The Decision Theory Framework

Statistical Theory

Guillaume Dehaene
Ecole Polytechnique Fédérale de Lausanne

Statistics as a Random Game

Statistics as a Random Game?

Nature and a statistician decide to play a game. What's in the box?

- A *family of distributions* $\mathcal{F}$, usually assumed to admit densities (frequencies). This is the variant of the game we decide to play.
- A *parameter space* $\Theta \subseteq \mathbb{R}^p$ which parametrizes the family $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$. This represents the space of possible *plays/moves* available to Nature.
- A *data space* $\mathcal{X}$, on which the parametric family is supported. This represents the space of possible outcomes following a play by Nature.
- An *action space* $\mathcal{A}$, which represents the space of possible *actions* or *decisions* or *plays/moves* available to the statistician.
- A *loss function* $\mathcal{L} : \Theta \times \mathcal{A} \to \mathbb{R}^+$. This represents how much the statistician has to pay nature when losing.
- A set $\mathcal{D}$ of *decision rules*. Any $\delta \in \mathcal{D}$ is a (measurable) function $\delta : \mathcal{X} \to \mathcal{A}$. These represent the possible strategies available to the statistician.

Statistics as a Random Game?

How the game is played:

- First we agree on the rules:
  1. Fix a parametric family $\{F_\theta\}_{\theta\in\Theta}$
  2. Fix an action space $\mathcal{A}$
  3. Fix a loss function $\mathcal{L}$
- Then we play:
  1. Nature selects (plays) $\theta_0 \in \Theta$.
  2. The statistician observes $\mathbf{X} \sim F_{\theta_0}$
  3. The statistician plays $\alpha \in \mathcal{A}$ in response.
  4. The statistician has to pay nature $\mathcal{L}(\theta_0, \alpha)$.

Framework proposed by A. Wald in 1939. Encompasses three basic statistical problems:

- Point estimation
- Hypothesis testing
- Interval estimation

Point Estimation as a Game

In the problem of point estimation we have:

1. Fixed parametric family $\{F_\theta\}_{\theta \in \Theta}$
2. Fixed an action space $\mathcal{A} = \Theta$
3. Fixed loss function $\mathcal{L}(\theta, \alpha)$ (e.g. $\|\theta - \alpha\|^2$)

The game now evolves simply as:

1. Nature picks $\theta_0 \in \Theta$
2. The statistician observes $\mathbf{X} \sim F_{\theta_0}$
3. The statistician plays $\delta(\mathbf{X}) \in \mathcal{A} = \Theta$
4. The statistician loses $\mathcal{L}(\theta_0, \delta(\mathbf{X}))$

Notice that in this setup δ is an *estimator* (it is a statistic $\mathcal{X} \to \Theta$).

The statistician always loses.

$\hookrightarrow$ Is there a good strategy $\delta \in \mathcal{D}$ for the statistician to restrict his losses?

$\hookrightarrow$ Is there an optimal strategy?

Risk (Expected Loss)

Risk of a Decision Rule

Statistician would like to pick strategy δ so as to minimize his losses. But losses are random, as they depend on $\mathbf{X}$.

Definition (Risk)

Given a parameter $\theta \in \Theta$, the *risk* of a decision rule $\delta : \mathcal{X} \to \mathcal{A}$ is the expected loss incurred when employing δ: $R(\theta, \delta) = \mathbb{E}_\theta \left[\mathcal{L}(\theta, \delta(\mathbf{X}))\right].$

Key notion of decision theory

decision rules should be compared by comparing their risk functions

Example (Mean Squared Error)

In point estimation, the mean squared error

$$MSE(\delta(\mathbf{X})) = \mathbb{E}_\theta[\|\theta - \delta(\mathbf{X})\|^2]$$

is the risk corresponding to a squared error loss function.

Coin Tossing Revisited

Consider the "coin tossing game" with quadratic loss:

- Nature picks $\theta \in [0,1]$
- We observe n variables $X_i \overset{iid}{\sim} \text{Bernoulli}(\theta)$.
- Action space is $\mathcal{A} = [0,1]$
- Loss function is $\mathcal{L}(\theta, \alpha) = (\theta - \alpha)^2$.

Consider 3 different decision procedures $\{\delta_j\}_{j=1}^3$:

1. $\delta_1(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i$
2. $\delta_2(\mathbf{X}) = X_1$
3. $\delta_3(\mathbf{X}) = \frac{1}{2}$

Let us compare these using their associated risks as benchmarks.

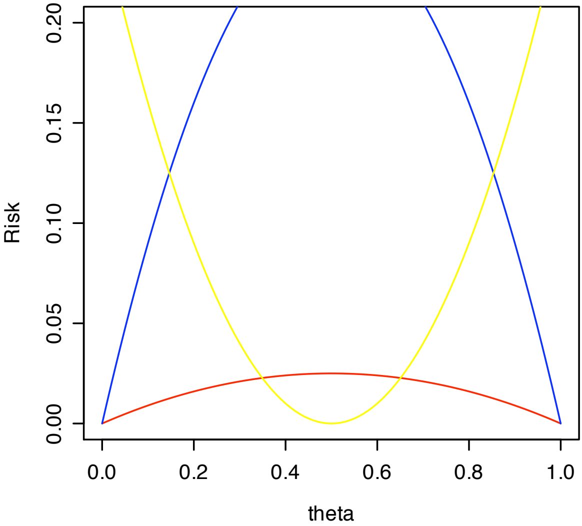
Coin Tossing Revisited

Risks associated with different decision rules:

$$R_j(\theta) = R(\theta, \delta_j(\mathbf{X})) = \mathbb{E}_\theta[(\theta - \delta_j(\mathbf{X}))^2]$$

- $R_1(\theta) = \frac{1}{n}\theta(1-\theta)$
- $R_2(\theta) = \theta(1-\theta)$
- $R_3(\theta) = \left(\theta - \frac{1}{2}\right)^2$

Coin Tossing Revisited – Every dog has its day



$R_1(\theta)$, $R_2(\theta)$, $R_3(\theta)$

Admissibility and Inadmisibility

Risk of a Decision Rule

Definition (Inadmissible Decision Rule)

Let δ be a decision rule for the experiment $(\{F_\theta\}_{\theta\in\Theta}, \mathcal{L})$. If there exists a decision rule δ^* that strictly dominates δ, i.e.

$$R(\theta,\delta^*) \leq R(\theta,\delta),\ \forall\theta\in\Theta \quad \& \quad \exists\ \theta'\in\Theta : R(\theta',\delta^*) < R(\theta',\delta),$$

then δ is called an *inadmissible decision rule*.

$R_2(\theta) > R_1(\theta)$ so $R_2(\theta)$ is inadmissible.

- An inadmissible decision rule is a "silly" strategy since we can find a strategy that always does at least as well and sometimes better.
- However "silly" is with respect to $\mathcal{L}$ and Θ. (it may be that our choice of $\mathcal{L}$ is "silly"!!!)
- If we change the rules of the game (i.e. different loss or different parameter space) then domination may break down.

Risk of a Decision Rule

Example (Exponential Distribution)

Let $X_1, ..., X_n \overset{iid}{\sim}$ Exponential(λ), $n \geq 2$. The MLE of λ is

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

with $\bar{X}$ the empirical mean. Observe that

$$\mathbb{E}_\lambda[\hat{\lambda}] = \frac{n\lambda}{n-1}.$$

It follows that $\tilde{\lambda} = (n-1)\hat{\lambda}/n$ is an unbiased estimator of λ. Observe now that

$$MSE_\lambda(\tilde{\lambda}) < MSE_\lambda(\hat{\lambda})$$

since $\tilde{\lambda}$ is unbiased and $\text{Var}_\lambda(\tilde{\lambda}) < \text{Var}_\lambda(\hat{\lambda})$. Hence the MLE is an inadmissible rule for quadratic loss.

Risk of a Decision Rule

Notice that the parameter space in this example is $(0, \infty)$. In such cases, quadratic loss tends to penalize over-estimation more heavily than under-estimation (the maximum possible under-estimation is bounded!).

Different loss function might change the result!

Risk of a Decision Rule

Example

If we consider another loss:

$$\mathcal{L}(a, b) = a/b - 1 - \log(a/b)$$

where, for each fixed a, $lim_{b \to 0} \mathcal{L}(a, b) = lim_{b \to \infty} \mathcal{L}(a, b) = \infty$. Now, for $n > 1$,

$$\begin{aligned} R(\lambda, \tilde{\lambda}) &= \mathbb{E}_\lambda \left[\frac{n\lambda\bar{X}}{n-1} - 1 - \log\left(\frac{n\lambda\bar{X}}{n-1}\right)\right] \\ &= \underbrace{\mathbb{E}_\lambda \left[\lambda\bar{X} - 1 - \log(\lambda\bar{X})\right]}_{R(\lambda,\hat{\lambda})} + \underbrace{\frac{\mathbb{E}_\lambda(\lambda\bar{X})}{n-1} - \log\left(\frac{n}{n-1}\right)}_{g(n)} \end{aligned}$$

where we wrote $\bar{X} = \frac{n-1}{n}\bar{X} + \frac{1}{n}\bar{X}$.

Example (Exponential Distribution)

Note that $\mathbb{E}_\lambda[\bar{X}] = \lambda^{-1}$, so

$$g(n) = \frac{1}{n-1} - \log\left(\frac{n}{n-1}\right).$$

We claim that $g(n) > 0$ for $n \geq 2$. Using $\log x = \int_1^x t^{-1} dt$, this follows if

$$\begin{aligned} \frac{1}{x} &> \log(x+1) - \log x, \qquad x > 1 \\ \iff \frac{1}{x} &> \int_x^{x+1} t^{-1} dt, \qquad x > 1 \end{aligned}$$

which holds by a rectangle area bound on the integral, as follows:

$$\frac{1}{x} = [(x+1) - x]\frac{1}{x} = \int_x^{x+1} \frac{1}{x} dt > \int_x^{x+1} \frac{1}{t} dt, \quad \text{when } x > 1$$

Consequently, $R(\lambda, \tilde{\lambda}) > R(\lambda, \hat{\lambda})$ and $\hat{\lambda}$ dominates $\hat{\lambda}$.

Criteria for Choosing Decision Rules

Definition (Admissible Decision Rule)

A decision rule δ is *admissible* for the experiment $(\{F_\theta\}_{\theta\in\Theta}, \mathcal{L})$ if it is not strictly dominated by any other decision rule.

- In non-trivial problems, it may not be easy at all to decide whether a given decision rule is admissible.
- Stein's paradox ("one of the most striking post-war results in mathematical statistics"-Brad Efron)

Admissibility is a minimal requirement - what about the opposite end (optimality) ?

- In almost any non-trivial experiment, there will be no decision rule that makes risk uniformly smallest over θ
- Narrow down class of possible decision rules by unbiasedness/symmetry/... considerations, and try to find *uniformly dominating* rules of all other rules (next week!).

Minimax Rules

Minimax Decision Rules

- Another approach to good procedures is to use global rather than local criteria (with respect to θ).

Rather than look at risk at every θ $\leftrightarrow$ Concentrate on maximum risk

Definition (Minimax Decision Rule)

Let $\mathcal{D}$ be a class of decision rules for an experiment $(\{F_\theta\}_{\theta\in\Theta}, \mathcal{L})$. If $\delta \in \mathcal{D}$ is such that

$$\sup_{\theta\in\Theta} R(\theta,\delta) \leq \sup_{\theta\in\Theta} R(\theta,\delta'), \quad \forall\, \delta' \in \mathcal{D},$$

then δ is called a minimax decision rule.

- A minimax rule δ satisfies $sup_{\theta\in\Theta} R(\theta,\delta) = \inf_{\kappa\in\mathcal{D}} \sup_{\theta\in\Theta} R(\theta,\kappa)$.
- In the minimax setup, a rule is *preferable* to another if it has smaller maximum risk.

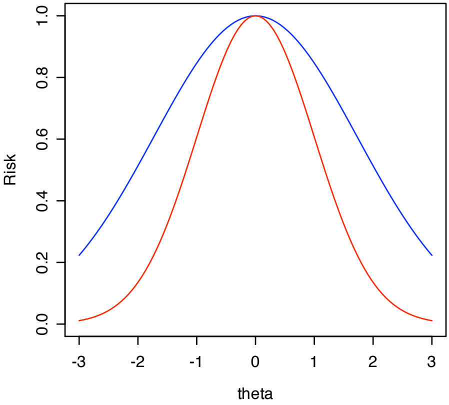
Minimax Decision Rules

A few comments on minimaxity:

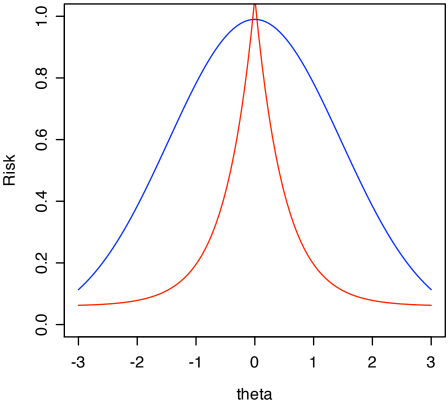
- Motivated as follows: we do not know anything about θ so let us insure ourselves against the worst thing that can happen.
- Makes sense if you are in a zero-sum game: if your opponent chooses θ to maximize $\mathcal{L}$ then one should look for minimax rules. But is nature really an opponent?
- If there is no reason to believe that nature is trying to "do her worst", then the minimax principle is overly conservative: it places emphasis on the "bad θ".
- Minimax rules may not be unique, and may not even be admissible. A minimax rule may very well dominate another minimax rule.
- A unique minimax rule is (obviously) admissible.
- Minimaxity can lead to counterintuitive results. A rule may dominate another rule, except for a small region in Θ, where the other rule achieves a smaller supremum risk.

Minimax Decision Rules

Inadmissible minimax rule



Counterintuitive minimax rule



Bayes Rules

Bayes Decision Rules

- Suppose we have some prior belief about the value of θ. How can this be factored in our risk-based considerations?

Rather than look at risk at every θ $\leftrightarrow$ Concentrate on average risk

Definition (Bayes Risk)

Let $\pi(\theta)$ be a probability density (frequency) on Θ and let δ be a decision rule for the experiment $(\{F_\theta\}_{\theta\in\Theta}, \mathcal{L})$. The π-Bayes risk of δ is defined as

$$r(\pi,\delta) = \int_\Theta R(\theta,\delta)\pi(\theta)d\theta = \int_\Theta\int_\mathcal{X} \mathcal{L}(\theta,\delta(\mathbf{x}))F_\theta[d\mathbf{x}]\pi(\theta)d\theta$$

The prior $\pi(\theta)$ places different emphasis for different values of θ based on our prior belief/knowedge.

Bayes Decision Rules

- Bayes principle: a decision rule is *preferable* to another if it has smaller Bayes risk (depends on the prior $\pi(\theta)$!).

Definition (Bayes Decision Rule)

Let $\mathcal{D}$ be a class of decision rules for an experiment $(\{F_\theta\}_{\theta\in\Theta}, \mathcal{L})$ and let $\pi(\cdot)$ be a probability density (frequency) on Θ. If $\delta \in \mathcal{D}$ is such that

$$r(\pi,\delta) \leq r(\pi,\delta') \quad \forall\ \delta' \in \mathcal{D},$$

then δ is called a *Bayes decision rule* with respect to π.

- The minimax principle aims to minimize the maximum risk.
- The Bayes principle aims to minimize the average risk
- Sometime no Bayes rule exists because the infimum may not be attained for any $\delta \in \mathcal{D}$. However in such cases $\forall \epsilon > 0\ \exists \delta_\epsilon \in \mathcal{D}$: $r(\pi,\delta_\epsilon) < \inf_{\delta\in\mathcal{D}} r(\pi,\delta) + \varepsilon$.

Admissibility of Bayes Rules

Rule of thumb: Bayes rules are nearly always admissible.

Theorem (Discrete Case Admissibility)

Assume that $\Theta = \{\theta_1, ..., \theta_t\}$ is a finite space and that the prior $\pi(\theta_i) > 0$, $i = 1, ..., t$. Then a Bayes rule with respect to π is admissible.

Proof.

Let δ be a Bayes rule, and suppose that κ strictly dominates δ. Then

$$\begin{aligned}
R(\theta_j, \kappa) &\leq R(\theta_j, \delta), \quad \forall j \\
R(\theta_j, \kappa)\pi(\theta_j) &\leq R(\theta_j, \delta)\pi(\theta_j), \quad \forall \theta \in \Theta \\
\sum_j R(\theta_j, \kappa)\pi(\theta_j) &< \sum_j R(\theta, \delta)\pi(\theta_j)
\end{aligned}$$

which is a contradiction (strict inequality follows by strict domination and the fact that $\pi(\theta_j)$ is always positive). □

Admissibility of Bayes Rules

Theorem (Uniqueness and Admissibility)

If a Bayes rule is unique, it is admissible.

Proof.

Suppose that δ is a unique Bayes rule and assume that κ strictly dominates it. Then,

$$\int_{\Theta} R(\theta, \kappa)\pi(\theta)d\theta \leq \int_{\Theta} R(\theta, \delta)\pi(\theta)d\theta.$$

as a result of strict domination and by $\pi(\theta)$ being non-negative. This implies that κ either improves upon δ, or κ is a Bayes rule. Either possibility contradicts our assumption. $\square$

Admissibility of Bayes Rules

Theorem (Continuous Case Admissibility)

Let $\Theta \subset \mathbb{R}^d$. Assume that the risk functions $R(\theta, \delta)$ are continuous in θ for all decision rules $\delta \in \mathcal{D}$. Suppose that π places positive mass on any open subset of Θ. Then a Bayes rule with respect to π is admissible.

Proof.

Let κ be a decision rule that strictly dominates δ. Let Θ_0 be the set on which $R(\theta, \kappa) < R(\theta, \delta)$. Given a $\theta_0 \in \Theta_0$, we have $R(\theta_0, \kappa) < R(\theta_0, \delta)$. By continuity, there must exist an $\epsilon > 0$ such that $R(\theta, \kappa) < R(\theta, \delta)$ for all theta satisfying $\|\theta - \theta_0\| < \epsilon$. It follows that Θ_0 is open and hence, by our assumption, $\pi[\Theta_0] > 0$. Therefore, it must be that

$$\int_{\Theta_0} R(\theta, \kappa)\pi(\theta)d\theta < \int_{\Theta_0} R(\theta, \delta)\pi(\theta)d\theta$$

Admissibility of Bayes Rules

Observe now that

$$
\begin{aligned}
r(\pi,\kappa) &= \int_{\Theta} R(\theta,\kappa)\pi(\theta)d\theta \\
&= \int_{\Theta_0} R(\theta,\kappa)\pi(\theta)d\theta + \int_{\Theta_0^c} R(\theta,\kappa)\pi(\theta)d\theta \\
&< \int_{\Theta_0} R(\theta,\delta)\pi(\theta)d\theta + \int_{\Theta_0^c} R(\theta,\delta)\pi(\theta)d\theta \\
&= r(\pi,\delta),
\end{aligned}
$$

since $\int_{\Theta_0^c} R(\theta,\kappa)\pi(\theta)d\theta \leq \int_{\Theta_0^c} R(\theta,\delta)\pi(\theta)d\theta$, while we have strict inequality on Θ_0, contradicting our assumption that δ is a Bayes rule. $\square$

- The continuity assumption and the assumption on π ensure that Θ_0 is not an isolated set, and has positive measure, so that it "contributes" to the integral.

Randomised Rules

Randomised Decision Rules

Given

- decision rules $\delta_1, ..., \delta_k$
- probabilities $\pi_i \geq 0$, $\sum_{i=1}^{k} p_i = 1$

we may define a new decision rule

$$\delta_* = \sum_{i=1}^{k} p_i \delta_i$$

called a *randomised decision rule*. Interpretation:

> Given data $\mathbf{X}$, choose a rule δ_i with probability p_i independently of $\mathbf{X}$. If δ_j is the outcome ($1 \leq j \leq k$), then take action $\delta_j(\mathbf{X})$.

$\rightarrow$ Risk of δ_* is average risk: $R(\theta, \delta_*) = \sum_{i=1}^{k} p_i R(\theta, \delta_i)$

- Appears artificial but often minimax rules are randomised
- Examples of randomised rules with $\sup_\theta R(\theta, \delta_*) < \sup_\theta R(\theta, \delta_i) \forall i$

Summary

Decision theory gives us a tool to compare different estimators / statistical procedures inside parametric models:

In order to use Decision Theory, we have to choose an appropriate loss function.

Comparing risk function is hard because there is no canonical ordering on positive functions ! We saw three possibilities:

- Admissibility : corresponding to a partial order.
- Minimax : ordering risk functions according to their maximum.
- Bayes' rules : corresponding to a weighting of the different θ.

Amazingly, Bayes' rules and admissible rules have a very close relationship.

We presented randomized decisions which might appear silly but are useful for minimaxity.