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..... iv

Acknowledgements..... v

Background to Bayesian Statistics.....1

1.1 Historical Development.....1

1.2 Bayes Theorem.....3

1.3 The Meaning of Probability.....6

1.3.1 Classical Probability.....6

1.3.2 The Frequency View.....7

1.3.3 Subjective Probability.....9

1.4 Bayesian Inference.....12

1.4.1 Estimation.....16

1.4.1.1 Point Estimation.....16

1.4.1.2 Interval Estimation.....17

1.4.2 Hypothesis Testing.....19

1.5 Conjugate Families of Prior Distributions.....24

1.6 Prior Information.....28

1.6.1 Subjective determination of prior density functions.....29

1.6.2 Substantial Prior Information.....29

1.6.2.1 The Grouping and Smoothing (Histogram) Technique.....30

1.6.2.2 The Relative Likelihood Approach.....31

1.6.2.3 Use of Location and Scale Parameters.....31

1.6.3 Vague or Nonexistent Prior Information.....32

1.6.3.1 Jeffreys Invariance Rule.....34

1.6.3.2 Data translated prior distributions.....38

1.7 Bayesian Decision Theory.....45

1.7.1 The Decision Theory Model.....46

1.7.1.1 Prior Knowledge Only.....47

1.7.1.2 Prior knowledge and sample data.....48

1.7.2 Selecting The Best Decision Rule.....50

1.7.2.1 The Bayes Decision Rule.....50

1.7.2.2 The minimax decision rule.....53

1.7.3 The Utility Concept.....55

1.7.3.1 The Assessment of Utilities.....57

1.7.3.2 Loss Functions.....58

1.7.3.3 Criticisms of Utility Functions.....59

1.7.4 The Decision-Theoretic approach to Statistical Inference.....60

1.7.4.1 Estimation.....60

1.7.4.2 Hypothesis Testing.....61

1.8 Further topics in Bayesian Statistics.....62

1.8.1 Exchangeability.....62

1.8.2 The likelihood principle.....64

1.8.3 Empirical Bayes Methods.....65

Introduction to the Applications of Bayesian Statistics.....69

2.1 Econometrics and Business Studies.....71

2.1.1 A Bayesian Approach to Real Estate Assessment.....74

2.1.2 Derivation of Predictive distribution using a Diffuse prior.....76

2.1.3 Predictive Distribution for a Non Diffuse Prior.....78

2.1.4 Loss minimising estimator.....79

2.1.5 The Data.....80

2.1.6 Comparative results and conclusions.....82

2.2 Medical, Scientific and Industrial Applications.....84

2.2.1 A Bayesian Modification of the Lincoln Index.....89

2.3 Education Applications.....92

2.4 Concluding Comments.....95

Appendix One.....97

Appendix Two.....98

Bibliography.....99

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!

1.4.1 Estimation

1.4.1.1 Point Estimation

In Classical statistics we encounter a wide variety of estimators for $\underline{\theta}$, many of which have certain optimal properties such as unbiasedness, most efficient, minimum variance, etc. However because Bayesian statistics do not utilise any long run frequency properties of estimators, concepts such as `correct on average`, or `minimum variance over all possible samples`, are no longer valid, since we are considering a particular sample that we have observed, not just one possible sample out of many that we might have observed. In this context, it would seem logical to choose as the best estimator for $\underline{\theta}$, the value which is most likely, given our prior knowledge and the observed data. In other words we should choose the value of $\underline{\theta}$ with the highest posterior probability of being correct (given the particular sample observed).

ie. $\underline{\theta}$ maximises $p(\underline{\theta}|\underline{x})$ for the specific sample $\underline{x}$ observed.

So that our estimate for $\underline{\theta}$ is the mode of the posterior distribution.

A criticism of the use of the mode as an estimator for $\underline{\theta}$ is that it is generally not invariant under transformation of the parameter space. A common refutation of this (Box and Tiao), is that although the choice of parameterisation is arbitrary, this arbitrariness exists in the specification of any statistical model and as long as the the conclusions provide a realistic approximation to the truth then the parameterisation is acceptable. Other point estimators which are used in Bayesian statistics are the mean and median, however these are mainly used in the area of Bayesian decision

theory.

1.4.1.2 Interval Estimation

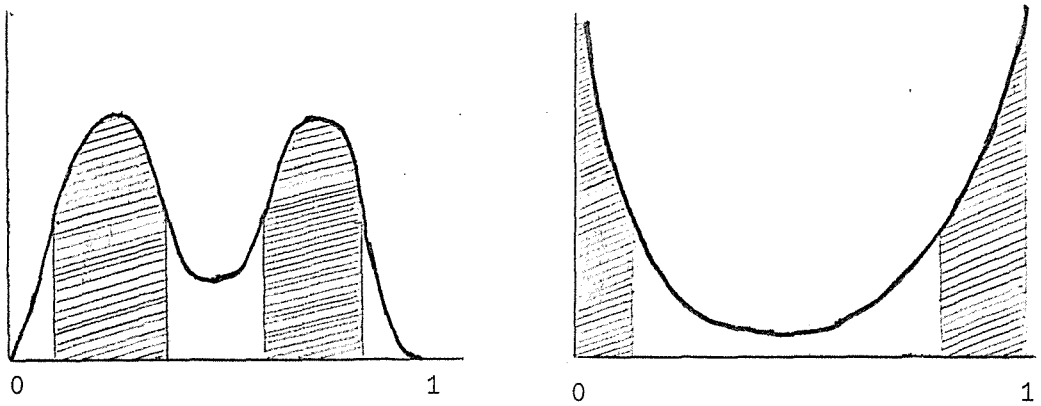
Because the posterior distribution $p(\underline{\theta}|\underline{x})$ contains all the information known about $\underline{\theta}$, it allows us to evaluate the probability of $\underline{\theta}$ lying in any given region or interval of Θ. We can define a Bayesian interval estimate of $\underline{\theta}$ as one which contains $\underline{\theta}$ with known probability $1-\alpha$. We call this a $100(1-\alpha)\%$ credible interval for $\underline{\theta}$. This interval is a subset C of Θ such that ...

$$\int_{\theta \in C} p(\underline{\theta}|\underline{x}) d\underline{\theta} \geq 1-\alpha$$

Classical intervals are based on such concepts as equal tail end probabilities, minimum width and greatest accuracy in some sense. These intervals are often designed with certain optimality properties such as being uniformly most powerful etc. However such properties are based on the Classical assessment of initial precision rather than final precision which the Bayesian method assesses, and as such can really have no relevence in the selection of Bayesian intervals.

Two properties do spring to mind as being desirable in a Bayesian credible interval. One property is that the posterior density for every point within the interval should be greater than for every point outside the interval. Secondly the interval should be as short as possible for a given probability of coverage (That is the probability that $\underline{\theta}$ lies in the interval given our prior knowledge and the observed sample). These two properties are in fact equivalent. A Bayesian credible interval which satisfies these properties is known as a Highest Posterior Density (HPD) credible interval. An interesting feature of this interval is that in the case of a bimodal

posterior density function it is possible for the HPD credible interval to to consist of two disjoint intervals. In such cases the HPD region although theoretically correct, produces some philosophical difficulties in its interpretation.



(a) Cauchy-type distribution. (b) Beta-type distribution.

Figure 1(a), 1(b) Illustration of disjoint HPD regions.

In some cases, notably that of no prior knowledge (see Section 1.6.3) and a symmetric likelihood function such as the normal distribution, the Bayesian HPD $100(1-\alpha)\%$ credible interval and the Classical $100(1-\alpha)\%$ confidence interval will coincide. However it is very important we realise that the interpretation of the two intervals is quite different. In such cases a Bayesian would say that the probability that his interval contains $\underline{\theta}$ (based on all his available information) is $1-\alpha$, whereas the frequentist would say that $100(1-\alpha)\%$ of similar intervals, computed from repeated sampling would contain $\underline{\theta}$. The latter, in basing his inferences on repeated sampling is able to say nothing about the probability that his particular interval contains $\underline{\theta}$. Conversely, the Bayesian can not impart to his inferences any long run frequency interpretation and any probability

statements that he makes can only relate to the specific situation (ie. this sample, this prior information). It is difficult to decide which interpretation is `best`, based as they are in their own philosophies of probability. However it is interesting to note that most laymen when confronted with a Classical Confidence Interval will interpret the probability statement not in terms of long run frequency but as a probability of coverage for the specific interval in a specific situation. If we are to believe that the role of the statistician is to produce information in a form which is easily interpretable to the `statistically ignorant` researcher, then I believe the Bayesian approach, if its subjective basis can be accepted, does offer a particularly strong challenge to Classical statistics!!

1.4.2 Hypothesis Testing

One of the most common problems in statistics is that of comparing two or more hypotheses concerning the value of some parameter $\underline{\theta}$. This is but another area where the Bayesian and Classical approaches to a problem differ vastly. The Classical approach to the so called hypothesis test, involves testing a null hypothesis concerning a parameter $\underline{\theta}$ against an alternative hypothesis concerning the same parameter. These hypotheses may relate to single values of $\underline{\theta}$, or to regions in which $\underline{\theta}$ is believed to lie. Generally the burden of proof lies with the alternative hypothesis in so far as it is the null hypothesis that is postulated as initially being true and must be discredited for the alternative to be prefered. The Classical test is typically based on the likelihood ratio test statistic and is formulated in terms of the prespecified probabilities of rejecting

the null hypothesis when it is really true (Type One error) and accepting the null hypothesis when it is false. (Type Two error) These probabilities are both measures of initial precision and as they are inversely proportional there must often be a subjective decision as to which error is more important in a given problem. Traditionally the Type One error probability is prespecified at some small value at the expense of the Type Two error probability. As is to be expected the Bayes procedure for Hypothesis testing makes no appeal whatsoever to initial precision. One merely calculates the posterior probability for each hypothesis and decides between them accordingly.

Consider the one tailed test of $Ho: \theta \leq \theta_o$ vs $Ha: \theta > \theta_o$.
In this case the Bayesian need only evaluate ...

$$\int_{\theta \leq \theta_o} p(\theta|\underline{x})d\theta \quad \text{and} \quad \int_{\theta > \theta_o} p(\theta|\underline{x})d\theta \quad \text{in the continuous case,}$$

$$\text{or} \sum_{\theta \leq \theta_o} p(\theta|\underline{x}) \quad \text{and} \sum_{\theta > \theta_o} p(\theta|\underline{x}) \quad \text{in the discrete case,}$$

and make his decision on the basis of these probabilities.

Since these probabilities are that of Ho and Ha being true (given the situation of the observed data and our prior knowledge) then they are measures of final precision rather than the initial precision of the Classical hypothesis test.

The decision as to the levels of posterior probability at which we will accept one hypothesis over another will often be made on the basis of the consequences of the decision. For this reason the area of Bayesian decision theory in which the costs of various actions are formulated explicitly in terms of loss or utility functions, has been

one of the fastest growing areas in Bayesian inference. This area of statistics is particularly relevant in fields such as econometrics where there is a definite loss, (often financial) from making an incorrect decision or estimate and the Bayesian framework itself is particularly useful, in that there is frequently only a limited amount of data available which can be most productively utilised only when augmented by other (prior) information available to the researcher. (In econometric problems the volume of such 'prior' information is often considerable)

Rejection or acceptance of hypotheses in Classical hypothesis testing will often depend on how conservatively the statistician wishes to regard the hypotheses of interest. As stated earlier, it is usually the null hypotheses that is regarded with most conservatism and this is reflected in the choice of significance level.

This conservatism also extends itself to Bayesian hypothesis testing, in the use of the posterior odds ratio. This is the ratio between the posterior probabilities of the two (or more) hypotheses of interest and represents the relative plausibility of the alternatives.

Consider hypotheses H1:$\theta \in C$ and H2:$\theta \notin C$, where C is some subset of the parameter space Θ.

The posterior odds ratio is written as ...

$$\frac{\int_{\theta \in C} p(\theta|\underline{x})d\theta}{\int_{\theta \notin C} p(\theta|\underline{x})d\theta} = \frac{\int_{\theta \in C} p(\theta|\underline{x})d\theta}{[1 - \int_{\theta \notin C} p(\theta|\underline{x})d\theta]} = R$$

In general a rejection region of the form $R \leq k$ is used where k is some critical value.

Difficulties may arise when one or more of the hypotheses is a point hypothesis, (ie. tests that $\underline{\theta}$ is equal to some specific value $\underline{\theta_o}$) and $\underline{\theta}$ has a continuous distribution. In such cases the Classical approach faces similar difficulties, even though much of Classical hypothesis testing is based around the fictional concept of a `point` null hypothesis. The problem with the Bayesian approach is that when we evaluate the posterior probability of the point hypothesis we find it is zero. This is because for a continuous density, the area under a point is zero. This result reflects the absurdity of the point hypothesis. Rather than asking if $\underline{\theta} = \underline{\theta}o$ exactly, it is often more reasonable to enquire if $\underline{\theta}$ is close to $\underline{\theta}o$ so our hypothesis should really be ... H:$\underline{\theta} \in (\underline{\theta}o - \underline{d}, \underline{\theta}o + \underline{d})$
where $\underline{d}$ is a suitable constant vector.

In some cases however, a point hypothesis will be perfectly reasonable, for example in testing that the concentration of a particular solvent is the same as that advertised. Also in the case that $\underline{\theta}$ is a discrete parameter it may often be reasonable to test if

$\underline{\theta}$ has a specific value. In these cases the general procedure is to assign a prior concentration of probability $\pi > 0$ at the value of $\underline{\theta} = \underline{\theta}o$ and spread the remaining probability $1-\pi$ amongst the values $\underline{\theta} \neq \underline{\theta}o$. The decision on whether to accept or reject this point hypothesis can now be made on the basis of the posterior odds ratio.

ie. $$\frac{p(\underline{\theta} = \underline{\theta}o|\underline{x})}{\int_{\theta \neq \theta o} p(\underline{\theta}|\underline{x})\,d\underline{\theta}} = \frac{\pi \cdot p(\underline{x}|\underline{\theta}o)}{(1-\pi)\int_{\theta \neq \theta o} p(\underline{x}|\underline{\theta}p(\underline{\theta})\,d\underline{\theta}}$$

where $\pi + (1-\pi)\int_{\substack{\theta \neq \theta o \\ \theta \varepsilon \Theta}} p(\underline{\theta})\,d\underline{\theta} = 1.$

1.5 Conjugate Families of Prior Distributions

Apart from the specification of prior information in a numerical form, one of the most crucial problems in Bayesian analysis is the evaluation of the posterior distribution.

Bayes Theorem is: $$p(\underline{\theta}|\underline{x}) = \frac{p(\underline{\theta}) \cdot p(\underline{x}|\underline{\theta})}{\int p(\underline{\theta}) \cdot p(\underline{x}|\underline{\theta}) d\underline{\theta}}$$

which can be conveniently expressed as:

$$\text{posterior density} = \frac{\text{prior density . sample likelihood}}{\int \text{prior density . sample likelihood}}$$

There may be considerable difficulty in the evaluation of the intergral in the denominator for some choices of $p(\underline{\theta})$ and $p(\underline{x}|\underline{\theta})$. Although numerical techniques now exist which will in most cases of practical interest, facilitate an approximate solution to this intergral, it would be useful if the prior distribution and likelihood function were expressed in such a form that enabled an exact solution to be determined. It is for this reason that the theory of conjugate families of distributions was developed.

Definition Let F denote the class of density functions $f(\underline{x}|\underline{\theta})$ (indexed by $\underline{\theta}$). A class P of prior distributions is said to be a conjugate family for F if $p(\underline{\theta}|\underline{x})$ is in the class P for all $f \in F$ and $p(\underline{\theta}) \in P$.

Such a class is found by examining the form of the likelihood function, (expressed in terms of the random variable $\underline{\theta}$) to find a wider family of distributions which has the likelihood function $\ell(\underline{\theta};\underline{x})$ as a member. This wider family is the family of Conjugate

Prior distributions. It is important in choosing a conjugate family that it is sufficiently `rich` so that one's prior distribution will be a member of the same family. It is often possible to enlarge further, the size of a conjugate family, by extending the domain of the parameter $\underline{r} = T(\underline{x})$, (a function of the sample $\underline{x}$), to include all values for which the kernel of $B(\underline{\theta};\underline{y})$ is non-negative.

Example

Consider the case of data generated by a Bernoulli process.

We observe x=0 with probability 1-p
and x=1 with probability p

$$f(x|p) = p^{x}.(1-p)^{1-x}$$

The likelihood function for a sample of size n is:

$$f(x_1 \ldots x_n|p) = p^{\sum x_i}.(1-p)^{n-\sum x_i}$$

$$= p^{r}.(1-p)^{n-r} \qquad \text{where } r = \Sigma x_i$$

Now in terms of the likelihood function r and n are discrete parameters. Extending the domain of these parameters to that of positive real numbers, we now see that $p^{r}.(1-p)^{n-r}$ is the kernel of the Beta density function

$$f_{\beta}(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha).\Gamma(b)} \; p^{\alpha-1}(1-p)^{\beta-1}$$

Where $\alpha = r+1$, $\beta = n-r$, and $n>r>0$.

And so since this enlarged family is closed under sampling, (ie. still conjugate in the sense of the above definition) then our conjugate prior density will be Beta with parameters α' and β' say. The Beta family is a very rich one, for by varying the parameters α'

and β! we can choose our prior distribution from a wide variety of possible shapes and scales ranging from uniform to symmetric, asymmetric, bimodal or unimodal.

The following figure (Hays and Winkler 1975), illustrates the versatility of the Beta distribution.

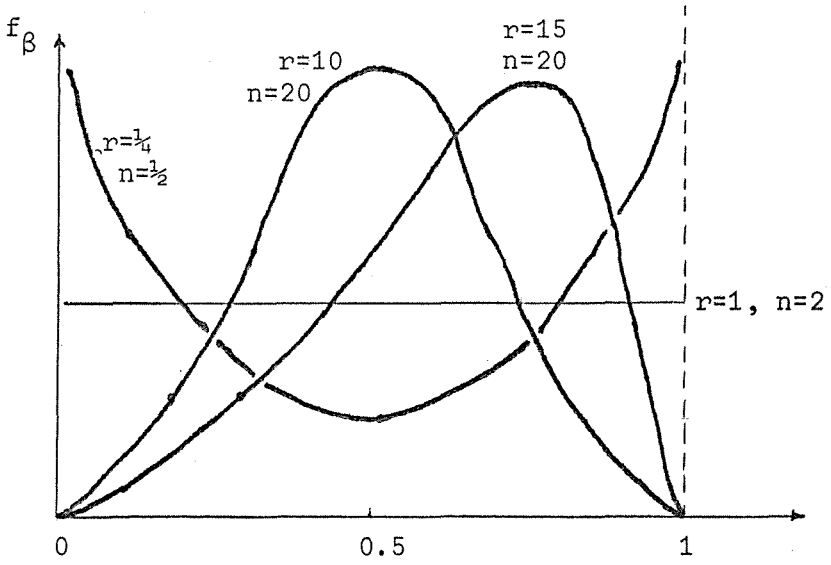


Figure 2. Some possible shapes of the Beta - Distribution.

Where $r > n/2$ results in a positively skewed curve.

$r < n/2$ results in a negatively skewed curve,

and $r = n/2$ results in a symmetric curve.

For the symmetric case $r = n/2$, the curve can have either 0, 1, or 2 modes, depending on whether $r > 1$, $r < 1$, or $r = 1$.

The hope is that at least one of the many forms of the distribution from the enormous range of those possible, will acceptably approximate our prior knowledge. Conjugate families of prior distributions have a number of desirable properties.

(1) They provide for ease of calculation of $p(\underline{\theta}|\underline{X})$, and have the appealing feature of allowing one to begin with a certain functional form, and end up with the same functional form but with the parameters updated by the sample information.

(2) They facilitate determination of expectations, since for most known families of distributions, expressions are already available for $E(\underline{\theta})$, $Var(\underline{\theta})$, etc.

(3) Because of the way parameters are updated, we have a way of seeing the relative weights of prior and sample information. It also enables us to utilise the concept of equivalent prior sample. That is the sample of given size that contains the same ammount of information as our prior knowledge.

(4) They make it unnecessary to solve for $p(\underline{x})$, since the parameters and form of the posterior distribution will be determined solely from those of the prior and the likelihood.

Without the existence of Conjugate families, the successful application of Bayesian techniques would clearly have been much more difficult, and so this concept has probably played an important role in the acceptance of Bayesian methods.

1.6 Prior Information

The desire to incorporate prior information into the inferential procedure is not unique to Bayesians. The well known Classical statisticians E S Pearson and J Neyman state in Savage (1962) ...

> `We were certainly aware that inferences must make use of prior information and that decisions must also take account of utilities, but after some considerable thought and discussion round these points we come to the conclusion, rightly or wrongly, that it was so rarely possible to give sure numerical values to these entities that our line of approach must proceed otherwise`

This inability to quantify prior opinion numerically, ultimately led Neyman and Pearson to develop a significant body of statistical methodology based on the frequentist or objective school of thought. Although this methodology, as well as that of other areas of Classical statistics do not admit any explicit use of prior information, there will always be an implicit use in terms of the significance level and and power of the statistical test used, and even the choice of test itself. In contrast to the implicit use of prior information made by Classical statisticans, Bayesians propose that prior information should be, and can be, used explicitly in the form of a prior probability distribution. The key thing about prior probabilities is that they must accurately reflect our knowledge to date! If this knowledge is based on earlier samples collected in a reasonably scientific manner, then our prior probabilities should be close to the observed relative frequency of these samples, and our

prior is said to be `data-based`. However our prior distribution must often be based on more subjective information such as `introspection, casual observation, or theoretical considerations` [Zellner 1971] and in these cases it is said to be non-data-based and appropriate methods must be used to specify this information numerically. In the case of a discrete parameter, methods for the assessment of subjective probabilities may be based on betting odds etc, as discussed in section 1.3, but in the case of assessing a prior probability distribution for a continuous parameter the problem becomes significantly more complex.

1.6.1 Subjective determination of prior density functions

The volume of available prior information, whether it be data-based or non-data-based or a combination, can range anywhere between a significant amount and none. It is useful in considering methods of determining prior distributions, to consider therefore, two possible states of knowledge, namely when the prior information ranges from vague to nonexistent, and when this information is moderate to substantial in volume. It will be useful to consider the latter case first.

1.6.2 Substantial Prior Information

When we have data-based prior information in the form of previous samples or experimental runs, we will often have a good idea of the functional form of the distribution of the parameter $\underline{\theta}$, of interest. However when our prior information is non-data-based, we may face great difficulties in the specification of the functional form of the

prior density function, because there is unfortunately no clear cut relationship between subjective judgements and mathematical functions! In any case we should like our prior distribution to represent our prior knowledge as precisely as possible while still retaining mathematical tractibility. It is this desire which provided much of the motivation for the development of the theory of conjugate families of distributions (see section 1.5). This theory is critically linked to the choice of a prior distribution, but usually is only the second step in its specification, since before any decision as to functional form can be made, we must first gain an idea of the overall shape and scale that prior knowledge suggests the distribution have. Several techniques are available for determination of this overall shape and the subsequent construction of the prior density function.

1.6.2.1 The Grouping and Smoothing (Histogram) Technique

This involves dividing the parameter space Θ into distinct intervals, and then subjectively assessing the probability of $\underline{\theta}$ lying in each interval. These intervals and their associated probabilities can then be used to construct a probability histogram to which a smooth curve can be fitted. We can then try to find a member of some conjugate family which provides a reasonable approximation to this curve. The principle difficulties with this method are that in some cases no conjugate family will exist that provides an adequate approximation to the `smoothed curve` and secondly, when Θ is unbounded our histogram will not help us to assess the probability in the `tails` of the distribution.

1.6.2.2 The Relative Likelihood Approach

This method is of most use when Θ is bounded. It involves assessing the relative likelihoods of various values of θ, perhaps using betting situations, and plotting the prior density directly from these values. As in the previous method, the approach faces great difficulty in determining the shape of the curve in the tail end areas, especially when Θ is unbounded.

1.6.2.3 Use of Location and Scale Parameters

In determining a pdf it is often useful to consider various summary measures of the distribution. Location parameters such as the mean, median and mode are useful to give an idea of the centre of the distribution. Similarly measures of dispersion would also be of great assistance. Measures such as the standard deviation although preferable for the specification of the parameters of a distribution are often difficult to assess subjectively especially if the assessor does not have enough experience with statistics to have a feel for these measures. More easily assessed measures of dispersion are the percentiles of the distribution or a selection of relevant credible intervals.Given subjective estimates of these location and dispersion measures we should be able to gain a good idea of the overall shape of the distribution. Unless the distribution is a very irregular one there will be a good chance of finding a member of some conjugate family that provides a satisfactory approximation to it.

There will inevitably be occasions when we are unable to find a suitable conjugate family with which to approximate our subjective

density function. In previous times this would probably have been a fatal blow to the use of Bayesian methodology in such situations. However now with the advent of high speed computers and numerical integration techniques this need no longer be a drawback in the determination of the posterior distribution and thus any inferences or decisions concerning $\underline{\theta}$.

It is important in our final choice of prior distribution that we know how sensitive our posterior (and hence our ensuing estimate or decision) is to variations in the prior density. Past research tends to suggest that in many cases, the estimation or decision procedure is insensitive to reasonably moderate changes in the prior distribution; however, whenever there is doubt about this, it is wise to consider a number of prior distributions before making a final choice.

1.6.3 Vague or Nonexistent Prior Information

Irrespective of the source of prior information there will be occasions when the amount of such information is very small. A case involving non-data-based prior information might be when an experimenter has only vague ideas about the value of some parameter, perhaps he expects that $\underline{\theta}$ is almost certain to be in a certain interval but he has no idea of the distribution of probability within that interval. In such a case the prior information would be overwhelmed by sample information and the statistician is said to have a diffuse or non-informative state of prior knowledge. It is important to realise that `diffuse` is a relative term. While his prior knowledge might be very little relative to a sample of size 50 say,

it might be quite informative relative to a sample of size 2!

When the prior distribution is diffuse relative to the likelihood function, the posterior distribution will depend almost solely on the sample likelihood.

ie. $p(\underline{\theta}|\underline{x}) \propto p(x|\underline{\theta})$

This is of course not strictly correct, since $p(\theta)$ will not normally be strictly uniform. However, providing $p(\theta)$ is fairly flat, with no spikes or peaks, in the region over which the likelihood is greatest and providing it does not take on large values outside that region, then the approximation will usually be adequate. Box and Tiao (1973) call such a prior locally uniform.

The problem of functionally specifying a state of prior ignorance has been tackled by many Bayesians and is a further cause for controversy within the Bayesian school. Bayes, in his 1763 paper considers the case of the parameter p of a binomial distribution and assigns a uniform distribution on (0,1) to represent his state of ignorance regarding p. Many people have subsequently taken this to suggest that Bayes intended as a general principle that when nothing is known a priori about the distibution of some parameter $\underline{\theta}$, then we should act as though all values of $\underline{\theta}$ were equally likely and thus be assigned equal probability. (ie. a uniform distribution in the continuous case) Modern commentaries on Bayes works suggest however, that he may have subsequently recanted on this point, that is if he intended it to be applied as a general principle at all!

Although Bayes "principle of insufficient reason" was, subsequent to his death,accepted by many statisticians such as Laplace, it has in

more recent times been rejected by many Bayesians on the grounds that the prior distributions which it suggests are not invariant to transformation of the parameters space. For example if the prior distribution of some continuous parameters $\underline{\theta}$ was taken as uniform, then the distribution of log $\underline{\theta}$ or some other transformation would not be uniform.

Most attempts at formulation of non-informative prior distributions, have therefore been aimed at overcoming this problem. Two main approaches stand out as being the most commonly applied. These are the "Invariance Rule" of Jeffreys (1961) and the "Data Translation Principle" of Box and Tiao (1973)

1.6.3.1 Jeffreys Invariance Rule.

Jeffreys proposed a rule for the representation of prior ignorance based on Fisher's information statistic. In the 1 dimensional case this statistic is ...

$$\Psi(\theta) = - \underset{y|\theta}{E} \left[\frac{\partial^2 \log p(y|\theta)}{\partial \theta^2} \right]$$

where $p(y|\underline{\theta})$ is the pdf of a single observation y, and the non-informative, or ignorance prior is given by ...

$$p(\underline{\theta}) \propto \Psi^{\frac{1}{2}}(\theta)$$

Example Consider the case of the binomial parameter θ. The result of a single trial is a Bernoulli random variable, taking on the values :

$$y = \begin{cases} 0 \text{ denoting a failure} \\ 1 \text{ denoting a success} \end{cases}$$

$$\text{so} \quad p(y|\theta) = \theta^{y}(1-\theta)^{1-y}$$

and $$\frac{\partial^2 \log p(y|\theta)}{\partial \theta^2} = -\frac{y}{\theta^2} - \frac{(1-y)}{(1-\theta)^2}$$

Thus $$\Psi(\theta) = -E\left[-\frac{y}{\theta^2} - \frac{(1-y)}{(1-\theta)^2}\right] = \left[\frac{E(y)}{\theta^2} + \frac{[1-E(y)]}{(1-\theta)^2}\right]$$

But $E(y) = \theta$. Therefore $$\Psi(\theta) = \frac{\theta}{\theta^2} + \frac{(1-\theta)}{(1-\theta)^2}$$

$$= \frac{1}{\theta} + \frac{1}{(1-\theta)}$$

$$= \frac{1}{\theta(1-\theta)}$$

and so ... $$p(\theta) \propto \Psi^{\frac{1}{2}}(\theta)$$

$$= \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$$

Jeffreys suggests two general principles for choosing a prior distribution, which cover the two most commonly encountered cases.

(1) If a parameter may have any value in a finite range or any value on $[-\infty, \infty]$, then its prior probability should be taken to be uniformly distributed.

(2) If the parameter may have any value on $[0, \infty]$, then the prior pdf of log θ should be taken as uniform, ie. $p(\theta) \propto 1/\theta$

When θ is unbounded above and/or below, these rules result in a prior pdf that is improper. ie. $\int p(\theta)d\theta = \infty$

The main justification for Jeffreys rule is on the basis of its invariance properties.

If $\phi = \phi(\theta)$ is a 1:1 transformation of θ then ...

$$\Psi(\phi) = -E_{y|\phi}\left[\frac{\partial^2 \log p(y|\phi)}{\partial \phi^2}\right]$$

$$= - E_{y|\theta} \left[\frac{\partial^2 \log p(y|\theta)}{\partial \theta^2} \cdot \frac{\partial \theta^2}{\partial \phi^2} \right]$$

$$= - E_{y|\theta} \left[\frac{\partial^2 \log p(y|\theta)}{\partial \theta^2} \right] \cdot \left(\frac{\partial \theta}{\partial \phi} \right)^2$$

$$= \Psi(\theta) \cdot \left(\frac{\partial \theta}{\partial \phi} \right)^2$$

and if $p(\underline{\theta})$ is a prior for $\underline{\theta}$ then $p(\phi)$ should be given by ...

$$p(\phi) = p(\theta) \cdot \left| \frac{\partial \theta}{\partial \phi} \right|$$

If $p(\theta) = \Psi^{\frac{1}{2}}(\theta)$

then $$p(\phi) = \Psi^{\frac{1}{2}}(\phi) = \left[\Psi(\theta) \cdot \left(\frac{\partial \theta}{\partial \phi} \right)^2 \right]^{\frac{1}{2}}$$

$$= \Psi^{\frac{1}{2}}(\theta) \left| \frac{\partial \theta}{\partial \phi} \right|$$

And so Jeffreys rule is indeed invariant under transformation of the parameter space.

Jeffreys rule has been extended to multiparameter cases and is given by ...

$$p(\theta) \propto \left| \Psi_n(\theta) \right|^{\frac{1}{2}}$$

where $\left| \Psi_n(\theta) \right|_{ij} = - E \left\{ \frac{\partial^2 L}{\partial \theta_i \partial \theta_j} \right\}$ is Fisher's Information Matrix.

As in the single parameter case, this rule is also based on the principle of invariance under parameter transformation.

A number of criticisms are commonly directed at Jeffreys rule. First, in the case of multiparameter problems especially, we may face great mathematical difficulties in the determination of $p(\underline{\theta})$.

Secondly, when different kinds of parameters are considered simultaneously, (eg location and scale parameters such as the mean and variance) the prior produced will very often be different from what we would expect if the parameters were considered individually.

For example in the case of independent normal mean and variance

we know ... $p(\mu, \sigma) = p(\mu) \cdot p(\sigma)$

and from the single parameter rule we have ...

$$p(\mu) \propto \text{constant} \quad \text{and} \quad p(\sigma) \propto 1/\sigma$$

and so we would expect ... $p(\mu, \sigma) \propto 1/\sigma$

However strict use of Jeffreys multi-parameter rule leads instead to ... $p(\mu, \sigma) \propto 1/\sigma^2$

This is because the rule does not allow for the prior independence of μ and σ.

The third, and some would say most serious criticism, is the use of the information function itself and specifically the sample space averaging that it incorporates, since many Bayesians believe that the likelihood function should be expressed only in terms of the actual sample observed, not in terms of any that might have been observed but were not!

A final criticism of Jeffreys rule is its reliance on the use of improper prior distributions, since their use forces us to use a degree of belief interpretation of probability. Although this is certainly not an undesirable feature in itself, it is not preferable in some cases when a frequency interpretation is reasonable and advisable.

The main method proposed as an alternative to Jeffreys invariance rule is that of Box and Tiao (1973). This method seeks a reparameterization of the problem so that different samples will affect only a translation of the likelihood function. Such likelihood functions are called data translated

1.6.3.2 Data translated prior distributions

Box and Tiao base their choice of non-informative prior distribution on the premise that if the data is to "speak for itself" then the ideal prior would be one that is rather flat compared with the likelihood function. (ie is dominated by the likelihood.) To be truly non-informative, they suggest that the prior should give us no reason for prefering one realisation of the likelihood function to another. In other words if we have little prior knowledge about $\underline{\theta}$ relative to the information supplied by the sample, then clearly we should be equally willing to accept the information from one experimental outcome as another.

The choice of non-informative prior is best understood by defining a $100(1-\alpha)\%$ likelihood interval for $\underline{\theta}$. This interval is defined in an identical manner to the $100(1-\alpha)\%$ HPD region in Section 1.4.1.2 except it has maximum density over the likelihood function rather than over the posterior pdf.

In the case of $\underline{\theta}$ being a one dimensional parameter, Box and Tiao suggest that a desirable characteristic for a non-informative prior, would be for θ to have equal prior probability, (or relative probability for an improper prior) of falling within the $100(1-\alpha)\%$

likelihood interval for every possible (sample) realisation of the likelihood function.

As an example, consider three possible realisations, L1, L2, and L3, (of an infinite number of possible ones) of a likelihood function $\ell(\theta|x)$. See figure 3.

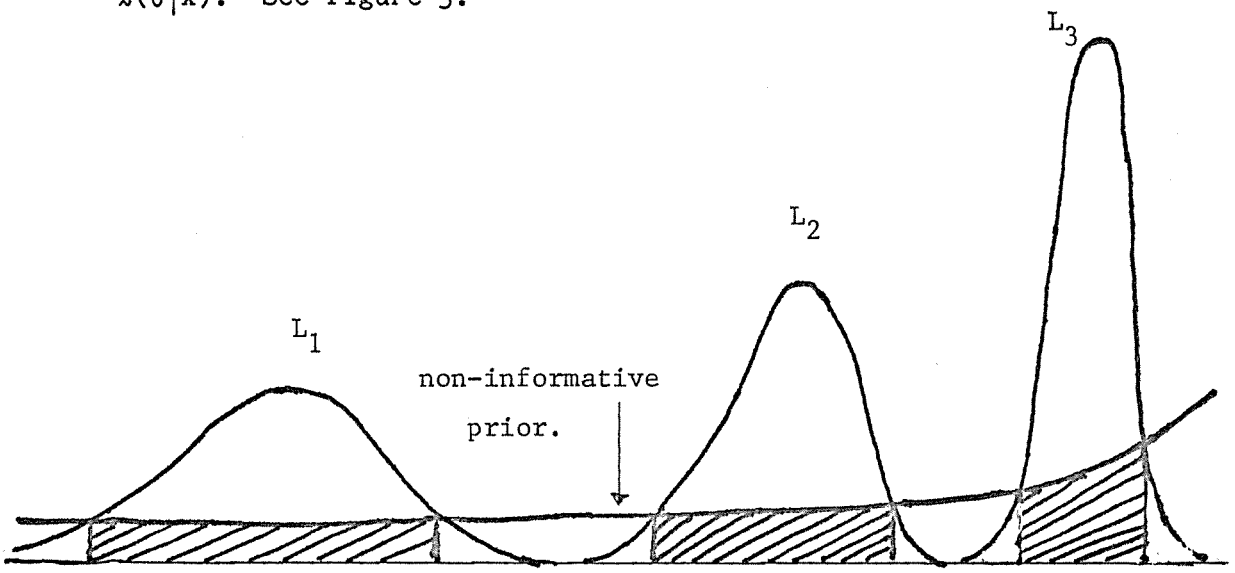


Figure 3. Central 100(1-α) Likelihood intervals -shaded areas equal.

The non-informative prior should be such that its integral over each 100(1-α)% interval,(any α) is the same for all realisations of the likelihood function. ie. The relative prior probability of θ being within each interval is the same.

Consider the simplified case where θ is a location parameter of the likelihood function and different sets of data affect only the location of the likelihood function. See figure 4. Such functions are said to be data translated. From our previous discussion we note that all 100(1-α)% likelihood intervals, (fixed α) for realisations of the likelihood function will have the same length, so that the

prior that assigns equal probabilities to all these intervals will simply be a uniform distribution.

ie. $p(\theta) \propto c$ (a constant)

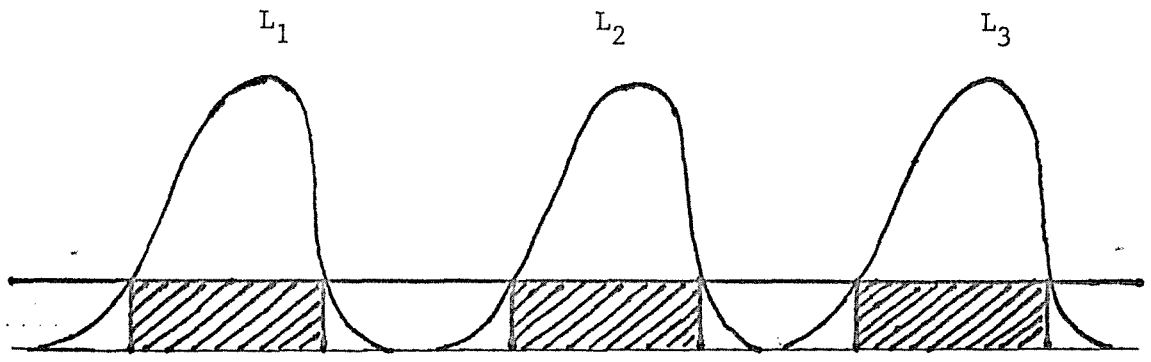


Figure 4. Central $100(1-\alpha)\%$ Likelihood Intervals for data translated parameter ϕ.

Consider again the general case of the parameter $\underline{\theta}$, whose likelihood is not data translated. Suppose it is possible to transform θ by some function $\phi(\theta)$, so that the likelihood for ϕ is now data translated. From our discussion of data translated likelihoods, we know that the non-informative prior for ϕ is simply a uniform distribution

ie. $p(\theta) \propto c$

And from our knowledge of transformations of random variables ...

$$p(\phi) = p(\phi(\theta)) \cdot \left| \frac{\partial \phi(\theta)}{\partial \theta} \right|$$

$$= \left| \frac{\partial \phi(\theta)}{\partial \theta} \right|$$

So given a transformation $\phi(\theta)$ exists which leaves the likelihood data translated, then the non-informative prior for θ is proportional to $\left| \frac{\partial \phi}{\partial \theta} \right|$

Unfortunately such transformations are not always available or apparent. However, it is often possible to find a transformation that leaves $\ell(\theta|\underline{x})$ approximately data translated. It can be shown (Appendix 1) that when the likelihood function obeys certain regularity conditions, and n is sufficiently large, then the likelihood function $\ell(\underline{\theta}|\underline{x})$ is approximately normal and distributed approximately as ...

$$\ell(\theta|x) \propto N(\hat{\theta},\ J^{-1}(\hat{\theta})/n) \qquad \ldots(1.1)$$

where $\hat{\theta}$ is the maximum likelihood estimator of θ,

and $$J(\hat{\theta}) = \left[\frac{1}{n}\ \frac{\partial^2 L}{\partial\theta^2}\right]_{\hat{\theta}}, \qquad \text{and } L(\theta|\underline{x}) = \log_e \ell(\theta|\underline{x}).$$

Consider $\phi(\theta)$, a 1:1 transformation of θ:

$$J(\phi) = \left[-\frac{1}{n}\ \frac{\partial^2 L}{\partial\phi^2}\right]_{\hat{\phi}=\phi(\hat{\theta})}$$

$$= \left[-\frac{1}{n}\ \frac{\partial^2 L}{\partial\theta^2}\ \frac{\partial\theta^2}{\partial\phi^2}\right]_{\hat{\phi}=\phi(\hat{\theta})}$$

$$= \left[-\frac{1}{n}\ \frac{\partial^2 L}{\partial\theta^2}\right]\left(\frac{\partial\theta}{\partial\phi}\right)^2_{\hat{\phi}=\phi(\hat{\theta})}$$

$$= J(\hat{\theta}) \cdot \left(\frac{\partial\theta}{\partial\phi}\right)^2$$

From Equation 1.1 we see that $\ell(\theta|\underline{x})$ will be approximately data translated if its variance is approximately constant.

ie. $J^{-1}(\hat{\phi})$ is approximately constant for given n.

and so $J^{-1}(\hat{\theta})\left(\frac{\partial\phi}{\partial\theta}\right)^2_{\hat{\theta}}$ is approximately constant.

and $\left(\frac{\partial\phi}{\partial\theta}\right)^2_{\hat{\theta}} \propto J(\hat{\theta})$

which implies $\left| \frac{\partial \phi}{\partial \theta} \right|_{\hat{\theta}} \propto J^{\frac{1}{2}}(\hat{\theta})$

Now if ϕ leaves the likelihood data translated we have already shown

that $p(\theta) \propto \left| \frac{\partial \phi}{\partial \theta} \right|$

and so the non-informative prior for θ is given by ...

$$p(\theta) \propto J^{\frac{1}{2}}(\theta) \Big|_{\theta=\hat{\theta}} = \left[-\frac{1}{n} \cdot \frac{\partial^2 \theta}{\partial \theta^2} \right]^{-\frac{1}{2}}_{\theta=\hat{\theta}}$$

This concept can be extended to multiparameter problems but as in the case of Jeffreys rule, the solution will often not be straightforward.

The technique of Box and Tiao provides a solution to the problem of which parameterization one should use when assigning a uniform prior distribution. Their method enables the determination of a "natural parameterization". This guarantees that whatever our initial parameterization, our prior will always lead to the same posterior distribution for θ. Most criticism of this method comes from the Pure Bayesian and is due to the fact that the form of the prior distribution will often depend on the form of the experiment.

eg. Fixed sample size, inverse sampling etc.

However Box and Tiao point out in refutation that, "although a non-informative prior doesn't represent an experimenter's <u>actual</u> state of mind, it should represent an <u>unprejudiced</u> state of mind" and, "knowing little can only have meaning relative to a <u>specific experiment</u>, and so for two different experiments, each of which can throw light on some parameter, the choice of non-informative prior can be different".

In many cases the exact form of non-informative prior may not be very crucial, in that for large enough sample size, providing the prior is "flat enough" the likelihood function will be dominant. To simplify analysis therefore, it is best to choose our prior to be a member of the same conjugate family as the likelihood.

The use of non-informative prior distributions has a number of distinct advantages in its favour. Firstly because of the dominance of the likelihood, any decisions or inferences will be based almost solely on the sample information expressed through the likelihood function. They will therefore be based on the same information as would be the corresponding Classical inferences. In such cases, the Bayesian method results in numerical results that are often very similar to Classical ones, however there is still the difference in interpretation pointed out in Section 1.4. That is, in one case $\underline{\theta}$ is viewed as a random variable about which probability statements can be made based on our prior knowledge of $\underline{\theta}$ and the observed sample, and in the other case $\underline{\theta}$ is regarded as a fixed quantity about which inferences are to be made based on the observed sample and its "sampling distribution".

Apart from its use in representing a state of diffuse prior knowledge, the non-informative prior offers a particularly useful tool in the reporting of scientific experiments. These are experiments carried out with no immediate action in mind. In reporting the results of a statistical analysis, the question often arises of what to report. Obviously the posterior distribution should be given because it is on this distribution that any inferences or decisions will be based. Similarly any loss function

used should also be given. It often happens that someone else, on examining the results of the study, will wish to carry out their own analysis based on their own (different) prior information, however unless the original report of the study contains the sample (or likelihood function), then such a analysis will often be impossible. The solution to this problem, is to present as part of the original report, an analysis using a diffuse prior distribution. This effectively removes prior judgement and enables anyone to apply their own prior distribution to the analysis. It also facilitates comparision with the approximate results that would be obtained using Classical techniques, and the effect of different priors on the results.

1.7 Bayesian Decision Theory

The field of Decision Theory as the name implies, is concerned with the problem of making decisions. Our whole lives consist of a sequence of decisions, from what to cook for tea, or what sort of car to purchase, to deciding how much money to invest in a particular company on the share market. Implicit in such decsions is the notion of consequence, what will happen if we make a certain decision. In the three cases mentioned above, the consequences might be; whether our guests will like the meal that we cook (eg. what proportion of people like tripe and onions!!), whether the car we decide to buy is good value for money, or whether the shares that we decide to purchase will rise or fall in price.

Another factor which we commonly utilize in decision making is data. This data could either be in the form of a sample which we might have taken to help us with our decision making, or it could be in the form of our own prior knowledge of the situation. We will very rarely make decisions in the absence of any data whatsoever.

Clearly the situations that we have been considering involve decision making under the condition of uncertainty as to which decision is truly best. In these cases statistical analysis of the problem will provide a clear aid in determining the "best decision"

The area of statistical decision theory was first developed by Abraham Wald in the 1940's. His book "Statistical Decision Functions" published in 1950, represented a landmark in its development of a theoretical basis for decision in the face of

uncertainty. Since 1950 the field has literally exploded with activity, and many statisticians such as De Groot (1970), Raiffa and Schlaifer (1961) and Ferguson (1967) have done much work in refining, generalizing and extending Wald's work. Not surprisingly the Bayesian contribution to the field has been far from small, to the extent that much of the current work in decision theory is now based on the Bayesian Standpoint.

1.7.1 The Decision Theory Model

The general decision making situation is concerned with making a decision in the face of some uncertainty. This uncertainty is generally represented by some unknown quantity $\theta \in \Theta$ which we call the state of nature. Clearly, different decisions will result from different states of nature. The set of possible decisions forms the decision space Δ. This space can either be discrete, as in the case of choosing between a number of well defined actions, or continuous, as in the case where the correct 'decision' is to obtain the best possible estimate of a parameter θ.

As we have mentioned earlier in the chapter, it is important that we evaluate the numerical consequences of each decision based, on the true state of nature. To do this we define a loss function $L(\theta,d)$ as a numerical measure of the loss incurred when decision d is chosen and θ is the true state of nature. Alternatively we can work in terms of gain functions or utility functions, (see Section 1.3.3) which measure the amount by which we benefit in choosing a particular decision. Clearly the optimal decision will be the one which minimizes this loss in some sense.

The Bayesian approach involves minimizing the expected loss relative to the prior and/or sample information. Two cases are worthy of consideration, namely where there is no observed data, ie. only prior information, and secondly where both sorts of information exist.

1.7.1.1 Prior Knowledge Only

In this situation the best we can do is to simply choose the decision for which we could expect the incurred loss to be a minimum. If our prior knowledge of θ is represented by $p(\theta)$, (see section 1.6) then we define ...

$$E[L(\theta,d)] = \int_{\Theta} L(\theta,d)\ p(\theta)\ d\theta$$

as the expected prior loss arising from making the decision d when the true state of nature is θ. (Here we assume θ is a continuous parameter. An analogous definition applies when θ is discrete.)

Example

An investor on the sharemarket learns that an oil company whose shares he owns, is about to make an important announcement regarding the success of their latest exploration venture. He wishes to decide whether or not to sell his shares for half their face value before the announcement is made or to wait for the announcement and then sell, doubling their face value if the results are favourable, or rendering them valueless if unfavourable. His previous experience of oil exploration suggests the following prior distribution for the probability of a favourable announcement

$$p(\theta) = \begin{cases} 0.3 & \text{favourable} \\ 0.7 & \text{unfavourable} \end{cases}$$

The losses that he estimates he will incur from his decision to sell now or later are represented in the following table:

		Announcement	
		Favourable	Unfavourable
Decision d_1:	Sell before Announcement	\$500	\$500
Decision d_2:	Sell after Announcement	-\$1000	\$1000

His optimal decision is the one which minimizes his expected prior loss.

For d_1 $E[L(\theta, d_1)] = \sum L(\theta, d_1)\, p(\theta)$

$= 0.3 \times 500 + 0.7 \times 500$

$= \$500$

For d_2 $E[L(\theta, d_2)] = \sum L(\theta, d_2)\, p(\theta)$

$= 0.3 \times -1000 + 0.7 \times 1000$

$= \$400$

And so his best decision given his available information is to hold onto his shares until after the announcement is made.

1.7.1.2 Prior Knowledge and Sample Data.

As well as having prior knowledge upon which to base our decisions, we will often have sample data in the form of observations of a random variable X whose probability distribution depends in a known way on the value of an unknown parameter θ. The problem of determining the best decision is now one

determining the best decision is now one of minimizing our expected loss over the observed sample and the prior distribution.

It is apparent that different data will lead us to make different decisions, and so rather than our optimal decision being fixed, as in the previous example, we now require a rule which tells us which decision to choose depending on the observed outcome.

Such a rule or decision function is denoted by $\delta(\underline{x})$, and is defined as a mapping $\delta: \chi \to \Delta$ from the sample space to the decision space. The number of possible decisions may be very large, in fact in many cases $\delta(\underline{x})$ will be a continuous function and so the number of possible decisions is in effect infinite. We therefore require a method for choosing the best of a series of decisions.

Definition

The Risk of a decision function is defined by ...

$$R(\theta,\delta) = \underset{x|\theta}{E}\left[L(\theta,\delta(x))\right]$$

$$= \int_X L(\theta,\delta(x))\, p(x|\theta)\, dx \qquad \text{if } x \text{ is continuous.}$$

$$= \sum_X L(\theta,\delta(x))\, p(x|\theta) \qquad \text{if } x \text{ is discrete.}$$

Clearly, the better a decision rule, the less its risk. We therefore define the following terms:

(1) A decision rule δ_1 is said to be as good as a decision rule δ_2, if $R(\theta,\delta_1) \leq R(\theta,\delta_2)$ for all $\theta \in \Theta$.

(2) A rule δ_1 is said to be better than a rule δ_2 if it is as good

as δ_2 and $R(\theta,\delta_1) < R(\theta,\delta_2)$ for at least one value of $\theta \in \Theta$.

(3) A rule is said to be admissible if there is no rule better and inadmissible otherwise.

The concept of one rule being better than all others can be regarded as analogous to that of the uniformly most powerful test in Classical hypothesis testing, and as in the case of the U.M.P test such rules are not common.

It is important to note that the fact that a rule is admissible does not imply that it is itself better, or even as good as, all other rules. There are usually many admissible rules for a given problem, each of which will be better than others over some particular range of θ.

1.7.2 Selecting The Best Decision Rule

There are a number of ways of choosing the optimal decision rule based on its risk and the criterion just outlined.

1.7.2.1 The Bayes Decision Rule

Define $r(\theta,\delta) = \int R(\theta,\delta)\, p(\theta)\, d\theta$

as the Bayes risk of a decision rule $\delta(x)$

The Bayes decision rule δ^0 is defined to be the decision rule that minimizes $r(\delta,\theta)$ over all decision rules.

ie. $\delta^0(x)$ is Bayes $\Leftrightarrow r(\delta^0(x),\theta) \leq r(\delta(x),\theta)$ for all $\delta(x)$.

Providing such rules are unique they will always be admissible.

We are now faced with the task of actually finding $\delta^0(x)$. Two methods are available, these are known as the normal and the extensive forms of analysis. The results produced by each are equivalent in most cases.

The Bayes decision $\delta^0(x)$ was defined as the decision function $\delta(x)$ that minimized $r(\theta,\delta)$.

i.e. $$\min_{\delta} \int_{\Theta} R(\theta,\delta)\, p(\theta)\, d\theta$$

$$= \min_{\delta} \int_{\Theta} \left[\int_{X} L(\theta,\delta(x))\, p(\underline{x}|\theta)\, dx \right] p(\theta)\, d\theta \qquad \ldots(1.2)$$

The normal form of analysis simply involves minimization of the double integral in (1.2). This form of analysis however, often faces computational difficulties, furthermore most Bayesians object to the sample space averaging which it incorporates.

An alternative form of analysis is the extensive form. This form is found by rearranging the integral in 1.2 to obtain ...

$$r(\theta,\delta) = \min_{\delta} \int_{X} \int_{\Theta} L(\theta,\delta(\underline{x}))\, p(\underline{x}|\theta)\, p(\theta)\, dx\, d\theta$$

But we already know that $p(\underline{x}|\theta)\, p(\theta) = p(\theta|\underline{x})\, p(\underline{x})$ so we can rewrite the above double integral as ...

$$\min_{\delta} \int_{\Theta} \int_{X} L(\theta,\, \delta(\underline{x}))\, p(\theta|\underline{x})\, p(\underline{x})\, dx\, d\theta$$

And if we rearrange the order of integration we have ...

$$\min_{\delta} \int_{X} \left[\int_{\Theta} L(\theta,\delta(\underline{x})) \cdot p(\theta|\underline{x}) \; d\theta \right] p(\underline{x}) \; dx$$

$$= \int_{X} \min_{a} \left[\int_{\Theta} L(\theta,a) \; p(\theta|\underline{x}) \; d\theta \right] p(\underline{x}) \; dx$$

And so the Bayes decision is simply the value $a(\underline{x})$ that minimises the inner integral, for the observed data $\underline{x}$.

i.e. $$\min_{a} \int_{\Theta} L(\theta,a) \; p(\theta|\underline{x}) \; d\theta \qquad \ldots(1.3)$$

The integral in 1.3 is known as the expected posterior loss, and its use in the extensive form of analysis clearly fits in well with the Bayesian philosophy.

Furthermore, the question of initial and final precision again arises. Berger (1980) notes ... " ... the only reasonable measure of the final precision of an action (decision) is its expected posterior loss, and while helpfull for other purposes, $r(\theta, \delta)$ is essentially a measure of initial precision and is not relevent to the choice of an action". The extensive form of analysis provides a much more straight forward way of obtaining a decision function $\delta(x)$ than does the normal form, and in fact in some cases where $r(\theta,\delta)$ is unbounded for all δ, only the extensive form can be used. Finally, the extensive form of analysis is fully consistent with the Bayesian approach to the no-data problem, one simply replaces the prior by the posterior distribution and proceeds as before.

1.7.2.2 The minimax decision rule

Amongst the proponants of decision theory the principle objections to the Bayes decision rules derived in the previous section lie in their use of a prior distribution $p(\theta)$.

The Classical approach to decision theory ignores prior information, except in the case where $p(\theta)$ has a clear frequency interpretation, and instead uses what is known as a minimax rule for choosing the optimal decision. Essentially the minimax rule involves choosing the decision rule for which the maximum possible risk is as small as possible.

Definition

$$\delta' \text{ is minimax} \iff \sup_{\theta} R(\theta,\delta') = \inf_{\delta} \sup_{\theta} R(\theta,\delta)$$

Example

Assume a decision-maker is faced with three possible decision rules and the risk $R(\theta,\delta)$ depends on the unknown value of the parameter θ, where $\theta \in \{ 1, 2, 3, 4 \}$. The risks for the respective decision rules depending on the true value of θ are given in the following table:

	State of Nature			
Decision Space	$\theta=1$	$\theta=2$	$\theta=3$	$\theta=4$
δ_1	10	4	8	3
δ_2	7	8	2	9
δ_3	5	6	5	4

$$\text{Sup } R(\theta,\delta) = \begin{cases} 10 \text{ if } \delta = \delta_1 \\ 9 \text{ if } \delta = \delta_2 \\ 6 \text{ if } \delta = \delta_3 \end{cases}$$

And $\underset{\delta}{\text{Inf}} \left[\underset{\theta}{\text{Sup}} R(\theta,\delta) \right] = 6$

So $\delta^{`} = \delta_3$ is the minimax decision rule.

The minimax rule is obviously a very pessimistic one since it sets out to optimize the worst that can happen. Like the Bayes the minimax rule it is usually, but not always, an admissible one, and in fact it can be shown that the minimax rule is, under certain conditions, also a Bayes rule. In particular it is the Bayes rule which has the highest possible Bayes risk $r(\theta,\delta)$ over any prior distribution $p(\theta)$. (In fact the prior distribution which corresponds to the maximum possible risk is known as the Least Favourable Distribution, and is heavily relied upon by Classical decision theorists.)

Unlike more standard Statistical Inference in which there is a great deal of controversy between the Bayesian and Classical approaches, there is less heated debate between the two schools, within the field of decision theory itself. This is because in most applications of decision thoery, prior information is available and quantifiable. Ironically, Classical minimax rules are often not easy to find without the use of Bayesian methods, and so even non-Bayesians must use these methods, without actually adopting a subjective Bayesian approach!

As in the case of standard statistical inference, Bayesian decision theory is often criticised on the basis of choice of prior

distribution $p(\theta)$, especially in the case where we have little prior knowledge. This question has already been dealt with fully in section 1.6 and will not be considered further here. The other criticism is if the use of the utility principle and in particular the assessment and validity of the loss or gain functions used in the analysis. This area wiil be considered in the following section.

1.7.3 The Utility Concept

Consider the case of a man offered the following gamble:

Win \$100 with probability p.

Lose \$20 with probability 1-p.

Don't play ie. win/lose nothing

The man's decision as to whether he should "play the game" will depend on a number of factors such as the value of p, whether the bet is fair, or whether it is favourable to him (ie. Is E(gain) >0 ?) It will also depend on the value of the bet to him.
If he were a millionaire he might accept the bet with almost any value of p, if however he was flat broke, he would have to think seriously about accepting any bet.

This illustrates one of the fundamental ideas of the utility concept. It provides us with a model for the way in which individuals choose between a number of different actions in any situation involving uncertainty. In order to make the best choice, utility theory attempts to provide a sensible method for assigning numerical values to the consequences of the different actions that we may choose. Clearly, such a task will often not be straight forward. In many cases there will be no clear-cut scale on which the consequences of

an action may be measured. For example the gratitude a client may feel when he receives a Christmas gift from a firm he patronizes. On the other hand such situations should certainly be amenable to consideration with regard to the value or utility of the consequences.

A firm, for example would clearly like to be able to assess the goodwill of their customers arising from Christmas giveaways, especially where this goodwill translates itself into increased sales.

The following conditions delineate the requirements for the existence of a utility function:

1. A series of actions are available, one of which must be chosen.
2. Conditional on each action is a series of one or more consequences.
3. The actual consequence arising, is a random event depending on the probabilities involved.
4. It is possible to order preferences for the various actions by assessing the relative benefit of each of the possibilities.
5. The assessor will obey the axioms of consistency and coherence in his choices of actions.

It appears from the above points that the concept of utility is a very subjective one. Although this is in fact true, there are many areas in which there are very clear-cut objective grounds for the choice of a utility fuction. For example, a bus company may wish to decide how many buses to use on a particular run each day given its knowledge of the average daily demand. In such a case the utility

function will have an obvious objective interpretation as the net gain from running a given number of buses when the average demand is θ say.

Regarding the consequencess themselves, Barnett (1973) makes the following points:

1. The consequences may not all be ameleorative.
2. They may consist of multiple componants. What is optimal for one componant, may be far from optimal for another and so there may have to be some form of trade off.
3. The preferences for the various consequences may be personal.

As in the case of the assessment of subjective probability, we are required to be coherent in our assessment and ranking of preferences. While the average man is not always rational in his assessments, it is assumed that once any inconsistency is made known to him he will modify his assessments, thus more closely approximating the ideal of the rational man.

1.7.3.1 The Assessment of Utilities

The aim of utility theory as we have mentioned, is to assign a numerical measure to the value of a consequence. Such measures are known as utilities. It has also been noted that there will often be uncertainty as to which of a series of consequences will actually occur. The actual assessment of utility functions is as complex a subject as that of the assessment of subjective probabilities and does not fall within the scope of this thesis. We will simply assume

such assessments are possible and introduce a number of utility functions that are commonly used.

1.7.3.2 Loss Functions

As defined in section 1.7.1, these are utility functions which specifically measure the loss arising from choosing a given action when a given state of nature θ exists. Many different loss functions have been developed for specific purposes, two of those used when θ is countable and a subinterval of the real line are:

Zero-One Loss

$$L(\theta_i, d) = \begin{cases} 0 & \text{when } \theta = \theta_i. \\ 1 & \text{when } \theta \neq \theta_i \end{cases}$$

Generalized Zero-One loss

$$L(\theta_i, d) \begin{cases} = 0 & \text{when } \theta = \theta_i. \\ > 0 & \text{when } \theta \neq \theta_i \end{cases}$$

Two of the most commonly used loss functions in practise however are:

Linear Loss

$$L(\theta, d) = \begin{cases} k_1(\theta - d) & \text{when } \theta \geq d \\ k_2(d - \theta) & \text{when } \theta < d \end{cases}$$

Quadratic Loss

This is seen in many modified forms, but is most simply stated as:

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

1.7.3.3 Criticisms of Utility Functions

Opponants of the decision-theoretic viewpoint criticise the assuption that individuals will always act rationally in their assessments, furthermore the use of any subjective procedure for formal decision making is rejected out of hand by "hard line objectivists". This, on purely progmatic grounds would seem to be most unreasonable as no other procedures are suggested which even attempt to provide a reasonable alternative. Furthermore, it is argued that throughout our lives, decisions are made on a subjective basis depending on our personal assessment of the consequences of various actions available to us. The theory of utility, in its attempt to model human behaviour provides the only reasonable response to the problem.

Another criticism of the use of utility concerns the difficulties inherent in the determination of utility functions. The counter criticism is that while misspecification of a loss function may lead to non-optimal decisions, such "bad" decisions are even more likely when we do not even attempt to explicitly consider the consequences of our actions.

Finally criticisms are made of the use of loss functions in problems of scientific inferences, such as estimation and hypothesis testing, where it is argued "personal prejudices should intrude as little as possible". The decision theoretic approach to such problems will briefly be considered in the next section.

1.7.4 The Decision-Theoretic approach to Statistical Inference.

1.7.4.1 Estimation

An estimation problem in decision theory is defined as one in which the decison function $\delta(\underline{x})$ is an estimate of the parameter (or state of nature) θ, conditional on the particular data $\underline{x}$ observed, and the decision space Δ. (Which will correspond to the parameter space Θ.)

In this situation the loss function measures the discrepency between the true, but unknown, value of θ, and the estimated value $\delta(\underline{x})$.
One of the loss structures most commonly associated with estimation problems is the quadratic, or squared error loss.

$$L(\theta, \delta(\underline{x})) = (\theta - \delta(\underline{x}))^2$$

Using this loss function, it can easily be shown that the optimal or "Bayes" decision, (ie. that which minimises the expected posterior loss), is simply the mean of the posterior distribution and the Bayes risk is the expected posterior variance with respect to the marginal distribution of $\underline{x}$.

ie. $\delta^{\circ}(\underline{x}) = E[\theta|\underline{x}]$

and $r(\theta,\delta(\underline{x})) = E[Var(\theta|\underline{x})]$ wrt to the marginal pdf of X.

The Bayesians use of this estimator is analogous to the use of the mean as the minimum variance estimator in classical theory, except, that in one case inferences are based on both prior and sample information and in the other case, only the sample is used.

1.7.4.2 Hypothesis Testing

In this case the Bayesian decision theoretic approach is much more clear cut.

Consider testing H_1: $\theta \in C$ vs H_2: $\theta \notin C$

where C is a subset of Θ, and so possible decisions are to choose either H_1 or H_2 as being the correct hypothesis. The Bayes decision in this case, simply corresponds to choosing the hypothesis for which the expected posterior loss is smaller. This procedure extends easily to the case of multiple hypotheses.

1.8 Further topics in Bayesian Statistics

There are a number of other topics that are extremely important to Bayesian Analysis. Concepts such as Exchangeability, and the Likelihood principle play a very important role in the development of Bayesian theory. There are also so-called fringe areas of Bayesian research such as Empirical Bayes methods which are worthy of consideration in order that we obtain a broad overview of Bayesian Statistics. Unfortunately each of these areas on their own would be substantial enough to write a thesis on! In this section, therefore, we briefly consider three important concepts which we feel are the most worthy of consideration.

1.8.1 Exchangeability

The concept of Exchangeability is a central one to the theory of Bayesian Statistics. It was developed by de Finetti in 1931 and provides a thoeretical justification for both the concept of differing subjective probabilities for the same event, and for the development of Bayesian inference.

Definition.

Consider an infinite sequence of events E_1, E_2, ... Such a set is said to be exchangeable if, for any subset of n distinct events, the probability that a specified r of them will occur, and the remaining n-r will not occur, depends solely on n and r, and not on the subset selected.

e.g. $P(E_1, E_2, E_3$ occur and E_4, E_5, E_6 don't occur)

$= P(E_8, E_{11}, E_{23}$ occur and $E_{87}, E_{900}, E_{1000}$ don't occur)

$= P(E_i, E_j, E_k$ occur and E_x, E_y, E_z don't occur)

for all distinct i, j, k, x, y, z.

According to Lindley (1971), "an exchangeable sequence of random variables (or events) acts like a random sample from some distribution". Furthermore " ... exchangeability can be a substitute for the randomness concept basic to the frequency school". This point is echoed by Barnett (1975), when he notes ... " Thus in a sense exchangeability plays a role in the subjective approach akin to von Mises' principle of randomness in the freguency approach".

The concept of exchangeability is a weaker one than that of independence. This point is best understood by considering the outcome of a simple experiment such as the roll of a die. Suppose we roll a die repeatedly, are the resulting events independent?

Lindley points out that if the die is known to be "fair", then the outcomes will be independent, but if the die is, contrary to our belief not fair, then the outcomes will not be independent, however in either event they will be exchangeable!

1.8.2 The likelihood principle.

The role of the likelihood function in Bayesian Statistics is, as previously stated, quite different to that in Classical Statistics. In Classical Statistics, it serves as a tool for the construction of particular inferential methods, whereas it forms the mainstay of the entire Bayesian approach, in so far as it is the only way through the sample can express itself. It is the likelihood principle that provides the means for the expression of this function.

Berger expresses this principle in the following way ...

> "In making inferences of decisions about θ after $\underline{x}$ is observed, all relevent sample information is contained in the likelihood function. Two likelihood functions are equivalent, (i.e. the corresponding samples contain the same information about θ), if they are proportional to each other for given $\underline{x}$.

One of the consequences of this principle that is directly at odds with Classical statistics, is what is known as the "irrelevence of the sampling rule".

Consider a sequence of independent Bernoulli trials indexed by the parameter θ. Asume that three successes were observed in five trials. The likelihood function therefore is proportional to $\theta^3(1 - \theta)^2$ and to the Bayesian is the only possible expression of the observed data.

Classical statistics however, requires knowledge of how the sample was observed. For example if 5 trials were made in which 3 successes

were observed, of if sampling continued until 3 successes were observed, in this case in 5 trials.

In the former case the likelihood function is ...

$$\binom{5}{3} \theta^3(1-\theta)^2 \qquad \text{i.e. Binomial.}$$

And in the latter case ...

$$\binom{4}{2} \theta^3(1-\theta)^2 \qquad \text{i.e. Negative Binomial.}$$

And so to the Classical Statistician, his inferences based on these functions may be different. (e.g. Different Confidence Intervals) Whereas, since the two functions are proportional, the Bayesian approach, as we have said, will lead to identical inferences. This reflects a basic difference in the two philosophies, on the one hand the likelihood principle rules supreme, and on the other, the method of sampling is held to be equally important.

1.8.3 Empirical Bayes Methods

In Section 1.6 we considered the problem of assigning a prior distribution to $\underline{\theta}$, particularly when our prior knowledge is non data-based (NDB) or vague.

There will often be cases however, when we do have data-based prior information which allows us to be fairly specific in our choice of prior distribution. Two cases are worthy of note.

Consider the situation where previous experimental runs of a similar kind have yielded the values $\theta_1, \theta_2, \ldots \theta_k$.

For example θ_i may denote the mean value of some characteristic, for

a given batch, the i'th, of a componant produced in some manufacturing process. In this case θ can truly be regarded as a random variable and can be used to formulate a prior distribution $p(\theta)$, based on the methods of section 1.6.2.

Unfortunately, although θ can often be interpreted as a random variable, it is rarely the case that the actual values of the θ_i's themselves are known. More commonly our information will only be in the form of limited samples of data $\underline{X}_i$, which arise from the process with parameter θ_i

Consider the previous example, where instead of the batch being examined and θ_i determined, only a sample was taken from each batch, and so the true value of each θ_i is unknown. This problem may arise when the measurement of some characteristic of a componant results in its ultimate destruction (eg. lifetimes of light bulbs) In such cases it is clearly only feasible and sensible to take a limited sample from each batch.

Techniques for the estimation of prior distibutions in such cases fall into the broad category of Empirical Bayes methods. The random variables $\theta_1 \ldots \theta_k$ arising from, for example, different experimental runs, clearly represent an independant identically distibuted set of observations from the unknown distribution $p(\theta)$.
Conversely $X_1 \ldots X_k$ cannot be regarded as independent identically distributed random variables with respect to the conditional density $p(\underline{X}_i|\theta_i)$, since the value of θ_i, on which the density is conditioned, will vary. However $\underline{X}_1 \ldots \underline{X}_k$ will be i.i.d with respect to the unconditional (marginal) density ...

$$p(\underline{x}_i) = \int_{\Theta} p(x_i|\theta_i)\, p(\theta)\, d\theta$$

Empirical Bayes methods are aimed at using the values of $\underline{x}_1 \ldots \underline{x}_k$, (each corresponding to a particular value of θ from the unknown prior density $p(\theta)$) to make inferences about the present value of θ, θ_{k+1}, say, corresponding to the present sample $\underline{x}_{k+1}$.

Example

Consider the case where x is an observation from a Binomal distribution. ie. $f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$ $x=0, \ldots n$ where n is fixed

Assume we wish to estimate the mean value of θ (perhaps under the assumption of quadratic loss as in section 1.7) where the prior distribution of θ, $g(\theta)$, is unknown.

Then $$E[\theta|x] = \frac{\int \theta\, f(x|\theta)\, g(\theta)\, d\theta}{f_g(x)}$$

$$= \frac{\int \theta \binom{n}{x} \theta^x (1-\theta)^{n-x} g(\theta)\, d\theta}{f_g(x)}$$

$$= \frac{\left(\frac{x+1}{n+1}\right) \int_{\Theta} \binom{n+1}{x+1} \theta^{x+1} (1-\theta)^{(n+1)-(x+1)} g(\theta)\, d\theta}{f_g(x, n)}$$

$$= \left(\frac{x+1}{n+1}\right) \frac{f_g(x+1, n+1)}{f_g(x, n)}$$

Where $f_g(x) = \int_{\Theta} f(x|\theta)\, g(\theta)\, d\theta$ is the marginal distribution of x, found by integrating out θ.

We now assume that previous data in the form of k earlier experiments, each of n trials, have enabled us to build up an empirical distribution of the number of successes in n trials. Assume the fre-

quency with which 'i successes in n trials' occurs, is given by $f_k(i)$ and so the Classical estimate of $f_g(x, n)$ is ...

$$\hat{f}_g(x, n) = \frac{f_k(i)}{k}$$

If we now include the current observation x, then our estimate becomes ...

$$\hat{\hat{f}}_g(x,n) = \begin{cases} \dfrac{f_k(i)}{k+1} & \text{If } i \neq x \\ \dfrac{f_k(i)+1}{k+1} & \text{If } i = x \end{cases}$$

And so $E[\theta|x]$ is estimated by ...

$$\frac{(x+1)}{(n+1)} \frac{[f_k(x+1)]}{f_k(x) + 1}$$

These methods were pioneered by Robbins (1964), Maritz (1970), and others, and are some what controversial due to their reliance on estimation of the marginal distribution $p(\underline{x}_1, \ldots, \underline{x}_k)$.

For this reason they have been criticised by many Bayesian purists as being 'seldom Bayesian in principle', and Lindley (1971), goes so far as to say that " ... hardly any Empirical Bayes procedures are Bayesian", and that the reference to Bayes in the title is almost solely due to the assumption that the θ_i's represent a random sample from some unknown distribution $p(\theta)$. Despite this criticism, we feel that on solely pragmatic grounds, we are likely to hear a lot more of this approach.

CHAPTER TWO

Introduction to the Applications of Bayesian Statistics.

Over the last 20 years, the scientific and research community has seen a tremendous expansion in the development of the theory and methodology of Bayesian statistics. The expanding volume of material dealing with aspects of the Bayesian viewpoint has led Professor A Houle of Universite' Laval in Canada to assemble an extensive Bibliography on the subject. In a recent paper (1983) relating to this Bibliography Professor Houle gives an interesting illustration of the growth of Bayesian Statistics.

> "Prior to 1935, we trace two books, 19 articles in periodicals and eight communications to the learned societies. From 1935 to 1950, 40 books, 50 articles in periodicals and eight communications to the learned societies were added. From 1975 to 1979, 207 books, 1600 articles in periodicals, 437 communications, 350 publications in research centres of universities and 260 doctoral dissertations further explore the field of Bayesian statistics."

Not only has the amount of material relating to Bayesian statistics skyrocketed in recent years, but also the range of subjects at which this research has been directed. Apart from a wealth of research into the theoretical development of inferential and decision making procedures, the Bayesian approach has been applied to a vast range of practical problems in subjects as diverse as medicine, education,

economics, meteorology, chemistry.... The list goes on and on! There have also been numerous conferences and seminars specialising in one or more aspects of Bayesian statistics, the most recent being the 1982 Institute of Statisticians conference on Practical Bayesian Statistics. The increasing volume of papers published, and conferences sponsored by learned societies, such as the I.O.S can only serve to further "spread the good news" of the Bayesian paradigm.

At present many of the practicioners of Bayesian methods are (non statisticians) # from disciplines which traditionally have relied heavily on Classical or "Berkeley" statistics, ## as Dennis Lindley so charmingly calls non-Bayesian statistics. Among statisticians themselves, however, there are many who, raised in the Classical school have hardly considered the use of Bayesian methods, let alone made an honest assessment of the approach based on its relative merits and disadvantages. The author believes that this must and will change.

Note here the distinction between the practicioner or applicator of Bayesian techniques and the statistician himself who is schooled not only in their application but in their theoretical basis.

He gives two reasons for the name. Firstly the statistics Department of the University of California at Berkeley is one of the best in the world of the non-Bayesian type. Secondly Bishop Berkeley was much criticised by Bayes for his views on Newton.

As the use of Bayesian techniques become more and more popular we would expect statisticians, especially those less entrenched in classical approach, to gradually appreciate the utility of Bayesian methods. This in no way suggests the imminent demise of Berkeley statistics. These techniques have been around for such a long time, and have been utilized in so many different disciplines, that it will take some time, and in some cases much more developmental work by Bayesian researchers, to produce techniques sufficiently attractive to replace Classical methods.

In this chapter we will be considering a number of disciplines in which Bayesian data analysis is proving itself useful, as well as investigating its continuing theoretical development. In each discipline we will give a resume' of the areas in which Bayesian techniques are exploited as well as giving a particular problem where the Bayesian approach has been found useful.

2.1 Econometrics and Business Studies

Before examining the role of Bayesian statistics in the field of econometrics,it is important that we first define what econometrics is.

Zellner (1981) defines it as ..."the field of study in which economic theory, statistical methods, and economic data, are combined in the investigation of economic problems."

He goes on to mention some of the problems commonly encountered such as measurement and forecasting problems, formulation and testing of

economic theories, decision and control problems, design and execution of surveys of humans and/or animals to shed light on scientific and practical policy problems. These problems involve analysis of many types of statistical models, from multivariate regression and time series models to simultaneous equation and Markov chain models. Many of the techniques used to solve such problems will clearly be equally applicable in other disciplines.

Bayesian methods have been attractive to econometricians for a number of reasons. The first relates to the unified nature of the Bayesian approach. Whatever the model to be analysed, the posterior pdf for the model parameters, (which essentially tells us everything we know about them) is always proportional to the product of the sample likelihood and the prior pdf. In comparision we find that there is no "standard" non-Bayesian approach and that different methods are designed to handle specific sets of problems. The second reason for the appeal of Bayesian techniques lies in the limited amount of data available in many econometric problems.

Zellner (1981) uses the example of a multivariate time series (autoregression) model of the form:

$$\underline{y}_t = A_1\underline{y}_{t-1} + A_2\underline{y}_{t-2} + \ldots + A_q\underline{y}_{t-q} + \underline{e}$$

where $\underline{y}_i$ is the observation made in the ith time period on the

p dimensional random variable $\underline{y} = (y_1 \ldots . y_p)$,

$A_j : j = 1 \ldots q$ are pxp matrices of coefficients

and $\underline{e}$ is the error term, distributed as a $N(\underline{0}, I\sigma^2)$ random variable. The system contains $qp + p(p + 1)/2$ parameters to be estimated. (The q matrices each with p elements plus the $p(p + 1)/2$ unique elements of the symmetric variance-covariance matrix.)

Consider the case of a "small" system where p=6 and q=10, there are 381 parameters to be estimated. If 20 years of quarterly data was available we would have pT = 6 x 4 x 20 = 480 observations and so the observation/parameter ratio would be very low ie. 480/381=1.3
In such cases, to get any sort of reasonable results it is crucial that we utilise all available prior information.

A further advantage of the Bayesian approach is in its decision making role when combined with a loss function. As mentioned in chapter One, this is particularly useful in business orientated analyses, where there is a financial consequence of making an incorrect decision

In the past, much of the work in Bayesian econometrics has been directed at the the development of general methodology rather than specific applications. However in recent years there has been a move towards the application of techniques. Morales (1970) used a Bayesian analysis of the simultaneous equation model to predict consumption in the Belgium beef market; Peck (1974) has used Bayesian estimation techniques to analyse the investment behaviour of firms in the electrical utility industry; Smith (1983) used Bayesian forecasting techniques to predict accident claims for an assurance company; Siczewicz (1981) has used a Bayesian method for grouping Massachusetts towns into territories for auto insurance purposes; and Varian (1974) has used a Bayesian decision theoretic approach for the assessment of property values for taxation (rating) purposes.

2.1.1 A Bayesian Approach to Real Estate Assessment

Varian (1974) uses a Bayesian approach for the estimation of market values of properties in the district of San Mateo, California. Annual valuations carried out by the assessors are determined by applying a regression equation whose coefficients are estimated from previous sales in the area. The reason for a Bayesian approach to the problem is that there may be significant financial losses to either the home owner or the local authority, if the estimated value of the property Ye differs from the actual value Ya. In Varian's study, he considers only the loss to the assessor's office. If the assessor underestimates the value of the property, then the loss to the assessor's office is the tax from under-assessment. If however, the property owner is over assessed then he has two options. he may either attempt to settle the matter "out of court" by conferring with the assessor's office and attempting to convince them of their error, or he may present an appeal to a statutory body which weighs all available evidence to produce a fair assessment. In either case if the property was over assessed, the assessor's office not only loses out on the tax but must pay the costs of the appellant. If the assessor's office wins the appeal it must still pay the court costs. The assessor's office would therefore wish to choose an estimate for home value that would minimise the total expected loss.

Varian reasons that because over-estimates tend to be more expensive than under-estimates then his loss function should be asymmetric. Various other factors lead him to choose the loss function ...

$$L(Ye,Ya) = b \exp [a(Ye - Ya)] - c(Ye - Ya) - b$$

which he calls a linex (linear exponential) loss function. For the purposes of this study, Varian uses a=0.0004, b=2500 and c=1, as estimates of the parameters of the loss function. The task of more accurate estimation is left to a later date.

Varian assumes that the selling price of a home is a linear function of the characteristics of the house and section plus an error term. Previous research has isolated nine variables as being good predictors of selling price.

X1 :	lot width	X2 :	view 1 = average or above 0 = below average
X3 :	age	X4 :	number of bedrooms
X5 :	quality class	X6 :	total living area
X7 :	fence cost	X8 :	flatwork cost (patios, driveways, etc)
X9 :	a constant term		

Given these variables the linear regression model can be stated as

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

where $\underline{y}$ is a m x 1 vector of selling prices

X is a m x k matrix of observed characteristics

$\underline{\beta}$ is a k x 1 vector of parameters

and $\underline{\varepsilon}$ is a m x 1 error term distributed as $N_m(\underline{0}, I\sigma^2)$.

Clearly for given X, $\underline{\beta}$ and σ, $\underline{y} \propto N_m(X\underline{\beta}, I\sigma^2)$.

i.e. $$p(\underline{y}|X, \underline{\beta}, \sigma) = [(2\pi)^k|\Sigma|]^{-1/2} \exp\{-\frac{1}{2}(\underline{y} - X\underline{\beta})' \Sigma^{-1}(\underline{y}-X\underline{\beta}) \}$$

where $$\Sigma = \sigma^2 I \text{ and } |\Sigma| = \sigma^{2m}$$

$\therefore$ $$p(\underline{y}|X,\underline{\beta},\sigma) \propto \sigma^{-m} \exp\{ -\frac{1}{2\sigma^2} (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta})\}$$

which reduces to $\sigma^{-m} \exp\{ -\frac{1}{2\sigma^2} [\nu s^2 + (\underline{\beta} - \underline{\hat{\beta}})' X'X(\beta - \hat{\beta})]\}$

where $\nu = m - k$

$\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y}$

and $s^2 = \frac{(\underline{y} - X\underline{\hat{\beta}})'(\underline{y} - X\hat{\beta})}{\nu}$ See Appendix Two for proof.

Bayes theorem now enables us to determine the posterior pdf of $\underline{\beta}$ and σ, given the data and prior knowledge.

ie. $p(\underline{\beta}, \sigma|\underline{y}, X) \propto p(\underline{y}|X, \underline{\beta}, \sigma) \cdot p(\underline{\beta}, \sigma)$ ... (2.1)

where $p(\underline{\beta}, \sigma)$ is the joint prior density of $\underline{\beta}$ and σ.

Varian uses this to construct a predictive pdf for the value Y_a, of a new property with characteristics $\underline{X}_a$. The predictive pdf of y_a given $\underline{X}_a$ and the previous sample data X and $\underline{y}$ is denoted by ...

$$p(y_a|\underline{X}_a, \underline{y}, X) = \int_{\sigma}\int_{\beta} p(y_a, \underline{\beta}, \sigma |\underline{X}_a, \underline{y}, X)\, d\underline{\beta}\, d\sigma$$

and $p(y_a, \underline{\beta}, \sigma |\underline{X}_a, \underline{y}, X) = p(y_a|\underline{\beta}, \underline{X}_a, \sigma) \cdot p(\underline{\beta}, \sigma |\underline{y}, X)$

Where y_a is a $N(\underline{X}_a\underline{\beta}, \sigma^2)$ random variable, for given $\underline{X}_a$, $\underline{\beta}$ and σ. Evaluation of $p(y_a|\underline{X}_a, \underline{y}, X)$ requires knowledge of $p(\underline{\beta}, \sigma |\underline{y}, X)$ which itself requires us to know $p(\underline{\beta}, \sigma)$.

Varian determines the predictive pdf for three different priors, a diffuse prior, a data-based prior and a non-data based (non diffuse) prior.

2.1.2 Derivation of Predictive distribution using a Diffuse prior.

Consider the diffuse prior $p(\underline{\beta},\sigma) = \frac{1}{\sigma}$

This gives $p(\underline{\beta},\sigma|\underline{y},X) \propto \sigma^{-m+1} \exp\{-\frac{1}{2\sigma^2} [\nu s^2 + (\hat{\beta} - \beta)' X' X(\hat{\beta} - \beta)]\}$ ...(2.2)

or equivalently $\sigma^{-(m+1)} \exp\{-\frac{1}{2\sigma^2} (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta})\}$

and so $p(y_a, \beta, \sigma | \underline{X}_a, y, X) = p(y_a | \underline{\beta}, \sigma, \underline{X}_a) \cdot p(\underline{\beta}\ \sigma | \underline{y}.X)$

$$\propto \sigma^{-1} \exp\{-\frac{1}{2\sigma^2}(y_a - \underline{X}_a\beta)^2\}$$

$$\times \sigma^{-(m+1)} \exp\{-\frac{1}{2\sigma^2}(\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta})\}$$

and $$p(y_a, \underline{\beta} | \underline{X}_a, \underline{y}, X) = \int_0^{\infty} p(y_a, \underline{\beta}, \sigma | \underline{X}_a, \underline{y}, X) d\sigma$$

$$\propto \int \frac{1}{\sigma^{m+2}} \exp\{-\frac{1}{2\sigma^2}[(y_a - \underline{X}_a\underline{\beta})^2 + (\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta})]\} d\sigma$$

$$\propto [(y_a - \underline{X}_a\underline{\beta})^2 + (\underline{y}-X\underline{B})'(\underline{y}-X\underline{B})]^{\frac{-m+1}{2}}$$

Integrating again w.r.t $\underline{\beta}$

$$p(y_a | \underline{X}_a, \underline{y}, X) = \int p(y_a, B | \underline{X}_a, \underline{y}, X) d\underline{\beta}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} p(y_a, \beta | \underline{X}_a, \underline{y}, X) d\beta_1, d\beta_2 .. d\beta_k$$

which it can be shown [Zellner (1971)] is proportional to

$$[\nu + (y_a - \underline{X}_a\hat{\underline{\beta}})\ H(y_a - \underline{X}_a\hat{\underline{\beta}})]^{-\frac{\nu+1}{2}}$$

where $$H = \frac{1}{s^2}(1 - \underline{X}_a M^{-1} \underline{X}_a')$$

and $$M = X'X + \underline{X}_a'\underline{X}_a$$

and so y_a has a multivariate t distribution with mean $\underline{X}_a\hat{\beta}$ and variance $H^{-1}(\frac{\nu}{\nu-2}) = \frac{\nu}{\nu-2}\ s^2(1 + \underline{X}_a(X'X)^{-1}\underline{X}_a')$ where $\nu = m - k$, and s^2 is the ordinary least squares estimate of σ^2 and $\hat{\underline{\beta}}$ is the (OLSQ) estimate of $\underline{\beta}$.

Varian notes that for a large enough sample of initial sales the t-distribution can be approximated by a normal one.

2.1.3 Predictive distribution for a non diffuse prior

Varian chooses a Normal-Inverted Gamma prior density for $\underline{\beta}$ and σ

$$p(\underline{\beta}|\sigma) \propto \frac{1}{\sigma^k} \exp[-\frac{1}{2\sigma^2}(\underline{\beta}-\bar{\underline{\beta}})'A(\underline{\beta}-\bar{\underline{\beta}})]$$

and $$p(\sigma) \propto \frac{1}{\sigma^{\nu_0+1}} \exp[\frac{-\nu_0 c_0^2}{2\sigma^2}] \qquad \nu_0 \geq 0.$$

i.e. $$\underline{\beta} \text{ given } \sigma \sim N(\bar{\underline{\beta}}, \sigma^2 A^{-1})$$

and $$\sigma \sim \text{Inverted Gamma } (\nu_0, c_0^2)$$

and $$p(\underline{\beta},\sigma) = p(\beta|\sigma)\, p(\sigma)$$

Now $$p(\underline{\beta},\sigma|\underline{y}\, X) \propto p(y|X,\underline{\beta},\sigma)\, p(\underline{\beta},\sigma)$$

$$\propto \frac{1}{\sigma^m} \exp\{-\frac{1}{2\sigma^2}(\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta})\}\, p(\underline{\beta},\sigma)$$

$$\propto \frac{1}{\sigma^{m+k+\nu_0+1}} \exp\{-\frac{1}{2\sigma^2}[\nu_0 c_0^2 + (\underline{\beta}-\bar{\underline{\beta}})'A(\underline{\beta}-\bar{\underline{\beta}}) + (\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta})\}$$

$$\propto \frac{1}{\sigma^{m'+k+1}} \exp\{-\frac{1}{2\sigma^2}[m'c^2 + (\underline{\beta}-\breve{\underline{\beta}})'(A+X'X)(\underline{\beta}-\breve{\underline{\beta}})] \quad \ldots (2.3)$$

where $m' = m + \nu_0$, $m'c^2 = \nu_0 c_0^2 + \underline{y}'\underline{y} + \bar{\underline{\beta}}' A \bar{\underline{\beta}} - \breve{\underline{\beta}}'(A+X'X)\breve{\underline{\beta}}$ and $\breve{\underline{\beta}} = (A+X'X)^{-1}(A\bar{\underline{\beta}} + X'\underline{y})$ which is the same form as equation 2.2.

And so the predictive pdf is the same form but with the parameters given in 2.3, i.e. mean $\underline{X}_a\breve{\underline{\beta}}$ and variance $\frac{m'c^2}{\nu-2}(1+\underline{X}_a(A+X'X)^{-1}\underline{X}_a')$ which will be normal for large enough m'.

2.1.4 Loss minimising estimator

Varian's next step, having found the predictive pdf of y_a is to find the estimator y_e of y_a which minimises the expected posterior linex loss.

The expected posterior loss is given by ...

$$E[L(y_e,y_a)] = \int_{-\infty}^{\infty} L(y_e,y_a)\ p(y_a|X,\ \underline{y},\ \underline{X}_a)\ dy_a$$

which it can be shown is given by ...

$$E[L(y_e,y_a)] = b\ \exp[a(y_e - u + \frac{va}{2})] - c(y_e - u)$$

where u and v are respectively the mean and variance of y_e and depend form of the prior. See equations 2.2 and 2.3.

To find the value of y_e which minimises this, we differentiate with respect to y_e and equate to zero.

$$ab\ \exp[ay_e + a(\frac{va}{2} - u)] - c = 0$$

$$ay_e + a(\frac{va}{2} - u) = \log_e(\frac{c}{ab})$$

and so $$y_e = \frac{1}{a}[\log_e(\frac{c}{ab}) - (\frac{va}{2} - u)]$$

$$= \frac{1}{a}\log_e(\frac{c}{ab}) - \frac{va}{2} + u$$

When the values of a, b and c given before are used the estimator reduces to ... $y_e = u - 0.0002\ v$

If the "correction" term, 0.0002 v is omitted, Varian points out that the estimate of y_e, the posterior mean, minimises quadratic loss instead of linex loss

2.1.5 The Data

Varian's data consisted of a sample of 168 observed sales of single family homes in the year 1965. He splits this into two groups, the first consisting of 125 observations was used for estimation, and the second of 43 observations was used for projection and comparision purposes. Using the first group he carried out a standard Bayesian regression, using both diffuse and non diffuse priors. The prior and posterior estimates of the elements of $\underline{\beta}$ are given in table 1.

Varian notes that the OLSQ coefficients (diffuse prior) all appear to be reasonable except for that for variable X4, the number of bedrooms, which for some reason is negative. The reasons he gives for this are: incorrectly specified model, missing variables or multicollinearity in the variables.
In his non data based prior Varian chooses a positive coefficient for the number of bedrooms, a prior variance of c = 2460, and uses for his SSCP matrix A = X'X. (The SSCP matrix from the data)

For this data-based prior, a OLSQ regression, (or Bayesian regressian with a diffuse prior) was carried out on the second group of 43 observations, and the resulting estimates of $\underline{\beta}$ and σ were used to form the prior and corresponding posterior estimates of $\underline{\beta}$. In comparison with the OLSQ estimates, Table 1 shows an increase of only 3.4% in the standard error of regression using the data based prior, compared with a 118% increase using the non-data based prior. Furthermore, inspection of the coefficients reveal a strong similarity between the data based Bayesian results and the OLSQ results, except for the value of the constant.

	Coefficient				
Variable.	OLSQ.	Non-data-based prior.	NDB Bayes-ian regr.	Data-based prior.	Data-based Bayesian regression
Lot width	33.11	50.00	41.55	98.08	53.88
View	1879.27	2000.00	1939.64	2405.24	2109.04
Effective years	-175.34	-200.00	-187.67	-139.22	-173.50
Number of Bedrooms	-722.48	1000.00	138.76	-12.04	-608.90
Quality Class	2000.54	2000.00	2002.73	2454.36	2119.04
Total living area	7.95	10.00	8.98	6.12	7.33
Fence cost	3.89	2.00	2.94	-0.31	3.51
Flatwork cost	5.05	5.00	5.03	4.52	4.48
Constant	1356.11	1000.00	1178.05	-4413.44	161.53
Standard Error	1827.16	——	3993.47	2081.64	889.34

TABLE 1

TABLE 1

Estimation Results for Varians 'Real-Estate Data'.

2.1.6 Comparative results and conclusions

The second group of 43 houses was utilised for prediction purposes. The average linex loss incurred was determined for valuation estimates using diffuse, data based, and non-data-based prior information, combined with both quadratic and linex loss functions. This information appears in Table 2.

The data based linex estimator is clearly superior in terms of its smaller average linex loss, with the next best being the diffuse linex estimator. The OLSQ estimator (given by the posterior mean $y_a = \underline{X}_a \underline{\beta}$ under quadratic loss with a diffuse prior) clearly performs quite poorly relative to the two previous estimates.

Prior	Correction	Average linex loss	Standard deviation
(1) Data Based	Linex	718.0	1367.0
(2) Diffuse (OLSQ)	Linex	790.0	1578.0
(3) Data Based	Quadratic	901.0	1922.0
(4) Diffuse (OLSQ)	Quadratic	999.0	2161.0
(5) Non Data Based	Linex	1278.0	2822.0
(6) Non Data Based	Quadratic	9351.0	14965.0

TABLE 2

Loss minimization Results for Varians Real-Estate Data

In particular the average linex loss from data-based Bayesian regression is only 72% of the loss resulting from OLSQ regression. With losses often measured in thousands of dollars, the use of data

based linex estimation could result in quite significant savings.

Varian proposes a follow up study aimed at more accurate determination of the form of the loss function and the values of its parameters. He also suggested work on a better specification of the regression model especially regarding the choice of variables, (eg number of bedrooms)

Varians study, and the others cited above, are just a small subset of the applications of Bayesian econometrics, but they do serve to indicate the wide range of econometric problems to which the Bayesian method has been found applicable. As Zellner (1981) points out, the progress of Bayesian econometrics depends critically on the quality of the results produced in practice. These results appear so far to be most satisfactory and this can only lead to the accelleration of the use of Bayesian Techniques in econometrics.

2.2 Medical, Scientific and Industrial Applications

The area of medical, scientific and industrial applications of Bayesian techniques is a very large one. Topics of study range from cancer and genetic research, forecasting of oil reserves and analysis of weather modification experiments, to life testing and analysis of system reliability. Problems range from straightforward estimation and hypothesis testing to more involved ones of prediction and decision making.

In the area of medical research, A F M Smith of the University of Nottingham uses the multiprocess Kalman Filter (Harrison and Stevens 1976) for the analysis of noisy time series data. Smith investigates the development of an on-line statistical procedure for monitoring the state of kidney function in transplant patients.

Although it is not possible to measure renal performance directly, it is possible to do so indirectly by monitoring the concentration of various substances in the bloodstream. Medical research has shown that the serum creatine chemical series is a good measure of the level of renal function, there being an inverse relationship between the level of creatine in the bloodstream, and renal performance. There is unfortunately a lag between the time that significant changes in renal behavior occur, (such as imminent rejection of the donor kidney), and the time taken for the creatine level to change significantly. Creatine level is also partially obscured by the presence of other factors in the bloodstream. This noise, combined with measurement and other errors must be filtered out, (hence the use of the Kalman Filter), for effective predictions to be made. The

aim of the research was to predict changes in kidney behavior that might indicate imminent rejection, before these changes become too advanced for preventative action to be taken.

Smith and Cook (1980) used a Bayesian analysis of a straight line model with a change point, (change in slope) to infer rejection crisis times for transplanted kidneys. They derived the posterior distribution for the time at which a marked change in kidney functioning occured. A more general time series model was developed by Smith and West (1982), based on Lindley and Smith's analysis of the linear model (1977). This was followed by a more decision orientated study (1983) aimed at taking the optimum action based on the predictions made.

Clearly in examples such as these, prior information takes the form of previously observed sample data. (Data Based Prior Knowledge) Data based prior information is also used by Du-Mouchel and Harris (1981), in their study of Bayesian methods for combining the results of cancer experiments in human and other species. They compare the effects of the known human carcinogens, roofing tar, coke oven emissions and cigarette smoke, on hamster embryo cells and also on two different species of mice. They then expose these species to other environmental carcinogens such as emissions from diesel and petrol engines and extrapolate the carcinogenic effects to the human species where results are not available.(And clearly we cannot experiment directly!!) The great advantage of this form of analysis is that all available information is used to reduce the uncertainty concerning the health risks of supposedly carcinogenic substances. Difficulties however arise in assessing the relevence of non-human

data to man.

A further apparently unrelated study, was carried out by Altshuler (1976), who uses the rate of observed cancers in test animals, combined with a subjective prior distribution, to predict an upper credible level for the rate of cancers in humans. His Non-Data-Based prior information incorporates his knowledge of the species transfer function relating the relative effects of non-human/human carcinogens on human/non-human species and also our belief of whether a certain carcinogen is species specific.

Bayes methods have also been applied in the area of genetic research. Steve and Rossi-Mori (1983) use a Bayesian approach for the development of a genetic screening programme, aimed at the detection of couples having a significant probability of giving birth to a homozygous individual afflicted with the serious genetic disease, Beta Thalassaemia.

They suggest the use of personal data, such as age, sex, demographic origin, number of children, geneological links with relatives who were known carriers,(heterozygous) or afflicted, (homozygous recessive) to derive a prior probability for an individual being a carrier. This probability is combined with haematological and fertility data to predict the risk of giving birth to a child and being a carrier. It is then suggested that this information is utilized in making the decision as to whether a definitive (and expensive) test should be carried out. Bearing in mind availability of finance, and suitable testing facilities. Finally, a comparison is made with screening programmes already in force, and it is

proposed that while an explicitly Bayesian approach is not at present being used, health planners are unconsciously Bayesian, despite their use of classical techniques such as discriminant functions.

Additionally the Bayesian method has been applied to the problem of estimating the linkage between genes. Tan (1980), used prior information concerning the value of the linkage parameter θ, ($0 \leq \theta \leq 0.5$), to derive a posterior estimate of θ. Classical estimators based on Maximum likelihood methods, neglect to utilize any prior information regarding θ, even the known fact that $0 \leq \theta \leq 0.5$.

Tan suggests a general prior distribution for θ, and proposes a Bayesian estimation procedure based on this prior and the quadratic loss structure. A simulation study was carried out as well as a comparison with Maximum Likelihood results. Results of this study, clearly indicate the superiority of Bayesian methods, especially when we have some prior knowledge about θ and when the sample size n is not large.

Other recent applications of Bayesian techniques to medical and biological problems include a study of the rate of parasite infestation in mice, Piccinato (1983); An analysis of optimal stopping rules in clinical trials, Freedman and Spiegelhalter (1983); and a method for the prediction of Ovulation times in women, Carter and Blight (1983). Aside from its application in medical and biological research there has been much interest in the development of Bayesian techniques for other research applications. In the area of industrial research, Baker and Lane-Joynt, (1983) of the Unilever research team, raise a number of interesting points regarding the

appeal of Bayesian techniques to their clients. This appeal, they propose is based on the following points.

(1) Clients are involved in discussion of the relationship between their prior knowledge and that gained from further experimentation.
(2) The method leads to an aggregation of new and prior knowledge.
(3) The method relates more to the client's own intuitive approach.

Furthermore, they note from their own experience as research statisticians, that we should always summarize existing knowledge, before undertaking further experimentation. This point has particular relevence in the industrial situation, where we often cannot afford to undertake unneccessary experimetation.

Particular applications in scientific and industrial areas include; A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification experiments, Simpson, Olsen and Eden (1975); A Discrimination problem in Forensic science, Lindley (1977); A Bayesian procedure for Forecasting Oil and Gas Discoveries in Mature Exploration Provinces, Meisner and Demirmen (1981); and the development of a Bayes estimate for population size in a Capture-recapture experiment, Gaskell and George (1972).

2.2.1 A Bayesian Modification of the Lincoln Index

Gaskell and George (1972) incorporate Bayesian methods into an approach for the estimation of population size in a capture-recapture experiment.

Consider a fixed population of unknown size N, of which R are marked or tagged in some way. Assume that a sample of S individuals is taken of which M are found to be marked.

Lincoln (1930) proposed the estimate ...

$$N = \frac{R.S}{M} \qquad \ldots(3.1)$$

for the unknown population size N.

Clearly, when M is small and RS is large enough, a change of only 1 in M may result in a large change in N. Furthermore, it is possible for M to equal zero, in which case N is undefined.

Bailey (1952) suggested the following modification to cope with this problem ...

$$N = \frac{R(S + 1)}{(M + 1)}$$

on the grounds that this estimate is less based.

Gaskell and George point out however, that this criterion is not a particularly valid one, since normally the experiment is to be performed once only. They note that while the experimenter will not know the value of N before the experiment commences, he will certainly have some rough ideas about its value. These ideas may only be in the form of a lower limit ($N \geq R$) and perhaps an upper limit, but when experimental evidence itself is weak, then the

inclusion of even this degree of information will be a considerable aid in producing a satisfactory estimate of N.

Assuming binomial sampling which is reasonable when $R < N$, then the sample likelihood is proportional to ...

$$\left(\frac{R}{N}\right)^{M}\left(1 - \frac{R}{M}\right)^{S-M}$$

irrespective of the sampling rule.

Let $\theta = \frac{N-R}{R}$

For given θ the estimate of N is $N = R(1 + \theta)$

The likelihood function is therefore proportional to $\theta^{S-M}(1+\theta)^{-S}$

Assuming a Beta prior distribution for θ we have ...

$$p(\theta) \propto \theta^{p-1}(1+\theta)^{-(p+q)} \quad \theta \geq 0.$$

which has mean $\frac{p}{q-1}$ and variance $\frac{p(p+q-1)}{(q-1)(q-2)}$

and yields the prior estimate of N as ...

$$A = R\left[\left(\frac{p}{q-1}\right) + 1\right]$$

And so the posterior distribution for θ is proportional to ...

$$\theta^{p-1}(1+\theta)^{-(p+q)} \cdot \theta^{S-M}(1+\theta)^{-S}$$

$$= \theta^{S-M+p-1}(1+\theta)^{-(p+q+S)}$$

which is Beta with mean $\theta' = \frac{S-M+p}{M+q-1}$.

And so the posterior estimate of N is ...

$$N' = R(1+\theta')$$

$$= R\left(1 + \frac{S-M+p}{M+q-1}\right)$$

$$= R\left(\frac{p+q+S-1}{M+q-1}\right)$$

$$= \frac{RS+AB}{M+B}$$

where A is the <u>prior</u> estimate of N and $B = q-1$

This estimate can be rewritten as ...

$$N' = L - \frac{B(L - A)}{M+B}$$

where L is the "Lincoln Index" $L = \frac{R.S}{M}$

Simulation studies suggest the value B = 2 as being most satisfactory in terms of the minimum average relative error, $\frac{N-N'}{N}$.

This leads to the estimate ...

$$N' = \frac{RS+2A}{M+2} \qquad \ldots(3.2)$$

Further simulations were also carried out to determine the sensitivity of to the prior estimate A, of N. These simulations indicate that (3.2) is robust, and should be used in preference to (3.1), whenever a prior estimate of N is available.

Finally, Gaskell and George note that in most practical applications, data from more than a single recapture is used. They propose a further study aimed at extending the method to such cases.

2.3 Education Applications

Over the last 15 years, there has been a great acceleration in the development of Bayesian methods in Educational Research. This growth, while not apparent in New Zealand, has been particularly noticeable in the United States.

In this country we are, at present, blessed with a system of higher education in which admission to most courses of study depends mainly on the students own area or academic interest, as does, in general the choice of which University the student wishes to attend. This is unfortunately not the case in many countries.

In the United States, for example, there are both private and state Universities, and population size places constraints on the number of students that a University can accept for a given course of study. In many U.S. Universities therefore, academic aptitude tests have been heavily used in the process of deciding whether to select one applicant over another. Novick (1971), notes that such tests have had the effect of making academic admissions dependent on academic promise rather than on status and influence.

Extensive studies have been undertaken to determine how these tests should be interpreted. Clearly it is in the best interests of both students and admissions officer, that both understand how the tests work. The admissions officer will benefit from having students more clearly informed of admission requirements (e.g. scores of aptitude test required), and the student will also benefit from the information he receives about himself, the university and particular

programme most relevant to his goals.

In 1964 the American College Testing Programme, (ACT), was set up to provide a guidance orientated testing programme for students. This programme provides the student with test scores and other information about himself, as well as predictions of his performances at the various colleges which he is interested in attending. Other programmes such as Comparative Guidance and Placement Program (C.G.P.), and a guidance orientated system, the Career Planning Profile (C.P.P.), have also heen developed by ACT.

Because of their ability to incorporate both prior and collateral information into the decision making process, Bayes methods have been widely used in Educational Testing. One of the foremost researchers in this area has been Professor Melvin Novick of the American College Testing Program. Professor Novick has, for many years, specialized in the development of Bayesian Methodology for the analysis of Educational Tests. Some specific topics that Novick has investigated include; Monitoring the performance of modular instruction programs (1975); development of Bayesian Guidance procedures, (1971); A Bayesian evaluation of methods for the elimination of cultural or racial bias in a Universities/Businesses selection of students/employees, (1976), (Complicated by the fact that institutions are compelled <u>legally</u> to discriminate <u>positively</u> towards some minority groups); and the development of an interactive computer program for Bayesian Data Analysis, Novick (1973).

Apart from the work of Novick, other researchers such as Jackson, (1980), Lindley, and Rubin, (1983), have also applied Bayesian

methods to the analysis of Educational data. The proceedings ot the 1982 I.O.S. Conference on practical Bayesian Statistics, also include three particularly interesting applications, (due to Rubin), which are well worth reading.

2.4 Concluding Comments

Applications mentioned in previous sections, represent of course, only a small subset of the practical uses of Bayesian methods. They do however, serve to illustrate the tremendous variety of problems for which viable Bayesian solutions have been found.

Aside from developmental work aimed at particular applications, there has been, in recent years, much general research into Bayesian Methods. Consider for example, the area of Multivariate Statistical Analysis. Sik-Yum Lee (1981) has investigated a Bayesian approach to the area of Confirmatory Factor Analysis, in which prior knowledge concerning the model parameters is available; Martin and MacDonald (1975) have proposed a Bayesian method for avoiding Heywood cases in ordinary factor analysis; and Binder (1978) has used a Bayesian Decision theoretic method to produce a solution to the problem of Cluster analysis, especially when the true number of clusters is unknown.

Other important area of statistical analysis such as Linear models, Regression analysis and general estimation techniques are also being studied. (From both Pure Bayesian and Bayesian Decision Theoretic viewpoints.) There is also much activity in the philosophical development of the field.

Like Classical Statistical Analysis, the field of Bayesian statistics is exceedingly large, and for this reason our presentation of some topics has neccessarily been only a summary of the total research to date. We have however, attempted to present a broad overview of all

the central concepts and techniques as well as pointing out a number of interesting applications.

It is hoped that after reading this thesis, the reader will realise the importance, and indeed the neccessity, of utilizing prior information in statistical analysis, through the use of Bayes theorem, and with a spirit of healthy criticism, come to develop a fuller appreciation of the value of Bayesian methods.

Appendix One

When the likelihood function obeys certain regularity conditions, and n is sufficiently large, then $\ell(\theta|\underline{x})$ is approximately normal.

Proof

Define $L(\theta|\underline{x}) = \log_e \ell(\theta|\underline{x})$

Expanding as a Taylor series about $\hat{\theta}$, the maximum likelihood estimate of θ ...

$$L(\theta|\underline{x}) \cong L(\hat{\theta}|\underline{x}) + (\theta-\hat{\theta}) \frac{\delta L}{\delta\theta}\bigg|_{\hat{\theta}} + \tfrac{1}{2}(\theta-\hat{\theta})^2 \frac{\delta^2 L}{\delta\theta^2}\bigg|_{\hat{\theta}} + \ldots$$

But at $\theta = \hat{\theta}$, $\dfrac{\delta L}{\delta\theta} = 0$ by definition.

So $$L(\theta|\underline{x}) \cong L(\hat{\theta}|\underline{x}) + \tfrac{1}{2}(\theta-\hat{\theta})^2 \frac{\delta^2 L}{\delta\theta^2}\bigg|_{\hat{\theta}}$$

$$\cong L(\hat{\theta}|\underline{x}) - \frac{n}{2}(\theta-\hat{\theta})^2 \frac{1}{n} \frac{\delta^2 L}{\delta\theta^2}\bigg|_{\hat{\theta}}$$

$$\cong L(\hat{\theta}|\underline{x}) - \tfrac{1}{2}(\theta-\hat{\theta})^2 / [J^{-1}(\hat{\theta})/n] \qquad \ldots A1.1$$

where $J(\hat{\theta}) = \dfrac{1}{n} \dfrac{\delta^2 L}{\delta\theta^2}\bigg|_{\hat{\theta}}$

Now in general if $X \sim N(\mu,\sigma^2)$,

then $\log_e p(x) = \text{constant} - \tfrac{1}{2}(x-\mu)^2/\sigma^2$, $\qquad \ldots A1.2$

where $p(x)$ is a Normal pdf.

Comparing A1.1 with A1.2, we see that ...

$$\ell(\theta|\underline{x}) \sim N(\hat{\theta}, J^{-1}(\hat{\theta})/n) \qquad \ldots A1.3$$

i.e. $\ell(\theta|\underline{x})$ is approximately Normal.

Appendix Two

$$(\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta}) = (\underline{y}-X\hat{\underline{\beta}}+X\hat{\underline{\beta}}-X\underline{\beta})'(\underline{y}-X\hat{\underline{\beta}}+X\hat{\underline{\beta}}-X\underline{\beta})$$

$$= [\underline{y}-X\hat{\underline{\beta}}] + [X\hat{\underline{\beta}}-X\underline{\beta}] \quad [\underline{y}-X\hat{\underline{\beta}}] + [X\hat{\underline{\beta}}-X\underline{\beta}]$$

$$= (\underline{y}-X\hat{\underline{\beta}})'(\underline{y}-X\hat{\underline{\beta}}) + 2(X\hat{\underline{\beta}}-X\underline{\beta})'(\underline{y}-X\hat{\underline{\beta}}) + (X\hat{\underline{\beta}}-X\underline{\beta})'(X\hat{\underline{\beta}}-X\underline{\beta})$$

...(A2.1)

Now let $(\underline{y}-X\hat{\underline{\beta}})'(\underline{y}-X\hat{\underline{\beta}}) = \nu s^2$

and so A2.1 reduces to ...

$$\nu s^2 + 2[X(\hat{\underline{\beta}}-\underline{\beta})]'[\underline{y}-X\hat{\underline{\beta}}] + [X(\hat{\underline{\beta}}-\underline{\beta})]'[X(\hat{\underline{\beta}}-\underline{\beta})]$$

$$= \nu s^2 + 2(\hat{\underline{\beta}}-\underline{\beta})'X'(\underline{y}-X\hat{\underline{\beta}}) + (\hat{\underline{\beta}}-\underline{\beta})'X'X(\hat{\underline{\beta}}-\underline{\beta}) \quad ...(A2.2)$$

But if we choose $\hat{\underline{\beta}}$ so that $\underline{y}-X\hat{\underline{\beta}} = 0$, then the second term vanishes.

This choice of $\hat{\underline{\beta}}$ is given by $\underline{y} = X\hat{\underline{\beta}}$,

where $X'\underline{y} = X'X\hat{\underline{\beta}}$. Assuming X is of full rank.

Premultiplying both sides by $(X'X)^{-1}$ we have ...

$$(X'X)^{-1}X'\underline{y} = (X'X)^{-1}X'X\hat{\underline{\beta}}$$

$$= I\hat{\underline{\beta}}$$

And so $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$

Equation A2.2 now reduces to ...

$$\nu s^2 + (\hat{\underline{\beta}}-\underline{\beta})'X'X(\hat{\underline{\beta}}-\underline{\beta}) \quad ...(A2.3)$$

where $s^2 = \dfrac{(\underline{y}-X\hat{\underline{\beta}})'(\underline{y}-X\hat{\underline{\beta}})}{\nu}$

Where $\nu = n-k$ and $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$.

Bibliography.

Aitchison J and Dunsmore I.R. (1975). Statistical Prediction Analysis. Cambridge University Press.

Altshuler B. A Bayesian Approach to Assessing Population Risks from Environmental Carcinogens. Institute of Environmental Medicine Report, New York University Medical Centre.

Aykak A. and Brumat C. (Eds). (1977). New Developments in Applications of Bayesian Methods. Contributions to Economic Analysis. North Holland.

Baker A.G. and Lane-Joynt D.R. (1983). Bayesian Statistics in an Industrial Research Laboratory. In Proceedings of 1982 I.O.S Annual Conference on Practical Bayesian Statistics. 32, 118-123.

Barnett V. (1973). Comparative Statistical Inference. London, Wiley.

Berger J.O. (1980). Statistical Decision Theory: Foundations, Concepts and Methods. Springer Series in Statistics. New York, Springer-Verlag.

Binder D.A. (1978). Bayesian Cluster Analysis. Biometrika, 65, 31-38.

Box G.E.P. and Tiao G.C. (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley.

Carter R.L. and Blight B.J.N. (1983). A Bayesian Change-Point Problem with an Application to the Prediction and Detection of Ovulation Times in Women. In Proceedings of 1982 I.O.S Annual Conference on Practical Bayesian Statistics. 32, 229-230.

DeGroot M.H. (1970). Optimal Statistical Decisions. McGraw-Hill.

DuMouchel W.H. and Harris J.E. (1981). Bayes Methods for Combining Cancer Experiments in Humans and other Species. Technical Report NSF-24, Statistics Centre, Massachusetts Institute of Technology.

DuMouchel W.H. (1983). The 1982 Massachusetts Automobile Insurance Classification Scheme. In Proceedings of 1982 I.O.S Annual Conference on Practical Bayesian Statistics. 32, 69-81.

Ferguson T.S. (1976). Mathematical Statistics: A Decision Theoretic Approach. New York, Academic Press.

Freedman L.S. and Spiegelhalter D.J. (1983). The Assessment of Subjective Opinion and its Use in Relation to Stopping Rules for Clinical Trials. In Proceedings of 1982 I.O.S Annual Conference on Practical Bayesian Statistics. 32, 153-160.

Jackson P.H. and Novick M.R. (1980). Adversary Preposterior Analysis for Simple Parametric Models. Bayesian Analysis in Econometrics and Statistics. A Zellner (Ed). North-Holland.

Gaskell T.J. and George B.J. (1972). A Bayesian Modification of the Lincoln Index. J. Applied Ecology., 9, 377-384.

Houle A. (1983). The Geneological Tree of Bayesians. In Proceedings of 1982 I.O.S Annual Conference on Practical Bayesian Statistics. 32, 214-215.

Lewis C. et al. (1975). Marginal Distributions for the Estimation of Proportions in m Groups. Psychometrika. 40, 63-75.

Lindley D.V. and Novick M.R. (1981). The role of Exchangebility in Inference. Annals of Statistics, 9, 45-58.

Lindley D.V. (1971). Bayesian Statistics, A Review. Regional Conference Series in Applied Mathematics, SIAM.

Lindley D.V. (1977). A Problem in Forensic Science. Biometrika, 64, 207-213.

Lindley D.V. (1983). The Theory and Practice of Bayesian Statistics. In Proceedings of 1982 I.O.S Annual Conference on Practical Bayesian Statistics. 32, 1-13.

Lindley D.V. and Smith A.F.M. (1972). Bayes Estimates for the Linear Model. J.R.S.S.

Maritz, J.S. (1970) Empirical Bayes Methods. London, Methuen.

Martin J.K. and McDonald R.P. (1975). Bayesian Estimation in Unrestricted Factor Analysis: A Treatment for Heywood Cases. Psychometrika, 40, 505-517.

Meisner J. and Demirmen F. (1981). The Creaming Method: A Bayesian Procedure to Forecast Future Oil and Gas Discoveries in Mature Exploration Provinces. J.R.S.S., Series A, 144, 1-31.

Naylor J.C. and Smith A.F.M. Applications of a Method for the Efficient Computation of Posterior Distibutions. To appear in Journal of Applied Statistics.

Novick M.R. (1971). Bayesian Considerations in Educational Information Systems. Educational Testing Service. From Proceedings of the 1970 Invitational Conference on Testing Problems.

Novick M.R. (1975?). Bayesian Methods in Educational Testing: A Third Survey. Paper of unknown origin.

Novick M.R. (1973). Educational Decisions with Limited Information. Invited Address at dedication of Lindquist Centre for Measurement, Iowa City, Iowa.

Novick M.R. (1973). High School Attainment: An Example of a Computer-Assisted Bayesian Approach to Data Analysis. International Statistical Review. 41,264-271.

Novick M.R. and Grizzle J.E. (1965). A Bayesian Approach to the Analysis of Data from Clinical Trials. J.A.S.A., 60, 81-96.

Novick M.R. and Jackson P.H. (1983). Bayesian Guidance Technology. Review of Educational Research. 40, .459-494.

Novick M.R. and Lindley D.V. (1978). The Use of More Realistic Utility Functions in Educational Applications. Journal of Educational Measurement. 15,181-191.

Novick M.R. et al. (1972). Estimating Multiple Regressions in m Groups: A Cross Validation Study. J. Math. Statist. Psychol. 25, 33-50.

Novick M.R. et al. (1973). The Estimation of Proportions in m Groups. Psychometrika. 38, 19-46.

Peck S.C. (1974). Alternative Investment Models for firms in the Electric Utility Industry. Bell Journal of Economic and Management Science. 5, 420-458.

Petersen N.S. and Novick M.R. (1976). An Evaluation of Some Methods fo Culture-Fair Selection. Journal of Educational Measurement. 13, 3-30.

Piccinato L. (1983) Some Statistical Problems in Parasite Experiments. In Proceedings of 1982 I.O.S Annual Conference on Practical Bayesian Statistics. 32, 138-143.

Raiffa H and Schlaifer R. (1967). Applied Statistical Decision Theory: Studies in Managerial Economics. Graduate School of Business Administration, Harvard University.

Robbins, H. (1964). The Empirical Bayes approach to Statistical decision problems. Annals of Mathematical Statistics, 35, 1-20.

Rothenberg T.J. (1974). The Bayesian Approach and Alternatives in Econometrics: Part Two. Studies in Bayesian Econometrics and Statistics, Fienberg and Zellner (Eds). North Holland.

Rubin D.H. (1983). Some Applications of Bayesian Statistics to Educational Data. In Proceedings of 1982 I.O.S Annual Conference on Practical Bayesian Statistics. 32, 55-68

Savage L.J. (1954). The Foundation of Statistics. Wiley Publication in Statistics.

Siczewicz P.J. (1981). A Procedure to Determine Massachusetts Automobile Insurance Territories. Technical Report No. 30. Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Sik-Yum Lee. (1981). A Bayesian Approach to Confirmatory Factor Analysis. Psychometrika, 46, 153-160.

Simpson J. and Olsen A. and Eden J.C. (1975). A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification. Technometrics, 17, 161-166.

Smith A.F.M. and Cook D.G. (1980). Straight Lines with a Change-point: A Bayesian Analysis of some Renal Transplant Data. J.R.S.S., Series C, 29, 180-189.

Smith A.F.M. and West M. (1982). Monitoring Renal Transplants: An Application of the Multi-process Kalman Filter. Research Report. Department of Mathematics, University of Nottingham.

Smith A.F.M. et al. (1983). Monitoring Kidney Transplant Patients. In Proceedings of 1982 I.O.S Annual Conference on Practical Bayesian Statistics. 32, 46-54.

Smith J.Q. (1983). Forecasting Accident Claims for an Assurance Company. In Proceedings of 1982 I.O.S Annual Conference on Practical Bayesian Statistics. 32, 109-115.

Spiegelhalter D.J. and Smith A.F.M. Decision Analysis and Clinical Decisions. Department of Mathematics, University of Nottingham.

Steve G. and Rossi-Mori A. (1983) Prescreening of Beta-Thalassaemia Carriers: a Comparison Between Bayyesian and Other Approaches. In Proceedings of 1982 I.O.S Annual Conference on Practical Bayesian Statistics. 32, 233-239.

Tan W.Y. (1980). Comparative Studies on the Estimation of Linkage by Bayesian Method and Maximum Likelihood Method. Commun. Statist.-Simula. Computa., B9(1), 19-41.

Varian H.R. (1974). A Bayesian Approach to Real Estate Assessment. Studies in Bayesian Econometrics and Statistics, Fienberg and Zellner (Eds). North Holland.

Winkler R.L. and Hays W.L. (1975). Statistics: Probability, Inference and Decision. 2nd ed. Holt, Rinehart and Winston.

Zellner A. (1971). An Introduction to Bayesian Inference in Econometrics. Wiley.

Zellner A. (1971). Applications of Bayesian Analysis in Econometrics. In Proceedings of 1982 I.O.S Annual Conference on Practical Bayesian Statistics. 32, 23-36.

Zellner A. (1981). The Current State of Bayesian Econometrics. H.G.B. Alexander Research Foundation Graduate School of Business Report, University of Chicago.

Zellner A. (1974). The Bayesian Approach and Alternatives in Econometrics: Part One. Studies in Bayesian Econometrics and Statistics, Fienberg and Zellner (Eds). North Holland.