STAT 830
Decision Theory and Bayesian Methods

Example: Decide between 4 modes of transportation to work:

- B = Ride my bike.
- C = Take the car.
- T = Use public transit.
- H = Stay home.

Costs depend on weather: R = Rain or S = Sun.

Ingredients of Decision Problem in the no data case

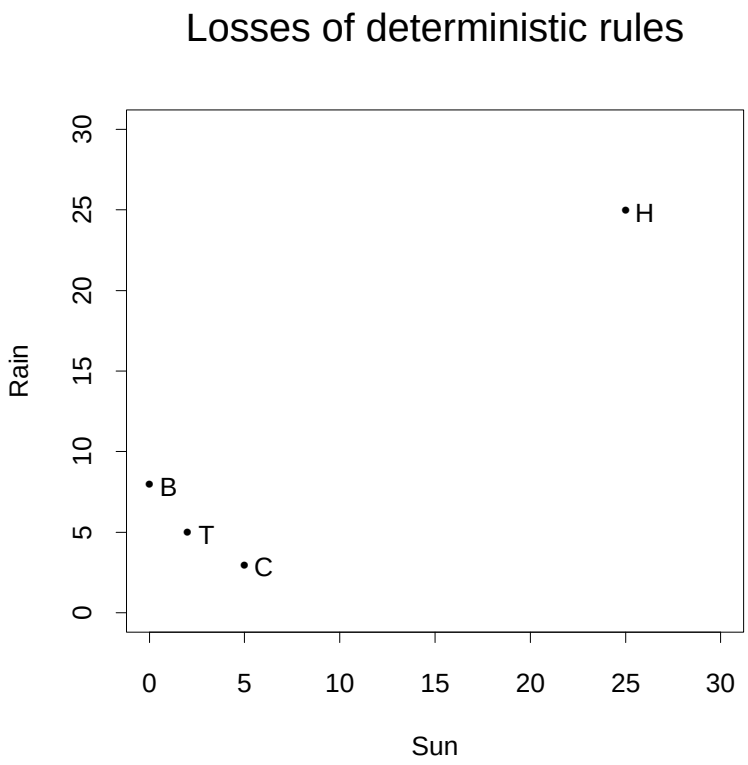
- Decision space $D = \{B, C, T, H\}$ of possible actions.
- Parameter space $\Theta = \{R, S\}$ of possible "states of nature".
- Loss function $L = L(d, \theta)$ loss incurred if do d and θ is true state of nature.

In the example we might use the following table for L:

	C	B	T	H
R	3	8	5	25
S	5	0	2	25

Notice that if it rains I will be glad if I drove. If it is sunny I will be glad if I rode my bike. In any case staying at home is expensive.

In general we study this problem by comparing various functions of θ. In this problem a function of θ has only two values, one for rain and one for sun and we can plot any such function as a point in the plane. We do so to indicate the geometry of the problem before stating the general theory.


Losses of deterministic rules
Rain
Sun
0
5
10
15
20
25
30
B
T
C
H

Statistical Decision Theory

Statistical problems have another ingredient, the data. We observe X a random variable taking values in say $\mathcal{X}$. We may make our decision d depend on X. A **decision rule** is a function $\delta(X)$ from $\mathcal{X}$ to D. We will want $L(\delta(X), \theta)$ to be small for all θ. Since X is random we quantify this by averaging over X and compare procedures δ in terms of the **risk function**

$$R_\delta(\theta) = E_\theta(L(\delta(X), \theta))$$

To compare two procedures we must compare two functions of θ and pick "the smaller one". But typically the two functions will cross each other and there won't be a unique 'smaller one'.

Example: In estimation theory to estimate a real parameter θ we used $D = \Theta$,

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

and find that the risk of an estimator $\hat{\theta}(X)$ is

$$R_{\hat{\theta}}(\theta) = E[(\hat{\theta} - \theta)^2]$$

which is just the Mean Squared Error of $\hat{\theta}$. We have already seen that there is no unique best estimator in the sense of MSE. How do we compare risk functions in general?

- **Minimax methods** choose δ to minimize the worst case risk:

  $$\sup\{R_\delta(\theta); \theta \in \Theta)\}.$$

  We call δ^* minimax if

  $$\sup_\theta R_{\delta^*}(\theta) = \inf_\delta \sup_\theta R_\delta(\theta)$$

  Usually the sup and inf are achieved and we write max for sup and min for inf. This is the source of "minimax".

- **Bayes methods** choose δ to minimize an average

  $$r_\pi(\delta) = \int R_\delta(\theta)\pi(\theta)d\theta$$

  for a suitable density π. We call π a **prior** density and r the **Bayes** risk of δ for the prior π.

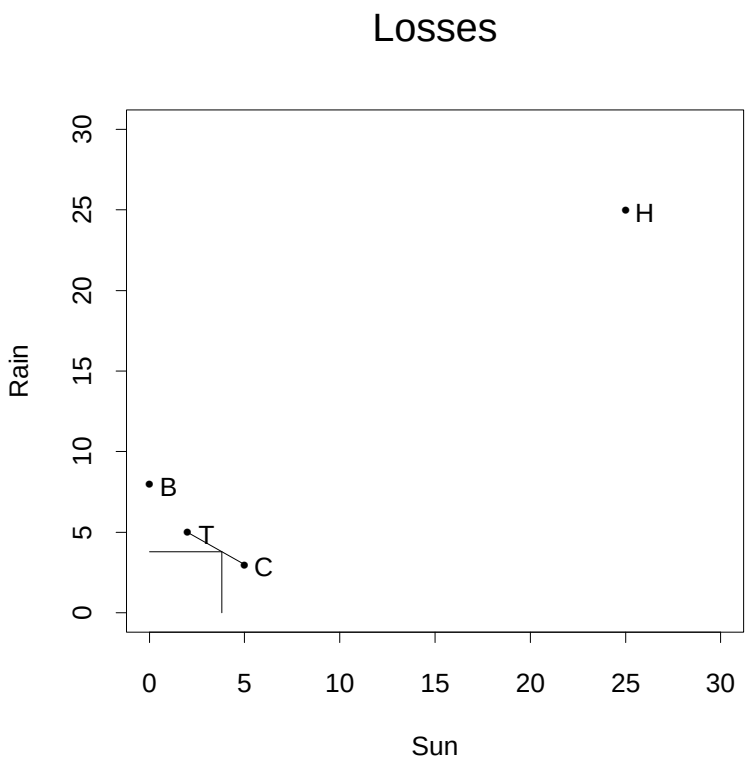
Example: My transportation problem has no data so the only possible (non-randomized) decisions are the four possible actions B, C, T, H. For B and T the worst case is rain. For the other two actions Rain and Sun are equivalent. We have the following table:

	C	B	T	H
R	3	8	5	25
S	5	0	2	25
Maximum	5	8	5	25

To get the smallest maximum: take car, or transit. Thus the minimax action is either to take the car or to take public transit.

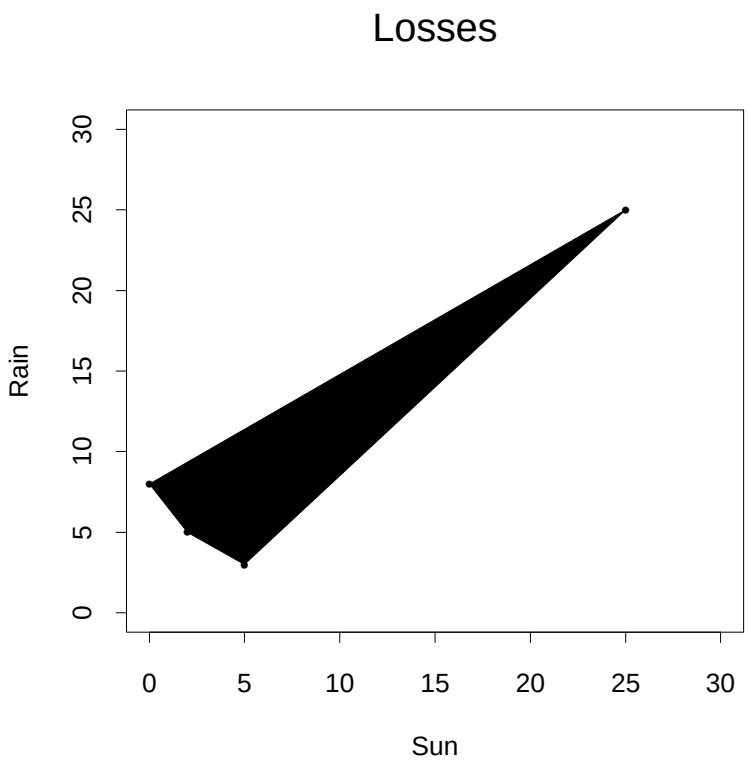
Now imagine I toss a coin with probability λ of getting Heads and take my car if I get Heads, otherwise take transit. The long run average daily loss would be $3\lambda + 5(1-\lambda)$ when it rains and $5\lambda + 2(1-\lambda)$ when it is Sunny. Call this procedure d_λ; add it to graph for each value of λ. Varying λ from 0 to 1 gives a straight line running from $(3, 5)$ to $(5, 2)$. The two losses are equal when $\lambda = 3/5$. For smaller λ worst case risk is for sun; for larger λ worst case risk is for rain.

Added to graph: loss functions for each d_λ, (straight line) and set of (x, y) pairs for which $\min(x, y) = 3.8$ — worst case risk for d_λ when $\lambda = 3/5$.


Losses
Rain
Sun
0
5
10
15
20
25
30
H
B
T
C

The figure then shows that $d_{3/5}$ is actually the minimax procedure when randomized procedures are permitted.

In general we might consider using a 4 sided coin where we took action B with probability λ_B, C with probability λ_C and so on. The loss function of such a procedure is a convex combination of the losses of the four basic procedures making the set of risks achievable with the aid of randomization look like the following:


Losses
Rain
0
5
10
15
20
25
30
Sun

Randomization in decision problems permits the assumption that the set of possible risk functions is convex — an important technical conclusion used to prove many basic decision theory results.

The graph shows that many points in the picture correspond to bad decision procedures. Rain or shine not taking my car to work has a lower loss than staying home; the decision to stay home is *inadmissible*.

Definition: A decision rule δ is **inadmissible** if there is a rule δ^* such that

$$R_{\delta^*}(\theta) \leq R_\delta(\theta)$$

for all θ and there is at least one value of θ where the inequality is strict. A rule which is not inadmissible is called **admissible**.

Admissible procedures have risks on lower left of graphs, i.e., lines connecting B to T and T to C are the admissible procedures.

Connection between Bayes procedures and admissible procedures

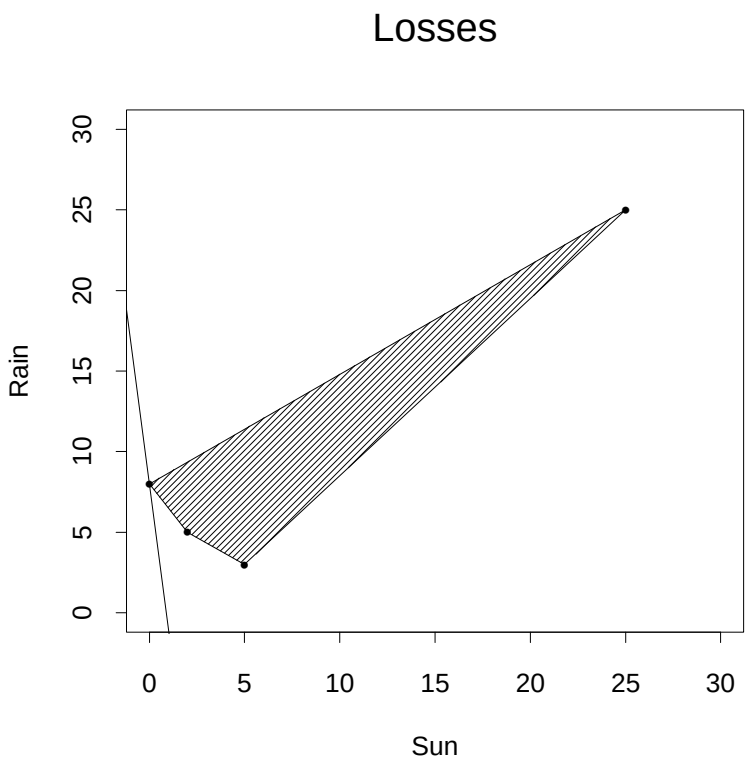
A prior distribution in the example is specified by two probabilities, π_S and π_R which add up to 1. If $L = (L_S, L_R)$ is the risk function for some procedure then the Bayes risk is

$$r_\pi = \pi_R L_R + \pi_S L_S.$$

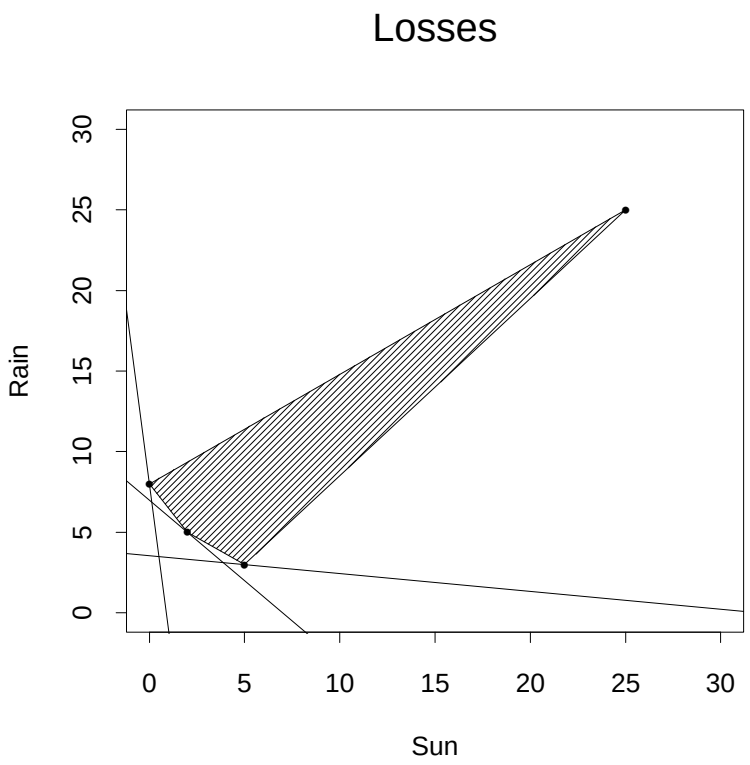
Consider the set of L such that this Bayes risk is equal to some constant. On our picture this is a line with slope $-\pi_S/\pi_R$.

Now consider three priors: $\pi_1 = (0.9, 0.1)$, $\pi_2 = (0.5, 0.5)$ and $\pi_3 = (0.1, 0.9)$. For π_1: imagine a line with slope -9 =0.9/0.1 starting on the far left of the picture and sliding right until it bumps into the convex set of possible losses in the previous picture. It does so at point B as shown in the next graph.

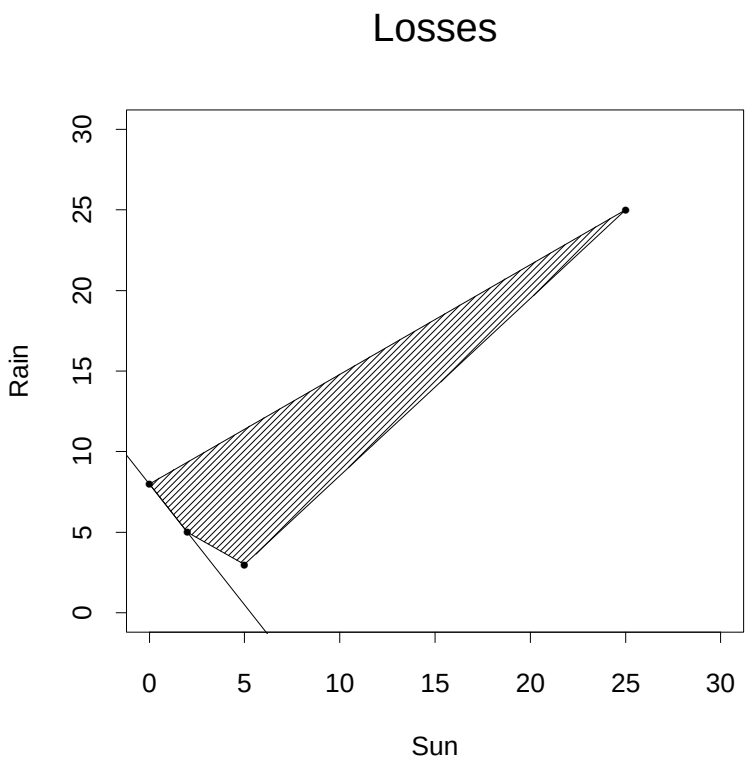
Sliding this line to the right corresponds to making r_π larger and larger so that when it just touches the convex set we have found the Bayes procedure.


Losses
Rain
Sun
0
5
10
15
20
25
30

Here is a picture showing the same lines for the three priors above.


Losses
Rain
Sun
0
5
10
15
20
25
30

The Bayes procedure for π_1 (a prior which says you're pretty sure it will be sunny) is to ride your bike. If it's a toss up between R and S you take the bus. If R is very likely you take your car. Prior $(0.6, 0.4)$ produces the line shown here:


Losses
Rain
Sun
0
5
10
15
20
25
30

Any point on line BT is Bayes for this prior.

Decision Theory and Bayesian Methods
Summary for no data case

- Decision space is the set of possible actions I might take. We assume that it is convex, typically by expanding a basic decision space D to the space $\mathcal{D}$ of all probability distributions on D.
- Parameter space Θ of possible "states of nature".
- Loss function $L = L(d, \theta)$ which is the loss I incur if I do d and θ is the true state of nature.
- We call δ^* minimax if

$$\max_\theta L(\delta^*, \theta) = \min_\delta \max_\theta L(\delta, \theta)\,.$$

- A **prior** is a probability distribution π on Θ,.
- The Bayes risk of a decision δ for a prior π is

$$r_\pi(\delta) = E_\pi(L(\delta, \theta)) = \int L(\delta, \theta)\pi(\theta)d\theta$$

  if the prior has a density. For finite parameter spaces Θ the integral is a sum.
- A decision δ^* is Bayes for a prior π if

$$r_\pi(\delta^*) \leq r_\pi(\delta)$$

  for any decision δ.
- For infinite parameter spaces: $\pi(\theta) > 0$ on Θ is a proper prior if $\int \pi(\theta)d\theta < \infty$; divide π by integral to get a density. If $\int \pi(\theta)d\theta = \infty$ π is an **improper** prior density.
- Decision δ is **inadmissible** if there is δ^* such that

$$L(\delta^*, \theta) \leq L(\delta, \theta)$$

  for all θ and there is at least one value of θ where the inequality is strict. A decision which is not inadmissible is called **admissible**.

- Every admissible procedure is Bayes, perhaps only for an improper prior. (Proof uses the Separating Hyperplane Theorem in Functional Analysis.)

- Every Bayes procedure with finite Bayes risk (for prior with density > 0 for all θ) is admissible.

  Proof: If δ is Bayes for π but not admissible there is a δ^* such that

  $$L(\delta^*, \theta) \leq L(\delta, \theta)$$

  Multiply by the prior density; integrate:

  $$r_\pi(\delta^*) \leq r_\pi(\delta)$$

  If there is a θ for which the inequality involving L is strict and if the density of π is positive at that θ then the inequality for r_π is strict which would contradict the hypothesis that δ is Bayes for π.

  Notice: the theorem actually requires the extra hypotheses: positive density, and risk functions of δ and δ^* continuous.

- A minimax procedure is admissible. (Actually there can be several minimax procedures and the claim is that at least one of them is admissible. When the parameter space is infinite it might happen that set of possible risk functions is not closed; if not then we have to replace the notion of admissible by some notion of nearly admissible.)

- The minimax procedure has constant risk. Actually the admissible minimax procedure is Bayes for some π and its risk is constant on the set of θ for which the prior density is positive.

Decision Theory and Bayesian Methods
Summary when there is data

- Decision space is the set of possible actions I might take. We assume that it is convex, typically by expanding a basic decision space D to the space $\mathcal{D}$ of all probability distributions on D.

- Parameter space Θ of possible "states of nature".

- Loss function $L = L(d, \theta)$: loss I incur if I do d and θ is true state of nature.

- Add data $X \in \mathcal{X}$ with model $\{P_\theta; \theta \in \Theta\}$: model density is $f(x|\theta)$.
- A *procedure* is a map $\delta : \mathcal{X} \mapsto \mathcal{D}$.
- The risk function for δ is the expected loss:

$$R_\delta(\theta) = R(\delta, \theta) = \mathrm{E}\left[L\{\delta(X), \theta\}\right].$$

- We call δ^* minimax if

$$\max_\theta R(\delta^*, \theta) = \min_\delta \max_\theta R(\delta, \theta).$$

- A **prior** is a probability distribution π on Θ,.
- **Bayes risk** of decision δ for prior π is

$$\begin{aligned} r_\pi(\delta) &= E_\pi(R(\delta, \theta)) \\ &= \int L(\delta(x), \theta) f(x|\theta) \pi(\theta) dx d\theta \end{aligned}$$

  if the prior has a density. For finite parameter spaces Θ the integral is a sum.
- A decision δ^* is Bayes for a prior π if

$$r_\pi(\delta^*) \leq r_\pi(\delta)$$

  for any decision δ.
- For infinite parameter spaces: $\pi(\theta) > 0$ on Θ is a **proper** prior if $\int \pi(\theta) d\theta < \infty$; divide π by integral to get a density. If $\int \pi(\theta) d\theta = \infty$ π is an **improper** prior density.
- Decision δ is **inadmissible** if there is δ^* such that

$$R(\delta^*, \theta) \leq R(\delta, \theta)$$

  for all θ and there is at least one value of θ where the inequality is strict. A decision which is not inadmissible is called **admissible**.
- Every admissible procedure is Bayes, perhaps only for an improper prior.

- If every risk function is continuous then every Bayes procedure with finite Bayes risk (for prior with density > 0 for all θ) is admissible.
- A minimax procedure is admissible.
- The minimax procedure has constant risk. The admissible minimax procedure is Bayes for some π; its risk is constant on the set of θ for which the prior density is positive.