CHAPTER 4

Bayesian Analysis

4.1. Introduction

This chapter differs from later chapters in scope, because Bayesian analysis is an essentially self-contained paradigm for statistics. (Later chapters will, for the most part, deal with special topics within frequentist decision theory.) In order to provide a satisfactory perspective on Bayesian analysis, we will discuss Bayesian inference along with Bayesian decision theory. Before beginning the study, however, we briefly discuss the seven major arguments that can be given in support of Bayesian analysis. (Later chapters will similarly begin with a discussion of justifications.) Some of these arguments will not be completely understandable initially, but are best placed together for reference purposes.

I. *Important Prior Information May be Available*

This point has already been discussed (cf. Section 1.1), but bears repeating; in a significant fraction of statistical problems, failure to take prior information into account can lead to conclusions ranging from merely inferior to absurd. Of course, most non-Bayesians would agree to the use of reliable and significant prior information, so the impact of this consideration for general adoption of the Bayesian paradigm is unclear. One of the advantages of adopting the Bayesian viewpoint, on the other hand, is that one will be far more likely to recognize *when* significant prior information is available. Also, when significant prior information is available, the Bayesian approach shows how to sensibly utilize it, in contrast with most non-Bayesian approaches. As a simple example, a common situation in statistics is to have a study on several different, but similar, populations, for each of which

is needed an estimate of variability. The question is—How should one use the important prior information that the populations are similar? Classical statistics is hard put to answer this. The usual approach, of deciding between separate estimates of variance or a pooled estimate (often based on some significance test) is an extremely crude utilization of the prior information. Bayesian analysis allows much more effective use of such prior information. (See Sections 4.5 and 4.6 for the Bayesian approach to this type of problem.)

II. *Uncertainty Should Be Quantified Probabilistically*

The business of statistics is to provide information or conclusions about uncertain quantities and to convey the extent of the uncertainty in the answer. The language of uncertainty is probability, and only the (conditional) Bayesian approach consistently uses this language to directly address uncertainty.

Consider, for instance, statistical hypothesis testing. The hypotheses are uncertain, and the result of a (conditional) Bayesian analysis will be simply the statement of the believed *probabilities* of the hypotheses (in light of the data and the prior information). In contrast, classical approaches provide "probabilities of Type I or Type II error" or "significance levels (P-values)," all of which are, at best, indirectly related to the *probabilities of the hypotheses* (see Subsection 4.3.3). As another example, we will see that when a Bayesian provides a "confidence set" (to be called a *credible set* in Bayesian language), the reported accuracy will be the believed *probability* that the set actually contains the unknown θ, in contrast to the classical coverage probability (see Subsection 1.6.2).

Of course, direct probability statements about uncertainty essentially *require* Bayesian analysis, and the thrust of classical statistics has been to find alternate ways of indicating accuracy. Indeed, even in classical elementary statistics courses it is common to spend a great deal of effort in pointing out that classical measures are *not* direct probability statements about uncertainty ("a 95% confidence interval is *not* to be interpreted as an interval that has probability 0.95 of containing θ").

There are two issues here, the first philosophically pragmatic and the second pragmatically pragmatic. The philosophically pragmatic issue is—What is the best method of quantifying uncertainty? The literature arguing in favor of direct probabilistic (Bayesian) quantification is vast (cf. Jeffreys (1961), Edwards, Lindman, and Savage (1963), deFinetti (1972, 1974, 1975), Box and Tiao (1973), Lindley (1982a), Good (1983), and Jaynes (1983)). And these contain not *just* philosophical arguments but also very compelling examples. (One such is discussed in Subsection 4.3.3, namely testing a point null hypothesis, where classical error probabilities or significance levels convey a completely misleading impression as to the validity of the null hypothesis.)

The second issue is the very practical issue of how statistical *users* (as opposed to professional statisticians or, at least, those with extensive statistical training) interpret statistical conclusions. Most such users (and probably the overwhelming majority) interpret classical measures in the direct probabilistic sense. (Indeed the only way we have had even moderate success, in teaching elementary statistics students that an error probability is not a probability of a hypothesis, is to teach enough Bayesian analysis to be able to demonstrate the difference with examples.) Among the formal evidence for this misinterpretation of classical measures is an amusing study in Diamond and Forrester (1983). If the majority of users are incapable of interpreting classical measures except, incorrectly, as Bayesian probabilities (whether through our teaching inadequacies or the inherent obscurity of the classical measures), can it be right to provide classical measures?

III. *The Conditional Viewpoint*

In Section 1.6 it was argued that analysis conditional on the observed data, as opposed to frequentist averaging over all potential data, is of crucial importance. There was also a brief mention of strong arguments supporting the Bayesian approach to conditional analysis. Some discussion of these arguments, with examples, will be given later in the chapter, after some needed Bayesian machinery has been developed. (See Berger and Wolpert (1984) for more complete discussion.)

IV. *Coherency and Rationality*

Note that, if one does have a loss function developed via utility theory and a prior distribution for θ, then one should (by the very nature of the utility construction) evaluate an action a by the Bayesian expected loss, and evaluate a decision rule δ by the Bayes risk. This presupposes the existence of a loss and prior, however, and thus may not be a very compelling argument for Bayesian analysis.

As with the axiomatic development of utility theory, however, one can develop various axiomatic bases for statistics itself. These involve the assumption that a preference ordering exists among actions, decision rules, inferences, or statistical procedures (depending on the perceived statistical goal), together with a set of axioms that any "coherent" or "rational" preference ordering should satisfy. Most people find these axioms quite believable (with some exceptions, such as LeCam (1977)), and yet it is invariably found that any rational preference ordering *must* correspond to some type of Bayesian preference ordering. This provides strong support for the Bayesian viewpoint, in that any approach which fails to correspond with a Bayesian analysis must violate some very "common sense" axiom of behavior.

We will not present any of the axiom systems here, partly because there are so many that it is hard to choose among them (Fishburn (1981) reviews over 30 different systems for decision theory alone), and partly because they are not that different in nature from the utility axioms. Some other references to axiom systems are Ramsey (1926) and Savage (1954) (two of the earliest—that in deFinetti (1972, 1974, 1975) was also developed quite early), Ferguson (1967) (a simple and easily understandable case), Rubin (1987) (the most general decision-theoretic system), and Lindley (1982a) and Lane and Sudderth (1983) (and the references therein) which consider "inference" axiom systems. A component of many of the non-decision-theoretic systems is "betting coherency" or the "Dutch book" argument, which will be discussed in Subsection 4.8.3.

There are several subtleties involved in the conclusion of many of these axiomatic developments. One is that the conclusion often requires only that the preference pattern correspond to a Bayesian analysis with respect to what is called a *finitely additive* prior; priors considered in this book are countably additive. Another subtlety is that the developments do not necessarily lead to a separation of the prior from the loss (cf. Savage (1954) and Rubin (1987)). Discussion of either of these issues is beyond the scope of the book.

There is a third subtlety which is very relevant, however, namely that these axiomatic developments do not say that "to be coherent or rational one must do a Bayesian analysis." Instead they say that "to be coherent or rational the analysis *must correspond to* a Bayesian analysis." This is an important logical difference, in that there could be rational methods of choosing a statistical procedure other than the Bayesian method of developing a prior (and loss) and doing a Bayesian analysis. In fact, the method (see Chapter 6 for explanation of terms)—for invariant decision problems with compact parameter space use the best invariant decision rule—can be shown to be "rational," since it corresponds to Bayesian analysis with respect to the prior which is the Haar measure on Θ (see Section 6.6). It is true, however, that no broadly applicable "rational" statistical paradigm, other than the Bayesian paradigm, has been found. And, even if such were found, it would be hard to argue that Bayesian analysis could be ignored: the method, to be rational, must yield an answer corresponding to that for a Bayesian analysis with respect to some prior, and it would be hard to justify the answer if the corresponding prior seemed unreasonable.

Another important aspect of the proper interpretation of the conclusion of the axiomatic developments is that the conclusion does not say that *any* Bayesian analysis is good. A Bayesian analysis may be "rational" in the weak axiomatic sense, yet be terrible in a practical sense if an inappropriate prior distribution is used. Thus Smith (1961) says,

> "Consistency is not necessarily a virtue: one can be consistently obnoxious."

Indeed there is no logical axiomatic guarantee that the best way to be

consistent *and* nonobnoxious is through Bayesian analysis. See also Kiefer (1977b) and LeCam (1977) on this matter of interpretation of the rationality developments.

Another common criticism of rationality axioms is that numerous studies (cf. Ellsburg (1961)) have shown that people *do not* act in accordance with these axioms; hence the axioms are supposedly suspect. This criticism misses the point. The purpose of developing statistical methodology is to *improve* the way people act in the face of uncertainty, not to model how they do act. Smith (1984) responds to this behavioral criticism of the axioms by saying,

> "It is rather like arguing against the continued use of formal logic or arithmetic on the grounds that individuals can be shown to perform badly at deduction or long division in suitable experiments."

We have saved for last the most telling criticism of the majority of the axiomatic developments, namely the assumption that a preference ordering on *all* actions (or whatever) even exists. This simply is not going to be the case in what could be termed finite reality, namely the time and calculational constraints of our minds. We delay, however, formal discussion of this until Subsection 4.7.1, since it ties in with the argument that the ideal method of analysis is robust Bayesian analysis.

In spite of the limitations and "weak spots" in the rationality and coherency developments, they provide very powerful evidence that "truth" lies in a Bayesian direction. They also provide devastating weapons in exposing the "irrationality" of many other purported truths in statistics.

V. *Equivalence of Classically Optimal and Bayes Rules*

When a classical optimality principle is proposed, it is natural to reduce consideration to the class of statistical procedures which are acceptable according to this principle. In decision theory, for instance, it is natural to consider only admissible decision rules. As a special case, in simple versus simple hypothesis testing, with the desire for small error probabilities being the optimality principle, it is natural to reduce consideration to the class of "most powerful" tests. In such situations, it has repeatedly been shown that the class of "acceptable" decision rules corresponds to the class of Bayes decision rules (or some subclass or limits thereof). Chapter 8 presents a number of results of this type.

The correspondence between acceptable rules and Bayes rules clearly *suggests* that one should choose from among the acceptable rules through consideration of prior information. Consider, for instance, the situation mentioned above of testing between two simple hypotheses. The class of most powerful tests does essentially correspond to the class of Bayes tests (see Subsection 8.2.4); let us contrast the classical and Bayesian methods of selecting a test from this class. The classical approach is to select a most

powerful test by choosing desired error probabilities. Although it is common to select the test having Type I error probability of either $\alpha = 0.05$ or $\alpha = 0.01$, this practice is not particularly endorsed by most statisticians, who instead tend to urge choice of α based on "careful consideration and comparison of the two hypotheses." Unfortunately, this is not much help; the way in which this "careful consideration" can suggest a choice of α is very obscure. Contrast this with the Bayesian (decision-theoretic) approach which says: (i) determine the prior probabilities of each of the two hypotheses; (ii) determine the relative harm in mistakingly concluding each hypothesis; and (iii) use the corresponding Bayes test. This Bayes test will correspond to a most powerful test at some α level, but there is explicit guidance concerning which most powerful test to use. To reiterate, the options here are to either make a *subjective* selection of prior probabilities and decision losses for the hypotheses, or to make a *subjective* selection of the α level; we have no idea how to intuitively do the latter, whereas the Bayesian inputs are intuitively accessible. And this argument can be reversed: the choice of an α level will correspond to certain prior and loss beliefs, and if these beliefs are unreasonable, how can that α level be reasonable? This type of situation is encountered very regularly in classical statistics; even after application of some optimality criterion, there are many possible procedures to use, and it seems hard to argue that an intuitive choice is actually better than attempted quantification of the factors that should be involved in the intuitive choice. Section 5.4 presents an interesting class of such examples.

The above argument, that classical optimality criteria themselves lead to consideration of the class of Bayes rules, is actually very similar to the coherency axiomatics. The big difference is that classical measures still form the basis of the development. Thus, in simple versus simple hypothesis testing, the reported accuracy would still presumably be the relevant error probability, even if α were chosen by Bayesian means (i.e., chosen so as to yield the Bayes test for the available prior information and loss). For this reason, the implications of this argument for Bayesian analysis are somewhat limited in scope. On the other hand, the fact that one seems to end up with Bayes rules, even when starting down a non-Bayesian route, is highly suggestive. Non-Bayesians tend to view this as a mathematical coincidence, but the history of science teaches us that coincidences are usually trying to tell us something.

VI. *Operational Advantages of Bayesian Analysis*

One response, to the above arguments that only Bayesian analysis is completely sensible, is to agree, but state that Bayesian analysis is too hard. After all, one must determine a prior distribution and (for decision theory) a loss function, and we have seen that such determinations are not easy.

Therefore, the argument goes, we should accept nonoptimal, but easier, classical analyses. (Classical decision-theorists clearly have a harder time making this argument, especially because most people find it easier to elicit priors than loss functions.)

This argument is not without some merit. Indeed, in Section 4.7 we will discuss certain situations in which a classical analysis seems to be the best bet because of formidable technical difficulties in conducting a Bayesian analysis (and because the classical analysis can be given a type of Bayesian validation). On the whole, however, we would argue that the situation is just the opposite: for a given investment of effort on a problem, it is Bayesian analysis that is most likely to yield the best answers. Until we see how Bayesian analysis works, of course, it is hard to compare its operational effectiveness with that of classical statistics. And even then the only way to really make a comparison is to go through a large number of situations and actually apply the two paradigms. Although we will see some such comparisons as we proceed, the book's emphasis on ideas (rather than applied methods) prevents a thorough comparison. (In some sense, in any case, one can only be convinced as to the operational advantages of Bayesian analysis by personally trying it on a number of problems, and seeing the clarity that results.) In spite of these provisos it will be helpful to list some of the significant operational advantages of Bayesian analysis:

1. Conditioning on the observed data x introduces great simplifications in the analysis, as mentioned in Section 1.6; one need only work with the observed likelihood function, rather than with averages over all $\mathscr{X}$. The advantages of this include:
   (a) Realistic models can more easily be chosen, since there is less need to have models which allow special frequentist calculations (cf. Rubin (1984)).
   (b) Robustness (of all types) can be dealt with more easily, since (by (a)) model variations cause no essential changes in the needed calculations (cf. Box (1980), Rubin (1984), and Smith (1984)).
   (c) Optional stopping (see Section 7.7) becomes permissible, and as Edwards, Lindman, and Savage (1963) say,

> "The irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design .... Many experimenters would like to feel free to collect data until they have either conclusively proved their point, conclusively disproved it, or run out of time, money, or patience."

   Note that classical statistics does *not* allow one to stop an experiment when some *unanticipated* conclusive evidence appears.
   (d) Various kinds of censoring of data cause no essential problem for Bayesian analysis, but serious difficulties for classical analysis (cf. Berger and Wolpert (1984)).

2. Bayesian analysis with noninformative priors will be seen to be simple and remarkably successful. If one desires to avoid subjective prior specification, it is rare that one can do better than a Bayesian noninformative prior analysis.
3. Bayesian analysis yields a final distribution (the posterior distribution) for the unknown θ, and from this a large number of questions can be answered simultaneously. For instance, one can not only estimate θ, but can (with little additional effort) obtain accuracy measures for the estimate (or, alternatively, can obtain Bayesian "credible sets" for θ). This is in contrast to classical statistics, for which obtaining estimates for θ and determining the accuracies of these estimates (or confidence sets) are two *very* different problems. (Examples will be given in Sections 4.3 and 4.6.) Another illustration of the ease with which a variety of answers can be obtained from the posterior distribution is multiple hypothesis testing. One can calculate the Bayesian probability of any number of hypotheses (from the posterior distribution), while classical multiple hypothesis testing becomes much more difficult as more hypotheses are involved.
4. Bayesian analysis is an excellent alternative to use of large sample asymptotic statistical procedures. Bayesian procedures are almost always equivalent to the classical large sample procedures when the sample size is very large (see Subsection 4.7.8), and are likely to be more reasonable for moderate and (especially) small sample sizes (where many classical large sample procedures break down). Indeed, unless there have been extensive studies establishing the small and moderate sample size validity of a particular classical large sample procedure, almost any plausible Bayesian analysis would seem preferable (unless calculationally too difficult).

VII. *Objectivity and Scientific Uncertainty*

The most frequent criticism of Bayesian analysis is that different reasonable priors will often yield different answers, a supposedly unappealing lack of objectivity. The issue of objectivity was addressed in Section 3.7, where it was argued that, in attempting to achieve objectivity, there is no better way to go than Bayesian analysis with noninformative priors. We will not repeat these arguments here, but should mention the other side of the coin—when different reasonable priors yield substantially different answers, can it be right to state that there *is* a single answer? Would it not be better to admit that there is scientific uncertainty, with the conclusion depending on prior beliefs?

None of the above seven arguments for Bayesian analysis is completely convincing by itself, although we feel that most of the arguments can be

made to be almost compelling, with appropriate "fleshing-out" (cf. the fleshing-out of IV in Berger and Wolpert (1984)). Taken as a whole, the arguments provide strong evidence, indeed, in support of the centrality of the Bayesian viewpoint to statistics. There are, of course, various criticisms of Bayesian analysis, some of which have already been mentioned (here and in Section 3.7), and others which will be encountered in Sections 4.7 through 4.12. These additional criticisms warn the Bayesian not to be too dogmatic. It is crucial to keep sight of the ideal Bayesian goal, but one should be pragmatic about how best to achieve this goal.

4.2. The Posterior Distribution

Bayesian analysis is performed by combining the prior information ($\pi(\theta)$) and the sample information (x) into what is called the posterior distribution of θ given x, from which all decisions and inferences are made. This section discusses the meaning and calculation of this distribution.

4.2.1. Definition and Determination

The *posterior distribution of θ given x* (or *posterior* for short) will be denoted $\pi(\theta|x)$, and, as the notation indicates, is defined to be the conditional distribution of θ given the sample observation x. Noting that θ and X have joint (subjective) density

$$h(x, \theta) = \pi(\theta) f(x|\theta),$$

and (as in Subsection 3.5.1) that X has marginal (unconditional) density

$$m(x) = \int_{\Theta} f(x|\theta) dF^{\pi}(\theta),$$

it is clear that (providing $m(x) \neq 0$)

$$\pi(\theta|x) = \frac{h(x, \theta)}{m(x)}.$$

The name "posterior distribution" is indicative of the role of $\pi(\theta|x)$. Just as the prior distribution reflects beliefs about θ *prior* to experimentation, so $\pi(\theta|x)$ reflects the updated beliefs about θ after (*posterior* to) observing the sample x. In other words, the posterior distribution combines the prior beliefs about θ with the information about θ contained in the sample, x, to give a composite picture of the final beliefs about θ. Note that the Likelihood Principle is implicitly assumed in the above statement, in that there is felt to be no sample information about θ other than that contained in $f(x|\theta)$ (for the given x).

In calculating the posterior distribution, it is often helpful to use the concept of sufficiency. Indeed if T is a sufficient statistic for θ with density $g(t|\theta)$, the following result can be established. (The proof is left as an exercise.)

Lemma 1. *Assume $m(t)$ (the marginal density of t) is greater than zero, and that the factorization theorem holds. Then, if $T(x) = t$,*

$$\pi(\theta|x) = \pi(\theta|t) = \frac{\pi(\theta)g(t|\theta)}{m(t)}.$$

The reason for determining $\pi(\theta|x)$ from a sufficient statistic T (if possible) is that $g(t|\theta)$ and $m(t)$ are usually much easier to handle than $f(x|\theta)$ and $m(x)$.

EXAMPLE 1. Assume $X \sim \mathcal{N}(\theta, \sigma^2)$, where θ is unknown but σ^2 is known. Let $\pi(\theta)$ be a $\mathcal{N}(\mu, \tau^2)$ density, where μ and τ^2 are known. Then

$$h(x, \theta) = \pi(\theta)f(x|\theta) = (2\pi\sigma\tau)^{-1} \exp\left\{-\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\tau^2} + \frac{(x-\theta)^2}{\sigma^2}\right]\right\}.$$

To find $m(x)$, note that defining

$$\rho = \tau^{-2} + \sigma^{-2} = \frac{\tau^2 + \sigma^2}{\tau^2\sigma^2}$$

and completing squares gives

$$\begin{aligned}
\frac{1}{2}\left[\frac{(\theta-\mu)^2}{\tau^2} + \frac{(x-\theta)^2}{\sigma^2}\right] &= \frac{1}{2}\left[\left(\frac{1}{\tau^2} + \frac{1}{\sigma^2}\right)\theta^2 - 2\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right)\theta + \left(\frac{\mu^2}{\tau^2} + \frac{x^2}{\sigma^2}\right)\right] \\
&= \frac{1}{2}\rho\left[\theta^2 - \frac{2}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right)\theta\right] + \frac{1}{2}\left(\frac{\mu^2}{\tau^2} + \frac{x^2}{\sigma^2}\right) \\
&= \frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right)\right]^2 - \frac{1}{2\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right)^2 \\
&\quad + \frac{1}{2}\left(\frac{\mu^2}{\tau^2} + \frac{x^2}{\sigma^2}\right) \\
&= \frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right)\right]^2 + \frac{(\mu-x)^2}{2(\sigma^2+\tau^2)}.
\end{aligned}$$

Hence

$$h(x, \theta) = (2\pi\sigma\tau)^{-1} \exp\left\{-\frac{1}{2}\rho\left[\theta - \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right)\right]^2\right\} \exp\left\{-\frac{(\mu-x)^2}{2(\sigma^2+\tau^2)}\right\}$$

and

$$m(x) = \int_{-\infty}^{\infty} h(x, \theta)\,d\theta = (2\pi\rho)^{-1/2}(\sigma\tau)^{-1} \exp\left\{-\frac{(\mu-x)^2}{2(\sigma^2+\tau^2)}\right\}.$$

It follows that

$$\pi(\theta|x)=\frac{h(x,\theta)}{m(x)}=\left(\frac{\rho}{2\pi}\right)^{1/2}\exp\left\{-\frac{1}{2}\rho\left[\theta-\frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)\right]^2\right\}.$$

Note, from the above equations, that the marginal distribution of X is $\mathcal{N}(\mu,(\sigma^2+\tau^2))$ and the posterior distribution of θ given x is $\mathcal{N}(\mu(x),\rho^{-1})$, where

$$\mu(x)=\frac{1}{\rho}\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)=\frac{\sigma^2}{\sigma^2+\tau^2}\mu+\frac{\tau^2}{\sigma^2+\tau^2}x=x-\frac{\sigma^2}{\sigma^2+\tau^2}(x-\mu).$$

As a concrete example, consider the situation wherein a child is given an intelligence test. Assume that the test result X is $\mathcal{N}(\theta,100)$, where θ is the true IQ (intelligence) level of the child, as measured by the test. (In other words, if the child were to take a large number of independent similar tests, his average score would be about θ.) Assume also that, in the population as a whole, θ is distributed according to a $\mathcal{N}(100,225)$ distribution. Using the above equations, it follows that, marginally, X is $\mathcal{N}(100,325)$, while the posterior distribution of θ given x is normal with mean

$$\mu(x)=\frac{(100)(100)+x(225)}{(100+225)}=\frac{400+9x}{13}$$

and variance

$$\rho^{-1}=\frac{(100)(225)}{(100+225)}=\frac{900}{13}=69.23.$$

Thus, if a child scores 115 on the test, his true IQ θ has a $\mathcal{N}(110.39,69.23)$ posterior distribution. Note that, as discussed in Subsection 3.5.1, the $\mathcal{N}(100,325)$ marginal distribution of X would be the anticipated distribution of actual test scores in the population.

EXAMPLE 2. Assume a sample $\mathbf{X}=(X_1,\ldots,X_n)$ from a $\mathcal{N}(\theta,\sigma^2)$ distribution is to be taken (σ^2 known), and that θ has a $\mathcal{N}(\mu,\tau^2)$ density. Since $\bar{X}$ is sufficient for θ, it follows from Lemma 1 that $\pi(\theta|x)=\pi(\theta|\bar{x})$. Noting that $\bar{X}\sim\mathcal{N}(\theta,\sigma^2/n)$, it can be concluded from Example 1 that the posterior distribution of θ given $\mathbf{x}=(x_1,\ldots,x_n)$ is $\mathcal{N}(\mu(\mathbf{x}),\rho^{-1})$, where

$$\mu(\mathbf{x})=\frac{\sigma^2/n}{(\tau^2+\sigma^2/n)}\mu+\frac{\tau^2}{(\tau^2+\sigma^2/n)}\bar{x}$$

and $\rho=(n\tau^2+\sigma^2)/\tau^2\sigma^2$.

EXAMPLE 3. A blood test is to be conducted to help indicate whether or not a person has a particular disease. The result of the test is either positive (denoted $x=1$) or negative (denoted $x=0$). Letting θ_1 denote the state of nature "the disease is present" and θ_2 denote the state of nature "no disease is present," assume it is known that $f(1|\theta_1)=0.8, f(0|\theta_1)=0.2, f(1|\theta_2)=0.3,$

and $f(0|\theta_2) = 0.7$. According to prior information, $\pi(\theta_1) = 0.05$ and $\pi(\theta_2) = 0.95$. Then

$$m(1) = f(1|\theta_1)\pi(\theta_1) + f(1|\theta_2)\pi(\theta_2) = 0.04 + 0.285 = 0.325,$$
$$m(0) = f(0|\theta_1)\pi(\theta_1) + f(0|\theta_2)\pi(\theta_2) = 0.01 + 0.665 = 0.675,$$
$$\pi(\theta|x=1) = \frac{f(1|\theta)\pi(\theta)}{m(1)} = \begin{cases} \frac{0.04}{0.325} = 0.123 & \text{if } \theta = \theta_1, \\ \frac{0.285}{0.325} = 0.877 & \text{if } \theta = \theta_2, \end{cases}$$

and

$$\pi(\theta|x=0) = \frac{f(0|\theta)\pi(\theta)}{m(0)} = \begin{cases} \frac{0.01}{0.675} = 0.0148 & \text{if } \theta = \theta_1, \\ \frac{0.665}{0.675} = 0.9852 & \text{if } \theta = \theta_2. \end{cases}$$

It is interesting to observe that, even if the blood test is positive, there is still only a 12.3% chance of the disease being present. Note that $m(1)$ and $m(0)$ give the overall proportions of positive and negative tests that can be anticipated. These might be useful for logistic purposes if, say, the positive tests were to be followed up with more elaborate testing; 32.5% of those initially tested would require the more elaborate testing.

In discrete situations, such as Example 3, the formula for $\pi(\theta|x)$ is commonly known as Bayes's theorem, and was discovered by Bayes (1763). The typical phrasing of Bayes's theorem is in terms of disjoint events $A_1, A_2, \ldots, A_n$, whose union has probability one (i.e., one of the A_i is certain to occur). Prior probabilities $P(A_i)$, for the events, are assumed known. An event B occurs, for which $P(B|A_i)$ (the conditional probability of B given A_i) is known for each A_i. Bayes's theorem then states that

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}.$$

These probabilities reflect our revised opinions about the A_i, in light of the knowledge that B has occurred. Replacing A_i by θ_i and B by x, shows the equivalence of this to the formula for the posterior distribution.

EXAMPLE 4. In airplanes there is a warning light that goes on if the landing gear fails to fully extend. Sometimes the warning light goes on even when the landing gear has extended. Let A_1 denote the event "the landing gear extends" and A_2 denote the event "the landing gear fails to extend." (Note that A_1 and A_2 are disjoint and one of the two must occur.) Let B be the event that the warning light goes on. It is known that the light will go on with probability 0.999 if A_2 occurs (i.e., $P(B|A_2) = 0.999$), while $P(B|A_1) = 0.005$. Records show that $P(A_1) = 0.997$ and $P(A_2) = 0.003$. It is desired to

determine the probability that the landing gear has extended, even though the warning light has gone on. This is simply $P(A_1|B)$, and from Bayes's theorem is given by

$$P(A_1|B) = \frac{(0.005)(0.997)}{(0.005)(0.997)+(0.999)(0.003)} = 0.62.$$

4.2.2. Conjugate Families

In general, $m(x)$ and $\pi(\theta|x)$ are not easily calculable. If, for example, X is $\mathcal{N}(\theta, \sigma^2)$ and θ is $\mathcal{C}(\mu, \beta)$, then $\pi(\theta|x)$ can only be evaluated numerically. A large part of the Bayesian literature is devoted to finding prior distributions for which $\pi(\theta|x)$ can be easily calculated. These are the so called *conjugate priors*, and were developed extensively in Raiffa and Schlaifer (1961).

Definition 1. Let $\mathcal{F}$ denote the class of density functions $f(x|\theta)$ (indexed by θ). A class $\mathcal{P}$ of prior distributions is said to be a *conjugate family* for $\mathcal{F}$ if $\pi(\theta|x)$ is in the class $\mathcal{P}$ for all $x \in \mathcal{X}$ and $\pi \in \mathcal{P}$.

Example 1 shows that the class of normal priors is a conjugate family for the class of normal (sample) densities. (If X has a normal density and θ has a normal prior, then the posterior density of θ given x is also normal.)

For a given class of densities $\mathcal{F}$, a conjugate family can frequently be determined by examining the likelihood functions $l_x(\theta) = f(x|\theta)$, and choosing, as a conjugate family, the class of distributions with the same functional form as these likelihood functions. The resulting priors are frequently called *natural* conjugate priors.

When dealing with conjugate priors, there is generally no need to explicitly calculate $m(x)$. The reason is that, since $\pi(\theta|x) = h(x, \theta)/m(x)$, the factors involving θ in $\pi(\theta|x)$ must be the same as the factors involving θ in $h(x, \theta)$. Hence it is only necessary to look at the factors involving θ in $h(x, \theta)$, and see if these can be recognized as belonging to a particular distribution. If so, $\pi(\theta|x)$ is that distribution. The marginal density $m(x)$ can then be determined, if desired, by dividing $h(x, \theta)$ by $\pi(\theta|x)$. An example of the above ideas follows.

EXAMPLE 5. Assume $\mathbf{X} = (X_1, \ldots, X_n)$ is a sample from a Poisson distribution. Thus $X_i \sim \mathcal{P}(\theta)$, $i = 1, \ldots, n$, and

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \left[\frac{\theta^{x_i} e^{-\theta}}{x_i!}\right] = \frac{\theta^{n\bar{x}} e^{-n\theta}}{\prod_{i=1}^{n} [x_i!]}.$$

Here, $\mathcal{F}$ is the class of all such densities. Observing that the likelihood function for such densities resembles a gamma density, a plausible guess for a conjugate family of prior distributions is the class of gamma distribu-

tions. Thus assume $\theta \sim \mathcal{G}(\alpha, \beta)$, and observe that

$$h(\mathbf{x}, \theta) = f(\mathbf{x}|\theta)\pi(\theta) = \frac{e^{-n\theta}\theta^{n\bar{x}}}{\prod_{i=1}^{n}[x_i!]} \cdot \frac{\theta^{\alpha-1}e^{-\theta/\beta}I_{(0,\infty)}(\theta)}{\Gamma(\alpha)\beta^{\alpha}}$$
$$= \frac{e^{-\theta(n+1/\beta)}\theta^{(n\bar{x}+\alpha-1)}I_{(0,\infty)}(\theta)}{\Gamma(\alpha)\beta^{\alpha}\prod_{i=1}^{n}[x_i!]}.$$

The factors involving θ in this last expression are clearly recognizable as belonging to a $\mathcal{G}(n\bar{x}+\alpha, [n+1/\beta]^{-1})$ distribution. This must then be $\pi(\theta|\mathbf{x})$. Since this posterior is a gamma distribution, it follows that the class of gamma distributions is indeed a (natural) conjugate family for $\mathcal{F}$.

In this example, $m(\mathbf{x})$ can be determined by dividing $h(\mathbf{x}, \theta)$ by $\pi(\theta|\mathbf{x})$ and cancelling factors involving θ. The result is

$$m(\mathbf{x}) = \frac{h(\mathbf{x}, \theta)}{\pi(\theta|\mathbf{x})} = \frac{(\Gamma(\alpha)\beta^{\alpha}\prod_{i=1}^{n}[x_i!])^{-1}}{\{\Gamma(\alpha+n\bar{x})[n+1/\beta]^{-(\alpha+n\bar{x})}\}^{-1}}.$$

Besides providing for easy calculation of $\pi(\theta|x)$, conjugate priors have the intuitively appealing feature of allowing one to begin with a certain functional form for the prior and end up with a posterior of the same functional form, but with parameters updated by the sample information. In Example 1, for instance, the prior mean μ gets updated by x to become the posterior mean

$$\mu(x) = \frac{\tau^2}{\sigma^2+\tau^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\mu.$$

The prior variance τ^2 is combined with the data variance σ^2 to give the posterior variance

$$\rho^{-1} = \left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}.$$

This updating of parameters provides an easy way of seeing the effect of prior and sample information. It also makes useful the concept of equivalent sample size discussed in Section 3.2.

These attractive properties of conjugate priors are, however, only of secondary importance compared to the basic question of whether or not a conjugate prior can be chosen which gives a reasonable approximation to the true prior. Many Bayesians say this can be done, arguing for example that, in dealing with a normal mean, the class of $\mathcal{N}(\mu, \tau^2)$ priors is rich enough to include approximations to most reasonable priors. Unfortunately, in Section 4.7 we will encounter reasons to doubt this belief, observing that using a normal prior can sometimes result in unappealing conclusions. Most of the examples and problems in this chapter will make use of conjugate priors, however, due to the resulting ease in calculations. And, at least for initial analyses, conjugate priors such as above can be quite useful in practice.

There do exist conjugate families of priors other than natural conjugate priors. One trivial such class is the class of *all* distributions. (The posterior is certainly in this class.) A more interesting example is the class of finite mixtures of natural conjugate priors. In Example 1, for instance, the class of all priors of the form (for fixed m, say)

$$\pi(\theta) = \sum_{i=1}^{m} w_i \pi_i(\theta), \tag{4.1}$$

where $\sum_{i=1}^{m} w_i = 1$ (all $w_i \geq 0$) and the π_i are normally distributed, can be shown to be a conjugate class—the proof will be left for an exercise. The use of such mixtures allows approximations to bimodal and more complicated subjective prior distributions, and yet preserves much of the calculational simplicity of natural conjugate priors. Development and uses of such mixture conjugate classes can be found in Dalal and Hall (1983) and Diaconis and Ylvisaker (1984). Jewell (1983) generalizes natural conjugate priors in a different direction.

4.2.3. Improper Priors

The analysis leading to the posterior distribution can formally be carried out even if $\pi(\theta)$ is an improper prior. For example, if $X \sim \mathcal{N}(\theta, \sigma^2)$ (σ^2 known) and the noninformative prior $\pi(\theta) = 1$ is used, then

$$h(x, \theta) = f(x|\theta)\pi(\theta) = f(x|\theta),$$

$$m(x) = \int_{-\infty}^{\infty} f(x|\theta)\, d\theta = (2\pi)^{-1/2}\sigma^{-1} \int_{-\infty}^{\infty} \exp\left\{\frac{-(x-\theta)^2}{2\sigma^2}\right\} d\theta = 1,$$

and

$$\pi(\theta|x) = \frac{h(x, \theta)}{m(x)} = (2\pi)^{-1/2}\sigma^{-1} \exp\left\{\frac{-(\theta - x)^2}{2\sigma^2}\right\}.$$

Hence the posterior distribution of θ given x is $\mathcal{N}(x, \sigma^2)$. Of course, this posterior distribution cannot rigorously be considered to be the conditional distribution of θ given x, but various heuristic arguments can be given to support such an interpretation. For example, taking a suitable sequence of finite priors $\pi_n(\theta)$, which converge to $\pi(\theta)$ as $n \to \infty$, it can be shown that the corresponding posteriors, $\pi_n(\theta|x)$, converge to $\pi(\theta|x)$. Other arguments using finitely additive probability measures can be given to support the informal interpretation of $\pi(\theta|x)$ as the conditional density of θ given x.

4.3. Bayesian Inference

Inference problems concerning θ can easily be dealt with using Bayesian analysis. The idea is that, since the posterior distribution supposedly contains all the available information about θ (both sample and prior informa-

tion), any inferences concerning θ should consist solely of features of this distribution. Several justifications for this view were mentioned in Sections 1.6 and 4.1.

Statistical inference is not the main subject of this book, so we will only indicate the basic elements of the Bayesian approach to inference. (For more thorough treatments of Bayesian inference, see Jeffreys (1961), Zellner (1971), and Box and Tiao (1973).) Of course, inference can also be treated as a decision-theoretic problem (see Subsection 2.4.3), and in Subsection 4.4.4 we briefly illustrate this possibility.

Statistical inference is often associated with a desire for "objectivity." This issue was discussed in Section 3.7, wherein it was briefly argued that the most reasonable method of "attempting" to be objective is to perform a Bayesian analysis with a noninformative prior. We will, therefore, especially emphasize use of noninformative priors in this section.

Some Bayesians maintain that inference should ideally consist of simply reporting the *entire* posterior distribution $\pi(\theta|x)$ (maybe for a noninformative prior). We do not disagree in principle, since from the posterior one can derive any feature of interest, and indeed a visual inspection of the graph of the posterior will often provide the best insight concerning θ (at least in low dimensions). More standard uses of the posterior are still helpful, however (especially for outside consumption), and will be discussed in this section. Note, once again, that all measures that will be discussed are conditional in nature, since they depend only on the posterior distribution (which involves the experiment only through the observed likelihood function).

4.3.1. Estimation

The simplest inferential use of the posterior distribution is to report a point estimate for θ, with an associated measure of accuracy.

I. *Point Estimates*

To estimate θ, a number of classical techniques can be applied to the posterior distribution. The most common classical technique is maximum likelihood estimation, which chooses, as the estimate of θ, the value $\hat{\theta}$ which maximizes the likelihood function $l(\theta)=f(x|\theta)$. The analogous Bayesian estimate is defined as follows.

Definition 2. The *generalized maximum likelihood* estimate of θ is the largest mode, $\hat{\theta}$, of $\pi(\theta|x)$ (i.e., the value $\hat{\theta}$ which maximizes $\pi(\theta|x)$, considered as a function of θ).

Obviously $\hat{\theta}$ has the interpretation of being the "most likely" value of θ, given the prior and the sample x.

EXAMPLE 1 (continued). When f and π are normal densities, the posterior density was seen to be $\mathcal{N}(\mu(x), \rho^{-1})$. A normal density achieves its maximum value at the mean, so the generalized maximum likelihood estimate of θ in this situation is

$$\hat{\theta} = \mu(x) = \frac{\sigma^2 \mu}{\sigma^2 + \tau^2} + \frac{\tau^2 x}{\sigma^2 + \tau^2}.$$

EXAMPLE 6. Assume

$$f(x|\theta) = e^{-(x-\theta)} I_{(\theta,\infty)}(x),$$

and $\pi(\theta) = [\pi(1+\theta^2)]^{-1}$. Then

$$\pi(\theta|x) = \frac{e^{-(x-\theta)} I_{(\theta,\infty)}(x)}{m(x)(1+\theta^2)\pi}.$$

To find the $\hat{\theta}$ maximizing this quantity, note first that only $\theta \le x$ need be considered. (If $\theta > x$, then $I_{(\theta,\infty)}(x) = 0$ and $\pi(\theta|x) = 0$.) For such θ,

$$\begin{aligned} \frac{d}{d\theta}\pi(\theta|x) &= \frac{e^{-x}}{m(x)\pi}\left[\frac{e^{\theta}}{1+\theta^2} - \frac{2\theta e^{\theta}}{(1+\theta^2)^2}\right] \\ &= \frac{e^{-x}}{m(x)\pi}\,\frac{e^{\theta}(\theta-1)^2}{(1+\theta^2)^2}. \end{aligned}$$

Since this derivative is always positive, $\pi(\theta|x)$ is increasing for $\theta \le x$. It follows that $\pi(\theta|x)$ is maximized at $\hat{\theta} = x$, which is thus the generalized maximum likelihood estimate of θ.

Other common Bayesian estimates of θ include the mean and the median of $\pi(\theta|x)$. In Example 1 these clearly coincide with $\mu(x)$, the mode. In Example 6, however, the mean and median will differ from the mode, and must be calculated numerically.

The mean and median (and mode) are relatively easy to find when the prior, and hence posterior, are from a conjugate family of distributions. In Example 5, for instance, $\pi(\theta|x)$ is $\mathcal{G}(\alpha + n\bar{x}, [n + 1/\beta]^{-1})$, which has mean $[\alpha + n\bar{x}]/[n + 1/\beta]$. The median can be found using tables of the gamma distribution.

The mean and median of the posterior are frequently better estimates of θ than the mode. It is probably worthwhile to calculate and compare all three in a Bayesian study, especially with regard to their robustness to changes in the prior.

As mentioned at the beginning of this section, Bayesian inference using a noninformative prior is often an easy and reasonable method of analysis. The following example gives a simple demonstration of this in an estimation problem.

EXAMPLE 7. A not uncommon situation is to observe $X \sim \mathcal{N}(\theta, \sigma^2)$ (for simplicity assume σ^2 is known), where θ is a measure of some clearly positive quantity. The classical estimate of θ is x, which is clearly unsuitable when x turns out to be negative. A reasonable way of developing an alternative estimate (assuming no specific prior knowledge is available) is to use the noninformative prior $\pi(\theta) = I_{(0,\infty)}(\theta)$ (since θ is a location parameter). The resulting posterior is

$$\pi(\theta \mid x) = \frac{\exp\{-(\theta - x)^2/2\sigma^2\} I_{(0,\infty)}(\theta)}{\int_0^\infty \exp\{-(\theta - x)^2/2\sigma^2\} d\theta}.$$

Making the change of variables $\eta = (\theta - x)/\sigma$, the mean of the posterior can be seen to be

$$\begin{aligned} E^{\pi(\theta|x)}[\theta] &= \frac{\int_0^\infty \theta \exp\{-(\theta - x)^2/2\sigma^2\} d\theta}{\int_0^\infty \exp\{-(\theta - x)^2/2\sigma^2\} d\theta} \\ &= \frac{\int_{-(x/\sigma)}^\infty (\sigma\eta + x)\exp\{-\eta^2/2\}\sigma \, d\eta}{\int_{-(x/\sigma)}^\infty \exp\{-\eta^2/2\}\sigma \, d\eta} \\ &= x + \frac{(2\pi)^{-1/2}\sigma \int_{-(x/\sigma)}^\infty \eta \exp\{-\eta^2/2\} d\eta}{1 - \Phi(-x/\sigma)} \\ &= x + \frac{(2\pi)^{-1/2}\sigma \exp\{-x^2/2\sigma^2\}}{1 - \Phi(-x/\sigma)}, \end{aligned}$$

where Φ is the standard normal c.d.f. This estimate of θ is quite simple and easy to use.

Example 7 is a simple case of a common type of statistical problem that is quite difficult to handle classically, namely the situation of a restricted parameter space. Restricted parameter spaces can be of many types. The situation in Example 7, where the parameter is known to be positive (or, in higher dimensional settings, where the signs of all coordinates of the parameter are known), is one important case. Among the practical problems which involve such a parameter space are "variance component" problems (cf. Hill (1965, 1977)). Another typical occurrence of restricted parameter spaces is when $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ and the θ_i are known to be ordered in some fashion (cf. Exercise 24). Much more complicated scenarios can also occur, of this same form that θ is known to be in $\Theta' \subset \Theta$.

As in Example 7, a restricted parameter space Θ' can be easily handled using noninformative prior Bayesian analysis. Simply let $\pi(\theta) = \pi_0(\theta) I_{\Theta'}(\theta)$ (recall that $I_{\Theta'}(\theta) = 1$ if $\theta \in \Theta'$ and equals zero, otherwise), where $\pi_0(\theta)$ is an appropriate noninformative prior for unrestricted θ, and proceed. There may be calculational difficulties in determining, say, the posterior mean, but there is no problem conceptually. (See Section 4.9 for discussion of some aspects of Bayesian calculation.)

Sometimes, one may be dealing with situations in which the parameters are not necessarily strictly ordered, but in which they tend to be ordered with "high probability." Such situations are virtually impossible to handle classically, but prior distributions incorporating such "stochastic ordering" can be constructed and used. See Proschan and Singpurwalla (1979, 1980) and Jewell (1979) for development.

II. *Estimation Error*

When presenting a statistical estimate, it is usually necessary to indicate the accuracy of the estimate. The customary Bayesian measure of the accuracy of an estimate is (in one dimension) the posterior variance of the estimate, which is defined as follows.

Definition 3. If θ is a real valued parameter with posterior distribution $\pi(\theta|x)$, and δ is the estimate of θ, then the *posterior variance of* δ is

$$V_\delta^\pi(x) \equiv E^{\pi(\theta|x)}[(\theta-\delta)^2].$$

When δ is the posterior mean

$$\mu^\pi(x) \equiv E^{\pi(\theta|x)}[\theta],$$

then $V^\pi(x) \equiv V_{\mu^\pi}^\pi(x)$ will be called simply the *posterior variance* (and it is indeed the variance of θ for the distribution $\pi(\theta|x)$). The *posterior standard deviation* is $\sqrt{V^\pi(x)}$.

It is customary to use $\sqrt{V_\delta^\pi(x)}$ as the "standard error" of the estimate δ. For calculational purposes, it is often helpful to note that

$$\begin{aligned} V_\delta^\pi(x) &= E^{\pi(\theta|x)}[(\theta-\delta)^2] = E[(\theta-\mu^\pi(x)+\mu^\pi(x)-\delta)^2] \\ &= E[(\theta-\mu^\pi(x))^2]+E[2(\theta-\mu^\pi(x))(\mu^\pi(x)-\delta)]+E[(\mu^\pi(x)-\delta)^2] \\ &= V^\pi(x)+2(\mu^\pi(x)-\delta)(E[\theta]-\mu^\pi(x))+(\mu^\pi(x)-\delta)^2 \\ &= V^\pi(x)+(\mu^\pi(x)-\delta)^2. \end{aligned} \tag{4.2}$$

Observe from (4.2) that the posterior mean, $\mu^\pi(x)$, minimizes $V_\delta^\pi(x)$ (over all δ), and hence is the estimate with smallest standard error. For this reason, it is customary to use $\mu^\pi(x)$ as the estimate for θ and report $\sqrt{V^\pi(x)}$ as the standard error.

EXAMPLE 1 (continued). It is clear that

$$V^\pi(x) = \rho^{-1} = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}.$$

Thus, in the example of intelligence testing, the child with $x=115$ would be reported as having as estimated IQ of $\mu^\pi(115)=110.39$, with associated standard error $\sqrt{V^\pi(115)}=\sqrt{69.23}=8.32$.

The classical estimate of θ for the general normal problem is just $\delta = x$, which (using (4.2)) has

$$\begin{aligned} V_\delta^\pi(x) &= V^\pi(x) + (\mu^\pi(x) - x)^2 \\ &= V^\pi(x) + \left(\frac{\sigma^2\mu}{\sigma^2+\tau^2} + \frac{\tau^2 x}{\sigma^2+\tau^2} - x\right)^2 \\ &= V^\pi(x) + \frac{\sigma^4}{(\sigma^2+\tau^2)^2}(\mu - x)^2. \end{aligned} \tag{4.3}$$

Note that, in the IQ example, the classical estimate $\delta = x = 115$ would have standard error (with respect to $\pi(\theta|x)$) of (using (4.2))

$$\sqrt{V_{115}^\pi(115)} = [69.23 + (110.39 - 115)^2]^{1/2} = \sqrt{90.48} = 9.49.$$

Of course, the *classical standard error* of $\delta = x$ is σ, the sample standard deviation. It is interesting to observe from (4.3) that, if $(\mu - x)^2 > (\sigma^2 + \tau^2)$, then $V_\delta^\pi(x) > \sigma^2$, so that a Bayesian who believed in π would feel *dishonest* in reporting the smaller number, σ, as the standard error.

In the above example, a Bayesian would estimate θ by $\mu^\pi(x)$ with standard error $\sqrt{V^\pi(x)} = \rho^{-1/2}$, which is *less than* σ, the classical standard error of the classical estimate $\delta = x$. This is usually (but not always) true for Bayesian estimation: because the Bayesian is using prior information as well as sample information to estimate θ, his estimate will typically have a (Bayesian) standard error that is smaller than the standard error that a classical statistician would report for the classical estimate. (Of course, someone who did not believe that π was a reasonable prior would view the smaller Bayesian standard error as misleading, so we are not claiming this to be an *advantage* of the Bayesian approach.) The major exceptions to this pattern of smaller Bayesian standard error occur when noninformative priors are used, and for certain flat-tailed priors (cf. O'Hagan (1981)).

EXAMPLE 8. In Subsection 4.2.3 it was shown that, if $X \sim \mathcal{N}(\theta, \sigma^2)$ (σ^2 known) and the noninformative prior $\pi(\theta) = 1$ is used, then the posterior distribution of θ given x is $\mathcal{N}(x, \sigma^2)$. Hence the posterior mean is $\mu^\pi(x) = x$, and the posterior variance and standard deviation are σ^2 and σ, respectively.

Example 8 is the first instance of what will be seen to be a common phenomenon: the report (here estimate and standard error thereof) from a noninformative prior Bayesian analysis is often *formally* the same as the usual classical report. The *interpretations* of the two reports differ, but the numbers are formally the same. Many Bayesians maintain (and with considerable justification) that classical statistics has prospered only because, in so many standard situations (such as that of Example 8), the classical numbers reported can be given a sensible noninformative prior Bayesian interpretation (which also coincides with the meaning that nonsophisticates

ascribe to the numbers, see Section 4.1). There are many situations in which the two reports do differ, however (and we will encounter several), and the classical report almost invariably suffers in comparison. See Pratt (1965) for further discussion of a number of these issues.

Another point (and this is one of the operational advantages of Bayesian analysis that was alluded to in Section 4.1) is that the calculation of $V^\pi(x)$ is rarely much more difficult than that of $\mu^\pi(x)$. When the calculation is a numerical calculation, there will be essentially no difference in difficulty. And when $\mu^\pi(x)$ can be written in a simple closed form, it is usually also possible to find a reasonably simple closed form for $V^\pi(x)$.

EXAMPLE 7 (continued). Writing

$$\psi(x) = \frac{(2\pi)^{-1/2}\sigma \exp\{-x^2/(2\sigma^2)\}}{1-\Phi(-x/\sigma)}.$$

it was shown earlier that

$$\mu^\pi(x) = x + \psi(x).$$

To calculate $V^\pi(x)$ easily, note from (4.2) (choosing $\delta = x$) that

$$\begin{aligned} V^\pi(x) &= V_x^\pi(x) - (\mu^\pi(x) - x)^2 \\ &= V_x^\pi(x) - [\psi(x)]^2, \end{aligned}$$

and that an integration by parts (in the numerator below) gives

$$\begin{aligned} V_x^\pi(x) &= E^{\pi(\theta|x)}[(\theta - x)^2] \\ &= \frac{\int_0^\infty (\theta-x)^2 \exp\{-(\theta-x)^2/2\sigma^2\}d\theta}{\int_0^\infty \exp\{-(\theta-x)^2/2\sigma^2\}d\theta} \\ &= \frac{-\sigma^2 x \exp\{-x^2/2\sigma^2\} + \sigma^2 \int_0^\infty \exp\{-(\theta-x)^2/2\sigma^2\}d\theta}{\int_0^\infty \exp\{-(\theta-x)^2/2\sigma^2\}d\theta} \\ &= -x\psi(x) + \sigma^2. \end{aligned}$$

Hence

$$V^\pi(x) = \sigma^2 - [x + \psi(x)]\psi(x). \tag{4.4}$$

The typically straightforward Bayesian calculation of standard error compares favorably (from an operational perspective) with the classical approach. A frequentist must propose an estimator $\delta(x)$, and calculate a "standard error" such as $\sqrt{\bar{V}}$ where, say,

$$\bar{V} = \sup_\theta E_\theta^X[(\delta(X) - \theta)^2]$$

(the maximum "mean squared error" of δ). This calculation can be difficult when no reasonable "unbiased" estimators are available. And sometimes $\bar{V}$ will be very large (even infinite) for all δ (because of the sup over θ), in which case the report of $\sqrt{\bar{V}}$ as the standard error seems highly questionable. Even when $\bar{V}$ appears to be reasonable, its use can be counter-intuitive in some respects, as the following example shows.

EXAMPLE 7 (continued). A natural frequentist estimator for θ is $\delta(x) = \max\{x, 0\}$, and it can be shown that $\bar{V} = \sigma^2$. Thus one can report $\delta(x)$, together with the frequentist standard error $\sqrt{\bar{V}} = \sigma$. The counter-intuitive feature of this report is that, were it not known that $\theta > 0$, a frequentist would likely use $\delta(x) = x$, which also has $\bar{V} = \sigma^2$. Thus the same standard error would be reported in either case, while (intuitively) it would seem that the knowledge that $\theta > 0$ should result in a smaller reported standard error (at least when x is near zero). Note that the Bayesian analysis, given earlier, does reflect the benefits of this added knowledge. Indeed, $V^{\pi}(x)$ is increasing in x, with $V^{\pi}(0) = (1 - 2/\pi)\sigma^2$ and $V^{\pi}(\infty) = \sigma^2$. Thus for small x, the knowledge that $\theta > 0$ is being used to report a substantially smaller standard error than that of the similar analysis in Example 8. It can be shown (using the Cramér-Rao inequality) that *no* estimator will yield frequentist standard error smaller than σ, so the frequentist approach can never take advantage of the additional knowledge. (Conceivably, the estimated frequentist approach mentioned in Subsection 1.6.3 could take advantage of the additional information, but implementation of this approach is very hard.)

III. *Multivariate Estimation*

Bayesian estimation of a vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ is also straightforward. The generalized maximum likelihood estimate (the posterior mode) is often a reasonable estimate, although existence and uniqueness difficulties are more likely to be encountered in the multivariate case. The posterior mean

$$\boldsymbol{\mu}^{\pi}(x) = (\mu_1^{\pi}(x), \ldots, \mu_p^{\pi}(x))' = E^{\pi(\boldsymbol{\theta}|x)}[\boldsymbol{\theta}]$$

is a very attractive Bayesian estimate (providing it can be calculated, see Section 4.9), and its accuracy can be described by the *posterior covariance matrix*

$$\mathbf{V}^{\pi}(x) = E^{\pi(\boldsymbol{\theta}|x)}[(\boldsymbol{\theta} - \boldsymbol{\mu}^{\pi}(x))(\boldsymbol{\theta} - \boldsymbol{\mu}^{\pi}(x))']. \tag{4.5}$$

(For instance, the standard error of the estimate $\mu_i^{\pi}(x)$ of θ_i would be $\sqrt{V_{ii}^{\pi}(x)}$, where $V_{ii}^{\pi}(x)$ is the (i, i) element of $\mathbf{V}^{\pi}(x)$; more sophisticated uses of $\mathbf{V}^{\pi}$ are discussed in the next subsection.)

The analog of (4.2), for a general estimate $\boldsymbol{\delta}$ of $\boldsymbol{\theta}$, can be shown to be

$$\begin{aligned} \mathbf{V}_{\boldsymbol{\delta}}^{\pi}(x) &= E^{\pi(\boldsymbol{\theta}|x)}[(\boldsymbol{\theta} - \boldsymbol{\delta})(\boldsymbol{\theta} - \boldsymbol{\delta})'] \\ &= \mathbf{V}^{\pi}(x) + (\boldsymbol{\mu}^{\pi}(x) - \boldsymbol{\delta})(\boldsymbol{\mu}^{\pi}(x) - \boldsymbol{\delta})'. \end{aligned} \tag{4.6}$$

Again, it is clear that the posterior mean "minimizes" $\mathbf{V}_{\boldsymbol{\delta}}^{\pi}(x)$.

EXAMPLE 9. Suppose $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and $\pi(\boldsymbol{\theta})$ is a $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{A})$ density. (Here $\boldsymbol{\mu}$ is a known p-vector, and $\boldsymbol{\Sigma}$ and $\mathbf{A}$ are known $(p \times p)$ positive definite matrices.) It will be left for the exercises to show that $\pi(\boldsymbol{\theta}|\mathbf{x})$ is a

$\mathcal{N}_p(\boldsymbol{\mu}^\pi(\mathbf{x}), \mathbf{V}^\pi(\mathbf{x}))$ density, where the posterior mean is given by

$$\boldsymbol{\mu}^\pi(\mathbf{x}) = \mathbf{x} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \mathbf{A})^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{4.7}$$

and the posterior covariance matrix by

$$\begin{aligned} \mathbf{V}^\pi(\mathbf{x}) &= (\mathbf{A}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} \\ &= \boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\mathbf{A} + \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}. \end{aligned} \tag{4.8}$$

More sophisticated multivariate Bayesian analyses will be considered in Sections 4.5 and 4.6 and Subsection 4.7.10.

4.3.2. Credible Sets

Another common approach to inference is to present a confidence set for θ. The Bayesian analog of a classical confidence set is called a credible set, and is defined as follows:

Definition 4. A $100(1-\alpha)\%$ *credible set* for θ is a subset C of Θ such that

$$1-\alpha \leq P(C|x) = \int_C dF^{\pi(\theta|x)}(\theta) = \begin{cases} \int_C \pi(\theta|x)\,d\theta & \text{(continuous case)}, \\ \sum_{\theta \in C} \pi(\theta|x) & \text{(discrete case)}. \end{cases}$$

Since the posterior distribution is an actual probability distribution on Θ, one can speak meaningfully (though usually subjectively) of the probability that θ is in C. This is in contrast to classical confidence procedures, which can only be interpreted in terms of coverage probability (the probability that the random X will be such that the confidence set $C(X)$ contains θ). Discussions of this difference were given in Sections 1.6 and 4.1.

In choosing a credible set for θ, it is usually desirable to try to minimize its size. To do this, one should include in the set only those points with the largest posterior density, i.e., the "most likely" values of θ. (Actually, this minimizes specifically the volume of the credible set. It may be desirable to minimize other types of size, as will be seen shortly.)

Definition 5. The $100(1-\alpha)\%$ HPD *credible set* for θ (HPD stands for highest posterior density), is the subset C of Θ of the form

$$C = \{\theta \in \Theta\colon \pi(\theta|x) \geq k(\alpha)\},$$

where $k(\alpha)$ is the largest constant such that

$$P(C|x) \geq 1-\alpha.$$

EXAMPLE 1 (continued). Since the posterior density of θ given x is $\mathcal{N}(\mu(x), \rho^{-1})$, which is unimodal and symmetric about $\mu(x)$, it is clear that

the $100(1-\alpha)\%$ HPD credible set is given by

$$C = \left(\mu(x) + z\left(\frac{\alpha}{2}\right)\rho^{-1/2}, \mu(x) - z\left(\frac{\alpha}{2}\right)\rho^{-1/2}\right),$$

where $z(\alpha)$ is the α-fractile of a $\mathcal{N}(0, 1)$ distribution.

In the IQ example, where the child who scores 115 on the intelligence test has a $\mathcal{N}(110.39, 69.23)$ posterior distribution for θ, it follows that a 95% HPD credible set for θ is

$$(110.39 + (-1.96)(69.23)^{1/2}, 110.39 + (1.96)(69.23)^{1/2}) = (94.08, 126.70).$$

Note that since a random test score X is $\mathcal{N}(\theta, 100)$, the classical 95% confidence interval for θ is

$$(115 - (1.96)(10), 115 + (1.96)(10)) = (95.4, 134.6).$$

EXAMPLE 8 (continued). Since the posterior distribution of θ given x is $\mathcal{N}(x, \sigma^2)$, it follows that the $100(1-\alpha)\%$ HPD credible set for θ is

$$C = \left(x + z\left(\frac{\alpha}{2}\right)\sigma, x - z\left(\frac{\alpha}{2}\right)\sigma\right).$$

This is exactly the same as the classical confidence set for θ, and is another instance of the frequent *formal* similarity of classical and noninformative prior Bayesian answers.

Bayesian credible sets are usually much easier to calculate than their classical counterparts, particularly in situations where simple sufficient statistics do not exist. The following example illustrates this.

EXAMPLE 10. Assume $\mathbf{X} = (X_1, \ldots, X_n)$ is an i.i.d. sample from a $\mathcal{C}(\theta, 1)$ distribution, and that $\theta > 0$. Suppose that a noninformative prior Bayesian analysis is desired. Since θ is a (restricted) location parameter, a reasonable noninformative prior would be $\pi(\theta) = 1$ (on $\theta > 0$). The posterior density of θ given $\mathbf{x} = (x_1, \ldots, x_n)$ is then given (on $\theta > 0$) by

$$\pi(\theta|\mathbf{x}) = \frac{\prod_{i=1}^{n}[1 + (\theta - x_i)^2]^{-1}}{\int_0^\infty \prod_{i=1}^{n}[1 + (\theta - x_i)^2]^{-1} d\theta}. \tag{4.9}$$

While this is not an overly attractive posterior to work with, finding a $100(1-\alpha)\%$ HPD credible set on a computer is a relatively simple undertaking. For instance, if $n = 5$ with $\mathbf{x} = (4.0, 5.5, 7.5, 4.5, 3.0)$, then the resulting 95% HPD credible set is the interval (3.10, 6.06). In contrast, it is not at all clear how to develop a good classical confidence procedure for this problem. Classical confidence procedures are usually developed with a variety of tricks which do not have general applicability.

The general idea behind numerical calculation of the HPD credible set, in a situation where $\pi(\theta|x)$ is continuous in θ (such as Example 10), is to set up a program along the following lines:

(i) Create a subroutine which, for given k, finds all solutions to the equation $\pi(\theta|x) = k$. The set $C(k) = \{\theta: \pi(\theta|x) \geq k\}$ can typically be fairly easily constructed from these solutions. For instance, if Θ is an infinite interval in R^1 (as in Example 10) and only two solutions, $\theta_1(k)$ and $\theta_2(k)$, are found, then $C(k) = (\theta_1(k), \theta_2(k))$.

(ii) Create a subroutine which calculates

$$P^{\pi(\theta|x)}(C(k)) = \int_{C(k)} \pi(\theta|x)\, d\theta.$$

(iii) Numerically solve the equation

$$P^{\pi(\theta|x)}(C(k)) = 1 - \alpha,$$

calling on the above two subroutines as k varies. (There are, undoubtedly, much more efficient ways to actually write such a program; the main purpose of this general outline has been to simply provide some insight into the process.)

It can happen that an HPD credible set looks unusual. For instance, in Example 10 it could happen that the HPD credible set consists of several disjoint intervals. In such situations, one could abandon the HPD criterion and insist on connected credible sets, but there are very good reasons *not* to do so. Disjoint intervals often occur when there is "clashing" information (maybe the prior says one thing, and the data another), and such clashes are usually important to recognize. Natural conjugate priors typically are unimodal and yield unimodal posteriors, and hence will tend to *mask* these clashes. (This is one of a number of reasons that we will see for being wary of conjugate priors.)

An often useful approximation to an HPD credible set can be achieved through use of the normal approximation to the posterior. It can be shown (see Subsection 4.7.8) that, for large sample sizes, the posterior distribution will be approximately normal. Even for small samples, a normal likelihood function will usually yield a roughly normal posterior, so the approximation can have uses even in small samples. The attractions of using the normal approximation include calculational simplicity and the fact that the ensuing credible sets will have a standard form (for, say, consumption by people who are only comfortable with sets having a familiar shape).

The most natural normal approximation to use, in the univariate case, is to approximate $\pi(\theta|x)$ by a $\mathcal{N}(\mu^\pi(x), V^\pi(x))$ distribution, where $\mu^\pi(x)$ and $V^\pi(x)$ are given in Definition 3. (For large samples, $\mu^\pi(x)$ and $V^\pi(x)$ can be estimated as discussed in Subsection 4.7.8.) The corresponding

approximate $100(1-\alpha)\%$ HPD credible set is then

$$C = \left(\mu^{\pi}(x) + z\left(\frac{\alpha}{2}\right)\sqrt{V^{\pi}(x)}, \mu^{\pi}(x) - z\left(\frac{\alpha}{2}\right)\sqrt{V^{\pi}(x)} \right). \tag{4.10}$$

EXAMPLE 10 (continued). The posterior density in (4.9) is clearly nonnormal and, with only five Cauchy observations, one might imagine that the normal approximation would be somewhat inaccurate. The posterior can be seen to be unimodal, however, and the normal approximation turns out to be excellent out to the 2.5% and 97.5% tails. A numerical calculation gave $\mu^{\pi}(\mathbf{x}) = 4.55$, $V^{\pi}(\mathbf{x}) = 0.562$, and actual and approximate percentiles as follows:

Table 4.1. Actual and Approximate Posterior Percentiles.

α	2.5	25	50	75	97.5
αth percentile of $\pi(\theta\|x)$	3.17	4.07	4.52	5.00	6.15
αth percentile of $\mathcal{N}(\mu^{\pi}, V^{\pi})$	3.08	4.05	4.55	5.06	6.02

In the extreme tails the approximation became quite inaccurate, and this indicates an important general limitation of the use of such approximations.

The approximate 95% HDP credible set (using (4.10)) is $C = (3.08, 6.02)$, which is very close to the actual 95% HPD credible set, (3.10, 6.06), calculated earlier. Also, this approximate C has *actual* posterior probability (under (4.9)) of 0.948, which is extremely close to its nominal probability. Similarly, the approximate 90% HPD credible set is (3.32, 5.78), and has *actual* posterior probability of 0.906.

Multivariate Case

For multivariate $\boldsymbol{\theta}$, the definition of an HPD credible set remains the same.

EXAMPLE 9 (continued). The posterior density is $\mathcal{N}_p(\boldsymbol{\mu}^{\pi}(\mathbf{x}), \mathbf{V}^{\pi}(\mathbf{x}))$ which is large when $(\boldsymbol{\theta} - \boldsymbol{\mu}^{\pi}(x))^t \mathbf{V}^{\pi}(\mathbf{x})^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}^{\pi}(\mathbf{x}))$ is small. Furthermore, this quadratic form has a chi-square distribution with p degrees of freedom, so the $100(1-\alpha)\%$ HPD credible set for $\boldsymbol{\theta}$ is the ellipse

$$C = \{\boldsymbol{\theta}: (\boldsymbol{\theta} - \boldsymbol{\mu}^{\pi}(\mathbf{x}))^t \mathbf{V}^{\pi}(\mathbf{x})^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}^{\pi}(\mathbf{x})) \le \chi_p^2(1-\alpha)\}, \tag{4.11}$$

where $\chi_p^2(1-\alpha)$ is the $(1-\alpha)$-fractile of the chi-square distribution.

In the multivariate case, the use of the normal approximation to the posterior is often especially valuable, because of the increased calculational difficulties. An example of such use is given in Subsection 4.7.10.

Alternatives to the HPD *Credible Set*

Though seemingly natural, the HPD credible set has the unnatural property of not necessarily being invariant under transformations.

EXAMPLE 6 (continued). Slightly modify this example by assuming that the parameter space is $\Theta = (0, \infty)$. Then the posterior becomes

$$\pi(\theta|x) = ce^{\theta}(1+\theta^2)^{-1} I_{(0,x)}(\theta).$$

This was shown to be increasing in θ, so that an HPD credible set will be of the form $(b(\alpha), x)$.

Suppose, now, that, instead of working with θ, we had used $\eta = \exp\{\theta\}$ as the unknown parameter. This is a one-to-one transformation, and should ideally have no effect on the ultimate answer. The posterior density for η, $\pi^*(\eta|x)$, can be found by simply transforming $\pi(\theta|x)$; note that $\theta = \log \eta$ so that $d\theta/d\eta = \eta^{-1}$ is the Jacobian that we must multiply by to obtain the transformed density. Thus

$$\begin{aligned}\pi^*(\eta|x) &= \eta^{-1} ce^{\log \eta}(1+(\log \eta)^2)^{-1} I_{(0,x)}(\log \eta) \\ &= c(1+(\log \eta)^2)^{-1} I_{(1,\exp\{x\})}(\eta).\end{aligned}$$

Note that $\pi^*(\eta|x)$ is decreasing on $(1, \exp\{x\})$, so that the $100(1-\alpha)\%$ HPD credible set for η will be an interval of the form $(1, d(\alpha))$. Transforming back to the θ coordinate system, this interval becomes $(0, \log d(\alpha))$.

The extreme conflict here is apparent; in the original parametrization, the *upper tail* of the posterior is the HPD credible set, while, for a monotonic reparametrization, it is the *lower tail.*

For more examples and discussion of the conflict, see Lehmann (1985). There is no clearcut resolution of the conflict, which is one reason that many Bayesians encourage reporting of the entire posterior, as opposed to just a credible set. Of course, there often is a "preferred" parametrization, in the sense that θ might be in some type of natural unit, and use of an HPD credible set in such a parametrization is not unreasonable.

More formally, one could define a secondary criteria to specify a credible set. One natural such criteria is size. Different measures of size could be considered, however; a quite general formulation of the size of a set C is given by

$$S(C) = \int_C s(\theta)\, d\theta,$$

where s is a nonnegative function. If $s(\theta) \equiv 1$ is chosen, $S(C)$ is just the usual Lebesgue measure of C ("length" of an interval, "volume" in higher dimensions).

The problem can then be stated as that of finding a $100(1-\alpha)\%$ credible set which has minimum size with respect to S; let us call such a set the

S-optimal $100(1-\alpha)\%$ *credible set.* It is easy to show that, in the continuous case (i.e., when $\pi(\theta|x)$ is a density with respect to Lebesgue measure), the S-optimal $100(1-\alpha)\%$ credible set is given (under mild conditions, see Exercise 35) by

$$C = \{\theta: \pi(\theta|x) > ks(\theta)\} \tag{4.12}$$

for some positive k. Note that if $s(\theta) \equiv 1$, then the S-optimal set is simply the HPD credible set, as indicated earlier. This does, therefore, provide a formal justification for the use of the HPD credible set when θ is a "preferred" parametrization, in that we might reasonably want to minimize volume (Lebesgue measure) in this preferred parametrization. An example in which a measure of size other than volume might be desirable is given in Exercise 34.

The utilization of S to specify an optimal confidence set preserves "optimality" under transformation, in that the choice of S will clearly be dependent on the parametrization chosen. Indeed, if S is deemed suitable in the θ parametrization, then the appropriate measure of size in a transformed parametrization is simply that induced by transforming S.

We will not pursue this subject further, because we do not view credible sets as having a clear decision-theoretic role, and therefore are leery of "optimality" approaches to selection of a credible set. We mainly view credible sets as an easily reportable crude summary of the posterior distribution, and, for this purpose, are not unduly troubled by, say, nonuniqueness of HPD credible sets.

4.3.3. Hypothesis Testing

In classical hypothesis testing, a null hypothesis H_0: $\theta \in \Theta_0$ and an alternative hypothesis H_1: $\theta \in \Theta_1$ are specified. A test procedure is evaluated in terms of the probabilities of Type I and Type II error. These probabilities of error represent the chance that a sample is observed for which the test procedure will result in the wrong hypothesis being accepted.

In Bayesian analysis, the task of deciding between H_0 and H_1 is conceptually more straightforward. One merely calculates the posterior probabilities $\alpha_0 = P(\Theta_0|x)$ and $\alpha_1 = P(\Theta_1|x)$ and decides between H_0 and H_1 accordingly. The conceptual advantage is that α_0 and α_1 are the actual (subjective) probabilities of the hypotheses in light of the data and prior opinions. The difficulties in properly interpreting classical error probabilities will be indicated later.

Although posterior probabilities of hypotheses are the primary Bayesian measures in testing problems, the following related concepts are also of interest. Throughout this section we will use π_0 and π_1 to denote the prior probabilities of Θ_0 and Θ_1, respectively.

Definition 6. The ratio α_0/α_1 is called the *posterior odds ratio* of H_0 to H_1, and π_0/π_1 is called the *prior odds ratio.* The quantity

$$B = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}} = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1} = \frac{\alpha_0\pi_1}{\alpha_1\pi_0}$$

is called the *Bayes factor* in favor of Θ_0.

Many people are more comfortable with "odds" than with probabilities, and indeed it is often convenient to summarize the evidence in terms of posterior odds. (Saying that $\alpha_0/\alpha_1 = 10$ clearly conveys the conclusion that H_0 is 10 times as likely to be true as H_1.) The interest in the Bayes factor is that it can sometimes be interpreted as the "odds for H_0 to H_1 that are *given by the data.*" This is a clearly valid interpretation when the hypotheses are simple, i.e., when $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, for then

$$\alpha_0 = \frac{\pi_0 f(x|\theta_0)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)}, \qquad \alpha_1 = \frac{\pi_1 f(x|\theta_1)}{\pi_0 f(x|\theta_0) + \pi_1 f(x|\theta_1)},$$

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi_0 f(x|\theta_0)}{\pi_1 f(x|\theta_1)} \quad \text{and} \quad B = \frac{\alpha_0\pi_1}{\alpha_1\pi_0} = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

In other words, B is then just the likelihood ratio of H_0 to H_1, which is commonly viewed (even by many non-Bayesians) as the odds for H_0 to H_1 that are given by the data.

In general, however, B will depend on the prior input. To explore this dependence (and for later developments), it is convenient to write the prior as

$$\pi(\theta) = \begin{cases} \pi_0 g_0(\theta) & \text{if } \theta \in \Theta_0, \\ \pi_1 g_1(\theta) & \text{if } \theta \in \Theta_1, \end{cases} \tag{4.13}$$

so that g_0 and g_1 are (proper) densities which describe how the prior mass is spread out over the two hypotheses. (Recall that π_0 and π_1 are the prior probabilities of Θ_0 and Θ_1.) With this representation, we can write

$$\frac{\alpha_0}{\alpha_1} = \frac{\int_{\Theta_0} dF^{\pi(\theta|x)}(\theta)}{\int_{\Theta_1} dF^{\pi(\theta|x)}(\theta)} = \frac{\int_{\Theta_0} f(x|\theta)\pi_0 dF^{g_0}(\theta)/m(x)}{\int_{\Theta_1} f(x|\theta)\pi_1 dF^{g_1}(\theta)/m(x)}$$
$$= \frac{\pi_0 \int_{\Theta_0} f(x|\theta) dF^{g_0}(\theta)}{\pi_1 \int_{\Theta_1} f(x|\theta) dF^{g_1}(\theta)}.$$

Hence

$$B = \frac{\int_{\Theta_0} f(x|\theta) dF^{g_0}(\theta)}{\int_{\Theta_1} f(x|\theta) dF^{g_1}(\theta)},$$

which is the ratio of "weighted" (by g_0 and g_1) likelihoods of Θ_0 to Θ_1. Because of the involvement of g_0 and g_1, this cannot be viewed as a measure of the relative support for the hypotheses provided solely by the data. Sometimes, however, B will be relatively insensitive to reasonable choices

of g_0 and g_1, and then such an interpretation is reasonable. The main operational advantage of having such a "stable" Bayes factor is that a scientific report (see Section 4.10) could include this Bayes factor, and any reader could then determine his personal posterior odds by simply multiplying the reported Bayes factor by his personal prior odds.

EXAMPLE 1 (continued). The child taking the IQ test is to be classified as having below average IQ (less than 100) or above average IQ (greater than 100). Formally, it is thus desired to test H_0: $\theta \leq 100$ versus H_1: $\theta > 100$. Recalling that the posterior distribution of θ is $\mathcal{N}(110.39, 69.23)$ a table of normal probabilities yields

$$\alpha_0 = P(\theta \leq 100|x) = 0.106, \qquad \alpha_1 = P(\theta > 100|x) = 0.894,$$

and hence the posterior odds ratio is $\alpha_0/\alpha_1 = 1/8.44$. Also, the prior is $\mathcal{N}(100, 225)$, so that $\pi_0 = P^{\pi}(\theta \leq 100) = \frac{1}{2} = \pi_1$ and the prior odds ratio is 1. (Note that a prior odds ratio of 1 indicates that H_0 and H_1 are viewed as equally plausible initially.) The Bayes factor is thus $B = \alpha_0\pi_1/(\alpha_1\pi_0) = 1/8.44$.

We next explicitly consider one-sided testing, testing of a point null hypothesis, and multiple hypothesis testing, pointing out general features of Bayesian testing and comparing the Bayesian and classical approaches in each case.

I. *One-Sided Testing*

One-sided hypothesis testing occurs when $\Theta \subset R^1$ and Θ_0 is entirely to one side of Θ_1. Example 1 (continued) above is an illustration of this. There are no unusual features of Bayesian testing here. What is of interest, however, is that this is one of the few testing situations in which classical testing, particularly the use of P-values, will sometimes have a Bayesian justification. Consider, for instance, the following example.

EXAMPLE 8 (continued). When $X \sim \mathcal{N}(\theta, \sigma^2)$ and θ has the noninformative prior $\pi(\theta) = 1$, we saw that $\pi(\theta|x)$ is $\mathcal{N}(x, \sigma^2)$. Consider now the situation of testing H_0: $\theta \leq \theta_0$ versus H_1: $\theta > \theta_0$. Then

$$\alpha_0 = P(\theta \leq \theta_0|x) = \Phi((\theta_0 - x)/\sigma),$$

where, again, Φ is the standard normal c.d.f.

The classical P-value against H_0 is the probability, when $\theta = \theta_0$, of observing an X "more extreme" than the actual data x. Here the P-value would be

$$P\text{-value} = P(X \geq x) = 1 - \Phi((x - \theta_0)/\sigma).$$

Because of the symmetry of the normal distribution, it follows that α_0 equals the P-value against H_0.

The use of a noninformative prior in this example is, of course, rather disturbing at first glance, since it gives infinite mass to each of the hypotheses (precluding, for instance, consideration of prior odds). It is, however, possible to justify the use of the noninformative prior as an approximation to very vague prior beliefs. Indeed, any proper prior density which is roughly constant over the interval $(\theta_0 - 2\sigma, x + 2\sigma)$ (here we assume $x > \theta_0$), and is not significantly larger outside this interval, would result in an α_0 roughly equal to the P-value.

In a number of other one-sided testing situations, vague prior information will tend to result in posterior probabilities that are similar to P-values (cf. Pratt (1965), DeGroot (1973), Dempster (1973), Dickey (1977), Zellner and Siow (1980), Hill (1982), and Good (1983, 1984)). This is not true for all one-sided testing problems, however. For instance, testing $H_0: \theta = 0$ (or $H_0: 0 \leq \theta \leq \varepsilon$) versus $H_1: \theta > 0$ (or $H_1: \theta > \varepsilon$) is a one-sided testing problem, but P-values and posterior probabilities will tend to differ drastically (in much the same way as in part II of this subsection; see also Berger and Sellke (1987)). Note, also, that classical probabilities of Type I and Type II error do *not* usually have any close correspondence to posterior probabilities of hypotheses; this may partly explain the preference of many classical practitioners for use of P-values instead of probabilities of Type I and Type II errors.

II. *Testing a Point Null Hypothesis*

It is very common in classical statistics (cf. the statistical literature survey in Zellner (1984a)) to conduct a test of $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$. Such testing of a point null hypothesis is interesting, partly because the Bayesian approach contains some novel features, but mainly because the Bayesian answers differ *radically* from the classical answers.

Before discussing these issues, several comments should be made about the entire enterprise of testing a point null hypothesis. First of all, as mentioned in Subsection 1.6.1, tests of point null hypotheses are commonly performed in inappropriate situations. It will virtually never be the case that one seriously entertains the possibility that $\theta = \theta_0$ *exactly* (cf. Hodges and Lehmann (1954) and Lehmann (1959)). More reasonable would be the null hypothesis that $\theta \in \Theta_0 = (\theta_0 - b, \theta_0 + b)$, where $b > 0$ is some constant chosen so that all θ in Θ_0 can be considered "indistinguishable" from θ_0. An example in which this might arise would be an attempt to analyze a chemical by observing some feature, θ, of its reaction with a known chemical. If it were desired to test whether or not the unknown chemical is a specific compound, with a reaction strength θ_0 known up to an accuracy of b, then it would be reasonable to test $H_0: \theta \in (\theta_0 - b, \theta_0 + b)$ versus $H_1: \theta \notin (\theta_0 - b, \theta_0 + b)$. A similar example involving forensic science can be found in Lindley (1977) (see also Shafer (1982c)). An example where b might be extremely

close to zero is in a test for extra sensory perception, with θ_0 reflecting the hypothesis of *no* extra sensory perception. (The only reason that b would probably be nonzero here is that the experiment designed to test for ESP would not lead to a perfectly well-defined θ_0.) Of course, there are also many decision problems that would lead to a null hypothesis of the above interval form with a *large* b, but such problems will rarely be well approximated by testing a point null.

Given that one should really be testing H_0: $\theta \in (\theta_0 - b, \theta_0 + b)$, we need to know when it is suitable to approximate H_0 by H_0: $\theta = \theta_0$. From the Bayesian perspective, the only sensible answer to this question is—the approximation is reasonable if the posterior probabilities of H_0 are nearly equal in the two situations. A very strong condition under when this would be the case is that the observed likelihood function be approximately constant on $(\theta_0 - b, \theta_0 + b)$. (A more formal statement of this is given in the Exercises.)

EXAMPLE 11. Suppose a sample $X_1, \ldots, X_n$ is observed from a $\mathcal{N}(\theta, \sigma^2)$ distribution, σ^2 known. The observed likelihood function is then proportional to a $\mathcal{N}(\bar{x}, \sigma^2/n)$ density for θ. This will be nearly constant on $(\theta_0 - b, \theta_0 + b)$ when b is small compared to $\sigma/\sqrt{n}$. For instance, in the interesting case where the classical test statistic $z = \sqrt{n}|\bar{x} - \theta_0|/\sigma$ is larger than 1, the likelihood function can be shown to vary by no more than 5% on $(\theta_0 - b, \theta_0 + b)$ if

$$b \leq (0.024) z^{-1} \sigma / \sqrt{n}.$$

When $z = 2$, $\sigma = 1$, and $n = 25$, this condition becomes $b \leq 0.0024$. Note that the bound on b will depend on $|\bar{x} - \theta_0|$, as well as on $\sigma/\sqrt{n}$.

The point null approximation will be satisfactory in substantially greater generality than the "constant likelihood" situation, but it is usually easier for a Bayesian to deal directly with the interval hypothesis than to check the adequacy of the approximation. Nevertheless, we will develop the Bayesian test of a point null, so as to allow evaluation of the classical test.

To conduct a Bayesian test of the point null hypothesis H_0: $\theta = \theta_0$, one cannot use a continuous prior density, since any such prior will give θ_0 prior (and hence posterior) probability zero. A reasonable approach, therefore, is to give θ_0 a positive *probability* π_0, while giving the $\theta \neq \theta_0$ the *density* $\pi_1 g_1(\theta)$, where $\pi_1 = 1 - \pi_0$ and g_1 is proper. One can think of π_0 as the mass that would have been assigned to the realistic hypothesis H_0: $\theta \in (\theta_0 - b, \theta_0 + b)$, were the point null approximation not being used.

The Bayesian analysis of this situation is quite straightforward, although one must be careful to remember that the prior has discrete and continuous parts. The marginal density of X is

$$m(x) = \int f(x|\theta) dF^{\pi}(\theta) = f(x|\theta_0)\pi_0 + (1 - \pi_0) m_1(x),$$

where

$$m_1(x) = \int_{\{\theta \neq \theta_0\}} f(x|\theta) dF^{g_1}(\theta)$$

is the marginal density of X with respect to g_1. Hence the posterior probability that $\theta = \theta_0$ is

$$\begin{aligned} \pi(\theta_0|x) &= \frac{f(x|\theta_0)\pi_0}{m(x)} \\ &= \frac{f(x|\theta_0)\pi_0}{f(x|\theta_0)\pi_0 + (1-\pi_0)m_1(x)} \\ &= \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}. \end{aligned} \tag{4.14}$$

Note that this is α_0, the posterior probability of H_0 in our earlier terminology, and that $\alpha_1 = 1 - \alpha_0$ is hence the posterior probability of H_1. Also, the posterior odds ratio can easily be shown to be (recalling that $\pi_1 = 1 - \pi_0$)

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi(\theta_0|x)}{1 - \pi(\theta_0|x)} = \frac{\pi_0}{\pi_1} \cdot \frac{f(x|\theta_0)}{m_1(x)},$$

so that the Bayes factor for H_0 versus H_1 is

$$B = f(x|\theta_0)/m_1(x). \tag{4.15}$$

EXAMPLE 11 (continued). Reduction to the sufficient statistic $\bar{X}$ yields the effective likelihood function

$$f(\bar{x}|\theta) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{n}{2\sigma^2}(\theta - \bar{x})^2\right\}.$$

Suppose that g_1 is a $\mathcal{N}(\mu, \tau^2)$ density on $\theta \neq \theta_0$. (Usually $\mu = \theta_0$ would be an appropriate choice, since θ close to θ_0 would often be deemed more likely *a priori* than θ far from θ_0.) Then, as in Example 1, m_1 is a $\mathcal{N}(\mu, \tau^2 + \sigma^2/n)$ density (the point $\theta = \theta_0$ not mattering in the integration since g_1 is a continuous density). We thus obtain

$$\begin{aligned} \alpha_0 = \pi(\theta_0|x) = \Bigg[1 + \frac{(1-\pi_0)}{\pi_0} \\ \cdot \frac{\{2\pi(\tau^2 + \sigma^2/n)\}^{-1/2} \exp\{-(\bar{x}-\mu)^2/(2[\tau^2 + \sigma^2/n])\}}{\{2\pi\sigma^2/n\}^{-1/2} \exp\{-(\bar{x}-\theta_0)^2/(2\sigma^2/n)\}}\Bigg]^{-1}. \end{aligned} \tag{4.16}$$

In the special case where $\mu = \theta_0$, this reduces to

$$\begin{aligned} \alpha_0 &= \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{\exp\{(\bar{x}-\theta_0)^2 n\tau^2/(2\sigma^2[\tau^2 + \sigma^2/n])\}}{\{1 + n\tau^2/\sigma^2\}^{1/2}}\right]^{-1} \\ &= \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{\exp\{\frac{1}{2}z^2[1 + \sigma^2/(n\tau^2)]^{-1}\}}{\{1 + n\tau^2/\sigma^2\}^{1/2}}\right]^{-1}, \end{aligned} \tag{4.17}$$

where $z = \sqrt{n}\,|\bar{x} - \theta_0|/\sigma$ is the usual statistic used for testing H_0: $\theta = \theta_0$. For later reference, note that

$$\alpha_0 \geq \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{\exp\{\frac{1}{2}z^2\}}{\{1 + n\tau^2/\sigma^2\}^{1/2}}\right]^{-1}. \tag{4.18}$$

Table 4.2 presents values of α_0 for various z (chosen to correspond to standard classical two-tailed P-values or significance levels for testing a point null) and n, when the prior is specified by $\mu = \theta_0$, $\pi_0 = \frac{1}{2}$, and $\tau = \sigma$. The numbers in Table 4.2 are astonishing. For instance, if one observed $z = 1.96$, classical theory would allow one to reject H_0 at level $\alpha = 0.05$, the smallness of which gives the impression that H_0 is very likely to be false. But the posterior probability of H_0 is quite substantial, ranging from about $\frac{1}{3}$ for small n to nearly 1 for large n. Thus $z = 1.96$ actually provides little or no evidence against H_0 (for the specified prior).

The conflict in the above example is of such interest that it deserves further investigation. The Bayesian analysis can, of course, be questioned because of the choice of the prior. The assumption of a normal form for g_1 can be shown to have no real bearing on the issue (unless $|\bar{x} - \theta_0|$ is large), and the choices $\mu = \theta_0$ and $\pi_0 = \frac{1}{2}$ are natural. (Indeed, if one were trying to be "objective," these choices of μ and π_0 would seem to be virtually mandatory.) One could certainly question the choice $\tau = \sigma$, however. Jeffreys (1961) argues for such a choice in "objective" testing, but the answer definitely depends crucially on the choice for τ. (Jeffreys actually prefers a Cauchy form for the prior, but again the form matters only in extreme cases. See also Raiffa and Schlaifer (1961), Lindley (1965, 1977), Smith (1965), Zellner (1971, 1984a), Dickey (1971, 1974, 1980), Zellner and Siow (1980), and Diamond and Forrester (1983) for similar analyses and generalizations.)

When the answer heavily depends on features of the prior such as τ, a Bayesian has no recourse but to attempt subjective specification of the features. (See Section 4.10 for useful graphical techniques of presenting the Bayesian conclusion as a *function* of such features, allowing easy processing by consumers.) We can, however, expose the enormity of the classical

Table 4.2. Posterior Probability of H_0.

	n						
z (P-value)	1	5	10	20	50	100	1000
1.645 (0.1)	0.42	0.44	0.49	0.56	0.65	0.72	0.89
1.960 (0.05)	0.35	0.33	0.37	0.42	0.52	0.60	0.80
2.576 (0.01)	0.21	0.13	0.14	0.16	0.22	0.27	0.53
3.291 (0.001)	0.086	0.026	0.024	0.026	0.034	0.045	0.124

Bayesian conflict in testing a point null by finding the *minimum* of α_0 over all τ. We will actually go one step further, and find the minimum over *all* g_1. Versions of the following theorem can be found in Edwards, Lindman, and Savage (1963), and Dickey (1973, 1977). (The proof of the theorem is easy and will be left for the Exercises.)

Theorem 1. *For* any *distribution* g_1 *on* $\theta \neq \theta_0$,

$$\alpha_0 = \pi(\theta_0|x) \geq \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{r(x)}{f(x|\theta_0)}\right]^{-1}, \tag{4.19}$$

where $r(x) = \sup_{\theta \neq \theta_0} f(x|\theta)$. (*Usually,* $r(x) = f(x|\hat{\theta})$, *where* $\hat{\theta}$ *is a maximum likelihood estimate of* θ.) *The corresponding bound on the Bayes factor for* H_0 *versus* H_1 *is*

$$B = \frac{f(x|\theta_0)}{m_1(x)} \geq \frac{f(x|\theta_0)}{r(x)}. \tag{4.20}$$

EXAMPLE 11 (continued). With $f(\bar{x}|\theta)$ being a $\mathcal{N}(\bar{x}, \sigma^2/n)$ likelihood, it is clear that the supremum over all $\theta \neq \theta_0$ is $f(\bar{x}|\bar{x})$, so that

$$r(x) = f(\bar{x}|\bar{x}) = \{2\pi\sigma^2/n\}^{-1/2}.$$

Thus

$$\begin{aligned}\alpha_0 &\geq \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{\{2\pi\sigma^2/n\}^{-1/2}}{\{2\pi\sigma^2/n\}^{-1/2} \exp\{-(\bar{x}-\theta_0)^2/(2\sigma^2/n)\}}\right]^{-1} \\ &= \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \exp\{\tfrac{1}{2}z^2\}\right]^{-1},\end{aligned}$$

and

$$B \geq 1/\exp\{\tfrac{1}{2}z^2\},$$

where $z = \sqrt{n}|\bar{x} - \theta_0|/\sigma$. Table 4.3 gives the values of these bounds for the various z in Table 4.2 (and with $\pi_0 = \frac{1}{2}$ for the bound on α_0). Although the bounds on α_0 are substantially smaller than the particular values given in Table 4.2, they are still substantially larger than the corresponding P-values. Indeed, it can be shown (see Berger and Sellke (1987)) that, for $\pi_0 = \frac{1}{2}$ and $z > 1.68$,

$$\alpha_0 \geq (P\text{-value}) \times (1.25)z. \tag{4.21}$$

Table 4.3. Bounds on α_0 and B.

z	P-value	Bound on α_0	Bound on B
1.645	0.1	0.205	1/3.87
1.960	0.05	0.127	1/6.83
2.576	0.01	0.035	1/27.60
3.291	0.001	0.0044	1/224.83

A pause for recapitulation is in order. It is well understood that a classical error probability or P-value is *not* a posterior probability of a hypothesis. Nevertheless, it is *felt* by the vast majority of users that a P-value of 0.05 means that one can be pretty sure that H_0 is wrong. The evidence, however, is quite to the contrary. Any reasonably fair Bayesian analysis will show that there is at best very weak evidence against H_0 when $z = 1.96$. And even the lower bound on α_0, which was obtained by a Bayesian analysis heavily slanted towards H_1 (the g_1 chosen being that which was most favorable towards H_1), indicates much less real doubt about H_0 than $\alpha = 0.05$ would seem to imply. It thus appears that, for the situation studied, classical error probabilities or P-values are completely misleading descriptions of the evidence against H_0.

For those who are highly suspicious of any Bayesian reasoning, there is still the bound in (4.20) to consider. This bound is clearly the minimum likelihood ratio of H_0 to H_1. Thus the bound on B in Table 4.3 for $z = 1.96$ means that the likelihood ratio of H_0 to H_1 will be at least $1/6.83$; it would seem ridiculous to conclude that the data favors H_1 by more than a factor of 6.83. For more discussion, and also presentation of the conflict in conditional frequentist terms, see Berger and Sellke (1987).

In the face of this overwhelming evidence that classical testing of a point null is misleading, we must seek a better approach. Of course, we basically recommend the subjective Bayesian approach alluded to earlier (especially if implemented as discussed in Section 4.10). It is interesting to ask if there is a more "objective" sensible analysis, however. Noninformative prior Bayesian analyses unfortunately do not work well for testing a point null hypothesis, making impossible an objective Bayesian solution. One can, however, argue for a *standardized* Bayesian solution. Indeed Jeffreys (1961) does so argue for the use of a particular proper prior, one chosen to reflect what he feels will be reasonable uncertainty in many scientific problems. (Zellner and Siow (1980) call these *reference informative priors.*) Although we feel that such a standardized Bayesian approach would definitely be far better than the standardized classical approach, it is much harder to defend than other objective Bayesian analyses.

Another possibility is to present the bounds on α_0 and B, from Theorem 1, as objective evidence. Unfortunately these bounds will typically be much smaller than α_0 and B that are actually obtained through subjective Bayesian calculations; the use of such bounds would thus be limited to showing that there is *no* reason to doubt H_0. In terms of objective appearance, reporting the bound on B is particularly attractive, since it is just a likelihood ratio. (Similar and more extensive developments of likelihood ratios as evidence can be found in Edwards, Lindman, and Savage (1963) and Dempster (1973).)

A quite promising possibility is to find more accurate lower bounds on α_0 and B, by restricting g_1 in some natural fashion. In many applications,

for instance, it would be reasonable to restrict consideration to

$$\mathcal{G} = \{g_1 : g_1 \text{ is symmetric about } \theta_0 \text{ and nonincreasing in } |\theta - \theta_0|\}. \tag{4.22}$$

Indeed, this could be argued to be a *requirement* of an objective Bayesian analysis. It turns out to be surprisingly easy to work with such $\mathcal{G}$. Indeed, the desired lower bounds on α_0 and B are found in Berger and Sellke (1987) for quite general situations (see also Theorem 5 in Subsection 4.7.9). We content ourselves here with stating the conclusion in the situation of Example 11.

EXAMPLE 11 (continued). It is shown in Berger and Sellke (1987) that, if $z \le 1$, then $\alpha_0 \ge \pi_0$ and $B \ge 1$ for *any* g_1 in (4.22). For $z > 1$, the lower bounds on α_0 and B for this class of priors are

$$\alpha_0 \ge \left\{1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{[\phi(k+z)+\phi(k-z)]}{2\phi(z)}\right\}^{-1}, \tag{4.23}$$

and

$$B \ge \frac{2\phi(z)}{[\phi(k+z)+\phi(k-z)]}, \tag{4.24}$$

where k is a solution to equation (4.112) in Section 4.7.9, which can be approximated by

$$k = z + \left[2\log\left(\frac{k}{\Phi(k-z)}\right) - 1.838\right]^{1/2} \tag{4.25}$$

when $z \ge 1.645$. (Here ϕ and Φ denote the standard normal density and c.d.f., as usual.) Equation (4.25) has the feature of allowing iterative calculation of k: plug in a guess for k on the right-hand side of (4.25), and use the result as a new guess for k to be plugged back in; repeat the process until the guesses for k stabilize, which usually takes only two or three iterations. (A good first guess is $k = z$.)

Table 4.4 gives these lower bounds on α_0 and B for the values of z used in Tables 4.2 and 4.3. The bounds are much larger than the corresponding bounds in Table 4.3, and indeed are close to some of the exact values of α_0 and B obtained from the Jeffreys-type analysis in Table 4.2. Hence these

Table 4.4. $\mathcal{G}$-Bounds on α_0 and B.

z	P-value	$\mathcal{G}$-Bound on α_0	$\mathcal{G}$-Bound on B
1.645	0.10	0.390	1/1.56
1.960	0.05	0.290	1/2.45
2.576	0.01	0.109	1/8.17
3.291	0.001	0.018	1/54.56

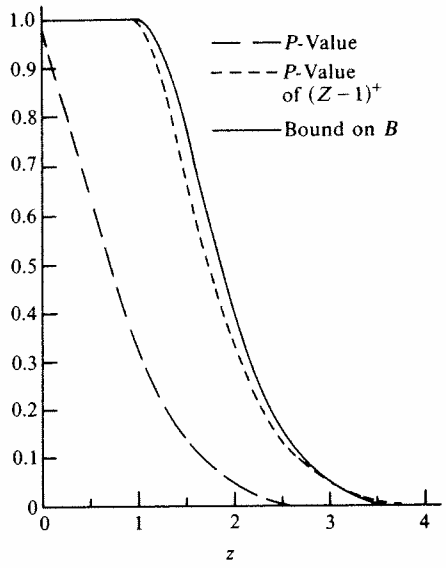

Figure 4.1

lower bounds are indeed somewhat reasonable as "objective" measures of the evidence against H_0, at least for small n.

A rather fascinating "empirical" observation follows from comparing the bound on B in (4.24) with the P-value of $(z-1)^+$ (the positive part of $(z-1)$). These are graphed in Figure 4.1, along with the P-value of z itself. (Note that the bound in (4.23) can be determined from the bound on B.) Clearly, the bound on B is roughly similar to the P-value of $(z-1)^+$. Since the bound on B can be interpreted as a bound on the comparative support of the data for the two hypotheses, the indication is that the common rule-of-thumb that

$$z = \begin{cases} 1 & \text{means only very mild evidence against } H_0, \\ 2 & \text{means significant evidence against } H_0, \\ 3 & \text{means highly significant evidence against } H_0, \\ 4 & \text{means overwhelming evidence against } H_0, \end{cases}$$

should, at the very least, be replaced by the rule-of-thumb

$$z = \begin{cases} 1 & \text{means no evidence against } H_0, \\ 2 & \text{means only very mild evidence against } H_0, \\ 3 & \text{means significant evidence against } H_0, \\ 4 & \text{means highly significant evidence against } H_0. \end{cases}$$

Although we have indicated that the lower bounds on B, or on α_0 when $\pi_0 = \frac{1}{2}$, can be interpreted as *lower bounds* on the comparative support of

the data for the two hypotheses, it is questionable if they can be used as actual measures of evidence of comparative support. The reason is that all the bounds are independent of n. For the choice of g_1 leading to (4.18), however, it is clear that $\alpha_0 \to 1$ as $n \to \infty$ with z fixed! (See also Table 4.2.) Thus the bounds on α_0 and B are all likely to be excessively small when n is very large. This phenomenon, that $\alpha_0 \to 1$ as $n \to \infty$ and the P-value is held fixed, actually holds for virtually any fixed prior and point null testing problem. Indeed, if g_1 has a continuous density at θ_0, then, for large (and even moderate) n, an accurate approximation to α_0 when x corresponds to a moderate P-value is

$$\alpha_0 \cong \left[1 + \frac{(1-\pi_0)}{\pi_0} \cdot \frac{g_1(\theta_0)}{f(x|\theta_0)}\right]^{-1}. \tag{4.26}$$

When n is large and the P-value is not too small, $f(x|\theta_0)$ will be large and α_0 close to one. Thus, for large n, the true "evidence" against H_0 is likely to be much less than that indicated by the lower bounds on α_0 or B. One is then left no recourse but to perform a subjective Bayesian analysis of the problem (but see Section 4.10). It is interesting that there is *no* suitable "noninformative prior" Bayesian analysis here.

This large n phenomenon provides an extreme illustration of the conflict between classical and Bayesian testing of a point null. One could classically reject H_0 with a P-value of 10^{-10}, yet, if n were large enough, the posterior probability of H_0 would be very close to 1. This surprising result has been called "Jeffreys' paradox" and "Lindley's paradox," and has a very long history (cf. Jeffreys (1961), Good (1950, 1965, 1983), Lindley (1957, 1961), and Shafer (1982c)). We will not discuss this "paradox" here because the point null approximation is rarely justifiable for very large n, and because we have already seen enough reasons, when n is small, to reject the use of classical error measures. It is, of course, useful to know that the problem becomes even worse for larger n.

We have concentrated, in this section, on the Example 11 situation of testing a normal mean, so as to provide a reasonably complete view of the issues. Most of the general formulas apply to other point null testing problems, however, and, in one dimension, the numerical results will be very similar. In higher dimensions the lower bounds for *all* g_1 in Theorem 1 can be very small, but lower bounds for $\mathcal{G}$ as in (4.22) will usually be comparable to the one-dimensional bounds. (An alternative to the use of $\mathcal{G}$ as in (4.22) is to replace $f(x|\theta)$ in the analysis by the density of the one-dimensional classical test statistic that would have been used, see Berger and Sellke (1987) for further discussion.) Additional references concerning point null analyses of the type we have considered, but in other situations, include Good (1950, 1958, 1983), Lempers (1971), Leamer (1978), and Smith and Spiegelhalter (1980).

III. *Multiple Hypothesis Testing*

The interesting feature of multiple hypothesis testing is that it is no more difficult, from a Bayesian perspective, than is testing of two hypotheses. One simply calculates the posterior probability of each hypothesis.

EXAMPLE 1 (continued). The child taking the IQ test is to be classified as having below average IQ (less than 90), average IQ (90 to 110), or above average IQ (over 110). Calling these three regions Θ_1, Θ_2, and Θ_3, respectively, and recalling that the posterior is $\mathcal{N}(110.39, 69.23)$, a table of normal probabilities can be used to show that $P(\Theta_1|x=115)=0.007$, $P(\Theta_2|x=115)=0.473$, and $P(\Theta_3|x=115)=0.520$.

4.3.4. Predictive Inference

In Subsection 2.4.4 we discussed the situation of trying to predict a random variable $Z\sim g(z|\theta)$ based on the observation of $X\sim f(x|\theta)$. We will again assume that X and Z are independent, and (for simplicity) that g is a density. (If X and Z are not independent, the necessary change in the following would be to replace $g(z|\theta)$ by $g(z|\theta, x)$.)

The idea of Bayesian predictive inference is that, since $\pi(\theta|x)$ is the believed (posterior) distribution of θ, then $g(z|\theta)\pi(\theta|x)$ is the joint distribution of z and θ given x, and integrating out over θ will give the believed distribution of z given x.

Definition 7. The *predictive density* of Z given x, when the prior for θ is π, is defined by

$$p(z|x)=\int_\Theta g(z|\theta)dF^{\pi(\theta|x)}(\theta).$$

EXAMPLE 12. Consider the linear regression model

$$Z=\theta_1+\theta_2 Y+\varepsilon, \tag{4.27}$$

where $\varepsilon\sim\mathcal{N}(0,\sigma^2)$ (σ^2 known for simplicity). Available is independent data $((z_1,y_1),\ldots,(z_n,y_n))$ from the regression. A sufficient statistic for $\boldsymbol{\theta}=(\theta_1,\theta_2)$ is the least squares estimator $\mathbf{X}=(X_1,X_2)$, where

$$X_2=\sum_{i=1}^n (Z_i-\bar{Z})(Y_i-\bar{Y})/SSY, \qquad X_1=\bar{Z}-X_2\bar{Y},$$

$$SSY=\sum_{i=1}^n (Y_i-\bar{Y})^2, \qquad \bar{Z}=\frac{1}{n}\sum_{i=1}^n Z_i, \quad \text{and} \quad \bar{Y}=\frac{1}{n}\sum_{i=1}^n Y_i.$$

Furthermore, $\mathbf{X} \sim \mathcal{N}_2(\boldsymbol{\theta}, \sigma^2 \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \frac{1}{SSY} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} y_i^2 & -\bar{y} \\ -\bar{y} & 1 \end{pmatrix}.$$

If the noninformative prior $\pi(\boldsymbol{\theta}) = 1$ is used for $\boldsymbol{\theta}$, the posterior distribution, $\pi(\boldsymbol{\theta}|\mathbf{x})$, is easily seen to be $\mathcal{N}_2(\mathbf{x}, \sigma^2 \boldsymbol{\Sigma})$.

Suppose now that one desires to predict a future Z corresponding (through (4.27)) to a given y. Clearly $g(z|\boldsymbol{\theta})$ is then $\mathcal{N}(\theta_1 + \theta_2 y, \sigma^2)$, and the joint density of $(z, \boldsymbol{\theta})$ given $\mathbf{x}$ (and for the given y) is also normal. The desired predictive distribution, $p(z|x)$, is the marginal distribution of Z from this joint normal posterior; this marginal must also be a normal distribution, and calculation gives a mean of $(x_1 + x_2 y)$ and variance $V = \sigma^2(1 + n^{-1} + (\bar{y} - y)^2/SSY)$.

This predictive distribution can be used for any desired Bayesian inference. For instance, a $100(1-\alpha)\%$ HPD credible set for Z would be

$$\left((x_1 + x_2 y) + z\left(\frac{\alpha}{2}\right)\sqrt{V}, \quad (x_1 + x_2 y) - z\left(\frac{\alpha}{2}\right)\sqrt{V} \right).$$

Note that this happens to be the same set that the classical approach suggests for predicting Z, providing another example of the formal similarity that often exists between answers from classical and noninformative prior Bayesian analyses.

We will not specifically pursue predictive problems in this book, although many of the issues and techniques discussed for posterior distributions will apply equally well to predictive distributions. Among the many good references on predictive inference are Roberts (1965), Geisser (1971, 1980, 1984b), and Aitchison and Dunsmore (1975).

4.4. Bayesian Decision Theory

In this section the influence of the loss function will be considered, as well as that of the prior information. It will be seen that the Bayesian approach to decision theory is conceptually very straightforward.

4.4.1. Posterior Decision Analysis

In Subsections 1.3.1 and 1.5.1 the conditional Bayes decision principle was discussed. One looks at the expected loss of an action for the believed distribution of θ at the time of decision making. We now know that this distribution should be the posterior distribution, $\pi(\theta|x)$. For completeness, we redefine the relevant concepts.

Definition 8. The *posterior expected loss* of an action a, when the posterior distribution is $\pi(\theta|x)$, is

$$\rho(\pi(\theta|x), a) = \int_{\Theta} L(\theta, a)\, dF^{\pi(\theta|x)}(\theta). \tag{4.28}$$

A (*posterior*) *Bayes action*, to be denoted by $\delta^{\pi}(x)$ (for consistency with later definitions), is any action $a \in \mathcal{A}$ which minimizes $\rho(\pi(\theta|x), a)$, or equivalently which minimizes

$$\int_{\Theta} L(\theta, a) f(x|\theta) dF^{\pi}(\theta). \tag{4.29}$$

The simplicity of the Bayesian approach follows from the fact that an optimal action can be found by a simple minimization of (4.28) or (4.29). (The possible advantage of using (4.29) is that $m(x)$ need not be calculated.) Of course, there could be several minimums, and, hence, several Bayes actions.

Another kind of Bayesian analysis was mentioned in Subsections 1.3.2 and 1.5.2, namely choice of a decision rule through minimization of the (frequentist) Bayes risk, $r(\pi, \delta)$. Although minimizing $r(\pi, \delta)$ appears to be a much more difficult problem than minimizing posterior expected loss (choosing a minimizing *function* is usually very hard), the two problems are essentially equivalent.

Result 1. *A Bayes rule δ^{π} (i.e., a rule minimizing $r(\pi, \delta)$) can be found by choosing, for each x such that $m(x) > 0$, an action which minimizes the posterior expected loss* (4.28) (*or equivalently* (4.29)). *The rule can be defined arbitrarily when $m(x) = 0$.*

Since (posterior) Bayes actions need not be unique, δ^{π} need not be unique (and can sometimes even be chosen to be randomized). Also, if $r(\pi, \delta) = \infty$ for *all* δ, then *any* decision rule is a Bayes rule (not just those which correspond to posterior Bayes actions). There are certain technical conditions needed for the validity of Result 1 (cf. Brown and Purves (1973) or Diaconis and Stein (1983)), but the result will hold in all cases of practical interest. Result 1 is an immediate consequence of the following result, which is stated separately for later use (and is only stated for nonrandomized estimators—the generalization to randomized estimators is left for the exercises).

Result 2. *If δ is a nonrandomized estimator, then*

$$r(\pi, \delta) = \int_{\{x:\, m(x)>0\}} \rho(\pi(\theta|x), \delta(x)) dF^{m}(x). \tag{4.30}$$

PROOF. By definition,

$$r(\pi, \delta) = \int_\Theta R(\theta, \delta) dF^\pi(\theta)$$
$$= \int_\Theta \int_{\mathscr{X}} L(\theta, \delta(x)) dF^{X|\theta}(x) dF^\pi(\theta).$$

Since $L(\theta, a) \geq -K > -\infty$ and all measures above are finite, Fubini's theorem can be employed to interchange orders of integration and obtain

$$r(\pi, \delta) = \begin{cases} \int_{\mathscr{X}} \left[\int_\Theta L(\theta, \delta(x)) f(x|\theta) dF^\pi(\theta) \right] dx, \\ \sum_{x \in \mathscr{X}} \left[\int_\Theta L(\theta, \delta(x)) f(x|\theta) dF^\pi(\theta) \right], \end{cases}$$

in the cases of continuous and discrete $\mathscr{X}$, respectively. Finally, noting that, if $m(x) = 0$, then $f(x|\theta) = 0$ almost everywhere with respect to π, the definitions of $\pi(\theta|x)$ and $\rho(\pi(\theta|x), \delta)$ yield the result. □

Result 1 easily follows from Result 2, since a minimizing δ can be found by minimizing the integrand in (4.30) for each x for which $m(x) > 0$, and this is precisely the earlier definition of a posterior Bayes action. The overall minimization of $r(\pi, \delta)$ has been called the *normal form* of Bayesian analysis, while minimization of $\rho(\pi(\theta|x), a)$ has been called the *extensive form* (cf. Raiffa and Schlaifer (1961)). Operationally, we will virtually always proceed by minimizing $\rho(\pi(\theta|x), a)$, and will for simplicity always call this result the Bayes rule $\delta^\pi(x)$ (to be interpreted as an action or decision rule, as implied by the context). The only case where the normal form of Bayesian analysis need be considered is when a restricted set of decision rules is to be used, so that the overall Bayes rule need not be in the restricted class. (Some examples are discussed in Subsection 4.7.7, see also Haff (1983).)

Note that, from the conditional perspective together with the utility development of the loss, the *correct* way to view the situation is that of minimizing $\rho(\pi(\theta|x), a)$. One should condition on what is known, namely x (incidentally following the Likelihood Principle, since $\pi(\theta|x)$ depends only on the observed likelihood function), and average the utility over what is unknown, namely θ. The desire to minimize $r(\pi, \delta)$ would be deemed rather bizarre from this perspective.

So far we have assumed that π is a proper prior. Even when π is improper, however, it will often make sense to minimize posterior expected loss.

Definition 9. If π is an improper prior, but $\delta^\pi(x)$ is an action which minimizes (4.28) or (4.29) for each x with $m(x) > 0$, then δ^π is called a *generalized Bayes rule.*

We now turn to some standard areas of application of Bayesian decision theory. It should be emphasized that Bayesian decision theory is by no

means limited to such standard applications. The methodology of choosing a prior and loss and minimizing the posterior expected loss can be applied in almost any situation.

4.4.2. Estimation

In Bayesian estimation of a real valued parameter θ, the loss which is easiest to deal with is squared-error loss ($L(\theta, a) = (\theta - a)^2$). The posterior expected loss is then

$$\int_\Theta (\theta - a)^2 dF^{\pi(\theta|x)}(\theta).$$

The value of a which minimizes this can be found by expanding the quadratic expression, differentiating with respect to a, and setting equal to zero. The result (assuming all integrals are finite) is

$$0 = \frac{d}{da}\left[\int_\Theta \theta^2 dF^{\pi(\theta|x)}(\theta) - 2a\int_\Theta \theta dF^{\pi(\theta|x)}(\theta) + a^2\int_\Theta dF^{\pi(\theta|x)}(\theta)\right]$$
$$= -2E^{\pi(\theta|x)}[\theta] + 2a.$$

Solving for a gives the following result.

Result 3. *If $L(\theta, a) = (\theta - a)^2$, the Bayes rule is*

$$\delta^\pi(x) = E^{\pi(\theta|x)}[\theta],$$

which is the mean of the posterior distribution of θ given x.

EXAMPLE 1 (continued). If f and π are normal, the posterior is $\mathcal{N}(\mu(x), \rho^{-1})$. This has mean $\mu(x)$, so the Bayes rule for squared-error loss is

$$\delta^\pi(x) = \mu(x) = \frac{\sigma^2\mu}{\sigma^2 + \tau^2} + \frac{\tau^2 x}{\sigma^2 + \tau^2}.$$

For weighted squared-error loss the following result holds. (The proof will be left as an exercise.)

Result 4. *If $L(\theta, a) = w(\theta)(\theta - a)^2$, the Bayes rule is*

$$\delta^\pi(x) = \frac{E^{\pi(\theta|x)}[\theta w(\theta)]}{E^{\pi(\theta|x)}[w(\theta)]}$$
$$= \frac{\int \theta w(\theta) f(x|\theta) dF^\pi(\theta)}{\int w(\theta) f(x|\theta) dF^\pi(\theta)}.$$

From the last expression in Result 4, it is interesting to note that the weight function, $w(\theta)$, plays a role analogous to that of the prior $\pi(\theta)$. This is important to note, in that robustness concerns involving $w(\theta)$ are thus the same as robustness concerns involving the prior.

For the quadratic loss $L(\boldsymbol{\theta}, \mathbf{a}) = (\boldsymbol{\theta} - \mathbf{a})^t \mathbf{Q}(\boldsymbol{\theta} - \mathbf{a})$ ($\boldsymbol{\theta}$ and $\mathbf{a}$ are now vectors and $\mathbf{Q}$ is a positive definite matrix), it can be shown that the Bayes estimator is still the posterior mean. (Interestingly, $\mathbf{Q}$ has no effect.)

For absolute error loss, the following result holds. Recall that a median is a $\frac{1}{2}$-fractile of a distribution. (In general, a point $z(\alpha)$ is an α-fractile of the distribution of a random variable X if $P(X \leq z(\alpha)) \geq \alpha$ and $P(X < z(\alpha)) \leq \alpha$.)

Result 5. *If $L(\theta, a) = |\theta - a|$, any median of $\pi(\theta|x)$ is a Bayes estimator of θ.*

PROOF. Let m denote a median of $\pi(\theta|x)$, and let $a > m$ be another action. Note that

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

from which it follows that

$$L(\theta, m) - L(\theta, a) \leq (m - a) I_{(-\infty, m)}(\theta) + (a - m) I_{(m, \infty)}(\theta).$$

Since $P(\theta \leq m|x) \geq \frac{1}{2}$, so that $P(\theta > m|x) \leq \frac{1}{2}$, it can be concluded that

$$\begin{aligned} E^{\pi(\theta|x)}[L(\theta, m) - L(\theta, a)] &\leq (m - a) P(\theta \leq m|x) + (a - m) P(\theta > m|x) \\ &\leq (m - a)\tfrac{1}{2} + (a - m)\tfrac{1}{2} = 0, \end{aligned}$$

establishing that m has posterior expected loss at least as small as a. A similar argument holds for $a < m$, completing the proof. □

In the IQ example (Example 1), the posterior was normal with mean $\mu(x)$, which, for a normal distribution, is also the median. Hence the Bayes estimator is the same in this example, whether squared-error or absolute-error loss is used. Indeed, when the posterior is unimodal and symmetric, it can be shown that, for any loss of the form $L(|\theta - a|)$ which is increasing in $|\theta - a|$, the Bayes estimator is the median of the posterior. This partially indicates that, when underestimation and overestimation are of equal concern (merely a verbal statement of the condition that L be a function of $|\theta - a|$), the exact function of $|\theta - a|$ which is used as the loss is not too crucial (i.e., the Bayes rule is robust with respect to this part of the loss). Recall, however, that any weight function, $w(\theta)$, in the loss can have a significant effect.

Perhaps the most useful standard loss is linear loss. The proof of the following result will be left as an exercise.

Result 6. *If*

$$L(\theta, a) = \begin{cases} K_0(\theta - a) & \text{if } \theta - a \geq 0, \\ K_1(a - \theta) & \text{if } \theta - a < 0, \end{cases}$$

any $(K_0/(K_0 + K_1))$-fractile of $\pi(\theta|x)$ is a Bayes estimate of θ.

EXAMPLE 1 (continued). In estimating the child's IQ, it is deemed to be twice as harmful to underestimate as to overestimate. A linear loss is felt to be appropriate, so the loss in Result 6 is used with $K_0 = 2$ and $K_1 = 1$. The $\frac{2}{3}$-fractile of a $\mathcal{N}(0, 1)$ distribution is about 0.43, so the $\frac{2}{3}$-fractile of a $\mathcal{N}(110.39, 61.23)$ distribution (which is $\pi(\theta|x)$) is

$$110.39 + (0.43)(61.23)^{1/2} = 113.97.$$

This is the Bayes estimate of θ.

It should again be emphasized that, while easy to work with, none of the above standard losses need be suitable for a given real problem. If this is the case, it will usually be necessary to calculate and minimize the posterior expected loss numerically. This can be done with a computer or, in many cases, with a programmable hand calculator.

4.4.3. Finite Action Problems and Hypothesis Testing

In estimation there are generally an infinite number of actions to choose from. Many interesting statistical problems involve only a finite number of actions, however. The most important finite action problem is, of course, hypothesis testing.

The finite action Bayesian decision problem is easily solved when considered in extensive form. If $\{a_1, \ldots, a_k\}$ are the available actions and $L(\theta, a_i)$ the corresponding losses, the Bayes action is simply that for which the posterior expected loss $E^{\pi(\theta|x)}[L(\theta, a_i)]$ is the smallest. Several specific finite action problems will now be considered.

In testing H_0: $\theta \in \Theta_0$ versus H_1: $\theta \in \Theta_1$, the actions of interest are a_0 and a_1, where a_i denotes acceptance of H_i. Actually, this is to a degree putting the cart before the horse, since in true decision problems the hypotheses are usually determined by the available actions. In other words, the decision maker is often faced with two possible courses of action, a_0 and a_1. He determines that, if $\theta \in \Theta_0$, then action a_0 is appropriate, while if $\theta \in \Theta_1$, then a_1 is best. While the distinction as to whether the hypotheses or actions come first is important in discussing reasonable formulations of hypothesis testing problems, it makes no difference in the formal analysis of a given problem.

When the loss is "0-1" loss ($L(\theta, a_i) = 0$ if $\theta \in \Theta_i$ and $L(\theta, a_i) = 1$ if $\theta \in \Theta_j, j \neq i$), then

$$E^{\pi(\theta|x)}[L(\theta, a_1)] = \int L(\theta, a_1) dF^{\pi(\theta|x)}(\theta) = \int_{\Theta_0} dF^{\pi(\theta|x)}(\theta) = P(\Theta_0|x),$$

and

$$E^{\pi(\theta|x)}[L(\theta, a_0)] = P(\Theta_1|x).$$

Hence the Bayes decision is simply the hypothesis with the larger posterior probability.

For the more realistic "0-K_i" loss,

$$L(\theta, a_i) = \begin{cases} 0 & \text{if } \theta \in \Theta_i, \\ K_i & \text{if } \theta \in \Theta_j (j \neq i), \end{cases}$$

the posterior expected losses of a_0 and a_1 are $K_0 P(\Theta_1|x)$ and $K_1 P(\Theta_0|x)$, respectively. The Bayes decision is again that corresponding to the smallest posterior expected loss.

It is useful to observe the relationship of these Bayesian tests with classical hypothesis tests. In the Bayesian test, the null hypothesis is rejected (i.e., action a_1 is taken) when

$$\frac{K_0}{K_1} > \frac{P(\Theta_0|x)}{P(\Theta_1|x)}. \tag{4.31}$$

Usually $\Theta_0 \cup \Theta_1 = \Theta$, in which case

$$P(\Theta_0|x) = 1 - P(\Theta_1|x).$$

Inequality (4.31) can then be rewritten

$$\frac{K_0}{K_1} > \frac{1 - P(\Theta_1|x)}{P(\Theta_1|x)} = \frac{1}{P(\Theta_1|x)} - 1,$$

or

$$P(\Theta_1|x) > \frac{K_1}{K_0 + K_1}.$$

Thus in classical terminology, the rejection region of the Bayesian test is

$$C = \left\{ x\colon P(\Theta_1|x) > \frac{K_1}{K_0 + K_1} \right\}.$$

Typically, C is of exactly the same form as the rejection region of a classical (say likelihood ratio) test. An example of this follows.

EXAMPLE 1 (continued). Assume f and π are normal and that it is desired to test H_0: $\theta \geq \theta_0$ versus H_1: $\theta < \theta_0$ under "0-K_i" loss. Noting that $\pi(\theta|x)$ is a $\mathcal{N}(\mu(x), \rho^{-1})$ density, the Bayes test rejects H_0 if

$$\frac{K_1}{K_0 + K_1} < P(\Theta_1|x) = \left(\frac{\rho}{2\pi}\right)^{1/2} \int_{-\infty}^{\theta_0} \exp\left\{\frac{-\rho(\theta - \mu(x))^2}{2}\right\} d\theta$$

$$= (2\pi)^{-1/2} \int_{-\infty}^{\rho^{1/2}(\theta_0 - \mu(x))} \exp\left\{\frac{-\eta^2}{2}\right\} d\eta$$

(making the change of variables $\eta = \rho^{1/2}(\theta - \mu(x))$). Letting $z(\alpha)$ denote the α-fractile of a $\mathcal{N}(0, 1)$ distribution, it follows that the Bayes test rejects

H_0 if

$$\rho^{1/2}(\theta_0 - \mu(x)) > z\left(\frac{K_1}{K_0 + K_1}\right).$$

Recalling that

$$\mu(x) = \frac{\tau^2}{\sigma^2 + \tau^2} x + \frac{\sigma^2}{\sigma^2 + \tau^2} \mu$$

and rearranging terms, gives the equivalent condition

$$x < \theta_0 + \frac{\sigma^2}{\tau^2}(\theta_0 - \mu) - \sigma^2 \rho^{1/2} z\left(\frac{K_1}{K_0 + K_1}\right).$$

The classical uniformly most powerful size α tests are of the same form, rejecting H_0 when

$$x < \theta_0 + \sigma z(\alpha).$$

In classical testing, the "critical value" of the rejection region is determined by α, while in the Bayesian test it is determined by the loss and prior information.

In situations such as that above, the Bayesian method can be thought of as providing a rational way of choosing the size of the test. Classical statistics provides no such guidelines, with the result being that certain "standard" sizes (0.1, 0.05, 0.01) have come to be most frequently used. Such an *ad hoc* choice is clearly suspect in a true decision problem. Indeed even classical statisticians will tend to say that, in a true decision problem, the size should be chosen according to "subjective" factors. This is, of course, precisely what the Bayesian approach does.

There are many decision problems with more than two possible actions. For instance, a frequently faced situation in hypothesis testing is the existence of an indifference region. The idea here is that, besides the actions a_0 and a_1, which will be taken if $\theta \in \Theta_0$ or $\theta \in \Theta_1$, a third action a_2 representing indifference will be taken if $\theta \in \Theta_2$. For example, assume it is desired to test which of two drugs is the most effective. Letting θ_1 and θ_2 denote the probabilities of cures using the two drugs, a reasonable way to formulate the problem is as a test of the three hypotheses H_0: $\theta_1 - \theta_2 < -\varepsilon$, H_1: $\theta_1 - \theta_2 > \varepsilon$, and H_2: $|\theta_1 - \theta_2| \leq \varepsilon$, where $\varepsilon > 0$ is chosen so that when $|\theta_1 - \theta_2| \leq \varepsilon$ the two drugs are considered equivalent.

Even in classical hypothesis testing there are usually three actions taken: a_0—accept H_0, a_1—accept H_1, and a_2—conclude there is not significant evidence for accepting either H_0 or H_1. The choice among these actions is classically made through an informal choice of desired error probabilities. Attacking the problem from a Bayesian decision-theoretic viewpoint (including the specification of $L(\theta, a_2)$) seems more appealing.

Another type of common finite action problem is the classification problem, in which it is desired to classify an observation as belonging to one of several possible categories. An example follows.

EXAMPLE 1 (continued). For the IQ example in which it is desired to classify the child as a_1—below average ($\theta < 90$), a_2—average ($90 \le \theta \le 110$), or a_3—above average ($\theta > 110$), the following losses are deemed appropriate:

$$L(\theta, a_1) = \begin{cases} 0 & \text{if } \theta < 90, \\ \theta - 90 & \text{if } 90 \le \theta \le 110, \\ 2(\theta - 90) & \text{if } \theta > 110, \end{cases}$$

$$L(\theta, a_2) = \begin{cases} 90 - \theta & \text{if } \theta < 90, \\ 0 & \text{if } 90 \le \theta \le 110, \\ \theta - 110 & \text{if } \theta > 110, \end{cases}$$

$$L(\theta, a_3) = \begin{cases} 2(110 - \theta) & \text{if } \theta < 90, \\ 110 - \theta & \text{if } 90 \le \theta \le 110, \\ 0 & \text{if } \theta > 110. \end{cases}$$

(These could arise, for example, if children are put into one of three reading groups (slow, average, and fast) depending on their IQ classification.) Since $\pi(\theta|x)$ is $\mathcal{N}(110.39, 69.23)$, the posterior expected losses are

$$\begin{aligned} E^{\pi(\theta|x)}[L(\theta, a_1)] &= \int_{90}^{110} (\theta - 90)\pi(\theta|x)\,d\theta + \int_{110}^{\infty} 2(\theta - 90)\pi(\theta|x)\,d\theta \\ &= 6.49 + 27.83 = 34.32, \\ E^{\pi(\theta|x)}[L(\theta, a_2)] &= \int_{-\infty}^{90} (90 - \theta)\pi(\theta|x)\,d\theta + \int_{110}^{\infty} (\theta - 110)\pi(\theta|x)\,d\theta \\ &= 0.02 + 3.53 = 3.55, \\ E^{\pi(\theta|x)}[L(\theta, a_3)] &= \int_{-\infty}^{90} 2(110 - \theta)\pi(\theta|x)\,d\theta + \int_{90}^{110} (110 - \theta)\pi(\theta|x)\,d\theta \\ &= 0.32 + 2.95 = 3.27. \end{aligned}$$

(The preceding integrals are calculated by first transforming to a $\mathcal{N}(0, 1)$ density, and then using normal probability tables and the fact that

$$\int_a^b \theta e^{-\theta^2/2}\,d\theta = -e^{-\theta^2/2}\Big|_a^b = e^{-a^2/2} - e^{-b^2/2}.)$$

Thus a_3 is the Bayes decision.

4.4.4. With Inference Losses

In Subsection 2.4.3 we discussed two of the ways in which decision theory can be useful for inference problems. First, many common inference measures can be given a formal decision-theoretic representation. For instance, the choice of "0-1" loss in testing leads to standard Bayesian

testing measures; the posterior expected loss of rejecting a hypothesis is then the posterior probability of the hypothesis. As another example, if

$$L(\theta, C(x)) = 1 - I_{C(x)}(\theta),$$

where $C(x) \subset \Theta$, then the posterior expected loss of $C(x)$ would be the posterior probability that θ is not in $C(x)$. Such formal relationships allow the use of decision-theoretic machinery in solving inference problems.

The second use of decision theory in inference, also discussed in Subsection 2.4.3, was the use of loss functions to represent the actual success of an inference in communicating information. Thus, in Example 4 of Subsection 2.4.3, it was suggested that, since one reports $\alpha(x)$ as a measure of the "confidence" with which it is felt that θ is in $C(x)$, it would be reasonable to measure the accuracy of the report by

$$L_C(\theta, \alpha(x)) = (I_{C(x)}(\theta) - \alpha(x))^2.$$

This loss could be used, decision-theoretically, to suggest a choice of the report $\alpha(x)$. For instance, if $C(x)$ is the credible set to be used in Bayesian inference, it would be reasonable to choose $\alpha(x)$ to minimize the posterior expected loss for L_C. As in the development of Result 3, it is easy to show that the Bayes choice of $\alpha(x)$ is then

$$\begin{aligned} \alpha^{\pi}(x) &= E^{\pi(\theta|x)}[I_{C(x)}(\theta)] \\ &= P^{\pi(\theta|x)}(\theta \in C(x)). \end{aligned}$$

Thus we have been led to the "obvious," that the best (to a Bayesian) report, for the confidence to be placed in $C(x)$, is the posterior probability of the set. More sophisticated (and interesting) uses of inference losses can be found in the references in Subsection 2.4.3.

4.5. Empirical Bayes Analysis

4.5.1. Introduction

An empirical Bayes problem, as discussed in Subsection 3.5.2, is one in which known relationships among the coordinates of the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ allow use of the data to estimate some features of the prior distribution. Such problems occur with moderate frequency in statistics, generally in one of two related situations. The most obvious such situation is when the θ_i arise from some common population, so what we can imagine creating a probabilistic model for the population and can interpret this model as the prior distribution. The simplest version of this situation is when the θ_i are i.i.d. from (the prior) π_0, which is the (perhaps partially