Lecture 3: Statistical Decision Theory II

25 January 2019

Frequentist Risk

Definition 1

A **decision rule** $\delta(x)$ is a function from Ω into $\mathscr{A}$. Given a particular realization $X = x$, $\delta(x)$ is the action that will be taken. Two decision rules, δ_1 and δ_1, are said to be equivalent if $P_\theta\left(\delta_1(X) = \delta_2(X)\right) = 1$ for all θ.

- In the absence of data, a decision rule is also called an action.

Example 1: Drug company

Example 1

- Let's consider the situation of a drug company that has to decide whether or not to market a new pain reliever.
- Suppose that the factor affecting its decision is the proportion of the market denoted by θ the drug will capture.
- The value of θ is unknown and needs to be estimated.
- The action taken is simply the choice of a number as an estimate for θ.
- Hence, the standard decision rule for estimating θ is $\delta(x) = \frac{x}{n}$

Example 2: Radio company

Example 2

- Let's consider a company that produces radios.
- It receives a shipment of transistors and randomly selects a sample of n transistors from the shipment for testing.
- Based on the number of defective transistors X in the sample, the shipment will be accepted or rejected.
- Hence, there are two possible actions:

  a_1 – accept the shipment

  a_2 – reject the shipment
- The decision rule is then:

$$\delta(x) = \begin{cases} a_1 & \text{if } \frac{x}{n} \leq 0.05 \\ a_2 & \text{if } \frac{x}{n} > 0.05 \end{cases}$$

The risk function of a decision rule

Definition 2

The **risk function** of a decision rule $\delta(x)$ is defined by:

$$R(\theta, \delta) = E_\theta^X[L(\theta, \delta(X))] = \int_\Omega L(\theta, \delta(x)) dF^X(x|\theta)$$

The risk function of a decision rule

- It is natural to use a decision rule $\delta(X)$ which has smallest $R(\theta, \delta)$.
- However, in contrast to the *Bayesian expected loss*, the risk is a function of θ, and hence it is not a single number.
- Since θ is unknown, the meaning of "smallest" is not clearly defined.

- Nevertheless, there exists a partial ordering of decision rules which will allow to define a "good" decision rule.

Definition 3

A decision rule δ_1, is *R-better* than a decision rule δ_2, if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all $\theta \in \Theta$, with strict inequality for some θ. A decision rule δ_1, is *R-equivalent* to a decision rule δ_2, if $R(\theta, \delta_1) = R(\theta, \delta_2)$ for all $\theta \in \Theta$.

Definition 4

A decision rule δ is said to be **admissible** if there does not exist *R-better* decision rule. A decision rule δ is **inadmissible** if there does exist an *R-better* decision rule.

- From Definition 4 it is clear that an **inadmissible** decision rule should not be used, because we can always find decision rule with smaller risk.

- However, the class of **admissible** decision rules for a given decision problem can be large.

- This means that there will be **admissible** rules with risk functions $R(\theta, \delta)$ that are "better" in some regions of the parameter space Θ, and "worse" in others, i.e. risk functions cross.

Example 3: Risk functions that cross

Example 3

Consider a random variable $X \sim N(\theta, 1)$. It is desired to estimate the unknown parameter θ under loss function $L(\theta, a) = (\theta - a)^2$. Let the decision rule be as follows: $\delta_c(x) = cx$.

The risk function is then as follows:

$$\begin{aligned} R(\theta, \delta_c) &= E_\theta^X L(\theta, \delta_c(X)) = E_\theta^X (\theta - cX)^2 \\ &= E_\theta^X (c[\theta - X] + [1 - c]\theta)^2 \\ &= c^2 E_\theta^X [\theta - X]^2 + 2c(1-c)\theta E_\theta^X [\theta - X] + (1-c)^2\theta^2 \\ &= c^2 + (1-c)^2\theta^2 \end{aligned}$$

Example 3: Risk functions that cross

Let $c = 1$, in which case $\delta_1 = x$, and the risk function is $R(\theta, \delta_1) = 1$

Clearly the decision rule δ_1 is *R-better* than δ_c for values $c > 1$.

$$R(\theta, \delta_1) = 1 < c^2 + (1-c)^2\theta^2 = R(\theta, \delta_c)$$

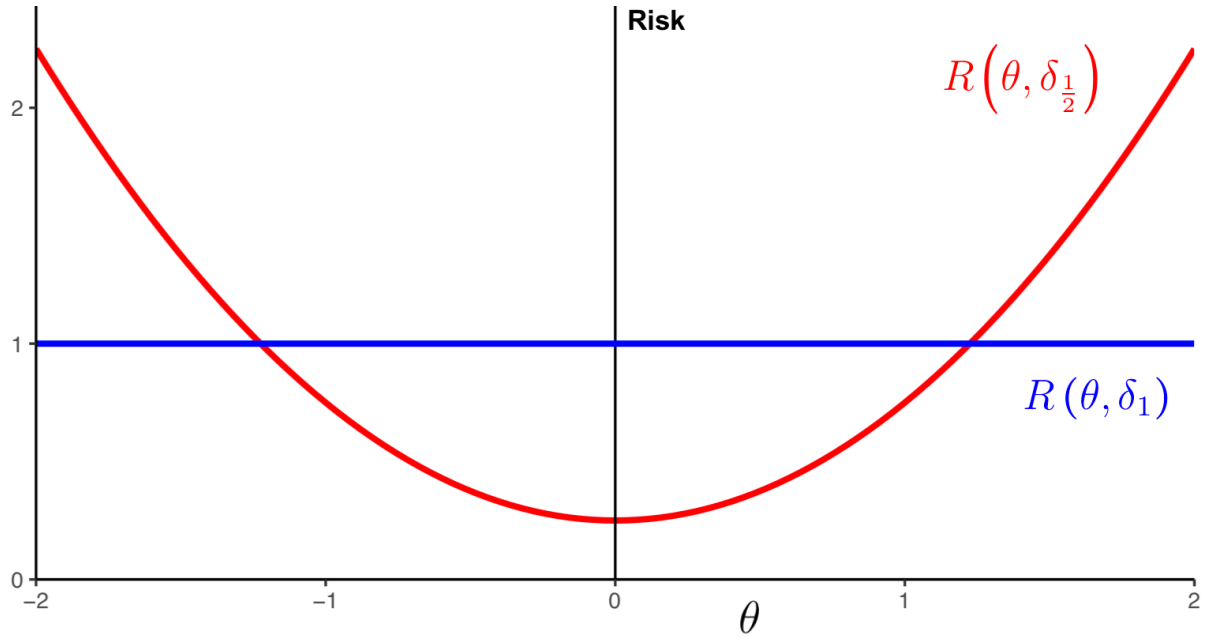
The decision rules $\{\delta_c\}$, are then said to be **inadmissible** for $c > 1$.

Example 3: Risk functions that cross

The decision rules are **non-comparable** for $0 \leq c \leq 1$, because the risk functions associated with them cross.

Let's consider: $c = 1$, then $R(\theta, \delta_1) = 1$

$c = 0.5$, then $R\left(\theta, \delta_{\frac{1}{2}}\right) = 0.5^2 + 0.5^2\theta^2$



Example 3: Risk functions that cross

- According to Definition 4, both decision rules δ_1 and $\delta_{\frac{1}{2}}$ are **admissible** since there are no *R-better* decision rules for $0 \leq c \leq 1$.
- In fact, all decision rules $\{\delta_c\}$ are **admissible** for $0 \leq c \leq 1$.
- This also includes a nonsensical decision rule $\delta_0 = 0$ with risk function $R(\theta, \delta_0) = \theta^2$, which estimates θ to be 0 completely ignoring the observed sample information $\mathbf{x}$.
- Hence, although for a decision rule the ***admissibility*** may be a desirable property, it does not guarantee that the decision rule will be reasonable.

Example 4: Risk functions in the absence of data

Example 4

- From Definition 2, we can see that for a no-data decision problem $R(\theta, \delta) = L(\theta, \delta)$.
- Consider the following loss matrix.

	a_1	a_2	a_3
θ_1	1	3	4
θ_2	-1	5	5
θ_3	0	-1	-1

Question 1: Which decision rules (actions) are **inadmissible**?

Question 2: Which decision rules (actions) are non-comparable?

Question 3: Which decision rules (actions) are **admissible**?

Example 4: Risk functions in the absence of data

Example 4

	a_1	a_2	a_3
θ_1	1	3	4
θ_2	-1	5	5
θ_3	0	-1	-1

Answer 1: The action a_2 is *R-better* than a_3 because $L(\theta_i, a_2) < L(\theta_i, a_3)$ for all i, but not a_1. Hence, only a_3 is **inadmissible**.

Answer 2: Actions a_1 and a_2 are non-comparable because $L(\theta_1, a_1) = 1 < 3 = L(\theta_1, a_2)$, and $L(\theta_3, a_1) = 0 >$ -1 $= L(\theta_3, a_2)$.

Answer 3: From Definition 4 it is clear that actions a_1 and a_2 are **admissible** since there are no *R-better* actions.

The Bayes risk of a decision rule

Definition 5

The Bayes risk of a decision rule δ, with respect to a prior distribution π on Θ, is defined as:

$$r(\pi, \delta) = E^{\pi}[R(\theta, \delta)]$$

Example 5: The Bayes risk of a decision rule

Example 5

Let the prior distribution $\pi(\theta)$ be $N(0, \tau^2)$. Then the Bayes risk of a decision rule δ_c is:

$$\begin{aligned} r(\pi, \delta_c) &= E^{\pi}\left[R(\theta, \delta_c)\right] \\ &= E^{\pi}\left[c^2 + (1-c)^2\theta^2\right] \\ &= c^2 + (1-c)^2 E^{\pi}\left[\theta^2\right] \\ &= c^2 + (1-c)^2\tau^2 \end{aligned}$$

Randomized decision rules

- So far, we have considered **deterministic** decision rules.
- That is, given a particular realization $X = x$, a **deterministic** decision rule $\delta(x)$ is a function from Ω into $\mathscr{A}$.
- However, there will be situations when decision will have to be taken in a randomised manner.
- These situations may arise in the presence of an intelligent competitor.

- A generalization of this concept is a randomized decision rule.

Definition 6

A randomized decision rule $\delta^*(x, \cdot)$ is a probability distribution on $\mathscr{A}$. That is, given that $X = x$ is observed, $\delta^*(x, A)$ is the probability that an action in $A \subseteq \mathscr{A}$ will be chosen.

- In the absence of data, a randomized decision rule is also called a randomized action, which is denoted as $\delta^*(\cdot)$.
- A randomized action is also a probability distribution on $\mathscr{A}$.

Example 6: Matching Pennies

Example 6

Consider the following game called "matching pennies".

- In this game, two players **A** and **B** uncover a coin simultaneously.
- If the two coins match, i.e. both heads or both tails, then player **A** wins £1 from his opponent.
- If the coins do not match, then player **B** wins £1 from his opponent.

Example 6: Matching Pennies

- The actions which are available to player **A** are:

  a_1 – choose heads
  a_2 – choose tails

- The possible states of nature are:

  θ_1 – the opponent's coin is a head
  θ_2 – the opponent's coin is a tail

- The loss matrix in this game is:

	a_1	a_2
θ_1	-1	1
θ_2	1	-1

Example 6: Matching Pennies

	a_1	a_2
θ_1	-1	1
θ_2	1	-1

- Here, both actions a_1, and a_2, are **admissible**.
- However, if the game is to be played many times, it would be a very poor idea for player **A** to always choose a_1 or a_2.
- This is because player **B** would quickly realize this, and could then develop a winning strategy.
- Similarly, any deterministic choice of a_1, and a_2 could be discerned by an intelligent opponent.

Example 6: Matching Pennies

	a_1	a_2
θ_1	-1	1
θ_2	1	-1

- To prevent loosing, choose a_1 and a_2 by some random mechanism.
- For example, choose a_1 with probability p, and a_2 with probability $(1-p)$ respectively.

- **Non-randomized** decision rules can be considered as a special case of **randomized** rules.
- It can be shown that they correspond to the **randomized** rules which choose a specific action with probability one for each x.
- Indeed if $\delta(x)$ is a non-randomized decision rule, let $\langle\delta\rangle$ denote the equivalent randomized rule given by:

$$\langle\delta\rangle(x, A) = I_A(\delta(x)) = \begin{cases} 1 & \text{if } \delta(x) \in A \\ 0 & \text{if } \delta(x) \notin A \end{cases}$$

Example 6: Matching Pennies (continued)

Example 6 (continued)

- The randomized action (no-data problem) discussed in the "matching pennies" example is defined by:

$$\delta^*(a_1) = p$$
$$\delta^*(a_2) = (1-p)$$

- This can also be expressed in a more convenient way using the previous notation of $\langle \delta \rangle$:

$$\delta^* = p\langle a_1 \rangle + (1-p)\langle a_2 \rangle$$

- Similar to **non-randomized** decision rules, there are costs associated with **randomized** decision rules.

Question: How can we define the loss function associated with a particular **randomized** rule?

- The natural way to define the loss function is in terms of expected loss.

Definition 7

The loss function $L(\theta, \delta^*(x))$ of the randomized rule $\delta^*(x, \cdot)$ is:

$$L(\theta, \delta^*(x)) = E^{\delta^*(x,\cdot)}[L(\theta, a)]$$

- Note that the expectation is taken over a.

- Similarly, we can define the **risk function** $R(\theta, \delta^*)$ of a randomized decision rule $\delta^*(x, \cdot)$ in terms of expected loss.

Definition 8

The **risk function** $R(\theta, \delta^*)$ of a randomized decision rule $\delta^*(x, \cdot)$ with the loss function $L(\theta, \delta^*(x))$ is:

$$R(\theta, \delta^*) = E_\theta^X [L(\theta, \delta^*(X))] = \int_\Omega L(\theta, \delta^*(x)) dF^X(x|\theta)$$

- Again, we can see that for a no-data decision problem $R(\theta, \delta^*) = L(\theta, \delta^*)$.

Example 6: Matching Pennies (continued)

Example 6 (continued)

- Because this is a no-data problem, the risk $R(\theta, \delta^*)$ is just the loss $L(\theta, \delta^*)$.

$$
\begin{aligned}
R(\theta, \delta^*) = L(\theta, \delta^*) &= E^{\delta^*}[L(\theta, a)] \\
&= \delta^*(a_1) L(\theta, a_1) + \delta^*(a_2) L(\theta, a_2) \\
&= p L(\theta, a_1) + (1-p) L(\theta, a_2) \\
&= \begin{cases} -p + (1-p) = 1 - 2p & \text{if } \theta = \theta_1 \\ p - (1-p) = 2p - 1 & \text{if } \theta = \theta_2 \end{cases}
\end{aligned}
$$

Example 6: Matching Pennies (continued)

Example 6 (continued)

$$R(\theta, \delta^*) = \begin{cases} -p + (1-p) = 1 - 2p & \text{if } \theta = \theta_1 \\ p - (1-p) = 2p - 1 & \text{if } \theta = \theta_2 \end{cases}$$

- It is evident that if player **A** chooses a_1 with probability $p = \frac{1}{2}$, and a_2 with $(1-p) = \frac{1}{2}$, the loss is zero no matter what player **B** does.
- Therefore, the randomized rule $\delta^*(\cdot)$ with $p = \frac{1}{2}$ guarantees an expected loss of zero.

Definition 9

Let $\mathscr{D}^*$ be the set of all randomized decision rules δ^* for which $R(\theta, \delta^*) < \infty$ for all θ.

A decision rule will be said to be **admissible** if there exists no *R-better* randomized decision rule in $\mathscr{D}^*$.

Is the use of randomized decision rule reasonable?

- Often decision problems do not involve an intelligent opponent.
- For example, when deciding whether to prescribe a particular drug to a patient, no intelligent opponent is involved.
- In these situations there seems to be no good reason to use randomized decision rules.
- On the contrary, our intuition would argue against it.

Is the use of randomized decision rule reasonable?

- Whenever possible, each possible action has to be evaluated in order to find the optimal action.
- If there is only one optimal action, then randomizing is of limited use.
- If there are 2 or more optimal actions, one could potentially choose at random, although the usefulness of doing so is questionable.
- Often, leaving the final choice of an action to be decided by some random mechanism just seems illogical.
- Therefore, the actual use of a randomized rule will rarely be recommended.

Frequentist Decision Principles

- We have seen that using risk functions to select a decision rule does not always produce a clear final choice.
- It may well be that there are many **admissible** decision rules.
- To overcome this limitation, we must introduce an additional principle in order to select a specific decision rule.
- For example, in classical statistics there are a number of such principles for developing statistical procedures:
  - least squares principles
  - the maximum likelihood
  - unbiasedness
  - minimum variance

Frequentist Decision Principles

- In decision theory, the three most important principles that can be used are:

  I. the Bayes risk principle
  II. the minimax principle
  III. the invariance principle

I. The Bayes Risk Principle

- We have already seen in Definition 5 that, we could obtain a (real) number associated with a particular decision rule instead of a risk function.
- This approach involved using the prior distribution π, and computing the Bayes risk of a decision rule δ:

$$r(\pi, \delta) = E^{\pi} R(\theta, \delta)$$

- Since this is a (real) number, we can simply find a decision rule which will minimize it.

I. The Bayes Risk Principle

- Thus, we can define the first frequentist decision principle.

Definition 10

A decision rule δ_1 is preferred to a rule δ_2 if:

$$r(\pi, \delta_1) < r(\pi, \delta_2)$$

A decision rule is said to be optimal if it minimizes $r(\pi, \delta)$. This decision rule is called a Bayes rule, and will be denoted δ^π .

The quantity $r(\pi) = r(\pi, \delta^\pi)$ is then called the Bayes risk for π.

I. The Bayes Risk Principle: Example 7

Example 7

- Consider previous Example 5.
- We have seen that the Bayes risk $r(\pi, \delta_c)$ of a decision rule δ_c when a prior distribution π is $N(0, \tau^2)$ was $c^2 + (1-c)^2\tau^2$.
- Minimizing with respect to c, we can establish that $c_0 = \frac{\tau^2}{(1+\tau^2)}$ is the optimal value.
- Hence, δ_{c_0} has the smallest Bayes risk among all estimators of the form δ_c.

- Therefore, δ_{c_0} is the Bayes rule, which has the following form:

$$\begin{aligned} r(\pi) &= r\left(\pi, \delta_{c_0}\right) \\ &= c_0^2 + \left(1 - c_0\right)^2 \tau^2 \\ &= \left(\frac{\tau^2}{1+\tau^2}\right)^2 + \left(\frac{1}{1+\tau^2}\right)^2 \tau^2 \\ &= \frac{\tau^2}{1+\tau^2} \end{aligned}$$

II. The Minimax Principle

- Analysis of decision problems using the **minimax principle** generally requires consideration of randomized decision rules.
- Hence, if $\delta^* \in \mathscr{D}^*$ is the randomized decision rule, then the worst case possible using this decision rule δ^* is:

$$\sup_{\theta \in \Theta} R(\theta, \delta^*)$$

- In order to protect from the worst case scenario, one should use the **minimax principle**.

Definition 11 (The Minimax Principle)

A decision rule δ_1^* is preferred to a rule δ_2^* if:

$$\sup_{\theta} R(\theta, \delta_1^*) < \sup_{\theta} R(\theta, \delta_2^*)$$

II. The Minimax Principle

Definition 12

A decision rule δ^{*M} is a **minimax decision rule** if it minimizes $\sup_{\theta \in \Theta} R(\theta, \delta^*)$ among all randomized rules in $\mathscr{D}^*$, that is, if:

$$\sup_{\theta \in \Theta} R\left(\theta, \delta^{*M}\right) = \underbrace{\inf_{\delta^* \in \mathscr{D}^*} \sup_{\theta \in \Theta} R\left(\theta, \delta^*\right)}_{\text{minimax value}}$$

- For a no-data decision problem, the **minimax decision rule** is simply called the **minimax action**.

II. The Minimax Principle: Example 8

Example 8

- There may be situations where it is of interest to determine the best **non-randomized** rule according to the **minimax principle**.
- When such a best rule exists, it will be called the **minimax non-randomized rule**
- In a no-data decision problem, the **minimax non-randomized** rule is simply called **minimax non-randomized action**.
- Consider previous Example 3.

$$r\left(\pi, \delta_c\right) = c^2 + (1-c)^2\theta^2$$

II. The Minimax Principle: Example 8

Example 8

- For the decision rules δ_c:

$$\sup_{\theta} R(\theta, \delta_c) = \sup_{\theta} \left[c^2 + (1-c)^2\theta^2 \right] = \begin{cases} 1 & \text{if } c = 1 \\ \infty & \text{if } c \neq 1 \end{cases}$$

- Therefore, according to the **minimax principle**, δ_1 is best among the decision rules δ_c.
- Hence, δ_1 is a minimax rule and that 1 is the minimax value for the decision problem.
- Recall that the decision rule is $\delta_1(x) = x$, and if it is used, one can ensure that the risk is no worse than 1
- It can also be noted that the minimax rule and the Bayes rule are different.

Example 6: Matching Pennies (continued)

Example 6 (continued)

- Consider again the randomized action δ^* (no-data problem) discussed in the "matching pennies" example:

$$\delta^* = p\langle a_1 \rangle + (1-p)\langle a_2 \rangle$$

  where a_1 is to be selected with probability p, and a_2 with $(1-p)$.
- The risk (which is also the loss) was shown to be:

$$R(\theta, \delta^*) = L(\theta, \delta^*) = \begin{cases} -p + (1-p) = 1-2p & \text{if } \theta = \theta_1 \\ p - (1-p) = 2p-1 & \text{if } \theta = \theta_2 \end{cases}$$

Example 6: Matching Pennies (continued)

Example 6 (continued)

$$R(\theta, \delta^*) = L(\theta, \delta^*) = \begin{cases} -p + (1-p) = 1 - 2p & \text{if } \theta = \theta_1 \\ p - (1-p) = 2p - 1 & \text{if } \theta = \theta_2 \end{cases}$$

- Therefore:

$$\sup_{\theta} R\left(\theta, \delta_p^*\right) = \max\{1 - 2p, 2p - 1\}$$

- Plotting these two functions, it is clear that the minimum value of $\max\{1 - 2p, 2p - 1\}$ is 0, which occurs at $p = \frac{1}{2}$.
- Therefore, $\delta^*_{1/2}$ is the **minimax action**, and 0 is the minimax value for the problem.

III. The invariance principle

- The invariance principle states that if two problems have identical structures (e.g. the same sample space, parameter space, probability distribution, and loss function), then the same decision rule should be used in each problem.

- This principle is employed in situations where the problem has been transformed (e.g. change of scale in the unit of measurement) but retained the identical structure.