Statistical inference
Part 3
Statistical Decision Theory

Michael Goldstein
Durham University
APTS December 2023

Decisions.

We have shown that, under various reasonable conditions, our analysis should obey the likelihood principle.

However, this does not tell us which analysis we should actually choose.

A natural way to identify possible choices is to suppose that there are consequences of our choices, i.e. to consider our statistical analysis as a decision problem.

This approach to the foundations of statistics is very productive. It has been influential historically in motivating and developing the formal basis for statistical inference and led to much useful methodology.

Plus, knowing how to support good decision making under uncertainty should be a key part of the statistician's toolkit.

Outline

Our plan for this section is

(i) to develop the general structure for the solution of decision problems under uncertainty

(ii) to consider statistical problems within this general framework.

(We will discover that our methods do indeed satisfy the likelihood principle.)

(iii) to consider some traditional statistical procedures within this view.

Rewards

In a simple decision problem, you must choose between decisions.

You then receive a reward.

Which reward you receive is uncertain.

Here's a simple decision problem. Should you take an umbrella or not when you go for a walk?

The reward set might be

$\mathcal{R} = \{\text{get wet, stay dry}\}$

Uncertainty over rewards is measured by probability.

The value of each reward is measured by a general concept, called **utility** which we shall now explain.

Insurance example

We start with assuming some set $\mathcal{R}$ of **basic rewards**. This set lists all of the things that you can receive at the end of the decision problem.

Suppose, as an example, that the decision involves insurance of an item against theft, where

	theft	no theft
	E_1	E_2
insured d_1	r_1	r_2
not insured d_2	r_3	r_4
probability	p	$1-p$

Then $\mathcal{R} = \{r_1, r_2, r_3, r_4\}$.

It is clear that you have preferences between the basic rewards.

Maybe you would prefer r_4 over r_2 over r_1 over r_3.

Preference notation

To denote preferences between basic rewards, we will use the following notation.

For any basic rewards r and s in $\mathcal{R}$, the notation

$$r \geq^* s$$

means that you think r is at least as good s.

Similarly, if you are indifferent between r and s (i.e. i.e. you consider either at least as good as the other), our notation is

$$r \sim^* s$$

And if you strictly prefer r to s, we write

$$r >^* s$$

Gambles

Our decisions do not just involve basic rewards. They also involve uncertainty about which rewards we are going to get.

Hence, we must consider random rewards, or **gambles**.

In the insurance example, if we decide to insure ourselves, then

we get r_1 (pay insurance fee, and receive insurance payback), probability p

and r_2 (pay insurance fee and that's it), probability $1 - p$.

Our notation for this random reward is $pr_1 \oplus (1 - p)r_2$

Gambles

Definition Let $\mathcal{R}$ be a set of basic rewards. A **gamble** over $\mathcal{R}$ is a random quantity g that takes a finite number of values in $\mathcal{R}$.

For a gamble g yielding r_1 with probability p_1, r_2 with probability p_2, ..., and r_n with probability p_n (where r_1,...,$r_n \in \mathcal{R}$ and $\Sigma_i p_i = 1$),

we write $g = p_1 r_1 \oplus \cdots \oplus p_n r_n$ or $g = \bigoplus_i p_i r_i$.

The set of all gambles over $\mathcal{R}$ is denoted by $G(\mathcal{R})$.

Preferences over gambles

	theft	no theft
	E_1	E_2
insured d_1	r_1	r_2
not insured d_2	r_3	r_4
probability	p	$1-p$

In the insurance example, you might have $r_4 >^* r_2 >^* r_1 >^* r_3$.

However, in a decision problem, we need to compare gambles over rewards.

In our case, we must compare $pr_1 \oplus (1-p)r_2$ with $pr_3 \oplus (1-p)r_4$.

Hence, considering a decision problem with a set of rewards $\mathcal{R}$, we need to specify a preference ordering on $G(\mathcal{R})$.

The way that we will do this is by specifying a utility function over rewards.

Utility

Definition Given a set $\mathcal{R}$ of basic rewards, and a preference order $\geq^*$ on $G(\mathcal{R})$, a **utility function**, U, is a real valued function on the set $G(\mathcal{R})$ which satisfies the following two conditions.

[1] $\forall g_1, g_2 \in G(\mathcal{R})$:

$$g_1 \geq^* g_2 \iff U(g_1) \geq U(g_2)$$

[2] $\forall g_1, g_2 \in G(\mathcal{R}) \, \forall p \in [0,1]$:

$$U(pg_1 \oplus (1-p)g_2) = pU(g_1) + (1-p)U(g_2)$$

Utility: property 1

$\forall g_1, g_2 \in G(\mathcal{R})$:

$$g_1 \geq^* g_2 \iff U(g_1) \geq U(g_2)$$

The interpretation of this property is that utilities agree with preferences.

The larger the utility value of a gamble, the more highly we value receiving this gamble.

Therefore, if we can construct a utility function over all of our gambles, then we should choose the decision which corresponds to the gamble with highest utility.

Utility property 2

Utility, property [2]: $\forall g_1, g_2 \in G(\mathcal{R})\, \forall p \in [0,1]:$

$$U(pg_1 \oplus (1-p)g_2) = pU(g_1) + (1-p)U(g_2)$$

From this property, we can deduce the following property of utility.

For any gamble $\bigoplus_{i=1}^{n} p_i r_i$

$$U\left(\bigoplus_{i=1}^{n} p_i r_i\right) = \sum_{i=1}^{n} p_i U(r_i)$$

(Proof: Exercise - use induction)

Expected and actual utility

$$U\left(\bigoplus_{i=1}^{n} p_i r_i\right) = \sum_{i=1}^{n} p_i U(r_i)$$

Therefore, the actual utility of any gamble (the left hand side of the equation)

equals the expected utility of that gamble (the right hand side of this equation).

Notice, in particular, that when we have specified the value of $U(r_i)$ for each reward $r_i \in \mathcal{R}$, then this automatically determines the utility for every gamble over the rewards in $\mathcal{R}$.

Utility and decision choice

Choosing between decisions is equivalent to choosing between gambles over rewards.

A utility function defined over gambles over rewards has the properties

(i) the more that you prefer a gamble, the higher the utility of that gamble

(ii) the expected utility of a gamble equals the actual utility of that gamble.

(i) and (ii) imply that
you should choose the decision with the highest expected utility

(as the highest expected utility equals the highest actual utility and this corresponds to the most preferred gamble).

Insurance example

	theft	no theft
	E_1	E_2
insured d_1	r_1	r_2
not insured d_2	r_3	r_4
probability	p	$1-p$

Suppose that $p = \frac{1}{20}$, and assume that (somehow) you have managed to identify the following utilities. (Note they agree with your earlier preferences.)

$$U(r_4) = 100$$
$$U(r_2) = 96$$
$$U(r_1) = 80$$
$$U(r_3) = 0$$

These assessments determine your decision.

Choosing the best decision

	theft	no theft
	E_1	E_2
insured d_1	r_1	r_2
not insured d_2	r_3	r_4
probability	p	$1-p$

Decision d_1 corresponds to the gamble $g_1 = pr_1 \oplus (1-p)r_2$, with utility

$$U(pr_1 \oplus (1-p)r_2) = pU(r_1) + (1-p)U(r_2) = \frac{1}{20}80 + \frac{19}{20}96 = 95.2$$

d_2 corresponds to $g_2 = pr_3 \oplus (1-p)r_4$ whose utility is

$$U(pr_3 \oplus (1-p)r_4) = pU(r_3) + (1-p)U(r_4) = \frac{1}{20}0 + \frac{19}{20}100 = 95$$

$U(d_1) > U(d_2)$ so choose d_1.

Questions

[1] When will a utility function on a reward set exist?

[2] If it does exist, how do we construct it?

Answers

[1] under certain very reasonable assumptions on our preferences, a utility function over the reward set will always exist,

[2] there is a straightforward way to construct this function

This approach provides a rigorous approach to the solution of decision problems. This method is termed **maximizing (subjective) expected utility**.

Comment A utility function is subjective. It will differ from person to person, because it reflects a person's preferences between gambles, and different people may have different preferences. Therefore utility maximization will (correctly) lead different people to different decision choices.

Constructing utility

We will give a constructive demonstration of the existence of utility by an informal example.

Let's suppose that you want to construct your utility on gambles involving 5 different types of cakes:

cheese cake, apple cake, banana cake, lemon cake, and fruit cake.

For example, your decision problem might be to decide which shop to go to, where each shop has different probabilities for having each type of cake in stock.

I am now going to show you how I might specify my utility over these cakes.

[You can follow along and use this method to specify your own utility over these cakes (which will be different from mine).]

[If you hate cakes, then you can replace these cakes by five different things that you do like and follow along - the method will be the same.]

Ranking the rewards

The first step is to rank the rewards.

Here is my ranking.

cheese $>^*$ apple $>^*$ banana $>^*$ lemon $>^*$ fruit

Therefore, for me, my utility function must satisfy the constraints

U(cheese) $>$ U(apple) $>$ U(banana) $>$ U(lemon) $>$ U(fruit)

Best and worst rewards

The next step is to set the utility of the best reward equal to one, and the worst reward equal to zero.

With my preference ordering
cheese $>^*$ apple $>^*$ banana $>^*$ lemon $>^*$ fruit

I therefore set

$U(\text{cheese}) = 1, \quad U(\text{fruit}) = 0$

The utilities for all of the other cakes will lie between zero and one.

I now must specify the utility of each of the other cakes.

(Note that this will automatically specify the utility of every gamble over cakes as the utility of each such gamble is equal to the expected utility of the gamble.)

Assigning the remaining utilities

Let's choose another cake, say lemon cake.

Suppose that I have two choices.

[1] choose gamble

receive cheese cake with probability p

receive fruit cake with probability (1-p)

or

[2] receive lemon cake for sure.

Which do I prefer? Depends on the value of p.

Assigning the remaining utilities

Let p_l be the value of p for which I am indifferent between lemon cake for sure or the gamble giving cheese cake with probability p_l and fruit cake with probability $1 - p_l$

i.e. I judge

$$\text{lemon} \sim^* p_l \text{ cheese} \oplus (1 - p_l) \text{ fruit}$$

I set my utility for lemon cake to be this value of p.

$$U(\text{lemon}) = p_l$$

Assigning the remaining utilities

I repeat this for each of the cakes.

Here are my assessments.

U(cheese) $=$ 1, U(apple)$=$ 0.8, U(banana) $=$ 0.25,

U(lemon) $=$ 0.1, U(fruit) $=$ 0

(so this means that I would need 80%, 25% and 10% chances of cheesecake to pass up the sure chance of eating the different cakes).

[You can carry out this procedure for your preferences to obtain your utility over these cakes.]

Discussion

This method will allow you to construct a numerical score over each reward in any reward set.

Suppose our rewards are $r_1, ..., r_m$.
We judge r_1 worst, r_m best and set $U(r_1) = 0, U(r_m) = 1$.
For any other reward r_i, we set $U(r_i) = p_i$, where we judge
$r_i \sim *p_i r_m \oplus (1 - p_i) r_1$.

This score will clearly agree with your preference ranking. (The more you like a reward, the higher the chance of the best reward you need in order to pass up having this reward for sure.)

Does this function satisfy the second property of utility, namely that expected utility equals actual utility?

Yes, as we will now demonstrate, given certain (reasonable) assumptions concerning your preferences over gambles.

Demonstration of property 2

Want to show that

$$U(pr_i \oplus (1-p)r_j) = pU(r_i) + (1-p)U(r_j)$$

As

$$r_i \sim^* p_i r_m \oplus (1-p_i)r_1,\ r_j \sim^* p_j r_m \oplus (1-p_j)r_1$$

we have

$$pr_i \oplus (1-p)r_j \sim^* p[p_i r_m \oplus (1-p_i)r_1] \oplus (1-p)[p_j r_m \oplus (1-p_j)r_1]$$

Therefore, the result follows as

$$pr_i \oplus (1-p)r_j \sim^* [pp_i + (1-p)p_j]r_m \oplus [1-[pp_i + (1-p)p_j]r_1$$

Assumptions

The general (reasonable!) assumptions on your preferences that we need to make this type of argument work are as follows.

Comparability

$\forall g_1, g_2 \in G(\mathcal{R})$, $g_1 <^* g_2$ or $g_1 \sim^* g_2$ or $g_1 >^* g_2$

Coherence

$\forall g_1, g_2, g_3 \in G(\mathcal{R})$

$g_1 \leq^* g_2$ and $g_2 \leq^* g_3 \Rightarrow g_1 \leq^* g_3$

Monotonicity

$\forall g_1, g_2 \in G(\mathcal{R})$, $\forall p < q \in [0, 1]$

$$g_1 <^* g_2 \Rightarrow \Big(pg_2 \oplus (1-p)g_1 <^* qg_2 \oplus (1-q)g_1\Big)$$

Substitutability

$\forall g_1, g_2, g_3 \in G(\mathcal{R})$, $\forall p \in [0, 1]$

$$(g_1 \sim^* g_2) \Rightarrow \Big(pg_1 \oplus (1-p)g_3 \sim^* pg_2 \oplus (1-p)g_3\Big)$$

Theorem: existence of utility

We have given an outline demonstration of the following fundamental result.

Let $\mathcal{R}$ be a set of basic rewards, and let $\geq^*$ be a preference ordering on the set $G(\mathcal{R})$ of all gambles over $\mathcal{R}$.

Suppose that $\geq^*$ satisfies the above assumptions on preferences over $G(\mathcal{R})$.

Then there is a real-valued function U on $G(\mathcal{R})$ which satisfies the properties of a utility function.

Theorem: uniqueness of utility

Given a preference ordering $\leq^*$ over $G(\mathcal{R})$, a utility function for $\leq^*$ is unique up to an arbitrary positive linear transformation. More precisely:

[1] if U is a utility function for $\leq^*$ then, for all real numbers a, b, with $a > 0$, it holds that $aU + b$ is a utility function for $\leq^*$ as well.

[2] Conversely, suppose that U and V are utility functions for the same preference ordering $\leq^*$.

Then there are real numbers a, b, with $a > 0$, such that

$$V(g) = aU(g) + b, \ \forall g \in G(\mathcal{R})$$

.

Solving decision problems

We have shown that, uinder reasonable assumptions on your preferences, you should solve decision problems by

(i) specifying your uncertainties as probabilities,

(ii) specifying your values as utilities

(iii) choosing the decision which maximises expected utility.

The method is powerful and very widely applicable.

Here's a good place to start reading about this
Lindley (1991) Making Decisions, 2nd edition, Wiley

and a good followup is
Smith, J. (2010). Bayesian Decision Analysis: Principle and Practice. Cambridge, UK: Cambridge University Press.

Statistical Decision Theory

In particular, this approach is very relevant to the development and assessment of good statistical procedures.

Statistical Decision Theory is concerned to assess our inferences by considering their consequences.

By axiomatising our attitude to uncertainty and preference, we identify the general form of the statistical procedures that it is rational for us to use.

This allows us to consider how to construct the Ev function that we have discussed in ways that reflect our needs, which will vary between problems.

We will see, in particular, how this constructive approach to what we should do is consistent with our previous analysis of the requirements that we should impose for Ev.

Loss functions

The set of possible inferences, or decisions, is termed the decision space, denoted $\mathcal{D}$.

For each $d \in \mathcal{D}$, we want a way to assess the consequence of how good or bad the choice of decision d was under the outcome, or parameter value, θ.

Rather than trying to maximise gain, statisticians prefer to minimise losses.

Therefore, we work with loss, which is minus utility.

A **loss function** is any function L from $\Theta \times \mathcal{D}$ to $[0, \infty)$.

The loss function measures the penalty or error, $L(\theta, d)$ of the decision d when the parameter takes the value θ. Thus, larger values indicate worse consequences.

We suppose that the loss function is expressed in minus utility units so our aim is to maximise expected utility i.e. to minimise expected loss.

Types of inference

Examples of the types of inference about θ that we might consider are

1. point estimation,
2. set estimation,
3. hypothesis testing.

It is a great conceptual and practical simplification that Statistical Decision Theory distinguishes between these three types simply according to their decision spaces.

Type of inference	Decision space $\mathcal{D}$
Point estimation	The parameter space, Θ.
Set estimation	A set of subsets of Θ.
Hypothesis testing	A specified partition of Θ, denoted $\mathcal{H}$.

Bayesian statistical decision theory

In a Bayesian approach, a statistical decision problem $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$ has the following ingredients.

1. The possible values of the parameter: Θ, the **parameter space**.
2. The set of possible decisions: $\mathcal{D}$, the **decision space**.
3. The **probability distribution** on Θ, $\pi(\theta)$. For example,
   (a) this could be a **prior distribution**, $\pi(\theta) = f(\theta)$.
   (b) this could be a **posterior distribution**, $\pi(\theta) = f(\theta \mid x)$ following the receipt of some **data** x.
4. The **loss function** $L(\theta, d)$.

In this setting, only θ is random and we can calculate the expected loss, or risk for different decision choices.

Risk

We compare inferences (decisions) through their risk (expected loss)

Definition (Risk)

The **risk** of decision $d \in \mathcal{D}$ under the distribution $\pi(\theta)$ is

$$\rho(\pi(\theta), d) \quad = \quad \int_{\theta} L(\theta, d)\pi(\theta)\, d\theta.$$

We choose d to minimise the risk.

(This is the same as maximising our expected utility).

Bayes risk

Good inferences have small risk. We make the following definition.

Definition (Bayes rules and Bayes risk)

The **Bayes risk** $\rho^*(\pi)$ minimises the expected loss,

$$\rho^*(\pi) \quad = \quad \inf_{d \in \mathcal{D}} \rho(\pi, d)$$

with respect to $\pi(\theta)$.

A decision $d^* \in \mathcal{D}$ for which $\rho(\pi, d^*) = \rho^*(\pi)$ is a **Bayes rule** against $\pi(\theta)$.

[The Bayes rule may not be unique, and in weird cases it might not exist.]

We solve $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$ by finding $\rho^*(\pi)$ and (at least one) d^*.

Example - quadratic loss

Suppose that $\Theta \subset \mathbb{R}$ and we wish to find a point estimate for θ. We consider the loss function $L(\theta, d) = (\theta - d)^2$.

The risk of decision d is

$$\begin{aligned} \rho(\pi, d) = \mathbb{E}\{L(\theta, d) \,|\, \theta \sim \pi(\theta)\} &= \mathbb{E}_{(\pi)}\{(\theta - d)^2\} \\ &= \mathbb{E}_{(\pi)}(\theta^2) - 2d\mathbb{E}_{(\pi)}(\theta) + d^2, \end{aligned}$$

where $\mathbb{E}_{(\pi)}(\cdot)$ denotes the expectation with respect to $\pi(\theta)$.

Differentiating with respect to d we have

$$\frac{\partial}{\partial d}\rho(\pi, d) = -2\mathbb{E}_{(\pi)}(\theta) + 2d.$$

So, the Bayes rule is $d^* = \mathbb{E}_{(\pi)}(\theta)$.

Example - quadratic loss (continued)

The corresponding Bayes risk, for $d^* = \mathbb{E}_{(\pi)}(\theta)$, is

$$\rho^*(\pi) = \rho(\pi, d^*) = \mathbb{E}_{(\pi)}(\theta^2) - 2d^*\mathbb{E}_{(\pi)}(\theta) + (d^*)^2 = Var_{(\pi)}(\theta)$$

where $Var_{(\pi)}(\theta)$ is the variance of θ computed with respect to $\pi(\theta)$.

If $\pi(\theta) = f(\theta)$, a prior for θ, then the Bayes rule of an immediate decision is $d^* = \mathbb{E}(\theta)$ with corresponding Bayes risk $\rho^* = Var(\theta)$.

If we observe sample data x then $\pi(\theta) = f(\theta \mid x)$.

Therefore, the Bayes rule given this sample information is $d^* = \mathbb{E}(\theta \mid x)$ with corresponding Bayes risk $\rho^* = Var(\theta \mid x)$.

Note the way that data enters into our decision is through Bayes theorem.

Generalised quadratic loss

A more flexible version of this loss function is generalised quadratic loss, which is of form

$$L(\theta, d) = g(\theta)(\theta - d)^2$$

and so allows errors in estimation to be different for different values of θ.

For this loss function, the Bayes decision and Bayes risk are

$$d^* = \frac{E_\theta(g(\theta)\theta)}{E_\theta(g(\theta))}$$

$$\rho^*(p) = E_\theta(\theta^2 g(\theta)) - \frac{(E_\theta(\theta g(\theta)))^2}{E_\theta(g(\theta))}$$

(We leave this as an exercise.)

Solving statistical decision problems

We now have a general way of solving statistical decision problems.

We specify, and work out how to solve, the immediate decision problem,

$$[\Theta, \mathcal{D}, f(\theta), L(\theta, d)],$$

When we see sample data x, we apply Bayes theorem to update prior $f(\theta)$ to posterior $f(\theta|x)$ and then solve

$$[\Theta, \mathcal{D}, f(\theta \,|\, x), L(\theta, d)]$$

Decision rules

We often want to consider the **risk of the sampling procedure**, before observing the sample, for example to decide whether or not to sample and, if so, how much.

We now consider both θ and X as random (because we do not know what value of X we will observe).

For each possible sample, we need to specify which decision to make.

Definition (Decision rule) A decision rule $\delta(x)$ is a function from $\mathcal{X}$ into $\mathcal{D}$,

$$\delta : \mathcal{X} \to \mathcal{D}.$$

If $X = x$ is the observed value of the sample information then $\delta(x)$ is the decision that will be taken.

Bayes decision rules

The collection of all decision rules is denoted by Δ so that $\delta \in \Delta \Rightarrow \delta(x) \in \mathcal{D}\, \forall x \in X$.

We wish to solve the problem $[\Theta, \Delta, f(\theta, x), L(\theta, \delta(x))]$.

Definition (Bayes (decision) rule and risk of the sampling procedure)
The decision rule δ^* is a **Bayes (decision) rule** when

$$\mathbb{E}\{L(\theta, \delta^*(X))\} \quad \leq \quad \mathbb{E}\{L(\theta, \delta(X))\}$$

for all $\delta(x) \in \mathcal{D}$.

The corresponding risk $\rho^* = \mathbb{E}\{L(\theta, \delta^*(X))\}$ is termed the **(Bayes) risk of the sampling procedure**.

If the sample information consists of $X = (X_1, \ldots, X_n)$ then ρ^* will be a function of n and so can be used to help determine sample size choice.

Bayes rules

We will now show that the Bayes decision rule is simply the rule which chooses, for each x, the Bayes rule against $f(\theta|x)$.

(This is as we would expect.)

We have the following theorem.

Bayes rule theorem

Suppose that a Bayes rule exists for $[\Theta, \mathcal{D}, f(\theta \mid x), L(\theta, d)]$.

Then

$$\delta^*(x) = \arg\min_{d \in \mathcal{D}} \mathbb{E}(L(\theta, d) \mid X = x).$$

Proof

Let δ be arbitrary. Then

$$\begin{aligned}
\mathbb{E}\{L(\theta, \delta(X))\} &= \int_x \int_\theta L(\theta, \delta(x)) f(\theta, x)\, d\theta dx \\
&= \int_x \int_\theta L(\theta, \delta(x)) f(\theta \,|\, x) f(x)\, d\theta dx \\
&= \int_x \left\{ \int_\theta L(\theta, \delta(x)) f(\theta \,|\, x)\, d\theta \right\} f(x)\, dx \\
&= \int_x \mathbb{E}\{L(\theta, \delta(x)) \,|\, X = x\} f(x)\, dx
\end{aligned}$$

Now, as $f(x) > 0$, the $\delta^* \in \Delta$ which minimises $\mathbb{E}\{L(\theta, \delta(X))\}$ may equivalently be found as the δ^* which satisfies

$$\rho(f(\theta), \delta^*) = \inf_{\delta(x) \in \mathcal{D}} \mathbb{E}\{L(\theta, \delta(x)) \,|\, X\},$$

giving the result. □

Bayes rules

We have shown that the minimisation of expected loss over the space of all functions from $\mathcal{X}$ to $\mathcal{D}$ can be achieved by the pointwise minimisation over $\mathcal{D}$ of the expected loss conditional on $X = x$.

The risk of the sampling procedure is $\rho^* = \mathbb{E}[\mathbb{E}\{L(\theta, \delta^*(x)) \mid X\}]$.

Example - quadratic loss

With the loss function

$$L(\theta, d) \;=\; (\theta - d)^2,$$

we have

$$\delta^*(X) = \mathbb{E}(\theta \mid X)$$

and

$$\rho^* = \mathbb{E}\{Var(\theta \mid X)\}.$$

Bayes rule and the likelihood principle

Suppose that we consider Δ, the set of decision rules, to be our possible set of inferences about θ when the sample is observed.

Therefore, $\mathsf{Ev}(\mathcal{E}, x)$ is $\delta^*(x)$.

We have the following result.

Theorem The Bayes rule for the posterior decision respects the strong likelihood principle.

Proof If we have two Bayesian models with the same prior distribution and loss function, then if

$$f_{X_1}(x_1 \,|\, \theta) = c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta)$$

the corresponding posterior distributions are the same and so the corresponding Bayes rule (and risk) is the same. □

Example

A production line produces items, each of which may (independently) be acceptable, with probability ω, or defective, with probability $1 - \omega$.

We judge that our prior probability distribution for ω may be approximated by a beta distribution with parameters α and β, for some $\alpha > 1$ and $\beta > 0$.

We need to estimate the value of ω and mistakes are more important for smaller ω. Thus we use the loss function

$$L(\omega, d) = \frac{(\omega - d)^2}{\omega}$$

To help us determine a good estimate for ω, we plan to take a sample of n items from the process, with cost nc $(c > 0)$.

We may choose any value of n, but the value must be decided before we take any observations.

Questions

1. Find the Bayes decision and Bayes risk without sampling.

2. Find the Bayes decision and Bayes risk after observing a particular sample (i.e., when we have chosen n items from the line and found k acceptable, and $n - k$ defective).

3. Find the Bayes risk of sampling, for a given value of n.

4. Find the optimal choice of n.

Example:Bayes rule and risk

Before sampling, your prior distribution for ω is $\mathrm{Be}(\alpha, \beta)$.
With loss function $L(\omega, d) = g(\omega)(\omega - d)^2$ the Bayes decision and Bayes risk are

$$d^* = \frac{E_\omega(g(\omega)\omega)}{E_\omega(g(\omega))}$$

$$\rho^*(p) = E_\omega(\omega^2 g(\omega)) - \frac{(E_\omega(\omega g(\omega)))^2}{E_\omega(g(\omega))}$$

In this case, $g(\omega) = \frac{1}{\omega}$.
Therefore, the Bayes decision is

$$d^* = \frac{E_\omega(\frac{1}{\omega}\omega)}{E_\omega(\frac{1}{\omega})} = \frac{1}{E_\omega(\frac{1}{\omega})}$$

Example:Bayes rule

As $\omega \sim \mathrm{Be}(\alpha, \beta)$,

$$
\begin{aligned}
E_\omega(\tfrac{1}{\omega}) &= \int_0^1 \frac{1}{\omega} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \omega^{\alpha-1}(1-\omega)^{\beta-1}\mathrm{d}\omega \\
&= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \omega^{\alpha-2}(1-\omega)^{\beta-1}\mathrm{d}\omega \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha-1)}{\Gamma(\alpha+\beta-1)} \underbrace{\int_0^1 \frac{\Gamma(\alpha+\beta-1)}{\Gamma(\alpha-1)\Gamma(\beta)} \omega^{(\alpha-1)-1}(1-\omega)^{\beta-1}\mathrm{d}\omega}_{\text{integral is } 1 \text{ as it is the integral of a beta density } \mathrm{Be}(\alpha-1,\beta)} \\
&= \frac{\alpha+\beta-1}{\alpha-1}
\end{aligned}
$$

Example: Bayes risk

Hence, the Bayes decision is

$$d^* = \frac{1}{E_\omega(\frac{1}{\omega})} = \frac{\alpha - 1}{\alpha + \beta - 1}$$

The Bayes risk is

$$\begin{aligned}
\rho^*(p) &= E_\omega(\omega^2 \tfrac{1}{\omega}) - \frac{\left(E_\omega(\omega \tfrac{1}{\omega})\right)^2}{E_\omega(\frac{1}{\omega}))} \\
&= \frac{\alpha}{\alpha + \beta} - \frac{\alpha - 1}{\alpha + \beta - 1} \\
&= \frac{\beta}{(\alpha + \beta)(\alpha + \beta - 1)}
\end{aligned}$$

Example: Posterior rule and risk

Let X denote the number of successes observed in n independent trials.

X has binomial distribution and, having observed $X = k$ successes, your posterior distribution for ω is $\mathrm{Be}(\alpha + k, \beta + n - k)$
[as the beta family of distributions is conjugate with respect to binomial sampling].

Thus, the Bayes decision and Bayes risk are exactly as in the previous part, but with α replaced by $\alpha + k$ and β by $\beta + n - k$.

Therefore, we have

$$\delta^*(k) = \frac{\alpha + k - 1}{\alpha + \beta + n - 1}$$

$$\rho^*(k) = \frac{\beta + n - k}{(\alpha + \beta + n)(\alpha + \beta + n - 1)}$$

(where $\rho^*(k)$ is a shorthand notation for $\rho(\delta^*(k), p(\omega|X = k))$).

Example: Risk of sampling procedure

The Bayes risk of the sampling procedure, for sample size n, is (not taking into account the sampling cost nc for now)

$$\rho_S^*(p) = E_X(\rho^*(X)) = \frac{\beta + n - E_X(X)}{(\alpha + \beta + n)(\alpha + \beta + n - 1)}$$

But, for a fixed value of ω and n, the number of successes X is binomially distributed with parameters n and ω: i.e. $X \sim \text{Bi}(n, \omega)$, so $E_X(X|\omega) = n\omega$. Thus, by the law of iterated expectation, we have

$$E_X(X) = E_\omega(E_X(X|\omega)) = E_\omega(n\omega) = n\frac{\alpha}{\alpha + \beta}$$

Hence,

$$\begin{aligned}\rho_S^*(p) &= \frac{\beta + n - n\frac{\alpha}{\alpha+\beta}}{(\alpha + \beta + n)(\alpha + \beta + n - 1)} \\ &= \frac{\beta}{(\alpha + \beta)(\alpha + \beta + n - 1)}\end{aligned}$$

Example: Optimal sample size

Including the sampling cost nc, the total risk of sampling is

$$f(n) = \frac{\beta}{(\alpha+\beta)(\alpha+\beta+n-1)} + nc$$

Total risk $f(n)$ is minimal for some zero of equation $\frac{df}{dn} = 0$. These are

$$n = \pm\sqrt{\frac{\beta}{(\alpha+\beta)c}} + 1 - \alpha - \beta$$

but the only minimum for $f(n)$ corresponds to

$$n^* = \sqrt{\frac{\beta}{(\alpha+\beta)c}} + 1 - \alpha - \beta$$

If $n^* \leq 0$, then do not sample. If $n^* > 0$, then we should sample: round up or down n^*, whichever gives smaller $f(n)$.

Classical statistical decision theory

So far, we have concentrated on Bayesian Statistical Decision Theory.

We now consider implications of a decision theoretic view for the classical approach to statistics.

Bayes rules rely upon a prior distribution for θ: the risk is a function of d only.

In classical statistics, there is no distribution for θ and so another approach is needed.

Admissible rules

Definition (The classical risk)

For a decision rule $\delta(x)$, the classical risk for the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)$ is

$$R(\theta, \delta) \quad = \quad \int_X L(\theta, \delta(x)) f_X(x \mid \theta)\, dx.$$

The classical risk is thus, for each δ, a function of θ.

Example

Let $X = (X_1, \ldots, X_n)$ where $X_i \sim N(\theta, \sigma^2)$ and σ^2 is known.

Suppose that $L(\theta, d) = (\theta - d)^2$ and consider a conjugate prior $\theta \sim N(\mu_0, \sigma_0^2)$.

Therefore, the posterior distribution of θ, given $\underline{x}$ is $N(\mu_n, \sigma_n^2)$ where

$$\mu_n = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2}\right),$$

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

Example

Possible decision functions include:

1. $\delta_1(x) = \overline{x}$, the sample mean.
2. $\delta_2(x) = \text{med}\{x_1, \ldots, x_n\} = \tilde{x}$, the sample median.
3. $\delta_3(x) = \mu_0$, the prior mean.
4. $\delta_4(x) = \mu_n$, the posterior mean where

$$\mu_n = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2}\right),$$

(the weighted average of the prior and sample mean accorded to their respective precisions.)

Example - continued

The respective classical risks are

1. $R(\theta, \delta_1) = \frac{\sigma^2}{n}$, a constant for θ, since $\overline{X} \sim N(\theta, \sigma^2/n)$.
2. $R(\theta, \delta_2) = \frac{\pi\sigma^2}{2n}$, a constant for θ, since $\tilde{X} \sim N(\theta, \pi\sigma^2/2n)$ (approximately).
3. $R(\theta, \delta_3) = (\theta - \mu_0)^2 = \sigma_0^2 \left(\frac{\theta-\mu_0}{\sigma_0}\right)^2$.
4. $R(\theta, \delta_4) = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-2} \left\{\frac{1}{\sigma_0^2}\left(\frac{\theta-\mu_0}{\sigma_0}\right)^2 + \frac{n}{\sigma^2}\right\}$.

Which decision do we choose? We observe that $R(\theta, \delta_1) < R(\theta, \delta_2)$ for all $\theta \in \Theta$ but other comparisons depend upon θ.

Comment The risk of δ_3 is high unless θ is near μ_0, when it is low.
The risk of δ_4 moves from the risk of δ_3 to the risk of δ_1 as n increases.

Admissibility

An accepted approach for classical statisticians is to narrow the set of possible decision rules by ruling out those that are obviously bad.

Definition (Admissible decision rule)

A decision rule δ_0 is **inadmissible** if there exists a decision rule δ_1 which **dominates** it, that is

$$R(\theta, \delta_1) \leq R(\theta, \delta_0)$$

for all $\theta \in \Theta$ with $R(\theta, \delta_1) < R(\theta, \delta_0)$ for at least one value $\theta_0 \in \Theta$.

If no such δ_1 exists then δ_0 is **admissible**.

Admissibility

If δ_0 is dominated by δ_1 then the classical risk of δ_0 is never smaller than that of δ_1 and δ_1 has a smaller risk for θ_0.

Thus, you would never want to use δ_0.

(This is assuming that all other considerations are the same in the two cases: e.g. for all $x \in \mathcal{X}$, $\delta_1(x)$ and $\delta_0(x)$ take about the same amount of resource to compute.)

Therefore, it seems natural to reduce the set of possible decision rules under consideration by only using admissible rules.

Bayes rules and admissibility

We now show that an admissible rule can be related to a Bayes rule δ^* for a prior distribution $\pi(\theta)$.

We have the folowing theorem.

Theorem

Suppose that a prior distribution $\pi(\theta)$ is strictly positive for all Θ with finite Bayes risk and the classical risk, $R(\theta, \delta)$, is a continuous function of θ for all δ.

Under these conditions, the Bayes rule δ^* is admissible.

[Therefore, if you select a Bayes rule according to some positive prior distribution $\pi(\theta)$ then you cannot ever choose an inadmissible decision rule.]

Proof

We have

$$\begin{aligned}
\mathbb{E}\{L(\theta, \delta(X))\} &= \int_x \int_\theta L(\theta, \delta(x)) f_X(x \mid \theta) \pi(\theta)\, dx d\theta \\
&= \int_\theta \left\{ \int_x L(\theta, \delta(x)) f_X(x \mid \theta)\, dx \right\} \pi(\theta)\, d\theta \\
&= \int_\theta R(\theta, \delta) \pi(\theta)\, d\theta
\end{aligned}$$

Suppose that the Bayes rule δ^* is inadmissible and dominated by δ_1.

Thus, in an open set C of θ, $R(\theta, \delta_1) < R(\theta, \delta^*)$ with $R(\theta, \delta_1) \leq R(\theta, \delta^*)$ elsewhere.

Consequently, $\mathbb{E}\{L(\theta, \delta_1(X))\} < \mathbb{E}\{L(\theta, \delta^*(X))\}$ which is a contradiction to δ^* being the Bayes rule. □

Complete class theorem

The relationship between a Bayes rule with prior $\pi(\theta)$ and an admissible decision rule is even stronger.

The following result was derived by Abraham Wald (1902-1950).

Wald's Complete Class Theorem, CCT

In the case where the parameter space Θ and sample space $\mathcal{X}$ are finite, a decision rule δ is **admissible if and only if it is a Bayes rule** for some prior distribution $\pi(\theta)$ with strictly positive values.

An illuminating proof of this result can be found in Cox and Hinkley (1974, Section 11.6).

There are generalisations of this theorem to non-finite decision sets, parameter spaces, and sample spaces but the results are highly technical.

We'll proceed assuming the more general result, which is that a decision rule is admissible if and only if it is a Bayes rule for some prior distribution $\pi(\theta)$, which holds for practical purposes.

Admissable rules and the SLP

Admissible decision rules respect the SLP.

This follows from the facts that

(i) admissible rules are Bayes rules,

(ii) Bayes rules respect the SLP.

This provides some support for using admissible decision rules.

Point estimation

We now look at possible choices of loss functions for different types of inference.

For **point estimation** the decision space is $\mathcal{D} = \Theta$, and the loss function $L(\theta, d)$ represents the (negative) consequence of choosing d as a point estimate of θ.

There is a need for generic loss functions which are acceptable over a wide range of applications, particularly in a decision theoretic context.

Suppose that Θ is a convex subset of $\mathbb{R}^p$. A natural choice is a convex loss function,

$$L(\theta, d) \;=\; h(d - \theta)$$

where $h : \mathbb{R}^p \to \mathbb{R}$ is a smooth non-negative convex function with $h(0) = 0$.

Loss functions

This type of loss function asserts that small errors are much more tolerable than large ones.

One possible further restriction is that h is an even function, $h(d-\theta) = h(\theta - d)$.

In this case, $L(\theta, \theta+\epsilon) = L(\theta, \theta-\epsilon)$ so that under-estimation incurs the same loss as over-estimation.

We saw previously, that for quadratic loss $\Theta \subset \mathbb{R}$, $L(\theta, d) = (\theta - d)^2$, the Bayes rule was the expectation of $\pi(\theta)$. As we will see, this attractive feature can be extended to more dimensions.

There are many situations where this is not appropriate and the loss function should be asymmetric and a generic loss function should be replaced by a more specific one.

Absolute loss

The **bilinear loss function** for $\Theta \subset \mathbb{R}$ is, for $\alpha, \beta > 0$,

$$L(\theta, d) = \begin{cases} \alpha(\theta - d) & \text{if } d \leq \theta, \\ \beta(d - \theta) & \text{if } d \geq \theta. \end{cases}$$

The Bayes rule is a $\frac{\alpha}{\alpha+\beta}$-fractile of $\pi(\theta)$.

If $\alpha = \beta = 1$ then $L(\theta, d) = |\theta - d|$, the absolute loss which gives a Bayes rule of the median of $\pi(\theta)$.

$|\theta - d|$ is smaller than $(\theta - d)^2$ for $|\theta - d| > 1$ and so absolute loss is smaller than quadratic loss for large deviations. Thus, it takes less account of the tails of $\pi(\theta)$ leading to the choice of the median.

If $\alpha > \beta$, so $\frac{\alpha}{\alpha+\beta} > 0.5$, then under-estimation is penalised more than over-estimation and so that Bayes rule is more likely to be an over-estimate.

Example

If $\Theta \in \mathbb{R}^p$, the Bayes rule δ^* associated with the distribution $\pi(\theta)$ and the quadratic loss

$$L(\theta, d) \quad = \quad (d-\theta)^T Q\,(d-\theta)$$

is the expectation $\mathbb{E}_{(\pi)}(\theta)$ for every positive-definite symmetric $p \times p$ matrix Q.

Example,

Suppose $Q = \Sigma^{-1}$

Suppose $X \sim N_p(\theta, \Sigma)$ where the known variance matrix Σ is diagonal with elements σ_i^2 for each i.

Then $\mathcal{D} = \mathbb{R}^p$.

A possible loss function is

$$L(\theta, d) \quad = \quad \sum_{i=1}^{p} \left(\frac{d_i - \theta_i}{\sigma_i} \right)^2$$

so that the total loss is the sum of the squared component-wise errors.

Example

As the Bayes rule for $L(\theta, d) = (d - \theta)^T Q\,(d - \theta)$ does not depend upon Q, it is the same for an uncountably large class of loss functions.

If we apply the Complete Class Theorem to this result we see that 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 $\pi(\theta)$.

The value, and interpretability, of the quadratic loss can be further observed by noting that, from a Taylor series expansion, an even, differentiable and strictly convex loss function can be approximated by a quadratic loss function.

Stein's Example

Let $X = (X_1, \ldots, X_p)^T$, $\theta = (\theta_1, \ldots, \theta_p)^T$ for $p \geq 3$.

Suppose that $X \mid \theta \sim N_p(\theta, I_p)$ where I_p is the $p \times p$ identity matrix.

Thus, given θ, the X_is are independent $N(\theta_i, 1)$.

For a single observation $X = x$ the maximum likelihood estimate is $\delta^0(x) = x = (x_1, \ldots, x_p)^T$. This is unbiased.

For quadratic loss $L(\theta, d) = (\theta - d)^T(\theta - d)$ the classical risk of δ^0 is

$$\begin{aligned} R(\theta, \delta^0) &= \mathbb{E}[L(\theta, \delta^0(X)) \mid \theta] \\ &= \sum_{i=1}^{p} \mathbb{E}[(\theta_i - X_i)^2 \mid \theta] \\ &= \sum_{i=1}^{p} Var(X_i \mid \theta) \;=\; p. \end{aligned}$$

It turns out that δ^0 is inadmissible.

James-Stein estimators

Consider the estimator

$$\delta^a(X) \;=\; \left(1 - \frac{p-2}{X^T X}\right) X$$

(This is an example of the class of **James-Stein estimators**)

We can show that

$$R(\theta, \delta^a) \;=\; R(\theta, \delta^0) - (p-2)^2 \mathbb{E}\left[\frac{1}{X^T X} \,\middle|\, \theta\right].$$

Therefore δ^a dominates δ^0.

So therefore δ^0 is inadmissible.

Comments

The ith term of $\delta^a(X) = \left(1 - \frac{p-2}{X^T X}\right) X$ is $\left(1 - \frac{p-2}{X^T X}\right) X_i$ and so depends on all $X_1, \ldots, X_p$ even though the X_is are independent.

This outcome, often called **Stein's Paradox**, can be shown to occur in many situations when comparing three or more populations.

It occurs because the loss function is dealing with simultaneous estimation of all parameters and so is an on average property, whose relevance must be considered for each individual problem.

See Efron, B. and C.Morris (1977). Stein's paradox in statistics. *Scientific American 236*(5), 119–127.
[Available at http://statweb.stanford.edu/ ckirby/brad/other/Article1977.pdf]
for more details.

Note that whilst its admissibility under quadratic loss is questionable, the MLE remains the dominant point estimator in applied statistics.

Set estimation

For set estimation the decision space is a set of subsets of Θ so that each $d \subset \Theta$.

There are two contradictory requirements for set estimators of Θ.

1. We want the sets to be small.
2. We also want them to contain θ.

A simple way to represent these two requirements is to consider the loss function

$$L(\theta, d) \;=\; |d| + \kappa(1 - I_{\theta \in d})$$

for some $\kappa > 0$ where $|d|$ is the volume of d. The value of κ controls the 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 Θ.

Level sets

For loss functions of the form $L(\theta, d) = |d| + \kappa(1 - I_{\theta \in d})$ there is a a simple necessary condition for a rule to be a Bayes rule.

Definition (Level set)
A set $d \subset \Theta$ is a **level set** of the posterior distribution exactly when $d = \{\theta : \pi(\theta \,|\, x) \geq k\}$ for some k.

Theorem (Level set property, LSP)
If δ^* is a Bayes rule for $L(\theta, d) = |d| + \kappa(1 - I_{\theta \in d})$ then it is a level set of the posterior distribution.

Proof Note that

$$\begin{aligned} \mathbb{E}\{L(\theta, d) \,|\, X\} &= |d| + \kappa(1 - \mathbb{E}(I_{\theta \in d} \,|\, X)) \\ &= |d| + \kappa \mathbb{P}(\theta \notin d \,|\, X). \end{aligned}$$

Proof continued

For fixed x, we show that if d is not a level set of the posterior distribution then there is a $d' \neq d$ which has a smaller expected loss so that $\delta^*(x) \neq d$.

Suppose that d is not a level set of $\pi(\theta \,|\, x)$. Then there is a $\theta \in d$ and $\theta' \notin d$ for which $\pi(\theta' \,|\, x) > \pi(\theta \,|\, x)$.

Let $d' = d \cup d\theta' \setminus d\theta$ where $d\theta$ is the tiny region of Θ around θ and $d\theta'$ is the tiny region of Θ around θ' for which $|d\theta| = |d\theta'|$.

Then $|d'| = |d|$ but

$$\mathbb{P}(\theta \notin d' \,|\, X) < \mathbb{P}(\theta \notin d \,|\, X)$$

Thus, $\mathbb{E}\{L(\theta, d') \,|\, X\} < \mathbb{E}\{L(\theta, d) \,|\, X\}$ showing that $\delta^*(x) \neq d$. □

Comments

The Level Set Property Theorem states that δ having the level set property is necessary for δ to be a Bayes rule for loss functions of the form $L(\theta, d) = |d| + \kappa(1 - I_{\theta \in d})$.

The Complete Class Theorem states that being a Bayes rule is a necessary condition for δ to be admissible.

Being a level set of a posterior distribution for some prior distribution $\pi(\theta)$ is a necessary condition for being admissible for loss functions of this form.

Bayesian HPD regions satisfy the necessary condition for being a set estimator.

Classical set estimators achieve a similar outcome if they are level sets of the likelihood function, because the posterior is proportional to the likelihood under a uniform prior distribution.

[In the case where Θ is unbounded, this prior distribution may have to be truncated to be proper.]