal{P} = \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$
- ▶ Estimate  $\theta$  from  $X$  under the squared error loss

## Two estimators

- ▶  $\hat{\theta}_1(x) = x$ , risk function  $R(\theta, \hat{\theta}_1) = 1$
- ▶  $\hat{\theta}_2(x) = 3$ , risk function  $R(\theta, \hat{\theta}_2) = (\theta - 3)^2$

Neither risk function dominates the other

## Example: Probability of Success in Bernoulli Trial

### Setting

- ▶ Observations  $X = (X_1, \dots, X_n) \in \{0, 1\}^n$
- ▶  $\mathcal{P} = \{\text{Bern}(\theta)^n : \theta \in (0, 1)\}$
- ▶ Estimate  $\theta$  from  $X$  under the squared error loss

### Two estimators

$$\hat{\theta}_1(x) = \bar{x} \quad R(\theta, \hat{\theta}_1) = \frac{\theta(1-\theta)}{n}$$

$$\hat{\theta}_2(x) = \frac{n\bar{x} + \sqrt{n}/2}{n + \sqrt{n}} \quad R(\theta, \hat{\theta}_2) = \frac{1}{4(1 + \sqrt{n})^2} \quad (\text{constant})$$

Neither risk function dominates the other

# Admissibility

**Setting:** General inference problem with family  $\mathcal{D}$  of candidate decision rules

**Definition:** A decision rule  $d \in \mathcal{D}$  is *inadmissible* if there is some  $d' \in \mathcal{D}$  such that

- (i)  $R(\theta, d') \leq R(\theta, d)$  for all  $\theta \in \Theta$
- (ii)  $R(\theta, d') < R(\theta, d)$  for some  $\theta \in \Theta$

If no such  $d'$  exists, then  $d$  is said to be *admissible*

- ▶ Admissibility depends on the family  $\mathcal{D}$  and the loss function  $\ell$
- ▶ A rule  $d$  is either admissible or inadmissible
- ▶ Admissible rules are candidates for good/reasonable rules
- ▶ There may be many admissible rules
- ▶ Admissibility is a weak criterion. Obviously silly rules can be admissible.

## Example

**Observations:**  $X_1, \dots, X_n$  i.i.d.  $\text{Bern}(\theta)$  with  $\theta \in (0, 1)$

**Goal:** Estimate of  $\theta$  under squared loss. Candidate estimators

▶  $\hat{\theta}_1(x) = \bar{x}$  with  $R(\theta, \hat{\theta}_1) = \theta(1 - \theta)/n$

▶  $\hat{\theta}_2(x) = x_1$  with  $R(\theta, \hat{\theta}_2) = \theta(1 - \theta)$

▶  $\hat{\theta}_3(x) = \frac{1}{2}$  with  $R(\theta, \hat{\theta}_3) = (\theta - \frac{1}{2})^2$

### Fact

1.  $\hat{\theta}_1$  is admissible
2.  $\hat{\theta}_2$  is inadmissible (bettered by  $\hat{\theta}_1$ )
3.  $\hat{\theta}_3$  is *admissible* (lazy, but unbeatable when  $\theta = \frac{1}{2}$ )

# Frequentist and Bayesian Perspectives on Inference

Different approaches stemming in part from different interpretations of probability

## Frequentist

- ▶ Probability defined through repetitions of a random experiment
- ▶ True parameter  $\theta$  is a fixed element of  $\Theta$ , but otherwise unknown
- ▶ Analysis and interpretation of inference based on (potentially unrealized) replications of basic experiment

## Bayesian

- ▶ Probability understood as a (potentially subjective) measure of belief
- ▶ Belief about true parameter before/after an experiment represented by prior/posterior distributions on the parameter space  $\Theta$
- ▶ Experiment regarded as unique. Inference based on updating prior based on data, without reference to other experiments or repetition

# Review of Bayesian Inference

## Ingredients

- ▶ Family  $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$  of distributions on  $\mathcal{X}$
- ★ Prior density  $\pi(\theta)$  on parameter space  $\Theta$
- ★ Joint density  $f(x, \theta) = f(x|\theta)\pi(\theta)$ , marginal density  $m(x) = \int f(x, \theta)d\theta$
- ★ Observation model: First generate  $\theta \sim \pi$ , then generate  $X \sim f(\cdot|\theta)$

**Idea:** Prior  $\pi(\theta)$  reflects information about parameters *before* the experiment. Once data  $x$  is obtained, prior updated using Bayes formula to obtain *posterior* distribution

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$$

**Key:** All inferences about  $\theta$  are based on the posterior distribution

## Comparing Decision Rules: Different Perspectives

**Recall:** Risk of decision rule  $d : \mathcal{X} \rightarrow \Theta$  under loss  $\ell$  summarized by risk *function*

$$R(\theta, d) = \mathbb{E}_\theta \ell(\theta, d(X))$$

**Question:** How should we compare two decision rules  $d_1$  and  $d_2$  based on their risk functions  $R(\theta, d_1)$  and  $R(\theta, d_2)$ ?

- ▶ Frequentist perspective: Consider *maximum risk* of each rule over all  $\theta \in \Theta$
- ▶ Bayesian perspective: Consider *average risk* of each rule relative to prior  $\pi$

# Maximum Risk and Bayes Risk

**Idea:** Single number summaries of overall risk derived from risk function  $R(\theta, d)$

**Definition:** Given family  $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$  and loss function  $\ell : \Theta \times \mathcal{A} \rightarrow \mathbb{R}$

(i) The *maximum risk* of a decision rule  $d : \mathcal{X} \rightarrow \mathcal{A}$  is

$$R_m(d) = \max_{\theta \in \Theta} R(\theta, d)$$

(ii) The *Bayes risk* of a decision rule  $d : \mathcal{X} \rightarrow \mathcal{A}$  under prior density  $\pi$  is

$$R_\pi(d) = \int R(\theta, d)\pi(\theta)d\theta$$

## Example: Probability of Success in Bernoulli Trial

**Recall:** Observe  $X_1, \dots, X_n \sim \text{Bern}(\theta)$ . Estimators  $\hat{\theta}_1, \hat{\theta}_2$  for  $\theta$  with

$$R(\theta, \hat{\theta}_1) = \frac{\theta(1-\theta)}{n} \quad R(\theta, \hat{\theta}_2) = \frac{1}{4(1+\sqrt{n})^2}$$

A. *Maximum risk:* Prefer estimator  $\hat{\theta}_2(x)$  as

$$R_m(\hat{\theta}_1) = \frac{1}{4n} > \frac{1}{4(1+\sqrt{n})^2} = R_m(\hat{\theta}_2)$$

B. *Bayes risk:* Under uniform prior  $\pi(\theta) = 1$ , prefer estimator  $\hat{\theta}_1$  for  $n \geq 20$  as

$$R_\pi(\hat{\theta}_1) = \frac{1}{6n} < \frac{1}{4(1+\sqrt{n})^2} = R_\pi(\hat{\theta}_2)$$

## Minimax and Bayes Rules for a Family $\mathcal{D}$

**Definition:** The *minimax risk* for a family of decision rules  $\mathcal{D}$  is

$$R_m^* = \min_{d \in \mathcal{D}} R_m(d) = \min_{d \in \mathcal{D}} \max_{\theta \in \Theta} R(\theta, d)$$

A rule  $d \in \mathcal{D}$  is said to be *minimax* if  $R_m(d) = R_m^*$ .

**Definition:** The optimal Bayes risk for a family of decision rules  $\mathcal{D}$  under a prior  $\pi$  is

$$R_\pi^* = \min_{d \in \mathcal{D}} R_\pi(d) = \min_{d \in \mathcal{D}} \int R(\theta, d) \pi(\theta) d\theta$$

A rule  $d \in \mathcal{D}$  is called a *Bayes rule* for  $\pi$  if  $R_\pi(d) = R_\pi^*$ . Note that  $R_\pi^*$  depends on  $\pi$

**Fact:** The minimax risk is always bounded below by the optimal Bayes risk: for every prior distribution  $\pi$  on  $\Theta$  one has  $R_m^* \geq R_\pi^*$

## Bayes Rules with Constant Risk are Minimax

**Theorem:** Let  $d_\pi$  be the Bayes rule for a family  $\mathcal{D}$  under a prior  $\pi$ . If the risk function  $R(\theta, d_\pi^*)$  is constant then  $d_\pi^*$  is minimax for  $\mathcal{D}$ .

Terminology: If  $d_\pi$  is minimax then  $\pi$  is said to be a *least favorable* prior

**Example:**  $X_1, \dots, X_n \sim \text{Bern}(\theta)$ . Let  $\mathcal{D}$  be all point estimators  $\hat{\theta} : \{0, 1\} \rightarrow (0, 1)$  and let  $\ell(\theta, \theta') = (\theta - \theta')^2$  be the squared loss. Consider estimator

$$\hat{\theta}_0(x) = \frac{n\bar{x} + \sqrt{n}/2}{n + \sqrt{n}}$$

- ▶ Have seen  $R(\theta, \hat{\theta}) = 1/4(1 + \sqrt{n})^2$  is constant
- ▶ Can show  $\hat{\theta}_0$  is the Bayes rule for  $\mathcal{D}$  under prior  $\pi_0 = \text{Beta}(\sqrt{n}/2, \sqrt{n}/2)$

Upshot:  $\hat{\theta}$  is minimax among all estimators of  $\theta$ , and  $\pi_0$  is least favorable prior

## Admissibility of Bayes Rules

**Fact:** Consider a Bayesian decision problem in which

- ▶  $\Theta \subseteq \mathbb{R}^p$  is open
- ▶  $\pi(\theta) > 0$  for every  $\theta \in \Theta$
- ▶  $R_\pi^*$  is finite
- ▶  $R(\theta, d)$  is a continuous function of  $\theta$  for each  $d \in \mathcal{D}$

Then the Bayes rule for  $\pi$  is admissible.