3
Statistical Decision Theory

3.1 Introduction

The basic premise of Statistical Decision Theory is that we want to make inferences about the parameter of a family of distributions (see Section 1.3). So the starting point of this chapter is a model for the observables $Y \in \mathcal{Y}$ of the general form

$$\mathcal{E} = \{\mathcal{Y}, \Omega, f\},$$

just as in Chapter 1 and Chapter 2. The value $f(y; \theta)$ denotes the probability that $Y = y$ under family member $\theta \in \Omega$, where θ is the parameter, and Ω is the parameter space. I will stick with my convention that $\mathcal{Y}$ is countable and Ω is uncountably infinite. I will assume throughout this chapter that $f(y; \theta)$ is easy to evaluate (see Section 1.2).

We accept as our working hypothesis that $\mathcal{E}$ is true (see Section 1.1), so that inference is learning about Θ, the true value of the parameter. More precisely, we would like to understand how to construct the 'Ev' function from Chapter 2, in such a way that it reflects our needs, which will vary from application to application. Statistical Decision Theory allows us to select an 'Ev' which is suitable for the type of inference we want to make, and which reflects the consequence of making a poor inference.

The the set of possible inferences is termed the *action set*, $\mathcal{A}$, with typical element a. The consequence of making a poor inference is specified as the *loss function* $L : \mathcal{A} \times \Omega \to \mathbb{R}$, with larger values indicating worse consequences. Thus $L(a, \theta)$ is the loss incurred by the statistician if action a is taken and Θ turns out to be θ. I will assume, as is natural, that L is bounded, but many results below also hold in the more general case.

Before making her choice of action, the statistician will observe y, a value for Y. Her choice should be some function of the value y, and this is represented as a *decision rule*, $\delta : \mathcal{Y} \to \mathcal{A}$. As we are taking the model $\mathcal{E}$ as given, $\delta(y)$ in this chapter is the analogue of $\text{Ev}(\mathcal{E}, y)$ from Chapter 2.

The three main 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 Section 3.5.
Set estimation	The set of all subsets of Ω, denoted 2^{Ω}. See Section 3.6.
Hypothesis testing	A specified partition of Ω, denoted $\mathcal{H}$ below. See Section 3.7.

All three of these types of inference are easily adapted to specified functions of Θ, say $g(\Theta)$. Thus point estimation would have $\mathcal{A} = g\Omega$; set estimation would have $\mathcal{A} = 2^{g\Omega}$, and hypothesis testing would have $\mathcal{A} =$ a specified partition of $g\Omega$. For example, if $\theta = (\theta_1, \theta_2)$ but θ_2 is nuisance parameter, then $g(\theta) = \theta_1$. In point estimation, $\mathcal{A} = \Omega_1$, and $L(a, \theta) = L_1(a, \theta_1)$, where θ_1 is the value of Θ_1, and $a \in \Omega_1$ is the point estimate of Θ_1.

The next three sections develop some general results for Statistical Decision Theory, applicable to all types of inference, and then the later sections consider each of the three types in more detail.

3.2 *Bayes rules*

In a Bayesian approach, Θ is treated as a random variable, and the model $\mathcal{E}$ is augmented by a prior probability density function (PDF) π, for which $\Pr(\Theta \in S) = \int_{\theta \in S} \pi(\theta)\, \mathrm{d}\theta$ for any well-behaved $S \in \Omega$; see Section 1.5. I will write the joint distribution of (Y, Θ) as

$$\mathrm{p}(y, \theta) = f(y; \theta)\, \pi(\theta).$$

From this joint distribution, we can also calculate, as needed, the marginal distribution $\mathrm{p}(y)$ and the posterior distribution $\mathrm{p}(\theta \mid y)$; the latter using Bayes's theorem.

Let $\mathcal{D}$ be the set of all possible decision rules. The decision rule δ^* is a *Bayes rule* exactly when

$$\mathrm{E}\{L(\delta^*(Y), \Theta)\} \leq \mathrm{E}\{L(\delta(Y), \Theta\} \quad \text{for all } \delta \in \mathcal{D}. \tag{3.1}$$

The value $\mathrm{E}\{L(\delta(Y), \Theta)\}$ is termed the *Bayes risk* of decision rule δ, and is always well-defined under the condition that L is bounded. Therefore a Bayes rule is any decision rule which minimizes the Bayes risk, for some action set, loss function, model, and prior distribution. There is a justly famous result which gives the explicit form for a Bayes rule.

Theorem 3.1 (Bayes Rule Theorem, BRT). *If $\mathcal{A}$ is finite, then a Bayes rule exists*[1] *and satisfies $\delta^* = \tilde{\delta}$, where*

$$\tilde{\delta}(y) := \operatorname*{argmin}_{a \in \mathcal{A}} \mathrm{E}\{L(a, \Theta) \mid Y = y\}. \tag{3.2}$$

[1] Finiteness of $\mathcal{A}$ ensures existence. Similar but more general results are possible, but they require tedious and distracting topological conditions to ensure that a minimum obtains within $\mathcal{D}$.

Proof. I will show that $\mathrm{E}\{L(\delta(Y),\Theta)\} \geq \mathrm{E}\{L(\tilde{\delta}(Y),\Theta)\}$ for all $\delta : \mathcal{Y} \to \mathcal{A}$; i.e. that $\tilde{\delta}$ minimises the Bayes risk. Let δ be arbitrary. Then

$$\begin{aligned}
\mathrm{E}\{L(\delta(Y),\Theta)\} &= \int \textstyle\sum_y L(\delta(y),\theta) \cdot \mathrm{p}(y,\theta)\,\mathrm{d}\theta \\
&= \textstyle\sum_y \int L(\delta(y),\theta)\,\mathrm{p}(\theta \mid y)\,\mathrm{d}\theta \cdot \mathrm{p}(y) \\
&\geq \textstyle\sum_y \min_a \left\{ \int L(a,\theta)\,\mathrm{p}(\theta \mid y)\,\mathrm{d}\theta \right\} \cdot \mathrm{p}(y) \qquad \text{as } \mathrm{p}(y) \geq 0 \\
&= \textstyle\sum_y \int L(\tilde{\delta}(y),\theta)\,\mathrm{p}(\theta \mid y)\,\mathrm{d}\theta \cdot \mathrm{p}(y) \qquad \text{from (3.2)} \\
&= \int \textstyle\sum_y L(\tilde{\delta}(y),\theta) \cdot \mathrm{p}(y,\theta)\,\mathrm{d}\theta \\
&= \mathrm{E}\{L(\tilde{\delta}(Y),\Theta)\}. \qquad \square
\end{aligned}$$

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

The next result will not be a surprise for those who have read Chapter 2.

Theorem 3.2. *Bayes rules respect the Likelihood Principle (LP, see Definition 2.5).*

Proof. Let $\mathcal{E}_1 = \{\mathcal{Y}_1, \Omega, f_1\}$ and $\mathcal{E}_2 = \{\mathcal{Y}_2, \Omega, f_2\}$ be different models with the same parameter Θ. Because they have the same parameter, they have the same prior distribution π. By Bayes's theorem,

$$\begin{aligned}
\mathrm{p}_1(\theta \mid y_1) &\propto f_1(y_1;\theta)\,\pi(\theta) \\
\mathrm{p}_2(\theta \mid y_2) &\propto f_2(y_2;\theta)\,\pi(\theta)
\end{aligned}$$

where the missing multiplicative constants are $\mathrm{p}_1(y_1)^{-1}$ and $\mathrm{p}_2(y_2)^{-1}$, respectively. Now suppose that

$$f_1(y_1;\bullet) = c(y_1,y_2) \cdot f_2(y_2;\bullet),$$

as in the condition for the LP. I will show that this implies $\delta_1^*(y_1) = \delta_2^*(y_2)$, as required by the LP. By the Bayes Rule Theorem (Theorem 3.1),

$$\begin{aligned}
\delta_1^*(y_1) &= \operatorname*{argmin}_a \mathrm{E}_1\{L(a,\Theta) \mid Y_1 = y_1\} \\
&= \operatorname*{argmin}_a \int L(a,\theta) \cdot f_1(y_1;\theta)\,\pi(\theta)\,\mathrm{d}\theta \\
&= \operatorname*{argmin}_a \int L(a,\theta) \cdot c(y_1,y_2)\,f_2(y_2;\theta)\,\pi(\theta)\,\mathrm{d}\theta \\
&= \operatorname*{argmin}_a \int L(a,\theta) \cdot f_2(y_2;\theta)\,\pi(\theta)\,\mathrm{d}\theta \\
&= \operatorname*{argmin}_a \mathrm{E}_2\{L(a,\Theta) \mid Y_2 = y_2\} \\
&= \delta_2^*(y_2). \qquad \square
\end{aligned}$$

To hark back to the analysis in Chapter 2, if your inference (i.e. your decision rule) does not respect the LP then you are either illogical or obtuse—please excuse me for being blunt. So Theorem 3.2 is a good reason for selecting a Bayes rule as your decision rule. You can also be sure that your decision rule respects the Conditionality Principle (CP, Definition 2.7) and the Stopping Rule Principle (SRP, Definition 2.11). To assert the contrapositive, if your decision rule does not respect the LP, CP, and SRP, then it cannot be a Bayes rule.

3.3 *Admissible rules*

As discussed in Section 1.4, Frequentist statisticians are averse to prior distributions. But it is not possible to construct Bayes rules without them, and so Frequentist statisticians need another approach to selecting their decision rule for some action set, loss function, and model.

The accepted approach 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

$$R(\delta, \theta) := \mathrm{E}\{L(\delta(Y), \theta); \theta\} = \textstyle\sum_y L(\delta(y), \theta) \cdot f(y; \theta). \qquad (3.3)$$

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

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

with a strict inequality for at least one $\theta \in \Omega$. If you had both δ and δ', you would never want to use δ'.[2] 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.

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. It turns out that this issue has a clear-cut answer.

Theorem 3.3 (Wald's Complete Class Theorem, CCT)**.** *Let $\mathcal{E} = \{\mathcal{Y}, \Omega, f\}$, $\mathcal{A}$, and L be given. In the case where Ω is finite, a decision rule δ is admissible* if and only if *it is a Bayes rule for some positive prior distribution π.*

The proof is given in Section 3.4. There are generalisations of this theorem to non-finite and uncountable Ω; however, the results are highly technical. See Ferguson (1967, ch. 2), 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. In the rest of this section, I will assume the more general result, which is that a decision rule is admissible if and only if it is a Bayes rule, which holds for practical purposes.

[2] 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.

So what does the CCT say? First of all, admissible decision rules obey the LP. This follows from the fact that admissible rules are Bayes rules, and Bayes rules respect the LP, by Theorem 3.2. Insofar as we think respecting the LP is a good thing, this provides support for using admissible decision rules, because we cannot be certain that inadmissible rules respect the LP.

Second, if you select a Bayes rule according to some positive prior distribution π then you cannot ever choose an inadmissible decision rule. So the CCT states that there is a very simple way to protect yourself from choosing an inadmissible decision rule. Finally, if you cannot produce a positive π for which your proposed rule δ is a Bayes Rule, then you cannot show that δ is admissible.

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 positive π. 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. Do not attempt to explore this line of reasoning with your client!

3.4 The Complete Class Theorem

This section can be skipped once the previous section has been read. It proves a very powerful result, Theorem 3.3 above, originally due to an iconic figure in Statistics, Abraham Wald.[3] The parameter space is assumed to be finite, so write it as

$$\Omega = \{\theta_1, \dots, \theta_k\}.$$

Denote the available decision rules as δ_i, for $i = 1, 2, \dots$; I am assuming that the set of rules is countable, but this is without loss of generality (we will shortly create an uncountable number of decision rules). For each decision rule, define the risk function as

$$R_{ij} := \mathrm{E}\{L(\delta_i(Y), \theta_j); \theta_j\} \quad \begin{cases} i = 1, 2, \dots \\ j = 1, \dots, k. \end{cases}$$

Thus R_{ij} is the expected loss for rule δ_i under parameter value θ_j.

I will give a blackboard proof for $k = 2$ which generalises to any finite k. Call $\delta_1, \delta_2, \dots$ the 'pure' rules, and $R_1, R_2, \dots$ the pure risks, where $R_i = (R_{i1}, \dots, R_{ik})$. Panel (a) in Figure 3.4 shows a set of pure risks when $k = 2$.

We must widen the set of available decision rules, to include rules selected randomly from the pure rules according to probabilities $w = (w_1, w_2, \dots)$. This is because a rule δ_i might not be

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

dominated by a pure rule but it might be dominated by a randomised rule; see Figure 3.1. Let $\Pr(I = i) = w_i$. Then the risk of randomised rule w is

$$R_{wj} = \mathrm{E}\left\{L(\delta_I(Y), \theta_j); \theta_j\right\} = \textstyle\sum_i R_{ij} \cdot w_i,$$

by the Law of Iterated Expectation (LIE). The set of all rules, pure and randomised, is termed the *risk set*, and it is the convex hull of $\{R_1, R_2, \dots\}$. Every point in the risk set is an attainable risk, for a suitable choice of w. See Panel (b) of Figure 3.4. From now on, we can refer to 'risks' rather than 'rules'.

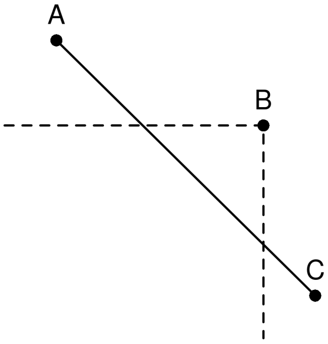


Figure 3.1: Rule B is not dominated by either A or δ, but it is dominated by some randomised rules based on A and δ, notably those with risks that lie in the facet between A and δ within the dashed lines.

Now consider the subset of the risk set which is admissible. A risk is dominated if there is another risk in its 'southwest' quadrant. So the only admissible risks in the risk set are on the southwest boundary, shown in Panel (c) of Figure 3.4. So we have identified the set of admissible risks: the pure risks on the southwest boundary, and the randomised risks which lie on the facets between the pure risks.

Now I show that this set of admissible risks is identical to the set of risks for Bayes rules for some positive prior probability. Fix $\pi = (\pi_1, 1 - \pi_1)$ with $0 < \pi_1 < 1$, and consider the set of risks with a specified Bayes risk a, i.e. the values (r_1, r_2) for which

$$\begin{aligned}
a &= \mathrm{E}\{L(\delta(Y), \Theta)\} && \text{defn of Bayes risk} \\
&= \mathrm{E}\left[\,\mathrm{E}\{L(\delta(Y), \Theta) \mid \Theta\}\right] && \text{by the LIE} \\
&= \mathrm{E}\{R(\delta, \Theta)\} && \text{defn of risk function} \\
&= \sum_{j=1}^{k} R(\delta, \theta_j) \cdot \pi_j && \Omega \text{ finite} \\
&= r_1 \cdot \pi_1 + r_2 \cdot (1 - \pi_1) && \text{for } k = 2.
\end{aligned}$$

On the panels in Figure 3.4, this is a straight line with equation

$$r_2 = \frac{a}{1 - \pi_1} + \frac{-\pi_1}{1 - \pi_1} r_1.$$

This line may pass below the risk set, in which case there is no attainable risk which has Bayes risk of a. So increase a until the line just touches the risk set, at risk $B(\pi)$ with Bayes risk b; see Panel (d) in Figure 3.4. $B(\pi)$ is the attainable risk which achieves the minimum Bayes risk for π, i.e. it is the risk of the Bayes rule for π. Varying π in the open interval $(0, 1)$ and repeating the exercise shows that the set of admissible risks and the set of risks for Bayes rules with positive prior probability are identical.

This proof generalises to any finite k according to the Supporting Hyperplane Theorem; see, e.g., Ferguson (1967, ch. 2) or Schervish (1995, ch. 3).

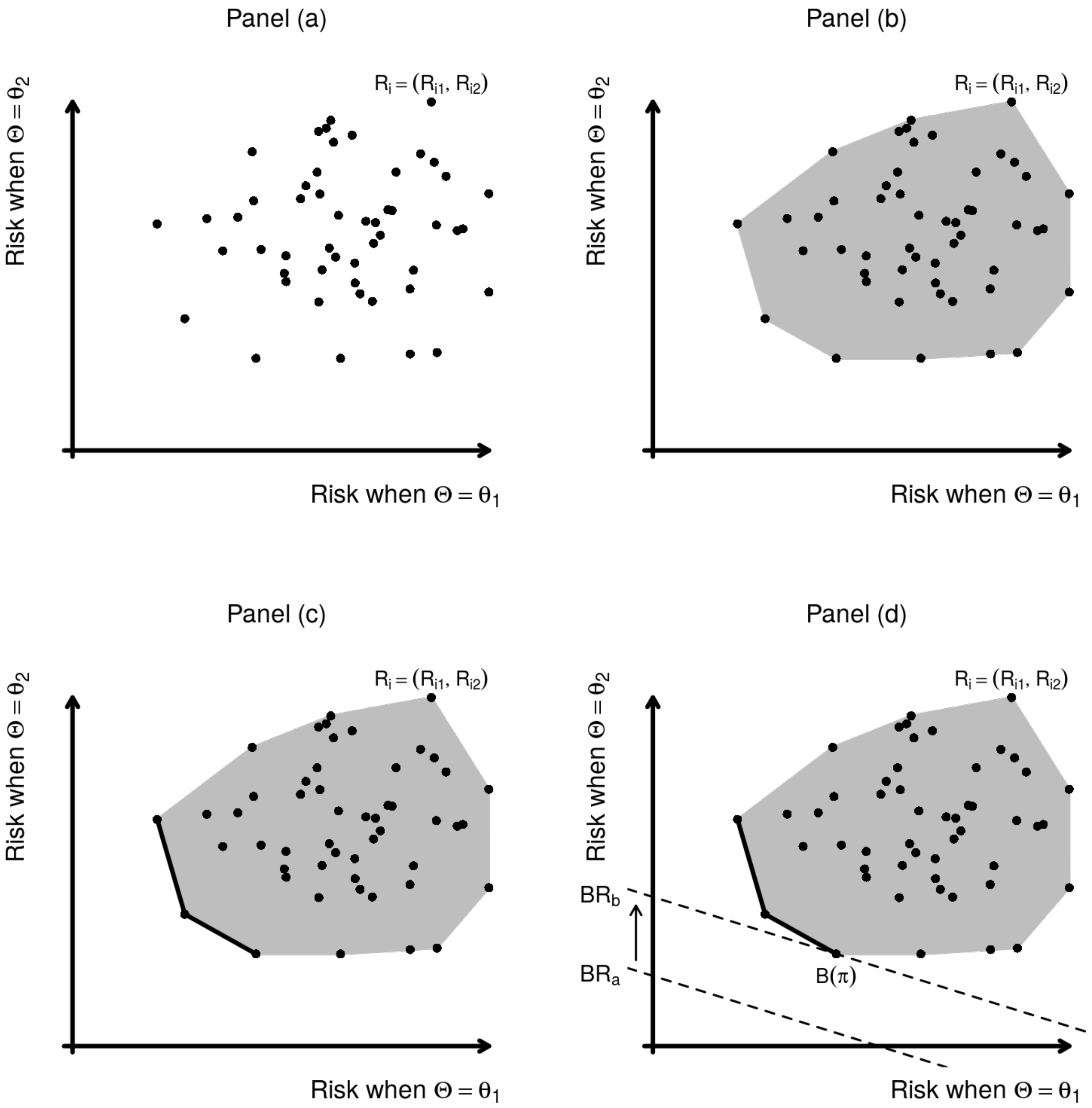


Figure 3.4. Blackboard proof of Theorem 3.3, with $\Omega = \{\theta_1, \theta_2\}$. Panel (a). The risks for a set of pure decision rules. Panel (b). The risk set: the convex hull of the pure risks, showing all risks that are attainable using randomised rules. Panel (c). The set of admissible risks is shown with a thick line. Panel (d). The dashed line 'BR_a' shows the set of risks which have Bayes risk a, for fixed probabilities $\pi = (\pi_1, 1 - \pi_1)$, where $0 < \pi_1 < 1$. None of the risks on BR_a are attainable. By increasing the Bayes risk to b, admissible pure risk $B(\pi)$ becomes attainable. $B(\pi)$ is the Bayes rule for π. Changing π changes the gradient of the dashed line, but it always just touches the set of attainable risks on the set of admissible risks.

3.5 *Point estimation*

For point estimation the action space is $\mathcal{A} = \Omega$, and the loss function $L(a, \theta)$ represents the (negative) consequence of choosing a as a point estimate of Θ, when in fact $\Theta = \theta$. A point estimate of Θ is often termed a *point prediction*, or just 'prediction'.

There will be situations where a function $L : \Omega \times \Omega \to \mathbb{R}$ is fairly easy to specify. Fox example, consider the Netflix challenge.[4] Netflix wants to make a prediction $a \in \Omega = \{1, 2, 3, 4, 5\}$ for a film that a client has not seen yet, but who will rate the film as Θ. Netflix suffers a reputational loss (which may lead to revenue loss) when a recommended film is rated below 5 by the client. But in fact Netflix will only recommend films that it predicts will be 5's, and so its loss function is something like

$$L(a, \theta) = \begin{cases} \epsilon \cdot (5 - a) & a = 1, 2, 3, 4 \\ a - \theta & a = 5 \end{cases}$$

where ϵ, which is a small positive value, is there to reflect that Netflix wants to make recommendations. In the Netflix challenge, the actual loss function was $L(a, \theta) = (a - \theta)^2$, which either goes to show that the people at Netflix are not very bright or, perhaps more likely, that the entire challenge was in fact a marketing exercise.

In many cases, however, specifying the loss function presents a challenge. 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*,

$$L(a, \theta) = h(a - \theta) \tag{3.4}$$

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.[5] This would assert that under-prediction incurs the same loss as over-prediction. There are many situations where an even function is *not* appropriate, but in these cases a generic loss function should be replaced by a more specific one.[6]

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

$$h(x) \propto x^T Q \, x \tag{3.5}$$

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

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

where the first 0 is because $h(\mathbf{0}) = 0$, the second 0 is because $\nabla h(\mathbf{0}) = 0$ since h is minimized at $x = \mathbf{0}$, and the third 0 is because

[4] See `https://en.wikipedia.org/wiki/Netflix_Prize`.

[5] I.e. $h(x) = h(-x)$.

[6] See, e.g., Milner and Rougier (2014), on predicting the weights of donkeys.

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 $x = \mathbf{0}$, if h is strictly convex and minimized 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 3.4. *Under a quadratic loss function, the Bayes rule for point estimation is the conditional expectation*

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

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 (3.5), 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 (Theorem 3.1) asserts that

$$\delta^*(y) = \operatorname*{argmin}_{a \in \Omega} \mathrm{E}\{L(a, \Theta) \mid Y = y\}. \tag{3.6}$$

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

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

Take expectations conditional on $Y = y$ to get

$$\mathrm{E}\{L(a, \Theta) \mid Y = y\} \propto (a - \psi(y))^2 + \mathrm{E}\{(\psi(y) - \theta)^2 \mid Y = y\}, \tag{†}$$

where the cross-product term is zero. Only the first term contains a, and this term is minimized over a by setting $a = \psi(y)$, as was to be shown.

The extension to vector θ with loss function (3.5) is straightforward, but involves more ink. It is crucial that Q in (3.5) is positive definite, because otherwise the first term in (†), which becomes $(a - \psi(y))^T Q\, (a - \psi(y))$, is not minimized if and only if $a = \psi(y)$. □

Now apply the CCT (Theorem 3.3) 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 positive prior distribution π.[7] Among the casualties of this conclusion is the Maximum Likelihood Estimator (MLE),

$$\hat{\theta}(y) := \operatorname*{arg\,max}_{\theta \in \Omega} f(y; \theta).$$

Stein's paradox showed that under quadratic loss, the MLE is not always admissible in the case of a Multinormal distribution with

[7] This is under the conditions of Theorem 3.3, or with appropriate extensions of them in the non-finite cases.

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. Persi Diaconis thought this was such a powerful result that he focused on it for his brief article on Mathematical Statistics in the *The Princeton Companion to Mathematics* (Ed. T. Gowers, 2008, 1056 pages). Nevertheless, the MLE is still the dominant point estimator in large areas of applied statistics, even though its admissibility under quadratic loss is questionable.

3.6 Set estimation

For set estimation the action space is $\mathcal{A} = 2^{\Omega}$, and the loss function $L(a, \theta)$ represents the (negative) consequences of choosing $a \subset \Omega$ as a set estimate of Θ, when the true value of Θ is θ.

There are two contradictory 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

$$L(a, \theta) = |a| + \kappa \cdot (1 - \mathbb{1}_{\theta \in a}) \quad \text{for some } \kappa > 0 \tag{3.7a}$$

where $|a|$ is the volume of a.[8] The value of κ controls the trade-off between the two requirements. If $\kappa \downarrow 0$ then the Bayes rule is the empty set, for all y. If $\kappa \uparrow \infty$ then the Bayes rule is Ω, for all y. For κ in-between, the Bayes rule will depend on beliefs about Y and the value y. Theorem 3.5 below continues to hold for the much more general set of loss functions

$$L(a, \theta) = g(|a|) + h(1 - \mathbb{1}_{\theta \in a}) \tag{3.7b}$$

where g is non-decreasing and h is strictly increasing. This is a large set of loss functions, which should satisfy most clients 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 (Theorem 3.4). For set estimators the situation is not so simple. However, for loss functions of the form (3.7) there is a simple necessary condition for a rule to be a Bayes rule. A set $a \subset \Omega$ is a *level set* of the posterior distribution exactly when $a = \{\theta : \text{p}(\theta \mid y) \geq k\}$ for some k.

Theorem 3.5. *If $\delta^* : \mathcal{Y} \to 2^{\Omega}$ is a Bayes rule for the loss function in* (3.7a)*, then it is a level set of the posterior distribution.*

Proof. The proof is by contradiction. For fixed y, I show that if a is not a level set of the posterior distribution, then there is an $a' \neq a$ which has a smaller expected loss; hence $\delta^*(y) \neq a$ according to the Bayes Rule theorem (BRT, Theorem 3.1).

First, note that

$$\text{E}\{L(a, \Theta) \mid Y = y\} = |a| + \kappa \cdot \Pr(\Theta \notin a \mid Y = y). \tag{†}$$

[8] Technically, Lebesgue measure, if Ω is a convex subset of $\mathbb{R}^d$.

Now suppose that a is not a level set of $\mathrm{p}(\theta \mid y)$. In that case there is a $\theta \in a$ and a $\theta' \notin a$ for which $\mathrm{p}(\theta' \mid y) > \mathrm{p}(\theta \mid y)$. Let $a' = a \cup \mathrm{d}\theta' \setminus \mathrm{d}\theta$.[9] Then $|a'| = |a|$, but

$$\Pr(\Theta \notin a' \mid Y = y) < \Pr(\Theta \notin a \mid Y = y).$$

Thus

$$\mathrm{E}\{L(a', \Theta) \mid Y = y\} < \mathrm{E}\{L(a, \Theta) \mid Y = y\}$$

from (†), showing that $\delta^*(y) \neq a$. □

Now relate this result to the CCT (Theorem 3.3). First, Theorem 3.5 asserts that δ being a level set of the posterior distribution is necessary (but not sufficient) for δ to be a Bayes rule for loss functions of the form (3.7). Second, the CCT asserts that being a Bayes rule is necessary (but not sufficient) for δ to be admissible.[10] So being a level set of a posterior distribution for some prior distribution π (which is *not* allowed to depend on y) is a necessary condition for being admissible under (3.7).

Now no one actually has (3.7) as their loss function; κ is a very inaccessible quantity. Eq. (3.7) is a generic loss function designed to help understand the features of a useful set estimator. Bayesian set estimators are usually *level* 95% *high posterior density (HPD) regions*. This is the level set of the posterior distribution which contains 95% of the posterior probability; other levels are also used.[11] So HPD regions satisfy the necessary condition for being a set estimator for the generic loss function (3.7).

Frequentist set estimators achieve a similar outcome if they are level sets of the likelihood function $\mathcal{L}(\bullet) \propto f(y; \bullet)$, because the posterior distribution is proportional to the likelihood function under a uniform prior distribution.[12] Frequentists do not need to actually adopt a unform prior distribution: they only need to point out that the uniform prior distribution justifies the admissibility of their 'level-sets of $\mathcal{L}$' estimator, via the CCT.

3.7 *Hypothesis tests*

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

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

Each element of $\mathcal{H}$ is termed a *hypothesis*; it is traditional to number the hypotheses from zero, where H_0 is termed the *null hypothesis*. The loss function $L(H_i, \theta)$ represents the (negative) consequences of choosing element H_i, when the true value of Θ is θ. It would be usual for the loss function to satisfy

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

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

[9] Here, $\mathrm{d}\theta$ is the tiny region of Ω around θ, and $\mathrm{d}\theta'$ is the tiny region of Ω around θ', for which $|\mathrm{d}\theta| = |\mathrm{d}\theta'|$.

[10] Necessary but not sufficient because being a Bayes rule AND having a positive prior distribution is equivalent to being admissible by the CCT, so being a Bayes rule without a condition on the prior distribution is necessary but not sufficient. As before, terms and conditions apply in the non-finite cases.

[11] HPD regions have the useful property of being nested for different levels.

[12] Or an almost-uniform prior distribution, in the case where Ω is unbounded, because the prior distribution will have to taper or be truncated in order to integrate to 1 over Ω.

There is one case where we have a complete theory of Bayes/admissible rules. Let $\Omega = \{\theta_0, \theta_1\}$, with $H_i = \{\theta_i\}$, for which the loss function will have the form

L	θ_0	θ_1
H_0	0	ℓ_1
H_1	ℓ_0	0

with $\ell_0, \ell_1 > 0$. Then it is straightforward to show that the Bayes rule for choosing between H_0 and H_1 has the form

$$\frac{f(y;\theta_0)}{f(y;\theta_1)} \begin{cases} < c & \text{choose } H_1 \\ = c & \text{toss a coin} \\ > c & \text{choose } H_0 \end{cases} \tag{3.8}$$

where $c = (\pi_1/\pi_0) \cdot (\ell_1/\ell_0)$. Thus the CCT states that a decision rule is admissible if and only if it has the form in (3.8) for some $c > 0$. This is effectively the Neyman-Pearson lemma, although it is usually expressed (and proved) differently.

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.[13]

There is a generic loss function for hypothesis tests, but it is hardly defensible. The *0-1 ('zero-one') loss function* is

$$L(H_i, \theta) = 1 - \mathbb{1}_{\theta \in H_i},$$

i.e., zero if θ 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.

* * *

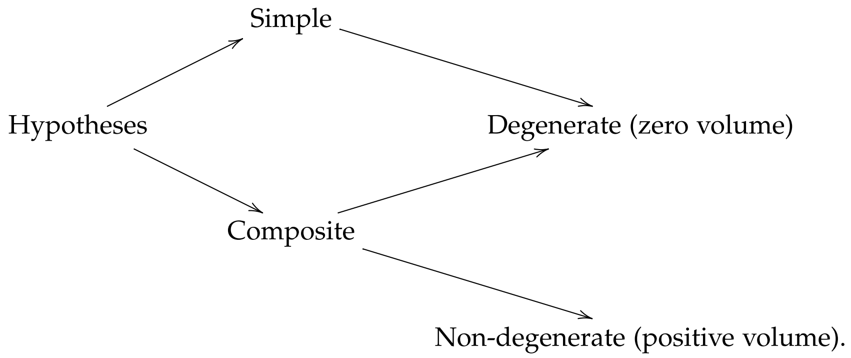
There is another approach which has proved much more popular. In fact, it is the dominant approach to hypothesis testing. This is to co-opt the theory of set estimators, for which there *is* a defensible generic loss function (see Section 3.6). The statistician can use her set estimator $\delta : \mathcal{Y} \to 2^{\Omega}$ 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.

[13] 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.

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.

In order to see how this approach plays out, we need to distinguish three types of hypothesis. The traditional distinction is between *simple hypotheses*, where $H_i = \{\theta_i\}$, a single element of Ω, and *composite hypotheses*, where H_i comprises more than a single element of Ω. Within composite hypotheses, though, we have *degenerate hypotheses*, which have zero volume, and *non-degenerate hypotheses*, which have positive volume; simple hypotheses always have zero volume. So here is the picture:



Obviously, it is effectively impossible to put a set inside a degenerate hypothesis, and so it is effectively impossible to accept a degenerate hypothesis using a set estimator—it is only possible reject it, or to be undecided.

To illustrate, suppose that the model is

$$\mathcal{E} = \left\{\mathbb{R}, (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++}, f\right\}$$

where f is the Normal probability density function. $H_1 : \{\mu = 0, \sigma^2 = 1\}$ would be a simple hypothesis; $H_2 : \{\mu = 0\}$ would be a composite degenerate hypothesis, and $H_3 : \{\mu > 0\}$ would be a composite non-degenerate hypothesis. It is possible to reject or be undecided about all three hypotheses, but it is only possible to accept H_3. Some statistics teachers seem to be confused about this, asserting that "it is never possible to accept the null hypothesis", or similar. This is not true in general, but it is true in the special case where the null hypothesis is degenerate (as is often the case in practice).

This set-estimator approach to hypothesis testing seems quite clear-cut, but we must end on two cautions. First, the statistician 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.

Second, in many situations, a hypothesis test is only superficially the right approach: attractive because of its simplicity, but limited for the same reason. For example, suppose that $H_0 : \{\mu \leq 0\}$ and $H_1 : \{\mu > 0\}$, where a positive value of μ indicates that a new type of drug does more good than harm. One could accept H_1 and yet the set estimate could be pressed close up against the line $\mu = 0$ without touching it, or one could be undecided about H_1 and yet most of the set estimate could be much larger than $\mu = 0$, with only a small tail crossing over. It is excessively crude to reduce a set estimate to a discrete choice between elements of $\mathcal{H}$, and for this reason many statisticians have never done a hypothesis test.[14] This is not a new revelation. Over fifty years ago, Edwards et al. (1963, p. 213) wrote

> No aspect of classical statistics has been so popular with psychologists and other scientists as hypothesis testing, though some classical statisticians agree with us that the topic has been overemphasized. A statistician of great experience told us, "I don't know much about tests, because I have never had occasion to use one."

Plus ça change, as they say.

[14] Including me, since I became a proper statistician.