16 Decision theory

To guess is cheap, to guess wrongly is expensive. (Old Chinese proverb)

16.1 The ideas of decision theory

We began the course with the following definition of Statistics:

> *a collection of procedures and principles for gaining and processing information in order to make decisions when faced with uncertainty.*

We have studied various ways to process data and to draw inferences from it: e.g., point estimation, interval estimation, hypothesis testing and regression modelling. There have been some key concepts, such as unbiasedness, the Neyman-Pearson lemma, and the fact that least squares estimators are the best linear unbiased estimators of regression parameters. But there are things that we have done which may seem to have been rather ad hoc, and which beg unanswered questions: e.g., do we always prefer unbiased estimators to biased ones? Do we care about estimators being linear? Sometimes we have have done things simply so that we can get an answer.

Decision theory attempts to provide Statistics with a satisfying foundation by placing everything within a unifying framework. In this framework the act of decision making is made central and ideas of optimality are introduced so that one can properly speak about making the 'best' inference. The conclusions are often the same as those reached by other means, but can also lead in new directions.

The decision theoretic approach begins with a careful definition of all the elements of a decision problem. It is imagined that there is a decision-maker who is to choose an *action* a from a set A. He is to do this based upon observation of a random variable, or *data* X. This X (typically a vector $X_1, \dots, X_n$) has a probability distribution which depends on an unknown parameter θ. Here θ denotes a *state of nature*. The set of all possible values of θ is the *parameter space* Θ.

The decision is to be made by a *statistical decision function* (or rule) d; this is a function which specifies $d(x)$ as the action to be taken when the observed data is $X = x$. On taking action $a = d(X)$ the decision-maker incurs a *loss* of $L(\theta, a)$. A good decision function is one that has a small value of the *risk function*

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

where this expectation is taken over X.

Clearly if $R(\theta, d_1) \le R(\theta, d_2)$ for all θ and $R(\theta, d_1) < R(\theta, d_2)$ for some θ then we would never want to use rule d_2, since d_1 can always do as well and sometimes better. We say d_2 is *inadmissible*.

Decision theory requires several lectures or a whole course to cover fully. Here we just give the flavour of some of the ideas.

Example 16.1 *In* Nature *(29 August, 1996, p. 766) Matthews gives the following table for various outcomes of Meteorological Office forecasts of weather covering* 1000 *one-hour walks in London.*

	Rain	No rain	Sum
Forecast of rain	66	156	222
Forecast of no rain	14	764	778
Sum	80	920	1000

Should one pay any attention to weather forecasts when deciding whether or not to carry an umbrella?

To analyse this question in a decision-theoretic way, let W, F and U be respectively the events that it is going to rain (be wet), that rain has been forecast, and that we carry an umbrella. The possible states of nature are W and W^c. The data is $X = F$ or $X = F^c$. Possible actions are chosen from the set $A = \{U, U^c\}$. We might present the loss function as

	W^c	W
U^c	L_{00}	L_{01}
U	L_{10}	L_{11}

For example, we might take $L_{01} = 4$, $L_{11} = 2$, $L_{10} = 1$, $L_{00} = 0$. Of course these are subjective choices, but most people would probably rank the four outcomes this way.

One possible decision function is given by $d_1(X) = U^c$, i.e., never carry an umbrella. It's risk function is

$$R(W^c, d_1) = L_{00}; \quad R(W, d_1) = L_{01}.$$

Another possible decision function is given by $d_2(F) = U$ and $d_2(F^c) = U^c$, i.e., carry an umbrella if and only if rain is forecast. The risk function is

$$R(W^c, d_2) = (764/920)L_{00} + (156/920)L_{10}; \quad R(W, d_2) = (66/80)L_{11} + (14/80)L_{01}.$$

We see that if $\theta = W^c$ then d_1 is better, but if $\theta = W$ then d_2 is better. Thus neither rule is uniformly better for both states of nature. Both d_1 and d_2 are admissible. By averaging over the states of nature we have the so-called *Bayes risk*, defined as

$$B(d) = \mathbb{E}[R(\theta, d)],$$

where the expected value is now taken over θ. For example, in our problem, $\mathbb{P}(W) = 0.08$ and $\mathbb{P}(W^c) = 0.92$, so $B(d) = 0.08R(W, d) + 0.92R(W^c, d)$.

The *Bayes rule* is defined as the rule d which minimizes the Bayes risk. Thus to find the Bayes rule for our problem, we must compare

$$B(d_1) = .08L_{01} + .92L_{00}$$

to

$$\begin{aligned}B(d_2) &= .08\big[(66/80)L_{11} + (14/80)L_{01}\big] + .92\big[(764/920)L_{00} + (156/920)L_{10}\big]\\ &= .066L_{11} + .014L_{01} + .764L_{00} + .156L_{10}.\end{aligned}$$

It follows that it is better to ignore weather forecasts and simply go for walks without an umbrella, if

$$B(d_1) < B(d_2) \iff \Delta := \frac{L_{01} - L_{11}}{L_{10} - L_{00}} < \frac{.156}{.066} = 2.364,$$

which can hold for reasonable values of the loss function, such as those given above, for which $\Delta = 2$. It all depends how you feel about getting wet versus the inconvenience of carrying an umbrella. Similar analysis shows that the commonly followed rule of always carrying an umbrella is better than doing so only if rain is forecast only if one is very adverse to getting wet, i.e., if $\Delta > 764/14 \doteqdot 53$.

16.2 Posterior analysis

In Lecture 5 we considered a decision theoretic approach to the point estimation problem. We used a loss function $L(\theta, a)$ to measure the loss incurred by estimating the value of a parameter to be a when its true value is θ. Then $\hat\theta$ was chosen to minimize $\mathbb{E}[L(\theta, \hat\theta)]$, where this expectation is over θ with respect to the posterior distribution $p(\theta \mid x)$.

Another way to think about the decision problem above is similar. We consider the expected loss under the posterior distribution. The posterior distribution for rain, given the data that there has been a forecast of rain, is $\mathbb{P}(W \mid F) = 66/222 \doteqdot 0.30$. (Note that this is less than 0.50!) Hence, given a forecast of rain, the expected loss if we carry an umbrella is

$$B(U \mid F) = (66/222)L_{11} + (156/222)L_{10},$$

whereas if we don't carry an umbrella the expected loss is

$$B(U^c \mid F) = (66/222)L_{01} + (156/222)L_{00}.$$

Not surprisingly, this leads to exactly the same criterion for choosing between d_1 and d_2 as we have already found above.

This is a general principle: the Bayes rule, d, can be determined as the action a which minimizes $\mathbb{E}_{\theta|X}[R(\theta, a)]$, this expectation being taken over θ with respect to the posterior distribution $p(\theta \mid x)$.

16.3 Hypothesis testing as decision making

We conclude by elucidating a decision theoretic approach to hypothesis testing. Consider the problem of testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative hypothesis $H_1 : \theta = \theta_1$. On the basis of an observation X we must decide in favour of H_0 (i.e., take action a_0) or decide in favour of H_1 (i.e., take action a_1).

For the case of so-called *0–1 loss* we take $L(\theta_0, a_0) = L(\theta_1, a_1) = 0$ and $L(\theta_0, a_1) = L(\theta_1, a_0) = 1$. I.e., there is unit loss if and only if we make the wrong decision. The risk function is then simply the probability of making the wrong decision, so $R(\theta_0, d) = \mathbb{P}(d(X) = a_1 \mid H_0)$ and $R(\theta_1, d) = \mathbb{P}(d(X) = a_0 \mid H_1)$.

Suppose we have prior probabilities on H_0 and H_1 of p_0 and p_1 respectively. This gives Bayes risk of

$$B(d) = p_0 R(\theta_0, d) + p_1 R(\theta_1, d).$$

As we have seen in the previous section the Bayes rule minimizes the posterior losses, so we should choose $d(X)$ to be a_1 or a_0 as

$$\frac{B(a_0 \mid x)}{B(a_1 \mid x)} = \frac{\mathbb{P}(H_1 \mid x)}{\mathbb{P}(H_0 \mid x)} = \frac{p_1 \mathbb{P}(x \mid H_1)}{p_0 \mathbb{P}(x \mid H_0)} = \frac{p_1}{p_0}\frac{f(x \mid \theta_1)}{f(x \mid \theta_0)}$$

is greater or less than 1.

This is of course simply a likelihood ratio test. Observe, however, that we have reached this form of test by a rather different route than in Lecture 6.

16.4 The classical and subjective points of view

The decision theoretic approach to statistical inference is appealing for the way it directly addresses issues such as loss, risk, admissibility, etc. These have intuitive interpretations in terms of the economics of decision making.

Decision theory also has the philosophical merit or dismerit, depending on your point of view, that it incorporates the Bayesian notions of prior and posterior beliefs. In the analysis of the hypothesis test above, we had to introduce a prior distribution on H_0 and H_1, as given by the probabilities p_0 and p_1. Some statisticians argue that this is fine; people always come to decision problems armed with prior beliefs, if only an uninformed belief expressed as $p_0 = p_1 = 1/2$. Others take the 'classical' line that statistical procedures should not depend upon the introduction of subjective prior beliefs on the part of the person analysing the data. They argue that only the data should matter: two people should automatically come to the exactly the same conclusion when presented with the same data. Most practising statisticians are happy to take the best of both viewpoints, letting the actual question under consideration decide which concepts and procedures are most helpful.