Lecture notes on statistical decision theory
Econ 2110, fall 2013

Maximilian Kasy

March 10, 2014

These lecture notes are roughly based on

> Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Verlag, chapter 2.

(Robert is very passionately Bayesian - read critically!)

You might also want to look at

> Casella, G. and Berger, R. (2001). *Statistical inference*. Duxbury Press, chapter 7.3.4.

What I would like you to take away from this section of class:

1. A general framework to think about what makes a "good" estimator, test, etc.
2. How the foundations of statistics relate to those of microeconomic theory.
3. In what sense the set of Bayesian estimators contains most "reasonable" estimators.

1 Basic definitions

A general statistical decision problem has the following components:

- Observed data X
- A statistical decision a
- A state of the world θ
- A loss function $L(a, \theta)$ (the negative of utility)
- A statistical model $f(X|\theta)$
- A decision function $a = \delta(X)$

The underlying state of the world θ influences the distribution of the observation X. The decision maker picks a decision a after observing X. She wants to pick a decision that minimizes loss $L(a, \theta)$, for the unknown state of the world θ. X is useful because it reveals some information about θ, at least if $f(X|\theta)$ does depend on θ. The problem of statistical decision theory is to find decision functions δ which are good in the sense of making loss small.

Examples of loss functions:

1. In **estimation**, we want to find an a which is close to some function μ of θ, such as for instance $\mu(\theta) = E[X]$. Loss is larger if the difference between our estimate and the true value is larger. A commonly used loss in this case is the **squared error**,

$$L(a, \theta) = (a - \mu(\theta))^2. \tag{1}$$

2. An alternative to squared error loss is **absolute error** loss,

$$L(a, \theta) = |a - \mu(\theta)|. \tag{2}$$

3. In **testing**, we want to decide whether some statement $H_0 : \theta \in \Theta_0$ about the state of the world is true. We might evaluate such a decision $a \in \{0, 1\}$ based on the loss

$$L(a, \theta) = \begin{cases} 1 & \text{if } a = 1,\ \theta \in \Theta_0 \\ c & \text{if } a = 0,\ \theta \notin \Theta_0 \\ 0 & \text{else.} \end{cases} \tag{3}$$

Risk function:
An important intermediate object in evaluating a decision function δ is the risk function R. It measures the expected loss for a given true state of the world θ,

$$R(\delta, \theta) = E_\theta[L(\delta(X), \theta)]. \tag{4}$$

Note the dependence of R on the state of the world - a decision function δ might be good for some values of θ, but bad for other values. The next section will discuss how we might deal with this trade-off.

Examples of risk functions:

1. We observe $X \in \{0, \dots, k\}$ multinomially distributed with $P(X = x) = f(x)$. We want to estimate $f(0)$ and loss is squared error loss. Taking the estimator $\delta(X) = \mathbf{1}(X = 0)$, we get the risk function

$$R(\delta, f) = E[(\delta(X) - f(0))^2] = \mathrm{Var}(\delta(X)) = f(0)(1 - f(0)). \tag{5}$$

2. We observe $X \sim N(\mu, 1)$, and want to estimate μ.[1] Loss is again squared error loss. Consider the estimator

$$\delta(X) = a + b \cdot X.$$

We get the risk function

$$\begin{aligned} R(\delta, \mu) &= E[(\delta(X) - \mu)^2] = \mathrm{Var}(\delta(X)) + \mathrm{Bias}(\delta(X))^2 \\ &= b^2 \,\mathrm{Var}(X) + (a + bE[X] - E[X])^2 = b^2 + (a + (b-1)\mu)^2. \end{aligned} \tag{6}$$

Choosing a and b involves a trade-off of bias and variance, and this trade-off depends on μ.

2 Optimality criteria

Throughout it will be helpful for intuition to consider the case where θ can only take two values, and to consider a 2D-graph where each axis corresponds to $R(\delta, \theta)$ for $\theta = \theta_0, \theta_1$.

[1] This example is important because of the asymptotic approximations we will discuss in the next part of class.

The ranking provided by the risk function is multidimensional; it provides a ranking of performance between decision functions for every θ. In order to get a global comparison of their performance, we have to somehow aggregate this ranking into a global ranking. We will now discuss three alternative ways to do this:

1. The partial ordering provided by considering a decision function better relative to another one if it is better for *every* θ,

2. the complete ordering provided if we trade off risk across θ with pre-specified weights, where the weights are given in the form of a prior distribution, and

3. the complete ordering provided if we evaluate a decision function under the worst-case scenario.

Admissibility:
A decision function δ is said to dominate another function δ' if

$$R(\delta, \theta) \leq R(\delta', \theta) \tag{7}$$

for all θ, and

$$R(\delta, \theta) < R(\delta', \theta) \tag{8}$$

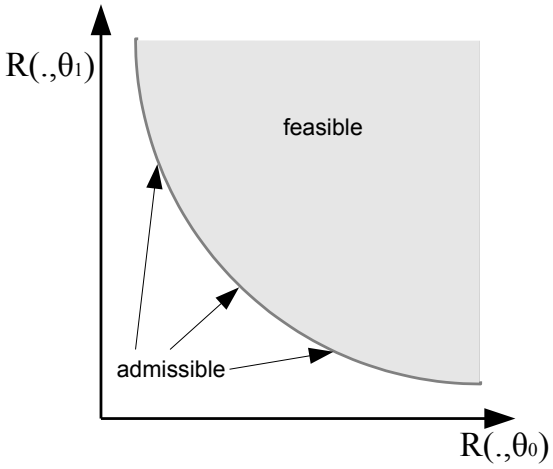
for at least one θ. It seems natural to only consider decisions functions which are not dominated. Such decision functions are called admissible, all other decision functions are inadmissible.
(picture with "Pareto frontier"!)

Note that dominance only generates a partial ordering of decision functions. In general there will be many different admissible decision functions.

Bayes optimality:
An approach which comes natural to economists is to trade off risk across different θ by assigning weights $\pi(\theta)$ to each θ. Such an approach evaluates decision functions based on the integrated risk

$$R(\delta, \pi) = \int R(\delta, \theta)\pi(\theta)d\theta. \tag{9}$$

Figure 1: Feasible and admissible risk functions



In this context, if π can be normalized such that $\int \pi(\theta)d\theta = 1$ and $\pi \geq 0$, then π is called a prior distribution. A Bayes decision function minimizes integrated risk,

$$\delta^* = \underset{\delta}{\operatorname{argmin}} \; R(\delta, \pi). \tag{10}$$

(picture with linear indifference curves!)

If π is a prior distribution, we can define posterior expected loss

$$R(\delta, \pi|X) := \int L(\delta(X), \theta)\pi(\theta|X)d\theta, \tag{11}$$

where $\pi(\theta|X)$ is the posterior distribution

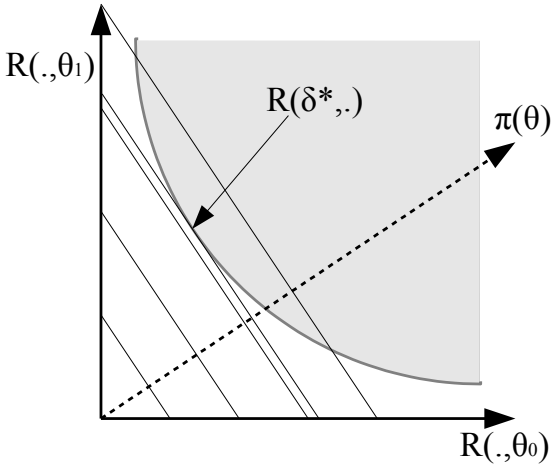
$$\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{m(X)}, \tag{12}$$

and $m(X) = \int f(X|\theta)\pi(\theta)dX$ is the normalizing constant given by the prior likelihood of m.

It is easy to see that any Bayes decision function δ^* can be obtained by minimizing $R(\delta, \pi|X)$ through choice of $\delta(X)$ for every X, since

$$R(\delta, \pi) = \int R(\delta(X), \pi|X)m(X)dX. \tag{13}$$

Figure 2: Bayes optimality



Minimaxity:
An approach which does not rely on the choice of a prior is minimaxity. This approach evaluates decisions functions based on the worst-case risk,

$$\overline{R}(\delta) = \sup_{\theta} R(\delta, \theta). \tag{14}$$

(picture with "Leontieff" indifference curves!)

A minimax decision function, if it exists, solves the problem

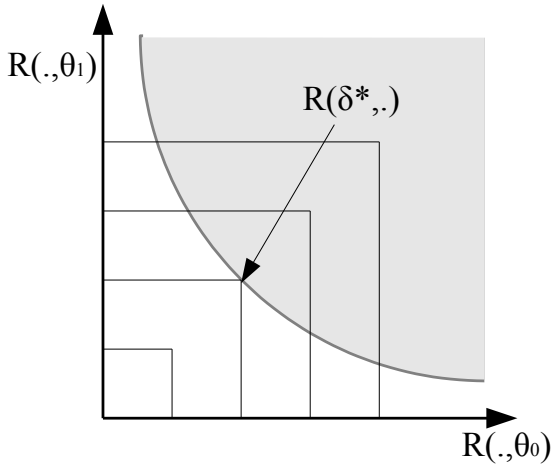
$$\delta^* = \operatorname*{argmin}_{\delta} \ \overline{R}(\delta) = \operatorname*{argmin}_{\delta} \ \sup_{\theta} R(\delta, \theta). \tag{15}$$

Some relationships between these concepts

Much can be said about the relationship between the concepts of admissibility, Bayes optimality, and minimaxity. We will discuss some basic relationships between them in this and the next section.

1. If δ^* is admissible with constant risk, then it is a minimax decision function. Proof:

Figure 3: Minimaxity



Suppose that δ' had smaller minimax risk than δ^*. Then

$$R(\delta', \theta') \leq \sup_\theta R(\delta', \theta) < \sup_\theta R(\delta^*, \theta) = R(\delta^*, \theta'),$$

where we used constant risk in the last equality. But this contradicts admissibility. (picture!)

2. If the prior distribution $\pi(\theta)$ is strictly positive and the Bayes decision function δ^* has finite risk and risk is continuous in θ, then it is admissible. Sketch of proof:
Suppose it is not admissible. Then it is dominated by δ'. But then

$$R(\delta', \pi) = \int R(\delta', \theta)\pi(\theta)d\theta < \int R(\delta^*, \theta)\pi(\theta)d\theta = R(\delta^*, \pi),$$

since $R(\delta', \theta) \leq R(\delta^*, \theta)$ for all θ with strict inequality for some θ. This contradicts δ^* being a Bayes decision function. (picture!)

3. We will prove the reverse of this statement in the next section.

4. The Bayes risk $R(\pi) := \inf_\delta R(\delta, \pi)$ is always smaller than the minimax risk $\overline{R} := \inf_\delta \sup_\theta R(\delta, \theta)$. Proof:

$$R(\pi) = \inf_{\delta} R(\delta, \pi) \leq \sup_{\pi} \inf_{\delta} R(\delta, \pi) \leq \inf_{\delta} \sup_{\pi} R(\delta, \pi) = \inf_{\delta} \sup_{\theta} R(\delta, \theta) = \overline{R}.$$

If there exists a prior π^* such that $\inf_\delta R(\delta, \pi^*) = \sup_\pi \inf_\delta R(\delta, \pi)$, it is called the least favorable distribution.

Analogies to microeconomics

Concluding this section note the correspondence between the concepts discussed here and those of microeconomics.

First, there is an analogy between statistical decision theory and social welfare analysis. The analogy maps different parameter values θ to different people i, and risk $R(., \theta)$ to individuals' utility $u_i(.)$. For this correspondence

1. Our dominance is analogous to Pareto dominance.
2. Admissibility, in particular, is analogous to Pareto efficiency.
3. Bayes risk corresponds to weighted sums of utilities as in standard social welfare functions in public finance. The prior corresponds to welfare weights (distributional preferences).
4. Minimaxity corresponds to Rawlsian inequality aversion.

Second, consider choice under uncertainty, and in particular choice in strategic interactions. For this correspondence

1. Dominance of decision functions corresponds to dominance of strategies in games.
2. Bayes risk corresponds to expected utility (after changing signs), and Bayes optimality corresponds to expected utility maximization.
3. Minimaxity is an extreme form of so called "ambiguity aversion". It is also familiar from early game theory, in particular from the theory of zero-sum games.

3 Two justifications of the Bayesian approach

In the last section we saw that, under some conditions, every Bayes decision function is admissible. An important result states that the reverse also holds true under certain conditions - every admissible decision function is Bayes, or the limit of Bayes decision functions. This can be taken to say that we might restrict our attention in the search of good decision functions to the Bayesian ones. We state a simple version of this type of result.

We call the set of risk functions that correspond to some δ the **risk set**,

$$\mathscr{R} := \{r(.) = R(., \delta) \text{ for some } \delta\}. \tag{16}$$

We will assume convexity of $\mathscr{R}$, which is no big restriction, since we can always randomly "mix" decision functions.
We say a class of decision functions δ is a **complete class** if it contains every admissible decision function δ^*.

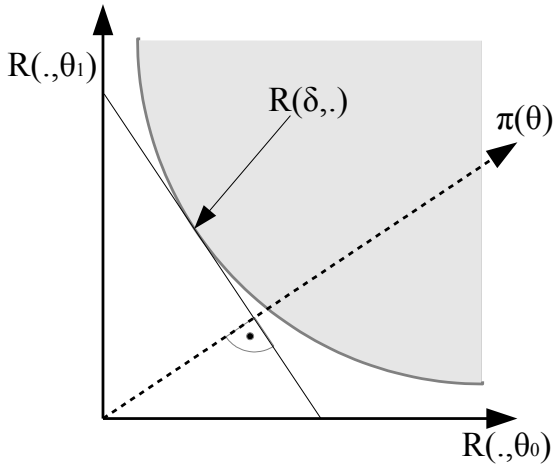
Complete class theorem
Suppose the set Θ of possible values for θ is compact, that the risk set $\mathscr{R}$ is convex, and that all decision functions have continuous risk. Then the Bayes decision functions constitute a complete class in the sense that for every admissible decision function δ^*, there exists a prior distribution π such that δ^* is a Bayes decision function for π.
Proof:
(picture!!) The proof is essentially an application of the separating hyperplane theorem, applied to the space of functions of θ with the inner product

$$\langle f, g\rangle = \int f(\theta)g(\theta)d\theta.$$

(For intuition, focus on the case where Θ is finite and the inner product is given by $\langle f, g\rangle = \sum_\theta f(\theta)g(\theta)$.)
Let δ^* be admissible. Then the corresponding risk function $R(., \delta^*)$ belongs to the lower boundary of the risk set $\mathscr{R}$ by the definition of admissibility. The separating hyperplane theorem (in combination with convexity of $\mathscr{R}$) implies that there exists a function $\tilde{\pi}$ (with finite integral) such that for all δ

$$\langle R(., \delta^*), \tilde{\pi}\rangle \leq \langle R(., \delta), \tilde{\pi}\rangle.$$

Figure 4: Complete class theorem



By admissibility $\tilde{\pi}$ can be chosen such that $\tilde{\pi} \geq 0$, and thus $\pi := \tilde{\pi} / \int \tilde{\pi}$ defines a prior distribution. δ^* minimizes

$$\langle R(., \delta^*), \pi \rangle = R(\delta^*, \pi)$$

among the set of feasible decision functions, and is therefore the optimal Bayesian decision function for the prior π.□

A second general justification for Bayesian decision theory is given by **subjective probability theory**, going back to Savage (1954) and Anscombe and Aumann (1963). This theory is discussed in chapter 6 of

> Mas-Colell, A., Whinston, M., and Green, J. (1995). *Microeconomic theory*. Oxford university press,

and maybe in Econ 2010.

Suppose a decision maker ranks risk functions $R(., \delta)$ by a preference relationship $\succeq$. If this preference relationship is complete, monotonic, and satisfies the independence condition

$$R^1 \succeq R^2 \Leftrightarrow \alpha R^1 + (1 - \alpha) R^3 \succeq \alpha R^2 + (1 - \alpha) R^3 \tag{17}$$

for all R^1, R^2, R^3 and $\alpha \in [0,1]$, then there exists a prior π such that

$$R(.,\delta^1) \succeq R(.,\delta^2) \Leftrightarrow R(\pi,\delta^1) \geq R^2(\pi,\delta^2).$$

Sketch of proof:
Using independence repeatedly, we can show that for all $R^1, R^2, R^3 \in \mathbb{R}^{\mathscr{X}}$, and all $\alpha > 0$,

1. $R^1 \succeq R^2$ iff $\alpha R^1 \succeq \alpha R^2$,
2. $R^1 \succeq R^2$ iff $R^1 + R^3 \succeq R^2 + R^3$,
3. $\{R : R \succeq R^1\} = \{R : R \succeq 0\} + R^1$,
4. $\{R : R \succeq 0\}$ is a convex cone.
5. $\{R : R \succeq 0\}$ is a half space.

The last claim requires completeness. It immediately implies the existence of π. Monotonicity implies that π is not negative.

4 Testing and the Neyman Pearson lemma

A special statistical decision problem is the problem of testing, where we want to decide whether some statement $H_0 : \theta \in \Theta_0$ about the state of the world is true, or whether alternatively $H_1 : \theta \in \Theta_1$ is true. A statistical test is then a decision function $\varphi : X \Rightarrow \{0,1\}$, where $\varphi = 1$ corresponds to rejecting the null hypothesis. We might more generally allow for randomized tests $\varphi : X \Rightarrow [0,1]$ which reject H_0 with probability $\varphi(X)$.

We can make two types of classification errors when we try to distinguish H_0 from H_1 :

1. We can decide that H_1 is true when if fact H_0 is true. This is called the *Type I error*.
2. We can decide that H_0 is true when if fact H_1 is true. This is called the *Type II error*.

Suppose that the data are distributed according to the density (or probability mass function) $f_\theta(x)$. Then the probability of rejecting H_0 given θ is equal to

$$\beta(\theta) = E_\theta[\varphi(X)] = \int \varphi(x) f_\theta(x) dx. \tag{18}$$

Suppose that θ has only two points of support, θ_0 and θ_1. Then

1. The probability of a Type I error is equal to $\beta(\theta_0)$. This is also called the level or significance of the test, and is often denoted α.

2. The probability of a Type II error is equal to $1 - \beta(\theta_1)$. $\beta(\theta_1)$ is called the power of a test, and is often denoted β.

We would like to make the probabilities of both types of errors small, that is to have a small α and a large β. In the case where θ has only two points of support, there is a simple characterization of the solution to this problem. Consider a 2D-graph where the first axis corresponds to $\beta(\theta_0)$ and the second to $1 - \beta(\theta_1)$. This is very similar to the risk functions we drew before.

We can restate the problem as finding a φ^* that solves

$$\max_{\varphi} \beta(\theta_1) \quad \text{s.t. } \beta(\theta_0) = \alpha$$

for a prespecified level α. The solution to this problem is given by the **Neyman-Pearson Lemma**:

$$\varphi^*(x) = \begin{cases} 1 \text{ for } f_1(x) > \lambda f_0(x) \\ \kappa \text{ for } f_1(x) = \lambda f_0(x) \\ 0 \text{ for } f_1(x) < \lambda f_0(x) \end{cases}$$

where λ and κ are chosen such that $\int \varphi^*(x) f_0(x) dx = \alpha$.
Proof: Let $\varphi(x)$ be any other test of level α, i.e. $\int \varphi(x) f_0(x) dx = \alpha$. We need to show that $\int \varphi^*(x) f_1(x) dx \geq \int \varphi(x) f_1(x) dx$.
Note that

$$\int (\varphi^*(x) - \varphi(x))(f_1(x) - \lambda f_0(x)) dx \geq 0$$

since $\varphi^*(x) = 1 \geq \varphi(x)$ for all x such that $f_1(x) - \lambda f_0(x) > 0$ and $\varphi^*(x) = 0 \leq \varphi(x)$ for all x such that $f_1(x) - \lambda f_0(x) < 0$. Therefore, using $\alpha =$

$\int \varphi(x) f_0(x) dx = \int \varphi^*(x) f_0(x) dx$,

$$\begin{aligned} & \int (\varphi^*(x) - \varphi(x))(f_1(x) - \lambda f_0(x)) dx \\ = & \int (\varphi^*(x) - \varphi(x)) f_1(x) dx \\ = & \int \varphi^*(x) f_1(x) dx - \int \varphi(x) f_1(x) dx \geq 0 \end{aligned}$$

as required. The proof in the discrete case is identical with all summations replaced by integrals.