

MATH 559: BAYESIAN THEORY AND METHODS

BASICS OF DECISION THEORY

The key components of a *decision problem* are as follows;

- a *decision* d is to be made, and the decision is selected from some set $\mathcal{D}$ of alternatives.
 - sometimes the terminology *action* (denoted a) is used.
- a *true state of nature*, $v(\theta)$, lying in set Υ , defined by the data generating model, $F_Y(y; \theta)$.
- a *loss function*, $L(d, v)$, for decision d and state v , which records the loss (or penalty) incurred when the true state of nature is v and the decision made is d .

Without loss of generality, the loss function is presumed to be non-negative.

Decisions without data: We can first consider the framework in the **deterministic** setting; ultimately, the decisions will be functions of data drawn from F_Y , with $d \equiv d(y)$. We may also simplify to a discrete case, that is, where Υ is a finite set. This will allow us to give concrete examples of the key concepts.

Example: Suppose there are four possible states of nature, $\{v_1, v_2, v_3, v_4\}$, and three possible decisions $\{d_1, d_2, d_3\}$. Note that precisely **one** of the states of nature is the true state (but we do not know which one). Suppose that $L(d_i, v_j) = l_{ij}$, for $i = 1, 2, 3, 4$, $j = 1, 2, 3$, that is, the loss for making the decision d_i when the state of nature is v_j is the numerical value l_{ij} . We may display this in a loss table

	d_1	d_2	d_3
v_1	2	7	12
v_2	3	4	20
v_3	1	2	5
v_4	5	6	9

For example, if the state of nature is v_1 , and decision d_2 is chosen, the loss incurred is $l_{12} = 7$.

Our task is to make the decision that leads to the smallest loss. The smallest loss in each row of the table is indicated in bold. We note from the table that decision d_1 has the smallest loss irrespective of the state of nature. Hence d_1 is the optimal decision.

Example: Suppose there are three possible states of nature, $\{v_1, v_2, v_3\}$, and five possible decisions $\{d_1, d_2, d_3, d_4, d_5\}$. We may display the numerical loss values in a loss table

	d_1	d_2	d_3	d_4	d_5
v_1	2	4	1	2	5
v_2	3	3	2	5	9
v_3	1	5	4	2	4

From the table, we can read off the optimal decisions:

- if the true state is v_1 , the optimal decision is d_3 ; a loss of 1 would be incurred.
- if the true state is v_2 , the optimal decision is d_3 ; a loss of 2 would be incurred.
- if the true state is v_3 , the optimal decision is d_1 ; a loss of 1 would be incurred.

There is not one single optimal decision that outperforms the others for all states of nature; in general $L(d, v)$ will not have a unique minimizing d across all v .

In this case, we would need to establish criteria via which the single optimal decision should be made.

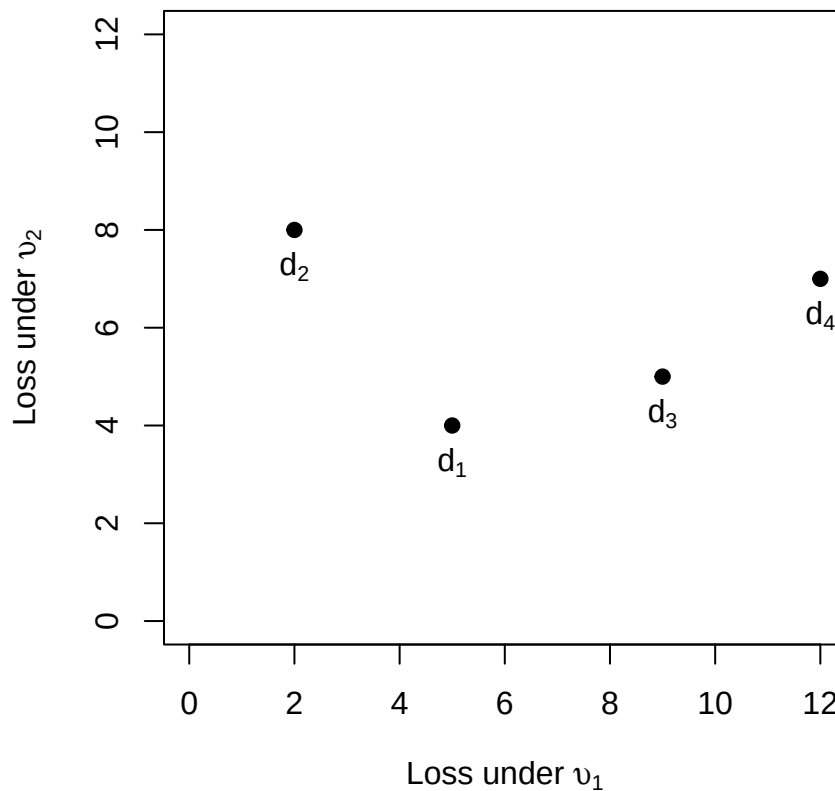
Example: Suppose there are two possible states of nature, $\{v_1, v_2\}$, and four possible decisions $\{d_1, d_2, d_3, d_4\}$. Suppose that the loss table takes the form

	d_1	d_2	d_3	d_4
v_1	5	2	9	12
v_2	4	8	5	7

- if the true state is v_1 , the optimal decision is d_2 ; a loss of 2 would be incurred.
- if the true state is v_2 , the optimal decision is d_1 ; a loss of 4 would be incurred.

We can depict this loss table in a plot using the columns as coordinates in (v_1, v_2) space.

```
u1.loss<-c(5,2,9,12); u2.loss<-c(4,8,5,7)
par(mar=c(5,4,1,2))
plot(u1.loss,u2.loss,pch=19,xlim=range(0,12),ylim=range(0,12),
     xlab=expression(paste('Loss under ',upsilon[1])),
     ylab=expression(paste('Loss under ',upsilon[2])),
     text(u1.loss,u2.loss-0.75,
          c(expression(d[1]),expression(d[2]),expression(d[3]),expression(d[4]))))
```



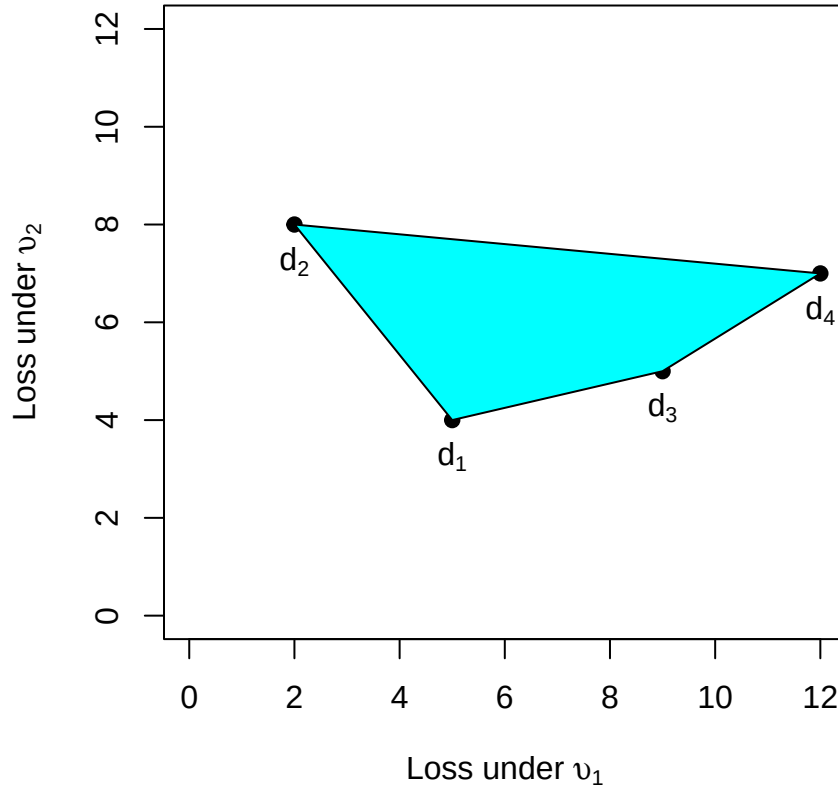
One strategy for choosing the decision would be to choose the one leads to the smallest possible loss, which would be d_2 , incurred when v_1 occurs. However, this would lead to potential difficulties, as the loss for this decision if v_2 occurs is 8, which is large.

This plot does clarify, however, that certain decisions should not be considered; d_3 and d_4 are inferior decisions irrespective of whether v_1 or v_2 is the true state. This encourages us to consider the 'lower left' boundary of the convex hull of the points.

```

par(mar=c(5,4,1,2))
plot(u1.loss,u2.loss,pch=19,xlim=range(0,12),ylim=range(0,12),
     xlab=expression(paste('Loss under ',upsilon[1])),
     ylab=expression(paste('Loss under ',upsilon[2])))
text(u1.loss,u2.loss-0.75,
     c(expression(d[1]),expression(d[2]),expression(d[3]),expression(d[4])))
polygon(u1.loss[c(1,2,4,3)],u2.loss[c(1,2,4,3)],col='cyan')

```



Admissibility: A decision rule d is *inadmissible* if there exists another rule d^* such that

$$R_n(d^*, v) \leq R_n(d, v) \quad \text{for all } v \in \Upsilon$$

and

$$R_n(d^*, v) < R_n(d, v) \quad \text{for at least one } v \in \Upsilon$$

that is, the decision rule d yields frequentist risk that is **at least as large** as the frequentist risk for d^* , for all possible states of nature, and is **greater** than the frequentist risk for d^* for at least one value of v . If d is not inadmissible, then it is *admissible*.

Randomized decisions: Consider the line segment connecting the d_1 and d_2 points, that is, the line constructed as

$$pd_1 + (1 - p)d_2 \quad 0 \leq p \leq 1$$

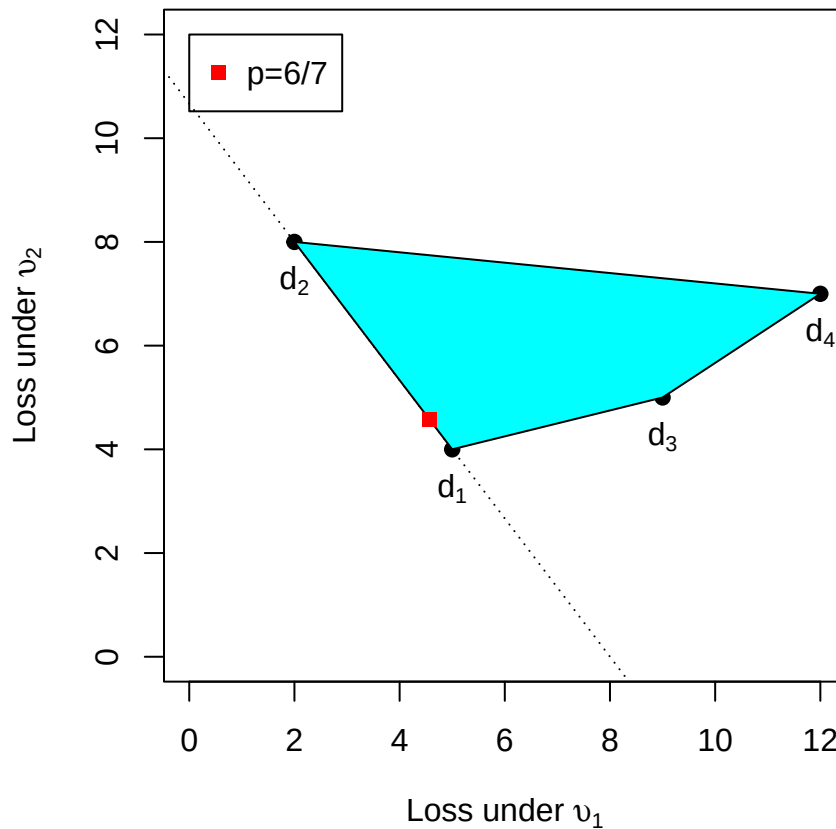
We can consider making the following *randomized* decision: for some p , we pick decision d_1 with probability p , and decision d_2 with probability $(1 - p)$. In this case, our expected loss will be

$$v_1 \text{ true} : pL(d_1, v_1) + (1 - p)L(d_2, v_1) = 5p + 2(1 - p) = 3p + 2$$

$$v_2 \text{ true} : pL(d_1, v_2) + (1 - p)L(d_2, v_2) = 4p + 8(1 - p) = -4p + 8$$

These two losses are equal if $p = 6/7$, when the loss is $32/7$. The randomized decision approach provides an alternative strategy to choosing a decision; we can generalize this to consider an arbitrary convex combination of all four decisions, and consider three probabilities (p_1, p_2, p_3) whose sum is not greater than 1, with $p_4 = (1 - p_1 - p_2 - p_3)$, then with probability p_j choose decision d_j . The randomized decision then falls within the *convex hull* of the four points. Plotting the location of the $p = 6/7$ randomized decision and the convex hull, we obtain the following plot:

```
par(mar=c(4,4,1,2))
plot(u1.loss,u2.loss,pch=19,xlim=range(0,12),ylim=range(0,12),
     xlab=expression(paste('Loss under ',upsilon[1])),
     ylab=expression(paste('Loss under ',upsilon[2])))
text(u1.loss,u2.loss-0.75,
     c(expression(d[1]),expression(d[2]),expression(d[3]),expression(d[4])))
slope<-(u2.loss[2]-u2.loss[1])/(u1.loss[2]-u1.loss[1])
intercept<-u2.loss[2]-slope*u1.loss[2]
abline(intercept,slope,lty=3)
polygon(u1.loss[c(1,2,4,3)],u2.loss[c(1,2,4,3)],col='cyan')
legend(0,12,c('p=6/7'),col='red',pch=15)
points((6/7)*u1.loss[1]+(1/7)*u1.loss[2],
       (6/7)*u2.loss[1]+(1/7)*u2.loss[2],pch=15,col='red')
```



Bayes decisions: Bayes decision rules posit a prior distribution over the space of states of nature, and computes the expected frequentist risk over that distribution. That is, in the discrete case, the Bayes risk associated with decision d is

$$R(d) = \mathbb{E}_{\pi_0}[L(d, v)] = \sum_{v \in \mathcal{Y}} L(d, v) \pi_0(v)$$

and we can compare the decisions directly in terms of this expected loss. In the continuous case we have

$$R(d) = \mathbb{E}_{\pi_0}[L(d, v)] = \int L(d, v) \pi_0(v) dv.$$

To make it explicit that this expected risk is prior-dependent, we may write $R^{\pi_0}(d)$.

Example: For the loss table above we can report the Bayes risk for each decision according to a prior

	d_1	d_2	d_3	d_4
v_1	5	2	9	12
v_2	4	8	5	7

that places $(\pi_{01}, 1 - \pi_{01})$ on the two states of nature. If the probabilities are $(0.2, 0.8)$, we have the result

	π_0	d_1	d_2	d_3	d_4
v_1	0.2	5	2	9	12
v_2	0.8	4	8	5	7
Bayes risk		4.2	6.8	5.8	8.0

whereas if the probabilities are $(0.9, 0.1)$, we have the result

	π_0	d_1	d_2	d_3	d_4
v_1	0.9	5	2	9	12
v_2	0.1	4	8	5	7
Bayes risk		4.9	2.6	8.6	11.5

We can plot a prior as lines on the plot above by considering points of equal risk, that is, where

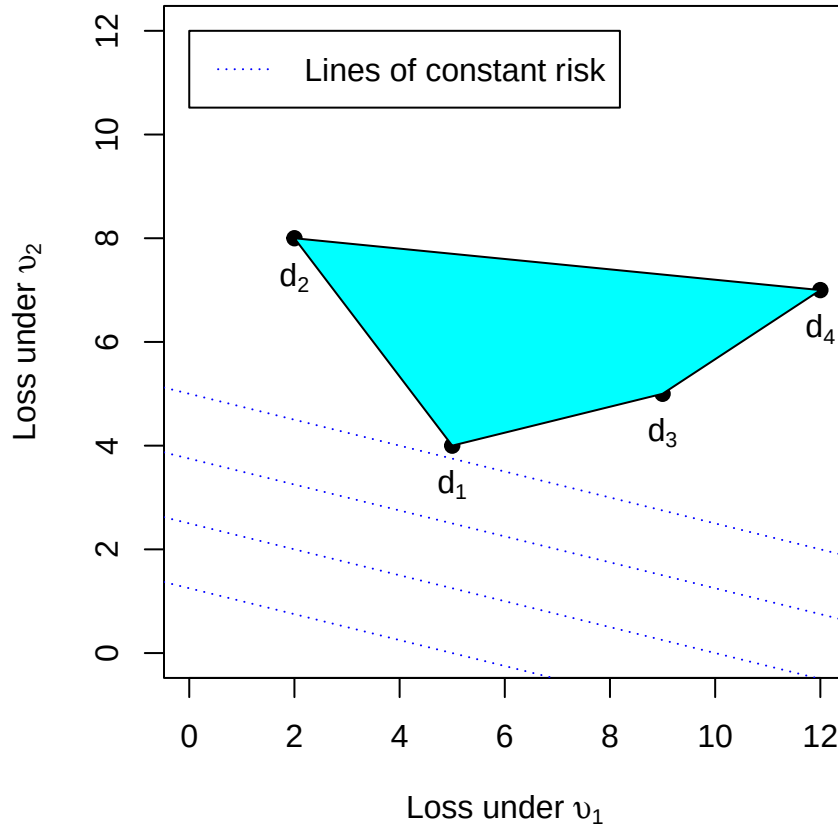
$$\pi_{01} L_1 + (1 - \pi_{01}) L_2 = r$$

say. In the following plot, we give the lines

$$y = \frac{r}{1 - \pi_{01}} - \frac{\pi_{01}}{1 - \pi_{01}} x$$

for different values of r , and $\pi_{01} = 0.2$.

```
par(mar=c(4,4,1,2))
plot(u1.loss,u2.loss,pch=19,xlim=range(0,12),ylim=range(0,12),
     xlab=expression(paste('Loss under ',upsilon[1])),
     ylab=expression(paste('Loss under ',upsilon[2])))
text(u1.loss,u2.loss-0.75,
     c(expression(d[1]),expression(d[2]),expression(d[3]),expression(d[4])))
polygon(u1.loss[c(1,2,4,3)],u2.loss[c(1,2,4,3)],col='cyan')
pi.01<-0.2
r<-1;abline(r/(1-pi.01),-pi.01/(1-pi.01),col='blue',lty=3)
r<-2;abline(r/(1-pi.01),-pi.01/(1-pi.01),col='blue',lty=3)
r<-3;abline(r/(1-pi.01),-pi.01/(1-pi.01),col='blue',lty=3)
r<-4;abline(r/(1-pi.01),-pi.01/(1-pi.01),col='blue',lty=3)
legend(0,12,c('Lines of constant risk'),lty=3,col='blue')
```



It is evident that for this prior, the optimal decision will be where the blue dotted lines touch the convex hull for the first time, when the decision will be d_1 , and the risk will be 4.2. Furthermore, d_1 will be optimal decision for **all** priors such that the slope $-\pi_{01}/(1 - \pi_{01})$ is less negative than the slope of the line connecting d_1 to d_2 ; this slope is $4/(-3) = -1.333$, so we need

$$\frac{\pi_{01}}{1 - \pi_{01}} > 1.333 \quad \implies \quad \pi_{01} > \frac{4}{7}.$$

If $\pi_{01} < 4/7$, the optimal decision is d_2 , and the risk will be 4.9. If $\pi_{01} = 4/7$, the two decisions give equal risk and the dotted blue line coincides with the line from d_1 to d_2 :

	π_0	d_1	d_2	d_3	d_4
v_1	4/7	5	2	9	12
v_2	3/7	4	8	5	7
Bayes risk		4.6	4.6	7.3	9.9

The *least favourable prior*, $\pi_0^\dagger$, is the prior for which the expected loss is as large as possible, that is, the prior for which

$$R^{\pi_0}(d) \equiv \mathbb{E}_{\pi_0}[L(d, v)] \leq \mathbb{E}_{\pi_0^\dagger}[L(d, v)] \equiv R^{\pi_0^\dagger}(d)$$

Note that here the least favourable prior is decision-dependent. For the example above, the expected loss for decision d_1 is

$$\pi_{01} \times 5 + (1 - \pi_{01})4 = 4 + \pi_{01} \times 1$$

so the least favourable prior is when π_{01} is as **large** as possible, that is, when $\pi_{01} = 4/7$. For decision d_2 , the expected loss is

$$\pi_{01} \times 2 + (1 - \pi_{01})8 = 8 + \pi_{01} \times (-6)$$

so the least favourable prior is when π_{01} is as **small** as possible, that is, again when $\pi_{01} = 4/7$.

Minimax decisions: The *minimax* decision is the one that **minimizes** the **maximum** frequentist risk after all states of nature have been considered. That is

$$\hat{d}_M = \min_d \max_v R(d, v).$$

More formally, to allow for the general nature of the two optimizations, we should write

$$\hat{d}_M = \inf_d \sup_v R(d, v).$$

although in most cases the first version is adequate.

Example: For the loss table

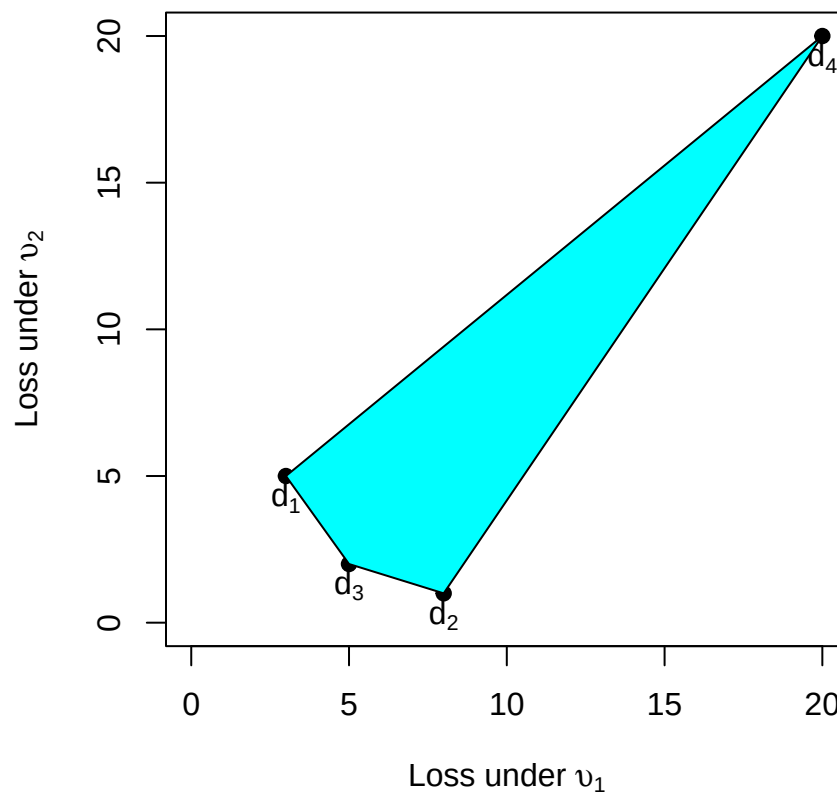
	d_1	d_2	d_3	d_4
v_1	5	2	9	12
v_2	4	8	5	7
Maximum	5	8	9	12

the minimax decision is d_1 , incurring losses 5 and 4 under the two states of nature.

Example: Consider the modified loss table

	d_1	d_2	d_3	d_4
v_1	3	8	5	20
v_2	5	1	2	20
Maximum	5	8	5	20

In this case, the (pure) minimax decisions are d_1 and d_3 with minimax risk 5.



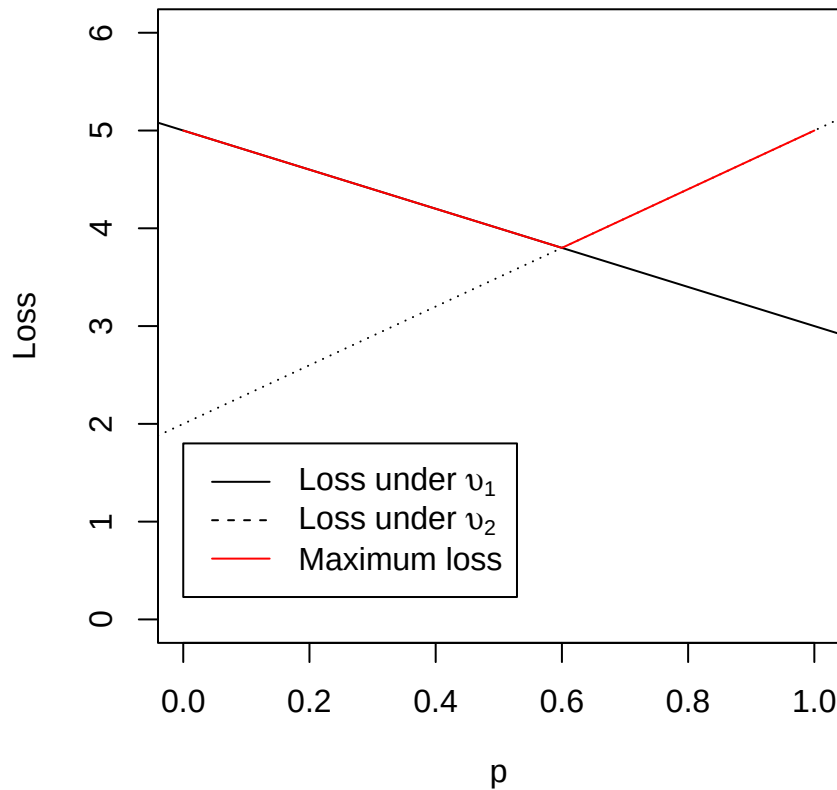
Now consider the randomized decision with probability p : this yields loss

$$v_1 \text{ true : } pL(d_1, v_1) + (1 - p)L(d_2, v_1) = 3p + 5(1 - p) = -2p + 5$$

$$v_2 \text{ true : } pL(d_1, v_2) + (1 - p)L(d_2, v_2) = 5p + 2(1 - p) = 3p + 2$$

and the two losses are equal when $5p = 3$, so $p = 3/5$, that is, when the loss is equal to $19/5$. This loss is **lower** than the non-randomized minimax risk. Furthermore, the randomized decision itself a minimax decision, as depicted in the next plot.

```
par(mar=c(4,4,2,2))
plot(0.5,0.5,xlim=range(0,1),ylim=range(0,6), type='n',
     xlab=expression(p),ylab='Loss')
abline(5,-2); abline(2,3,lty=3); lines(c(0,0.6,1),c(5,3.8,5),col='red')
legend(0,1.8,c(expression(paste('Loss under ',upsilon[1])),
     expression(paste('Loss under ',upsilon[2])),
     'Maximum loss'),lty=c(1,2,1),col=c('black','black','red'))
```



The red line indicates the maximum loss across the two states of nature as p varies; the minimum loss achievable is $19/5 = 3.8$, whereas the pure minimax loss is 5.

Statistical Decision Theory: We now acknowledge that the decisions may be functions of observed data. The true state of nature will be some function of the data generating model parameter θ , so for convenience we will write $v(\theta) \equiv \theta$, and note that the true state of nature will in general lie in some region of $\mathbb{R}^d$ that is not simply a finite set. Our task is to identify the optimal decision function, $d(\cdot)$.

- **Frequentist risk:** The *frequentist risk* or *loss* associated with decision denoted $d \equiv d(\mathbf{Y})$ is the expected loss associated with $d(\mathbf{Y})$, with the expectation taken over the distribution of $\mathbf{Y}$ given θ :

$$R_n(d, \theta) = \mathbb{E}_{F_{\mathbf{Y}}} [L(d(\mathbf{Y}), \theta); \theta] = \int_{\mathbf{y}} L(d(\mathbf{y}), \theta) f_{\mathbf{Y}}(\mathbf{y}; \theta) d\mathbf{y}$$

- **Bayes risk:** The (*average*) *Bayes risk* for $d(\mathbf{Y})$ is the expected frequentist risk over *prior* distribution, $\pi_0(\theta)$, of θ : in the continuous case

$$\begin{aligned} R_n(d) &= \mathbb{E}_{\pi_0} [R_n(d, \theta)] = \mathbb{E}_{\pi_0} [\mathbb{E}_{F_{\mathbf{Y}}} [L(d(\mathbf{Y}), \theta); \theta]] && \text{iterated expectation} \\ &= \int_{\Theta} \left\{ \int_{\mathbf{y}} L(d(\mathbf{y}), \theta) f_{\mathbf{Y}}(\mathbf{y}; \theta) d\mathbf{y} \right\} \pi_0(\theta) d\theta \\ &= \int_{\mathbf{y}} \left\{ \int_{\Theta} L(d(\mathbf{y}), \theta) \pi_n(\theta) d\theta \right\} f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} && \text{Bayes theorem} \end{aligned}$$

where the Bayes theorem identity is $\pi_n(\theta) = f_{\mathbf{Y}}(\mathbf{y}; \theta) \pi_0(\theta) / f_{\mathbf{Y}}(\mathbf{y})$. We can see that if we choose the decision mechanism $d(\cdot)$ such that the inner integral

$$\int_{\Theta} L(d(\mathbf{y}), \theta) \pi_n(\theta) d\theta$$

is minimized for each $\mathbf{y}$, then this decision will **necessarily** minimize $R_n(d)$. That is, we seek to find the decision function to minimize the posterior expected loss; note that this decision is optimal with respect to the chosen prior $\pi_0(\theta)$.

Note: In the computation of the Bayes risk, it is assumed that π_0 is *proper* (that is, it integrates to one), however the calculation may be performed with an *improper* prior (that is, a prior that is not integrable) also, in which case the terminology *generalized Bayes risk* is used. It is sometimes useful to consider a sequence of proper priors whose limiting form is improper: for example if $\pi_0(\theta) \equiv \text{Normal}(0, \lambda_0^{-1})$, for $\lambda_0 > 0$, this prior is proper, but it becomes improper in the limit as $\lambda_0 \rightarrow 0$.

Example: If $f_{\mathbf{Y}}(\mathbf{y}; \theta) \equiv \text{Normal}(\mu, \sigma^2)$, then the sample mean $\bar{Y}$ is the limit of a sequence of Bayes estimators under quadratic loss if the sequence of priors is $\text{Normal}(0, \lambda_0^{-1})$ with $\lambda_0 \rightarrow 0$

It can be shown that *Bayes rules* (the decisions that minimize the Bayes risk) have desirable properties:

Theorem Every Bayes decision made with respect to prior π_0 such that $\pi_0(\theta) > 0$ for all $\theta \in \Theta$, and such that $R_n(d) < \infty$, is *admissible*.

Proof. By contradiction: suppose $d \equiv d(\mathbf{y})$ is the Bayes decision, but is inadmissible. Then there exists another decision d^* such that we have

$$L(d^*, \theta) \leq L(d, \theta) \quad \forall \theta.$$

Multiplying both sides by $\pi_0(\theta)$ and integrating both sides, we have that

$$R_n(d^*) \leq R_n(d) < \infty$$

by assumption. By assumption of inadmissibility, there must exist a θ such that $L(d^*, \theta) < L(d, \theta)$ for that θ . But by assumption this value receives positive prior density, so in fact we must have that $R_n(d^*) < R_n(d)$. This contradicts the assumption that d is the Bayes decision. ■

This result can be augmented to confirm the desirable features of Bayes decision rules: for example, we can show that

- Every admissible decision rule is a (generalized) Bayes decision rule for some prior π_0 (which may be improper).
- If every risk function is continuous, then every Bayes procedure with finite Bayes risk (for prior with positive density for all $\mathbb{R}$) is admissible.

Connection between Bayes and minimax:

Theorem Suppose that prior π_0 is proper, and denote by d^{π_0} the Bayes decision rule computed with respect to π_0 . Then the Bayes risk for d^{π_0} is no greater than the minimax risk.

Proof. We have for the Bayes risk for d^{π_0}

$$\begin{aligned} R_n(d^{\pi_0}) &= \inf_d R_n(d) = \inf_d \int R_n(d, \theta) \pi_0(\theta) d\theta \\ &\leq \inf_d \int \sup_t R_n(d, t) \pi_0(\theta) d\theta \\ &= \inf_d \sup_{\theta} R_n(d, \theta) \end{aligned}$$

■

Theorem Suppose that prior π_0 is proper, and denote by d^{π_0} the Bayes decision rule computed with respect to π_0 . Suppose that d^{π_0} has constant risk on Θ . Then d^{π_0} is minimax, and if d^{π_0} is the unique Bayes decision, it is the unique minimax decision.

Proof. Suppose that d is any other decision. Then

$$\begin{aligned} \sup_{\theta} R_n(d, \theta) &= \int \sup_t R_n(d, t) \pi_0(\theta) d\theta \\ &\geq \int R_n(d, \theta) \pi_0(\theta) d\theta \\ &\geq \int R_n(d^{\pi_0}, \theta) \pi_0(\theta) d\theta && \text{as } d^{\pi_0} \text{ is Bayes} \\ &= \sup_{\theta} R_n(d^{\pi_0}, \theta) && \text{as } d^{\pi_0} \text{ has constant risk} \end{aligned}$$

If d^{π_0} is the unique Bayes decision, then the second inequality is strict, and the minimax decision is unique. ■

- A unique minimax procedure is admissible.
- The minimax procedure has constant risk across Θ .
- An admissible minimax procedure is Bayes for some prior; its risk is constant on the set of $\mathbb{R}$ for which the prior density is positive.
- The prior corresponding to an admissible minimax procedure is the least favourable prior.

References: These texts give much more detail of statistical decision theory:

- Robert, CP (2004), *The Bayesian Choice*, Chapter 2.
- Lehmann, EL and Casella, G, (1998), *Theory of Point Estimation*, Chapters 4-6.
- Shao, J (2003), *Mathematical Statistics*, Chapters 2-4.