The Bayesian Approach to Statistics:

A review of methodology with selected applications.

A thesis presented in partial fulfilment

of the requirements for the degree of

Master of Science

in

Statistics

at

Massey University.

Shane Byram Wood.

1984

Table of Contents

Abstract

In this thesis we present a review of the Bayesian approach to Statistical Inference. In Chapter One we develop the theory and methodology behind the approach. Starting from its basis in subjective probability we outline the Bayesian philosophy towards such problems as Point and Interval estimation, Hypothesis testing and Decision Theory. For each of these areas, we indicate the corresponding Classical approach and comment on the differences between this and the Bayesian one. We then develop the idea of conjugate families of prior distributions which is central to the practice of Bayesian statistics, and follow this with a section on the assessment of subjective probability distributions, their functional specification and the problem of mathematically representing a state of 'ignorance'. The Decision Theoretic approach to statistical analysis is then integrated into the Bayesian framework, and reference is made to the assessment of 'loss' functions, and their subjective nature. Finally we consider the concepts of Empirical Bayes, Exchangeability, and Likelihood, and their relevence to the Bayesian scheme.

Chapter Two consists of a review of areas such as econometrics, medicine, industry, and education, where Bayesian methods have been applied, accompanied by a number of particularly interesting applications which illustrate the principles outlined in chapter one.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr Howard Edwards, for the help and encouragement given me in the preparation of this thesis. I would also like to thank my sister Odette, who patiently spent many long hours typing the manuscript.

CHAPTER ONE

Background to Bayesian Statistics

1.1 Historical Development

Folks 1981, gives an interesting introduction to the work of Bayes:

> About half a mile north of Gresham College, where Karl Pearson gave his lectures on probability and statistics, lies Bunhill Fields. Across City Road from the chapel and home of John Wesley, Bunhill Fields, burial ground for 120,000 souls, is the famous cemetery of the nonconformists, disused since 1852. Here are the graves of John Bunyan, Daniel Defoe, Isaac Watts, and Susannah Wesley, among many other famous people. Here also are the graves of Richard Price and Thomas Bayes.
>
> Readers of current statistics textbooks and journals may search in vain for the names of Fisher and Pearson but may see repeated references to Bayes. Beginning students might well conclude that Bayes, not Pearson, was the founder of modern statistics! In fact, Bayes made almost no contribution to statistics. Why, then, the repeated references to Bayes? The answer seems to lie in a paper by Bayes that has become the focal point for the mode of reasoning called Bayesian statistics.

This paper, published in 1763, was submitted after Bayes' death by his friend, Richard Price. It was entitled "An essay towards solving a problem in the doctrine of chances" and has earned him a crucial and somewhat controversial position in the development of statistical inference. Although there are diverging opinions over precisely what Bayes was proposing in his Essay, two ideas do stand out as being clearly understood.

These are: (1) Bayes Theorem

(2) Bayes Principle of insufficient reason

The former is simply a statement in conditional probabilities, and as such, few could find fault with it. The latter is essentially a statement that if we have no reason to believe that one event or hypothesis from a set of possible events (hypotheses), is more likely to arise than any other, then we should assume all events (hypotheses) are equally likely. This use of principle, therefore enables one to make a quantitative description of the state of ones `ignorance`.

Much of the controversy surrounding Bayes' work and that of his followers relates to the subjective or degree of belief view of probability which must be adopted for so called Bayes' methods to provide their strongest challenge to Classical Statistics.

1.2 Bayes Theorem

Consider a mutually exclusive, exhaustive set of events A1 ... Ak. and let B be some other event of interest. Assume the probabilities P(Ai) i=1...k, of each event Ai are known, and $\sum_i P(Ai) = 1$. Also known are the conditional probabilities P(B|Ai) i=1...k, that is the probability of the event B occurring given that the event Ai has occurred. Then the conditional probability of any Ai i=1...k, given that B has occurred, is given by ...

$$P(Ai|B) = \frac{P(B|Ai) \cdot P(Ai)}{\sum_i P(B|Ai) \cdot P(Ai)} \qquad i=1...k$$

Proof

$$P(B|Ai) \cdot P(Ai) = P(B \cup Ai)$$

$$= P(Ai|B) \cdot P(B)$$

$$\sum_i P(B|Ai) \cdot P(Ai) = \sum_i P(B \cup Ai)$$

$$= P(B \cup A1) + \ldots + P(B \cup Ak)$$

$$= P(B)$$

And $P(Ai|B) \cdot P(B) = P(B|Ai) \cdot P(Ai)$

so $P(Ai|B) = \frac{P(B|Ai) \cdot P(Ai)}{P(B)}$

$$= \frac{P(B|Ai) \cdot P(Ai)}{\sum_i P(B|Ai) \cdot P(Ai)} \qquad i=1...k$$

As previously stated, this theorem as written, is widely used and invokes no controversy regarding its interpretation.

The controversy arises when instead of considering a series of events A1...Ak we consider a set of hypotheses H1...Hk, concerning what constitutes an acceptable model for some practical situation.

The `event` B now represents the data observed in the same situation. Before B is observed, the probabilities P(Hi) that the hypotheses Hi represent an adequate specification of the model are known for

i=1...k. These probabilities are called prior probabilities and constitute a secondary source of information. The probabilities $P(B|Hi)$ i=1...k of observing the data {B}, when Hi is the correct model specification, represent nothing more than the likelihoods of the sample data.

Bayes theorem can now be reinterpreted as a method of updating our prior knowledge concerning the model, through the use of the sample data which expresses itself in the likelihood function. This updated specification of the model is represented by the posterior probabilities $P(Hi|B)$ i=1...k of the hypotheses H1...Hk being `true` after we have utilised the information contained in the sample data.

Example

A game is to be played using a biased coin. Before tossing the coin, it is believed that the proportion of heads, θ, has the following distribution.

$$p(\theta) = \begin{cases} 0.2 & \text{if } \theta = 0.25 \\ 0.3 & \text{if } \theta = 0.50 \\ 0.5 & \text{if } \theta = 0.75 \end{cases}$$

The coin is then tossed five times and three heads are observed. What is the posterior distribution for the proportion of heads observed ?

Let X=3 be the event that three heads are observed in five tosses of the coin. The distribution fof the number of head is given by ...

$$p(X=3|\theta) = \binom{5}{3} \theta^3(1-\theta)^2$$

So $p(X|\theta=0.25) = 0.0879$ where $p(\theta=0.25) = 0.2$

$p(X|\theta=0.50) = 0.3125$ where $p(\theta=0.50) = 0.3$

and $p(X|\theta=0.75) = 0.2637$ where $p(\theta=0.75)=0.5$

And the posterior probabilities are given by ...

$$p(\theta|X=3) = \frac{p(X=3|\theta)\ p(\theta)}{\Sigma\ p(X=3|\theta)\ p(\theta)}$$

i.e. $$p(\theta=0.25|X=3) = \frac{0.2 \times 0.0879}{0.2 \times 0.0879 + 0.3 \times 0.3125 + 0.5 \times 0.2637}$$

$$= \frac{0.01758}{0.24318} = 0.072$$

Also $P(\theta=0.50|X=3) = 0.386$

and $P(\theta=0.75|X=3) = 0.542$

i.e. $$p(\theta|X=3) = \begin{cases} 0.072 & \text{if } \theta = 0.25 \\ 0.386 & \text{if } \theta = 0.50 \\ 0.542 & \text{if } \theta = 0.75 \end{cases}$$

1.3 The Meaning of Probability

Throughout our everyday lives we constantly encounter statements about probability,likelihood or chance. These are words that everyone has in their vocabulary.

We all have some experience of statements such as : "It is unlikely that the All-Blacks will be beaten by Fiji in the upcoming tour" , "It will probably rain tomorrow", "The probability that this coin will land heads up is one half". Often such statements are made without a conscious interpretation let alone any numerical assessment of the uncertainty inherent in the statement. In most `everyday' uses they merely express a personal conviction that the proposition under consideration lies in some intermediate position on the impossibility-inevitability scale.

1.3.1 Classical Probability

People have been trying to derive numerical measures to adequately express their convictions since the 16th century when games of chance became popular. However, at this stage, most interest centred on the evaluation of probability rather than definition of the concept. It was not until the time of De Moivre and later on Laplace, that any serious attempt was made to define it.

Laplace defined probability as "the ratio of the number of outcomes favourable to the event, to the total number of possible outcomes, each assumed to be equally likely".

This Classical definition of probability, as Barnett calls it, was accepted until early this century.

The reasons for which this definition was rejected relate to its roots in the concept of equally likely outcomes.

What does equally likely mean ?

Our first answer to this question would be on the basis of equal probability but this clearly involves a circular argument.

How do we recognise equally likely outcomes ?

Here we must resort to symmetry arguments, but on what grounds does physical symmetry imply equal probabilities? And even if this argument does hold, can there be any such thing as perfect symmetry!

But the most serious criticism is the restricted number of situations in which the concept of equally likely events can be applied. If we were to accept this concept of probability we would find that most areas of human enquiry would lie outside the scope of probability theory.

1.3.2 The Frequency View

The view of probability which replaced the classical one was the so called frequency one. The frequency concept was first formally defined by Venn in the 19th century as the "limiting value of relative frequencies in infinite sequences of repeatable and essentially identical situations". However it was von Mises who early this century provided the Mathematical basis for the frequency view and in doing so put the final nail in the `Classical` coffin.

As Laplace based his definition of probability on the concept of equally likely events, von Mises based his on his idea of the `collective`.

He defines this as:

> "...a mass phenomenon or an unlimited sequence of observations fulfilling the following two conditions:
>
> (i) the relative frequencies of particular attributes within the collective tend to fixed limits.
>
> (ii) these fixed limits are not affected by any place selection. That is to say, if we calculate the relative frequency of some attribute not in the original sequence, but in a partial set, selected according to some fixed rule, then we require that the relative frequency so calculated should tend to the same limit as it does in the original set."

It is this `collective` among other things which provides the basis for much of the criticism of the frequency view. As in the case of Classical probability it can be forcefully argued that the frequency definition of probability is far too restrictive, in that only experiments that can be viewed as repeatable under essentially identical conditions lie within the scope of statistical enquiry.

1.3.3 Subjective Probability

A further interpretation of probability is the subjective one. This is the interpretation which the majority of people would attach to probability. Statements such as "The odds are 50:50 of USA regaining the Americas Cup in 1985", or "I have a chance of two in three of getting this job" appear to relate to probabilities but it is not at all clear how we can attach a long run relative frequency interpretation to their occurrance. They give a persons degree of belief about an event that will occur once only and so there can be no possibility of repeated trials and thus no frequency interpretation. This is in fact the unique feature of Subjective Probability that distinguishes it from the frequency approach. For a Subjectivist and therefore a Bayesian, a probability is interpreted as a degree of belief held by a particular individual about some hypothetical event or uncertain quantity. It is obvious from this statement that probability is a personal thing and that different individuals may have different probabilistic assessments of the same situation. This is a natural consequence of the fact that different people will have different personal experiences that they will utilise in their assessment of probability. This is one of the main criticisms of the subjective viewpoint, in that, since subjective judgement is involved in the determination of probability, then the resulting probabilities cease to be objective but are in some respects arbitrary.

There are however, objective procedures for the assessment of probability which nullify much of this criticism. These procedures revolve around the use of betting situations. The probability that an individual assigns to an event E occuring, can be defined as the odds at which he would be willing to bet that E will happen. For example, if we were to say that "The odds are two to one that it will rain tomorrow",

this implies P(Rains tomorrow | Personal Experiences) = 2/3.

and P(Doesn't Rain | Personal Experiences) = 1/3.

and so we would be willing to accept the bet:

A bets $2 that it will rain tomorrow.

B bets $1 that it won't rain tomorrow

as a fair bet in the sense that we would be equally willing to take either side of the bet.

ie. Taking A's side of the bet:

We expect to win $1 (B loses) with probability 2/3.

and we expect to lose $2 with probability 1/3.

and so our expected gain is 2/3.$1 - 1/3.$2 = 0.

Similarly B can also expect to gain $0. The difficulty with this sort of bet is that while we may view the above as a fair bet, if we were to change the value of the bet to:

A bets $2000 that it will rain tomorrow.

B bets $1000 that it won't rain tomorrow.

Then we may no longer accept this bet as fair. If we don't have $2000 we would probably not accept any bet with this at stake. This leads us on to the concept of utility, the value which an individual attaches to an amount of money. This concept is central to the application of Bayesian techniques in decision theory.

Efforts have been made to utilise betting procedures for the assessment of probabilities without the intrusion of the utility question. These efforts are centred on the existance of an ethically neutral proposition (Ramsay) which can be used as a reference with which to compare our own proposition. For example we may be offered the chance of risking a loss if it rains tomorrow, or the same loss if a fair die gives a 1, 2, 3 or 4. If we accept this bet we assign the same probability to it raining tomorrow as to the die coming up 1, 2, 3 or 4. ie. 2/3. If we are to use such betting situations to assess our personal probability, we must require ourselves to behave reasonably or coherently in terms of the odds or bets we will be prepared to accept! The principle of Coherence requires that we would not accept any bet where we will lose whether or not the event of interest occurs. Further more it requires us to order our preferences for bets. If we prefer bet A to bet B and bet B to bet C then we must prefer bet A to bet C. The acceptance of these principles enables the development of a set of `probability laws` similar to those developed from more traditional approaches.

It should be apparent from this discussion that the concept of subjective probability is built on a substantially different framework to that of its main rival, the frequency approach. The debate over which concept is most valid is still a cause for much heated argument between Statisticians. However the writer believes that in so far as the subjective approach greatly expands the realms of possible scientific investigation as well as imparting a deeper personal appreciation of the meaning of probability, then this approach is not to be rejected lightly.

1.4 Bayesian Inference

Before proceeding further in our discussion it is useful to rewrite Bayes Theorem in terms of probability density functions.

Suppose $\underline{x} = (x_1, x_2 \ldots, x_n)$ is a vector of n observations on a random variable X, and its probability density function depends on the value of k parameters, $\theta_1 \ldots \theta_k$. We will represent this pdf by $p(\underline{x}|\underline{\theta})$ where $\underline{\theta}=(\theta_1 \ldots \theta_k)$.

Now suppose $\underline{\theta}$ itself has a density function given by $p(\underline{\theta})$

then $$p(\underline{x}|\underline{\theta}) \cdot p(\underline{\theta}) = p(\underline{x},\underline{\theta})$$

is the joint pdf of $\underline{x}$ and $\underline{\theta}$.

But we can also write ... $$p(\underline{\theta}|\underline{x}) \cdot p(\underline{x}) = p(\underline{x},\underline{\theta})$$

and so ... $$p(\underline{x}|\underline{\theta}) \cdot p(\underline{\theta}) = p(\underline{\theta}|\underline{x}) \cdot p(\underline{x})$$

which gives ... $$p(\underline{\theta}|\underline{x}) = \frac{p(\underline{x}|\underline{\theta}) \cdot p(\underline{\theta})}{p(\underline{x})}$$

Clearly $p(\underline{x})$ is just the marginal pdf of $\underline{x}$ found by integrating $p(\underline{x},\underline{\theta})$ wrt $\underline{\theta}$.

$$p(\underline{x}) = \int_{\theta} p(\underline{x}|\underline{\theta})p(\underline{\theta})d\underline{\theta}$$

And so we have $$p(\underline{\theta}|\underline{x}) = \frac{p(\underline{x}|\underline{\theta})p(\underline{\theta})}{\int_{\theta} p(\underline{x}|\underline{\theta})p(\underline{\theta})d\underline{\theta}}$$

which is Bayes Theorem for continuous random variables.

We can further simplify this expression.

Since the quantity $p(\underline{x})$ is merely a normalizing constant wrt $\underline{\theta}$, which ensures that $p(\underline{\theta}|\underline{x})$ integrates to 1,

we now have ... $$p(\underline{\theta}|\underline{x}) \propto p(\underline{x}|\underline{\theta}) \cdot p(\underline{\theta})$$

Where $p(\underline{\theta})$ tells us what is known about $\underline{\theta}$ a' priori, (before sampling) and $p(\underline{\theta}|\underline{x})$ tells us what is known about $\underline{\theta}$ posterior to the

sampling process.

It is the specification and interpretation of $p(\underline{\theta})$ that is the basis for most of the criticism leveled at Bayesian methods. This important area will be considered in more depth in Section 1.6.

As we have previously stated, the relationship given by Bayes Theorem provides a means of updating our prior knowledge of $\underline{\theta}$, with the use of the sample likelihood function, to produce a more complete expression of our knowledge concerning $\underline{\theta}$. However this updating need not be a one-off process. Bayes Theorem enables us to continuously update our knowledge of $\underline{\theta}$ as more observations are taken.

Assume an initial sample $\underline{x}_1$ when combined with $p(\underline{\theta})$ yields the posterior ... $P(\underline{\theta}|\underline{x}_1) \propto p(\underline{x}_1|\underline{\theta}) \cdot p(\underline{\theta})$

If a further sample $\underline{x}_2$ distributed independently of $\underline{x}_1$ is taken, then our posterior is now given by ...

$$p(\underline{\theta}|\underline{x}_1,\underline{x}_2) \propto p(\underline{x}_1|\underline{\theta}) \cdot p(\underline{x}_2|\underline{\theta}) \cdot p(\underline{\theta})$$
$$\propto p(\underline{\theta}|\underline{x}_1) \cdot p(\underline{x}_2|\underline{\theta})$$

And so the posterior in the first stage of sampling becomes the prior in the second stage, or in other words, the distribution of $\underline{\theta}$ posterior to the observation of $\underline{x}_1$, becomes the distribution of $\underline{\theta}$ prior to the observation of $\underline{x}_2$. Naturally this process can be extended as many times as required to represent at any stage our current state of knowledge regarding $\underline{\theta}$. In this way, the use of Bayes Theorem enables us to "learn from experience".

It is apparent that the prime motivation behind Bayesian methods is the desire to base any possible inference or decision on all the

available information, whether this be from the sample or some other source.

This philosophy is particularly relevant in decision-theoretic appplications in such areas as marketing and industry where the financial cost of an incorrect decision may be very large, and prior knowledge of $\underline{\theta}$ is usually available and acceptable.

In contrast, the only information that classical statistics allows as input to any inferential procedure is the sample itself and the choice of initial precision. (e.g. In terms of Type One and Two errors.) The choice of the latter is essentially a subjective one itself. It is interesting to note here that the use of the sampling distribution on which the frequentist bases his inferences is not excluded from the Bayesian approach. Providing the Bayesian can accept subjectively any assumptions such as normality, known mean or variance etc, upon which the frequentist bases his distribution, then he will face no conflict in giving the distribution a subjective interpretation. It appears that the frequentist, in his subjective assessment of prior errors, and the Bayesian, in his utilisation of classical sampling distributions, (albeit with his own interpretation) are both to some extent, utilising some of the concepts on which the other's arguments are based!

Because the posterior distribution $p(\underline{\theta}|\underline{x})$, is supposed to be a complete representation of all available knowledge about $\underline{\theta}$, then any inferences which are to be made concerning $\underline{\theta}$ should be based solely on this. (Except in Decision Theory applications which will be considered later.) The two main areas of classical inference, that is estimation and hypothesis testing, have not been ignored by the Bayesian Statistician. However, because of the substantial

differences in the Bayesian and Classical philosophies, the Bayesian approach to these problems is quite different in some cases. Indeed there are subtle differences even within the Bayesian school itself. Some authors, notably Box and Tiao contend that the best way to convey information about $\underline{\theta}$ is merely to give the posterior distribution itself, rather to resort to Classical ideas such as estimation of $\underline{\theta}$. However, although it is true that the posterior distribution does contain all the known information about $\underline{\theta}$, it is also true that it is often easier (especially for a layman) to comprehend the significance of a single estimate of $\underline{\theta}$, rather than the functional form of the entire pdf!