AMSI
AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

National Collaboration in the Mathematical Sciences

Statistical Decision Theory

Wai Him Pun
School of Mathematical Sciences
University of Adelaide
Supervised by Professor Patty Solomon

February 27, 2014

Acknowledgement

First, I would like to thank my supervisor Professor Patty Solomon. Without her patient guidance, I could not have completed this project at this standard. Also, I would like to thank the Australian Mathematical Sciences Institute and the University of Adelaide for providing this invaluable research experience and the scholarship.

1 Introduction

Statistical decision theory is a framework for inference for any formally defined decision-making problem. For example (Berger 1985), suppose a drug company is deciding whether or not to sell a new pain reliever. Then the question is how much of the drug to produce. However, decision-making processes usually involve uncertainty. In this case, the proportion of the population that will buy this pain reliever is not known. If we over-produce, there will be an excess of drugs which cannot be sold; if we under-produce, we cannot maximise our profit. This is the potential loss for the company. The main goal of this study is to find an action or decision rule that reduces our loss as much as possible.

This report will consist of three sections. First, ideas about loss functions and decision rules in decision theory will be introduced. Decision rules or actions are chosen according to Bayesian expected loss, admissibility and some decision principles.

Second, we will introduce Bayesian statistics to handle the uncertainty in decision making. The basic ideas of Bayesian statistics, such as choice of prior distributions and calculations of posterior distributions, will be discussed. Third, combining Bayesian analysis with decision theory, we will see how the optimal decision rule, called the Bayes rule, can be derived from the posterior expected loss.

2 Basic elements

To work with the problem mathematically, it is necessary to employ some notation. The unknown quantity which affects the decision process is called the *state of nature* and commonly denoted θ. The parameter space, which contains all the possible values of θ, is denoted Θ. We also let a denote the *action* we take and let $\mathscr{A}$ denote the set of all of the possible actions.

2.1 Loss function

In order to make the right decision, we need to understand the consequences of taking an action under the uncertainty. This information is summarised in a *loss function.*

Definition 1. *The loss function,* $L : \Theta \times \mathscr{A} \to \mathbb{R}$ *represents the loss when an action* a *is employed and* θ *turns out to be the true nature of state.*

We express the consequences in term of loss. If the consequence of an action is a reward, we multiply the value of the reward by -1. Therefore, maximising rewards becomes minimising the loss.

Returning to the drug example from the introduction (Berger 1985), let θ be the proportion of the population that people will buy this drug, thus $\Theta = [0, 1]$. The action in this problem would be the estimate of θ. Hence, $a \in [0, 1]$. The company defines the loss function as

$$L(\theta, a) = \begin{cases} \theta - a & \text{if } \theta - a \geq 0, \\ 2(a - \theta) & \text{if } \theta - a < 0. \end{cases}$$

In other words, the company considers that over-estimation of the demand will be twice as costly as under-estimation. This kind of loss function is called *weighted linear loss.*

The main goal of the decision making is to find an action which incurs the least loss. A decision-making is said to be a *no-data problem* when there is no data available. To deal with the no-data problem, we measure how good an action is by taking the expected value of the loss function. Hence, it gives rise to the *Bayesian expected loss* and *conditional Bayes principle.*

Definition 2. *Let π be the probability distribution of θ. The Bayesian expected loss of an action a is,*

$$\rho(\pi, a) = E_\pi[L(\theta, a)] = \int_\Theta L(\theta, a)\pi(\theta)d\theta,$$

the expectation of the loss function with respect to the probability density of θ when we employed action a.

The Condtional Bayes Principle. *Choose an action a which minimises the $\rho(\pi, a)$. If such an action a exists, we call it Bayes action and will be denoted a^π.*

2.2 Decision rule

In the decision-making process, experiments are usually performed to better understand the unknown quantity θ. The random variables are denoted $X = (X_1, X_2, \cdots, X_n)$ from a common distribution independently and we denote $x = (x_1, x_2, \cdots, x_n)$ as the observed values. One of the goals of decision theory is to construct a *decision rule*, which summarises all the data and suggests an action to take.

Definition 3. *A decision rule $\delta : X \to \mathscr{A}$ is a function which maps the data to an action. When the data $X = x$, $\delta(x)$ is the action to take. If $p(\delta_1(X) = \delta_2(X)) = 1$ for all θ, then two decision rules are equivalent.*

We also denote $\mathscr{D}$ to be the set containing all possible decision rules δ. The idea of a decision rule is to determine the action using the data. In every experiment, we sample and obtain different sets of data. We would have different suggested actions by the decision rule.

In the drug example (Berger 1985), to estimate the proportion of the demand, the company selects a sample of n potential customers and observes x will buy their drugs. Therefore, a decision rule $\delta(x) = \frac{x}{n}$ can be adopted. This decision rule is an unbiased estimator for the proportion of the demand. However, it does not consider the loss function. We also wish to deal with the problem of over-estimation. We will see how to construct a better decision rule later.

It is necessary to understand, for given θ, the expected loss for employing a decision rule $\delta(x)$ repeatedly for changing X (Berger 1985). Therefore, we will use the *risk function* to represent the expected loss.

Definition 4. *The risk function of a decision rule $\delta(x)$ is defined by*

$$R(\theta,\delta) = E_X[L(\theta,\delta(X))] = \int_{\mathscr{X}} L(\theta,\delta(x))\pi(x|\theta)dx.$$

An elementary analysis can be performed by simply comparing the expected loss with respect to particular values of θ.

Definition 5. *A decision rule δ_1 is R-better than a decision rule δ_2 if $R(\theta,\delta_1) \leq R(\theta,\delta_2)$ for all $\theta \in \Theta$. They are R-equivalent if equality holds for any θ. A decision rule δ is admissible if there exists no R-better decision rule. It is inadmissible otherwise.*

We should obviously employ the R-better decision rule. For example, in the left plot of Figure 1, we should use the decision rule δ_2 because it leads to a smaller loss than δ_2 for any θ. However, if the risk functions of two decision rules cross each other, such as the right plot of Figure 1, the decision rules are not R-better than each other so they are both admissible. Then it would be impracticable to compare them in this way because we do not always have a unique admissible decision rule.

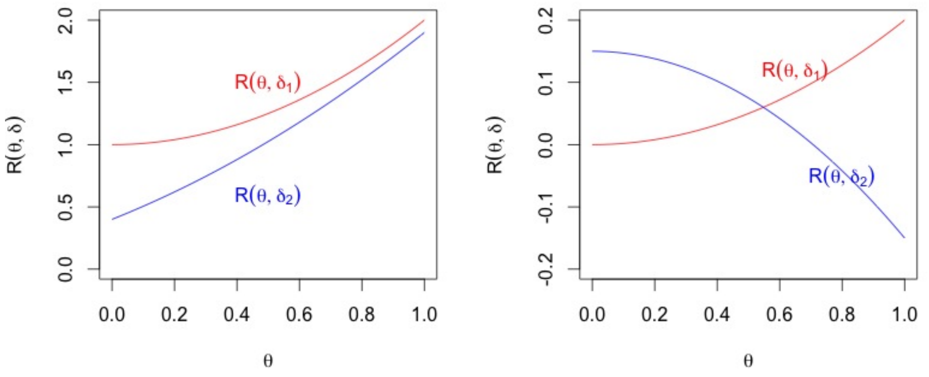

Figure 1. *A plot of an R-better risk function (left); a plot of crossing risk functions (right).*

In this subsection, we need other methods to select the decision rule. We now consider the expected loss of the risk function. That is, the *Bayes risk* of a decision rule, and we will then choose the decision rule by the *Bayes risk Principle*.

Definition 6. *The Bayes risk of a decision rule δ, with respect to the prior distribution $\pi(\theta)$ on Θ, is defined as*

$$r(\pi, \delta) = E_\pi[R(\theta, \delta)] = \int_\Theta R(\theta, \delta)\pi(\theta)d\theta.$$

The Bayes Risk Principle. *A decision rule δ_1 is preferred to a rule δ_2 if*

$$r(\pi, \delta_1) < r(\pi, \delta_2).$$

A decision rule $\delta \in \mathscr{D}$ which minimises $r(\pi, \delta)$ is called a Bayes rule δ^π. The quantity $r(\pi) = r(\pi, \delta^\pi)$ is then called the Bayes risk for π.

Searching for a minimising function in $\mathscr{D}$ to minimise $r(\pi, \delta(x))$ sounds more difficult than choosing the least loss action by the conditional Bayes principle. However, the Bayes risk principle and conditional Bayes principle are strongly related. In no-data problem, Bayes risk principle always gives the same answer as the conditional Bayes decision principle. Moreover, the Bayes action chosen with the posterior distribution is equivalent to the Bayes rule and this result will be discussed later. We now need the posterior distributions.

3 Bayesian Statistics

Bayesian statistics provides a good way to understand the unknown quantities in a decision making problem. The main goal of Bayesian data analysis is to construct a probability model for the unknown quantities of interest given the data (Gelman, et al., 2003). We call this model the posterior distribution. In this section, we will discuss how to calculate and interpret the posterior distribution to reach the ultimate inference of interest.

3.1 Bayes' Theorem

The Bayes' theorem is the foundation of the Bayesian statistics. When it is extended to inference problems with data, we use this theorem to calculate the posterior distributions.

Theorem 1. *Bayes' theorem states that given $P(B) > 0$, then,*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Now, we consider inference for an unknown parameter θ with known values of data x. Then, we reformulate the rule as

$$\begin{aligned}\pi(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{m(x)} \\ &\propto f(x|\theta)\pi(\theta).\end{aligned}$$

The prior distribution $\pi(\theta)$ is our belief about the parameter θ before we have considered the data. The data $x = (x_1, \cdots, x_n)$ are considered to a joint probability distribution depending on θ. The joint probability density of the fixed data x given a common parameter θ is called the likelihood, $f(x|\theta)$. We gather the information and update our understanding about θ by conditioning on the data x. The probability distribution which represents the latest knowledge about θ is called the posterior distribution $\pi(\theta|x)$. The marginal density of the data $m(x)$ is considered a constant. Combining the prior density $\pi(\theta)$ and the likelihood $f(x|\theta)$ yields the unnormalised posterior density.

3.2 Choice of prior distributions

The prior distribution $\pi(\theta)$ represents our knowledge of θ prior to collecting the data. There are different approaches to choosing a prior distribution. A subjective prior distribution is determined by the information before the experiment. If there is no information available, then we should use a noninformative prior distribution which has limited influence on the posterior distribution. We will discuss the choice of prior distributions with examples below.

3.2.1 Subjective Prior distribution

When there is some information about the unknown θ available prior to the experiment, we can include this information in the inference through the prior distribution. For instance, we are interested in estimating the population proportion θ of binomial data, that is $x|\theta \sim \text{Bin}(n, \theta)$. Researchers believe the proportion is about 10% and use a $\text{Beta}(1, 9)$ as the prior because now the expected value of the prior distribution is 0.1.

For the binomial data, the likelihood is

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Combining this likelihood with the prior density $\pi(\theta) = \frac{\Gamma(10)}{\Gamma(1)\Gamma(9)}(1-\theta)^8$, the posterior density function is given by

$$\begin{aligned}\pi(\theta|y) &\propto f(y|\theta)\pi(\theta)\\ &\propto \theta^x(1-\theta)^{n-x} \times (1-\theta)^8,\end{aligned}$$

which is Beta$(x+1, n-x+9)$.

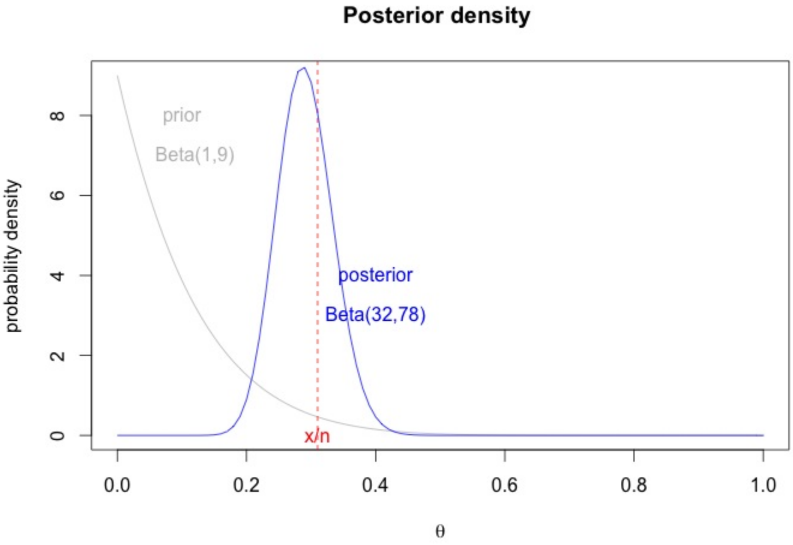

Figure 2. *The plot of posterior distribution with prior Beta*$(1,9)$*.*

The researchers observed 31 successes in 100 independent trials in an experiment. With the prior Beta$(1,9)$, we found that the posterior is a Beta$(32,78)$. We show the plots of the prior and posterior densities in Figure 2. We started with a prior distribution which reflected the prior belief that the true unknown θ is about 10%. However, in the experiment, it was observed that there are actually more successes in trials and the true θ should be greater than this. Thus, the data moves the posterior inference to a greater value.

The red line is $\frac{x}{n}$, an unbiased estimator for θ. However, the posterior distribution is not concentrated at this estimate because we also include our conservative belief about θ in the inference through the prior distribution. The posterior distribution is a compromise between the prior distribution and the data.

3.2.2 Conjugate Prior distributions

A prior distribution is said to be conjugate to a sampling distribution if the prior and posterior densities follow the same parametric distribution. Conjugate prior distributions are usually computationally easier because we just simply need to update the parameters of the distributions when we are updating our understanding about the unknown parameters with the data. In the example of binomial data, the Beta distribution is a conjugate prior distribution for the binomial likelihood. Using a Beta(α, β) as a prior distribution, we have

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Then, combining the prior density with the likelihood, we have

$$\begin{aligned} \pi(\theta|x) &\propto \theta^{x}(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}, \end{aligned}$$

which is a Beta$(\alpha + x, \beta + n - x)$ distribution.

3.2.3 Noninformative prior distributions

Noninformative prior distributions are constructed to avoid bias in posterior inference. The ideal shape of this kind of distribution is flat, diffuse and has large variance so that the density is not in favour of any value. One of the approaches often used is the *Jeffrey's invariance principle.*

Theorem 2. *Jeffrey's invariance principle states that a noninformative prior density $\pi(\theta)$ can be determined by*

$$\pi(\theta) \propto [I(\theta)]^{\frac{1}{2}},$$

where $I(\theta)$ is the Fisher information for θ.

The Fisher information is defined by $I(\theta) = E\left[-\frac{\partial^2 l(x|\theta)}{\partial\theta^2}\right]$, where $l(x|\theta) = \log f(x|\theta)$ is the log likelihood function. For binomial data, after evaluating the second derivative of the log likelihood and taking its expected value, we obtain the Fisher information

$$I(\theta) = \frac{n}{\theta(1-\theta)}.$$

The Jeffrey's Prior distribution is $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, which is a Beta$(\frac{1}{2}, \frac{1}{2})$.

One of the alternatives to choosing a noninformative prior is to use an improper prior distributions. Such prior distributions do not integrate to a bounded constant or are often undefined. However, conditioned on at least one data point using the Bayes' theorem, the posterior distributions become a valid probability density. Estimation of the mean of normal distributions is an example.

Suppose the data are normally distributed with unknown mean θ and known variance σ^2. The likelihood of one data point is given by

$$\pi(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\theta)^2}.$$

We first consider the posterior distribution with one data point and then generalise it to multiple data points. The likelihood is in an exponential quadratic form. Thus, we can use the normal distribution $\mathrm{N}(\mu_0, \tau_0)$ as a conjugate prior distribution. Thus, we have

$$\begin{aligned}\pi(\theta|x) &\propto f(x|\theta)\pi(\theta)\\ &\propto \exp(-\frac{1}{2\sigma^2}(x-\theta)^2)\exp(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2)\\ &= \exp(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2),\end{aligned}$$

where

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}x}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}, \qquad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}.$$

Hence, we have $\theta|x \sim \mathrm{N}(\mu_1, \sigma_1^2)$.

The sufficient statistic for θ is $\bar{x} = \frac{1}{n}\sum_{i=1} x_i$. It is equivalent that we obtain the posterior distribution by conditioning on the sufficient statistic, that is,

$$\pi(\theta|x_1, \ldots, x_n) = \pi(\theta|\bar{x}).$$

We know that $\bar{x}|\theta \sim \mathrm{N}(\theta, \frac{\sigma^2}{n})$ and treat $\bar{x}$ as a single normal distribution. Then, we obtain $\theta|x \sim \mathrm{N}(\mu_n, \tau_n^2)$ where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \qquad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}.$$

If τ_0^2 is large, the prior distribution is normally distributed with flat tails and its influence becomes increasingly limited on the posterior inference. If we let $\tau_0^2 \to \infty$, the prior distribution is a uniform distribution $U(-\infty, \infty)$, which is an improper prior distribution (Gelman, et al., 2003). However, conditioned on at least one data point, we have $\mu_n \to \bar{x}$ and $\tau_n \to \frac{\sigma^2}{n}$ as $\tau^2 \to \infty$, yielding the posterior distribution $\theta|x \sim \mathrm{N}(\bar{x}, \frac{\sigma^2}{n})$.

4 Bayesian Decision Theory

It appears that it is difficult to search for a decision rule which can minimise the Bayes risk in $\mathscr{D}$. However, this problem can be easily dealt with using Bayesian approach. Also, in this section, the optimal decision rules for some standard loss functions will be discussed.

4.1 Posterior expected loss

We discussed in section 3 how to obtain the posterior distribution, the updated knowledge about the unknown parameter θ. Then we now consider the expected loss with respect to the posterior distribution.

Definition 7. *The posterior expected loss is defined by*

$$\rho(\pi(\theta|x), a) = E[L(\theta, a)|x] = \int_{\Theta} L(\theta, a)\pi(\theta|x)d\theta.$$

By the conditional Bayes principle, we should employ the action which minimises the posterior distribution. The Bayes action is obtained by minimising the conditional expectation of the loss function on x, therefore, the Bayes action is a function of x, thus a decision rule. So, we will denote the Bayes action $\delta^*(x)$ (Berger 1985).

Theorem 3. *The Bayes action $\delta^*(x)$ found by minimising the posterior expected loss $\rho(\pi(\theta|x), a)$, minimises the Bayes risk $r(\pi, \delta(x))$.*

Proof. By definition,

$$\begin{aligned}
r(\pi, \delta) &= \int R(\theta, \delta)\pi(\theta) d\theta \\
&= \int_{\Theta} \int_{\mathscr{X}} L(\theta, \delta(x)) f(x|\theta)\pi(\theta) d\theta dx \\
&= \int_{\mathscr{X}} \int_{\Theta} L(\theta, \delta(x))\pi(\theta|x) d\theta \ m(x) dx \\
&= \int_{\mathscr{X}} \rho(\pi(\theta|x), \delta(x)) m(x) dx \\
&= E_X[\rho(\pi(\theta|x), \delta(x))].
\end{aligned}$$

$\delta^*(x)$ is found by minimising the posterior expected loss for any x. So, we have

$$\begin{aligned}
&\rho(\pi(\theta|x), \delta^*(x)) \leq \rho(\pi(\theta|x), \delta(x)) \\
\implies &E_X[\rho(\pi(\theta|x), \delta^*(x))] \leq E_X[\rho(\pi(\theta|x), \delta(x))] \\
\implies &r(\pi, \delta^*(x)) \leq r(\pi, \delta(x)).
\end{aligned}$$

Therefore, δ^* also minimises the Bayes risk (Berger 1985). □

4.2 Optimal decision rules for some standard loss functions

We now know that the optimal decision rule can be derived using Bayesian analysis. In this section, some standard loss functions and their results will be discussed.

One kind of standard loss functions is the square error loss, $L(\theta, a) = (\theta - a)^2$. When this loss function is considered, the risk function of the decision rule becomes

$$R(\pi, \delta) = E[L(\theta, \delta(X))] = E[(\theta - \delta(X))^2].$$

The decision making is now equivalent to choosing an estimator to minimise the mean squared error and the optimal decision rule is the Bayesian estimator, $E[\theta|x]$ (Berger 1985).

We can also apply a weighted function $w(\theta)$ such that $w(\theta) > 0$ for any θ to represent that given error of estimation varies according to different values of θ. The squared error loss $L(\theta, a) = (\theta - a)^2$ is then the special cases of weighted squared error loss with $w(\theta) = 1$ for any θ (Berger 1985). We will discuss the Bayes rule of this kind of weighted squared error loss functions.

Result 1. *If $L(\theta, a) = w(\theta)(\theta - a)^2$, then $\frac{E[\theta w(\theta)|x]}{E[w(\theta)|x]}$ is the unique Bayes action.*

Proof.

$$\begin{aligned} E[L(\theta, a)|x] &= \int L(\theta, a)\pi(\theta|x)d\theta \\ &= \int w(\theta)(\theta - a)^2 \pi(\theta|x)d\theta. \end{aligned}$$

Now, we can differentiate with respect to a. Since this is an integral over θ, we bring the differential operator in the integral and we have

$$\begin{aligned} \frac{\partial}{\partial a} E[L(\theta, a)|x] &= \frac{\partial}{\partial a} \int w(\theta)(\theta - a)^2 \pi(\theta|x)d\theta \\ &= \int w(\theta) \left(\frac{\partial}{\partial a} (\theta - a)^2 \right) \pi(\theta|x)d\theta \\ &= -2 \int w(\theta)(\theta - a)\pi(\theta|x)d\theta. \end{aligned}$$

Then a unique zero is attained at

$$a = \frac{\int \theta w(\theta)\pi(\theta|x)d\theta}{\int w(\theta)\pi(\theta|x)d\theta} = \frac{E[\theta w(\theta)|x]}{E[w(\theta)|x]}.$$

□

Another kind of commonly used loss function is the weighted linear loss functions. It is often used to represent the importance of over-estimation and under-estimation, as in the drug company example. Losses incurred by the error of estimation are approximately linear (Berger 1985).

Result 2. *If k_0 and k_1 are positive integers, and*

$$L(\theta, a) = \begin{cases} k_0(\theta - a) & \textit{if } \theta \geq a, \\ k_1(a - \theta) & \textit{if } \theta < a, \end{cases}$$

then the Bayes action is the $\frac{k_0}{k_1 + k_0}$-quantile of the posterior distribution.

Proof.

$$\begin{aligned}E[L(a|y)] &= \int L(\theta|x)\pi(\theta|x)d\theta \\ &= \int_{-\infty}^{a} k_1(a-\theta)\pi(\theta|x)d\theta + \int_{a}^{\infty} k_0(\theta-a)\pi(\theta|x)d\theta.\end{aligned}$$

We can minimise this function by obtaining the zero of the first derivative of this function with respect to a using the Fundamental Theorem of Calculus. Since $\pi(\cdot)$ is a probability density, the integral must converge at infinity. Therefore, we have,

$$E[L(\theta,a)|x)] = k_1 a\int_{-\infty}^{a}\pi(\theta|x)d\theta - k_1\int_{-\infty}^{a}\theta\pi(\theta|x)d\theta - k_0 a\int_{a}^{\infty}\pi(\theta|x)d\theta + k_0\int_{a}^{\infty}\theta\pi(\theta|x)d\theta$$

$$\begin{aligned}\implies \frac{\partial}{\partial a}E[L(a|x)] &= k_1\int_{-\infty}^{a}\pi(\theta|x)d\theta + k_1 a\pi(a|x) - k_1 a\pi(a|x) \\ &\qquad - k_0\int_{a}^{\infty}\pi(\theta|x)d\theta + k_0 a\pi(a|x) - k_0 a\pi(a|x)d\theta \\ &= k_1\int_{-\infty}^{a}\pi(\theta|x)d\theta - k_0\int_{a}^{\infty}\pi(\theta|x)d\theta.\end{aligned}$$

Therefore, setting $k_1\int_{-\infty}^{a}\pi(\theta|y)d\theta = k_0\int_{a}^{\infty}\pi(\theta|y)d\theta$, a unique zero is attained if a is the $\frac{k_0}{k_1+k_0}$-quantile of the posterior distribution.

□

The loss function in the drug company example is a special case of this problem, with $k_0 = 1$ and $k_1 = 2$,

$$L(\theta,a) = \begin{cases}\theta - a & \text{if } \theta - a \geq 0, \\ 2(a-\theta) & \text{if } \theta - a < 0.\end{cases}$$

Therefore, the Bayes action would be the third-quantile of the posterior distribution to avoid the greater loss caused by over-estimation.

5 Conclusion

Statistical decision theory is a study of decision making under uncertainty. We quantify the potential loss in the loss function $L(\theta,a)$ and employ a decision rule $\delta(x)$ to select an action depending on the observed data x. However, searching for the Bayes rule (the optimal decision rule) is a difficult problem. To deal with this problem, we need some Bayesian statistics. In this report, we discussed the basic concepts of Bayesian inference including choosing prior distributions and calculating posterior distributions. Combining the Bayesian analysis with decision theory, we found that the Bayes rule can be obtained by minimising the posterior expected loss.

Future work on this topic will be to analyse a dataset of mice with gastric cancer collected by the Adelaide Proteomics Centre. We will decide to handle the missing values in replicated mass spectra of mice. This will involve the study of hierarchical models to deal with the hierarchical structure of this dataset.

6 References

A. Gelman, J. Carlin, H. Stern and D. Rubin, 2003, Bayesian Data Analysis, Third Edition, Chapman and Hall/CRC.

J. O. Berger, 1985. Statistical Decision Theory and Bayesian Analysis, Second Edition, Springer.