4
Statistical Decision Theory

The basic premise of Statistical Decision Theory is that we want to make inferences about the parameter of a family of distributions. So the starting point of this chapter is a family of distributions for the observables $\boldsymbol{Y} := (Y_1, \ldots, Y_n)$, of the general form

$$\boldsymbol{Y} \sim f(\cdot\,; \theta) \quad \text{for some } \theta \in \Omega,$$

where f is the 'model', θ is the 'parameter', and Ω the 'parameter space', just as in Chapter 3. Nothing in this chapter depends on whether $\boldsymbol{Y}$ is a scalar or a vector, and so I will write Y throughout. The parameter space Ω may be finite or non-finite, possibly non-countable; generally, though, I will treat it as finite, since this turns out to be much simpler. The value $f(y; \theta)$ denotes the probability of $Y \doteq y$ under family member θ. I will assume throughout this chapter that $f(y; \theta)$ is easily computed.

These basic premises, (i) that we are interested in the value of the parameter θ, and (ii) that $f(y; t)$ is easily computed, are both restrictive, as was discussed in Chapter 3. But in this chapter and the next we are exploring the challenges of Frequentist inference, which operates in a more restrictive domain than modern Bayesian inference.

4.1 General Decision Theory

There is a general theory of decision-making, of which Statistical Decision Theory is a special case. Here I outline the general theory, subject to one restriction which always holds for Statistical Decision Theory (to be introduced below). In general we should imagine the statistician applying decision theory on behalf of a client, but for simplicity of exposition I will assume the statistician is her own client.

There is a set of random quantities X with domain $\mathcal{X}$; as above I treat these as a scalar quantity, without loss of generality. The statistician contemplates a set of *actions*, $a \in \mathcal{A}$. Associated with each action is a consequence which depends on X. This is quantified in terms of a *loss function*, $L : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$, with larger values indicating worse consequences. Thus $L(a, x)$ is the loss incurred by the statistician if action a is taken and X turns out to be x.

Before making her choice of action, the statistician will observe $Y \in \mathcal{Y}$. Her choice should be some function of the value of Y, and this is represented as a *decision rule*, $\delta : \mathcal{Y} \to \mathcal{A}$. Of the many ways in which she might choose δ, one possibility is to minimise her expected loss, and this is termed the *Bayes rule*,

$$\delta^* := \operatorname*{argmin}_{\delta \in \mathcal{D}} \mathrm{E}\{L(\delta(Y), X)\},$$

where $\mathcal{D}$ is the set of all possible rules. The value $\mathrm{E}\{L(\delta(Y), X)\}$ is termed the *Bayes risk* of decision rule δ, and therefore the Bayes rule is the decision rule which minimises the Bayes risk.

There is a justly famous result which gives the explicit form for a Bayes rule. I will give this result under the restriction anticipated above, which is that the PMF $\mathrm{p}(x \mid y)$ does not depend on the choice of action. Decision theory can handle the more general case, but it is seldom appropriate for Statistical Decision Theory.

Theorem 4.1 (Bayes Rule Theorem, BRT)**.** *A Bayes rule satisfies*

$$\delta^*(y) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathrm{E}\{L(a, X) \mid Y \doteq y\} \tag{4.1}$$

whenever $y \in \operatorname{supp} Y$.[1]

[1] Recollect that $\operatorname{supp} Y$ is the subset of $\mathcal{Y}$ for which $\mathrm{p}(y) > 0$, termed the 'support' of Y.

This astounding result indicates that the minimisation of expected loss over the space of all functions from $\mathcal{Y}$ to $\mathcal{A}$ can be achieved by the pointwise minimisation over $\mathcal{A}$ of the expected loss conditional on $Y \doteq y$. It converts an apparently intractable problem into a simple one.

Proof. As usual, we take expectations to be completely coherent. Then the FTP (Thm 1.2) asserts the existence of a PMF for (X, Y), which we can factorise as

$$\mathrm{p}(x, y) = \mathrm{p}(x \mid y)\, \mathrm{p}(y)$$

using the notation and concepts from Chapter 1. Now take any $\delta \in \mathcal{D}$, for which

$$\begin{aligned}
\mathrm{E}\{L(\delta(Y), X)\} &= \textstyle\sum_y \sum_x L(\delta(y), x) \cdot \mathrm{p}(x \mid y)\, \mathrm{p}(y) && \text{by the FTP} \\
&\geq \textstyle\sum_y \left\{ \operatorname{argmin}_a \sum_x L(a, x)\, \mathrm{p}(x \mid y) \right\} \mathrm{p}(y) \\
&= \textstyle\sum_y \left\{ \sum_x L(\delta^*(y), x)\, \mathrm{p}(x \mid y) \right\} \mathrm{p}(y) && \text{from (4.1) and the CFTP, (3.6)} \\
&= \textstyle\sum_y \sum_x L(\delta^*(y), x) \cdot \mathrm{p}(x \mid y)\, \mathrm{p}(y) \\
&= \mathrm{E}\{L(\delta^*(Y), X)\} && \text{FTP again.}
\end{aligned}$$

Hence δ^* provides a lower bound on the expected loss, over all possible decision rules. Note that the sum over y can actually be over $\operatorname{supp} Y$ if there are y for which $\mathrm{p}(y) = 0$, which ensures that the conditional expectation inside the curly brackets is always well-defined. □

4.2 *Inference about parameters*

Now consider the special case of Statistical Decision Theory, in which inference is not about some random quantities X, but about the parameter θ. For simplicity I will assume that the parameter space is finite.[2] Furthermore, because nothing in this chapter depends on whether each element of the parameter space is a scalar or a vector, I will treat θ as a scalar and write

$$\Omega := \{\theta_1, \ldots, \theta_k\},$$

rather than my usual notation for elements of sets, which is to use superscripts in parentheses (i.e. I will write θ_j rather than $\theta^{(j)}$). A word about notation. I will write 'θ_j' to indicate one of the elements of Ω, and 'θ' to indicate the unknown index of Ω (Frequentist) or the random variable with realm Ω (Bayesian). This is clearer than letting one symbol represent several different things, which is unfortunately a common practice.

[2] See the comment after Thm 4.2 for extensions.

The three types of inference about θ are (i) point estimation, (ii) set estimation, and (iii) hypothesis testing. It is a great conceptual and practical simplification that Statistical Decision Theory distinguishes between these three types simply according to their action sets, which are:

Type of inference	Action set $\mathcal{A}$
Point estimation	The parameter space, Ω. See Sec. 4.4.
Set estimation	The set of all subsets of Ω, denoted 2^Ω. See Sec. 4.5.
Hypothesis testing	A specified partition of Ω, denoted $\mathcal{P}$ below. See Sec. 4.6.

One challenge for Statistical Decision Theory is that finding the Bayes rule requires specifying a *prior distribution* over Ω, which I will denote

$$\boldsymbol{\pi} := (\pi_1, \ldots, \pi_k) \in \mathbb{S}^{k-1}$$

where $\mathbb{S}^{k-1}$ is the $(k-1)$-dimensional unit simplex, see (1.4). Applying the BRT (Thm 4.1),

$$\begin{aligned} \delta^*(y) &= \operatorname*{argmin}_{a \in \mathcal{A}} \mathrm{E}\{L(a,\theta) \mid Y \doteq y\} \\ &= \operatorname*{argmin}_{a \in \mathcal{A}} \textstyle\sum_j L(a,\theta_j) \cdot \mathrm{p}(\theta_j \mid y) \qquad \text{by the CFTP,} \end{aligned}$$

where the conditional PMF is

$$\mathrm{p}(\theta_j \mid y) = \frac{f(y;\theta_j) \cdot \pi_j}{\Pr(Y \doteq y)} = \frac{f(y;\theta_j) \cdot \pi_j}{\sum_{j'} f(y;\theta_{j'}) \cdot \pi_{j'}} \tag{4.2}$$

by Bayes's Theorem. So the Bayes rule will not be an attractive way to choose a decision rule for Frequentist statisticians, who are reluctant to specify a prior distribution for θ. These statisticians need a different approach to choosing a decision rule.

The accepted approach for Frequentist statisticians is to narrow the set of possible decision rules by ruling out those that are obviously bad. Define the *risk function* for rule δ as

$$\begin{aligned} R(\delta, \theta_j) &:= \mathrm{E}\{L(\delta(Y), \theta_j); \theta_j\} \\ &= \textstyle\sum_y L(\delta(y), \theta_j) \cdot f(y; \theta_j). \end{aligned} \tag{4.3}$$

That is, $R(\delta, \theta_j)$ is the expected loss from rule δ when $\theta = \theta_j$. A decision rule δ *dominates* another rule δ' exactly when

$$R(\delta, \theta_j) \leq R(\delta', \theta_j) \quad \text{for all } \theta_j \in \Omega,$$

with a strict inequality for at least one $\theta_j \in \Omega$. If you had both δ and δ', you would never want to use δ'.[3] A decison rule is *admissible* exactly when it is not dominated by any other rule; otherwise it is *inadmissible*. So the accepted approach is to reduce the set of possible decision rules under consideration by only using admissible rules.

[3] Here I am assuming that all other considerations are the same in the two cases: e.g. $\delta(y)$ and $\delta'(y)$ take about the same amount of resource to compute.

It is hard to disagree with this approach, although one wonders how big the set of admissible rules will be, and how easy it is to enumerate the set of admissible rules in order to choose between them. This is the subject of Sec. 4.3. To summarise,

Theorem 4.2 (Wald's Complete Class Theorem, CCT)**.** *In the case where both the action set $\mathcal{A}$ and the parameter space Ω are finite, a decision rule δ is admissible if and only if it is a Bayes rule for some prior distribution $\boldsymbol{\pi}$ with strictly positive values.*

There are generalisations of this theorem to non-finite realms for Y, non-finite action sets, and non-finite parameter spaces; however, the results are highly technical. See Schervish (1995, ch. 3), Berger (1985, chs 4, 8), and Ghosh and Meeden (1997, ch. 2) for more details and references to the original literature.

So what does the CCT say? First of all, if you select a Bayes rule according to some prior distribution $\boldsymbol{\pi} \gg \mathbf{0}$ then you cannot ever choose an inadmissible decision rule.[4] So the CCT states that there is a very simple way to protect yourself from choosing an inadmissible decision rule. Second, if you cannot produce a $\boldsymbol{\pi} \gg \mathbf{0}$ for which your proposed rule δ is a Bayes Rule, then you cannot show that δ is admissible.

[4] Here I am using a fairly common notion for vector inequalities. If all components of $\boldsymbol{x}$ are non-negative, I write $\boldsymbol{x} \geq \mathbf{0}$. It in addition at least one component is positive, I write $\boldsymbol{x} > \mathbf{0}$. If all components are positive I write $\boldsymbol{x} \gg \mathbf{0}$. For comparing two vectors, $\boldsymbol{x} \geq \boldsymbol{y}$ exactly when $\boldsymbol{x} - \boldsymbol{y} \geq \mathbf{0}$, and so on.

But here is where you must pay close attention to logic. Suppose that δ' is inadmissible and δ is admissible. It does not follow that δ dominates δ'. So just knowing of an admissible rule does not mean that you should abandon your inadmissible rule δ'. You can argue that although you know that δ' is inadmissible, you do not know of a rule which dominates it. All you know, from the CCT, is the family of rules within which the dominating rule must live: it will be a Bayes rule for some $\boldsymbol{\pi} \gg \mathbf{0}$. This may seem a bit esoteric, but it is crucial in understanding modern parametric inference. Statisticians sometimes use inadmissible rules according to standard loss functions. They can argue that yes, their rule δ is or

may be inadmissible, which is unfortunate, but since the identity of the dominating rule is not known, it is not wrong to go on using δ. Nevertheless, it would be better to use an admissible rule.

4.3 *The Complete Class Theorem*

This section can be skipped once the previous section has been read. But it describes a very beautiful result, Thm 4.2 above, originally due to an iconic figure in Statistics, Abraham Wald.[5] I assume throughout this section that all sets are finite: the realm $\mathcal{Y}$, the action set $\mathcal{A}$, and the parameter space Ω.

[5] For his tragic story, see https://en.wikipedia.org/wiki/Abraham_Wald.

The CCT is if-and-only-if. Let $\boldsymbol{\pi}$ be any prior distribution on Ω. Both branches use a simple result that relates the Bayes Risk of a decision rule δ to its Risk Function:

$$\begin{aligned} \mathrm{E}\{L(\delta(Y),\theta)\} &= \textstyle\sum_j \mathrm{E}\{L(\delta(Y),\theta_j);\theta_j\}\cdot\pi_j \quad \text{by (1.28) and (1.26)} \\ &= \textstyle\sum_j R(\delta,\theta_j)\cdot\pi_j. \end{aligned} \tag{†}$$

The first branch is easy to prove.

Theorem 4.3. *If δ is a Bayes rule for prior distribution $\boldsymbol{\pi} \gg \mathbf{0}$, then it is admissible.*

Proof. By contradiction. Suppose that the Bayes rule δ is not admissible; i.e. there exists a rule δ' which dominates it. In this case

$$\begin{aligned} \mathrm{E}\{L(\delta(Y),\theta)\} &= \textstyle\sum_j R(\delta,\theta_j)\cdot\pi_j && \text{from (†)} \\ &> \textstyle\sum_j R(\delta',\theta_j)\cdot\pi_j && \text{if } \boldsymbol{\pi} \gg \mathbf{0} \\ &= \mathrm{E}\{L(\delta'(Y),\theta)\} \end{aligned}$$

and hence δ cannot have been a Bayes rule, because δ' has a smaller expected loss. The strict inequality holds if δ' dominates δ *and* $\boldsymbol{\pi} \gg \mathbf{0}$. Without it, we cannot deduce a contradiction. □

The second branch of the CCT is harder to prove. The proof uses one of the great theorems in Mathematics, the Supporting Hyperplane Theorem (SHT, given below in Thm 4.5).

Theorem 4.4. *If δ is admissible, then it is a Bayes rule for some prior distribution $\boldsymbol{\pi} \gg \mathbf{0}$.*

For a given loss function L and model f, construct the *risk matrix*,

$$R_{ij} := R(\delta_i,\theta_j)$$

over the set of all decision rules. If there are m decision rules althogether (m is finite because $\mathcal{Y}$ and $\mathcal{A}$ are both finite), then R represents m points in k-dimensional space, where k is the cardinality of Ω.

Now consider *randomised rules*, indexed by $\boldsymbol{w} \in \mathbb{S}^{m-1}$. For randomised rule $\boldsymbol{w}$, actual rule δ_i is selected with probability w_i.

The risk for rule $\boldsymbol{w}$ is

$$\begin{aligned} R(\boldsymbol{w}, \theta_j) &:= \textstyle\sum_i \mathrm{E}\{L(\delta_i(Y), \theta_j); \theta_j\} \cdot w_i \qquad \text{by the (1.28)} \\ &= \textstyle\sum_i R(\delta_i, \theta_j) \cdot w_i. \end{aligned}$$

If we also allow randomised rules—and there is no reason to disallow them, as the original rules are all still available as special cases—then the set of risks for all possible randomised rules is the *convex hull* of the rows of the risk matrix R, denoted $[R] \subset \mathbb{R}^k$, and termed the *risk set*.[6] We can focus on the risk set because every point in $[R]$ corresponds to at least one choice of $\boldsymbol{w} \in \mathbb{S}^{m-1}$.

[6] If $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ are m points in $\mathbb{R}^k$, then the convex hull of these points is the set of $\boldsymbol{x} \in \mathbb{R}^k$ for which $\boldsymbol{x} = w_1\boldsymbol{x}^{(1)} + \cdots + w_m\boldsymbol{x}^{(m)}$ for some $\boldsymbol{w} \in \mathbb{S}^{m-1}$.

Only a very small subset of the risk set will be admissible. A point $\boldsymbol{r} \in [R]$ is admissible exactly when it is on the lower boundary of $[R]$. More formally, define the 'quantant' of $\boldsymbol{r}$ to be the set

$$Q(\boldsymbol{r}) := \{\boldsymbol{x} \in \mathbb{R}^k : \boldsymbol{x} \leq \boldsymbol{r}\}$$

(see footnote 4). By definition, $\boldsymbol{r}$ is dominated by every $\boldsymbol{r}'$ for which $\boldsymbol{r}' \in Q(\boldsymbol{r}) \setminus \{\boldsymbol{r}\}$. So $\boldsymbol{r} \in [R]$ is admissible exactly when $[R] \cap Q(\boldsymbol{r}) = \{\boldsymbol{r}\}$. The set of $\boldsymbol{r}$ for satisfying this condition is the lower boundary of $[R]$, denoted $\lambda(R)$.

Now we have to show that every point in $\lambda(R)$ is a Bayes rule for some $\boldsymbol{\pi} \gg \mathbf{0}$. For this we use the SHT, the proof of which can be found in any book on convex analysis.

Theorem 4.5 (Supporting Hyperplane Theorem, SHT)**.** *Let $[R]$ be a convex set in $\mathbb{R}^k$, and let $\boldsymbol{r}$ be a point on the boundary of $[R]$. Then there exists an $\boldsymbol{a} \in \mathbb{R}^k$ not equal to $\mathbf{0}$ such that*

$$\boldsymbol{a}^T\boldsymbol{r} = \min_{\boldsymbol{r}' \in [R]} \boldsymbol{a}^T\boldsymbol{r}'.$$

So let $\boldsymbol{r} \in \lambda(R)$ be any admissible risk. Let $\boldsymbol{a} \in \mathbb{R}^k$ be the coefficients of its supporting hyperplane. Because $\boldsymbol{r}$ is on the lower boundary of $[R]$, $\boldsymbol{a} \gg \mathbf{0}$.[7] Set

[7] Proof: because if $\boldsymbol{r}$ is on the lower boundary, the slightest decrease in any component of $\boldsymbol{r}$ must move $\boldsymbol{r}$ outside $[R]$.

$$\pi_j := \frac{a_j}{\sum_{j'} a_{j'}} \quad j = 1, \ldots, k,$$

so that $\boldsymbol{\pi} \in \mathbb{S}^{k-1}$ and $\boldsymbol{\pi} \gg \mathbf{0}$. Then the SHT asserts that

$$\textstyle\sum_j r_j \cdot \pi_j \leq \sum_j r'_j \cdot \pi_j \quad \text{for all } \boldsymbol{r}' \in [R]. \qquad (\ddagger)$$

Let $\boldsymbol{w}$ be any randomised strategy with risk $\boldsymbol{r}$. Since $\sum_j r_j \cdot \pi_j$ is the expected loss of $\boldsymbol{w}$ (see †), (‡) asserts that $\boldsymbol{w}$ is a Bayes rule for prior distribution $\boldsymbol{\pi}$. Because $\boldsymbol{r}$ was an arbitrary point on $\lambda(R)$, and hence an arbitrary admissible rule, this completes the proof of Thm 4.4.

4.4 *Point estimation*

For point estimation the action space is $\mathcal{A} = \Omega$, and the loss function $L(\theta_j, \theta_{j'})$ represents the (negative) consequence of choosing θ_j

as a point estimate of θ, when the 'true' value of θ is $\theta_{j'}$. Note that this is questionable, if θ does not correspond to an operationally-defined quantity such as the population mean. If the model and its parameters are a convenient abstraction, then there is no 'true' value. I return to this issue in **??**.

There will be situations where an obvious loss function $L : \Omega \times \Omega \to \mathbb{R}$ presents itself. But not very often. Hence the need for a generic loss function which is acceptable over a wide range of situations. A natural choice in the very common case where Ω is a convex subset of $\mathbb{R}^d$ is a *convex loss function*,[8]

$$L(\theta_j, \theta_{j'}) \leftarrow h(\theta_j - \theta_{j'}) \tag{4.4}$$

where $h : \mathbb{R}^d \to \mathbb{R}$ is a smooth non-negative convex function with $h(\mathbf{0}) = 0$. This type of loss function asserts that small errors are much more tolerable than large ones. One possible further restriction would be that h is an even function.[9] This would assert that under-prediction incurs the same loss as over-prediction. There are many situations where this is *not* appropriate, but in these cases a generic loss function should be replaced by a more specific one.

[8] If Ω is convex then it is uncountable, and hence definitely not finite. But this does not have any disturbing implications for the following analysis.

[9] I.e. $h(-\boldsymbol{x}) = h(\boldsymbol{x})$.

Proceeding further along the same lines, an even, differentiable and strictly convex loss function can be approximated by a *quadratic loss function*,

$$h(\boldsymbol{x}) \propto \boldsymbol{x}^T Q \boldsymbol{x} \tag{4.5}$$

where Q is a symmetric positive-definite $d \times d$ matrix. This follows directly from a Taylor series expansion of h around $\mathbf{0}$:

$$h(\boldsymbol{x}) = 0 + 0 + \tfrac{1}{2}\boldsymbol{x}^T \nabla^2 h(\mathbf{0})\, \boldsymbol{x} + 0 + O(\|\boldsymbol{x}\|^4)$$

where the first 0 is because $h(\mathbf{0}) = 0$, the second 0 is because $\nabla h(\mathbf{0}) = 0$ since h is minimised at $\boldsymbol{x} = \mathbf{0}$, and the third 0 is because h is an even function. $\nabla^2 h$ is the *hessian matrix* of second derivatives, and it is symmetric by construction, and positive definite at $\boldsymbol{x} = \mathbf{0}$, if h is strictly convex and minimised at $\mathbf{0}$.

In the absence of anything more specific the quadratic loss function is the generic loss function for point estimation. Hence the following result is widely applicable.

Theorem 4.6. *Under a quadratic loss function, the Bayes rule for point prediction is the conditional expectation*

$$\delta^*(y) = \mathrm{E}(\theta \mid Y \doteq y).$$

A Bayes rule for a point estimation is known as a *Bayes estimator*. Note that although the matrix Q is involved in defining the quadratic loss function in (4.5), it does not influence the Bayes estimator. Thus the Bayes estimator is the same for an uncountably large class of loss functions. Depending on your point of view, this is either its most attractive or its most disturbing feature.

Proof. Here is a proof that does not involve differentiation. The BRT (Thm 4.1) asserts that

$$\delta^*(y) = \operatorname*{argmin}_{t \in \Omega} \mathrm{E}\{L(t, \theta) \mid Y \doteq y\}. \tag{4.6}$$

So let $\psi(y) := \mathrm{E}(\theta \mid Y \doteq y)$. For simplicity, treat θ as a scalar. Then

$$\begin{aligned} L(t,\theta) &\propto (t-\theta)^2 \\ &= (t-\psi(y)+\psi(y)-\theta)^2 \\ &= (t-\psi(y))^2 + 2(t-\psi(y))(\psi(y)-\theta) + (\psi(y)-\theta)^2. \end{aligned}$$

Take expectations conditional on $Y \doteq y$ to get

$$\mathrm{E}\{L(t,\theta) \mid Y \doteq y\} \propto (t-\psi(y))^2 + \mathrm{E}\{(\psi(y)-\theta)^2 \mid Y \doteq y\}. \quad (\dagger)$$

Only the first term contains t, and this term is minimised over t by setting $t \leftarrow \psi(y)$, as was to be shown.

The extension to vector θ with loss function (4.5) is straightforward, but involves more ink. It is crucial that Q in (4.5) is positive definite, because otherwise the first term in (†), which becomes $(t-\psi(y))^T Q\,(t-\psi(y))$, is not minimised if and only if $t = \psi(y)$. □

Note that the same result holds in the more general case of a point prediction of random quantities X based on observables Y: under quadratic loss, the Bayes estimator is $\mathrm{E}(X \mid Y \doteq y)$.

* * *

Now apply the CCT (Thm 4.2) to this result. For quadratic loss, a point estimator for θ is admissible if and only if it is the conditional expectation with respect to some prior distribution $\boldsymbol{\pi} \gg \mathbf{0}$.[10] Among the casualties of this conclusion is the Maximum Likelihood Estimator (MLE),

$$\hat{\theta}(y) := \operatorname*{argmax}_{t \in \Omega} f(y;t).$$

Stein's paradox showed that under quadratic loss, the MLE is not admissible in the case of a Multinormal distribution with known variance, by producing an estimator which dominated it. This result caused such consternation when first published that it might be termed 'Stein's bombshell'. See Efron and Morris (1977) for more details, and Samworth (2012) for an accessible proof. Interestingly, the MLE is still the dominant point estimator in applied statistics, even though its admissibility under quadratic loss is questionable.

[10] This is under the conditions of Thm 4.2, or with appropriate extensions of them in the non-finite cases.

4.5 *Set estimators*

For set estimation the action space is $\mathcal{A} = 2^\Omega$, and the loss function $L(C,\theta_j)$ represents the (negative) consequences of choosing $C \subset \Omega$ as a set estimate of θ, when the 'true' value of θ is θ_j. The points made at the start of Sec. 4.4 also apply here; see **??**.

There are two contrary requirements for set estimators of θ. We want the sets to be small, but we also want them to contain θ. There is a simple way to represent these two requirements as a loss function, which is to use

$$L(C,t) \leftarrow |C| + \kappa \cdot (1 - \mathbb{1}_{t \in C}) \quad \text{for some } \kappa > 0 \qquad (4.7a)$$

where $|C|$ is the cardinality of C.[11] The value of κ controls the

[11] Here and below I am treating Ω as countable, for simplicity.

trade-off between the two requirements. If $\kappa \downarrow 0$ then minimising the expected loss will always produce the empty set. If $\kappa \uparrow \infty$ then minimising the expected loss will always produce Ω. For κ in-between, the outcome will depend on beliefs about Y and the value y.

It is important to note that the crucial result, Thm 4.7 below, continues to hold for the much more general set of loss functions

$$L(C, t) \leftarrow g(|C|) + h(1 - \mathbb{1}_{t \in C}) \tag{4.7b}$$

where g is non-decreasing and h is strictly increasing. This is a large set of loss functions, which should satisfy most statisticians who do not have a specific loss function already in mind.

For point estimators there was a simple characterisation of the Bayes rule for quadratic loss functions (Thm 4.6). For set estimators the situation is not so simple. However, for loss functions of the form (4.7) there is a simple necessary condition for a rule to be a Bayes rule.

Theorem 4.7. *Under a loss function of the form* (4.7), $\delta : \mathcal{Y} \to 2^{\Omega}$ *is a Bayes rule* only if*:*

$$\forall y, \forall \theta_j \in \delta(y) \quad \theta_{j'} \notin \delta(y) \implies \mathrm{p}(\theta_{j'} \mid y) \leq \mathrm{p}(\theta_j \mid y) \tag{4.8}$$

where $\mathrm{p}(\theta_j \mid y)$ *was defined in* (4.2)*.*

Proof. The proof is by contradiction. Fix y and let $C \leftarrow \delta(y)$. We show that if (4.8) does not hold, then C does not minimise the expected loss conditional on $Y \doteq y$, as required by the BRT (Thm 4.1). Now,

$$\mathrm{E}\{L(C, \theta) \mid Y \doteq y\} = |C| + \kappa \cdot (1 - \Pr\{\theta \in C \mid Y \doteq y\}) \tag{†}$$

using (4.7a), for simplicity. Let $\theta_j \in C$, and let $\theta_{j'} \notin C$, but with $\mathrm{p}(\theta_{j'} \mid y) > \mathrm{p}(\theta_j \mid y)$, contradicting (4.8). In this case, θ_j and $\theta_{j'}$ could be swapped in C, leaving the first term in (†) the same, but decreasing the second. Hence C could not have minimised the expected loss conditional on $Y \doteq y$, and δ could not have been a Bayes rule. □

To give condition (4.8) a simple name, I will refer to it as the 'level set' property, since it almost asserts that $\delta(y)$ must always be a level set of the probabilities $\{\mathrm{p}(\theta_j \mid Y \doteq y) : \theta_j \in \Omega\}$.[12] Chapter 5 provides a tighter definition of this property.

Now relate this result to the CCT (Thm 4.2). First, Thm 4.7 asserts that δ having the level set property for all y is necessary (but not sufficient) for δ to be a Bayes rule for loss functions of the form (4.7). Second, the CCT asserts that being a Bayes rule is a necessary (but not sufficient) condition for δ to be admissible.[13] So unless δ has the level set property for all y then it is impossible for δ to be admissible for loss functions of the form (4.7). This result is embodied in Bayesian approaches to set estimation for θ.

[12] I can only say 'almost' because the property is ambiguous about the inclusion of θ_j and $\theta_{j'}$ for which $\mathrm{p}(\theta_j \mid Y \doteq y) = \mathrm{p}(\theta_{j'} \mid Y \doteq y)$, while a level set is unambiguous.

[13] As before, terms and conditions apply in the non-finite cases.

Definition 9 (High Posterior Probability (HPP) set)**.** *The rule* $\delta : \mathcal{Y} \to 2^\Omega$ *is a level-*$(1-\alpha)$ *HPP set exactly when it is the smallest set for which* $\Pr(\theta \in \delta(y) \mid Y \doteq y) \geq 1-\alpha$.

This definition acknowledges that for a given level, say $(1-\alpha) \leftarrow 0.95$, it might not be possible to find a set C for which $\Pr(\theta \in C \mid Y \doteq y) = 0.95$, so instead we settle for the smallest set whose probability is at least 0.95.[14] The requirement that $\delta(y)$ is the smallest set automatically ensures that it satisfies the level set property.

[14] If Ω is uncountable, then it is usually possible to hit 0.95 exactly, in which case C is an 'exact' 95% *High Posterior Density (HPD)* set.

Now it is *not* the case that the collection of, say, level 0.95 HPP sets (taken over all $y \in \mathcal{Y}$) is consistent with the Bayes rule for (4.7) for some specified κ. So the level 0.95 HPP sets cannot claim to be a Bayes rule for (4.7). But they satisfy the necessary condition to be admissible for (4.7), which is a good start. Moreover, the level of an HPP set is much easier to interpret than the value of κ.

Things are trickier for Frequentist approaches, which must proceed without a prior distribution for $\theta \in \Omega$, and thus cannot compute $\mathrm{p}(\theta_j \mid Y \doteq y)$. Frequentist approaches to set estimation are based on confidence procedures, which are covered in detail in Chapter 5. We can make a strong recommendation based on Thm 4.7. Denote the Frequentist model as $\{f, \Omega\}$, for which a prior distribution π would imply

$$\mathrm{p}(\theta_j \mid Y \doteq y) = \frac{f(y;\theta_j) \cdot \pi_j}{\sum_{j'} f(y;\theta_{j'}) \cdot \pi_{j'}}.$$

Clearly, if $\pi_j = 1/k$ for all j, then $\mathrm{p}(\theta_j \mid Y \doteq y) \propto f(y;\theta_j)$, which which implies that they have the same level sets. So the recommendation is

- Base confidence procedures on level sets of $\{f(y;\theta_j) : \theta_j \in \Omega\}$.

This recommendation ensures that confidence procedures satisfy the necessary condition to be admissible for (4.7). I will be adopting this recommendation in Chapter 5.

4.6 Hypothesis tests

For hypothesis tests, the action space is a partition of Ω, denoted

$$\mathcal{H} := \{H_0, H_1, \ldots, H_d\}.$$

Each element of $\mathcal{H}$ is termed a *hypothesis*; it is traditional to number the hypotheses from zero. The loss function $L(H_i, \theta_j)$ represents the (negative) consequences of choosing element H_i, when the 'true' value of θ is θ_j. It would be usual for the loss function to satisfy

$$\theta_j \in H_i \implies L(H_i, \theta_j) = \min_{i'} L(H_{i'}, \theta_j)$$

on the grounds that an incorrect choice of element should never incur a smaller loss than the correct choice.

I will be quite cavalier about hypothesis tests. If the statistician has a complete loss function, then the CCT (Thm 4.2) applies,

a $\pi \gg \mathbf{0}$ must be found, and there is nothing more to be said. The famous *Neyman-Pearson (NP) Lemma* is of this type. It has $\Omega = \{\theta_0, \theta_1\}$, with $H_i = \{\theta_i\}$, and loss function

L	θ_0	θ_1
H_0	0	ℓ_1
H_1	ℓ_0	0

with $\ell_0, \ell_1 > 0$. The NP Lemma asserts that a decision rule for choosing between H_0 and H_1 is admissible if and only if it has the form

$$\frac{f(y;\theta_0)}{f(y;\theta_1)} \begin{cases} < c & \text{choose } H_1 \\ = c & \text{toss a coin} \\ > c & \text{choose } H_0 \end{cases}$$

for some $c > 0$. This is just the CCT (Thm 4.2).[15]

[15] In fact, $c = (\pi_1/\pi_0) \cdot (\ell_1/\ell_0)$, where (π_0, π_1) is the prior probability for which $\pi_1 = 1 - \pi_0$.

The NP Lemma is particularly simple, corresponding to a choice in a family with only two elements. In situations more complicated than this, it is extremely challenging and time-consuming to specify a loss function. And yet statisticians would still like to choose between hypotheses, in decision problems whose outcome does not seem to justify the effort required to specify the loss function.[16]

[16] Just to be clear, *important* decisions should not be based on cut-price procedures: an important decision warrants the effort required to specify a loss function.

There is a generic loss function for hypothesis tests, but it is hardly defensible. The *o-1 ('zero-one') loss function* is

$$L(H_i, \theta_j) \leftarrow 1 - \mathbb{1}_{\theta_j \in H_i},$$

i.e., zero if θ_j is in H_i, and one if it is not. Its Bayes rule is to select the hypothesis with the largest conditional probability. It is hard to think of a reason why the 0-1 loss function would approximate a wide range of actual loss functions, unlike in the cases of generic loss functions for point estimation and set estimation. This is not to say that it is wrong to select the hypothesis with the largest conditional probability; only that the 0-1 loss function does not provide a very compelling reason.

* * *

There is another approach which has proved much more popular. In fact, it is the dominant approach to hypothesis testing. This is to co-opt the theory of set estimators, for which there *is* a defensible generic loss function, which has strong implications for the selection of decision rules (see Sec. 4.5). The statistician can use her set estimator $\delta : \mathcal{Y} \to 2^\Omega$ to make at least some distinctions between the members of $\mathcal{H}$, on the basis of the value of the observable, y^{obs}:

- 'Accept' H_i exactly when $\delta(y^{\text{obs}}) \subset H_i$,
- 'Reject' H_i exactly when $\delta(y^{\text{obs}}) \cap H_i = \emptyset$,
- 'Undecided' about H_i otherwise.

Note that these three terms are given in scare quotes, to indicate that they acquire a technical meaning in this context. We do not use

the scare quotes in practice, but we always bear in mind that we are not "accepting H_i" in the vernacular sense, but simply asserting that $\delta(y^{\text{obs}}) \subset H_i$ for our particular choice of δ.

Looking at the three options above, there are two classes of outcome. If we accept H_i then we must reject all of the other hypotheses. But if we are undecided about H_i then we cannot accept any hypothesis. One very common case is where $\mathcal{H} = \{H_0, H_1\}$, which is known as *Null Hypothesis Significance Testing (NHST)*, where H_0 is the *null hypothesis* and H_1 is the *alternative hypothesis*. There are two versions of NHST. In the first, known as a *two-sided test* (or 'two-tailed test'), the H_0 is a tiny subset of Ω, too small for $\delta(y^{\text{obs}})$ to get inside. Therefore it is impossible to accept H_0, and all that we can do is reject H_0 and accept H_1, or be undecided. In the second case, known as a *one-sided test* (or 'one-tailed test'), H_0 is a sizeable subset of Ω, and then it is possible to accept H_0 and reject H_1.

For example, suppose that $Y \sim f(\cdot\,; \mu, \sigma^2)$ where μ and σ^2 are respectively the expectation and variance of X, and $(\mu, \sigma^2) \in \mathbb{R}^2_{++}$. Consider two different NHSTs:

Test A	Test B
$H_0 : \kappa = c$	$H_0 : \kappa \geq c$
$H_1 : \kappa \neq c$	$H_1 : \kappa < c$

where $\kappa := \sigma/\mu \in \mathbb{R}_{++}$, known as the 'coefficient of variation', and c is some specified constant. Test A is a one-sided test, in which it is impossible to accept H_0, and so there are only two outcomes: to reject H_0, or to be undecided, which is usually termed 'fail to reject H_0'. Test B is a two-sided test in which we can accept H_0 and reject H_1, or accept H_1 and reject H_0, or be undecided.

In applications we usually want to do a one-sided test. For example, if μ is the performance of a new treatment relative to a control, then we can be fairly sure *a priori* that $\mu = 0$ is false: different treatments seldom have identical effects. What we want to know is whether the new treatment is worse or better than the control: i.e. we want $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. In this case we can find in favour of H_0, or in favour of H_1, or be undecided. In a one-sided test, it would be sensible to push the upper bound of H_0 above $\mu = 0$ to some value $\mu_0 > 0$, which is the *minimial clinically significant difference (MCSD)*.

NHST is practiced mainly by Frequentist statisticians, and so I will continue in a Frequentist vein. In the Frequentist approach, it is conventional to use a 95% confidence set as the set estimator for hypothesis testing. Other levels, notably 90% and 99%, are occasionally used. If H_0 is rejected using a 95% confidence set, then this is reported as "H_0 is rejected at a significance level of 5%" (occasionally 10% or 1%). Confidence sets are covered in detail in Chapter 5.

This confidence set approach to hypothesis testing seems quite clear-cut, but we must end on a note of caution. First, the statisti-

cian has not solved the decision problem of choosing an element of $\mathcal{H}$. She has solved a different problem. Based on a set estimator, she may reject H_0 on the basis of y^{obs}, but that does not mean she should proceed as though H_0 is false. This would require her to solve the correct decision problem, for which she would have to supply a loss function. So, first caution:

- Rejecting H_0 is not the same as deciding that H_0 is false. Significance tests do not solve decision problems.

Second, loss functions of the form (4.7) may be generic, but that does not mean that there is only one 95% confidence procedure.[17] As Chapter 5 will show, there are an uncountable number of ways of constructing a 95% confidence procedure. In fact, there are an uncountable number of ways of constructing a 95% confidence procedure based on level sets of the likelihood function. So the statistician still needs to make and to justify two subjective choices, leading to the second caution:

- Accepting or rejecting a hypothesis is contingent on the choice of confidence procedure, as well as on the level.

[17] The same point can be made about 95% HPP sets, for which there is one for each prior distribution over Ω.