Springer Series in Statistics

Francisco J. Samaniego

A Comparison of the Bayesian and Frequentist Approaches to Estimation

Springer

Springer Series in Statistics

Advisors:
P. Bickel, P. Diggle, S. Fienberg, U. Gather,
I. Olkin, S. Zeger

Springer Series in Statistics

For other titles published in this series, go to
www.springer.com/series/692

Francisco J. Samaniego

A Comparison of the Bayesian and Frequentist Approaches to Estimation

Springer

Francisco J. Samaniego
Department of Statistics
University of California
1 Shields Avenue
Davis, CA 95616
USA
fjsamaniego@ucdavis.edu

ISSN 0172-7397
ISBN 978-1-4419-5940-9 e-ISBN 978-1-4419-5941-6
DOI 10.1007/978-1-4419-5941-6
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010929747

Dedication

To my family: *Mary*, my compass, for forty years of love and support; *Monica* and *Elena*, who have brought me nothing but love and enormous pride; *Keb*, for his friendship and contagious positive outlook; *Jack* and *Will*, for the joy they constantly bring to their Papa's life; and my sister *Margarita*, whose constant encouragement, since we were toddlers, gave me the courage to dream impossible dreams and seek to make them a reality;

and,

To three friends who are primarily responsible for sparking my interests in Bayesian Statistics: *Thomas Ferguson*, my teacher and mentor in graduate school and beyond; *Dennis Lindley*, whose visit to Davis as a Regent's Professor in the 1980s really rocked my boat; and *Nozer Singpurwalla*, whose creativity and generosity did much to sustain and expand these interests. One qualification: each of them should be absolved of any responsibility for the views and opinions put forward in this monograph.

Preface

The main theme of this monograph is "comparative statistical inference." While the topics covered have been carefully selected (they are, for example, restricted to problems of statistical estimation), my aim is to provide ideas and examples which will assist a statistician, or a statistical practitioner, in comparing the performance one can expect from using either Bayesian or classical (aka, frequentist) solutions in estimation problems. Before investing the hours it will take to read this monograph, one might well want to know what sets it apart from other treatises on comparative inference. The two books that are closest to the present work are the well-known tomes by Barnett (1999) and Cox (2006). These books do indeed consider the conceptual and methodological differences between Bayesian and frequentist methods. What is largely absent from them, however, are answers to the question: "which approach should one use in a given problem?" It is this latter issue that this monograph is intended to investigate.

There are many books on Bayesian inference, including, for example, the widely used texts by Carlin and Louis (2008) and Gelman, Carlin, Stern and Rubin (2004). These books differ from the present work in that they begin with the premise that a Bayesian treatment is called for and then provide guidance on how a Bayesian analysis should be executed. Similarly, there are many books written from a classical perspective. Prominent among these are the texts by Ferguson (1967) and Lehmann and Casella (1998). These books do treat Bayesian methods, but not from the comparative perspective to be taken here. My aim is to present both approaches to estimation from the perspective of a disinterested third party, someone who is open to executing either method in a given problem but would like to give serious thought to the questions of which might be preferable, and why. Robert's (2001) book, *The Bayesian Choice*, has similarities to the present work in that the author seeks to determine whether one should be a Bayesian or a frequentist. That book uses a decision-theoretic framework to motivate the author's conclusion that one should choose to be a Bayesian. The main difference between our books is that I come to a different conclusion! But the difference is a nuanced rather than an obvious one: my conclusion might be summarized as: one should "often," but not always, choose to be a Bayesian. My goal is to shed some light on the nature of the dividing line separating

Bayesian analyses which tend to be superior to frequentist alternatives and Bayesian analyses that tend to be inferior. Throughout, the criterion for making the recommended choice is performance based. In short, I will seek to describe the types of problems in which the Bayesian tends to have the advantage in estimating an unknown parameter. As the reader can infer from that statement, the frequentist tends to have the advantage in complementary cases.

In this monograph, we will focus our attention on the fundamental statistical problem of point estimation. Although multiparameter, multivariate problems will be discussed, and certain asymmetric frameworks will be treated, we will fix the main ideas in our comparative analysis of Bayesian and frequentist estimation by considering, first, the problem of estimating an unknown scalar parameter θ. The estimators considered are based on a random sample drawn from a population of values X whose distribution is indexed by θ. We initially will take "squared error" as the criterion for comparing two estimators. The frequentist or classical school of Statistics has a long history of attacking this problem with some success, with "least squares," "minimum distance," "method of moments," "minimum variance unbiased" and "maximum likelihood" estimators collectively playing a major role in the way experimental data has been analyzed for over a century. The Bayesian school is also deeply rooted in statistical history, dating back, at least, to the posthumous publication in 1763 of the Reverend Thomas Bayes' influential "Essay Towards Solving a Problem in the Doctrine of Chances." While these two methodologies have the same basic goal, the philosophical and practical distance between them is enormous, perhaps even enormous squared!

This monograph begins with a review of the fundamental ideas and notation of Statistical Decision Theory. This provides an avenue for introducing Bayesian estimation methods in the context of "decision making under uncertainty." Separate chapters follow on the basic elements of the frequentist and the Bayesian approaches to estimation. We then embark upon the comparison of frequentist and Bayesian estimators in a variety of statistical contexts. Chapter 4 reviews the traditional arguments made in favor of one method or against the other and arrives at the position that the overall argument is inconclusive, with both methods having certain potential advantages but also certain failings. The "threshold problem," the problem of identifying the boundary between the circumstances in which Bayes estimators tend to outperform frequentist estimators and the complementary circumstances in which the opposite is true, is then introduced. A criterion for comparing Bayes and frequentist estimators is proposed and it is argued that it is natural, relevant and sensible; it also serves to make the threshold problem well defined. One-parameter problems estimated under squared error loss are considered in Chapter 5, and an explicit solution to the threshold problem, applicable to exponential families of sampling distributions and conjugate families of prior distributions, is presented. The surprising breadth of the class of Bayes estimators which dominate frequentist competitors is noted, and the characteristics of prior modeling which can provide the Bayesian with an advantage (as well as those which tend to be unfavorable to the Bayesian) are discussed in detail. In Chapter 6, the notion of conjugacy is further explored, both in the light of the concept of Bayesian self-consistency and as a tool in treating the Bayesian

consensus problem. In Chapter 7, the treatment is extended to the estimation of a vector-valued parameter. More specifically, a multivariate version of the threshold problem is developed for comparing Bayesian and frequentist shrinkage. Generalizations of the threshold problem to estimation under an asymmetric loss function are considered in Chapter 8.

Chapter 9 deals with special topics in which the Bayesian viewpoint is essential in the development of solutions. While the frequentist approach is ill-suited for handling models with nonidentifiable parameters, Bayesian methods are applicable and are amenable to careful study. In Chapter 9, the efficacy of Bayes estimators of nonidentifiable parameters is examined through a concrete example in which a fully Bayesian version of the threshold problem is treated for a nonidentifiable Binomial model. Nonparametric estimation in the context of competing risks and the estimation of the parameters of a nonidentifiable stress-strength model in reliability are also discussed.

In Chapter 10, both Bayesian and frequentist estimation are treated in contexts similar to the classical empirical Bayes framework in which one seeks to learn from *similar* past experiments in the process of estimating an unknown parameter in a current experiment. Prescriptions are given for improving upon a Bayes estimator and for improving upon a frequentist estimator when data are available from one or more past experiments satisfying the empirical Bayes sampling assumptions. It is shown that such an improvement is always possible. In Chapter 11, we examine estimation problems in which data are available from several *related*, rather than *similar*, experiments. In the context studied, it is shown that the strategy of borrowing strength from past experiments provides an avenue for improved estimation in the "current experiment."

The final chapter contains a summary and synthesis of the main themes of the monograph and provides a general set of conclusions and recommendations regarding the types of problems in which the Bayesian approach to estimation stands to provide reliable and preferred solutions (distinguishing them from the types of problems in which they don't). The chapter concludes with comments on open problems of interest and promising directions for future research.

Who is the intended audience for this monograph? I see the target audience as potentially quite broad. The minimal prerequisite for understanding its contents is a one-year, calculus-based undergraduate course in probability and mathematical statistics. The first three chapters of the monograph consist of a review and overview of decision-theoretic concepts and the basic tools and ideas of frequentist and Bayesian estimation. An appendix contains a list of standard univariate models with the parameterizations used in the book. Cumulatively, this material is intended to make the monograph relatively self-contained. Naturally, readers with more advanced training and experience in statistical theory and practice will be better prepared to appreciate the more subtle or technical aspects of the comparative analyses considered. To make the monograph suitable as a text on Bayesian methods or on comparative statistical inference, I have included a collection of exercises of varying degrees of difficulty. These can be easily augmented by problems from related texts or problems of interest to the instructor. The monograph, with or without augmenta-

tion, would be appropriate as a text either for an advanced undergraduate course or for a graduate-level course or seminar. In the latter context, it might serve as the text for a capstone course which addresses a type of comparative analysis that would generally not be covered in standard graduate-level offerings on classical or on Bayesian methods. A one-quarter course can be based on Chapters 1–10 and 12, with a light discussion (say, one lecture apiece) on the highlights of Chapters 7, 8 and 12. A manual with solutions to a healthy selection of the book's exercises is available from the publisher for instructors who adopt the monograph for a course of any size.

Beyond its potential use as a text, it is my hope that professional statisticians, be they academics or practitioners, will find the monograph stimulating and of use as a resource in the area of comparative statistical inference. The monograph is especially aimed at statisticians who are open to using either frequentist or Bayesian methods in selected problems, but would like to have a defensible basis for using one or the other. But "steadfast" Bayesians and "steadfast" frequentists should also find ample food for thought in these pages.

I would like to express my appreciation to Dr. Harry Chang, and more generally, to the Army Research Office, for the moral and financial support they have offered throughout the development of this monograph. The ARO has supported both this project, much of the research that preceded its writing and the new research findings that it contains. I also gratefully acknowledge the support I received from the University of California, Davis, for a sabbatical leave during which much of the monograph was composed. I am indebted to Barry Arnold, Richard A. Johnson, and a third (anonymous) reviewer who contributed greatly to the improvement of the initial version of the monograph. I thank Michael McAssey for innumerable helpful questions and insightful comments over the past year, and also for his assistance in putting the manuscript into its final form. Finally, I thank the students in my course on Bayesian inference at UC Davis in Winter Quarter, 2010, for their help in improving the penultimate version of the monograph.

Francisco J. Samaniego
Davis, California
March 2010

Contents

1

Point Estimation from a Decision-Theoretic Viewpoint

1.1 Tennis anyone? A glimpse at Game Theory

True story: I was out for lunch with two friends recently. I didn't care what restaurant we went to, but my friends John and Marsha had strong preferences, one for Indian food and the other for Chinese. When it became clear that neither one was going to yield to the other in a reasonable amount of time, I proposed to settle the argument by picking a random digit between 1 and 9 (using the random number generator on the fancy-dan cell phone I always carry on my belt) and having each of them try to guess its value. Whoever was closest to my number would get to choose the restaurant. John made the gentlemanly but ill-advised gesture of letting Marsha guess first. Marsha immediately guessed "5" and guaranteed herself an advantage, since no matter what John guessed, there were at least five out of nine numbers that she would be closer to than John. When John unexpectedly won the game, he made the gentlemanly but ill-advised gesture of choosing Marsha's restaurant. It was John's misfortune to end the day with a nontrivial case of food poisoning. Fortunately, Marsha and I managed to dodge that bullet.

Games, including formal ones like chess, Scrabble and blackjack, informal ones like the guessing game above and athletic contests like tennis or golf, are part of American culture (and many others) and pop up with some frequency in our daily lives. While these games each involve some strategizing, most of us don't go to the trouble to think out the best possible available strategy. Indeed, some games are sufficiently complex that the "optimal" strategy is either unknown or quite difficult to implement. But the general principles of *Game Theory* are worth knowing and keeping in mind. Playing in general conformance with these principles will usually keep us from getting trounced.

In this section, we will review the basic elements of the Theory of Games. *Decision Theory* can be viewed as an extension of *Game Theory*, and its foundations are firmly grounded in the well-known treatises by von Neumann and Morganstern (1944), Wald (1950) and Blackwell and Girshick (1954). The usual starting point is a two-person, zero-sum game in which two rational opponents each seek to maximize their gains (or, equivalently, minimize their losses). The "zero-sum" feature of such

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_1,

games simply stipulates that in a given trial of the game, nothing is lost to a third party, so that the gain experienced by one player is equal in magnitude to the loss experienced by the other. In a given game, each player may choose from a set of available actions and each experiences the corresponding and complementary gain or loss. It will be convenient, for future developments, to denote the action spaces for players 1 and 2 as Θ and A, respectively. The game is well-defined as soon as a particular real-valued loss function $L : \Theta \times A \to R$ is specified. The value of $L(\theta, a)$ is interpreted as the amount that player 2 loses, that is, pays to player 1, when player 1 chooses action θ and player 2 chooses action a. A negative loss for a particular play represents a gain for player 2.

As simple as the framework above may seem, a good deal of complexity may arise in the analysis of a particular game. The subjects of primary interest are whether the game favors a particular player and whether the game has a fixed worth or "value" V. The existence of V means that both players have strategies which ensure that, on average, player 2 can't lose more than the amount V and player 1 can't gain more than V. There are games in which no such number exists, though the so-called Fundamental Theorem of Game Theory asserts that (most) finite games, that is, most games in which Θ and A are finite sets, do have a finite value V and that both players have strategies that guarantee that each can attain the outcome V. The basic issues arising in games with finite action spaces are entertainingly explained by Williams (1954), and the mathematical elements of Game Theory are nicely surveyed by Ferguson (1967). The 2×2 game with the loss function specified in Table 1.1 below is an example of a game in which each player has a strategy which bounds his opponent's loss (or gain) at a common value V.

Table 1.1. A game with value $V = 5$

		Player 2	
		a_1	a_2
Player 1	θ_1	3	4
	θ_2	5	6

You'll notice that, in the game defined by the table above, player 1's gain is at least 3 when he chooses action θ_1 and is at least 5 when he chooses action θ_2. He can maximize his minimal gain by taking action θ_2; this action is thus appropriately called his *maximin* strategy. Thinking along similar lines, player 2 will lose at most 5 when she takes action a_1 and will lose at most 6 when she chooses action a_2. She can minimize her maximal loss by taking action a_1; this action is thus appropriately called her *minimax* strategy. Since it is apparent that player 1 will gain at least 5 in this game while player 2 will lose no more than 5, the game clearly has the value $V = 5$. Games with the properties exhibited here are said to have a "saddle point," a condition that can be summarized by the equation

$$\max_{\theta_i} \min_{a_j} L(\theta_i, a_j) = 5 = \min_{a_j} \max_{\theta_i} L(\theta_i, a_j) \; . \tag{1.1}$$

Games such as these are easy to analyze. In fact, the analysis above can be replaced by the simple observation that, for player 1, the action θ_2 is uniformly more profitable than the action θ_1, so that player 1 will clearly take action θ_2. Knowing this, player 2 will surely choose action a_1. The outcome is thus preordained to be a payoff of 5 to player 1.

Most games are not quite this simple. The game of "odd or even" (drawn from Ferguson's (1967) text) nicely illustrates the additional complexity that typically arises in finite games. Each of two players will simultaneously show either one or two fingers (in the same fashion that the game "rock, paper, scissors" is played), and the payoff varies, favoring player 1 if the sum is odd and player 2 if the sum is even. The payoff matrix is displayed in the table below, where the subscript of each available action represents the number of fingers shown.

Table 1.2. A payoff matrix for the game "odd or even"

		Player 2	
		a_1	a_2
Player 1	θ_1	-2	3
	θ_2	3	-4

It is clear by inspection that the game of odd or even does not have a saddle point, that is, for this game

$$\max_{\theta_i} \min_{a_j} L(\theta_i, a_j) \neq \min_{a_j} \max_{\theta_i} L(\theta_i, a_j) \,. \tag{1.2}$$

The new element that the analysis of this game requires is the notion of a *mixed strategy*. This type of "action" involves randomization achieved through the process of placing a probability distribution of the set of available actions. Note that if player 1 uses a "mixed strategy" which chooses actions θ_1 and θ_2 with probabilities 7/12 and 5/12, respectively, player 2 will lose, on average, the amount 1/12 to player 1, regardless of the strategy player 2 employs. This implies that $V \geq 1/12$. On the other hand, if player 2 chooses her strategy according to probability distribution $(7/12, 5/12)$ on the actions a_1 and a_2, then player 2's expected loss will be 1/12, regardless of the action that player 1 chooses. This implies that $V \leq 1/12$. These strategies are thus the maximin and minimax strategies for the two players, and the value of the game is $V = 1/12$.

Exercise 1.1. Consider the two-person, zero-sum game with the payoff matrix pictured below:

		Player 2	
		a_1	a_2
Player 1	θ_1	3	6
	θ_2	5	4

Find the minimax strategy for player 2, the maximin strategy for player 1, and the value of the game.

1.2 Experimental data, decision rules and the risk function

The extension from games to decision problems involves the inclusion of additional information in the form of a statistical experiment. There is also an essential change in the view taken toward the two players. In a decision-theoretic setting, the role of player 2 is taken by "the statistician" whose aim is to minimize her losses to player 1, while player 1 is seen as "nature" whose action space now represents the possible states of nature that may be operational at a given point in time. Nature is not seen as a rational opponent, but rather as an administrator who chooses an action θ in an impartial manner (much like Mother Nature may choose to let it rain tomorrow).

The new element in a decision problem is that the statistician is able to observe an experiment which contains some information about the choice that nature has made. (To continue with the analogy above, the statistician might gather "weather data" on regions west of her location and use that information to formulate a guess as to whether or not it will rain tomorrow). Let us suppose that the datum available to the statistician consists of a (possibly vector valued) random variable X which takes values according to a distribution F_θ which depends on the state of nature θ that is in effect when his decision must be made. Often, the available data can be appropriately modeled as n independent and identically distributed (i.i.d.) observations, in which case we will write $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$. Unless otherwise stated, this latter assumption is made throughout the remainder of the present chapter. We refer to the set of values that a particular observation X may take on as the sample space, and we will denote it by $\mathscr{X}$. The vector of observations (or, in more general, non-i.i.d. settings, the data set available to the statistician) will be denoted by $\mathbf{X}$.

Assume that nature has chosen its "state." Once $\mathbf{X}$ has been observed, the statistician will choose an action from his action space A. The decision problem is then resolved by referring to the specified loss function. The statistician's process of selecting an action is equivalent to the selection of a decision rule d. In the i.i.d. case highlighted above, the vector of observations is drawn from the Cartesian product $\mathscr{X}^n$. Thus, the decision rule selected by the statistician is simply a mapping $d : \mathscr{X}^n \to A$ which identifies a desired action for any given observation the statistician might make. In the game defined by Table 1.2 above, note that both players were "allowed" to choose an action at random according to some probability distribution on their respective action spaces. Randomized decision rules are "allowed" in the same way, that is, they are available to the statistician if she cares to use them. Let us denote by D the space of all "nonrandomized" decision rules (that is, mappings from $\mathscr{X}^n$ to A), and let D^* be the set of all "randomized" decision rules (that is, probability distributions on the space D). The latter may be viewed as the space of all probability distributions on D. For a detailed treatment of randomized decision rules, see Ferguson (1967).

In Section 1.1, we alluded to the loss function L, one of the essential elements of game theory. It was evident from the game defined by Table 1.2 that, even in resolving a game and determining its value, we may need to evaluate an "expected loss," that is, a loss function averaged over a probability distribution on the action space. Decision problems involve an additional element of randomness, this being due to the experimental data available to the statistician. We will thus need the following extension of the framework discussed thus far. When the statistician uses the decision rule $d \in D$ and observes the data $\mathbf{X} = \mathbf{x}$, she will employ the action $d(\mathbf{x}) \in A$ and, if the true state of nature is θ, she will incur the loss $L(\theta, d(\mathbf{x}))$. But what might the statistician lose, on average, in the overall process of using experimental data to reach a decision? The expected loss, averaged over all possible outcomes of the experiment, weighted by their appropriate likelihood, is called the *risk function* $R(\theta, d)$ and is defined by

$$R(\theta, d) = E_{F_\theta} L(\theta, d(\mathbf{X})) \, . \tag{1.3}$$

The expectation in (1.3) has the usual interpretation, being a weighted sum for discrete $\mathbf{X}$ and an integral relative to the distribution of $\mathbf{X}$ for continuous $\mathbf{X}$. If $\delta \in D^*$ is a randomized decision rule corresponding to the probability distribution P_δ on D, then the risk function of δ is given by

$$R(\theta, \delta) = E_{P_\delta} E_{F_\theta} L(\theta, d(\mathbf{X})) = E_{P_\delta} R(\theta, d) \, . \tag{1.4}$$

In discussing decision problems further, we note that the space D^* of randomized decision rules contains all degenerate distributions on D and thus may be seen as properly containing D. We will thus focus, for now, exclusively on the space D^*. The risk function of a decision rule δ can be thought of as its primary measure of merit in the decision problem of interest. Comparing the risk functions of two decision rules may help us determine whether one is better than the other. The decision rule δ_2 is said to be *inadmissible* if there exists a rule $\delta_1 \in D^*$ such that

$$R(\theta, \delta_1) \leq R(\theta, \delta_2) \quad \text{for all } \theta \in \Theta \, , \tag{1.5}$$

with strict inequality in (1.5) for at least one θ. It is clear that one would never wish to use an inadmissible rule δ in a given decision problem. Not only might it be considered bad etiquette, it would make no sense, as there exists a decision rule whose expected performance is always as good as δ and is, for at least some θ, actually better than δ. If no "better rule" than δ exists (in the sense above), then δ is said to be *admissible*. Admissibility is a good property for a decision rule δ to have, in the sense that it is a whole lot better than being inadmissible. It is, nonetheless, a very weak endorsement of δ, since all it says is that there is no decision rule that is always as good or better. Clearly this does not preclude the possibility that an admissible rule will be terrific, indeed uniquely best, for one particular value of θ or some small set of θ values and have quite poor, or even unspeakably gruesome, performance for other values of θ. There are many "admissible" rules that one wouldn't be caught dead using. So the concept is mostly useful in the other direction, that is, in the exclusion from consideration of decision rules which are inadmissible.

Let us return to the example of the game of odd or even discussed in the preceding section. Again, following Ferguson (1967), suppose that player 1 is viewed as "Nature," a player that selects the action θ_1 or θ_2 without any particular strategy in mind. Suppose, further, that the experimental datum X, with the following probability distribution depending on the true value of θ, is available to player 2 (the statistician):

$$P(X = 1 \mid \theta = \theta_1) = 3/4 \text{ and } P(X = 2 \mid \theta = \theta_1) = 1/4\,, \tag{1.6}$$

while

$$P(X = 1 \mid \theta = \theta_2) = 1/4 \text{ and } P(X = 2 \mid \theta = \theta_2) = 3/4\,. \tag{1.7}$$

It is clear that the experiment is informative, as when we observe $X = 1$, we would be inclined to believe that $\theta = \theta_1$, while when $X = 2$ is observed, we would tend to believe, instead, that $\theta = \theta_2$. Given an observation $X \in \{1, 2\}$, the statistician has four decision rules $d_i \in D$, $i = 1, \ldots, 4$, from which to choose:

$$d_1(1) = 1,\ d_1(2) = 1; \quad d_2(1) = 1,\ d_2(2) = 2;$$

$$d_3(1) = 2,\ d_3(2) = 1; \text{ and } d_4(1) = 2,\ d_4(2) = 2.$$

The decision rules d_1 and d_4 ignore the outcome of the experiment altogether, and are not expected to be very good. The risk functions $(R(\theta_1, d),\ R(\theta_2, d))$ for the four decision rules are $(-2, 3)$, $(-3/4, -9/4)$, $(7/4, 5/4)$ and $(3, -4)$, respectively, for d_1, d_2, d_3 and d_4. From this, it is clear that the rules d_1, d_2 and d_4 are admissible.

The risk function seems like the perfect measure of "goodness" of a decision rule, as it captures the expected performance of the rule across all values that the state of nature θ may take on. While it clearly does serve this purpose, it is also true that risk functions are quite imperfect as tools for comparing pairs of decision rules. Except in the fairly rare circumstance in which the inadmissibility of a seemingly good rule can be established, the comparison of the risk functions of two "reasonable" decision rules of interest tends to be inconclusive, with the first rule dominating the second for some values of θ and the second rule dominating the first of other values of θ. The fact is that, in most situations in which one wishes to compare a pair of decision rules, the rules turn out to be incomparable. It will often be the case that both decision rules are admissible. What are we to do in circumstances like that? We'll investigate this question in the next section.

Exercise 1.2. In the decision problem above based on the game of odd or even and supplemented by the experiment modeled in (1.6) and (1.7), show that the randomized decision rule $d_0 \in D^*$ which selects d_1 with probability 3/13 and d_2 with probability 10/13 is the unique minimax rule for the statistician, achieving the smallest possible maximum risk of $-27/26$. (**Hint:** Graph the "risk set," that is, the smallest convex set containing the risk points of the four nonrandomized rules. Decision rules whose risk points are on the lower boundary of the risk set are the admissible rules in this problem. Show that the rule d_0 has risk point $(-27/26, -27/26)$ and that the risk point of any other rule $\delta \in D^*$ has a larger maximum.)

1.3 Point estimation as a decision problem; approaches to optimization

While the question posed above can be pondered in the context of any decision problem of interest, it seems most reasonable to consider the question in the particular context that will be the center of attention in the remainder of this monograph. The problem of point estimation is quite well understood in statistical theory and practice, so that its embedding in a decision-theoretic context may seem to be, to those well versed in estimation theory, something akin to gilding the lily. But the main issues that interest us, that is, answers to the questions "why is the risk function not enough?" and "what can be done about that?" surface with special clarity in the context of point estimation. So this is the problem to which we will now turn.

Most problems of point estimation involve a continuous parameter θ (the problem of estimating the unknown size N of a finite population being an interesting exception). In most problems arising in practice, the parameter to be estimated lies in a specified interval of real numbers, the three intervals $[0,1]$, $(0,\infty)$ and $(-\infty,\infty)$ being the most common, as they arise, for example, in the problems of estimating a proportion, an expected lifetime or the expected "profit" of an investment. Viewing a point estimation problem as a game, one would stipulate that Nature's action space Θ is an interval of real numbers. Further, since the aim of point estimation is to "guess" the exact value of θ, the statistician's action space A would be taken to be that same interval. Indeed, it is the equality $\Theta = A$ that defines a decision problem as one of estimation.

Let us suppose, for simplicity, that the spaces Θ and A are both the entire real line. The problem of interest might be that of estimating the mean of a normal distribution. The decision rules available to the statistician are literally countless, as the statistician is free to use literally any formula based on the data $\mathbf{X}$ in guessing the value of the parameter. It is customary to denote a particular decision rule in such problems as $\widehat{\theta}$, suppressing its dependence on $\mathbf{X}$ (except when needed for clarity's sake). For this decision problem to be fully defined, one must specify the loss function to be used. In estimation problems, it is natural for the loss to be a function of the distance between the true value of the parameter θ and its estimated value a. The most widely-used loss criterion in one-parameter estimation problems is "squared error loss," that is,

$$L(\theta,a) = (\theta - a)^2 \,. \tag{1.8}$$

Squared error loss is a symmetric function that penalizes overestimation and underestimation equally, and takes the value zero when the estimate is right on target. Squared error of course predates decision-theoretic thinking, and has been used for many generations, going back, at least, to Gauss and his theory of least squares. It has an obvious connection to measures like "variance" and "mean squared error" in standard estimation theory, and has a long history of useful application. There are, of course, many alternatives to this choice. Among them, the absolute error loss function $L(\theta,a) = |\theta - a|$ may be considered as a reasonable alternative, but mathematical analysis based on absolute error is often substantially less tractable than that

based on squared error. Both loss functions capture the basic idea that one is penalized, in a symmetric fashion, for the distance between an estimator and its target.

There are many versions of asymmetric loss functions. A V-shaped loss function with different slopes on either side of the minimum may be appropriate in certain problems. One of the more widely used forms of asymmetric loss is the Linex (or, "linear-exponential") loss function, the most common version of which is given by

$$L(\theta,a) = \exp\{c(a-\theta)\} - c(a-\theta) - 1\,, \tag{1.9}$$

where $\exp A = e^A$ and c is a fixed and known constant. The Linex loss function, in the form above, has the following properties: The loss function is equal to zero when $a = \theta$, and is otherwise positive, so it achieves its minimum value at $a = \theta$. Further, the Linex loss function may be written as $h(\Delta) = \exp\{c\Delta\} - c\Delta - 1$, which is a convex function for $\Delta \equiv (a-\theta) \in (-\infty,\infty)$, is decreasing for $\Delta \in (-\infty,0)$ and increasing for $\Delta \in (0,\infty)$. When c is positive, $h(\Delta)$ grows exponentially in positive Δ, but behaves approximately linearly for negative values of Δ. Thus, when $c > 0$, the Linex loss function imposes a substantial penalty for overestimation. The mirror image of this asymmetry may be achieved by setting $c < 0$.

If two estimators of θ are to be compared relative to a chosen loss criterion, the obvious "first-order" comparison would be to examine their risk (or expected loss) functions side by side. A typical comparison of this type is pictured in Figure 1.1.

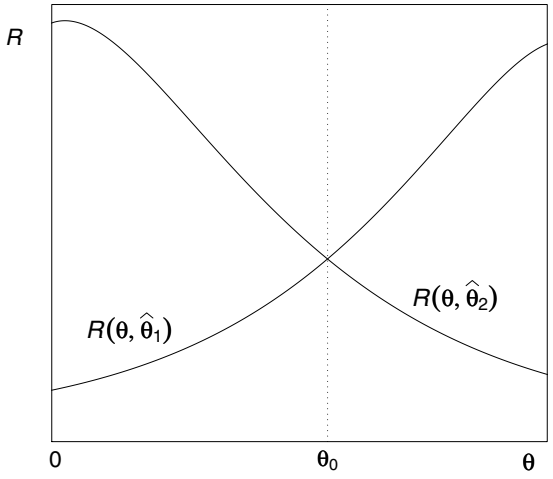

Fig. 1.1. The risk functions $R(\theta,\widehat{\theta}_1)$ and $R(\theta,\widehat{\theta}_2)$

The comparison pictured in Figure 1.1 is by no means uncommon. Shown is a situation in which the estimator $\widehat{\theta}_1$ is better than $\widehat{\theta}_2$ for values of $\theta < \theta_0$, with the opposite being true if $\theta > \theta_0$. Unless the value of θ is known (which of course it is not,

by definition, in any estimation problem of interest), one cannot determine which estimator will provide a better expected performance. The case of inadmissibility provides the sole circumstance in which one estimator may be declared superior to another on the basis of the comparison of their risk functions. In most problems of practical interest, comparisons based on risk functions will be inconclusive. This holds true even in extreme situations when we know that using a particular estimator would be highly ill-advised. Consider, for example, the problem of estimating a location parameter μ (which we may think of as the unknown population mean). Suppose $\mathbf{X}$ is a vector of i.i.d. observations drawn from the distribution F_μ. One possible estimator of μ is the decision rule $\delta(\mathbf{X}) = 10$. This decision rule ignores the data altogether and, in all circumstances, estimates μ to be 10. Assuming, for simplicity, that the chosen loss criterion is squared error, the risk function of such a rule is the quadratic $R = (\mu - 10)^2$. This function shoots off to ∞ as the distance between μ and 10 grows. In contrast to this estimator, estimators of a location parameter μ with bounded risk functions can generally be found, and such estimators would be far better choices than the estimator $\widehat{\mu} = 10$, except, of course, for a small interval of values of μ near the number 10. In spite of the fact that the estimator $\widehat{\mu}$ above is highly risky (please excuse the pun), it is an admissible estimator of μ and cannot be excluded from consideration purely on the basis of a risk function comparison.

The question that naturally arises at this juncture is: are there reasonable ways of ranking available estimators, ways that would lead to a unique estimator (or group of estimators) that could be described as optimal in a given statistical setting? The type of problem we face here is not uncommon in mathematical work. Risk functions are complex objects, and it should not be surprising that one might be unable to rank estimators solely on the basis of these functions. Many pairs of estimators would necessarily be judged to be incomparable on that basis. Risk functions provide a *partial ordering* among estimators, and while it is true that some estimators have uniformly smaller risk functions than others, there will be, in many statistical problems, a host of reasonable estimators, none of which can be judged as better than any of the others on the basis of their risk functions. What is required, or at least desired, is a *total ordering* among estimators. If, for example, every estimator could be assigned a numerical score on some reasonable basis, then the estimator with the "best" score could legitimately be referred to as optimal. In spite of the fact that this seems like a tall order, there are several approaches to meeting this requirement. Our discussion of optimality will be fairly brief, but the basic ideas on the formulation of total orderings among estimators, or among certain restricted classes of estimators, will reveal themselves in the following illustrations.

One approach to achieving a total ordering among all estimators is to utilize the so-called *minimax principle*. Suppose that a given estimator $\widehat{\theta}$ has risk function $R(\theta, \widehat{\theta})$. Then one might consider the numerical summary

$$R(\widehat{\theta}) = \max_{\theta} R(\theta, \widehat{\theta})$$

to be a sensible measure of the performance of the estimator. Estimators for which $R(\widehat{\theta})$ is finite might well be preferred over estimators for which $R(\widehat{\theta})$ is infinite.

This, for example, seems like a good reason to set an estimator like $\widehat{\mu} = 10$ aside in the example discussed above. What would one judge to be best using this criterion? An estimator $\widehat{\theta}_{\text{mm}}$ is said to be the *minimax estimator* if it has the smallest possible maximum risk, that is, if

$$\max_{\theta} R(\theta, \widehat{\theta}_{\text{mm}}) = \min_{\widehat{\theta}} \max_{\theta} R(\theta, \widehat{\theta}) \, . \tag{1.10}$$

Minimax estimators need not be unique. When they are, they are the unambiguously best estimator according to the minimax principle. If this criterion yields a set of estimators with the same maximal risk, then they would be considered equivalent in the minimax sense. In either case, one would have identified one or more estimators that have this optimality property.

It should be recognized that the minimax criterion is very conservative, protecting the statistician from the worst possible outcome. One might say that it identifies an estimator corresponding to the best worst case. The criterion does not necessarily lead to an excellent estimator. However, in all but quite pathological problems, it will lead to estimators that are admissible, a conclusion which follows from the fact that if a minimax estimator was inadmissible, the estimator that beats it would also be minimax. If a minimax estimator is unique, it is necessarily admissible.

Another approach to obtaining a total ordering among estimators would be to consider some specific weighted average of risk functions which yields a numerical score. This approach is, in fact, equivalent to the derivation of Bayes estimators. We will treat Bayesian estimation in some detail in Chapter 3, and we will thus limit the discussion here to some basic ideas and notation.

Suppose that G is a probability distribution on the parameter space Θ and that the parameter θ receives weight according to the distribution G. In other words, suppose θ is treated as a random variable with distribution G. Although the risk functions R of competing estimators (or, more generally, competing decision rules) may not be comparable when examined over the whole parameter space, the weighting of the parameter space according to a chosen G yields a single numerical score for every estimator and thus leads to a total ordering among them. In Bayesian theory and practice, the distribution G is called the *prior distribution* on θ, and is viewed as a summary of the statistician's opinion about θ prior to the execution of the planned statistical experiment.

If we assume that $\theta \sim G$ and we expect to observe the data $\mathbf{X}_1, \ldots, \mathbf{X}_n \mid \theta \overset{iid}{\sim} F_\theta$, the *Bayes risk* of the decision rule δ with respect to the prior distribution G is given by

$$r(G, \delta) = E_G R(\theta, \delta) \, , \tag{1.11}$$

where R is the risk function of δ given in (1.4). The Bayes rule δ_G with respect to G is the decision rule that minimizes $r(G, \delta)$, that is, the rule given by

$$\delta_G = \operatorname*{argmin}_{\delta \in D^*} r(G, \delta) \, . \tag{1.12}$$

From a decision-theoretic perspective, the Bayes rule δ_G represents the optimal decision rule relative to a certain weighting (represented by G) that the statistician has

chosen to ascribe to $\theta \in \Theta$. As with all formulations of total orderings in statistical problems, the ordering associated with the Bayes risk relative to a fixed G comes with a price. In this case, the price is associated with the assumption that G is an appropriate weighting for θ. What if G is a poor representation of the truth? The best decision rule relative to a weighting G that has little or no relation to reality may be quite a poor choice in the application of interest. This suggests that in using a Bayes rule in a given problem, the prior G should be chosen with care and the effects of the chosen G, as compared with other plausible choices of a prior distribution, should be scrutinized.

On a more positive note, the Fundamental Theorem of Decision Theory (usually called *the Complete Class Theorem*) roughly states that, under specific conditions, the (slightly expanded) class of Bayes rules in a given decision problem contains a decision rule that is as good or better than any decision rule outside the class. This would seem to suggest that restricting attention to Bayes rules in such problems should suffice, as any rule one would wish to use belongs to this class. In the context of point estimation, the Bayes risk of the estimator $\widehat{\theta}$ will be written as $r(G, \widehat{\theta})$. It is discussed in greater detail in Chapter 3, as are the various issues raised above, and many others.

A different sort of optimality theory may be devised by starting with a (hopefully reasonable) restriction on the estimators to be considered in a given problem. Perhaps the best-known example of this approach is based on the concept of unbiasedness. Given the available data $\mathbf{X}$ from an experiment indexed by the unknown parameter θ, an estimator $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$ is said to be an *unbiased estimator* of θ if

$$E\widehat{\theta} = \theta \, . \tag{1.13}$$

Since the estimator $\widehat{\theta}$ is a function of the random outcome of a statistical experiment, it is itself a random variable. The probability distribution of $\widehat{\theta}$ is generally referred to as its sampling distribution. The unbiasedness property simply stipulates that the target parameter θ is the mean of the sampling distribution. One can think of this property as simply saying that the estimator is properly calibrated, that is, it is aimed at the right place. The property is not a logically compelling one; this is because most people would agree that being close to the target is more important than being aimed at the target. An archer whose arrows are symmetrically distributed on the outer ring of a round target would not be judged to be better than an archer whose arrows are always in the inner ring, albeit not symmetrically distributed around the center. Still, since a universally best estimator (that is, one whose risk function beats all others) will not exist in any nontrivial statistical problem, imposing the unbiasedness restriction would seem, at least at first view, to be a reasonable way to proceed.

Consider, now, the class of unbiased estimators. Among unbiased estimators, it seems sensible to prefer estimators that are close to the target over estimators that are not. Suppose we compare the risk functions of estimators that are unbiased. We are then led to a simplified and quite appealing "new" measure of merit — the estimator's variance. If $\widehat{\theta}$ is an unbiased estimator of θ, then its risk function, under squared error loss, reduces to its variance, that is,

$$R(\theta,\widehat{\theta}) = E(\widehat{\theta}-\theta)^2 = V_\theta(\widehat{\theta})\,. \tag{1.14}$$

Variance functions are still curves, of course, and as we have seen, comparing curves will often lead to inconclusive results. In the present context, however, there is an important collection of problems in which a best unbiased estimator can be found. The theory associated with *uniformly minimum variance unbiased estimators* (UMVUEs) will be discussed in the next chapter.

There are a variety of other ways of restricting the class of estimators in a given problem. The approaches that have been found to be useful in leading to "optimal" estimators under particular restrictions are the ones that end up receiving some space and discussion in standard texts in Mathematical Statistics. Prominent among the popular methods of estimation within restricted classes are *best linear unbiased estimators* (BLUEs), when best unbiased estimators are out of reach, and *best invariant estimators*. These methods of estimation are also discussed in Chapter 2.

Exercise 1.3. LetΘ, D^* and L be the parameter space, the space of randomized decision rules and the loss function, respectively, in a given decision problem. Prove the following claims:

(a) The identity $\min_\delta \max_\theta R(\theta,\delta) = \min_\delta \max_G r(G,\delta)$ holds, where G represents a proper prior on θ. (**Hint:** In one direction, keep in mind that $r(G^*,\delta) = R(\theta^*,\delta)$ for the prior G^* that places probability 1 on θ^*.)

(b) If a decision rule $d \in D^*$ is Bayes with respect to a proper prior G, and

$$R(\theta,d) \le r(G,d) \quad \text{for all } \theta \in \Theta\,,$$

then d is minimax.

Exercise 1.4. Let $X \sim \mathscr{B}(1,p)$, the Bernoulli distribution. A nonrandomized decision rule d may be represented as the pair (a,b), where $a = d(1)$ and $b = d(0)$. Find values of a and b for which the rule d is minimax. Assume $L(p,a) = (p-a)^2$. (**Hint:** Write $R(p,d)$ as a function of a and b and find values for which the risk function is independent of p. Show that this rule is Bayes with respect to a particular Beta prior on p. The minimaxity of d then follows from the well-known fact: If an equalizer rule is Bayes, it is minimax.)

Exercise 1.5. Consider a decision problem based on the elements Θ, A, L and $X \sim F_\theta$. Let G be a prior distribution on the parameter θ, and suppose that the Bayes rule δ_G with respect to G is unique (that is, uniquely minimizes $r(G,\delta)$ among $\delta \in D^*$). Show that δ_G is admissible.

Exercise 1.6. Let X be a random variable with finite mean μ and variance σ^2. Show that, under squared error loss, the decision rule $d(x) = ax + b$ is inadmissible as an estimator of μ if the constant a exceeds 1.

Exercise 1.7. A decision rule whose risk function is constant is called an *equalizer rule*. Show that if an equalizer rule δ is Bayes with respect to a prior distribution G whose support set is the entire parameter space Θ, then δ is admissible under either of the following conditions: (i) Θ is finite, or (ii) for any $\delta \in D^*$, the risk function $R(\theta,\delta)$ is continuous in θ.

Exercise 1.8. Consider a decision problem with $\Theta = [0,1]$, $A = [0,1]$ and $L(\theta,a) = (\theta - a)^2$. Suppose that $X \sim \mathscr{B}(1,\theta)$, the Bernoulli distribution with parameter θ, and that the prior distribution on θ is $G = \mathscr{U}[0,1]$. A nonrandomized decision rule d in this problem is determined by the pair (a,b), where a and b both take values in $[0,1]$ and represent your estimates of θ based on the observation X, with $d(0) = a$ and $d(1) = b$. Calculate the Bayes risk of the rule d as a function of a and b and derive the Bayes rule d_G with respect to the prior G.

2

An Overview of the Frequentist Approach to Estimation

2.1 Preliminaries

The frequentist will often make the assumption that the available data is a random sample of i.i.d. variables. DeGroot (1988) articulates the view that the i.i.d. assumption is logically untenable. Are i.i.d. observations really possible? Are the conditions under which we toss a coin several times ever truly identical? Of course the answer is no. In general, identical trials are physically impossible. If repeated trials were *truly identical*, wouldn't we necessarily obtain the same result in each trial? These criticisms notwithstanding, experience suggests that the i.i.d. assumption is often an excellent approximation to reality, and, in many statistical contexts, making this assumption is relatively harmless. While one might take issue with the i.i.d. assumption in modeling the available data, this is not a central issue in the disagreements between frequentists and Bayesians. Both schools will often make this assumption when describing the data available for study. In this section, we will make the assumption, for the sake of simplicity and clarity, that the available experimental data satisfies an i.i.d. assumption and may thus be represented as the random sample $X_1, X_2, \ldots, X_n \sim F_\theta$.

There is a second and more fundamental reason that Bayesians might express concern about independent, identically distributed observations. This concern is not so easily dismissed. As will be seen repeatedly as the present chapter unfolds (and is, of course, quite well known), classical methods tend to be judged on the basis of their theoretical average performance. For example, under squared error loss, the risk function of an estimator $\widehat{\theta}$ is simply $E(\widehat{\theta} - \theta)^2$, its *mean squared error* (MSE), and frequentist estimators are often compared on the basis of their MSEs. The MSE is interpreted as the squared error one would expect, on average, in many identical trials of the experiment. The Bayesian school takes the position that such averages are inappropriate for judging the merits of statistical procedures. One quite compelling reason supporting this position is that identical repetitions of an experiment are impossible, thus rendering the frequentist criterion an unrealistic abstraction. Further, the Bayesian paradigm includes adherence to the *likelihood principle*, to be discussed in Section 3.6, which postulates that one's statistical inferences should depend on an experiment only through the data that are actually observed. On the issue of ap-

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_2,

propriate criteria for judging the value of a statistical procedure, the only apparent resolution would seem to be that Bayesians and frequents must simply agree to disagree. The Bayesian school will eschew the process of averaging loss functions over the sample space of an experiment, while frequentists will defend the practice as a theoretical analysis that represents a reasonable approximation of reality. Since our interest here will revolve around the comparative performance of estimators, however derived, as judged by an impartial third party, we will not be forced to take sides in this particular debate.

In the problem of estimating the parameter θ, frequentists will typically select between two approaches for identifying viable estimators: (i) optimizing relative to a risk-based criterion for a fixed sample size n or (ii) optimizing relative to some asymptotic measure of performance (as $n \to \infty$). In what follows, I will present a brief overview of the main frequentist options under each of these viewpoints. I present no proofs here, and my presentation is not meant to be comprehensive, but I do intend to survey the main ideas, methods and jargon of frequentist estimation in preparation for the comparative analyses pursued in later chapters. We will touch on the standard approaches to restricting the class of estimators considered and the most commonly used asymptotic methods, and we will briefly discuss the issue of robustness. For a detailed treatment of estimation theory from the classical perspective, see Bickel and Doksum (2001), Ferguson (1967) or Lehmann and Casella (1998).

2.2 Minimum variance unbiased estimators

An unbiased estimator $\widehat{\theta}$ of a parameter θ is defined above by the property in equation (1.13). Unbiasedness is one of a variety of intuitively appealing *ad hoc* conditions that might be placed on an estimator. Once the restriction to unbiased estimators has been made, the search for the best such estimator commences. The standard theory associated with the behavior of unbiased estimators leads to the identification of an optimal estimator in certain special circumstances. The statistical concepts of sufficiency and completeness play important roles in this theory.

A statistic $T = T(\mathbf{X})$ is said to be a *sufficient* statistic for the unknown parameter θ if the conditional distribution of the data $\mathbf{X}$, given $T = t$, does not depend on θ. Typical examples of sufficient statistics in one-parameter problems include the sample proportion $\widehat{p}$ of successes based on a random sample of Bernoulli trials $X_1, X_2, \ldots, X_n$ with common distribution $\mathscr{B}(1, p)$, where p is the probability of success in a single trial, and the sample mean $\overline{X}$ based on a random sample from a normal population with unknown mean and known variance. The maximum observation $X_{(n)}$ from a random sample with an underlying uniform distribution $\mathscr{U}[0, \theta]$ is also a sufficient statistic for θ. While $\widehat{p}$ and $\overline{X}$ are also unbiased estimators of their respective target parameters, $X_{(n)}$ is a biased estimator of θ in the uniform example. A sufficient statistic carries all the information about the unknown parameter that the data $\mathbf{X}$ themselves contain, so that all inference about the parameter can and should be based on that statistic. It is often the case that a simple transformation of a biased sufficient statistic will be both sufficient and unbiased for the parameter of interest,

so that, when restricting to unbiased estimators, one would typically consider the transformed statistic for further investigation. In the uniform example, the statistic $T = \frac{n+1}{n} X_{(n)}$ is both sufficient and unbiased for θ.

A sufficient statistic $T = T(\mathbf{X})$ is said to be *complete* for the parameter θ if whenever the equation $E_\theta g(T) = 0$ holds for a given function g, then $g(T) = 0$ with probability one. Completeness ensures that there is only one function of the sufficient statistic T that is unbiased for θ, for if there were two, their difference would have expected value 0, so the two estimators would, in fact, be one and the same. Some authors use the term *completeness* in a different but equivalent way. One may speak of a family of distributions $\mathscr{F} = \{F_\theta, \theta \in \Theta\}$ as being a complete family. We will take this language as conveying the fact that there exists a statistic T based on one or more observations $X \sim F_\theta$ such that T is a complete sufficient statistic for $\theta \in \Theta$.

Now suppose we are searching for a "good" estimator within the class of unbiased estimators. Any two unbiased estimators would have distributions with mean θ, that is, they would both be "aimed" at the right place. In many repetitions of the sampling process, both estimators would have an average value that would be very close to θ. The preferred estimator, however, would be the one that tends to be closer to the target parameter. One would naturally prefer the estimator with the smaller variance, as its average squared distance from θ would be smaller than that of the other estimator. In general, one would seek the estimator with the smallest possible variance. Three theoretical results that generally come into play is this search are the following.

Theorem 2.1 (Rao–Blackwell). *Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$, and that T, a function of $X_1, \ldots, X_n$, is a sufficient statistic for θ. Let $S = S(\mathbf{X})$ be an unbiased estimator of θ, and define the estimator $\widehat{\theta} = \widehat{\theta}(T) = E_\theta(S|T)$. Then the estimator $\widehat{\theta}$ is unbiased for θ, and*

$$V_\theta(\widehat{\theta}) \leq V_\theta(S) \text{ for all } \theta \in \Theta. \tag{2.1}$$

That $\widehat{\theta}$ is a legitimate estimator of θ follows from the sufficiency of T. The unbiasedness of $\widehat{\theta}$ follows from the identity $E(X) = E(E(X|Y))$ and the variance inequality follows from the identity $V(X) = E(V(X|Y)) + V(E(X|Y))$, both valid for arbitrary random variables X and Y under the quite mild assumptions that the expectations in these equations exist and that the order of the sums or integrals implicit in them may be interchanged. The Rao–Blackwell Theorem tells us that, in searching for good unbiased estimators in a given problem, one needn't go beyond those that are functions of the sufficient statistic T, as any other unbiased estimator may be replaced, and generally improved upon, by an unbiased estimator based on T.

Theorem 2.2 (Lehmann–Scheffe). *Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$, and that T, a function of $X_1, \ldots, X_n$, is a complete sufficient statistic for θ. Let $\widehat{\theta} = h(T)$ be an unbiased estimator of θ. Then, for arbitrary $\theta \in \Theta$, $\widehat{\theta}$ has the smallest possible variance among all unbiased estimators of θ, that is, $\widehat{\theta}$ is the uniformly minimum variance unbiased estimator (UMVUE) of θ.*

The Lehmann–Scheffe Theorem gives us an explicit recipe for finding the best unbiased estimator: identify a sufficient statistic T, confirm that it is complete and then find an unbiased function of T. The result will be the UMVUE. While recipes are, in general, nice to have, this one comes with a couple of sidebars. First, it should be mentioned that, when a sufficient statistic T is complete, there can only be one unbiased function of it, so the final step involves finding the "best" estimator in a class of size 1. Second, the completeness of a sufficient statistic is not something that is pervasive in statistical estimation problems. For example, the smallest and largest observations $X_{(1)}$ and $X_{(n)}$ in a sample of size n from the uniform distribution $\mathscr{U}[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$ are sufficient for the parameter θ, but the pair $(X_{(1)}, X_{(n)})$ is not a complete sufficient statistic for θ. It is nevertheless useful to have a concrete process which generally leads to the best unbiased estimator. The routine works in a context that arises with some frequency in statistical work, that is, when sampling from a distribution belonging to an "exponential family." An elementary version of this modeling concept, applicable to one-parameter models, is briefly discussed in the following paragraphs.

A family of distributions $\{F_\theta, \theta \in \Theta\}$ is said to be an *exponential family* if each member has a density function or probability mass function of the form

$$f_\theta(x) = c(\theta)h(x)e^{\sum_{i=1}^{k} r_i(\theta)t_i(x)}, \quad \theta \in \Theta. \tag{2.2}$$

Exponential families have a number of notable properties. For example, these families are complete. Given a random sample $X_1, X_2, \ldots, X_n$ from a distribution belonging to an exponential family, the statistic $T = (\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i))$ is a complete, sufficient statistic for θ. Moreover, the complete sufficient statistic T is of fixed dimension, independent of the sample size n. Further, T itself has a distribution belonging to an exponential family. An important distinguishing property of exponential families is that the support set of densities within a given family, that is, the set $\{x \mid f_\theta(x) > 0\}$, does not depend on the parameter θ. Many popular models may be written in the form (2.2) and thus enjoy the properties above. Examples include the binomial, negative binomial, Poisson, normal, gamma, and beta distributions. The multivariate normal, multinomial and Dirichlet distributions are examples of models to which the natural multivariate, multiparameter extension of (2.2) applies.

One further theoretical result is often presented in discussions of unbiased estimation. The Cramér–Rao inequality provides a lower bound on the variance of unbiased estimators in a given problem. The potential utility of such a result is immediately evident. If one has the lower bound in hand, and if one finds an unbiased estimator whose variance is equal to that bound, then the estimator is, of necessity, the best unbiased estimator. The Cramér–Rao Inequality holds under a set of conditions on the model which is assumed to govern the available random sample. These are generally referred to as regularity conditions. The reader is referred to Lehmann and Casella (1998) for exact statements. All we shall say about them here is that they include three particularly important requirements; first, that the support set of the model is an open interval of real numbers which is independent of θ, second, that the expectations to be discussed below exist, and third, that one may pass derivatives

under integral signs as needed, that is, that the equation

$$\frac{\partial}{\partial\theta}\int s(x)f_\theta(x)\,\mathrm{d}x = \int s(x)\frac{\partial}{\partial\theta}f_\theta(x)\,\mathrm{d}x \tag{2.3}$$

holds for any integrable function s. Under the so-called Cramér–Rao regularity conditions, the inequality below, and a variety of other results in the classical theory of estimation, can be shown to hold. It is of special interest to note that exponential families of probability distributions satisfy these regularity conditions.

Before presenting the Cramér–Rao inequality and giving an example of its use, we take a brief digression to define another fundamental notion in statistical theory. Consider a probability distribution F_θ with density or probability mass function $f_\theta(x)$. The *Fisher Information* $I_X(\theta)$, reflecting the information content about θ in the single observation $X \sim F_\theta$, is defined by

$$I_X(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\ln f_\theta(X)\right]^2\right\}. \tag{2.4}$$

Students of Statistics, upon seeing this definition for the first time, often find it quite intimidating. Indeed, there is nothing intuitive in the formula, and few would recognize it as an essential measure of how much your data tells you about the parameter of your model. The fact that Ronald Fisher saw it as such is one of many reflections of his genius. Under the Cramér–Rao regularity conditions, it is easily shown that $I_X(\theta)$ may be calculated by the following alternative formula:

$$I_X(\theta) = -E\left\{\frac{\partial^2}{\partial\theta^2}\ln f_\theta(X)\right\}. \tag{2.5}$$

The following simple example helps to see why "information" is a good name for the quantity in (2.4). Suppose you have a single observation from a normal population, that is, suppose that $X \sim \mathcal{N}(\theta,\sigma^2)$. One may then easily verify that

$$\frac{\partial^2}{\partial\theta^2}\ln f_\theta(X) = -\frac{1}{\sigma^2},$$

from which it follows that $I_X(\theta) = 1/\sigma^2$. Now consider the value of X as a piece of information about θ. When the variance of X is small, its information content about θ is large, as X will no doubt be quite close to θ under such circumstances. The reverse is true when the variance of X is large. A measure which tracks when an observation provides precise rather than imprecise information about an unknown parameter stands to be useful in statistical work. If $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$, it is easy to confirm that $I_{\mathbf{X}}(\theta) = nI_X(\theta)$, a fact that demonstrates that the more data you collect, the more information you have regarding the unknown parameter θ and the more precision you can expect in estimating it.

Theorem 2.3 (Cramér–Rao). *Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$, a probability distribution with density or probability mass function $f_\theta(x)$ for all θ in some open interval*

of real numbers Θ. Suppose, further, that the model F_θ is "regular" in the sense described above. Let $\widehat{\theta}$ be an unbiased estimator of θ based on $X_1, \ldots, X_n$. Then

$$V(\widehat{\theta}) \geq \frac{1}{nI_X(\theta)} \,. \tag{2.6}$$

We know that the Fisher Information of an observation $X \sim \mathcal{N}(\theta, \sigma^2)$ is $I_X(\theta) = 1/\sigma^2$. It thus follows that no unbiased estimator of θ can have a variance smaller than σ^2/n. But the sample mean $\overline{X}$ has precisely this variance, and is thus necessarily the UMVUE. This is no surprise, of course, since $T = \sum_{i=1}^n X_i$ is a complete sufficient statistic for θ in the normal model and $\overline{X}$ is the unique function of T that is unbiased for θ.

Exercise 2.1. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{P}(\lambda)$, a Poisson distribution with mean λ. Let $\theta = e^{-\lambda}$. Note that $\theta = P(X_1 = 0)$, so that the estimator $\widehat{\theta} = 1$ if $X_1 = 0$ and $\widehat{\theta} = 0$ if $X_1 > 0$ is an unbiased estimator of θ. The statistic $T = \sum_{i=1}^n X_i$ is complete and sufficient for θ. Use the Rao–Blackwell and Lehmann–Scheffe Theorems to identify the UMVUE of θ. (**Hint:** Show that, conditional on $T = t$, X_1 has the binomial distribution $\mathcal{B}(t, 1/n)$. Then, evaluate $E(\widehat{\theta}|T)$.)

Exercise 2.2. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$, where the family $\{F_\theta, \theta \in \Theta\}$ satisfies the Cramér–Rao regularity conditions. Suppose $\widehat{h}(\mathbf{X})$ is an estimator of $h(\theta)$ with bias $b(\theta) = E\widehat{h}(\mathbf{X}) - h(\theta)$. Assuming that h and b are differentiable functions, show that the variance of the estimator $\widehat{h}(\mathbf{X})$ satisfies the inequality

$$V(\widehat{h}(\mathbf{X})) \geq \frac{(\partial h/\partial\theta + \partial b/\partial\theta)^2}{nI(\theta)} \,.$$

2.3 Best linear unbiased estimators

When one strays a bit from the i.i.d. framework, the UMVUE of a parameter of interest may not exist or may be analytically inaccessible. When the available data are independent but have nonidentical distributions, or when some form of dependency is present in the data, attempts to obtain the UMVUE may be futile. Linear unbiased estimators often provide a reasonable alternative. Linear estimators have the virtue of utilizing all the data and have the flexibility of allowing the statistician to place different weights on different observations, thereby taking account of their individual precision. The best linear unbiased estimator (BLUE) is, quite simply, the linear unbiased estimator with the smallest variance. There are many examples in the statistical literature of the use of BLUEs, but perhaps the best-known application is in the framework of multiple linear regression. For the linear model $\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ is a $(k+1)$-dimensional vector of unknown parameters, $\mathbf{X}$ is a fixed design matrix and $\boldsymbol{\varepsilon}$ is a vector of uncorrelated errors with common variance (often modeled as i.i.d. $\mathcal{N}(0, \sigma^2)$ variables), the standard estimation technique employed is "ordinary least

squares," by which is meant that the estimator of $\boldsymbol{\beta}$ is the vector $(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k)$ minimizing the sum of squares $\sum_{j=1}^{n}(Y_j - \beta_0 - \sum_{i=1}^{k}\beta_i X_{ij})^2$. The Gauss–Markov Theorem famously asserts that, under the standard linear model with uncorrelated errors having common finite variance σ^2, least squares estimators (LSEs) are the best linear unbiased estimators of the elements of the vector $\boldsymbol{\beta}$. A good deal more can be said about BLUEs, but the above will suffice for our purposes. For more details, see, for example, Rao (1973).

Exercise 2.3. Let $X_1, X_2, \ldots, X_n$ be independent normally distributed random variables, where $Xi \sim \mathcal{N}(\mu, \sigma_i^2)$, with σ_i^2 known for $i = 1, 2, \ldots, n$. Obtain the best linear unbiased estimator of μ.

Exercise 2.4. It is sometimes known *a priori* that a regression line goes through the origin, that is, the y-intercept β_0 in the simple linear regression model is equal to zero. A salesperson's monthly commission, as a function of monthly sales, would be an example. Suppose that the appropriate model for a particular experiment is

$$Y_i = \beta X_i + \varepsilon_i, \quad \text{for } i = 1, 2, \ldots, n,$$

where ε_i are taken to be uncorrelated random errors with mean 0 and common variance σ^2. Show that the least squares estimator $\widehat{\beta} = \sum_{i=1}^{n} k_i Y_i$ of β is BLUE, where $k_i = X_i \big/ \sum_{i=1}^{n} X_i^2$.
(**Hint:** Try a "variational" argument: (a) show that $\sum k_i X_i = 1$ and $\sum k_i^2 = 1/\sum X_i^2$, (b) show that $\widehat{\beta}$ is unbiased and that $V(\widehat{\beta}) = \sigma^2/\sum X_i^2$, (c) show that $\widetilde{\beta} = \sum c_i Y_i$ is unbiased if and only if $\sum c_i X_i = 1$ and (d) assuming that $c_i = k_i + d_i$, where at least one $d_i \neq 0$, show that $\sum k_i d_i = 0$, so that $V(\widetilde{\beta}) = \sigma^2(\sum k_i^2 + \sum d_i^2) > \sigma^2 \sum k_i^2 = V(\widehat{\beta})$.)

2.4 Best invariant estimators

The invariance of an estimator under transformations of the data in a given problem is another intuitive property that may be used to restrict the class of estimators to be considered. The notion of invariance is quite simple, though the theoretical development of this notion and the behavior of estimators having such a property, is embedded in the theory of algebraic groups and, in its most abstract form, requires some fairly sophisticated mathematics. Our treatment is necessarily superficial, given our aim of providing a quick overview, but the basic definitions involved, and some of the main results of the theory, are quite easy to understand. Let's suppose that we have a single observation $X \sim F_\theta$, which may be thought of as the sufficient statistic for θ in the problem of interest. Consider a family of transformations $G = \{g : \mathcal{X} \to \mathcal{X}\}$ on the sample space onto itself. Typical examples include location or scale changes, for which $g(x) = x + c$ or $g(x) = cx$, respectively. In problems to which the theory of invariance applies, the functions in the family G are one-to-one and have the properties of an algebraic group, that is, G is closed under composition, its members satisfy the associative property and G contains the identity transformation $g(x) = x$ and the inverse function g^{-1} of every member $g \in G$. The distribution of the variable $Y = g(X)$

will of course differ from that of X, but it will often be the case that it belongs to the same family as that governing X, that is, there is a parameter value $\widetilde{g}(\theta)$ for which $Y \sim F_{\widetilde{g}(\theta)}$ when $X \sim F_\theta$. The group of transformations that G induces on the parameter space may be denoted as $\widetilde{G}$. A loss function $L(\theta, a)$ is said to be invariant if for any $g \in G$, and for all $a \in A$, there exists a value $a^* \in A$ such that

$$L(\widetilde{g}(\theta), a^*) = L(\theta, a) \; . \tag{2.7}$$

The group of transformations induced on the action space by (2.7) will be denoted by $\overline{G}$. When a loss function L satisfies (2.7), the two estimation problems in which one observes either X or $Y = g(X)$ may be considered equivalent in the sense that any estimator of θ based on X has a natural counterpart based on Y, and the estimator based on Y will have precisely the same statistical properties. An estimator $\widehat{\theta} = \widehat{\theta}(X)$ is said to be *invariant* with respect to the group G if it satisfies the equation

$$\widehat{\theta}(g(x)) = \overline{g}(\widehat{\theta}(x)) \; . \tag{2.8}$$

This equation simply stipulates that if a statistician is prepared to estimate the parameter θ by $\widehat{\theta}(x)$ when $X = x$ is observed, then she should be comfortable estimating θ by $\widetilde{g}(\widehat{\theta}(x))$ when she observes $g(x)$ instead of x. If you were trying to estimate the mean μ of a normal population, and your data are centered around the number 5, then you'd probably decide to estimate μ as 5. But suppose every data point was moved to the right by 7 units. Wouldn't you then be inclined to estimate μ as $5+7$ or 12? That's all that invariance amounts to. It stipulates that any estimators you use should be consistent relative to a group of natural transformations. If $\widehat{\theta}$ is thought to be a reasonable estimator of θ when you observe x, then you should estimate θ to be $g(\overline{\theta})$ when $g(x)$ is observed.

Invariance is an intuitively pleasing framework, and it does lead to its own optimality theory. When a decision problem is invariant, one can often find the best invariant rule, that is, the estimator which has the smallest possible risk function among all invariant estimators. The search for the best invariant rule is substantially simplified by the following fundamental property of invariant decision rules. If $\widehat{\theta}$ is an invariant estimator of the parameter θ, the risk function of $\widehat{\theta}$ is "constant on orbits of θ," that is, $R(\widetilde{g}(\theta), \widehat{\theta}) = R(\theta, \widehat{\theta})$ for all $\widetilde{g} \in \widetilde{G}$. In location and scale parameter problems, there is only one orbit (since there is a function $\widetilde{g}$ that will map any value of θ to any other value), so the risk function of an invariant estimator is constant over the entire parameter space. (This, by the way, is why some authors who write on this topic prefer to use the word "equivariant" rather than the word "invariant" when referring to these estimators.) Estimators with constant risk are generally called "equalizer rules," and the best among them in a given problem tends to have the minimax property. In problems in which the risk function of an invariant estimator is constant, one has but a simple one-variable minimization problem to solve in order to identify the best invariant estimator. Further, the standard theory of invariant estimators identifies conditions under which the best invariant estimator is admissible or minimax or both. In the problem of estimating a location parameter based on the i.i.d. sample $X_1, X_2, \ldots, X_n$, the Pitman estimator, given by

$$\widehat{\theta}(\mathbf{X}) = X_1 - E_{\theta=0}(X_1 \mid X_2 - X_1, \ldots, X_n - X_1), \tag{2.9}$$

is known to be the best invariant estimator under squared error loss. For further details on invariance, see Ferguson (1967, Chapter 4).

Exercise 2.5. Let $X \sim \Gamma(\alpha, \theta)$, where α is known. Consider estimating θ relative to the loss function $L(\theta, a) = \left(\frac{a}{\theta} - 1\right)^2$. Let G be the group of scale transformations on X, that is, $G = \{g : g(x) = cx,\ c > 0\}$, with $\widetilde{G}$ and $\overline{G}$ being the corresponding groups of scale changes on Θ and A. Find the best invariant estimator of θ.

2.5 Some comments on estimation within restricted classes

While restricting the class of estimators that will be entertained in a given problem may have some intuitive appeal, there is no guarantee that the best estimator within the restricted class is a good estimator in general. Examples in which the UMVUE or the best invariant estimator is inadmissible are not difficult to find. Perhaps the best-known example of an inadmissible UMVUE is the so-called sample variance, given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \,, \tag{2.10}$$

based on the random sample $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Insisting that an estimator be unbiased may run counter to other worthy goals in estimation, one of which is precision. Under squared error loss, the risk function of an estimator, generally referred to in this case as its mean squared error (MSE), may be written as follows, in terms of the estimator's variability and its bias:

$$R(\theta, \widehat{\theta}) = E(\widehat{\theta} - \theta)^2 = V_\theta(\widehat{\theta}) + (E_\theta \widehat{\theta} - \theta)^2 \,. \tag{2.11}$$

It is often possible to find biased estimators with a substantially smaller variance than the best unbiased estimator; the trade-off may well result in a smaller MSE. In the case of estimating the variance σ^2 of a normal distribution, the estimator

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{2.12}$$

has a uniformly smaller mean squared error than s^2, so that the latter estimator is inadmissible. Actually, the estimator in (2.12) is itself inadmissible. The best estimator of σ^2 having the form $c\sum_{i=1}^{n}(X_i - \overline{X})^2$ is the one corresponding to the constant $c = 1/(n+1)$.

The example above is just the tip of the iceberg. The list of well-known examples of inadmissible UMVUEs includes estimators of positive parameters which, for certain realizations of the experiment they depend on, can take on negative values. There are, in fact, examples in which the UMVUE is not only inadmissible but patent nonsense. Ferguson (1967, p.136) gives a lovely example of the estimation of a certain

probability based on a Poisson variable X in which the UMVUE of that probability is $(-1)^X$.

I'll give one brief example of similar happenings in the world of linear unbiased estimation. The following problem is treated by Samaniego and Kaiser (1978). A collection of increasing "bids" are accepted in a progressive auction, the observed sequence of bids being $X_1 < X_2 < \cdots < X_n$. It is hypothesized that the bidders are knowledgeable about the worth θ of the item at auction, but they are also interested in obtaining it at the best price possible. The sequence is thus presumed to be bounded above by θ, and is modeled as follows: $X_1 \sim \mathscr{U}[0,\theta], X_2|X_1 \sim \mathscr{U}[X_1,\theta], \ldots, X_n|X_{n-1} \sim \mathscr{U}[X_{n-1},\theta]$. Unlike the order statistic model that this framework resembles, the maximum observation X_n is not a sufficient statistic for θ; in fact, no data reduction at all is available via sufficiency (that is, the entire sample is a "minimal" sufficient statistic for θ). This fact is often taken as a signal that linear unbiased estimators are worth considering. The best linear unbiased estimator of θ is identified by Samaniego and Kaiser to be

$$\widehat{\theta} = \frac{u^T \Sigma^{-1} X}{u^T \Sigma^{-1} u}, \tag{2.13}$$

where u is the n-dimensional vector with ith element equal to $(1 - 1/2^i)$ and Σ is the symmetric $n \times n$ matrix with elements $\sigma_{ij} = (1/2^{j-i}3^i - 1/2^{i+j})$ for $1 \le i \le j \le n$. Now, since θ is a scale parameter for the model, the possible treatment of the estimation of θ using invariance considerations also comes to mind. Samaniego and Kaiser obtain the best invariant estimator of θ with respect to the scale-invariant loss function

$$L(\theta, a) = (a/\theta - 1)^2 . \tag{2.14}$$

The problem above is instructive on several levels. It appears that one has two viable estimators based on quite different restrictions and resulting from two different loss criteria. Interestingly, they are directly comparable by virtue of the fact that the BLUE is itself a scale-invariant estimator. It thus has a constant risk function, under the loss function in (2.14), which is uniformly larger than that of the best invariant estimator. Interestingly, the best invariant estimator is superior to the BLUE under squared error loss as well, since for the loss function in (2.14), the risk function may be written $R(\theta, \widehat{\theta}) = (1/\theta^2)E(\widehat{\theta} - \theta)^2$. Thus, the best invariant estimator has a smaller mean squared error than the BLUE, so that the best invariant estimator beats the BLUE on its own turf. In either scenario, the BLUE is inadmissible.

Best invariant estimators are, of course, not immune to this type of failing. The most famous example of an inadmissible best invariant estimator occurs in a multivariate setting. In a 1956 paper, Stein shocked the statistical community with the news that the mean $\overline{\mathbf{X}}$ of a sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ of vectors drawn from a k-dimensional normal distribution (with, for simplicity, covariance matrix $\Sigma = I$), with k larger than 2, is inadmissible as an estimator of the population mean $\boldsymbol{\mu}$. The sample mean $\overline{\mathbf{X}}$ is, in this context, the best invariant estimator under generalized squared error loss and the transformation groups G, $\widetilde{G}$ and $\overline{G}$ associated with changes in location. Stein showed that the risk function of the estimator

$$\widehat{\boldsymbol{\mu}}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) = \overline{\mathbf{X}} \left(1 - \frac{k-2}{|\overline{\mathbf{X}}|^2} \right) \tag{2.15}$$

is uniformly smaller (as a function of $\boldsymbol{\mu}$) than that of the estimator $\overline{\mathbf{X}}$, where, for $\mathbf{c} \in \mathbb{R}^k$, $|\mathbf{c}| = \sum_{i=1}^k c_i^2$. The Stein estimator is often referred to as a "shrinkage estimator," since the net effect of the adjustment of $\overline{\mathbf{X}}$ through (2.15) is to "shrink" $\overline{\mathbf{X}}$ toward the origin $\mathbf{0}$. The origin is of course an arbitrary choice in this problem, as shrinking toward any fixed k-dimensional vector $\mathbf{C}$ also produces an estimator that uniformly improves upon $\overline{\mathbf{X}}$, with the greatest improvement occurring at values of $\boldsymbol{\mu}$ that are close to $\mathbf{C}$. There have been many subsequent studies of shrinkage estimators that have attempted to shed light on the makeup and behavior of the Stein estimator and its variants. The papers of Efron and Morris (1971, 1972a, 1972b, 1973a, 1973b, 1975, 1976) are deservedly prominent within this literature. A less-cited paper, but truly a *tour de force* in mathematical statistics, is the beautiful "unification" paper by L. D. Brown (1971) which explicitly shows the connection between the inadmissibility of $\overline{\mathbf{X}}$ in dimensions $k \geq 3$ and the nonrecurrence of Brownian motion for $k \geq 3$. Although our discussion of inadmissible invariant estimators has focused on the Stein effect, there are, of course, simpler examples of best invariant rules that are inadmissible. See, for example, Blackwell (1951).

The examples above demonstrate that point estimation problems involving a restriction on the class of estimators must be viewed as mixed blessings. One may be able to find the best estimator in the restricted class, but it may turn out that all estimators in this class are inferior to alternative estimators outside of the class. The lesson to be learned from this, for those who are inclined to consider estimation within restricted classes, is that it would be worthwhile to do a little sniffing around outside of the class to convince oneself that the restriction is not overly confining.

Exercise 2.6. In the "uniform auction" considered above, show that the estimator $\widehat{\theta}$ given in (2.13) is the BLUE of θ.

Exercise 2.7. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathscr{E}(\mu)$, the exponential distribution with mean μ. Confirm that the sample mean $\overline{X}$ is the UMVUE of μ. Show that $\overline{X}$ is an inadmissible estimator of μ under squared error loss.

2.6 Estimators motivated by their behavior in large samples

The methods of point estimation that are most widely used in practice tend to draw their "validity" from an examination of their asymptotic properties. In examining such methods, one must take leave of the decision-theoretic framework, as there is no fixed decision problem to focus on. This notwithstanding, we propose to discuss asymptotic ideas, as leaving this topic out of our discussion would leave a gaping hole in our claimed overview of frequentist methods. Further, as we shall see, there are interesting connections with decision-theoretic ideas when one thinks about sample sizes that are fixed but large, and there do exist asymptotic analogs to decision-theoretic considerations (for example, asymptotic minimaxity) which the reader may wish to explore (even though I don't explore them further here).

The limiting behavior of an estimator $\widehat{\theta}_n = \widehat{\theta}(X_1, X_2, \ldots, X_n)$ as the sample size $n \to \infty$ is often taken as justification for its use. The estimator $\widehat{\theta}_n$ is said to be a *consistent* estimator of the parameter θ if $\widehat{\theta}_n \to \theta$ in some appropriate stochastic sense (e.g., "in probability" or "almost surely"). Consistency is an optimality property of sorts, but a very weak one, as most reasonable point estimators in a given problem will enjoy this property. Among the different consistent estimators one might identify, one would typically examine the rate at which each converges to the target parameter, and among consistent estimators that converge to θ at the same "best" rate, one would wish to recommend for use the estimator that has the best (asymptotic) precision. In problems that satisfy the usual regularity conditions, one is typically able to establish that the estimators of interest are asymptotically normal, that is, that when suitably standardized, they converge in distribution to normal variables. One typically writes this as

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{D} Y \sim \mathcal{N}(0, V) \quad \text{as } n \to \infty\,. \tag{2.16}$$

The value V in (2.16) is referred to as the asymptotic variance of $\widehat{\theta}_n$, and the estimator with the smallest possible asymptotic variance is said to be *best asymptotically normal* (BAN). Such an estimator is said to be an asymptotically efficient estimator of θ. One may think of (2.16) as constituting an approximation result. If n is sufficiently large, then the fraction

$$Z = \frac{\widehat{\theta}_n - \theta}{\sqrt{V/n}} \tag{2.17}$$

should behave, approximately, like a standard normal variable, and $\widehat{\theta}_n$ should have, approximately, mean θ and variance V/n. This suggests that a BAN estimator might be thought of as being, approximately, an unbiased estimator with the best possible variance. Given a sufficiently large sample, one would have identified a reasonable substitute for the UMVUE in the problem under study.

In the late nineteenth century, Karl Pearson advocated the use of estimators obtained by the so-called method of moments. This is one of the better known methods in a class of estimators described by the term "analog." The basic idea behind analog estimators is that the available random sample should exhibit roughly the same features as the population from which the sample was drawn. The general theory of analog estimators is laid out in the monograph by Manski (1988). The method of moments exploits the fact that sample moments and the corresponding population moments should be reasonably close to each other (as the latter is the expected value of the former, and averaging increases precision). Finding a method of moments estimator (MME) involves deriving the solution(s) of one or more "moment equations," that is, equations of the form

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k = EX^k \tag{2.18}$$

for selected integers k. For example, for a sample of geometric random variables $\{X_i\}$ modeling the number of trials needed to obtain the first success in a sequence

of Bernoulli trials with probability θ of success, the first moment equation is

$$\overline{X} = \frac{1}{\theta} ,$$

an equation which identifies $\widehat{\theta} = 1/\overline{X}$ as an MME of θ.

The method of moments has a number of drawbacks. One is that the method doesn't lead to a unique estimator. For example, in the geometric case above, the second moment equation leads to a different estimator. One would thus need to determine which of them is better. A second inconvenient truth is that, while MMEs are consistent and asymptotically normal (provided population moments of sufficiently high order are finite), it is often the case that they do not have the smallest possible asymptotic variance. Thus, MMEs are generally seen as easily-obtained preliminary estimators which require further examination before being recommended for use.

Ronald Fisher is credited with finding a method of estimation which is generally superior to the method of moments. Karl Pearson didn't take the news well. Indeed, the ensuing feud between Pearson and Fisher has become one of the most famous in our discipline. (Actually, Fisher and Pearson weren't particularly fond of each other even before their debate about the quality of their estimation techniques.) The basic idea behind Fisher's recommendation of maximum likelihood estimation is simple and quite intuitive. Given the outcomes $x_1, x_2, \ldots, x_n$ of an i.i.d. sample with underlying distribution F_θ having density or probability mass function $f_\theta(x)$, the likelihood function $L(\theta)$ is defined as

$$L(\theta) = L(\theta \mid x_1, x_2, \ldots, x_n) = f_\theta(x_1, x_2, \ldots, x_n) , \tag{2.19}$$

that is, as the joint density or probability mass function of $\mathbf{X}$ evaluated at the observed $\mathbf{x}$. (Actually, the likelihood is usually considered to be the portion of the RHS of (2.19) that depends on θ; then, $L(\theta)$ in (2.19) is the constant multiple of the likelihood which standardizes it so that it sums or integrates to one. For simplicity, and since it does no harm, we will continue to refer to the function L in (2.19) as the likelihood function. But one may also refer to any constant multiple of this function as the likelihood.) If $L(\theta)$ is given, you might then ask "what value of the parameter θ would give the sample you actually observed the highest chance of happening?" When the variable X is discrete, $L(\theta)$ truly represents the probability of observing $X_1 = x_1, \ldots, X_n = x_n$. In the continuous case, the actual probability associated with the observed values is zero, but the question may be rephrased in terms of a small open set centered at the observed data point $(x_1, \ldots, x_n)$. The natural answer is the value of θ for which the density or mass function in (2.19) is maximized. In terms of probability, one may think of the value of θ for which small intervals (say, $(x_i - \varepsilon, x_i + \varepsilon)$) centered at the data values receive the largest possible probability. In either case, the function of the data $\widehat{\theta} = \widehat{\theta}(x_1, x_2, \ldots, x_n)$ that maximizes $L(\theta)$ above is referred to as the *maximum likelihood estimator* (MLE). Given the data, the value of the parameter which makes what you have seen as likely as possible would seem, intuitively, to be a very reasonable estimator of the unknown parameter. But the important matter of its statistical behavior remains to be discussed.

Fisher developed the asymptotic theory of maximum likelihood estimators. Under suitable regularity conditions (see Lehmann and Casella (1998), Chapter 6), the MLE $\widehat{\theta}_{ML}$ is best asymptotically normal, with its limiting distribution given by

$$\sqrt{n}(\widehat{\theta}_{ML} - \theta) \xrightarrow{D} Y \sim \mathcal{N}(0, I^{-1}(\theta)) \,. \tag{2.20}$$

The asymptotic variance V of any estimator that converges to θ at the same rate (usually written as $O(1/\sqrt{n})$) can be no smaller than $I^{-1}(\theta)$, except, possibly, on a subset of Θ of measure zero.

Exponential families have a prominent place among families of distributions that satisfy the standard regularity conditions under which $\widehat{\theta}_{ML}$ is BAN. But the method of maximum likelihood is quite general, and it is often the frequentist method of choice in problems in which the model is not among the well-known collection of exponential families. In parametric models of this sort, the asymptotic justification of the estimator reduces to checking that the model satisfies the regularity conditions. The ML approach has in fact been extended to nonparametric problems in which the aim is to estimate the underlying distribution F of the sample (where F is allowed to be any distribution on the support set of the available observations). This extension was developed by Keifer and Wolfowitz (1956), and it has since led to many interesting findings in the field of nonparametric statistics. The empirical distribution F_n (that is, the step function which takes a jump of $1/n$ at each of the observations in a random sample of size n) is the nonparametric maximum likelihood estimator (NPMLE) of a general distribution F. The Kaplan–Meier estimator is the NPMLE of F given failure time data subject to right-censoring.

Given that there are no guarantees outside the domain of regular models, one might expect that MLEs can, on occasion, behave rather badly. This is, of course, possible. There are many examples of inconsistent MLEs in the literature. Perhaps the best-known parametric example is that of Neyman and Scott (1948). Boyles, Marshall and Proschan (1985) showed that the NPMLE of a distribution F known to have the "increasing failure rate average" (IFRA) property is inconsistent as an estimator of F. Rojo and Samaniego (1991) show that NPMLE is inconsistent as an estimator of a distribution F known to satisfy a uniform stochastic ordering constraint. It is often the case that parametric and nonparametric MLEs behave poorly in problems in which the target parameter is assumed to satisfy a specific restriction. But the misbehavior of MLEs in nonregular problems is by no means universal. Consider, for example, the estimation of the parameter θ based on a random sample from the uniform distribution $\mathcal{U}[0,\theta]$. This model is of course nonregular due to its support set's dependence on θ. The observed maximum $X_{(n)}$ is the MLE of θ. As is well known,

$$Y_{(n)} = \frac{X_{(n)}}{\theta} \sim \text{Be}(n,1) \,,$$

so that $EY_{(n)} = n/(n+1)$ and $V(Y_{(n)}) = n/(n+1)^2(n+2)$. It follows that $V(\sqrt{n}(X_{(n)} - \theta))$ tends to zero as $n \to \infty$, and thus that $\sqrt{n}(X_{(n)} - \theta) \to 0$ as $n \to \infty$. In this problem, the MLE converges to the target parameter faster than it does when it is based on a random sample from a "regular" model. It is thus clear that one cannot assume that the MLE will misbehave in nonregular cases, even though it often does.

A persistent question that arises when using estimators justified by their asymptotic behavior is how large the sample size n must be for the distribution of the standardized MLE to be well approximated by a standard normal distribution. The question has no general analytical answer, as there are problems in which any predetermined value of n will not be large enough. This is evident in the case of a random sample of Bernoulli trials, as the normal approximation for the distribution of the MLE, the sample proportion $\widehat{p}$, will be quite unlike the true distribution of $\widehat{p}$ unless the product np is sufficiently large. If p happens to be very close to 0, an enormous sample size n is required for $\sqrt{n}(\widehat{p}-p)$ to be approximately normally distributed. No matter how large n may be, $\sqrt{n}(\widehat{p}-p)$ may be quite nonnormal when p is very small. Fortunately, the latter circumstance is not typical in applied work. In many applications, samples of size 30, 50 or 100 might prove entirely satisfactory. Often, the validity of the normal approximation to the distribution of the MLE is investigated via simulation, and one may be able to assure oneself that the normal approximation of this distribution is justifiable in samples of moderate size.

Exercise 2.8. Let $X_1,\ldots,X_n \overset{iid}{\sim} F$. Show that the empirical distribution F_n is the nonparametric maximum likelihood estimator of F, that is, show that the distribution F_n assigns the largest possible probability to the observed sample values $x_1,\ldots,x_n$.

Exercise 2.9. Prove the "invariance property" of maximum likelihood estimators: If $\widehat{\theta}$ is the MLE of the parameter θ, and h is a continuous one-to-one function, then $h(\widehat{\theta})$ is the MLE of the transformed parameter $h(\theta)$.

Exercise 2.10. Suppose $X_1,X_2,\ldots,X_n \overset{iid}{\sim} \mathscr{E}(\mu)$, the exponential distribution with mean μ. The Xs represent the observed failure times of n identical items placed on test, and the experiment is carried out under type I censoring at the fixed time horizon T. Thus, the observed outcomes of the life-testing experiment are $Y_1,Y_2,\ldots,Y_n$, where $Y_i=\min\{X_i,T\}$. Suppose that every X_i exceeds T. The likelihood function is then the probability of the observed event, that is,

$$L=\prod_{i=1}^{n} \mathrm{e}^{-T/\mu}=\mathrm{e}^{-nT/\mu}\ .$$

Show that the maximum likelihood estimator of μ does not exist.

Exercise 2.11 (Neyman and Scott). For each fixed $i=1,\ldots,n$ and each $j=1,\ldots,k$, suppose that $X_{ij}\sim\mathscr{N}(\mu_i,\sigma^2)$. Show that the estimator

$$\widehat{\sigma}^2=\frac{\sum_{i,j}(X_{ij}-\overline{X}_i)^2}{kn}$$

is the MLE of the parameter σ^2, where $\overline{X}_i=\frac{1}{k}\sum_{j=1}^{k}X_{ij}$, and that, for any fixed value of k, it is inconsistent for σ^2 as $n\to\infty$.

2.7 Robust estimators of a population parameter

The concern that the assumed model for the available data is misspecified, or that the data may be contaminated in some fashion (for example, by recording errors or by occasional occurrences of data drawn under conditions that differ from the assumed experimental conditions), has led to the study of methods of estimation that are unaffected (or only mildly affected) by such deviations from the model. While this is an important development in statistical work which enjoyed a notable surge of interest in the decade of the 1970s (largely stimulated by the path-breaking study of Andrews *et al*. (1972) on robust estimators of location) and continues to be of substantial interest (witness the increased attention given to nonparametric inference today), it has only a rather modest connection with issues raised in the sequel. Its treatment here will therefore be quite brief.

There is a natural tension between estimating the parameters of the postulated model for one's data and protecting oneself from errors due to various types of violations of the model. Since win-win situations are rarely available in Statistics, it is typical here for a compromise to be struck between good estimation under the original modeling assumptions and protection against various forms of model misspecification. It is instructive to review how one might attack the problem of estimating a location parameter, say the population mean μ. In many problems, and certainly in the problem of estimating the mean of a normal distribution, one would be inclined to use the time-honored estimator $\overline{X}$. If the underlying distribution of the sample is indeed $\mathcal{N}(\mu, \sigma^2)$, then $\overline{X}$ has a good deal to recommend it. But what if there is an occasional misrecording, so that the true distribution of the data is actually the mixture model with distribution $F(x) = \varepsilon M(x) + (1-\varepsilon)N_{\mu,\sigma^2}(x)$, where $M(x)$ describes the behavior of the "maverick observations" (including outliers) and ε is some suitably small value, perhaps 0.05. Since the estimator $\overline{X}$ is highly affected by outliers, it immediately loses its appeal as an estimator of μ. Many alternatives have been studied for this scenario and for a variety of other versions of an alternative true model. The median ($X_{\left(\frac{n+1}{2}\right)}$ when n is odd) is a robust estimator of μ. Of course it is inferior to $\overline{X}$ when the true model is normal, but it is the estimator that is the most highly resistant to the influence of outliers, when they are present, and it easily outperforms $\overline{X}$ when outliers occur with some frequency. This is one of the reasons that the reported "average income" of a group of interest will typically be the median rather than the mean income. In spite of the simplicity and utility of the sample median, there are better-performing compromises available. Trimmed means, for which a fixed percentage of large observations and of small observations are simply ignored, and the remaining observations are averaged, are a well-studied class of robust estimators. The monograph by Andrews *et al.* (1972) studies and compares a host of other options, including so-called adaptive procedures which include a preestimation procedure that is meant to detect the extent of outlier-protection that appears to be needed in a given application.

The literature on the robust estimation of location parameters includes the study of *L*-estimators, *M*-estimators and *R*-estimators. Lehmann and Casella (1998) give an excellent account of the highlights of the theory of these three particular approaches

to robust estimation. It might be mentioned that both the median and the trimmed mean mentioned above are examples of L-estimators, the L standing for "linear combination of order statistics." Other topics of interest in robustness theory include the influence function, which measures the impact of one or more outliers on a given estimator, and the breakdown point, which identifies the number of outliers an estimator can accommodate without being compromised as an estimator of the parameter of the assumed model. For a detailed treatment of robustness topics, see Huber (1981).

Exercise 2.12. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$, where $F_\theta(x) = F(x-\theta)$, $F(0) = 1/2$ (that is, θ is the median of F) and F has finite variance. Suppose, further, that when $\theta = 0$, the density f of F is positive at $x = 0$. Let $\widetilde{\theta}_n$ be the median of the sample of size n. The following asymptotic result is well known:

$$\sqrt{n}(\widetilde{\theta}_n - \theta) \xrightarrow{D} Y \sim \mathcal{N}\left(0, \frac{1}{4f^2(0)}\right).$$

(a) Show that, when sampling from a normal population, the asymptotic efficiency of $\widetilde{\theta}_n$ relative to the MLE $\widehat{\theta}_n = \overline{X}$ (that is, the ARE $= AV(\widehat{\theta}_n)/AV(\widetilde{\theta}_n)$) is 0.637. (b) Show that if F_θ is a t-distribution with three degrees of freedom and median θ (a t-distribution with finite variance), then the ARE of $\widetilde{\theta}_n$ relative to $\overline{X}$ is 1.62.

3

An Overview of the Bayesian Approach to Estimation

3.1 Bayes' Theorem

In the subsections below, I will go into considerable detail on the philosophy, methodology and characteristics of the Bayesian approach to statistical estimation. It seems appropriate to begin the discussion by presenting the famous theorem by Thomas Bayes which underpins the entire enterprise. Its most common form involves a two-stage experiment. Consider an event A of interest as a possible outcome of the first stage of the experiment and an event B, a possible outcome of the second stage. If, for example, one is drawing marbles at random from an urn containing red and white marbles, A might be the event of drawing a red marble on the first draw and B might be the event of drawing a white marble on the second draw. Such experiments are often represented by a "tree" such as that in Figure 3.1.

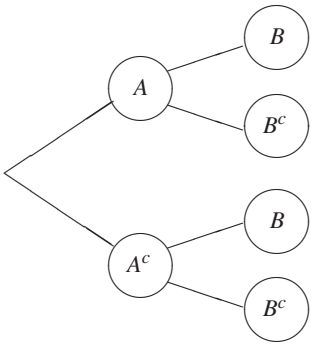

Fig. 3.1. A tree for a two-stage experiment

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_3,

The well-known "multiplication rule" is used to compute the probabilities of the possible outcomes of the experiment as a whole. For example, $P(A \cap B) = P(A)P(B|A)$. There is a natural temporal progression in this experiment, with the second stage following upon the outcome of the first stage. The question that Bayes asked and answered was: If the outcome of the second stage is known, what would be the probability of a particular first-stage outcome? Bayes' Theorem simply provides a formula for making that calculation. Specifically, we could make that computation using the (now) standard definition of conditional probability, that is,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}\,, \tag{3.1}$$

assuming, of course, that $P(B) > 0$. Bayes' formula is often written in a more extensive but equivalent form. If the first stage has n possible outcomes $A_1, \ldots, A_n$ and the second stage has m possible outcomes $B_1, \ldots, B_m$, then for $1 \leq i \leq n$, and for $1 \leq j \leq m$, such that $P(B_j) > 0$,

$$P(A_i \mid B_j) = \frac{P(A_i)P(B_j \mid A_i)}{\sum_{k=1}^{n} P(A_k)P(B_j \mid A_k)}\,. \tag{3.2}$$

The question posed by Bayes was more than a curiosity. It raises intriguing philosophical questions and it calls attention, as well, to a practical tool for calculating certain conditional probabilities of interest. On the philosophical level, consider the following apparently conflicting views. If the two-stage experiment above has been performed, and you happen to be informed that the event B occurred in the second stage, is it appropriate to talk about the probability that A occurred? After all, the experiment already happened, and the outcome of the first stage, while unknown to you, has in fact already occurred. Perhaps it is known to someone who witnessed the experiment, in which case, whether or not A occurred can simply be determined by asking the question. This leads to the view that, since the occurrence or nonoccurrence of A is now a historical fact, it is not an event whose probability we should be contemplating. The fact that we can indeed ask and answer questions about the likelihood that the event A occurred identifies an interesting and new proposition: we can discuss probabilities based on our *uncertainty* about the occurrence of a particular event. The number $P(A|B)$ obtained via (3.1) is precisely such a probability. Bayes' Theorem essentially opens the door to the consideration of a new form of probability, one that is personal and subjective and conditioned on what you happen to know. The theorem is also a useful tool, as the following example shows.

Example 3.1. Suppose that it is known that 2% of the population has a certain disease, a fact that we record as $P(D) = 0.02$. As is often the case, there is a simple, noninvasive diagnostic test that is pretty effective in detecting the presence of the disease when a person actually has it. Let's assume the probability of detecting the disease is 0.95 in such cases, that is, that $P(+|D) = 0.95$. Diagnostic tests are never perfect, and there is generally some small probability of a "false positive." Let's assume, here, that $P(+|D^c) = 0.1$. If the test is administered to a random individual, say

you, and the test result is positive, you would most certainly want to know what the chances are that you actually have the disease. Bayes' Theorem provides the answer: $P(D|+) = 0.019/(0.019+0.098) = 0.1624$. You would no doubt be quite relieved to learn that the chances that you have the disease are relatively low. A (probably less comfortable) follow-up test would be used to determine more definitively whether or not you have the disease. The surprisingly low value of $P(D|+)$ often raises some eyebrows. It is explained by the fact that, even though the probability of a false positive is low, the proportion of the population without the disease is very large. Thus, nearly 85% of the positive test results come from that segment of the population.

Putting the example above in Bayesian language, the first-stage probabilities represent "prior" information about the disease. The second stage consists of the available experimental data. Given the experimental outcome "+," one may compute the "posterior" probability of having the disease, that is, one may "update" the prior information on the basis of the data to obtain a revised perspective about someone having the disease. Bayesian analysis is, in general, about the updating of prior information in light of available experimental data. A Bayesian formulation of an estimation problem involves a prior distribution $G(\theta)$ for the unknown parameter θ, a model F_θ for the observable data and a posterior distribution $G(\theta|\mathbf{X} = \mathbf{x})$ for the parameter θ given the experimental data $\mathbf{X}$.

Exercise 3.1. Lie detector tests are of course imperfect. But it is well documented that they have proven useful in the context of criminal investigations. Suppose that it is known that when a criminal suspect gives an answer to a relevant question, he/she answers truthfully (T) only 30% of the time. Suppose, further, that a lie detector test will classify a truthful answer as a lie (L) 10% of the time, and will classify an untruthful answer (T^c) as a lie 80% of the time. On a given question that the lie detector test classifies as a lie, what's the probability that the suspect is telling the truth?

3.2 The subjectivist view of probability

The foundations of Bayesian inference cannot be properly understood without a discussion of subjective probability. The foremost proponents of this perspective on probability include Frank Ramsey (1930), Bruno De Finneti (1974), Morris DeGroot (1970), I.J. Good (1950), Dennis Lindley (1985) and L.J. Savage (1954). The brief treatment I give this topic cannot begin to do justice to the celebrated tomes just cited, but I will nevertheless bravely dive in and try to summarize the main ideas. Let's begin with the off-the-wall question "What is the probability that a stranger named Isaac Newton knocks on your door tomorrow?" At first view, this question doesn't seem to be appropriately placed in the domain of probability. But one speaks of the chances of the occurrence of such events all the time. While the frequency theory of probability relies, at least informally, on the idea that a probability of an event is determined by the long-run relative frequency of the event in repeated trials,

this view is not helpful when dealing with one-time happenings and, in general, in dealing with situations when no related experimental data at all is available. Further, there are many things in life about which we are uncertain, and we often use the language of probability, if not its formal tools, in thinking and talking about them. The subjectivist view of probability is based on the following premise: every individual should be able to determine, through introspection or via consultation with experts, his beliefs about the odds that an event A will occur or not, that is, his beliefs about the ratio $P(A)/(1-P(A))$. A common formulation of this premise is that one could determine one's subjective assessment of the value of the ratio above by considering how much one is willing to wager on the occurrence of the event A. There are various versions leading to the subjective assessment of the value of this ratio. We will outline below the development favored by DeGroot (1970, 1988) which is based on an individual's basic intuition about the relative likelihood of any pair of events. What is important here is to understand that this individual's assessment is indeed *subjective*, representing the individual's belief system concerning the events in question. It is thus quite possibly different from another individual's assessment, and this really is of no consequence in what follows. What does matter is that the way an individual assesses probabilities in an uncertain situation reflects that individual's subjective beliefs and is internally logical and consistent. The probabilities assigned to the possible events (generally, the members of a σ-field of subsets of a universal set S) in an experiment of interest (or an occasion involving uncertainty) are assumed to obey the standard laws of the calculus of probability, namely, that (i) $P(S)=1$, (ii) $P(A)\geq 0$ for any event A and (iii) when $A_i \cap A_j = \varnothing$ for all $i \neq j$, $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$. These three laws are applicable to all forms of probability assessment, subjective or otherwise, and are satisfied, for example, when probability assignments are made on the basis of the relative frequency of occurrence of subsets of S in a fixed number of trials of a real experiment.

How might an individual proceed with his subjective assignment of probabilities to the events of interest? Unlike the frequentist paradigm in which one loosely relies on tradition or experience, assigning probabilities to events either intuitively or being guided by empirical evidence from repetitions of the experiment of interest, the Bayesian process is based on one's subjective assessments about the relative likelihood of any two events. It is, however, assumed that one's assessments conform to several intuitively evident requirements. For example, it is assumed that an (admittedly idealized) individual is endowed with a relation $\preceq$ which determines his subjective judgment about the relative likelihood of any pair of events. More specifically, for the events A and B, $A \preceq B$ is interpreted as signifying that that event B is at least as likely to occur as event A, with $A \prec B$ meaning that B is considered to be more likely than A. If $A \preceq B$ and $A \succeq B$, then we write $A \sim B$, signifying that the events A and B are considered to be equally likely. Five specific assumptions are made about the relation $\preceq$. The first simply asserts that any two events are comparable.

Axiom 1. For any pair of events A and B, either $A \succ B$, $A \sim B$ or $A \prec B$.

The second axiom stipulates that the relation $\preceq$ must satisfy a certain logical ordering restriction.

Axiom 2. Let $\{A_i, B_i,\ i = 1,2\}$ be events for which $A_1 \cap A_2 = \varnothing$ and $B_1 \cap B_2 = \varnothing$. If $A_1 \preceq B_1$ and $A_2 \preceq B_2$, then $A_1 \cup A_2 \preceq B_1 \cup B_2$. Further, if either $A_1 \prec B_1$ or $A_2 \prec B_2$, then $A_1 \cup A_2 \prec B_1 \cup B_2$.

It is easily shown that Axioms 1 and 2 together imply that the relation $\preceq$ is transitive, that is, given events A, B and C, if $A \preceq B$ and $B \preceq C$, then $A \preceq C$. The third axiom asserts that the empty event $\varnothing$ is the least likely among all possible events and that the certain event (or universe) S is more likely than $\varnothing$. This axiom serves to exclude the trivial experiment in which $\varnothing \sim S$.

Axiom 3. For any event A, $\varnothing \preceq A$. Further, $\varnothing \prec S$.

An immediate consequence of Axiom 3 is that if $A \subseteq B$, then $A \preceq B$. The fourth axiom resembles the continuity property of probability measures.

Axiom 4. If every event in the decreasing sequence of events $A_1 \supseteq A_2 \supseteq \cdots \supseteq A_n \supseteq \cdots$ is at least as likely as the event B, then $\bigcap_{i=1}^{\infty} A_i \succeq B$.

The axioms above seem both natural and quite innocuous, and most readers would accept them as appropriate characteristics of any particular execution of the process of assessing the relative likelihood of pairs of events in an uncertain environment. One might wonder whether they are, by themselves, sufficient to ensure the existence of a probability distribution P that is consistent with the relation $\preceq$ in the sense that $P(A) \leq P(B)$ if and only if $A \preceq B$. And if so, would such a probability distribution be unique? These questions are answered in Kraft, Pratt and Seidenberg (1959), wherein a particular finite experiment is exhibited in which the chosen relation $\preceq$ satisfies Axioms 1–4, and yet no probability distribution P exists that agrees with the specified ordering. An additional axiom is needed to ensure that such a P exists. This axiom posits the existence of a random variable X with the uniform distribution $\mathscr{U}[0,1]$ which assigns every interval $I \subseteq [0,1]$ a probability equal to its length. Further, it asserts that any event in the experiment of interest can be compared to any interval I. Specifically, we will adjoin the following assertion to the axioms stated thus far.

Axiom 5. There exists a random variable X with distribution $\mathscr{U}[0,1]$, and for any event A and any interval $I \subseteq [0,1]$, either $A \prec I$, $A \sim I$ or $A \succ I$.

The fifth axiom assumes that any individual judging the relative likelihood of the events in a given experiment can also compare the likelihood of any of these events to the likelihood of the event $\{X \in I\}$ for an arbitrary interval $I \subseteq [0,1]$, where $X \sim \mathscr{U}[0,1]$. This axiom essentially asserts that the individual who is assessing relative likelihoods is thinking probabilistically, and can thus assess how the likelihood of any event A compares to any value $p \in [0,1]$. This axiom is clearly the boldest and most complex. Is it a reasonable assumption? For the idealized individual about whom we are thinking, the axiom seems justifiable. This individual is able to compare any two events and decide which is the more likely. It seems reasonable to suppose that this individual has some numerical measure in mind when comparing

two likelihoods. It also seems reasonable that this measure be associated with the chances of an event's occurrence and thus take values in the interval $[0,1]$. As for the existence of the variable X in Axiom 5, one can simply imagine a well-oiled spinning wheel as yielding such an outcome in $[0,1]$ or think of X as the outcome of one of the many random number generators that routinely pass the test of uniformity.

Axioms 1–5 are precisely the necessary and sufficient conditions which imply the existence of a unique probability distribution P that is consistent with the relation $\preceq$. This is stated in the following. For a proof, see DeGroot (1970).

Theorem 3.1. *Let $\preceq$ be a relation which serves as a total ordering, in terms of relative likelihood, for all the events in a random experiment (or other situation involving uncertainty), and assume that Axioms 1–5 hold. Then, for every event A, there is a unique value $x \in [0,1]$ such that $A \sim [0,x]$.*

Orthodox Bayesians interpret Theorem 3.1 as follows. Under the assumptions made (namely, the axioms), there is only one way for an individual to treat uncertainty. The relation $\preceq$ which this individual uses to assess relative likelihood of a pair of events may, in fact, be replaced by a probability distribution. Whether the uncertainty involved pertains to the outcomes of a real, physical experiment or pertains to the value of an unknown constant (think parameter), one must, in the end, quantify this uncertainty using a probability measure and the associated calculus of probability. Informally, the developments above may be thought of this way: if Axioms 1–5 seem sensible to you, then Theorem 3.1 says that you are a Bayesian. When one adopts these axioms as self-evident truths, one is then committed to the use of probability to quantify uncertainty. Since one is always uncertain about the exact value of a parameter one is trying to estimate, it follows that one should place a probability distribution on that parameter. Acting in accordance with this prescription is referred to in the Bayesian literature as *coherence*. Violations of this prescription are then, of course, incoherent. (A personal aside — I have always thought that this was a very clever choice of words. Given the options offered, who among us would want to be thought of as "incoherent?")

There are other notable principles in the Bayesian gospel, and we will now touch on them briefly. The subject that deals with quantifying and comparing the gains and losses associated with a particular decision in a statistical context is called *utility theory*. The utility of a given decision is a measure of its value, and one would typically seek to maximize the utility in selecting a decision. Utility Theory has its own axiomatic development leading from a system of "preferences" among possible payoffs associated with the available decisions to the existence of a utility function (that's unique up to linear transformations) consistent with one's preferences. The loss functions used in the standard presentation of decision theory are "negative utilities." We will view the process of minimizing expected losses as equivalent to the process of maximizing expected utility. The Bayesian bottom line is that a coherent, rational statistician will quantify uncertainty using probability and will optimize her decision making by maximizing the expected utility of her decision. More details on utility theory may be found in DeGroot (1970) and Ferguson (1967).

Exercise 3.2. Show that Axioms 1 and 2 imply that the relation $\preceq$ is transitive.

Exercise 3.3. Show that Axiom 3 implies that if A and B are events such that $A \subseteq B$, then $A \preceq B$.

3.3 The Bayesian paradigm for data analysis

We now will turn our focus, as we did in Chapter 2, to the problem of point estimation. We will assume that the available data are modeled as $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$ and that our statistical goal is to estimate the unknown parameter θ with respect to an agreed upon loss function L. Many authors have pointed out that the choice of model F_θ is a more subjective process than generally recognized. Savage (1962), for example, remarked that the specification of F_θ is, in practice, often tentative and that such choices are generally treated as "rough practical ways to get on with the problem"; while noting its lack of definitiveness, he goes on to assert that the choice of the model F_θ "tends to have a quality that might cautiously be called 'objectivity.'" We will take the view here that the statistician, frequentist or Bayesian, has chosen the particular F_θ above as her working model for the observable data. As indicated above, the first step in any Bayesian treatment of a situation involving uncertainty is to quantify that uncertainty through the use of probability assessment and then using, as needed, elements of the calculus of probability. In the context of the estimation problem of interest here, the remaining "uncertainty" in the problem has to do with the parameter θ. Of course, the data will ultimately shed some light on the possible value of θ, but when one encounters this estimation problem in its early stages, a Bayesian will attempt to quantify his uncertainty about θ in the form of a *prior distribution* G usually, but not always, assumed to have a density function $g(\theta)$.

If one is serious about doing a Bayesian analysis, and serious about the problem to which the resulting inferences are to be applied, the specification of a prior G is not an easy matter, nor should it be done in a casual or frivolous way. Bayesian writers often refer to the process of *introspection*, that is, of thinking very carefully about one's experience and its relevance to the unknown parameter one is trying to model. While it is not a viewpoint that I will emphasize in this monograph, I should mention that, in the subjectivist's view, whatever one comes up with through introspection can't be wrong, since the object of that process is representing one's beliefs. While one's beliefs may be wrong (for example, when one strongly believes that θ is -25 when the true value of θ is 25), one's subjective views are generally considered to be correctly expressed by one's chosen prior. What if one's introspection leads to results that are judged too tentative, vague or perhaps even questionable in their trustworthiness? This may often be the case when dealing with a parameter θ describing some aspect of a technical experiment. In such cases, the statistician would be wise to consult with one or more experts in the subject-matter area under study. This process is usually referred to as *probability elicitation*, and has been the subject of considerable discussion in the Bayesian literature (see, for example, O'Hagan *et*

al. (2006)). A popular method of elicitation involves a series of questions regarding several percentiles of the distribution of θ (like, what value "c" would you be 95% sure is larger that the true value of θ?). Another approach is to ask questions about one's best guess at θ and the level of confidence that one has in that guess. While one can't prove many theorems about the elicitation process (though I will discuss in Section 3.9 a rather famous one due to Blackwell and Dubins), it is generally acknowledged that it plays an important role in Bayesian analysis and that those who do it well are generally rewarded by better inferences. The elicitation process is particularly valuable in problems drawn from fields (like engineering, the medical sciences and experimental social sciences, to name a few) in which reliable expertise is abundantly available. Whatever process is used for the determination of the prior density g, I will now assume that both $g(\theta)$ and $f_\theta(x)$ have been selected and will proceed with a description of what a Bayesian does with these two models.

Let us, for simplicity, imagine that we are working with a single observation X. In many problems of practical interest, this is not a heavy imposition, as X can be taken to be the one-dimensional sufficient statistic in the problem, in which case F_θ is simply the sampling distribution of that statistic. The joint density for the datum X and the parameter θ (both modeled, for convenience, as continuous variables) is given by

$$f(x,\theta) = f(x|\theta)g(\theta)\ . \tag{3.3}$$

Once the experiment has yielded a specific observation, that is, once we have observed that $X = x$, we are left with the assessment of the remaining uncertainty in the problem. It is natural to turn to the conditional distribution of θ, given $X = x$, the so-called *posterior distribution* of θ, whose density (in continuous problems) is given by

$$g(\theta|x) = f(x|\theta)g(\theta)/f(x)\ , \tag{3.4}$$

where $f(x)$ is the marginal density of X, that is,

$$f(x) = \int f(x,\theta)\,\mathrm{d}\theta\ . \tag{3.5}$$

From the Bayesian perspective, any and all statistical inference one might consider, including estimation, hypothesis testing and prediction, flows solely from the posterior distribution of the parameter(s) given the data. In problems of point estimation, a Bayesian seeks to minimize the *posterior expected loss* given by $E_{\theta|X=x}L(\theta,\widehat{\theta}(x))$. The Bayes estimator $\widehat{\theta}(x)$ may thus be represented as

$$\widehat{\theta}(x) = \underset{\widetilde{\theta}(x)}{\operatorname{argmin}} E_{\theta|X=x}L(\theta,\widetilde{\theta}(x))\ . \tag{3.6}$$

There are several well-known theoretical results which show us how this minimization is done for particular choices of L. Readers to whom these results are unfamiliar should try to prove them. The first and third proofs are straightforward. The second proof flows ever-so-smoothly after a quick glance at the appendix of Chernoff and Moses' 1959 text.

Theorem 3.2. *Assume that (X,θ) has a joint density as specified in* (3.3)*, that is, based on the sampling distribution F_θ and the prior distribution G. When estimating the parameter θ under the squared error loss function $L(\theta,a)=(\theta-a)^2$, the Bayes estimator $\widehat{\theta}(x)$ with respect to the prior G is the mean of the posterior distribution of θ, that is,*

$$\widehat{\theta}(x)=E(\theta \mid X=x)\,, \tag{3.7}$$

provided that the distribution of $\theta \mid X=x$ has a finite second moment.

Theorem 3.3. *Assume that (X,θ) has a joint density as specified in* (3.3)*, that is, based on the sampling distribution F_θ and the prior distribution G. When estimating the parameter θ under the absolute error loss function $L(\theta,a)=|\theta-a|$, the Bayes estimator $\widehat{\theta}(x)$ with respect to the prior G is the median of the posterior distribution of θ, that is,*

$$\widehat{\theta}(x)=\text{ median of } G(\theta \mid X=x)\,, \tag{3.8}$$

provided that the distribution of $\theta \mid X=x$ has a finite first moment.

Theorem 3.4. *Assume that (X,θ) has a joint density as specified in* (3.3)*, that is, based on the sampling distribution F_θ and the prior distribution G. When estimating the parameter θ under the Linex loss function $L(\theta,a)=\exp\{c(a-\theta)\}-c(a-\theta)-1$, the Bayes estimator $\widehat{\theta}(x)$ with respect to the prior G is given by*

$$\widehat{\theta}(x)=-\frac{1}{c}\ln\left\{E_{\theta|X=x}\mathrm{e}^{-c\theta}\right\}\,, \tag{3.9}$$

provided that the moment generating function of $\theta \mid X=x$ exists and is finite.

Let's take a look at a simple example of these three results.

Example 3.2. Suppose that $X|\theta \sim \mathscr{U}[0,\theta]$ and that θ has the gamma prior distribution $G=\Gamma(2,1)$, that is, the prior distribution with density

$$g(\theta)=\theta\mathrm{e}^{-\theta}I_{(0,\infty)}(\theta)\,,$$

where $I_A(x)$ is the indicator function of the set A taking the value 1 if $x\in A$ and the value 0 if $x\notin A$. The joint density of X and θ is thus

$$f(x,\theta)=\mathrm{e}^{-\theta},\quad \text{for } 0<x<\theta<\infty\,.$$

The marginal density of X is that of the exponential (or $\text{Exp}(1)\equiv\Gamma(1,1)$) distribution with mean 1; thus, the posterior density of θ is

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

the density of the "translated exponential distribution" or, alternatively, the density of $x+Y$, where $Y\sim\text{Exp}(1)$. The mean and median of Y are 1 and $\ln 2$, respectively, and the moment generating function of Y is $m_Y(t)=(1-t)^{-1}$ for $t<1$. It follows that, under squared error loss, the Bayes estimate of θ with respect to (wrt) G is $\widehat{\theta}(x)=x+1$, under absolute error loss, the Bayes estimate of θ wrt G is $\widehat{\theta}(x)=x+\ln 2$ and under Linex loss (with $c>0$, say, which penalizes overestimation), the Bayes estimate of θ wrt G is $\widehat{\theta}(x)=x+[\ln(1+c)]/c$. ■

Exercise 3.4. Prove Theorem 3.2.

Exercise 3.5. Prove Theorem 3.3.

Exercise 3.6. Prove Theorem 3.4.

Exercise 3.7. Let $X|p \sim \mathscr{B}(n,p)$, the binomial distribution with parameters n and p, and suppose $p \sim G = \mathrm{Be}(\alpha,\beta)$, the beta distribution with mean $\alpha/(\alpha+\beta)$. Assume that one wishes to estimate p with squared error loss. (a) Show that the marginal pmf of X is the "beta-binomial" distribution on the integers $\{0,1,\ldots,n\}$; (b) Obtain the posterior density of $p|X=x$ as the ratio $f(x,p)/f(x)$. Note that $g(p|x)$ can also be identified "by inspection," as $g(p|x) \propto p^{\alpha+x}(1-p)^{\beta+n-x}$; (c) Using Theorem 3.2, identify the Bayes estimate $\widehat{p}_G$ of p.

Exercise 3.8. Let X be a Bernoulli variable with distribution $X|p \sim \mathscr{B}(1,p)$, and take the prior distribution of p to be the discrete uniform distribution G on the set $S = \{0,1/n,2/n,\ldots,1\}$. Obtain the posterior distribution of p, given $X=x$, and derive the Bayes estimator $\widehat{p}_G$ of p relative to squared error loss.

Exercise 3.9. Let $X|\theta \sim \mathscr{N}(\theta,\sigma_0^2)$, the normal distribution with mean θ and known variance σ_0^2, and suppose $\theta \sim G = \mathscr{N}(\mu_0,\tau_0^2)$, the normal distribution with known mean μ_0 and known variance τ_0^2. Assume that one wishes to estimate θ with squared error loss. (a) Identify the marginal density of X. (b) Obtain the posterior density of $\theta|X=x$ as the ratio $f(x,\theta)/f(x)$. Note that $g(\theta|x)$ can also be identified "by inspection." (c) Using Theorem 3.2, identify the Bayes estimate $\widehat{\theta}_G$ of θ.

Exercise 3.10. Ralph Lauren Inc. produces N polo shirts (a known number) on any given weekday. Historical records show that, on average, a known proportion p of these shirts are defective. Let X be the number of defective shirts in last Friday's batch. Assume that $X \mid N,p \sim \mathscr{B}(N,p)$. Assume that the value of X is unknown. Suppose that the Ralph Lauren store at Pavilions in Sacramento received a known number n of polo shirts from last Friday's batch. Let Y be the number of defective shirts received by the Pavilions' store. Assuming a random distribution scheme, Y has the hypergeometric distribution, that is, $Y \mid N,x,n \sim \mathscr{HG}(N,x,n)$ with probability mass function (pmf) given by

$$P(Y=y \mid N,x,n) = \frac{\binom{x}{y}\binom{N-x}{n-y}}{\binom{N}{n}} \quad \text{for } 0 \le y \le x\,.$$

Your aim is to derive the Bayes estimator of the unknown (parameter) x with respect to squared error loss based on the observed value $Y=y$.

a) Show that the marginal distribution of Y is binomial, that is, show that $Y \mid n,p \sim \mathscr{B}(n,p)$.
b) Identify $p(x|y)$, the conditional pmf of X given $Y=y$. (Hint: Note that if $Y=y$, the range of X is $y \le x \le N-n+y$. The pmf of $X-y \mid Y=y$ is more easily recognized than that of $X \mid Y=y$.)
c) Find $E(X \mid Y=y)$, the Bayes estimator of X, relative to squared error loss, based on the observation $Y=y$. This represents your best guess at X when p is known in advance and Y is observed to be y.

3.4 The Bayes risk

Given $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} F_\theta$ and $\theta \sim G$, the Bayes risk of the decision rule δ is given by

$$r(G, \delta) = E_G R(\theta, \delta) \,, \tag{3.10}$$

where R is the risk function of δ given in (1.4). The Bayes risk sits squarely on the cusp between Bayesian and frequentist thinking, as it is clearly a subjective measure of performance depending on the statistician's prior assessment G concerning the value of θ, and yet it also includes the process of averaging over all possible values that the data $\mathbf{X}$ could take on. The definition of r in (3.10) is written in terms of a certain ordering in the averaging (that is, summing or integrating) process, first wrt F_θ and then wrt G:

$$r(G, \delta) = E_\theta E_{\mathbf{X}|\theta} L(\theta, \delta(\mathbf{X})) \,. \tag{3.11}$$

Under the mild conditions which allow interchanging the order of these operations (for continuous problems, consult Fubini's Theorem), we may also write the Bayes risk of the rule δ as

$$r(G, \delta) = E_{\mathbf{X}} E_{\theta|\mathbf{X}=\mathbf{x}} L(\theta, \delta(\mathbf{x})) \,. \tag{3.12}$$

The reader will recognize the inner expected value in (3.12) as the posterior expected loss of the decision rule δ. Of course the Bayes rule with respect to the prior G is precisely the rule δ^* that minimizes the posterior expected loss for each fixed $\mathbf{x}$. If δ^* has this latter property, then averaging its posterior expected loss over the marginal distribution of $\mathbf{X}$ (that is, executing the outer expectation in (3.12)) will necessarily yield the smallest average. Thus, the Bayes rule δ^* with respect to G, that is, the rule given by

$$\delta^*(\mathbf{x}) = \underset{\widetilde{\delta}}{\operatorname{argmin}} E_{\theta|\mathbf{X}=\mathbf{x}} L(\theta, \widetilde{\delta}(\mathbf{x})) \,, \tag{3.13}$$

is the rule that minimizes $r(G, \delta)$ as well, and vice versa. From this, one can conclude that if the frequentist were to accept the Bayesian premise that one should deal with an unknown parameter by placing a particular probability distribution on it, then the frequentist would be led to the Bayes rule δ^* as the optimal rule in the frequentist sense. In what follows, we will focus on problems of estimation and will therefore write $r(G, \widehat{\theta})$ for the Bayes risk $E_\theta E_{\mathbf{X}|\theta} L(\theta, \widehat{\theta}(X))$ of the estimator $\widehat{\theta}$.

Exercise 3.11. (DeGroot (1970)) Let G_1 and G_2 be two proper prior distributions on the parameter space Θ. Show that for any number $\alpha \in (0, 1)$,

$$\mathbf{r}^*(\alpha G_1 + (1-\alpha)G_2) \geq \alpha \mathbf{r}^*(G_1) + (1-\alpha)\mathbf{r}^*(G_2),$$

where $\mathbf{r}^*(G) = \inf_\delta \mathbf{r}(G, \delta)$ is the Bayes risk of the Bayes rule wrt G, and $\alpha G_1 + (1-\alpha)G_2$ represents the distribution function that is equal to G_1 with probability α and is equal to G_2 with probability $1-\alpha$. Such distributions are typically referred to as "mixtures."

3.5 The class of Bayes and "almost Bayes" rules

It should be quite clear from the above how a Bayesian goes about his business, at least in principle. There are, of course, practical issues that arise in the implementation of Bayesian inference, such as how the prior should be chosen and how to do the calculus involved. The latter question has been resolved to almost everyone's satisfaction by the host of modern iterative methods (Markov chain Monte Carlo methods and their many cousins) that render the process of approximating posterior distributions reliably quite feasible. The *Complete Class Theorem* in Decision Theory says (roughly) that, under certain fairly weak assumptions, every "good" decision rule is a Bayes rule with respect to some prior distribution. (For a formal statement and proof of two versions of the theorem, see Ferguson (1967, Chapter 2); for a general treatment, see Le Cam (1955).) In light of this theorem, one might wonder why one would bother looking elsewhere. A closer look at the completeness of the class of Bayes rules in statistical estimation problems gives an important clue. It turns out that the completeness of the class of Bayes rules in standard estimation problems (where Θ is an interval) includes a hitch — one has to include decision rules that are "almost Bayes" in the following sense.

Definition 3.1. *An estimator $\widehat{\theta}$ is said to be an* extended Bayes rule *if $\widehat{\theta}$ is ε-Bayes for every $\varepsilon > 0$, that is, if for arbitrary $\varepsilon > 0$, there exists a prior distribution $G = G(\varepsilon)$ such that*

$$r(G, \widehat{\theta}) \leq \inf_{\widetilde{\theta}} r(G, \widetilde{\theta}) + \varepsilon \, . \tag{3.14}$$

General forms of the Complete Class Theorem state the class of extended Bayes rules (which of course includes the class of all Bayes rules with respect to probability distributions G on Θ) is an "essentially complete class" (the latter class being defined as a class which contains a decision rule as good or better than any decision rule outside the class). Another common formulation of the concept of "almost Bayes" rules is defined as follows.

Definition 3.2. *Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$, and let τ be a σ-finite measure on the parameter space Θ for which $\tau(\Theta) = \infty$. For a fixed loss function $L(\theta, a)$, the estimator $\widehat{\theta} = \widehat{\theta}(\mathbf{x})$ is a* generalized Bayes rule *with respect to τ if*

$$\int L(\theta, \widetilde{\theta}(\mathbf{x})) f_\theta(x_1, x_2, \ldots, x_n) \, \mathrm{d}\tau(\theta) \tag{3.15}$$

takes its minimum when $\widetilde{\theta}(\mathbf{x}) = \widehat{\theta}(\mathbf{x})$.

The measure τ in the definition above is typically taken to be Lebesgue measure in the problem of estimating a location parameter and is taken as a measure for which $\mathrm{d}\tau(\theta)/\mathrm{d}\theta$ is proportional to $1/\theta$ when θ is a scale parameter. Since such measures do not have probabilistic interpretations, the estimators which result from minimizing (3.15) are not associated with a subjective Bayesian analysis. We will refer to priors whose densities integrate to one as *proper priors*, and prior measures

τ which assign infinite weight to the parameter space as *improper priors*. When first considered, improper priors were seen as a way to represent "prior ignorance." Since Lebesgue measure on the real line gives every interval of the same length equal weight, one can see why this term might seem apt. Over time, the nomenclature evolved, due in part, perhaps, to the fact that no statistical practitioner would be particularly happy to begin an analysis with the label "ignorant." For a time, such priors were called "noninformative" priors, but today, they tend to be referred to as "objective" priors, both because the name has a certain appealing ring to it, and because the term is the natural complement of existing alternatives, i.e., subjective priors. One motivation for entertaining improper priors is the following interesting fact.

Theorem 3.5. *Consider estimating the parameter θ under squared error loss. If $\widehat{\theta}$ is an unbiased estimator of θ, then $\widehat{\theta}$ is not a Bayes rule with respect to any proper prior G.*

This theorem may be proven by establishing the following contradiction: the unbiasedness of the estimator $\widehat{\theta} = E(\theta|\mathbf{X})$ will imply that $r(G,\widehat{\theta}) = 0$. However, in reality, the Bayes risk of $\widehat{\theta}$ must be the positive number $E_G[V(X|\theta)]$. Now, suppose we are interested in estimating the mean μ of a normal population or the proportion p of items of a certain type when sampling with replacement from a finite population. The standard estimators $\overline{X}$ of μ and $\widehat{p}$ of p are well known for their unbiasedness and, in fact, are the UMVUEs of μ and p, respectively, under the corresponding normal and Binomial models. Both are admissible (in the univariate problems under discussion), and both are widely used. However, according to Theorem 3.5, neither is a Bayes estimator with respect to any prior probability distribution under squared error loss. But both estimators are easily shown to be extended Bayes and they are also generalized Bayes estimators with respect to improper priors. Thus, even though they are not estimators which can be derived through a subjective Bayesian analysis in the contexts described above, each is included in the essentially complete class about which the Complete Class Theorem speaks.

Exercise 3.12. Prove Theorem 3.5.

Exercise 3.13. Let $X|p \sim \mathscr{B}(n,p)$, the binomial distribution with parameters n and p. Assume that one wishes to estimate p with squared error loss. Show that the estimator $\widehat{p}_G = X/n$, the sample proportion of "successes" in the binomial experiment, is a generalized Bayes rule wrt the improper prior measure G with $\mathrm{d}G/\mathrm{d}p = g(p) = 1/p(1-p)$.

Exercise 3.14. Let $X|\theta \sim F_\theta$ and suppose that θ has the prior distribution G. Show that if G is an improper prior, then the marginal distribution of X is also improper. (For simplicity, assume that $X|\theta$ has a density or probability mass function f_θ and that G has derivative g.) Verify this property when $X|\theta \sim \mathscr{B}(n,\theta)$ and G is the improper prior with $g(\theta) = \theta^{-1}(1-\theta)^{-1}$. Comment on this property of improper priors in general, and also in the special case in which the random variable X has bounded support.

Exercise 3.15. Let $X|\theta \sim \mathcal{N}(\theta, \sigma_0^2)$, the normal distribution with mean θ and known variance σ_0^2. Assume that one wishes to estimate θ with squared error loss. Show that the estimator $\widehat{\theta}_G = X$ is a generalized Bayes rule wrt Lebesgue measure G with $\mathrm{d}G/\mathrm{d}\theta = g(\theta) = 1$.

Exercise 3.16. In general, Bayes estimators do not enjoy an "invariance" property that would guarantee that $h(\widehat{\theta})$ is the Bayes estimator of $h(\theta)$ when $\widehat{\theta}$ is the Bayes estimator of θ. Demonstrate this fact by showing that $\widehat{p}^2$ is not the Bayes estimator of p^2 relative to squared error loss and the uniform prior $\mathcal{U}[0,1]$ on p, where $\widehat{p} = \frac{X+1}{n+2}$, with $X|p \sim \mathcal{B}(n,p)$.

3.6 The likelihood principle

It is perhaps a good time to introduce an important companion of the axioms of "coherent" behavior. The *Likelihood Principle* is a fundamental tenet of Bayesian inference, and examples of its violation in the course of executing a frequentist procedure constitute a healthy proportion of the settings in which such procedures are judged to be incoherent. Berger and Wolpert's (1988) monograph is the definitive reference on the subject. These authors point out that, while the principle is not the exclusive province of the Bayesian school, it is most staunchly defended by that school and is often mentioned by Bayesians as the primary reason they adopted the Bayesian view.

The likelihood function $L(\theta)$ was defined in (2.19). It is to be viewed as a function of the parameter θ alone, as the data in the formula in (2.19) are taken to be fixed and known. The "principle" of interest here may be formally stated as follows. It is stated in the i.i.d. case for simplicity, but its generalization to more complex data schemes is straightforward.

The Likelihood Principle. Suppose that an experiment is performed involving a random sample $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$, where θ is an unknown parameter. Suppose that the outcome of the experiment is $\mathbf{X} = \mathbf{x}$. Let $L(\theta)$ be the likelihood function given by

$$L(\theta) = L(\theta \mid x_1, x_2, \ldots, x_n) = f_\theta(x_1, x_2, \ldots, x_n)\,. \tag{3.16}$$

Then $L(\theta)$ contains all the information about θ that can be gleaned from the experiment. All inferences about θ should depend on the observed data only through L.

The implications of the likelihood principle are quite broad. Among them is the fact that our inferences about θ should not take into account what might have been observed (values of $\mathbf{X}$ that didn't occur) but, instead, should be based solely on what we did observe. Choosing an estimator based on properties it might have when its performance is averaged over the whole sample space X is a strategy that violates the likelihood principle. Using artificial randomization (for example, using a randomized decision rule in an estimation problem or using randomization to achieve a certain desired significance level in a hypothesis test) violates the likelihood principle. One

is violating the likelihood principle when one makes different inferences about an unknown parameter θ in a situation in which two different experimental designs lead to the same likelihood. The latter issue is well illustrated by the following example.

Example 3.3. One might get the outcome 9 heads and 3 tails in 12 coin tosses in a wide variety of different ways. One might have performed a Binomial experiment in which $n = 12$ tosses were made and $X = 9$ heads were observed. One might obtain 9 heads and 3 tails in an inverse sampling framework in which a coin was tossed repeatedly until 9 heads occurred (calling for the negative binomial model). One might have been planning on tossing the coin 100 times, but stopped when dinner was ready. One might have planned to stop tossing if and when the proportion of heads was at least 0.75. In all of these instances, the likelihood from the observed data is proportional to $\theta^9(1-\theta)^3$, where θ represents the probability of heads. The likelihood principle implies that we should make precisely the same inference about θ in each of these situations. In sequential analysis, it can be explicitly proven that, given a prior distribution G on an unknown parameter θ, the decision rule that minimizes the posterior expected loss plus cost (that is, the Bayes rule, as an estimator of θ) does not depend on the stopping rule, provided that the stopping rule itself is noninformative about θ. ■

Interestingly, frequentist methods often violate the likelihood principle. For example, in testing the hypothesis $H_0 : \theta = 1/2$ against $H_1 : \theta > 1/2$ in the coin tossing context of Example 3.3, the rejection region of the uniformly most powerful test will be different at certain significance levels, depending on whether the experiment actually involved a binomial or a negative binomial setup. The notion of a confidence interval is itself a violation of the likelihood principle, as such intervals are based on what one might observe in some future experiment rather than solely on what was observed in the experiment one actually performed. The following example illustrates this point.

Example 3.4. Let $X_1, X_2, X_3 \overset{iid}{\sim} U[\theta - 1, \theta + 1]$, and let $X_{(1)} < X_{(2)} < X_{(3)}$ be the corresponding order statistics. Since $P(X_i < \theta,\ i = 1,2,3) = 1/8$ and $P(X_i > \theta,\ i = 1,2,3) = 1/8$, we have that $P(X_{(1)} < \theta < X_{(3)}) = 3/4$. Thus, the interval $(X_{(1)}, X_{(3)})$ is a 75% confidence interval for θ. The usual interpretation of this statement is that if we were to repeat this experiment many times, the interval $(X_{(1)}, X_{(3)})$ would capture θ about 75% of the time. But what does that actually tell us about θ based on the experiment we just performed? If one observes $X_{(1)} = 1.5$ and $X_{(3)} = 3$, then the interval $(x_{(1)}, x_{(3)}) = (1.5, 3)$ contains θ with certainty. If, instead, $X_{(1)} = 1.5$ and $X_{(3)} = 1.6$, one would have very little confidence that the interval $(x_{(1)}, x_{(3)}) = (1.5, 1.6)$ has captured θ. A Bayesian would call the whole process "incoherent," and in this particular example, it would be difficult to disagree with this characterization. ■

Let us examine why, in light of the Bayesian premise that one should use probabilistic thinking in handling uncertainty, it follows that Bayesian procedures with respect to proper prior distributions will obey the likelihood principle. When an experiment is planned, but not yet performed, there are two sources of uncertainty. The

first concerns the possible outcome of the experiment. Let us assume that a stochastic model for the experiment has been selected, and for the purposes here, ignore the fact that there is invariably an element of subjectivity in that selection. For simplicity, let's take the modeling of the data which will be observed to be captured in the statement: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$. The second source of uncertainty concerns the unknown value of the parameter θ. When the experiment is performed, and the observed values $x_1, x_2, \ldots, x_n$ of the random variables $X_1, X_2, \ldots, X_n$ are known, the first source of uncertainty is eliminated. What remains is the uncertainty about θ, given the observed data. The Bayesian has a well-prescribed process for dealing with the remaining uncertainty. It is, of course, to update his prior information, encapsulated in his prior distribution G, and derive the posterior distribution $G(\theta|\mathbf{x})$ of θ, given $\mathbf{X} = \mathbf{x}$. The posterior density of θ (in the continuous case) may be written as

$$g(\theta|\mathbf{x}) = L(\theta \mid x_1, x_2, \ldots, x_n) g(\theta) / \int L(\theta \mid x_1, \ldots, x_n) g(\theta)\, \mathrm{d}\theta \,,$$

and clearly depends on the experiment involved only through the likelihood function. Since all Bayesian inference is based upon the posterior distribution of θ, it is squarely aligned with the likelihood principle. Further, it includes a prescription for what to do with the likelihood function, as maximizing expected utility is the required next step.

It is natural to discuss, next, whether generalized Bayes rules obey the likelihood principle. It should be clear from (3.15) that they do. Estimators that are Bayes with respect to improper priors do in fact depend on the experiment that is performed only through the likelihood function. Where this approach runs afoul of the laws of coherent Bayesian inference is in its failure to use probability assessments in the quantification of uncertainty. The measure τ associated with the generalized Bayes rule in Definition 3.2 has no probabilistic interpretation. It is impossible to discuss the relative likelihood of subsets of Θ like $\{\theta \leq \theta_0\}$ and $\{\theta > \theta_0\}$ in terms of τ, and the axioms of Bayesian inference, and their consequences, are inapplicable. Thus, the process of utilizing an "objective" or generalized prior τ must be classified as incoherent in Bayesian terms.

Ronald Fisher, a frequentist with a capital F, was one of the early proponents of the likelihood principle, though his approach to it differed from the above. Fisher (1925, 1932) proved, for example, that the function of the data on which the random likelihood function $L(\theta \mid X_1, X_2, \ldots, X_n)$ depends, is a minimal sufficient statistic for θ, and thus contains all the information about θ that the experiment has to offer. In (most) estimation problems, Fisher advocated the use of maximum likelihood estimators of θ, i.e., estimators that maximized $L(\theta \mid x_1, x_2, \ldots, x_n)$, and were thus consonant with the likelihood principle as stated above. G. A. Barnard (1949, 1962) was another early advocate of the likelihood principle, albeit from a frequentist perspective. But the principle has not become a hallmark of the classical approach to statistical inference as it has for the Bayesian approach. It must be recognized that, at its core, the likelihood principle is a statement based on logic and intuition. Indeed, it has been argued that it can be deduced from unassailable first principles. (Notable developments with the latter aim include the derivation of the Likelihood Principle

as a consequence of the "Sufficiency Principle" and the "Conditionality Principle"; see Robert (2001).) It nonetheless has the status similar to that of an axiom, and there are statisticians who don't find the axiom compelling. There are also a variety of practical concerns that have been raised. Among them is the fact that the principle tacitly depends on a fixed model for the data that will be available, and as discussed above, the model, in a given practical application, is often thought of as tentative and approximate. Berger and Wolpert (1988) propose generalizations of the likelihood principle aimed at addressing this concern, and discuss, quite comprehensively, the arguments for and against the adoption of the principle. We refer the reader to that monograph for further details.

Exercise 3.17. Consider the experiment in Example 3.3, and suppose that θ represents the probability of heads. Suppose one is interested in testing the hypothesis $H_0 : \theta \leq 1/2$ against $H_1 : \theta > 1/2$ at significance level $\alpha = 0.05$. Show that the outcome $X = 9$, where X is the number of heads observed in 12 tosses of a coin, leads to the "acceptance" of H_0 if $X \sim \mathscr{B}(12, \theta)$, a binomial experiment based on 12 trials, but leads to the "rejection" of H_0 if $X \sim \mathscr{NB}(9, \theta)$, a negative binomial experiment which terminates the sampling process as soon as 9 heads are obtained. Note that the two experiments give rise to the same likelihood (proportional to $\theta^9(1-\theta)^3$). Conclude that this constitutes an example of a classical hypothesis testing procedure that violates the likelihood principle.

3.7 Conjugate prior distributions

In the context of Bayesian inference, the term *conjugacy* is used to describe a particular type of relationship between the model that governs the available data and the prior distribution of the unknown parameter. Conjugate prior families $\Gamma = \{G_\lambda, \lambda \in \Lambda\}$ of distributions are usually thought of in terms of a "closure" property they obey. This property is simply the requirement that the posterior distribution be a member of Γ whenever the prior distribution is a member of Γ. While this property cannot uniquely identify conjugate prior families for a given model for one's data (for example, the class of all distributions on Θ is closed in this sense), there exists a well-known "standard" conjugate class for many stochastic models in common use. More specifically, given a model $\{F_\theta,\ \theta \in \Theta\}$ for a random variable X, we will be interested in a family Γ of prior distributions having the property that $G_{\theta|X=x} \in \Gamma$ whenever $G_\theta \in \Gamma$. Additional restrictions on the family Γ will make the class unique.

A simple and frequently used example of this closure property is the relationship between the Binomial and beta distributions. If $X \mid \theta \sim \mathscr{B}(n, \theta)$ and $\theta \sim \mathrm{Be}(\alpha, \beta)$, then $\theta \mid X = x \sim \mathrm{Be}(\alpha + x, \beta + n - x)$. Diaconis and Ylvisaker (1979) made the following quite pertinent observation. If $g(\theta \mid \alpha, \beta)$ represents the beta density and h is a bounded measurable function on the unit interval, the family $\{ch(\theta)g(\theta \mid \alpha, \beta)\}$, where

$$c = 1 \Big/ \int h(\theta) g(\theta \mid \alpha, \beta)\, \mathrm{d}\theta\ ,$$

also has the closure property; $\theta \mid X = x$ having density $ch(\theta)g(\theta \mid \alpha + x, \beta + n - x)$. It is thus clear that there are uncountably many closed parametric families that may serve as conjugate families for the binomial model for X. When the distribution of X belongs to an exponential family, Diaconis and Ylvisaker prove that only one closed family enjoys the additional property of having a linear posterior mean, that is, satisfies the equation

$$E\{E(X \mid \theta) \mid X = x\} = ax + b\,. \tag{3.17}$$

When $X \sim \mathscr{B}(n, \theta)$, the $\mathrm{Be}(\alpha, \beta)$ distribution is the unique closed family satisfying (3.17), with $a = n/(\alpha + \beta + n)$ and $b = \alpha/(\alpha + \beta + n)$.

In sampling situations dealing with a one-parameter exponential family, we will refer to the closed family of prior distributions with linear posterior expectations as "standard conjugate priors." Table 3.1 includes the most frequently encountered examples. In each case, it is assumed that a random sample $X_1, X_2, \ldots, X_n$ is drawn from the distribution F_θ with density f_θ. Parameters subscripted by "0" are taken as known.

Table 3.1. Conjugate priors for selected one-parameter exponential families

$f_\theta(x)$	$g(\theta)$	$g(\theta \mid x)$
$\mathscr{B}(n, \theta)$	$\mathrm{Be}(\alpha, \beta)$	$\mathrm{Be}(\alpha + x, \beta + n - x)$
$\mathscr{P}(\theta)$	$\Gamma(\alpha, \beta)$	$\Gamma\left(\alpha + \sum_{i=1}^{n} x_i,\ \dfrac{\beta}{n\beta + 1}\right)$
$\mathscr{NB}(r_0, \theta)$	$\mathrm{Be}(\alpha, \beta)$	$\mathrm{Be}\left(\alpha + r_0 n,\ \beta + \sum_{i=1}^{n} x_i - n\right)$
$\mathscr{E}\left(\frac{1}{\theta}\right)$	$\Gamma(\alpha, \beta)$	$\Gamma\left(\alpha + n,\ \left[(1/\beta) + \sum_{i=1}^{n} x_i\right]^{-1}\right)$
$\mathscr{N}(\theta, \sigma_0^2)$	$\mathscr{N}(\nu, \tau^2)$	$\mathscr{N}\left(\dfrac{\nu\sigma_0^2 + \tau^2 \sum_{i=1}^{n} x_i}{\sigma_0^2 + n\tau^2},\ \dfrac{\sigma_0^2 \tau^2}{\sigma_0^2 + n\tau^2}\right)$
$\mathscr{N}(\mu_0, 1/\theta)$	$\Gamma(\alpha, \beta)$	$\Gamma\left(\alpha + n/2,\ \left[(1/\beta) + \dfrac{1}{2}\sum_{i=1}^{n} (x_i - \mu_0)^2\right]^{-1}\right)$

In addition to the one-parameter exponential families featured in Table 3.1, the following "conjugate pairs" of models are well known: the uniform distribution on the interval $[0, \theta]$ and a Pareto prior distribution with location parameter θ_0 and shape

parameter α, the uniform distribution on $[\theta_1, \theta_2]$ and a bilateral Pareto prior distribution, the multivariate normal distribution with known covariance matrix and a multivariate normal prior distribution, the multinomial distribution and a Dirichlet prior distribution, and the multivariate normal distribution with a known mean vector and an inverse Wishart prior distribution. Among these additional examples, the latter three are covered by regularity conditions in the Diaconis–Ylvisaker treatment, and may thus be regarded as uniquely conjugate in the sense of their paper.

Conjugacy is, of course, a mathematical convenience, as the closed form of the posterior density, and of some important functionals depending on it, facilitates the analytical study of their properties. But it is worth noting that conjugacy is also a notion of substantial practical value. Conjugate priors tend to have a good deal of interpretive value, often allowing one to summarize the prior information that is brought into the analysis in terms of one's "best guess" at the unknown parameter and the weight one would wish to place on that guess as compared to the weight that the experimental data deserves. In the binomial–beta pairing, for example, one may parametrize the beta prior in terms of the *prior mean* $\theta_0 = \alpha/(\alpha+\beta)$ and the *prior sample size* $\omega = \alpha+\beta$. With this parametrization, the Bayes estimator of θ is $\widehat{\theta}(X) = (\omega/(\omega+n))\theta_0 + (n/(\omega+n))(X/n)$, so that the estimator is an easily interpretable mixture of one's prior guess at θ and the data-driven guess at θ, the sample proportion X/n of "successes" in n Bernoulli trials.

What makes a class of prior distributions useful for Bayesian inference? One sometimes hears the term "richness" used in this regard. What is meant by "richness" is that the class contains a reasonable array of different characteristics — flexibility of the choice of center and the amount of dispersion, perhaps different shapes and a wide range for, say, the model's coefficient of variation. What such richness provides is the ability to capture prior information or intuition without severe constraints. When one chooses a prior density $g(\theta)$ for a population proportion from the family of beta distributions, one has the option of choosing a U-shaped density, a uniform density, a monotone (either increasing or decreasing) density or a unimodal density tied down at 0 and 1 (i.e., with $f(0) = 0 = f(1)$). The richness of the class notwithstanding, it might be inadequate for handling prior information in a particular problem. If a bimodal prior density (with $f(0) = 0 = f(1)$) is deemed necessary in capturing prior opinion about a population proportion, then clearly one must look outside the beta class for a prior model. Properly reflecting prior knowledge about the unknown parameter is the primary consideration in choosing a prior distribution.

Exercise 3.18. Verify the posterior distributions in Table 3.1 for the negative binomial and the exponential models.

Exercise 3.19. Let $X_1, X_2, \ldots, X_n \mid \theta \overset{iid}{\sim} \mathscr{U}[0,\theta]$, the uniform distribution on $(0,\theta)$, and let $\theta \sim \text{Par}(\alpha, \theta_0)$, the Pareto distribution with density $g(\theta) = \alpha\theta_0^\alpha/\theta^{\alpha+1}$ for $\theta > \theta_0$. Verify that the posterior density of θ, given $x_1, x_2, \ldots, x_n$, is the Pareto distribution $\text{Par}(\alpha+n, \theta_0^*)$, where $\theta_0^* = \max(\theta_0, x_1, x_2, \ldots, x_n)$.

Exercise 3.20. Let $X_1, X_2, \ldots, X_n$ be a random sample from the two-parameter Uniform distribution of the interval (θ_1, θ_2). Take the prior distribution on (θ_1, θ_2) to be

the Bilateral Bivariate Pareto (BBP) distribution with parameters $r_1 < r_2$ and $\alpha > 0$, and density function

$$g(\theta_1, \theta_2 \mid r_1, r_2, \alpha) = \frac{\alpha(\alpha+1)(r_2 - r_1)^{\alpha}}{(\theta_2 - \theta_1)^{\alpha+2}} \quad \text{for } \theta_1 < r_1 \text{ and } \theta_2 > r_2 \, .$$

Identify the posterior density of $\theta_1, \theta_2 \mid x_1, x_2, \ldots, x_n$. (Be sure to keep careful track of the constraints on the range of the θs, given the xs.)

3.8 Bayesian robustness

The term *robustness* is used quite differently by a Bayesian than it is by a frequentist. A Bayesian's estimate of an unknown parameter θ is of course influenced by his choice of prior G, so that a robust Bayesian procedure would be one which is fairly stable when the prior is moderately perturbed around G (or that gives fairly close answers when the chosen prior is replaced by "reasonable" alternative priors). Without formally defining the term, those who seek a robust Bayes estimator would typically examine an appropriate neighborhood Ω of the prior G and determine the extent to which Bayesian quantities change as the prior changes within Ω. This process is usually referred to as "sensitivity analysis." Analytically, sensitivity analysis is not an easy assignment for any among a reasonable collection of choices for Ω and for any Bayesian quantity one might study. Possible choices of Ω include distribution functions which are suitably close to a prior distribution G (e.g., all F such that $\sup_x |F(x) - G(x)| < \varepsilon$ for some small ε), distributions for which certain percentiles are suitably close to those of G, distributions for which certain moments are suitably close to those of G or distributions in a fixed class of prior distributions for which the prior parameters are suitably close to those of G. There are many other possibilities, including that of simply examining an alternative class of prior distributions, say translated t-distributions instead of normal distributions as priors for a normal mean. Possible choices of Ω are discussed at some length by Berger (1985).

The second element of a "Bayesian robustness check" involves the specification of a measure by which Bayesian output may be compared. Monitoring the effect of perturbations of the prior on (i) the posterior distribution of θ, (ii) the Bayes estimator of θ, and (iii) the Bayes estimator's risk function are commonly considered options in gauging the stability of a prior. When one restricts attention to a conjugate family of priors, the comparison of the effect of perturbing the prior can generally be done analytically. Otherwise, some form of simulation would typically be needed. In either case, the outcome of such an investigation involves some *ad hoc* choices and thus will not be definitive; it can, however, allay concerns about possible prior "misspecification." A sensitivity analysis is judged to be confirmatory if the chosen measure of a prior's influence on the analysis varies only moderately as the prior varies within a reasonable class of alternatives to the prior distribution that is proposed for use.

Savage (1962) pointed out that in many statistical problems, the sample size is sufficiently large to make the sampling distribution F_θ of any reasonable estimator of

θ quite concentrated, while the densities of the prior distributions considered "reasonable" in the problem are quite flat in the interval to which F_θ assigns most of its mass. If, in addition, the prior densities under consideration aren't unduly large outside of this interval (it suffices, for example, for them to be bounded), then the influence of the prior on the posterior will be moderate and, more importantly, will not vary a great deal from one prior to another. Savage referred to this as the *principle of precise measurement*. In such cases, the influence of the prior distribution on the resulting Bayesian inference will be fairly modest, and the statistician will largely be free of any worries about prior misspecification.

The logical conundrum in Bayesian robustness is the fact that it is precisely when a prior model seems to contain weighty information (and thus differs from other priors one could use, particularly ones which are relatively noninformative) that Bayesian inference stands to be the most useful. Thus, noting that the prior one has chosen has a strong influence on the posterior analysis is not sufficient reason, by itself, to alter one's prior specification. The utility of a sensitivity analysis thus goes beyond the possibility of confirming that reasonable alternative priors give roughly the same answer; it also may serve as a bellwether which informs the statistician that a particular prior has considerable influence on the analysis. The latter insight is useful, both for sharing with any potential consumers of the analysis and also for the purpose of leading to a careful scrutiny of the introspection and consultation that went into the determination of one's prior. I would be remiss if I didn't mention a suggestion that appears in certain corners of the Bayesian literature on Bayesian robustness. I have seen, more than once, the suggestion that, to be sure that one's prior isn't way off the mark, one should take a peek at the data and decide on one's prior distribution on the unknown parameter after that. Nothing could be more incoherent than such a practice! It's the moral equivalent of choosing what hypothesis to test after taking a look at what's "provable" from the observed data. If one is going to rely on the data to choose one's prior, it seems appropriate and reasonable to go all the way and let the data do all the talking through a carefully selected frequentist procedure. The practice of data peeking is in direct conflict with the Bayesian paradigm.

We close this section with a brief examination of a quite different approach to Bayesian robustness. Take Γ to represent a class of distributions $\{G\}$, all of which are considered viable candidates as priors for an unknown parameter θ. One way to deal with the difficulty of choosing a specific prior $G \in \Gamma$ is to adopt a strategy that protects the statistician against the entire class of priors in Γ. A decision rule that does so is defined below.

Definition 3.3. *A decision rule δ_0 is said to be* Γ-minimax *if*

$$\sup_{G\in\Gamma} r(G,\delta_0) = \inf_{\delta} \sup_{G\in\Gamma} r(G,\delta)\ . \tag{3.18}$$

A given estimator $\widehat{\theta}$ may have a small Bayes risk against one prior in the class Γ but a large Bayes risk against another. Consider the worst result possible, that is, the maximum Bayes risk that $\widehat{\theta}$ experiences against all priors in Γ. A Γ-minimax estimator

has the smallest possible "maximum Bayes risk" against priors in the class, and thus offers the best protection against the class as a whole. It should be acknowledged that this is a frequentist approach to Bayesian robustness in that the "optimal" estimator is chosen based on its expected behavior relative to the entire sample space. This of course violates the likelihood principle, and thus is beyond the scope of the orthodox Bayesian. Defenders of the approach might argue that the approach is Bayesian in spirit, as it strikes a compromise resulting in an estimator that has "reasonable" performance over a range of possible priors. A Γ-minimax estimator $\widehat{\theta}$ might not be Bayes with respect to any particular prior $G \in \Gamma$, but it has better performance (in terms of posterior expected loss), against at least one $G \in \Gamma$, than any alternative estimator. If Γ consists of a single distribution G, then $\widehat{\theta}$ is Bayes with respect to G, while if Γ contains all degenerate distributions, then $\widehat{\theta}$ is a minimax estimator of θ.

An elementary example of a Γ-minimax estimator (based on $X \sim \mathscr{B}(1,\theta)$) is given in Samaniego (1975). There, the estimator $\widehat{\theta} = (1+2X)/4$ is shown to be Γ-minimax, under squared error loss, with respect to the class of priors Γ on $[0,1]$ with mean $= 1/2$. Berger (1985) derives the Γ-minimax estimator of the mean of the normal distribution $\mathscr{N}(\theta,1)$, under squared error loss, with respect to the class of all priors with a given finite mean μ and variance σ^2.

Exercise 3.21. Suppose that in a given estimation problem, $\widehat{\theta}_1$ is the unique Γ-minimax estimator for the class of priors Γ on an unknown parameter θ. Show that, for any alternative estimator $\widehat{\theta}_2$ of θ, there exists a prior $G^* \in \Gamma$ such that $r(G^*, \widehat{\theta}_1) < r(G^*, \widehat{\theta}_2)$.

Exercise 3.22. In a given decision problem (Θ, D, L), suppose the decision rule δ^* is an equalizer rule (that is, has constant risk function). Let Γ be a class of prior distributions on Θ. Show that if δ^* is a Bayes rule with respect to some prior $G^* \in \Gamma$, then δ^* is Γ-minimax.

3.9 Bayesian asymptotics

Bayesians are not generally concerned about the asymptotic performance of Bayes procedures. The likelihood principle dictates that one's attention should always be focused on the observed data in the experiment at hand, and thus any musing about how the procedure might behave if the sample size were allowed to grow to infinity lies well beyond the scope of a Bayesian treatment of the problem. This notwithstanding, mathematical statisticians have wondered how the Bayesian approach fares in large samples, and of course, in the limit, as $n \to \infty$. What would one guess the answer to be?

First, one might think about what would happen, as n grows, to two Bayesians who started out with quite different opinions about an unknown parameter θ. One might conjecture that if the two priors had different support sets, then the two Bayes estimators might converge to two different values of θ. It is quite easy to confirm that the support of the posterior distribution of θ must be a (possibly proper) subset

of the support set of the prior distribution of θ. That being the case, one sees that there could be unresolvable disagreements between Bayesians with priors supported on disjoint subsets of Θ. In estimation problems, it is virtually always the case that a "thoughtful Bayesian" will choose a prior whose support set is the entire parameter space Θ. This is a mild restriction indeed, as one's prior can still be almost totally concentrated on a particular subset Θ_0 of Θ while placing miniscule, though non-zero, weight on parameter values in the complementary subset $\Theta - \Theta_0$.

Suppose, now, that two Bayesians have followed the practice above, that is, have selected prior distributions whose support set is Θ. Then it can be shown that the two Bayes estimates will converge to the true value of θ. Writings on this problem are sometimes referred to as part of the "merging of opinion" literature, taking this name from the famous paper by Blackwell and Dubins (1962). Although shedding insight into prior elicitation was not the intended focus of this paper, it is relevant to the elicitation problem in that it alerts the Bayesian to the wisdom in selecting priors that do not preclude the possibility that the true θ lies in any given subset of the parameter space, and it also indicates that all such priors, even ones that might be considered "wrong" in some objective sense, are eventually overwhelmed by the information provided by the sample, so that any pair of such Bayes estimators give roughly the same answer when the sample size is sufficiently large.

One can say considerably more than this under the same regularity conditions under which the MLE is guaranteed to be the best asymptotically normal estimator of θ. The main result in this area is generally known as the Bernstein–von Mises Theorem (see Ferguson, 1996). Suppose that the prior distribution G has density $g(\theta) > 0$ for all $\theta \in \Theta$, and let $\widehat{\theta}_n$ be the MLE of θ based on a random sample of size n from F_θ. Then the conditional expectation of $\sqrt{n}(\theta - \widehat{\theta}_{\mathrm{ML}})$, given the data, converges to 0 almost surely. Indeed, the assumed conditions imply that the Bayes estimator $\widehat{\theta}_G$ of θ under squared error loss is asymptotically normal with distribution specified by

$$\sqrt{n}(\widehat{\theta}_G - \theta_0) \xrightarrow{D} Y \sim \mathcal{N}(0, I^{-1}(\theta_0)) \,, \tag{3.19}$$

where θ_0 represents the true value of the parameter θ. This result asserts that, under standard regularity conditions, the posterior distribution of θ has the same limiting form as the distribution of the MLE and that $\widehat{\theta}_G$ is thus a strongly consistent, efficient estimator of θ, that is, like the MLE, it is also best asymptotically normal.

Exercise 3.23. Let $X|p \sim \mathcal{B}(n, p)$, the binomial distribution with parameters n and p. Assume that one wishes to estimate p with squared error loss. Let $\widehat{p}_1$ and $\widehat{p}_2$ be the Bayes estimators with respect to the beta prior distributions $\mathrm{Be}(\alpha_1, \beta_1)$ and $\mathrm{Be}(\alpha_2, \beta_2)$, respectively, where $\alpha_i > 0$ and $\beta_i > 0$ for $i = 1, 2$. Show that both of these estimators converge in probability to the parameter p as the sample size $n \to \infty$.

3.10 Bayesian computation

Let us examine where the difficulty lies in the analytical derivation of a Bayes estimator in a particular problem. Though the difficulty we'll discuss is significantly

magnified in higher dimensions, its character reveals itself quite clearly in the univariate setting we've been discussing. Assume that $X|\theta \sim F_\theta$, a distribution with density $f_\theta(x)$, and that $\theta \sim G$, a prior distribution with density g. We then can express the posterior density of θ as

$$g(\theta \mid x) = f(x \mid \theta)g(\theta)/f(x) \,. \tag{3.20}$$

Now obtaining a closed form expression for the posterior density requires that the marginal density $f(x)$ be obtained in closed form, a process that requires the integration of the joint density of the random pair (X, θ) with respect to θ. This is the simplest version of a problem that caused Bayesians massive headaches in the era preceding the ready availability of high-speed computation. In many problems of practical interest, the integrals needed in an exact Bayesian analysis could not be obtained analytically, and some were even difficult to approximate numerically in a reliable way. Not having $f(x)$ in hand translates into not having the posterior distribution of θ in hand, and thus the Bayesian solution to the problem could only be described in theory but its derivation in that problem, given real data, could not be implemented.

I heard the following story from Jay Kadane. It concerns a wise old owl and a whining centipede. The centipede was complaining to the owl about foot pain. Having so many feet, it seemed like pain was unavoidable, since at any given time, some of his feet would be bruised and sore. The owl gave the matter some thoughtful introspection and then triumphantly announced that he had a solution. What the centipede needed to do was to walk about a half inch above the ground. The centipede immediately conceded that this would solve his problem, but added that he had no idea how to implement the solution. The owl replied "Neither do I, but what I've done is solve your problem in principle. It's up to you to find a way to put the solution into practice!"

This fable describes a situation not unlike that in which Bayesians found themselves prior to the introduction of computer-intensive methods in Statistics. Efron's (1979) introduction of the bootstrap played a major role in this revolution in the discipline. Since then, the practical utility of computer-driven analyses in Statistics has become increasingly obvious and the range of their applications continues to increase at an ever-accelerating pace. In Bayesian statistics, three papers that played especially influential roles in demonstrating the utility of computer-intensive analyses in Bayesian inference are those by Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990).

In the one-parameter setting we have focused on up to now, the exact derivation of the posterior density $g(\theta|x)$ might pose analytical problems. There are a number of approaches that have proven useful in the approximation of this density. Numerical integration is often feasible. Analytical approximations include the development of large-sample normal approximations to the posterior density. An example of this approach is given in Yee, Johnson and Samaniego (2002). A widely used alternative is known as Laplace's method and exploits the Taylor series expansions of both $f_\theta(x)$ and $g(\theta)$ as functions of θ, resulting in efficient second-order approximation

of the posterior mean. The latter method may be generalized to obtain an approximation of $g(\theta)$ itself. Details may be found in Tierney and Kadane (1986) and Tierney, Kass and Kadane (1989). Monte Carlo methods (using, for instance, "rejection sampling") for obtaining the posterior distribution of θ have also been developed (see, for example, Devroye (1986)).

Markov chain Monte Carlo (MCMC) methods are the methods of choice in multiparameter problems in which exact results are intractable. These methods are based on iterations which utilize the (full) conditional densities (known, at least, up to normalizing constants), of each parameter given the data and all other parameters. They depend on sampling from these conditionals repeatedly to obtain a sequence of parameter values which, in the limit, behave like random draws from the posterior distribution of θ. There is an enormous literature on MCMC methods; for simplicity, we describe the basic idea here for a particular two-parameter case.

Suppose that the model of interest has components X with density $f(x \mid \theta_1, \theta_2)$ and a parameter pair (θ_1, θ_2) with prior density $g(\theta_1, \theta_2)$. In many Bayesian applications involving two or more unknown parameters, problems arise in the derivation and/or management of the posterior distribution of the pair (θ_1, θ_2). In some cases, the derivation is simply intractable, and some kind of approximation of it is required. In other cases, $g(\theta_1, \theta_2 \mid \mathbf{x})$ can be explicitly derived but takes a complex, unfamiliar form which makes the derivation of related quantities (like the posterior mean, median or mode) difficult. MCMC methods will be applicable to such problems if the full conditionals $g(\theta_1 \mid x, \theta_2)$ and $g(\theta_2 \mid x, \theta_1)$ can be obtained (up to a normalizing constant). MCMC iterations begin with an initial choice of a value of one of the parameters, say $\theta_1^{(1)}$. Having initialized the iterative process, one then draws a random θ_2 from the density $g(\theta_2 \mid x, \theta_1^{(1)})$ and denotes the value obtained as $\theta_2^{(1)}$. Repeated iterations yield $\theta_1^{(i)} \sim g(\theta_1 \mid x, \theta_2^{(i-1)})$ and $\theta_2^{(i)} \sim g(\theta_2 \mid x, \theta_1^{(i)})$. After an appropriate "burn in" period, the successive iterations of the pairs $(\theta_1^{(t)}, \theta_2^{(t)})$, $(\theta_1^{(t+1)}, \theta_2^{(t+1)}), \ldots$ behave, approximately, like a sample from the posterior density $g(\theta_1, \theta_2 \mid \mathbf{x})$, the stationary distribution of the associated Markov chain. Thus, the posterior density, and all posterior quantities of interest, may be approximated from the burned-in Markov chain. An elementary proof of the convergence of the chain in an example involving two parameters is given in Casella and George (1992). General convergence results may be found in Tanner and Wong (1987) and a special treatment involving exponential families and conjugate priors appears in Diaconis *et al.* (2008). For a comprehensive treatment of MCMC methods, see the recent book by Robert and Casella (2004).

Example 3.5. As an illustration of the process above, let's consider the estimation of the mean μ of a normal distribution with unknown precision $\tau = 1/\sigma^2$ based on the single observation x. Let's suppose that their prior distributions are $\mathcal{N}(a, b)$ and $\Gamma(c, d)$, respectively, where a, b, c and d are known constants, with μ and τ assumed to be independent *a priori*. The posterior distribution of μ and τ is easily determined up to a scalar constant, that is,

$$g(\mu,\tau|x) \propto \sqrt{\tau}\exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\}\exp\left\{-\frac{1}{2b}(\mu-a)^2\right\}\tau^{c-1}\exp\left\{-\frac{\tau}{d}\right\}. \quad (3.21)$$

But since the marginal posterior distribution of μ takes the unrecognizable form

$$g(\mu \mid x) \propto \frac{1}{[(1/2)(x-\mu)^2+1/d]^{1/2+c}}\exp\left\{-\frac{1}{2b}(\mu-a)^2\right\},$$

obtaining posterior quantities of interest (such as $E(\mu|X)$) poses an analytical challenge. On the other hand, from (3.21), we may infer that the full conditionals may be identified in closed form as the normal and gamma distributions given by

$$\mu \mid x,\tau \sim \mathscr{N}\left(\frac{x\tau+a/b}{\tau+1/b},\ \frac{1}{\tau+1/b}\right) \quad (3.22)$$

and

$$\tau \mid x,\mu \sim \Gamma\left(1/2+c,\ \frac{1}{(1/2)(x-\mu)^2+1/d}\right). \quad (3.23)$$

Using (3.22) and (3.23), MCMC iterations will lead to, among other things, a reliable approximation of the marginal posterior of μ and of its mean. ■

Another context in which MCMC methods have proven very useful is in Bayesian analysis of data based on models with a hierarchical structure. Suppose, for example, $X|\theta$ has density $f(x \mid \theta)$, where θ has prior density $g(\theta \mid \eta)$ and η has prior density $p(\eta)$, i.e., η is a hyperparameter on which the density of θ depends. It is also assumed that $f(x|\theta,\eta) = f(x|\theta)$. Then $g(\theta|x,\eta)$ is proportional to $f(x|\theta)g(\theta|\eta)$ and $g(\eta|x,\theta) = p(\eta|\theta)$. (See Carlin and Louis (2008), Section 5.4, for a concrete example.) MCMC iterations begin with a single draw from $p(\eta)$, which we denote by $\eta^{(1)}$. We then draw a random θ from the density $g(\theta \mid x,\eta^{(1)})$, and denote the value obtained as $\theta^{(1)}$. Repeated iterations yield $\eta^{(i)} \sim g(\eta|x,\theta^{(i-1)})$ and $\theta^{(i)} \sim g(\theta|x,\eta^{(i)})$. After appropriate "burn in," successive iterations $\theta^{(t)},\theta^{(t+1)},\ldots$ may be used to approximate the marginal posterior density $g(\theta|x)$ and related posterior quantities.

Exercise 3.24 (Casella and George). Let X and Y be Bernoulli random variables with joint probability mass function given by $P(X=i,Y=j) = p_{ij}$, for $i,j \in \{0,1\}$, where $p_{ij} > 0$ and $\sum_{i=0,1;j=0,1} p_{ij} = 1$. Note that the marginal distribution of X is $\mathscr{B}(1, p_{10}+p_{11})$. For any given x or $y \in \{0,1\}$, the conditional pmfs for Y or X, respectively, are specified in the matrices $A_{y|x}$ and $A_{x|y}$ implicit in the tables below:

$\mathbf{Y}\|\mathbf{X}$	$Y=0$	$Y=1$
$X=0$	$\frac{p_{00}}{p_{00}+p_{01}}$	$\frac{p_{01}}{p_{00}+p_{01}}$
$X=1$	$\frac{p_{10}}{p_{10}+p_{11}}$	$\frac{p_{11}}{p_{10}+p_{11}}$

and

$\mathbf{X}\|\mathbf{Y}$	$X=0$	$X=1$
$Y=0$	$\frac{p_{00}}{p_{00}+p_{10}}$	$\frac{p_{10}}{p_{00}+p_{10}}$
$Y=1$	$\frac{p_{01}}{p_{01}+p_{11}}$	$\frac{p_{11}}{p_{01}+p_{11}}$

Consider the Markov chain $\{X_i, i = 1,2,3,\ldots\}$ with transition matrix $A_{xx} = A_{y|x} \cdot A_{x|y}$. Let $\mathbf{p}_0 = (p_0(0), p_0(1))$ be the initial probability distribution of X, and let $\mathbf{p}_k = \mathbf{p}_0(A_{xx})^k$. Show that the chain's stationary distribution is $\mathbf{p}_X = (p_{00} + p_{01},\ p_{10} + p_{11})$, the marginal distribution of X. This demonstrates that the "Gibbs sampler" for X, with any initial nondegenerate distribution $\mathbf{p}_0$, converges to the appropriate pmf $(p_x(0), p_x(1))$. (**Hint:** Show that $\mathbf{p}_X A_{y|x} = (p_{00}+p_{10},\ p_{01}+p_{11})$ and that $(p_{00}+p_{10},\ p_{01}+p_{11})A_{x|y} = \mathbf{p}_X$.)

3.11 Bayesian interval estimation

Although I will not pursue the matter further in this monograph, I will close this chapter with some brief comments about how Bayesians do interval estimation. The frequentist notion of a confidence interval, often summarized by statements such as "we're 95% sure that the interval (L,U) contains the true value of the parameter," seems to have a Bayesian flavor. However, from its initial introduction, Neyman (1938) made clear that the interpretation of a confidence interval involved no probabilistic conclusion about the unknown parameter but, rather, was to be interpreted in terms of the relative frequency that the process by which the interval was generated would capture the unknown parameter in repeated trials of the experiment. As seen in Example 3.4, an interpretation such as this, which depends on unobserved data rather than solely on the experimental data in hand, can lead to conclusions that are untenable when applied to the experiment of interest.

For the Bayesian, the notion of interval estimation is simpler and manages to avoid potential conflicts with the observed data. The posterior distribution of the parameter θ comprises the basis for all Bayesian inference about θ. The Bayesian counterpart of a confidence interval for θ is called a *credibility interval* for θ, and is obtained from the posterior distribution by selecting an interval corresponding to the probability level desired. For example, any interval (θ_L, θ_U) for which

$$\int_{\theta_L}^{\theta_U} g(\theta|\mathbf{x})\, \mathrm{d}\theta = 1 - \alpha$$

is a $100(1-\alpha)\%$ credibility interval for θ, where $g(\theta|\mathbf{x})$ is the posterior density of θ. The credibility interval used most often is the central one in which the limits θ_{L} and θ_{U} are chosen to satisfy

$$\int_{-\infty}^{\theta_{\mathrm{L}}} g(\theta|\mathbf{x})\,\mathrm{d}\theta = \alpha/2 = \int_{\theta_{\mathrm{U}}}^{\infty} g(\theta|\mathbf{x})\,\mathrm{d}\theta\ .$$

Credibility intervals represent the statistician's posterior judgment about intervals that contain θ with a given probability. They are clearly in harmony with the likelihood principle.

Exercise 3.25. Suppose $X_1, X_2, \ldots, X_8 \mid \mu \overset{iid}{\sim} \mathcal{N}(\mu, 1)$, the normal distribution with known variance 1, and suppose the prior distribution on μ is $\mathcal{N}(6,1)$. Suppose the mean of the observed sample is $\bar{x} = 4.875$. Identify the posterior distribution of $\mu|\mathbf{x}$ and obtain a central 95% credibility interval for μ. Compare this interval to the classical 95% confidence interval for μ.

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 $\text{Be}(100, 100)$ 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 $\text{Be}(110, 100)$, 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 $\widehat{\theta}_{\mathrm{ML}}$ 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 $\widehat{\theta}_{\mathrm{ML}}$, 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 pa-

rameter, 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 $\mathbf{X}$ 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 $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$. Further, let's suppose that the distribution F_θ belongs to an exponential family. Finally, let L be squared error loss, and let $\mathcal{G}$ 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.

5

Comparing Bayesian and Frequentist Estimators of a Scalar Parameter

5.1 Introduction

As should be evident from the discussion in the three preceding chapters, both the Bayesian and the frequentist approaches to estimation have positive attributes, and yet both also have vulnerabilities that can lead to poor and misleading inferences. The Bayesian paradigm appears to have the advantage in terms of pure logic, both in its foundations and in the methodology that's built upon them. We have noted, however, that a logically consistent analysis might rightly be judged to be inadequate when it leads to a conclusion that is off the mark. The Frequentist school, on the other hand, has an apparent edge in terms of the notion of "objectivity," as it proceeds on the basis of a data-driven model and does not utilize "subjective" inputs concerning unknown population parameters whose influence is often difficult to identify and may, in some circumstances, be detrimental. But "objectivity" has been seen to be a two-edged sword, as simple examples make it abundantly clear that subjective inputs can, at times, save an analyst from disaster. Our examination of asymptotic methods in Statistics leads to the conclusion that, under reasonably broad conditions, the two theories of estimation result in solutions that may be described as equivalent (albeit with respect to a frequentist measure of merit). Ease of application has been discussed, and while it is hardly a criterion one would want to place undue weight on when choosing an approach in any serious application, the issue does help us understand why frequentist methods might be the more popular options in certain kinds of applications. In modern computing environments, Bayesian analyses are now feasible in a wide range of models and problems, and the "ease of application" issue might well be considered a draw at this point in time.

Some Bayesians find comfort in versions of the Complete Class Theorem, a result that suggests that, in certain specific problem types, one never needs to consider statistical procedures other than Bayesian (or "almost" Bayesian) ones. We have argued, however, that this motivation for Bayesian methods is not really helpful in a given real-life statistical problem. The property under consideration is that *there exists* a Bayesian procedure that is as good as or better than anything else one might want to use. But, unfortunately, existence theorems are of little assistance in *finding*

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_5,

a good procedure. Consider the extension of the argument to the class of all decision rules in a particular problem — the whole class D^*, as we have referred to it in Chapter 1. Should we feel good about the selection of a particular decision rule just because the class it was drawn from, namely, D^*, contains all decision rules we would want to use? At this point in time, it seems fair to say that the Complete Class Theorem remains a result that is of theoretical rather than of practical interest.

Certain statistical problems are highly complex, dealing either with huge models or huge data sets or, quite often, both. I have argued that the frequentist school has the edge in multiparameter estimation problems, not because of the inherent qualities of their estimators, but because the approach often provides useable and defensible answers. It has also been noted, however, that frequentist solutions in this area are often developed under multivariate normality, and similar results are relatively sparse outside of these frameworks. Add to these arguments the well-known examples of incoherent, inadmissible and sometime plain silly "optimal" frequentist procedures and of potentially quite misleading (prior-dominated) Bayesian solutions and one is left with the uneasy feeling that the comparison between Bayesians and frequentists is unresolved and, perhaps, unresolvable.

There is, of course, one form of comparison that is yet to be discussed. None of the bases for comparison discussed above seek to determine which estimator tends to give better answers in particular problems of interest. Let us now consider this issue. We will do so in the context of point estimation, since the comparison of interest is most clearly defined there. Our approach has been described, somewhat abstractly, in Chapter 4. We will now turn our attention to a particular version of the threshold problem. We will demonstrate that in the well-known and widely encountered context of sampling distributions belonging to a one-parameter exponential family, one can indeed distinguish, quite explicitly, the circumstances in which the Bayesian approach to point estimation stands to outperform frequentist estimators from the circumstances in which the opposite is true. The approach taken here was first advocated by Samaniego and Reneau in a 1994 JASA paper. In this chapter, I'll present the main ideas of that paper, emphasizing its primary empirical and theoretical findings and highlighting the insights and interpretations that are especially relevant to statistical practice.

5.2 The word-length experiment

I'll begin with a discussion of a real experiment. Ninety-nine students in an elementary statistics class at the University of California, Davis, were asked to participate in an experiment involving an observed binomial variable with an unknown probability p of "success." The population from which data were to be drawn was the collection of "first words" on the 758 pages of a particular edition of Somerset Maugham's 1915 novel *Of Human Bondage*. Ten pages were to be sampled randomly, with replacement, and the number X of long words (i.e., words with six or more letters) was to be recorded. After a brief introduction to the Bayesian approach to estimation (with emphasis on the problem of estimating an unknown proportion), each student

was asked to provide a Bayes estimate of the unknown proportion p of long words. The elicitation of the students' beta priors was accomplished by obtaining each student's best guess p^* at p and the weight η he or she wished to place on the sample proportion $\widehat{p} = X/10$, with weight $(1-\eta)$ placed on the prior guess p^*. The prior specifications $\{(p^*, \eta)\}$ obtained from the students are displayed in Figure 5.1.

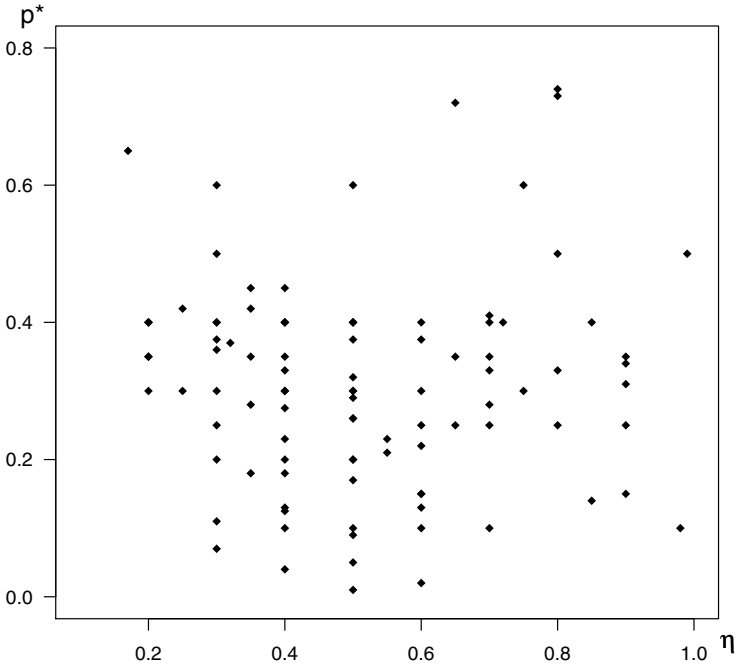

Fig. 5.1. Scatter plot of (p^*, η) values in the Word-Length Experiment

As Figure 5.1 suggests, the 99 students who participated in the word-length experiment seem to have rather diverse views about the word usage of early twentieth-century British authors, and they also appear to have quite variable confidence in their prior opinions on the matter. The scatter plot in Figure 5.1 looks a bit like a shotgun blast into the unit square. So it's natural to ask: how many of these nouveau Bayesians would tend to be closer to the true value of p than a statistician using the sample proportion $\widehat{p}$ as an estimator of p? It may surprise the reader to learn that about 90% of these Bayesians have an advantage over a frequentist who uses $\widehat{p}$ to estimate p. We will provide more detail about the outcomes of this word-length experiment in Section 5.3. In the following section, a theoretical development is pre-

sented which explains why the outcome mentioned above might have been expected in advance.

Exercise 5.1. Find a thumbtack. Toss it in the air 10 times and determine the proportion of times it landed on its flat side (i.e., with the point facing up). Call this proportion $\widehat{p}$. Assume that you guess the true probability p should be about 0.45, but that you are not too sure about your guess. Suppose that you use a Bayes estimator of p that places weight 0.2 on your guess and 0.8 on the observed $\widehat{p}$ (that is, you estimate p by $\widehat{p}_G = (0.2)(0.45) + (0.8)\widehat{p}$). If the true value of p happens to be 1/3 (as I was told it was, in confidence, by the manager of a thumbtack factory), which estimator, $\widehat{p}$ or $\widehat{p}_G$, turned out to be closer to the true value of p? (**Note:** When the whole class has done this experiment, we'll find out what fraction of the Bayes estimators outperformed $\widehat{p}$.)

5.3 A theoretical framework

We will focus our attention on the following statistical setting. Assume that a random sample $X_1, X_2, \ldots, X_n$ is drawn from a distribution F_θ which belongs to a one-parameter exponential family. We will be interested in estimating the scalar parameter θ using the squared error loss function $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$. We posit the existence of a statistic $\widehat{\theta}$ that is sufficient for θ and is an unbiased estimator of θ. In the applications of primary interest, $\widehat{\theta}$ can be thought of as the uniformly minimum variance unbiased estimator (UMVUE) of θ. We will refer to $\widehat{\theta}$ as "the best frequentist estimator," since in the contexts of primary interest, the standard alternative frequentist estimators (namely, maximum likelihood, method of moments, best linear unbiased and least squares estimators) will also be equal to $\widehat{\theta}$. Prototypical examples of such estimators include the sample proportion $\widehat{p}$ as an estimator of a binomial parameter p and the sample mean $\overline{X}$ as an estimator of the mean of a normal population.

The Bayes estimators to which the best frequentist estimator will be compared are assumed to be Bayes rules (under squared error loss) with respect to proper prior distributions on the parameter space. This restriction is a common one in "subjective" Bayesian analysis, where the prior model is a probability distribution representing the Bayesian's *a priori* assessment of his uncertainty about the value of the unknown parameter θ. "Coherent" Bayes theory requires that the measure placed on the parameter space be a probability measure. It is a known fact that an unbiased estimator cannot be a Bayes rule with respect to a proper prior and squared error loss (see Theorem 3.5). This identifies statistics such as $\widehat{p}$ and $\overline{X}$ above as estimators available to the frequentist but unavailable to the Bayesian.

Using the language introduced in Chapter 4, I will refer to the problem to which we now direct our attention as "the threshold problem," that is, the problem of identifying the threshold that separates Bayes estimators (or alternatively, their corresponding prior distributions) whose performance is superior to the best frequentist estimator from Bayes estimators (or priors) for which the reverse is true. We view the comparison to be made as a contest between a Bayesian statistician, using the

"operational prior" G, and a frequentist statistician who employs the best frequentist estimator. The true value of θ is of course unknown to both statisticians. As discussed in Chapter 4, we find it useful to model the unknown θ as a random variable, and we refer to its distribution G_0 as the "true prior distribution." The prior G_0 simply represents the existing physical (though unknown) truth. In most problems of interest, the parameter θ is simply an unknown constant and the distribution G_0 would in fact be degenerate at a point. But the generality employed here allows us to deal with problems in which the parameter θ actually varies in time or space and its value at the instant it is to be estimated can reasonably be thought of as a random realization from some unknown probability distribution G_0.

As the criterion for judging the performance of either estimator, we will take its average squared distance from the true value of θ, a criterion which translates into an estimator's Bayes risk $\mathbf{r}$ relative to the "true prior" G_0. The Bayesian statistician will use his own prior G, obtained by introspection or consultation with appropriate experts in the application of interest, while the frequentist statistician will use a frequentist estimator that is assumed here to be sufficient and unbiased. We will seek to compare $r(G_0, \widehat{\theta}_G)$, the Bayes risk of the Bayesian's estimator $\widehat{\theta}_G$, with respect to the true prior G_0, to the corresponding Bayes risk $r(G_0, \widehat{\theta})$ of the frequentist's estimator. The Bayes risk is a frequentist measure that represents the average or expected loss relative to all the randomness in the problem. A detailed defense of the criterion $r(G_0, \widehat{\theta})$ was given in Chapter 4. Having put forward the case for using the Bayes risk of an estimator, relative to the true prior distribution G_0, as an appropriate and relevant measure of its quality, I now present this chapter's main result.

Theorem 5.1. *Assume that a random sample is drawn from a distribution F_θ. Let $\widehat{\theta}_G$ be the Bayes estimator of θ under squared error loss, relative to the operational prior G. If $\widehat{\theta}_G$ has the form*

$$\widehat{\theta}_G = (1-\eta)E_G\theta + \eta\widehat{\theta}\,, \tag{5.1}$$

where $\widehat{\theta}$ is a sufficient and unbiased estimator of θ and $\eta \in [0,1)$, then for any fixed distribution G_0 for which the expectations exist,

$$\mathbf{r}(G_0, \widehat{\theta}_G) \le \mathbf{r}(G_0, \widehat{\theta}) \tag{5.2}$$

if and only if

$$V_{G_0}(\theta) + (E_G\theta - E_{G_0}\theta)^2 \le \frac{1+\eta}{1-\eta}\mathbf{r}(G_0, \widehat{\theta})\,. \tag{5.3}$$

Proof. For a fixed but arbitrary θ, the mean squared error of the Bayes estimator $\widehat{\theta}_G$ may be written as

$$E_{F_\theta}(\widehat{\theta}_G - \theta)^2 = E_{F_\theta}(\eta(\widehat{\theta} - \theta) + (1-\eta)(E_G\theta - \theta))^2\,. \tag{5.4}$$

Using the assumed unbiasedness of $\widehat{\theta}$, we may rewrite (5.4) as

$$E_{F_\theta}(\widehat{\theta}_G - \theta)^2 = \eta^2 E_{F_\theta}(\widehat{\theta} - \theta)^2 + (1-\eta)^2(E_G\theta - \theta))^2\,. \tag{5.5}$$

Taking the expectation of both sides of (5.5) with respect to the distribution G_0, we have

$$r(G_0, \widehat{\theta}_G) = \eta^2 r(G_0, \widehat{\theta}) + (1-\eta)^2 E_{G_0}(\theta - E_G\theta)^2 \,. \tag{5.6}$$

Viewing $E_{G_0}(\theta - E_G\theta)^2$ as the mean squared error of the variable θ as an estimator of $E_G\theta$, we may replace $E_{G_0}(\theta - E_G\theta)^2$ in (5.6) by

$$V_{G_0}(\theta) + (E_G\theta - E_{G_0}\theta)^2 \,.$$

It then follows that the inequality in (5.2) is equivalent to

$$(1-\eta)^2(V_{G_0}(\theta) + (E_G\theta - E_{G_0}\theta)^2) \le (1-\eta^2)\mathbf{r}(G_0, \widehat{\theta}) \,, \tag{5.7}$$

an inequality that may equivalently be written as the conclusion in (5.3), namely,

$$V_{G_0}(\theta) + (E_G\theta - E_{G_0}\theta)^2 \le \frac{1+\eta}{1-\eta}\mathbf{r}(G_0, \widehat{\theta}) \,. \qquad \blacksquare$$

Remark 5.1. The reader will note that the proof of Theorem 5.1 does not make explicit use of the fact that the estimator $\widehat{\theta}$ is a sufficient statistic for θ. The theorem would appear to apply more broadly, that is, the inequality in (5.3) will hold for any Bayes estimator of the form (5.1) and any unbiased estimator $\widehat{\theta}$ of θ. The sufficiency of $\widehat{\theta}$ enters Theorem 5.1 in a somewhat subtle way. The hypotheses of the theorem are vacuous without the assumption! This follows from the well-known fact that the Bayes estimator of θ is necessarily a function of the sufficient statistic T for θ (whether T consists of the entire sample or offers some measure of data reduction to a function of lower dimension) since the Bayes estimator depends on the data only through the likelihood function L in (3.16), and the likelihood may be written as a scalar multiple of a function of θ and T. As mentioned previously, we refer to the estimator $\widehat{\theta}$ of θ as the best frequentist estimator in the exponential family context, this being the estimator which results from virtually all standard frequentist analyses. Further, because of the Rao–Blackwell Theorem, the variance (or equivalently, the mean squared error) of any unbiased estimator that is not sufficient can be improved, a fact which implies that its Bayes risk with respect to any prior distribution can be improved. It thus becomes apparent that the comparison on which Theorem 5.1 is focused, that is, where $\widehat{\theta}$ is assumed to be sufficient, is the only comparison with potential applications and utility.

Theorem 5.1 has a considerable amount of interpretive value. One striking fact that leaps out from (5.3) is that the left-hand side (LHS) of (5.3) can be made equal to zero, while the right-hand side (RHS) of (5.3) is necessarily positive. When the LHS of (5.3) is zero, the Bayesian will win the contest with certainty. Two other insights that may be drawn from (5.3) are worth mentioning: (i) the variance on the LHS of (5.3) is the variance of the true prior, not the variance of the operational prior, a fact which suggests that in the typical estimation problem in which $V_{G_0}(\theta)$ may be considered to be zero, the Bayes estimator $\widehat{\theta}_G$ looks especially promising and (ii) since the weight η placed on $\widehat{\theta}$ lies in the interval $[0,1)$, the ratio $(1+\eta)/(1-\eta)$ on the

RHS of (5.3) takes values in the interval $[1,\infty)$, Thus, it can never be smaller than 1, and it can be made arbitrarily large by taking η sufficiently close to 1. Our interest in point (i) above derives from the fact that it goes beyond our natural intuition on when Bayes estimators should be good. Our intuition tells us that, in estimating an unknown constant θ, one would expect the Bayesian to outperform the frequentist whenever the mean of the operational prior is close to the true value of θ *and* the operational prior has a small variance. This intuition is articulated, for example, in the following statement by Diaconis and Freedman (1986): "*A statistician who has sharp prior knowledge of these parameters* (sic) *should use it, according to Bayes' Theorem.... On this point, there seems to be general agreement in the statistical community* (emphasis added)." But notice that the LHS of (5.3) makes no mention of the variance of the operational prior. That variance enters the picture only through the value of η on the RHS of (5.3). The statement quoted above is thus seen to be unduly conservative. The inequality in (5.3) suggests, instead, that the Bayesian will outperform the frequentist whenever the mean of the operational prior is "sufficiently close" to the true value of θ (or more accurately, is close to the mean of G_0) and the true prior has a small variance. That's interesting! In most problems of practical interest in which the statistical framework under study occurs, the true prior has variance zero. This brings into focus the role of an interesting feature of an operational prior which we formally define as follows:

Definition 5.1. *In the context discussed above with G and G_0 being, respectively, the "operational" and "true" priors of the random variable θ, the operational prior G is said to be* mean correct *if $E_G\theta = E_{G_0}\theta$.*

Corollary 5.1. *Under the hypotheses of Theorem 5.1, a Bayes estimator with respect to a mean-correct prior G has a smaller Bayes risk than the best frequentist estimator $\widehat{\theta}$ if and only if*

$$V_{G_0}(\theta) \leq \frac{1+\eta}{1-\eta}\mathbf{r}(G_0,\widehat{\theta})\,. \tag{5.8}$$

Further, if the true prior distribution G_0 is degenerate at a point, any Bayes estimator with respect to a mean-correct operational prior is superior to the best frequentist estimator.

It should thus be clear that Theorem 5.1 really does go well beyond the usual intuition about "sharp" (that is, accurate and precise) priors. Few in the statistical community would consider the uniform distribution on the interval $[0,1]$ to be a sharp prior on a population proportion p, and fewer still would consider a U-shaped beta prior like $\mathrm{Be}(0.1,0.1)$ to be appropriately described as "sharp," but in the event that they are mean correct (i.e., the true p is equal to 1/2), the Bayes estimators with respect to either of these priors will outperform the best frequentist estimator. Further, it should be noted that mean correctness isn't really necessary for Bayesian superiority. When the true prior is degenerate, there is clearly an interval containing the true θ (or, for nondegenerate G_0, containing $E_{G_0}\theta$) such that, when this interval contains the mean $E_G\theta$ of the operational prior, the Bayes estimator is necessarily superior. In fact, even when $E_G\theta$ lies well outside that interval, the Bayes estimator

will still be superior to the best frequentist estimator $\widehat{\theta}$ provided that the weight η that the Bayes estimator places on $\widehat{\theta}$ is not too small. Unless you've given this issue a good deal of thought in the past, you may well find this outcome surprising. Imagine that you are estimating a binomial proportion p, and that you use a Bayes estimator with a mean near 0 when in fact the true value of p is close to 1. You couldn't have been more wrong! One might expect that the price to be paid would diminish if the weight placed on your prior guess is suitably small. But why should a convex combination of $\widehat{\theta}$ and a terrible guess at θ ever be superior to the estimator $\widehat{\theta}$ alone? Theorem 5.1 tells us precisely when this will happen.

The inequality (5.3) suggests that the Bayes estimator will be superior to $\widehat{\theta}$ unless the Bayesian statistician miscalculates on two fronts simultaneously, that is, makes a particularly poor prior guess at θ and also puts considerable weight on that guess. Interestingly, neither of these negative characteristics alone will necessarily cause the Bayesian to lose his advantage. If, for example, the heights of a particular (say female) human population are (reasonably) modeled as normally distributed, the Bayes estimator relative to a normal prior distribution with mean 1000 ft. will actually outperform the frequentist estimator $\overline{X}$ if the weight placed on the prior mean is sufficiently small. Such phenomena can be understood by examining the damping effect of the weight $(1-\eta)$ placed on the prior mean, as seen in the following expression:

$$MSE(\widehat{\theta}_G(\eta)) = \eta^2 V(\widehat{\theta}) + (1-\eta)^2(\theta - E_G\theta)^2 \,. \tag{5.9}$$

Clearly, there exists a value η^* such that if $\eta > \eta^*$, then $MSE(\widehat{\theta}_G(\eta)) < V(\widehat{\theta})$, which in turn implies that $\mathbf{r}(G_0, \widehat{\theta}_G(\eta)) < \mathbf{r}(G_0, \widehat{\theta})$.

I mentioned earlier that the notion of "true prior distribution" is not in the Bayesian vernacular. It thus seems useful to state our main result without reference to the true prior G_0, that is, with the true value of the parameter θ considered simply as an unknown constant θ_0. In that case, Theorem 5.1 can be seen to imply

Corollary 5.2. *Under the hypotheses of Theorem 5.1, a Bayes estimator with respect to the prior G is closer, on average, to the true parameter value θ_0 than the unbiased and sufficient estimator $\widehat{\theta}$ if and only if*

$$(E_G\theta - \theta_0)^2 \leq \frac{1+\eta}{1-\eta} MSE_{\theta_0}(\widehat{\theta}) \,. \tag{5.10}$$

From this version of (5.3), which uses the mean squared error of an estimator as the measure of its merit, one may again conclude that the Bayesian who uses a mean-correct prior distribution cannot lose, and also that, no matter how poor a Bayesian's prior guess might be, the Bayesian's estimator would still be superior to the best frequentist estimator unless the weight η that he places on $\widehat{\theta}$ is too small. From either of the corollaries above, the Bayesian's winning strategy becomes quite clear: (1) careful attention to one's prior guess is worth the effort, since when that specification is done well, you can't lose, and (2) overstating one's confidence in a prior guess can lead to inferior performance, so conservative prior modeling is indicated; if one is to err in specifying the weight η one places on the frequentist estimator $\widehat{\theta}$, it's better to

err on the high side, thereby understating the confidence associated with one's prior guess.

The reader will note that Theorem 5.1 does not make the assumption that the family of distributions $\{F_\theta,\ \theta \in \Theta\}$ is an exponential family, that G belongs to the conjugate family of priors or that $\widehat{\theta}$ is the UMVUE of θ. The theorem actually applies more broadly. For example, it solves a version of the threshold problem for any Bayes estimator of the form (5.1) and any sufficient, unbiased estimator $\widehat{\theta}$ of θ. One situation it immediately applies to is the case in which F_θ and $\widehat{\theta}$ are unrestricted and the operational prior G is degenerate at a point. But the theorem is clearly intended to apply to sampling distributions from exponential families and their standard conjugate prior families. Indeed, this is precisely the type of problem in which Bayes estimators are necessarily linear and may be expressed in the form (5.1). While the theorem can shed some light on other problems, it is tailor made to treat the case of exponential families, and in that context, separates the class of standard conjugate prior distributions into priors which give the Bayesian the advantage and priors that don't.

In the context just mentioned, that is, in the case of sampling distributions belonging to exponential families together with the corresponding families of standard conjugate prior distributions, Bayes estimates of parameters of interest often take the special form in equation (5.1). The weight η placed on $\widehat{\theta}$ can typically be written as the fraction

$$\eta = \frac{n}{n+\omega}\ , \tag{5.11}$$

where n is the size of the sample drawn from F_θ and ω, generally called the prior sample size, represents the weight one attaches to one's prior guess (that is, the number of observations that one believes one's prior guess is worth, in contrast to the size n of the sample one is actually able to observe).

We will now turn our attention to an "asymptotic" interpretation of Theorem 5.1 as $n \to \infty$. This result provides insight into the outcome of the Bayes vs. frequentist contest as the data available to the two statisticians grows without bound.

Corollary 5.3. *Let $I(\theta)$ be the Fisher Information in a single observation X from the distribution F_θ. Suppose the hypotheses of Theorem 5.1 hold and that, in addition,*

(i) *for some fixed positive number ω and for any fixed n, $\eta = \dfrac{n}{n+\omega}$,*
(ii) *the model F_θ satisfies the Cramér–Rao regularity conditions (see Lehmann and Casella (1998)), and*
(iii) *the estimator $\widehat{\theta}$ is an efficient estimator of θ.*

Then it follows that the Bayes estimator $\widehat{\theta}_G$ is superior to the estimator $\widehat{\theta}$ as $n \to \infty$ if and only if

$$V_{G_0}(\theta) + (E_G\theta - E_{G_0}\theta)^2 \le \frac{2}{\omega} E_{G_0} I^{-1}(\theta)\ . \tag{5.12}$$

Proof. Note that when $\eta = \dfrac{n}{n+\omega}$, the fraction $\dfrac{1+\eta}{1-\eta}$ in (5.3) may be written as $\dfrac{\omega+2n}{\omega}$. Under conditions (ii) and (iii), the expression $\mathbf{r}(G_0, \widehat{\theta}) = \frac{1}{n} E_{G_0} I^{-1}(\theta)$ fol-

lows from the Cramér–Rao inequality. Substituting these two expressions in (5.3) yields

$$V_{G_0}(\theta) + (E_G\theta - E_{G_0}\theta)^2 \leq \frac{2+\omega/n}{\omega} E_{G_0} I^{-1}(\theta)\,. \tag{5.13}$$

Letting $n \to \infty$ in (5.13), we obtain (5.12). ∎

Corollary 5.3 has a number of interesting implications. Note, first, that the RHS of (5.13) is a decreasing function of n. This implies that if a Bayes estimator $\widehat{\theta}_G$ corresponding to the prior specification $(E_G\theta, \eta)$ satisfies the inequality in (5.12), it will satisfy the inequality in (5.13) for all values of n. Thus, a Bayes estimator that is asymptotically superior to the best frequentist estimator is superior to the latter estimator for any fixed sample size. In general, there exists an extended nonnegative integer n^* such that the Bayes estimator $\widehat{\theta}_G$ will be superior to the best frequentist estimator for all $n < n^*$. If B_n represents the collection of Bayes estimators (relative to prior specifications $(E_G\theta, \eta)$) that are superior to the best frequentist estimator when the sample size is n, then $B_1 \supseteq B_2 \supseteq \cdots B_n \supseteq \cdots \supseteq B_\infty$. We will call attention to this phenomenon when we return to our examination of the word-length experiment.

A second inference that may be drawn from Corollary 5.3 concerns the ratio of Bayes risks of the Bayes and frequentist estimators given by

$$\rho_n = \frac{r_n(G_0, \widehat{\theta}_G)}{r_n(G_0, \widehat{\theta})}\,. \tag{5.14}$$

Since $n \to \infty$ implies that $\eta \to 1$, the fact that $\rho_n \to 1$ as $n \to \infty$ follows from the identity in (5.6) (since $(1-\eta)^2 \to 0$ at the rate $(1/n)^2$). The new insight made available by Corollary 5.3 is that, for Bayes estimators $\widehat{\theta}_G \in B_\infty$, the ratio ρ_n approaches 1 from below! Such Bayes estimators are superior to $\widehat{\theta}$ for any and all fixed sample sizes. Thus, the size and shape of the collection B_∞ (or equivalently, the set of prior specifications $(E_G\theta, \eta)$ for which $\widehat{\theta}_G \in B_\infty$) will be of special interest.

A final consequence of Theorem 5.1, and perhaps the most important, is that in the context of exponential families, conjugate prior distributions and squared error loss, it provides an explicit solution to the threshold problem. The characterization of Bayesian superiority in this latter context is contained in the following:

Corollary 5.4. *Under the hypotheses of Theorem 5.1, the Bayes estimator $\widehat{\theta}_G$ and the frequentist estimator $\widehat{\theta}$ have the same Bayes risk with respect to the true prior G_0 for any operational prior G corresponding to the prior parameters (Δ, η) satisfying the hyperbolic equation*

$$\Delta\eta + \eta(\mathbf{r}(G_0, \widehat{\theta}) + V_{G_0}(\theta)) - \Delta + (\mathbf{r}(G_0, \widehat{\theta}) - V_{G_0}(\theta)) = 0\,, \tag{5.15}$$

where $\Delta = (E_G\theta - E_{G_0}\theta)^2$ and $\eta \in [0,1)$ is the constant specified in (5.1)*. Further,*

(i) *for given operational prior G with mean $\theta^* = E_G\theta$ and weight parameter η, there exists a constant $\eta^* \in [0,1)$ such that the Bayes estimator $\widehat{\theta}_G = (1-\eta)\theta^* + \eta\widehat{\theta}$ is superior to the best frequentist estimator $\widehat{\theta}$ if and only if $\eta > \eta^*$,*

(ii) *for any* $\eta \in [0,1)$*, there exists a constant* $\Delta^* \in [0,\infty)$ *such that the Bayes estimate* $(1-\eta)\theta^* + \eta\widehat{\theta}$ *is superior to the estimator* $\widehat{\theta}$ *if and only if* $\Delta = (\theta^* - E_{G_0}\theta)^2 < \Delta^*$ *and*

(iii) *for each fixed* $\eta \in [0,1)$*, the Bayes risk of the Bayes estimator* $\widehat{\theta}_G$ *with respect to* G_0 *is minimized by the choice of prior mean* θ^* *equal to* $E_{G_0}\theta$.

Proof. It is easy to verify that (5.15) is algebraically equivalent to

$$V_{G_0}(\theta) + (E_G\theta - E_{G_0}\theta)^2 = \frac{1+\eta}{1-\eta}\mathbf{r}(G_0,\widehat{\theta})\,.$$

This equation characterizes the circumstances, under the assumptions of Theorem 5.1, in which the Bayes risk, relative to G_0, of the best frequentist estimator $\widehat{\theta}$ is equal to that of the Bayes estimator with respect to the operational prior G. It therefore provides an explicit solution to the threshold problem in the context of Theorem 5.1. The additional claims above may be argued as follows.

(i) For $\eta \in [0,1)$, multiplying both sides of the inequality (5.3) by $(1-\eta)$ yields an inequality of the form

$$(1-\eta)A \le (1+\eta)B\,, \tag{5.16}$$

where A and B are nonnegative. Equality is achieved in (5.16) when

$$\eta(B+A) + B - A = 0\,. \tag{5.17}$$

If $C = (A-B)/(A+B)$, it is clear that (5.16) holds, with a strict inequality, if $\eta > C$. The claim in (i) follows.

(ii) Upon rewriting the inequality (5.3) as

$$\Delta \le \frac{1+\eta}{1-\eta}\mathbf{r}(G_0,\widehat{\theta}) - V_{G_0}(\theta)\,, \tag{5.18}$$

it is clear that that the Bayes estimator $\widehat{\theta}_G$ is superior to the best frequentist rule precisely when Δ is sufficiently small. When $V_{G_0}(\theta) = 0$, as it is in most problems of interest, the RHS of (5.18) is a positive number we may designate as Δ^*. If the RHS of (5.18) is negative, then claim (ii) holds with $\Delta^* = 0$.

(iii) This claim follows from the fact that, for any fixed η, the Bayes risk of $\widehat{\theta}_G$, written as a function of the mean $E_G\theta$ of the operational prior, is given by

$$r(G_0,\widehat{\theta}_G) = \eta^2 r(G_0,\widehat{\theta}) + (1-\eta)^2(V_{G_0}(\theta) + (E_G\theta - E_{G_0}\theta)^2)\,,$$

and clearly is minimized at the value $E_G\theta = E_{G_0}\theta$. ∎

A graph of the threshold identified in (5.15) is easily drawn for any particular fixed values of $V = V_{G_0}(\theta)$ and $r = \mathbf{r}(G_0,\widehat{\theta})$. For the case of common interest, that is, when $V = 0$, the threshold may be written as

$$\Delta \equiv \left(E_G\theta - E_{G_0}\theta\right)^2 = \frac{(1+\eta)r(G_0,\widehat{\theta})}{1-\eta}\,,$$

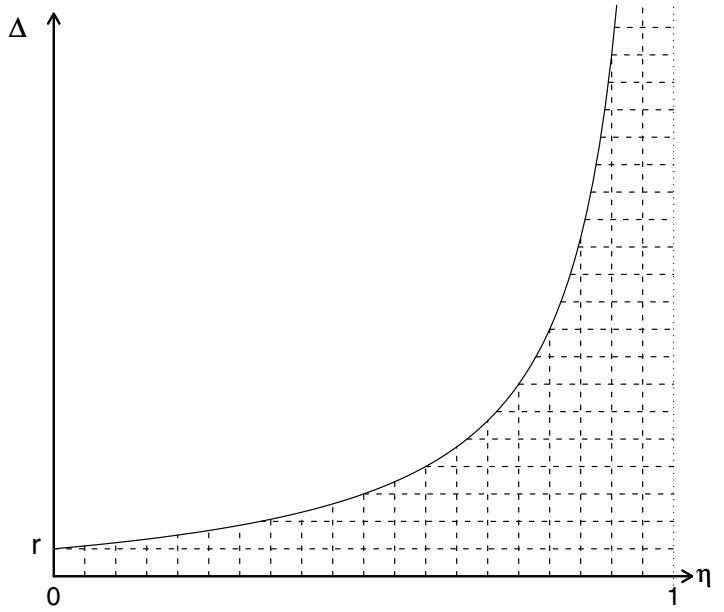

Fig. 5.2. Graph of the threshold (5.15) and of the region of Bayesian superiority when $V_{G_0}(\theta) = 0$

the graph of which is shown in Figure 5.2. Superior Bayes estimators correspond to (η, Δ) pairs below the threshold. More informative graphs, which identify superior Bayes specifications $(E_G\theta, \eta)$ directly, are displayed in the next section.

Exercise 5.2. (a) Let h be a concave function on the real line, and let X be a random variable for which $EX < \infty$. Prove that $h(EX) \geq Eh(X)$. (b) Assume that the hypotheses of Corollary 5.3 hold, and suppose that $I^{-1}(\theta)$ is concave for $\theta \in \Theta$. Prove that the percentage of Bayes estimators (that is, pairs $(E_G\theta, \omega)$) that are asymptotically superior to the "best frequentist estimator" $\widehat{\theta}$ is maximal when the true prior G_0 is degenerate at a point. (**Hint:** Apply part (a) to the RHS of the inequality in (5.12).)

5.4 Empirical results

In Section 5.1, I described the word-length experiment in which 99 students provided Bayes estimates, relative to their individually elicited beta priors, of the proportion of "long words" among the first words appearing on the 758 pages of a copy of Maugham's novel *Of Human Bondage*. We posed a question about the expected performance of these 99 Bayesians in comparison with that of a statistician using the

frequentist estimator $\widehat{p}$, with both estimators utilizing an available sample of 10 first words, drawn with replacement, from the novel's 758 pages. We mentioned that a strong majority of the Bayes estimators tend to outperform $\widehat{p}$. We now examine this experiment in light of the theoretical results above. The true proportion of long words in the population was determined to be $p = 228/758 = 0.3008$.

Table 5.1 records each student's prior specification (p^*, η), the prior sample size $\omega = (n/\eta) - n$ (with $n = 10$) of the corresponding beta prior and the ratio of the Bayes risk of each Bayes estimator to that of $\widehat{p}$. Bayes estimators are superior in 88 of 99 cases.

In light of the quite evident diversity of the 99 prior opinions about the target parameter p in the word-length experiment, the strong domination of the Bayes over the frequentist estimators apparent from Table 5.1 would, without any preceding discussion, likely be greeted by some measure of surprise. On the other hand, the theoretical developments in Section 5.2 give us substantial guidance regarding what should be expected here. In the word-length experiment, the true prior distribution G_0 is the degenerate distribution at the point 0.3008. From the inequality (5.3), it follows that the Bayes estimator based on the prior mean p^* (and irrespective of the prior weight parameter η) will be superior to $\widehat{p}$, relative to our Bayes risk criterion, whenever

$$(p^* - 0.3008)^2 \leq \mathbf{r}(G_0, \widehat{p}) = 0.02103\,. \tag{5.19}$$

Thus, a Bayesian has a fairly generous window of opportunity for outperforming the frequentist; the Bayes estimator will prevail if the Bayesian's prior mean is within 0.145 of the true value 0.3008 of p. Given this circumstance, the fact that 66 of the 99 Bayes estimators dominated the best frequentist estimator on the basis of a "good" prior guess might well be anticipated. For each of the Bayes estimators for which $|p^* - 0.3008| > 0.145$, the Bayes estimator will be superior to the best frequentist estimator provided that

$$(p^* - 0.3008)^2 < \frac{1+\eta}{1-\eta}(0.02103)\,. \tag{5.20}$$

For each value of p^* in this latter class, Bayesian superiority involves a direct relationship between the distance $|p^* - 0.3008|$ and the prior weight η that the Bayesian places on the sample proportion $\widehat{p}$. Of the 33 Bayes estimators for which the prior means p^* were far enough from the true mean 0.3008 so that the value of η actually plays a role in determining the direction of superiority, 22 of them chose a value of η that was large enough to satisfy the inequality (5.20). In the end, 88 out of 99, or 89%, of the Bayes estimators stand to outperform the sample proportion $\widehat{p}$.

Both the theoretical and empirical results above point to the following basic principle. There are two characteristics that serve to diminish the effectiveness of the Bayesian approach to estimation. The Bayesian is clearly penalized for being *misguided*, as the Bayes risk $r(G_0, \widehat{\theta}_G)$ is an increasing function of the distance between the prior mean and the true value of θ (or the value of $E_{G_0}\theta$). The Bayesian can also suffer from being *stubborn* about his prior opinion, a situation characterized by placing an unduly large weight on the prior mean. Fortunately for the Bayesian, neither of these characteristics need be fatal by itself. In the word-length experiment,

Table 5.1. Performance of Bayes vs. Frequentist Estimates in the Word-Length Experiment

ID	p^*	η	ω	r_G/r	ID	p^*	η	ω	r_G/r
1	0.350	0.35	18.571	0.171	42	0.350	0.70	4.286	0.500
2	0.150	0.90	1.111	0.821	43	0.250	0.65	5.385	0.438
3	0.150	0.60	6.667	0.533	44	0.500	0.99	0.101	0.980
4	0.375	0.50	10.000	0.315	45	0.300	0.25	30.000	0.063
5	0.260	0.50	10.000	0.270	46	0.110	0.30	23.333	0.938
6	0.200	0.50	10.000	0.371	47	0.720	0.65	5.385	1.446
7	0.400	0.20	40.000	0.340	48	0.600	0.30	23.333	2.176
8	0.350	0.20	40.000	0.114	49	0.450	0.35	18.571	0.570
9	0.030	0.20	40.000	2.271	50	0.100	0.60	6.667	0.667
10	0.030	0.40	15.000	1.415	51	0.250	0.70	4.286	0.501
11	0.040	0.40	15.000	1.324	52	0.180	0.40	15.000	0.410
12	0.250	0.30	23.333	0.150	53	0.280	0.70	4.286	0.492
13	0.400	0.40	15.000	0.328	54	0.150	0.60	6.667	0.533
14	0.400	0.30	23.333	0.319	55	0.500	0.80	2.500	0.715
15	0.400	0.20	40.000	0.340	56	0.290	0.50	10.000	0.251
16	0.450	0.40	15.000	0.541	57	0.130	0.60	6.667	0.582
17	0.350	0.65	5.385	0.437	58	0.200	0.30	23.333	0.327
18	0.600	0.50	10.000	1.314	59	0.360	0.30	23.333	0.172
19	0.020	0.60	6.667	0.960	60	0.420	0.35	18.571	0.408
20	0.330	0.70	4.286	0.494	61	0.250	0.80	2.500	0.645
21	0.050	0.50	10.000	0.998	62	0.280	0.35	18.571	0.131
22	0.300	0.30	23.333	0.090	63	0.340	0.90	1.111	0.811
23	0.400	0.50	10.000	0.367	64	0.310	0.90	1.111	0.810
24	0.250	0.90	1.111	0.811	65	0.130	0.40	15.000	0.659
25	0.400	0.85	1.765	0.733	66	0.300	0.50	10.000	0.250
26	0.100	0.40	15.000	0.850	67	0.400	0.60	6.667	0.435
27	0.100	0.50	10.000	0.729	68	0.375	0.60	6.667	0.402
28	0.320	0.50	10.000	0.254	69	0.730	0.80	2.500	0.990
29	0.410	0.70	4.286	0.541	70	0.100	0.70	4.286	0.663
30	0.220	0.60	6.667	0.410	71	0.330	0.80	2.500	0.642
31	0.090	0.50	10.000	0.778	72	0.300	0.50	10.000	0.250
32	0.350	0.20	40.000	0.114	73	0.210	0.55	8.182	0.382
33	0.740	0.80	2.500	1.007	74	0.010	0.50	10.000	1.255
34	0.650	0.17	48.824	4.023	75	0.300	0.40	15.000	0.160
35	0.250	0.60	6.667	0.380	76	0.125	0.40	15.000	0.689
36	0.275	0.40	15.000	0.171	77	0.180	0.35	18.571	0.416
37	0.500	0.30	23.333	1.015	78	0.400	0.20	40.000	0.340
38	0.100	0.98	0.204	0.961	79	0.230	0.55	8.182	0.351
39	0.300	0.75	3.333	0.563	80	0.400	0.40	15.000	0.328
40	0.140	0.85	1.765	0.750	81	0.375	0.30	23.333	0.218
41	0.420	0.25	30.000	0.443	82	0.400	0.50	10.000	0.367

Table 5.1. (continued)

ID	p^*	η	ω	r_G/r	ID	p^*	η	ω	r_G/r
83	0.400	0.40	15.000	0.328	92	0.350	0.90	1.111	0.811
84	0.300	0.60	6.667	0.360	93	0.200	0.50	10.000	0.371
85	0.600	0.75	3.333	0.829	94	0.070	0.30	23.333	1.331
86	0.400	0.70	4.286	0.532	95	0.400	0.50	10.000	0.367
87	0.400	0.30	23.333	0.319	96	0.330	0.40	15.000	0.175
88	0.350	0.40	15.000	0.201	97	0.260	0.50	10.000	0.270
89	0.400	0.72	3.899	0.555	98	0.200	0.40	15.000	0.334
90	0.230	0.40	15.000	0.246	99	0.170	0.50	10.000	0.453
91	0.370	0.32	21.250	0.208					

it is apparent that a misguided Bayesian such as student #85, whose prior guess of 0.6 was quite far off the mark, did not exhibit stubbornness about the guess, placing weight $\eta = 0.75$ on $\widehat{p}$, allowing him to outperform the frequentist. An example of a Bayesian who is stubborn but not misguided is student #8 whose prior guess of 0.35 was sufficiently close to the true value of p that his undoubtedly inflated confidence in that guess (placing weight 0.8 on it) caused him no difficulty. As mentioned earlier, this latter Bayesian would outperform the frequentist even if he had been supremely stubborn, placing weight 1 on his prior guess. In contrast to these students, a misguided and stubborn Bayesian such as student #34 has little chance of success in this experiment.

In Figure 5.3, the threshold between "superior" and "inferior" Bayes estimators in the word-length experiment is displayed. The curve shown in Figure 5.3 is the collection of priors $G(p^*, \eta)$, where $p^* = E_G\theta$, for which the Bayes estimator wrt G has the same Bayes risk with respect to the true prior G_0 as the sample proportion $\widehat{p}$, that is, for which

$$(p^* - 0.3008)^2 = \frac{1+\eta}{1-\eta}(0.02103)\,. \tag{5.21}$$

The graph shows quite vividly that the collection of Bayes estimators that are superior to the frequentist estimator $\widehat{p}$ constitutes a nonnegligible fraction of the unit square (being all points (p^*, η) above the threshold). It is apparent from the graph in Figure 5.3 that all Bayes estimators for which the prior mean $p^* \in (0.1558, 0.4458)$ are necessarily superior to $\widehat{p}$, regardless of the weight η the Bayesian places on $\widehat{p}$, and that all Bayes estimates which place weight $\eta > 0.9601$ are necessarily superior to $\widehat{p}$, regardless of the prior mean p^* selected by the Bayesian. The percentage of the unit square taken up by prior specifications (p^*, η) corresponding to Bayes estimates that are superior to $\widehat{p}$ is 0.55. The proper interpretation of this percentage is that, if a Bayesian were to pick a prior specification (p^*, η) completely at random according to a uniform distribution on the unit square (something akin to the aforementioned shotgun blast), the Bayes estimator would outperform $\widehat{p}$, the frequentist estimator, 55% of the time. For Bayesians who have "useful" prior information available about the parameter p, one would expect even better performance, on the average. The fact

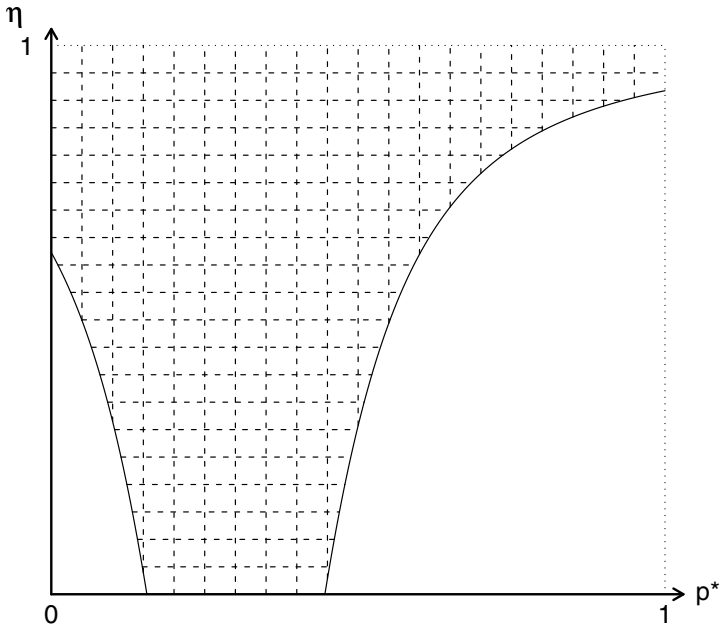

Fig. 5.3. Graph of the threshold (5.21), and the region of Bayesian superiority, in the Word-Length Experiment

that 89% of the students participating in the word-length experiment specified Bayes estimators that outperformed $\widehat{p}$ suggests that most of them did have such information, even though most of their prior distributions could hardly be referred to as "sharp."

The flip side of the story above must also be mentioned, as it is an important part of the overall lesson learned. We noted that, both in theory and in practice, the frequentist estimator $\widehat{p}$ sometimes beats the Bayes estimator $\widehat{p}_G$. In the particular experiment of interest, and in similar settings in which Theorem 5.1 applies, neither estimator will be uniformly superior. This suggests that the statistician should remain open to using one approach or the other, depending on which promises to provide better performance. In the concluding section of this chapter, we will elaborate on this and related issues.

It is natural to wonder what the effect is of the chosen sample size n in the comparison of Bayes and frequentist estimators. In the word-length experiment in which the size of the available sample was stipulated to be 10, one might conjecture that even weak prior information might in fact be useful, as a sample of size 10 is not itself very informative about p. To examine this question, we have sought to determine what percentage of the stipulated Bayes estimators would be superior to $\widehat{p}$ as the sample size n varies. In this extended comparison, we have retained the same beta prior distribution for each student, as in stating his or her prior, each student implicitly

specified both their best guess at p and the prior sample size they deemed appropriate for that guess. In Table 5.2, we display the percentage of Bayes estimators that would be superior to the frequentist estimator $\widehat{p}$ for the word-length experiment with different sample sizes. Note that the percentage of superior Bayes estimators is at least 80%, irrespective of the sample size n.

Table 5.2. The percentage of Bayes estimators (SBEs) that are superior to the sample proportion in the word-length experiment, as a function of n

Sample size n	Number of SBEs	Percentage of SBEs
1	99	100%
5	93	94%
10	88	89%
50	83	84%
100	81	82%
500	80	81%
1,000	80	81%
5,000	79	80%
10,000	79	80%
∞	79	80%

Table 5.2 demonstrates that the size of the experiment has a relatively small impact of the comparisons we have made. We noted above that size of the collection B_n of Bayes estimators (relative to prior specifications $(E_G\theta, \eta)$) that are superior to the best frequentist estimator when the sample size is n decreases as n grows. It is apparent from the above that this shrinkage occurs rather slowly, and that the size of the limiting set B_∞ is still formidable. To get a better idea of what B_∞ looks like, we provide in Figure 5.4 a graph of the threshold separating SBEs (superior Bayes estimators) from IBEs (inferior Bayes estimators) for the domain containing all the Bayes estimators employed by the participants in the word-length experiment. The reader will note that every prior sample size specification ω utilized by a student in the word-length experiment was in the range $(0, 50)$, an outcome that might be expected when thinking about the worth of one's prior guess when combined with a sample proportion based on 10 observations. The threshold for the case $n = \infty$, as a function of the prior specification (p^*, ω), is given by

$$(p^* - 0.3008)^2 = (2/\omega)(0.02103)\,. \tag{5.22}$$

Figure 5.4 admits to an interpretation similar to that of Figure 5.1. The percentage of space $[0,1] \times [0,50]$ taken up by prior specifications (p^*, η) corresponding to Bayes estimates that are superior to $\widehat{p}$ is 33%. If a Bayesian were to pick a prior specification (p^*, η) completely at random according to a uniform distribution on that space, the Bayes estimator would outperform the frequentist estimator 33% of the time. For Bayesians who have "useful" prior information available about the parameter

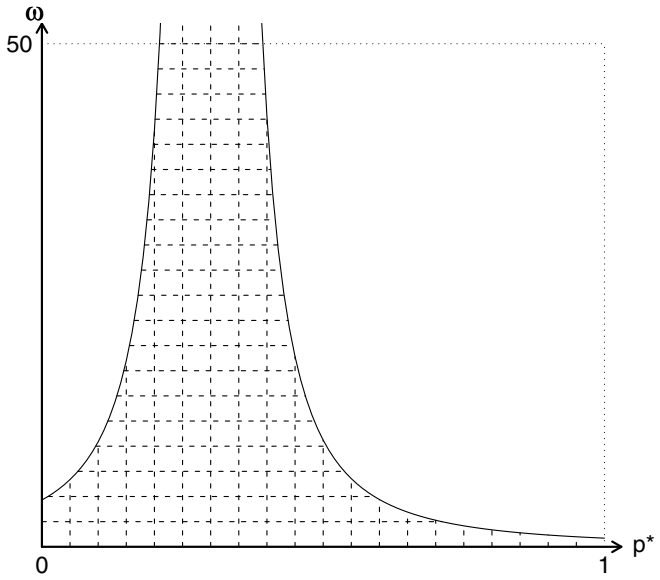

Fig. 5.4. Graph of the threshold in (5.22), where $n = \infty$ and $(p^*, \omega) \in [0,1] \times [0,50]$, and the region of Bayesian superiority

p, one would expect better performance, on the average. In the word-length experiment, 80% of the students who participated specified Bayes estimators that would outperform $\widehat{p}$, even for an arbitrarily large sample size n, again suggesting that most of them did have information that is reasonably described as "useful." We note that only one student specified a prior sample size larger than 40, and only 7 specified a prior sample size larger than 30, in the word-length experiment. Table 5.3 provides additional relevant information about Bayesian superiority for large n.

Table 5.3. The percentage of "random" Bayes estimators (SRBEs) in the word-length experiment which are superior to $\widehat{p}$ at $n = \infty$ for various upper bounds on the prior sample size ω

Upper Bound on ω	Percentage of SRBEs
50	0.327
40	0.360
30	0.407
20	0.480
10	0.621

Exercise 5.3. To assess the extent of the domination of Bayes estimators over the sample proportion $\widehat{p}$ in the word-length experiment, it is useful to examine those cases in which the domination is substantial, in one direction or the other. Identify the students for which $|\ln(r_G/r)| > 1$ (that is, the cases in which the Bayes risk of one estimator is roughly three times (or more) better that the other. Among these cases, what percentage of the Bayes estimators were superior to $\widehat{p}$?

Exercise 5.4. Suppose you are asked to estimate a binomial parameter p based on $X \sim \mathscr{B}(20, p)$. Your performance will be judged by someone who, unlike you, actually knows that the true value of p is 0.6. The criterion for judging an estimator is the Bayes risk, under squared error loss, of the estimator wrt the true prior G_0 (which is degenerate at 0.6). You are toying with 5 quite different stances regarding an operational prior. (You probably should have consulted an expert or two, but it's too late for that now.)

(a) Calculate the Bayes risks (wrt G_0) of the Bayes estimator wrt the 5 operational priors: i) Be(0.5,0.5), ii) Be(10,10), iii) Be(0.6,0.4), iv) Be(100,100) and v) Be(400,600).
(b) Calculate the Bayes risk (wrt G_0) of the sample proportion $\widehat{p}$.
(c) Rank these 6 estimators from best to worst relative to the criterion above.

5.5 Potpourri

A variety of other issues and perspectives were considered by Samaniego and Reneau (1994). Here, I will mention just two. The first has to do with whether or not the frequentist actually has any wiggle room in the problems examined above. When I presented this work in a seminar at Berkeley in 1992, Lucien LeCam raised the following interesting issue. He conjectured that the level of Bayesian success in the word-length experiment might actually be due to the fact that the frequentist was using the wrong estimator. Perhaps the frequentist would have fared better if he had used the minimax estimator $\widehat{p}_{mm}$ given by

$$\widehat{p}_{mm} = \frac{n\widehat{p} + \sqrt{n}/2}{n + \sqrt{n}} \tag{5.23}$$

instead of the estimator $\widehat{p}$. I responded as follows. Trustworthy frequentist alternatives to the sample proportion $\widehat{p}$ as an estimator of p are truly hard to come by. The estimator in (5.23) isn't actually a candidate, since it is clearly a convex combination of $\widehat{p}$ and the constant 1/2 and is, in fact, the Bayes estimator relative to the beta prior Be$(\sqrt{n}/2, \sqrt{n}/2)$. Professor LeCam was correct in guessing that it outperforms $\widehat{p}$ in the word-length experiment. But this just means that it joins the ranks of the "Superior Bayes Estimators." One could take the view that the minimax estimator in (5.23) is actually a frequentist rule, but it's very composition suggests that it represents Bayesian thinking, as one is clearly hedging one's bets between $\widehat{p}$ and the prior intuition that $p = 1/2$ is a reasonable guess at the true value of p. In our formulation

of the threshold problem, I have taken the position that "you are what you eat," that is, in any given problem, you are classified as a Bayesian or a frequentist based on the estimator you choose to use — quite irrespective of your statistical philosophy or outlook. Since the estimator $\widehat{p}_{mm}$ is a Bayes estimator of p, a statistician who uses it in estimating an unknown proportion p is being a Bayesian on that particular occasion. If, in problems such as the word-length experiment, there's a frequentist estimator that is preferable to $\widehat{p}$, it has not as yet surfaced in the open literature!

A second issue worth mentioning is the extent to which Theorem 5.1 generalizes to frameworks other than those studied above. One framework that is quite distant from the exponential family setting we have emphasized is the fully nonparametric problem of estimating the distribution F of a random sample, where F is completely unrestricted. Ferguson (1973) developed a methodology for Bayesian estimation of F based on what he called the Dirichlet Process. I will use Ferguson's notation in presenting the nonparametric analog of Theorem 5.1. Let F and $\widehat{F}$ be two cumulative distribution functions (cdfs) on the real line. We will consider the problem of estimating F using the integrated squared error loss function given by

$$L(F,\widehat{F}) = \int_{-\infty}^{\infty} (\widehat{F}(x) - F(x))^2 \, \mathrm{d}W(x) \,, \tag{5.24}$$

where W is a fixed, finite measure on (R,B), with B being the σ-field of Borel sets on R. If G is a "cdf process," that is, a stochastic process whose realizations are cdfs on R with probability 1, then the Bayes risk of the estimator $\widehat{F}$ of F, wrt the "prior" G, is given by

$$\mathbf{r}(G,\widehat{F}) = E_G E_F L(F,\widehat{F}) \,. \tag{5.25}$$

In analogy with the preceding results and discussion, we will denote the "true prior distribution" of F by G_0. The following result may be established in this context.

Theorem 5.2. *For any non-null, finite measure π on (R,B), let $D(\pi)$ represent the Dirichlet process with parameter π. Let F be a realization of the cdf process G_0, and suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$. If $\widehat{F}_\pi$ is the Bayes estimator of F with respect to the prior distribution $D(\pi)$ and $\widehat{F}_n$ is the empirical cdf, then for some $\eta \in (0,1)$,*

$$\widehat{F}_\pi(x) = \eta \widehat{F}_n(x) + (1-\eta) E_\pi F(x) \quad \textit{for all } x \in R \,, \tag{5.26}$$

and when the expectations involved exist,

$$\mathbf{r}(G_0,\widehat{F}_\pi) \le \mathbf{r}(G_0,\widehat{F}_n) \tag{5.27}$$

if and only if

$$\int_{-\infty}^{\infty} V_{G_0}(F(x)) \, \mathrm{d}W(x) + \int_{-\infty}^{\infty} (E_\pi F(x) - E_{G_0} F(x))^2 \, \mathrm{d}W(x) \le \frac{1+\eta}{1-\eta} \mathbf{r}(G_0,\widehat{F}_n). \tag{5.28}$$

The representation in (5.26) was, of course, established by Ferguson (1973). Theorem 5.2 demonstrates that the threshold problem for nonparametric estimation via

Dirichlet process priors gives rise to solutions similar to those obtained in the parametric contexts we have studied. For example, when the true prior is degenerate at a particular distribution F_0, a mean-correct Bayes estimator is guaranteed to be superior to F_π. Further, the threshold separating good and bad prior distributions may be identified explicitly through (5.28).

Exercise 5.5. Show that, relative to squared error loss, the risk function of the estimator $\widehat{p}_{mm}$ in (5.23) is a constant independent of p. Verify that $\widehat{p}_{mm}$ is the Bayes estimator wrt the prior

$$\text{Be}(\sqrt{n}/2, \sqrt{n}/2) \, .$$

Conclude that $\widehat{p}_{mm}$ is minimax as an estimator of p. (**Hint:** First, show that if an equalizer rule is Bayes wrt a proper prior distribution, it is minimax.)

Exercise 5.6. Prove Theorem 5.2.

5.6 Discussion

I will be brief, as the results above more or less speak for themselves. The paper in which the ideas above were first laid out was entitled "Towards a Reconciliation of the Bayesian and Frequentist Approaches to Point Estimation." The word "reconciliation" admits to a variety of interpretations. One of the more common connotations of the word is that a form of agreement has been found such that two parties can both adopt the same action or viewpoint without sacrificing their respective principles. One might think of "objective Bayesian analysis" as constituting this type of reconciliation, since the approach uses Bayesian techniques (with improper priors) yet is, in essence, a frequentist methodology (yielding estimators like $\widehat{p}$ and $\overline{X}$ in problems of the sort we have been discussing). In the paper to which I have alluded, we used the word "reconciliation" in quite a different sense.

Many statisticians view the Bayesian and frequentist approaches to statistical inference to be mutually exclusive. The two schools are based on different views regarding the goals and legitimate methods of statistical analysis. Thus, the orthodox Bayesian and the committed frequentist will not agree, in principle or in practice, regarding the application of Statistics to a particular problem. The reconciliation proposed here can be summarized as follows: both frequentists and Bayesians should acknowledge that there are circumstances, in any estimation problem of the type considered here, in which the "other approach" will give better results. If our role as statisticians is to provide reliable conclusions from data, then we cannot set aside either approach in treating the data we have available. Instead, we should ask the question: is this a problem which is better treated by a Bayesian or by a frequentist approach?

This latter attitude naturally points to the threshold problem. The threshold separating superior from inferior Bayes procedures is of course elusive. It depends on features of the problem that are unknown to the statistician. It is nonetheless true that certain general principles can be helpful. From the results above, it is reasonable to

conclude that prior information is useful in the estimation process under conditions that are really quite broad: a carefully chosen prior guess at an unknown parameter, together with a conservative assessment of one's confidence in that guess, appears to be a reliable recipe for good performance. This argues for careful introspection (or consultation) about the value of an unknown parameter and a moderately diffuse prior with that value as its mean. This prescription stops short of actually using a diffuse (improper) prior; we have seen that this leaps all the way to a frequentist estimator that might in fact be inferior. But the leap to a frequentist estimator may indeed be the rational choice when the available prior information is too vague or ill-defined for one to be able to implement with confidence the recipe above for obtaining a "superior" Bayes estimator.

The empirical and theoretical results of this chapter support the conclusion that the Bayesian approach to estimation is surprisingly resilient, providing superior results even in cases in which the operational prior distribution used might, on the basis of some sort of impartial analysis, be considered to be quite weak. To be perfectly honest, this finding caught me by surprise. In the study culminating in the 1994 JASA paper, I set out thinking that we would find a theoretical framework which more or less substantiated the statement I've quoted from Diaconis and Freedman (1986). Instead, our findings indicated that Bayes procedures work well a lot more often than we (and most other people) suspected. The reader might wonder if that study in fact turned me into a Bayesian. Most of my colleagues think of me as one, but the outcomes above, together with studies such as those considered in the next three chapters, made it clear to me that a single statistical paradigm was not sufficient to handle all statistical situations well. What these studies did, instead, was turn me into a "Bayesian sympathizer." My current state of mind is that Bayesian methods are intellectually satisfying and, often, very effective. Thus, I'm very open to using them, while recognizing that frequentist alternatives can produce better results in selected circumstances.

6

Conjugacy, Self-Consistency and Bayesian Consensus

6.1 Another look at conjugacy

The notion of "conjugacy" has long been a staple of Bayesian inference, though the use of conjugate priors is much less prominent in current forms of Bayesian analysis than it was, say, twenty years ago. There are a number of reasons for the devaluation of conjugacy over the last two decades. The most obvious ones concern those associated with the major advances in Bayesian computation in the decade of the 1990s which have made possible the approximation of posterior distributions and related quantities for a broad array of prior models, rendering the use of priors which facilitate straightforward, closed-form posterior analysis less necessary, if still convenient. With the development and refinement of the Gibbs sampler and similar Markov chain Monte Carlo algorithms, one can now execute a well-approximated Bayesian analysis based on virtually any prior model.

Another reason for the decline in the use of conjugate priors is the growth in popularity of the "objective" Bayesian approach. Conjugate priors are tools that facilitate subjective Bayesian analysis, where the prior is taken to be a reflection of prior knowledge or expert opinion regarding an unknown quantity (generally a parameter of the model governing the observable quantities in a given problem). Since objective or noninformative priors are "improper" in most standard applications, they lie beyond the realm of conjugacy (except perhaps as limits) and, indeed, lie outside the domain of subjective Bayesian inference. Instead of seeking to capture what one knows or believes about an unknown parameter in a given application, an objective Bayesian analysis attempts to minimize or eliminate the influence of "prior information" while retaining the formalism of the use of "Bayesian updating" when calculating the posterior distribution of the parameter(s) of interest given the available experimental outcomes.

These two developments have led to a reduced reliance on conjugate priors in executing Bayesian analyses. Indeed, some Bayesian practitioners consider conjugate priors as anachronistic "crutches" which may now (largely) be set aside. In this monograph, I have taken a more generous view of conjugate priors. In the present chapter, I will add to the motivation for doing so by a further exploration of their

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_6,

properties as well as by noting a special role they are able to play in the traditionally thorny "consensus problem" — the problem of utilizing input from several experts with varying prior opinions in the formulation of a Bayesian procedure. However, criticizing "modern" Bayesian tools and approaches is far from my purpose here. I will thus mention only briefly some issues that might serve as food for thought before or while utilizing them.

With regard to recent developments in Bayesian computation, I should acknowledge that it seems foolish to see them as anything but a blessing, as they open up the possibility of carrying out a Bayesian analysis with great flexibility in one's prior modeling and without dealing with the often imposing analytical difficulties involved in getting exact solutions. But are these tools represented correctly as the missing link that finally rendered Bayesian inference complete? This is where the introduction of some caveats seems appropriate. The fact that certain prior specifications have led to improper and therefore uninterpretable posteriors via MCMC has been pointed out by Hobert and Casella (1996). A more fundamental source of potential difficulty, I think, is the implicit suggestion in much modern Bayesian work that the computational feasibility of an analysis somehow justifies the procedure followed and the answers obtained therefrom. Many practitioners seem quite satisfied when diagnostic measures appear to confirm the convergence of an MCMC process, in spite of the fact that such convergence says nothing whatsoever about the quality of the corresponding statistical inference (in the sense of the threshold problem as discussed in Chapters 4 and 5). However, using a highly complex prior, just because one now can, may in fact make little sense in a given application, and may also lead to inferior statistical performance (for the same reasons that overfitting can have negative effects in regression analysis). The principle of parsimony suggests that a simpler prior, perhaps a conjugate one, may actually yield better results.

In fairness, it should be noted that many users of MCMC methods have not simply abandoned the use of conjugate priors in favor of increased flexibility. It has been observed, for example, that in a variety of applications, both the mixing properties and the rate of convergence of the Gibbs sampler are positively affected when conjugate priors are employed. In this general context, the recent study by Diaconis, Khari and Saloff-Coste (2008) merits special mention. The aim of Diaconis *et al.* was to identify modeling scenarios in which sharp rates of convergence of the Gibbs sampler can be established. Their main results demonstrate that this aim is most definitely realized in the context of standard exponential families with their conjugate prior families. Since my brief comments on the paper will only scratch its surface, I should add that this comprehensive, insightful and authoritative work should be carefully read by anyone interested in understanding how, why and when the Gibbs sampler works.

While Diaconis *et al.* do treat some general models, special emphasis is given to the six exponential families with quadratic variance structure. They also restrict their attention to bivariate problems based on a joint density $f(x, \theta)$, but their results are highly suggestive and provide helpful guidance regarding what might be expected in problems of higher dimension. In the problems they consider, they obtain striking results such as the following: in the binomial–beta model based on n Bernoulli tri-

als, the number of steps required for convergence of the Gibbs sampler is of order n (typically some small integer multiple of n), while in the Poisson-gamma model, the number of steps required for convergence of the Gibbs sampler is of order $\log n + c$. The primary tool used in obtaining such results is the singular value decomposition of an operator T associated with the "x-chain." It is shown that T takes orthogonal polynomials for the mean function $m(x)$, leading to exact computations which ultimately reveal the rate of convergence of the chain. Two important lessons that may be gleaned from this work are that the analytical treatment of MCMC properties is actually feasible for exponential families with conjugate prior distributions and that, for these models, the convergence of the Gibbs sampler tends to be (surprisingly) rapid. It is true, of course, that restricting to traditional conjugate priors may sometimes be overly constraining and that some applications demand more complex prior modeling.

The use of "objective" or "noninformative" priors is by no means a new phenomenon in statistical analysis. The concept can in fact be traced to Bayes and Laplace themselves in the earliest writings on the approach we now call Bayesian. But the early attempts to use priors that assume little or nothing about the unknown parameter used proper probability distributions to represent prior "ignorance." Later developments used general measures as priors in spite of the fact that they had no probabilistic interpretation. One of the motivations of the latter approach was the obvious dilemma which arises when using proper priors and attempting to preserve prior ignorance under transformations of the parameter of interest. Clearly, if one is ignorant about the parameter θ, then one must also be ignorant about θ^2, yet when one specifies that θ has a uniform distribution, this implies that θ^2 does not. This unsatisfactory situation was largely resolved by Jeffreys (1961) who, in short, proposed and defended the principle that there was a unique form of the unknown parameter — essentially the one that most closely resembles a location parameter — that should be modeled by a uniform or "flat" prior. Jeffreys' rule is still widely used today, its broad utility having been ably demonstrated by Box and Tiao (1973), among others. Bernardo (1979) advocated a different but related approach to noninformative prior modeling. His "reference priors" are now also in widespread use. See Berger, Bernardo and Sun (2009) for a rigorous and comprehensive treatment of the latter topic.

The inclination to use a noninformative (albeit improper) prior is perhaps quite understandable. The usual defense is simply that subjective Bayesian inference involves the interjection of prior beliefs or intuition into a statistical analysis. The fact that doing so may give one a good answer — possibly a better answer than other approaches one might take — often fails to relieve the anxiety one might have about introducing personal biases or misconceptions into the analysis. The old maxim "let the data speak for itself" seems to argue against introducing subjective elements into the statistical treatment of experimental data. This is true with special force in the context of scientific inference, a context in which the prior opinions of the analyst are traditionally excluded. The elimination of all sources of potential bias is a natural first step in developing an analysis that is deemed to be credible and worthy of consideration. This seems to be prominent among the reasons that leaders in the field

such as Berger (2006) and Efron (2005) argue that objective rather than subjective Bayesian analysis is the appropriate mode of Bayesian inference in the statistical analysis of scientific data.

Whatever the merits of an objective Bayesian analysis might be, one should recognize that the approach is patently non-Bayesian. The usual axiom system (see, e.g., Section 3.2) that drives Bayesian inference leads to the necessary conclusion that the unique way to quantify uncertainty is through the assignment of probabilities to uncertain events. The use of improper priors violates these axioms and is therefore among the variety of procedures a statistician might use that "orthodox" Bayesians would describe as *incoherent*. One might not be opposed, in principle, to the use of "objective Bayesian analysis" (after all, a variety of widely respected estimators result from such an analysis), but the classification of the procedure as Bayesian is, in the view of orthodox Bayesians, inappropriate. It is, instead, a frequentist procedure properly belonging to the classical school of statistics. While it is true that something resembling Bayes' Theorem is used in such analyses to update the improper prior, this process is a mathematical formality rather than a probabilistic calculation, since Bayes' Theorem, as a result lodged within the calculus of probability, only applies to prior and posterior probabilities. Since an "objective prior" brings essentially nothing to the table, save the possibility of some analytical or computational feasibility, it might be thought of simply as another data analytic tool, in the same class, say, as bootstrapping and uses of the EM algorithm — both important and effective procedures that are, similarly, non-Bayesian. In the grand scheme of things, it cannot be deemed unreasonable to take the view that "objective Bayesian procedures" are legitimate potential analyses that one might (perhaps even should) consider using. Their merits should be judged, however, by their real or expected performance in particular applications as compared to other procedures one might carry out, including real (proper) Bayesian analyses.

It seems worthwhile spending a moment reflecting upon what one sacrifices in restricting attention to "objective" Bayes methods. An important practical concern one might have about "objective Bayesian procedures" is simply the fact that, by restricting oneself to this approach, one is ignoring the potentially substantial advantages available in exploiting what one knows (or what one or more experts know) about a particular experimental situation. In the simple yet compelling example of estimating the bias of a freshly minted coin, it is quite difficult to justify ignoring the fact that we all know a priori that the probability of heads is most certainly quite close to 1/2. The classical (and the objective Bayesian) estimate of the probability of heads would be quite unacceptable if one happened to observe the unlikely but still possible string of ten consecutive heads in ten tosses of that coin.

Further, one can legitimately ask what one is really doing when one postulates an improper prior for an unknown parameter. Does doing so have a natural interpretation? It might be argued that the desired and actual interpretation of improper priors is "prior ignorance," that is, the state in which no given value of the unknown parameter receives more "weight" than any other. But does an "objective" prior achieve that goal? When Lebesgue measure is used as a prior measure for a location parameter, one places infinitely more weight on the complement of the interval

$(-3,000,000,000, +3,000,000,000)$ than one places on the interval itself — clearly a peculiar quantification of "prior ignorance."

The foregoing discussion is not meant to be definitive with regard to the pros and cons of modern implementations of Bayesian analysis. Instead, it is meant to motivate the thesis that subjective Bayesian analysis and, in particular, the use of conjugate priors in subjective Bayesian inference, deserves a reexamination. In the section that follows, I will argue that the apparently unexplored property of "Bayesian self-consistency" is a characteristic that one might reasonably take as a requirement of a Bayesian estimation procedure. Doing so certainly elevates the status of traditional families of conjugate priors, as they are a broad and useful class of priors possessing this property. Our study of possible "solutions" to the consensus problem provides further evidence of the utility of conjugate prior models. In the latter problem, I will examine how the estimators I will propose for use in the consensus problem compare to frequentist estimators in the familiar context of sampling distributions belonging to exponential families.

Exercise 6.1. Jeffreys (1961) proposed the use of a prior distribution whose "density" g is given by $g(\theta) \propto [I(\theta)]^{1/2}$, where $I(\theta)$ is the Fisher information corresponding to a single observation X from F_θ. Jeffreys' prior is typically improper; when it isn't, g is standardized so as to integrate to one. Show that Jeffreys' prior is invariant under smooth 1-1 transformations on θ, that is, show that if $\lambda = h(\theta)$, where h is 1-1 and differentiable, then

$$[I(\lambda)]^{1/2} = [I(h^{-1}(\lambda))]^{1/2} \left| \frac{\mathrm{d}\theta}{\mathrm{d}\lambda} \right| .$$

Exercise 6.2. Show that the Jeffreys' prior on a location parameter θ is constant.

Exercise 6.3. If $X|\theta \sim \mathscr{N}(\theta, \sigma^2)$, show that Jeffreys' prior for θ is $g(\theta) \propto 1$ for $-\infty < \theta < \infty$ and that the Jeffreys' prior for σ is $g(\sigma) \propto 1/\sigma$ for $\sigma > 0$.

Exercise 6.4. Let $X|\theta \sim \mathscr{B}(n, \theta)$. Show that Jeffreys' prior for θ is $\mathrm{Be}(1/2, 1/2)$. (**Note:** In problems in which the posterior distribution of θ is asymptotically normal, Bernardo's "reference prior" reduces to the Jeffreys' prior.)

Exercise 6.5. Consider a simple one-way random effects model with

$$Y_{ij} = \beta + u_i + \varepsilon_{ij}, \quad i = 1, \ldots, k \quad \text{and} \quad j = 1, \ldots, J ,$$

where $u_1, \ldots, u_k \overset{iid}{\sim} \mathscr{N}(0, \sigma^2)$ and $\{\varepsilon_{ij}\} \overset{iid}{\sim} \mathscr{N}(0, \sigma_\varepsilon^2)$, and the us and εs are assumed to be independent. A Bayesian treatment of this model would place prior distributions on the model parameters β, σ^2 and σ_ε^2. Denote the associated priors' "densities" as $g(\beta)$, $g(\sigma^2)$ and $g(\sigma_\varepsilon^2)$, respectively. The posterior "density" may then be represented as

$$g(\sigma^2, \sigma_\varepsilon^2, \mathbf{u}, \beta \mid \mathbf{y}) \propto f(\mathbf{y} \mid \beta, \mathbf{u}, \sigma_\varepsilon^2) f(\mathbf{u}|\sigma^2) g(\beta) g(\sigma^2) g(\sigma_\varepsilon^2) .$$

Tiao and Tan (*Biometrika*, 1965) propose the following improper priors for the model parameters:

$$g(\beta) \propto 1, \quad g(\sigma^2) \propto 1/\sigma^2 \quad \text{and} \quad g(\sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon^2 \, .$$

Show that, for this choice of prior model, the posterior above is improper. (**Note:** If you need guidance on how this may be shown, see Hill (*JASA*, 1965).)

6.2 Bayesian self-consistency

We begin by describing the basic inferential framework within which our discussion and results will be lodged. It will be assumed that a random (i.i.d.) sample of size n is available from an exponential family of distributions indexed by a scalar parameter θ. We will be interested in the problem of estimating θ relative to a squared error loss criterion. Assume that $\widehat{\theta}$ is sufficient for θ and that it is an unbiased estimator of θ.

In the parametrization to be used here, prior distributions will be indexed by two particular parameters, the first being

$$\theta^* = E_\pi \theta \, , \tag{6.1}$$

the prior mean, with the second parameter ω representing what is generally referred to as "the prior sample size." These parameters are often represented as specific functions of the parameters of the prior in its most common form (for example, for the $\text{Be}(\alpha, \beta)$ prior, $\theta^* = \alpha/(\alpha+\beta)$ and $\omega = \alpha+\beta$), though they may also be written as specific functions of the mean μ and the variance σ^2 of the prior model. When the beta prior $\text{Be}(\alpha, \beta)$ on the proportion θ has its standard parametrization, the mean of θ is $\mu = \theta^* = \alpha/(\alpha+\beta)$ and the variance is $\sigma^2 = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$. In this instance, one may write $\theta^* = \mu$ and $\omega = \{\mu(1-\mu) - \sigma^2\}/\sigma^2$; the beta model may be parametrized in terms of θ^* and ω by replacing α and β by $\omega\theta^*$ and $\omega(1-\theta^*)$ in the standard parametrization. Unless otherwise stated, conjugate priors referred to below are assumed to be parametrized in terms of θ^* and ω. Let $\pi(\theta^*, \omega)$ be a "standard" conjugate prior model for θ. We will denote the family of priors of a parametric form of interest by Π. Typical examples of such families are displayed in Table 3.1.

When a conjugate prior is used for a specific sampling distribution in an exponential family with parameter θ, the posterior mean is often a convex combination of the prior mean θ^* and the UMVUE $\widehat{\theta}$ and the weights on each are, respectively, proportional to ω and the sample size n. Specifically,

$$E(\theta|\widehat{\theta}) = \frac{\omega}{n+\omega} \, \theta^* + \frac{n}{n+\omega} \, \widehat{\theta} \, . \tag{6.2}$$

Even though the term "prior sample size" would seem to require that the parameter ω be a positive integer, we will use the term, as is customary, for any real $\omega > 0$. It is

clear from (6.2) where ω gets its name. As in the preceding chapter, the weight this estimator places on $\widehat{\theta}$ will be denoted by $\eta \in (0,1)$ and is given by

$$\eta = \frac{n}{n+\omega} \,. \tag{6.3}$$

The best known and most useful property of a family of conjugate priors is its closure under Bayesian updating, that is, the property that the posterior distribution of the parameter belongs to the same family Π as the prior distribution. Indeed, this property is often used as the definition of conjugacy, as in, for example, Gelman *et al.* (2004). These authors also acknowledge, however, that this definition leaves something to be desired, as the family of all distributions on θ is also conjugate in this sense. Diaconis and Ylvisaker (1979) showed that, under mild regularity conditions, the properties of closure and a linear posterior mean characterize the "standard" conjugate prior families.

The notion of self-consistency of inferential procedures has arisen in a number of statistical contexts. In a survival-analysis setting, Efron (1967) defined self-consistency in terms of a natural recursive relationship that nonparametric estimators of a survival function might satisfy. Efron then showed that the Kaplan–Meier estimator is the unique self-consistent estimator of the survival function on the interval containing all deaths and censoring times. Tsai and Crowley (1985) identified self-consistent estimators as the unique fixed points of nonparametric EM algorithms. In this paper, we will look at self-consistency from a Bayesian perspective.

Definition 6.1. *Given a fixed sufficient and unbiased estimator $\widehat{\theta}$ of the scalar parameter θ and a prior distribution G with mean θ^*, the Bayes estimator of θ with respect to G, relative to squared error loss, is said to be* self-consistent *if*

$$E(\theta \mid \widehat{\theta} = \theta^*) = \theta^* \,. \tag{6.4}$$

Remark 6.1. Self-consistency is equivalent to the requirement that the prior mean θ^* be a fixed point of the posterior mean function. Equation (6.4) says: if your experimental outcome agrees with your prior opinion about θ, then the experiment shouldn't change your opinion.

Remark 6.2. The definition above clearly depends on the estimator $\widehat{\theta}$ of θ that one chooses to use as a summary of the available data. It would thus seem that one should speak of self-consistency relative to this estimator rather than in general. While this is technically true, the reference to $\widehat{\theta}$ will be subsumed in the sequel, since for sampling distributions belonging to exponential families (which is our focus here), there can only be one unbiased sufficient statistic, and special reference to the estimator $\widehat{\theta}$ is unnecessary.

Remark 6.3. Given that the self-consistency property is a reasonable expectation to have in one's prior modeling, it is natural to ask how broad the self-consistency property is. We are not able to provide a definitive answer to this question at present, but three specific claims can be substantiated. First, in the context of sampling distributions belonging to exponential families, the traditional conjugate families $\{\pi(\theta^*, \omega)\}$

are self-consistent. This is apparent from the fact that the posterior mean in such situations enjoys the convexity property in equation (6.2). Second, the property goes beyond conjugate families, though such extensions appear to be rather limited. An example of a self-consistent nonconjugate prior is mentioned at the end of the next section in a problem involving a sampling distribution that is not a member of an exponential family. Finally, it is clear that no improper prior can be self-consistent, as the notion of prior mean is then vacuous.

There are a variety of circumstances in which estimators which are approximately Bayes in some sense (relative to G) are advocated for use. Further, since the definition in (6.4) deals only with Bayes rules relative to squared error loss, some appropriate extension of the definition would seem to be in order. One natural extension of the notion of self-consistency that applies to such possibilities is the following:

Definition 6.2. *Let G be a prior distribution for a scalar parameter θ, and denote the mean of G by θ^*. Let $\widehat{\theta}$ be a sufficient, unbiased estimator for θ and let $T_G(\widehat{\theta})$ be an estimator of θ that approximates (in some sense) the Bayes estimator $\widehat{\theta}_G$ relative to a fixed loss criterion $L(\theta, \widehat{\theta})$. The estimator $T_G(\widehat{\theta})$ is said to be* generalized self-consistent *relative to the prior G if it satisfies the equation*

$$T_G(\theta^*) = \theta^* \,. \tag{6.5}$$

Definition 6.2 mimics Definition 6.1 in requiring that the prior mean be a fixed point of the estimating function. They both quantify the notion that one's prior and posterior opinion about the parameter θ should be the same when the data and your prior opinion of θ agree.

Remark 6.4. It may not be immediately apparent from equations (6.4) or (6.5), but both versions of self-consistency depend on the precise circumstances under which it will be applied. More specifically, since the posterior mean of θ, given $\widehat{\theta}$, depends on the size n of the available sample, a given prior distribution may be self-consistent for a fixed sample size n_1 and yet not satisfy the self-consistency equation for an alternative sample size n_2. This might very well seem to the reader to be a disturbing characteristic of self-consistency. Yet the dependence of the prior model on the size of one's experiment is clearly anathema to the subjectivist Bayesian. One answer to this apparent dilemma is that the designation of a prior distribution is a process that takes place in the context of a given experiment, with the sample size fixed and known. Thus, concerns about Bayesian inference in an alternative experiment one doesn't actually have in hand should be set aside. A seemingly more satisfactory approach one could take would be to restrict attention to priors that are self-consistent for all sample sizes in the experimental setting under study. It should be noted that standard conjugate priors satisfy such a restriction.

Remark 6.5. When the estimator $\widehat{\theta}$ is a discrete random variable, equation (6.4) is well defined only if the mean θ^* of the prior under consideration lies within the support set $S(\widehat{\theta})$ of $\widehat{\theta}$. When $\widehat{\theta}$ is discrete, we will consider a prior to be self-consistent if equation (6.4) is not violated. Thus, for example, the beta prior with parameters

$\alpha = 1$ and $\beta = \sqrt{2}$ is self-consistent when estimating θ based on the binomial observation $X \sim \mathscr{B}(n,\theta)$ since the mean θ^* of this prior is irrational and thus $\theta^* \notin S(\widehat{\theta})$ for any given fixed value of n. The point of self-consistency is that one's prior and posterior opinions about θ should not be contradictory when the data confirms your prior opinion. When $\widehat{\theta}$ is continuous, self-consistency is fully characterized by equation (6.4). Similar remarks apply to the notion of generalized self-consistency.

A question that arises quite naturally concerns the breadth of the notion of Bayesian self-consistency. We mentioned above that in the context of sampling distributions belonging to exponential families and under squared error loss, standard conjugate prior distributions are self-consistent. Is it possible that the self-consistency equation (6.4) actually characterizes standard prior distributions in the setting just outlined? The following example provides a negative answer to this question.

Example 6.1. Consider a binomial experiment yielding the outcome $X \sim \mathscr{B}(10,\theta)$, and let $\pi^{(a)}$ represent the prior distribution that is the mixture of two specific Beta distributions, namely, the prior $a\text{Be}(3,2) + (1-a)\text{Be}(2,3)$, where $a \in [0,1]$. The mean of this prior is $\theta^* = 2/5 + a/5$. Given the distributions of $X|\theta$ and of θ as specified above, the posterior distribution of $\theta \mid X = x$ has a density function given by

$$f(\theta|x) = \left[a\theta^{2+x}(1-\theta)^{11-x} + (1-a)\theta^{1+x}(1-\theta)^{12-x}\right] \,/\, A(x)$$

where

$$A(x) = a\frac{\Gamma(3+x)\Gamma(12-x)}{\Gamma(15)} + (1-a)\frac{\Gamma(2+x)\Gamma(13-x)}{\Gamma(15)}\,.$$

It follows that the posterior mean of θ, given $X = x$, is

$$E(\theta \mid X = x) = \frac{1}{15}\frac{a(3+x)(2+x) + (1-a)(2+x)(12-x)}{a(2+x) + (1-a)(12-x)}\,.$$

The self-consistency equation in (6.4) may be written here as

$$E\left(\theta \mid \widehat{p} = \frac{2}{5} + \frac{a}{5}\right) = \frac{2}{5} + \frac{a}{5}\,, \tag{6.6}$$

where $\widehat{p} = X/10$. We would like to identify all values of a, that is, all mixtures of the form $\pi^{(a)}$ for which (6.6) holds. Making the substitution $y = 2/5 + a/5$ or equivalently, setting $a = 5y - 2$, equation (6.6) becomes

$$\frac{1}{15}\frac{(5y-2)(3+10y)(2+10y) + (3-5y)(2+10y)(12-10y)}{(5y-2)(2+10y) + (3-5y)(12-10y)} = y\,. \tag{6.7}$$

It is clear that (6.7) is equivalent to a cubic equation in y with three potential roots. The fact that the distributions Be(3,2) and Be(2,3) are standard conjugate priors implies that two of those roots are $y = 2/5$ and $y = 3/5$ (corresponding to the values $a = 0$ and $a = 1$). It is a simple matter to confirm that the third root also resides in the unit interval. Indeed, (6.7) is satisfied when $y = 1/2$, showing that the prior

$(1/2)\text{Be}(3,2) + (1/2)\text{Be}(2,3)$ also satisfies equation (6.4) in this problem. Thus, while the class of priors satisfying the self-consistency equation is certainly quite limited, it is apparent that the self-consistency equation does not apply exclusively to the standard conjugate priors. ■

An interesting question that remains open is whether or not, when sampling distributions are exponential families and the loss criterion is squared error, the self-consistency property characterizes the standard prior distributions among prior distributions belonging to an exponential family. The example above does not shed light on this question.

We now turn our attention to an investigation of the "consensus problem." As we shall see, the notions of conjugacy and of self-consistency arise naturally in the approach we take to this problem. The "consensus estimator" to be recommended for use in Section 6.3 will be seen to be generalized self-consistent relative to the consensus prior distribution considered there.

Exercise 6.6. Consider a normal experiment yielding the outcome $X \sim \mathcal{N}(\mu, 1)$, and let $\pi^{(a)}$ represent the prior distribution $a\mathcal{N}(1,1) + (1-a)\mathcal{N}(-1,1)$, where $a \in [0,1]$. For what values of the mixing constant a does the prior $\pi^{(a)}$ satisfy the self-consistency equation (6.4)?

6.3 An approach to the consensus problem

The elicitation of prior opinion is an important element of the practice of subjective Bayesian analysis. While a statistician can always arrive at a prior distribution through some form of introspection (or may decide on a given prior because of its convenience), the major advantages that can be gained from a Bayesian analysis often come from consultation with experts in the subject matter pertaining to a given application. Here, we will consider the problem that arises when several (say k) experts are consulted, and they have different, perhaps conflicting, prior opinions about the parameter of interest. We will again assume that the sampling distribution involved belongs to an exponential family. Because of the simplicity of using conjugate prior distributions in elicitation (by asking the expert for a best guess at the parameter and the weight he/she would wish to place on it), we will assume that each of the k experts has responded with a conjugate prior, resulting in the collection $\{\pi(\theta_i^*, \omega_i),\ i = 1, \ldots, k\}$.

The question we now consider is how the statistician should deal with this prior information. One approach would be to engage the experts in conversations that might lead to convergence toward agreement about a single prior distribution for use in the problem. This approach is laden with practical difficulties, including the possibility that achieving such agreement proves to be impossible. The approach we take here places the burden of finding a consensus prior opinion on the statistician's shoulders. It is perhaps natural to consider mixtures of the prior opinions, and in fact that will be our starting point. Let us digress briefly to establish some notation.

In the problem considered here, I will restrict attention to finite mixtures. Let us consider the k distinct conjugate priors in the problem of estimating the scalar parameter θ of an exponential family based on a random sample of size n. Let $\widehat{\theta}$ be a sufficient, unbiased estimator of θ, and let $\mathbf{a} = (a_1, a_2, \ldots, a_k)$ be a vector satisfying the conditions

$$a_i \geq 0 \text{ for } i = 1, \ldots, k \tag{6.8}$$

and

$$\sum_{i=1}^{k} a_i = 1 \, . \tag{6.9}$$

Let π_a be the k-fold mixture of the priors $\pi(\theta_i^*, \omega_i)$ according to the mixing distribution $\mathbf{a}$, that is, let

$$\pi_{\mathbf{a}} = \sum_{i=1}^{k} a_i \pi(\theta_i^*, \omega_i) \, . \tag{6.10}$$

Use of the prior family of k-fold mixtures of conjugate priors offers some immediate advantages. A number are listed here. (a) The mean of the prior $\pi_{\mathbf{a}}$ is $\theta_{\mathbf{a}}^* = \sum_{i=1}^{k} a_i \theta_i^*$, the corresponding mixture of the means of the component conjugate priors. Its closed form and natural interpretability are a boon to prior modeling. (b) The posterior distribution is again a k-fold mixture of conjugate priors, giving the prior model a form of conjugacy. It should be noted, however, that the mixing distribution of the posterior model will typically be data dependent. (c) These models are appropriately fashioned to play a natural role in representing the consensus prior when several experts are consulted in the process of eliciting prior opinions; the mixing distribution $\mathbf{a}$ affords the statistician the flexibility of assigning each expert the weight or influence that seems due. While the mixture in (6.10) has appeal as a prior model, the posterior distribution is a complex mixture of the component conjugate posteriors, and the Bayes estimator of θ relative to $\pi_{\mathbf{a}}$ can be quite complex. In addition, it loses the intuitive appeal and convenience of linearity and generally fails to satisfy the self-consistency condition in (6.4).

A natural alternative to the Bayes estimator in the consensus problem, relative to the prior in (6.10), is the class of convex combinations of the individual Bayes estimators with respect to the component conjugate priors. Let us consider for further study the class of estimators of the form

$$\widetilde{\theta}_{\pi} = \sum_{i=1}^{k} a_i \widehat{\theta}_i \, , \tag{6.11}$$

where $\widehat{\theta}_i = (1 - \eta_i)\theta_i^* + \eta_i \widehat{\theta}$, the Bayes estimator with respect to $\pi(\theta_i^*, \omega_i)$, with $\eta_i = n/(n + \omega_i)$. The estimator in (6.11) has a number of practical advantages. First, the estimator is a linear function of the sufficient unbiased estimator $\widehat{\theta}$. Linear estimators have been found useful in Bayesian inference, and have been espoused in selected circumstances (see, e.g., Hartigan (1969) and Ericson (1969, 1970)) for both their tractability and their interpretability. Second, this estimator is ideally constructed as a consensus estimator. The statistician, having elicited opinions from k experts,

can use the mixture in (6.11) to assign the weight to each expert's opinion that he feels the input merits. Finally, under specific conditions given in the results below, we will see that consensus estimators enjoy both generalized self-consistency and a desirable convexity property, and, in fact, will be the Bayes estimator relative to a conjugate prior with the same mean as the mixture of conjugate priors in (6.10). Let us proceed to the theoretical results that establish these facts. The following result provides conditions under which the linear estimator in (6.11) has the generalized self-consistency property.

Theorem 6.1. *Let $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ be a sufficient, unbiased estimator of θ, the parameter of an exponential family, and let $\widetilde{\theta}_\pi$ be the linear estimator in* (6.11)*. If $\theta^*_{\mathbf{a}} \in S(\widehat{\theta})$, then the estimator $\widetilde{\theta}_\pi$ is generalized self-consistent (GSC) if and only if*

$$\sum_{i=1}^{k} a_i \eta_i (\theta^*_{\mathbf{a}} - \theta^*_i) = 0 \,. \tag{6.12}$$

If $\theta^*_{\mathbf{a}} \notin S(\widehat{\theta})$, then $\widetilde{\theta}_\pi$ is generalized self-consistent in the sense of Remark 6.5.

Proof. The GSC equation in (6.5), with $T(\widehat{\theta}) = \widetilde{\theta}_\pi$, may be written as

$$\sum_{i=1}^{k} a_i [(1-\eta_i)\theta^*_i + \eta_i \theta^*_{\mathbf{a}}] = \theta^*_{\mathbf{a}} \,. \tag{6.13}$$

Since $\theta^*_{\mathbf{a}} = \sum_{i=1}^{k} a_i \theta^*_i$, equation (6.13) may be rewritten as

$$\sum_{i=1}^{k} a_i \theta^*_i - \sum_{i=1}^{k} a_i \eta_i \theta^*_i + \sum_{i=1}^{k} a_i \eta_i \theta^*_{\mathbf{a}} = \sum_{i=1}^{k} a_i \theta^*_i \,. \tag{6.14}$$

Upon the cancellation of the common term $\sum_{i=1}^{k} a_i \theta^*_i$ on both sides of (6.14), we have

$$\sum_{i=1}^{k} a_i \eta_i (\theta^*_{\mathbf{a}} - \theta^*_i) = 0 \,. \qquad \blacksquare \tag{6.15}$$

The necessary and sufficient conditions for the generalized self-consistency of $\widetilde{\theta}_\pi$ given in Theorem 6.1 (namely, either (a) $\theta^*_{\mathbf{a}} \notin S(\widehat{\theta})$ or (b) $\theta^*_{\mathbf{a}} \in S(\widehat{\theta})$ and (6.12) holds) allow for a broad range of possible generalized self-consistent estimators. It is of particular interest to identify circumstances in which a given prior is self-consistent for any experiment in the class under consideration, that is, for any value of the sample size n. If $\theta^*_{\mathbf{a}} \in S(\widehat{\theta})$ for all n, either of the conditions (i) $\theta^*_i = \theta^*_{\mathbf{a}}$ for $i = 1, \ldots, k$ or (ii) $\eta_i = \eta$ for $i = 1, \ldots, k$ is sufficient for ensuring the generalized self-consistency of θ_π for all n. Condition (ii) has some practical value in that it can be implemented by the statistician by eliciting only the prior mean from each expert and then assigning each a fixed, common prior sample size ω. This prior sample size might, for example, represent the weight the statistician wishes to place on the combined prior input. This interpretation is borne out in equation (6.17) below.

The following example illustrates the dependence of generalized self-consistency on condition (6.12).

Example 6.2. Suppose that $X \sim \mathscr{B}(10,\theta)$ and $\widehat{\theta} = X/10$. Let $k = 2$ and $\mathbf{a} = (0.64, 0.36)$, and take $\pi_{\mathbf{a}} = 0.64$ Beta(3,1) + 0.36 Beta(1,2). We thus have that $\theta_1^* = 3/4$ and $\theta_2^* = 1/3$ and that $\theta_{\mathbf{a}}^* = 6/10$, while $\eta_1 = 10/14$ and $\eta_2 = 10/13$. We note that condition (6.12) fails to hold under these circumstances. The estimator $\widetilde{\theta}_\pi$ of θ is given by

$$\widetilde{\theta}_\pi = 0.64[3/14 + (10/14)\widehat{\theta}] + 0.36[1/13 + (10/13)\widehat{\theta}] = 15/91 + (334/455)\widehat{\theta}\,.$$

If $\widehat{\theta} = \theta_{\mathbf{a}}^* = 0.6$, then $\widetilde{\theta}_\pi = 0.1648 + 0.7341(0.6) = 0.6052$. Thus, when equation (6.12) fails, the otherwise natural estimator $\widetilde{\theta}_\pi$ can take on a numerical value that differs from the common value suggested by one's prior distribution and by the data themselves. ■

What conditions on the parameters $\{a_i, \theta_i^* \text{ and } \eta_i,\ i = 1,\ldots,k\}$ of the prior model will guarantee that the estimator $\widetilde{\theta}_\pi$ is a convex combination of the prior mean $\theta_{\mathbf{a}}^*$ and the estimator $\widehat{\theta}$? Because of its interpretability, this convexity property is one of the most appealing features of standard conjugate priors. It allows one to pursue prior elicitation through the two most natural questions one might ask: "what is one's best guess at the value of the unknown parameter?" and "how much weight would one put on that guess relative to the weight one would put on the estimator $\widehat{\theta}$?" The answer to the question concerning convexity is provided in the result below.

Theorem 6.2. . *Let $\widehat{\theta} = \widehat{\theta}(X_1,\ldots,X_n)$ be a sufficient, unbiased estimator of θ, the parameter of an exponential family, and let $\widetilde{\theta}_\pi$ be the linear estimator of θ in* (6.11). *The estimator $\widetilde{\theta}_\pi$ may be written as*

$$\widetilde{\theta}_\pi = (1-c)\theta_{\mathbf{a}}^* + c\widehat{\theta} \tag{6.16}$$

for some $c \in (0,1)$ if and only if the generalized self-consistency equation (6.12) *holds.*

Proof. The estimator $\widetilde{\theta}_\pi$ will have the form in (6.16) if and only if $c = \sum_{i=1}^k a_i\eta_i$. It follows that (6.16) holds if and only if

$$\sum_{i=1}^k a_i(1-\eta_i)\theta_i^* = \left(1 - \sum_{i=1}^k a_i\eta_i\right)\theta_{\mathbf{a}}^*\,,$$

an equation which reduces to (6.12). ■

The two theorems above demonstrate the somewhat surprising fact that condition (6.12) simultaneously guarantees two important properties of the associated consensus estimator $\widetilde{\theta}_\pi$, namely, that the estimator will be self-consistent in the generalized sense and that it can be written as a convex combination of the "prior guess" $\theta_{\mathbf{a}}^*$ and the sufficient, unbiased estimator $\widehat{\theta}$ of θ. The latter fact has, by itself, a further interesting implication: we may infer from it that a consensus estimator $\widetilde{\theta}_\pi$ which satisfies condition (6.12) is in fact a Bayes estimator! Indeed, it is the Bayes estimator of θ

with respect to the conjugate prior $\pi(\theta^*_{\mathbf{a}}, \omega)$ and squared error loss, where the prior sample size ω is given by

$$\omega = n\left(1 - \sum_{i=1}^{k} a_i \eta_i\right) \Bigg/ \sum_{i=1}^{k} a_i \eta_i \ . \tag{6.17}$$

It is thus apparent that, when estimating the parameter θ of an exponential family with squared error loss, an estimator $\widetilde{\theta}_\pi$ of θ satisfying condition (6.12) will also satisfy the assumptions of Theorem 5.1. It follows that, for the class of consensus estimators under study, that is, for any $\widetilde{\theta}_\pi$ having the form in (6.11) and satisfying the condition (6.12), it is possible to directly compare the expected performance of the Bayes estimator $\widetilde{\theta}_\pi$ and the best frequentist estimator $\widehat{\theta}$. The availability of such comparisons provides strong motivation for the use of the proposed consensus estimators, as the assessment of the comparative performance of alternative Bayes estimators (including Bayes rules wrt mixtures of standard conjugate priors) in the consensus problem is generally intractable, while the performance characteristics of consensus estimators defined by (6.11) and (6.12) relative to that of the best frequentist estimator can be described quite precisely, leading to rather explicit guidance regarding the potential efficacy of the estimator $\widetilde{\theta}_\pi$. Applying Theorem 5.1, one may characterize the conditions under which the proposed consensus estimators are superior to their frequentist counterparts. For any fixed distribution G_0 for which the expectations exist,

$$\mathbf{r}(G_0, \widetilde{\theta}_\pi) \leq \mathbf{r}(G_0, \widehat{\theta}) \tag{6.18}$$

if and only if

$$V_{G_0}(\theta) + \left(\sum_{i=1}^{k} a_i \theta_i^* - E_{G_0}\theta\right)^2 \leq \frac{1 + \sum_{i=1}^{k} a_i \eta_i}{1 - \sum_{i=1}^{k} a_i \eta_i}\, \mathbf{r}(G_0, \widehat{\theta}) \ . \tag{6.19}$$

It follows that, in the context of exponential families and squared error loss, when using a consensus estimator $\widetilde{\theta}_\pi$ satisfying equation (6.12), the Bayesian will have an advantage over the frequentist in estimating θ unless the Bayesian is both misguided and stubborn, that is, unless he utilizes a prior mean $\theta^*_{\mathbf{a}}$ that is substantially distant from the true (though unknown) value of θ and he places a considerable amount of weight on that prior mean.

The statistical framework studied here has been restricted to estimation problems involving squared error loss and sampling distributions belonging to one-parameter exponential families. Do the types of results obtained here have wider applicability? In Chapters 7 and 8, we will investigate the "threshold problem" in broader contexts — estimation of a vector-valued parameter and estimation with asymmetric loss. Regarding the breadth of applicability of the notion of Bayesian self-consistency, it is easy to verify that, under squared error loss, the Bayes estimator of θ when $X|\theta \sim \mathscr{U}[0, \theta]$ and θ has prior distribution $\Gamma(2, 1)$ is self-consistent, demonstrating that the concept extends beyond sampling distributions from exponential families coupled with standard conjugate priors or selected mixtures thereof.

Exercise 6.7. Two metallurgists are consulted about the mean tensile strength of a newly developed alloy when subjected to a random stress. The strength, measured as the strain at peak load (a standard proxy for the strength of a material), is modeled as an exponential variable with mean θ. If a sample of size 25 yields a mean $\overline{X} = 4$ and the two experts have provided prior guesses $\theta_1^* = 5$ and $\theta_2^* = 6$ (with $\eta_1 = \eta_2$ set equal to 1/2), identify the class of consensus estimators $\widetilde{\theta}_\pi$ that are superior to the frequentist estimator $\overline{X}$ (in the sense of (6.19)) if the true value of θ happens to be 4.25.

Exercise 6.8. Prove the claim in this chapter's last sentence.

7

Bayesian vs. Frequentist Shrinkage in Multivariate Normal Problems

7.1 Preliminaries

This chapter is dedicated to the comparison of Bayes and frequentist estimators of the mean $\boldsymbol{\theta}$ of a multivariate normal distribution in high dimensions. For dimension $k \geq 3$, the James–Stein estimator specified in (2.15) (and its more general form to be specified below) is usually the frequentist estimator of choice. The estimator is known to improve uniformly upon the sample mean vector $\overline{\mathbf{X}}$ as an estimator of $\boldsymbol{\theta}$ when $k \geq 3$, and while it is also known that it is not itself admissible, extant alternative estimators with smaller risk functions are known to offer only very slight improvement. For this and other reasons, the James–Stein estimator is widely used among estimators which exploit the notion of shrinkage. In the results described in this chapter, I will use the form of the James–Stein estimator which shrinks $\overline{\mathbf{X}}$ toward a (possibly nonzero) distinguished point. This serves the purpose of placing the James–Stein estimator and the Bayes estimator of $\boldsymbol{\theta}$ with respect to a standard conjugate prior distribution in comparable frameworks, since the latter also shrinks $\overline{\mathbf{X}}$ toward a distinguished point. It is interesting to note that the James–Stein estimator has a certain Bayesian flavor that goes beyond the empirical Bayes character highlighted in the writings of Efron and Morris (1973, etc.) in that the act of shrinking toward a particular parameter vector suggests that the statistician using this estimator is exercising some form of introspection in determining a good "prior guess" at $\boldsymbol{\theta}$. The Bayesian of course goes further, specifying, *a priori*, the weight he wishes to place on the prior guess. What results in the latter case is an alternative form of shrinkage, one that leads to a linear combination of $\overline{\mathbf{X}}$ and the prior guess, with weights influenced by the prior distribution rather than by the observed data. Since $\overline{\mathbf{X}}$ is a sufficient statistic for the mean of a multivariate normal distribution with known variance-covariance matrix $\boldsymbol{\Sigma}$, I will henceforth, without loss of generality, take the sample size n to be 1.

A general treatment of a comparison between Bayesian and frequentist shrinkage in estimating the mean vector of the distribution $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ remains, as of the present exposition, intractable. I will begin by defining the threshold problem in the general case. I will then turn to the special case from which considerable insight and intuition

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_7,

can be gleaned and on which the work presented in this chapter is focused. The core of this chapter is based on two papers in which the solution of the version of the threshold problem treated here is given. The reader is referred to Vestrup and Samaniego (2004a, 2004b) for detailed proofs and further discussion.

As in Chapters 4 and 5, we will posit the existence of a "true prior distribution" G_0 representing the true state of nature in a given estimation problem. As before, the special case in which G_0 is degenerate at a point $\boldsymbol{\theta}_0$ — the true but unknown value of the target parameter $\boldsymbol{\theta}$ — will play a prominent role in the analysis pursued here, as this assumption is appropriate in most applications of interest, where $\boldsymbol{\theta}$ is simply an unknown k-dimensional vector. In a general treatment of the threshold problem, one would typically make the following assumptions regarding the true prior G_0, the operational prior G, the sampling distribution of $\mathbf{X}$ and the loss function L:

$$G_0 : \boldsymbol{\theta} \sim \mathcal{N}_k(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0) \tag{7.1}$$

$$G : \boldsymbol{\theta} \sim \mathcal{N}_k(\boldsymbol{\theta}_G, \boldsymbol{\Sigma}_G) \tag{7.2}$$

$$F_{\mathbf{X}|\boldsymbol{\theta}} : \mathbf{X}|\boldsymbol{\theta} \sim \mathcal{N}_k(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \tag{7.3}$$

where $\boldsymbol{\Sigma}$ is a known positive definite matrix, and

$$L(\boldsymbol{\theta}, \mathbf{a}) = (\boldsymbol{\theta} - \mathbf{a})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \mathbf{a}) \,. \tag{7.4}$$

The Bayes risk $E_{G_0} E_{\mathbf{X}|\boldsymbol{\theta}} L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_G)$, relative to the true prior G_0, of the Bayes estimator $\widehat{\boldsymbol{\theta}}_G$ wrt the operational prior G is shown in Vestrup and Samaniego (2003) to be

$$r(G_0, \widehat{\boldsymbol{\theta}}_G) = \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1/2} \mathbf{A} \boldsymbol{\Sigma} \mathbf{A}' \boldsymbol{\Sigma}^{-1/2}\right) + \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1/2} \mathbf{B} \boldsymbol{\Sigma}_0 \mathbf{B}' \boldsymbol{\Sigma}^{-1/2}\right) + \left\| \boldsymbol{\Sigma}^{-1/2} \mathbf{B} (\boldsymbol{\theta}_G - \boldsymbol{\theta}_0) \right\|^2 , \tag{7.5}$$

where $\mathbf{A} = \boldsymbol{\Sigma}_G (\boldsymbol{\Sigma}_G + \boldsymbol{\Sigma})^{-1}$ and $\mathbf{B} = \mathbf{I} - \mathbf{A}$. The James–Stein estimator which shrinks $\mathbf{X}$ toward the constant vector $\boldsymbol{\theta}^*$ will be denoted by $\widehat{\boldsymbol{\theta}}_{JS,\boldsymbol{\theta}^*}$. The Bayes risk of the $\widehat{\boldsymbol{\theta}}_{JS,\boldsymbol{\theta}^*}$ (in its general form, i.e., applicable to estimating the mean $\boldsymbol{\theta}$ in the model (7.3)) relative to the conjugate prior G_0 in (7.1) is also derived in Vestrup and Samaniego (2004b), and is shown to be given by an infinite series involving expectations of rather complex functions of an infinite collection of Poisson random variables. Inspection of these two expressions for Bayes risk makes clear that the determination of the class of operational priors G for which the Bayes estimator $\widehat{\boldsymbol{\theta}}_G$ is superior to the James–Stein estimator $\widehat{\boldsymbol{\theta}}_{JS,\boldsymbol{\theta}^*}$ which shrinks $\mathbf{X}$ toward the vector $\boldsymbol{\theta}_G$, that is, for which

$$r\left(G_0, \widehat{\boldsymbol{\theta}}_G\right) \le r\left(G_0, \widehat{\boldsymbol{\theta}}_{JS,\boldsymbol{\theta}^*}\right) \tag{7.6}$$

is not a tractable exercise. (I have nonetheless listed this version of the threshold problem as Exercise 7.1, and I invite the highly motivated, long-suffering reader to mail me his or her solution!)

The following simplifications do lead to a definitive solution which yields substantial insight. Consider the following special case of the framework in (7.1)–(7.4):

$$G_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ with probability 1} \tag{7.7}$$

$$G : \boldsymbol{\theta} \sim \mathcal{N}_k\left(\boldsymbol{\theta}_G, \sigma_G^2 \mathbf{I}\right) \tag{7.8}$$

$$F_{\mathbf{X}|\boldsymbol{\theta}} : \mathbf{X}|\boldsymbol{\theta} \sim \mathcal{N}_k\left(\boldsymbol{\theta}, \sigma^2 \mathbf{I}\right) \tag{7.9}$$

where σ^2 is a known constant, and

$$L(\boldsymbol{\theta}, \mathbf{a}) = \frac{1}{\sigma^2} \sum_{i=1}^{k} (\theta_i - a_i)^2 \,. \tag{7.10}$$

The framework in (7.8)–(7.10) corresponds to Robbins' (1951) original formulation of the estimation problem of interest as a compound decision problem and is the context in which Stein (1956) first demonstrated the inadmissibility of the sample mean in dimension $k \geq 3$. The specification of the true prior in (7.7) is the most common description of the "truth" in applications involving the estimation of a multivariate normal mean. In this simpler context, the Bayes estimator of $\boldsymbol{\theta}$ with respect to the operational prior G is

$$\widehat{\boldsymbol{\theta}}_G = \frac{\sigma_G^2}{\sigma^2 + \sigma_G^2} \mathbf{X} + \frac{\sigma^2}{\sigma^2 + \sigma_G^2} \boldsymbol{\theta}_G \,, \tag{7.11}$$

while the James–Stein estimator which shrinks $\mathbf{X}$ toward $\boldsymbol{\theta}_G$ is given by

$$\widehat{\boldsymbol{\theta}}_{JS,\boldsymbol{\theta}_G} = (\mathbf{X} - \boldsymbol{\theta}_G)\left(1 - \frac{\sigma^2(k-2)}{\|\mathbf{X} - \boldsymbol{\theta}_G\|^2}\right) + \boldsymbol{\theta}_G \,. \tag{7.12}$$

In the framework of (7.7)–(7.10), the Bayes risks of the two estimators, relative to the degenerate prior G_0, reduce to

$$r(G_0, \widehat{\boldsymbol{\theta}}_G) = \frac{k\sigma_G^4 + \sigma^2 \|\boldsymbol{\theta}_G - \boldsymbol{\theta}_0\|^2}{(\sigma^2 + \sigma_G^2)^2} \tag{7.13}$$

and

$$r\left(G_0, \widehat{\boldsymbol{\theta}}_{JS,\boldsymbol{\theta}_G}\right) = k - (k-2)E\left(\frac{1}{k-2+2T}\right) \,, \tag{7.14}$$

where $T \sim P(\|\boldsymbol{\theta}_G - \boldsymbol{\theta}_0\|^2 / 2\sigma^2)$. The latter expression matches that derived by James and Stein (1961). In the developments that follow, it will be convenient to denote the Euclidean distance between the prior mean and the true value of $\boldsymbol{\theta}$ (or mean of G_0) by

$$\Delta = \|\boldsymbol{\theta}_G - \boldsymbol{\theta}_0\|^2 \,. \tag{7.15}$$

Taking $T \sim P(\Delta / 2\sigma^2)$, the inequality $r(G_0, \widehat{\boldsymbol{\theta}}_G) \leq r(G_0, \widehat{\boldsymbol{\theta}}_{JS,\boldsymbol{\theta}_G})$ may then be written as

$$\frac{k\sigma_G^4 + \sigma^2\Delta}{(\sigma^2 + \sigma_G^2)^2} \leq k - (k-2)E\left(\frac{1}{k-2+2T}\right) \,. \tag{7.16}$$

Exercise 7.1. Characterize the solutions $\{G\}$ such that $r(G_0, \widehat{\boldsymbol{\theta}}_G) \leq r(G_0, \widehat{\boldsymbol{\theta}}_{JS,\boldsymbol{\theta}_G})$ under the assumptions (7.1)–(7.4). Mail your result to the author. ☺

Exercise 7.2. Under the sampling assumption in (7.9) and the loss function in (7.10), verify that the estimator $\widehat{\boldsymbol{\theta}} = \mathbf{X} \sim F_{\mathbf{X}|\boldsymbol{\theta}}$ has a constant risk function.

Exercise 7.3. Suppose the true prior G_0 is degenerate at the point $\boldsymbol{\theta}_0$ and that the operational prior G in (7.8) is mean correct. Taking $n = 1$ and $\widehat{\theta} = X$ in the framework of (7.7)–(7.10), show that $r(G_0, \widehat{\boldsymbol{\theta}}_G) \leq r(G_0, \widehat{\boldsymbol{\theta}})$ uniformly in the prior variance σ_G^2. (**Note:** This result implies that any Bayes estimator with respect to a mean-correct prior will uniformly outperform the sample mean $\overline{\mathbf{X}}$ based on an MVN sample of arbitrary size n.)

7.2 A solution to the threshold problem

Characterizing the prior specifications $(\boldsymbol{\theta}_G, \sigma_G^2)$ for which the inequality in (7.16) holds (that is, characterizing the Bayes estimators which outperform the James–Stein estimator under the modeling assumptions (7.7)–(7.10)) will involve a detailed study of a collection of functions of the elements k, Δ and σ^2. The first of these functions is represented by an expression that does not admit to a closed form but can be evaluated via simulation with arbitrarily high precision. Let

$$A(k, \Delta, \sigma^2) = E\left(\frac{1}{k-2+2T}\right), \tag{7.17}$$

where $T \sim P(\Delta/2\sigma^2)$. Using this notation, we may rewrite (7.16) as $B(\sigma_G^2) \leq 0$, where $B(\sigma_G^2)$ is the quadratic function of the prior variance σ_G^2 given by

$$\begin{aligned} B(\sigma_G^2) = {} & (\sigma_G^2)^2[(k-2)^2 A(k, \Delta, \sigma^2)] + \sigma_G^2[2\sigma^2((k-2)^2 A(k, \Delta, \sigma^2) - k)] \\ & + \sigma^2[\Delta - \sigma^2(k - (k-2)^2 A(k, \Delta, \sigma^2))]. \end{aligned} \tag{7.18}$$

The coefficient of $(\sigma_G^2)^2$ in (7.18) being positive, we may describe, for any fixed values of k, Δ and σ_G^2, the values of σ_G^2 for which (7.16) holds to be precisely the collection of positive numbers between the two roots of the equation $B(\sigma_G^2) = 0$. Further progress toward a solution requires that we check that the discriminant associated with this equation is positive. Vestrup (2001) showed that this discriminant reduces to

$$C(k, \Delta, \sigma^2) = 4\sigma^4[k^2 - (k-2)^2(k + \Delta/\sigma^2) A(k, \Delta, \sigma^2)]. \tag{7.19}$$

A proof of the following result may be found in Vestrup and Samaniego (2004b).

Theorem 7.1. *For any $k \geq 3$ and any fixed value of σ^2, the function $C(k, \Delta, \sigma^2)$ in* (7.19) *is a positive function which increases from $8k\sigma^2$ when $\Delta = 0$ to $16(k-1)\sigma^4$ as $\Delta \to \infty$.*

A deeper investigation into the class of "superior Bayes estimators" in this problem requires some additional ideas and notation. My intention is to provide the main

results obtained by Vestrup and Samaniego (2004a, 2004b), along with the motivation, intuition and interpretation of these results, but I will state them without proofs, as the details of these proofs are excruciatingly tedious (and only possible, really, when one of the authors desperately wants to finish his doctoral dissertation). Having obtained in Theorem 7.1 the fact that the quadratic function $B(\sigma_G^2)$ has two real roots, we will wish to ascertain that the interval bounded above and below by these roots contains an interval of positive values. The results below confirm this fact and provide, in addition, a number of other useful facts about this latter interval. Collectively, the theoretical results discussed here show that (i) for any fixed dimension k and choice of prior mean $\boldsymbol{\theta}_G$, there is an interval in the positive real line such that Bayesian shrinkage outperforms the James–Stein estimator if and only if the prior variance σ_G^2 lies in that interval, (ii) when the operational prior G is mean-correct, that is, when $E_G\boldsymbol{\theta} = E_{G_0}\boldsymbol{\theta}$, this interval can be specifically identified as $\left(0,\ \frac{2+\sqrt{2k}}{k-2}\sigma^2\right)$, (iii) for any fixed value of k and σ^2, both the lower bound L and the upper bound U on σ_G^2 which define the interval of Bayesian superiority tend to ∞ as $\Delta \to \infty$, (iv) for any fixed value of k and σ^2, the length of the interval which guarantees Bayesian superiority also tends to ∞ as $\Delta \to \infty$ and (v) the ratio $R = (U-L)/U$ decreases as $\Delta \to \infty$. We will return to the interpretation and discussion of these outcomes. First, we will define the additional notation that allows us to make these claims precise.

Let us now introduce the following additional functions of the triplet (k, Δ, σ^2). For $k \geq 3$, $\Delta \geq 0$ and $\sigma^2 \geq 0$, define

$U(k,\Delta,\sigma^2)$ = the larger root of the equation $B(\sigma_G^2) = 0$,
$H(k,\Delta,\sigma^2)$ = the smaller root of the equation $B(\sigma_G^2) = 0$,
$L(k,\Delta,\sigma^2) = \max(0,\ H(k,\Delta,\sigma^2))$,
$I(k,\Delta,\sigma^2) = U(k,\Delta,\sigma^2) - L(k,\Delta,\sigma^2)$, and
$R(k,\Delta,\sigma^2) = I(k,\Delta,\sigma^2)/U(k,\Delta,\sigma^2)$.

The following results concerning these functions were proven by Vestrup and Samaniego (2004a). The first result shows that the interval of values of σ_G^2 for which $\widehat{\boldsymbol{\theta}}_G$ dominates $\widehat{\boldsymbol{\theta}}_{JS,\boldsymbol{\theta}_G}$ is a nonempty interval of positive real numbers.

Theorem 7.2. *For every $k \geq 3$ and $\sigma^2 \geq 0$, the function $U(k,\Delta,\sigma^2)$ is increasing for $\Delta \in [0,\infty)$, with*

$$U(k,0,\sigma^2) = \frac{2+\sqrt{2k}}{k-2}\sigma^2 \quad \textit{and} \quad \lim_{\Delta\to\infty} U(k,\Delta,\sigma^2) = +\infty\,. \tag{7.20}$$

Theorem 7.3. *For every $k \geq 3$, $\Delta \geq 0$ and $\sigma^2 \geq 0$, the inequality*

$$r\left(G_0, \widehat{\boldsymbol{\theta}}_G\right) \leq r\left(G_0, \widehat{\boldsymbol{\theta}}_{JS,\boldsymbol{\theta}_G}\right) \tag{7.21}$$

is equivalent to the inequality

$$L(k,\Delta,\sigma^2) \leq \sigma_G^2 \leq U(k,\Delta,\sigma^2)\,, \tag{7.22}$$

where $0 \leq L < U \leq \infty$.

Theorem 7.4. *For every $k \geq 3$ and $\sigma^2 \geq 0$, the function $L(k,\Delta,\sigma^2)$ is nondecreasing for $\Delta \in [0,\infty)$ from 0 at $\Delta = 0$ to ∞ as $\Delta \to \infty$. Also, there exists a $\Delta^* = \Delta^*(k,\sigma^2) \in [0,\infty)$ such that $L(k,\Delta,\sigma^2) = 0$ if and only if $\Delta \in [0,\Delta^*]$.*

Theorem 7.5. *For every $k \geq 3$ and $\sigma^2 \geq 0$, the function $I(k,\Delta,\sigma^2)$ is increasing for $\Delta \in [0,\infty)$, with*

$$I(k,0,\sigma^2) = \frac{2+\sqrt{2k}}{k-2}\sigma^2 \quad \textit{and} \quad \lim_{\Delta\to\infty} I(k,\Delta,\sigma^2) = +\infty\,. \tag{7.23}$$

Theorem 7.6. *For every $k \geq 3$ and $\sigma^2 \geq 0$, the function $R(k,\Delta,\sigma^2)$ is decreasing for $\Delta \in [0,\infty)$, with*

$$R(k,0,\sigma^2) = 1 \quad \textit{and} \quad \lim_{\Delta\to\infty} R(k,\Delta,\sigma^2) = \frac{4\sqrt{k-1}}{k+2\sqrt{k-1}}\,. \tag{7.24}$$

Exercise 7.4. Verify the lower bound $\frac{2+\sqrt{2k}}{k-2}\sigma^2$ for U in Theorem 7.2 and the upper bound 1 for R in Theorem 7.6.

7.3 Discussion

The goal of this chapter is to compare Bayesian and frequentist shrinkage in the context of estimating a multivariate normal mean. While general formulations of this problem pose seemingly intractable analytical challenges, quite interesting answers and insights can be obtained in the special framework we have considered. For the sake of executing a concrete analysis, we have restricted attention to a formulation in which the true prior distribution G_0 is degenerate at a point $\boldsymbol{\theta}_0$, the normal sampling distribution is assumed to have the diagonal covariance matrix $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ (with σ^2 assumed known), and the operational prior G is assumed to be the standard conjugate prior, again with a diagonal covariance matrix $\boldsymbol{\Sigma}_G = \sigma_G^2\mathbf{I}$.

The problem studied here is admittedly constrained, but as has been mentioned, it is nonetheless of some interest in its own right. First, the restriction to a degenerate true prior is in line with the standard view of the target parameter in an estimation problem as simply a fixed unknown vector. While a more general G_0 might be called for in selected (though perhaps somewhat rare) circumstances, the restriction made here is an innocuous one in most real applications. Regarding the assumption of independence and a common variance for the components of the sampling distribution, we simply note that these are common assumptions in the multivariate context studied, patterned after the assumptions made in the foundational work by Robbins (1951) and by Stein (1956). The framework studied has ample historical precedents as well as many counterparts that are widely used in modern-day statistical analyses (such as the assumption of independent errors with common variance made in the theory and applications of the general linear model). But perhaps the most important reason that we should not be unduly concerned about the restrictions under which the analysis we have discussed was undertaken is the interesting conclusion to which this

work leads. As I will explain in more detail below, even in the restricted comparison described above, one finds that while there is, as expected, a threshold separating good and bad Bayesian procedures, the picture for the Bayesian is not nearly as rosy as it is in the one-parameter case examined in Chapter 5. In high-dimensional problems, the Bayesian has a strikingly narrower window for selecting an estimator that outperforms the James–Stein estimator. If that is the case in the framework encapsulated in (7.7)–(7.10), then the performance of Bayesian shrinkage in the general framework in (7.2)–(7.4) might well be expected, in most applications, to be inferior to that of the James–Stein estimator, that is, one can reasonably expect that Bayesian shrinkage will prevail over the frequentist alternative only if the prior distribution belongs to a relatively limited subspace of the space of available priors.

Let us consider some numerical examples that support the claim made in the last sentence of the preceding paragraph. When, for example, $k = 7$, one can ascertain that the bounds L and U for which $\sigma_G^2 \in (L,U)$ ensure the superiority of Bayesian shrinkage over frequentist (i.e., James–Stein) shrinkage obey, for large Δ, the approximate relationship

$$L(7,\Delta,\sigma^2) \approx (0.177)U(7,\Delta,\sigma^2)\,. \tag{7.25}$$

On the other hand, when $k = 101$, this relationship is

$$L(7,\Delta,\sigma^2) \approx (0.75)U(7,\Delta,\sigma^2)\,. \tag{7.26}$$

Clearly, the conditions for Bayesian superiority are strikingly reduced in high dimensions. In the latter instance, Bayesian shrinkage will be inferior to frequentist shrinkage if either $\sigma_G^2 \in (0,(0.75)U)$ or $\sigma_G^2 \in (U,\infty)$. These are not very attractive odds for the Bayesian! Further, let's consider the situation in which the Bayesian's operational prior is mean-correct (i.e., $\Delta = 0$), a case in which the Bayesian would expect to do quite well. Here, Theorems 7.2 and 7.3 imply that Bayesian shrinkage outperforms the James–Stein estimator (which is also assumed to shrink $\mathbf{X}$ toward $\boldsymbol{\theta}_0$, the true value of $\boldsymbol{\theta}$) if and only if

$$0 \le \sigma_G^2 \le \frac{2+\sqrt{2k}}{k-2}\sigma^2\,. \tag{7.27}$$

When $k = 8$, for example, Bayesian shrinkage outperforms James–Stein shrinkage if and only if

$$\sigma_G^2 \le \sigma^2\,. \tag{7.28}$$

To prevail over the James–Stein estimator in this situation, the Bayesian must be mean correct (no mean feat, if I may be permitted another pun), and must also explicitly exhibit a high level of confidence in the quality of his prior guess. Conservative prior modeling will not serve his interests well under these circumstances. This outcome lies in stark contrast to the comparison of Bayes and frequentist estimators in the one-parameter problems treated in Chapter 5 where the Bayes estimator with respect to a mean-correct operational prior will uniformly dominate the best frequentist estimator. It is perhaps worth noting that a mean-correct prior ensures uniform

Bayesian superiority over the sample mean $\overline{\mathbf{X}}$ in estimating the mean of a multivariate normal distribution (see Vestrup (2001)). In a simplified version of the problem, this latter result is stated in Exercise 7.3.

Lest the reader conclude from the above that one should not attempt Bayesian shrinkage in estimating a multivariate normal mean, let me comment on some of the positives of a Bayesian analysis in such problems. First, note that the inequality in (7.27) indicates that "sharp" prior information about the true value of $\boldsymbol{\theta}$ is clearly of value and can cede the advantage to the Bayesian over the frequentist. Further, Theorem 7.5 indicates that, even when the distance Δ between one's prior guess and the true value of the parameter is large, there is a fairly generous window of possible specifications of the prior variance σ_G^2 which ensure Bayesian superiority. Indeed, this latter theorem indicates that the length $U - L$ of the interval associated with Bayesian domination actually tends to ∞ as $\Delta \to \infty$. Clearly, the misguided Bayesian whose prior guess is quite distant from the truth still has the opportunity to outperform the frequentist, provided his prior variance is properly chosen. However, I would not wish to leave the impression that the latter choice is an easy one. Bayesian shrinkage will outperform James–Stein shrinkage in such situations only when the prior variance σ_G^2 is neither too small nor too large, making the specification of σ_G^2 a rather delicate matter.

What are the take-home lessons of this chapter? In a general and somewhat abstract sense, this chapter underscores the fact that Bayesian estimation of a high-dimensional parameter is a difficult enterprise — a fact that is not particularly surprising, given that the specification of a prior model which leads to inferences that are superior to notable frequentist alternatives is quite challenging. It is clear that even in the simplified comparison described here, where the prior model involves only $(k+1)$ parameters, the opportunity of inferior Bayesian inference is far from negligible. In the general framework specified in (7.2)–(7.4), that opportunity is no doubt larger. The situation becomes all the more imposing once we transition to the real problem one would typically face in practice, the problem of estimating the mean of a normal distribution with an unknown covariance matrix $\boldsymbol{\Sigma}$. What undoubtedly remains true in all versions of this problem is the fact that there is a threshold separating good priors from bad priors. From the investigation above, it appears that the relative size of the collection of good priors among the space of available priors grows smaller as the dimension of the estimation problem grows, and the identification of a prior distribution which will lead to better inferences than a good frequentist estimator becomes increasingly difficult. This suggests that Bayesian estimators of vector-valued parameters must be used with considerable care, as the importance of "good prior modeling" becomes substantially magnified as the dimension of the problem of interest grows.

8

Comparing Bayesian and Frequentist Estimators under Asymmetric Loss

8.1 Introduction

While many estimation problems involving one or more parameters are treated using a symmetric loss function which gives equal weight to estimation errors that are the same "distance" from the true parameter value, there are clearly problems in which estimation errors in a particular direction are considered more serious than errors in another direction. In univariate problems, it may well be the case that overestimation has potential repercussions that underestimation does not (or, of course, vice versa). For example, Varian (1975) motivated the use of asymmetric loss functions in estimation problems arising in real estate assessment, where the overestimation of a property's value might cause it to remain on the market unsold for an extended period, ultimately costing the seller inordinate and unnecessary expenses. The estimation of peak water flow in the construction of dams or levies clearly has asymmetric consequences; overestimation might lead to increased construction costs while underestimation might lead to the much more serious consequence of subsequent overflows which can seriously threaten lives and property in adjacent communities. Further examples of contexts requiring the asymmetric treatment of estimation errors are given in papers by Shao and Chow (1991), who treat estimation problems regarding release dates of certain pharmaceutical products, by Thompson and Basu (1996), who treat asymmetric problems arising in reliability and by Zellner and Palm (1974), who considers a variety of problems in the area of econometrics.

In this chapter, I will consider extensions to the comparison of Bayes and frequentist estimators to the case of asymmetric loss. While there are a variety of possible loss functions one could consider in such a study, I will restrict attention to the Linex loss function introduced in Section 1.3. This is the formulation of asymmetry that appears to be the most widely used, and it represents, as well, a loss function under which the associated estimation problems are manageable analytically. The reader will recall the form of the Linex loss function from our earlier discussion:

$$L(\theta, \widehat{\theta}) = \mathrm{e}^{c(\widehat{\theta}-\theta)} - c(\widehat{\theta}-\theta) - 1\,, \tag{8.1}$$

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_8,

where a is a fixed and known constant. As noted earlier, the Linex loss function achieves its minimum 0 when $\widehat{\theta} = \theta$ and is a convex function of the difference $\Delta = (\widehat{\theta} - \theta) \in (-\infty, \infty)$, being decreasing for $\Delta \in (-\infty, 0)$ and increasing for $\Delta \in (0, \infty)$. When a is positive, Linex loss grows exponentially in positive Δ, but behaves approximately linearly for negative values of Δ. Thus, when $c > 0$, Linex loss imposes a substantial penalty for overestimation, with the opposite being true when $c < 0$. Under Linex loss, the form of the Bayes estimator of a scalar parameter θ with respect to the prior G is given in Theorem 3.4.

The framework to be studied in succeeding sections resembles that found in Chapters 5 and 7. Specifically, my aim will be to identify, in two quite different settings, the circumstances in which the Bayes estimator of a parameter of interest, with respect to a fixed "operational" prior G, outperforms the classical (frequentist) estimator of that parameter, with the Bayes risk of each estimator relative to the "true prior distribution" G_0 being the basis for comparison. This of course defines a threshold problem in the scenarios to be investigated. In the problems to be examined here, the classical estimator on which we will focus is the maximum likelihood estimator of the unknown parameter.

Section 8.2 is dedicated to the comparison of Bayes estimators and the MLE of the mean of a k-variate normal distribution (for any fixed $k \geq 1$) under Linex loss. Our main goal is to determine the extent to which the main conclusions of Chapters 5 and 7 regarding the characteristics of "good" Bayes procedures under squared error loss hold true when the loss function is asymmetric. For $k \geq 2$, a multicomponent version of Linex loss, namely,

$$L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{k} L(\theta_i, \widehat{\theta}_i) \,, \tag{8.2}$$

is employed. In Section 8.3, we investigate similar questions in a linear regression setting in which the target parameter is a predetermined linear combination of regression coefficients. The final section of the chapter summarizes the main findings in these investigations, elaborates on the need for further research regarding the comparison of Bayes estimators with alternative frequentist estimators in the two problems treated here and discusses the potential for generalizations of these results to other parametric paradigms. The results presented in this chapter are largely drawn from Bhattacharya, Samaniego and Vestrup (2002), referred to hereafter as BSV (2002). The reader is referred to that paper for the various technical details that are omitted from the overview of that work presented here.

8.2 Estimating the mean of a normal distribution under Linex loss

While, under squared error loss, the sample mean $\overline{\mathbf{X}}$ of a random sample from a multivariate normal distribution $\mathcal{N}_k(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ is inadmissible as an estimator of the population mean in dimension $k \geq 3$, the situation is considerably less clear when its

performance is examined under other loss functions. For most asymmetric loss functions, the admissibility of the sample mean in an arbitrary dimension k is an open question. Thus, based on its asymptotic optimality, its interpretive value and its simple closed form, the maximum likelihood estimator $\overline{\mathbf{X}}$ of the population mean tends to be the frequentist estimator of choice in such problems. In this section, the comparison made will focus on the performance of Bayes estimators with respect to standard conjugate priors relative to the performance of the MLE. Performance is measured, as in earlier chapters, by the Bayes risk of a given estimator with respect to a hypothesized true prior G_0. Since, under asymmetric loss functions, there are in fact alternative frequentist estimators which can be identified as legitimate contenders to the MLE in estimating a normal mean, we will address the issue of threshold problems relative to other frequentist estimators in the chapter's concluding section. In Sections 8.2 and 8.3, we restrict attention to the MLE while recognizing that the associated threshold problem is just one of several such problems that one might wish to consider.

As in earlier chapters, we will be interested here in three particular scenarios in which this comparison is of special significance. In the most general scenario, where the mean $\boldsymbol{\theta}_G$ of the operational prior and the mean $\boldsymbol{\theta}_{G_0}$ of the true prior are arbitrary and may be quite different, we will be interested in whether Bayes estimators will be superior to the MLE when the operational prior is sufficiently diffuse. Second, we will be interested in the circumstances in which a mean-correct operational prior will provide performance for the Bayes estimator that is superior to that of the MLE. Third, we will wish to determine what can be said about the comparative performance of a Bayes estimator and the MLE in the case of primary practical interest in which the true prior G_0 is degenerate at a point $\boldsymbol{\theta}_{G_0}$. The results developed in this section provide definitive answers to these three questions.

The modeling assumptions under which we will proceed resemble those of Chapter 7, although we will permit slightly more generality than in conditions (7.6)–(7.8). Specifically, we make the following explicit assumptions:

$$G_0 : \boldsymbol{\theta} \sim \mathcal{N}_k\left(\boldsymbol{\theta}_{G_0},\ \sigma^2_{G_0}\mathbf{I}\right) \tag{8.3}$$

$$G : \boldsymbol{\theta} \sim \mathcal{N}_k\left(\boldsymbol{\theta}_{G},\ \sigma^2_{G}\mathbf{I}\right) \tag{8.4}$$

$$F_{\mathbf{X}|\boldsymbol{\theta}} : \mathbf{X}|\boldsymbol{\theta} \sim \mathcal{N}_k\left(\boldsymbol{\theta},\ \sigma^2\mathbf{I}\right) \tag{8.5}$$

where σ^2 is a known constant, and the loss function to be utilized is the natural multivariate extension of Varian's Linex loss, that is,

$$L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{k} \left[\mathrm{e}^{c_i(\widehat{\theta}_i - \theta_i)} - c_i(\widehat{\theta}_i - \theta_i) - 1 \right] . \tag{8.6}$$

In this section, the Bayes estimator with respect to the operational prior G will be denoted by $\widehat{\boldsymbol{\theta}}^G = \left(\widehat{\theta}_1^G, \widehat{\theta}_2^G, \ldots, \widehat{\theta}_k^G\right)$. It may be inferred from Theorem 3.4 that

$$\widehat{\theta}_i^G = -\frac{1}{c_i} \log E_{\boldsymbol{\theta}|\mathbf{X}=\mathbf{x}}\{\mathrm{e}^{-c_i\theta_i}\} \quad \text{for } i = 1, \ldots, k . \tag{8.7}$$

It is easy to verify that, under assumptions (8.4) and (8.5), the posterior distribution of $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta}|\mathbf{X}=\mathbf{x}\sim\mathscr{N}_k\left(\frac{\sigma_G^2}{\sigma_G^2+\sigma^2}\mathbf{x}+\frac{\sigma^2}{\sigma_G^2+\sigma^2}\boldsymbol{\theta}_G,\ \frac{\sigma_G^2\sigma^2}{\sigma_G^2+\sigma^2}\mathbf{I}\right). \tag{8.8}$$

Thus, for $i=1,2,\ldots,k$, we have that

$$\theta_i|\mathbf{X}=\mathbf{x}\sim\mathscr{N}_k\left(\frac{\sigma_G^2}{\sigma_G^2+\sigma^2}x_i+\frac{\sigma^2}{\sigma_G^2+\sigma^2}\theta_i^G,\ \frac{\sigma_G^2\sigma^2}{\sigma_G^2+\sigma^2}\right). \tag{8.9}$$

From (8.9), one may easily calculate

$$E_{\theta_i|\mathbf{X}=\mathbf{x}}\{\mathrm{e}^{-c_i\theta_i}\}=\exp\left\{-c_i\left[\frac{\sigma_G^2}{\sigma_G^2+\sigma^2}x_i+\frac{\sigma^2}{\sigma_G^2+\sigma^2}\theta_i^G\right]+\frac{a_i^2\sigma_G^2\sigma^2}{2(\sigma_G^2+\sigma^2)}\right\}. \tag{8.10}$$

It follows that, under the generalized Linex loss function in (8.6), the Bayes estimator with respect to the operational prior G is

$$\widehat{\boldsymbol{\theta}}^G=\frac{\sigma_G^2}{\sigma_G^2+\sigma^2}\mathbf{x}+\frac{\sigma^2}{\sigma_G^2+\sigma^2}\boldsymbol{\theta}^G-\frac{\sigma_G^2\sigma^2}{2(\sigma_G^2+\sigma^2)}\mathbf{c}\,. \tag{8.11}$$

The risk function of the Bayes estimator in (8.11) is calculated in a practically useful form by BSV (2002), and its Bayes risk with respect to the true prior G_0 is shown to be

$$r\left(G_0,\widehat{\boldsymbol{\theta}}^G\right)=\sum_{i=1}^k\left[\exp\left\{t_i\Delta_i+\tfrac{1}{2}t_i^2(\sigma_{G_0}^2-\sigma_G^2)\right\}-t_i\Delta_i+\tfrac{1}{2}c_it_i\sigma_G^2-1\right], \tag{8.12}$$

where $\exp(A)=\mathrm{e}^A$ and, for $i=1,2,\ldots,k$,

$$\Delta_i=\theta_i^G-\theta_i^{G_0}\quad\text{and}\quad t_i=\frac{c_i\sigma^2}{\sigma_G^2+\sigma^2}\,. \tag{8.13}$$

We wish to compare the expression in (8.12) to the Bayes risk of the MLE of $\boldsymbol{\theta}$. Since the risk function of the MLE is constant, its Bayes risk is easily found to be

$$r(G_0,\overline{\mathbf{X}})=\sum_{i=1}^k\left[\mathrm{e}^{\frac{1}{2}c_i^2\sigma^2}-1\right]. \tag{8.14}$$

We are now in the position to prove

Theorem 8.1. *For arbitrary values of the true and operational prior means $\boldsymbol{\theta}_{G_0}$ and $\boldsymbol{\theta}_G$, the Bayes estimator $\widehat{\boldsymbol{\theta}}^G$ is superior to the MLE $\overline{\mathbf{X}}$ if the operational prior is sufficiently diffuse.*

Proof. Note that, from (8.13), we have, as $\sigma_G^2 \to \infty$, that $t_i \to 0$, $t_i\sigma_G^2 \to c_i\sigma^2$ and $t_i^2\sigma_G^2 \to 0$. Using these facts and the inequality $x < e^x - 1$, which holds for all real x, it follows that

$$\lim_{\sigma_G^2 \to \infty} r\left(G_0, \widehat{\boldsymbol{\theta}}^G\right) = \sum_{i=1}^{k} \frac{c_i^2\sigma^2}{2} < \sum_{i=1}^{k} \left[e^{\frac{1}{2}c_i^2\sigma^2} - 1\right] = r(G_0, \overline{\mathbf{X}}) \,, \tag{8.15}$$

an inequality which confirms that $\widehat{\boldsymbol{\theta}}^G$ is superior to $\overline{\mathbf{X}}$ if σ_G^2 is sufficiently large. ■

The implications of the mean-correctness of the operational prior G are as follows:

Theorem 8.2. *If the prior G is mean correct, that is, if $\boldsymbol{\theta}_G = \boldsymbol{\theta}_{G_0}$, the Bayes estimator $\widehat{\boldsymbol{\theta}}^G$ is superior to the MLE $\overline{\mathbf{X}}$ whenever $\sigma_G^2 > \sigma_{G_0}^2$.*

Proof. Under the assumptions $\boldsymbol{\theta}_G = \boldsymbol{\theta}_{G_0}$ and $\sigma_G^2 > \sigma_{G_0}^2$, we may write

$$\begin{aligned}
r\left(G_0, \widehat{\boldsymbol{\theta}}^G\right) &= \sum_{i=1}^{k} \left[e^{\frac{1}{2}t_i^2(\sigma_{G_0}^2 - \sigma_G^2)} + \tfrac{1}{2}c_i t_i \sigma_G^2 - 1\right] \\
&< \sum_{i=1}^{k} \left[1 + \tfrac{1}{2}c_i t_i \sigma_G^2 - 1\right] \\
&= \sum_{i=1}^{k} \left[\tfrac{1}{2}c_i t_i \sigma_G^2\right] \\
&< \sum_{i=1}^{k} \left[\tfrac{1}{2}c_i^2\sigma^2\right] \qquad \text{(by (8.13))} \\
&< \sum_{i=1}^{k} \left[e^{\frac{1}{2}c_i^2\sigma^2} - 1\right] \\
&= r\left(G_0, \overline{\mathbf{X}}\right) \,. \qquad \blacksquare
\end{aligned} \tag{8.16}$$

Corollary 8.1. *If the conditions of Theorem 8.2 hold, and if the true prior distribution is degenerate at the point $\boldsymbol{\theta}_{G_0}$, that is, $\sigma_{G_0}^2 = 0$, then for all values of $\sigma_G^2 > 0$,*

$$r\left(G_0, \widehat{\boldsymbol{\theta}}^G\right) < r\left(G_0, \overline{\mathbf{X}}\right) \,. \tag{8.17}$$

This section constitutes a complementary treatment, under an asymmetric loss function, of the estimation problem considered in Chapter 7. Because the dimension k of the problem is permitted to be an arbitrary positive number, the results provide comparisons, as well, to the univariate problem of estimating a normal mean in a context complementary to that treated in Chapter 5. Without benefit of the study in this section, one may have conjectured that the solutions to the threshold problem, as treated in earlier chapters, are somehow driven by the assumed symmetry (perhaps even by the precise formulation of these problems as estimation under squared

error loss), and that similar phenomena might not occur when such assumptions are relaxed. The results above serve to demonstrate that the "threshold" phenomena encountered under a symmetric loss structure surface in very similar ways in asymmetric settings. Further, the general characteristics of the Bayes estimators which stand to be superior to the maximum likelihood estimator are essentially the same under the symmetric and asymmetric loss functions we have considered. In both circumstances, the Bayesian approach involving careful prior elicitation, with a focus on approximate mean correctness, together with conservative prior modeling, in the form of operational priors that are "reasonably diffuse," appear to result in a reliable recipe for good performance relative to that of the MLE. In Section 8.4, alternative versions of the threshold problem are discussed.

Exercise 8.1. Derive the Bayes estimator whose general form is given in equation (8.7).

Exercise 8.2. Verify that, under assumptions (8.4) and (8.5), the posterior distribution of $\boldsymbol{\theta}$ is given by the expression in (8.8).

Exercise 8.3. Derive the Bayes risk expression in (8.12).

Exercise 8.4. Show that, under condition (8.5), the risk function R of the MLE of a multivariate normal mean is the constant on the RHS of equation (8.14).

8.3 Estimating a linear combination of regression parameters

In this section, we will treat an estimation problem that has received some attention in the econometrics literature. The problem is typically treated under the standard assumptions of the general linear model. We will be interested in both the Bayesian and classical treatment of the problem, and we thus specify below the models that we have heretofore referred to as the operational and the true prior distribution of the unknown parameters. The relevant models are given by

$$G_0 : \boldsymbol{\beta} \sim \mathcal{N}_p \left(\boldsymbol{\beta}_{G_0}, \boldsymbol{\Sigma}_{G_0} \right) , \tag{8.18}$$

$$G : \boldsymbol{\beta} \sim \mathcal{N}_p \left(\boldsymbol{\beta}_{G}, \boldsymbol{\Sigma}_{G} \right) , \tag{8.19}$$

$$F_{\mathbf{Y}|\boldsymbol{\beta}} : \mathbf{Y}|\boldsymbol{\beta} \sim \mathcal{N}_k \left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I} \right) , \tag{8.20}$$

where $k > p, \sigma^2$ is, for convenience, assumed to be a known constant and $\mathbf{X}$ is the fixed and known $k \times p$ "design matrix" of full rank p. The problem of interest is the estimation of a linear function of the elements of the vector $\boldsymbol{\beta}$, that is, $\theta = \mathbf{w}'\boldsymbol{\beta}$, where $\mathbf{w}$ is a known $p \times 1$ vector. We note that the problem posed above subsumes the problem of estimating individual regression parameters $\{\beta_i, i = 1, \ldots, p\}$, for which the vector $\mathbf{w}$ is taken to be a unit vector (with one element equal to one), as well as the common problem of estimating differences between two such parameters. Here, the vector $\mathbf{w}$ can be chosen arbitrarily, so the problem to be treated is substantially

more general than these particular examples. Since the problem entails the estimation of a scalar parameter θ, we will use the univariate form of the Linex loss function specified in (8.1).

As indicated above, it is our intention to compare the performance of the Bayes estimator $\widehat{\theta}^G$ of θ wrt the operational prior G with that of the maximum likelihood estimator $\widehat{\theta}$ of θ. We begin by identifying the Bayes estimator which, from earlier discussions, is easily shown to be

$$\widehat{\theta}^G = -\frac{1}{c}\log E_{\mathbf{w}'\boldsymbol{\beta}|\mathbf{Y}=\mathbf{y}}\{\mathrm{e}^{-c\mathbf{w}'\boldsymbol{\beta}}\}\,. \tag{8.21}$$

Using well-known normal theory and appropriate matrix algebra, the posterior distribution of $\mathbf{w}'\boldsymbol{\beta}|\mathbf{Y}=\mathbf{y}$ may be identified as a univariate normal distribution with mean

$$\mu = \mathbf{w}'\boldsymbol{\Sigma}_G\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}_G\mathbf{X}'+\sigma^2\mathbf{I}_k)^{-1}\mathbf{y}+\mathbf{w}'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}_G\mathbf{X}'+\sigma^2\mathbf{I}_k)^{-1}\mathbf{X}\boldsymbol{\beta}_G \tag{8.22}$$

and variance

$$\sigma^2 = \mathbf{w}'\boldsymbol{\Sigma}_G\mathbf{w}-\mathbf{w}'\boldsymbol{\Sigma}_G\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}_G\mathbf{X}'+\sigma^2\mathbf{I}_k)^{-1}\mathbf{X}\boldsymbol{\Sigma}_G\mathbf{w}\,. \tag{8.23}$$

From the posterior distribution identified above, we can obtain the Bayes estimator $\widehat{\theta}^G$ in (8.21) in the following closed form:

$$\widehat{\theta}^G = \boldsymbol{\lambda}_1\mathbf{y}+\boldsymbol{\lambda}_2\boldsymbol{\beta}_G-\frac{c}{2}\lambda_3\,, \tag{8.24}$$

where $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$ and λ_3 are, respectively, the $1\times k$ matrix, the $1\times p$ matrix and the scalar given by

$$\boldsymbol{\lambda}_1 = \mathbf{w}'\boldsymbol{\Sigma}_G\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}_G\mathbf{X}'+\sigma^2\mathbf{I}_k)^{-1}\,, \tag{8.25}$$

$$\boldsymbol{\lambda}_2 = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}_G\mathbf{X}'+\sigma^2\mathbf{I}_k)^{-1}\mathbf{X} \tag{8.26}$$

and

$$\lambda_3 = \mathbf{w}'\boldsymbol{\Sigma}_G\mathbf{w}-\mathbf{w}'\boldsymbol{\Sigma}_G\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}_G\mathbf{X}'+\sigma^2\mathbf{I}_k)^{-1}\mathbf{X}\boldsymbol{\Sigma}_G\mathbf{w}\,. \tag{8.27}$$

The development of an expression for the Bayes risk $r(G_0,\widehat{\theta}^G)$ in the precise form that proves useful in the comparisons of interest here is an arduous endeavor, one that requires a substantial amount of matrix manipulation and fairly subtle analytical work. For the details of this development, the reader is referred to BSV (2002). A key element of the argument is the alternative expression for a matrix that occurs prominently in analytical work involving the λs in (8.25)–(8.27):

$$\mathbf{X}'(\mathbf{X}\boldsymbol{\Sigma}_G\mathbf{X}'+\sigma^2\mathbf{I}_k)^{-1}\mathbf{X} = (\sigma^2(\mathbf{X}'\mathbf{X})^{-1}+\boldsymbol{\Sigma}_G)^{-1}\,. \tag{8.28}$$

This identity is proven as "Matrix Lemma 3" in BSV (2002, p.251). Using (8.28) and additional algebraic argumentation, the following expression for the desired Bayes risk of $\widehat{\theta}^G$ is obtained:

$$r\left(G_0,\widehat{\theta}^G\right) = \mathrm{e}^{c\gamma_1}\mathrm{e}^{c^2\gamma_2/2}-c\gamma_1+\frac{c^2}{2}\mathbf{w}'\boldsymbol{\gamma}_3\mathbf{w}-1\,, \tag{8.29}$$

where γ_1 and γ_2 are the scalars, and $\boldsymbol{\gamma}_3$ is the $p \times p$ matrix, given below:

$$\gamma_1 = \mathbf{w}'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \boldsymbol{\Sigma}_G)^{-1}(\boldsymbol{\beta}_G - \boldsymbol{\beta}_0)\,, \tag{8.30}$$

$$\gamma_2 = \boldsymbol{\lambda}_2\boldsymbol{\Sigma}_{G_0}\boldsymbol{\lambda}_2' + \sigma^2\boldsymbol{\lambda}_1\boldsymbol{\lambda}_1' - \lambda_3 \tag{8.31}$$

and

$$\boldsymbol{\gamma}_3 = \boldsymbol{\Sigma}_G - \boldsymbol{\Sigma}_G(\sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \boldsymbol{\Sigma}_G)^{-1}\boldsymbol{\Sigma}_G\,. \tag{8.32}$$

We now turn to similar developments for the maximum likelihood estimator of θ. The MLE of θ is $\widehat{\theta} = \mathbf{w}'\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is the least squares estimator of $\boldsymbol{\beta}$. It is easy to verify that the risk function of $\widehat{\theta}$ is a constant which does not depend on $\boldsymbol{\beta}$. This leads to the Bayes risk expression

$$r\left(G_0, \widehat{\theta}\right) = \mathrm{e}^{\frac{c^2\sigma^2}{2}\mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}} - 1\,. \tag{8.33}$$

We are now in a position to examine the proposition that the Bayes estimator of θ will outperform the MLE if the operational prior is sufficiently diffuse. To treat this question, we need to be able to quantify the growth in diffusion of a covariance matrix. In the univariate case, it suffices to postulate that $\sigma_G^2 \to \infty$. If we were considering only diagonal covariance matrices $\boldsymbol{\Sigma}_G$, the univariate case would be properly generalized if the variance of every component of the random vector $\boldsymbol{\beta}$ tended to infinity. In the general case, we will say that a sequence of positive definite matrices are growing more diffuse if the three conditions given in the following definition are satisfied.

Definition 8.1. *Let* $\mathbf{A}^{(n)} = \left[a_{ij}^{(n)}\right]$, $n = 1, 2, \ldots$, *be a sequence of* $p \times p$ *positive definite matrices. If the elements of* $\{\mathbf{A}^{(n)}\}$ *increase in magnitude in such a way that, as* $n \to \infty$,

(a) $\min_i a_{ii}^{(n)} \to \infty$ *for all* i ,
(b) $\max_{i \neq j} a_{ij}^{(n)} = o\left(\min_i a_{ii}^{(n)}\right)$, *and*
(c) all diagonal elements $a_{ii}^{(n)}$ *tend to* ∞ *at the same rate (in* n*)* ,

then the sequence $\{\mathbf{A}^{(n)}\}$ *is said to be* growing more diffuse. *If the sequence* $\{\mathbf{A}^{(n)}\}$ *is growing more diffuse, we will write* $\mathbf{A}^{(n)} \to \infty$.

Given the definition above, the following result is established by BSV (2002). It confirms that in the problem of estimating a linear combination of regression parameters, the Bayes estimator with respect to a sufficiently diffuse operational prior will outperform the MLE of that target parameter.

Theorem 8.3. *Regardless of the size of* $\|\boldsymbol{\beta}_G - \boldsymbol{\beta}_{G_0}\|^2$, *the Bayes risks of the Bayes estimator* $\widehat{\theta}^G$ *and the MLE* $\widehat{\theta}$ *of* $\theta = \mathbf{w}'\boldsymbol{\beta}$ *under Linex loss will satisfy the inequality*

$$\lim_{\boldsymbol{\Sigma}_G^{(n)} \to \infty} r\left(G_0,\, \widehat{\theta}^G\right) < r(G_0,\, \widehat{\theta})\,, \tag{8.34}$$

where $\left\{\boldsymbol{\Sigma}_G^{(n)}\right\}$ *are variance-covariance matrices satisfying conditions (a)–(c) of Definition 8.1.*

Partial Proof: The proof of (8.34) is based on the facts, established in BSV (2002), that $\lim_{\mathbf{\Sigma}_G \to \infty} \gamma_i = 0$ for $i = 1, 2$ and $\lim_{\mathbf{\Sigma}_G \to \infty} \gamma_3 = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$. Applying these facts to the Bayes risk expression in (8.29) leads to the identity

$$\lim_{\mathbf{\Sigma}_G \to \infty} r\left(G_0,\, \widehat{\theta}^G\right) = \frac{c^2\sigma^2}{2} \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}\,, \tag{8.35}$$

which, by virtue of the fact that $x < \mathrm{e}^x - 1$ for all $x > 0$, is smaller than

$$\exp\left\{\frac{c^2\sigma^2}{2} \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}\right\} - 1 = r\left(G_0, \widehat{\theta}\right)\,. \qquad \blacksquare \tag{8.36}$$

The impact of mean correctness and the degeneracy of the true prior distribution are summarized in the following.

Theorem 8.4. *If the operational prior G is mean-correct (that is, $\boldsymbol{\beta}_G = \boldsymbol{\beta}_{G_0}$) and the true prior G_0 is degenerate (that is, $\mathbf{\Sigma}_{G_0} = \mathbf{0}$), then the Bayes estimator $\widehat{\theta}^G$ is uniformly superior to the MLE $\widehat{\theta}$ as an estimator of $\theta = \mathbf{w}'\boldsymbol{\beta}$ under Linex loss, that is, for any operational covariance matrix $\mathbf{\Sigma}_G$,*

$$r\left(G_0,\, \widehat{\theta}^G\right) < r\left(G_0, \widehat{\theta}\right)\,. \tag{8.37}$$

Exercise 8.5. Verify that the posterior distribution of $\mathbf{w}'\boldsymbol{\beta}$, given $\mathbf{Y} = \mathbf{y}$, is the univariate normal distribution with mean μ and variance σ^2 given in (8.22) and (8.23).

Exercise 8.6. Show that the MLE $\widehat{\theta}$ of $\theta = \mathbf{w}'\boldsymbol{\beta}$ is an equalizer rule with constant risk given by

$$R\left(\boldsymbol{\beta}, \widehat{\theta}\right) = \exp\left\{\frac{c^2\sigma^2}{2} \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}\right\} - 1\,.$$

Exercise 8.7. Construct a sequence of 2×2 positive definite matrices $\{\mathbf{A}_n\}$ that are growing more diffuse in the sense of Definition 8.1.

8.4 Discussion

The goal of this chapter is to investigate the extent to which the characteristics of Bayes estimators found to be "superior" to the frequentist estimator of choice in symmetric estimation problems (in which squared error loss was used) carry over into problems in which it is deemed more appropriate to use an asymmetric loss function. We have demonstrated that, in the specific problems examined here, the threshold problem is well-defined and its treatment is analytically tractable. More importantly, the solutions obtained here are remarkably similar to those obtained earlier, giving strong support to the proposition that the class of threshold problems to which solutions of this sort obtain is quite broad. In this section, we will briefly review our findings in our treatment of estimation under the Linex loss function, and

also provide some indication of the potential range of possible applications beyond those we have treated above.

In Section 8.2, we presented evidence that under asymmetric loss (specifically, under the Linex loss function and its multivariate generalizations), the Bayesian will have the advantage over the frequentist using the maximum likelihood estimator of the mean $\boldsymbol{\mu}$ of a k-dimensional normal distribution whenever his operational prior is sufficiently diffuse, regardless of the distance between his prior guess $\boldsymbol{\theta}_G$ and the mean $\boldsymbol{\theta}_{G_0}$ of the true prior. This advantage is accentuated under the assumption that the operational prior is mean-correct, and, in this latter case, it is universal (relative to the Bayesian's choice of the variance σ_G^2 of his operational prior) in the important scenario in which the true prior is degenerate at the point $\boldsymbol{\theta}_{G_0}$. In the univariate case, that is, when $k = 1$, these results constitute direct analogs to the results obtained in Chapter 5. In higher dimensions, a mean-correct prior does not guarantee Bayesian superiority, but the domain of Bayesian superiority is large (namely, whenever $\sigma_G^2 > \sigma_{G_0}^2$). In the end, we see that the threshold problem is well-defined and analytically tractable for an important class of models and asymmetric loss functions. The full extent of the generality of our conclusions is at present unknown. Clearly, there will be some versions of the threshold problem that can only be attacked using numerical methods or simulation. But the results above, coupled with those in earlier chapters, provide strong evidence that the general conclusions drawn from the threshold problems we have examined are quite robust. It appears that, in an interesting variety of settings, there is a reasonably large subclass of prior distributions which provide the Bayesian the advantage in problems of point estimation.

Section 8.3 treats an estimation problem with a considerable amount of structure. In the context of the multiple linear regression model where $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with i.i.d. normally distributed errors, the problem of estimating a fixed linear combination $\theta = \mathbf{w}'\boldsymbol{\beta}$ of regression parameters is addressed. Our findings are similar to those above. Regardless of the extent of prior misspecification of the true value of $\boldsymbol{\beta}$, the Bayes estimator $\widehat{\theta}^G$ of θ will outperform the MLE $\widehat{\theta}$ of θ under Linex loss, provided that the operational prior G is sufficiently diffuse. The notion of "growing diffusion" in a sequence of matrices was made explicit in Definition 8.1. It was also shown that, under Linex loss, the Bayes estimator $\widehat{\theta}^G$ is uniformly superior to the MLE $\widehat{\theta}$ as an estimator of $\theta = \mathbf{w}'\boldsymbol{\beta}$ when the operational prior G is mean correct and the true prior G_0 is degenerate. We thus find, again, that priors that are approximately mean-correct and reasonably diffuse, that is, priors which are well-calibrated but reflect conservative prior modeling, generally provide the Bayesian with an advantage in this estimation problem over a frequentist using the maximum likelihood estimator.

The reader will surely have noticed that, while the assumption of symmetry of the loss function was relaxed in this chapter, the estimation problems considered in the chapter all assume that the data follows a normal model. Aha! Maybe that's the trick! The problem was broadened in one direction but narrowed in another. So how general can we really expect the types of solutions seen here to be?

I offer the following example, drawn from BSV (2002), as evidence that there appears to be a general pattern underlying the problems considered above, and that one can expect to see solutions of the threshold problem with the same flavor in other

estimation problems with asymmetric loss. Suppose that we have a random sample from a Poisson distribution with mean θ, that is, $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathscr{P}(\theta)$. The conjugate prior for θ is the gamma distribution $\Gamma(\alpha, \beta)$. It is useful, in comparing the Bayes risks of the Bayes estimator $\widehat{\theta}^G$ of θ to that of the MLE $\widehat{\theta}$, to parametrize the operational and true priors of θ in terms of the means θ_G and θ_{G_0} of these distributions. We thus specify these models as

$$G_0 : \ \theta \sim \Gamma(\theta_{G_0}/\beta_{G_0}, \ \beta_{G_0}) \tag{8.38}$$

and

$$G : \ \theta \sim \Gamma(\theta_G/\beta_G, \ \beta_G) \ . \tag{8.39}$$

It is then possible to represent a true prior distribution that is degenerate at the value θ_{G_0} as equivalent to the true prior that results from allowing the parameter β_{G_0} to tend to 0. Similarly, we may represent an operational prior with mean θ_G which becomes increasingly diffuse by the process of allowing the parameter β_G to grow to ∞. With such a parametrization, it is established by BSV (2002) that

$$\lim_{\sigma_{G_0} \to 0, \sigma_G \to \infty} r\left(G_0, \widehat{\theta}^G\right) < \lim_{\sigma_{G_0} \to 0} r\left(G_0, \widehat{\theta}\right) \ . \tag{8.40}$$

The inequality in (8.40) indicates that, if the true prior distribution G_0 is degenerate at θ_{G_0}, then, under the Linex loss function, the Bayes estimator with respect to the operational prior with mean θ_G will outperform the MLE $\widehat{\theta}$ as an estimator of θ, irrespective of the distance between the mean of the operational prior and the true value of the parameter θ, provided that the operational prior G is sufficiently diffuse.

As mentioned in Section 8.2, the status of the MLE $\overline{\mathbf{X}}$ as the frequentist estimator of choice when the loss function is asymmetric is open to question. Zellner (1986) treated this matter in considerable detail. Among his findings was the interesting result that, under the Linex loss function given in (8.2), the mean $\overline{X}$ of a univariate sample from a normal population with known variance σ^2 was inadmissible as an estimator of the population mean μ. The frequentist estimator

$$\widetilde{\theta}_1 = \overline{X} - \frac{c\sigma^2}{2n} \ , \tag{8.41}$$

an estimator which may be derived as the generalized Bayes estimator with respect to Lebesgue measure, has a uniformly smaller risk function than $\overline{X}$. The superiority of the latter estimator, or something similar to it, might have been predicted in a problem in which the overestimation of μ is severely penalized. It is clear that the threshold problem comparing the performance of $\widetilde{\theta}_1$ to that of Bayes estimators with respect to proper priors would be of interest in the one-dimensional case. Further, to my knowledge, the admissibility of $\widetilde{\theta}_1$ is still an open question, suggesting that other frequentist estimators might well be considered when $k = 1$. These questions are worthy of investigation. The results in Section 8.2, as they apply to the univariate problem, show that a certain subclass of Bayes estimators competes well with the MLE. The domain of Bayesian superiority will, of course, shrink somewhat in

threshold problems involving better frequentist alternatives. It seems reasonable to conjecture that the "shape" of the domain will resemble that identified in Section 8.2, though the boundaries will surely change. In higher dimensions (i.e., for $k > 1$), the situation is considerably fuzzier, though it seems likely that other versions of the threshold problem than that considered in Section 8.2 will also be of interest. In this regard, the James–Stein estimator which shrinks $\overline{\mathbf{X}}$ toward zero would seem to be a promising alternative to $\overline{\mathbf{X}}$ in problems in which overestimation is severely penalized. It seems reasonable to conjecture that the relative performance of $\widehat{\theta}^G$ and $\widehat{\theta}^{\mathrm{JS}}$ in such problems will bear a greater resemblance to that encountered in Chapter 7, where Bayesian estimation in high dimensions was found to be a dicey proposition.

The problem considered in Section 8.3 admits to a similar type of discussion. Zellner (1986) shows that, in the regression problem under consideration, the MLE $\widehat{\theta} = \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ of $\theta = \mathbf{w}'\boldsymbol{\beta}$ is inadmissible, its risk function being uniformly larger than that of the frequentist estimator

$$\widetilde{\theta}_2 = \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - (c\mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}\sigma^2)/2\,, \tag{8.42}$$

an estimator which is also the generalized Bayes estimator with respect to a diffuse prior on $\boldsymbol{\beta}$. It is therefore clear that solving the threshold problem which compares the performance of $\widetilde{\theta}_2$ to that of Bayes estimators with respect to proper priors would also be of interest. It is, at present, an open problem.

Exercise 8.8. Consider the scenario above involving the estimation of a Poisson parameter.

(a) Derive the Bayes estimator $\widehat{\theta}^G$ of the parameter θ with respect to the prior G in (8.39).
(b) Obtain an expression for the Bayes risk $r(G_0, \widehat{\theta}^G)$.
(c) Obtain an expression for the Bayes risk $r(G_0, \widehat{\theta})$ of the MLE $\widehat{\theta} = \overline{X}$ of θ.
(d) Verify the inequality in (8.40).

9

The Treatment of Nonidentifiable Models

9.1 The classical viewpoint.

The identifiability of a statistical model, or of the parameters that serve as an index for the model, is one of the pillars on which the classical approach to statistical estimation is based. For parametric classes of distributions represented as $\{F_\theta, \theta \in \Theta\}$, the parameter θ is said to be *identifiable* if different values of the parameter, say θ_1 and θ_2, give rise to different distributions F_{θ_1} and F_{θ_2} of the observable variable X drawn from a distribution in the class. Without identifiability, a classical estimator $\widehat{\theta}$ of the unknown parameter θ would necessarily be ambiguous, and thus of little use. The data can only help "identify" an equivalence class in which the parameter appears to reside, but they cannot provide a specific numerical value that would play the role of one's best guess of the true value of the target parameter. In classical statistical estimation theory, the estimation of a nonidentifiable parameter is viewed, quite simply, as an ill-posed problem.

Interestingly, the occurrence of nonidentifiability in statistical problems of some importance is by no means uncommon. While classical methods are inapplicable in treating such problems directly, there are several available options. These options amount to treating a different but related problem to which classical methods do apply. Among these options are (a) placing additional restrictions on the original model, rendering the parameters of the restricted model identifiable, (b) focusing on the estimation of a function of the original parameters that is in fact identifiable and (c) expanding the model to include additional data which, together with the original data, makes the original parameters identifiable. I give examples of each of these strategies below.

Consider, first, the simple problem of estimating the parameter vector (μ_1, μ_2) from a univariate random sample $X_1, X_2, \ldots, X_n$ assumed to be normally distributed with distribution $\mathcal{N}(\mu_1 + \mu_2, \sigma^2)$. The sample mean $\overline{X}$ serves as a perfectly reasonable estimator of the sum $\mu_1 + \mu_2$, but it provides no useful information about the individual parameters μ_1 and μ_2. The parameter vector (μ_1, μ_2) is nonidentifiable in this problem since all such vectors with the same sum give rise to the same distribution for the available data. A version of this problem to which classical methods

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_9,

apply is the same estimation problem under the assumption that the value of, say, μ_1 is known. In that case, one could consider the pair $(\mu_1, \overline{X} - \mu_1)$ as an "estimator" of (μ_1, μ_2). This example of strategy (a) above solves the problem, in a sense, but has some obvious drawbacks. It is clearly not a solution to the original problem; also some justification is required for the assumption that μ_1 is known. These failings are typical of the approach taken; the assumption which renders the model identifiable cannot be investigated or tested using the data available in the original problem. If the assumption happens to be seriously wrong, the resulting "estimator" will be quite unsatisfactory.

As an example of strategy (b) above, consider the following problem. Suppose k Bernoulli variables are taken to represent the success or failure of the k independent components of an engineered system of interest. Further, assume that each of these variables is governed by its own parameter p_i. Given $X_i \sim \mathscr{B}(1, p_i)$, for $i = 1, \ldots, k$, system data is available in the form of a set of n independent observations of the variable $Y = \sum_{i=1}^{k} X_i$. The distribution of Y is the k-fold convolution of the Bernoulli distributions above. Samaniego and Jones (1981) noted that, while the parameter vector $\mathbf{p} = (p_1, \ldots, p_k)$ is nonidentifiable, the vector of ordered ps, that is, $\mathbf{p}^* = (p_{(1)}, \ldots, p_{(k)})$, where $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}$, is an identifiable function of $\mathbf{p}$ and can thus be estimated by standard methods. They then derived the maximum likelihood estimator of $\mathbf{p}^*$ based on the observed Ys and described its asymptotic properties. Again, while the problem of estimating the parameter $\mathbf{p}$ could not be treated by classical methods, a related problem of some practical interest is both well-defined and tractable.

In a similar vein, Tsiatis (1975) famously demonstrated that the problem of estimating the multiple decrement function $P(X_1 > x_1, X_2 > x_2, \ldots, X_k > x_k)$ nonparametrically in the biostatistical context of competing risks (where the observable data consist of pairs $(\min\{X_i\}, \delta)$, with X_i representing the time of failure due to "cause i" when that cause is acting alone and δ being an indicator function for the index of the observed X) suffers from nonidentifiability. Various researchers attempted to "fix" this deficiency by placing additional restrictions on the model, though most workers in the area acknowledged that the best that one could do in the original problem, within the classical framework, was to provide estimates of different parameters (for example, the cause-specific hazard functions) which are identifiable in the competing risks framework.

As an example of strategy (c) above, consider the problem of estimating the parameter pair (p, F) of the imperfect repair model of Brown and Proschan (1983). The model postulates that for any one of several identical systems that are put on test, the system is replaced, upon failure, by a new system (that is, it is repaired "perfectly") with probability p. With probability $(1 - p)$, the system is minimally repaired, that is, it is restored to its condition just prior to failure. The lifetime distribution of a new system is modeled nonparametrically as F, while a system which fails at time t_0 and is minimally repaired has the survival function $\overline{F}(t|t_0) = \overline{F}(t + t_0)/\overline{F}(t_0)$. The available data consists of the interfailure times for each of the systems on test. Whitaker and Samaniego (1989) noted that the parameter pair (p, F) of the Brown–Proschan model was not an identifiable parameter for the observed interfailure times from

fielded systems. They then showed that the parameter pair could be rendered identifiable if interfailure times were augmented by data on the mode of repair following each failure. Utilizing the augmented data, a consistent estimator for (p, F) was obtained. As with the other strategies mentioned above, the solution derived in this problem solves a problem that is related to, but different from, the original problem. In all the examples above, the original estimation problem posed remained unsolved.

Exercise 9.1. Given the independent observations $X_1, X_2, \ldots, X_k$, with $X_i \sim \mathscr{B}(1, p_i)$, show that the ordered parameter $(p_{(1)}, p_{(2)}, \ldots, p_{(k)})$ is an identifiable parameter of the distribution of $Y = \sum_{i=1}^{k} X_i$.

9.2 The Bayesian treatment of nonidentifiability

In contrast with the frequentist approach to estimation in the presence of nonidentifiability, the Bayesian paradigm has no difficulty in treating nonidentifiable parameters. The idea here is very simple. A Bayesian will begin the treatment of an estimation problem by stipulating a prior distribution on the parameters of the model of interest. Now, in a model with nonidentifiable parameters, the data available to the statistician are "defective" in the sense that they do not provide unambiguous information about the model's parameters. It is nonetheless the case that the data observed in such problems are still *informative* about these parameters. The updating of the prior distribution on the basis of the observed data is thus both feasible and meaningful, resulting in a posterior distribution on which inference can be based. While this circumstance has been recognized in the Bayesian literature for some time, it has not been widely exploited.

Nonidentifiability is an issue that occurs with some frequency in econometric modeling. An early example of the Bayesian treatment of such problems is the paper by Lindley and El-Sayyad (1968), where the Bayesian solution is derived to the problem of estimating parameters subject to a linear functional constraint. Other developments in this general domain include papers by Dreze (1975), Clayton and Kaldor (1987) and Besag, York and Mollie (1991). But the fact that Bayesian analysis is possible in such problems, while the frequentist can offer no solution, is not, by itself, reason to proceed with or trust the Bayesian solution. Bayesian estimation of a nonidentifiable parameter involves some potential dangers. For example, unlike the case of estimating an identifiable parameter, a Bayes estimator of a nonidentifiable parameter is, of necessity, highly dependent on the prior model. In addition, Bayes estimators in the latter circumstance will not, in general, be consistent estimators of the true value of the parameter θ. Under mild assumptions, they can be guaranteed to converge to a parameter value within an equivalence class containing the true parameter. But even as the sample size n tends to infinity, a Bayes estimator may not converge to a value that is "acceptably close" to the true value of θ.

The question of whether Bayesian inference is "efficacious" remains to be examined. Now the choices available are limited, but there are clearly two possible estimators of θ that should be compared. Under squared error loss, the mean of the

posterior distribution of θ is the option of primary interest. What might that estimator be compared to? If θ were identifiable, the prior mean would not be considered a potential estimator, as it's simply a guess at θ that makes no use of the experimental data. In standard Bayesian inference, the prior distribution is just a "seed" that gets the Bayesian process going. But the prior mean is more than that when θ is nonidentifiable. In the latter case, the prior mean can actually be closer to the true value of θ than the posterior mean, even when the information available through sampling is unlimited. In such cases, one would prefer the prior guess to the posterior guess. In essence, the available data would have served, in that instance, to mislead the statistician into an inference that was inferior to her starting point.

This circumstance leads naturally to a new form of the threshold problem, one that is focused on the comparison of two different Bayes estimators — the no-data estimator θ_G based solely on the prior distribution G and the Bayes estimate $\widehat{\theta}_G$ based on the posterior distribution of θ given the available data. Such a comparison seeks to assess the efficacy of Bayesian updating, and its ultimate goal would be to characterize a class of prior distributions for which Bayesian updating provides an improvement over an estimator based solely on the prior distribution. In Section 9.3, the "efficacy" problem will be studied in the context of a simple nonidentifiable (binomial) model — a toy problem, if you will — but the goal of the developments in that section is to shed light on the general character of solutions to the threshold problem in this new context. In Sections 9.3 and 9.4, we treat an estimation problem based on a nonidentifiable binomial model. The treatment here is drawn from Neath and Samaniego (1997). The discussion on the efficacy of Bayes estimators in the nonparametric competing risks problem (Section 9.5) is based on Neath and Samaniego (1996a, 1996b). In Section 9.6, I discuss a nonidentifiable version of an estimation problem based on stress–strength testing in Reliability. For that, I will draw upon the developments in Samaniego (2007).

9.3 Estimation for a nonidentifiable binomial model

Suppose that the observed data in an experiment of interest is a binomial random variable with distribution

$$X \sim \mathscr{B}(n, p_1 + p_2)\,, \tag{9.1}$$

where $p_1 \geq 0$, $p_2 \geq 0$ and $0 \leq p_1 + p_2 \leq 1$. The model would be appropriate in situations in which there are two mutually exclusive causes of "success" in a sequence of n Bernoulli trials, and these causes are indistinguishable without costly or infeasible follow-up. The target of the estimation problem of interest is the pair (p_1, p_2), a parameter pair which is, of course, nonidentifiable on the basis of the available data. A Bayesian analysis of this estimation problem would typically begin with the specification of a Dirichlet prior. Specifically, let us take, as a prior model G, the Dirichlet distribution of (p_1, p_2, p_3), where $p_3 = 1 - p_1 - p_2$. This distribution will be denoted by $\mathscr{D}(\alpha_1, \alpha_2, \alpha_3)$, with $\alpha_i > 0$ for $i = 1, 2, 3$, and has the bivariate density function given by

$$g(p_1,p_2)=\frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}p_1^{\alpha_1-1}p_2^{\alpha_2-1}(1-p_1-p_2)^{\alpha_3-1}\,, \tag{9.2}$$

where $p_1>0$, $p_2>0$, with $0<p_1+p_2<1$. The mean of this distribution is the pair

$$(p_1^G,p_2^G)=\left(\frac{\alpha_1}{\alpha_1+\alpha_2+\alpha_3},\,\frac{\alpha_2}{\alpha_1+\alpha_2+\alpha_3}\right)\,. \tag{9.3}$$

The pair (p_1^G,p_2^G) represents the Bayesian statistician's prior guess at the parameter pair (p_1,p_2). It is easy to verify (you guessed it, it's an exercise for your amusement and edification) that the posterior distribution of (p_1,p_2), given $X=x$, is the mixture of Dirichlet distributions given by

$$f(p_1,p_2|x)=\sum_{k=0}^{x}a_k\mathscr{D}(\alpha_1+k,\alpha_2+x-k,\alpha_3+n-x)\,, \tag{9.4}$$

where

$$a_k=\frac{\binom{x}{k}\Gamma(\alpha_1+k)\Gamma(\alpha_2+x-k)}{\sum_{j=0}^{x}\binom{x}{j}\Gamma(\alpha_1+j)\Gamma(\alpha_2+x-j)}\,. \tag{9.5}$$

From (9.4), one may identify the Bayes estimator of (p_1,p_2), given $X=x$, as

$$(\widehat{p}_1^G,\widehat{p}_2^G)=\left(\sum_{k=0}^{x}a_k\frac{\alpha_1+k}{A+n},\,\sum_{k=0}^{x}\alpha_k\frac{\alpha_2+x-k}{A+n}\right)\,, \tag{9.6}$$

where $A=\alpha_1+\alpha_2+\alpha_3$. One might compare the prior Bayes estimator in (9.3) with the posterior Bayes estimator in (9.6) for any size experiment, that is, for any value of n. But the most telling comparison is the limiting case. As $n\to\infty$, the Bayesian will obtain as much information about the parameter (p_1,p_2) as the experiment can offer. It is thus of special interest to ask whether, when taking maximal advantage of the experimental input, the process of Bayesian updating can provide improvement over the prior guess (p_1^G,p_2^G). With that goal in mind, we turn to the development of the asymptotic form of the Bayes estimator $(\widehat{p}_1^G,\widehat{p}_2^G)$ as $n\to\infty$. This is accomplished in the following theorem. The comparisons to be made in the sequel can be placed in the context of earlier developments by specifying a true prior distribution G_0 on the pair (p_1,p_2). I will henceforth consider the true prior G_0 to be degenerate at the point (p_1^*,p_2^*). I will refer to this point repeatedly, but make no further mention of G_0.

Theorem 9.1. *Let $X_n\sim\mathscr{B}(n,p_1+p_2)$, and let (p_1^*,p_2^*) be the true but unknown value of the parameter pair (p_1,p_2). Suppose the operational prior distribution G of (p_1,p_2) is the Dirichlet distribution $\mathscr{D}(\alpha_1,\alpha_2,\alpha_3)$. As $n\to\infty$, the posterior distribution of p_1, given $X_n=x$, is a rescaled beta distribution, that is,*

$$p_1|X_n=x\xrightarrow{D}cW\,, \tag{9.7}$$

where $W\sim Be(\alpha_1,\alpha_2)$ and $c=p_1^*+p_2^*$, and the posterior distribution of p_2, given $X_n=x$, is the complementary rescaled beta distribution, that is,

$$p_2|X_n = x \xrightarrow{D} cV \ , \tag{9.8}$$

where $V \sim Be(\alpha_2, \alpha_1)$ *and* $c = p_1^* + p_2^*$.

Proof. Note that the Dirichlet prior on (p_1, p_2) can be transformed into a prior distribution on the parameter pair (θ_1, θ_2), where $\theta_1 = p_1/(p_1 + p_2)$ and $\theta_2 = p_1 + p_2$. The variables θ_1 and θ_2 are independent, with $\theta_1 \sim \text{Be}(\alpha_1, \alpha_2)$ and $\theta_2 \sim \text{Be}(\alpha_1 + \alpha_2, \alpha_3)$. Given $X_n = x$, it is easy to confirm that the posterior distributions of θ_1 and θ_2 are

$$\theta_1|X_n = x \sim \text{Be}(\alpha_1, \alpha_2) \text{ and } \theta_2|X_n = x \sim \text{Be}(\alpha_1 + \alpha_2 + x, \alpha_3 + n - x) \ . \tag{9.9}$$

As $n \to \infty$, the limiting posterior distribution of θ_1 is $\text{Be}(\alpha_1, \alpha_2)$ while the limiting posterior distribution of θ_2 is the distribution that is degenerate at the constant $c = p_1^* + p_2^*$. It follows that the limiting posterior distribution of the parameter $p_1 = \theta_1 \cdot \theta_2$ is the rescaled beta distribution in (9.7). The symmetric roles played by p_1 and p_2 in the model above imply that (9.8) must also hold. ■

Theorem 9.1 implies that the limiting form of the Bayes estimator $(\widehat{p}_1^G, \widehat{p}_2^G)$ of (p_1, p_2), with respect to the operational prior $G = \mathscr{D}(\alpha_1, \alpha_2, \alpha_3)$ and loss equal to the sum of squared errors, is

$$(\widehat{p}_1^G, \widehat{p}_2^G) = \left(\frac{c\alpha_1}{\alpha_1 + \alpha_2}, \frac{c\alpha_2}{\alpha_1 + \alpha_2} \right) \ , \tag{9.10}$$

where $c = p_1^* + p_2^*$. From (9.10), it is clear that the Bayes estimator $(\widehat{p}_1^G, \widehat{p}_2^G)$ is a consistent estimator of (p_1, p_2) if and only if

$$\frac{\alpha_1}{\alpha_1 + \alpha_2} = \frac{p_1^*}{p_1^* + p_2^*} \ , \tag{9.11}$$

a restriction on the prior distribution that will virtually never be satisfied in practice. The effect of Bayesian updating on the operational prior distribution $\mathscr{D}(\alpha_1, \alpha_2, \alpha_3)$ with mean $(a, b) = (\alpha_1/(\alpha_1 + \alpha_2 + \alpha_3),\ \alpha_2/(\alpha_1 + \alpha_2 + \alpha_3))$ is shown in Figure 9.1. While the consistency of a Bayes estimator is obviously an unachievable goal in this problem (and in any problem involving nonidentifiability), the efficacy of Bayesian updating is still amenable to study. In the next section, a complete characterization is presented of the subclass of Dirichlet priors G whose corresponding estimators $\left(\widehat{p}_1^G, \widehat{p}_2^G\right)$ are uniformly superior, asymptotically, to the prior estimate (p_1^G, p_2^G). A related simulation study shows that, for randomly chosen values of the prior mean and of (p_1^*, p_2^*), the percentage of limiting Bayes estimators that are superior to the prior estimator of the parameter pairs (p_1, p_2) satisfying the constraint $p_1 + p_2 = c$, can range from a low of 51.55% of the class of Dirichlet priors in the worst case (when c, the true value of the sum $p_1 + p_2$, is close to 1) to percentages arbitrarily close to 100% in the best case (when c is close to 0). Sharp upper and lower bounds are established for the ratio of Euclidean distances between each of the estimators (p_1^G, p_2^G) and $(\widehat{p}_1^G, \widehat{p}_2^G)$ and the true value of the pair (p_1, p_2).

Exercise 9.2. Confirm that the density in (9.4) is the posterior density of (p_1, p_2), given the observation $X = x$, where $X \sim \mathscr{B}(n, p_1 + p_2)$ and the prior on (p_1, p_2) is $\mathscr{D}(\alpha_1, \alpha_2, \alpha_3)$.

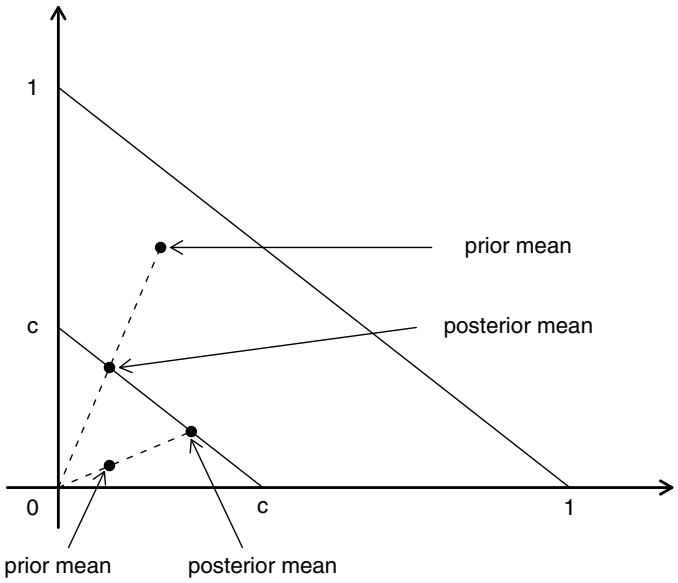

Fig. 9.1. Prior and limiting posterior means in the (p_1, p_2) plane

9.4 On the efficacy of Bayesian updating in the binomial model

In this section, we will index the class of Dirichlet distributions on (p_1, p_2) by their means $\{(a,b) \in (0,1)^2\}$. This simplification sacrifices no generality since the limiting form of the corresponding Bayes estimators, and their distances from the true value of (p_1, p_2), depend on the prior only through its mean. As we have seen in Section 9.2, the limiting Bayes estimate of the parameter (p_1, p_2) maps the prior mean (a,b) onto the posterior mean $(\gamma a, \gamma b)$, where $\gamma = c/(a+b)$, with the constant c being the true value $p_1^* + p_2^*$ of the sum $p_1 + p_2$. The point $(\gamma a, \gamma b)$ lies on the line $p_1 + p_2 = c$.

I will now proceed with a detailed examination of the issue of the relative closeness of the points (a,b) and $(\gamma a, \gamma b)$ to the true value of the pair (p_1, p_2). The Euclidean distance between the two points (u,v) and (x,y) in the plane will be denoted by

$$D[(u,v),(x,y)] = \sqrt{(u-x)^2 + (v-y)^2}\,. \tag{9.12}$$

Now, consider an operational Dirichlet prior distribution on (p_1, p_2) with mean (a,b) and the associated limiting posterior mean $(\gamma a, \gamma b)$. These points represent the prior estimator and the posterior estimator of (p_1, p_2) given this particular operational prior. It is clear that the posterior estimator will necessarily be closer to the true value (p_1^*, p_2^*) of (p_1, p_2) than the prior estimator, regardless of what value (p_1^*, p_2^*) takes on the line $p_1 + p_2 = c$, if and only if

$$D[(\gamma a, \gamma b), (c,0)] < D[(a,b), (c,0)] \tag{9.13}$$

and

$$D[(\gamma a, \gamma b), (0,c)] < D[(a,b), (0,c)] \, . \tag{9.14}$$

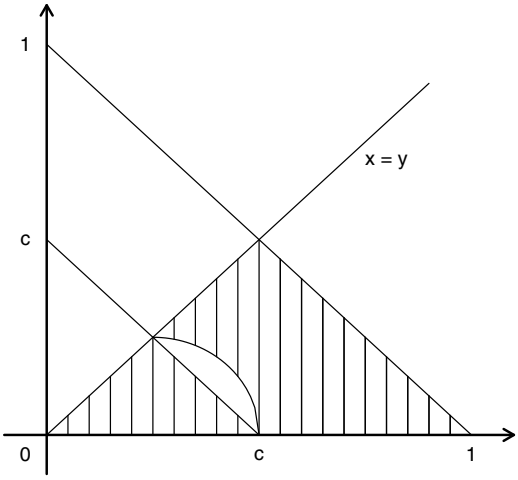

Fig. 9.2. Prior means (a,b) with $a > b$ for which the limiting posterior mean $(\gamma a, \gamma b)$ is closer to the point $(c,0)$

To identify the prior distributions for which (9.13) and (9.14) hold, we will consider the following two cases separately.

Case 1: Characterizing prior means $(a,b) \in (0,1)^2$ that are farther from the point $(c,0)$ than the limiting posterior means $(\gamma a, \gamma b)$, where $c \in (0,1)$ is a fixed constant.

(i) The subcase in which the prior mean (a,b) is such that $a > b$ and $a+b < c$ is particularly simple, since the angle formed by the line segments joining the points (a,b), $(\gamma a, \gamma b)$ and $(c,0)$ exceeds 90°, so that $(\gamma a, \gamma b)$ is necessarily closer than (a,b) to $(c,0)$.

(ii) Now, let us attempt to identify the points (a,b), with $1 > a > b > 0$ and $a+b > c$, for which

$$D[(\gamma a, \gamma b), (c,0)] = D[(a,b), (c,0)] \, . \tag{9.15}$$

Consider the class of concentric circles centered at the point $(c,0)$ and with radius r satisfying $0 \le r \le c/\sqrt{2}$. Each of these circles will intersect the line segment joining the points $(c,0)$ and $(c/2, c/2)$. For any point (x,y) such

that $x > y$ and $x + y > c$, there is a unique point at which the circle with radius

$$r = D\left[\left(\frac{cx}{x+y}, \frac{cy}{x+y}\right), (c,0)\right]$$

will intersect the line segment through the points $(0,0)$ and (x,y). This point of intersection represents a prior mean (a,b) which is the same distance from $(c,0)$ as the limiting posterior mean $(\gamma a, \gamma b)$. Such a point (a,b) is the unique solution of the equation

$$(s-c)^2(s+t)^2 + t^2(s+t)^2 - 2c^2t^2 = 0 \tag{9.16}$$

among points (s,t) on the line passing through $(0,0)$ with slope y/x, where (x,y) is a fixed point satisfying the inequalities $x > y$ and $x + y > c$. There are uncountably many such solutions as (x,y) varies among pairs satisfying the constraints above; the set of all such solutions constitutes a curve in the simplex $\{(u,v) \mid u \geq 0, v \geq 0, u+v \leq 1\}$. It is easy to confirm that the points $(c/2, c/2)$ and $(c,0)$ satisfy equation (9.16) when this line has slope 1 or 0, respectively, a fact that implies that the set of solutions to (9.16) intersects the line $x + y = c$ at these two points. The region between the collection of solutions of (9.16) and the line $x + y = c$ corresponds to prior means (a,b) which are closer to the point $(c,0)$ than are the posterior means $(\gamma a, \gamma b)$. The prior means which improve through Bayesian updating for cases (i) and (ii) are pictured in Figure 9.2.

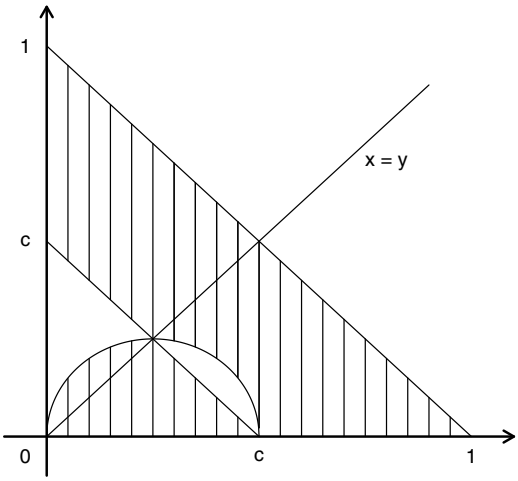

Fig. 9.3. Prior means (a,b) for which the limiting posterior mean $(\gamma a, \gamma b)$ is closer to the point $(c,0)$

(iii) When $a < b$, prior means (a,b) satisfying $a+b > c$ are obviously farther from the point $(c,0)$ than the limiting posterior mean $(\gamma a, \gamma b)$, since the angle formed by the line segments joining the points (a,b), $(\gamma a, \gamma b)$ and $(c,0)$ exceeds 90°.

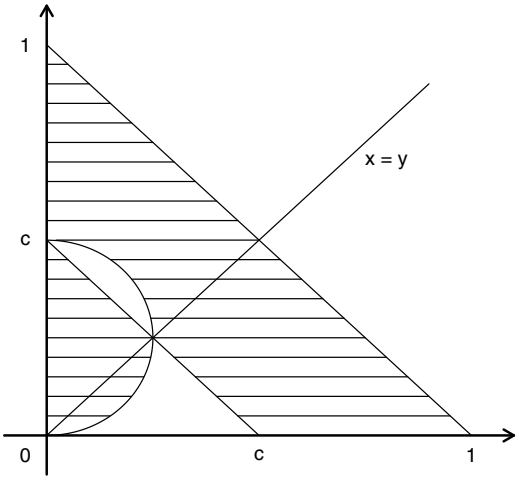

Fig. 9.4. Prior means (a,b) for which the limiting posterior mean $(\gamma a, \gamma b)$ is closer to the point $(0,c)$

(iv) Consider, finally, prior means for which $a < b$ and $a+b < c$. To identify the boundary separating prior means for which Bayesian updating is effective from the remaining prior means (a,b) satisfying these constraints, consider the class of concentric circles centered at the point $(c,0)$ and with radius r satisfying $c/\sqrt{2} \le r \le c$. Each of these circles will intersect the line segment joining the points $(c/2, c/2)$ and $(0,c)$. For any point (x,y) such that $x < y$ and $x+y < c$, there is a unique point (a,b) at which the circle with radius

$$r = D\left[\left(\frac{cx}{x+y}, \frac{cy}{x+y}\right), (c,0)\right]$$

will intersect the line segment through the points $(0,0)$ and (x,y). The points (a,b) and $(\gamma a, \gamma b)$ are equidistant from $(c,0)$. The point (a,b) is the unique solution of the equation (9.16) among points (s,t) on the line in the unit square passing through $(0,0)$ with slope y/x, where (x,y) is a fixed point satisfying the inequalities $x < y$ and $x+y < c$. It is easy to confirm that the points $(c/2, c/2)$ and $(0,0)$ satisfy equation (9.16) when this line has slope 1 or passes through the point $(1-c/\sqrt{2}, c/\sqrt{2})$, respectively, a fact that implies that the collection of solutions of (9.16) intersects with the line

$x = y$ at these two points. The region between the solution set to (9.16) and the line $x = y$ correspond to prior means (a,b), with $a < b$ and $a + b < c$, which are farther from the point $(c,0)$ than the posterior means $(\gamma a, \gamma b)$.

When the prior means (a,b), with $a < b$, which are farther from the point $(c,0)$ than the limiting posterior mean $(\gamma a, \gamma b)$ are added to the prior means pictured in Figure 9.2, one obtains the complete collection of prior means (a,b) in the unit square which are farther from the point $(c,0)$ than is the limiting posterior mean $(\gamma a, \gamma b)$. This collection is pictured in Figure 9.3.

Case 2: Characterizing prior means (a,b) that are farther from the point $(0,c)$ than the limiting posterior means $(\gamma a, \gamma b)$.

The problem of characterizing the collection of prior means (a,b) which are farther from the point $(0,c)$ than the limiting posterior mean $(\gamma a, \gamma b)$ is the mirror image of the problem considered as Case 1. The symmetric nature of these complementary problems leads to the conclusion that the reflection, across the diagonal line of the unit square, of the solution pictured in Figure 9.3 constitutes the solution to the problem in Case 2. These prior means are displayed in Figure 9.4.

Now, consider the set of prior means $(a,b) \in (0,1)^2$ which are farther from both $(c,0)$ and $(0,c)$ than the limiting posterior mean $(\gamma a, \gamma b)$. This set is simply the intersection of the prior means obtained as solutions in each of the two cases considered above. The collection of prior means for which Bayesian updating guarantees universal improvement is pictured in Figure 9.5.

When the true value of the sum $p_1 + p_2$ is small, it is clear from Figure 9.5 that the collection of prior means for which Bayesian updating ensures universal improvement, in the limit, is quite large relative to the size of the parameter space. In such problems, even if the prior mean was determined by randomly selecting a point from the unit square, the chances of improving one's prior estimate through Bayesian updating would be quite high. Such is not the case when the true value of the sum $p_1 + p_2$ is close to 1. Figure 9.6 illustrates this fact for the value $c = 1$.

Besides investigating the question of when the limiting Bayes estimator $(\gamma a, \gamma b)$ is closer to the true value of (p_1, p_2) than the prior mean (a,b), regardless of the location of the true value (p_1^*, p_2^*) in the equivalence class $\{(p_1,p_2) \mid p_1 + p_2 = c\}$, it is reasonable to ask how often one might expect Bayesian updating to be efficacious in the limit. For any prior mean (a,b), there are at least some potential true values of the parameter (p_1,p_2) for which Bayesian updating will be asymptotically superior. Neath and Samaniego (1997) examined this problem via simulation based upon random sampling in which the prior mean (a,b) is chosen at random from the simplex $\{(u,v) \mid u \geq 0, v \geq 0, u+v \leq 1\}$ and the "true value" of (p_1,p_2) is chosen at random from the set $\{(p_1,p_2) \mid p_1 + p_2 = c\}$ for a variety of fixed values of c ranging from 0 to 1. Table 9.1 sheds light on the behavior of the probability

$$\Pi_c = P(D_1 > D_2 \mid p_1^* + p_2^* = c)\,, \tag{9.17}$$

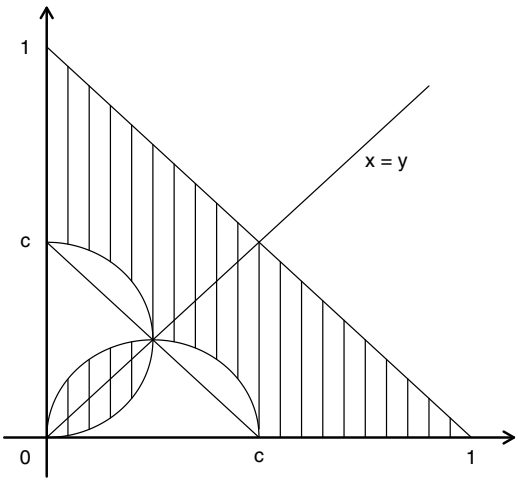

Fig. 9.5. Prior means (a,b) for which the limiting posterior mean $(\gamma a, \gamma b)$ is closer to all the points on the line $x+y=c$

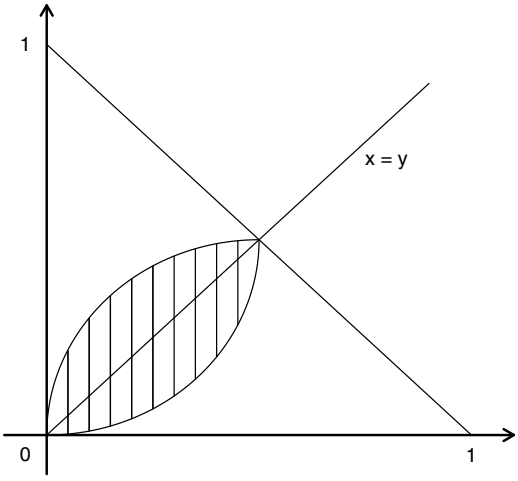

Fig. 9.6. Prior means (a,b) for which the limiting posterior mean $(\gamma a, \gamma b)$ is closer to any parameter value on the line $p_1+p_2=1$

where

$$D_1 = D[(a,b),(p_1^*,p_2^*)] \quad \text{and} \quad D_2 = D[(\gamma a,\gamma b),(p_1^*,p_2^*)]\,. \tag{9.18}$$

Table 9.1 indicates that Bayesian updating will, in the limit, almost always be efficacious when the true value of the sum p_1+p_2 is small, and that, even in the worst case

scenario, where $p_1 + p_2$ is close to 1, Bayesian updating improves upon the prior estimate of (p_1, p_2), in the limit, over 50% of the time. These claims apply when the prior mean is chosen at random within the unit square. When the prior distribution is chosen with care, based on useful prior information that may be available in the application of interest, the efficacy of Bayesian updating can be expected to be reasonably large. One additional insight that may be gleaned from Figures 9.5 and 9.6 is that a prior model whose mean (a,b) is sufficiently close to the diagonal line, that is, for which $a \approx b$, will always lead to efficacious Bayesian updating in the limit.

Table 9.1. The simulated probability Π_c of asymptotic superiority of Bayesian updating

c	Π_c	c	Π_c	c	Π_c
0.00	1.000	0.35	0.937	0.70	0.729
0.05	0.999	0.40	0.915	0.75	0.690
0.10	0.996	0.45	0.892	0.80	0.647
0.15	0.989	0.50	0.866	0.85	0.608
0.20	0.981	0.55	0.834	0.90	0.576
0.25	0.970	0.60	0.799	0.95	0.545
0.30	0.952	0.65	0.767	1.00	0.516

One final issue that we will examine in this problem regards the extent to which either the prior mean (a,b) or the limiting posterior mean $(\gamma a, \gamma b)$ dominates the other for an arbitrary choice of (a,b). The following result shows that the potential for improving upon the prior mean is essentially unlimited, while when the prior mean turns out to be superior to the limiting posterior mean, it is only slightly so.

Theorem 9.2. *Let D_1 and D_2 be the Euclidean distances defined in* (9.18)*. Then*

$$\frac{\sqrt{2}}{2} \leq \frac{D_1}{D_2} \leq \infty\,, \tag{9.19}$$

and these bounds are sharp.

Proof. Let (p_1^*, p_2^*) be the true value of (p_1, p_2) and let $D_1 = D[(a,b),(p_1^*, p_2^*)]$ and $D_2 = D[(\gamma a, \gamma b),(p_1^*, p_2^*)]$. It is clear that, if the prior mean (a,b) happens to be chosen so that $a/b = p_1^*/p_2^*$, then the limiting posterior mean $(\gamma a, \gamma b)$ is (p_1^*, p_2^*), and the distance $D_2 = 0$. This establishes the RHS of the inequality in (9.19), and it shows that the upper bound is sharp. To establish the lower bound in (9.19), one needs to consider separately the subcases (i)–(iv) defined within "Case 1" above. We give a formal proof of the lower bound for subcase (iv), as the subcase is typical and the other subcases may be treated by similar arguments.

Let us suppose that $a < b$ and that $a + b < c$. If the point (p_1^*, p_2^*) lies above and to the left of $(\gamma a, \gamma b)$, it is clear that $D_1/D_2 > 1$. Consider the remaining possibility, that is, the case in which $p_1^* > \gamma a$ and $p_2^* < \gamma b$. The latter situation is pictured in Figure 9.7, where the interior angles of the triangle joining the points (a,b), $(\gamma a, \gamma b)$

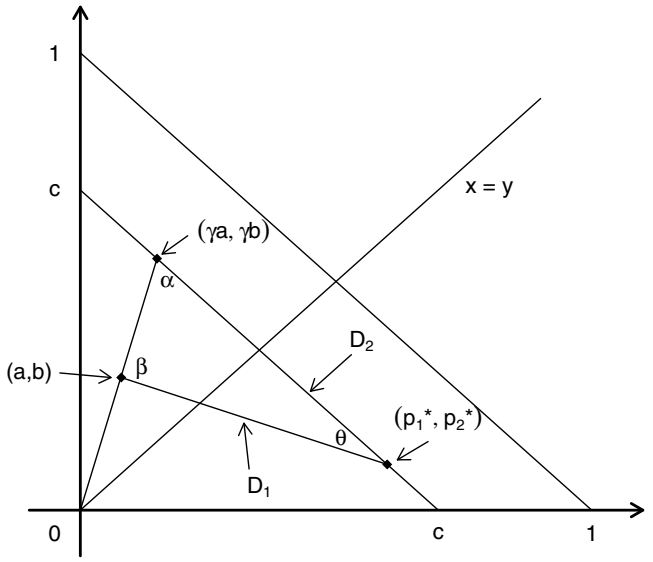

Fig. 9.7. Triangle connecting the points (a,b), $(\gamma a, \gamma b)$ and (p_1^*, p_2^*) in subcase (iv): $a < b$ and $a+b < c$

and (p_1^*, p_2^*) are denoted by α, β and θ. Since in this case we have $45^\circ < \alpha < 90^\circ$, it follows from the law of sines that

$$\frac{D_1}{D_2} = \frac{\sin\alpha}{\sin\beta} \geq \sin\alpha \geq \sin 45^\circ = \frac{\sqrt{2}}{2}\,. \tag{9.20}$$

This establishes the lower bound in (9.19). That the lower bound is sharp follows from the fact that it is the exact value of D_1/D_2 for the prior mean $(0,b)$. This completes the proof of the theorem for subcase (iv). Subcase (iii) may be proven by a similar argument, and subcases (i) and (ii) follow by symmetry. ■

Exercise 9.3. To disabuse the reader of the impression that the discussion above is merely of theoretical interest, consider the following application to a problem involving medical screening tests. The results of a collection of Canadian HIV screening tests, as reported by Nusbacher *et al.* (1986), were as follows: in 94,496 blood samples, there were 405 positive tests. In subsequent testing of these 405 individuals, 14 were determined to be true positives. Suppose this information is used as the basis for prior modeling of HIV screening data elsewhere. A large-scale HIV screening test executed in the UK resulted in 373 positive tests in 3,122,556 trials (see Johnson and Gastwirth (1991)). Using $(a,b) = (0.000148, 0.004138)$ as the mean of our Dirichlet prior, treating the sample size n in the UK test as effectively infinite and using the realized proportion of positive tests in the UK as the true value $c = 0.000119$ of the

sum $p_1 + p_2$, obtain the limiting posterior Bayes estimate of (p_1, p_2). Verify that this estimate is closer to the true value (p_1^*, p_2^*) of (p_1, p_2) than the prior mean, regardless of the exact value of (p_1^*, p_2^*). When the subjects who tested positive in the UK were retested, it was found that the number of true positives was 64. Thus, the true value of (p_1, p_2) was determined to be $(0.000020, 0.000099)$. Evaluate the extent of improvement afforded by Bayesian updating by calculating the ratio D_1/D_2.

9.5 On the efficacy of Bayesian updating in the nonparametric competing risks problem

In the general competing risks problem, it is assumed that a given subject is exposed to k possible causes of failure. Let $X_1, X_2, \ldots, X_k$ be positive random variables, where X_i represents the theoretical survival time of a subject until failure due to the ith risk when that risk is assumed to be acting alone. In reality, a subject will survive up to the time $Z = \min\{X_1, X_2, \ldots, X_k\}$, his or her "actual survival time." In a life testing experiment, the available data consists of pairs (Z, δ), where $\delta = j$ when $Z = X_j$. The "crude survival probability" $P(Z > z, \delta = j)$ may be estimated by the empirical relative frequency of the event $\{Z > z, \delta = j\}$ in the sample. The more interesting problem, and one whose solution would have greater importance and broader interpretability, is that of estimating the "multiple decrement function" (MDF) $S(x_1, x_2, \ldots, x_k) = P(X_1 > x_1, X_2 > x_2, \ldots, X_k > x_k)$ based on the sample of "identified minima" $\{(Z_1, \delta_1), (Z_2, \delta_2), \ldots, (Z_n, \delta_n)\}$. Even in early treatments of this problem, it was clear that the question of the identifiability of S needed to be resolved.

Berman (1963) proved the identifiability of S as the overarching survival function governing the behavior of identified minima under the assumption that the Xs were mutually independent. However, since the assumption of independence cannot be tested from a sample of identified minima, the estimation of S under that assumption remained a somewhat chancy strategy. In an important and highly influential paper, Tsiatis (1975) demonstrated that, without the independence assumption, the nonparametric multiple decrement function S was not identifiable from observed identified minima. It thus became clear that S could not be estimated by classical methods. However, the possibility of the development of Bayes estimators of S remained open. Phadia and Susarla (1983) derived the Bayes estimator of S with respect to a Dirichlet process (DP) prior (see Ferguson (1973) for a formal definition and basic properties), though they did not determine the posterior distribution. Arnold *et al.* (1984) identified the posterior distribution as a mixture of Dirichlet processes (MDP) (see Antoniak (1974) for details on MDPs) and demonstrated the general inconsistency of the corresponding Bayes estimators of S, a failing attributable to the nonidentifiability of the model. Neath and Samaniego (1996a, 1996b) considered the issue of the efficacy of Bayesian estimation in the competing risks context described above.

As in the binomial problem treated in Section 9.4, it is possible that the Bayes estimator of S might in fact be poorer than the prior estimator, the mean of the prior distribution, even when the statistician has unlimited data in the form of identified minima. In order to make the comparison of interest, that is, the comparison between

S^G, the prior estimate of S and $\widehat{S}^G_\infty$, the limiting posterior estimate of S, as the sample size n grows to ∞, one needs to identify the limiting posterior distribution of S, given the infinite sequence of independent observations $(Z_1,\delta_1),(Z_2,\delta_2),\ldots,(Z_n,\delta_n),\ldots$. For simplicity, Neath and Samaniego (1996a) restrict attention to the bivariate case in which there are two competing risks with corresponding survival times X and Y. Their Theorem 3.2 obtains the posterior distribution of S, based on a finite sample of identified minima, given a Dirichlet process prior $D(\alpha)$, with α a continuous, finite measure over $(\mathbb{R}^2,B)$, where $\mathbb{R}^2$ is the Euclidean plane and B is the σ-field of Borel sets. I record this result as

Theorem 9.3. *Without loss of generality, take $x>y$. If $S\sim D(\alpha)$, then the posterior distribution of $S(x,y)$, given $(Z_1,\delta_1),(Z_2,\delta_2),\ldots,(Z_n,\delta_n)$, with $\delta_j=I_{X_i}(Z_j)$, is a mixture of beta distributions which can be represented as the following beta distribution with a random parameter:*

$$\begin{aligned} S(x,y)\mid(Z_1,\delta_1),(Z_2,\delta_2),\ldots,(Z_n,\delta_n)\\ \sim Be(A(x,y)+m_1+U_n, A-A(x,y)+m_2+n-U_n)\,, \end{aligned} \tag{9.21}$$

where

$$\begin{aligned} A(x,y) &= \alpha((x,\infty),(y,\infty)),\\ m_1 &= \sum_{i=1}^n I_{(x,\infty)}(Z_i),\\ m_2 &= \sum_{i=1}^n I_{(0,y)}(Z_i)+\sum_{i=1}^n I_{(y,x)}(Z_i)I_{\{1\}}(\delta_i),\\ n &= \sum_{i=1}^n I_{(y,x)}(Z_i)I_{\{0\}}(\delta_i) \end{aligned}$$

and $\{U_n\}$ is a stochastic process defined by: $W_1,W_2,\ldots,W_n$ are i.i.d. random variables drawn from the distribution of $Y\mid y<Y<x,\ Y<X$; $U_0=0$ wp1, and for $k=1,2,\ldots,n$, $U_k-U_{k-1}\mid U_{k-1},W_1,\ldots,W_n\sim\mathscr{B}(1,P(X>x\mid X>W_k,Y=W_k))$, where the probability P is computed with respect to the operational prior distribution on S.

The representation in Theorem 9.3 facilitates the derivation of the limiting posterior distribution of S. In Theorem 4.1 of Neath and Samaniego (1996a), this distribution is shown to be degenerate at a point, that is, at a distinct multiple decrement function. For arbitrary $0<y<x$, the limiting posterior estimator $\widehat{S}^G_\infty(x,y)$ was shown to be

$$\widehat{S}^G_\infty(x,y)=P^*(X>x,Y>y)+P^*(y<Y<x,X>Y)\int P(w)\,\mathrm{d}H(w)\,, \tag{9.22}$$

where H is the distribution of the i.i.d. Ws in Theorem 9.3 and P^* is the "true probability measure" associated with the random variables X and Y. The expression for the

limiting posterior estimator of $S(x,y)$ for the case $0 < x < y$ is completely analogous to (9.22).

The point estimator in (9.22) is seen to be an intuitively reasonable way of combining prior information and the available experimental data. Identified minima provide reliable information about two key aspects of the multiple decrement function S, leading, in effect, to estimates of the probability of survival beyond a fixed time x and also of the joint probability that $y < Y < x$ and $X < Y$. On the other hand, the limiting Bayes estimator $\widehat{S}_{\infty}^{G}$ must rely on the prior distribution for estimating the conditional survival probability of X when $y < Y < x$ is observed and $X > Y$. The correctness of the latter information is at the crux of the potential consistency of the sequence of Bayes estimators. A necessary and sufficient condition for consistency of the sequence of Bayes estimators of S, as $n \to \infty$, is the property

$$\int P(w)\, \mathrm{d}H(w) = \int P^*(w)\, \mathrm{d}H(w)\,, \tag{9.23}$$

where P represents the operational prior model associated with $S(x,y)$, P^* represents the true model and H is defined as in (9.22).

It is of interest to find prior distributions $D(\alpha)$ for which the limiting posterior estimator of $S(x,y)$ in the competing risks problem is closer to the true value $S^*(x,y)$ of the multiple decrement function than the prior estimate. Now when $\alpha(\mathbb{R}^2) < \infty$, the mean of the prior $D(\alpha)$ on a survival function S is the survival function $\overline{G}(x,y) = \alpha((x,\infty),(y,\infty)/\alpha(\mathbb{R}^2)$. We will now turn our attention to a discussion regarding a particular circumstance in which a productive investigation of possible Bayesian superiority can be made. For simplicity, we will again restrict our investigation to the bivariate case. However, the main result below generalizes easily to the multivariate version of this problem.

The target parameter of our investigation is the MDF $S(x,y) = P(X > x, Y > y)$, where X and Y are nonnegative variables representing theoretical survival times under two potential risks, were these risks operating alone. Tsiatis showed that the function S is nonidentifiable on the basis of a sample of identified minima $\{(Z_i,\delta_i),\ i = 1,\ldots,n\}$, where $\delta_j = I_{X_j}(Z_j)$. Tsiatis pointed out, for example, that the marginal survival functions $S(x) = P(X > x)$ and $S(y) = P(Y > y)$ are not identifiable functions of the multiple decrement function and thus cannot be estimated by classical methods on the basis of a sample of identified minima. Tsiatis demonstrated, through an example with

$$S(x) = \mathrm{e}^{-\lambda x}, \quad x > 0 \tag{9.24}$$

and

$$S(y) = \mathrm{e}^{-\lambda y - \eta y^2}, \quad y > 0\,, \tag{9.25}$$

where $\lambda > 0$ and $\eta > 0$, that different bivariate models can give rise to the same distribution of identified minima. The developments described below will be focused on Bayesian estimation of the nonidentifiable parameter $S(x)$. Similar developments are possible for estimating $S(y)$ and for estimating $S(x,y)$.

We will model the true prior G^* as degenerate at the bivariate exponential (BVE) survival function

$$S^*(x,y) = \mathrm{e}^{-\lambda_1^* x - \lambda_2^* y - \lambda_3^* \max(x,y)}, \quad x > 0,\ y > 0\,. \tag{9.26}$$

The operational prior distribution on $S(x,y)$ is a Dirichlet process whose prior measure α is a scalar multiple of the BVE survival function given by

$$\overline{G}(x,y) = \mathrm{e}^{-\lambda_1 x - \lambda_2 y - \lambda_3 \max(x,y)}, \quad x > 0,\ y > 0\,, \tag{9.27}$$

with the parameters in $\{\lambda^*\}$ and $\{\lambda\}$ all nonnegative. Neath and Samaniego (1996b) demonstrate that the limiting posterior estimator $\widehat{S}_\infty^G(x)$ of $S(x)$ is superior to the prior estimate $\overline{G}(x)$ uniformly in $x \in (0,\infty)$ and uniformly in the vectors $\boldsymbol{\lambda}^*$ and $\boldsymbol{\lambda} \in (0,\infty)^3$. Their result to this effect is stated below.

Theorem 9.4. *Let $\{(Z_i, \delta_i)\}$ be an infinite sequence of i.i.d. identified minima drawn from the bivariate multiple decrement function S^* given in* (9.26)*, where $\lambda_i^* \geq 0$ for $i = 1,2,3$ and $\lambda^* = \sum_{i=1}^3 \lambda_i^*$. Suppose that the operational prior distribution on S is the Dirichlet process prior with parameter measure proportional to the BVE distribution with multiple decrement function $\overline{G}$ given in* (9.27)*, with $\lambda_i \geq 0$ for $i = 1,2,3$. For $x > 0$, denote the corresponding marginal survival functions for X as $S^*(x)$ and $\overline{G}(x)$, where*

$$S^*(x) = \mathrm{e}^{-(\lambda_1^* + \lambda_2^*)x} \quad \textit{and} \quad \overline{G}(x) = \mathrm{e}^{-(\lambda_1 + \lambda_2)x}\,. \tag{9.28}$$

The limiting posterior estimate of $S(x)$ is given by

$$\widehat{S}_\infty(x) = \mathrm{e}^{-\lambda^* x} + \frac{\lambda_2^*}{\lambda^* - (\lambda_1 + \lambda_3)}\left[\mathrm{e}^{-(\lambda_1+\lambda_3)x} - \mathrm{e}^{-\lambda^* x}\right] \quad \textit{if } \lambda^* \neq \lambda_1 + \lambda_3\,, \tag{9.29}$$

and

$$\widehat{S}_\infty(x) = \mathrm{e}^{-\lambda^* x} + \lambda_2^* x \mathrm{e}^{-(\lambda_1+\lambda_3)x} \quad \textit{if } \lambda^* = \lambda_1 + \lambda_3\,. \tag{9.30}$$

Moreover,

$$\left|S^*(x) - \widehat{S}_\infty(x)\right| \leq \left|S^*(x) - \overline{G}(x)\right|$$

for all $x \in (0,\infty)$, $\lambda^* \in (0,\infty)$ and $\boldsymbol{\lambda} \in (0,\infty)^3$. Thus, Bayesian updating provides universal improvement over the prior estimator of $S(x)$ for all values of x, λ^* and $\boldsymbol{\lambda}$.

The result above establishes the uniform superiority of the limiting posterior Bayes estimator of the nonidentifiable parameter $S(x)$ over the prior estimate of that parameter in the special case in which the true multiple decrement function is a BVE distribution and the prior distribution $D(\alpha)$ is centered on a distribution in the BVE class. While there are many questions left to investigate, and there has been, to date, no success in obtaining a general solution to the asymptotic threshold problem comparing prior and limiting posterior Bayes estimators in a competing risks framework, the result above both highlights the distinguished role played by multivariate exponential distributions in that general problem and demonstrates the efficacy of Bayesian updating in this important special case.

It should, of course, be noted that a real threshold does indeed exist in the general problem. Suppose, for example, one wishes to estimate the bivariate multiple decrement function $S^*(x,y)$ when $S^*(x,y)$ has the BVE$(\lambda_1^*, \lambda_2^*, \lambda_3^*)$ distribution and the

Dirichlet prior measure is proportional to $\mathrm{BVE}(\lambda_1, \lambda_2, \lambda_3)$. If, for $x > y$, the vectors $\boldsymbol{\lambda}^*$ and $\boldsymbol{\lambda}$ satisfy the conditions

$$(\lambda_1 + \lambda_3)x + \lambda_2 y = (\lambda_1^* + \lambda_3^*)x + \lambda_2^* y \quad \text{and} \quad \lambda_1 + \lambda_3 \neq \lambda_1^* + \lambda_3^* , \tag{9.31}$$

then the prior estimate $\overline{G}(x,y)$ will be equal to the true multiple decrement function $S^*(x,y)$ while the limiting posterior Bayes estimate $\widehat{S}_\infty(x,y)$ will generally differ from $S^*(x,y)$, being dependent on the exact specification of the conditional distribution of X given $Y \in (y,x)$ and $X > Y$.

Exercise 9.4. Show that the function in (9.27) is a valid survival function, evaluate the probability that $X = Y$, and obtain the marginal density functions of X and of Y.

9.6 Bayesian estimation of a nonidentifiable parameter in a reliability context

The classical theory of stress–strength testing in reliability theory is generally formulated as follows. The (breaking) strength of a given material, such as a welded steel bar (or "rebar") used in the construction of buildings, bridges and the like, is modeled as a positive random variable Y. The random stress to which the material is subjected is modeled as a positive random variable X, independent of Y. The reliability of the material is then captured by $R = P(X < Y)$, the probability that the material survives the stress that is placed on it. If $X \sim F$ and $Y \sim G$, then R may be calculated as

$$R = \int_0^\infty F(y)\, \mathrm{d}G(y) . \tag{9.32}$$

Given independent samples of stresses and strengths, that is, given $X_1, X_2, \ldots, X_m \overset{iid}{\sim} F$ and $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} G$, the statistic U/mn is an unbiased nonparametric estimator of the reliability R, where U is the Mann–Whitney statistic given by

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} I\{X_i \leq Y_j\} . \tag{9.33}$$

Birnbaum (1956) appears to have been the first to study the estimation of R carefully, providing a lower confidence bound for R; in Birnbaum and McCarty (1958), formulae were given for calculating the required sample size associated with a certain desired precision in one's estimate. Subsequently, a good deal of work has appeared treating other versions of the stress–strength problem, including a variety of parametric formulations and including both classical and Bayesian approaches to it. Johnson (1988) and Kotz *et al.* (2003) provide excellent overviews of this work.

The version of this problem that I will discuss here is one which is "plagued" with model nonidentifiability. Identifiability problems in the model (F, G) arise when either the stresses (X) or the strengths (Y), or possibly both, cannot be directly observed. Such a circumstance might occur when a structure fails when it is subjected

to various stresses. A bridge collapse following an earthquake would be a "concrete" example (People are always asking for concrete examples!). The issue to be investigated here is the extent to which the stress and/or strength distributions can be meaningfully estimated on the basis of autopsy data, that is, on the basis of data obtained from the examination of stressed material following a structure's failure. Let us first consider an alternative formulation of the stress–strength problem in which the observable quantity available for analysis consists of random pair (Z,Y), where, as before, Y measures an item's breaking strength and Z is a binary variable which records whether or not an item survived the stress to which it was subjected. The connection between the models for (X,Y) and for (Z,Y) will now be examined.

Let Y be a continuous variable with distribution G, and let Z, given $Y = y$, be a Bernoulli random variable, that is, $Z|Y = y \sim \mathscr{B}(1, p(y))$. The parameter $p(y)$ is interpreted as the conditional probability of surviving a random stress to which an item having known strength $Y = y$ is subjected. The joint probability function of the pair (Z,Y) is

$$f(z,y) = g(y)[p(y)]^z[1-p(y)]^{1-z}\,, \tag{9.34}$$

where g is the density function of Y and $p(y) = P(Z = 1 \mid Y = y)$. The model in (9.34) has no necessary connection to the original model (F,G), but if one is trying to model the same problem, then we must acknowledge that $P(Z = 1 \mid Y = y) = P(X < Y \mid Y = y) = F(y)$. We thus may reformulate the autopsy model as

$$Y \sim G \quad \text{and} \quad Z \sim \mathscr{B}(1,F(y))\,, \tag{9.35}$$

where F is the distribution of the latent variable X representing a random stress. The observation (Z,Y) carries less information than the observation (X,Y), as given X and Y, the variable Z is degenerate, taking the value 1 if $X < Y$ and the value 0 if $X \geq Y$.

We now turn to the estimation problem of interest. The main issues to be examined here relate to identifiability, and the rather simple models we will consider will serve us well in illustrating this particular feature and in outlining the Bayesian solution to the estimation of the model's nonidentifiable parameters. We will take both F and G above to be exponential distributions, that is, we shall assume that

$$F(x) = 1 - \mathrm{e}^{-x/\nu}, \quad x > 0 \tag{9.36}$$

and

$$G(y) = 1 - e^{-y/\mu}, \quad y > 0\,. \tag{9.37}$$

Now suppose that $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} G$ and that $Z_1, Z_2, \ldots, Z_n$ are Bernoulli random variables whose conditional distributions are given by

$$Z_i|Y_i = y_i \sim \mathscr{B}(1,F(y_i)) \quad \text{for } i = 1,\ldots,n\,. \tag{9.38}$$

We will treat the problem of estimating the unknown parameters of the model above based on the autopsy data $\{Z_i\}$ alone, as in the context we have mentioned earlier, where data is collected on the material (say, rebar) in a collapsed structure

such as a bridge, an elevated highway or a building. Assume that neither the Xs nor the Ys are available for observation. We therefore must work with the marginal distribution of the Zs. It is a well-known fact that the Zs are marginally independent and have the common distribution $\mathscr{B}(1,R)$, where

$$\begin{aligned} R &= P(X<Y) \\ &= \int_0^\infty P(Y \geq x) f(x)\, \mathrm{d}x \\ &= \int_0^\infty \frac{1}{\nu} \exp\left\{ -\frac{(\mu+\nu)x}{\mu\nu} \right\} \mathrm{d}x \\ &= \frac{\mu}{\mu+\nu}\,. \end{aligned} \tag{9.39}$$

It is clear that the parameter pair (μ,ν) is not an identifiable parameter of the distribution of the observed Zs. Indeed, it is not possible to distinguish among parameter pairs in the equivalence class $\{(\mu,\nu) \mid \mu = M\nu\}$, where M is a fixed, known constant. Thus, a Bayesian treatment of this estimation problem is the only available option. We will take as the operational prior a particular version of the bivariate Pareto distribution having density

$$f(u,w) = \frac{\alpha(\alpha+1)}{(u+w+1)^{\alpha+2}}, \quad u>0,\ w>0\,. \tag{9.40}$$

In many applications of the stress–strength model, there is an assumed lower bound on the monitored stress to which the material under study would be subjected. The stress levels of interest are generally quite a bit higher than the stress applied on a daily basis. An example of these two ideas is the stress to which a bridge is subjected due to daily traffic in contrast with the stress due to an earthquake or other "act of God." We will employ a version of the model above which allows the mean strain μ to be unconstrained in the interval $(0,\infty)$ but we will assume the mean stress ν has a known lower bound (taken here, without loss of generality, to be 1). The prior model for the pair (μ,ν) can be obtained from the Pareto model in (9.40) by the transformation $\mu = u$ and $\nu = w+1$. The resulting joint density of (μ,ν) is then

$$f(\mu,\nu) = \frac{\alpha(\alpha+1)}{(\mu+\nu)^{\alpha+2}}, \quad \mu>0,\ \nu>1\,. \tag{9.41}$$

Given the autopsy data $Z_1, Z_2, \ldots, Z_n$, one may restrict one's attention to the sufficient statistic $S = \sum_{i=1}^n Z_i$ for the parameter R. The joint density of (S,μ,ν) is thus given by

$$f(s,\mu,\nu) = \binom{n}{s} \frac{\alpha(\alpha+1)\mu^s \nu^{n-s}}{(\mu+\nu)^{\alpha+n+2}}, \quad \text{for } s=0,1,\ldots,n,\ \mu>0,\ \nu>1\,. \tag{9.42}$$

From (9.42), we may identify the Bayes estimator of (μ,ν), with respect to squared error loss, as

$$E(\mu|s) = \frac{I(s+1, n-s, \alpha+1)}{I(s, n-s, \alpha+2)} \tag{9.43}$$

and

$$E(\nu|s) = \frac{I(s, n-s+1, \alpha+1)}{I(s, n-s, \alpha+2)} \tag{9.44}$$

where

$$I(a,b,c) = \int_0^\infty \int_1^\infty \frac{\mu^a \nu^b}{(\mu+\nu)^{a+b+c}} \, \mathrm{d}\nu \mathrm{d}\mu \, , \tag{9.45}$$

with $a, b, c \geq 0$. To evaluate the latter integral, we use the transformation $t = \mu/(\mu+\nu)$ and $z = \nu$, for which the absolute value of the Jacobian is $|J| = z/t^2$, to obtain

$$I(a,b,c) = \int_0^1 \int_1^\infty \frac{t^{b+c-2}(1-t)^a}{z^{c-1}} \, \mathrm{d}z \, \mathrm{d}t = \frac{1}{c-2} \frac{\Gamma(b+c-1)\Gamma(a+1)}{\Gamma(a+b+c)} \, . \tag{9.46}$$

The Bayes estimators of μ and ν are then easily identified as

$$\widehat{\mu} = \frac{\alpha}{\alpha-1} \times \frac{s+1}{\alpha+n-s} \tag{9.47}$$

and

$$\widehat{\nu} = \frac{\alpha}{\alpha-1} \, , \tag{9.48}$$

formulae that are applicable provided that $\alpha > 1$. We thus see that the Bayesian approach yields estimators of the parameters μ and ν, and also yields, informally, an estimator $\widehat{M}$ of the scalar multiple M that defines the equivalence class containing the true value of (μ, ν):

$$\widehat{M} = \frac{\widehat{\nu}}{\widehat{\mu}} = \frac{\alpha+n-s}{s+1} \, . \tag{9.49}$$

Since M is an identifiable parameter of the exponential stress–strength model based on the data $\{Z_i\}$, the estimator $\widehat{M}$ may be compared to the MLE $(n-s)/s$ of M.

The fact that the prior model plays a major role in this estimation problem is clear from (9.47) and (9.48), as the estimator of ν is independent of the data and depends only on the prior in (9.41). A different choice of operational prior can make both estimates data dependent, but the strong influence of the prior will still manifest itself, albeit in a different way. The prior utilized above would be recommended for use only when the experimenter has substantial confidence in his prior information about ν and thus is primarily interested in estimating the unknown mean strength μ.

The reader will note that we have not included here a comparison of the prior and posterior estimates of the pair (μ, ν). This threshold problem is quite tractable and is left as an exercise.

Exercise 9.5. Consider the exponential model for stress–strength testing in Section 9.5. Compare the performance of the Bayes estimator of (μ, ν) in (9.47) and (9.48) to the prior estimator (μ_0, ν_0), the mean of the Pareto prior given in (9.41).

10

Improving on Standard Bayesian and Frequentist Estimators

10.1 The empirical Bayes framework

Suppose that an experiment of interest will yield the outcome X, where X has been modeled as having the probability distribution F_θ depending on the scalar parameter θ. The statistician is prepared to estimate θ by the estimator $\widehat{\theta} = \widehat{\theta}(X)$, where $\widehat{\theta}$ might be either a Bayesian or a frequentist estimator, depending on the statistician's inclination. Suppose that before this estimation process is completed, the statistician becomes aware of the outcome Y of a "similar" experiment. The question then naturally arises: can the information obtained in the other experiment be exploited to provide a better estimator of θ than $\widehat{\theta}$? If it can, then the opportunity afforded to the statistician to improve upon his initial strategy should not be squandered! The situation I have outlined here is both intriguing and seductive, but it is also vague, involving the as-yet-undefined term "similar" as well as dependent on suppositions that might, in many circumstances, be found to be unrealistic. In this chapter, we will examine a scenario in which opportunities of the sort above tend to arise. The empirical Bayes framework, the quintessential statistical setting in which one may reliably learn from similar experiments, was introduced by Robbins at the Berkeley Symposium on Probability and Statistics in 1955, and was developed further in Robbins (1964). This section is aimed at presenting the empirical Bayes "model" and discussing its statistical implications.

Let us posit the existence of a sequence of $(k+1)$ independent experiments, with $k \geq 1$, in which the distributions of the observable quantities $X_1, \ldots, X_k, X_{k+1}$ may vary. We will think of the first k experiments as having occurred in the past, and we will refer to the $(k+1)$st experiment as "the current experiment." These experiments will also be postulated to be "similar" in the following sense. We will assume that the data are distributed as

$$X_i \sim F_{\theta_i}, \quad i = 1, \ldots, k, k+1 \tag{10.1}$$

where the parameters $\{\theta_i,\ i = 1, \ldots, k+1\}$ are viewed as independent random draws from a true but unknown "prior" distribution G_0, that is, we will assume that

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_10,

$$\theta_1, \ldots, \theta_k, \theta_{k+1} \overset{iid}{\sim} G_0 \,. \tag{10.2}$$

The independence of the $(k+1)$ experiments alluded to above is interpreted simply as the independence of the random pairs $\{(X_i, \theta_i),\ i = 1, \ldots, k+1\}$. The similarity of the experiments is now clear. The parameters $\{\theta_i,\ i = 1, \ldots, k+1\}$ are "similar" in the sense that they are "generated" from a single fixed random process (e.g., they have the same median, along with other shared characteristics), and if we were to venture a guess at the value of θ in any of these experiments, we would consider the index of the experiment quite irrelevant. Typically, the distributions $\{F_{\theta_i}\}$ are assumed to be members of the same parametric family (though this is just a convenient and generally realistic assumption, not an essential one). Together, the assumptions in (10.1) and (10.2) provide the reasonable expectation that the $(k+1)$ experiments above are similar enough for one to be able to utilize the information in the observations $\{X_1, \ldots, X_k\}$ to improve upon an estimator of θ_{k+1} based on the datum X_{k+1} alone. Doing so is often referred to as "borrowing strength" from similar experiments.

Let us now consider the goal of estimating the parameter θ_{k+1} of the current experiment based on all the available information. We will denote such an estimator by $d_k(\mathbf{X})$, where d_k is a decision rule based on $\{X_1, \ldots, X_{k+1}\}$ in the problem of estimating the unknown parameter θ_{k+1}. Following Robbins, we take our loss function to be squared error and the criterion for assessing the quality of an estimator d as its Bayes risk $r(G_0, d)$ with respect to the unknown true prior G_0, where it is assumed that the distributions G_0 and $\{F_{\theta_i},\ i = 1, \ldots, k, k+1\}$ have finite second moments. Robbins (1956) made a simple observation with quite profound statistical implications. He noted that, while the Bayes estimator $d_{G_0}(X_{k+1})$ with respect to G_0, the best one could hope to do in estimating θ_{k+1} based on the current data, was impossible to identify, one could approximate this estimator and its statistical performance with arbitrary accuracy, as $k \to \infty$, by an estimator $d_k(\mathbf{X})$ depending on all the data. In Robbins' leading example, where F_θ was taken as the Poisson model with mean θ, the Bayes estimator d_{G_0} of θ may be written as

$$d_{G_0}(x) = \frac{(x+1)p_{G_0}(x+1)}{p_{G_0}(x)} \,, \tag{10.3}$$

where $p_{G_0}(\cdot)$ is the marginal probability mass function (pmf) of X. Robbins noted that one could mimic the form of this estimator by its "empirical Bayes" counterpart

$$d_k(x_{k+1}) = \frac{(x_{k+1}+1)p_k(x_{k+1}+1)}{p_k(x_{k+1})} \,, \tag{10.4}$$

where $p_k(\cdot)$ is the empirical pmf based on the observations $X_1, \ldots, X_k$. It is clear from equations (10.3) and (10.4) that for all integers $x \geq 0$, $d_k(x) \to d_{G_0}(x)$ as $k \to \infty$. Johns (1957) showed that d_k was indeed asymptotically optimal in the Bayes sense, that is, that

$$r(G_0, d_k) \to r(G_0, d_{G_0}) \quad \text{as } k \to \infty \,. \tag{10.5}$$

It is worth noting that the estimator in (10.4) was never intended to be used when k is small, as the estimator would clearly be undefined whenever the $(k+1)$st observation happens to differ from any previous observation. This problem vanishes in the limit.

Examples such as that above, and similar developments for broader modeling scenarios, serve to establish the fact that empirical Bayes (EB) estimators will behave well (that is, perform as well as the Bayes estimator with respect to the unknown true prior G_0 based on the current experiment alone) when the number of past experiments is sufficiently large. Note, however, that this assertion differs from the type of asymptotic claim one would generally make about an estimator of interest. It should be noted explicitly that the EB framework permits, and generally includes, multiple observations within each experiment. Thus, a more complete representation of the EB setup would be that given below. The sample available from the ith experiment is

$$X_{i1}, \ldots, X_{in_i} \overset{iid}{\sim} F_{\theta_i} \quad \text{for } i = 1, \ldots, k, k+1 \,. \tag{10.6}$$

Now a result of the sort given in (10.5), with appropriate adjustments made for the precise form of d_k, will generally hold as $k \to \infty$. Interestingly, for a fixed k, such results say nothing about the behavior of the empirical Bayes estimator d_k as the samples sizes $\{n_i\}$ grow. Indeed, for fixed k, even if $n_i \to \infty$ for $i = 1, \ldots, k+1$, the convergence in (10.5) will not generally occur. Still, the potential for exploiting past information in the estimation of the current parameter was made stunningly clear by Robbins, and much productive work associated with the framework he introduced has followed. We will routinely assume, in the remainder of this chapter, that the "similarity" assumptions encapsulated in (10.2) and (10.6) hold. Applications in which these assumptions are relaxed from "similar" to "related" experiments are discussed in Chapter 11. In this section of the present chapter, we will be interested in the characteristics and performance of EB procedures. In subsequent sections, we address the question about when and how past experiments may be used, under EB assumptions, to improve upon a standard estimator (Bayesian or frequentist) under consideration in the current experiment. Before turning to performance questions, we take a brief digression to discuss some philosophical and theoretical issues.

The term "empirical Bayes" suggests that the methodology is a close cousin to Bayesian methods. This is not actually the case. Inference based on the EB framework is firmly rooted within the classical school of statistics. There is, for instance, no subjective (nor even objective) prior input about unknown parameters. An explicit example of the connection between Bayes and empirical Bayes analyses, and one that was especially insightful regarding the interpretation of the Stein effect, appeared in Efron and Morris's papers (1973a,b). There, it is shown that the James–Stein estimator is in fact an empirical Bayes estimator of a multivariate normal (MVN) mean in which the parameters of the conjugate MVN prior are replaced by a known mean and a particular data-based estimate of the shrinkage parameter. A characteristic of EB analysis is that the consideration of the model parameters being random serves the purpose of quantifying the notion of "similarity," but their common distribution G_0 remains unknown and unknowable throughout the analysis. The potential for defining "similarity" in this fashion becomes clear when one recognizes that the case in which this "true prior" is degenerate represents the case in which the same experiment is replicated $(k+1)$ times. Certainly if G_0 is not too diverse, the experiments involved in (10.6) are indeed similar. But even if G_0 has substantial dispersion, one

might still reasonably consider the $(k+1)$ experiments to be sufficiently similar to be of use; of course, the rate of convergence in (10.5) can be expected to be slower in the latter scenario.

A fully Bayesian version of Robbins' EB framework would place prior distributions on all unknown parameters of the sampling distribution. Assuming (10.2) as in Robbins (1956), with G_0 taken as unknown, a Bayes empirical Bayes (BEB) procedure would put a prior distribution on G_0. Early and notable examples of BEB treatments of statistical problems are Berry and Christensen's (1979) use of a Dirichlet process prior on G_0 in the problem of estimating a binomial parameter and the work of Deely and Lindley (1981) which introduces hierarchical models in the general EB framework, essentially looking at the distribution G_0 as governed by a random parameter. The latter paper had a strong influence in the evolution of a wide-ranging literature on the use of hierarchical modeling in Bayesian inference.

The first issue that comes to mind in the practical application of EB methods in statistical problems is the actual comparative performance of an EB estimator relative to other options (for example, relative to the MLE or the UMVUE based on the $(k+1)$st sample alone). What is generally known is that EB estimators tend to outperform the standard frequentist estimators when k is sufficiently large. A related issue of interest is that of identifying the value k^* defined as

$$k^* = \min_k \, r(G_0, d_k) \leq r(G_0, d^*) \,, \tag{10.7}$$

where d^* represents the frequentist estimator of choice. The body of theoretical results on this latter question is rather sparse. Insights into this matter have largely been derived from simulation studies. Maritz and Lwin (1989, pp. 84–89), for example, compared the performance of seven different EB estimators to that of the MLE and of the Bayes estimator under the modeling assumptions that F_θ is the Poisson distribution with mean θ and the true prior G_0 is a gamma distribution with two unknown parameters. The problem studied set $k = 50$ and $n_i = 1$ for all i. Not surprisingly, the performance of the Robbins estimator d_{50} in (10.4) was poorer than the MLE $d^*(x_{51}) = x_{51}$. However, certain alternative EB estimators, especially those involving some "smoothing," outperformed the MLE. From Maritz and Lwin's study, it appears that, for the right EB estimator, the value of k^* in (10.7) is substantially smaller than 50. In other comparative simulation studies, Canovos (1973) and Bennett (1977) showed that, for certain classes of prior distributions, smooth EB estimators tended to outperform maximum likelihood estimators of the failure rates of exponential and Weibull models, even when the number k of past experiments was quite small. In the next two sections, we will turn to the theoretical study of these types of comparisons. The presentation here is based on the papers by Samaniego and Neath (1996) and Samaniego and Vestrup (1998).

I will conclude this section with a brief overview of the problems which will be investigated in the two sections that follow. I will focus on the modeling scenario treated in Chapter 5 in which a sufficient, unbiased statistic exists for the scalar parameter θ of a distribution belonging to an exponential family and θ is to be estimated under squared error loss. In Section 10.2, we will compare, under an assumed

EB framework, the performance of a Bayes estimator with respect to an operational prior G to the performance of a Bayes estimator with respect to a new prior which is a revised form of the operational prior G which takes the information from past experiments into account. The latter estimator is in the class of BEB estimators which utilize both subjective prior information, through the operational prior G, and data from past experiments, in the development of an estimator of θ. In Section 10.3, we compare the performance of a standard frequentist estimator $\widehat{\theta}$ of θ, based on the random sample drawn in the current experiment, to the performance of alternative estimators that utilize data from past experiments. I will denote the EB alternative to the Bayes estimator d_G as d_{G_k} and the EB alternative to the frequentist estimator $\widehat{\theta}$ as $\widehat{\theta}^{(k)}$. The Bayes risks of the four estimators above relative to the true but unknown prior G_0 are graphed in Figure 10.1 as functions of the number k of past experiments.

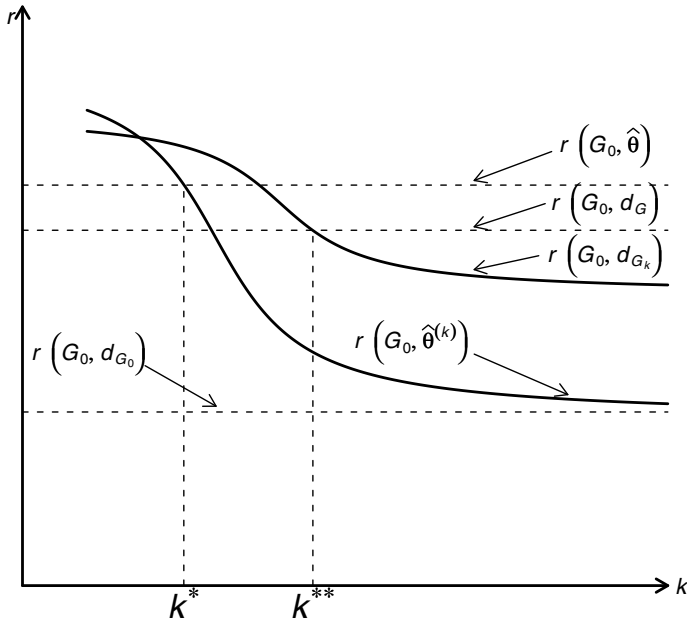

Fig. 10.1. Bayes risks, relative to the true prior G_0, for the Bayes estimator d_{G_0}, for the standard and EB frequentist estimators $\widehat{\theta}$ and $\widehat{\theta}^{(k)}$ and for the Bayes and BEB estimators d_G and d_{G_k} based on the operational prior G

The Bayes risks of these four estimators are pictured in Figure 10.1 against the baseline $r(G_0, d_{G_0})$, the Bayes risk of the optimal estimator wrt the true prior G_0 known. For concreteness, we have assumed in this figure that $r(G_0, d_G) < r(G_0, \widehat{\theta})$, although we know from Chapter 5 that this inequality might be reversed for some

choices of the operational prior G. The ordering of these two risks will, however, play no role in the problems to be investigated in the sequel. Of special interest in Figure 10.1 are the values of the integers k^* and k^{**} along the x-axis. The integer k^* represents the smallest integer for which the EB alternative $\widehat{\theta}^{(k)}$ has a smaller Bayes risk relative to the true prior G_0 than that of the frequentist estimator of choice $\widehat{\theta}$. The integer k^{**} represents the smallest integer for which the BEB estimator d_{G_k} has a smaller Bayes risk wrt G_0 than that of the Bayes estimator d_G wrt the operational prior G. It will be shown in the next two sections that for particular choices of EB and BEB estimators in the problem above, both k^* and k^{**} are equal to one! Thus, even with just a single past experiment that is similar to the current experiment in Robbins' sense, the statistician who properly exploits past data has the opportunity to improve upon standard estimators in the current experiment. In addition to the explicit assumptions made in the theorems that follow, I will tacitly take as given that interchanges in the order of integration, typically justified via Fubini's Theorem, are valid in the context under study.

Exercise 10.1. Let $(X_1,\lambda_1),\ldots,(X_k,\lambda_k),(X_{k+1},\lambda_{k+1})$ be a collection of independent random pairs, where $\lambda_1,\ldots,\lambda_k,\lambda_{k+1} \overset{iid}{\sim} G_0$, the true but unknown distribution of the λs, and, given λ_i, X_i has a geometric distribution with pmf parametrized as

$$p(x|\lambda) = (1-\lambda)\lambda^x \quad \text{for } x = 0,1,2,\ldots \text{ and } 0 < \lambda < 1\,.$$

Show that, under squared error loss, the Bayes estimator of λ with respect to G_0, based on a single geometric observation $X = x$, may be written as

$$\widehat{\lambda}_{G_0} = \frac{p_{G_0}(x+1)}{p_{G_0}(x)}\,,$$

where

$$p_{G_0}(x) = \int_0^1 (1-\lambda)\lambda^x \, \mathrm{d}G_0(\lambda)\,.$$

Obtain an empirical Bayes estimator of λ_{k+1} based on the empirical marginal pmf of the observations in the first k experiments.

10.2 How to be a better Bayesian

Let us suppose that data is available from two similar experiments, one "past" and the other "current," each depending on a scalar parameter θ. We will be interested in estimating the current parameter value using the standard Bayesian approach. Adapting the notation above to the present situation, it is assumed that

$$\theta_1,\theta_2 \overset{iid}{\sim} G_0 \tag{10.8}$$

and that, given θ_i,

$$X_{i1},\ldots,X_{in_i} \overset{iid}{\sim} F_{\theta_i} \text{ for } i = 1,2\,. \tag{10.9}$$

It is also assumed that the pairs $(\theta_1, \mathbf{X}_1)$ and $(\theta_2, \mathbf{X}_2)$ are independent. The setting in which the estimation of θ_2 will be studied is the familiar one featured in Chapter 5, that is, the case in which the distribution F_θ is a member of a one-parameter exponential family and the parameter θ_2 is to be estimated relative to squared error loss. Further, we will assume that the operational prior G is chosen from the standard family of distributions that is conjugate to the class $\{F_\theta\}$. Although the theorem below is stated without these formal assumptions, it is clear from the theorem's assumptions that this distributional scenario is the primary domain of the result's intended applications. We proceed to the section's main result, which is a modest generalization of Theorem 1 from Samaniego and Neath (1996).

Theorem 10.1. *Let $(\theta_1, \mathbf{X}_1)$ and $(\theta_2, \mathbf{X}_2)$ be independent random vectors governed by the distributional assumptions in* (10.8) *and* (10.9)*, where $\{F_{\theta_i},\ i = 1, 2\}$ and G_0 are assumed to have finite second moments. Suppose that G_0 is the true but unknown prior distribution of θ. Let G represent the "operational" prior distribution on θ_2 parametrized by its mean $E_G\theta$ and a weighting parameter α which is characterized by the fact that the Bayes estimator of θ_2 wrt G under squared error loss is given by*

$$\widehat{\theta}_G = \widehat{\theta}_G(\mathbf{X}_2) = \eta\,\widehat{\theta}_2 + (1-\eta)E_G\theta\ , \tag{10.10}$$

where $\eta \in [0, 1)$ and $\widehat{\theta}_2$ is a sufficient and unbiased estimator of θ_2 based on the observation $\mathbf{X}_2$. For $c \in (0, 1)$, let $G^{(1)}$ be the "adjusted" operational prior with mean $(c\widehat{\theta}_1 + (1-c)E_G\theta)$ and weighting parameter η, i.e., $G^{(1)}$ is the prior on θ_2 for which the Bayes estimator of θ_2 wrt $G^{(1)}$ under squared error loss is given by

$$\widehat{\theta}_{G^{(1)}} = \widehat{\theta}_{G^{(1)}}(\mathbf{X}_1, \mathbf{X}_2, c) = \eta\,\widehat{\theta}_2 + (1-\eta)(c\widehat{\theta}_1 + (1-c)E_G\theta)\ , \tag{10.11}$$

where $\widehat{\theta}_1$ is a sufficient and unbiased estimator of θ_1 based on the observation $\mathbf{X}_1$. Then

$$r\left(G_0, \widehat{\theta}_{G^{(1)}}\right) < r\left(G_0, \widehat{\theta}_G\right) \tag{10.12}$$

for any value of the constant c satisfying

$$0 < c < \frac{2(E_{G_0}\theta - E_G\theta)^2}{(E_{G_0}\theta - E_G\theta)^2 + V(\widehat{\theta}_1)}\ ,$$

and $r\left(G_0, \widehat{\theta}_{G^{(1)}}\right)$ is minimized, as a function of c, by

$$c^* = \frac{(E_{G_0}\theta - E_G\theta)^2}{(E_{G_0}\theta - E_G\theta)^2 + V(\widehat{\theta}_1)}\ . \tag{10.13}$$

Proof. Simple algebra shows that $r\left(G_0, \widehat{\theta}_{G^{(1)}}\right)$ may be expressed as a function of c as follows, where the expectation with respect to the distribution of $(\mathbf{X}_i, \theta_i)$ is written as E_i:

$$r(c) = E_1E_2\left(\widehat{\theta}_{G^{(1)}}(\mathbf{X}_2) - \theta_2\right)^2$$
$$= \eta^2 E_1E_2(\widehat{\theta}_2 - \theta_2)^2 + (1-\eta)^2 E_1E_2[(c\widehat{\theta}_1 + (1-c)E_G\theta) - \theta_2]^2 \,, \quad (10.14)$$

an expression which follows from the fact that the cross product term in the quadratic expansion of $r(c)$ vanishes due to the assumed unbiasedness of $\widehat{\theta}_2$ as an estimator of θ_2. The first term in (10.14) may be identified as $\eta^2 E_{\theta_2} V(\widehat{\theta}_2|\theta_2)$ and is independent of the constant c. To show that $r(c)$ is decreasing in c for c in an interval of the form $(0,C)$, it suffices to study the behavior of the second term on the RHS of (10.14). To that end, we rewrite that term as

$$\begin{aligned}
&(1-\eta)^2 E_1E_2[c\widehat{\theta}_1 + (1-c)E_G\theta - \theta_2]^2 \\
&= (1-\eta)^2 E_1E_2[c(\widehat{\theta}_1 - E_G\theta) + (E_G\theta - \theta_2)]^2 \\
&= (1-\eta)^2[c^2 E_2E_1(\widehat{\theta}_1 - E_G\theta)^2 + 2cE_1(\widehat{\theta}_1 - E_G\theta)E_2(E_G\theta - \theta_2) \\
&\qquad\qquad + E_1E_2(\theta_2 - E_G\theta)^2] \\
&= (1-\eta)^2\{c^2[(E_{G_0}\theta - E_G\theta)^2 + V(\widehat{\theta}_1)] - 2c(E_{G_0}\theta - E_G\theta)^2 + V_{G_0}(\theta) \\
&\qquad\qquad + (E_{G_0}\theta - E_G\theta)^2\} \qquad (10.15)
\end{aligned}$$

From (10.15), we obtain

$$r'(c) = \frac{\partial}{\partial c} r(c) = (1-\eta)^2\{2c[(E_{G_0}\theta - E_G\theta)^2 + V(\widehat{\theta}_1)] - 2(E_{G_0}\theta - E_G\theta)^2\} \,. \quad (10.16)$$

If the operational prior is mean correct, that is, if $E_G\theta = E_{G_0}\theta$, then $r'(c) > 0$ for $c > 0$. It follows that the Bayes risk is uniquely minimized, among $c \geq 0$, at $c = 0$, implying that the Bayes estimator $\widehat{\theta}_G$ wrt the original operational prior G outperforms the BEB estimator wrt to the adjusted operational prior $G^{(1)}$, regardless of the choice of $c > 0$. (We note, in passing, that the virtual impossibility of exact mean correctness in any practical application will imply that improvement over the original estimator $\widehat{\theta}_G$ is virtually always possible.) If, on the other hand, $E_G\theta \neq E_{G_0}\theta$, then we may conclude from (10.16) that $r(c)$ is decreasing at $c = 0$, that it achieves its unique minimum value at

$$c^* = \frac{(E_{G_0}\theta - E_G\theta)^2}{(E_{G_0}\theta - E_G\theta)^2 + V(\widehat{\theta}_1)} \,, \quad (10.17)$$

and that the BEB estimator $\widehat{\theta}_{G^{(1)}}$ in (10.11) is superior to the Bayes estimator $\widehat{\theta}_G$ for any value of the constant $c \in (0, 2c^*)$. This completes the proof. ■

For completeness, I will interject here a brief discussion of the fact that the distribution $G^{(1)}$ may legitimately be viewed as a "prior distribution" in the problem of estimating the parameter θ_2. It is not uncommon, in the practice of Bayesian inference, to select one's prior distribution on the basis of the total amount of intuition, experience and expert opinion available before one examines the experimental data

one is given to analyze. In the EB framework, the available experimental data is precisely the data $X_{21}, \ldots, X_{2n_2}$ from the current experiment. Thus, the data from past experiments, here $X_{11}, \ldots, X_{1n_1}$, is simply an additional (perhaps newly found) component of our "experience" which is to be combined with other prior information, represented by G, to form an adjusted prior $G^{(1)}$ aimed at improving our inference. Its legitimacy within the context of Bayesian inference comes from the fact that the consideration of past data chronologically and operationally precedes our treatment of the data associated with the current experiment.

Theorem 10.1 asserts that, when Robbins' EB assumptions hold in the familiar setting of exponential families of sampling distributions, squared error loss and standard conjugate priors, one can essentially always improve upon a given Bayes estimator of the parameter of the current experiment through a process which borrows strength from past experiments by adjusting the prior to incorporate past information (the only possible exception occurring when the original operational prior is mean correct). The alternative prior $G^{(1)}$ in Theorem 10.1 is not necessarily the best one can do, but it does serve to demonstrate that improvement is virtually always possible. It can thus rightly be claimed that, under EB assumptions, a statistician who would use a Bayes estimator to estimate the current parameter based on current data alone has the opportunity to be a "better Bayesian" by exploiting the information available from past experiments.

Theorem 10.1 makes clear that, except in the (virtually impossible) event that the operational prior G is mean correct, there exists a collection of empirical Bayes estimators that will improve upon the Bayes estimator $\widehat{\theta}_G$ relative to the Bayes risk criterion in (10.12). While undoubtedly not unique in this regard, the specific EB estimators that are shown in this theorem to afford the Bayesian a measure of improvement are an immediately attractive class from an intuitive standpoint, as these estimators simply change the prior mean slightly, shrinking it a bit toward the estimator obtained in a similar experiment. The "similarity" gives one confidence that the data utilized from the first experiment carries some useful information about the second. Thus, using it to recalibrate the central tendency of the original prior seems eminently reasonable. But improvement over this approach may well be possible, and is worthy of further investigation.

Another issue that should be stated regarding the interpretation of Theorem 10.1 is that it is most definitely an "existence theorem" rather than a practical prescription for identifying improved estimators. While some practical guidance can be gleaned from the theorem, it must be acknowledged that the constant c^* which gives the greatest improvement and the interval $(0, 2c^*)$ over which $\widehat{\theta}_{G^{(1)}}$ improves upon $\widehat{\theta}_G$ depends on unknown aspects of the experiment. While it may be possible to estimate $V(\widehat{\theta}_1)$ reliably, the distance $|E_{G_0}\theta - E_G\theta|$ between the means of the true and operational priors is simply unknown. These impediments do not preclude placing some reliance on Theorem 10.1 in devising a BEB estimator which is quite likely to improve upon the Bayes estimator $\widehat{\theta}_G$. For example, with a moderate sample size in the first experiment, $V(\widehat{\theta}_1)$ is likely to be small relative to $(E_{G_0}\theta - E_G\theta)^2$. When this is the case, a value of c close to 1 for $\widehat{\theta}_{G^{(1)}}$ in (10.11) can be expected to produce a BEB

estimator that improves upon $\widehat{\theta}_G$. Even in the extreme case in which the size of the term for c^* in (10.17) is difficult to assess, it is worth noting that, under EB assumptions, the BEB estimator $\widehat{\theta}_{G^{(1)}}$ outperforms the EB estimator $\widehat{\theta}_G$ provided that the constant c in (10.11) is sufficiently close to 0. Thus, using a sufficiently conservative guess at the value of c^* will always produce improvement over $\widehat{\theta}_G$.

Finally, let us consider the general case in which the EB framework involves k past experiments. It is already clear that BEB estimators can provide improvement over EB estimators, as Theorem 10.1 indicates that we can toss out all but one of these past experiments and still improve upon $\widehat{\theta}_G$. But, clearly, one should be able to do better. In the result below, we indicate how data from each of k past experiments can be utilized in improving upon the Bayes estimator based solely on data from the current experiment. Let us, then, assume that, for $i = 1, \ldots, k, k+1$, the data available from the ith experiment is given by

$$X_{i1}, \ldots, X_{in_i} \overset{iid}{\sim} F_{\theta_i} , \tag{10.18}$$

where, for each i, the parameters governing these experiments satisfy

$$\theta_1, \ldots, \theta_k, \theta_{k+1} \overset{iid}{\sim} G_0 . \tag{10.19}$$

It is also assumed that the random vectors $\{(\theta_i, \mathbf{X}_i),\ i = 1, \ldots, k, k+1\}$ are mutually independent. I now state, without proof, an extension of Theorem 10.1 to the case of k past experiments.

Theorem 10.2. *Let $\{(\theta_i, \mathbf{X}_i),\ i = 1, \ldots, k, k+1\}$ be mutually independent random vectors satisfying the assumptions in* (10.18) *and* (10.19)*, where $\{F_{\theta_i}\}$ and G_0 are assumed to have finite second moments. Let G represent a prior distribution on θ_{k+1} parametrized by its mean $E_G\theta$ and a weighting parameter α which is characterized by the fact that the Bayes estimator of θ_{k+1} under squared error loss is given by*

$$\widehat{\theta}_G = \widehat{\theta}_G(\mathbf{X}_{k+1}) = \eta\,\widehat{\theta}_{k+1} + (1-\eta)E_G\theta , \tag{10.20}$$

where $\eta \in [0,1)$ and $\widehat{\theta}_{k+1}$ a sufficient and unbiased estimator of θ_{k+1} based on the observation $\mathbf{X}_{k+1}$. For $c \in (0,1)$, let $G^{(k)}$ be the "adjusted" operational prior with mean $(c\widehat{\theta}^* + (1-c)E_G\theta)$ and weighting parameter η, that is, let $G^{(k)}$ be the prior on θ_{k+1} for which the Bayes estimator of θ_{k+1} under squared error loss is given by

$$\widehat{\theta}_{G^{(k)}} = \widehat{\theta}_{G^{(k)}}(\mathbf{X}_1, \ldots, \mathbf{X}_{k+1}, c) = \eta\,\widehat{\theta}_{k+1} + (1-\eta)(c\widehat{\theta}^* + (1-c)E_G\theta) \tag{10.21}$$

with

$$\widehat{\theta}^* = \frac{\sum_{i=1}^k n_i \widehat{\theta}_i}{\sum_{i=1}^k n_i} , \tag{10.22}$$

where, for $i = 1, \ldots, k$, $\widehat{\theta}_i$ is a sufficient and unbiased estimator of θ_i based on the observation $\mathbf{X}_i$. Then, under squared error loss,

$$r\left(G_0, \widehat{\theta}_{G^{(k)}}\right) < r(G_0, \widehat{\theta}_G) \tag{10.23}$$

for any value of the constant c satisfying

$$0 < c < \frac{2(E_{G_0}\theta - E_G\theta)^2}{(E_{G_0}\theta - E_G\theta)^2 + V(\widehat{\theta}^*)}\,, \tag{10.24}$$

and $r\left(G_0, \widehat{\theta}_{G^{(k)}}\right)$ *is minimized, as a function of c, by*

$$c^* = \frac{(E_{G_0}\theta - E_G\theta)^2}{(E_{G_0}\theta - E_G\theta)^2 + V(\widehat{\theta}^*)}\,.$$

Exercise 10.2. Prove Theorem 10.2.

10.3 How to be a finer frequentist

As in the preceding section, we shall assume that data is available from two similar experiments, one "past" and the other "current," each depending on a scalar parameter θ. Here, we will treat a problem that is complementary to the problem studied above, that is, we will assume that the statistician is interested in estimating the current parameter value using frequentist methods. We will retain the basic EB assumptions in (10.8) and (10.9) regarding the parameters θ_1 and θ_2 and regarding the available data $\mathbf{X}_1$ and $\mathbf{X}_2$. It is also assumed that the experiments associated with the random pairs $(\theta_1, \mathbf{X}_1)$ and $(\theta_2, \mathbf{X}_2)$ are independent. The setting in which the estimation of θ_2 will be studied is the familiar one in which the distribution F_θ is a member of a one-parameter exponential family and the parameter θ_2 is to be estimated relative to squared error loss. As in previous results involving comparisons between estimators, these assumptions are not stated explicitly in the results below, but the intended application of these results is to situations satisfying these modeling assumptions. The theorems below differ from the preceding comparative results in that the two estimators we will be interested in are frequentist in nature, with the estimators differing on the basis of whether or not they make use of information from past experiments. The two-experiment EB problem is treated in the following result. Extensions of this theorem to general EB problems follow, with several results stated without proof. The interested reader is referred to Samaniego and Vestrup (1998) for a complete treatment of the latter case. I will now state and prove the main result.

Theorem 10.3. *Let* $(\theta_1, \mathbf{X}_1)$ *and* $(\theta_2, \mathbf{X}_2)$ *be independent random vectors governed by the distributional assumptions in* (10.8) *and* (10.9)*, where the distributions* $\{F_{\theta_i},\ i = 1, 2\}$ *and* G_0 *have finite second moments. Suppose that* $\widehat{\theta}_1$ *and* $\widehat{\theta}_2$ *are sufficient, unbiased estimators of* θ_1 *and* θ_2*, respectively. Consider the estimation of* θ_2 *under squared error loss. Then the linear empirical Bayes estimator* $\widehat{\theta}_c$ *of* θ_2 *given by*

$$\widehat{\theta}_c(\mathbf{X}_2) = c\widehat{\theta}_1 + (1-c)\widehat{\theta}_2\,, \tag{10.25}$$

for $c \in (0, 1)$*, is superior to the standard frequentist estimator* $\widehat{\theta}_2$*, that is,*

$$r(G_0, \widehat{\theta}_c) < r(G_0, \widehat{\theta}_G) \tag{10.26}$$

provided that the constant c satisfies

$$0 < c < \frac{2EV(\widehat{\theta}_2|\theta_2)}{EV(\widehat{\theta}_1|\theta_1) + EV(\widehat{\theta}_2|\theta_2) + 2V(\theta)} , \tag{10.27}$$

where θ is a generic random variable having distribution G_0. The optimal value of the constant c is given by

$$c^* = \frac{EV(\widehat{\theta}_2|\theta_2)}{EV(\widehat{\theta}_1|\theta_1) + EV(\widehat{\theta}_2|\theta_2) + 2V(\theta)} . \tag{10.28}$$

Proof. We may write

$$\begin{aligned} r(c) &= r(G_0, c\widehat{\theta}_1 + (1-c)\widehat{\theta}_2) \\ &= E_{\boldsymbol{\theta},\mathbf{X}} \left(c\widehat{\theta}_1 + (1-c)\widehat{\theta}_2 - \theta_2 \right)^2 \\ &= E_{\boldsymbol{\theta},\mathbf{X}} \left[c(\widehat{\theta}_1 - \theta_2) + (1-c)(\widehat{\theta}_2 - \theta_2) \right]^2 \\ &= c^2 E_{\boldsymbol{\theta},\mathbf{X}} (\widehat{\theta}_1 - \theta_2)^2 + (1-c)^2 E_{\boldsymbol{\theta},\mathbf{X}} (\widehat{\theta}_2 - \theta_2)^2 . \end{aligned} \tag{10.29}$$

The last equality follows from the fact that the cross-product term missing from (10.29) equals zero due to the unbiasedness of $\widehat{\theta}_2$. Elementary calculus shows that the Bayes risk in (10.29) is uniquely minimized by the constant $c^* \in (0,1)$ given by

$$c^* = \frac{E_{\boldsymbol{\theta},\mathbf{X}} (X_2 - \theta_2)^2}{E_{\boldsymbol{\theta},\mathbf{X}} (X_1 - \theta_2)^2 + E_{\boldsymbol{\theta},\mathbf{X}} (X_2 - \theta_2)^2} . \tag{10.30}$$

Indeed, $r(c) < r(0)$ for all $c \in (0, 2c^*)$. The proof will be completed by demonstrating that $2c^*$ is equal to the expression on the RHS of (10.27). Note, first, that by the independence of $(\theta_1, \mathbf{X}_1)$ and $(\theta_2, \mathbf{X}_2)$ and the unbiasedness of $\widehat{\theta}_2$, we have

$$\begin{aligned} E_{\boldsymbol{\theta},\mathbf{X}} (\widehat{\theta}_2 - \theta_2)^2 &= E_{\theta_2} E_{\mathbf{X}_2|\theta_2} (\widehat{\theta}_2 - \theta_2)^2 \\ &= E_{\theta_2} V_{\mathbf{X}_2|\theta_2} (\widehat{\theta}_2|\theta_2) . \end{aligned} \tag{10.31}$$

On the other hand, by the unbiasedness of $\widehat{\theta}_1$, we have

$$\begin{aligned} E_{\boldsymbol{\theta},\mathbf{X}} (\widehat{\theta}_1 - \theta_2)^2 &= E_{\boldsymbol{\theta},\mathbf{X}} (\widehat{\theta}_1 - \theta_1)^2 + E_{\boldsymbol{\theta},\mathbf{X}} (\theta_1 - \theta_2)^2 \\ &= E_{\theta_1} V_{\mathbf{X}_1|\theta_1} (\widehat{\theta}_1|\theta_1)^2 + V_{G_0} (\theta_1 - \theta_2) \\ &= E_{\theta_1} V_{\mathbf{X}_1|\theta_1} (\widehat{\theta}_1|\theta_1)^2 + 2V_{G_0} (\theta) . \end{aligned} \tag{10.32}$$

Substituting (10.31) and (10.32) into (10.30) and subsuming subscripts, we have (10.28). ∎

The optimal constant c^* in Theorem 10.3 reveals that the EB estimator $\widehat{\theta}_c$ should place a substantial amount of weight on the estimator $\widehat{\theta}_1$ based on the past experiment when either the variability in θ is small or the estimator $\widehat{\theta}_2$ is much more variable than $\widehat{\theta}_1$. The case in which the distribution G_0 is degenerate at a point is again of special interest. By virtue of the identity $V(X) = EV(X|Y) + V[E(X|Y)]$, we may, in general, identify the constant c^* above as

$$c^* = \frac{V(\widehat{\theta}_2) - V_{G_0}(\theta)}{V(\widehat{\theta}_1) + V(\widehat{\theta}_2)} \,. \tag{10.33}$$

When the true prior G_0 is degenerate, the optimal weight c^* depends entirely on the relative variability of the estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$. In the common case in which $V(\widehat{\theta}_1)$ and $V(\widehat{\theta}_2)$ take the form σ^2/n_1 and σ^2/n_2, respectively, $c^* = n_1/(n_1+n_2)$, and when the sample sizes are equal, $c^* = 1/2$. This would of course be as expected, since when G_0 is degenerate, the data from both experiments are independent and identically distributed with common distribution F_θ.

Example 10.1. Consider a parametric EB treatment of Robbins' Poisson problem. For $i = 1, 2$, let $X_i|\theta_i \sim P(\theta_i)$, and assume that $G_0 = \Gamma(\alpha, 1)$, a one-parameter gamma distribution, where α is unknown and is both the mean and the variance of the distribution. The posterior distribution of θ_2, given the datum $X_2 = x_2$, is $\Gamma(\alpha + x_2, 1/2)$. If α were taken as known, the (useable) Bayes estimate of θ_2 based on $X_2 = x_2$ would be

$$\widehat{\theta}_2^{G_0} = \frac{\alpha + x_2}{2} \,. \tag{10.34}$$

The EB framework of course assumes that G_0 (or in this case α) is unknown. The "parametric EB" approach to estimating θ_2 involves the estimation of unknown prior parameters using all available data, past and current. In the present example, the mean of the marginal distribution of X is α, so that

$$\widehat{\alpha} = \frac{x_1 + x_2}{2} \tag{10.35}$$

is a natural estimate of α. The corresponding EB estimator of θ_2 results from substituting $\widehat{\alpha}$ above into (10.34), yielding

$$\widehat{\theta}_2^{\widehat{G}_0} = \frac{x_1 + 3x_2}{4} \,. \tag{10.36}$$

It's natural to ask whether the EB estimator in (10.36) outperforms the standard estimator X_2 of θ_2. Since the former is a convex combination of x_1 and x_2, Theorem 10.3 provides the answer. In fact, the optimal constant in Theorem 10.3 is $c^* = 1/4$, so that the EB estimator above not only outperforms X_2 but is the best possible estimator of θ_2 among convex combinations of X_1 and X_2.

We now turn our attention to EB problems in which the number of similar past experiments exceeds 1. The result above guarantees the existence of EB estimators

that will outperform the standard frequentist estimator, as it is always possible to utilize a single past experiment to do so. The following results shed light on how the entire ensemble of past experiments can be used productively in improving upon the estimator $\widehat{\theta}_{k+1}$. The following results are slight generalizations of those developed in Samaniego and Vestrup (1998), but the proofs of both versions are essentially the same. I will thus state these results without proof. The first simply provides an explicit expression for the Bayes risk of the convex combinations of estimators that will be of interest in the sequel.

Theorem 10.4. *Let* $(\mathbf{X}_1, \theta_1), \ldots, (\mathbf{X}_{k+1}, \theta_{k+1})$ *be mutually independent, real-valued random vectors satisfying the following assumptions:*

(i) $\theta_1, \ldots, \theta_k, \theta_{k+1} \overset{iid}{\sim} G_0$*, where* G_0 *has a finite second moment,*
(ii) For $i = 1, \ldots, k, k+1$*,* $X_{i1}, \ldots, X_{in_i} \overset{iid}{\sim} F_{\theta_i}$*, where* F_{θ_i} *has a finite second moment,*
(iii) For $i = 1, \ldots, k, k+1$*,* $\widehat{\theta}_i = \widehat{\theta}_i(\mathbf{X}_i)$ *is an unbiased estimator of* θ_i*.*

Then, under squared error loss, the Bayes risk of the empirical Bayes estimator $\widehat{\theta}_{\mathbf{c}}$ *of* θ_{k+1} *defined as*

$$\widehat{\theta}_{\mathbf{c}} = \sum_{i=1}^{k+1} c_i \widehat{\theta}_i \,, \tag{10.37}$$

with $\sum_{i=1}^{k+1} c_i = 1$*, is*

$$r(G_0, \widehat{\theta}_{\mathbf{c}}) = \sum_{i=1}^{k} c_i^2 [EV(\widehat{\theta}_i|\theta_i) + V(\theta)] + c_{k+1}^2 EV(\widehat{\theta}_{k+1}|\theta_{k+1}) + (1 - c_{k+1})^2 V(\theta) \,, \tag{10.38}$$

where θ *is a generic variable with distribution* G_0*.*

Given the explicit expression for $r(G_0, \widehat{\theta}_{\mathbf{c}})$ in Theorem 10.4, it is now possible, using, for example, the method of Lagrange multipliers, to identify the convex combination of the estimators $\{\widehat{\theta}_i,\ i = 1, \ldots, k, k+1\}$ that minimizes the Bayes risk wrt to G_0.

Theorem 10.5. *Assume that conditions (i)–(iii) of Theorem 10.4 hold. Denote the simplex* $\left\{\mathbf{c} \in [0,1]^{k+1} \;\middle|\; \sum_{i=1}^{k+1} c_i = 1\right\}$ *by* S_{k+1}*. Let* $a_i = EV(\widehat{\theta}_i|\theta_i) + V(\theta)$ *for* $i = 1, \ldots, k$*,* $a_{k+1} = EV(\widehat{\theta}_{k+1}|\theta_{k+1})$ *and* $V = V(\theta)$*. The vector* $\mathbf{c} \in S_{k+1}$ *that minimizes* $r(G_0, \widehat{\theta}_c)$ *is given by*

$$c_i^* = \frac{a_{k+1}}{a_i \left[1 + (a_{k+1} + V) \sum_{j=1}^{k} \frac{1}{a_j}\right]} \,, \quad i = 1, \ldots, k \tag{10.39}$$

and

$$c_{k+1}^* = \frac{1 + V \sum_{j=1}^{k} \frac{1}{a_j}}{1 + (a_{k+1} + V) \sum_{j=1}^{k} \frac{1}{a_j}} \,. \tag{10.40}$$

Theorem 10.5 is an existence theorem. It guarantees that, relative to the Bayes risk criterion $r(G_0,\cdot)$, there is always an EB estimator of θ_{k+1} (in the form of a convex combination of the $(k+1)$ estimators $\{\widehat{\theta}_i,\ i=1,\ldots,k,k+1\}$) that will outperform the standard frequentist estimator $\widehat{\theta}_{k+1}$ that is based solely on the current experiment. However, since the optimizing vector $\mathbf{c}^*$ above depends on unknown parameters, the result can, at best, only provide some useful guidance in the use of the "linear combination strategy" in estimating θ_{k+1}. The following result shows that the standard frequentist estimator $\widehat{\theta}_{k+1}$ is dominated by all estimators of the form $\widehat{\theta}_{\mathbf{c}}$ which place sufficient weight on $\widehat{\theta}_{k+1}$.

Theorem 10.6. *Assume that conditions (i)–(iii) of Theorem 10.4 hold and let V and the constants $\{a_i,\ i=1,\ldots,k,k+1\}$ be defined as in Theorem 10.5. If*

$$c_{k+1} \in \left(\frac{a^* + V - a_{k+1}}{a^* + V + a_{k+1}},\ 1 \right) , \tag{10.41}$$

where $a^* = \max\{a_1,\ldots,a_k\}$, then

$$r(G_0, \widehat{\theta}_{\mathbf{c}}) < r(G_0, \widehat{\theta}_{k+1}) . \tag{10.42}$$

Exercise 10.3. Consider an EB problem in which $X_1 \sim \mathscr{B}(n_1,p_1)$ and $X_2 \sim \mathscr{B}(n_2,p_2)$. Suppose the true prior G_0 is modeled as a one-parameter Beta distribution with mean μ, that is, $G_0 = \mathrm{Be}(\mu K,(1-\mu)K)$, where μ is unknown and K is taken as a fixed known constant. The standard frequentist estimator of p_2 is $\widehat{p}_2 = X_2/n_2$. Show that the EB estimator of the form

$$\widehat{p}_2^{\widehat{G}_0}(\mathbf{X}) = c(X_1/n_1) + (1-c)(X_2/n_2)$$

has a smaller Bayes risk wrt G_0 than $\widehat{p}_2$ for any value of c satisfying

$$0 < c < \frac{2n_1K}{(n_1+n_2)K + 2n_1n_2} .$$

Exercise 10.4. Consider the following version of the problem of estimating a normal mean μ_{k+1} given data from k similar past experiments. Specifically, let $\mathbf{X} \sim \mathscr{N}_{k+1}(\boldsymbol{\mu},\mathbf{I})$, and take the unknown true prior distribution G_0 to be the univariate normal distribution $\mathscr{N}(\mu_0,1)$. Use Theorem 10.5 to show that the EB estimator

$$\widehat{\mu}_{k+1} = \frac{k}{2k+2}\overline{X}_k + \frac{k+2}{2k+2}X_{k+1} ,$$

where $\overline{X}_k = \sum_{i=1}^k X_i/k$ is the best convex EB estimator of μ_{k+1}, that is, show that, in this problem, the optimal vector $\mathbf{c} \in S_{k+1}$ in Theorem 10.5 is

$$c^* = \left(\frac{1}{2k+2},\ldots,\frac{1}{2k+2},\frac{k+2}{2k+2} \right) .$$

11

Combining Data from "Related" Experiments

11.1 Introduction

It is not uncommon that the experimental data that is available in a particular statistical investigation is accompanied by collateral information drawn from other sources. The possibility of exploiting auxiliary information, be it empirical or subjective, in order to improve one's inferences is a challenge that has intrigued statisticians for decades. Indeed, the fields of Bayesian statistics, empirical Bayes inference and meta-analysis each focus on particular prescriptions for appropriately combining information from disparate sources, and each can be thought of as a way of exploiting information auxiliary to the experiment of current interest. Excellent overviews of these varied approaches to combining information include Gaver *et al.* (1992), a treatise which covers general approaches, and the monographs by Hedges and Olkin (1985) on meta-analytic techniques and by Maritz and Lwin (1989) on empirical Bayes methods. Notable contributors to this literature include Fisher (1932), Cochran (1937), Savage (1954), Robbins (1956), Glass (1978) and Deely and Lindley (1981).

The problem to be treated in this chapter bears a strong resemblance to problems that are often treated by EB methods. As in the EB setting, we will assume here that one has data from k past experiments and that the statistician is primarily interested in estimating the parameter associated with the $(k+1)$st (or "current") experiment. However, as we will see, there is an important difference between the two problem types. The following example, which arises in the context of the military acquisitions process (and often, in industrial settings as well), serves as a good illustration of estimation problems involving 'related' rather than 'similar' experiments.

The processes of Developmental Testing (DT) and Operational Testing (OT) in military acquisitions programs are central to the development and adoption of an engineered system which either addresses an application of interest *ab initio* or is thought of as a potential improvement over an existing system. In either case, there are two phases of testing aimed at, first, the development of a new system and, second, the assessment of that system's performance under anticipated field conditions and, ultimately, of the determination of the system's suitability for deployment in the field. In the DT phase, it is common to test and fix (possibly successive) prototypes

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_11,

and to make engineering changes, as needed, with the aim of improved performance. The testing done during DT is often refereed to as "bench testing." One may think of it as the type of testing done under laboratory conditions where only modest efforts (if any at all) are made to simulate the field conditions in which the system would be used, if adopted. During the latter part of the DT phase, it is typically the case that a single "best" prototype has been developed, and that some form of "final" testing is done on that prototype to confirm that it is ready for independent testing under field conditions. If the system passes these final checks, a predetermined number of copies are made and provided for use in the OT phase of the acquisitions process. The OT phase resembles standard statistical practice resulting in test data produced under "normal use" conditions which might involve road conditions, speeds, temperatures, altitudes, etc. that are not feasible to incorporate in the DT phase. The aim of operational testing is to determine whether the performance of the new prototype under normal use (or field) conditions is "effective" and "suitable," terms that generally mean that the prototype will adequately fulfill its intended mission. When there is a system in the field that the new system was developed to replace, OT is also aimed at confirming that the new system will improve upon the performance of the existing system.

In the scenario described above, there are clearly two sources of data concerning the system under discussion. The most relevant data are those produced in the final stages of DT and the data produced during the OT phase. It seems reasonable to seek to draw inferences about the performance of this system using both of these data sources. How to do so, however, presents some new challenges. Individuals who have experience with the military acquisitions process will tell you that the two experiments are not similar in the usual (EB) statistical sense. Indeed, it is almost always the case that systems perform more poorly in OT than they do in DT. This is no doubt due to the harsher conditions under which the OT phase is carried out. While the DT and OT results are not expected to be similar, they are certainly related in some (possibly quantifiable) way, and if the two sources of data on system performance are to be used for any inferential purpose, the relationship between the two experiments would have to be concretely specified.

In the general environment of "related" experiments, it is often the case that the relationships among the experiments are difficult to specify with any confidence. In such circumstances, inferences about the current experiment might well be based solely on the data associated with the current experiment. A frequentist analysis, for example, will often rely upon the MLE or UMVUE of the current parameter based on the current data alone, thus ignoring past data altogether. In the comparisons made later in this chapter, we will examine the efficacy of that strategy as compared to the possibility of using estimators that employ data from past related experiments in a Bayesian fashion. We begin our formal discussion of such questions by describing the basic scenario in which the treatment of related experiments will be developed.

Suppose that the data $X_1, \ldots, X_k, X_{k+1}$ are drawn from $(k+1)$ experiments, the last of which is viewed as the "current" experiment, with the others being "past" experiments. These data are taken to be univariate, though in typical applications, they would represent sufficient statistics for the parameters which conditionally define

their distributions. We will define these $(k+1)$ experiments to be related if the following conditions are satisfied. In this chapter, we will use the notation $G^{(0)}$ rather than G_0 for the true prior distribution of the parameter vector, with subscripts reserved to denote marginal distributions.

(C1) The random pairs $\{(X_i, \theta_i),\ i = 1, \ldots, k+1\}$ associated with $(k+1)$ "related" experiments are independent.

(C2) The independent parameters $\{\theta_i,\ i = 1, \ldots, k+1\}$ have potentially nonidentical "true prior distributions," with $\theta_i \sim G_i^{(0)}$.

(C3) For $i = 1, \ldots, k+1$, $V_{G_i^{(0)}}(\theta_i) = (\sigma_i^{(0)})^2 < \infty$ and $E_{G_i^{(0)}}\theta_i = \mu_i^{(0)}$.

(C4) For each i, the conditional distribution of X_i is represented as $X_i|\theta_i \sim F_{\theta_i}$, and X_i is assumed to be a sufficient statistic for θ_i.

(C5) For $i = 1, \ldots, k+1$, $V(X_i|\theta_i) < \infty$ and $E(X_i|\theta_i) = \theta_i$.

(C6) $V_i^{(0)} = E_{G_i^{(0)}}(V(X_i|\theta_i)) < \infty\ \forall\ i$.

(C7) The $(k+1)$ experiments are linked through a functional relationship of the form $h_j(\boldsymbol{\mu}) = 0$ for $j = 1, \ldots, m$, where m is a fixed positive integer in the set $\{1, \ldots, k\}$.

There are several notable differences between the conditions above and those that define the EB framework considered in Chapter 10. Although the parameters $\{\theta_i\}$ are modeled as independent random variables, the potential difference in their individual distributions is acknowledged in conditions (C2) and (C3). More importantly, a formal relationship among the $(k+1)$ experiments is acknowledged through condition (C7).

While related experiments for arbitrary k will be studied in the next section, the application of primary interest in this chapter involves the DT/OT scenario discussed above. Section 11.3 is focused entirely on the case in which $k = 1$, that is, on the estimation of the parameter θ_2 based on data from the current experiment and, possibly, on data available from a single related past experiment. To aid in our discussion, it will be useful to have a data set to think about and analyze. In Table 11.1, simulated data are shown that were drawn from two distinct exponential distributions (with a specific relationship that will be revealed later). Table 11.2 provides a basic summary of the data in Table 11.1. These data resemble typical outcomes from the OT/DT process in two particular ways. First, they capture the fact that lifetime of a prototype system tends to be larger under DT than under OT. Second, the available OT data is sparser, a common occurrence given the expense involved in OT. Because of this sparseness, effective methods of combining information from DT and OT in estimating OT parameters may be of special importance.

I will return to the analysis of the data below in Section 11.3. I will be especially interested in the comparative performance of two particular estimators of θ_2: the linear Bayes estimator of θ_2 with respect to a given operational prior G, that is, the estimator of the form $\widehat{\theta}_{\mathbf{c}} = c_0 + c_1\overline{X}_1 + c_2\overline{X}_2$ that minimizes $r(G, \widehat{\theta}_{\mathbf{c}})$, and the standard frequentist estimator $\widehat{\theta}_2 = \overline{X}_2$ of θ_2 based on the OT data alone. Before dealing with that comparison formally, the basic tools for this comparative analysis will be

Table 11.1. Simulated exponential life testing data

DT Data		OT Data
28.7335	18.0050	13.4764
21.7593	1.5495	18.6327
6.0077	35.5350	4.5435
46.6829	22.0601	23.5081
7.5756	2.5790	5.3412
11.2651	20.8876	8.3927
16.0805	7.1455	39.9724
8.0645	10.1876	7.7885
9.9661	67.0262	33.1363
41.6649	7.7921	6.1353

Table 11.2. Summary of simulated life testing data

DT Data	OT Data
Model: $\Gamma(1,\theta_1)$	Model: $\Gamma(1,\theta_2)$
$n_1 = 20$	$n_2 = 10$
$\overline{X}_1 = 19.63$	$\overline{X}_2 = 16.09$

developed. In Section 11.2, a general expression (where the number k of past experiments is fixed but arbitrary) is derived for the Bayes risk $r(G^{(0)}, \widehat{\theta}_{\mathbf{c}})$ for arbitrary linear estimators of θ_2. In the main result of Section 11.2, an explicit relationship between the Bayes risks (relative to G) of the best linear Bayes estimator of θ_2 and the standard frequentist estimator of θ_2 is derived. The ratio of these two Bayes risks is a constant which can be explicitly identified. This fact proves useful in assessing the potential for improvement over the frequentist estimator by the use of linear Bayes estimators as tools for combining information in the estimation of the parameter of the current experiment. In Section 11.3, I will explore in detail the estimation of the parameter of the current experiment in the DT/OT problem in which the number k of past experiments is equal to 1. In the final section of this chapter, I summarize the basic findings in Sections 11.1–11.3 and discuss some possible alternative approaches to the estimation problem considered, including the approximation of unrestricted Bayes estimates with respect to Bayes hierarchical priors.

The work to be described, while heretofore unpublished, was part of the long-term collaborative investigations I carried out with my former student Hien Tran and with Dr. Duane Steffey during his frequent visits to UC Davis. I wish to acknowledge their many contributions, often in the form of useful feedback and constructive criticism, to the development of the ideas and results to be presented. Another former student, Eric Vestrup, participated in a more formal way in the specific material covered below. On one of his several summer visits to Davis while he was on the faculty at De Paul University, he worked on generalizing results I had obtained under certain

simplifying assumptions. His essential contributions to the results of Section 11.2 are gratefully acknowledged.

11.2 A linear Bayesian approach to treating related experiments.

I will take conditions (C1)–(C7) above as the defining characteristics of *related statistical experiments*. Condition (C7), when made concrete, may be utilized in drawing inferences based on data from specific related experiments. An example of this process is treated in the next section. Here, I will concentrate on the implications of the structure embedded in conditions (C1)–(C6). Our first order of business is the development of an explicit expression for the Bayes risk of linear estimators of the form

$$\widehat{\theta}_{\mathbf{c}} = c_0 + \sum_{i=1}^{k+1} c_i X_i \tag{11.1}$$

with respect to a given operational prior distribution. In all results in Sections 11.2 and 11.3 in which Bayes risks are computed or discussed, I make the tacit assumption that the underlying criterion for gauging estimation error is *squared error loss*. Regarding notation, we make note of the fact that whenever the summation $\sum_{i\neq j}$ is utilized in the sequel, it is assumed that the sum is taken over all unequal values of i and j in the set $\{1,\ldots,k+1\}$.

Theorem 11.1. *Suppose that the random pairs* $(\mathbf{X}_1,\theta_1),\ldots,(\mathbf{X}_{k+1},\theta_{k+1})$ *obey the conditions in Section 11.1 defining* $(k+1)$ *related experiments, with* $G^{(0)}$ *replaced by the operational prior G in (C3) and (C6). Under conditions (C1)–(C6), the Bayes risk of the linear estimator in* (11.1) *wrt a prior G on* $\boldsymbol{\theta}$ *for which* $E_G\theta_i^2 < \infty\ \forall\ i$ *is given by*

$$\begin{aligned} r(G,\widehat{\theta}_{\mathbf{c}}) = c_0^2 + \sum_{i=1}^{k+1} c_i^2 V_i + \sum_{i=1}^{k+1} c_i^2(\sigma_i^2+\mu_i^2) + \sum_{i\neq j} c_i c_j \mu_i \mu_j + \sigma_{k+1}^2 + \mu_{k+1}^2 \\ + 2c_0 \sum_{i=1}^{k+1} c_i\mu_i - 2c_0\mu_{k+1} - 2c_{k+1}\sigma_{k+1}^2 - 2\sum_{i=1}^{k+1} c_i\mu_i\mu_{k+1}\ , \end{aligned} \tag{11.2}$$

where the parameters $\{\mu_i,\sigma_i,\ i=1,\ldots,k+1\}$ *and the values* $\{V_i,\ i=1,\ldots,k+1\}$ *are defined as in (C3) and (C6), but relative to G rather than to* $G^{(0)}$.

Proof. By definition,

$$r(G,\widehat{\theta}_{\mathbf{c}}) = E_{\boldsymbol{\theta}} E_{\mathbf{X}|\boldsymbol{\theta}} \left\{ c_0 + \sum_{i=1}^{k+1} c_i X_i - \theta_{k+1} \right\}^2 . \tag{11.3}$$

We may evaluate the inner expectation (or risk function $R(\boldsymbol{\theta},\widehat{\theta}_{\mathbf{c}})$) in (11.3) as

$$R(\boldsymbol{\theta}, \widehat{\theta}_{\mathbf{c}}) = c_0^2 + \sum_{i=1}^{k+1} c_i^2 E_{\mathbf{X}|\boldsymbol{\theta}}(X_i^2) + \sum_{i \neq j} c_i c_j E_{\mathbf{X}|\boldsymbol{\theta}}(X_i X_j) + \theta_{k+1}^2 + 2c_0 \sum_{i=1}^{k+1} c_i \theta_i - 2c_0 \theta_{k+1} - 2\theta_{k+1} \sum_{i=1}^{k+1} c_i \theta_i \,. \tag{11.4}$$

Since

$$E_{\mathbf{X}|\boldsymbol{\theta}}(X_i^2) = V_{\mathbf{X}|\boldsymbol{\theta}}(X_i) + \theta_i^2$$

and, by assumptions (C1) and (C5),

$$E_{\mathbf{X}|\boldsymbol{\theta}}(X_i X_j) = \theta_i \theta_j \,,$$

the expression in (11.4) may be rewritten as

$$R(\boldsymbol{\theta}, \widehat{\theta}_{\mathbf{c}}) = c_0^2 + \sum_{i=1}^{k+1} c_i^2 V_{\mathbf{X}|\boldsymbol{\theta}}(X_i) + \sum_{i=1}^{k+1} c_i^2 \theta_i^2 + \sum_{i \neq j} c_i c_j \theta_i \theta_j + \theta_{k+1}^2 + 2c_0 \sum_{i=1}^{k+1} c_i \theta_i - 2c_0 \theta_{k+1} - 2\theta_{k+1} \sum_{i=1}^{k+1} c_i \theta_i \,. \tag{11.5}$$

In applying the expectation wrt the distribution G to the risk function in (11.5), we note that, using condition (C1) and the notation of (C3) and (C6),

$$E_{\boldsymbol{\theta}} V_{\mathbf{X}|\boldsymbol{\theta}}(X_i) = V_i, \quad E_{\boldsymbol{\theta}} \theta_i^2 = \sigma_i^2 + \mu_i^2 \quad \text{and} \quad E_{\boldsymbol{\theta}}(\theta_i \theta_j) = \mu_i \mu_j \text{ for } i \neq j \,. \tag{11.6}$$

The desired result then follows by applying the expectation $E_{\boldsymbol{\theta}}$ to the expression in (11.5), making appropriate substitutions:

$$\begin{aligned} r(G, \widehat{\theta}_{\mathbf{c}}) &= E_{\boldsymbol{\theta}} R(\boldsymbol{\theta}, \widehat{\theta}_{\mathbf{c}}) \\ &= c_0^2 + \sum_{i=1}^{k+1} c_i^2 V_i + \sum_{i=1}^{k+1} c_i^2 (\sigma_i^2 + \mu_i^2) + \sum_{i \neq j} c_i c_j \mu_i \mu_j + \sigma_{k+1}^2 + \mu_{k+1}^2 \\ &\quad + 2c_0 \sum_{i=1}^{k+1} c_i \mu_i - 2c_0 \mu_{k+1} - 2c_{k+1} \sigma_{k+1}^2 - 2 \sum_{i=1}^{k+1} c_i \mu_i \mu_{k+1} \,. \end{aligned}$$ ■

The main result of this section provides a simple connection between the Bayes risk wrt G of the linear Bayes estimator of θ_{k+1} and the Bayes risk of the standard frequentist estimator X_{k+1} of θ_{k+1}.

Theorem 11.2. *Suppose that the conditions in Theorem 11.1 hold. Let $\widehat{\theta}_{\mathbf{c}^*}$ be the linear Bayes estimator of θ_{k+1} wrt the prior G, that is, the estimator of the form $\widehat{\theta}_{\mathbf{c}}$ in* (11.1) *that minimizes $r(G, \widehat{\theta}_{\mathbf{c}})$. Then*

$$r(G, \widehat{\theta}_{\mathbf{c}^*}) = c_{k+1}^* r(G, X_{k+1}) \,. \tag{11.7}$$

Proof. Straightforward calculus and algebra show that the values of $c_0, c_1, \ldots, c_{k+1}$ that minimize $r(G, \widehat{\theta}_{\mathbf{c}^*})$ are the unique solutions $c_0^*, c_1^*, \ldots, c_{k+1}^*$ of the following $(k+2)$ equations:

$$c_0^* = \mu_{k+1} - \sum_{i=1}^{k+1} c_i^* \mu_i \tag{11.8}$$

$$c_i^* = (V_i + \sigma_i^2 + \mu_i^2)^{-1} \left\{ \mu_i \mu_{k+1} - c_0^* \mu_i - \sum_{\substack{j \neq i \\ j=1,\ldots,k+1}} c_j^* \mu_i \mu_j \right\} \text{ for } i = 1, \ldots, k \tag{11.9}$$

and

$$c_{k+1}^* = (V_{k+1} + \sigma_{k+1}^2 + \mu_{k+1}^2)^{-1} \left\{ \sigma_{k+1}^2 + \mu_{k+1}^2 - c_0^* \mu_{k+1} - \sum_{j=1}^{k} c_j^* \mu_{k+1} \mu_j \right\} . \tag{11.10}$$

From (11.2), we may write the Bayes risk of the linear Bayes estimator $\widehat{\theta}_{\mathbf{c}^*}$ as

$$\begin{aligned} r(G, \widehat{\theta}_{\mathbf{c}^*}) = c_0^{*2} + \sum_{i=1}^{k+1} c_i^{*2} V_i + \sum_{i=1}^{k+1} c_i^{*2} (\sigma_i^2 + \mu_i^2) + \sum_{i \neq j} c_i^* c_j^* \mu_i \mu_j + \sigma_{k+1}^2 + \mu_{k+1}^2 \\ + 2c_0^* \sum_{i=1}^{k+1} c_i^* \mu_i - 2c_0^* \mu_{k+1} - 2c_{k+1}^* \sigma_{k+1}^2 - 2 \sum_{i=1}^{k+1} c_i^* \mu_i \mu_{k+1} , \end{aligned} \tag{11.11}$$

where $\{c_i^*,\ i = 1, \ldots, k+1\}$ are the coefficients of $\widehat{\theta}_{\mathbf{c}^*}$. In simplifying the expression in (11.11), we will utilize the following easily verified identities:

$$c_0^{*2} = \left(\mu_{k+1} - \sum_{i=1}^{k+1} c_i^* \mu_i \right)^2 , \tag{11.12}$$

$$\begin{aligned} \sum_{i=1}^{k+1} c_i^{*2} (V_i + \sigma_i^2 + \mu_i^2) = \sum_{i=1}^{k} c_i^* \left\{ \mu_i \mu_{k+1} - c_0^* \mu_i - \sum_{\substack{j \neq i \\ j=1,\ldots,k+1}} c_j^* \mu_i \mu_j \right\} \\ + c_{k+1}^* \left\{ \sigma_{k+1}^2 + \mu_{k+1}^2 - c_0^* \mu_{k+1} - \sum_{j=1}^{k} c_j^* \mu_{k+1} \mu_j \right\} \end{aligned} \tag{11.13}$$

and

$$c_0^* \left(\sum_{i=1}^{k+1} c_i^* \mu_i - \mu_{k+1} \right) = -c_0^{*2} . \tag{11.14}$$

In addition to these identities, we recall the definition of V_i as it pertains to the current experiment, that is,

$$V_{k+1} = E_{\theta_{k+1}} V_{X_{k+1}|\theta_{k+1}} (X_{k+1} | \theta_{k+1}) = E_{\theta_{k+1}} E_{X_{k+1}|\theta_{k+1}} (X_{k+1} - \theta_{k+1})^2 = r(G,\, X_{k+1}) . \tag{11.15}$$

From (11.14), we see that $2c_0^*\sum_{i=1}^{k+1} c_i^*\mu_i - 2c_0^*\mu_{k+1}$ in (11.11) may be replaced by $-2c_0^{*2}$, yielding the slightly simplified expression

$$r(G,\widehat{\theta}_{\mathbf{c}^*}) = -c_0^{*2} + \sum_{i=1}^{k+1} c_i^{*2}(V_i+\sigma_i^2+\mu_i^2) + \sum_{i\neq j} c_i^* c_j^* \mu_i\mu_j + \sigma_{k+1}^2 + \mu_{k+1}^2 - 2c_{k+1}^*\sigma_{k+1}^2 - 2\sum_{i=1}^{k+1} c_i^*\mu_i\mu_{k+1}\,. \tag{11.16}$$

Upon applying (11.13), we are led to the equivalent expression

$$\begin{aligned} r(G,\widehat{\theta}_{\mathbf{c}^*}) = -c_0^{*2} &+ \sum_{i=1}^{k} c_i^* \left\{ \mu_i\mu_{k+1} - c_0^*\mu_i - \sum_{\substack{j\neq i \\ j=1,\ldots,k+1}} c_j^*\mu_i\mu_j \right\} \\ &+ c_{k+1}^* \left\{ \sigma_{k+1}^2 + \mu_{k+1}^2 - c_0^*\mu_{k+1} - \sum_{j=1}^{k} c_j^*\mu_{k+1}\mu_j) \right\} + \sum_{i\neq j} c_i^* c_j^* \mu_i\mu_j \\ &+ \sigma_{k+1}^2 + \mu_{k+1}^2 - 2c_{k+1}^*\sigma_{k+1}^2 - 2\sum_{i=1}^{k+1} c_i^*\mu_i\mu_{k+1}\,. \end{aligned} \tag{11.17}$$

It is useful to expand (11.17) into the following expression in ten terms:

$$\begin{aligned} r(G,\widehat{\theta}_{\mathbf{c}^*}) = -c_0^{*2} &+ \sum_{i=1}^{k} c_i^*\mu_i\mu_{k+1} - c_0^*\sum_{i=1}^{k} c_i^*\mu_i - \sum_{i=1}^{k} c_i^* \sum_{\substack{j\neq i \\ j=1,\ldots,k+1}} c_j^*\mu_i\mu_j \\ &+ c_{k+1}^*(\sigma_{k+1}^2+\mu_{k+1}^2) - c_0^*c_{k+1}^*\mu_{k+1} - c_{k+1}^*\sum_{j=1}^{k} c_j^*\mu_{k+1}\mu_j \\ &+ \sum_{i\neq j} c_i^*c_j^*\mu_i\mu_j + (1-2c_{k+1}^*)(\sigma_{k+1}^2+\mu_{k+1}^2) - 2\sum_{i=1}^{k} c_i^*\mu_i\mu_{k+1}\,. \end{aligned} \tag{11.18}$$

Combining the second term in (11.18)) with the fifth and the third term with the sixth, and noting that the fourth, seventh and eighth terms sum to zero, we obtain the somewhat more manageable Bayes risk expression below:

$$r(G,\widehat{\theta}_{\mathbf{c}^*}) = -c_0^{*2} + c_{k+1}^*\sigma_{k+1}^2 + \sum_{i=1}^{k+1} c_i^*\mu_i\mu_{k+1} - c_0^*\sum_{i=1}^{k+1} c_i^*\mu_i + (1-2c_{k+1}^*)(\sigma_{k+1}^2+\mu_{k+1}^2) - 2\sum_{i=1}^{k} c_i^*\mu_i\mu_{k+1}\,. \tag{11.19}$$

The latter expression can be further simplified via the following algebraic steps:

$$r(G,\widehat{\theta}_{\mathbf{c}^*}) = -c_0^{*2} + c_{k+1}^*(\sigma_{k+1}^2+\mu_{k+1}^2) - \sum_{i=1}^{k} c_i^*\mu_i\mu_{k+1}$$
$$- c_0^* \sum_{i=1}^{k+1} c_i^*\mu_i + (1-2c_{k+1}^*)(\sigma_{k+1}^2+\mu_{k+1}^2) \tag{11.20}$$
$$= -c_0^{*2} + \sigma_{k+1}^2 + \mu_{k+1}^2 - \sum_{i=1}^{k} c_i^*\mu_i\mu_{k+1} - c_0^* \sum_{i=1}^{k+1} c_i^*\mu_i$$
$$- c_{k+1}^*(\sigma_{k+1}^2+\mu_{k+1}^2) \tag{11.21}$$

which, upon adding and subtracting $c_0^*\mu_{k+1}$, yields

$$r(G,\widehat{\theta}_{\mathbf{c}^*}) = -c_0^{*2} + \sigma_{k+1}^2 + \mu_{k+1}^2 - c_0^*\mu_{k+1} - \sum_{i=1}^{k} c_i^*\mu_i\mu_{k+1}$$
$$+ c_0^*\mu_{k+1} - c_0^* \sum_{i=1}^{k+1} c_i^*\mu_i - c_{k+1}^*(\sigma_{k+1}^2+\mu_{k+1}^2) \tag{11.22}$$
$$= -c_0^{*2} + c_{k+1}^*(V_{k+1}+\sigma_{k+1}^2+\mu_{k+1}^2) + c_0^{*2} - c_{k+1}^*(\sigma_{k+1}^2+\mu_{k+1}^2)\,, \tag{11.23}$$

the latter equality obtained via the use of equation (11.10) which defines c_{k+1}^* and equation (11.14). Cancellations in (11.23) yield $r(G,\widehat{\theta}_{\mathbf{c}^*}) = c_{k+1}^* V_{k+1}$, or equivalently, by (11.15),

$$r(G,\widehat{\theta}_{\mathbf{c}^*}) = c_{k+1}^* r(G,X_{k+1})\,. \qquad \blacksquare$$

The identity established in Theorem 11.2 merits some comment. Although I am unable to cite a reference for the result, it would not surprise me to learn that the result is known and has merely been rediscovered here. The result seems to be too basic to have needed to wait until 2010 to appear in print. The result is reminiscent of the fact that, in the simplified setting of $(k+1)$ i.i.d. observations from a common distribution F_μ, the variance of the BLUE $\overline{X}$ of μ is equal to $(1/(k+1))V(X_{k+1})$. Theorem 11.2 might thus be regarded as a Bayesian counterpart to this fact. Since the estimator X_{k+1} is a member of the class of linear estimators $\{\widehat{\theta}_{\mathbf{c}}\}$, the value of the coefficient c_{k+1}^* will necessarily be no greater than 1, but its exact value will depend on the specifics of the application of interest and, in particular, on the first two moments of the operational prior distributions $\{G_i\}$. We make special note of the fact that the comparison in Theorem 11.2 says nothing about the relative performance of the two estimators wrt the true prior $G^{(0)}$, though one might reasonably expect that the inequality would hold, approximately, with $G^{(0)}$ in place of G, if the operational prior G suitably approximates $G^{(0)}$. The main result of the next section provides some support for this conjecture.

One further comment about Theorem 11.2 seems warranted. We have noted that the relationship among the $(k+1)$ experiments stipulated in condition (C7) is not utilized in the theorem or its proof. This suggests that the coefficient c_{k+1}^*, while necessarily in the interval $[0,1]$, may in fact be quite large when the experiments are related but markedly different. Knowing the true relationships among the $(k+1)$

experiments would afford the statistician the opportunity to reduce the Bayes risk of the linear Bayes estimator of θ_{k+1}, thus improving its performance relative to G over that of the standard frequentist estimator X_{k+1}.

I close this section with a brief discussion of a similar but slightly simpler class of restricted Bayes estimators that can be quite useful in selected applications. They are closely related to the linear Bayes estimators treated above, and they have similar properties. Let us consider linear combinations of the available data, that is, linear estimators without the constant c_0 which appears in the estimator $\widehat{\theta}_{\mathbf{c}}$. To distinguish these estimators from those treated above, we will index them by the $(k+1)$-dimensional vector $\mathbf{d}$. The estimators we now study are specified as

$$\widehat{\theta}_{\mathbf{d}} = \sum_{i=1}^{k+1} d_i X_i \,. \tag{11.24}$$

We will be interested in the estimator of the form in (11.24) which minimizes the Bayes risk $r(G, \widehat{\theta}_{\mathbf{d}})$ under conditions (C1)–(C6), with G in place of $G^{(0)}$ in (C3) and (C6). This estimator will play a central role in our treatment of the DT/OT problem based on the simulated data in Table 11.1, the problem to which Section 11.3 is dedicated. In that application, the Bayes risk will take a somewhat simpler form (obtained by setting the constant c_0 equal to zero). For the case $k = 1$, where there is but one past experiment, the coefficients which minimize the Bayes risk $r(G, \widehat{\theta}_{\mathbf{d}})$ of the estimator $d_1 X_1 + d_2 X_2$ are given by

$$d_1^* = \frac{V_2 \mu_1 \mu_2}{(V_1 + \sigma_1^2 + \mu_1^2)(V_2 + \sigma_2^2 + \mu_2^2) - \mu_1^2 \mu_2^2} \tag{11.25}$$

and

$$d_2^* = 1 - \frac{V_2 (V_1 + \sigma_1^2 + \mu_1^2)}{(V_1 + \sigma_1^2 + \mu_1^2)(V_2 + \sigma_2^2 + \mu_2^2) - \mu_1^2 \mu_2^2} \,. \tag{11.26}$$

These coefficients clearly depend on parameters of the operational prior G. Concrete estimators of the form $\widehat{\theta}_{\mathbf{d}}$ may be obtained when either these parameters are taken as known or when an appropriate hierarchical Bayes model is specified.

Under the framework stipulated in Section 11.1, the following result may be established. Its proof, which follows the same lines of argument as in the proof of Theorem 11.2, is left as an exercise.

Theorem 11.3. *Suppose that the conditions in Theorem 11.1 hold. Let* $\widehat{\theta}_{\mathbf{d}^*}$ *be the restricted linear Bayes estimator of* θ_{k+1} *wrt the prior G, that is, the estimator of the form* $\widehat{\theta}_{\mathbf{d}}$ *in* (11.24) *that minimizes* $r(G, \widehat{\theta}_{\mathbf{d}})$. *Then*

$$r(G,\ \widehat{\theta}_{\mathbf{d}^*}) = d_{k+1}^* r(G,\ X_{k+1}) \,. \tag{11.27}$$

Exercise 11.1. For the case $k = 1$, verify that the Bayes risk in (11.28) is minimized by $\widehat{\theta}_{\mathbf{d}^*}$, where d_1^* and d_2^* are given in (11.25) and (11.26).

Exercise 11.2. Prove Theorem 11.3. (**Hint**: Begin by confirming the following expression for the Bayes risk of $\widehat{\theta}_{\mathbf{d}}$ wrt the prior G:

$$r(G,\, \widehat{\theta}_{\mathbf{d}}) = \sum_{i=1}^{k+1} d_i^2 (V_i + \sigma_i^2 + \mu_i^2) + 2 \sum_{1 \le i < j \le k+1} d_i d_j \mu_i \mu_j - 2 \sum_{i=1}^{k} d_i \mu_i \mu_{k+1} + (1 - 2d_{k+1})(\sigma_{k+1}^2 + \mu_{k+1}^2)\ .) \tag{11.28}$$

11.3 Modeling and linear Bayesian inference for data from related life testing experiments.

Let us consider the simulated life testing data in Table 11.1, data that represents what one might expect to observe from independent developmental and operational tests in the context of a military acquisitions program. Let us assume that we have modeled these data as independent samples from exponential distributions with means θ_1 and θ_2, respectively. Letting $T_i = \sum_{j=1}^{n_i} X_{ij}$ be the total time on test from the ith experiment, we have that

$$T_i \sim \Gamma(n_i, \theta_i) \text{ for } i = 1, 2\ . \tag{11.29}$$

As usual, we stipulate that the parameters θ_1 and θ_2 are governed by an unknown, possibly degenerate, true prior distribution $G^{(0)}$. In the context of the related experiments of interest, it has also been assumed (in (C1)–(C3)) that the true parameters, while possibly random, are independent and that each has a finite second moment. The form of the relationship in (C7) between the two experiments may be stipulated, without loss of generality, as $\mu_2^{(0)} = \kappa_0 \mu_1^{(0)}$, as this relationship must in fact be true for some value of the constant κ_0. To specify the true prior distribution fully, we will assume that the parameters $\mu_1^{(0)}$ and κ_0 are independent random variables with finite second moments. In the special case in which the true prior distribution is taken to be degenerate, we simply take $G^{(0)}$ to be the distribution that is degenerate at the constant $(\mu_1^{(0)}, \kappa_0 \mu_1^{(0)})$ in the Euclidean plane.

The operational prior on (θ_1, θ_2) will be specified as a hierarchical model with two first-stage parameters and four second-stage parameters. Specifically, in stage 1, we take the operational prior on the (independent) parameters θ_1 and θ_2 to be given by

$$\theta_i \sim \Gamma\left(S_i,\, \frac{\mu_i}{S_i}\right) \text{ for } i = 1, 2\ , \tag{11.30}$$

where μ_i is the (unknown) mean of θ_i and S_i is the shape parameter, assumed to be fixed and known. Thus, the operational prior distribution models the variability in θ_i by a one-parameter gamma distribution with unknown mean μ_i and with a known parameter S_i governing the dispersion in the model. Large S_i corresponds to quite precise priors for θ_i and smaller S_i corresponds to more diffuse priors for θ_i. We acknowledge the linkage between the two experiments to be correctly represented by the equation

$$\mu_2 = \kappa\mu_1 \tag{11.31}$$

for some fixed constant κ.

The model above is a reasonable facsimile of models actually used in developmental and operational testing. Experienced military personnel who work with DT and OT data on a regular basis often refer to the "kappa factor," a proportionality constant that reflects the amount of deterioration in performance one will see in a system as it moves from the DT phase to the OT phase of the testing process. In fact, it is often claimed that the value of κ is known (approximately, but with substantial confidence) in certain DT/OT settings which occur with some regularity.

In our treatment of the simulated data in Table 11.1, we will, for simplicity, utilize estimators of the OT parameter θ_2 of the form (11.24), that is, we will seek to identify the estimator $\widehat{\theta}^*_{\mathbf{d}} = d_1^*\overline{X}_1 + d_2^*\overline{X}_2$ of θ_2 which minimizes the Bayes risk with respect to the operational prior G (treated as a gamma distribution with an unknown mean). We will explore circumstances in which $\widehat{\theta}_{\mathbf{d}^*}$ tends to outperform $\overline{X}_2$. In our first pass at analyzing these data, we will suppose that we *know* the true value of κ which, in the simulation, was set at the value 0.75. In our second pass at the problem, we shall see that this simplification is by no means essential. Because of (11.31), we may write

$$\theta_1 \sim \Gamma\left(S_1,\ \frac{\mu}{S_1}\right) \text{ and } \theta_2 \sim \Gamma\left(S_2,\ \frac{\kappa\mu}{S_2}\right). \tag{11.32}$$

In the full hierarchical model, the operational prior will specify, in stage 2, that both κ and μ are random variables with finite first and second moments (K_1, K_2) and (M_1, M_2), respectively. As mentioned above, we will first treat the case in which the true value of κ is known (and thus, set K_2 equal to K_1^2 in stage 2 of our hierarchical model). Taking κ as known and μ as fixed but unknown, it is easy to confirm that

$$\sigma_1^2 = \frac{\mu^2}{S_1} \text{ and } \sigma_2^2 = \frac{\kappa^2\mu^2}{S_2} \tag{11.33}$$

and

$$V_1 = \left(\frac{S_1+1}{S_1}\right)\frac{\mu^2}{n_1} \text{ and } V_2 = \left(\frac{S_2+1}{S_2}\right)\frac{\kappa^2\mu^2}{n_2}. \tag{11.34}$$

Letting

$$r_i = \left(\frac{S_i+1}{S_i}\right) \text{ for } i = 1, 2, \tag{11.35}$$

the coefficients of the restricted linear Bayes estimator $\widehat{\theta}_{\mathbf{d}^*}$, relative to the models in (11.32), are given by

$$d_1^* = \frac{n_1 r_2 \kappa}{r_1 r_2 (n_1+1)(n_2+1) - n_1 n_2} \tag{11.36}$$

and

$$d_2^* = 1 - \frac{r_1 r_2 (n_1+1)}{r_1 r_2 (n_1+1)(n_2+1) - n_1 n_2}. \tag{11.37}$$

We make special note of the fact that the restricted linear Bayes estimator $\widehat{\theta}_{\mathbf{d}^*}$ above depends on the operational prior only through the parameter κ. Since we have taken κ to be known, $\widehat{\theta}_{\mathbf{d}^*}$ can be computed from the data. If the parameter θ_2 was estimated by the mean $\overline{X}_2$ of the OT data, one would obtain

$$\widehat{\theta}_2 = 16.09\,. \tag{11.38}$$

As an example of an alternative Bayesian analysis, with the parameter $\kappa = 0.75$ assumed to be known, suppose that our prior model specified the constants S_i as

$$S_1 = 50 \quad \text{and} \quad S_2 = 100\,. \tag{11.39}$$

These latter assumptions correspond to the standard errors for θ in the range $(1.5,\ 3.0)$. With these choices, the coefficients of the optimal estimator $\widehat{\theta}_{\mathbf{d}^*} = d_1^*\overline{X}_1 + d_2^*\overline{X}_2$ are given by

$$d_1^* = 0.3989 \quad \text{and} \quad d_2^* = 0.4303\,,$$

and the restricted linear Bayes estimator of θ_2 is thus

$$\widehat{\theta}_2^* = 14.7539\,. \tag{11.40}$$

We now reveal that the true parameter values of the θs used in the simulated data in Table 11.1 are

$$\theta_1 = 20 \quad \text{and} \quad \theta_2 = 15\,. \tag{11.41}$$

The example above suggests that a rather striking reduction in the error of estimation is possible when the data from DT and OT are combined. This example is no accident; from Theorem 11.3, we have that under the assumptions made above,

$$r\left(G,\ \widehat{\theta}_{\mathbf{d}^*}\right) = d_{k+1}^* r(G,\ X_{k+1})\,, \tag{11.42}$$

suggesting that, in the hierarchical scenario considered above, one might realize, on the average, as much as a 60% improvement in the precision of $\widehat{\theta}_{\mathbf{d}^*}$ over the standard frequentist estimator $\overline{X}_2$ based on the current experiment alone, as measured by the Bayes risk criterion wrt G. If G approximates the true prior $G^{(0)}$ "reasonably well," one might expect a similar level of improvement relative to the truth.

While having some knowledge concerning the value of the kappa factor in a DT/OT setting may often be a reasonable expectation, the assumption that κ is known may, in other circumstances, be considered as inappropriate. We thus develop below an alternative analysis without that assumption. Under the modeling assumptions made in (11.29)–(11.31), and the additional stipulation that the parameters μ and κ are independent a priori, each having distributions with finite first and second moments (K_1, K_2, M_1 and M_2, respectively), it follows from (11.28), with $k = 1$, that the Bayes risk of the restricted linear estimator $\widehat{\theta}_{\mathbf{d}}$ is given by

$$\begin{aligned} r\left(G,\ \widehat{\theta}_{\mathbf{d}}\right) = \Bigg[& d_1^2 \frac{r_1(n_1+1)}{n_1} + d_2^2 \frac{r_2(n_2+1)}{n_2} K_2 \\ & + 2d_1 d_2 K_1 - 2d_1 K_1 + (1-2d_2) r_2 K_2 \Bigg] \cdot M_2\,, \end{aligned} \tag{11.43}$$

where G represents the joint operational prior on the parameters θ_1, θ_2, μ and κ, with the operational prior in (11.30) governing the behavior of θ_1 and θ_2 and independent operational priors governing the behavior of μ and κ. It is clear from (11.43) that the restricted linear Bayes estimator $\widehat{\theta}_{\mathbf{d}^*}$ of θ_2 does not depend on the distribution of μ, and it depends on the distribution of κ only through its first two moments. By minimizing the Bayes risk in (11.43) with respect to d_1 and d_2, we find that the optimal restricted linear Bayes estimator $\widehat{\theta}_{\mathbf{d}^*}$ of θ_2 is the estimator whose coefficients are given by

$$d_1^* = \frac{n_1 r_2 K_1 K_2}{r_1 r_2 (n_1+1)(n_2+1)K_2 - n_1 n_2 K_1^2} \tag{11.44}$$

and

$$d_2^* = 1 - \frac{r_1(n_1+1)}{n_1 K_1} d_1^* \,. \tag{11.45}$$

The derivations above demonstrate that obtaining Bayes estimators of the form $\widehat{\theta}_{\mathbf{d}}$ is perfectly feasible under the modeling assumptions in (11.29)–(11.31) without assuming that the kappa factor is known. The alternative analysis assumes, instead, a fully hierarchical model which specifies that κ is a random variable with known first and second moment. The latter relaxation of the assumption that κ is known allows the investigator to model his uncertainty about κ rather than assigning κ a fixed value.

There is, of course, a related threshold problem: for what specifications of (K_1, K_2) and (M_1, M_2) will the restricted Bayes estimator $\widehat{\theta}_{\mathbf{d}^*}$ of θ_2 outperform the estimator $\widehat{\theta}_2 = \overline{X}_2$? We treat this problem below in the special case in which the true prior distribution $G^{(0)}$ of $(\theta_1,\ \theta_2)$ is degenerate at a point and the operational prior distributions of κ and μ are mean correct (a case which has received special attention in all preceding comparative analyses). In the modeling scenario adopted in (11.29)–(11.31), these latter assumptions imply that the true prior $G^{(0)}$ on $(\theta_1,\ \theta_2)$ is degenerate at the point $(\mu_0,\ \kappa_0\mu_0)$, that the operational prior on μ has first moment $M_1 = \mu_1^{(0)}$ and that the operational prior on κ has first moment $K_1 = \kappa_0$. For simplicity, we will henceforth use an abbreviated notation for the true value $(\mu_1^{(0)}, \kappa_0\mu_1^{(0)})$ of (θ_1, θ_2), denoting $(\mu_1^{(0)}, \kappa_0\mu_1^{(0)})$ by $(\mu_0, \kappa_0\mu_0)$. We will thus be interested in comparing the mean squared errors of the estimators $\widehat{\theta}_{\mathbf{d}}$ and $\overline{X}_2$ at the true value $(\theta_1,\ \theta_2) = (\mu_0,\ \kappa_0\mu_0)$; these MSEs are, of course, the Bayes risks of these two estimators with respect to the degenerate true prior $G^{(0)}$. We have

$$r(G^{(0)},\, \overline{X}_2) = E(\overline{X}_2 - \kappa_0\mu_0)^2 = V(\overline{X}_2) = \frac{\kappa_0^2 \mu_0^2}{n_2} \,, \tag{11.46}$$

while, under the assumption that the operational priors on μ and κ are mean correct,

$$\begin{aligned} r\left(G^{(0)},\, \widehat{\theta}_{\mathbf{d}^*}\right) &= E(d_1^*\overline{X}_1 + d_2^*\overline{X}_2 - \kappa_0\mu_0)^2 \\ &= V(d_1^*\overline{X}_1 + d_2^*\overline{X}_2) + (d_1^*\mu_0 + d_2^*\kappa_0\mu_0 - \kappa_0\mu_0)^2 \\ &= \left(\frac{d_1^{*2}}{n_1} + \frac{d_2^{*2}\kappa_0^2}{n_2} + (d_1^* + d_2^*\kappa_0 - \kappa_0)^2\right)\mu_0^2 \,. \end{aligned} \tag{11.47}$$

We are thus led to consider the inequality

$$\frac{d_1^{*2}}{n_1} + \frac{d_2^{*2}\kappa_0^2}{n_2} + (d_1^* + d_2^*\kappa_0 - \kappa_0)^2 \leq \frac{\kappa_0^2}{n_2} \,. \tag{11.48}$$

Now, as is apparent from (11.44) and (11.45), the coefficients of the restricted linear Bayes estimator of θ_2 depend on a number of constants: r_1 and r_2, which are taken to be known parameters of the operational prior on θ, the sample sizes n_1 and n_2, and the first two moments K_1 and K_2 of the operational prior on κ. We now explore the question: assuming that the true prior $G^{(0)}$ is degenerate at a point, when does the mean correctness of the operational priors on μ and κ provide the Bayesian using the estimator $\widehat{\theta}_{\mathbf{d}^*}$ an advantage over the frequentist estimator $\overline{X}_2$ based on the current experiment alone? The following result provides a definitive answer: *always*, regardless of the specific values of the constants r_1, r_2, n_1, n_2 and of the second moments M_2 and K_2 of the operational priors on μ and κ! Interestingly, this domination is shown to persist even without the assumption of mean correctness of the operational prior on μ.

Theorem 11.4. *Assume that data is available from two related experiments satisfying conditions (C1)–(C7) of Section 11.1, and assume that these experiments obey the models specified in* (11.29) *and the relationship specified in* (11.31)*. If the true prior distribution $G^{(0)}$ of $(\theta_1,\ \theta_2)$ is degenerate at a point $(\mu_0,\ \kappa_0\mu_0)$ and the operational prior distributions of μ and κ are mean correct, that is, $M_1 = E_G\mu = \mu_0$ and $K_1 = E_G\kappa = \kappa_0$, then for fixed but arbitrary values of the sample sizes n_1 and n_2, of the constants r_1 and r_2 and of the second moments M_2 and K_2 governing the operational priors of μ and κ, the restricted linear Bayes estimator $\widehat{\theta}_{\mathbf{d}^*}$ satisfies*

$$r\left(G^{(0)},\ \widehat{\theta}_{\mathbf{d}^*}\right) < r(G^{(0)},\ \overline{X}_2) \,. \tag{11.49}$$

Further, (11.49) holds even without the assumption of the mean correctness of the operational prior on μ.

Proof. Substituting (11.44) and (11.45) into (11.48) and setting $K_1 = \kappa_0$ yields the inequality

$$\begin{aligned} &\frac{1}{n_1}\left[\frac{n_1 r_2 K_2}{D}\right]^2 + \frac{1}{n_2}\left[1 - \frac{r_1 r_2 (n_1+1) K_2}{D}\right]^2 \\ &\quad + \left[\frac{n_1 r_2 K_2}{D} + \frac{D - r_1 r_2 (n_1+1) K_2}{D} - 1\right]^2 < \frac{1}{n_2} \,, \end{aligned} \tag{11.50}$$

where $D = r_1 r_2 (n_1+1)(n_2+1)K_2 - n_1 n_2 \kappa_0^2$. Multiplying both sides of (11.50) by $n_2 D^2$ and simplifying, we obtain the equivalent inequality

$$\begin{aligned} &n_1 n_2 r_2 K_2^2 - 2 r_1^2 r_2 (n_1+1)^2 (n_2+1) K_2^2 + 2 r_1 (n_1+1) n_1 n_2 \kappa_0^2 K_2 \\ &\quad + r_1^2 r_2 (n_1+1)^2 K_2^2 + n_2 r_2 K_2^2 [(n_1+1) r_1 - n_1]^2 < 0 \,. \end{aligned} \tag{11.51}$$

It is apparent that the LHS of the inequality in (11.51) is quadratic in K_2. Further, we note that the constant term of this quadratic is equal to zero, so that one of its roots is equal to zero. If we write (11.51) in the form

$$aK_2^2 + bK_2 < 0\,, \tag{11.52}$$

then the conditions

$$a < 0 \quad \text{and} \quad -\frac{b}{a} < \kappa_0^2 \tag{11.53}$$

will imply that the inequality (11.52) holds, as the first of these conditions implies that the quadratic in (11.52) is concave while the second indicates that the second root of that quadratic is smaller than κ_0^2. The latter implication ensures that for any possible value of K_2 (which must, by definition, be greater than or equal to κ_0^2), the quadratic on the LHS of (11.52) must be negative. After some simplification, the coefficients a and b in (11.52) can be written as

$$a = -r_2^2(n_1+1)\{r_1^2(n_1+n_2+1) + n_1n_2[(r_1+1)^2 - 2]\} \tag{11.54}$$

and

$$b = 2r_1r_2(n_1+1)n_1n_2\kappa_0^2\,. \tag{11.55}$$

Since $r_1 > 1$, the coefficient a in (11.54) is clearly negative. Further, the inequality $-(b/a) < \kappa_0^2$ is equivalent to

$$r_1^2r_2(n_1+n_2+1) + n_1n_2[r_2(r_1^2-1) + 2r_1(r_2-1)] > 0\,. \tag{11.56}$$

The latter inequality clearly holds for all values of r_1, r_2, n_1 and n_2 in their natural domains, that is, for positive integers n_1 and n_2 and real numbers $r_1 > 1$ and $r_2 > 1$. The fact that (11.51) holds for all values of $K_2 > \kappa_0^2$ shows that the domination of Bayes risks in (11.49) holds for all mean-correct operational priors on κ. Further, the argument above is unaffected by the mean correctness of the operational prior on μ or by the value of its second moment M_2. We thus conclude that, under the assumption of mean correctness of the operational prior on κ, the domination of the restricted linear Bayes estimator $\widehat{\theta}_{\mathbf{d}^*}$ over the frequentist estimator $\overline{X}_2$ is universal, irrespective of the exact values of the constants r_1, r_2, n_1, n_2, K_2, M_1 and M_2. ■

Because the difference of the Bayes risks in (11.49) is a continuous function of the first and second moments of the operational priors on μ and κ, there is an open interval containing the value κ_0 such that the domination in (11.49) holds, universally in the values of r_1, r_2, n_1, n_2, K_2, M_1 and M_2, when the mean K_1 of the operational priors of κ lies within that interval. Thus, mean correctness of the operational prior of κ is a sufficient, but not a necessary, condition for Bayesian superiority in the problem studied above. While Theorem 11.4 is narrow in its scope, applying only to restricted linear Bayes estimators of θ_2, it does serve the purpose of demonstrating that a Bayesian approach to the combination of data from related experiments can be efficacious.

Since the class of general linear estimators $\{\widehat{\theta}_{\mathbf{c}} : \widehat{\theta}_{\mathbf{c}} = c_0 + c_1\overline{X}_1 + c_2\overline{X}_2\}$ contains the class of restricted linear estimators $\{\widehat{\theta}_{\mathbf{d}} : \widehat{\theta}_{\mathbf{d}} = d_1\overline{X}_1 + d_2\overline{X}_2\}$, it follows that

the Bayes risks of the respective Bayes estimators wrt the operational prior G are ordered, that is, $r(G, \widehat{\theta}_{\mathbf{c}^*}) \leq r(G, \widehat{\theta}_{\mathbf{d}^*})$. This of course does not by itself imply that $r(G^{(0)}, \widehat{\theta}_{\mathbf{c}^*}) \leq r(G^{(0)}, \widehat{\theta}_{\mathbf{d}^*})$, an inequality that would ensure that, under the assumptions of Theorem 11.4, the linear Bayes estimator $\widehat{\theta}_{\mathbf{c}^*}$ satisfies

$$r\left(G^{(0)}, \widehat{\theta}_{\mathbf{c}^*}\right) < r\left(G^{(0)}, \overline{X}_2\right) . \tag{11.57}$$

It is clear from the proof of Theorem 11.4 that establishing the validity of the inequality in (11.57) under similar assumptions is an imposing algebraic challenge. Exercising great discipline and internal fortitude, I have resisted the temptation to include an investigation into this inequality as an exercise. Instead, I will simply state the following as a reasonable *conjecture* motivated by the preceding theorem: Under the assumptions of Theorem 11.4, the inequality in (11.57) holds for fixed but arbitrary values of the sample sizes n_1 and n_2, of the constants r_1 and r_2 and of the second moments M_2 and K_2 of the mean-correct operational priors utilized for modeling μ and κ.

Exercise 11.3. Under conditions (C1)–(C7) of Section 11.1 and under the model in (11.29), show that the linear estimator $\widehat{\theta}_{\mathbf{d}^*}$ of θ_2 whose coefficients are given in (11.36) and (11.37) minimizes the Bayes risk with respect to the operational priors specified by (11.30) and (11.31) among all linear combinations of the observed means $\overline{X}_1$ and $\overline{X}_2$.

Exercise 11.4. Confirm that the inequality in (11.51) is equivalent to the inequality in (11.50).

Exercise 11.5. Confirm that the values of the coefficients a and b in (11.52) are those displayed in equations (11.54) and (11.55).

11.4 Discussion

The primary aim of this chapter is to introduce a formulation of the notion of related experiments and to put forth a framework within which one can think statistically about the combination of data from such experiments. Although other statistical objectives are possible, I have focused exclusively on the possibility of using data from one or more related past experiments, in combination with the data available in the current experiment, to estimate the parameter governing the latter experiment. This setup is, of course, patterned after Robbins' original formulation of the empirical Bayes approach to statistics. The estimation problem posed is a challenging one, with perhaps the greatest challenge being the reliable modeling of the relationship among current and past experiments. It should be obvious that this must be done with considerable care, and that, even when care is taken, one must acknowledge the risk of misspecification of that relationship, just as one would be concerned about misspecifying the stochastic models one assumes for the available data or the prior distributions one might use in a Bayesian analysis of that data. The risk that is added

when modeling the relationship among experiments might reasonably lead a statistician, in some instances, to rely exclusively on the current experiment for drawing inferences about the current parameter. The framework, results and examples of this chapter are meant to draw attention to the fact that there is real potential for improving the quality of one's inferences by exploiting the information provided by past related experiments. It is not my intent, however, to leave the impression that the use of data related in some way to the experiment of current interest can be carried out in a broad array of applications. Situations in which the approach we have examined may be usefully applied will most certainly occur in practice, but they cannot be expected to be common.

Let me briefly summarize this chapter's contents. The notion that two or more experiments may be "related" rather than "similar" (the latter term used in the EB sense) was motivated by the problem of dealing with data from developmental and operational tests in the context of the military acquisitions process. As a way of thinking about such data, a set of conditions defining related statistical experiments is laid out in Section 11.1. In Section 11.2, the general characteristics of linear functions of the data from a current experiment and k past experiments are explored. For two specific classes of such functions, explicit expressions are given for their Bayes risks wrt a given operational prior distribution G and squared error loss. Results are also obtained which provide a direct comparison between the performance of linear functions of the data which minimize the Bayes risk wrt G and the standard frequentist estimator based on the current experiment alone. The fact that the ratio of their Bayes risks wrt G is an explicitly displayed constant is a useful tool in making the comparisons of interest. In Section 11.3, I turn to the analysis of the case in which $k = 1$, focusing on problems resembling the DT/OT framework. Under particular modeling assumptions (which include a degenerate true prior $G^{(0)}$, a mean correct prior distribution of the parameter κ and a specific class of operational priors), the universal superiority of restricted linear Bayes estimators of the form $\widehat{\theta}_{\mathbf{d}} = d_1\overline{X}_1 + d_2\overline{X}_2$ over the standard frequentist estimator based solely on current data is demonstrated.

Although I have focused, as mentioned above, on comparisons of Bayesian modes of data combination and the option of using the standard frequentist estimator based on the current data alone, it should be acknowledged that the frequentist has other available options. One may gain some useful insights on this fact through a reexamination of the problem treated in Section 11.3. Let us return to the simulated data in Table 11.1 and ask what else a frequentist might wish to do with the two data sets. Let us assume that two particular modeling assumptions are considered reasonable, namely, the modeling of the data in (11.29) and the modeling of the relationship between the two experiments given in (11.31). The Bayesian analysis carried out in Section 11.3 treats two quite different scenarios, the first with the parameter κ assumed known and the second with it taken as random with an unspecified distribution with known first and second moments. The frequentist may also consider two cases, where κ is taken as either known or unknown. In the first case, one has two independent, unbiased estimators of the parameter μ (the mean of the past data), $T_1/n_1 = \overline{X}_1$ and $T_2/\kappa n_2 = \overline{X}_2/\kappa$, with differing variances. The parameter θ_2, the theoretical mean in the current experiment, may thus be estimated as $\kappa\widehat{\mu}$,

where $\widehat{\mu}$ is the BLUE of μ. Obtaining this estimator for the data in Table 11.1 is left as an exercise. It will be apparent from this exercise that, when the statistician actually knows the precise nature of the relationship between the DT and OT experiments (that is, knows (11.31)) and, moreover, actually knows the kappa factor, the BLUE stands to perform very well, and is clearly an effective approach to frequentist data combination.

Let us consider the setting above again, but now without the assumption that κ is known. The framework of two independent experiments with means μ and $\kappa\mu$, where both μ and κ are unknown, is simply a reparametrization of the same two experiments with unknown means θ_1 and θ_2. The likelihood function of the pair $(\theta_1,\ \theta_2)$, given $T_1 = t_1$ and $T_2 = t_2$, is

$$L(\theta_1,\theta_2 \mid t_1,t_2) = \frac{1}{\Gamma(n_1)\theta_1^{n_1}\Gamma(n_2)\theta_2^{n_2}} t_1^{n_1-1} t_2^{n_2-1} \mathrm{e}^{-\frac{t_1}{\theta_1}-\frac{t_2}{\theta_2}} . \tag{11.58}$$

It follows that $\widehat{\theta}_2 = \overline{X}_2$ is the maximum likelihood estimator of θ_2 (as well as the UMVUE of θ_2, of course). Thus, in contrast with the Bayesian approach, the frequentist solution to this estimation problem when the kappa factor is unknown is to ignore the DT data and employ an estimator that depends on the current experiment alone. It is in this circumstance that the Bayesian has the opportunity, using carefully chosen priors on the hyperparameters μ and κ, to provide an estimator that is superior to $\widehat{\theta}_2$. When the value of (θ_1,θ_2) is an unknown constant, Theorem 11.4 provides assurance that there is a collection of operational priors for which Bayesian superiority will occur.

Finally, let us examine the Bayesian approach to the DT/OT problem a bit further. The estimators considered in Section 11.3 are linear functions of the two sample means and are not Bayes estimators in general, but are rather Bayes estimators subject to linearity constraints. Each has the advantage of being available in closed form and of providing potential improvement over the frequentist estimator $\overline{X}_2$ of θ_2. Suppose we wished to obtain the Bayes estimator of θ_2 with respect to a fully specified hierarchical model. Suppose we have two experiments satisfying conditions (C1)–(C7) of Section 11.1. Let us assume that the operational prior on the parameters θ_1 and θ_2 depend on hyperparameters μ and κ, and given μ and κ, the parameters θ_1 and θ_2 are independent with densities $f(\theta_1|\mu)$ and $f(\theta_2|\mu,\kappa)$, respectively. If, in addition, the joint prior density $f(\mu,\kappa)$ has been specified, then inference about the parameter is based on the marginal posterior density given by

$$f(\theta_2 \mid x_1,x_2) = \iint f(\theta_2 \mid x_2,\mu,\kappa) f(\mu,\kappa \mid x_1,x_2)\, \mathrm{d}\mu \mathrm{d}\kappa ,$$

and the (Hierarchical) Bayes estimator of θ_2 under squared error loss is the posterior mean

$$\widehat{\theta}_2^{HB} = E(\theta_2 \mid x_1,x_2) .$$

Since closed form expressions for $\widehat{\theta}_2^{HB}$ are typically unavailable, its evaluation requires the use of approximation or iterative methods. For details on the fully hierarchical Bayes approach to this problem, see Steffey, Samaniego and Tran (2000).

Exercise 11.6. Let T_1 and T_2 be the total time on test statistics from the two independent samples displayed in Table 11.1. Assume that (11.29) holds and that the two population means θ_1 and θ_2 may be expressed as $\theta_1 = \mu$ and $\theta_2 = \kappa\mu$, where κ is known to be equal to 0.75. Find the general form of the BLUE of μ based on T_1 and T_2, and using it, compute the BLUE of θ_2.

12

Fatherly Advice

12.1 Where do I get off?

Fatherly advice is pretty heavy stuff. Am I being presumptuous in offering advice under such a pretence? Perhaps I should state my credentials. This, of course, would be an unusual tactic, since the traditional approach at this point would be to mush on with one's conclusions and leave it to the book's reviewers to judge the strength of the case for people reading what you have to say and for taking it seriously. I believe the book largely makes its own case. So I won't bore you with a list of accomplishments or awards that might appear to give me an air of authority. Instead, I'd say that my main qualification for offering fatherly advice is that I am old. Older people tend to accumulate information and insights that younger people might not yet have gotten to. My own journey in the field of Statistics includes an evolving appreciation for Bayesian methods, in spite of an ever-present skepticism about the appropriateness of their universal applicability. But for quite a long time, the question of "when" one should be a Bayesian seemed to me to be quite elusive. The formulation of the comparative performance of Bayesian and frequentist procedures as a "threshold problem," with performance measured relative to the "true state of nature," provided (for me, at least) some interesting inroads into the answer to this vexing question. Having explored various versions of this and closely related problems over the last twenty years, I believe that I really do have some fatherly advice to share.

I have to admit that I myself have not always taken the advice of "the elders" I have known. I have always, however, tried to give them my thoughtful attention. I am assuming that you, the reader, having reached this final chapter, have afforded me the same courtesy. That is not to say that I assume that I've converted you to the point of view I have put forward. And I want to assure you that this final chapter is not intended to coax you into taking that final step. It is, instead, intended to encourage you to keep thinking about the problems and issues that I have raised. I cannot claim to have come upon the final and definitive answers to the questions investigated here. But to the extent that I have succeeded in stimulating you to continue the investigation of these questions, I will consider this effort to have been worthwhile.

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_12,

Jerzy Neyman, who is of course best known for his seminal contributions to mathematical statistics (including, for example, his introduction of the notions of optimality in hypothesis testing, the concept of interval estimation and his prescription for optimal sample size allocation in stratified sampling), was also well known for his administrative talents, his serious interest in statistical applications and his quirky sense of humor. In the latter regard, the following advice on writing a research paper (advice equally applicable, in my view, to monographs such as this) is attributed to Neyman. "First, tell them what you are going to tell them; then tell them; then tell them what you've told them." I feel that I have satisfied the first two of Neyman's three prescriptions. Now for the third. In the next section, I will offer a brief but fairly comprehensive overview of the material covered in this monograph. This is followed by a section which presents my views regarding the general conclusions and practical recommendations to which this material leads me. In the final section of this chapter, I will discuss some loose ends, some open problems that seem to me to be interesting and important and some conjectures. It is my earnest hope that the ambitious reader will wish to pursue some of these latter issues further.

12.2 An overview

I don't think I am alone in subscribing to the conviction that the subject of Decision Theory was a huge advance in the theoretical development of mathematical statistics. Decision Theory is not really in vogue in today's graduate curriculum in Statistics, nor is it a subject in which there is vigorous ongoing research. But the formulation of statistical work in terms of the quality of the procedures considered for use, as compared to other things one could do, has a continuing and important impact on modern statistical thinking. It must be recognized, of course, that the complexity of the models used, and of the data collected, in many applications of current interest tend to overtax the analytical reach of decision-theoretic treatments. The formal use of methods of optimization has largely been replaced by less formal approaches based on simulation, performance comparisons on collections of "test data sets," numerical methods and iterative techniques. But our goal remains the same: to find the best thing to do or to determine which of several options promises to have the best general performance. This is Decision Theory's legacy! It appears that further advances in Decision Theory must, for the most part, await the development of more powerful mathematical tools. Fortunately, the development of increasingly powerful computational tools and methods has served us well, both in the execution of complex statistical procedures and in the evaluation of their performance.

This monograph begins with a brief discussion of decision-theoretic ideas, in part because these ideas surface with some regularity in the sequel and in part because of the need to introduce the decision-theoretic notation that is used throughout the monograph. This discussion is followed by a careful examination of the primary frequentist methods of estimation. The material in the second chapter will be familiar to many readers, but I have always felt that "redundancy" is a highly useful property, as undervalued as it might be. Our review of frequentist methods, while by no

means exhaustive, covers many of the ideas and methods that a frequentist will tend to utilize in problems of point estimation. Since this monograph is primarily focused on "comparative analyses" and, in particular, on the comparison of the Bayesian and frequentist approaches to point estimation, this review is intended to provide the essential features of frequentist methods that constitute the frequentist's "toolbox" for attacking such problems. Chapter 3 is meant to do the same for the Bayesian approach to point estimation. Since Bayesian methods are accompanied by philosophical stances as well as logical and technical developments, this chapter treats all three of these aspects of the approach.

Chapter 4 reviews the standard arguments regarding the potential superiority of one approach over the other. In the end, these arguments appear to be helpful, to some extent, in clarifying the differences between the two approaches, but can reasonably be judged to be inconclusive when applied to the problem of seeking a generally favored approach. It seems clear that alternative criteria are needed to discuss the superiority of one method over the other in a meaningful way. The general "threshold problem" is introduced in Chapter 4. Its most distinctive feature is what I have referred to as "modeling the truth." While the true value of the parameter of interest is unknown, one may nonetheless consider that parameter as having a "true prior" distribution. This distribution has been allowed to be general, though it has been recognized that in most point estimation problems of interest, the true prior distribution is degenerate at a single point — the true value of the unknown parameter. This monograph could have been written under the assumption that the so-called true prior is always degenerate, though the more general framework under which I have proceeded causes no harm, does have some potential applications and actually helps to shed light on when a Bayesian or a frequentist procedure might be preferred.

Our initial consideration of the criterion for comparing Bayesian and frequentist estimators takes an estimator's average squared distance from the truth as the ultimate measure of its quality. One may view this criterion as a generalized version of an estimator's mean squared error, with the evaluation made after averaging over the (possibly degenerate) true prior distribution. The criterion, in general, is simply the Bayes risk of an estimator, frequentist or Bayesian, with respect to the true prior distribution. When the true prior is in fact degenerate, the criterion function reduces to the mean squared error of an estimator evaluated at the true value of the target parameter — no doubt one of the most relevant measures one might consider in judging an estimator's worth. In later chapters, the threshold problem is examined under other loss criteria, including a widely used class of asymmetric loss functions. The fact that the true prior is unknown will seem, early on, as a substantial obstacle to reaching useful conclusions. Later developments show that this obstacle is by no means impenetrable, and that useful practical insights into the comparisons of interest can in fact be realized.

In Chapters 5, 6, 7 and 8, solutions to different versions of the threshold problem are obtained. The most concrete solution is derived in a framework of one-parameter exponential families, conjugate prior distributions and squared error loss. In Chapter 5, under these particular assumptions, it is shown that there is a sizable subclass of operational priors available which provide a Bayesian statistician with an estima-

tor which outperforms the "best frequentist estimator." From these developments, the following practical principle emerges: a Bayesian will tend to do well, as compared to a frequentist, in the estimation problems considered here unless the Bayesian is both *misguided* and *stubborn*, that is, he makes a "prior guess" at the unknown parameter which is quite distant from its true value and he places a good deal of weight on it (thereby expressing considerable confidence in the quality of his poor guess). It is later confirmed that this principle applies in a variety of other modeling scenarios.

Chapter 6 introduces the new concept of Bayesian self-consistency (SC). While the concept is interesting in its own right, it is introduced largely for the purpose of assisting in the examination of a particular proposed solution to the Bayesian consensus problem. The SC concept is an intuitively appealing property that is easily put into words: if your experiment confirms your prior opinion (that is, your prior guess) about the value of an unknown target parameter, then your posterior opinion should remain the same, that is, your prior and posterior guesses at the target parameter should match. It is noted that conjugate prior distributions associated with exponential families of sampling distributions are self-consistent, but that the occurrence of self-consistency outside of this framework is rather rare. There are interesting open questions regarding the notion of Bayesian self-consistency, some of which will be discussed in the sequel. For now, I will focus on the relevance of the concept in the context of a rather famous scenario known as the Bayesian consensus problem.

The consensus problem is a real dilemma in Bayesian inference. We know that, in general, the consultation with experts in a particular field of application tends to play an important role in the determination of the Bayesian strategy that should be used in addressing a statistical problem in that field. When several experts are consulted, and they have different prior opinions about an unknown parameter, the statistician is confronted with the challenging issue of whom to trust or, perhaps, the equally challenging issue of how to use all of this prior input in the inferential problem he is trying to solve. There is, of course, a sizable literature on how one should proceed. The strategy that is advocated in Chapter 6 involves the mixing of all the prior guesses one has collected and the determination of precisely how much confidence should be placed on that mixture. Under the same exponential family framework introduced in the preceding chapter, a class of pseudo-Bayesian "consensus estimators" are proposed, and their properties are investigated. Conditions are identified under which the proposed estimator will enjoy a generalized form of self-consistency. Further, it is shown that the exact same conditions ensure that the proposed estimators may be expressed as convex mixtures of certain linear combinations of the expert's prior guesses and the standard frequentist estimator in the problem of interest. This latter property implies that the proposed estimators are actually the Bayes estimators with respect to particular conjugate priors. This naturally leads to the application of the solution of the threshold problem considered in Chapter 5, resulting in a theorem which provides a direct comparison of the performance of the proposed consensus estimators to that of the standard frequentist estimator, using the Bayes risk of an estimator with respect to the true prior distribution as the criterion for comparison. To my knowledge, this result is unique in the literature on the Bayesian consensus prob-

lem, as such performance comparisons tend to be intractable when other Bayesian approaches are used.

Chapter 7 deals with a version of the threshold problem which is somewhat more resistant to solution than the class of the problems treated in Chapter 5. In Chapter 7, the problem of interest is the estimation of a multivariate normal mean. Most modern-day approaches to this problem involve some form of shrinkage. The James–Stein estimator is chosen as the standard-bearer for the frequentist school. As is well known, this estimator shrinks the sample mean $\overline{\mathbf{X}}$ in the direction of a distinguished point, with the extent of the shrinkage being driven by the observed data. The Bayes estimator of a normal mean relative to the conjugate normal prior distribution is also a "shrinkage estimator," but differs from the James–Stein estimator in that the distinguished point is the prior mean, the resulting estimator is a convex combination of the sample mean and this distinguished point and, finally, the extent of shrinkage of $\overline{\mathbf{X}}$ toward the distinguished point is a function of the prior distribution and is not data-dependent. We treat a rather stylized version of the problem, stipulating that the variance-covariance matrices of the operational prior and of the sampling distribution are both scalar multiples of the identity matrix. It is also assumed that the true prior is degenerate at a point. One of the primary results of the chapter is that the threshold separating good and bad operational priors exists and that its general characteristics can be explicitly described. Thus, just as in the one-parameter case, there is indeed a collection of operational prior distributions which will enable the Bayesian to outperform the frequentist who uses the James–Stein estimator. But an additional insight, one substantially different from those drawn from our examination of one-parameter problems, is that the Bayesian's "window of opportunity" for achieving superiority is relatively modest. An example is given in which the Bayesian will prevail only if the common variance σ_G^2 of the priors on the elements of the mean vector is sufficiently small, so that conservative prior modeling does not serve the Bayesian well. Another example shows the domain of Bayesian superiority to consist of σ_G^2 in an interval of strictly positive numbers, so that neither priors that are too precise nor priors that are too diffuse lead to Bayesian superiority. The unavoidable conclusion is that Bayesian point estimation of high-dimensional parameters is a very challenging enterprise. Unless the prior distribution is selected with a great deal of care (and perhaps even some luck), Bayesian point estimation may be expected to provide a level of performance that is inferior to that available using selected frequentist alternatives.

Chapter 8 treats the threshold problem in two specific problems, the estimation of a multivariate normal mean, and the estimation of a linear function of regression parameters, under asymmetric loss. In both cases, the loss is taken to be the well-known Linex loss which places a penalty that is essentially linear in the estimation error for underestimating the parameter and a penalty that is essentially exponential in the estimation error for overestimating the parameter (or vice versa, if desired). The performance of the Bayes estimator with respect to a fixed operational prior is compared to that of the maximum likelihood estimator of the target parameter. In estimating a normal mean, results obtained include the fact that the Bayes estimator will always be superior if the prior is sufficiently diffuse and that a Bayes estimator uniformly outperforms the MLE when the operational prior is mean correct. Similar results are

obtained in the problem of estimating a linear function of regression parameters in the general linear model. These results confirm the existence of a threshold separating good and bad priors in these problems. It is noted, however, that in both of the problems above, the maximum likelihood estimators of the target parameters are known (or suspected) to be inadmissible, often being dominated by frequentist estimators that are generalized Bayes with respect to particular improper priors. Other threshold problems in which Bayes estimators are compared to alternative frequentist estimators are open problems requiring further investigation.

The results of Chapters 7 and 8 make it clear that the formulation and solution of the threshold problem treated in Chapters 4 and 5, that is, concerning the estimation of scalar parameters under squared error loss, is not a phenomenon that applies solely in this restricted setting. The extension of our treatment to the threshold problem in higher dimensions in Chapter 7 and to asymmetric loss functions in Chapter 8 shows that the framework is quite general and that solutions are tractable for selected extensions in both of these directions. The full extent of the generality of these "threshold problems" and their associated solutions remains to be characterized.

The "curse" of nonidentifiability has posed difficulties in the classical theory of Statistics since its inception. This term refers to a certain innate ambiguity in the model used to describe the observable data. When a model is not identifiable, the distribution of the data drawn therefrom is not well explained by a single parameter value, but instead, only by a collection of possible parameter values. From the classical viewpoint, this renders the standard problem of point estimation as fundamentally unsolvable. In Chapter 9, I discuss the various "end-runs" that a frequentist might take in dealing with such problems. None of these "alternative solutions" addresses the original problem, though each may, in particular settings, shed useful light on the problem. The main focus of Chapter 9 is, however, a discussion of Bayesian approaches to nonidentifiability. Unlike the classical approach, the Bayesian treatment of nonidentifiability is fully interpretable and, generally, quite straightforward. One begins with a prior distribution on the model's parameters. One then updates the prior on the basis of the available data. While the data is admittedly "imperfect," it is nonetheless the case that it contains information about the unknown parameters that may well add something useful to one's prior opinions about them. The posterior distribution is, as usual, the basis for inferences about the parameters.

The behavior of Bayes estimates of nonidentifiable parameters differs from their general behavior in the case of standard, identifiable models. For example, they need not converge to the true values of the unknown parameters. However, the potential (indeed, typical) lack of consistency of the Bayes estimators of a nonidentifiable parameter need not be considered to be a fatal flaw, especially in view of the fact that there really are no available (frequentist) competitors to these estimators. But one should certainly investigate the "efficacy" of Bayes estimators in such circumstances. Even though the development of a Bayesian solution is feasible, one needs to have some way of checking that the process is worth carrying out. At first view, this may seem like an imposing challenge, as there appears to be no natural competitors to a standard Bayesian analysis. But, interestingly, there is in fact a way to measure the utility of that analysis. In Chapter 9, we investigate the question of when

a traditional Bayesian analysis will improve upon the prior (no-data) estimator of the unknown parameters — the mean of the prior distribution. It is a curious but quite real possibility that a full Bayesian analysis may fail to improve upon one's prior guess at the parameters of a nonidentifiable model, that is, that the available data may serve to mislead the statistician and result in inferences that are inferior to the guess based on one's prior opinion. A characterization of the circumstances in which this may happen will thus shed useful light on whether or not a full Bayesian analysis is worth pursuing.

The problem studied in Section 9.2 is meant as a prototype of this sort of comparison. Data is assumed to be drawn from the Binomial model $\mathscr{B}(1, p_1 + p_2)$, a model in which the parameter pair (p_1, p_2) is not identifiable. Using Dirichlet distributions as the operational priors on the pair (p_1, p_2), a full characterization is obtained of the class of priors (indexed by their mean vectors (a,b)) for which the limiting Bayes estimator (as the sample size $n \to \infty$) of the parameter pair (p_1, p_2) is closer to the true value (p_1^*, p_2^*) than is the prior mean (a,b). Our treatment may be viewed as a version of the threshold problem in which two competing Bayes estimators are compared. This comparison is extended to an investigation of how much improvement is possible when one of these estimators dominates the other. If D_1 and D_2 are, respectively, the Euclidean distances between the prior guess (a,b) and the true value (p_1^*, p_2^*) of (p_1, p_2), and between the limiting Bayes estimator $(\gamma a, \gamma b)$ and (p_1^*, p_2^*), it is shown that the ratio D_1/D_2 lies between strict upper and lower bounds, that is, $\sqrt{2}/2 \leq D_1/D_2 \leq \infty$. The bottom line in this study may be roughly summarized as follows: the limiting Bayes estimator of (p_1, p_2) will be closer to the true value (p_1^*, p_2^*) of the pair for most choices of the prior mean (a,b), with the percentage associated with this domination being close to 100% when the true value of $p_1 + p_2$ is small and being a value slightly larger than 50% when $p_1^* + p_2^* = 1$. Further, when the limiting Bayes estimator of (p_1, p_2) is farther from the true value (p_1^*, p_2^*) than the prior mean (a,b), the distance D_1 will be no less than 70% of D_2, while when the limiting Bayes rule provides improvement over (a,b), the percentage associated with the improvement can be large and is, in general, unbounded. Also studied in Chapter 9 is the efficacy of Bayesian estimation in the nonparametric competing risks problem, where the multiple decrement function is known to be nonidentifiable, and the efficacy of Bayes estimators in a nonidentifiable version of the model for stress–strength testing in the context of engineering reliability.

Although Herbert Robbins contributed to the development of Mathematical Statistics in many different ways and through many inspired ideas, his introduction of the empirical Bayes approach to Statistics in 1955 was perhaps his most influential creation. The idea that one could learn from, and productively utilize, data that was only "loosely related" to the data of interest in a current experiment was a breakthrough in the field that has had many useful applications. (More precisely, in Robbins' EB framework, data from a past experiment is modeled as conditionally independent from the current data and is governed by a different parameter value.) In Chapter 10, two particular threshold problems are treated, one which compares the performance of two Bayes estimators and one which compares the performance of two frequentist estimators. In the first scenario, it is shown that there is (essen-

tially) always an opportunity to improve upon one's inferences based on data from the current experiment alone when one incorporates the information drawn from past similar experiments. In the second scenario, a similar result is obtained regarding the use of a frequentist estimator based on current data as compared to alternative frequentist estimators which utilize data from past similar experiments. The bottom line is that Bayesians and frequentists who restrict their attention to the data from the current experiment have the opportunity to be "better Bayesians" or "finer frequentists" by exploiting the information provided by past similar experiments. This conclusion presumes, of course, that the modeling assumptions of the empirical Bayes framework are applicable to the problem of interest.

There is no comprehensive theory available at this time for the treatment of "related experiments." While a sequence of such experiments may bear some resemblance to experiments that are properly modeled within the empirical Bayes framework, they differ in several respects. The most important of these differences is that there is some form of functional relationship among the parameters of the various experiments that precludes the possibility of describing them as "similar" in the EB sense. Chapter 11 represents an initial attempt to treat such problems within the Bayesian paradigm. While various frequentist approaches are also possible, and are mentioned in the discussion section of this chapter, the primary focus of the chapter is the development of a Bayesian estimator of the parameter of the current experiment based on both current and past data and the comparison of its performance to that of the standard frequentist estimator based solely on current data. Although the framework described is applicable to an arbitrary number of past experiments, my main goal is to treat data that might be considered typical in the motivating two-sample problem in which the two experiments of interest are drawn from the processes of developmental and operational testing that generally arise in certain military and industrial applications. For a simple but fairly realistic model for the typical DT/OT data one might observe in the military acquisitions process, results are obtained which provide a direct comparison between the performance of linear functions of the data which minimize the Bayes risk with respect to a fixed operational prior distribution and the standard frequentist estimator based on the current experiment alone. Under the modeling assumptions made (which include a degenerate true prior and a certain "mean-correctness" assumption), the universal superiority of restricted linear Bayes estimators over the standard frequentist estimator based solely on current data is established.

I now turn my attention to some general reflections about the findings described above and the implications they may have on general statistical practice.

12.3 Implications

Perhaps I should begin by acknowledging the fact that debates in the scientific community about competing approaches to various classes of problems will never be completely resolved. Among the reasons for this is that those with opposing views will sometimes dig in their heals and insist that, at least for them, the preferred ap-

proach is abundantly and immutably clear, and that further discussion of alternative approaches is unnecessary and, perhaps, quite irrelevant. In the context of the present developments, this translates into the fact that there are some in the Bayesian camp, and also some in the frequentist camp, who find no need for, or justification of, possible crossovers into the alternative methodology. The main theme of this monograph is that neither the Bayesian nor the frequentist approach to the problem of statistical point estimation is universally superior and that the context of the statistical problem at hand should therefore guide one's decision about the approach that promises to give better performance. I believe that the main results of the monograph amply support this latter view. In my discussion of these results in preceding sections, I have highlighted certain situations in which Bayesian methods are particularly promising, certain situations in which frequentist methods are particularly promising, certain situations in which the Bayesian approach is uniquely applicable and certain situations in which, even when one is staunchly committed to one or the other approach, one has the opportunity to improve the performance of one's estimator using Bayesian or pseudo-Bayesian ideas. I now boldly proceed to a statement of my position on the question of how all statisticians should approach their work. While my "fatherly advice" officially pertains solely to estimation problems, I nonetheless believe that roughly comparable advice is relevant to statistical work in general, though this latter belief should properly be viewed as a hunch rather than a well-defended position.

Whether or not they work well in a given application or in general, Bayesian methods will always be vulnerable to the criticism that some form of subjective information has been injected into the analysis of data, thereby risking the potential introduction of errors or biases that would not be there otherwise. A relevant response, when confronted with this criticism, is the question: is the risk worth taking? After all, we are faced with risks all the time, and we are constantly forced to determine for ourselves whether or not we should take a particular risk or pass on it. Just as in real life, the appropriate answer to the statistical question posed will vary with the particular circumstances in which the risk is presented. A key premise shared by those who would confidently espouse a (proper) Bayesian analysis in a given problem is that there is useful prior information available in the problem and that utilizing that information stands to improve the quality of one's inferences. Studying that premise carefully for the last couple of decades has led me to new understandings about its meaning. I've come to the conclusion that a proper interpretation of the word "useful" is essential in making, in Robert's (1998) language, the Bayesian choice or, alternatively, in making the frequentist choice.

In the above survey of the main findings that have been described in this monograph, I repeatedly referred to the existence of a threshold separating good and bad priors. This is, of course, simply alternative language about separating problems in which useful prior information exists from those in which it doesn't. In problems in which the collection of "good priors" is relatively small, the risk involved in using a Bayesian procedure might be considered too large and thus not worth taking. But even in such circumstances, there may be solid justification (for example, confidence that one's own prior information is really good) for proceeding with a Bayesian analysis. One of the take-home lessons of this monograph is that, in using a Bayes estima-

tor in a given problem, one should give careful thought to the question of whether or not one's prior information is of sufficient quality to justify that risk one encounters when employing it.

Let's discuss the word "useful." It is helpful to recall the word-length experiment. That empirical demonstration, together with the theoretical results that help us understand the general outcome encountered there, point to an insight about Bayesian treatments of that and similar problems which many statisticians, and students of Statistics, might consider surprising. In one-parameter problems such as those considered in Chapter 5, the word "useful" has a substantially broader interpretation than might have been ascribed to it in advance. In the experiment itself, one finds that Bayesians using prior information that appears to be quite off the mark may nonetheless produce estimators of an unknown proportion p that tend to outperform the time-honored frequentist estimator, the sample proportion $\widehat{p}$. Further, this phenomenon is only mildly affected by the size of the sample upon which the sample proportion is based. The reader will recall that certain participants in the word-length experiment had quite miserable prior guesses at p and yet, by putting a rather modest amount of weight on their prior guesses, managed to outperform the imaginary frequentist. That this should happen is counterintuitive. We might imagine that placing small weight on a very poor guess would reduce the impact of a poorly aimed prior, but the fact that, using a convex combination of that poor guess and the estimator that the frequentist will use, the Bayesian actually can still win, in spite of his embarrassing *faux pas*, demonstrates a rather amazing resiliency in the Bayesian approach. It appears that Bayesian methods are much more forgiving than they are generally perceived to be. Recall the take-home lesson from Chapter 5: A Bayesian will generally outperform a frequentist in the type of problems considered there unless he is both misguided (having a poor prior guess) and stubborn (placing a substantial amount of weight on his guess). Interestingly, neither of these clearly negative features need be fatal by itself. It is clear that it is quite possible for an imperfect Bayesian to outperform a frequentist, provided that his imperfections are "univariate" rather than "bivariate" in nature. In the end, careful attention to the calibration of one's prior (that is, approximate mean-correctness) and conservative prior modeling are generally the keys to the Bayesian's success, at least in the one-parameter problems discussed early on. The careless Bayesian, who chooses a prior distribution without much introspection or consultation, perhaps choosing it with "convenience" as his main objective, cannot be expected to estimate unknown parameters well. In my view, this type of Bayesian gives the Bayesian community a bad name.

The treatment of high-dimensional parameters presents a picture that is quite different than the picture painted in the one-parameter case. If one thoughtfully reflects upon general multiparameter estimation problems, one might well anticipate the outcomes we have seen. First, it must be recognized that the modeling of prior information in multiparameter problems represents an imposing challenge. Not only are the prior models which might be used more limited, our understanding of them tends to be more shallow, and the ways we might capture expert opinion through their use is often quite unclear. Second, there is much more room for "prior misspecification" in multiparameter problems. Third, since one gross error among many individual es-

timates can be quite costly, the consequences of poor prior modeling, even in only one or two directions, may be substantial. The findings discussed in Chapter 7 generally argue against the use of Bayesian methods in estimating a high-dimensional normal mean vector. We have noted that a threshold separating good and bad prior distributions does exist in such problems, but the relative abundance of bad priors, compared to the "size" of the collection of good priors, must surely give one pause. As I mentioned earlier, careful prior modeling, and a dose of good luck, might lead to Bayesian superiority, but the odds are increasingly stacked against Bayesian estimation as the dimension of the problem increases. I would go so far as to assert that I would personally be quite hesitant to use a Bayes estimator of a high-dimensional parameter, though I must admit that I would be tempted to do so in a problem in which I had access to special prior information in which I had "well-justified" confidence.

The linchpin that holds the results in the monograph together and upon which my conclusions and "advice" are based is the somewhat unusual performance criterion that I have employed throughout. The idea of modeling the truth is not entirely novel, though it is not an idea that statisticians generally give much thought to. Regarding its natural predecessors, it is fair to say that the unknown prior in Robbins' empirical Bayes framework is the same object that I have called the true prior distribution. Although Robbins used that prior in a different way and had quite different statistical goals in mind, the Bayes risk with respect to the true prior is clearly the underlying measure by which he chose to judge the performance of EB estimators. Since I use this criterion with somewhat different goals in mind, and because Robbins was not particularly vocal about defending the criterion as a general performance measure in an arbitrary point estimation problem, I will address this issue here, repeating, for emphasis, some of the arguments I put forward in Chapter 4. (You will perhaps recall that I'm something of a fan of "redundancy.") But since so much depends on this choice, I would be remiss if I did not include a comprehensive defense of the choice in my summation.

First, let us recognize that, in any given statistical estimation problem, there is something we can legitimately call "the truth." If we didn't believe this, we would probably not engage in the estimation process, or, alternatively, we would have to admit that doing so was a not a well-defined exercise. Even in the case of a nonidentifiable model, we tend to believe in the existence of "the truth," while acknowledging that the data at our disposal provides ambiguous information about it. In most problems of practical interest, the truth may be thought of as the true numerical value θ_0 of the unknown parameter θ we are attempting to estimate. In this monograph, we have allowed the truth to be represented by a "true prior distribution" G_0 on the parameter space, though in many instances, we have focused our attention on the case of special practical interest, that is, the case in which G_0 is degenerate at a point. The potential randomness of the true value of θ doesn't really need a defense since, if it bothers you, you can simply concentrate on the degenerate case. I have nonetheless mentioned that, just as in Robbins' EB framework, one might reasonably consider the true value of the parameter as random in situations in which a particular experiment is replicated with some frequency and one wishes to recognize the variability

of the parameter value that perhaps is due to the moderately varying conditions under which these experiments are carried out.

This brings us to the core of the matter, the particular criterion $r(G_0, \widehat{\theta})$ that I have routinely used as the ultimate measure of the quality of a particular estimator's performance. There are three specific aspects of the choice, as our criterion, of the Bayes risk wrt the true prior that require some careful reflection. The first is the fact that the worth of an estimator is based on its closeness to "the truth," the interpretation of "closeness" of course depending on the loss function employed. The second is the fact that the criterion is a quantity that is unknown at the time of the experiment and, in many instances, will never be known. The third is that the measure is an average, or more accurately, an expected value which measures a chosen estimator's expected loss over both a random outcome of the experiment and the possibly random true value of the target parameter. These three issues are discussed in turn in the following paragraphs.

It was noted early on that the risk function $R(\theta, \widehat{\theta})$ attempts to measure the quality of estimation as the parameter varies over the entire parameter space. It is a clumsy and coarse measure of quality, since in most real problems, there are a host of estimators that are good for some values of θ and bad for others. It is often the case that two particular estimators will be incomparable on the basis of their risk functions alone. On the other hand, the Bayes risk of an estimator with respect to a given "operational prior" G resolves one problem while creating another. While it renders all estimators comparable, and it leads to an identification of a "best" estimator (relative to G), the real quality and reliability of that estimator is unknown, as its quality is highly dependent on the choice of G, a distribution that may have little connection or resemblance to the truth. The closeness of an estimator to someone's choice of "weight function" on the parameter space cannot serve as our ultimate measure of quality simply because it can vary within any given problem and does not have any inherent validity relative to the truth.

We are thus led to the measure we have chosen, the Bayes risk with respect to the truth. Set aside for now the fact that the truth is unknown. If it was the case that the truth would eventually be known, then there can be no doubt that we would go back and judge the quality of an estimator by its closeness to the truth. When we are operating in the here and now, we must average over uncertain quantities, looking for what is expected to provide good performance relative to the truth and with respect to what one might observe in the type of experiment with which one is dealing. The Bayes risk $r(G_0, \widehat{\theta})$ does exactly this. The use of this criterion for comparing estimators does not interfere with a Bayesian's adherence to the likelihood principle or other tenets of the Bayesian approach, as the Bayesian is allowed, indeed expected, to abide by these tenets in formulating his estimator. Nor does the criterion interfere with the choices a frequentist might make. But once the Bayesian or the frequentist has decided on an estimator, it is only reasonable to ask whether the answer they've proposed has objective merit in the problem that is being addressed. This latter question can only be answered be assessing their performance relative to the truth. For either statistician, our criterion amounts to the process of an impartial third party gauging how close the statistician would tend to come to the "right answer."

In the important special case of estimating an unknown constant relative to squared error loss, the Bayes risk $r(G_0, \widehat{\theta})$ reduces to the mean squared error of the estimator evaluated at the true value of the parameter. Without declaring this measure to be the only or the best way to judge the quality of an estimator, it seems easy to make the case that the measure is highly relevant to such judgments and certainly ranks highly among the measures one would wish to examine in comparing the estimation strategies that one might use in the problem under study.

Consider, now, the fact that "the truth," represented here by the true prior distribution G_0, is unknown. This fact clearly precludes the possibility of identifying the best possible estimator in a given problem. However, as we have seen, we can discern a good deal by studying the problem in the abstract. In the various forms of the threshold problem considered in this monograph, we have typically been able to identify the general characteristics of Bayes estimators which stand to give the Bayesian the advantage over the frequentist. When the collection of "good" Bayes priors is large relative to the size of the class of priors considered for use, one might justifiably have substantial confidence that the use of a Bayes estimator will produce results that are superior to those obtainable by the use of a frequentist alternative. In the complementary case, the justification for using a Bayes estimator is more tenuous, as its use might well lead to results that are inferior to those a frequentist estimator will attain. With G_0 unknown, it is impossible to make the right choice with certainty, but the guidance provided by the solution to the related threshold problem can be quite helpful in making a rational choice in a given problem. The determination of which side of the threshold one's prior lies in will always be a challenging part of one's analysis, but giving careful thought to this issue stands to be of some help in avoiding the potentially substantial risk of making a poor choice. It should increase the likelihood of taking the best approach to particular point estimation problems and of achieving a higher level of performance than strict adherence to a single statistical paradigm stands to provide.

In general, the findings laid out in this monograph suggest that, in problems of point estimation, Bayes procedures will often give answers that are superior to what a frequentist procedure can be expected to provide. We have noted that, in certain contexts, the collection of Bayes estimators which outperform frequentist alternatives is substantially large, suggesting a certain natural robustness of Bayesian inference relative to the choice of one's prior distribution. When the dimension of the parameter space is low, even imperfect Bayesians can do well, and the careful and studious Bayesian is quite likely to do well most of the time. But there are also problems in which the prospects of the Bayesian are not so rosy. The estimation of a high-dimensional parameter is one, and problems in which there is a paucity of useful or certifiably reliable prior information is another. As stated earlier, I am unabashedly a "Bayesian sympathizer," ready to execute a Bayesian procedure whenever I and my collaborators in a given problem have solid justification for our prior beliefs about model parameters. At the same time, I am quick to question the utility of the prior information I might have, and I am prepared to utilize a standard frequentist analysis when I have nagging doubts that the chosen prior is appropriate. When I am prepared to use a Bayes estimator in a given problem, I try to keep my eyes on the

prize, keeping in mind the potential benefits of mean correctness and of choosing to err on the conservative side when deciding on how much confidence I have in my prior estimate. I regard both philosophical and the technical underpinnings of the Bayesian approach with respect, while recognizing the need to consider classical methods when indicated by a problem's particulars. Being open to both the Bayesian and frequentist approaches to estimation cannot guarantee success, but it is, I believe, an important ingredient in securing successful outcomes on a regular basis.

12.4 Desiderata

Kai Lai Chung once remarked to a colleague that it was "academic suicide" to prove the last theorem in a particular field of study. While obtaining the definitive result in a given problem area has its obvious positives, not the least of which is the pride and satisfaction one would feel from having done it, the fact that there is nothing left to do means that the area is essentially closed and will thus not attract other researchers. The negative repercussions of this circumstance include, in Chung's words, the fact that "no one will cite your work." While Chung's advice was given with tongue in cheek, there is some (perhaps a bit perverse) validity to Chung's assessment. But, of course, I have nothing to fear regarding the state to which he referred. There are tons of things left to be done.

In this monograph, our treatment of the threshold problem has been dedicated, almost exclusively, to estimation problems involving sampling distributions that belong to univariate or multivariate exponential families. I have made occasional reference to results that may be obtained in fully nonparametric settings, and have treated a few parametric models that lie outside the exponential family framework, but this work only scratches the surface of the alternative modeling that could and should be investigated. I conjecture that, while the "look" of the solutions will vary, the basic premise of the monograph that a threshold separating good and bad priors will virtually always exist, and may be usefully described, can and will be confirmed in more general settings. I expect that, while analytical solutions may prove harder to obtain, investigations through numerical means or via simulation will show that the threshold problem which compares the performance of Bayes and frequentist estimators relative to "truth" can be tackled in more general settings and will provide useful guidance on when a Bayesian or a frequentist estimator is to be preferred.

I have argued that the Bayes risk of an estimator with respect to the "true prior distribution" is a highly appropriate criterion for judging performance of an estimator and for comparing their expected performance in a given estimation problem. It would be difficult to argue that the true state of nature is irrelevant to our judgments about performance. Some readers may feel that the idea of using a criterion that involves averaging over the sample space is somewhat disturbing. On this score, perhaps it is useful to state that an "*impartial* third party" who is commissioned to compare the performance of Bayes and frequentist estimators should not be committed to either paradigm, and thus should not be held to either paradigm's traditions nor be criticized for not adhering to them. This "judge's" only concern is the ques-

tion: which approach tends to give better answers in a given problem or for a class of estimation problems of interest? In that context, the average distance from the truth, however that "distance" is defined, is certainly a measure that would seem to address that question fairly and directly.

Much of the work presented here uses squared error loss (or its natural generalizations) as the distance criterion for measuring closeness to the truth. Were it not for Chapter 8, one might justifiably remain skeptical about the threshold phenomenon in settings using alternative loss criteria. Chapter 8 shows that the phenomenon surfaces in quite a similar form under loss criteria that are dramatically asymmetric. Arguing about loss functions is, however, somewhat reminiscent of some of the conjectures that followed Stein's demonstration of the inadmissibility of the sample mean in high dimensions: was this really a "squared error loss" related phenomenon? That it is not has been confirmed in a variety of studies. As mentioned earlier, Zellner (1986) showed that, even in the univariate case, the sample mean was inadmissible, under Linex loss, as an estimator of a normal mean. This seems intuitively plausible, even in the absence of a technical demonstration, since when overestimation is heavily penalized, shrinking $\overline{X}$ toward zero seems like a highly sensible strategy. All the above notwithstanding, I believe there is both room for, and value to be derived from, examining the threshold problem under a broad cross section of loss criteria. How would the boundary between good and bad Bayes estimators change if squared error was replaced by absolute error? Perhaps the general class of loss functions $\{L_r(\theta,a) = (\theta - a)^r,\ r > 0\}$ merits investigation. Finally, it is legitimate to ask "What is the effect on the threshold problem of other forms of asymmetric loss?"

The highly perceptive reader might have picked up the fact that I am sort of partial to conjugate prior distributions in problems dealing with exponential families. OK, OK, all of you have picked that up. But my reasons have little to do with "convenience," which was a popular reason for their use in the days that preceded the computational revolution in Bayesian inference. There is an unrivaled interpretability of Bayes estimators with respect to conjugate priors which is most definitely useful in prior elicitation and is equally useful for the transparency it brings to the estimators themselves, due in large part to the convexity property that they typically enjoy.

Add to this the intriguing property of Bayesian self-consistency. Ask yourself if you are comfortable using priors without that property. Finally, consider the principle of parsimony. Typically, a couple of pieces of information suffice to identify the conjugate prior one would wish to use in a given application. Suppose you have the opportunity to gather ten pieces of information from an expert in a certain application area. One has to wonder whether the prior that is fit to this information will actually lead to better performance than the conjugate prior that was identified from the first two questions asked. This is a difficult question to answer analytically, though a careful formulation and analytical solution of that problem would be most interesting. But the problem does lend itself to empirical study, either through real experiments or simulations. My guess is that, when work of this type has begun to accumulate, it will be found that there is definitely a point of diminishing returns in prior elicitation. I predict that conjugate priors will be shown, in most problems

of interest, to provide Bayes estimators that tend to outperform Bayes estimators with respect to complex priors, with results possibly extending to the domination of conjugacy over prior modeling that is only moderately more complex. In case it escaped the reader's notice, let me mention that this is a whole new type of "threshold problem!"

The idea of self-consistency has been introduced and utilized in this monograph, but the full extent of its utility is not yet known. On the theoretical side, it is of interest to characterize the class of priors that enjoy this property. A somewhat easier problem, but still of some interest, is the question of whether, when sampling distributions belong to exponential families and the loss criterion is squared error, the self-consistency property characterizes the standard conjugate prior distributions among prior distributions belonging to an exponential family. Further, one might ask whether self-consistency is a property that can be derived from the basic principles on which Bayes theory is based. And certainly more can be said about the practical implications of self-consistency. Perhaps ignoring the property is justifiable in certain applications. But a statistical consultant might ask the question: how would I explain the absence of self-consistency in an estimator to a client who asked for a Bayesian analysis?

The Bayesian treatment of nonidentifiable models is a somewhat controversial subject. The content of Chapter 9 might be viewed as a template for further studies regarding the potential payoffs and risks involved in using Bayes estimates of nonidentifiable parameters. In the particular problem we studied in some detail, it appears that the Bayesian approach can be said to be, more often than not, efficacious. While a Bayesian might be inclined to bask in this sunshine for a few moments, he should keep in mind that all that has been established is that Bayesian inference tends to be better than "just guessing" in this particular problem. This is far from a ringing endorsement. Still, when confronted with an estimation problem in which an answer is urgently needed, and the parameter of interest happens to be nonidentifiable, a full Bayesian treatment may well be the only available alternative to just guessing. In that sense, its efficacy will be of practical importance. What's left to do in this area? The most obvious "to do's" are the examination of the literature on Bayes estimates of nonidentifiable parameters and the consideration of the question of "efficacy" in each of these problems. Another area of investigation that would be of some practical importance is the examination of frequentist solutions of estimation problems in which the model employed is "made" identifiable through the addition of nonverifiable side conditions on the parameters. Econometric models often fit this description. A Bayesian treatment of such problems, with the side conditions set aside, may well be a more defensible and justifiable analysis. Efficacy questions will, of course, accompany these alternative analyses. Finally, the performance of Bayes estimators of high-dimensional nonidentifiable parameters is a problem that is largely uninvestigated and certainly merits further study.

In the one-parameter models treated in Chapter 5, we used the phrase "best frequentist estimator" with wild abandon. I made the claim that this was justified by the observation that when that sampling distribution belongs to a one-parameter exponential family, virtually all frequentists would utilize the same estimator, one that sat-

isfied all the usual frequentist goals, being simultaneously the UMVUE and the MLE of the target parameter, among other things. A frequentist would be hard pressed, for example, to recommend an alternative to the sample proportion $\widehat{p}$ as an estimator of a population proportion p. However, we have seen in subsequent developments that there may be more than one frequentist estimator worthy of consideration in particular problems of interest. The frequentist estimators on which we concentrate in Chapters 7 and 8 are notable examples. These estimators are widely used, but they are also known to be inadmissible. Thus, the frequentist, who may not have a "best" estimator to put forward, may nonetheless wish to consider some frequentist alternatives. For each such alternative, there is an associated threshold problem, and these problems have not been addressed in this monograph. This suggests that there is a collection of open problems that remains to be investigated. In each such problem, it is of interest to characterize the threshold that would separate good priors from bad priors in comparisons to the frequentist estimator of choice.

I should add, as a general remark, that most of the comparisons in this monograph involve Bayes estimators of a given known form and frequentist estimators that are unbiased estimators of the target parameter. The threshold problem will often be of interest in other settings. The treatment of threshold problems in which the Bayes estimator of the parameter of interest is unavailable in closed form, and in which there are several different frequentist estimators that merit consideration, will be especially challenging. Solutions to problems of this type might only be accessible through approximations involving numerical or iterative techniques. Such problems certainly occupy a prominent place within the agenda for future investigations.

Chapter 10 is dedicated to the theory and applications of empirical Bayes methods for using both "past" and "current" data in estimating the parameter of the current experiment. While the original framework for empirical Bayes analysis leads to frequentist procedures, later developments by Deely and Lindley (1981) and others introduced Bayesian counterparts which added the use of subjective inputs to Robbins' original formulation. In Chapter 10, we focus on the potential of empirical Bayes ideas for improving standard estimators (be they Bayesian or frequentist) of the current parameter. The chapter's main contributions are proofs of the existence of improved procedures. Left for future investigations is the determination of the various ways in which this improvement can be achieved in a broad class of problems of practical interest. When tractable, the characterization of the extent of improvement possible would also be of use. There is clearly a good deal left to be done in this area.

Chapter 11, on the treatment of related experiments, constitutes the most tentative discussion in the monograph. While a formulation of related experiments is proposed, I cannot claim that it is a unique or even best way of modeling such problems. Further, general results in this chapter are rather sparse, as its main focus is the treatment of the particular two-sample problem that motivated the chapter. Generalizations that would be of interest include the exploration of more general forms of the potential relationships among experiments than the simple scalar-multiple relationship I consider, the fully Bayesian treatment of the estimation of the current parameter relative to appropriate hierarchical operational priors on model parameters, the treatment of other versions of the threshold problem, including those in

which the true prior distribution is allowed to be nondegenerate and, finally, alternative formulations of the notion of related experiments which may be better suited to other statistical applications.

Even though I've already outlined a well-packed agenda for future research in the paragraphs above, I feel that I must formally recognize the large collection of complementary problems which the present treatment has completely ignored. I have focused exclusively on point estimation, giving other modes of inference (for example, interval estimation, hypothesis testing, prediction, sequential analysis) virtually no attention. I have done this largely because this has allowed me to present a fairly comprehensive examination of one important problem area, the existing theory in other problem areas being notably less developed. But this suggests that there are a myriad of problems and applications in which a properly adapted version of the threshold problem remains to be investigated. This is clearly a mission well beyond my own levels of energy and ambition. But consider this a call to arms! As Johnny Appleseed's guardian angel said to him in the first scene of the Disney cartoon, "there's a lot of work out there to do." So . . . go forth and multiply! Divide and conquer! Add your imprint to this work! You are cordially and enthusiastically invited to the party.

Appendix: Standard Univariate Probability Models

I. Discrete Models

Distribution	Notation	Probability Mass Function	Mean
Bernoulli	$\mathscr{B}(1,\theta)$	$p(x)=\theta^x(1-\theta)^{1-x}$, $x\in 0,1$; $\theta\in[0,1]$	θ
Binomial	$\mathscr{B}(n,\theta)$	$p(x)=\binom{n}{x}\theta^x(1-\theta)^{n-x}$, $x\in 0,1,\ldots,n$; $\theta\in[0,1]$	$n\theta$
Geometric	$G(\theta)$	$p(x)=\theta(1-\theta)^{x-1}$, $x=1,2,\ldots$; $\theta\in[0,1]$	$1/\theta$
Negative Binomial	$\mathscr{NB}(r,\theta)$	$p(x)=\binom{r+x-1}{x}\theta^r(1-\theta)^{x-1}$, $x=1,2,\ldots$; $\theta\in[0,1]$	r/θ
Poisson	$\mathscr{P}$	$p(x)=\theta^x e^{-\theta}/x!$, $x=0,1,2,\ldots$; $\theta>0$	θ

II. Continuous Models

Distribution	Notation	Density Function	Mean
Uniform	$\mathscr{U}(\theta_1, \theta_2)$	$f(x) = \dfrac{1}{(\theta_2 - \theta_1)}$, $-\infty < \theta_1 < x < \theta_2 < \infty$	$(\theta_1 + \theta_2)/2$
Beta	$\mathrm{Be}(\alpha, \beta)$	$f(x) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, $0 < x < 1$; $\min(\alpha, \beta) > 0$	$\alpha/(\alpha+\beta)$
Exponential	$\mathscr{E}(\theta)$	$f(x) = \frac{1}{\theta}\mathrm{e}^{-x/\theta}$, $x > 0$; $\theta > 0$	θ
Gamma	$\Gamma(\alpha, \beta)$	$f(x) = \dfrac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1}\mathrm{e}^{-x/\beta}$, $x > 0$; $\min(\alpha, \beta) > 0$	$\alpha\beta$
Normal	$\mathscr{N}(\mu, \sigma^2)$	$f(x) = \dfrac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-\mu)^2/2\sigma^2}$, $-\infty < x < \infty$; $\mu \in (-\infty, \infty)$, $\sigma^2 > 0$	μ
Pareto	$\mathrm{Par}(\alpha, \theta_0)$	$f(x) = \dfrac{\alpha\theta_0^{\alpha}}{x^{\alpha+1}}$, $\alpha < x < \infty$; $\min(\alpha, \theta_0) > 0$	$\alpha\theta_0/(\alpha-1)$

References

1. Andrews, D. F., Bickel. P. J., Hampel, F. R., Huber, P. J., Rogers, W.H. and Tukey, J. W.: *Robust Estimates of Location: Survey and Advances*, Princeton, NJ.: Princeton University Press (1972)
2. Antoniak, C. E.: Mixtures of Dirichlet processes with applications to nonparametric problems, *Annals of Statistics*, **2**, 1152–1174 (1974)
3. Arnold, B., Brockett, P. L., Torrez, W. and Wright, A. L.: On the inconsistency of Bayesian non-parametric estimators in competing risks/multiple decrement models, *Insurance: Mathematical Economics*, **3**, 49–55 (1984)
4. Barnard, G. A.: Statistical Inference, *Journal of the Royal Statistical Society, Series B*, **11**, 115–149 (1949)
5. Barnard, G. A., Jenkins, G. M and Winsten, C. B.: Likelihood Inference and Time Series, *Journal of the Royal Statistical Society, Series A*, 125:321–372 (1962)
6. Barnett, V. D.: *Comparative Statistical Inference*, 3rd Edition, Chichester, UK: Wiley and Sons (1999)
7. Bayes, T.: An essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society of London*, **53**, 370–418 (1763)
8. Bennett, G. K.: Basic concepts of empirical Bayes Methods with some results for the Weibull distribution, in *Theory and Applications of Reliability*, Volume 2 (Tsokos, C. and Shimi, I, editors), New York: Academic Press (1977)
9. Berger, J. O.: *Statistical Decision Theory and Bayesian Analysis*, 2nd Edition, New York: Springer-Verlag (1985)
10. Berger, J. O.: The case for objective Bayesian analysis (with Discussion). *Bayesian Analysis*, **1**, 385–402 (2006)
11. Berger, J. O., Bernardo, J. M. and Sun, D.: The formal definition of reference priors, *The Annals of Statistics*, **37**(2), 905–938 (2009)
12. Berger, J. O. and Wolpert, R.: *The Likelihood Principle*, 2nd Edition, Hayward, CA: IMS Monograph Series (1988)
13. Berman, S. M.: Note on extreme values, competing risks and semi-Markov processes, *Annals of Mathematical Statistics*, **34**, 1104–1106 (1963)
14. Bernardo, J.: Reference posterior distributions for Bayesian inference (with Discussion), *Journal of the Royal Statistical Society, Series B*, **41**, 113–147 (1979)
15. Berry, D. and Christensen, R.: Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes, *Annals of Statistics*, **7**, 558–568 (1979)

16. Besag, J., York, J. C. and Mollie, A.: Bayesian image restoration, with two applications in spatial statistics (with Discussion), *Annals of the Institute of Statistical Mathematics*, **43**, 1–59 (1991)
17. Bhattacharya, D., Samaniego, F. J. and Vestrup, E. M.: On the Comparative Performance of Bayesian and Classical Point Estimators under Asymmetric Loss, *Sankhya, Series B*, 239–266 (2002)
18. Bickel, P. J. and Doksum, K. A.: *Mathematical Statistics: Basic Ideas and Selected Topics, Vol.1*, 2nd Edition, Pearson Prentice Hall (2007)
19. Birnbaum, Z. W.: On the use of the Mann-Whitney statistic, in *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability*, **1**, Berkeley: University of California Press, 13–17 (1956)
20. Birnbaum, Z. W. and McCarty, R. C.: A distribution free upper confidence bound for $P(Y < X)$ based on independent samples of X and Y, *Annals of Mathematical Statistics*, **59**, 558–562 (1958)
21. Blackwell, D.: On the translation parameter problem for discrete variables, *Annals of Mathematical Statistics*, **22**, 393–399 (1951)
22. Blackwell, D. and Dubins, L.: Merging of opinions with increasing information, *Annals of Mathematical Statistics*, **33**, 882–886 (1962)
23. Blackwell, D. and Girshick, M. A.: *Theory of Games and Statistical Decisions*, New York: Wiley (1954)
24. Box, G. E. P. and Tiao, G.: *Bayesian Inference in Statistical Analysis*, London: Addison-Wesley (1973)
25. Boyles, R. A., Marshall, A. W. and Proschan, F.: Inconsistency of the maximum likelihood estimator of a distribution having increasing failure rate average, *Annals of Statistics*, **13**, 413–417 (1985)
26. Breslow, N. E.: Biostatistics and Bayes (with Discussion), *Statistical Science*, **5**, 269–298 (1990)
27. Brown, L. D.: Admissible estimators, recurrent diffusions and insoluble boundary value problems, *Annals of Mathematical Statistics*, 42, 855–903 (1971)
28. Brown, M. and Proschan, F.: Imperfect Repair, *Journal of Applied Probability*, **20**, 851–859 (1983)
29. Canovos, G. C.: An empirical Bayes approach for the Poisson life distribution, *IEEE Transactions on Reliability*, **TR-22**, 91–96 (1973)
30. Carlin, B. P. and Louis, T. A.: *Bayes and Empirical Bayes Methods for Data Analysis*, 3rd Edition, London: Chapman and Hall (2008)
31. Casella, G. and George, E.: Explaining the Gibbs sampler, *The American Statistician*, **46**, 167–174 (1992)
32. Chernoff, H. and Moses, L. E.: *Elementary Decision Theory*, New York: Wiley (1959)
33. Clayton, D. G. and Kaldor, J. M.: Empirical Bayes estimates of age-standardized relative risk for use in disease mapping, *Biometrics*, **43**, 671–681 (1987)
34. Cochran, W. G.: Problems arising in the analysis of a series of similar experiments, *Journal of the Royal Statistical Society* (Supplement), **4**, 102–118 (1937)
35. Cox, D. R.: *Principles of Statistical Inference*, Cambridge: Cambridge University Press (2006)
36. Deely, J. J. and Lindley, D. V.: Bayes empirical Bayes, *Journal of the American Statistical Association*, **76**, 833–841 (1981)
37. De Finneti, B.: *Theory of Probability*, New York: Wiley (1974)
38. DeGroot, M. H.: *Optimal Statistical Decisions*, New York: McGraw-Hill (1970)

39. DeGroot, M.: Modern aspects of utility and probability, in *Accelerated Life testing and Experts' Opinions in Reliability* (Clarotti, C. A. and Lindley, D. V., Editors), Amsterdam: North Holland, 3–24 (1988)
40. Devroye, L.: *Non-Uniform Random Variate Generation*, New York: Springer-Verlag (1986)
41. Diaconis, P. and Freedman, D.: On the consistency of Bayes estimates, *Annals of Statistics*, **14**, 1–25 (1986)
42. Diaconis, P., Khari, S. and Saloff-Coste, L.: Gibbs sampling, exponential families and orthogonal polynomials, *Statistical Science*, **23**, 151–191 (with Discussion) (2008).
43. Diaconis, P. and Ylvisaker, N. D.: Conjugate priors for exponential families, *Annals of Statistics*, **7**, 269–281 (1979)
44. Dreze, J.: Bayesian theory of identification in simultaneous equation models, in *Studies in Bayesian Econometrics and Statistics* (Feinberg, S. and Zellner, A., Editors), Amsterdam: North Holland, 159–174 (1975)
45. Efron, B.: The two-sample problem with censored data, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **4**, Berkeley: University of California Press, 831–853 (1967)
46. Efron, B.: Computers and the theory of statistics: thinking the unthinkable, *SIAM Review*, **21**, 460–480 (1979)
47. Efron, B.: Why isn't everyone a Bayesian? *The American Statistician*, **40**, 1–5 (1986)
48. Efron, B.: Bayesians, frequentists, and scientists, *Journal of the American Statistical Association*, **100**, 1–5 (2005)
49. Efron, B. and Morris, C. N.: Limiting the risk of Bayes and empirical Bayes estimators–Part I, the Bayes case, *Journal of the American Statistical Association*, **66**, 807–815 (1971)
50. Efron, B. and Morris, C. N.: Limiting the risk of Bayes and empirical Bayes estimators–Part II, the empirical Bayes case, *Journal of the American Statistical Association*, **67**, 130–139 (1972a)
51. Efron, B. and Morris, C. N.: Empirical Bayes on vector observations: an extension of Stein's method, *Biometrika*, **59**, 335–347 (1972b)
52. Efron, Bradley and Morris, Carl N.: Stein's estimation rule and its competitors—an empirical Bayes approach, *Journal of the American Statistical Association*, **68**, 117–130 (1973a)
53. Efron, B. and Morris, C. N.: Combining possibly related estimation problems (with Discussion), *Journal of the Royal Statistical Society, Series B*, **35**, 379–421 (1973b)
54. Efron, B. and Morris, C. N.: Data analysis using Stein's estimator and its generalizations, *Journal of the American Statistical Association*, **70**, 311–319 (1975)
55. Efron, B. and Morris, C. N.: Families of minimax estimators of the mean of a multivariate normal distribution, *Annals of Statistics*, **4**, 11–21 (1976)
56. Ericson, W. A.: Subjective Bayesian models in sampling finite populations (with Discussion), *Journal of the Royal Statistical Society, Series B*, **31**, 195–233 (1969)
57. Ericson, W. A.: On the posterior mean and variance of a population mean, *Journal of the American Statistical Association*, **65**, 649–652 (1970)
58. Ferguson, T. S.: *Mathematical Statistics, a Decision Theoretic Approach*, New York: Academic Press (1967)
59. Ferguson, T. S.: A Bayesian analysis of some nonparametric problems, *Annals of Statistics*, **1**, 209–230 (1973)
60. Ferguson, T. S.: *A Course in Large Sample Theory*, London: Chapman & Hall (1996)
61. Fishburn, P. C.: The axioms of Subjective Probability (with Discussion), *Statistical Science*, **1**, 335–358 (1986)

62. Fisher, R. A.: Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society*, **22**, 700–725 (1925)
63. Fisher, R. A.: *Statistical Methods for Research Workers*, 4th Edition, London: Oliver & Boyd (1932)
64. Gaver, D., Draper, D., Goel, P. K., Greenhouse, J., Hedges, L. V., Morris, C. N. and Waternaux, C.: *Combining Information: Statistical Issues and Opportunities for Research*, Washington, DC: National Academy Press (1992)
65. Gelfand, A. E. and Smith, A. F. M.: Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398–409 (1990)
66. Gelman, A., Carlin, J., Stern, H. and Rubin, D.: *Bayesian Data Analysis*, 2nd Edition, London: Chapman & Hall (2004)
67. Geman, S. and Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741 (1984)
68. Geweke, J.: Bayesian econometrics and forecasting, *Journal of Econometrics*, **100**, 11–15 (2001)
69. Geweke, J. and Terui, N.: Bayesian threshold autoregressive models for nonlinear time series, *Journal of Time Series Analysis*, **14**, 441–454 (1993)
70. Glass, G. V.: Integrating findings: the meta-analysis of research, in *Review of Research in Education* (Schulman, L. S., Editor), **5**, 351–379, Itasca, IL: F. E. Peacock (1978)
71. Good, I. J.: *Probability and Weighting of Evidence*, London: Griffin (1950)
72. Hartigan, J.: *Bayes Theory*, New York: Springer-Verlag (1969)
73. Hedges, L. and Olkin, I.: *Statistical Methods for Data Analysis*, New York: Academic Press (1985)
74. Hill, B. Inference about variance components in the one-way model, *Journal of the American Statistical Association*, **65**, 806–825 (1965)
75. Hobert, J. P. and Casella, G.: The effect of improper priors on Gibbs sampling in hierarchical linear mixed models, *Journal of the American Statistical Association*, **91**, 1461–1473 (1996)
76. Huber, P.: *Robust Statistics*, New York: Wiley (1981)
77. James, W. and Stein, C.: Estimation with quadratic loss, in *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability*, **1**, Berkeley: University of California Press, 361–379 (1961)
78. Jeffreys, H.: *Theory of Probability*, 3rd Edition, Oxford: Oxford University Press (1961)
79. Jiang, J.: *Linear and Generalized Linear Mixed Models and Their Applications*, New York: Springer (2007)
80. Johns, V.: Non-parametric empirical Bayes procedures, *Annals of Mathematical Statistics*, **28**, 649–669 (1957)
81. Johnson, R. A.: Stress-strength models for reliability, in *Handbook of Statistics, 7 (Quality Control and Reliability)*, Krishnaiah, P. R., Editor, New York: Elsevier, 27–54 (1988)
82. Johnson, W. O. and Gastwirth, J. L.: Bayesian inference for medical screening tests: Approximations useful for the analysis of AIDS, *Journal of the Royal Statistical Society, Series B*, **53**, 427–439 (1991)
83. Kadane, J. B., Dickey, J.M., Winkler, R.L., Smith, W. S. and Peters, S. C.: Interactive elicitation of opinion for a normal linear model, *Journal of the American Statistical Association*, **75**, 845–854 (1980)
84. Kaplan, E.L. and Meier, P.: Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481 (1958)

85. Kass, R. E. and Steffey, D. L.: Approximate Bayesian inference in conditionally independent hierarchical models, *Journal of the American Statistical Association*, **84**, 717–726 (1989)
86. Keifer, J., and Wolfowitz, J.: Consistency of the maximum likelihood estimates in the presence of infinitely many incidental parameters, *Annals of Mathematical Statistics*, **27**, 887–906 (1956)
87. Kotz, S., Lumelskii, Y. and Pensky, M.: *The Stress-Strength Model and its Generalizations*, Hackensack, NJ: World Scientific (2003)
88. Kraft, C.H., Pratt, J.W. and Seidenberg, A.: Intuitive Probability on Finite Sets, *Annals of Mathematical Statistics*, **30**, 408–419 (1959)
89. Le Cam, L.: An extension of Wald's theory of statistical decision functions, *Annals of Mathematical Statistics*, **26**, 69–81 (1955)
90. Lehmann, E.: *Theory of Point Estimation*, 2nd Edition, Pacific Grove, CA: Wadsworth & Brooks Cole (1983)
91. Lehmann, E. and Casella, G.: *Theory of Point Estimation*, 2nd Edition, New York: Springer-Verlag (1998)
92. Lindley, D. V.: *Bayesian Statistics: A Review*, Philadelphia: SIAM (1972)
93. Lindley, D. V.: *Making Decisions*, 2nd Edition, New York: Wiley (1985)
94. Lindley, D. V. and El-Sayyad, G. M.: The Bayesian estimation of a linear functional relationship, *Journal of the Royal Statistical Society, Series B*, **30**, 190–202 (1968)
95. Lindley, D. V. and Novick, M. R.: The role of exchangeability in inference, *Annals of Statistics*, **9**, 45–58 (1981)
96. Lindley, D. V. and Smith, A. F. M.: Bayes estimates for the linear model (with Discussion), *Journal of the Royal Statistical Society, Series B*, **34**, 1–41 (1973)
97. Manski, C. F.: *Analog Estimation Methods in Econometrics*, London: Chapman & Hall (1988)
98. Maritz, J. S. and Lwin, T.: *Empirical Bayes Methods*, London: Chapman & Hall (1989)
99. Marshall, A. and Olkin, I.: A multivariate exponential distribution, *Journal of the American Statistical Association*, **62**, 30–44 (1967)
100. Maugham, W. S.: *Of Human Bondage*, New York: Doubleday (1915) (Reprinted in 1961 by New York: Vintage Books, Random House)
101. McCullagh, P. and Nelder, J.: *Generalized Linear Models*, 2nd Edition, London: Chapman & Hall (1989)
102. Neath, A. A. and Samaniego, F. J.: On Bayesian Estimation of the Multiple Decrement Function in the Competing Risks Problem, *Statistics and Probability Letters*, **31**, 75–83 (1996a)
103. Neath, A. A. and Samaniego, F. J.: On Bayesian Estimation in the Competing Risks Problem: An Example of the Distinguished Role Played by the Multivariate Exponential Distribution, *Statistics and Probability Letters*, **31**, 69–74 (1996b)
104. Neath, A. A. and Samaniego, F. J.: On the Efficacy of Bayesian Inference for Nonidentifiable Models, *The American Statistician*, **51**, 325–332 (1997)
105. Neyman, J.: Contribution to the theory of sampling human populations, *Journal of the American Statistical Association*, **33**, 101–116 (1938)
106. Nusbacher, J., Chiavetta, J., Naiman, R., Buchner, B., Scalia, V. and Horst, R.: Evaluation of a confidential method of excluding blood donors exposed to human immunodeficiency virus, *Transfusion*, **26**, 539–541 (1986)
107. Neyman, J. and Scott, E.: Consistent estimates based on partially consistent observations, *Econometrika*, **16**, 1–32 (1948)

108. O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J. , Oakley, J. E. and Rakow, T.: *Uncertain Judgements: Eliciting Experts' Probabilities*, New York: Wiley (2006)
109. Phadia, E. and Susarla, V.: Nonparametric Bayesian estimation of a survival curve with dependent censoring mechanism, *Annals of the Institute of Statistical Mathematics*, **35A**, 389–400 (1983)
110. Ramsey, F.: Truth and Probability, reprinted in *Studies in Subjective Probability* (Kyburg, H. and Smaklar, H., Editors, New York: Wiley) (1930)
111. Rao, C. R.: *Linear Statistical Inference and Its Applications*, New York: Wiley (1973)
112. Robbins, H.: Asymptotically subminimax solutions of compound statistical decision problems, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 131–149 (1951)
113. Robbins, H.: An empirical Bayes approach to statistics, in *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability*, **1**, Berkeley: University of California Press, 157–164 (1956)
114. Robbins, H.: The empirical Bayes approach to statistical decision problems, *Annals of Mathematical Statistics*, **20**, 1–20 (1964)
115. Robert, C.: *The Bayesian Choice: A Decision Theoretic Motivation*, 2nd Edition, London: Chapman and Hall (2001)
116. Robert, C. and Casella, G.: *Monte Carlo Statistical Methods*, 2nd Edition, New York: Springer-Verlag (2004)
117. Rojo, J. and Samaniego, F. J.: On Nonparametric Maximum Likelihood Estimation of a Distribution Uniformly Stochastically Smaller than a Standard, *Statistics and Probability Letters*, **11**, 267–271 (1991)
118. Samaniego, F. J.: On T-minimax Estimation, *The American Statistician*, **29**, 168–169 (1975)
119. Samaniego, F. J.: Estimation based on autopsy data from stress-strength experiments, *Journal of Quality Technology and Quality Management*, Special Issue on Reliability, **4**, 1–15 (2007)
120. Samaniego, F. J. and Jones, L. E.: Maximum Likelihood Estimation for a Class of Multinomial Distributions Arising in Reliability, *Journal of the Royal Statistical Society, Series B*, **43**, 45–52 (1981)
121. Samaniego, F. J. and Kaiser, L.: Estimating Value in a Uniform Auction, *Naval Research Logistics Quarterly*, **25**, 621–632 (1978)
122. Samaniego, F. J. and Neath, A. A.: How to be a better Bayesian, *Journal of the American Statistical Association*, **91**, 733–742 (1996)
123. Samaniego, F. J. and Reneau, D. M.: Toward a reconciliation of the Bayesian and frequentist approaches to point estimation, *Journal of the American Statistical Association*, **89**, 947–957 (1994)
124. Samaniego, F. J. and Vestrup, E. M.: On improving standard estimators via linear empirical Bayes methods, *Statistics and Probability Letters*, **44**, 309–318 (1998)
125. Savage, L. J.: *The Foundations of Statistics*, New York, Wiley (1954)
126. Savage, L. J.: *The Foundations of Statistical Inference–A Discussion*, London: Methuen (1962)
127. Shao, J. and Chow, S. C.: Constructing release targets for drug products: a Bayesian decision theory approach, *Applied Statistics*, **40**, 381–390 (1991)
128. Steffey, D., Samaniego, F. J. and Tran, H.: Hierarchical Bayesian inference in related reliability experiments, in *Recent Advances in Reliability* (Limnios, N. and Nikulin, M., Editors), Boston: Birkauser, 379–390 (2000)

129. Stein, C.: Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, in *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability*, **1**, Berkeley: University of California Press, 197–206 (1956)
130. Tanner, M. A. and Wong, W. H.: The calculation of posterior distributions by data augmentation (with Discussion), *Journal of the American Statistical Association*, **82**, 528–550 (1987)
131. Thompson, R. D. and Basu, A.: Asymmetric loss functions for estimating system reliability, in *Bayesian Analysis in Statistics and Econometrics* (Berry, D., Chaloner, K. and Geweke, J., Editors), New York: Wiley (1996)
132. Tiao, G. and Tan, W.: Bayesian analysis of random-effect models in the analysis of variance. I. Posterior distribution of variance components, *Biometrika*, **52**, 53–57 (1965)
133. Tierney, L. and Kadane, J. B.: Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association*, **81**, 82–86 (1986)
134. Tierney, L., Kass, R. and Kadane, J. B.: Fully exponential Laplace approximations to expectations and variances of nonpositive functions, *Journal of the American Statistical Association*, **84**, 710–716 (1989)
135. Tsai, W. Y. and Crowley, J.: A Large Sample Study of Generalized Maximum Likelihood Estimators from Incomplete Data Via Self-Consistency, *Annals of Statistics*, **13**, 1317–1334 (1985)
136. Tsiatis, A.: A nonidentifiability aspect of competing risks, *Proceedings of the National Academy of Sciences*, **72**, 20–22 (1975)
137. Varian, H. R.: A Bayesian Approach to Real Estate Assessment, in *Studies in Bayesian Econometrics and Statistics in Honor of L.J. Savage* (Fienberg, S.E. and Zellner, A., Editors), Amsterdam: North-Holland, 195–208 (1975)
138. Vestrup, E. M.: *Bayesian vs Frequentist Estimation of Multivariate Parameters*, Unpublished doctoral dissertation, Department of Statistics, University of California, Davis (2001)
139. Vestrup, E. M. and Samaniego, F. J.: Bayes versus frequentist shrinkage in multivariate normal problems, *Sankhya*, **66**, 109–139 (2004a)
140. Vestrup, E. M. and Samaniego, F. J.: Auxiliary results for "Bayes versus frequentist shrinkage in multivariate normal problems," Technical Report #339, Department of Statistics, University of California, Davis (2004b)
141. von Neumann, J. and Morganstern, O.: *Theory of Games and Economic Behavior*, 3rd Edition, Princeton, NJ: Princeton University Press (1944)
142. Wald, A.: *Statistical Decision Functions*, New York: Wiley (1950)
143. Whitaker, L. R. and Samaniego, F. J.: On Estimating the Reliability of Systems Subject to Imperfect Repair, *Journal of the American Statistical Association*, **84**, 301–309 (1989)
144. Williams, J. D.: *The Compleat Strategyst*, New York: McGraw-Hill (1954)
145. Yee, J., Johnson, W. O. and Samaniego, F. J.: Asymptotic Approximations to Posterior Distributions via Latent-Data Conditional Moment Equations, *Biometrika*, **89**, 755–767 (2002)
146. Zellner, A.: Bayesian estimation and prediction using asymmetric loss functions, *Journal of the American Statistical Association*, **81**, 446–451 (1986)
147. Zellner, A. and Palm, F.: Time Series Analysis and Simultaneous Equation Models, *Journal of Econometrics*, 17–54 (1974)

Index