§1 Decision Problem: Frequentist Approach

§1.1 Motivating example

1.1.1 Let $\theta \in (0,1)$ be an unknown parameter. A statistician has to choose between two possible actions, a_0 or a_1, which will incur different *losses* under different values of θ, as specified below:

Action		Loss under θ
a_0	$\to$	6
a_1	$\to$	$11 - 6\theta^6$

1.1.2 The statistician also observes two independent realisations, (x_1, x_2), of a random variable $X \in \{1, 2, \ldots, 8\}$, which has a distribution depending on θ and may help him make a good decision. Specifically, it is given that

$$\mathbb{P}_\theta(X = j) = \begin{cases} \theta^{(j-1)j/2} - \theta^{j(j+1)/2}, & 1 \le j \le 7, \\ \theta^{28}, & j = 8. \end{cases}$$

The joint mass function of $\boldsymbol{x} = (x_1, x_2)$ is then given by

$$f(\boldsymbol{x}|\theta) = f(x_1, x_2|\theta) = \mathbb{P}_\theta(X = x_1)\mathbb{P}_\theta(X = x_2).$$

The family $\big\{f(\cdot|\theta) : \theta \in (0,1)\big\}$ is an example of a *parametric model* of infinitely many members, each indexed by an element θ in the *parameter space* $\Theta = (0,1)$.

1.1.3 More formally, the above decision problem consists of the following key components:

1. Possible actions: $\mathcal{A} = \{a_0, a_1\}$ $\leftarrow$ *action space*.
2. Possible scenarios: $\Theta = (0,1)$ $\leftarrow$ *parameter space*.

   Note: true value of $\theta \in \Theta$ is **unknown** and **unobservable**.
3. Available observation: $\boldsymbol{x} = (x_1, x_2) \in \{1, \ldots, 8\} \times \{1, \ldots, 8\}$ $\leftarrow$ *sample space*.

   Note: *data* $\boldsymbol{x}$ is assumed to be a realisation of a random vector $\boldsymbol{X} = (X_1, X_2)$ with (joint) mass function $f(\cdot|\theta)$.
4. Loss incurred by action a taken under scenario θ: $L(\theta, a)$ $\leftarrow$ *loss function*, specified as

   $$L(\theta, a) = 6\,\mathbf{1}\{a = a_0\} + (11 - 6\theta^6)\,\mathbf{1}\{a = a_1\},$$

   where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

1.1.4 Ideally, if we know θ, then we

should pick action $a \in \mathcal{A}$ such that $L(\theta, a)$ is **minimum**.

Without knowing the true value of θ and assisted only by the observation x, how should the statistician pick an action to minimise his loss?

Decision depends on information about θ conveyed by the observed data x.

This leads to a *statistical inference* problem.

§1.2 General setup

1.2.1 *Parameter space* Θ (usually a set of k-dimensional real vectors $\subset \mathbb{R}^k$).

The true parameter is some **unknown** $\theta \in \Theta$.

1.2.2 *Sample space* $\mathcal{S}$

— collection of all possible *realisations* $\boldsymbol{x}$ of a random vector $\boldsymbol{X}$.

Note: "realisation" means the assignment of an observed value to a random "variate". Alternatively, we say, "$\boldsymbol{X}$ is observed to be $\boldsymbol{x}$."

Usually, $\boldsymbol{x}$ takes the form of

a sample of size n: $\boldsymbol{x} = (x_1, \ldots, x_n)$,

so that $\mathcal{S}$ consists of all possible realisations $\boldsymbol{x} = (x_1, \ldots, x_n)$ of $\boldsymbol{X} = (X_1, \ldots, X_n)$.

1.2.3 *Statistical model (a link between Θ and $\mathcal{S}$)*

— a family of probability functions $\{f(\cdot|\theta) : \theta \in \Theta\}$ defined on $\mathcal{S}$,

proposed as candidates for the "true" probability function of the random vector $\boldsymbol{X}$.

Note: $f(\cdot|\theta)$ is a *mass function* for *discrete* $\boldsymbol{X}$ and a *probability density function* (pdf) for *continuous* $\boldsymbol{X}$.

1.2.4 Assume that $\boldsymbol{X}$ is distributed under $f(\cdot|\theta)$, for some unknown value $\theta \in \Theta$.

Statistical inference attempts to infer about the "true" value of θ based on observed data $\boldsymbol{X} = \boldsymbol{x}$.

1.2.5 *Action space* $\mathcal{A}$ — collection of all possible actions under consideration.

Examples of action spaces for common statistical inference problems:

- *(point estimation)* — pick a vector of k numbers to estimate an unknown parameter $\in \mathbb{R}^k$
  $\longrightarrow \mathcal{A} = \mathbb{R}^k =$ set of all k-dimensional vectors
- *(interval estimation)* — pick an interval to estimate an unknown positive parameter
  $\longrightarrow \mathcal{A} =$ set of all possible intervals $[u, v]$ with $0 \le u < v$
- *(hypothesis test)* — reject or accept a hypothetical statement about an unknown parameter
  $\longrightarrow \mathcal{A} = \{$rejection, acceptance$\}$

1.2.6 *Loss function* $L : \Theta \times \mathcal{A} \to \mathbb{R}$

— $L(\theta, a)$ is *loss* incurred by taking action a when θ is the true parameter.

Note: negative loss implies positive *gain*.

§1.3 Frequentist approach

1.3.1 Operating according to the *repeated sampling principle*, the *frequentist approach*

- formulates potential solutions to a decision problem in the form of *decision rules*,
- identifies optimal decision rule by considering losses incurred by a decision rule if this rule were repeatedly applied to samples (hypothetically) generated from $f(\cdot|\theta)$, for each $\theta \in \Theta$.

1.3.2 *Decision rule* $d : \mathcal{S} \to \mathcal{A}$ *(a map from* $\mathcal{S}$ *into* $\mathcal{A}$*)*

— prescribes an action, i.e. $d(\boldsymbol{x}) \in \mathcal{A}$, to be taken when $\boldsymbol{X}$ is observed to be $\boldsymbol{x}$ $(\in \mathcal{S})$.

1.3.3 For any given decision problem, there exist numerous choices of $d(\cdot)$. Frequentists attempt to find the "best" rule $d(\cdot)$ by considering $L(\theta, d(\boldsymbol{X}))$ under repeated sampling of $\boldsymbol{X}$ from $f(\cdot|\theta)$.

This entails consideration of the *risk function* of a decision rule.

Definition. The *risk function* of the decision rule $d(\cdot)$ is the *expected loss* incurred by adopting decision rule d under each possible $\theta \in \Theta$, i.e.

$$R(\theta, d) = \mathbb{E}_\theta\big[L(\theta, d(\boldsymbol{X}))\big].$$

— if $\boldsymbol{X}$ is continuous, $R(\theta, d) = \displaystyle\int_{\mathcal{S}} L(\theta, d(\boldsymbol{x}))f(\boldsymbol{x}|\theta)\,d\boldsymbol{x}$;

— if $\boldsymbol{X}$ is discrete, $R(\theta, d) = \displaystyle\sum_{\boldsymbol{x}\in\mathcal{S}} L(\theta, d(\boldsymbol{x}))f(\boldsymbol{x}|\theta)$.

The risk function characterises the performance of rule d under each possible $\theta \in \Theta$.

1.3.4 Examples of loss and risk functions:

- point estimation — to estimate some unknown $\theta \in \mathbb{R}^k$.

  Action space $\mathcal{A} = \mathbb{R}^k$.

  Decision rule $d \equiv$ *estimator* (e.g. sample mean, sample median).

  For any $\boldsymbol{y} = (y_1, \dots, y_k)^\top \in \mathbb{R}^k$, define L_p norm of $\boldsymbol{y}$ to be $\|\boldsymbol{y}\|_p = \big(\sum_{j=1}^k |y_j|^p\big)^{1/p}$.

  For any $\theta, a \in \mathbb{R}^k$,

  (i) loss function $L(\theta, a) = \|\theta - a\|_2^2 \longrightarrow$ (*squared loss*)
  $\longrightarrow$ risk function $R(\theta, d) = \mathbb{E}_\theta \|\theta - d(\boldsymbol{X})\|_2^2$ (*mean squared error, MSE*);

  (ii) loss function $L(\theta, a) = \|\theta - a\|_1 \longrightarrow$ (*absolute deviation loss*)
  $\longrightarrow$ risk function $R(\theta, d) = \mathbb{E}_\theta \|\theta - d(\boldsymbol{X})\|_1$ (*mean absolute deviation, MAD*).

- hypothesis test — to test $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \notin \Theta_0$.

  Action space $\mathcal{A} = \{a_0, a_1\}$: $\quad a_0 \longrightarrow$ accept H_0, $\quad a_1 \longrightarrow$ reject H_0.

  Decision rule d is characterised by a *critical region* $\mathcal{C}$ (for rejecting H_0), defined as

  $$\mathcal{C} = \{\boldsymbol{x} \in \mathcal{S} : d(\boldsymbol{x}) = a_1\}.$$

  Possible loss function L (*unit loss*):

  $$L(\theta, a) = \mathbf{1}\{a = a_1,\, \theta \in \Theta_0\} + \mathbf{1}\{a = a_0,\, \theta \notin \Theta_0\} \quad \leftarrow \quad \text{1-unit loss for wrong decision,}$$

  where $\mathbf{1}\{\cdot\}$ denotes the *indicator function*, i.e. $\mathbf{1}\{E\} = 1$ if E is true and 0 if E is false.

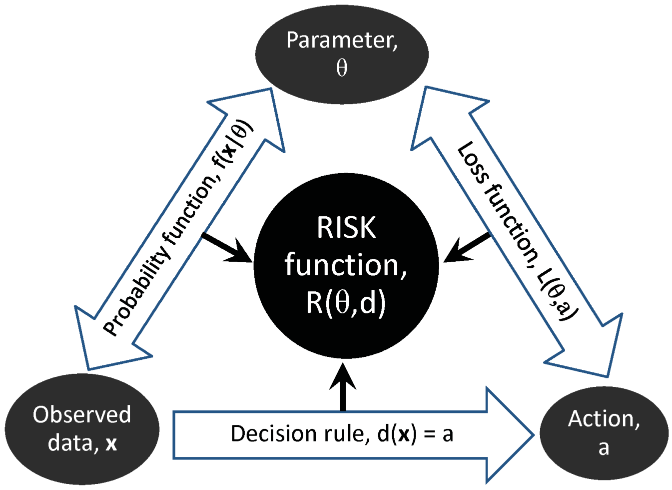
  Under the above unit loss function,

  $$R(\theta, d) = \mathbb{E}_\theta\big[L(\theta, d(\boldsymbol{X}))\big] = L(\theta, a_1)\, \mathbb{P}_\theta(d(\boldsymbol{X}) = a_1) + L(\theta, a_0)\, \mathbb{P}_\theta(d(\boldsymbol{X}) = a_0)$$
  $$= \begin{cases} \mathbb{P}_\theta(d(\boldsymbol{X}) = a_1) = \mathbb{P}(\text{reject } H_0 \,|\, H_0 \text{ correct}) = \textit{Type I error} \text{ probability}, & \theta \in \Theta_0, \\ \mathbb{P}_\theta(d(\boldsymbol{X}) = a_0) = \mathbb{P}(\text{accept } H_0 \,|\, H_0 \text{ wrong}) = \textit{Type II error} \text{ probability}, & \theta \notin \Theta_0. \end{cases}$$

1.3.5 Frequentists compare different decision rules $d(\cdot)$ in terms of their respective risk functions.

Ultimate goal $\longrightarrow$ choose one rule $d(\cdot)$ that is "best" (has "small" risk in **some** sense).

A schematic chart of the frequentist approach to a decision problem is shown below.



Even in a very simple problem where there exist only a finite number of possible decision rules, it might not be obvious that any one of them is "best". We need some "sensible" criteria to guide our choice.

§1.4 Example: quality control

1.4.1 From a batch of 10 batteries, draw 1 at random and test it. Either it is defective or OK. This is the only observation upon which our decision about the whole batch has to rely.

1.4.2 Let X be the outcome:

$$X = 1 \text{ if defective}, \qquad X = 0 \text{ if OK}.$$

Sample space $\mathcal{S} = \{0, 1\}$.

1.4.3 Let θ be no. of defective batteries in the batch: $\theta \in \Theta = \{0, 1, \ldots, 10\}$.

1.4.4 After the test, we take either of the following two actions,

- a_0: sell all 10 batteries, at \$15 each, but it costs us \$30 for each defective battery sold
- a_1: scrap all 10 batteries at total fixed cost of \$10.

1.4.5 Define the loss function to be the cost, i.e.

$$L(\theta, a_0) = 30\theta - 150, \qquad L(\theta, a_1) = 10.$$

1.4.6 Only 4 possible decision rules:

rule	$X = 0$	$X = 1$		
$d_1(X)$	a_0	a_1	$\longleftarrow$	sensible
$d_2(X)$	a_1	a_0	$\longleftarrow$	silly
$d_3(X)$	a_0	a_0	$\longleftarrow$	reckless
$d_4(X)$	a_1	a_1	$\longleftarrow$	pessimistic

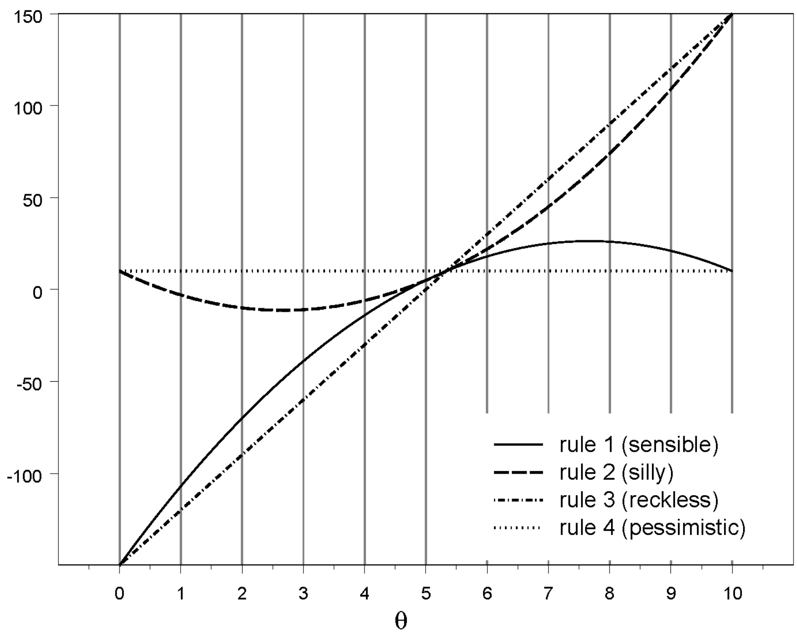
1.4.7 Statistical model:

$$\text{(Bernoulli model)} \quad \mathbb{P}_\theta(X = 1) = 1 - \mathbb{P}_\theta(X = 0) = \theta/10, \qquad \theta \in \{0, 1, \dots, 10\}.$$

1.4.8 Risk functions:

$$\begin{aligned} R(\theta, d_1) &= \mathbb{E}_\theta\big[L(\theta, d_1(X))\big] = \mathbb{P}_\theta(X = 0)\, L(\theta, d_1(0)) + \mathbb{P}_\theta(X = 1)\, L(\theta, d_1(1)) \\ &= (1 - \theta/10)(30\theta - 150) + (\theta/10)(10) = -3\theta^2 + 46\theta - 150, \\ R(\theta, d_2) &= (1 - \theta/10)(10) + (\theta/10)(30\theta - 150) = 3\theta^2 - 16\theta + 10, \\ R(\theta, d_3) &= L(\theta, a_0) = 30\,\theta - 150, \\ R(\theta, d_4) &= 10. \end{aligned}$$

The diagram below plots the risk functions of the 4 decision rules $d_1, \dots, d_4$.



Ideally, the best decision rule is one with the smallest risk function at **all** $\theta = 0, 1, \ldots, 10$. The plots show that this does **not** exist, e.g.

for small θ, d_1, d_3 better than d_2, d_4 (d_3 best);
for large θ, d_1, d_4 better than d_2, d_3 (d_4 best), etc.

Although d_1 is commonsensical rule, it is **never** actually the best since its risk function is bounded below by those of d_3 and d_4 for small and large θ, respectively.

We need more concrete criteria for selecting decision rules...

§1.5 Admissibility

1.5.1 **Definition.** A rule d *strictly dominates* another rule d^* if

$$R(\theta, d) \leq R(\theta, d^*) \;\text{ for all } \theta \in \Theta \quad \text{ and } \quad R(\theta', d) < R(\theta', d^*) \;\text{ for some } \theta' \in \Theta.$$

1.5.2 If d strictly dominates d^*, then obviously d is the better choice.

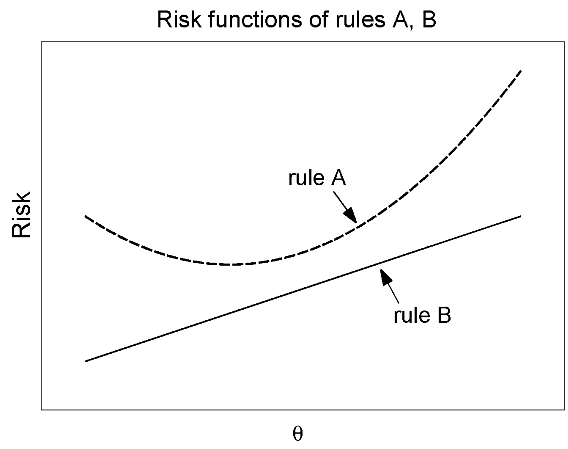
1.5.3 **Definition.** A rule strictly dominated by another rule is *inadmissible*.

1.5.4 **Definition.** If d is not inadmissible, then it is said to be *admissible*.

1.5.5 An inadmissible rule is obviously stupid! An admissible rule is **not** obviously stupid but may still be stupid!

1.5.6 Admissibility is a rather weak (but sensible) condition on choice of decision rules. It provides the first criterion for discarding obviously stupid rules.

1.5.7 The diagram below shows that rule A is strictly dominated by rule B, hence inadmissible.



1.5.8 **Example §1.4** *(cont'd).*

Rule d_2 is strictly dominated by d_1 and is inadmissible. The other 3 rules, d_1, d_3, d_4, are admissible.

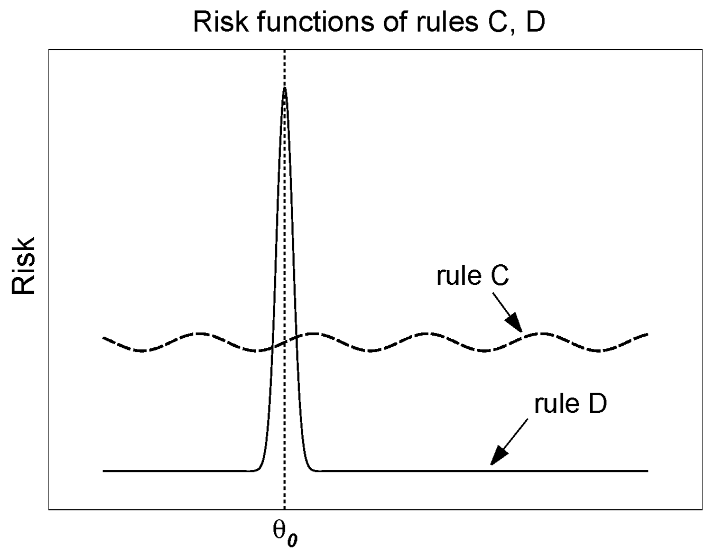
§1.6 Minimaxity

1.6.1 **Definition.** A rule d is *minimax* if, for all possible rules d',

$$\sup\{R(\theta, d') : \theta \in \Theta\} \geq \sup\{R(\theta, d) : \theta \in \Theta\}.$$

1.6.2 A minimax rule has the minimum "worst-case" risk among the "worst-case" risks of all rules.

1.6.3 In the following diagram, $\sup\{R(\theta, D) : \theta \in \Theta\} > \sup\{R(\theta, C) : \theta \in \Theta\}$ (rule D has bigger "worst-case" risk than rule C).



But, in practice, would you prefer rule C to rule D?

Not *if the true θ is unlikely to lie close to θ_0, at which rule D has maximum risk.*

1.6.4 **Example §1.4** *(cont'd).*

We see from the plots of risk functions in §1.4.8 that rule d_4 is minimax.

§1.7 Unbiased rule

1.7.1 **Definition.** A rule d is *unbiased* if $\mathbb{E}_\theta L(\theta', d(\boldsymbol{X})) \geq R(\theta, d)$ for all $\theta, \theta' \in \Theta$.

1.7.2 An unbiased rule $d(\boldsymbol{X})$ incurs the smallest expected loss if the loss is measured with respect to the true θ generating the data $\boldsymbol{X}$.

1.7.3 **Example §1.4** *(cont'd).*

Note that for any $\theta, \theta' \in \{0, 1, \ldots, 10\}$,

$$\begin{aligned} &\mathbb{E}_\theta L(\theta', d_1(X)) - R(\theta, d_1) = (1 - \theta/10)\, 30(\theta' - \theta), && \mathbb{E}_\theta L(\theta', d_3(X)) - R(\theta, d_3) = 30(\theta' - \theta), \\ &\mathbb{E}_\theta L(\theta', d_2(X)) - R(\theta, d_2) = (\theta/10)\, 30(\theta' - \theta), && \mathbb{E}_\theta L(\theta', d_4(X)) - R(\theta, d_4) = 0. \end{aligned}$$

Clearly, d_4 is unbiased.

For $d = d_1, d_2$ or d_3, there exist θ, θ' with $\mathbb{E}_\theta L(\theta', d(X)) - R(\theta, d) < 0$, so d_1, d_2, d_3 are not unbiased.

1.7.4 *Point estimation* —

Let $L(\theta, d(\boldsymbol{X})) = \|\theta - d(\boldsymbol{X})\|_2^2$ (squared loss).

Assume that $\varphi(\theta) \equiv \mathbb{E}_\theta\left[d(\boldsymbol{X})\right] \in \Theta$.

Then d is an unbiased rule iff

$$\mathbb{E}_\theta \|\theta' - d(\boldsymbol{X})\|_2^2 \geq \mathbb{E}_\theta \|\theta - d(\boldsymbol{X})\|_2^2 \quad \text{for all } \theta, \theta' \in \Theta$$

iff

$$\|\theta' - \varphi(\theta)\|_2^2 \geq \|\theta - \varphi(\theta)\|_2^2 \quad \text{for all } \theta, \theta' \in \Theta$$

iff

$$\varphi(\theta) = \theta \quad \text{for all } \theta \in \Theta$$

iff $d(\boldsymbol{X})$ is an *unbiased estimator* of θ.

1.7.5 It may be intractable to find the **best** rule in the class of **all** rules, but we may be able to find the **best** rule in the subclass of **unbiased** rules.

The best unbiased rule may even turn out to be the best rule.

§1.8 Bayes rule

1.8.1 Major problem in comparing two rules, d_1 and d_2, is that we may find

$$R(\theta, d_1) > R(\theta, d_2) \text{ for some } \theta, \text{ but } R(\theta, d_1) < R(\theta, d_2) \text{ for other } \theta.$$

1.8.2 Sometimes, although we do not know the true θ, somehow we may have a **subjective** opinion about how the true θ is like (e.g. extreme cases of θ are less likely than moderate cases).

This subjective opinion may be based on past experience, intuition, empirical data, religious belief etc. We call this *prior knowledge* about the unknown θ.

Such prior knowledge may be used to reduce the risk function to a scalar, based on which we may compare the risks of different decision rules in an unambiguous way.

1.8.3 Let $\pi(\theta)$ be a non-negative *prior weight function* defined on the parameter space Θ. For each $\theta \in \Theta$, the "prior weight" $\pi(\theta)$ reflects how much "belief" we have in θ being the true parameter value according to our prior knowledge.

Note: If $\int_\Theta \pi(\theta)\, d\theta = 1$ or $\sum_{\theta\in\Theta} \pi(\theta) = 1$, then $\pi(\cdot)$ is usually referred to as a *prior probability function* — an important device in the Bayesian approach to inference.

1.8.4 **Definition.** The *Bayes risk* of the rule d (with respect to prior π) is

$$r(\pi, d) = \begin{cases} \int_\Theta R(\theta, d)\, \pi(\theta)\, d\theta, & \theta \text{ continuous,} \\ \sum_{\theta\in\Theta} R(\theta, d)\, \pi(\theta), & \theta \text{ discrete.} \end{cases}$$

1.8.5 **Definition.** The *Bayes rule* is the rule d that has the smallest Bayes risk, i.e.

$$r(\pi, d) = \min\big\{r(\pi, d^*) : d^* \in \text{class of rules}\big\}.$$

1.8.6 *Goal* — given prior weight function $\pi(\cdot)$, seek the Bayes rule (with respect to prior π).

Note: Since the choice of $\pi(\theta)$ is subjective, Bayes rules may differ from person to person.

1.8.7 **Example §1.4**: (cont'd)

Exercise — What are the Bayes rules with respect to the following prior weight functions?

(i) $\pi(\theta) = \theta^2$, (ii) $\pi(\theta) = (10-\theta)^2$, (iii) $\pi(\theta) \equiv 1$ (*non-informative).*

Can you find a prior π with respect to which d_2 is the Bayes rule?

1.8.8 In practice, the Bayes rule can be found conveniently as follows.

Assume for brevity that both θ and $\boldsymbol{X}$ are continuous (the other cases where θ and/or $\boldsymbol{X}$ are discrete follow analogously).

For each $\boldsymbol{x} \in \mathcal{S}$, let $d(\boldsymbol{x})$ be the action $a \in \mathcal{A}$ which minimises $\int_\Theta L(\theta, a)\pi(\theta) f(\boldsymbol{x}|\theta)\, d\theta$.

Then, for any rule d^*,

$$\begin{aligned} r(\pi, d^*) &= \int_\Theta \pi(\theta) \int_\mathcal{S} L(\theta, d^*(\boldsymbol{x})) f(\boldsymbol{x}|\theta)\, d\boldsymbol{x}\, d\theta = \int_\mathcal{S} \left\{ \int_\Theta L(\theta, d^*(\boldsymbol{x}))\pi(\theta) f(\boldsymbol{x}|\theta)\, d\theta \right\} d\boldsymbol{x} \\ &\geq \int_\mathcal{S} \left\{ \int_\Theta L(\theta, d(\boldsymbol{x}))\pi(\theta) f(\boldsymbol{x}|\theta)\, d\theta \right\} d\boldsymbol{x} = r(\pi, d). \end{aligned}$$

This confirms that d is a Bayes rule with respect to the prior π.

§1.9 Revisit of motivating example §1.1

1.9.1 Recall that a decision rule d is a map from $\{1, \ldots, 8\} \times \{1, \ldots, 8\}$ to $\{a_0, a_1\}$.

There are altogether $\mathbf{2^{8\times 8}}$ possible decision rules.

The risk function of a rule d is given by

$$R(\theta, d) = L\big(\theta, a_0\big)\mathbb{P}_\theta\big(d(\boldsymbol{X}) = a_0\big) + L\big(\theta, a_1\big)\mathbb{P}_\theta\big(d(\boldsymbol{X}) = a_1\big) = 6 + \mathbb{P}_\theta\big(d(\boldsymbol{X}) = a_1\big)(5 - 6\theta^6).$$

In what follows we denote by d_j the rule of "always taking action a_j", $j = 0, 1$, i.e.

$$d_0(\boldsymbol{x}) \equiv a_0, \qquad d_1(\boldsymbol{x}) \equiv a_1.$$

1.9.2 **Admissibility**

Is d_1 admissible?

Let d^* be any arbitrary rule distinct from d_1. Then we have

$$\begin{aligned} R(\theta, d_1) - R(\theta, d^*) &= \big\{\mathbb{P}_\theta\big(d_1(\boldsymbol{X}) = a_1\big) - \mathbb{P}_\theta\big(d^*(\boldsymbol{X}) = a_1\big)\big\}(5 - 6\theta^6) \\ &= \mathbb{P}_\theta\big(d^*(\boldsymbol{X}) = a_0\big)(5 - 6\theta^6) < 0 \quad \text{for } \theta > (5/6)^{1/6}, \end{aligned}$$

which implies that d_1 cannot be strictly dominated by d^*. Thus we conclude that

$\boxed{d_1 \textit{ is admissible.}}$

1.9.3 Minimaxity

It is clear that $R(\theta, d_0) \equiv 6$ for any $\theta \in (0,1)$, hence $\sup_{\theta\in(0,1)} R(\theta, d_0) = 6$.

For any decision rule d and any $0 < \psi \le (5/6)^{1/6}$, we have

$$\sup_{\theta\in(0,1)} R(\theta, d) \ge 6 + \mathbb{P}_\psi\big(d(\boldsymbol{X}) = a_1\big)(5 - 6\psi^6) \ge 6 = \sup_{\theta\in(0,1)} R(\theta, d_0).$$

Thus, $\boxed{d_0 \text{ is minimax.}}$

1.9.4 Unbiasedness

For any $\theta_1, \theta_2 \in (0,1)$ and any rule d, consider

$$\begin{aligned}\mathbb{E}_{\theta_1} L(\theta_2, d(\boldsymbol{X})) - R(\theta_1, d) &= \big\{6 + \mathbb{P}_{\theta_1}\big(d(\boldsymbol{X}) = a_1\big)(5 - 6\theta_2^6)\big\} - \big\{6 + \mathbb{P}_{\theta_1}\big(d(\boldsymbol{X}) = a_1\big)(5 - 6\theta_1^6)\big\} \\ &= 6(\theta_1^6 - \theta_2^6)\,\mathbb{P}_{\theta_1}\big(d(\boldsymbol{X}) = a_1\big) \begin{cases} = 0, & d = d_0, \\ < 0, & d \ne d_0 \;\;\&\;\; \theta_1 < \theta_2. \end{cases}\end{aligned}$$

Thus, $\boxed{d_0 \text{ is the only unbiased rule.}}$

1.9.5 Bayes rule

(i) Consider prior weight $\pi(\theta) = 0$ for $0 < \theta < (5/6)^{1/6}$. The Bayes risk of d is

$$\begin{aligned} r(\pi, d) &= \int_{(5/6)^{1/6}}^1 R(\theta, d)\pi(\theta)\,d\theta = \int_{(5/6)^{1/6}}^1 \big\{6 + \mathbb{P}_\theta\big(d(\boldsymbol{X}) = a_1\big)(5 - 6\theta^6)\big\}\pi(\theta)\,d\theta \\ &\ge 6 + \int_{(5/6)^{1/6}}^1 (5 - 6\theta^6)\,\pi(\theta)\,d\theta = r(\pi, d_1).\end{aligned}$$

Thus, $\boxed{d_1 \textit{ is the Bayes rule.}}$

(ii) Consider prior weight $\pi(\theta) = 0$ for $(5/6)^{1/6} < \theta < 1$. The Bayes risk of d is

$$r(\pi, d) = \int_0^{(5/6)^{1/6}} \big\{6 + \mathbb{P}_\theta\big(d(\boldsymbol{X}) = a_1\big)(5 - 6\theta^6)\big\}\pi(\theta)\,d\theta \ge 6 = r(\pi, d_0).$$

Thus, $\boxed{d_0 \textit{ is the Bayes rule.}}$

(iii) Consider prior weight function $\pi(\theta) = \theta^{c-1}$ for $0 < \theta < 1$ and some constant $c > 0$. How can we derive the Bayes rule d that minimises the Bayes risk

$$r(\pi, d) = \int_0^1 \big\{6 + \mathbb{P}_\theta\big(d(\boldsymbol{X}) = a_1\big)(5 - 6\theta^6)\big\}\theta^{c-1}\,d\theta = ?$$

Based on the data $\boldsymbol{x} = (x_1, x_2)$ from the probability function

$$f(x_1, x_2|\theta) = \prod_{i=1}^{2} \mathbb{P}_\theta(X = x_i) = \prod_{i=1}^{2} \left[\theta^{(x_i-1)x_i/2} - \theta^{x_i(x_i+1)/2}\mathbf{1}\{x_i < 8\}\right]$$
$$= \theta^{(x_1^2+x_2^2-x_1-x_2)/2}\left[1 + \mathbf{1}\{x_1, x_2 < 8\}\theta^{x_1+x_2} - \mathbf{1}\{x_1 < 8\}\theta^{x_1} - \mathbf{1}\{x_2 < 8\}\theta^{x_2}\right]$$

and the loss function

$$L(\theta, a) = 6\,\mathbf{1}\{a = a_0\} + (11 - 6\,\theta^6)\,\mathbf{1}\{a = a_1\},$$

we have

$$\int_0^1 L(\theta, a_1)\pi(\theta)f(x_1, x_2|\theta)\,d\theta - \int_0^1 L(\theta, a_0)\pi(\theta)f(x_1, x_2|\theta)\,d\theta$$
$$= \int_0^1 (5 - 6\,\theta^6)\theta^{(x_1^2+x_2^2-x_1-x_2)/2}$$
$$\times\left[1 + \mathbf{1}\{x_1, x_2 < 8\}\theta^{x_1+x_2} - \mathbf{1}\{x_1 < 8\}\theta^{x_1} - \mathbf{1}\{x_2 < 8\}\theta^{x_2}\right]\theta^{c-1}\,d\theta.$$

Thus, Bayes rule $d(\boldsymbol{x}) = a_1$ if and only if the above integral is negative. In particular, if $x_1 = x_2 = 8$, then the integral equals $-(c + 26)/\{(c + 56)(c + 62)\} < 0$, implying that $d(8, 8) = a_1$.

§1.10 Randomized decision rule

1.10.1 **Definition.** A *randomized* decision rule is a probability mixture of rules.

Suppose we have s decision rules $d_1, d_2, \ldots, d_s$.

Fix probabilities $p_1, p_2, \ldots, p_s \geq 0$ with $p_1 + p_2 + \cdots + p_s = 1$.

Define a new rule d^* by

$$d^*(\boldsymbol{X}) = d_i(\boldsymbol{X}) \text{ with probability } p_i, \qquad \text{for } i = 1, 2, \ldots, s.$$

Then d^* is called a *randomized* decision rule.

Note: By contrast, the rules $d_1, \ldots, d_s$ are *non-randomized*, or *deterministic*.

1.10.2 The risk function of the randomized rule d^* is

$$R(\theta, d^*) = \mathbb{E}_\theta L(\theta, d^*(\boldsymbol{X})) = \mathbb{E}_\theta\Big[\sum_{i=1}^{s} p_i L(\theta, d_i(\boldsymbol{X}))\Big] = \sum_{i=1}^{s} p_i R(\theta, d_i),$$

which is a *convex combination* of the risk functions of the individual rules d_i's.

1.10.3 Given a prior $\pi(\cdot)$, the Bayes risk of the randomized rule d^* is

$$r(\pi, d^*) = \int R(\theta, d^*)\,\pi(\theta)\,d\theta = \sum_{i=1}^{s} p_i r(\pi, d_i) \geq \min\big\{r(\pi, d_1), \ldots, r(\pi, d_s)\big\}.$$

Thus, the risk of a Bayes rule chosen among $\{d_1, \ldots, d_s\}$ cannot be reduced further by mixing them into a randomized rule.

§1.11 Risk set

1.11.1 Consider a simple case where $\Theta = \{\theta_1, \theta_2\}$ has only two possible parameter values.

Definition. The *risk set* is

$$\mathcal{R} = \Big\{\big(R(\theta_1, d), R(\theta_2, d)\big) : d \text{ is a (possibly randomized) decision rule}\Big\} \subset \mathbb{R}^2.$$

Note: Each point in $\mathcal{R}$ can be identified with the risk function of a (possibly randomized) rule.

1.11.2 **Lemma.** $\mathcal{R}$ is *convex*.

Proof:

Take $r_1, r_2 \in \mathcal{R}$, $p_1, p_2 \geq 0$, $p_1 + p_2 = 1$.

By definition, there exist decision rules d_1, d_2 such that

$$r_1 = \big(R(\theta_1, d_1), R(\theta_2, d_1)\big), \quad r_2 = \big(R(\theta_1, d_2), R(\theta_2, d_2)\big).$$

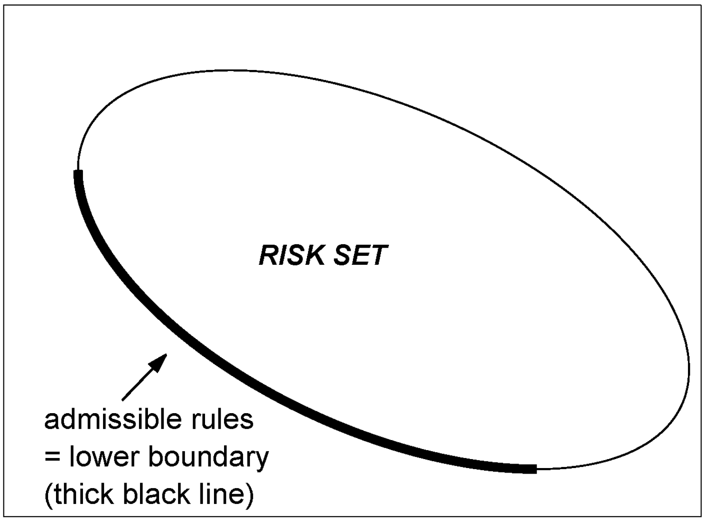
Let d be the randomized rule such that $d = d_i$ with probability p_i, $i = 1, 2$. Then

$$R(\theta, d) = p_1 R(\theta, d_1) + p_2 R(\theta, d_2) \quad \textit{for } \theta = \theta_1 \textit{ or } \theta_2.$$

Thus $p_1 r_1 + p_2 r_2 = \big(R(\theta_1, d), R(\theta_2, d)\big) \in \mathcal{R}$. ∎

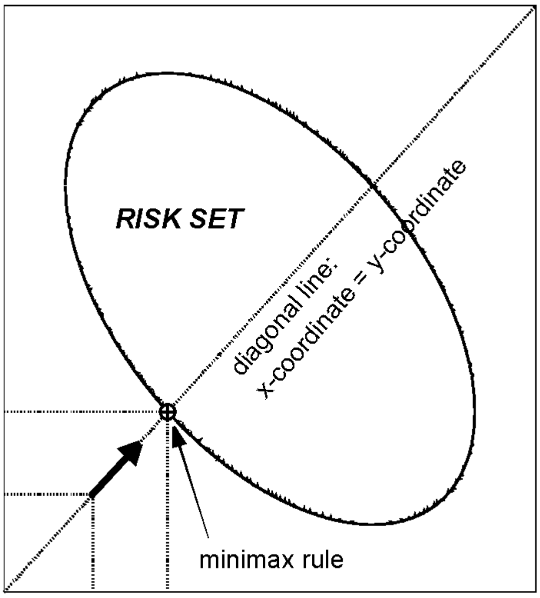
1.11.3 We can draw $\mathcal{R}$ on a 2D plane, such that the x- and y-coordinates give the risks at θ_1 and θ_2 respectively.

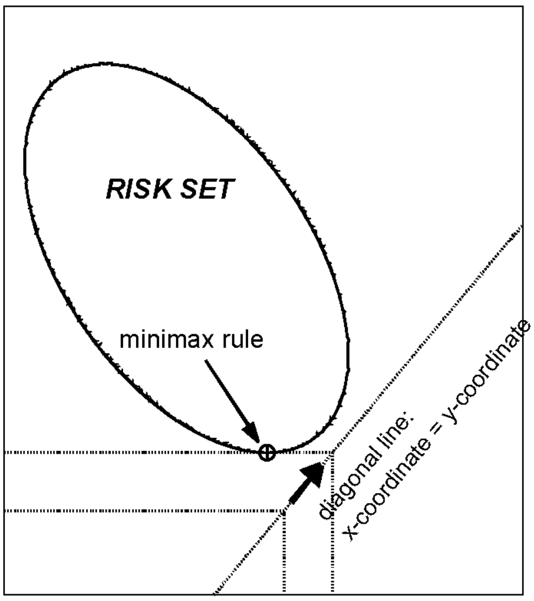
1.11.4 Admissible rules correspond to points on the "lower boundary" of $\mathcal{R}$, which consists of points (a, b) in $\mathcal{R}$ such that the region $\{(x, y) : x \leq a,\ y \leq b\}$ touches $\mathcal{R}$ at (a, b) only:



1.11.5 Minimax rules can be found as follows:

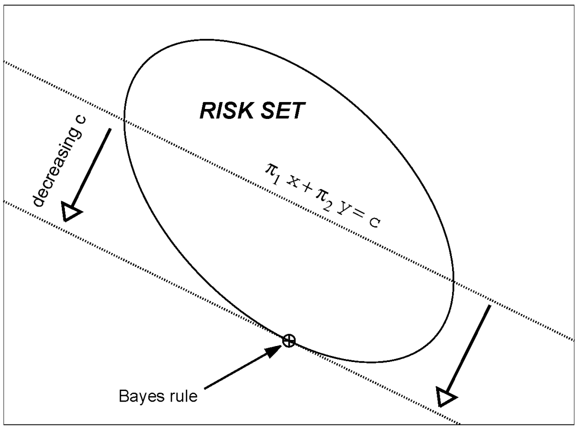
- consider the region $\{(x,y) : \max\{x,y\} \le c\}$ for a small c such that the region has no intersection with $\mathcal{R}$,
- move the region upwards by increasing c until it just touches the lower boundary of $\mathcal{R}$,
- the point at which the two regions touch identifies a minimax rule.





1.11.6 Given a prior weight function $\pi(\cdot)$: $\pi(\theta_1) = \pi_1 > 0$, $\pi(\theta_2) = \pi_2 > 0$,

every rule whose risk function lies on the line $\ell : \pi_1 x + \pi_2 y = c$ has the **same** Bayes risk c. Decreasing c means sliding ℓ downwards without altering its orientation. The point at which ℓ touches the lower boundary of $\mathcal{R}$ corresponds to the Bayes rule with respect to the prior π:



If ℓ touches $\mathcal{R}$ at more than one points, then any of those points corresponds to a Bayes rule, which is therefore not unique.

1.11.7 **Example §1.11.1** A single observation X is taken from $N(\theta, 1)$, where θ is known to belong to $\{0, 1\}$. Consider testing

$$H_0 : \theta = 0 \quad \text{vs} \quad H_1 : \theta = 1.$$

Under decision problem formulation,

- action space $\mathcal{A} = \{a_0, a_1\}$:

  $a_0 \longrightarrow$ accept H_0, $\quad a_1 \longrightarrow$ reject H_0;

- unit loss function L:

  $L(0, a_0) = L(1, a_1) = 0 \longleftarrow$ make the right decision,
  $L(0, a_1) = L(1, a_0) = 1 \longleftarrow$ make the wrong decision.

1. Bayes rules —
   For a prior weight function $\pi(\cdot)$ and a fixed $x \in \mathbb{R}$, minimise

$$L(0, a_j)\pi(0)\phi(x) + L(1, a_j)\pi(1)\phi(x-1) = \begin{cases} \pi(1)\phi(x-1), & j = 0, \\ \pi(0)\phi(x), & j = 1, \end{cases}$$

w.r.t. $j \in \{0,1\}$, where $\phi(u) = (2\pi)^{-1/2}\exp(-u^2/2)$ denotes the standard normal pdf.

Bayes rule: take action a_1 iff

$$\pi(1)\phi(x-1) > \pi(0)\phi(x), \quad \text{i.e. } x > \ln\frac{\pi(0)}{\pi(1)} + \frac{1}{2}.$$

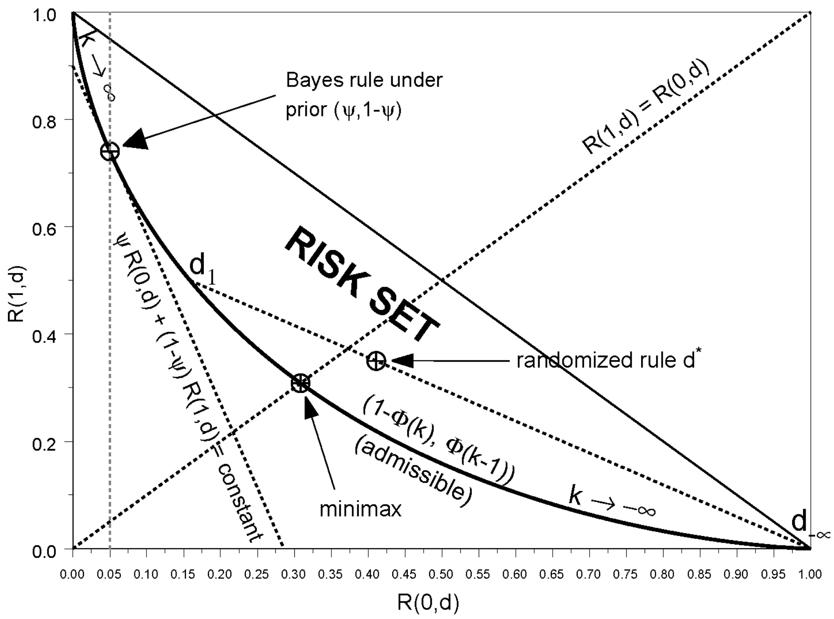
For any $k \in [-\infty, \infty]$, denote by d_k the decision rule $d_k(x) = \begin{cases} a_1, & x > k, \\ a_0, & x \le k. \end{cases}$

Thus, each d_k corresponds to a Bayes rule w.r.t. prior weight ratio $\pi(0)/\pi(1) = e^{k-1/2}$.

Risk function of d_k:

$$\begin{cases} R(0, d_k) = \mathbb{E}_0\big[L(0, d_k(X))\big] = \mathbb{P}_0(d_k(X) = a_1) = \mathbb{P}_0(X > k) = 1 - \Phi(k), \\ R(1, d_k) = \mathbb{E}_1\big[L(1, d_k(X))\big] = \mathbb{P}_1(d_k(X) = a_0) = \mathbb{P}_1(X \le k) = \Phi(k-1). \end{cases}$$

The diagram below displays the risk set $\big\{(R(0,d), R(1,d))\big\}$ of all randomized rules d formed by d_k, for $k \in [-\infty, \infty]$, where the lower arc of the risk set corresponds to the risks of d_k: $\big(1-\Phi(k), \Phi(k-1)\big)$.



2. Admissible rules —

Consider any fixed $k \in (-\infty, \infty)$. Suppose there exists a rule $d^\dagger$ strictly dominating d_k.

W.r.t. prior weight $\big(\pi(0), \pi(1)\big) = (e^{k-1/2}, 1)$, d_k is the Bayes rule and the Bayes risks of $d^\dagger$ and d_k satisfy

$$r(\pi, d^\dagger) = e^{k-1/2} R(0, d^\dagger) + R(1, d^\dagger) \geq r(\pi, d_k) = e^{k-1/2} R(0, d_k) + R(1, d_k),$$

so that

$$0 \geq e^{k-1/2}\big\{R(0, d^\dagger) - R(0, d_k)\big\} \geq R(1, d_k) - R(1, d^\dagger) \geq 0,$$

which implies $\big(R(0, d^\dagger), R(1, d^\dagger)\big) = \big(R(0, d_k), R(1, d_k)\big)$, contradicting the assumption that $d^\dagger$ strictly dominates d_k. Thus, d_k is admissible for $k \in (-\infty, \infty)$.

Consider next d_∞, which has risks $\big(R(0, d_\infty), R(1, d_\infty)\big) = (0, 1)$. If it is strictly dominated by some rule $d^\dagger$, then $d^\dagger$ must have risks $(0, r)$, for some $r \in [0, 1)$. Since $d^\dagger$ cannot strictly dominate d_k for any $k \in (-\infty, \infty)$, we must have $1 > r > R(1, d_k) = \Phi(k - 1)$ for all $k \in (-\infty, \infty)$, which is a contradiction. Thus, d_∞ is admissible. Similar arguments show that $d_{-\infty}$ is also admissible.

We conclude that d_k is admissible for any $k \in [-\infty, \infty]$.

3. Minimax rule –
The minimax rule corresponds to the intersection between the lower arc of the risk set and the diagonal line $R(0, d) = R(1, d)$, i.e. the rule d_{k^*} where k^* satisfies

$$1 - \Phi(k^*) = \Phi(k^* - 1) \quad \Rightarrow \quad k^* = 0.5 \quad \Rightarrow \quad \text{rule } d_{0.5} \text{ is minimax.}$$

Note that $d_{0.5}$ has risks $(0.3085, 0.3085)$.

4. Unbiased rules –
Note that

$$\mathbb{E}_\theta L(1 - \theta, d(X)) \geq R(\theta, d) \quad \text{for } \theta \in \{0, 1\}$$

iff

$$\begin{cases} \mathbb{P}_0(d(X) = a_0) \geq \mathbb{P}_0(d(X) = a_1) \\ \mathbb{P}_1(d(X) = a_1) \geq \mathbb{P}_1(d(X) = a_0) \end{cases}$$

iff

$$R(0, d) \leq 0.5 \quad \text{and} \quad R(1, d) \leq 0.5.$$

Thus, the unbiased rules correspond to all those rules with risks ≤ 0.5 at both $\theta = 0$ and $\theta = 1$.

5. Randomized rule —
   Consider the randomized rule d^* given by

$$d^* = \begin{cases} d_1 & \text{with probability } 0.7, \\ d_{-\infty} & \text{with probability } 0.3. \end{cases}$$

   On the risk set,

   d_1 corresponds to the point $\big(1 - \Phi(1), \Phi(0)\big) = (0.1587, 0.5)$,
   $d_{-\infty}$ corresponds to the point $\big(1 - \Phi(-\infty), \Phi(-\infty)\big) = (1, 0)$,
   d^* corresponds to the point

$$\big(R(0, d^*), R(1, d^*)\big) = 0.7(0.1587, 0.5) + 0.3(1, 0) = (0.4111, 0.35).$$

6. It is common to calibrate a hypothesis test to have a size (type I error probability) 5%. It would be interesting to ask:

   to what prior belief does this size 5% test correspond if viewed as a Bayes rule?

   To find the size 5% test d_k, we solve the equation

$$\text{size of test} = R(0, d_k) = 1 - \Phi(k) = 0.05 \quad \Rightarrow \quad k = \Phi^{-1}(0.95).$$

   The above d_k corresponds to a Bayes rule w.r.t. prior weight ratio

$$\pi(0)/\pi(1) = e^{k-1/2} = \exp\big\{\Phi^{-1}(0.95) - 1/2\big\}.$$

   Normalising the prior weights to have sum one, we get

$$\begin{aligned} \big(\pi(0), \pi(1)\big) = (\psi, 1 - \psi) &= \left(\frac{\exp\big\{\Phi^{-1}(0.95) - 1/2\big\}}{1 + \exp\big\{\Phi^{-1}(0.95) - 1/2\big\}}, \frac{1}{1 + \exp\big\{\Phi^{-1}(0.95) - 1/2\big\}}\right) \\ &= (0.75857, 0.24143). \end{aligned}$$

   The 5% size required of the test thus corresponds to a prior belief that places weights $(0.75857, 0.24143)$ on the two hypotheses (H_0, H_1).

1.11.8 Exercise:

- parameter space $\Theta = \{\theta_0, \theta_1\}$, action space $\mathcal{A} = \{a_0, a_1\}$,
- loss function $L(\theta, a)$:

$\theta\backslash a$	a_0	a_1
θ_0	-1	2
θ_1	2	1

- data $X \in \{0,1\}$ $\rightarrow$ model: $\mathbb{P}_{\theta_0}(X=1)=0.4, \quad \mathbb{P}_{\theta_1}(X=1)=0.9.$

1. Draw the risk set generated by all the non-randomized and randomized rules.
2. Identify the admissible rules.
3. Identify the minimax rules.
4. Identify the unbiased rules.
5. Identify the Bayes rules under the prior weights $\pi(\theta_0)=\pi_0 \geq 0$, $\pi(\theta_1)=\pi_1 \geq 0$, with $\pi_0+\pi_1>0$.