4

The Threshold Problem

4.1 Traditional approaches to comparing Bayes and frequentist estimators

Both Bayesian and frequentist methods of inference have qualities which would seem to recommend them for use. They both also have apparent deficiencies. Both schools can find, without great difficulty, reasons to support the position they have chosen as well as reasons to critique the methodologies espoused by the other school. Many professional statisticians see themselves as being in one camp or the other but, in practice, remain open to using either of the methodologies when a particular application seems to call for them. A common example is a Bayesian's use of the standard methods of linear model theory (regression analysis, for example), because the methodology is so well developed and easily interpretable; he might do so while, at the same time, being quite adamant about the use of the Bayesian approach to estimation and testing in other settings. Similarly, one often encounters staunch frequentists who are happy to use Bayesian methods on occasion (especially those labeled as "objective") because of the enticing computational tools available for executing the Bayesian approach.

Since "convenience" and "feasibility" are ever-present realities in the world of applied statistics, it seems unfair to criticize "crossovers" such as those mentioned above. Still, it is no doubt useful to seek principles that might generally guide one's choices, even while acknowledging that in actual practice, one might occasionally "sin" and sneak in an analysis that's not entirely in line with these principles. In this chapter, we begin with a survey of the varied traditional arguments supporting either the Bayesian or the classical school. In doing so, we will seek to assess whether one or the other position is stronger on a particular point. More importantly, we will explore the question of whether, in their totality, these arguments point to a clear winner. The reader may well conclude, as I have, that, on the basis of traditional means of comparison, neither methodology decisively dominates the other. I will then introduce the "threshold problem," which frames the comparison of these methodologies in a useful, if nontraditional, way. This leads to what we all perhaps knew from the

F.J. Samaniego, *A Comparison of the Bayesian and Frequentist Approaches to Estimation*,
Springer Series in Statistics, DOI 10.1007/978-1-4419-5941-6_4,

beginning: the question we should be asking is not *whether* one approach is better than the other, but *when* one is better than the other.

4.1.1 Logic

At the pinnacle of a Bayesian's defense of his methodology (on the intellectual, if not the practical, level) is the argument based on pure logic. As we have seen, the Bayesian paradigm is based on a system of axioms. The nature of axioms, of course, is that they are "self-evident" truths that one simply accepts as reasonable. There are a variety of versions of the axioms of Bayesian inference (see Fishburn (1986)), and while one can nit-pick about one or another, it is not as easy to summarily dismiss these assumptions. Once an axiom system is accepted and set in place, one has no choice but to accept its consequences. In the case of axioms such as those in Section 3.2, the primary consequence is the basic premise of all Bayesian inference — that the only way to deal with uncertainty in a rational and coherent manner is through the assignment of probabilities to uncertain events. This is a powerful statement, one that merits careful thought by every practicing statistician. It is made all the more powerful by the fact it leads to a method of inference that it is in perfect harmony with the likelihood principle, a statement that, by itself, has drawn many people of better-than-average intelligence into the Bayesian camp. It thus seems reasonable to ask the question: is the logic of Bayesian inference compelling in the sense that the argument about B vs. F is essentially over?

Before addressing this question, let's take a look at the logical underpinnings of frequentist inference. Surprise, there are none! One might think, at first view, that decision theory is an attempt to bring logic to the frequentist approach. As we have seen, however, decision theory serves (primarily) to "organize" the approach rather than to render it "logical." It sets up a framework for thinking about optimality, and it does give clear guidance about procedures that should be avoided. On the other hand, it generally fails in leading us to broadly optimal procedures. Further, a large portion of statistical practice, including the quite heavy use of asymptotic techniques, is simply beyond the scope of decision theory. The best one-word description of the classical school of Statistics is "opportunistic." As seen in Chapter 2, there are a good many distinct frequentist approaches to point estimation (and this characteristic extends to other forms of inference), and the approach one might choose on a given occasion is not infrequently selected because of its analytical or numerical feasibility rather than because of explicit knowledge about its local or global superiority. While asymptotic considerations can suggest some form of "approximate optimality" in the (fixed-sample-size) problem at hand, a definitive answer to the question of "what's best" remains unattained. Frequentist methods are generally motivated by appealing intuitive considerations, but the choice among available methods is virtually always *ad hoc* and tends to be defended on intuitive or practical grounds rather than on some logical basis.

So, does the Bayesian win on the basis of logic? It might seem that the answer is clearly "yes." But there is a nagging worry in closing the deal on this basis alone. Think of the issue this way. Bayesian coherence is really about internal consistency.

As admirable as it might be to be perfectly consistent, we should recognize that this offers no protection from the possibility of being consistently wrong. In the context of point estimation, a Bayesian who is truly dismal at introspection and/or probability elicitation may end up being perfectly coherent but also patently inferior to most reasonable alternatives to estimating unknown parameters. Logical consistency is not enough to guarantee good results in statistical estimation. It thus seems clear that the argument between Bayesians and frequentists can't be settled on the basis of logic alone.

While the discussion above graciously assumes that the axioms of Bayesian inference are unassailable, it seems only fair to mention some potential difficulties. I will mention just one which pertains to the specific axiomatic developments described in Section 3.2. Let us reexamine Axiom 5. Few would find the postulated existence of uniform random variables troublesome. However, the axiom goes beyond this in postulating that one can compare the relative likelihoods of an event A of interest and the event that a random variable $X \sim \mathscr{U}[0,1]$ will take a value in any given interval $I \subset [0,1]$. From this assumption, together with Axioms 1–4, one can easily prove that one may identify $P(A)$ as a unique value $p \in [0,1]$. Interestingly, if we assume the latter fact, one can then prove that one can compare the relative likelihoods of the event A and any arbitrary interval $I \subset [0,1]$. This fact suggests that the conclusion that one must assign probabilities to uncertain events is actually imbedded in the axioms of Bayesian inference, that is, it is itself an axiom rather than a derived result. This of course changes the interpretation of Theorem 3.1. Viewed in this light, the theorem just says that if one is a Bayesian, then one is a Bayesian, clearly diminishing the clout of the result. One might still argue that Axioms 1–5 of Section 3.2 are in fact self-evident. If one accepts this as one's starting point, then there is no question that one should adopt the Bayesian approach to statistical modeling and inference. But one would also have to accept the conclusion that the adoption of the Bayesian approach is a choice rather than a logical imperative. As pointed out by DeGroot (1970), all axiomatic developments of the Bayesian approach require an assumption equivalent to Axiom 5. Thus, there appears to be no way around the circularity of the logical defense of Bayesian inference. The positive contribution made by axiom systems such as the one considered in Chapter 3 is that they make the assumptions behind the Bayesian choice abundantly clear.

4.1.2 Objectivity

There's a popular saw among statistical practitioners: "One should let the data speak for themselves." On these grounds, the frequentists appear to have the upper hand. Subjective Bayesians clearly bring something extra to their data analysis and are prepared to "alter" the inferences that the data themselves might lead to by infusing some subjectively determined "prior information" into the analysis. This could well worry someone interested in the scientific interpretation of the outcome of a planned experiment, as it seems dangerous, and perhaps even unethical, to pepper one's data analysis with one's own subjective opinions. Indeed, in research studies in the sciences, there has been a traditional (though not universal) aversion to the use of Bayesian methods, with a desired “objectivity” given as the primary reason. As a counterpoint to such reservations, Breslow (1990) argued that there was a compelling need for the development of Bayesian approaches in a variety of important statistical problems in the health sciences. The opposing view has been voiced more frequently. Rob Easterling, a good applied statistician of a strongly frequentist persuasion, stated at a conference in 2000 that Bayesian methods leave the door wide open for “statistical mischief.” Efron (1986) famously stated that the frequentist school had clearly staked its claim on “the high road of statistical objectivity.” Both Easterling and Efron are quite right — the subjective Bayesian approach has a potential failing — it allows for the (possibly intentional, though typically unintentional) infusion of misleading information through the use of a prior distribution. The frequentist approach does not have this particular failing, at least not to the extent that the Bayesian does. Regarding objectivity, it is only fair to state that the frequentist approach does contain subjective components, the most obvious of which is the selection of a model for the observable data. But it must also be recognized that the subjective Bayesian will also need to select a model for the data. It thus remains true that the Bayesian brings “more subjectivity” to the analysis of data than does the frequentist.

What about those who do an “objective Bayesian analysis”? In the view of orthodox (coherent) Bayesians, such “objective” approaches are frequentist rather than Bayesian procedures. They contain no subjective input, they are incoherent in the Bayesian sense, and they often lead to the same procedures that would be obtained by standard frequentist approaches. We might agree that the approach involves less subjectivity than the approach an orthodox Bayesian would take, but this “objectivity” has been purchased at the cost of abandoning the opportunity of using subjective input in cases in which it might be quite relevant and useful. In the end, it appears that the classical school (as well as “objective Bayesians”) make a good point — their approaches to point estimation enjoy a greater extent of objectivity than does the orthodox Bayesian approach. But is this really the unassailable virtue that it might seem to be?

In certain (some would say, in many) problems, the opportunity to utilize prior information in a formal way represents a great boon to the statistician and is precisely the vehicle that can guarantee reliable inference. Letting the data speak for themselves may not be the panacea it is often thought to be. A simple, oft-used example makes this point quite unambiguously. Suppose a freshly minted coin is tossed ten times, and we wish to estimate the probability p that represents the chances of the coin coming up heads in any single toss. If we obtain 10 heads in 10 tosses, the standard, universally recommended frequentist estimator of p would be $\widehat{p} = 1$. We all know, however, that p is no doubt close to $1/2$, and if we were to have done a Bayesian analysis, we would probably have used a beta prior like $\mathrm{Be}(100,100)$ which is quite heavily concentrated around its mean $1/2$. When we observe 10 heads in 10 tosses of the coin, we would adjust our prior opinion, and estimate p on the basis of the posterior distribution $\mathrm{Be}(110,100)$, that is, we would estimate p to be 0.5238. While we thought, initially, that the coin was fair, we are in fact affected by the surprising result of the experiment. We no longer believe the coin is fair, but our posterior opinion properly moderates our initial thoughts, resulting in a small estimated bias. As simple as this example is, it reveals an essential truth. A Bayesian analysis here doesn't introduce questionable subjectivity, intentional mischief or any other form of arbitrary alteration of the data. What it does is use some pretty solid prior knowledge to save us from the embarrassment of making a ridiculously poor inference. The take-home lesson here seems to be that the use of prior information can be extremely useful. The challenge in more complex estimation problems is to determine whether or not "useful" prior information is in fact available.

Let's discuss the notion of "useful prior information" further. When can we expect that such will be available? Curiously, it is in technical, scientific investigations that one is likely to be able to identify useful prior information. Why? Because the cumulative experience of researchers and practitioners in various scientific specialties provides substantial intuition regarding the processes that they study. In other words, expert opinion is not a rare commodity in science and engineering, and the elicitation of such opinions stands to put the statistician in an excellent position to produce creditable and effective inferences based on Bayesian methods. Thus, in the very areas in which objectivity is most revered, subjective input into a statistical analysis stands to be the most helpful.

Those who would eschew the use of subjective Bayesian inference in a scientific context will often readily admit that the Bayesian approach causes them much less concern in the context of "decision making." In the problems and issues that are encountered in everyday life, virtually all of us behave like Bayesians. In the practical problems we face in a typical day, we begin by assessing what we know about the problem, we update our prior opinion with whatever current information is available and we reach a conclusion. (Try that model out for yourself the next time you contemplate the possibility of jaywalking.) The difference between decision making and scientific inference is that in the former, we are making personal judgments whose consequences are largely personal rather than public, while in the latter, we seek to advance the general understanding of a scientific problem and therefore, at least implicitly, are asking others to rely on our subjective opinions about the problem. Taking the position that one should not engage in such practices is understandable, but it also might be criticized as perhaps too rigid a position to take as a universal principle.

A quite different "principle" that appropriately governs scientific inference is that one's assumptions should be clearly articulated so that the basis for the inferences drawn is transparent and can be carefully scrutinized. The prior distribution adopted by the Bayesian may properly be viewed as one of the assumptions of his analysis. When so viewed, the questions that remain are whether or not the assumption is reasonable and/or useful. These, of course, are challenging questions, ones which would seem to be quite difficult to resolve. Investigating these questions, and obtaining answers and insights of some practical value in problems of point estimation, constitutes the primary aims of the next four chapters. As we will see, the term "useful prior information" appears to admit to a considerably broader interpretation than has generally been ascribed to it; we shall also see that the term has some natural bounds. These findings lead to the conclusion that an objective (i.e., frequentist) statistical analysis may sometimes, but will not always, lead to superior inference in a given estimation problem, and that the availability of "useful prior information" can give the Bayesian the advantage.

So, is there a winner in the objectivity debate? It seems not. While frequentist estimators can legitimately be said to have captured a little more of the holy grail of "objectivity," one also must recognize that "objectivity" is not in fact sacred and that the subjective elements of a statistical analysis may turn out to be hugely important in producing good answers in some (yet to be characterized) class of problems. So the question of whether to execute an objective or a subjective analysis in a given problem must be considered, at least for now, as an issue requiring further thought and discussion.

4.1.3 Asymptotics

Let's turn our attention to another arena in which frequentists appear to have an edge. The asymptotic theory for a wide variety of frequentist procedures has been fully developed. In the sizable class of "regular" problems in which one wishes to estimate an unknown parameter θ, the maximum likelihood estimator $\widehat{\theta}_{\text{ML}}$ reigns supreme, in an asymptotic sense, being a strongly consistent estimator of the true value of θ, converging to the true θ at an optimal rate and being asymptotically normal with the smallest possible asymptotic variance. These credentials are hard (in fact, impossible) to beat! But they can be tied. In these same problems, Bayes estimates with respect to a large class of prior distributions (that is, priors whose support set is Θ) are asymptotically equivalent to $\widehat{\theta}_{\text{ML}}$, sharing all the good properties mentioned above. While the Bayesian school has not devoted as much attention to asymptotic analysis as has the frequentist school (as is understandable in light of the likelihood principle, which renders such musings irrelevant), it has nonetheless been shown that Bayes procedures tend to have the same asymptotic behavior as the best frequentist alternatives. Further, the concern that the Bayesian approach can lead to strikingly different answers when the prior distributions used in two separate analyses are substantially different is allayed, to a large degree, by the "merging of opinion" literature which indicates that, in typical applications, the difference will shrink to zero as the sample size grows.

As with the considerations in earlier subsections of this chapter, it appears that one cannot declare a clear winner on the basis of asymptotic comparisons. It should be mentioned that the asymptotic behavior of Bayesian nonparametric estimators has been shown to be a bit more spotty, requiring greater care on the part of the Bayesian to ensure good asymptotic performance than is the case in parametric problems. The interested reader is referred to Diaconis and Freedman (1986) for details.

4.1.4 Ease of application

In a 1986 paper, Bradley Efron posed the question "Why isn't everyone a Bayesian?" and he suggested several answers, among which two stood out. The issue of objectivity was one, an issue that seemed to favor the classical school of Statistics. We have discussed this issue above, making note of the proposition that the reliability of an “objective” analysis can at times be questionable. A second characteristic of frequentist procedures that Efron saw as contributing to their popularity was their ease of application. Efron pointed out that many frequentist estimators could be derived in closed form, and that a good deal was known about their behavior, either in fixed sample sizes or asymptotically. Efron’s paper of course predated the computational revolution in the Bayesian community, as the broad implications and impact of Geman and Geman’s 1984 paper had not yet taken hold. It is fair to say that today, the tide has turned, with the intractable integrations of earlier Bayesian treatments replaced by iterative methods aimed at precise approximations of posterior distributions and related quantities. Interestingly, some within the frequentist community have turned to the tools of Bayesian computation to solve problems originating from a frequentist perspective. Efron himself, in his ASA Presidential address in 2005, made note of the convergence of Bayesian and frequentist thinking and opined that “objective Bayesian analysis” would play an increasingly important role in scientific investigations in the decades ahead. So the “ease of application” issue, while hardly being a principle on which one would want to take firm stand, is an issue that is, today, by no means settled, with both frequentist and Bayesian analysis more and more often relying on high-speed computation with ease of application that is reasonably described as quite comparable.

4.1.5 Admissibility

Standard versions of the Complete Class Theorem in Statistical Decision Theory indicate that, in many problems of interest, the class of Bayes and extended Bayes rules is essentially complete. One apparent consequence of the theorem is the fact that, in any problem to which the theorem applies, one may restrict attention to this class since the performance of any decision rule outside the class can be matched or beaten by some rule in the class. There is no equivalent result which applies to a well-known class of frequentist estimators. Is, then, the proper conclusion of the Complete Class Theorem that one might as well be a Bayesian, as the (slightly expanded) class of Bayes rules contains all the decision rules that one would want to use?

There are a number of reasons why this conclusion is less than compelling. The first is simply that, in any nontrivial estimation problem, the (unexpanded) class of Bayes rules is typically not itself complete. Secondly, the collection of extended Bayes rules is not an innocuous addition to the class of Bayes rules. For example, they include decision rules that are incoherent, that is, are not Bayes with respect to any proper prior distribution. Thirdly, we must keep in mind that admissibility (a property that Bayes estimators tend to enjoy) is an extremely weak property. The estimator $\widehat{\theta}(\mathbf{X}) \equiv c$ is Bayes with respect to the prior G that is degenerate at the constant c and is an admissible estimator of θ, but it would never be considered for practical use. While it is true that one would never wish to use an inadmissible estimator, it is also true that the admissibility of an estimator does not provide sufficient justification to recommend it for use. While restricting attention to admissible estimators does make operational sense (since we would automatically toss out an estimator that was inadmissible), this restriction would naturally include frequentist (though extended and generalized Bayes) estimators like $\overline{X}$ as an estimator of a normal mean μ. Typically, the complete class of all admissible estimators in a given problem will contain both Bayes rules and frequentist rules, and restricting attention to one or the other subclass is unjustified. Finally, along the same lines, one should note that there do exist decision problems in which some Bayes rules are inadmissible. In such problems, the complete class with which we started would also contain some decision rules that one would not wish to use.

The most compelling reason for disregarding complete class theorems in a given decision problem is the fact that the quality of the decision rule chosen has very little to do with the class from which it was chosen. The quality of a Bayes rule, for instance, has a good deal to do with whether the prior distribution carries "useful" information about the true state of nature θ. There are good and bad Bayes estimators (measured, say, by how close the answer will be to the true value of the target parameter), so that simply resolving to use a Bayes rule in a particular problem is of no help in identifying a good decision rule.

Exercise 4.1. Suppose the risk set in a particular decision problem is the unit square. Confirm the fact that there are uncountably many Bayes rules, but that only one of them is admissible.

4.1.6 The treatment of high-dimensional parameters

The field of Multivariate Analysis has a storied history and is a well-established subfield within the discipline of Statistics. Until fairly recently, the great majority of this work was frequentist in nature. The early barriers to Bayesian inference in multi-parameter problems were in large measure due to the substantial difficulty of executing Bayesian methods involving many parameters. The analytical difficulties involved in the evaluation of the integrals on which the posterior distribution of the parameters depends were, at best, imposing, and were often completely overwhelming. In fairness, it must also be recognized that the classical approach to multivariate analysis has its limitations. The well-known methods of classical multivariate analysis tend to assume that data follow a multivariate normal model. In continuous problems, rather little has been done with other parametric models, mostly because of the paucity of tractable alternatives to the normal. For example, the most widely used applied statistical methods — regression analysis and the analysis of variance — tend to rely on the assumption of multivariate normality (with special structure). While the utility of such analyses has been proven repeatedly in a wide array of applications in the experimental sciences, the appropriateness of the analyses certainly depends on the modeling assumptions made. The multivariate Central Limit Theorem, and techniques such as "variance stabilizing transformations," may sometimes be used to justify the use of traditional multivariate analysis in large samples, but when serious concerns arise about the normality of the data, frequentists are often reduced to tentative (or descriptive rather than inferential) solutions and *ad hoc* approximations. The one glowing exception to this is the area of discrete multivariate (or categorical) data analysis where impressive analytical and practical advances have been made, largely through the theory and applications associated with generalized linear fixed-effects and mixed-effects models (see McCulloch and Nelder (1989) and Jiang (2007)).

Bayesian multivariate analysis has made substantial strides over the last several decades. On the analytical side, Bayesian treatments of linear models (see Lindley and Smith (1973) and Kadane *et al.* (1980)) have opened the door for Bayesian ANOVA and regression, though the prior modeling in typical applications would clearly benefit from a broadening of options. On another front, the analytical treatment of Bayesian time series and econometric models (see Geweke (2001), for example) has rendered other problems with multidimensional parameters amenable to a Bayesian treatment. But the best news for the Bayesian has been the arrival and maturation of MCMC methods, since now the analytical intractability of a Bayesian analysis in many modeling frameworks may be counterbalanced by reliable iterative methods.

It is clear that a definitive, flexible treatment of multivariate problems continues to be an elusive goal to both Bayesians and frequentists. For frequentists, methods that are applicable beyond a limited array of parametric families remain a challenge, while for the Bayesian, perhaps the greatest challenge is that of developing a meaningful way to identify "useful" prior information on a vector or matrix of parameters. It is rare that one can quantify real prior intuition in a multiparameter problem, and thus, simplifications (like prior independence and flat priors) are commonplace. The consequences (and the unexploited benefits of alternative prior modeling) are not well understood at this point in time, with the efficacy of a Bayesian analysis in such problems and its possible comparative advantage over a frequentist treatment remaining largely unexplored. In the end, both schools can boast some real successes in multivariate analysis, but neither appears to occupy a position of dominance in the area.

4.1.7 Shots across the bow

In the debate between frequentists and Bayesians over the years, each school has discovered examples in which one side looked good while the other looked silly. The Bayesian school has no difficulty finding examples of frequentist methods that are incoherent. Several instances are mentioned in Chapter 3. One on which little has been said, as yet, is the question of extraneous randomization. While randomizing among one's options may seem innocuous, it is clear that it violates the likelihood principle. The Bayesian would argue that it should not be necessary. Suppose that a randomized decision rule δ minimizes the Bayes risk $r(G,\delta)$. Then the associated probability distribution P on the space D of nonrandomized rules can only give weight to rules d for which $r(G,d) = r(G,\delta)$, so that such nonrandomized rules are also Bayes, and the rule δ is not needed. Aside from the ability to set aside randomized rules, the Bayesian may point to occurrences of randomization in frequentist procedures that seem misguided. It is well known in classical hypothesis testing, for example, that randomized tests are sometimes the uniformly best tests of hypotheses H_0 vs. H_1 at a given prespecified significance level α. Suppose you go to your doctor and are tested for skin cancer. Your doctor gets the test results and finds the outcome of your test cannot be resolved at the desired level of significance (say 5%) (perhaps because of the discreteness of the observable random variable). At your next appointment with your doctor, you sheepishly ask him for the results. Your doctor tosses a coin in the air, observes that the outcome is heads, and joyfully proclaims “whew, you don’t have skin cancer at the 5% significance level.” You should be quite happy to hear that, but who among us would not be just a little disturbed by the randomization involved?

While the axiomatic development of Bayesian inference may appear to provide a solid foundation on which to build a theory of inference, it is not without its problems. Suppose, for example, a stubborn and ill-informed Bayesian puts a prior on a population proportion p that is clearly terrible (to all but the Bayesian himself). The Bayesian will be acting perfectly logically (under squared error loss) by proposing his posterior mean, based on a modest size sample, as the appropriate estimate of p. This is no doubt the greatest worry that the frequentist (as well as the world at large) would have about Bayesian inference — that the use of a “bad prior” will lead to poor posterior inference. This concern is perfectly justifiable and is a fact of life with which Bayesians must contend. Unfortunately, being “coherent” is not enough! Being “right,” or very close to right, is also necessary, and in fact, is the more important characteristic in any real statistical application.

We have discussed other issues, such as the occasional inadmissibility of the traditional or favored frequentist method and the fact that frequentist methods don’t have any real, compelling logical foundation. We have noted that the specification of a prior distribution, be it through introspection or elicitation, is a difficult and imprecise process, especially in multiparameter problems, and in any statistical problem, suffers from the potential of yielding poor inferences as a result of poor prior modeling. All of these considerations leave unresolved the question of which school of statistical inference is to be preferred. The “debate” between Bayesians and frequentists, at least as represented by the foregoing commentary, ends up in an uncomfortably inconclusive state. The reader will notice that, while both sides have been rather carefully examined, one specific question has been left untouched. Which method stands to give “better answers” in real problems of practical interest? This is the question to which we now turn, and the question on which much of the remaining content of this monograph is focused.

Exercise 4.2. State, in your own words, the advantages and disadvantages you see in both the Bayesian and the frequentist approaches to estimation. Can you think of any additional pros or cons that have not been mentioned above?

4.2 Modeling the true state of nature

One may take the view that comparisons between frequentist and Bayesian statisticians are contests between two adversaries, each trying to optimize relative to some performance criterion based on an agreed-upon loss function. This is the view that permeates the discussion in Section 4.1 and as we have seen (specifically in problems of point estimation, but by natural inference, more generally), it tends to lead to inconclusive results when examined in ways that I have referred to as "traditional." The purpose of this section is to point out that there's an elephant in the room! (Actually, and more accurately, there's a third "player" in the room.) When this third relevant party is identified and formally dealt with, we will see that the comparison between Bayesian and frequentist estimators may be brought into much sharper focus. We will refer to the third party as the "Truth." It's certainly obvious to anyone interested in comparing two competing estimators that, if only they knew the true value of the target parameter, they would have some compelling evidence in favor of one estimator or the other. At the same time, we know that the "truth" is, and almost always remains, unknown. The luxury of knowing the truth is never really available. Interestingly, it is possible to make some real progress in the comparison of competing estimators by simply positing the existence of an unknown truth and taking its existence into account. I will refer to this latter process as that of "modeling the truth."

Let us focus on an estimation problem, given data $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$ and a fixed loss function $L(\theta, a)$. Suppose that a frequentist statistician is prepared to estimate the unknown parameter θ by the estimator $\widehat{\theta}$ and that a Bayesian statistician is prepared to estimate θ by the estimator $\widehat{\theta}_G$, the Bayes estimator relative to his chosen prior distribution G. How should the "truth" be modeled? I shall, henceforth, consider the true value of θ to be a random variable, and I will call its distribution G_0 the "true prior." Now in many problems of interest, θ is not random at all; it's just an unknown constant. One might refer to such problems as "almanac problems." If we had access to the right almanac, we could just look up the true value of θ. In such problems, it is appropriate to take G_0 to be a degenerate distribution which gives probability one to θ_0, the true value of θ. In other settings, as when θ is the proportion of defective items in today's production lot (a value which varies from day to day), it may be appropriate to consider G_0 to be nondegenerate. In either case, we take G_0 to be a description of "what is," the actual (random or fixed) state of nature. Think of it as God's prior, unknown to us and to the two statisticians who are trying to estimate the parameter θ. Accounting for the unknown state of nature in this way gives no one any particular advantage, as the exact form of G_0 is unknown and unknowable in any real estimation problem. We will nonetheless find that recognizing the existence of G_0 is useful.

Before moving on, I should acknowledge that the notion of a "true prior distribution" is not part of the Bayesian vernacular. To an orthodox (subjective) Bayesian, a prior distribution is simply a summary of his prior opinion about the unknown state of nature before a relevant experiment is performed. As a subjective opinion, it can't be wrong, provided it conforms with his intuition about θ, a fact that is, in general, tacitly assumed. The intuition itself may be misguided, but the prior nonetheless represents the Bayesian's sense of the truth, and must be considered correct from his personal perspective on the problem at hand. The term "true prior," as used above, is a separate quantity that differs from, and is independent of, any particular Bayesian's prior distribution and is not associated with the inference process that any Bayesian would actually pursue. Still, in any problem in which there is an unknown target parameter, the term “true prior” serves the purpose of quantifying the truth about that parameter.

Exercise 4.3. If θ is a random variable rather than a constant, the problem of “estimating” it is usually referred to as a “prediction” problem. Suppose that θ and X are dependent random variables and that you wish to predict θ from an observed X. Show that, when the loss criterion is squared error, the best predictor of θ based on X is the predictor $\widehat{\theta} = E(\theta|X = x)$.

4.3 A criterion for comparing estimators

We now examine the possibility of using the Bayes risk of an estimator, relative to the true prior G_0, as a criterion for judging the superiority of one estimator over another. For a fixed loss function L, the Bayes risk of an estimator $\widehat{\theta}$ with respect to the true prior G_0 is given by $r(G_0, \widehat{\theta}) = E_\theta E_{\mathbf{X}|\theta} L(\theta, \widehat{\theta}(\mathbf{X}))$, where the outer expectation is taken with respect to G_0. While this criterion can be defended for any choice of loss function, we will, for the sake of clarity and simplicity, provide such a defense for the particular choice of squared error loss, that is, for $L(\theta, a) = (\theta - a)^2$.

Let us consider the interpretation of the criterion $r(G_0, \widehat{\theta})$ for each of two statisticians, the frequentist and the Bayesian. In the classical theory of estimation, the choice of squared error loss is not only common but in fact quite prevalent. The mean squared error of the estimator $\widehat{\theta}$ is, without doubt, the criterion that is most widely used in assessing the performance of an estimator. The Bayes risk $r(G_0, \widehat{\theta})$ is simply the mean squared error averaged relative to the objective truth in the estimation problem of interest, and is thus a highly relevant measure of the estimator’s worth. In the most frequently encountered case in which the parameter θ is simply an unknown constant, the Bayes risk $r(G_0, \widehat{\theta})$ is precisely the mean squared error of the estimator $\widehat{\theta}$ evaluated at the true value of θ. In this case, the measure reduces to the most relevant measure of all, the actual and true mean squared error of the estimator. When G_0 is nondegenerate, the measure is equally relevant, as it is the global mean squared error relative to the truth. Setting aside the fact that G_0 is not known, our interest in this measure seems quite appropriate.

If the Bayesian statistician was able to discern the actual true prior G_0, then he would undoubtedly use it in estimating the parameter θ. The estimator $\widetilde{\theta}$ which minimizes the Bayes risk $r(G_0, \widehat{\theta})$, and thus also minimizes the posterior expected loss $E_{\theta|\mathbf{X}=\mathbf{x}} L(\theta, \widetilde{\theta}(\mathbf{x}))$, is the Bayes estimator with respect to G_0 and is thus the very best that the Bayesian could hope for in the problem of estimating θ. Since this scenario is a virtual impossibility, the Bayesian will select a prior G, henceforth referred to as his “operational prior,” in order to carry out his estimation. But how should the quality of this Bayes estimator be judged? The estimator is optimal with respect to the prior G, as it minimizes the posterior expected loss relative to G as well as the Bayes risk $r(G, \widehat{\theta})$. But G is not a representation of the truth; it is, rather, a representation of the Bayesian’s best *a priori* guess at the truth. The Bayes risk $r(G, \widehat{\theta})$ only measures how well the Bayesian did relative to his prior intuition, and, of course, he did very well indeed, minimizing his average risk relative to his chosen prior. How well the Bayesian did relative to the truth is measured, instead, by $r(G_0, \widehat{\theta})$. The Bayesian's estimation process is not driven by the true prior G_0, but there can be no question that an impartial adjudicator would be interested in $r(G_0, \widehat{\theta})$ rather than in $r(G, \widehat{\theta})$, as it is the former measure, rather than the latter, which pertains to how well the Bayesian did in estimating the true value of θ.

One other consideration is worth mentioning. As discussed in Chapter 3, the Bayes risk is a frequentist measure, involving the process of averaging losses over the entire sample space X, which of course includes potential, but unobserved, data values. It is important to recognize that the criterion we are examining has nothing whatsoever to do with how the Bayesian carries out his inference. The Bayesian is expected to obtain an estimator that is coherent in the Bayesian sense. It is only in the evaluation of the Bayesian's performance (taking the true state of nature into account) that the Bayes risk wrt G_0 comes into play. Consider the following allegory.

In a certain benign monarchy, the enlightened King has decided to retain a court statistician to do all of the kingdom's official point estimation. Two highly regarded statisticians apply, one a frequentist, the other a Bayesian. The King proposes that they undertake a series of estimation exercises, the goal of which, of course, is to determine who is likely to do a better job. The King happens to know the characteristics of his subjects well, a result of years of careful study by the King and his closest advisors. Put another way, the King happens to know the exact answers to certain questions (about common characteristics like age, gender, occupation) in advance of any experiments. After agreeing to a model for each experiment, the statisticians jointly design a sampling plan and collect the data from which each question will be answered. They then provide their estimates of each of the parameters of interest. Which of the two is likely to become the court statistician? Certainly the King would be looking for which statistician tended to be closest to the true value of the parameter. If, for example, the frequentist was closer to the target in eight of ten experiments, the King would probably select the frequentist for the available opening. If the experiments were of the same sort, then the average distance between the estimator and the true parameter value could also be a reasonable basis for comparison. Both of these metrics are based on an essential characteristic of the estimation process: closeness of the estimator to the true parameter value. The Bayes risk $r(G_0, \widehat{\theta})$ is the quintessential measure of closeness to the truth. If the two statisticians, on day one, before seeing any data, simply submitted their estimators of choice (formulaically) to the King, the measure $r(G_0, \widehat{\theta})$ would serve the King well in making his selection between the two competitors.

Finally, it should be mentioned that the general Bayes risk criterion has proven useful in certain Bayesian contexts. For example, if a Bayesian finds himself in the position of having to choose an estimator in a somewhat automated fashion, that is, before any experimental data is available for inspection, then the Bayes risk $r(G, \widehat{\theta})$ is a logical criterion for making a selection. This has been acknowledged in the Bayesian literature. Such a process was referred to as "pre-posterior analysis" by Lindley (1972).

Let us now examine the question of whether the proposed criterion, the Bayes risk $r(G_0, \widehat{\theta})$ with respect to the true prior G_0, is one that is fair to both the Bayesian and the frequentist if one were to use this criterion in comparing the performance of their estimators. Neither statistician is privy to the actual distribution G_0, so both are equally disadvantaged by not knowing it. Performance relative to the "truth" is certainly an important measure to both statisticians (or at least it should be), but it is most assuredly an important measure to their clients or to anyone with any interest in the estimation problem with which the two statisticians are engaged. If the Bayesian happens to be good enough or lucky enough to choose a prior that is, in some sense, close to G_0, then the Bayesian is likely to achieve a level of performance that is superior to that of the frequentist. But that is as it should be, since the selection of a prior distribution is an extremely important part of the Bayesian's inference process, and Bayesians who do that selection well should rightly be rewarded for it. On the other hand, the frequentist has nothing to fear in subjecting his inference to the criterion $r(G_0, \widehat{\theta})$, as it simply represents a generalized form of his estimator's mean squared error, being the squared error of his estimator averaged over all the randomness in the problem or, in many cases, the mean squared error of his estimator evaluated at the true value of the target parameter.

4.4 The threshold problem

I will now define the essence of the approach to be taken in the comparison of Bayesian and frequentist point estimators. I'll begin with a general treatment of the threshold problem and then turn to a special case in which we will be especially interested. I will assume, as before, that the distribution of the available data **X** has a known form indexed by a parameter θ (which, for now, may be thought of as either scalar or vector-valued), and that a loss function L has been specified. In the preceding section, I have argued that the Bayes risk $r(G_0, \widehat{\theta})$ of a point estimator $\widehat{\theta}$ with respect to the true prior distribution G_0 is a reasonable and meaningful measure of the estimator's performance. Now consider the class $\mathscr{G} = \{G\}$ of all possible prior distributions that a Bayesian might use in deriving a Bayes estimator $\widehat{\theta}_{\mathrm{G}}$ of θ. By the "threshold problem," we will mean the problem of determining the boundary which divides the class $\mathscr{G}$ into the subclass of priors for which

$$r(G_0, \widehat{\theta}_{\mathrm{G}}) < r(G_0, \widehat{\theta})\,, \tag{4.1}$$

where $\widehat{\theta}$ represents a given frequentist estimator, from the subclass of priors for which

$$r(G_0, \widehat{\theta}_{\mathrm{G}}) > r(G_0, \widehat{\theta})\,. \tag{4.2}$$

As formulated above, the threshold problem may seem entirely intractable. Reasons for this include (i) the class $\mathscr{G}$ is enormous and not analytically manageable, (ii) the problem is defined in terms of a particular frequentist estimator $\widehat{\theta}$, and so that, for any given estimation problem, there is not just one threshold problem to consider but a sizable collection of them, (iii) even if particular threshold problems were solvable (for different estimators), it seems quite likely that the solutions would vary from one version to another (as the frequentist estimator of choice varies), so that a "global" solution (that is, one which characterizes priors which satisfy (4.1) for all frequentist estimators $\widehat{\theta}$ under consideration) might be difficult to identify or might not lend a great deal of insight and (iv) the true prior distribution is unknown and any solution of (4.1) will not only depend on the particular G_0 considered but may not be meaningful in light of our inability to specify what G_0 actually is. All these are imposing difficulties, and together, they would seem to render the general threshold problem as both an unrealistic and unmanageable abstraction. Can any headway be made on the problem? It is perhaps somewhat surprising that the answer is "yes." To gain entrée into the problem, we will need to modify it, rendering it less abstract, more manageable and, ultimately, solvable. Further, as we shall see, solutions to the versions of the threshold problem to be considered in the sequel turn out to lend considerable insight, notwithstanding the fact that the true prior G_0 is unknown.

Although the following reformulation of the threshold problem is applicable to the estimation of vector-valued parameters, I will, for simplicity, initially present the new formulation for estimators of a scalar parameter. I will also make some additional simplifications. Let's assume that our data consist of a random sample from a distribution indexed by θ, that is, assume that $X_1, X_2, \dots, X_n \overset{iid}{\sim} F_\theta$. Further, let's suppose that the distribution F_θ belongs to an exponential family. Finally, let L be squared error loss, and let $\mathscr{G}$ be the class of standard conjugate priors corresponding to the distribution F_θ. These restrictions are not absolutely necessary to make the threshold problem well defined and manageable, but they will suffice in doing so. Now, when we consider the dual outcomes represented by (4.1) and (4.2), a number of simplifications are possible.

Regarding the existence of a whole host of possible frequentist estimators to be considered, we will be able to restrict attention to just one, the estimator $\widehat{\theta}$ that I will refer to as the "best frequentist estimator." Our ability to restrict attention to $\widehat{\theta}$ derives from the fact that exponential families are endowed with complete sufficient statistics, and for the usual target parameters of interest, UMVUEs generally exist. Not only are these the typical frequentist estimators of choice in such problems, all the alternative reputable estimators are one and the same; that is, the same estimator arises whether one approaches the problem by finding theUMVUE, the MME, the MLE, the BLUE or the LSE of θ. Thus, one can consider the apparent host of threshold problems defined by (4.1) and (4.2) to be equivalent to a single basic problem. In any situations in which such equivalence fails to hold, the solutions to the threshold problem considered in the sequel apply, specifically, to the unbiased estimator $\widehat{\theta}$ that is a sufficient statistic for θ.

As we have seen, the standard conjugate families to exponential families of sampling distributions are families indexed by a fixed number of parameters. Thus, the characterization of conjugate priors for which (4.1) holds reduces to a search over a finite-dimensional space of prior parameters. Thirdly, under squared error loss (and selected alternatives), Bayes estimators with respect to conjugate priors take particularly simple closed-form expressions, and the calculation of their Bayes risk is generally straightforward. What results from these assumptions are the manageable forms of the threshold problem whose solutions are treated in detail in Chapters 5, 6, 7 and 8. In Chapter 5, we consider the estimation of a scalar parameter. In Chapter 6, we treat a common version of the consensus problem in Bayesian estimation, that is, the problem of estimating the scalar parameter θ of an exponential family when prior opinions are elicited from several experts, each inclined to place a different prior distribution on θ. In that context, we obtain a solution of the threshold problem which compares a particular subclass of "consensus estimators" to the best frequentist estimator. In Chapter 7, we consider the quintessential multivariate estimation problem, namely, the estimation of the mean of a multivariate normal distribution, and we treat the threshold problem revolving around the comparison of frequentist and Bayesian shrinkage in that context. In Chapter 8, we consider the threshold problem in the more general setting in which the loss function is asymmetric.

Exercise 4.4. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathscr{B}(1, p)$. Derive the MLE, the MME, the BLUE and the LSE of the parameter p.

Exercise 4.5. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathscr{N}(\mu, 1)$. Derive the MLE, the MME, the BLUE and the LSE of the parameter μ.