Statistical games

In statistical inference, decisions about populations, such as what are the mean or variance of some characteristic, are based on sample data. Statistical inference can therefore be regarded as a game between Nature, which controls the relevant features of a population, and the statistician, who is trying to make a decision about the population.

One way in which statistical games differ from game theory is that in game theory each player chooses a strategy without knowing what the opponent will do. In a statistical game the statistician has some sample data, which will give some information about Nature's choice.

Example

A statistician is told that a coin has either a head on one side and a tail on the other or it has two heads. The statistician cannot inspect the coin but can observe a single toss of the coin and see whether it shows a head or a tail. The statistician must then decide whether or not the coin is two-headed. If the statistician makes the wrong decision there is a penalty of £1. There is no penalty (or reward) if the right decision is made.

Ignoring the fact that the statistician can observe one toss of the coin the problem could be regarded as the following game:

		Statistician (Player A)	
		a_1	a_2
Nature (Player B)	θ_1	0	1
	θ_2	1	0

where

θ_1 = "the state of nature" is that the coin is two-headed

θ_2 = "the state of nature" is that the coin is balanced

a_1 = Statistician's decision is that the coin is two-headed

a_2 = Statistician's decision is that the coin is balanced

However the statistician knows what has happened on the toss of the coin, *ie* the statistician knows whether a random variable x (say) has taken the value $x = 0$ (heads) or $x = 1$ (tails). The statistician will want to use this information in choosing between a_1 and a_2 and needs a "decision function" setting out the action to take in each case.

One possible decision function would be the function $d_1(x)$, where:

$$d_1(x) = \begin{cases} a_1 \text{ when } x = 0 \\ a_2 \text{ when } x = 1 \end{cases}$$

or $d_1(0) = a_1$ and $d_1(1) = a_2$

The purpose of the subscript is to distinguish different decision functions. Other possible decision functions might be

$d_2(0) = a_1$ and $d_2(1) = a_1$

***ie* always to choose a_1 regardless of the outcome of the experiment.**

or $d_3(0) = a_2$ and $d_3(1) = a_2$

or $d_4(0) = a_2$ and $d_4(1) = a_1$

These are the only possible decision functions if the statistician's value is based purely on the observed value x and randomised choices are not allowed.

Some of these decision functions may not be very sensible in practice.

The entries in the table give the corresponding values of the loss function thus:

		Statistician	
		a_1	a_2
Nature	θ_1	$L(a_1, \theta_1)$	$L(a_2, \theta_1)$
	θ_2	$L(a_1, \theta_2)$	$L(a_2, \theta_2)$

Consider again our original decision function d_1 and the resulting expected loss.

One option is to choose a_1 when $x = 0$ and a_2 when $x = 1$, which can be expressed as

$$R(d_1, \theta_j) = E[L(d_1(x), \theta_j)]$$

where $L(d_1(x), \theta_j)$ is the loss function when decision $d_1(x)$ is taken and the strategy of Nature is θ_j.

R is called a risk function. It gives us the value of the expected loss from a particular decision function and state of nature.

The expectation is taken with respect to the random variable x and

under θ_1 $P(x = 0) = 1$

$P(x = 1) = 0$

under θ_2 $P(x = 0) = ½$

$P(x = 1) = ½$

This gives

$$R(d_1, \theta_1) = 1L(a_1, \theta_1) + 0L(a_2, \theta_1) = 1 \times 0 + 0 \times 1 = 0$$

$$R(d_1, \theta_2) = ½L(a_1, \theta_2) + ½L(a_2, \theta_2) = ½ \times 1 + ½ \times 0 = ½$$

Similarly for the second decision function:

$$R(d_2, \theta_1) = 1L(a_1, \theta_1) + 0L(a_1, \theta_1) = 1 \times 0 + 0 \times 0 = 0$$

$$R(d_2, \theta_2) = ½L(a_1, \theta_2) + ½L(a_1, \theta_2) = ½ \times 1 + ½ \times 1 = 1$$

Question 1.5

Write down the risk functions for the decision functions d_3 and d_4.

This gives the following 4×2 zero sum two person game where the payoffs equal the corresponding values of the risk function

		Statistician			
		d_1	d_2	d_3	d_4
Nature	θ_1	0	0	1	1
	θ_2	½	1	0	½

Note that this is a slightly different type of matrix from the previous ones. It shows the expected loss for different decision rules, rather than the actual payoffs for different courses of action.

By inspection it can be seen that d_1 dominates d_2 and d_3 dominates d_4 so d_2 and d_4 can be discarded (they are said to be inadmissible). This is not surprising as they imply that a_1 (that the coin is two-headed) is accepted even though a tail was observed.

The game thus reduces to a 2×2 zero-sum two person game:

		Statistician	
		d_1	d_3
Nature	θ_1	0	1
	θ_2	½	0

The optimum strategy for the statistician is to adopt a randomised strategy with probabilities $\frac{2}{3}$ and $\frac{1}{3}$ for d_1 and d_3 respectively.

To verify that these are the correct probabilities, first assign a probability of p to strategy d_1 and a probability of $1-p$ to strategy d_3. We can calculate the expected loss for each of Nature's "strategies" θ_1 and θ_2. We have:

$$E(\text{Loss} \mid \theta_1) = 0 \times p + 1 \times (1-p) = 1-p$$

and:

$$E(\text{Loss} \mid \theta_2) = \tfrac{1}{2} \times p + 0 \times (1-p) = \tfrac{1}{2} p$$

If we equate these expressions we get:

$$\tfrac{1}{2} p = 1-p \Rightarrow p = 2/3$$

So the randomised strategy that minimises the maximum expected loss is to choose d_1 two thirds of the time and d_3 one third of the time.

The value of the game (the expected risk) is 33p. This is the value of both of the expected loss functions when $p = 2/3$.

Example

A statistician is observing values from a $Bin(2, p)$ distribution. He knows that p is equal either to ¼ or to ½, and he is trying to choose between these two values. He observes a single value x from the distribution. He proposes to use one of the following four decision functions:

$d_1(x)$: Set $p = ¼$ when $x = 0$

Set $p = ½$ when $x = 1$ or 2

$d_2(x)$: Set $p = ¼$ when $x = 0$ or 1

Set $p = ½$ when $x = 2$

$d_3(x)$: Set $p = ¼$ when $x = 0, 1$ or 2

$d_4(x)$: Set $p = ½$ when $x = 0, 1$ or 2

If he incorrectly concludes that $p = ¼$, he suffers a loss of 1. If he incorrectly concludes that $p = ½$, he suffers a loss of 2.

Find the risk function for each decision function, and find the decision function that minimises the maximum expected loss.

Solution

Consider first the decision function $d_1(x)$. If p actually is ¼, there is a loss of 0 if $x = 0$ and a loss of 2 if $x = 1$ or 2. So:

$$\begin{aligned} R(d_1, p = ¼) &= 0 \times P(X = 0 \mid p = ¼) + 2 \times P(X = 1 \text{ or } 2 \mid p = ¼) \\ &= 2 \times \left[2 \times ¼ \times ¾ + (¼)^2 \right] = 7/8 \end{aligned}$$

Similarly:

$$\begin{aligned} R(d_1, p = ½) &= 1 \times P(X = 0 \mid p = ½) + 0 \times P(X = 1 \text{ or } 2 \mid p = ½) \\ &= 1 \times (½)^2 = ¼ \end{aligned}$$

We now repeat the process with the other risk functions.

For $d_2(x)$ we get the following:

$$\begin{aligned} R(d_2, p=¼) &= 0\times P(X=0 \text{ or } 1 \mid p=¼)+2\times P(X=\ 2 \mid p=¼) \\ &= 2\times(¼)^2 = 1/8 \end{aligned}$$

and

$$\begin{aligned} R(d_2, p=½) &= 1\times P(X=0 \text{ or } 1 \mid p=½)+0\times P(X=\ 2 \mid p=½) \\ &= 1\times\left[(½)^2+2\times½\times½\right]=¾ \end{aligned}$$

For d_3 we always choose $p=¼$. So there will be a loss of 0 if p actually is ¼, and a loss of 1 if $p=½$:

$$\begin{aligned} R(d_3, p=¼) &= 0 \\ R(d_3, p=½) &= 1 \end{aligned}$$

Similarly:

$$\begin{aligned} R(d_4, p=¼) &= 2 \\ R(d_4, p=½) &= 0 \end{aligned}$$

If we look at the risk functions, we find that the maximum loss under each of the functions is 7/8 for d_1, ¾ for d_2, 1 for d_3 and 2 for d_4. So the decision function that minimises the maximum expected loss is d_2.

Question 1.6

Write down the values of the risk function in the previous example as a matrix. Are any of the strategies inadmissible?

Decision criteria

In general it is possible to find the best decision function only in respect of some criteria. Two criteria will be considered here.

The minimax criterion

Under the minimax criterion the decision function d chosen is that for which $R(d,\theta)$, maximised with respect to θ, is a minimum.

Note here that we're now applying the minimax criterion to a whole strategy, rather than to the choices on each turn.

Applying the minimax criterion to the example in Section 2 with d_2 and d_4 discarded, the maximum risk for

d_1 is ½ and for d_3 is 1

and so d_1 minimises the maximum risk.

> ***Question 1.7***
>
> How does the minimax criterion apply to the example on page 19?

The Bayes criterion

If Θ is regarded as a random variable under the Bayes criterion the decision function chosen is that for which $E[R(d,\Theta)]$ is a minimum where the expectation is taken with respect to Θ. The criterion needs Θ to be regarded as a random variable with a given distribution.

Applying the Bayes criterion to the example in Section 2 requires probabilities to be attached to Nature's two strategies θ_1 and θ_2. If $P(\theta_1) = p$ and $P(\theta_2) = 1 - p$ then the Bayes risk for

d_1 is $0.p + ½(1 - p) = ½(1 - p)$

d_3 is $1.p + 0(1 - p) = p$

Thus when $p > \frac{1}{3}$ the Bayes risk of d_1 is less than the Bayes risk of d_3, and d_1 is preferred to d_3.

When $p < \frac{1}{3}$ the Bayes risk of d_3 is less than the Bayes risk of d_1 and d_3 is preferred to d_1.

When $p = \frac{1}{3}$ the two Bayes risks are equal and either d_1 or d_3 can be chosen.

Chapter 1 Summary

In this chapter we study zero-sum two-player games. If we call our players Alice and Bob then the term "zero-sum" tells us that whatever Alice loses, Bob must win; there are no payments to or receipts from third parties.

Both Alice and Bob have a number of different strategies open to them. The payoffs from each combination of strategies can be represented in a matrix. The payoffs associated with Alice's available strategies (labelled I, II, III, *etc*) determine the columns of the matrix. The payoffs associated with Bob's available strategies (labelled 1, 2, 3, *etc*) determine the rows of the payoff matrix.

One strategy is said to dominate another if the first strategy is at least as good as the second and in some cases better. Dominated strategies can always be discarded.

Problems in Decision Theory usually involve the determination of optimum strategies and the corresponding payoff, or value, of the game. Two criteria that are often used to determine optimum strategies are the minimax criterion and the Bayes criterion.

Under the minimax criterion, each player will adopt the strategy that minimises their maximum loss (or maximises their minimum gain). The minimax criterion can be thought of as a 'best-of-all-evils' approach.

A saddle point is the name given to an entry of a payoff matrix that is the largest in its column and the smallest in its row. If a saddle point exists then each player will adopt the pure strategy, with the options that correspond to the saddle point being chosen.

If there is no saddle point, a randomised strategy can be adopted to enable a player to minimise his/her maximum expected loss. This means that the player will vary his or her choice of strategy in a random fashion but in accordance with some fixed set of probabilities.

The Bayes criterion is often used in the context of statistical games. An example of a statistical game is where a statistician wishes to determine a parameter value for a particular distribution, with nature acting as his/her "opponent". Under the Bayes criterion, the optimum strategy is the one that minimises the statistician's expected loss.

A decision function specifies a strategy for the outcome of an event, *eg* the sampling of a value from a particular distribution. A risk function can also be chosen to assign a value to the loss incurred if a wrong decision is made. In order to calculate the expected loss, probabilities must be assigned to each state of nature.

In classical statistics θ is a fixed but unknown quantity. This leads to difficulties such as the careful interpretation required for classical confidence intervals, where it is the interval that is random. As soon as the data are observed and a numerical interval is calculated, there is no probability involved. A statement such as $P(10.45 < \theta < 13.26) = 0.95$ cannot be made because θ is not a random variable.

Remember that in Subject CT3 we said that the correct way to express such statements was as follows:

"The interval $[g_1(X,\theta), g_2(X,\theta)]$ contains θ with probability 0.95, and hence the confidence interval is $(10.45, 13.26)$." This way we avoid making a statement that is meaningless. In classical statistics θ either lies within the interval or it does not. There can be no probability associated with such a statement.

In Bayesian statistics no such difficulties arise and probability statements can be made concerning the values of a parameter θ.

This means that it is quite possible to calculate a Bayesian confidence interval for a parameter. Although we shall not do this in this chapter, it is quite a common procedure in Bayesian statistics.

Another advantage of Bayesian statistics is that it enables us to make use of any information that we already have about the situation under investigation. Often researchers investigating an unknown population parameter have information available from other sources in advance of the study that provides a strong indication of what values the parameter is likely to take. This additional information might be in a form that cannot be incorporated directly in the current study. The classical statistical approach offers no scope for the researchers to take this additional information into account. However, the Bayesian approach does allow additional information to be taken into account when trying to estimate a population parameter.

An example of this would be where an insurance company is reviewing its premium rates for a particular type of policy and has access to results from other insurers, as well as from its own policyholders. This information cannot be taken into account directly because the terms and conditions of the policies for other companies may be slightly different. However, these additional data might contain a lot of useful information, which should not be ignored.

Bayes' Theorem

If $B_1, B_2, \ldots, B_k$ constitute a partition of a sample space S and $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A in S such that $P(A) \neq 0$

$$P(B_r|A) = \frac{P(A|B_r)P(B_r)}{P(A)} \text{ where } P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

for $r = 1, 2, \ldots, k$.

A partition of a sample space is a collection of events that are mutually exclusive and exhaustive, *ie* they do not overlap but cover the entire range of possible outcomes.

The result above is known as Bayes' Theorem. It can be proved as follows.

Proof

The result follows immediately if we insert an intermediate step:

$$P(B_r \mid A) = \frac{P(B_r \,\&\, A)}{P(A)} = \frac{P(A \mid B_r)\ P(B_r)}{\sum_i P(A \mid B_i)\ P(B_i)}$$

The first equality follows from the definition of the conditional probability $P(B_r \mid A)$.

The second equality follows by using the definition of $P(A \mid B_r)$ in the numerator and by conditioning on each of the possible events B_i in the denominator.

Bayes' formula allows us to "turn round" a conditional probability, *ie* to find $P(B|A)$ if we know $P(A|B)$.

An example

Three manufacturers supply clothing to a retailer. 60% of the stock comes from manufacturer 1, 30% from manufacturer 2 and 10% from manufacturer 3. 10% of the clothing from manufacturer 1 is faulty, 5% from manufacturer 2 is faulty and 15% from manufacturer 3 is faulty. What is the probability that a faulty garment comes from manufacturer 3?

Solution

Let A be the event that a garment is faulty. Let B_i be the event that the garment comes from manufacturer i.

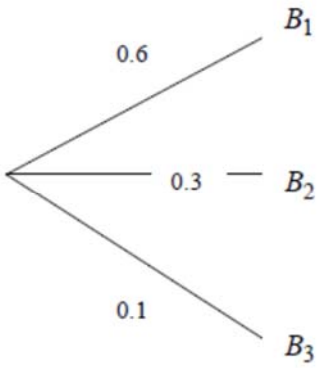
Substituting the figures into the formula for Bayes' Theorem,

$$P(B_3 \mid A) = \frac{(0.15)(0.1)}{(0.1)(0.6) + (0.05)(0.3) + (0.15)(0.1)} = \frac{0.015}{0.09} = 0.167$$

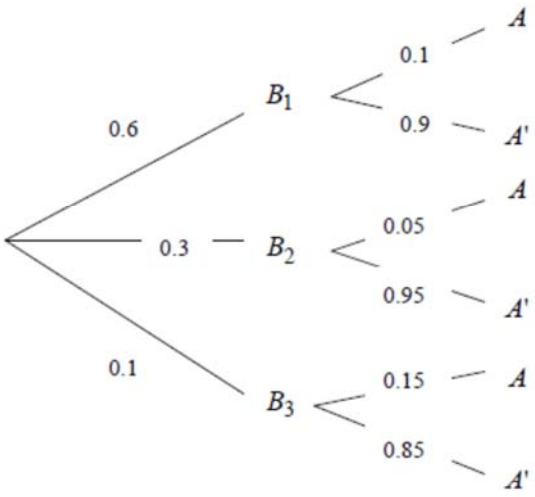
Although manufacturer 3 supplies only 10% of the garments to the retailer, nearly 17% of the faulty garments come from that manufacturer.

An alternative way of tackling this problem is to draw a tree diagram.

There are 3 manufacturers so we start with 3 branches in our tree and mark on the associated probabilities:



Each garment is either faulty (A) or perfect (A'). These outcomes and their (conditional) probabilities are now added to the diagram:



We want to work out $P(B_3 \mid A)$. The rule of conditional probability tells us that:

$$P(B_3 \mid A) = \frac{P(B_3 \text{ and A})}{P(A)}$$

From the diagram we can see that $P(B_3 \text{ and A}) = 0.1 \times 0.15 = 0.015$. (Just multiply together the appropriate branch weights.) We can also see that there are 3 ways in which event A can occur. Since these are mutually exclusive, we can calculate $P(A)$ by summing the 3 associated probabilities. Hence:

$$P(A) = (0.6 \times 0.1) + (0.3 \times 0.05) + (0.1 \times 0.15) = 0.09$$

and it follows that:

$$P(B_3 \mid A) = \frac{0.015}{0.09} = 0.167$$

as before.

Example

Claims arising from a particular class of general insurance business can be classified into three mutually exclusive types: S, M and L. The proportions of claims in each category are 80%, 15% and 5%.

The distribution of the amounts of individual claims in each category can be modelled by the probability density function $f_X(x) = 2\theta^2 / x^3$, $(x > \theta)$ where X represents the size of an individual claim. The parameters for the three categories are $\theta_S = £100$, $\theta_M = £1{,}000$, and $\theta_L = £2{,}500$.

Given only that the amount of an individual claim was £5,000, find the probability that it belongs to each of the three categories.

Solution

Define the following events (where h is a small quantity):

B_S: An individual claim is of type S

B_M: An individual claim is of type M

B_L: An individual claim is of type L

A: An individual claim falls in the range £5,000 to £$(5{,}000 + h)$

For a given value of θ, the probability of a claim falling in the range specified by A is:

$$P(A|\Theta = \theta) = \int_{5{,}000}^{5{,}000+h} \frac{2\theta^2}{x^3}\, dx$$

We can approximate this integral using the expression:

"*typical value of function* × *width of interval*"

This gives:

$$P(A|\Theta = \theta) \approx \frac{2\theta^2}{5{,}000^3} h$$

This gives us the following probabilities:

$$P(A|\Theta=\theta_S) = \frac{2\times 100^2}{5{,}000^3}h = 0.00000016h$$
$$P(A|\Theta=\theta_M) = \frac{2\times 1{,}000^2}{5{,}000^3}h = 0.000016h$$
$$P(A|\Theta=\theta_L) = \frac{2\times 2{,}500^2}{5{,}000^3}h = 0.0001h$$

We are told that:

$$P(\Theta=\theta_S) = 0.80$$
$$P(\Theta=\theta_M) = 0.15$$
$$P(\Theta=\theta_L) = 0.05$$

We can now use Bayes' formula (or construct a tree diagram) to obtain the required probabilities:

$$P(\Theta=\theta_S|A)$$
$$= \frac{P(A|\Theta=\theta_S)P(\Theta=\theta_S)}{P(A|\Theta=\theta_S)P(\Theta=\theta_S)+P(A|\Theta=\theta_M)P(\Theta=\theta_M)+P(A|\Theta=\theta_L)P(\Theta=\theta_L)}$$
$$= \frac{0.00000016h\times 0.80}{0.00000016h\times 0.80+0.000016h\times 0.15+0.0001h\times 0.05} = 0.0170$$

Similarly:

$$P(\Theta=\theta_M|A) = 0.3188 \qquad \text{and} \qquad P(\Theta=\theta_L|A) = 0.6642$$

So the posterior probabilities of types S, M and L are 1.70%, 31.88% and 66.42%.

Prior and posterior distributions

Suppose $\underline{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from a population specified by the density or probability function $f(x;\theta)$ and it is required to estimate θ.

Note that we are using f for both the density function of a continuous distribution and the probability function of a discrete distribution.

As a result of the parameter θ being a random variable, it will have a distribution. This allows the use of any knowledge available about possible values for θ before the collection of any data. This knowledge is quantified by expressing it as the prior distribution of θ.

Then after collecting appropriate data, the posterior distribution of θ is determined and this forms the basis of all inference concerning θ.

The information from the random sample is contained in the likelihood function for that sample. So the Bayesian approach combines the information obtained from the likelihood function with the information in the prior distribution. Both sources of information are combined to obtain a posterior estimate for the required population parameter.

Notation

As θ is a random variable, it should really be denoted by the capital Θ, and its prior density written as $f_\Theta(\theta)$. However, for simplicity no distinction will be made between Θ and θ, and the density will simply be denoted by $f(\theta)$. Note that referring to a density here implies that θ is continuous. In most applications this will be the case, as even when X is discrete (like the binomial or Poisson), the parameter (p or λ) will vary in a continuous space ((0, 1) or $(0,\infty)$, respectively).

The range of values taken by the prior distribution should also reflect the possible parameter values. So if we want a prior distribution for a parameter that can only take values in the range (0,1), then it would be silly to use, say, a gamma distribution for the prior in this case (which is defined on the interval from zero to infinity).

Also the population density or probability function will be denoted by $f(x|\theta)$ rather than the earlier $f(x;\theta)$ as it represents the conditional distribution of X given θ.

Determining the posterior density

Suppose that $\underline{X}$ is a random sample from a population specified by $f(x|\theta)$ and that θ has the prior density $f(\theta)$.

The posterior density of $\theta|\underline{X}$ is determined by applying the basic definition of a conditional density:

$$f(\theta|\underline{X}) = \frac{f(\theta,\underline{X})}{f(\underline{X})} = \frac{f(\underline{X}|\theta)f(\theta)}{f(\underline{X})}.$$

Note that $f(\underline{X}) = \int f(\underline{X}|\theta)f(\theta)d\theta$. This result is like a continuous version of Bayes' theorem from basic probability.

It is often convenient to express this result in terms of the value of a statistic, such as $\overline{X}$, rather than the sample values $\underline{X}$. So, for example,

$$f(\theta|\overline{X}) = \frac{f(\overline{X}|\theta)f(\theta)}{f(\overline{X})}.$$

In practice these are equivalent.

The notation can get a little confusing here. Remember that $\overline{X}$ is a sample mean, whereas $\underline{X}$ refers to the whole group of values of X *ie* a vector containing all the values in the sample.

A useful way of expressing the posterior density is to use proportionality. $f(\underline{X})$ does not involve θ and is just the constant needed to make it a proper density that integrates to unity, so

$$f(\theta|\underline{X}) \propto f(\underline{X}|\theta)f(\theta).$$

θ is not involved because we "integrated out" θ when we calculated $f(\underline{X})$ from the integral.

Also note that $f(\underline{X}|\theta)$, being the joint density of the sample values, is none other than the likelihood. So the posterior is proportional to the product of the likelihood and the prior.

THE POSTERIOR IS PROPORTIONAL TO THE PRIOR TIMES THE LIKELIHOOD.

This idea of proportionality is important. It enables us to do many questions on Bayesian methods relatively easily. We will demonstrate this with an example, which we will solve using a "first principles" approach, and also using a proportionality argument.

Example

You bought a box of light bulbs from a market stall a few months ago. You know that the bulbs are all either short-life bulbs with a mean life of 500 hours or long-life bulbs with a mean life of 2,500 hours, but you cannot tell which because there was no label on the box.

As you have not shopped at this stall before, you initially have no opinion as to whether you have been sold long-life bulbs or the cheaper alternative.

After approximately 300 hours, the 5 bulbs you have been using are all still going. Assuming that the life of an individual light bulb has an exponential distribution, how would you now assess the probability that you bought long-life bulbs?

Solution (first principles approach)

In this example, we initially consider each of the alternatives to be equally likely, so the prior distribution for λ, the parameter for the distribution of the lifetimes, is the discrete distribution:

$$\lambda = \begin{cases} 1/500 & \text{with probability } 1/2 \\ 1/2500 & \text{with probability } 1/2 \end{cases}$$

The probability that an individual light bulb with parameter λ (*ie* mean lifetime $1/\lambda$) will still be "alive" after 300 hours is: $\int_{300}^{\infty} \lambda e^{-\lambda x} dx = e^{-300\lambda}$

So the likelihood that all 5 will still be alive at that time is: $\left(e^{-300\lambda}\right)^5 = e^{-1500\lambda}$

So, we have:

$$P(5 \text{ alive} \mid \lambda = 1/500) = e^{-1500/500} = e^{-3} = 0.04979$$

$$P(5 \text{ alive} \mid \lambda = 1/2500) = e^{-1500/2500} = e^{-0.6} = 0.54881$$

Applying Bayes' formula, we then have:

$$P(\lambda = 1/500 \mid 5 \text{ alive})$$

$$= \frac{P(5 \text{ alive} \mid \lambda = 1/500) P(\lambda = 1/500)}{P(5 \text{ alive} \mid \lambda = 1/500) P(\lambda = 1/500) + P(5 \text{ alive} \mid \lambda = 1/2500) P(\lambda = 1/2500)}$$

$$= \frac{0.04979 \times 1/2}{0.04979 \times 1/2 + 0.54881 \times 1/2} = 0.08317$$

and

$$P(\lambda = 1/2500 \mid 5 \text{ alive})$$

$$= \frac{P(5 \text{ alive} \mid \lambda = 1/2500) P(\lambda = 1/2500)}{P(5 \text{ alive} \mid \lambda = 1/500) P(\lambda = 1/500) + P(5 \text{ alive} \mid \lambda = 1/2500) P(\lambda = 1/2500)}$$

$$= \frac{0.54881 \times 1/2}{0.04979 \times 1/2 + 0.54881 \times 1/2} = 0.91683$$

Note that the denominators for both probabilities for the posterior distribution were the same. So we could have just calculated the numerators and then used the fact that the posterior probabilities must total 1 to find the actual probabilities. If we look at the numerator, we see that it is just the likelihood $P(5 \text{ alive} \mid \lambda = 1/500)$ times the prior $P(\lambda = 1/500)$. So, we could have solved this problem more simply by writing down the figures in the table on the next page.

Solution (proportional approach)

	Prior ×	*Likelihood*	∝	*Posterior (proportional)*	*Posterior (actual)*
$\lambda = 1/500$:	1/2	$e^{-1500/500} = 0.04979$		0.02490	0.08319
$\lambda = 1/2500$:	1/2	$e^{-1500/2500} = 0.54881$		0.27441	0.91681
Total				0.29931	1.00000

The figures in the last column were found by dividing those in the previous column by 0.29931. They disagree slightly with those on the previous page due to rounding error.

Continuous prior distributions

The same logic underlying the proportional method applies in the case where the unknown parameter is assumed to have a continuous distribution. So, again, we can use the formula "*Posterior* ∝ *Prior* × *Likelihood*" to find the posterior parameter distribution.

The steps involved in finding the posterior distribution are:

Step 1 ***(selecting a prior distribution)***

Write down the prior distribution of the unknown parameter. Remember that the unknown parameter takes the place of the "x" in the PDF (or PF).

Step 2 ***(determining the likelihood function)***

Write down the (joint) likelihood function for the observation(s).

Step 3 ***(determining the posterior parameter distribution)***

Multiply the prior parameter distribution and the likelihood function to find the form of the posterior parameter distribution. You can ignore any multipliers that don't contain the unknown parameter. These will be "absorbed" by the proportional sign.

Step 4 ***(identifying the posterior parameter distribution)***

Either

- look for a standard distribution that has a PDF with the same algebraic form and range of values as the posterior distribution you have found *eg* by comparing with the PDFs in the *Tables*

or (if your posterior distribution doesn't match any of the standard distributions)

- "integrate out" (or "sum out") the unknown parameter to find the constant that makes the integral (or sum) of the PDF (or PF) of the posterior distribution equal to 1.

Very often in exam questions the first of these two methods will be successful (and the first method is usually easier than the second).

Question 2.4

If $x_1,\ldots,x_n$ is a random sample from an $Exp(\lambda)$ distribution where λ is an unknown parameter, find the posterior distribution for λ, assuming the prior distribution is $Exp(\lambda')$.

Conjugate priors

For a given likelihood, if the prior distribution leads to a posterior distribution belonging to the same family as the prior distribution, then this prior is called the conjugate prior for this likelihood.

The likelihood function determines which family of distributions will lead to a conjugate pair, *ie* a prior and posterior distribution that come from the same family. Conjugate distributions can be found by selecting a family of distributions that has the same algebraic form as the likelihood function, treating the unknown parameter as the random variable.

Example

Suppose that $x_1, x_2, \ldots, x_n$ are IID (independent and identically distributed) observations from a geometric distribution with parameter p, *ie* a distribution having probability function:

$$P(X = x) = p(1-p)^{x-1}, \quad x = 1, 2, 3, \ldots$$

where p is unknown. Find a family of distributions that would result in conjugate prior and posterior distributions.

Solution

The likelihood function is: $\prod_{i=1}^{n} p(1-p)^{x_i-1} = p^n(1-p)^{\sum x_i - n}$

So, we need a family of functions of the form $p^{something}(1-p)^{something}$ where $0 < p < 1$,

ie we need to use a beta distribution.

Using conjugate distributions often makes Bayesian calculations simpler. They may also be appropriate to use where there is a family of distributions that might be expected to provide a "natural" model for the unknown parameter *eg* in the previous example where the probability parameter p had to lie in the range $0 < p < 1$ (which is the range of values over which the beta distribution is defined).

Improper prior distributions

Sometimes it is useful to use an *uninformative* prior distribution, which assumes that an unknown parameter is equally likely to take *any* value. For example, we might have a sample from a normal population with mean μ where we know nothing at all about μ. This leads to a problem in this example because we would need to assume a $U(-\infty,\infty)$ distribution for μ, which doesn't make sense, since the PDF of this distribution would be 0 everywhere.

We can easily get round this problem by using the distribution $U(-N,N)$ where N is a very large number and then letting N to tend to infinity. The PDF of this distribution is $1/2N$ *ie* it is constant. If we use the "proportional to" method described above, with the prior distribution proportional to 1, everything works out nicely, even though the range of values is, in this case, infinite.

We have in fact used an uninformative prior already – in the lightbulbs example and in the packs of cards example.

The loss function

To obtain an estimator of θ, a loss function must first be specified. This is a measure of the "loss" incurred when $g(\underline{X})$ is used as an estimator of θ. A loss function is sought which is zero when the estimation is exactly correct, that is, $g(\underline{X}) = \theta$, and which is positive and does not decrease as $g(\underline{X})$ gets further away from θ. There is one very commonly used loss function, called quadratic or squared error loss. Two others are also used in practice.

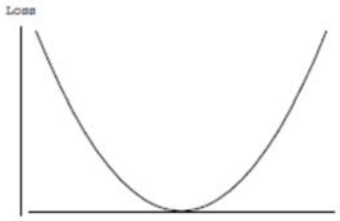
Then the Bayesian estimator is the $g(\underline{X})$ that minimises the expected loss with respect to the posterior distribution.

The main loss function is quadratic loss defined by

$$L(g(\underline{x}),\theta) = [g(\underline{x}) - \theta]^2,$$

and it is related to mean square error from classical statistics.

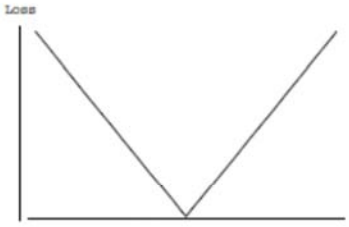
The formula for the squared error loss implies that as we move further away from the true parameter value, the loss increases at an increasing rate. The graph of the loss function is a parabola with a minimum of zero at the true parameter value.



A second loss function is absolute error loss defined by

$$L(g(\underline{x}),\theta) = |g(\underline{x}) - \theta|$$

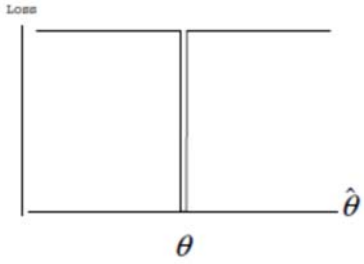
Here the graph of the loss function will be two straight lines which meet at the point (θ, 0). So as we move away from the true value in either direction, our loss increases at a constant rate.



A third loss function is "0/1" or "all-or-nothing" loss defined by

$$L(g(\underline{x}),\theta) = 0 \text{ if } g(\underline{x}) = \theta$$
$$1 \text{ if } g(\underline{x}) \neq \theta$$

In this case there is a constant loss for any parameter estimate that is not equal to the true underlying parameter value. If we hit the parameter value exactly, then the loss is zero.



The Bayesian estimator that arises by minimising the expected loss for each of these loss functions in turn is the mean, median and mode, respectively, of the posterior distribution, each of which is a measure of location of the posterior distribution.

The expected posterior loss is

$$EPL = E\{L(g(\underline{x}), \theta)\} = \int L(g(\underline{x}), \theta)\, f(\theta|\underline{x})\, d\theta$$

Remember that the expected value of any function is found by multiplying the function by the appropriate PDF and integrating over the whole range of values. Here the integral is assumed to cover the range $(-\infty, \infty)$.

Quadratic loss

For simplicity, g will be written instead of $g(\underline{x})$. So g just represents the number we come up with as our estimate of θ.

So $EPL = \int (g-\theta)^2 f(\theta|\underline{x})\, d\theta$

$$\frac{d}{dg} EPL = 2 \int (g-\theta)\, f(\theta|\underline{x})\, d\theta$$

More details on differentiating an integral with respect to a parameter are given in the ActEd FAC course.

Equating to zero $\Rightarrow g \int f(\theta|\underline{x})\, d\theta = \int \theta\, f(\theta|\underline{x})\, d\theta$

but $\int f(\theta|\underline{x})\, d\theta = 1$

since this is just the integral of the density function of the posterior distribution.

$$\therefore\ g = \int \theta\, f(\theta|\underline{x})\, d\theta = E(\theta|\underline{x})$$

Clearly this minimises *EPL*.

You can see this by differentiating the EPL a second time to get:

$$2\int f(\theta \mid \underline{x})\, d\theta = 2 > 0.$$

The Bayesian estimator under quadratic loss is the mean of the posterior distribution.

Question 2.6

Ten IID observations from a $Poisson(\lambda)$ distribution gave 3, 4, 3, 1, 5, 5, 2, 3, 3, 2. Assuming an $Exp(0.2)$ prior distribution for λ, find the Bayesian estimator of λ under squared error loss.

Absolute error loss

Again, g will be written instead of $g(\underline{x})$.

So $EPL = \int |g - \theta| f(\theta|\underline{x}) d\theta$

Assuming the range for θ is $(-\infty, \infty)$, then:

$$EPL = \int_{-\infty}^{g} (g - \theta) f(\theta|\underline{x}) d\theta + \int_{g}^{\infty} (\theta - g) f(\theta|\underline{x}) d\theta$$

We need to split the integral into two bits like this so that we can deal with the modulus signs. The first integral covers the range where $\theta \le g$. Here $|g-\theta| = g-\theta$ and so we can remove the modulus signs. The second integral covers the range where $\theta \ge g$. Here $|g-\theta| = \theta-g$, and again we can remove the modulus signs provided that we replace $|g-\theta|$ by $\theta-g$.

$$\therefore \frac{d}{dg} EPL = \int_{-\infty}^{g} f(\theta|\underline{x}) d\theta - \int_{g}^{\infty} f(\theta|\underline{x}) d\theta$$

[Recall that $\frac{d}{dy} \int_{a(y)}^{b(y)} f(x,y)dx = \int_{a(y)}^{b(y)} \frac{\partial}{\partial y} f(x,y)dx + b'(y)f(b(y),y) - a'(y)f(a(y),y)$**]**

(This is Leibniz' formula. See the ActEd FAC course.)

If we replace x by θ and y by g in the above result, we see that both of the last two terms reduce to zero for each integral, giving the required expression for the derivative (the same result was also used in a much more simple form in the quadratic loss case in the previous section).

Equating to zero $\Rightarrow \int_{-\infty}^{g} f(\theta|\underline{x}) d\theta = \int_{g}^{\infty} f(\theta|\underline{x}) d\theta$

that is, $P(\theta \le g) = P(\theta \ge g)$,

which specifies the median of the posterior distribution.

All-or-nothing loss

Here the differentiation approach cannot be used. Instead a direct approach will be used with a limiting argument.

Consider $L(g(\underline{x}), \theta) =$ 0 if $g - \varepsilon < \theta < g + \varepsilon$

1 otherwise,

so that, in the limit as $\varepsilon \to 0$, this tends to the required loss function.

$$EPL = 1 - \int_{g-\varepsilon}^{g+\varepsilon} f(\theta|\underline{x})d\theta = 1 - 2\varepsilon.f(g|\underline{x}) \quad \text{for small } \varepsilon.$$

This is saying that for a narrow strip the area under the function is approximately equal to the area of a rectangle whose width is 2ε and whose height is equal to the value of the function. We used a similar argument when calculating the probabilities in the example on Page 6.

This is minimised by taking g to be the mode of $f(\theta|\underline{x})$.

To minimise the expression, we need to maximise $2\varepsilon f(g|x)$. This occurs when $f(g|x)$ is maximised, *ie* at the mode of the posterior distribution.

This is known as all-or-nothing loss, or zero-one loss.

Question 2.7

$x_1, x_2, \ldots, x_n$ are IID observations from a $Gamma(\alpha, \lambda)$ distribution, where λ is unknown, but α is known. The prior distribution of λ is $Exp(\theta)$ where θ is a known constant. Find the Bayesian estimator of λ under zero-one error loss.

An example

For the estimation of a binomial probability θ from a single observation X with the prior distribution of θ being beta with parameters α and β, investigate the form of the posterior distribution of θ and determine the Bayesian estimator of θ under quadratic loss.

Solution

The proportionality argument will be used and any constants simply omitted as appropriate.

Prior: $f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$, **omitting the constant** $\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$.

Likelihood: $f(X|\theta) \propto \theta^{x}(1-\theta)^{n-x}$, **omitting the constant** $\binom{n}{x}$.

$$\therefore\ f(\theta \mid X) \propto \theta^{x}(1-\theta)^{n-x} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}.$$

Now it can be seen that, apart from the appropriate constant of proportionality, this is the density of a beta random variable. Therefore the immediate conclusion is that the posterior distribution of θ given X is beta with parameters $x+\alpha$ and $n-x+\beta$.

It can also be seen that the posterior density and the prior density belong to the same family of distributions. Thus the conjugate prior for the binomial distribution is the beta distribution.

The Bayesian estimator under quadratic loss is the mean of this distribution, that is, $\dfrac{x+\alpha}{(x+\alpha)+(n-x+\beta)} = \dfrac{x+\alpha}{n+\alpha+\beta}$.

Question 2.8

What would be the estimate using all-or-nothing loss?

Likelihood of IID Observations $X_1,\ldots,X_n$	Unknown Parameter	Distribution of parameter	
		Prior	Posterior
$Poisson(\lambda)$	$\lambda > 0$	$Exp(\lambda')$	$Gamma(\sum x+1, n+\lambda')$
$Exp(\lambda)$			$Gamma(n+1, \sum x+\lambda')$
$Gamma(\alpha,\lambda)$			$Gamma(n\alpha+1, \sum x+\lambda')$
$Exp(\lambda)$	$\lambda > 0$	$Gamma(\alpha',\lambda')$	$Gamma(n+\alpha', \sum x+\lambda')$
$Gamma(\alpha,\lambda)$			$Gamma(n\alpha+\alpha', \sum x+\lambda')$
$B(m,p)$	$0 < p < 1$	$Beta(\alpha',\beta')$	$Beta(\sum x+\alpha', nm-\sum x+\beta')$
$Geo(p)$			$Beta(n+\alpha', \sum x+\beta')$
$N(\mu,\sigma^2)$	$-\infty < \mu < \infty$	$N(\mu',\sigma'^2)$	$N\left(\frac{\frac{\sum x}{\sigma^2}+\frac{\mu'}{\sigma'^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma'^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma'^2}}\right)$
$LogN(\mu,\sigma^2)$	$-\infty < \mu < \infty$	$U(-\infty,\infty)$	$N\left(\frac{1}{n}\sum \log x, \frac{\sigma^2}{n}\right)$
$N(\mu,\sigma^2)$			$N\left(\frac{1}{n}\sum x, \frac{\sigma^2}{n}\right)$
$Exp(\lambda)$	$\lambda > 0$	$U(0,\infty)$	$Gamma(n+1, \sum x)$
$Gamma(\alpha,\lambda)$			$Gamma(n\alpha+1, \sum x)$
$Geo(p)$	$0 < p < 1$	$U(0,1)$	$Beta(n+1, \sum x+1)$
$B(m,p)$			$Beta(\sum x+1, nm-\sum x+1)$
$NB(k,p)$			$Beta(nk+1, \sum x+1)$

Chapter 2 Summary

A common problem in statistics is to estimate the value of some unknown parameter θ.

The classical approach to this problem is to treat θ as a fixed, but unknown, constant and use sample data to estimate its value. For example, if θ represents some population mean then its value may be estimated by a sample mean.

The Bayesian approach is to treat θ as a random variable. In the absence of any sample data, the knowledge available about the possible values of θ can be summarised in a prior distribution.

A likelihood function is then determined, based on a set of observations $x_1, x_2, \ldots, x_n$. The likelihood function is just the same as the joint density (or, in the discrete case, the joint probability) of $X_1, X_2, \ldots, X_n|\theta$. However, the likelihood is considered to be a function of θ rather than one of $x_1, x_2, \ldots, x_n$. Since we are assuming that the random variables $X_1|\theta, \ldots, X_n|\theta$ are independent, the joint density function $f_{\underline{X}|\theta}(x_1, x_2, \ldots, x_n)$ is equal to the product of the individual density functions $f_{X_i|\theta}(x_i)$. The likelihood function is therefore:

$$L(\theta) = \prod_{i=1}^{n} f_{X_i|\theta}(x_i)$$

The prior density and the likelihood function are then combined to obtain the density function of the posterior distribution for θ.

A loss function, such as quadratic (mean square) error loss, absolute error loss or zero-one loss, is chosen that specifies the seriousness of an incorrect estimate.

Under squared error loss, the Bayesian estimate that minimises the expected loss is the mean of the posterior distribution. Under absolute error loss, it is the median of the posterior distribution that minimises the expected loss. Using a 0-1 loss function, the mode of the posterior distribution minimises the expected loss.

In many situations there is a natural prior distribution to use that leads to conjugate prior and posterior distributions.

Chapter 2 Formulae

Bayes' Formula

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_i P(A|B_i)P(B_i)} \qquad \text{(discrete form)}$$

$$f(\theta|A) = \frac{P(A|\theta)f(\theta)}{\int P(A|\theta)f(\theta)\,d\theta} \qquad \text{(continuous form)}$$

Posterior distribution

Posterior $\propto$ Prior $\times$ Likelihood

Loss function	Loss	Bayesian estimator
Squared error	$(\hat{\theta}-\theta)^2$	mean
Absolute error	$\lvert\hat{\theta}-\theta\rvert$	median
Zero-one	0 if $\hat{\theta}=\theta$ 1 if $\hat{\theta}\neq\theta$	mode

#