Introduction to Decision Theory

Nishant D. Gurnani

July 4, 2020

The Big Picture

$$\text{financial markets} \rightarrow \text{data} \rightarrow \text{statistics} \rightarrow \text{inferences}$$

As statisticians, we are tasked with turning the large amount of data generated by experiments and observations into inferences about the world.

This gives rise to a number of core statistical questions:

1. Modeling: How do we capture the uncertainty in our data and the world that produced it?

2. Methodology: What are the right mathematical and computational tools that allow us to draw these statistical inferences?

3. Analysis: How do we compare and evaluate the statistical inferences we make and the procedures we use to make them? In particular, how do we do optimal inference?

Decision Theory Framework

Decision theory provides a framework to answer all of the core questions. In particular, it allows us to formalize the notion of inference as a **decision problem** consisting of three key ingredients:

1. A **statistical model** is a family of distributions P, indexed by a parameter θ. We write

$$P = \{\mathbb{P}_\theta : \theta \in \Omega\}$$

   Here θ is the parameter, Ω is the parameter space, and each $\mathbb{P}_\theta$ is a distribution.

   P is the class of distributions to which we believe our random sample X belongs. In other words, we assume that the data X come from some $\mathbb{P}_\theta \in P$ but that the true θ is unknown.

   The fact that we don't know θ captures our uncertainty about the problem.

Decision Theory Framework

Example 1 (Weighted coin flips)

Observe a sequence of coin flips $X_1, \ldots, X_n \in \{0, 1\}$ where 0 encodes tailes and 1 encodes heads.

It's a weighted coin, so I don't know how often I expect heads to arrive. The goal is to estimate the probability of heads given the observations.

To do this we model this process as independent draws from a Bernoulli distribution:

$$P = \{\text{Ber}(\theta) : \theta \in [0, 1] = \Omega\}$$

In this case, $\mathbb{P}_\theta(X_i = 1) = \theta$.

Decision Theory Framework

2. A **decision procedure** δ is a map from $\mathcal{X}$ (the sample space) to the decision space $\mathcal{D}$

   Example 2 (Weighted coin flips)

   Taking $P = \{\text{Ber}(\theta) : \theta \in [0,1] = \Omega\}$ as before, we may be interested in estimating θ or testing hypotheses based on θ.

   (a) Estimating θ: the decision space is $\mathcal{D} = [0,1]$, and the decision procedure might be $\delta(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$. This procedure is an example of an **estimator**.

   (b) Accept or rejecting the hypothesis $\theta > 1/2$: the decision space is $\mathcal{D} = \{\text{accept},\ \text{reject}\}$, and one possible decision procedure is $\delta(X) =$ "reject if$\frac{1}{n}\sum_{i=1}^{n} X_i \leq 1/2$, accept otherwise" This procedure is an example of a **hypothesis test**.

Decision Theory Framework

3. A **loss function** is a mapping $L : \Omega \times \mathcal{D} \to \mathbb{R}^+$.

   $L(\theta, d)$ represents the penalty for making the decision d when θ is in fact the true parameter for the distribution generating the data.

   The goal is to assign penalties for bad decisions.

   Example 3 (Squared-error loss).

   For estimating a real-valued parameter θ with decision $d \in \mathbb{R} = \mathcal{D}$, a common loss function is the squared-error loss

$$L(\theta, d) = (\theta - d)^2$$

Analyzing Procedures

Decision theory is useful because it allows us to analyze statistical procedures.

The three components of a decision problem together give rise to our primary basis for evaluation - **the risk function**.

$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(X))]$$

The risk is $R(\theta, \delta)$ is the average loss incurred when the decision procedure δ is used over many draws of the data from its generating distribution $\mathbb{P}_\theta$.

The risk function gives us a way to compare and rule out procedures.

δ is **inadmissible** if there exists δ' such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all θ and $R(\theta', \delta') < R(\theta', \delta)$ for some θ'.

Analyzing Procedures

Example 4 (Weighted coin flips)

For estimating the probability of heads θ, let us consider three possible decision procedures:

1. $\delta_n(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean of the first n data points
2. $\delta_1(X) =$ naive estimator that uses the outcome of just one flip
3. $\delta_{\text{goofy}}(X) = \frac{1}{2}$ which is a fixed constant

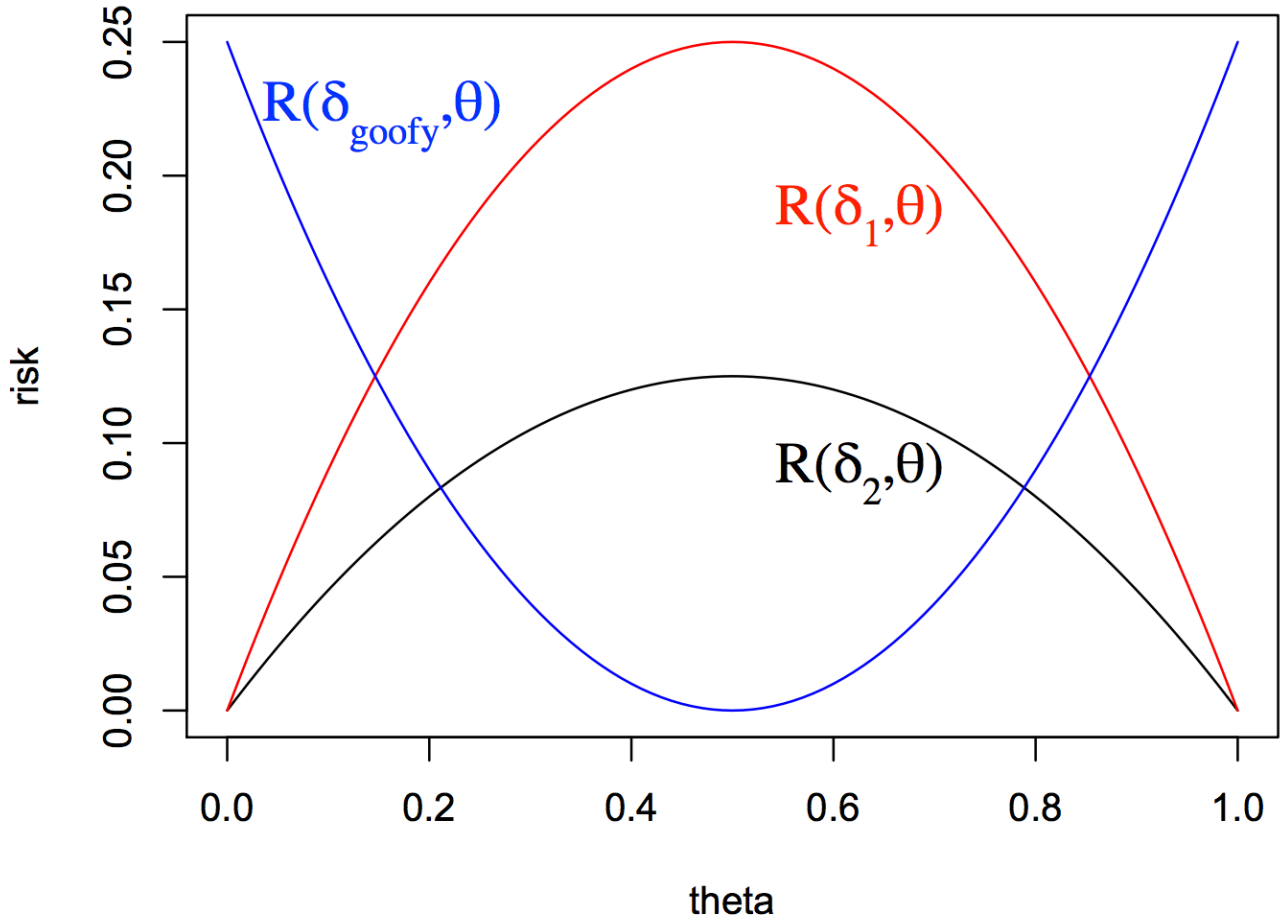
Under the loss function $L(\theta, d) = (\theta - d)^2$ the risk of each is:

1. $R(\theta, \delta_n) = \mathbb{E}_\theta[(\theta - \delta_n(X))^2] = \frac{\theta(1-\theta)}{n}$
2. $R(\theta, \delta_1) = \mathbb{E}_\theta[(\theta - \delta_1(X))^2] = \theta(1-\theta)$
3. $R(\theta, \delta_{\text{goofy}}) = (\theta - \frac{1}{2})^2$

Which one is the best procedure?

Analyzing Procedures

Clearly there is no uniformly best procedure, to develop concrete notions of optimality we must induce an optimizable problem.



Optimal Inference

To develop our theory of optimality we change the requirements of our decision problem.

In particular, we try to induce an optimizable problem by taking one of the following actions:

1. **Constrain** the set of decision procedures under consideration, by requiring our procedures to satisfy criteria like unbiasedness or invariance.

   (a) Unbiased estimators: we say that δ is unbiased for estimating $g(\theta)$ if $\mathbb{E}_\theta[\delta(X)] = g(\theta)$

   (b) Equivariance or invariance: enforce symmetries in the decision procedure. For example, location invariance requires an estimator to satisfy $\delta(X + c) = \delta(X) + c$

Optimal Inference

2. **Collapse** the risk function into a single numerical summary and minimize this overall summary of risk instead of requiring uniformly lower risk.

   (a) Bayes procedures minimize the average risk $\int R(\theta, \delta) d\Lambda(\theta)$ where Λ is a probability distribution (the prior distribution) over Ω.

   (b) Minimax procedures minimize the worst-case risk, $\sup_{\theta \in \Omega} R(\theta, \delta)$, and hence achieve the best worst-case performance.

These principles are foundational to the theory of optimal inference and apply both in the context of point estimation and hypothesis testing.